Modelo de análisis multivariante aplicado a la predicción de la tendencia del tipo de cambio euro-dólar

VNiVERSiDAD D SALAMANCA

Humberto Mauricio Argotty Erazo

Departamento de Estadística
Universidad de Salamanca

Tesis presentada como requisito parcial para optar al grado de:
Doctor en Estadística Multivariante Aplicada.

Salamanca - España Julio 2023

Modelo de análisis multivariante aplicado a la predicción de la tendencia del tipo de cambio euro-dólar

VNiVERSiDAD D SALAMANCA

Humberto Mauricio Argotty Erazo

Director:
Ph.D. Antonio Blázquez Zaballos

Departamento de Estadística
Universidad de Salamanca

Tesis presentada como requisito parcial para optar al grado de:
Doctor en Estadística Multivariante Aplicada.

Salamanca - España Julio 2023

Universidad de Salamanca
Departamento de Estadística

DR. ANTONIO BLÁZQUEZ ZABALLOS

Profesor Contratado Doctor del Departamento de Estadística de la Universidad de Salamanca

---

CERTIFICA que **D. Humberto Mauricio Argotty Erazo**, Ingeniero Industrial, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al título de Doctor presenta con el título "**Modelo de análisis multivariante aplicado a la predicción de la tendencia del tipo de cambio euro-dólar"**, autorizando expresamente su lectura y defensa.

Y para que conste, firman el presente certificado en Salamanca a 20 de julio de 2023.

BLAZQUEZ ZABALLOS ANTONIO - 07879049P

Firmado digitalmente por BLAZQUEZ ZABALLOS ANTONIO - 07879049P
Fecha: 2023.07.20 14:11:59 +02'00'

Fdo. Antonio Blázquez Zaballos

VNiVERSiDAD
D SALAMANCA
Departamento de Estadística

Modelo de análisis multivariante aplicado a la predicción de la tendencia del tipo de cambio euro-dólar.

Memoria que para optar al grado de: **Doctor en Estadística Multivariante Aplicada**, por el Departamento de Estadística de la Universidad de Salamanca, presenta: **Humberto Mauricio Argotty Erazo.**

Fdo. Humberto Mauricio Argotty Erazo.

Salamanca - España
Julio 2023.

A mis padres, de quienes
también son estos frutos.
A mi campeón y mi nena.

En los dominios del análisis multivariante,
los modelos despliegan su elegancia interpretativa,
precisión y sofisticada simplicidad cuando guían
las decisiones basadas en información oculta en los datos.
Mauricio Argotty Erazo.

Agradecimientos

Gracias, Dios Padre, por darme los dones más preciados: la vida, la salud, el valor, el amor y la paz en el corazón. Gracias por iluminar mi entendimiento, por infundir en mi la disciplina y la humildad que han hecho posible el cumplimiento de este gran propósito.

Quiero agradecer a mi hogar y mi familia, mis mayores tesoros, porque sin su comprensión y su valiosa ayuda esto no hubiese sido posible.

Quiero dar mis más sinceros agradecimientos al programa de: **"Becas internacionales Universidad de Salamanca-Banco Santander de España para la movilidad en estudios de doctorado destinados a estudiantes latinoamericanos"**. Gracias por brindarme los recursos y los medios necesarios para cumplir con este gran sueño.

Quiero agradecer a mi director de tesis: Ph.D. Antonio Blázquez Zaballos, quien con sus conocimientos y enseñanzas me condujo hacia el final de este camino.

También quiero dar un especial agradecimiento al Grupo de investigación: **SDAS Research Group**, cuya valiosa ayuda ha sido fundamental para alcanzar la meta.

Quiero dar las gracias a mis grandes amigos Amanda y Félix Cuadra, porque con su ejemplo han demostrado que el verdadero valor de las personas se manifiesta a través de las buenas acciones que nacen del corazón.

Finalmente, quiero dar las gracias a todas las instituciones y personas, que de alguna manera, contribuyeron al buen desarrollo de este importante trabajo.

A todos y todas, mi más profunda gratitud por su apoyo, comprensión y aliento a lo largo de este camino. El apoyo de todos y cada uno de ustedes ha sido fundamental para hacer posible este logro y ha dejado una huella indeleble en mi formación académica y personal. Gracias por formar parte de este sueño hecho realidad.

Información de la memoria del trabajo de grado

Título:

Modelo de análisis multivariante aplicado a la predicción de la tendencia del tipo de cambio euro-dólar

Autor:

Humberto Mauricio Argotty Erazo.

Director:

Ph.D. Antonio Blázquez Zaballos.
Departamento de Estadística.
Universidad de Salamanca. España.

Miembros del jurado:

Índice general

Lista de figuras

Lista de tablas

Resumen

Predecir los precios y las tendencias de los instrumentos financieros para mejorar la efectividad de las decisiones de inversión es un reto clave para la industria financiera y los agentes implicados. Aunque se han utilizado muchas técnicas eficaces de inteligencia artificial en el análisis de las series de tiempo, el problema de predecir la dirección del movimiento de los tipos de cambio en el mercado Forex aún requiere soluciones parsimoniosas, descifrables y precisas. Desde la perspectiva del análisis técnico, esta investigación presenta una metodología novedosa para clasificar la dirección de la tendencia de los tipos de cambio. La metodología utiliza puntos de inflexión y datos de mercado que miden la acción de los precios, junto con las diferencias multidimensionales entre tendencias, para construir una función lineal discriminante (LDA).

La metodología propuesta consta de cinco fases: preparación de datos, selección de características, detección de estructuras subyacentes, formulación de una función discriminante lineal y evaluación del desempeño del modelo con datos dentro y fuera de la muestra. Los experimentos se realizaron con datos de mercado del tipo de cambio euro-dólar en marcos de tiempo de 15 minutos y 1 semana, y una colección de puntos de inflexión del mercado (ET) definidos por un modelo de negociación algorítmico. El período de muestra va desde enero de 1999 hasta abril de 2023.

En contraste con algunos trabajos notables publicados en la literatura científica como la Memoria a Corto Plazo (LSTM), el Aprendizaje Profundo por Refuerzo (DRL), el Análisis Wavelet (WA), el Análisis de Sentimiento de Contenido Textual, las Máquinas de Vectores de Soporte (SVM) y los Algoritmos Genéticos (GA), la metodología propuesta logró una precisión de clasificación del 98.77% con datos fuera de muestra. Estos resultados respaldan la construcción de métodos de clasificación interpretables, generalizables, precisos y parsimoniosos, lo que sugiere mejoras significativas en el rendimiento financiero y la reducción del riesgo en las estrategias de negociación. Además, esta metodología es aplicable en la selección de variables y se adapta fácilmente a otros activos financieros.

Palabras clave

Análisis Discriminante Lineal (LDA), Sistemas de negociación algoritmicos, Aprendizaje de Máquina, Aprendizaje Automático Supervisado, Pronóstico de Series de Tiempo, Mercado de Intercambio de Divisas (Forex).

Abstract

Predicting the prices and trends of financial instruments to enhance the effectiveness of investment decisions is a key challenge for the financial industry and stakeholders involved. While many effective artificial intelligence techniques have been utilized in time series analysis, the problem of predicting the direction of exchange rate movements in the Forex market still requires parsimonious, interpretable, and accurate solutions. From the perspective of technical analysis, this research introduces an innovative methodology for classifying the direction of exchange rate trends. The methodology leverages inflection points and market data measuring price action, along with multidimensional differences between trends, to construct a linear discriminant function (LDA).

The proposed methodology consists of five phases: data preparation, feature selection, detection of underlying structures, formulation of a linear discriminant function, and evaluation of model performance with in-sample and out-of-sample data. Experiments were conducted using market data for the euro-dollar exchange rate at 15-minute and 1-week timeframes, and a collection of market inflection points (ET) defined by an algorithmic trading model. The sample period spans from January 1999 to April 2023.

In contrast to notable works published in the scientific literature, such as Long Short-Term Memory (LSTM), Deep Reinforcement Learning (DRL), Wavelet Analysis (WA), Sentiment Analysis of Textual Content, Support Vector Machines (SVM), and Genetic Algorithms (GA), the proposed methodology achieved a classification accuracy of 98.77% with out-of-sample data. These results support the development of interpretable, generalizable, precise, and parsimonious classification methods, suggesting significant improvements in financial performance and risk reduction in trading strategies. Additionally, this methodology is applicable in variable selection and easily adaptable to other financial assets.

Keywords

Linear Discriminant Analysis (LDA), Algorithmic Trading Systems, Machine Learning, Supervised Machine Learning, Time Series Forecasting, Foreign Exchange Market (Forex).

Notación

Variables y funciones

$\boldsymbol{X}$	Matriz de datos de i *observaciones* $\times$ j *variables*
$\boldsymbol{X}^\top$	Matriz transpuesta de $\boldsymbol{X}$
$\boldsymbol{x}_i$	i-ésima observación de la matriz de datos $\boldsymbol{X}$
$\bar{X}_{ij}$	Media de la variable i en un grupo j
X_{ijk}	Valor de la variable i para el caso k en el grupo j
S_{ij}^2	Varianza de la variable i para cada grupo j
S_i^2	Varianza de la variable i
$\tilde{S}$	Rango de la matriz $\boldsymbol{X}$
$(\boldsymbol{C_b/C_c})$	Tipo de cambio de la moneda base C_b y la moneda cotizada C_c
$\mathcal{I}C_i$	Tipo de interés de la moneda i
ΔER_t	Variación del tipo de cambio $(\boldsymbol{C_b/C_c})$ entre los instantes t y $t-1$
$\boldsymbol{S_i}$, $\boldsymbol{R_i}$	Niveles de Soporte y de Resistencia i
`ET`	Punto de inflexión y de entrada en el mercado
$\bar{U}_i$	Media de las variaciones al alza para el periodo i
$\bar{D}_i$	Media de las variaciones a la baja para el periodo i
$\overline{\Delta cp}$	Media de las variaciones entre los precios de cierre cp
XT	Salida del mercado
SL	Limite tolerable de perdidas
Z	Función objetivo
T	Nivel de tolerancia especificado
ψ_i	Valor de tolerancia obtenido para un predictor potencial i
$\mathbf{F}_i$	Valor F de entrada/salida para la variable i
F_ε	Nivel de entrada especificado en la selección de variables
F_r	Nivel de remoción especificado en la selección de variables
Λ	Lambda de Wilks
λ	Valor propio
F_a	Prueba F aproximada para lambda (R de Rao)
$M^2_{\mathtt{a\ b}}$	Distancia de Mahalanobis al cuadrado
a_i	Marcadores fila para representación de observaciones i
b_j	Marcadores columna para representación de variables j
r_{ij}	Coeficiente de correlación de Pearson entre las variables x_i y x_j
$d^2(x_i, x_j)$	Distancia Euclídea entre las variables x_i y x_j
$\mathfrak{Vt}_X$	Suma de cuadrados de los elementos de la matriz $\boldsymbol{X}$
$CR.E_jF_l$	Contribución relativa del elemento columna j al factor l
$CR.F_lE_j$	Contribución relativa del factor l al elemento columna j
ϑ_i	Retorno entre precios de cierre cp
γ	Coeficiente de asimetría
κ	Coeficiente de curtosis

Variables y funciones

Operadores matemáticos

$\|\cdot\|$	Norma euclidiana
$\mathrm{d}(\cdot,\cdot)$	Medida de distancia o disimilitud
$\lvert\cdot\rvert$	Valor absoluto
$\|\cdot\|_F$	Norma de Frobenius
$\mu(\cdot)$	Media aritmética
$\langle\cdot,\cdot\rangle$	Producto escalar
$\mathrm{tr}(\cdot)$	Traza de una matriz
$\mathrm{diag}(\cdot)$	Diagonal de la matriz de su argumento
$\mathrm{Diag}(\cdot)$	Matriz diagonal formada por el vector de su argumento

Abreviaturas

AD:	Estadístico Anderson-Darling
AHPR:	Media aritmética de las variaciones de capital por operación cerrada
AI:	Inteligencia artificial
APT:	Modelo de valoración de activos (Teoría de precios de arbitraje)
BCE:	Banco Central Europeo
CAPM:	Modelo de valoración de activos de capital o Modelo de fijación de precios de los activos de capital
CNN:	Red neuronal artificial
Cos:	Coseno del ángulo entre dos marcadores columna b_i y b_j
DL:	Aprendizaje profundo
EEA:	Espacio Económico Europeo (Por sus siglas en inglés)
EMH:	Hipótesis de los mercados eficientes
EMU:	Unión Monetaria y económica Europea (Austria, Belgium, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Luxembourg, Malta, Netherlands, Portugal, Slovak Republic, Slovenia, Spain)
ESG:	Medio ambiente, Sociedad y Gobernanza
ETF:	Fondos de inversión (Fondos negociados en bolsa)
EU:	Union Europea (Formalmente Comunidad Europea); Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovak Republic, Slovenia, Spain, Sweden
EUR:	Euro
EUR/USD:	Cruce euro-dólar americano
FOREX:	Mercado de Divisas
FX:	Divisas
GBM:	Gradient Boosting Machine (Algoritmo de refuerzo)
GHPR:	Media geométrica de las variaciones de capital por operación cerrada
GSE:	Bolsa de valores de Ghana
HDA:	Análisis Discriminante Heterocedastico
HDDA:	Análisis Discriminante de Alta-Dimension
KMO:	Kaiser-Meyer-Olkin medida de adecuación de muestreo
LCC:	Configurador Clasificador Lineal
LDA:	Análisis Discriminante Lineal
LR:	Correlación entre la línea de balance de la cuenta y la línea de regresión
LSTM:	Memoria a corto plazo
MCC:	Coeficiente de correlación de Matthews
ML:	Aprendizaje automático
MND:	Distribución Normal Multivariante
OHLCV:	Datos de mercado, precios de apertura, máximo, mínimo, cierre y volumen
OTC:	Over The Counter (Operaciones realizadas sin intermediarios)
PCA:	Análisis de Componentes Principales
PIB:	Producto Interno Bruto
QDA:	Análisis Discriminante cuadrático
RBF:	Función Kernel de base radial gaussiana
RDA:	Análisis Discriminante Regularizado
ROI:	Retorno sobre la inversión
RSI:	Índice de Fuerza Relativa
S-LDA:	Análisis Discriminante Lineal Sparse
SML:	Supervised Machine Learning
SVD:	Descomposición de Valores Singulares
UML:	Unsupervised Machine Learning (Aprendizaje automático no supervisado)

Abreviaturas

USD:	Dólar americano
VaR:	Valor en Riesgo
WT:	Transformada de wavelet

Parte I

Preliminares

1. Introducción

El mercado de divisas Forex (*FX*), comúnmente conocido como OTC (*Over The Counter*), es el mercado global y descentralizado más líquido y con el mayor volumen de negociación de los mercados financieros del mundo. Es uno de los mercados más complejos y volátiles en el que participan grandes y pequeños inversores, donde se determina el valor de las monedas y favorece el intercambio de divisas. El par de divisas euro-dólar americano (*EUR/USD*) ha demostrado ser el más atractivo a lo largo de los años para los inversores y es uno de los tipos de cambio más negociados debido a la fortaleza de las divisas y la posibilidad de beneficiarse de la volatilidad de los precios. En los últimos años, el fácil acceso al comercio de divisas, el rápido crecimiento del mercado y la incorporación de nuevos desafíos tecnológicos en los modelos de negociación han impulsado en profesionales, inversores e investigadores la necesidad de entender el comportamiento del mercado y el interés por predecir con éxito y precisión la tendencia del movimiento futuro de los tipos de cambio. Siendo previsibles desde este enfoque dos tipos de formaciones, movimiento de precios con tendencia al alza (*Bullish trend*) y movimiento de precios con tendencia a la baja (*Bearish trend*). Últimamente debido a la volatilidad de los tipos de cambio, la influencia de las fuerzas del mercado en la formación de los precios y la interacción de variables económicas, psicológicas, políticas y no fundamentales [1, 2, 3, 4, 5], las tendencias de los tipos de cambio y los puntos de inflexión del mercado deben ser identificados, supervisados y gestionados continuamente por los agentes que intervienen en las negociaciones [6, 7], para mejorar la toma de decisiones financieras. De forma que mejore el desempeño de las negociaciones, se reduzca la exposición al riesgo y se facilite el desarrollo de operaciones de cobertura, protegiendo al mismo tiempo el valor de las inversiones.

Por otra parte, las resiliencias financiera y operativa producidas por la crisis financiera de 2008 y la pandemia del *Covid-19* reafirman la necesidad de vigilar el comportamiento dinámico del mercado para identificar, anticipar y mitigar posibles movimientos adversos. Actualmente, la previsión y gestión del tipo de cambio es un tema de interés para gobiernos, autoridades monetarias, bancos centrales, empresas privadas y organismos internacionales [8]. Las autoridades monetarias, supervisores y

agentes de cálculo necesitan monitorizar y anticipar con precisión los cambios en los tipos de cambio para desarrollar pautas que se utilicen para establecer, revisar y divulgar los tipos de cambio de referencia con las partes interesadas [9]. De modo que estos valores reflejen la realidad económica del mercado, sean una representación confiable de la moneda y ayuden a tomar decisiones informadas [10]. En este contexto, la previsión del movimiento futuro del tipo de cambio puede brindar información útil para entender el panorama del mercado, reducir la incertidumbre sobre su comportamiento futuro y prever movimientos acusados, repentinos y adversos [11, 12]. Además, se puede utilizar como herramienta de planificación para diseñar y mejorar estrategias de negociación, gestión de carteras de inversión, asignación de activos y gestión de fondos de cobertura [13, 14].

Usualmente, la previsión del valor de los activos se ha enmarcado en dos vertientes principales: el análisis técnico y el análisis fundamental [15]. El primer enfoque utiliza precios de mercado históricos para identificar patrones ocultos o señales de mercado para predecir los precios de los activos o la dirección futura del movimiento. Estos estudios se distinguen por el hecho de que no utilizan información económica para estimar el valor futuro del activo [16]. En cambio, el análisis fundamental determina el valor intrínseco de un activo basado en factores económicos o variables fundamentales [17]. Los primeros estudios de análisis técnico predicen los precios de los activos basándose en el reconocimiento visual de patrones y formas detectados en series históricas de precios [18]. Posteriormente, el precio del activo se predice a través de datos cuantitativos obtenidos de indicadores técnicos que miden el movimiento histórico del precio. Por otro lado, en el campo del análisis fundamental, trabajos notables presentados en [19, 20, 21, 22, 23, 24] apoyan *La Hipótesis del Mercado Eficiente* [25]. Aunque estos estudios indican que los precios de los activos financieros reflejan toda la información de mercado disponible y, por lo tanto, no pueden predecir el comportamiento futuro con base en datos históricos, cabe señalar que investigaciones posteriores [26, 27, 28] muestran que los precios y los rendimientos de los instrumentos financieros se pueden predecir a partir de los rendimientos históricos. En consecuencia, desde el enfoque del análisis técnico y fundamental, la recopilación y preparación de grandes conjuntos de datos (a partir de información histórica basada en datos de mercado y datos fundamentales) para predecir la dirección de los precios de los activos financieros sigue siendo un problema desafiante y controvertido que permanece abierto a la investigación.

Por otro lado, en [29] se informa sobre el tipo de variables predictoras que se utilizan actualmente en la construcción de modelos predictivos. En general, las métricas

más utilizadas se enmarcan en seis dimensiones: valor del activo, tiempo, volumen de negociación, ratios (combinación de dos o más variables), indicadores técnicos (de tendencia y osciladores) e información heterogénea (noticias, anuncios, sentimiento de los inversores). Estas métricas generalmente se construyen sobre la base de datos de mercado (precios y valores derivados de sus cálculos) para diferentes marcos de tiempo. La configuración y los parámetros de cálculo se ajustan para cada estudio en función del desempeño obtenido con los modelos de predicción. Otras categorías especiales corresponden a indicadores calculados para fines específicos según necesidades y enfoques utilizados en cada estudio. Esto significa que la definición de variables, la configuración de los parámetros de cálculo, la selección de las mejores variables predictoras es también una tarea desafiante y de gran interés para la formulación de modelos predictivos [30].

En el campo de la inteligencia artificial (AI), en los últimos años ha aumentado el interés por predecir el precio de los instrumentos financieros [31]. Una revisión de la literatura científica muestra que este tema desafiante se ha explorado utilizando métodos estadísticos, algoritmos de aprendizaje automático (ML), métodos híbridos [32] y técnicas de aprendizaje profundo (*Deep Learning* DL) [33]. Si bien, los autores en [34] destacan una alta preferencia por el uso de DL, nuevos enfoques basados en el análisis de sentimiento (de noticias económicas y financieras) y la minería de textos (en publicaciones de redes sociales) se consolidan para predecir la dirección del precio [35, 36, 37]. Según estudios realizados en [38, 39], las DL son las más utilizadas en aplicaciones financieras y económicas debido al rendimiento alcanzado en el proceso de clasificación y predicción. A su vez, los algoritmos de ML también han demostrado su eficacia en tareas de clasificación. Los resultados obtenidos en [40, 41] superan el rendimiento obtenido por algunas aplicaciones financieras donde se utilizan determinados modelos de referencia (*GARCH models*).

En cuanto al Análisis Discriminante Lineal (LDA), es una técnica estadística multivariante y paramétrica ampliamente utilizada desde que fue publicada por Fisher [42, 43, 44]. Una revisión de la literatura en [45] muestra que el LDA se utiliza con éxito en varias aplicaciones económicas y financieras para tareas de clasificación y predicción. Aunque se han propuesto y reportado muchos métodos en la literatura científica, existen al menos cinco enfoques notables, a saber: Heterocedastico (HDA) [46], Cuadrático (QDA) [47], Regularizado (RDA) [48], Análisis Discriminante de Alta-Dimensión ($HDDA$) [49], y Análisis Discriminante Lineal Sparse (S-LDA) [50]. La principal ventaja del LDA, frente a otros métodos de clasificación, además de identificar variables candidatas, si pueden discriminar y producir diferencias significativas

entre grupos, es la capacidad de incorporar en la formulación del modelo múltiples variables cuantitativas (reducción de la dimensionalidad) o utilizar un subconjunto de variables predictoras (previamente seleccionadas mediante el uso de otras técnicas). En cualquiera de estos casos la función lineal discriminante maximiza las diferencias entre clases y en un subespacio de dimensión reducida lleva a cabo el proceso de clasificación.

La eficacia del LDA se ha demostrado en tareas de clasificación. Trabajos especializados que predicen el valor en riesgo (Value-at-Risk VaR) y los precios de algunos activos financieros se caracterizan por el uso de datos y métricas que mejoran el desempeño en la clasificación. Algunas de las colecciones de datos más notables son: titulares de noticias [51, 52], datos basados en la información del inversor [53], ratios financieros [54], datos fundamentales derivados de noticias financieras (twits) [55] y estadísticas descriptivas de variables predictoras [56], entre otros. En consecuencia, el uso de datos valiosos, con alto poder discriminante, mejora la precisión y el poder predictivo de los modelos de clasificación. Los resultados informados en [55, 57, 58, 59], superan el desempeño obtenido por varias técnicas de ML.

En general, la predicción de la dirección de los precios de los instrumentos financieros ha demostrado ser una valiosa herramienta de planificación para que los participantes del mercado tomen decisiones informadas sobre qué y cuándo negociar (comprar y/o vender). Aunque se han introducido muchas metodologías para predecir la dirección del movimiento de los precios, la formulación y evaluación del desempeño de los modelos de previsión para la toma de decisiones informadas no es una tarea fácil, principalmente debido a la complejidad, la fiabilidad y el nivel de riesgo que implica para las instituciones, los inversores y los agentes interesados.

La intervención de este problema requiere un profundo conocimiento del mercado, la preparación de los datos para el análisis, la definición de las variables predictoras, la elección de las técnicas para la selección de características y la formulación de los modelos de clasificación y predicción. Además de la definición de los indicadores que se utilizan en los procesos de verificación y validación del desempeño de los modelos con datos dentro y fuera de la muestra de estudio.

Uno de los mayores inconvenientes presentes en los modelos de predicción reportados en la literatura científica especializada es el *problema de la caja negra*. Es decir, la incapacidad de los modelos para explicar la forma en que intervienen las variables predictoras en el proceso de clasificación, lo que los hace indescifrables e ilegibles. Esta dificultad inherente a las técnicas de AI ha hecho que el foco de atención en este tipo de estudios se dirija a la precisión alcanzada en la clasificación y por tanto se deje de lado

la capacidad explicativa de las variables y su influencia en la predicción. Esta condición hace que este problema esté abierto a la investigación y por lo tanto, existe una gran necesidad de desarrollar modelos de predicción sencillos, parsimoniosos y altamente precisos que puedan explicar cómo las variables predictoras afectan la acción del precio para poder predecirlo [60].

Además, el problema de predecir la dirección del movimiento de los tipos de cambio, según las publicaciones reportadas en la literatura científica, ignora la capacidad discriminante inherente a los atributos que definen la naturaleza multidimensional de las diferencias entre movimientos tendenciales al alza y a la baja y, por lo tanto, se desaprovecha el alto potencial predictivo de las variables que pueden ofrecer el mejor desempeño en los procesos de clasificación y predicción.

Teniendo en cuenta este contexto y estableciendo una ruta a seguir, este trabajo de investigación presenta un método novedoso para predecir, a corto plazo (un periodo futuro en un marco de tiempo de 15 minutos), el movimiento direccional del tipo de cambio euro-dólar. La metodología propuesta se basa en el uso de datos de mercado y en el análisis discriminante junto con las etapas de preparación de datos, selección de variables predictoras y detección de estructuras subyacentes para validar la técnica y las variables predictoras seleccionadas.

La predicción de la dirección del movimiento se aborda como un problema de clasificación de aprendizaje supervisado. El poder discriminante del modelo de clasificación se evalúa mediante pruebas paramétricas que aportan pruebas estadísticamente significativas de que la función lineal discriminante contribuye a la diferenciación entre movimientos alcistas y bajistas. Por último, el poder predictivo del modelo se evalúa con datos dentro y fuera de la muestra de estudio utilizando los indicadores de desempeño más reportados en la literatura científica.

1.1 Planteamiento del Problema

1.1.1 Descripción del problema

En un esfuerzo por comprender y descifrar el panorama futuro de los mercados financieros, la predicción de los precios de los activos se ha convertido en un tema desafiante para expertos, inversores e investigadores. Aunque se dedican numerosos esfuerzos a esta difícil y compleja tarea, la cuestión de predecir con eficacia y exactitud la dirección de los tipos de cambio en el mercado FOREX requiere en realidad

soluciones más precisas, comprensibles y parsimoniosas.

Dada la relevancia que ha ganado últimamente la preparación y uso de enormes bases de datos (*Big Data*) y la construcción de modelos financieros complejos para tomar mejores decisiones de inversión, el análisis y la predicción de series de tiempo basadas en patrones, fluctuaciones de precios e irregularidades del mercado es un proceso desafiante y monótono que requiere un conocimiento profundo del mercado y cuya previsión debe estar respaldada por estrategias comerciales consistentes diseñadas sobre un adecuado análisis técnico y fundamental.

Históricamente, la predicción del valor de los activos se ha realizado a partir del análisis de datos de mercado, la detección de patrones, formaciones chartistas y el uso de indicadores técnicos [18]. Inicialmente, el análisis se orientaba al reconocimiento visual de patrones en el comportamiento histórico de los precios para predecir su movimiento futuro. El alcance limitado de este enfoque se debía principalmente a la escasa potencia de procesamiento de los computadores y a la insuficiente capacidad para realizar análisis estadísticos de grandes volúmenes de datos. Sin embargo, estudios tempranos en [61, 62, 63, 64, 65] examinan el impacto del análisis técnico sobre el movimiento de los tipos de cambio. Los resultados sugieren que el análisis no fundamentalista, sin tener en cuenta ningún tipo de análisis económico, ayuda a formar opiniones de corto plazo, para distintos marcos de tiempo, con un mayor nivel de confianza que se invierte con el análisis fundamentalista, que define el valor intrínseco de los activos en función de factores económicos para un horizonte de largo plazo [17].

En el análisis a corto plazo, Cheung et al. [66] señala que, si bien los factores fundamentales observables son importantes, los factores fundamentales subyacentes son los que realmente dominan los cambios en los tipos de cambio. En cualquiera de estos escenarios, el valor futuro de un activo se estima a partir de mediciones basadas en eventos pasados. Aunque ambos enfoques tienen un sesgo de confianza que se invierte según el horizonte de análisis, los dos enfoques se complementan. El análisis técnico se enfoca en el corto plazo (con variables de tipo técnico), mientras que el análisis fundamental se centra en el largo plazo (basado en variables macroeconómicas). Sin embargo, la literatura científica revela una mayor cantidad de estudios que favorecen los intervalos de pronóstico con enfoques a corto plazo [29].

Desde la esfera del análisis fundamental, algunos trabajos en favor de la hipótesis del mercado eficiente [67, 68, 69, 70, 71, 72] acentúan un claro escepticismo hacia el análisis técnico de los mercados financieros. Estos trabajos enfatizan que la predictibilidad de los rendimientos y precios de los activos financieros es un tema muy controvertido

porque en un mercado eficiente el precio recoge toda la información disponible sobre el valor del activo y, por tanto, la información contenida en los precios históricos de los instrumentos financieros no tiene poder predictivo sobre el valor futuro.

Según la hipótesis del mercado eficiente, estos métodos de previsión carecen de valor, ya que los precios siguen un camino aleatorio y son imprevisibles. En consecuencia, la recopilación y preparación de grandes conjuntos de datos (a partir de información histórica basada en datos de mercado y datos fundamentales) para predecir la dirección de los precios y el valor futuro de los activos financieros aún siguen siendo un problema de investigación controvertido, desafiante y abierto a la investigación.

Por otra parte, en estudios recientes [29, 73] se informa sobre el uso de mediciones cuantitativas, datos de mercado e información heterogénea para la formulación de modelos de predicción. La selección de variables obedece a procedimientos y criterios específicos definidos por cada estudio. De manera similar, la configuración y los parámetros de cálculo de las métricas varían entre los estudios y también están determinados por las especificidades de cada estudio. En consecuencia, la definición, la configuración de parámetros, el cálculo y la selección de variables predictoras es un tema desafiante y abierto a la investigación [30]. Esto debido a que, dada la variedad y volumen de información potencialmente utilizable, desde el enfoque del análisis técnico se desconocen las variables que determinan la dirección del movimiento de los tipos de cambio y que mejores diferencias producen entre grupos.

Los resultados de los trabajos que han explorado este tema le han conferido especial atención al método de predicción. Las variables predictoras más utilizadas en la construcción de modelos predictivos incluyen indicadores técnicos, datos de mercado y variables fundamentales. Sin embargo, al examinar el contexto del mercado sobre el cual se predice la dirección del movimiento futuro, los puntos de inflexión donde se produce el cambio de dirección de la tendencia precedente no han recibido la atención adecuada. En consecuencia, los trabajos sobre este tema ignoran este contexto y, por tanto, eluden el estudio de los momentos previos a la ocurrencia de estos puntos de referencia donde se produce el comienzo de los movimientos tendenciales.

En el ámbito de la AI, la predicción de precios de activos financieros se trata como un problema de clasificación supervisado y no supervisado. El aprendizaje automático supervisado es uno de los enfoques más empleados en la previsión del mercado de valores [74, 75], las criptomonedas [76] y los tipos de cambio [77]. En los últimos años, debido al rápido crecimiento del comercio electrónico de valores, también ha aumentado el interés por entender el comportamiento de los mercados y pronosticar la tendencia

de los precios de los instrumentos financieros [78, 79, 80], las criptomonedas [31] y materias primas [81].

Un análisis de la literatura científica revela que este tema desafiante se ha explorado con el uso de tecnologías basadas en AI, métodos DL, algoritmos de ML, minería de textos, análisis de sentimiento, métodos estadísticos y modelos híbridos [32, 35, 36, 33, 37]. Actualmente existe una alta preferencia por el uso de técnicas DL debido a la precisión que se logra en las tareas de clasificación y predicción [34, 38, 39]. No obstante, uno de los problemas más comunes con este tipo de modelos es la interpretabilidad [82].

Respecto a los métodos de ML, aunque la eficacia conseguida en los procesos de clasificación y predicción ha sido ampliamente demostrada, superando incluso a algunos modelos econométricos [40, 41], el sobreajuste en la definición de modelos es uno de los mayores problemas [83]. La selección de variables predictoras, el procedimiento utilizado para la configuración y definición de parámetros de cálculo (hyper-parameters) son algunos de los problemas que representan un esfuerzo desafiante [84] frente al análisis discriminante que no es demasiado sensible a la inestabilidad de los hyper-parámetros cuando hay una clara diferenciación entre clases [57].

Asimismo, en el campo de la AI, la construcción de modelos predictivos se enfrenta a uno de los mayores retos, el "problema de la caja negra". Este problema desafiante hace indescifrable el procedimiento que el modelo utiliza para clasificar o predecir la pertenencia de cada observación a un grupo [82, 85]. Por lo tanto, no puede explicar cómo interactúan las variables de entrada, utilizadas en la formulación del modelo, para predecir la dirección de los precios [85].

En cuanto al análisis discriminante, su uso en la predicción del movimiento direccional de los precios y de los rendimientos logarítmicos de los activos financieros es menos frecuente respecto a los métodos de AI [74]. Tal vez se deba a uno de los problemas más comunes, frecuentemente reportados en varios trabajos, la naturaleza asimétrica y leptocúrtica de los datos debido a su distribución no normal [81, 86]. Esta característica, inherente a este tipo de datos, hace que estas medidas (rendimientos logarítmicos) no puedan adoptar una distribución normal multivariante [87, 88]. En consecuencia, si se violan estos supuestos, el uso de métodos estadísticos paramétricos y sus interpretaciones no son confiables. Este problema también ha sido explorado con el uso de modelos econométricos, no obstante, es un tema que aún está abierto a la discusión [89, 90, 91, 92, 93, 94, 95, 96, 97].

En resumen, desde el punto de vista del análisis técnico y haciendo uso de datos de mercado, existe una gran necesidad de pronosticar el movimiento direccional del tipo

de cambio euro-dólar a partir de modelos legibles, parsimoniosos y altamente precisos. La efectividad del modelo no solo se mide por la tasa de precisión lograda en las tareas de clasificación y predicción, sino que también debe permitir explicar cómo las variables predictoras influyen sobre la acción del precio para estimar la dirección del movimiento futuro [60]. En este sentido, la precisión en la clasificación es un efecto explicativo de la legibilidad e interpretabilidad de las variables que intervienen en la predicción.

En consecuencia, para intervenir de manera efectiva este problema, esta tesis se ha propuesto superar cuatro desafíos principales: (1) Comprender el proceso que siguen los precios en la formación de los puntos de inflexión del mercado para utilizarlos en la medición, clasificación y predicción de la dirección del movimiento futuro. (2) Superar el sesgo natural de los datos (debido a su disposición asimétrica y leptocúrtica), mediante la extracción de una muestra de datos con distribución normal multivariante, para asegurar la fiabilidad y validez del método y de las interpretaciones. (3) Abordar el proceso de definición de los determinantes del movimiento direccional del tipo de cambio como un problema de selección de características observables. (4) Abordar la predicción de la dirección del tipo de cambio como un problema de clasificación de aprendizaje supervisado en donde la formulación del modelo se realiza en función de los datos muestreados y las variables predictoras seleccionadas.

Finalmente, la evaluación de la capacidad predictiva del modelo se confirma utilizando datos dentro y fuera de la muestra de estudio. La efectividad de los resultados y las interpretaciones realizadas se ratifican con el poder estadístico de las pruebas paramétricas. La eficacia de los predictores seleccionados y la idoneidad de la técnica utilizada en la construcción del modelo predictivo se corroboran desde un enfoque multivariante mediante la detección y análisis de estructuras subyacentes y la consecución de una tasa de precisión en la clasificación que supera los umbrales alcanzados por las publicaciones reportadas en la literatura científica especializada.

1.1.2 Formulación del problema

Desde el punto de vista del análisis técnico y estadístico, y haciendo uso de los datos históricos del mercado y del carácter multidimensional de las diferencias entre movimientos alcistas y bajistas, la preparación de datos, la selección de atributos, así como la detección y el análisis de estructuras subyacentes (que validan la capacidad discriminante de los predictores seleccionados y la conveniencia de utilizar el análisis discriminante) permitirán formular un modelo de clasificación parsimonioso, sencillo y muy

preciso, que explique la influencia de las variables predictoras en la acción del precio y contribuya a la diferenciación entre grupos, para predecir a corto plazo la dirección de la tendencia del tipo de cambio euro-dólar.

1.2 Objetivos

1.2.1 Objetivo general

Proponer y validar una metodología eficaz aplicable a la formulación de modelos de clasificación parsimoniosos, sencillos y altamente precisos, que permita explicar y predecir en el corto plazo la dirección de la tendencia del tipo de cambio euro-dólar.

1.2.2 Objetivos específicos

- Diseñar un procedimiento de preparación y disposición de colecciones de datos con el fin de hacerlas analizables y relevantes en el cumplimiento de los propósitos establecidos en las fases de selección de atributos, detección de estructuras y formulación de modelos.

- Diseñar un método de selección de atributos bajo un enfoque estadístico exploratorio - confirmatorio capaz de identificar las variables predictoras independientes con mayor poder discriminante, que mejor expliquen la acción del precio, y produzcan el mejor desempeño en la clasificación.

- Diseñar desde una perspectiva exploratoria multivariante y considerando la naturaleza y estructura de las diferencias entre clases, un procedimiento para validar el poder discriminante de las variables predictoras seleccionadas y determinar la idoneidad del método de clasificación a utilizar en la predicción.

- Formular, validar y evaluar una función discriminante lineal basada en datos de mercado muestreados y variables predictoras seleccionadas, para predecir la dirección del movimiento futuro del tipo de cambio euro-dólar en un horizonte de tiempo de corto plazo.

1.3 Contribuciones de esta tesis

- Diseño de una metodología eficaz, aplicada a la formulación de modelos de clasificación, que permite explicar y predecir a corto plazo la dirección de la tendencia de los precios de los instrumentos financieros para la toma de decisiones financieras informadas.

- Diseño de una metodología novedosa, orientada a la formulación de modelos de clasificación, que integra la preparación de los datos, la selección de características y la validación de la eficacia del método y los predictores seleccionados.

- Diseño de un modelo de negociación algorítmico de seguimiento de tendencia en un marco de tiempo de 15 minutos del tipo de cambio euro-dólar para identificar los puntos de inflexión en el mercado.

- Diseño de un modelo de selección de atributos para la construcción de factores comunes como ayuda a la detección y análisis de estructuras subyacentes, capaz de revelar en el subconjunto de variables el poder discriminante de los predictores seleccionados y la idoneidad del método utilizado.

- Diseño de un modelo de optimización multiobjetivo para la selección de instancias que garanticen el cumplimiento de los supuestos de normalidad multivariante e igualdad de matrices de varianza-covarianza intragrupos.

- Diseño de un método de selección de atributos bajo un enfoque estadístico exploratorio - confirmatorio capaz de identificar las variables predictoras independientes con mayor poder discriminante y que producen el mejor rendimiento en tareas de clasificación.

- Desarrollo de una metodología eficaz para identificar y seleccionar los determinantes de la dirección del tipo de cambio euro-dólar (rendimiento medio de los precios de cierre y pendiente de la línea de regresión de estas variaciones) en los puntos de inflexión del mercado.

1.4 Estructura del documento

Este documento consta de 10 capítulos distribuidos en 6 partes: preliminares, marco de referencia, materiales y métodos, discusión de resultados, comentarios finales y anexos.

A continuación se detalla el contenido de cada capítulo:

- En el capítulo 1 se hace una breve introducción al problema de estudio de esta tesis. Se presenta el planteamiento del problema, los objetivos, las aportaciones de esta tesis y la estructura del documento.

- El capítulo 2 es un análisis del contexto en el que se ofrece una visión teórica sobre la predicción y negociación de tipos de cambio. Además, se abordan aspectos esenciales como el análisis técnico, la definición de estrategias de negociación y el papel de los juicios de valor en la construcción de previsiones.

- El capítulo 3 trata sobre el estado del arte de la construcción de modelos de clasificación para predecir la dirección de los precios. Incluye aspectos técnicos como la selección y recopilación de datos, el preprocesamiento, la construcción de modelos de clasificación, la evaluación del rendimiento de los modelos y algunas consideraciones generales.

- El capítulo 4 corresponde al marco teórico del estudio, aquí se tratan conceptos técnicos y teóricos relacionados con los tipos de cambio, la formación de precios de mercado, los enfoques de resolución de problemas, la negociación algorítmica, la selección de características, el análisis biplot y el análisis discriminante.

- En el capítulo 5 dedicado a los datos aborda aspectos técnicos específicos del campo de estudio. Se hace referencia a los datos del mercado, los puntos de inflexión en los que se produce el cambio de dirección de los precios y el uso de datos de operaciones de mercado.

- El capítulo 6 dedicado a la metodología propuesta describe los procedimientos diseñados para alcanzar los objetivos propuestos. Aborda los procedimientos diseñados para la preparación de datos, la selección de características, el análisis de estructuras, la construcción del modelo de clasificación y la evaluación del rendimiento del modelo.

- El capítulo 7 dedicado a la configuración de experimentos trata del diseño de las pruebas, la definición de parámetros y medidas de rendimiento utilizadas en el desarrollo de los experimentos.

- El capítulo 8 de resultados presenta y discute los hallazgos obtenidas con el desarrollo experimental. Este capítulo incluye el análisis de resultados obtenidos con

la selección de características, el análisis de estructuras, el análisis discriminante y la evaluación del rendimiento del modelo de predicción.

- En el capítulo 9 se presentan las conclusiones obtenidas del proceso de investigación, y el trabajo futuro expone las líneas de investigación a seguir en base a los resultados obtenidos.

- Finalmente, en el capitulo 10 se presenta la producción académica, que abarca diversos aspectos de relevancia científica como la participación en congresos, y la publicación de un artículo de investigación en una revista científica especializada.

Parte II

Marco referencial

2. Análisis de contexto

2.1 Introducción

La negociación de activos en los mercados financieros en condiciones de incertidumbre ha atraído cada vez más la atención de los participantes del mercado [98]. El interés por su estudio se ha intensificado con las crisis, debido a sus efectos adversos sobre el valor de las inversiones. Actualmente, la compleja ambigüedad de las señales producidas por los mercados ha acentuado el papel de los juicios de valor en la predicción de la dirección de los precios de los activos. Aunque el análisis predictivo ha avanzado mucho, especialmente con el uso de una gran potencia informática, métodos de AI y enormes bases de datos (*Big Data*), la toma de decisiones de inversión sigue dependiendo no directamente de las previsiones, sino de las conclusiones generadas mediante juicios de valor [99]. Desde el contexto de la negociación y la predicción de los precios de los activos financieros, esta sección ofrece una visión holística de los mercados y el entorno de la negociación para comprender el papel que desempeña la construcción de modelos de clasificación y la influencia que tienen estas predicciones en la construcción de juicios de valor en la toma de decisiones de inversión.

Los resultados de esta sección se han inspirado en los estudios publicados en el ámbito de las finanzas conductuales, el análisis fundamental y el análisis técnico. El estudio de este contexto conduce a importantes consideraciones que deben tenerse en cuenta en la construcción de modelos de negociación y clasificación (para predecir, a corto plazo, y beneficiarse de la dirección del movimiento del tipo de cambio euro-dólar). Siguiendo la Figura 2.1, esta sección está organizada en cinco campos de estudio: Riesgo e incertidumbre del mercado, Irracionalidad de los mercados financieros, Análisis técnico, Estrategias de negociación, Previsiones y Juicios de valor.

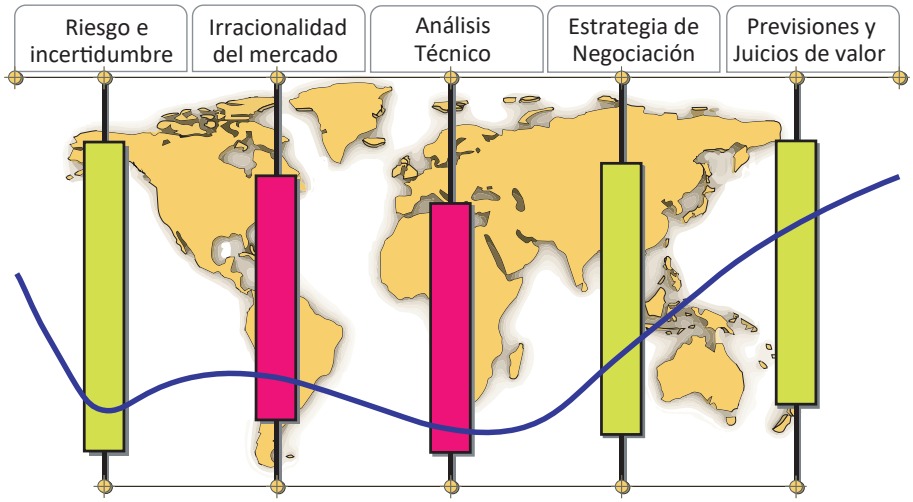


Figura 2.1: Esquema simplificado del contexto de predicción del valor de los activos. Se destaca la influencia de la incertidumbre y el riesgo de mercado en las decisiones de inversión. El papel de los modelos de predicción y valoración de activos desde un enfoque fundamental permeado por la irracionalidad del mercado y la influencia del análisis técnico. La identificación de aspectos comunes entre los modelos de previsión y de negociación respecto a la detección, predicción y capitalización de las anomalías del mercado.

Este capítulo se ha organizado en cinco secciones. La sección 2.2 resume brevemente los efectos de los acontecimientos que han impactado el comportamiento de los mercados y la respuesta de los expertos para hacer frente a estas adversidades. La sección 2.3 examina el papel de la previsión en el contexto de la valoración de los activos financieros. La sección 2.4 ofrece una breve descripción sobre el análisis técnico en la detección de patrones para predecir y explotar las anomalías del mercado. La sección 2.5 analiza los tipos de estrategias de negociación de activos y los elementos comunes que comparten con los modelos de predicción y valoración de activos. Finalmente, la sección 2.6 trata de la influencia de los modelos predictivos en la construcción de juicios de valor en la toma de decisiones de inversión. Este apartado se ha dividido en seis partes: (2.6.1)Juicios de valor; (2.6.2) Precisión de la previsión; (2.6.3) Heurística de juicio y sesgo de previsión; (2.6.4) Datos de mercado e información de contexto; (2.6.5) Comunidad inversora y previsiones de expertos; y (2.6.6) Horizonte de tiempo en las previsiones.

2.2 Riesgo e incertidumbre del mercado

Los recientes acontecimientos han afectado la estabilidad económica y la confianza en los mercados financieros. La percepción generalizada de un mayor nivel de riesgo en las

decisiones de inversión ha contribuido a crear una sensación de incertidumbre entre los inversores. Esta condición del mercado no favorece la previsibilidad del valor de los activos y el impacto negativo causado por estos acontecimientos también ha suscitado un gran debate. Se trata del verdadero significado que ha adquirido la confianza cuando se toman decisiones de inversión en condiciones de riesgo e incertidumbre. Según *Keynes* [100], la confianza en los mercados explica no sólo la estabilidad económica sino también el origen de las crisis financieras. Si los agentes del mercado pierden la confianza en él, el valor de los activos se desploma. En este sentido, este apartado refleja los efectos de los acontecimientos que han impactado en el comportamiento de los mercados y la capacidad de respuesta de los expertos financieros para hacer frente a estas adversidades.

La Gran Depresión, que comenzó en 1929 y terminó en 1939, se originó en Estados Unidos. La caída de los rendimientos en 1920 condujo a una economía inestable y vulnerable [101]. La competencia desenfrenada, la especulación y el afán de lucro, combinados con la rápida caída del consumo, el pánico y un marco político y reglamentario erróneo, contribuyeron al deterioro de la producción, al aumento del desempleo y a la quiebra. Este efecto recesivo además de haberse extendido y afectado a otras economías del mundo, llevó a la implementación de un nuevo orden basado en la intervención, la regulación y el control [102]. El papel de los analistas afectados por el sentimiento del mercado en el período previo a estos acontecimientos puso de manifiesto las importantes diferencias entre las señales del mercado y la valoración de los activos basada en fundamentales [103]. La evidencia empírica ha demostrado que, al examinar la relación *inversión-valoración-fundamentales*, el rol de los expertos financieros en la valoración del mercado es limitado. Especialmente cuando el análisis, para la toma de decisiones inmediatas, se basa en datos fundamentales. Así, incluso en ausencia de riesgo fundamental, los precios del mercado pueden desviarse de los valores fundamentales [104]. En la actualidad, estas discrepancias suscitadas en condiciones de incertidumbre siguen afectando a la valoración y predicción del valor de los activos a corto plazo en la toma de decisiones de inversión. A la luz de los mercados financieros, entre otras cuestiones que siguen cobrando importancia, se encuentra el tratamiento de la incertidumbre generada por estos acontecimientos. Una cuestión muy debatida hasta la fecha es si la incertidumbre puede transformarse en factores de riesgo calculables [105, 106]. Cabe señalar que, entre las posibles soluciones a este difícil reto, las medidas más estrictas impuestas en la década de 1930 fueron en su mayoría de carácter normativo y su finalidad era generar confianza en los mercados y en los colectivos im-

plicados en las negociaciones. Sin embargo, la tensión política en Europa que provocó la entrada de oro en Estados Unidos, como resultado de la expansión monetaria, acabó conduciendo a finales de 1939 a la *Segunda Guerra Mundial* [102].

Algunos estudios han indicado que los efectos de los acontecimientos históricos se reflejan en los precios de los activos [107]. En Europa tras el estallido de la Segunda Guerra Mundial y los cambios de soberanía, los bonos emitidos por el gobierno alemán parecen anticipar en sus precios el efecto bélico de la guerra y la caída de los programas de Hitler [108]. Los efectos de estos acontecimientos reducen la precisión y fiabilidad de las previsiones. Incluso el uso de información real, disponible y clara en condiciones de incertidumbre es insuficiente para la construcción de juicios de valor para la realización de estimaciones bajo reserva. Tras la *Gran Depresión* y el final de la *Segunda Guerra Mundial*, la política por defecto sigue siendo el restablecimiento del crecimiento económico mundial, el fomento de la producción y del consumo [109]. En tiempos de transición, el interés de los agentes del mercado, además de identificar la dirección de los rendimientos entre los activos y el mercado, se centra en detectar qué activos son sistémicamente más importantes que otros. Los analistas redefinen sus estrategias de negociación, equilibrando sus carteras y reduciendo el riesgo de estas, lo que conduce a una reducción/incremento de los precios de los activos de alto/bajo riesgo [110, 111].

Luego del período de auge de finales de los años 90 en Estados Unidos, se generalizó entre el público la visión optimista del mercado. El crecimiento sistémico y desbordado del crédito y otras formas de riesgo se hicieron cada vez más evidentes antes de la crisis bancaria [112, 113]. La racionalidad limitada y el contagio social de grupos mal informados, pero motivados por el sesgo cognitivo de invertir, fomentaron la inversión en activos de alto riesgo [114]. Aunque la incertidumbre y las señales contradictorias del mercado iban en contra de los intereses del colectivo, fue este conjunto de factores subyacentes el que desencadenó en parte la *Crisis financiera mundial de 2008*. Los estudios empíricos han revelado, desde el punto de vista de las finanzas conductuales, que el exceso de confianza afecta a la toma de decisiones de inversión [115]. Los participantes del mercado motivados por la ilusión monetaria, el sesgo del statu quo y la disonancia cognitiva rechazan la información relevante, contradictoria con sus intereses, para tomar decisiones no óptimas [116]. La negociación de instrumentos complejos y de alto riesgo basados en productos de crédito suele estar asociada a problemas de liquidez. Aunque la confianza en las calificaciones crediticias, propor-

cionadas por agencias especializadas, puede crear una falsa sensación de seguridad. Las agencias de calificación han utilizado las calificaciones crediticias, en la relación entre riesgo e incertidumbre, para transformar la incertidumbre en riesgo [117]. Sin embargo, los problemas y las limitaciones de estas prácticas ampliamente aceptadas y utilizadas en la gestión del riego crediticio volvieron a relucir con la crisis financiera de 2008 [118].

A finales del 2019, en pacientes con neumonía, vinculados a un mercado de mariscos en *Wuhan* (China), se descubrió un *betacoronavirus* desconocido. El nuevo *coronavirus 2019-nCoV* fue el agente causante de los brotes de síndrome respiratorio agudo severo que se propagó por todo el mundo [119]. Además de ralentizar la actividad económica internacional y afectar a los mercados financieros, los efectos de la pandemia produjeron la mayor contracción del *PIB* mundial (3.1%), a pesar de las medidas aplicadas para contrarrestar la recesión en las actividades productivas de la economía real [120]. Los mercados de valores han sufrido pérdidas multimillonarias, el sistema financiero mundial ha sido expuesto a la mayor prueba de resiliencia y las organizaciones se han visto obligadas a revisar sus proyecciones de crecimiento para años posteriores [121]. Mientras que los participantes del mercado en períodos de crisis hacen uso de la información disponible para tomar las mejores decisiones y salvaguardar el valor de sus inversiones. El interés por descubrir el impacto que el coronavirus ha tenido en los rendimientos financieros ha aumentado rápidamente en estos años [122]. Los primeros estudios descubrieron que los mercados financieros se deterioraban a medida que aumentaba el pánico provocado por las noticias durante la fase de transición de epidemia a pandemia. A medida que la pandemia se extendía de Asia a Europa y luego a América, la fuerte caída de los mercados financieros reportó rendimientos negativos, incluso en activos de menor riesgo como el oro. Tras las medidas adoptadas, los mercados chinos mostraron signos de recuperación mientras el virus se extendía en Estados Unidos [123]. Otros estudios, además de encontrar una clara asociación entre el pánico mediático (producido por los medios de comunicación) y la volatilidad de los mercados financieros mundiales, detectaron un fuerte impacto en los sectores más afectados [124].

En los últimos años, la *crisis de Ucrania* y la invasión rusa del 24 de febrero de 2022 han despertado un interés especial por el impacto que ha tenido a nivel mundial. El profundo desacuerdo entre la UE y Rusia debido al *conflicto UE-Rusia* por Ucrania y otros países europeos se ha enmarcado en el orden político, económico y de seguridad europeo e internacional [125]. Las sanciones occidentales sobre la economía política rusa, espe-

cialmente en los sectores financiero, energético y de defensa, y el acercamiento de Rusia a Asia están redefiniendo y posibilitando la integración de Rusia en un nuevo sistema de economía política [126]. Mientras esto ocurre en la esfera política, el aumento de los precios de las materias primas, las interrupciones y la escasez de suministros están afectando las economías europeas. Por otra parte, la situación crítica de los refugiados, la disminución de la confianza de los inversionistas de estos países y la incertidumbre de los mercados hacen que la definición del valor de los activos, que normalmente se basa en las condiciones del mercado, no tenga ningún respaldo en este momento.

En resumen, la incertidumbre y sus efectos adversos debido al impacto disruptivo de las grandes crisis han influido en parte en la volatilidad de los mercados y en el aumento del riesgo de inversión. Al final de las crisis, los datos de mercado reflejan el impacto de los acontecimientos y ayudan a dilucidar la versatilidad de las estrategias y los modelos de previsión utilizados para superar la irracionalidad del mercado. Dada esta condición, es el aprendizaje por retroalimentación el que ayuda a mejorar la eficacia de los métodos de previsión y de negociación utilizados por los inversores. Si bien existe un nutrido cuerpo de literatura que ha estudiado los efectos de diversos eventos y su influencia con las estrategias de negociación, los modelos predictivos y el comportamiento de los precios para diversos instrumentos financieros [127, 111, 128], el análisis de estas interdependencias ayuda a comprender mejor los mecanismos utilizados por los participantes para actuar en el mercado [129].

El interés por medir y anticipar el riesgo para minimizar los efectos indirectos de la incertidumbre se ha intensificado con la especulación. La valoración y previsión de los precios en los activos a propendido por identificar los mercados de alto riesgo para trasladar los recursos escasos y de uso optativo a mercados de bajo riesgo [130, 111]. Sin embargo, las crisis han demostrado el uso intensivo de estrategias de cobertura, por parte de los inversores y gestores de carteras, para reducir la exposición al riesgo debido a la volatilidad de los precios y los mercados [131]. Algunos detractores han cuestionado este tipo de prácticas. *Alexander Engel* [132] señala que estas operaciones sólo contribuyen a aumentar la exposición al riesgo de precios. Esto se debe a que el propio riesgo se ha transformado en una mercancía. Por otro lado, esto le ha llevado a cuestionar el papel de los mercados de futuros, dado que no fueron creados para ofrecer servicios de gestión de riesgos al público sino para su mercantilización. Además, el exceso de confianza de los inversores y la inadecuada gestión de riesgos subestimados

en condiciones de incertidumbre son, en parte, los inductores de la variabilidad de los precios y del comportamiento volátil e irracional del mercado. Estas condiciones, combinadas con un entorno competitivo dominado por la presión y el estrés, hacen que la previsión precisa del valor de los activos sea una tarea compleja y desafiante.

2.3 Irracionalidad del mercado

Durante décadas, las finanzas han dominado los estudios sobre la racionalidad de los mercados y las teorías sobre la valoración y la previsión del valor de los instrumentos financieros. La elección racional con efectos inciertos es el paradigma, basado en la *Teoría de la Utilidad Esperada*, que sigue prevaleciendo en la toma de decisiones de inversión bajo riesgo [133, 134]. Aunque las finanzas conductuales intentan explicar los sesgos y la irracionalidad de las decisiones de inversión ante las anomalías del mercado, en los últimos años, la comunidad académica se ha interesado cada vez más por los modelos de precios basados en el comportamiento de los individuos [135]. Este creciente interés, en el que se integran de forma multidisciplinar diferentes ramas del conocimiento como la economía, las finanzas, la psicología, la sociología y la neurología, se debe fundamentalmente a la necesidad de poder explicar el movimiento de los mercados y cómo los expertos financieros toman las decisiones de inversión. A este respecto, a continuación, se describe el contexto en el que han surgido los modelos de valoración de activos. Así como la influencia que han tenido en la predicción y definición de precios con fines de inversión y especulación.

A lo largo de *La Teoría Financiera*, la necesidad de comprender cómo se forman los precios de los activos en los mercados financieros ha dado lugar a diversos enfoques teóricos. La hipótesis de los mercados eficientes [25] sentó los primeros precedentes en la definición de modelos para encontrar los mejores procedimientos de inversión. Dado el fuerte escepticismo hacia la previsión de las variables financieras (precios, rendimientos, volatilidad, dirección del movimiento, entre otras), subraya que las variaciones históricas de los precios de los activos no aportan información útil. Por tanto, no tienen poder predictivo sobre los precios futuros porque supone que el precio del activo refleja toda la información disponible del mercado. Sin embargo, dada la presencia de anomalías que violan estos supuestos financieros y dan lugar a retornos anormales, el creciente cuerpo de investigación empírica se ha enfocado en demostrar que las variables financieras son predecibles.

Los resultados de las primeras investigaciones han revelado la presencia de anomalías en el mercado que contradicen la teoría de los mercados eficientes. Estas se producen cuando la acción del precio del activo negociado no se comporta como se espera. Las evidencias han mostrado la existencia de tres tipos [136]. Las anomalías fundamentales tienden a producir mayores rendimientos después de que los activos hayan alcanzado niveles de sobreventa, mientras que, tras alcanzar niveles de sobrecompra, los precios tienden a describir movimientos a la baja. En cuanto a las anomalías técnicas, *Murphy* [137] señala que, en determinadas condiciones, los mercados financieros no son totalmente eficientes, lo que hace que los movimientos futuros de los precios sean predecibles a partir de patrones o comportamientos históricos que se repiten y validan en el tiempo. Por otro lado, las anomalías debidas al efecto calendario son las que se producen en momentos concretos dando lugar a rendimientos anormales.

El creciente interés por comprender estas anomalías que subyacen a la irracionalidad de la comunidad inversora es el principal catalizador para encontrar respuestas sobre el comportamiento predecible de los movimientos de los precios [138]. El propósito básico de los agentes que participan en la negociación, además de buscar beneficios, es identificar y capitalizar los hallazgos que mejor puedan explicar cómo mejorar sus decisiones de inversión. Es en este contexto en el que los expertos financieros se centran en identificar, por asociación, patrones en los precios capaces de producir rendimientos consistentes para ser considerados en la construcción de modelos de previsión. Por otro lado, en las últimas décadas, los investigadores y expertos financieros se han centrado en la teoría del comportamiento para descubrir los procesos heurísticos que utilizan los agentes del mercado al tomar decisiones de inversión [139]. El análisis de los sesgos de previsión y la influencia que el uso de la información y la publicación de datos fundamentales (estadísticas o indicadores que reflejan los hechos económicos de los mercados financieros, de un sector específico o de la economía de un país) tienen en los juicios de valor [140]. Sin embargo, en el proceso de toma de decisiones hay una serie de factores y variables que influyen en el comportamiento de los analistas y expertos financieros que pueden ser investigados. Quizá muchos de estos factores sociales, culturales, demográficos y socioeconómicos, entre otros, puedan influir en las decisiones de inversión y explicar mejor las anomalías del mercado.

Las contribuciones de *Bachelier, Samuelson, Fama, Ross, Tobin* y *Shiller* han con-

tribuido a sentar las bases de las finanzas modernas [141]. Aunque no existe una acepción universal entre la comunidad académica, lo que está claro es que estos argumentos le han acercado a la búsqueda de una mejor comprensión del proceso que sigue el precio en la formación del valor de los activos. A principios del siglo XX, Bachelier se centró en desarrollar el concepto de paseo aleatorio. Sus aportaciones simplificaron el movimiento de los precios de las acciones a un proceso aleatorio cuya variación entre períodos es independiente [142]. *Harry Markowitz* [143], con la teoría moderna de portafolios, demostró a finales de los años cincuenta que el riesgo de un activo se valora en función de su contribución al riesgo total de la cartera. En la misma línea, Sharpe, además de sus aportaciones al método binomial de valoración de opciones [144], estableció el modelo de valoración de activos *CAPM* [145] y el ratio de Sharpe [146] para el análisis del rendimiento de las inversiones. La eficacia del modelo CAPM se ha evaluado a través de un número creciente de estudios realizados en los últimos cincuenta años. Entre los resultados más notables se encuentra que las fluctuaciones de los rendimientos esperados de los activos no pueden explicarse únicamente por el riesgo sistémico. Los factores que intervienen son, de hecho, variados e inciertos. Esto ha llevado al desarrollo de diversos enfoques para predecir el rendimiento de los instrumentos financieros [147].

Fama había observado que la distribución de los precios no cumplía los supuestos de normalidad [69]. Estos primeros hallazgos sentaron las bases de su tesis doctoral [141]. Más importante aún, la hipótesis de los mercados eficientes sentó los precedentes de la *Teoría Financiera* [25]. En los años 70, *Fischer Black* y *Myron Scholes* [148] redefinieron el estudio de Bachelier. Su trabajo dio lugar a la fórmula que se utilizaría en la valoración teórica de las opciones. Paul Samuelson a favor del mercado eficiente *"Prueba de que los precios correctamente anticipados fluctúan de forma aleatoria"* [149]. Además de sugerir que no es posible prever ningún cambio adicional en los precios, porque incorporan toda la información disponible, detecto un comportamiento estocástico con propiedades de *Martingala*. James Tobin, el último de los pioneros de la *Teoría Financiera Moderna*, para entender cómo el precio de mercado de un activo refleja su valor intrínseco (valor contable), propuso la ***Q*** *de Tobin* [150]. Este índice le permitió detectar si un activo está *sobrevalorado* o *infravalorado*. Mas adelante, el modelo de valoración de activos *APT* [151] *(Arbitrage Pricing Theory)* de Stephen Ross propuesto a principios de la década de 1970, consiguió integrar tres supuestos: fijación de precios, arbitraje y mercados eficientes. La principal aportación de este en-

foque predictivo, además de eliminar algunos de los supuestos más problemáticos del modelo CAPM, es que gestiona mejor el *Riesgo Sistémico* y el *Riesgo Idiosincrático* de un sector en concreto. Stephen Ross pudo demostrar que los cambios en los precios de las acciones están correlacionados con las noticias macroeconómicas, los precios de las materias primas y el consumo [141]. Al final de este periodo, además de consolidarse los modelos financieros y econométricos dominantes [152], se introdujeron importantes mejoras en el uso de las técnicas, la inclusión de supuestos menos restrictivos y la obtención de valor dada la naturaleza de los datos. Sin embargo, tanto las finanzas neoclásicas como las cuantitativas aún tienen algunas cuestiones sin resolver. Parte de estos problemas, relacionados con el comportamiento de los precios de mercado, no pueden ser explicados todavía por la teoría financiera anteriormente expuesta.

En los últimos años, la atención se ha centrado en las *Finanzas Conductuales* debido al interés por explicar la influencia de las *Anomalías del Mercado* en la valoración de los activos financieros [153, 154]. La incertidumbre en el comportamiento de los precios ha influido, en parte, en la búsqueda de nuevos procedimientos de valoración y en la previsión de sus precios [155, 156, 157, 158, 159]. Así, desde una perspectiva conductual mucho más amplia, estos modelos se basan en la irracionalidad del inversor para intentar explicar los movimientos del mercado [160]. Esto supone la ocurrencia de patrones no estándar que afectan y explican la desviación del precio de mercado respecto al valor fundamental del activo. Además, la posibilidad de incluir nuevas variables de comportamiento en la valoración de los activos constituye un gran reto [161]. Especialmente debido a los enormes volúmenes de datos disponibles para el análisis [162]. Robert Shiller cuestiona la eficiencia de los mercados financieros desde una perspectiva conductista [163]. Cree que el comportamiento del mercado está determinado por la influencia de factores psicológicos, culturales y sociológicos. Además, en sus estudios destaca que no todos los agentes que participan en los mercados actúan de forma racional, sino que muchos están influenciados por la *"exuberancia irracional"* [164].

En el ocaso de la teoría financiera neoclásica, *Kahneman*, en su crítica a la *Teoría de la Utilidad Esperada*, en contraste con los modelos neoclásicos de valoración de activos, desarrolla a finales de los años 70 un modelo alternativo *"La Teoría de la Perspectiva"* [165]. Este enfoque basado en el análisis de decisiones de alternativas de inversión bajo riesgo describe muchos de los problemas de elección en los que las

preferencias violan los supuestos de la teoría de la utilidad [166]. En la misma línea, *Gigerenzer* [167, 168] encontró que el razonamiento rápido centrado en la elección, en condiciones de incertidumbre, es el desencadenante del uso de la heurística para la toma de decisiones financieras [169]. Aunque su uso suele ser muy práctico, también puede conducir a errores sistémicos y predecibles [116]. Esto ha despertado el interés por mejorar su eficacia a la hora de emitir juicios de valor, especialmente en la toma de decisiones de inversión en condiciones de incertidumbre [170, 171]. Sin embargo, la publicación de noticias y datos fundamentales además de influir en la heurística de juicio también atrae la atención de los investigadores. Sobre todo, porque el uso de información actualizada incide en la formación de expectativas lo que podría afectar la creación de juicios de valor debido a los sesgos de previsión [172]. Los estudios empíricos sobre el análisis de sentimiento han demostrado que, dadas las creencias que tienen los expertos financieros sobre el valor fundamental del activo y las expectativas originadas por la publicación de noticias, pueden provocar un exceso de optimismo o pesimismo sobre la evolución futura de los precios [173, 174, 175]. Así, el uso eficaz de la información, la previsión, el análisis del sentimiento, la heurística y el juicio de valor en la toma de decisiones de inversión, aunque pueda estar sujeto a algún tipo de sesgo, es prácticamente la base contextual sobre la que se estructura la negociación de activos en los mercados financieros.

Por último, es preciso mencionar que, al no existir una teoría unificadora entre los modelos clásico y conductual, estos dos grandes campos de las finanzas seguirán divididos. No obstante, los recursos de uso limitado seguirán el curso de la inversión con aquellos enfoques que, a pesar de la incertidumbre del mercado, procuren la mayor rentabilidad y la menor exposición al riesgo.

2.4 Análisis Técnico

Los inicios del análisis técnico actual se remontan al año 1900 con los estudios de Dow. Se cree que *Charles Dow* fue uno de los primeros en estudiar el comportamiento de los precios en los mercados financieros. Estableció el primer *"barómetro bursátil"* utilizando una media de los precios de los valores más representativos del mercado, lo que le permitió definir la tendencia de los mercados [176]. Al sentar las bases de la *Teoría de Dow*, se convirtió en lo que hoy se consideraría una herramienta de predicción y guía de inversión para los analistas financieros. Al explorar los hallazgos descubiertos

en este campo de estudio, esta sección ayuda a desvelar la transversalidad compartida entre el análisis fundamental, el análisis técnico y las finanzas conductuales. Además, se mencionan importantes consideraciones a tener en cuenta para la construcción de modelos de previsión y negociación.

En los últimos años ha crecido el interés por el uso del análisis técnico en el estudio de los mercados financieros [177, 178, 27]. Además de ser útil para detectar patrones asociados a las anomalías del mercado [26], permite predecir la dirección de los precios, basándose en el análisis de datos históricos (precios y volumen negociado) [179]. Este enfoque se fundamenta en el análisis visual de patrones, la detección de puntos de inflexión, la formación de ciclos y el cambio de régimen en el movimiento de los precios. Estas pautas se forman en el movimiento dinámico de los precios y tienden a repetirse en el tiempo [18]. El uso de *Indicadores Técnicos* (expresiones matemáticas utilizadas sobre los precios y el volumen del activo negociado) se incrementa con la identificación de patrones de comportamiento para mejorar la precisión de las predicciones con fines de negociación [16].

Los primeros puntos de vista no fundamentalistas de Keynes [100] sobre el comportamiento del mercado acabaron siendo confirmados por las finanzas conductuales. Estos resultados empíricos mostraron que las variables no fundamentalistas también influyen en el comportamiento de los precios [180]. Los efectos irracionales de los sesgos de comportamiento de los operadores también intervienen en la formación de los precios de los activos [181]. Mas tarde, en el periodo de transición posterior a la Segunda Guerra Mundial, los economistas fundamentalistas siguen viendo con cierto recelo el ámbito del análisis técnico [20, 19]. La falta de confianza en este enfoque se debe en parte a los postulados de la hipótesis del mercado eficiente [25]. En teoría, los expertos que especulan con instrumentos financieros sin basarse en los fundamentos serán rápidamente *"liquidados"*, ya que la eficiencia del mercado lo descuenta todo [182]. Ante las dudas que despertó el análisis fundamental sobre el análisis técnico, varios autores han explorado la importancia de este enfoque en el estudio de los movimientos de los precios. Al analizar la sobrevaloración del dólar americano en la década de 1980, [183, 184, 185, 186] sugirieron que el análisis técnico, influido por sesgos de comportamiento, podría haber incidido en esos movimientos, dado que los datos fundamentales mostraban lo contrario. Otros investigadores [187, 188, 189] también han sugerido que el desplome experimentado por los mercados financieros en 1987 fue inducido en parte por el análisis técnico y el comportamiento irracional de los inversores.

Entre un amplio segmento de expertos financieros *"operadores de ruido"* parece haber un creciente interés en el uso del análisis técnico para capitalizar las anomalías del mercado [190, 191, 192, 193]. Algunos trabajos empíricos han cuestionado el uso de información racional en la formación de expectativas. Varios autores [187, 188, 194, 195] sugieren un mayor uso de los métodos de análisis técnico, por parte de los participantes del mercado, para predecir la dirección de los precios de los activos.

En las últimas décadas, también ha aumentado el interés por determinar la rentabilidad de las estrategias de negociación basadas en el análisis técnico [177, 178]. Aunque los rendimientos generados por los modelos de negociación varían en función de los activos, los enfoques y los procedimientos de valoración utilizados, los primeros estudios sobre tipos de cambio (1960-1987) han reportado beneficios considerables. Sin embargo, cabe mencionar que, a diferencia de los estudios más actualizados, los estudios empíricos anteriores a los años 90 tenían algunas limitaciones. Los modelos no se validan fuera de la muestra de estudio, los resultados carecen de pruebas estadísticas, se ignora la evaluación del riesgo y no se optimizan los parámetros de las reglas de decisión [178]. En general, los estudios empíricos que no demuestran la viabilidad financiera de las reglas de decisión no tienen en cuenta los costes de transacción y la exposición al riesgo [196]. En este punto, la fiabilidad del análisis técnico, además de ser cuestionada, puede quedar oculta bajo el llamado *"sesgo de publicación"* [197]. Debido a los escasos incentivos que supone la publicación de los grandes descubrimientos, es más plausible patentar y proteger esa novedad para luego explotarla económicamente. Por otro lado, algunos estudios empíricos, basados en datos de encuestas [198], muestran un uso bastante extendido del análisis técnico en el comercio de divisas. De hecho, el *Análisis Chartista* es el más utilizado en la detección de patrones para predecir la dirección de los precios a corto plazo [199]. En cuanto al uso de métodos de inteligencia artificial, aplicados en el análisis técnico de estrategias de negociación, se han detectado algunas deficiencias. El uso de redes neuronales no es fácilmente replicable, lo que limita la reproducibilidad de los resultados con otros activos y en otros horizontes temporales [197]. La efectividad de los métodos de aprendizaje profundo (LSTM), en la definición de las reglas de decisión, se cuestiona al no considerar los costes de transacción y la exposición al riesgo de los modelos [200].

En cuanto a los estudios de análisis técnico para la previsión de los mercados financieros, los enfoques computacionales evolutivos [201], las técnicas de optimización

(inspiradas en la naturaleza [202]) y la inteligencia artificial han ganado protagonismo en los últimos años [203]. El uso de redes neuronales [204, 205] y técnicas neuro-fuzzy, para predecir el comportamiento del mercado de valores [206], también han demostrado ser algunas de las técnicas de *"informática suave"* más aceptadas. En lo que respecta a la construcción de sistemas de negociación inteligentes, [202] además de identificar, en los estudios publicados entre 2009 y 2015, el uso de técnicas de preprocesamiento de datos y de minería de textos (a partir de noticias [207, 203]) sintetiza sus metodologías en un enfoque genérico basado en pasos. Es preciso señalar que los avances realizados hasta la fecha en la predicción de la dirección de los precios de los activos revelan unos índices de precisión muy prometedores. Sin embargo, hay algunas cuestiones críticas que aún no se han resuelto y que, por tanto, deberían ser objeto de consideración e intervención. Entre los aciertos más notables están el uso de métodos de predicción legibles, la definición de puntos de referencia en el mercado para predecir la dirección de los precios y el uso de información crítica que explique en el modelo de predicción su influencia en la formación de los precios.

Por último, hay que destacar que, dada la evidente complementariedad del análisis técnico y fundamental en el ámbito de las finanzas conductuales, las anomalías del mercado son capitalizables. Los sesgos de comportamiento subyacentes en los precios históricos son detectables a partir del análisis técnico de patrones. La fiabilidad y consistencia de estas anomalías sirven de base para la construcción de reglas de negociación y modelos de predicción. Así, la precisión en la predicción de la dirección del movimiento de los precios de los activos es una medida implícita del desempeño de los modelos de negociación. Esto sugiere que las anomalías y los patrones en los precios son el denominador común de los modelos predictivos y las estrategias de negociación.

2.5 Estrategias de negociación

Tras la publicación de la hipótesis de los mercados eficientes, son muchos los resultados de trabajos publicados en los que se explotan las anomalías del mercado con el diseño y la aplicación de estrategias de negociación. El éxito de los modelos de negociación viene determinado por la precisión de las previsiones realizadas sobre los patrones de precios. Estas estrategias de negociación emergieron en el campo del *Arbitraje Estadístico* con el comercio y la especulación bursátil, especialmente en el seno de los fondos de cobertura y los bancos de inversión [208]. Más tarde, con el diseño de indicadores técnicos,

se extendieron a otros tipos de instrumentos financieros. Al explorar las evidencias de los estudios empíricos aquí reportados, esta sección muestra que el uso eficaz de los patrones de movimiento de los precios de los activos influye en el desempeño de las reglas de decisión, las estrategias de negociación y los modelos de previsión. Además, este análisis sugiere que el uso compartido de estos patrones en dichos modelos, más allá de ayudar a predecir las anomalías del mercado, permite capitalizar los beneficios del *Sesgo de Sobrerreacción*.

La creciente literatura sobre las estrategias de negociación con dos o más activos está impulsada principalmente por el arbitraje estadístico [209]. En finanzas, el arbitraje es una forma de negociación en la que se compran y venden simultáneamente uno, dos o más activos en diferentes mercados y a diferentes precios, cuando el beneficio se hace plausible el arbitraje prevalece [210]. Teniendo en cuenta este esquema de negociación, existen básicamente tres tipos de enfoques, las estrategias de *Impulso*, *Contrarias* y de *Reversión* [211]. En la negociación de activos por *"pares"*, estas estrategias toman de un conjunto de valores, por ejemplo acciones, dos subconjuntos denominados *"ganadores"* y *"perdedores"*(Uno de los criterios más utilizados para establecer estas carteras es el rendimiento de las acciones, medido durante un periodo determinado [212]). Así, la cartera de coste cero se forma comprando las acciones "ganadoras" y la cartera de coste se forma vendiendo las acciones "perdedoras". Las carteras formadas en distintos momentos se mantienen simultáneamente durante un periodo de tiempo, tras el cual se liquidan. Los resultados financieros vienen determinados por el tipo de estrategia empleada en la negociación y el tiempo de tenencia de los activos.

Las estrategias de impulso (*Momentum*) fueron estudiadas por *Jegadeesh* y *Titman* [212]. Una medida de momentum es la aceleración/desaceleración del cambio de precios en el tiempo y suele utilizarse para definir los instantes de entrada y salida del mercado [213]. Las operaciones de mercado ejecutadas con este enfoque se basan en la *Hipótesis de la Subreacción*. Esto supone que los inversores reaccionan a la información con menos fuerza de la que realmente deberían. Como los precios no se ajustan adecuadamente, los inversores aprovechan los rendimientos producidos por los activos en los movimientos a corto plazo. Uno de muchos enfoques sugiere operar en largo/corto en activos cuya tendencia a corto plazo está en la misma dirección que la tendencia a largo plazo. Los resultados reportados con estas estrategias, evaluadas en períodos

de hasta 20 años, han sido muy favorables [214, 215, 216, 217]. Por otro lado, cuando se trata de operaciones especulativas, para marcos temporales intradía, la negociación por *Ciclos de Retroalimentación Positiva*[1] es más evidente cuando se opera a favor de la tendencia dominante.

Las estrategias contrarias fueron definidas por *De Bondt* y *Thaler* [220] y *Lakonishok* [221]. Estas estrategias aprovechan la sobrerreacción generada por los inversionistas a la información por encima de lo procedente, acentuando el movimiento de los precios. En las estrategias de impulso y contrarias la respuesta errónea inicial es rectificada posteriormente por los inversores, de manera que los precios se vuelven a equilibrar. En líneas generales, los inversores contrarios, que operan con estrategias de impulso, esperan que el movimiento direccional del precio sea seguido por movimientos en la dirección opuesta. El inversor podrá beneficiarse con estas dos estrategias tomando decisiones de inversión contrarias. La selección de valores para formar los grupos de "ganadores" y "perdedores", basada en el análisis de datos históricos, se realiza mediante el análisis de relación entre el precio de mercado y el valor intrínseco del activo o los beneficios reportados. La estrategia, estructurada según la dinámica de los precios históricos, sugiere operar simultáneamente en corto/largo con el grupo de acciones ganadoras/perdedoras una vez que estén sobrevaloradas/infravaloradas, ya que es probable que las direcciones de los precios se inviertan y se conviertan en perdedoras/ganadoras. Según De Bondt y Thaler [220] los resultados mostraron efectivamente que el grupo de "perdedores", a diferencia de los "ganadores", resulta ser rentable 36 meses después de la formación de la cartera. Esto indica que la sobrerreacción tras alcanzar niveles de resistencia/soporte es teóricamente predecible. Los resultados presentados en [222, 223] demuestran el desempeño generado por las estrategias contrarias.

Las estrategias de cambio de tendencia o de reversión inicialmente exploradas por *Lehmann* [224] actúan de forma opuesta a las estrategias de impulso. En este caso, el subconjunto de valores ganadores está formado por acciones de menor rendimiento, y

[1]Negociación por ciclos de retroalimentación positiva: Es un sesgo de comportamiento muy común entre los "operadores de ruido" en la especulación de activos. Se caracteriza por replicar, según patrones que validan un comportamiento histórico repetitivo, posiciones basadas en la tendencia dominante del mercado. Este tipo de operativa, además de aprovechar la tendencia de los precios, impulsa a otros analistas a realizar las mismas operaciones en busca de resultados similares, lo que produce el efecto rebaño [218]. En términos generales, la negociación a través de ciclos de retroalimentación positiva (negativa) consiste en ejecutar operaciones de compra (venta) del activo negociado cuando el valor del activo se aprecia y se vende (compra) cuando el valor del activo se deprecia [219].

el subconjunto de valores perdedores está formado por acciones de mayor rendimiento, analizándose ambos grupos en el mismo horizonte temporal. En este tipo de estrategia, el supuesto utilizado en la negociación es que las acciones ganadoras, valores recientes de menor rendimiento situados en la *Zona de Sobreventa*, se convertirán en "ganadores". Las acciones "perdedoras", valores recientes de mayor rendimiento situados en la *Zona de Sobrecompra*, perderán impulso y se convertirán en perdedores. El rendimiento de estas estrategias también ha sido demostrado en horizontes temporales de largo plazo y su desempeño es favorable [225, 226, 227].

En las estrategias de impulso y de reversión los grupos de acciones ganadoras y perdedoras se definen de forma opuesta. En [228] se indica que estas estrategias funcionan bien siempre que las medidas de impulso y los periodos de mantenimiento se calculen correctamente a partir de datos históricos y en ventanas temporales adecuadas. Aunque hay una gran diversidad de criterios para seleccionar grupos de acciones ganadoras y perdedoras [229, 230], *Asness* [231] señala que las estrategias de impulso proporcionan rendimientos sólidos para diferentes clases de activos, incluidas las materias primas y los tipos de interés. Este tipo de estrategia se ha convertido en una de las más utilizadas en la negociación de tipos de cambio y criptomonedas, su desempeño ha quedado ampliamente demostrado [232, 217, 233].

Las estrategias de impulso, basadas en el análisis de datos históricos, pueden estructurarse en función de una amplia gama de medidas de impulso [234, 213, 212]. Especialmente para la definición de los momentos de entrada y salida del mercado (reglas de decisión). El uso de estas medidas también tiene como objeto predecir el comportamiento futuro de los precios de los activos para un período de tiempo determinado.

El uso vinculante de las medidas de impulso con patrones de precios específicos permite detectar y explotar las anomalías del mercado. Estos patrones subyacentes a la estructura multidimensional de los datos son quizás determinantes de las diferencias entre el comportamiento alcista y bajista de un activo. Estas pautas predecibles, que suelen atribuirse a las reacciones erróneas de los inversores, y que se deben a sesgos de comportamiento, se producen con el cambio de dirección de los precios [235]. Así, es posible diferenciar entre movimientos direccionales alcistas y bajistas tomando como base la clasificación. Este enfoque predictivo que tradicionalmente se ha utilizado para abordar problemas de predicción es potencialmente útil para validar patrones para la definición de estrategias de negociación.

Hoy en día, la construcción de estrategias de negociación se ha visto facilitada en parte por el uso de técnicas avanzadas de análisis de datos, una alta capacidad de procesamiento y la disponibilidad de grandes volúmenes de datos [236]. Lo que conduce teóricamente a la estructuración de una estrategia de negociación óptima que va más allá de la simple definición de reglas de decisión genéricas de *Comprar-Mantener-Vender* o bien de *Vender-Mantener-Comprar*. Esto implica que la eficacia de la estrategia viene determinada por la idoneidad del activo seleccionado, la precisión de los patrones utilizados en las reglas de decisión y los métodos empleados en la construcción y optimización. Además, la fiabilidad y consistencia de la estrategia deben reflejarse en las posiciones exitosas, con menor exposición al riesgo, que superan en número y beneficios a las posiciones fallidas.

En general, existe una estrecha relación entre las estrategias de negociación y los modelos que predicen la dirección de los precios de los activos [237, 238, 239]. Los patrones subyacentes a la estructura multidimensional de los datos de mercado son el elemento común sobre el que se identifican y capitalizan las anomalías del mercado. Estos patrones, dados por las medidas de impulso, son los que se utilizan concomitantemente en la construcción de estrategias de negociación y modelos predictivos.

Los estudios empíricos muestran que la estrategia de negociación de impulso, además de ser una de las más utilizadas en la negociación de instrumentos financieros, permite capitalizar los beneficios generados por una de las anomalías más estudiadas del mercado (sobrerreacción y subreacción) [240]. La identificación de pautas asociados a las reacciones erróneas de los inversores debido a los sesgos de comportamiento, además de ser potencialmente predecibles, su uso adecuado permitiría superar al mercado y capitalizar los beneficios.

Los trabajos empíricos han demostrado que la sobrerreacción y la subreacción son un efecto del sesgo de comportamiento que puede producirse tanto a largo como a muy corto plazo [234]. De hecho, la literatura señala que el impulso va seguido de retrocesos, por lo que los inversores pueden reaccionar de forma insuficiente o exagerada ante la publicación de datos fundamentales u otras cuestiones no informativas. Además, estos fenómenos son independientes el uno del otro.

Por otro lado, los avances tecnológicos han propiciado el desarrollo de estrategias de negociación algorítmica en los mercados financieros [232]. La negociación de alta frecuencia, con una precisión extremadamente fina, genera un gran volumen de datos irregulares útiles para comprender mejor la dinámica de los mercados. Las estrategias

de negociación de alta frecuencia influyen en el comportamiento de los precios de los activos y afectan la dinámica de los mercados debido a la velocidad de colocación y ejecución de ordenes de mercado. Este tipo de operativa ha despertado gran interés en la industria financiera, especialmente para los operadores, los reguladores y los responsables de la toma de decisiones [241]. A la luz de los recientes avances en este tema, los resultados indican que durante la apertura y el cierre de los mercados se produce un elevado volumen de posiciones ejecutadas por estrategias de negociación de alta frecuencia. Esto podría llevar, por un lado, a proporcionar liquidez y controlar la volatilidad del mercado en beneficio de los inversores minoristas y, por otro, afectar a las estrategias de negociación de todos los inversores y al funcionamiento del mercado [242, 243]. Aunque todavía se requieren estudios formales sobre estos efectos, el volumen de datos que reportan este tipo de operaciones, además de proporcionar información detallada sobre los activos que pueden resultar atractivos para los inversores y especuladores, podría revelar pautas en los precios que pueden ayudar a predecir a corto plazo la evolución y la dirección de los mercados.

2.6 Previsiones y juicios de valor

El mejoramiento de la eficacia de las decisiones de inversión se ha convertido en un tema de creciente interés en la negociación de instrumentos financieros [244]. La previsión y la construcción de juicios de valor para la toma de decisiones financieras cobra cada vez más importancia en entornos competitivos y mercados en condiciones de incertidumbre. En especial porque se trata de juicios de aprobación o desaprobación elaborados a partir de información limitada y en ambientes bajo presión sobre los cuales se toman decisiones informadas [245]. En esta dirección y basándose en los resultados de la literatura más reciente, esta sección examina el papel de las previsiones en la construcción de juicios de valor. A pesar de los importantes avances en el análisis predictivo, los expertos financieros siguen basando sus decisiones de inversión en juicios de valor, como se discutirá en este documento. En la sección 2.6.1 se analiza el papel de las previsiones en la elaboración de juicios de valor. La sección 2.6.2 explora la influencia de algunos factores en la precisión de las previsiones. En la sección 2.6.3 se analiza la heurística de juicio y el sesgo de previsión en la calidad de las previsiones. La sección 2.6.4 presenta una breve descripción del papel de la información crítica en el desempeño de las previsiones. En la sección 2.6.5 se analizan algunas de las características exigidas por la comunidad inversora en materia de previsiones. En la última parte 2.6.6 se analiza

el horizonte de previsión en la construcción de modelos predictivos.

2.6.1 Juicios de valor

Tradicionalmente, el juicio de valor es una apreciación subjetiva, construida a partir de ciertos criterios y en un contexto determinado, para facilitar la toma de decisiones informadas. La estructura de estos juicios de valor además de que reflejan la experiencia de los analistas, su construcción también puede verse afectada por la naturaleza de la tarea, el uso de información irrelevante, el entorno de previsión y los sesgos cognitivos de los analistas. Así, las previsiones basadas en juicios de valor pueden proporcionar resultados fiables superiores a las previsiones obtenidas únicamente por métodos cuantitativos [99]. Aunque la precisión de estas previsiones también depende en parte de la experiencia del analista, cuando es capaz de adaptarse a las circunstancias cambiantes del entorno de previsión [246, 247], estas mejoran considerablemente cuando el horizonte de previsión se reduce al corto plazo [248, 249].

En los últimos años se han logrado avances significativos en la previsión gracias a la gran potencia de cálculo, la inteligencia artificial y el uso de enormes volúmenes de datos. Sin embargo, las decisiones de inversión basadas en previsiones siguen dependiendo de los juicios de valor de los expertos financieros [250]. Estos juicios de valor, además de ser subjetivos e irreproducibles por los algoritmos, tienen una influencia indirecta (en la definición de los métodos, los datos, los parámetros del modelo y las variables predictoras a utilizar) en los modelos predictivos y su composición refleja la influencia probabilística de sus partes en la estructuración de los juicios de valor [251].

Durante los últimos años, la eficacia de los sistemas de apoyo a la toma de decisiones de inversión también se ha visto impulsada por los avances tecnológicos. Las mejoras se han centrado en el análisis exhaustivo de las variables que rodean el entorno de previsión y negociación [252]. Así como en la identificación de problemas que cambian con rapidez y que no pueden detectarse con facilidad de manera anticipada. Sin embargo, la eficacia de estos enfoques no sólo viene determinada por las habilidades de los expertos financieros (en la construcción de juicios de valor), sino también por el tipo y la calidad de la información (obtenida en tiempo real) que se utiliza en el proceso [253].

La incertidumbre es una de las mayores causantes de errores en la previsión. Además de que impide detectar con claridad las señales de los patrones que se reflejan en los

precios, restringe la posibilidad de emitir juicios de valor sobre los posibles cambios y la evolución futura de las variables financieras. En este contexto, la previsión en condiciones de incertidumbre es una de las mayores dificultades en la formación de juicios de valor para la toma de decisiones de inversión [254, 255].

La construcción de juicios de valor se basa en el uso de predicciones estadísticas numéricas o categóricas y, en cualquier caso, su uso tiene una connotación probabilística [256, 257]. Así, las previsiones generadas por los modelos de clasificación son uno de los diversos componentes de información utilizados en la construcción de juicios de valor. Los trabajos en [258, 259, 260, 261] muestran que la precisión de las previsiones basadas en juicios de valor es superior cuando sus componentes se combinan con previsiones estadísticas. Las estructuras de estos juicios se desglosan en pesos subjetivos que definen implícitamente el grado de creencia del pronóstico en la ocurrencia del evento en cuestión. En consecuencia, estas estructuras, además de ser informativas, son útiles porque reflejan la confianza y la incertidumbre del pronosticador y, a través de las probabilidades establecidas, se define el grado de incertidumbre en la ocurrencia del resultado [262].

La planificación por escenarios es uno de los enfoques más utilizados en la construcción y comunicación de juicios sobre el valor futuro de las variables financieras [263]. La estructura de los juicios de valor se desagrega en sus componentes y la definición de juicios suele realizarse mediante estimaciones probabilísticas por escenarios. Así, los juicios de valor estructurados en función de los escenarios no solo ayudan a comprender las posibles situaciones adversas, ya que anticipan los efectos asociados a los patrones de previsión actuales, sino que también prevén diferentes escenarios con posibles acciones a realizar para abordar el problema en cuestión y con ello reducir la exposición al riesgo [264, 265].

En general, los juicios de valor se han convertido en una valiosa herramienta de previsión a la hora de tomar decisiones informadas. Aunque la precisión de los modelos de previsión ha mejorado en las últimas décadas, impulsada por la inteligencia artificial, las decisiones de inversión basadas en previsiones siguen dependiendo de los juicios de valor de los expertos financieros. De hecho, el escepticismo sobre las predicciones ilegibles basadas en modelos de previsión de *"caja negra"* hace que el papel del juicio humano en la toma de decisiones informadas sea relevante y siga siendo un tema de investigación.

Sin embargo, los nuevos sistemas de negociación algorítmica de instrumentos financieros en los que cada operación de mercado se escala automáticamente en términos de gestión de capital y riesgo. Interpreta y reacciona a los datos del mercado emitiendo órdenes y, por tanto, abre o cierra operaciones de compra o venta. El efecto diferencial de estos enfoques demuestra que en las reglas de decisión se reflejan los juicios de valor derivados de los patrones evaluados en el mercado. En consecuencia, los resultados producidos con estas estrategias de negociación sugieren que el conocimiento y la experiencia relacionados con la inversión se extraen de los patrones del mercado que, cuando se detectan y validan a lo largo del tiempo, producen rendimientos estables y consistentes.

2.6.2 Precisión de la previsión

Los estudios de las últimas décadas han aportado información importante sobre la predicción de la dirección de los tipos de cambio. Aunque su dificultad ha quedado claramente demostrada, se siguen incorporando nuevos enfoques a este desafiante tema que sigue abierto a la investigación [38, 266, 267]. Desde el campo de la inteligencia artificial, el creciente interés por construir modelos predictivos basados en el análisis técnico está ganando protagonismo frente a los enfoques basados en las finanzas conductuales y la teoría de los mercados eficientes [266, 211, 213]. La definición de reglas de decisión en la construcción de estrategias de negociación, para anticiparse a los futuros movimientos de los precios, es también una tarea compleja, incluso con el uso de modelos probabilísticos y métodos de inteligencia artificial. A pesar de ello, algunos estudios, tras validar los patrones detectados en los precios históricos del mercado, han logrado una precisión de clasificación muy prometedora [268, 269, 270]. La investigación existente reconoce el papel crítico de la precisión en las tareas de clasificación y predicción; para más detalles, véase el capítulo **??** sobre el estado del arte.

Los resultados obtenidos en estudios anteriores muestran que las diferencias observadas en la precisión de las previsiones se deben a la volatilidad de los tipos de cambio [271]. Esta variabilidad de los precios, producida por la influencia de diversos factores, suele presentarse con fuertes variaciones cuando el periodo en el que la publicación de datos fundamentales coincide con el periodo en el que se hace la previsión. En este sentido, la publicación de noticias influye en las expectativas del mercado, lo que provoca una gran variabilidad de los precios que es difícil de predecir a favor del tipo

de posición a realizar [234, 272]. Esto ha llevado a muchos operadores a mantenerse al margen de la publicación de noticias durante la negociación.

La eficacia del vínculo entre la precisión de la predicción y la imparcialidad del juicio de valor en la toma de decisiones de inversión es lo que viabiliza la estrategia de negociación. Sin embargo, desde una perspectiva conductivista multivariante, existen muchos factores que pueden incidir en la racionalidad del proceso decisional empleado por los expertos en la negociación. Para mitigar este efecto adverso, se han empleado varias técnicas de inteligencia artificial que emplean métodos multiagente, modelos híbridos y métodos ensemble basados en la votación [273, 274, 275, 276, 277, 278]. Estos enfoques, aunque ayudan a mejorar la precisión de las previsiones, siguen presentando problemas de legibilidad a la hora de construir juicios de valor. De manera que este sigue siendo uno de los mayores desafíos a los que se enfrentan profesionales, inversores e investigadores. Por lo tanto, todavía hay una acentuada necesidad de continuar investigando en la formulación de modelos predictivos parsimoniosos, precisos y altamente interpretativos. En cuanto al horizonte de predicción, la literatura científica ha demostrado que los resultados obtenidos con la aplicación de diferentes enfoques muestran menos discrepancias en la precisión de la predicción en horizontes de tiempo de corto plazo, mientras que las mayores diferencias se producen para los horizontes temporales a largo plazo [279].

En general, los estudios sobre previsión se han centrado en identificar los factores críticos que contribuyen a mejorar la precisión de las previsiones. Entre los elementos más destacados están el entorno de negociación, la experticia del analista, el horizonte de previsión, el uso de información crítica disponible y la elaboración de previsiones oportunas con explicaciones que justifiquen su aplicación. Históricamente, se ha demostrado que las previsiones obtenidas por los expertos son más precisas que las producidas por métodos de series temporales [280]. Estos resultados han confirmado que tales niveles de rendimiento se deben en parte a la experiencia del analista y al uso de heurísticos de juicio en la construcción de las previsiones antes de tomar la decisión de inversión.

2.6.3 Heurística de juicio y sesgo de previsión

Las finanzas conductuales han recibido mucha atención en los últimos años, especialmente con el estudio de los procesos que conducen a la toma de decisiones de inversión

[281]. Aunque los modelos predictivos se basan en el uso de "datos críticos", que han validado históricamente patrones que contribuyen a la consecución de beneficios, esta información fiable no es suficiente para tomar decisiones acertadas. Así, en este proceso intervienen dos elementos que definen la calidad de la previsión, la heurística del juicio y el sesgo de previsión [282, 283, 284]. El primero suele referirse a las reglas que utilizan los expertos financieros para formarse un juicio y tomar una decisión, el segundo corresponde a las discrepancias generadas entre el valor real y el valor previsto. En este sentido, antes de negociar y con información de calidad disponible, los analistas que utilizan métodos heurísticos tratan de formarse juicios sobre el comportamiento histórico y futuro del mercado. Luego, a fin de tomar una decisión de inversión favorable, entran en el mercado con la menor exposición al riesgo para reducir el sesgo de previsión [285, 286].

Desde el punto de vista psicológico, se han publicado muchos estudios que analizan el comportamiento de los expertos financieros y el uso de la heurística de juicio en la toma de decisiones de inversión [287, 116, 288, 170]. Sin embargo, siguen existiendo algunas cuestiones críticas sobre la forma en que utilizan un conjunto de *"reglas sencillas"* en la estructuración de los juicios de valor para la toma de decisiones informadas. Los hallazgos relativos a la heurística de juicio en estudios recientes han demostrado que el uso de estos *"atajos mentales"* hace que el analista se centre en algunas cuestiones importantes sobre el problema y deje de lado aspectos menos relevantes [289, 170]. Esto puede significar que, en la construcción de juicios de valor, las previsiones puras, obtenidas por métodos cualitativos o cuantitativos, ya sea individualmente o en grupo, pueden ser utilizadas directamente o ajustadas bajo el criterio y la experiencia del analista antes de tomar la decisión. Aunque los juicios de valor pueden asumir diversas tipologías (previsión numérica sobre una variable financiera, previsión por rangos y previsión de un evento) y sus estimaciones pueden ir acompañadas o no de probabilidades, la incertidumbre del mercado y la volatilidad de los precios pueden conducir al sesgo de previsión basado en conclusiones erróneas extraídas de los juicios de valor [290].

Teniendo en cuenta este contexto, se puede argumentar que en los enfoques de previsión de variables financieras que utilizan la heurística de juicio, el sesgo de previsión siempre estará presente. Más aún en entornos en los que la ambigüedad e incertidumbre de las señales generadas por los movimientos de los precios puede llevar a decisiones erróneas con un alto nivel de riesgo. Por otro lado, es la presión del entorno la que

fomenta el uso de la heurística de juicio, lo que hace que el experto financiero siempre este expuesto al sesgo de previsión [291]. En este sentido, la construcción de modelos predictivos debe tener en cuenta la naturaleza competitiva del entorno en el que se realiza la previsión [292]. Esto significa que, durante la negociación, las métricas utilizadas por el modelo de predicción no sólo deben confirmar y predecir los patrones formados en los precios del mercado, sino también explicar la dirección del movimiento futuro [60, 85]. Por lo tanto, la rapidez con la que se construye y se aborda el juicio de valor en la toma de decisiones de inversión viene determinada por la eficacia del modelo de previsión, la estructura del juicio de valor, la información de contexto, la experiencia y la heurística de juicio empleada por el agente que participa en la negociación.

2.6.4 Datos de mercado e información de contexto

En la era digital, el interés por mejorar la eficiencia en el uso de la información ha crecido rápidamente en los últimos años. La identificación de información crítica para predecir la dirección de los precios de los activos financieros a partir de grandes volúmenes de datos, no estructurados y que se producen constantemente a un ritmo acelerado, se ha convertido en una de las tareas más desafiantes para los participantes del mercado [293, 294, 295, 296]. La preparación de datos también está fomentando el desarrollo y la aplicación de nuevos métodos que facilitan su tratamiento, análisis y visualización. Así, los analistas que buscan mejores resultados financieros en sus operaciones están basando sus decisiones en un enfoque de intervención más analítico y basado en datos [74, 297]. Esto los ha llevado a prestar más atención al contexto económico, a una revisión exhaustiva de la información, a preferir el uso de técnicas de previsión legibles y con gran capacidad interpretativa y explicativa del fenómeno en cuestión, y a evaluar su posible desempeño e impacto.

Tradicionalmente, en la fase de negociación, en lo que se refiere a la construcción de estructuras de juicio de valor para la toma de decisiones de inversión, los expertos financieros se apoyan en el uso de distintas fuentes de información [298]. Su uso varía en función de las condiciones del contexto y de la heurística de juicio empleada por el analista. En la práctica, el experto financiero asigna pesos a cada uno de estos componentes para destacar la información más relevante a la hora de construir la previsión y así poder tomar la mejor decisión [299]. Por otro lado, la incertidumbre de las previsiones, a la hora de tomar decisiones de inversión, puede llevar a los expertos financieros

a confiar en datos e información de contexto, incluso sobre bases cuestionables. Sin embargo, en cuanto a la forma de realizar la previsión, son decisivos la condición del mercado y el contexto en el que se realizará la inversión. En consecuencia, el ejercicio de previsión puede basarse en la construcción de juicios de valor, en el ajuste del valor esperado mediante juicios de valor, o incluso puede basarse en el uso de información adicional. En cualquier caso, el experto financiero elegirá teóricamente el escenario más adecuado para realizar el análisis antes de tomar la decisión [300].

En los últimos años, también ha habido un gran interés por mejorar la precisión y el desempeño de las previsiones obtenidas con los modelos de apoyo a la toma de decisiones de inversión. Los estudios sugieren que la precisión de las previsiones viene determinada por la idoneidad y la calidad de la información de contexto utilizada en el proceso. Mientras que las discrepancias entre los valores de las previsiones y los datos reales se deben al uso de información irrelevante y al acceso limitado a información actualizada [301, 302]. En estas condiciones, la precisión alcanzada por los juicios de valor, debido al uso eficiente de información crítica (específica), es superior a las precisiones producidas con el uso de información dispersa (no restringida). Sin embargo, la influencia que tiene la información actualizada durante la negociación puede afectar la precisión de las previsiones. Algunos analistas técnicos prefieren operar al margen de los acontecimientos que puedan producirse durante la negociación. Esto se debe a la volatilidad que se produce en los precios de los activos debido a la influencia de la publicación de datos fundamentales [303, 304].

En general, se ha destacado que la eficacia de las decisiones de inversión viene determinada, en parte, por la calidad de la información utilizada en la previsión y la construcción de juicios de valor. Sin embargo, se ha prestado poca atención a la identificación de la información crítica que debe utilizarse en estos procesos para garantizar el mejor resultado. El uso de información crítica para construir estos modelos es lo que permite la automatización de procesos complejos para mejorar la toma de decisiones. Así, la información crítica juega un papel fundamental en la construcción de modelos predictivos y juicios de valor a la hora de tomar decisiones informadas. Debido al enorme volumen de datos y a la capacidad de procesamiento, así como a los métodos estadísticos y de inteligencia artificial disponibles en la actualidad, es preciso descubrir la información crítica del mercado para mejorar el desempeño de las previsiones y de los modelos de negociación.

2.6.5 Comunidad inversora y previsiones de expertos

En la comunidad inversora, se sabe que los expertos financieros son capaces de aprovechar los retos globales y las futuras tendencias del mercado para convertirlos en oportunidades. El uso de un enfoque prospectivo basado en la evaluación de riesgos y el establecimiento de planes de acción son elementos clave para generar confianza en la comunidad. Estas buenas prácticas generan confianza en las previsiones elaboradas y comunicadas por los expertos. La forma en que se lleva a cabo su comunicación y la manera en que los usuarios entienden estas previsiones tiene notables implicaciones en la toma de las mejores decisiones de inversión [305, 306].

La construcción de modelos eficaces de previsión, además de requerir una profunda comprensión del funcionamiento del mercado y de los procesos heurísticos que los expertos utilizan para construir juicios de valor, tiene importantes implicaciones para la toma de decisiones de inversión. Hay que tener en cuenta que la naturaleza incierta y volátil de los mercados hace que la utilización de juicios de valor, a pesar de su carácter subjetivo, desempeñe un papel primordial en la construcción de las previsiones del mercado. Incluso hay algunos participantes del comercio que atribuyen más importancia y confianza a las previsiones generadas por las técnicas de predicción que a los juicios de valor emitidos por los propios expertos [307, 308].

La definición de las reglas de negociación es otro ámbito de investigación, bastante inquietante dentro de la comunidad inversora, para la construcción de sistemas automáticos de negociación. El creciente interés por la formulación de reglas de negociación, con un carácter mucho más elaborado, se ha centrado en la forma en que los expertos financieros formulan sus previsiones. La principal ventaja de incorporar en las reglas de decisión las previsiones basadas en la construcción de juicios de valor radica en el hecho mismo de incorporar la experiencia del negociador, que es en definitiva el activo más valioso a la hora de negociar en los mercados financieros.

El análisis de los resultados producidos con la previsión tras la negociación permite evaluar y mejorar la eficacia de los procesos que van desde la construcción de juicios de valor y sus componentes hasta la toma de decisiones de inversión. Incluso en este contexto dinámico de negociación, el valor de la información crítica utilizada en dichos procesos puede cambiar con el tiempo. Además, la forma en que el experto financiero comunica y divulga las previsiones a la comunidad de inversores y usuarios es crucial en un contexto en el que la desinformación está muy extendida. Así, una

comunicación eficaz debe garantizar la veracidad y claridad de las previsiones, mucho más en contextos dinámicos en los que los canales de comunicación deben facilitar la retroalimentación.

En general, los usuarios de las previsiones económicas y financieras tienden a ajustarlas o utilizarlas directamente en función de sus expectativas y percepciones de calidad sobre estas. Esto significa que, además de requerir de previsiones oportunas y precisas, su uso debe estar claramente justificado. Así, los costes de estos servicios son indiferentes a los intereses de los usuarios. Por otra parte, hay que señalar que cuando las previsiones carecen de explicaciones justificables, los usuarios expertos, utilizando la heurística de juicio, ajustan y aplican estas previsiones basándose en sus conocimientos. No es el caso cuando estas previsiones provienen de fuentes fiables y gozan de explicaciones justificadas basadas en los fundamentos.

2.6.6 Horizonte de tiempo en las previsiones

Tradicionalmente, el horizonte de previsión ha sido una cuestión crítica que siempre se ha tenido en cuenta en la construcción de modelos predictivos de variables financieras [309, 310]. Esta variable decisiva tiene una gran influencia en la precisión y el desempeño en las tareas de clasificación de los modelos financieros. Existe una cantidad considerable de literatura que analiza la influencia que tiene el horizonte de previsión en el desempeño de las predicciones [311]. En estos estudios, los modelos predictivos han mostrado un mayor desempeño en términos de precisión en horizontes de previsión a corto plazo, mientras que se observa un efecto contrario en las previsiones a largo plazo, lo que hace que su uso sea menos frecuente [312, 313, 314]. Del mismo modo, el interés del especulador por obtener beneficios con la mínima exposición al riesgo está más orientado a la negociación a corto plazo que a la de largo plazo.

La heterogeneidad de los horizontes temporales en los que se negocian los activos financieros ha llevado a evaluar el tipo y la eficacia de la información crítica que debe utilizarse en la previsión y la toma de decisiones de inversión. Si el enfoque de la negociación es a largo plazo, la atención de los expertos financieros se centrará en el análisis y la previsión de las variables financieras a partir de los datos fundamentales. Por el contrario, si los operadores buscan beneficiarse de los movimientos del mercado a corto plazo, centrarán su atención en el uso de patrones y datos de mercado para ejercer y aprovechar las previsiones a corto plazo [315, 316].

En general, en la búsqueda de beneficios, el análisis del momentum en los precios de los activos ha llevado a los inversores a identificar puntos de entrada y salida en las tendencias del mercado. Estos puntos de referencia, basados en patrones de precios, ayudan a capitalizar los movimientos producidos por los sesgos de subreacción y sobrerreacción de los inversores. La precisión en la predicción de estos movimientos direccionales ha demostrado ser eficaz cuando los horizontes de previsión son a corto plazo. Esto se debe a que la precisión de la predicción está inmersa en el horizonte temporal en el que ocurren estos sesgos. Una vez que se produce el ajuste de los precios, puede dar lugar a cambios de dirección que podrían ir en contra de las previsiones. Por lo tanto, la eficacia de los modelos de previsión y de los sistemas de negociación dependen de la correcta identificación de los puntos de inflexión del mercado, así como de la determinación de la duración de los sesgos de subreacción y sobrerreacción para poder capitalizarlos.

3. Estado del arte

3.1 Introducción

Predecir la tendencia de los precios de los activos para mejorar la eficacia de las negociaciones se ha convertido en un tema de creciente interés para los participantes del mercado [317]. En un esfuerzo por anticiparse a los movimientos del mercado y beneficiarse de la volatilidad de los precios, se percibe un renovado empeño por construir modelos de clasificación más precisos. La Figura 3.1 muestra en términos generales el procedimiento utilizado, según la bibliografía más reciente revisada, para construir modelos de clasificación con fines predictivos [74].

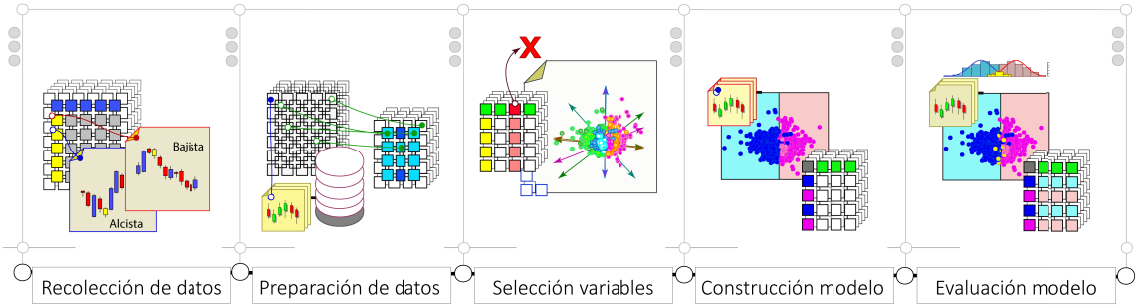


Figura 3.1: Esquema general para predecir la tendencia de los precios de los activos. Marco general utilizado en la construcción de modelos de predicción. El flujo de trabajo descrito por diferentes autores se resume en: (1) Recolección de datos y medición de variables, (2) Preprocesamiento de datos. (2.1) Preparación y disposición de datos. (2.2) Selección y/o extracción de características. (3) Elección y construcción del modelo de clasificación, y (4) Evaluación del rendimiento de la predicción.

Según la revisión bibliográfica, la construcción de modelos de clasificación consta al menos de cuatro etapas fundamentales: (**1**) Recolección de datos y medición de variables, (**2**) Preprocesamiento de datos. En esta etapa se llevan a cabo dos actividades esenciales (**2.1**) Preparación y disposición de datos, y (**2.2**) Selección y/o extracción de características). Le siguen (**3**) Elección y construcción del modelo de clasificación, y (**4**) Evaluación del rendimiento de la predicción.

En la primera fase, la identificación y recopilación de datos puede realizarse desde tres ámbitos. En primer lugar, tomando información del mercado (precios de apertura, máximos, mínimos, cierres y volumen) y obteniendo señales de ellos con fines analíticos y predictivos mediante el cálculo de indicadores técnicos (para distintos marcos temporales). En segundo lugar, recopilando datos e informes macroeconómicos (para el análisis fundamental) o extrayendo noticias financieras publicadas en las redes sociales. En tercer lugar, recopilando y combinando datos de análisis técnico y fundamental o una combinación de todos ellos. En cualquiera de estos ámbitos, como se analiza con más detalle en la Sección 3.2, la recopilación y medición de variables con información de mercado, técnica y fundamental es crucial para predecir los precios y las tendencias de los activos. Esto hace que el proceso de toma de decisiones sea una tarea difícil.

En la segunda fase, los datos se preparan de forma que se garantice su suficiencia para la correcta identificación de patrones y cumplan los requisitos de calidad exigidos por los métodos de clasificación. Esta etapa de preprocesamiento, que se analiza con más detalle en la Sección 3.3, además de las actividades de preparación y limpieza que garantizan la conformidad de los datos para su uso, puede tomar dos caminos: la selección de características que producen el mejor rendimiento de clasificación o la extracción de nuevas características (combinaciones lineales) que se utilizarán como variables predictoras. En cualquiera de estos casos, la elección se realiza a favor del subconjunto de atributos que ofrecen el mejor desempeño en la clasificación.

En la tercera fase, tras identificar y seleccionar el mejor método de clasificación, generalmente se entrena el modelo y se valida de forma cruzada con una muestra de entrenamiento y otra de evaluación. En la Sección 3.4 se revisan los métodos de clasificación utilizados en la predicción de la tendencia del precio de los activos, con especial énfasis en el uso del análisis discriminante. Además, la Sección 3.4.6 aborda el problema de la definición de hiperparámetros y su influencia con el análisis discriminante lineal.

Por último, en la fase final, tal y como se muestra en la Figura 3.1, la revisión bibliográfica muestra que tras definir las métricas que medirán objetivamente el comportamiento predictivo del modelo, se evalúa el rendimiento alcanzado en las tareas de clasificación. La Sección 3.5 aborda esta importante cuestión, en la que algunos estudios utilizan datos fuera de muestra para medir y verificar este desempeño y garantizar la fiabilidad del modelo propuesto. Dada esta estructura, la sección 3.6 conduce a consideraciones importantes a tener en cuenta en la construcción de modelos de clasificación (para predecir, a corto plazo, la dirección del movimiento del tipo de cambio

euro-dólar).

3.2 Selección de mediciones y recolección de datos

La correcta selección de los datos y métricas a utilizar es decisiva en la construcción de modelos de clasificación más precisos. Los datos y el tipo de métricas por elegir dependerán en parte del grado de conocimiento que se tiene del activo y del mercado, así como de la influencia que dichas métricas ejercen sobre la acción del precio para predecir su evolución futura.

Tradicionalmente, el principal reto al que se han enfrentado los investigadores es el desconocimiento de la capacidad discriminante y predictiva de la multiplicidad de variables que potencialmente pueden utilizarse para diferenciar y predecir la dirección de los precios. Los estudios realizados en las últimas décadas han proporcionado información importante sobre el tipo de variables de entrada utilizadas en la formulación de modelos de clasificación [318].

La Tabla 3.1 resume las variables de entrada comúnmente utilizadas en la construcción de modelos de clasificación con fines predictivos. Entre los elementos más importantes se encuentran el instrumento financiero abordado, el tipo de información utilizada como variables de entrada, el horizonte de análisis de la muestra de estudio, el alcance de la predicción y el método empleado.

Entre los tipos de información utilizados como variables de entrada se encuentran: información de mercado [319], indicadores técnicos [317], indicadores económicos [55, 320] y noticias financieras [321, 322] extraídas de publicaciones en redes sociales y el uso combinado de estos datos.

Algunos modelos predictivos se construyen únicamente a partir de datos históricos del mercado (**OHLCV**), es decir, precios de apertura (O), máximos (H), mínimos (L), cierre (C) [323] y volumen (V). Aunque esta información no es suficiente, puede ser útil para modelar el cambio de precios (en la plataforma de negociación).

Algunos trabajos en [324] han sugerido que el uso limitado de esta información puede generar resultados inexactos y proyectar una idea falsa de la presunta eficacia del modelo. Sin embargo, el uso de estos datos en periodos históricos más pequeños, marcos temporales de 1 minuto o en *ticks*, dada la precisión de las fluctuaciones, los cambios en la definición del precio del activo se emulan con mayor exactitud, lo que significa que las pruebas y los análisis son más auténticos y, por tanto, precisos.

Tabla 3.1: Datos de entrada.

Autor	[a] Símbolo	[b] Datos entrada	Muestra
Das et al.(2022) [325]	GBP/INR, AUD/INR*, USD/INR	IM, IT	2006 - 2021
Dash et al.(2022) [326]	USD/EUR, AUD/JPY, CHF/INR	IM, IT, ME	2015 - 2021
Ortu et al.(2022) [324]	BTC/USD, ETH/USD*	IM, IT, RS	2017 - 2021
Hao et al.(2022) [315]	BRN, WTI*	IM	2017 - 2020
Padhi et al.(2022) [327]	DHR, NVO, UNH, DPHL, AFX, MRK, CVS, EWL, MTO, TN8	IM, IT	1987 - 2021
Ozer et al.(2022) [328]	BTC/USD, ETH/USD, LTC/USD	IM, IF, IT	2017 - 2021
Sadeghi et al.(2021) [329]	EUR/USD	IM, IT	2014 - 2019
Uras et al.(2021) [330]	BTC/USD	IM, IT, IE	2015 - 2019
Ampomah et al.(2020) [331]	BAC, XOM, MSFT, KMX, TATASTEEL, HCLTECH, S&P 500, DJIA	IM, IT,	2005 - 2019
Kwon et al.(2019) [332]	BTC/USD, ETH/USD, XRP/USD, BCH/USD, LTC/USD, DASH/USD, ETC/USD, KRW/USD	IM	2017 - 2018
Fischer et al.(2018) [333]	S&P 500	IM	1992 - 2015
Hu et al.(2018) [334]	S&P 500, DJIA	IM, RS	2010 - 2015
Bustos et al.(2017) [335]	COLCAP Index	IT	2010 - 2017
Chakraborty et al.(2017) [336]	DJIA index, AAPL	RS	2016 - 2016
Coyne et al.(2017) [337]	JNJ*, JPM, AMZN, NFLX, BAC, INTC, AAPL, TSLA, GS	RS	2016 - 2017
Dingli et al.(2017) [338]	S&P 500, DJIA, NASDAQ-100	IM, IT, IE	2003 - 2016
Huang et al.(2017) [339]	TWSE	IM	2015 - 2016
Dang et al.(2016) [340]	VN30 Index	RS	2014 - 2015
Di et al.(2016) [341]	S&P 500	IM	1950 - 2016
Ghanavati et al.(2016) [342]	HSCI	IT, NF	2014 - 2015
Chai et al.(2015) [343]	CSI 300	IE	2009 - 2012
Dash et al.(2015) [344]	BSE SENSEX*, S&P 500*	IT	2010 - 2014
Feuerriegel et al.(2015) [345]	Valores de empresas alemanas	NF	2004 - 2011
Gonzalez et al.(2015) [346]	Ibovespa*, S&P 500, DJIA, FTSE 100, DAX, NIKKEI 225, HANG SENG, USD, EUR, CNY	IM, IT, IE	1989 - 1998

[a] **Índices bursátiles:** HSCI Índice Compuesto Hang Seng, China. Ibovespa Índice de la bolsa de valores de Sao Paulo, Brasil. TWSE Índice de la bolsa de valores de Taiwan. CSI 300 Índice de valores de China. DJIA Índex Dow Jones Industrial Average. VN30 Índice de la bolsa de valores de Vietnam. BSE SENSEX Índice de la Bolsa de Valores de Bombay, India. S&P 500 Índice Standard & Poor's 500. NASDAQ-100 índice bursátil de compañías tecnológicas de Estados Unidos. FTSE 100 Índice bursátil de la Bolsa de Valores de Londres, Reino Unido. DAX Índice de la Bolsa de valores de Fráncfort, Alemania. NIKKEI 225 Índice bursátil de la Bolsa de valores de Tokio, Japón. **Tipos de cambio:** EUR/USD par euro/dólar americano. GBP/INR par libra esterlina/rupia india. AUD/INR par dólar australiano/rupia india. USD/INR par dólar americano/rupia india. AUD/JPY par dólar australiano/yen japonés. CHF/INR par franco suizo/rupia india. USD/CNY par dólar americano/yuan chino. **Criptodivisas:** BTC/USD par bitcoin/dólar americano. ETH/USD par ethereum/dólar americano. LTC/USD par litecoin/dólar americano. XRP/USD par ledger/dólar americano. BCH/USD par bitcoin cash/dólar americano. DASH/USD par dash/dólar americano. ETC/USD par ethereum classic/dólar americano. KRW/USD par KRW/dólar americano. **Materias primas (Commodities):** BRN Petróleo crudo Brent. WTI Petróleo crudo WTI. **Valores empresas estadounidenses:** AAPL Apple Inc. JNJ Johnson & Johnson. JPM JPMorgan Chase & Co. AMZN Amazon. NFLX Netflix, Inc. BAC Bank of America Corp. INTC Intel Corporation. TSLA Tesla, Inc. GS Goldman Sachs Group Inc. **Valores empresas varios países:** DHR Danaher Corporation. NVO Novo Nordisk A/S. UNH UnitedHealth Group Incorporated. DPHL Dechra Pharmaceuticals PLC. AFX Alpha FX Group plc. MRK Merck & Co. CVS CVS Health Corporation. EWL iShares. MTO Mitie Group plc. TN8 Thermo Fisher Scientific Inc. BAC Bank of America Corporation. XOM Exxon Mobil Corporation. MSFT Microsoft Corporation. KMX CarMax, Inc. TATASTEEL Tata Steel Limited. HCL TECH Technologies Ltd. *Instrumento financiero con mejores resultados de previsión. Los métodos de clasificación y los valores de exactitud (Accuracy) se muestran en las Tablas 3.2 y 3.3 respectivamente.

[b] **Información sobre datos de entrada:** IM Datos de mercado (OHLCV). IT Indicadores técnicos. IE Indicadores económicos. ME Medidas estadísticas. RS Redes sociales (Indicadores sociales, Twits, Google trends). NF Noticias financieras.

Varios estudios han utilizado, como variables de entrada, tanto los indicadores técnicos como variables fundamentales [323]. Mientras que los primeros están asociados con la identificación de pautas en los precios, los segundos reflejan el estado del país para el que se calculan. Así, describen el nivel de actividad de las mayores economías mundiales (USA, Unión Europea, Reino Unido, Japón, Alemania, Australia, Canadá, Suiza, China, Nueva Zelanda, entre otras). Estos indicadores también han demostrado su eficacia en la determinación del valor y la salud de las empresas que cotizan en los mercados financieros. Esto hace que esta información sea especialmente útil en la negociación de instrumentos de renta variable, renta fija, índices bursátiles, tipos de cambio, productos derivados, futuros, entre otros.

Los estudios presentados en [347, 348] utilizan datos de mercado (OHLCV), incluidos los precios y el volumen en el cálculo de indicadores técnicos. Estos valores obtenidos a lo largo de varios periodos y en un marco temporal específico se utilizan como datos de entrada para predecir la tendencia de los precios. Según la literatura científica, hay muchos estudios empíricos que se centran en marcos temporales diarios con los que coincide el horizonte de previsión [321, 349]. Por otro lado, basándose en las señales proporcionadas por estos indicadores técnicos, las estrategias y modelos de negociación toman decisiones sobre cómo y cuándo abrir o cerrar una posición en el mercado. En los modelos predictivos los indicadores técnicos también se utilizan para predecir los cambios futuros en los precios de los activos. En cualquiera de estos ámbitos, el uso de indicadores técnicos ha demostrado ser una herramienta valiosa que ayuda a simplificar el proceso de toma de decisiones del analista para anticiparse al mercado y mejorar los resultados de las negociaciones.

Entre los indicadores técnicos, los indicadores de tendencia [350] y los osciladores [351] han sido los más utilizados. El primer grupo de indicadores ha demostrado su eficacia para identificar la dirección de los precios, detectando con retraso los puntos de inflexión del mercado. Los osciladores, por su parte, a diferencia del primer grupo, ayudan a identificar dichos puntos de inflexión con antelación [352]. Otras categorías que también han demostrado ser muy utilizadas son los indicadores basados en el volumen y los indicadores Bill Williams. En general, estos indicadores se han convertido en herramientas útiles que ayudan a analizar la dinámica de los precios para mejorar la toma de decisiones en la negociación de activos.

Entre los indicadores de tendencia, los más utilizados están: SAR Parabolico [353], Momentum, como oscilador de seguimiento de tendencia [354], Media Móvil de Convergencia/Divergencia (MACD) [355], Media Móvil Adaptiva (AMA), "Kaufman's Adap-

tive moving average" [350], Índice Direccional Medio (ADX), "Average Directional movement Index" [325], Bandas de Bollinger (BB), "Bollinger Bands" [356], Media Móvil Exponencial Doble (DEMA), "Double Exponential Moving Average" [357] Envelopes [358], Media Móvil Adaptiva Fractal (FRAMA), "Fractal Adaptive Moving Average" [359], Ichimoku Kinko Hyo [360], Media Movil Simple (SMA) [350], Media Movil Exponencial (EMA), "Exponential Moving Average" [361], Media Movil Suavizada (SMMA), "Smoothed Moving Average" [362], Media Movil Ponderada Lineal (LWMA), "Linear Weighted Moving Average" [363], Desviación Estándar (SD) [364], Media Móvil Exponencial Tripe (TEMA), "Triple Exponential Moving Average" [365], Índice Variable de Media Dinámica (VIDYA), "Variable Index Dynamic Average" [366],

Entre los osciladores están: Momentum, como indicador adelantado (se mueve muy por delante de los precios) [367], Rango Verdadero Medio (ATR), "Average True Range" [357], Fuerza de los osos (BP), "Bears Power" [368], Fuerza de los toros (BP), "Bulls Power" [369], Oscilador de Chaikin (CHO), "Chaikin Oscillator" [370], índice de Canal de Mercancías (CCI) "Commodity Channel Index" [350], Indicador Demarker (DeM), "DeMarker" [369], índice de Fuerza (FRC), "Force Index" [371], Media Móvil de Convergencia/Divergencia (MACD) [372], Oscilador de promedio móvil (OsMA), "Moving Average of Oscillator" [373], Índice de Fuerza Relativa (RSI), "Relative Strength Index" [374], Índice de Vigor Relativo (RVI), "Relative Vigor Index" [375], Oscilador estocástico (Stochastic), "Stochastic Oscillator" [376], Media Exponencial Triple (TRIX), "Triple Exponential Average" [377], Rango porcentual de Williams (%R), "Williams' Percent Range Technical Indicator" [378],

Entre los indicadores de volumen están: Acumulación/Distribución (A/D), "Accumulation/Distribution Technical Indicator" [379], Índice de Flujo Monetario (MFI), "Money Flow Index" [380], Volumen en Balance (OBV), "On Balance Volume" [354], Volúmenes, "Volumes" [381],

Entre los indicadores Bill Williams se destacan: Oscilador Acelerador, "Accelerator Oscillator" [382], Alligator [383], Oscilador Asombroso (AO), "Bill Williams's Awesome Oscillator" [384], Fractales, "Fractals" [385], Oscilador Gator, "Gator Oscillator" [386], Índice de Facilidad del Mercado (MFI), "Market Facilitation Index" [387],

Por otra parte, el uso de datos fundamentales y de información macroeconómica se ha extendido entre la comunidad fundamentalista desde que se publicó la hipótesis de los mercados eficientes. Varios estudios han evaluado la influencia que tiene el uso de datos fundamentales como variables de entrada en la construcción de mode-

los de predicción [388]. La publicación de noticias sobre indicadores macroeconómicos tiene impactos significativos en la volatilidad de los precios. Sin embargo, la literatura científica no muestra evidencia de estudios que analicen la duración del efecto de la publicación de datos fundamentales sobre los movimientos de los precios [55]. Estos vacíos señalan la posibilidad de evaluar el impacto que tiene en el movimiento de los precios de los activos en el tiempo, la publicación de los diferentes indicadores macroeconómicos. Esto permitiría identificar los datos fundamentales que mejor influyen en el precio y utilizarlos en la construcción de modelos predictivos. El uso combinado de distintos tipos de indicadores como variables de entrada en la formulación de modelos también ha aumentado considerablemente en los últimos años. Los casos más usuales incluyen datos de mercado e indicadores técnicos [389]. Estudios con un enfoque más amplio integran datos de mercado, indicadores técnicos y datos fundamentales [390].

Otras variantes muy utilizadas en investigaciones que utilizan el análisis de sentimiento para predecir la tendencia de los activos consideran la extracción de información financiera y económica de las publicaciones en redes sociales [391]. Aunque los resultados obtenidos son prometedores todavía es evidente el desconocimiento de las variables que efectivamente influyen desde el punto de vista técnico y fundamental en el comportamiento de los precios. Esta condición tradicionalmente se ha conservado en la mayoría de las publicaciones, en donde un método simplemente se alimenta de una multiplicidad de datos en espera de encontrar las variables que además de explicar su influencia sobre la acción del precio permita predecir la dirección de los precios de los instrumentos financieros.

Al observar en la Tabla 3.1 el intervalo de fechas de los datos de entrada utilizados en la formulación del modelo, se refleja implícitamente una idea de la eficacia presunta del modelo. Cuanto mayor sea el horizonte de tiempo y el número de observaciones históricas consideradas en la formulación de modelos estos serán más consistentes y precisos. Otra posibilidad es la extracción de una muestra de datos representativa cuyos resultados puedan ser evaluados y generalizados con datos fuera de la muestra. Estos intervalos de datos se pueden emplear de manera aleatoria o predefinida para la formulación y evaluación cruzada de la eficacia del modelo.

En general, los datos de entrada utilizados en la formulación de modelos de clasificación, además de reflejar el grado de conocimiento que se tiene sobre el activo y el mercado, son determinantes para la precisión, legibilidad e interpretabilidad del modelo. La eficacia de la elección depende de la capacidad de identificar y seleccionar las variables discriminantes que mejoran el rendimiento de la clasificación.

3.3 Preprocesamiento de datos

Los modelos de predicción de la dirección de los precios de los activos son una valiosa herramienta para realizar juicios de valor y tomar decisiones de inversión. La eficacia de estos modelos está determinada en parte por la calidad de los datos, los métodos de preprocesamiento y las actividades de preparación que los hacen útiles para el análisis. Aunque la introducción de la minería de datos y el aprendizaje automático también ha contribuido al progreso de las técnicas de preprocesamiento de datos, aún existen algunos interrogantes por responder. Uno de ellos consiste en determinar la verdadera influencia que tiene la aplicación de estos métodos, dada la naturaleza y la calidad de los datos, para que resulten útiles y se ajusten a los requisitos de las técnicas de clasificación. En este sentido, la sección 3.3.1 describe brevemente el efecto de la aplicación de técnicas de preparación de datos en la construcción de modelos de clasificación. La sección 3.3.2 aborda el problema de la selección de características en la construcción de modelos de clasificación. Por último, la sección 3.3.3 describe brevemente la extracción de características, aunque sin profundizar demasiado porque no es de interés para los objetivos de este estudio.

3.3.1 Preparación y disposición de datos

Tradicionalmente, la construcción de modelos predictivos eficaces ha estado influenciada por la calidad de los datos utilizados en su formulación. En términos generales, esto ha llevado no sólo a cuestionar la idoneidad de los datos en la construcción de los modelos, sino también a detectar si son adecuados para el fin para el que se van a utilizar. Los últimos avances han demostrado que la disponibilidad de datos como variables de entrada en la formulación de modelos no sólo debe ser suficiente en cantidad, sino también en calidad, utilidad y estructura. Además de ser adecuados según los requisitos de la técnica y la finalidad para la que se utilicen. En este sentido, la preparación de datos se ha convertido en un paso obligatorio, mediante el cual los "datos inútiles" se transforman en "datos útiles" para que los métodos no sólo funcionen, sino que produzcan resultados de calidad. Una revisión de la literatura muestra que el uso de actividades de preparación y ajuste de datos contribuye a la construcción de modelos más precisos. Entre las actividades más comunes que se comentan a continuación se encuentran la conservación de datos, las actividades de preparación de datos (recuperación, limpieza, eliminación de registros por falta de datos, corrección

de datos, discretización y almacenamiento de datos), la normalización de datos y, por último, el equilibrado de clases no balanceadas.

A. Conservación de datos

En la literatura se han reportado muchas actividades de preparación de datos, especialmente como preámbulo a la construcción de modelos predictivos. Entre las actividades más importantes están las que mejoran la calidad de los datos, asegurando su suficiencia y el cumplimiento de las especificidades y requisitos de calidad exigidos por los métodos [355, 392, 331]. Sin embargo, hay que tener en cuenta que la calidad es una condición inherente a los datos que se materializa y se preserva desde el momento en que se conciben, producen y almacenan. Por otro lado, cuando se garantiza la seguridad de los datos y se restringe el uso de información redundante, los riesgos de inconsistencia y pérdida de integridad de los datos son prácticamente reducidos, lo que implica, una vez construido el modelo, que la producción de resultados sea con altos niveles de calidad. No obstante, los autores en [393] señalan que la calidad de los datos de entrada mejora con la extracción de nuevos atributos a partir del espacio de características original.

B. Actividades de preparación

Algunas rutinas de preparación de datos, aunque son muy comunes en la mayoría de los casos, hacen que las colecciones de datos sean más eficientes para el análisis. Entre las acciones más frecuentes, que se aplican a los datos en bruto, están la recuperación, la limpieza, la eliminación de datos inútiles o de registros con valores perdidos, la corrección de datos erróneos, el tratamiento de datos atípicos, la discretización y el almacenamiento de datos [394, 395, 355, 396, 397]. Además, uno de los problemas más comunes y ampliamente reportados es la presencia de ruido en los datos de entrada [398, 329, 396]. Entre las intervenciones más destacadas están la incorporación de nuevos atributos (formados por la media y la desviación estándar de los datos) [329] y la transformada de Wavelet [396].

C. Normalización de datos

La normalización consiste en ajustar los valores de las mediciones dispuestas en diferentes escalas de medida a una escala común. Al examinar las especificidades de los

modelos predictivos, tanto las actividades de preparación como los métodos de preprocesamiento responden a la naturaleza de los datos para hacer viable su aplicación. Trabajos recientes han demostrado que la normalización evita la influencia de variables con escalas de medición superiores. En consecuencia, el uso de estos procedimientos se ha hecho obligatorio para algunas técnicas cuando se buscan resultados más precisos. Los enfoques más recientes, basados en el aprendizaje profundo (DL) [399, 395], los métodos de conjunto [396, 400, 401, 331, 402], los modelos híbridos [355], las máquinas de vectores de soporte (SVM) [372], las redes neuronales artificiales (ANN) [403] y las redes neuronales convolucionales (CNN) [392], hacen uso del escalado y la normalización de los datos de entrada. Los estudios empíricos han demostrado que la normalización de los datos no solo ayuda a mitigar la influencia de las variables con valores de mayor magnitud, sino que también mejora el aprendizaje durante el entrenamiento y mejora la precisión de las predicciones con los algoritmos de aprendizaje automático. En la extracción de características, para evitar que las primeras combinaciones lineales retengan la mayor variabilidad debido a las diferencias de escala, los datos se normalizan [404].

D. Equilibrio de clases desbalanceadas

Uno de los problemas más recurrentes en el uso de métodos de aprendizaje supervisado es el desequilibrio de instancias dentro de los grupos [394, 405]. En general, el análisis de eventos con observaciones desiguales se ha planteado como un problema de clasificación multiclase con datos desbalanceados [406]. Dada la complejidad de los datos, el desequilibrio entre clases puede afectar en algunos casos el entrenamiento de los modelos de aprendizaje automático [407], lo que repercute en la precisión de la clasificación. La intervención de los problemas de clasificación multiclase con clases desequilibradas puede admitir múltiples soluciones. La construcción de métodos de conjunto [402, 329, 396, 331], de sistemas híbridos [392, 398, 355] o la mejora de los algoritmos existentes [406]. El uso de métodos de balanceo por submuestreo o sobre muestreo de instancias [408, 409, 410]. Entre estos métodos, SMOTE es una de las técnicas de sobremuestreo más utilizadas para equilibrar las clases minoritarias y mejorar el rendimiento del modelo [327]. Otra medida muy practica es la ponderación de clases, que consiste en asignar pesos más altos a la clase con instancias deficitarias frente a la clase dominante [411]. [412] mediante el sobremuestreo de clases minoritarias logran una precisión del 74% en la predicción de la dirección del índice BIST100. [413]

pondera las clases seleccionadas logrando una precisión del 90.33% en la predicción de la tendencia bursátil.

En resumen, el volumen y las características de los datos han fomentado el uso de técnicas de preparación para mejorar la calidad de los datos y la eficacia de los modelos de clasificación. La mayoría de los datos del "mundo real", al estar influidos por el ruido, los datos ausentes, incoherentes, irrelevantes, redundantes y de baja calidad, dificultan aún más su explotación. Así, la preparación de los datos, además de conferirles valor y hacerlos aptos para el análisis, permite utilizar eficazmente los métodos de clasificación para construir modelos predictivos más precisos.

Por otra parte, la construcción de modelos predictivos más precisos requiere el uso de acciones específicas de preprocesamiento y preparación de datos en función de la técnica que se vaya a utilizar. El examen de los resultados de estos estudios muestra, dada la naturaleza y la calidad de los datos, la magnitud de los problemas abordados para lograr los mejores resultados de clasificación.

3.3.2 Selección de características

La selección de características es una cuestión clave en la construcción de modelos de predicción. Tradicionalmente, la definición de variables de entrada en la construcción de modelos de clasificación supervisados se ha enfocado como un problema de selección de características. En esencia, la selección consiste en descartar características redundantes e irrelevantes que no son útiles en la clasificación. Por tanto, la elección del mejor subconjunto de atributos es aquél que teóricamente discrimina e interpreta mejor las diferencias entre grupos en el momento de la clasificación. Teniendo en cuenta estos aspectos en esta sección se examina y sintetiza el estado actual del conocimiento sobre la selección de variables en la construcción de modelos que predicen la tendencia de los precios de los activos. Las interpretaciones que se describen a continuación proceden de notables trabajos publicados en el ámbito de la selección de características y se basan en los enfoques de filtrado (*filtering method*), envoltura (*wrapping method*) e incrustación (*embedding method*). El análisis de estas publicaciones conduce a observaciones que deben tenerse en cuenta en la definición de procedimientos en la selección de características.

A. Generalidades

La selección de características ha demostrado ser una herramienta valiosa para definir el mejor subconjunto de variables candidatas en la construcción de modelos de clasificación. Además de descartar la inclusión de variables irrelevantes y redundantes, ayuda a generar modelos predictivos legibles, interpretables, precisos y menos inestables. Siempre que se conozca o no el grupo de pertenencia de cada observación, la selección de características puede aplicarse a problemas de clasificación supervisada [414], no supervisada [415, 416] y semisupervisada [413]. La selección de atributos en problemas de clasificación supervisada, a diferencia de la clasificación no supervisada, implica conocer previamente el grupo al que pertenece cada observación [417]. A diferencia de las dos anteriores, la selección de atributos en la clasificación semi-supervisada utiliza observaciones de la misma población, con y sin etiquetado de clase, para identificar las características discriminativas más relevantes [418].

Aunque los modelos de aprendizaje no supervisado han demostrado su utilidad para detectar estructuras de correlación entre observaciones sin membresía de clase, y los enfoques semi-supervisados son valiosos para identificar el grupo de pertenencia de las observaciones mayoritarias sin etiquetas, estos enfoques no se abordan en este documento porque no son atractivos para los fines señalados.

B. Medidas de selección de características

En los problemas de clasificación del mundo real, tradicionalmente se desconocen las variables predictoras que proporcionan la mejor discriminación entre grupos [419, 420, 421]. En un esfuerzo por identificar las variables que mejor explican la acción del precio y predicen la tendencia de los activos financieros, se han propuesto varios enfoques de selección basados en algoritmos de aprendizaje automático [317, 422, 318, 423, 424]. Su correcta aplicación implica revisar y conocer de antemano los detalles operativos requeridos por estos métodos según las necesidades de cada problema de selección. Aunque los detalles operativos son muy diversos, al menos pueden resumirse en tres pasos [425], la definición de subconjuntos de variables, el análisis de estos subconjuntos utilizando un criterio de selección y la validación de las variables predictoras elegidas. Aunque quedan por resolver algunas cuestiones críticas sobre la definición y el uso de criterios en la selección de características. El problema de definir y medir la relevancia como criterio de selección sigue siendo objeto de debate [426]. Más aún cuando las medidas de estos criterios están presentes entre variables cuyas interacciones dificultan su

medición y detección individual [427]. La relevancia puede entenderse como un calificativo asignado subjetivamente a un atributo en un contexto determinado. Sin embargo, desde un punto de vista estadístico, este calificativo se hace medible en cuanto se identifican las características menos relevantes. Así, en el proceso de comparación, en el que se analiza la dependencia e independencia entre variables, un atributo es capaz de proporcionar información sobre el otro. Esto hace que se puedan identificar las variables relevantes a partir de las variables redundantes. En consecuencia, las variables independientes continuas que influyen en el comportamiento de las categorías de la variable dependiente son variables relevantes [428]. Del mismo modo, las variables que están muy correlacionadas entre sí no aportan información adicional en la construcción de modelos predictivos. Por lo tanto, se vuelven irrelevantes porque pueden generar problemas de multicolinealidad y de singularidad, especialmente cuando se utilizan algunos métodos estadísticos [429, 430, 431, 432]. En este contexto, los métodos utilizados en la selección de características se basan en la identificación de variables relevantes mediante mecanismos de comparación y de descarte de variables redundantes.

C. Métodos y medidas de selección

La revisión de la literatura muestra un creciente cuerpo de investigación sobre métodos de selección de atributos utilizados en la construcción de modelos predictivos [433, 420, 426, 434, 428]. Durante las últimas décadas, la selección de características en problemas de clasificación supervisada se ha abordado utilizando modelos de filtro [420, 426], modelos envolventes [435, 436, 437], modelos integrados [438]. Aunque todavía no existe un consenso claro sobre qué métodos de selección son los más adecuados, se han intensificado los esfuerzos para mejorar la eficacia de estos enfoques con miras a construir modelos de clasificación más precisos [437].

En términos generales, en los métodos de filtrado, la selección de características se basa en el uso de distintos parámetros de proximidad entre variables. Tradicionalmente, estas medidas se han utilizado como criterios de selección. De hecho, el análisis de dependencia e independencia univariantes y multivariantes entre variables es uno de los criterios más utilizados en la identificación del mejor subconjunto de características. Además, su definición se realiza independientemente del método de clasificación que se vaya a utilizar para llevar a cabo la predicción. Esto significa que, en cierto modo, la selección ignora el desempeño que este subconjunto de variables puede tener en la construcción de los modelos de clasificación [439]. Los métodos envolventes, a difer-

encia de los métodos anteriores, se han caracterizado porque utilizan la precisión de la clasificación como criterio de selección de características. Mediante un algoritmo de clasificación, que se asigna por defecto, el método de selección elige el mejor subconjunto de características en función de los mejores resultados obtenidos en la clasificación [440]. Lo que se convierte en un problema de alta complejidad computacional cuando se analizan matrices de datos multivariantes. Los métodos integrados en los que se combinan métodos de filtro y envoltura se caracterizan porque, después de identificar los mejores subconjuntos de variables utilizando distintos criterios de proximidad, se elige el mejor subconjunto de características usando la precisión de la clasificación como criterio de selección [441].

D. Métodos de filtro

Al abordar estos enfoques con más detalle, en los métodos de filtrado, entre las medidas de proximidad más utilizadas como criterios de selección de características, se destacan las basadas en la información [442, 443, 444], la similaridad [445, 446], la distancia [447], la consistencia [448, 449, 450], la dependencia [451, 452], la entropía [453, 454] y las medidas estadísticas (correlación [455], Chi-cuadrado [456] y las puntuaciones de Fisher [457]). Entre las medidas más relevantes utilizadas como criterios de selección se encuentran las medidas de independencia y de diferenciación de grupos [458]. Desde el punto de vista estadístico, el análisis de independencia entre variables se posibilita mediante el coeficiente de correlación lineal de Pearson, mientras que el análisis de diferencias entre grupos se garantiza con el criterio de Fisher [459]. El análisis de correlación permite identificar las variables que son relevantes para las categorías de la variable dependiente. Asimismo, identifica las variables correlacionadas, lo que ayuda a descartar las variables redundantes [460, 428]. En estos casos, cuando la presencia de muchas variables redundantes es evidente, una sola variable es suficiente para describir el comportamiento de los grupos en los datos. Por otra parte, la media de los datos es el estimador estadístico utilizado para evaluar las diferencias entre grupos [429] y el criterio de Fisher evalúa el nivel de discriminación; cuanto mayor sea la puntuación, mayor será el poder discriminatorio de las variables [461]. El cumplimiento de estas condiciones hace que las variables candidatas sean útiles y relevantes en la construcción de modelos de clasificación legibles, parsimoniosos y más precisos, de modo que las demás variables pueden descartarse porque no añaden información. Aunque la literatura revela numerosos esfuerzos en el diseño de métodos de ayuda en la búsqueda de variables

relevantes para la construcción de modelos de clasificación, aparentemente son pocos los trabajos que utilizan el poder discriminatorio y la independencia en la selección de variables legibles para la construcción de modelos predictivos [32, 29, 34, 36, 37, 77, 33].

Algunos estudios han explorado el uso de distintos criterios de selección para mejorar la eficacia de los modelos predictivos. Aunque la complejidad computacional también se ha visto afectada, la precisión en las tareas de clasificación ha alcanzado valores más competitivos. Utilizando métodos de filtrado, destacados estudios recurren a criterios de selección basados en la información mutua [454, 412, 462], Ratio de ganancia [426, 463], Relief [426], ReliefF [437], Ganancia de información [464, 465], incertidumbre simétrica [464, 437], Chi-cuadrado [465, 412], la ponderación de características [466, 467, 468, 413] y el análisis de redundancia [469, 470, 471] para predecir la dirección de los precios de los activos.

El rendimiento obtenido con el uso de estos métodos y de diferentes medidas de selección de características se pone de manifiesto en los resultados obtenidos con los modelos de clasificación. Los autores en [426] seleccionan los indicadores técnicos más informativos utilizando métodos de filtrado con puntuaciones de relevancia, ratio de ganancia y algoritmo Relief. Además, utilizando el algoritmo GBM como clasificador obtiene una precisión del 59% prediciendo el rendimiento diario de las acciones de la Bolsa de Estambul (BIST). [393] utiliza la correlación lineal de Pearson como criterio de selección de las características más relevantes para construir modelos de regresión de bosque aleatorio (RFR) y de regresión de gradiente extremo (XGBR) y predecir la tendencia del índice Nifty50. Sin embargo, hay que señalar que algunos estudios [393] consideran que la calidad de los datos de entrada se mejora ampliando el espacio de características con atributos "más relevantes" derivados de los atributos originales. Lo que puede llevar a problemas de multicolinealidad si no hay un proceso previo de selección o extracción de características relevantes. El uso de predictores, altamente correlacionados con sus variables originales, puede dar lugar a modelos de clasificación inestables.

Por otra parte, la predicción de la dirección de los precios a partir de noticias financieras ha planteado también un enorme desafío, dada la baja precisión de las previsiones señalada en la bibliografía [465]. Los autores en [412] utilizando la información mutua equilibrada y la relevancia como criterio de selección de características resuelve el problema de desequilibrio de clases. Además, alcanza una precisión del 74% en la predicción de la dirección del índice BIST 100. Este rendimiento es superior al proporcionado por las características seleccionadas utilizando información

mutua y Chi-cuadrado. Algunos estudios han sugerido que enormes espacios de características conducen a la construcción de modelos de clasificación inexactos [462]. El mejor rendimiento en la predicción de la dirección del mercado bursátil, con una precisión de hasta el 52.3%, se consigue tras la selección, el uso de noticias financieras y las máquinas de vectores de soporte. Los resultados sugieren que cuanto menor es el número de características que intervienen en la construcción del modelo de clasificación y más discriminativas son, mayor es la precisión. [465] utilizando cinco criterios de selección y una SVM encontró que 300/800 variables candidatas proporcionan un F-Score del 84% en la predicción bursátil a partir de noticias financieras. La exactitud de la predicción (PA) es el criterio de selección que supera a la frecuencia de documentos (DF), la frecuencia de términos-invertir frecuencia de documentos (TF-IDF), la ganancia de información (IG) y el estadístico Chi-cuadrado. [466] alcanza una precisión del 97.03% en la clasificación del sentimiento de las noticias financieras para predecir las tendencias del mercado de valores. Este rendimiento se consigue con una máquina de vectores de soporte (SVM) como clasificador y la reducción del espacio de características utilizando la frecuencia de documentos (DF) como criterio de selección y el enfoque n-gram con diferentes pesos para la extracción de características. Utilizando este mismo enfoque [467, 468, 413] señalan que los métodos de selección y ponderación de características pueden desempeñar un papel importante en la clasificación de sentimientos. Luego de seleccionar las variables de entrada más relevantes, [468] obtiene hasta un 83.33% de precisión al predecir la tendencia intradía del tipo de cambio, mientras que [413] logra un 90.33% de precisión en la predicción de la tendencia bursátil, en ambos casos utilizando titulares de noticias financieras.

E. Métodos de envoltura

Los métodos envolventes y el renovado interés por el uso del aprendizaje supervisado en la construcción de modelos predictivos han dado resultados prometedores [436, 437, 463, 464]. En comparación con los métodos de filtro, los métodos de envoltura también han producido resultados comparables [472]. Esto se debe principalmente a que su rendimiento viene determinado por la definición y solución de una función objetivo. El resultado de la clasificación es la función objetivo que se utiliza para evaluar los subconjuntos de variables independientes. Así, cuando la función objetivo maximiza la precisión de la clasificación, con el mínimo número de características, prácticamente identifica el subconjunto óptimo de variables relevantes potencialmente utilizables en

los modelos de clasificación. Sin embargo, hay que señalar que las variables seleccionadas con estos enfoques no se basan en propiedades de independencia y diferenciación de clases. Por lo tanto, su uso en la construcción de modelos, dada la posibilidad de utilizar variables redundantes, podría dar lugar a modelos de clasificación inestables y sobre ajustados [473].

F. Métodos integrados

La elección del subespacio de características siempre ha sido un factor determinante en la eficacia de los modelos de predicción. Su elección, en cierta medida, puede entrar en conflicto con los méritos de los métodos de clasificación si éstos no se ajustan a los principios que rigen la selección. Para explotar estas debilidades se han propuesto diversos enfoques integrados. En los últimos años, el creciente interés por construir métodos integrados [424] y modelos híbridos [464, 422, 318] se ha centrado en el uso de técnicas computacionalmente costosas. Estas técnicas han demostrado su eficacia en la selección y extracción de características. Sin embargo, la revisión de la literatura muestra que sigue siendo difícil encontrar mejores medidas para la selección de características. En particular, cuando se desconocen los principios y méritos de los modelos utilizados en la selección y clasificación. Especialmente cuando se desea mejorar la precisión de las predicciones [389].

Algunos estudios empíricos que predicen la dirección de los precios de los activos han hecho hincapié en el desarrollo de métodos híbridos de selección de características para mejorar el rendimiento de la predicción. [464] propone el método denominado Puntuación F y búsqueda secuencial hacia delante admitida (F_SSFS). Combinando los métodos de filtro y envoltura en la selección de un subconjunto óptimo de características (17/30 variables candidatas), SVM supera a una red neuronal de retro propagación (BPNN) con una tasa de precisión del 87.3% en la predicción de la tendencia del índice NASDAQ. Otros estudios basados en enfoques híbridos combinan la correlación como método de filtrado con el método envolvente de selección de características de evolución diferencial (DEFS) [422]. Este método despliega diferentes clasificadores basados en la distancia como criterio de evaluación. La media móvil simple de 5 días es uno de los indicadores técnicos más relevantes seleccionados. Mientras que los datos sobre materias primas, índices bursátiles y datos de mercado son minoritarios. Utilizando KNN como clasificador, la predicción de la dirección del índice bursátil S&P 500 alcanza una precisión de hasta el 90.18%.

G. Búsqueda de subconjuntos de características

Los enfoques envolventes, así como los enfoques integrados e híbridos, como se ha mencionado, han demostrado ser computacionalmente costosos y exhaustivos. En particular, en tareas de búsqueda de subconjuntos de características antes de su evaluación y selección. Para mitigar esta dificultad, la literatura especializada revela que las búsquedas heurísticas ayudan a mejorar la selección de las variables de entrada y, en consecuencia, la precisión de la clasificación [474, 475, 436, 468]. Otros enfoques como el uso de algoritmos evolutivos, aplicados como estrategias de búsqueda de soluciones y de optimización, han demostrado su utilidad en el análisis de enormes conjuntos de datos [84, 476]. Entre los algoritmos evolutivos computacionalmente viables para abordar este tipo de tareas se encuentran los algoritmos genéticos [473, 402] y la optimización por enjambre de partículas (PSO) [84]. Además, el uso de métodos de búsqueda y selección secuenciales ha mejorado el rendimiento de este tipo de enfoques para la selección de características [421]. Otros enfoques de búsqueda, como Best-First, Hill-climbing [477], Branch-and-Bound no son ampliamente utilizados en la búsqueda de subconjuntos de características. Concretamente, en la construcción de modelos de clasificación para predecir la tendencia de los precios de los activos.

H. Análisis discriminante como método de selección

La eficacia del proceso de clasificación reside en la capacidad de identificar y seleccionar las variables discriminantes e independientes que mejores diferencias producen entre las categorías de la variable dependiente. La revisión de la literatura científica muestra que el análisis discriminante como herramienta de selección de características no se utiliza con tanta frecuencia como los métodos de selección basados en inteligencia artificial [57]. Aunque su uso como método de clasificación se ha extendido a todas las ramas del conocimiento, su aplicación en la predicción de la evolución de variables financieras aún no ha recibido la importancia que merece. En parte, esto puede deberse a que la fiabilidad de su aplicación requiere que las variables predictoras asuman una distribución normal multivariante [81, 86]. En algunos casos, estos requisitos difíciles de cumplir limitan su aplicación, lo que no ocurre con los métodos de aprendizaje automático que son ampliamente utilizados [87, 88]. Sin embargo, su eficacia ha quedado plenamente demostrada en algunos estudios empíricos en los que se utiliza como referencia para evaluar el rendimiento de otros modelos de clasificación [55, 478, 479, 391, 480]. Otros enfoques se han beneficiado de la reducción de la dimensionalidad de los datos

extrayendo características y utilizándolas en la construcción de modelos [481].

La selección de variables por pasos utilizando análisis discriminante es una extensión del modelo original que ayuda a identificar las variables que ofrecen el mejor rendimiento de clasificación. Aunque este procedimiento es demasiado exhaustivo y costoso desde el punto de vista computacional, la literatura científica no proporciona pruebas de su uso en la selección de características para predecir la tendencia de variables financieras. Tampoco se observa su aplicación en la detección de variables que mejoren la precisión de los algoritmos de clasificación. Sin embargo, los estudios publicados hasta la fecha sí sugieren que el uso de grandes volúmenes de datos para la selección y extracción de características se debe al desconocimiento de las variables que mejor discriminan entre tendencias [465]. En estos enfoques, la selección de características se basa en el rendimiento obtenido con las tareas de clasificación y el uso de medidas de filtrado como criterios de selección. Aunque estos enfoques también han demostrado ser computacionalmente costosos, a diferencia del análisis discriminante por pasos, su uso si se ha intensificado.

Por otro lado, existe un importante subconjunto de métricas útiles para medir la similitud o diferencia entre observaciones. Estas medidas de distancia o separación entre clases han demostrado ser ideales para detectar si la característica medida por esa variable contribuye o no a la diferenciación de las categorías de la variable dependiente. Entre ellas se encuentran la lambda de Wilks, la distancia de Mahalanobis, la varianza no explicada y la traza de Lawley-Hotelling. Otras medidas, como el límite inferior de Cramer-Rao (CRLB), también han demostrado su valor en la diferenciación de clases [482]. Sin embargo, al revisar la literatura científica, no se informa del uso de medidas de diferenciación interclase como criterios de selección de características. Por lo tanto, el uso de estas medidas como criterios de selección en la detección de características discriminativas, que contribuyen a la diferenciación entre grupos, mejoraría el rendimiento de los modelos de clasificación.

En resumen, los problemas del mundo real se caracterizan por el desconocimiento de las variables que mejor explican el comportamiento de la variable objeto de estudio. Tradicionalmente, la identificación de variables relevantes se ha abordado como un problema de selección de características, especialmente en la construcción de modelos de clasificación. En general, estos procesos utilizan criterios de selección y algoritmos de búsqueda para detectar y descartar las variables menos relevantes. Los estudios empíricos han demostrado que la inestabilidad y el sobreajuste de los modelos de clasificación vienen determinados por la calidad de las variables y sus relaciones de

interdependencia entre las variables utilizadas en su construcción. Los enfoques de selección más notables propuestos hasta la fecha incluyen el uso de técnicas de aprendizaje automático y métodos estadísticos. Entre ellos se destacan los métodos de filtro, los métodos de envoltura y los modelos integrados. Sin embargo, en la predicción de tendencias de activos financieros no hay pruebas del uso de medidas de separación de clases como criterio de selección para la detección de variables discriminantes y su utilización en la construcción de modelos de clasificación.

La selección de características ha demostrado ser un procedimiento clave en el preprocesamiento de datos para construir modelos de clasificación más eficaces. Aunque hasta la fecha la bibliografía ha recogido un importante número de contribuciones, aún no existe un consenso claro sobre qué criterios y métodos utilizados en la selección de características son los más adecuados [483]. En la práctica, estos procedimientos de selección se han estructurado en función de las particularidades de cada problema bajo un contexto determinado. Esa falta de generalización es la que ha dado lugar a un gran número de contribuciones divulgadas hasta la fecha. Los resultados de estos estudios demuestran que los criterios utilizados en la selección de características dependen no sólo del tipo de datos y de la interdependencia multivariante de los atributos, sino también de los méritos de las técnicas de clasificación empleadas en la construcción de modelos predictivos. Llegados a este punto, los criterios a utilizar en la selección de características, además de adaptarse a la estructura de los datos, deben permitir seleccionar el subconjunto de variables que mejor ayuden a explicar y predecir el comportamiento de los datos según el propósito de cada estudio.

Por último, cabe señalar que la identificación y selección de características requiere un criterio de selección que ayude a medir y detectar las variables más relevantes entre los datos. Una vez definido este criterio, se puede desarrollar un procedimiento para identificar el subconjunto de variables predictoras que mejor se ajustan a los criterios utilizados en la selección. También hay que tener en cuenta que, en estos análisis de relevancia, las variables importantes por sí solas no son muy informativas, pero cuando se comparan con otras, la información que proporcionan ayuda a identificar y descartar las variables menos relevantes.

3.3.3 Extracción de características

La alta dimensionalidad de los datos se ha abordado clásicamente como un problema de extracción de características. La extracción, a diferencia de la selección de car-

acterísticas, consiste en obtener dos o más combinaciones lineales de las variables originales con la menor pérdida de información posible. Estas nuevas características ortogonalmente independientes suelen utilizarse como variables predictoras en la construcción de modelos de clasificación. La extracción ha demostrado ser una potente herramienta en la definición del mejor subconjunto de variables predictoras. Además de reducir el ruido generado por la presencia de variables irrelevantes, reduce la complejidad computacional, el tiempo de entrenamiento y ayuda a producir modelos más estables. Aunque las combinaciones lineales son estructuras de correlaciones entre variables observables, el problema de la legibilidad e interpretabilidad en su uso como variables predictoras es evidente. Tradicionalmente el análisis de componentes principales (PCA) y el análisis discriminante lineal (LDA) han sido las técnicas más utilizadas en la extracción de características [484, 400]. Como se ha mencionado, cuando se utilizan estos enfoques como métodos de extracción de características, las nuevas combinaciones lineales, aunque pueden utilizarse como variables predictoras independientes, la legibilidad e interpretabilidad de los modelos de clasificación se ve dificultada por su uso. Por lo tanto, este tipo de enfoque no se aborda en este trabajo porque no se ajusta a los objetivos de este.

3.4 Construcción del modelo de clasificación

En un esfuerzo por predecir la dirección del mercado de divisas y beneficiarse de la volatilidad de los tipos de cambio, los participantes del mercado están cada vez más interesados en construir modelos de clasificación más eficaces. Aunque se han obtenido resultados muy prometedores en el campo de la inteligencia artificial, sigue habiendo una gran necesidad de desarrollar modelos de previsión parsimoniosos, legibles y más precisos. Teniendo en cuenta estas directrices, a continuación, se presenta una breve revisión del estado de la técnica en la construcción de modelos de clasificación. Las obras citadas describen la eficacia y versatilidad que ofrece el análisis discriminante en la construcción de modelos predictivos. La revisión bibliográfica de los modelos de clasificación utilizados en la predicción de la tendencia de los precios aborda una amplia variedad de instrumentos financieros. La sección 3.4.1 hace referencia a las criptomonedas, la sección 3.4.2 a los tipos de cambio, la sección 3.4.3 a los valores de renta variable, la sección 3.4.4 a las carteras de inversión y la sección 3.4.5 a los índices bursátiles. Por último, la Sección 3.4.6 describe brevemente el estado del arte en la definición de hiperparámetros para la configuración de métodos de clasificación y su

influencia en el uso del análisis discriminante lineal. Predecir la dirección de las variaciones de los tipos de cambio para tomar decisiones de inversión acertadas se ha convertido en uno de los retos más desafiantes del sector financiero. Principalmente por la complejidad y los riesgos que supone para instituciones, inversores y partes interesadas. La literatura científica demuestra que la dirección de los precios de los instrumentos financieros se predice para distintos periodos de tiempo y utilizando métodos y variables de entrada muy diversos. Según los estudios divulgados hasta la fecha, la construcción de modelos de clasificación se basa en métodos estadísticos, algoritmos de aprendizaje supervisado y el uso de técnicas de inteligencia artificial [485, 76, 75, 486, 487, 488]. Como ya se ha mencionado, entre las variables de entrada más utilizadas, según la sección 3.2, se encuentran los datos de mercado, los indicadores técnicos, los fundamentales y la información heterogénea (noticias financieras y publicaciones en redes sociales).

El auge de la inteligencia artificial, dado que su aplicación no requiere el cumplimiento de supuestos estadísticos en los conjuntos de datos, ha desplazado en parte el uso de algunos enfoques tradicionales. Algunos de estos métodos convencionales incluyen el análisis de series temporales, aplicado a la construcción de modelos predictivos mediante métodos $ARIMA$ univariantes y multivariantes. Aunque estas técnicas han arrojado resultados prometedores, el rendimiento ofrecido por los nuevos enfoques ha relegado su uso. Estudios recientes, como ya se ha mencionado, han mostrado una marcada preferencia por el uso de técnicas de aprendizaje automático supervisadas y no supervisadas. Los hallazgos reportados en [75] ponen de manifiesto que, entre las técnicas más utilizadas para predecir la evolución de los precios de los instrumentos de renta variable, los tipos de cambio y los índices bursátiles se encuentran los modelos de clasificación basados en máquinas de vectores de soporte (SVM), los modelos difusos, las redes neuronales artificiales (ANN) y las redes neuronales de aprendizaje profundo (DL). Algunos trabajos notables, de previsión de la evolución de los tipos de cambio se basan en técnicas de inteligencia artificial. Entre los enfoques más avanzados se encuentran el aprendizaje profundo por refuerzo [489], la memoria a largo plazo (LSTM) [490], el análisis de sentimiento a partir de contenido textual [491, 492], el análisis wavelet [493], los algoritmos genéticos [494], las máquinas de vectores de soporte y otros enfoques [495]. En un intento por mejorar la precisión de estos algoritmos, los investigadores han propuesto técnicas para combinarlos y formar sistemas más robustos. Entre ellos se encuentran los métodos híbridos [84], los métodos ensemble [325, 329] y los métodos de predicción conjunta por refuerzo (Boosting) [496, 497].

La Tabla 3.2 presenta algunas de estas técnicas de clasificación y la Tabla 3.3 informa sobre el desempeño en la predicción obtenido con estos métodos. El rendimiento en la clasificación, con estos enfoques, alcanza valores de precisión competitivos.

Tabla 3.2: Método de clasificación.

Autor	[a] Método	Instrumento
Das et al.(2022) [325]	DL	Tipos de cambio
Dash et al.(2022) [326]	Híbrido	Tipos de cambio
Ortu et al.(2022) [324]	DL, MLP, LSTM*, ALSTM, CNN	Criptomonedas
Hao et al.(2022) [315]	Ensemble	Materias primas
Padhi et al.(2022) [327]	Fusión de modelos	Portafolios
Ozer et al.(2022) [328]	ML, LR*, KNN, SVM, RF, XGBoost, CatBoost	Criptomonedas
Sadeghi et al.(2021) [329]	Ensemble SVM	Tipos de cambio
Uras et al.(2021) [330]	SVM, XGBoost (XGB), CNN*, LSTM	Criptomonedas
Ampomah et al.(2020) [331]	Ensemble	Índices bursátiles y Valores
Kwon et al.(2019) [332]	DL, LSTM*	Criptomonedas
Fischer et al.(2018) [333]	DL	Portafolios
Hu et al.(2018) [334]	ANN	Índices bursátiles
Bustos et al.(2017) [335]	SVM*, ANN	Valores
Chakraborty et al.(2017) [336]	LR, SVM*, DT, BT, RF	Índices bursátiles y Valores
Coyne et al.(2017) [337]	ANN (MLP)	Valores
Dingli et al.(2017) [338]	DL	Índices bursátiles
Huang et al.(2017) [339]	GA	Valores
Dang et al.(2016) [340]	SVM	Índices bursátiles
Di et al.(2016) [341]	DL, Ensemble	Índices bursátiles
Ghanavati et al.(2016) [342]	SVM, LMNN, FuzzyML-SVM*	Índices bursátiles
Chai et al.(2015) [343]	Híbrido	Índices bursátiles
Dash et al.(2015) [344]	RELM	Índices bursátiles
Feuerriegel et al.(2015) [345]	DL	Valores
Gonzalez et al.(2015) [346]	Ensemble SVMs*, Bagging, Boosting y RF	Índices y Tipos de cambio

[a]DL Deep Learning = Aprendizaje profundo, LSTM Long-Short Term Memory = Memoria a corto plazo, ALSTM Attention-based Short-Term Memory = Memoria a corto plazo basada en la atención, CNN Convolutional Neural Network = Red neuronal convolucional, Ensemble methods = Combinación de varios modelos en un modelo predictivo óptimo, ML Machine Learning = Métodos de aprendizaje automático, LR Logistic Regression = Regresión logística, KNN k-nearest neighbors = Algoritmo de clasificación basado en la agrupación de k vecinos más cercanos, SVM Support Vector Machine = Máquina de vectores de soporte, RF Random Forest = Bosque aleatorio, XGBoost Extreme Gradient Boosting = Algoritmo de clasificación basado en árboles de decisión, CatBoost Categorical Boosting = Algoritmo de clasificación basado en el refuerzo del gradiente, ANN Artificial Neural Network = Red neuronal artificial, DT Decision Tree = Árbol de decisión, BT Boosted Tree = Árbol reforzado, MLP Multi-layer perceptron classifier = Clasificador perceptrón multicapa, GA Genetic Algorithm = Algoritmo genético, LMNN Large Margin Nearest Neighbor = Clasificación por vecino más próximo de gran margen, FuzzyML = Método de aprendizaje automático basado en lógica difusa, RELM Extreme Learning Machine Ridge = Máquina de aprendizaje extremo ridge, Bagging Bootstrap AGGregatING = Método ensemble basado en árboles de decisión ajustados por muestras diferentes del mismo conjunto de datos, Boosting = método ensemble basado en refuerzo secuencial de modelos mejorados. *Métodos de clasificación con mejor desempeño, los valores de exactitud (Accuracy) están referidos en la Tabla 3.3.

Trabajos basados en el análisis multidimensional de las diferencias entre grupos muestran que el análisis discriminante es una de las técnicas estadísticas más aplicadas

en el ámbito financiero [498]. En concreto, en el análisis del riesgo de insolvencia [499, 500, 501], la formulación de modelos de puntuación crediticia [502, 58, 503, 504, 505] y el análisis de series de tiempo [506, 507], entre otros. Estos trabajos han mostrado los mejores resultados en su campo de aplicación. Investigaciones recientes que utilizan técnicas de aprendizaje automático e inteligencia artificial en el campo financiero muestran que el análisis discriminante no se usa tan ampliamente en la predicción de precios como los algoritmos de aprendizaje automático [498]. Esto se debe probablemente al hecho de que la estructura de los datos viola los supuestos de normalidad multivariante exigidos por el método para dar fiabilidad a las interpretaciones. A pesar de estas dificultades señaladas por distintos autores en [54, 508, 509], se han realizado notables investigaciones sobre la previsión de la evolución de los precios utilizando el análisis discriminante a partir de datos de mercado, datos fundamentales y datos con información heterogénea [51, 58, 510, 59, 511, 55, 57].

3.4.1 Criptodivisas

Predecir la evolución de los precios de las criptodivisas es una tarea compleja dado su comportamiento altamente volátil en periodos de tiempo relativamente cortos. Aunque son muchos los factores que pueden afectar al comportamiento de sus precios, dada la influencia de la demanda, la regulación, los cambios tecnológicos, el sentimiento de los inversores, entre otros aspectos, el interés por predecir su evolución futura ha aumentado en un intento de reducir el riesgo de inversión. El trabajo de [53] analiza 146 activos digitales que comenzaron a cotizar a fines de 2014 y se valoraron hasta terminar 2018. Los autores construyen un modelo lineal discriminante que clasifica las criptomonedas en dos categorías: activo con riesgo de impago y activo sin riesgos de impago con información del inversor. El modelo haciendo uso del poder discriminativo de las variables produce una puntuación en la clasificación del 87%. Un trabajo reciente basado en el procesamiento del lenguaje natural (NPL) [51] predice la dirección del precio de Bitcoin (BTC) con un día de antelación. Haciendo uso del análisis de sentimiento (una rama emergente del ML) identifica información relevante a partir de titulares de noticias financieras y precios de Bitcoin para crear un modelo discriminante lineal (LDA). El rendimiento obtenido en las tareas de clasificación alcanza una precisión del 58.5%. [58] predice la tendencia de los precios del Bitcoin mediante ingeniería dimensional de muestras. El estudio compara el desempeño de los algoritmos de aprendizaje automático (memoria a largo plazo (LSTM), análisis discriminante cuadrático

(QDA), máquina de vectores de soporte (SVM), bosques aleatorios (RF) y XGBoost) con los métodos estadísticos tradicionales (análisis discriminante lineal (LDA) y regresión logística (LR)). Los resultados de los experimentos (predicción del Bitcoin en marcos de tiempo de 5 minutos) sugieren que los métodos estadísticos superan a los algoritmos de aprendizaje automático en la clasificación diaria con una precisión del 66%.

En general, los resultados de estos estudios sugieren que el análisis discriminante, dependiendo de la naturaleza de los datos, es estadísticamente adecuado si, además de superar los supuestos estadísticos, proporciona altos niveles de precisión en la clasificación. Aunque la precisión es un buen indicador del ajuste del modelo a la estructura de los datos, las variables utilizadas, dada su capacidad discriminante, son esencialmente fundamentales para obtener dichos niveles de desempeño.

3.4.2 Tipos de cambio

Predecir la tendencia futura de los tipos de cambio dada la volatilidad e incertidumbre de los mercados es también una tarea difícil. Aunque los tipos de cambio están influidos por factores fundamentales, técnicos y de comportamiento, los cambios en estos factores hacen que predecir con exactitud la tendencia futura sea una tarea compleja y arriesgada. En cuanto a los pares de divisas más negociados en el mercado Forex, un trabajo especializado investiga las ventajas de un algoritmo de ventana secuencial de Parzen que se basa en representaciones discretas [510]. El modelado mapea los precios, en una representación de cadena de mercado, utilizando casi 8 años de precios en un marco temporal de una hora de los cuatro pares de divisas más negociados (CHF/USD, EUR/USD, GBP/USD y AUD/USD). La tendencia de los tipos de cambio se predice utilizando el Análisis Discriminante de Fisher (FDA), una Máquina de vectores de soporte (SVM), y una Ventana de Parzen (PW). El desempeño de la clasificación arrojó una precisión de hasta el 53.25%. El desempeño de un modelo que analiza el efecto de la discretización de los datos fundamentales en la predicción diaria del movimiento de los tipos de cambio se presenta en [512]. Los autores evalúan el desempeño predictivo de la regresión lineal, el análisis discriminante y el uso de una red neuronal. Además, describen el proceso utilizado en la transformación de datos y la selección del modelo.

En general, aunque son pocos los estudios que predicen la tendencia de los tipos de cambio utilizando el análisis discriminante, la versatilidad de este método ha demostrado ser potencialmente funcional. Además de predecir la pertenencia de una

observación a una de varias categorías, también se ha demostrado su eficacia para evaluar el rendimiento de otros métodos de clasificación.

3.4.3 Valores

Predecir la tendencia de las acciones, dada la volatilidad a la que están sujetas debido a diversos factores, requiere un profundo conocimiento del mercado, lo que la convierte en una tarea difícil. En este contexto [55] analizan la influencia de los datos fundamentales en los movimientos del precio de las acciones de las empresas de telecomunicaciones sudafricanas basándose en el análisis de sentimiento. Utilizando datos fundamentales extraídos de titulares de noticias financieras y publicaciones de redes sociales (twits), los autores evaluaron el desempeño de distintos métodos tradicionales de clasificación. La función discriminante lineal (LDA) alcanza una precisión de clasificación del 94% al 96% en los experimentos, superando a la máquina de vectores de soporte (SVM), los bosques aleatorios (RF) y los árboles de decisión (DT). En [57] se presenta un nuevo enfoque que predice la tendencia de la prima de renta variable estadounidense utilizando 14 medidas de rendimiento financiero. Los autores evalúan cuatro grupos de técnicas de clasificación (métodos probit binarios penalizados, métodos probit binarios, árboles de clasificación y regresión (CART) y análisis discriminante). El análisis discriminante de alta dimensión (HDDA) produjo una precisión de clasificación del 67%, superando estadísticamente a los demás métodos. Al evaluar el rendimiento económico generado por la predicción, el análisis discriminante cuadrático (QDA) superó a los demás modelos generando el mayor retorno acumulado (3.51). Un novedoso trabajo presentado en [52] analiza el impacto de los anuncios de noticias en la predicción de retornos financieros. La utilización de variables cuantitativas extraídas de las noticias del mercado y la inclusión de indicadores de sentimiento (construidos con el análisis discriminante) como regresores exógenos de la volatilidad ayudan a reducir el número de violaciones reales del VaR con respecto a los datos esperados. Se propone un modelo de lógica difusa (n-fuzzy logic) con una máquina de vectores de soporte (SVM) en [513] con datos de mercado para predecir la tendencia bursátil. Los autores comparan el rendimiento de varios métodos, como el análisis discriminante múltiple (MDA), el árbol de clasificación y regresión (CART), la máquina de vectores de soporte (SVM), la regresión logística (LR), la red neuronal artificial (ANN) y los modelos SVM difusos. Además, los métodos estadísticos se utilizaron como referencia para evaluar el rendimiento de los demás modelos de clasificación. Se analiza el rendimiento de una

máquina de vectores de soporte (SVM) y de dos métodos estadísticos (Naive Bayes (NB) y análisis discriminante lineal (LDA)) en [514] para predecir los movimientos diarios de los valores máximos de las acciones vinculadas al índice S&P 500. Otro estudio en [515] utiliza técnicas estadísticas y métodos de aprendizaje automático supervisado en la predicción de datos de las acciones de ICICI Bank.

En general, el uso de información crítica (basada en titulares de noticias financieras y publicaciones en redes sociales) ha demostrado ser capaz de reflejar adecuadamente las diferencias entre clases (tendencias alcistas y bajistas, retornos financieros positivos y negativos, entre otras categorías). Esto significa que, cuando se dispone de información altamente discriminativa, una función lineal discriminante es estadísticamente suficiente para diferenciar entre grupos y facilitar la clasificación con altos índices de precisión.

3.4.4 Carteras de inversión

La estructuración de portafolios de inversión, con activos potencialmente generadores de rentabilidad y menor exposición al riesgo, es un proceso de diversificación que también ha demostrado su eficacia con el uso del análisis discriminante. En [56] se propone un enfoque de clasificación de carteras. El modelo predice el déficit esperado (ES) y el valor en riesgo (VaR) de 1500 carteras bivariantes formadas con información de materias primas, acciones y futuros sobre divisas. Basándose en una función discriminante lineal, los autores definen un modelo paramétrico de riesgo (VaR) y (ES) utilizando los estadísticos descriptivos de la distribución bivariante de las carteras como variables independientes y las categorías (VaR) y (ES) como variable dependiente. El rendimiento de la clasificación arrojó una precisión del 67.53% y el 53.80%, respectivamente. En [516], se presenta un nuevo método de selección de valores, basado en el análisis discriminante. El modelado se centra en la creación de una herramienta de ayuda a la configuración y diversificación de carteras de inversión. Otro método de selección de carteras se presenta en [508]. El análisis discriminante se utiliza para seleccionar ratios financieros y comparar su poder discriminatorio con el de otros enfoques. Los autores en [517] formulan una función discriminante lineal para clasificar los valores según su rendimiento en valores con menor y mayor rentabilidad financiera. Utilizando el análisis discriminante (LDA) y el análisis envolvente de datos (DEA), los autores obtienen una precisión de clasificación del 85%. Un enfoque novedoso propuesto en [54] predice la tendencia de los precios de las acciones utilizando un modelo

de regresión logística con parámetros definidos mediante análisis discriminante. Los resultados mostraron que los ratios financieros influyen en los precios más que los indicadores macroeconómicos. En [518] se presenta un modelo que detecta la manipulación del precio de las acciones en el mercado bursátil turco. El estudio evalúa el rendimiento de distintos métodos de aprendizaje automático. Entre las técnicas utilizadas figuran la regresión logística (LR), el análisis discriminante lineal (LDA), la máquina de vectores soporte (SVM) y la red neuronal artificial (ANN).

En general, al analizar los resultados de estos estudios, se observa que el uso de información crítica es esencial para lograr una alta precisión en las tareas de discriminación y/o clasificación. Esto sugiere que el bajo rendimiento obtenido en la clasificación se debe en parte al uso de información irrelevante y de técnicas que no se ajustan a la estructura de los datos. Por lo tanto, la obtención de los niveles más altos de precisión se debe principalmente al descubrimiento y uso de información crítica y al uso de métodos que mejor se ajusten a la estructura de los datos.

3.4.5 Índices bursátiles

Predecir la dirección en la que se moverán los índices bursátiles es una tarea compleja dada la influencia que ejercen múltiples factores en su comportamiento. No obstante, algunos trabajos destacados se han basado en el análisis discriminante. [59] evalúa el rendimiento de algunos métodos de clasificación en la predicción de la tendencia de los índices bursátiles FTSE 100, NIKKEI 225 y S&P 500. El análisis discriminante lineal (LDA) produce el mejor resultado con el índice NIKKEI 225, superando a las redes neuronales probabilísticas y a los modelos logit y probit, con una precisión de clasificación del 68%. Otro enfoque presentado en [511] predice la dirección de la tendencia semanal del índice NIKKEI 225. El modelo combinado de análisis discriminante lineal (LDA), máquina de vectores soporte (SVM), redes neuronales de retro propagación de Elman y análisis discriminante cuadrático (QDA) genera el mayor rendimiento en la predicción, con un porcentaje de aciertos del 75% (Hit ratio). El artículo en [519] predice tres índices bursátiles (LITIN, LITIN-A y LITIN-VVP) del mercado de valores lituano LNSE utilizando redes neuronales artificiales (ANN). Los autores utilizan el análisis discriminante lineal (LDA) como punto de referencia en la comparación de los modelos de tendencia autorregresivo, autorregresivo causativo y ANN causativo. Los mejores resultados de predicción se obtienen con diferentes enfoques ANN de retro propagación y diferentes ajustes de los parámetros.

En general, dada la complejidad de los mercados financieros, ha crecido el interés por predecir con precisión la dirección de los índices bursátiles utilizando diferentes métodos de clasificación. Los estudios aquí analizados demuestran que las funciones discriminantes lineales no adolecen de problemas de sobreajuste cuando las variables discriminantes son independientes y las observaciones de las muestras de estudio son suficientes para validar la eficacia de los modelos. Un trabajo notable que describe los métodos de aprendizaje automático más utilizados en el análisis de eventos financieros se resume en [520]. Los autores describen, entre otros temas de interés, los métodos de aprendizaje automático más utilizados para predecir los precios de los instrumentos financieros.

A la vista de los trabajos descritos, existe un interés creciente por predecir la dirección de los precios de los instrumentos financieros, como ayuda a la toma de decisiones informadas. Las contribuciones recogidas en la literatura ponen de manifiesto el interés por mejorar la precisión de las tareas de clasificación. Algunos estudios han reportado niveles de precisión de hasta el 96%, sin embargo, la multiplicidad de criterios disponibles para mejorar la eficiencia de estas predicciones sigue siendo un problema abierto a la investigación [473, 521]. Aunque se han realizado avances significativos en este aspecto, es innegable la necesidad de formular modelos legibles e interpretables para predecir y mejorar la eficacia de los juicios de valor en la toma de decisiones de inversión. Cabe señalar que la versatilidad del análisis discriminante en la clasificación de la tendencia de los precios ha quedado plenamente demostrada. Algunos estudios han utilizado incluso funciones lineales discriminantes como referencia para evaluar el rendimiento de otros modelos de predicción. Además, los esfuerzos por anticipar la dirección futura de los precios de los activos no se han visto respaldados por estrategias y modelos de negociación. Esto sugiere que debe avanzarse hacia la inclusión de estos modelos de previsión en las estrategias de negociación, con el fin de evaluar la eficacia de estas soluciones.

En resumen, la revisión bibliográfica muestra una fuerte predisposición hacia la construcción de modelos predictivos mediante métodos de inteligencia artificial. Este creciente interés puede deberse en parte a la necesidad de formular modelos más precisos, aunque se sacrifique la legibilidad e interpretación de las estimaciones. En cambio, el aparente desinterés por el uso de métodos estadísticos paramétricos multivariantes puede deberse en parte a la naturaleza asimétrica y leptocúrtica de los datos. Este

vacío de conocimiento, además de que supone la necesidad de superar las limitaciones derivadas por la disposición de los datos, también sugiere un mayor esfuerzo en la formulación de modelos predictivos parsimoniosos, legibles y más precisos. Incluso esto plantea un mayor desafío a los enfoques basados en inteligencia artificial, cuando se decida superar el problema de la caja negra, con la formulación de modelos de clasificación legibles e interpretables. Por otro lado, los modelos discriminantes lineales, dado su grado de simplicidad, son valiosos en términos de legibilidad, interpretabilidad y precisión, especialmente cuando se ajustan a los datos y utilizan medidas altamente discriminantes. Aunque se requiere que los datos satisfagan los supuestos de normalidad multivariante para proporcionar interpretaciones fiables, estos métodos no son sensibles a la definición de hiperparámetros si las variables predictoras separan claramente los grupos. En consecuencia, dado que estos modelos son combinaciones lineales de las variables predictoras, maximizan la distancia de separación entre clases y mejoran la precisión de la clasificación cuando la estructura de los datos lo permite.

3.4.6 Definición de hiperparámetros

La definición de los hiperparámetros de las técnicas utilizadas en la construcción de modelos predictivos ha sido durante años uno de los principales retos a los que se han enfrentado los investigadores [522]. La precisión de la clasificación depende en gran medida de los valores de los parámetros asignados. Aunque el número de parámetros, en la mayoría de los modelos, viene determinado por la naturaleza de cada técnica, se sabe por los datos de encuestas que en cada técnica hay al menos más de dos hiperparámetros [523]. Varios estudios han revelado que, para definir los parámetros de los métodos de aprendizaje automático, los métodos heurísticos y metaheurísticos (MH) son los más eficaces y ampliamente utilizados [522]. Aunque la definición es compleja y costosa desde el punto de vista computacional, muchos estudios se han visto obligados a idear métodos que ayuden a encontrar los valores de los hiperparámetros sin afectar la precisión y el tiempo computacional de las tareas de clasificación [524, 525]. Teniendo en cuenta estos aspectos [526] mediante la optimización de un problema dual define los parámetros de la función Kernel de base radial gaussiana (RBF) según las especificidades de [527] para construir el modelo (SVR) y predecir el precio y la tendencia del mercado de valores. [528] utilizan un algoritmo genético para optimizar los parámetros de una máquina de vectores de soporte de mínimos cuadrados (LS-SVM). La predicción de la tendencia de las acciones arrojo una tasa de clasificación

correcta fuera de la muestra de hasta el 94.5%. [529] mejoran el aprendizaje y logran predicciones eficientes de índices bursátiles utilizando una máquina de vectores de soporte (SVM) con optimización de parámetros con un algoritmo genético.

Los estudios empíricos basados en arquitecturas de redes neuronales artificiales (ANN) y aprendizaje profundo (DL) han demostrado que son difíciles de configurar y muy sensibles a los valores de los hiperparámetros [522, 530, 531, 519]. Los autores en [519] mediante la optimización de parámetros investigaron la mejor configuración de las redes neuronales de retropropagacion ANN para predecir los índices del mercado de valores de Lituania. [532] optimizan los datos de entrada y los parámetros de una red neuronal artificial (RNA) para mejorar la precisión de la predicción de una red neuronal NARX que predice las acciones de CIMB (Commerce International Merchant Bankers) en Malasia. [533] optimizan los parámetros y variables de entrada de una red neuronal artificial reparadora (RANN) generando una precisión de clasificación de hasta el 98.37% en los índices bursátiles Nifty50, Nifty Bank, Nifty Pharma, BSE IT y BSE Oil and Gas. [330], utilizando la técnica de búsqueda en cuadrícula "Grid Search", definieron los mejores valores de los hiperparámetros de una máquina de vectores de soporte (SVM), XGBoost (XGB), una red neuronal convolucional (CNN) y una red neuronal de memoria a largo plazo (LSTM) para predecir la dirección del bitcoin. [332] aplicaron una técnica de validación cruzada k-fold basada en la búsqueda en cuadrícula para encontrar los parámetros de mejor ajuste del modelo LSTM y predecir la dirección de las criptomonedas.

Por otra parte, varios estudios han revelado que el desarrollo de métodos heurísticos puede, mediante la resolución de problemas de optimización en la definición de hiperparámetros, mejorar el rendimiento de los algoritmos. [534] proponen un algoritmo de organización de conjuntos de transformación (2P-TO), en el que la definición de los parámetros de cardinalidad y el número de variables correlacionadas mejora la selección de atributos para predecir la tendencia de las acciones del índice NASDAQ. [535] con el algoritmo de enseñanza y aprendizaje (TLBO) optimizan los hiperparametros de una red neuronal convolucional hibrida con memoria a corto plazo (HPT-HCLSTM) para predecir el precio de las acciones y reducir el error de predicción. [166] introduce un método de optimización de parámetros para SVM a fin de mejorar el rendimiento de la predicción de precios de futuros de EUA (European allowance). [536] desarrollan una CNN y un algoritmo selector de características para optimizar los parámetros de una arquitectura híbrida de redes neuronales profundas CNN y memoria a largo plazo LSTM y logran una precisión del 98.31% en datos bursátiles de la Bolsa de Ghana

(GSE).

En general, el ajuste y definición de hiperparámetros es un proceso clave en la construcción de modelos de selección y clasificación [84], especialmente cuando se basan en técnicas de aprendizaje supervisado. Aunque existen múltiples métodos para encontrar y definir los valores de los hiperparámetros, los algoritmos metaheurísticos han demostrado su eficacia, principalmente cuando se trata de construir modelos de clasificación más precisos. En comparación con el análisis discriminante, la definición de hiperparámetros no supone un gran reto, ya que la técnica no es sensible a estos valores si las variables predictoras promueven una clara diferencia entre clases [57].

3.5 Evaluación de desempeño del modelo

Durante décadas, el problema de la predicción de la dirección de los precios de los activos se ha abordado desde el campo de la inteligencia artificial como un problema de clasificación supervisada. A lo largo de los años, ha prevalecido el uso de medidas cuantitativas como criterios de evaluación para mejorar el rendimiento de los modelos de clasificación. Este enfoque basado en indicadores calculados a partir de los valores extraídos de la matriz confusión es el más utilizado. Aunque la elección de las medidas a utilizar en el proceso de evaluación no es infalible y depende de cada problema específico, el uso de estos indicadores revela la eficacia del proceso de clasificación y el grado de ajuste del modelo a la estructura de los datos. En esta dirección, se presenta a continuación una breve descripción del estado del arte sobre el uso de medidas para evaluar y mejorar el rendimiento de los modelos de clasificación en la predicción de la tendencia de los precios de los activos.

Las medidas de desempeño son métricas cuantitativas que ayudan a evaluar y mejorar de manera objetiva el rendimiento de los modelos predictivos. Generalmente, la predicción de la evolución de los precios de los activos se ha tratado como un problema de clasificación de aprendizaje supervisado [38, 537, 499]. Estos movimientos direccionales se han clasificado tradicionalmente en dos escenarios, tendencia alcista y tendencia bajista [487]. Aunque algunos estudios también han incluido y explorado los movimientos laterales como parte de un problema de clasificación multiclase [330, 329, 337, 340, 345]. Las métricas utilizadas, en la medición del desempeño de los modelos de clasificación, se calculan con los valores de la matriz confusión [538]. A saber, verdaderos positivos (Tp), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (Tn) (Para más detalles, véase la sección 7.2 Medidas de desempeño).

Tabla 3.3: Medidas de desempeño de la clasificación.

[a] Autor	Predicción		Medidas	Acc
Das et al.(2022) [325]	A/B	Diaria	Acc	0.99*
Dash et al.(2022) [326]	A/B	Diaria	-	-
Ortu et al.(2022) [324]	A/B	Diaria/Intradía*	Acc, F1-S, Pr, Recall	0.84*
Hao et al.(2022) [315]	-	Intradía	DA	0.73*
Padhi et al.(2022) [327]	A/B	Diaria	Acc, F1-S, Pr, Recall, HL	0.99
Ozer et al.(2022) [328]	A/B	Diaria/Intradía	Acc, ROI	0.63
Sadeghi et al.(2021) [329]	A/B/L	Diaria	Acc, Recall, ROI, Drawdown	0.80
Uras et al.(2021) [330]	A/B/L	Intradía	Acc, F1-S, Pr, Recall, SD	0.54*
Ampomah et al.(2020) [331]	A/B	Diaria	AUC-ROC, W Test, Acc, Pr, F1-S	0.85
Kwon et al.(2019) [332]	A/B	Intradía	F1-S*, Pr, Recall	0.68*
Fischer et al.(2018) [333]	A/B	Diaria	Rm/d, SR, SD, CP/d, Acc	0.54
Hu et al.(2018) [334]	A/B	Diaria	HR	0.89
Bustos et al.(2017) [335]	A/B	Diaria	Acc	0.78*
Chakraborty et al.(2017) [336]	A/B	Diaria	Acc, F1-S, Pr, Recall	0.79*
Coyne et al.(2017) [337]	A/B/L	Diaria	Acc	0.78*
Dingli et al.(2017) [338]	A/B	Mensual	Acc	0.65
Huang et al.(2017) [339]	A/B	Intradía	Acc, Pr	0.55 - 1
Dang et al.(2016) [340]	A/B/L	4 meses	Acc, F1-S, Pr, Recall	0.73
Di et al.(2016) [341]	A/B	Diaria	Acc	0.55
Ghanavati et al.(2016) [342]	A/B	Mensual	Acc, ER, F1-S	0.85*
Chai et al.(2015) [343]	A/B	Diaria	HR	0.79
Dash et al.(2015) [344]	A/B	Diaria	Acc, F1-S, Pr, Recall	0.80*
Feuerriegel et al.(2015) [345]	A/B/L	Publicación	Acc, F1-S, Pr, Recall	0.56
Gonzalez et al.(2015) [346]	A/B	Semanal	Acc	0.72*

[a]Predicción: A = Tendencia Alcista, B = Tendencia Bajista, L = Tendencia Lateral. Medidas de desempeño: Acc = Accuracy, F1-S = F1-Score, Pr = Precision, DA = Directional accuracy, HL = Hamming Loss, ROI = Return on investment, Recall = Sensitivity, SD = Standard Deviation, W Test = Kendall W Test, Rm/d = Mean return per period, SR = Sharpe Ratio, CP/d = Cumulative payouts on average invest per period, HR = Hit Ratio, ER = Error Rate. * = Mejor desempeño obtenido en la clasificación.

Entre las medidas de evaluación del rendimiento más utilizadas, según las publicaciones de los últimos años, se encuentran Accuracy (Exactitud) [335, 338, 341], Recall (Sensibilidad) [340, 345], F1-Score [327, 332, 344] y Precision [324, 336, 339]. Otros estudios han recurrido al uso de métricas más especializadas, como las mencionadas en la Tabla 3.3, para resaltar la magnitud de los resultados obtenidos. Entre ellas cabe citar las siguientes: Proporción de aciertos HR [334, 343], Retorno medio por periodo Rm/d [333], Ratio de Sharpe SR [539, 333], Desviación estándar de los retornos SD [330, 333], Pagos acumulativos por inversión media diaria CP/d [333], Tasa de error ER [342], Retorno sobre la inversión ROI [328, 329], Reducción de fondos Drawdown [329], Precisión direccional DA [315] y Perdida Hamming HL [327].

Aunque no hay unanimidad en cuanto al tipo de métricas que deben utilizarse de forma estandarizada para medir el rendimiento de los modelos de clasificación, cada estudio define a discreción los criterios que mejor informan los resultados obtenidos con sus modelos. Algunos estudios más especializados han utilizado otros indicadores con fines específicos. G-mean es una medida de precisión particularmente útil para evaluar el rendimiento de modelos entrenados con clases desequilibradas [540]. Además de ser poco utilizada en este ámbito en comparación con otros dominios, [541] simplemente hace uso de este criterio para medir y comparar el rendimiento de 14 clasificadores ensemble para predecir los rendimientos de las acciones de la bolsa de Teherán (TSE).

Kappa es un Indicadores estadístico que compara el desempeño del modelo propuesto con respecto a un modelo de referencia [542]. Al revisar el desempeño del meta-clasificador (GBM) propuesto por [543] para predecir los movimientos bursátiles de la bolsa de Nairobi, los resultados obtenidos dejan algunas dudas. Aunque se alcanza una precisión de 0.78 y un área bajo la curva ROC de 0.82, el valor Kappa de 0.55 lejos de 1, sugiere un rendimiento aceptable del modelo propuesto con respecto a un modelo de referencia, como puede analizarse en [542]. Otro indicador bastante interesante y poco utilizado en este campo es el coeficiente de correlación de Matthews (MCC) [544]. Se trata de un indicador de referencia que produce puntuaciones más altas cuando los valores de la matriz confusión son correctos. Citando las conclusiones de [545], tras utilizar este índice como medida sintética de la calidad de las previsiones obtenidas en el mercado bursátil estadounidense. Constata que ni siquiera el uso de un gran conjunto de datos y de sofisticadas herramientas de aprendizaje automático profundo conduce a un resultado satisfactorio en la práctica. Por otra parte, el uso de pruebas no paramétricas para el análisis de correlaciones entre rangos de observaciones está ganando aceptación en los análisis. [546] evalúa la correlación entre el perfil de un usuario de Twitter y la popularidad de las criptomonedas mediante el coeficiente W de Kendall. [331] utiliza este estadístico como medida de concordancia para clasificar el rendimiento de los algoritmos utilizados en la predicción de índices bursátiles y valores del mercado estadounidense. Entre otras medidas se encuentra el área bajo la curva (AUC) [331], una medida ampliamente aceptada para evaluar la precisión de los modelos en problemas de clasificación. [547, 548, 549] encontraron que a medida que aumenta este valor, también lo hace la capacidad predictiva de los modelos. Otra medida utilizada para medir el rendimiento de los modelos cuando hay clases desequilibradas es la pérdida de Hamming (HL). Aunque técnicamente es una proporción de casos mal clasificados, en la práctica refleja el impacto del desequilibrio de clases en el

entrenamiento del modelo [327].

Varios estudios han revelado asimismo que la precisión de la clasificación depende también del número de periodos históricos considerados en la predicción. En [394] se observa que a medida que aumenta el número de datos que preceden al periodo de predicción, mejora el rendimiento de las métricas de precisión y F1-Score. Este efecto positivo puede deberse en parte a que, a medida que aumenta el número de periodos históricos, disminuye la variabilidad de los datos y mejora la precisión de la clasificación. [394] encontró mejores resultados de precisión en la clasificación tomando precios históricos de 17 a 20 periodos anteriores al periodo de estimación. Según los análisis realizados por los autores, a medida que aumenta el número de periodos históricos por encima de este umbral, se produce un proceso de degeneración en la precisión de la clasificación. [394].

Algunos estudios empíricos han demostrado que un mayor índice de precisión en la predicción no siempre produce mayores retornos en los modelos de negociación, [328] descubrió que una precisión en la predicción del 63.3% no siempre produce mayores rendimientos. Esto demuestra que las pautas de entrada y salida del mercado deben validarse primero con estrategias de negociación fiables y consistentes en el tiempo y, luego, basándose en esa información, se predice la dirección de los precios de los activos. En este sentido, los patrones de precios que generan rendimientos positivos y que han sido validados a lo largo del tiempo son susceptibles de predicción y, por tanto, pueden proporcionar al menos los niveles de rentabilidad que han producido históricamente.

La evidencia empírica demuestra que las estrategias agregadas, que integran modelos de negociación y de previsión, son más eficaces que los enfoques basados únicamente en modelos de previsión. Los autores de [327] demuestran que la previsión de activos de bajo riesgo y alto rendimiento elegidos con la teoría de carteras de Markovitz [143] mejora la precisión de la clasificación (99%). Al analizar este enfoque, se observa que la selección de activos de baja volatilidad, si bien plantea un problema de desequilibrio de clases manejable [327], su elección es estratégica para la negociación y la previsión. El uso de estos activos reduce el sesgo de previsión y, en consecuencia, aumenta la precisión de la clasificación, dada la baja volatilidad del precio del activo. Por lo tanto, al considerar activos más volátiles (de mayor riesgo), la previsión de la tendencia difícilmente puede alcanzar los mismos niveles de precisión que con activos menos volátiles. Estos resultados sugieren que el diseño de los modelos de negociación y de previsión deberían ser modelos agregados, es decir, además de compartir las mismas señales de mercado, los modelos de negociación deberían validar su viabilidad financiera, mientras que los

modelos de previsión, utilizando los mismos patrones de precios, deberían ser capaces de predecir y, en consecuencia, anticipar estas entradas y salidas del mercado para activos de diferentes niveles de riesgo.

Varios estudios han sugerido que la elección y el ajuste correcto del método al comportamiento de los datos, dada la estructura y tipología de estos, proporciona un alto rendimiento en la clasificación. Además, el uso de información crítica y actualizada también tiene un impacto significativo. [324] mejoró la precisión de la predicción de tendencias de criptomonedas al incluir información adicional de redes sociales e indicadores técnicos en los modelos entrenados con datos de mercado. Estos resultados permiten señalar que un diseño específico y un ajuste fino en la arquitectura de los modelos de inteligencia artificial, si bien mejoran el rendimiento de la clasificación, la ilegibilidad de estos enfoques no da cuenta de la influencia que las variables de entrada tienen en la evolución de los precios para predecir la tendencia del mercado.

En resumen, los estudios empíricos han demostrado el uso preferente de cuatro medidas de rendimiento, entre las que se incluyen Accuracy, Recall, Precision y F1-Scores. Aunque la elección adecuada de las medidas de rendimiento viene determinada en parte por el problema de previsión a resolver, el uso de estos criterios de evaluación no es infalible para construir y mejorar la eficacia y precisión de los modelos de clasificación. Estas medidas sólo revelan el grado de ajuste generado por la técnica de clasificación a la estructura de los datos e implícitamente revelan la capacidad discriminativa y predictiva de las variables utilizadas en la construcción. Por lo tanto, el nivel de ajuste del modelo no depende del tipo de medidas utilizadas en la evaluación, sino de las variables y la técnica empleadas en la construcción del modelo de clasificación en función de la naturaleza de los datos.

Desde hace varios años crece el interés de los participantes en la industria financiera por comprender el panorama futuro del mercado y reducir el riesgo de inversión. La previsión de la dirección de los precios de los activos se ha intensificado con la construcción de modelos predictivos más precisos. La complejidad de los datos, su tipología y sus características asimétricas y leptocúrticas han impulsado cada vez más el uso de algoritmos de inteligencia artificial. Sin embargo, la construcción de modelos más precisos sin comprometer su legibilidad es una necesidad urgente. Se han realizado numerosos estudios para intentar descubrir las mejores variables, mejorar la calidad de los datos, utilizar mejores técnicas de preprocesamiento y definición de parámetros y descubrir los modelos que mejor se ajustan a los datos para lograr una mayor precisión en la

predicción. La revisión de la literatura confirma que todos estos factores influyen en la búsqueda del modelo de predicción más preciso.

3.6 Consideraciones generales

En resumen, un análisis adecuado de los datos suele proporcionar información valiosa, y por lo general es evidente cuando se utilizan datos de calidad. Hay muchos problemas que pueden deteriorar la calidad de los datos y los resultados generados en la construcción de modelos predictivos. La redundancia es uno de ellos, además de afectar a la integridad, la coherencia y la calidad de los datos, la integridad de los datos puede producir resultados erróneos.

El problema asociado al uso de datos redundantes, por un lado, no radica en el aumento de la dimensionalidad, sino en la contribución al riesgo de inconsistencia y pérdida de integridad de los datos causado por la actualización de la información. Por otro lado, en la construcción de modelos predictivos, el uso de variables con información redundante y altamente correlacionada puede provocar problemas de multicolinealidad y, por tanto, generar problemas de inestabilidad con el uso de algunos modelos. Esto ha impulsado el desarrollo de medidas de mejora, en primer lugar, para aumentar y preservar la calidad de los datos, desde el momento en que se producen, transfieren y preparan hasta que se almacenan y, en segundo lugar, el diseño de métodos para identificar y eliminar las variables con valores irrelevantes y redundantes.

La dimensionalidad de los datos como concepto no es un problema para su tratamiento, selección y análisis. Sin embargo, para algunas técnicas de análisis estadístico multivariante cuando en matrices de datos el número de variables supera al número de observaciones esta condición puede contribuir a la ocurrencia de cierto tipo de problemas (singularidad).

Aunque los grupos de clases con instancias desequilibradas no plantean ningún problema en sí mismos [394], algunos estudios empíricos señalan que estos desequilibrios crean grandes dificultades en los algoritmos de aprendizaje automático. Frente a esta condición, se han propuesto múltiples soluciones para ayudar a mejorar el desempeño de los modelos en las tareas de clasificación [397]. Por otro lado, el avance de los métodos heurísticos en la definición y optimización de los hiperparámetros de los métodos de aprendizaje automático también ha contribuido notablemente a mejorar la precisión en la clasificación.

La selección de características es, en definitiva, un proceso que ayuda a filtrar las

variables de entrada de acuerdo con criterios de selección particulares que resultan útiles para la construcción de modelos de clasificación. En este contexto, la literatura científica revela con estos estudios un profundo desconocimiento del mercado y de la influencia que las variables técnicas y fundamentales pueden tener en el comportamiento de la acción del precio para estimar su dirección futura. Los enfoques propuestos se basan en la extracción de nuevas características para generar modelos de predicción más precisos, lo que compromete la legibilidad e interpretabilidad de los modelos de clasificación. Así, resulta difícil comprender cómo influyen estas variables de entrada en los movimientos de los precios para predecir y explicar su dirección futura.

La mayoría de los problemas que surgen durante la formulación de los modelos predictivos es que están relacionados con la naturaleza de los datos, y las soluciones que se aplican, proporcionadas por los métodos de preprocesamiento y las actividades de preparación de los datos, dependen en gran medida de las peculiaridades de cada estudio. Además, las actividades de preprocesamiento y preparación de datos permiten el uso conjunto de varios métodos de predicción, que de forma complementaria conducen a modelos mejorados y producen resultados más precisos. Los modelos conjuntos y los métodos híbridos no se limitan en cierto modo a los datos, sino que pueden utilizarse en la medida en que el preprocesamiento y la preparación de los datos, dada su naturaleza, lo permitan las distintas técnicas.

El uso de métodos de aprendizaje profundo en los últimos años ha adquirido cada vez más relevancia en la construcción de modelos predictivos. Los considerables avances en las técnicas de inteligencia artificial han influido en la construcción de modelos ensemble e híbridos con niveles de precisión cada vez más prometedores. También es cada vez más frecuente diseñar y aplicar métodos heurísticos y modelos de optimización en la definición de los parámetros del modelo. El escalado y la normalización de los datos es una prioridad con el uso de redes neuronales artificiales para garantizar un entrenamiento eficaz y mejorar la precisión de la predicción. Por último, los modelos de previsión que han alcanzado índices de precisión superiores al 90% son los que mejor se ajustan a los datos y, por tanto, tienen potencial para producir resultados más fiables en la práctica. Esto es especialmente cierto porque en entornos inciertos en los que la volatilidad de los precios ejerce una fuerte influencia, alcanzar altos niveles de precisión en las previsiones se hace más difícil. Además, esto también sugiere un mayor esfuerzo en la formulación de modelos predictivos parsimoniosos, legibles y más precisos. Incluso esto plantea un mayor desafío a los enfoques basados en inteligencia artificial, cuando se decida superar el problema de la caja negra, con la formulación de

modelos de clasificación legibles e interpretables.

4. Marco teórico

4.1 Introducción

Predecir la dirección del tipo de cambio ha sido siempre uno de los mayores retos para los agentes implicados en las negociaciones financieras. Aunque muchos trabajos basados en el uso de técnicas de inteligencia artificial han explorado este tema, lo cierto es que, en aras de mejorar la precisión de las previsiones, se ha prescindido de la legibilidad de las interpretaciones.

Teniendo en cuenta este contexto, a continuación se presenta una breve fundamentación teórica sobre los principales componentes a considerar para abordar el problema objeto de estudio. La originalidad y novedad de la metodología en la que se basa esta investigación se fundamenta en la negociación algorítmica de los tipos de cambio discutida en la sección 4.5, la selección de características tratada en el apartado 4.6, el uso del análisis biplot explicado en la sección 4.7 y la aplicabilidad del análisis discriminante tratada en la sección 4.8. Este desarrollo teórico constituye el punto de partida hacia la construcción de un modelo de clasificación que permite predecir, en el corto plazo, la tendencia del tipo de cambio euro-dólar.

Además, el desarrollo teórico de este trabajo se basa en el tipo de cambio abordado en la sección 4.2, el proceso de formación de precios de mercado presentado en la sección 4.3, y los enfoques de resolución de problemas presentados en la sección 4.4. El acercamiento a estos fundamentos teóricos proporciona el soporte teórico necesario para tratar los problemas planteados en esta tesis.

4.2 Tipo de cambio

El mercado de divisas es uno de los mercados financieros más volátiles y dinámicos del mundo. Factores económicos, técnicos y de comportamiento influyen en el movimiento de los tipos de cambio. Estas características han dificultado la predicción de la dirección de los movimientos en este tipo de instrumento financiero. De ahí que predecir la

dirección de los tipos de cambio se haya convertido en un reto tanto para académicos como para la industria financiera.

En esta sección se analizan algunos fundamentos teóricos relacionados con el tipo de cambio euro-dólar. Estos aspectos deben tenerse en cuenta en la previsión de la dirección de los precios. Esta sección consta de tres partes que se describen a continuación.

4.2.1 Fundamentos

El tipo de cambio es el precio de una divisa en relación con otra. Básicamente, hay dos formas de obtener el valor del tipo de cambio, al contado (*Spot*) y a plazo (*Forward*). El tipo de cambio al contado es el valor actual de una divisa en comparación con otra. Este tipo suele utilizarse en transacciones inmediatas y su valor está prácticamente establecido en el mercado de divisas. La expresión 4.1 determina el valor del tipo de cambio al contado.

$$Tipo\ de\ cambio\ de\ contado\ (spot) = \frac{Precio\ moneda\ base\ C_b}{Precio\ moneda\ cotizada\ C_c} \tag{4.1}$$

De acuerdo con la formula anterior, la divisa base C_b corresponde a la divisa que se compra o se vende, y la divisa que se cotiza C_c es la divisa utilizada en la transacción.

Por su parte, el tipo de cambio a plazo, negociado hacia futuro (*forward*), corresponde al valor convenido con el que se negocia una divisa con respecto a otra para una determinada fecha en el futuro. Este tipo de operación se realiza para evitar y protegerse de la volatilidad del mercado con el fin de anticiparse a sus posibles fluctuaciones en el futuro. Matemáticamente el tipo de cambio hacia futuro F para el tipo de cambio (C_b/C_c) se obtiene a partir de la ecuación 4.2:

$$F = \text{Spot} + T(\mathcal{I}(C_c) - \mathcal{I}(C_b)) \tag{4.2}$$

Obsérvese que el tipo de cambio hacia futuro F se obtiene en función del diferencial entre el tipo de interés de la moneda base $\mathcal{I}C_b$ y el tipo de interés de la moneda contraparte $\mathcal{I}C_c$ multiplicado por el tiempo T hasta el cual ocurre el vencimiento del contrato a plazo. Este procedimiento de cálculo es útil para los inversores que operan en el mercado de divisas. En efecto, les permite calcular el ingreso o el coste que esperarían obtener si el tipo de cambio se materializa en una fecha futura. Si el tipo de interés de la divisa cotizada C_c es superior al de la divisa base C_b, entonces se obtiene

una prima $forward$ (a plazo). En caso contrario, si el tipo de interés $forward$ (a plazo) es inferior al tipo de interés *spot* (al contado), entonces se obtiene un descuento $forward$ (a plazo).

El análisis de los tipos de cambio al contado y a plazo permite a los analistas técnicos detectar pautas y tendencias en el mercado de divisas. Además, esta información les permite tomar decisiones de inversión, especialmente cuando el tipo de cambio a plazo es superior al tipo de cambio al contado. Esta forma de predecir la tendencia del tipo de cambio puede indicar que existen expectativas de que la divisa se aprecie en el futuro, lo que puede representar una oportunidad de compra en el momento actual.

Los inversores y especuladores también pueden utilizar el análisis de los tipos de cambio al contado y a plazo junto con otros indicadores técnicos. Por ejemplo, los niveles de soporte y resistencia, así como los indicadores de impulso, pueden utilizarse simultáneamente para confirmar posibles decisiones de inversión y aumentar las probabilidades de éxito.

4.2.2 Régimen cambiario

El régimen cambiario corresponde al conjunto de reglas y normas que regulan el intercambio de divisas a nivel internacional. Este intercambio internacional de divisas comprende todas las transacciones que implican el pago y la transferencia de moneda extranjera o de valores representativos de dichas monedas. Existen tres tipos de regímenes cambiarios, el régimen de cambio flotante, el fijo y el semi-fijo.

A. Régimen de cambio flotante

Este tipo de cambio se define por la acción de la oferta y la demanda en el mercado de divisas. En este tipo de régimen, además de que el precio de las divisas fluctúa libremente, los bancos centrales no intervienen en la definición de su valor. Este tipo de cambio es muy común en economías desarrolladas como Estados Unidos, la zona euro y Japón.

B. Régimen de cambio fijo

Las autoridades monetarias de un determinado país definen este tipo de cambio frente a otras monedas. En este tipo de régimen, además de que los bancos centrales intervienen activamente en el mercado, sus acciones están encaminadas a mantener un tipo de

cambio oficial dentro de un nivel más o menos constante. Este tipo de regulación es muy común en países con economías emergentes, como Panamá y China.

C. Régimen cambiario semifijo

El valor del tipo de cambio viene definido por la actuación de las autoridades monetarias de un país en particular. Este régimen cambiario se define dentro de una banda de precios en la que el tipo de cambio puede fluctuar libremente en función de la oferta y la demanda del mercado. Tradicionalmente, este tipo de régimen es útil para aquellos países cuyas economías se encuentran en un periodo de transición hacia la definición de un régimen cambiario flotante.

Por ultimo, cabe señalar que Los tipos de cambio de las principales economías, como Estados Unidos y la zona del euro, se basan en un régimen de tipo de cambio flotante. Aunque los bancos centrales de ambas economías pueden intervenir en determinadas condiciones, generalmente no lo hacen para no influir en el valor del tipo de cambio. En este caso, el valor del par de divisas eur/usd se define principalmente por la acción de la oferta y la demanda del mercado.

4.3 Formación de precios de mercado

El precio de mercado de un activo financiero es el valor al que los agentes pueden adquirir ese activo en un momento dado. Aunque la definición de este valor se establece mediante un proceso de interacción entre oferentes y demandantes, el valor del activo viene determinado por una serie de factores fundamentales, técnicos y conductuales. En la misma línea, si bien la moneda es la unidad de valor que representa la soberanía monetaria de un país, el valor que asume con respecto a sus divisas homólogas varía en función de muchos factores. Entre los factores macroeconómicos más importantes están el nivel de actividad económica de las naciones que representan, la inflación, el consumo interno, los tipos de interés, la estabilidad política, la deuda pública, la salud económica, el déficit por cuenta corriente, la especulación y la balanza comercial, entre otros aspectos. Por tanto, la formación del precio de mercado al que se negocia un tipo de cambio, además de considerar los aspectos mencionados, refleja en el precio de mercado el punto de equilibrio entre oferentes y demandantes alcanzado durante la sesión de negociación. Matemáticamente, si el tipo de cambio se denota como (C_b/C_c), donde C_b es la moneda base y C_c la moneda de contrapartida, el valor a pagar con la

moneda base C_b por una unidad de la moneda de contrapartida C_c viene definido por la siguiente expresión:

$$C_c \quad = C_b \cdot \frac{1}{(C_b/C_c)} \tag{4.3}$$

De este modo, al revisar el valor del tipo de cambio del euro frente al dólar estadounidense (EUR/USD). Si, por ejemplo, el valor cotizado para un período de tiempo dado es 1.06652, significa que por cada euro de la divisa base C_b se negocian 1.06652 dólares estadounidenses de la divisa contraparte C_c. Esto significa que un dólar estadounidense puede intercambiarse por 0.93783 euros. En consecuencia, el valor asumido por el tipo de cambio (C_b/C_c) obedece a un proceso dinámico y continuo, cuya formación de precios puede cambiar en función de la evolución de las circunstancias que le afectan. Además, observando las variaciones entre los precios de cierre *cp* y de apertura *op* a lo largo del tiempo t, el valor asumido por el tipo de cambio puede describir movimientos tendenciales al alza o a la baja. La expresión 4.4 permite detectar, en términos generales, si se ha producido un aumento o una disminución del valor de la moneda base C_b con respecto al valor de la moneda de contrapartida C_c. Nótese que este valor se calcula entre el precio de apertura *op* y el precio de cierre *cp* en un intervalo de tiempo t en el que se produce la negociación.

$$\Delta ER_t = \begin{cases} cp_t - op_{t-1} > 0, & \text{Tipo de cambio apreciado} \\ cp_t - op_{t-1} < 0, & \text{Tipo de cambio depreciado .} \end{cases} \tag{4.4}$$

Donde ΔER_t, determina la variación del tipo de cambio (C_b/C_c) al final del periodo de negociación, entre el instante t y el periodo inicial $t-1$. Estas variaciones, definidas en estados de apreciación y depreciación, pueden reconocerse fácilmente en los gráficos de precios basados en velas japonesas. En consecuencia, en las Figuras (4.1a) y (4.1b), dado el proceso de formación de una vela alcista y bajista, dicho estado de apreciación o depreciación del tipo de cambio se reconoce fácilmente por el color de la vela.

Las velas japonesas son una herramienta de análisis técnico que se utiliza para representar gráficamente los precios de mercado de un instrumento financiero tras su negociación dentro de una sesión bursátil. Tradicionalmente, una vela alcista se indica en color blanco o verde. Esto significa, de acuerdo con la ecuación 4.4, que tras finalizar la sesión bursátil el precio de cierre *cp* es mayor que el precio de apertura *op*. Por el contrario, una vela bajista roja o negra adquiere esta connotación tras el final de la sesión bursátil, cuando el precio de cierre *cp* es inferior al de apertura *op*. Además, en la

Figura 4.1 se detallan algunos patrones de velas que ayudan a reconocer algunas pautas en los precios que, una vez confirmadas, pueden provocar un cambio de dirección en la evolución del precio y, por tanto, su reconocimiento es fundamental desde el punto de vista del análisis técnico antes de proceder a tomar decisiones de inversión.

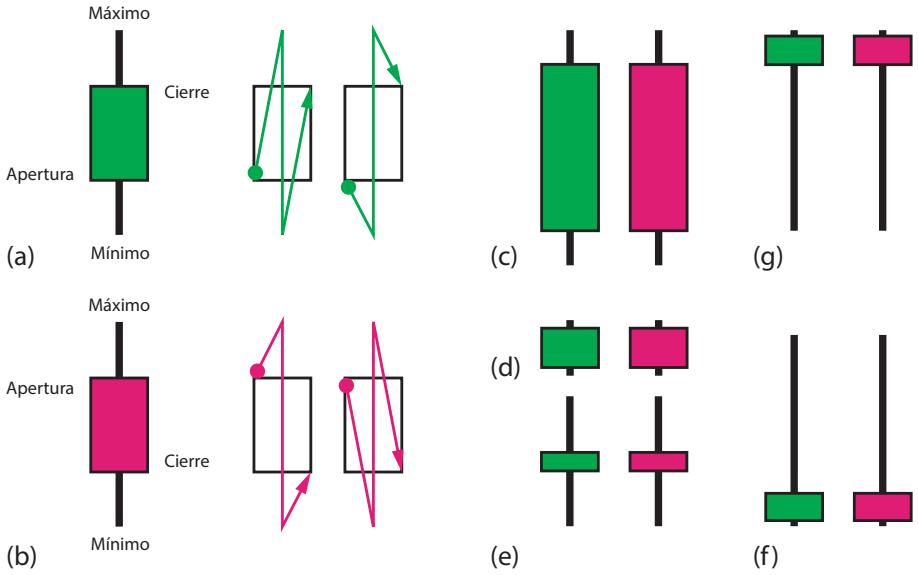


Figura 4.1: Formación de precios y representación en velas.
Patrones de velas que ayudan a reconocer pautas en los precios para predecir cambios de dirección. (**a**) Formación de precios en velas alcistas *"apreciación"*. (**b**)Formación de precios en velas bajistas *"depreciación"*. (**c**) Velas alcistas y bajistas grandes. (**d**) Velas alcistas y bajistas de cuerpos y sombras pequeñas *"Doji"*. (**e**) Doji de sombras largas. (**f**) Martillo invertido y estrella fugaz. (**g**) Martillo y hombre colgado.

Note que el análisis técnico, basado en el uso de velas japonesas, resume en una sola vela los precios de mercado $OHLC$ obtenidos durante la sesión de negociación realizada en un marco de tiempo k. Las Figuras (4.1a) y (4.1b) detallan el proceso de formación de una vela alcista (verde) y bajista (roja) en una sesión de negociación k. Estas sesiones, también denominadas marcos temporales de negociación, pueden ser de 1, 5, 15 y 30 minutos; 1 y 4 horas; y 1 día, 1 semana y 1 mes. Además de la forma de la vela, el proceso descrito en su formación permite identificar el papel asumido por oferentes y demandantes en cada sesión de negociación. Esta información, dadas las señales producidas en el mercado, es útil para comprender el contexto de la negociación y ayuda a confirmar patrones en el movimiento de los precios antes de tomar una decisión de inversión.

La formación de velas con cuerpos largos y sombras superior e inferior cortas, como

se muestra en la Figura (4.1c) indica que cuanto mayor sea la longitud de la vela, más fuerte será la presión alcista o bajista según el caso. En consecuencia, una vela alcista de gran longitud ayuda a validar los niveles de soporte históricos y, por lo general, a romper los niveles históricos de resistencia. Una vela larga bajista ayuda a confirmar los niveles de resistencia históricos o rompe los niveles históricos de soporte. Para más información sobre las zonas de soporte y resistencia, véase la sección 4.3.2.

La conformación de velas de cuerpos pequeños y con sombras superior e inferior cortas, conforme se muestra en la Figura (4.1d) es indicativo de que mientras más corta sea la vela, menor será el movimiento del precio y, por lo tanto, se producirá la consolidación o movimiento lateral del mismo. Al analizar la falta de movimiento del precio puede ser una indicación de una pausa temporal en el movimiento del mercado o un cambio en la dirección de la tendencia del precio. Por otra parte, la definición de velas de cuerpo pequeño con largas sombras superior e inferior, como se muestra en la Figura (4.1e), podría sugerir indecisión en el mercado. Esto significa que tanto oferentes como demandantes controlaron la negociación durante la sesión. Lo que sugiere que el precio del mercado puede haber alcanzado un punto de equilibrio momentáneo antes de continuar o cambiar de dirección.

La formación de velas con cuerpos pequeños y sombras superiores largas, como se muestra en la Figura (4.1f), denota durante la sesión un control por parte de los compradores, cuya influencia mueve al alza el precio de las cotizaciones. Sin embargo, al final de la sesión, los vendedores intervienen y ejercen el control sobre el mercado, lo que da lugar a la formación de una larga sombra superior. Esta señal podría sugerir, tras un largo movimiento al alza, un cambio en la dirección de los precios. Esto implica la recogida de beneficios para las posiciones largas o la apertura de posiciones cortas.

Las formaciones de velas con cuerpos pequeños y sombras inferiores largas, como se muestra en la Figura (4.1g), denotan que los vendedores tenían el control de la operación durante la sesión. Sin embargo, al final de la sesión, los compradores intervinieron y consiguieron impulsar el precio al alza, cerrando la vela por encima del precio de apertura. La aparición de estas señales, en el mercado, sugiere prestar atención a las zonas de soporte y resistencia cercanas. En este sentido, si un patrón alcista de este tipo se forma cerca de la zona de soporte, puede ser una señal de que el precio seguirá subiendo. Por el contrario, si un patrón alcista aparece cerca de la zona de resistencia, puede sugerir que el precio se moverá a la baja.

4.3.1 Régimen de mercado

El régimen de mercado es el conjunto de condiciones o circunstancias persistentes que influyen en el comportamiento del mercado y, por tanto, pueden afectar al rendimiento de las estrategias de negociación e inversión. En efecto, en la actualidad, el análisis de las condiciones del mercado permite reconocer su comportamiento a lo largo del tiempo antes de invertir. Tradicionalmente, se ha observado que el nivel de actividad de una economía se refleja en el comportamiento del mercado bursátil. Aunque no existe una teoría que vincule el nivel de actividad económica con el mercado bursátil, la comparación entre el sector real de la economía y el mercado financiero se ha basado principalmente en enfoques empíricos que han ayudado a determinar el estado actual de los mercados.

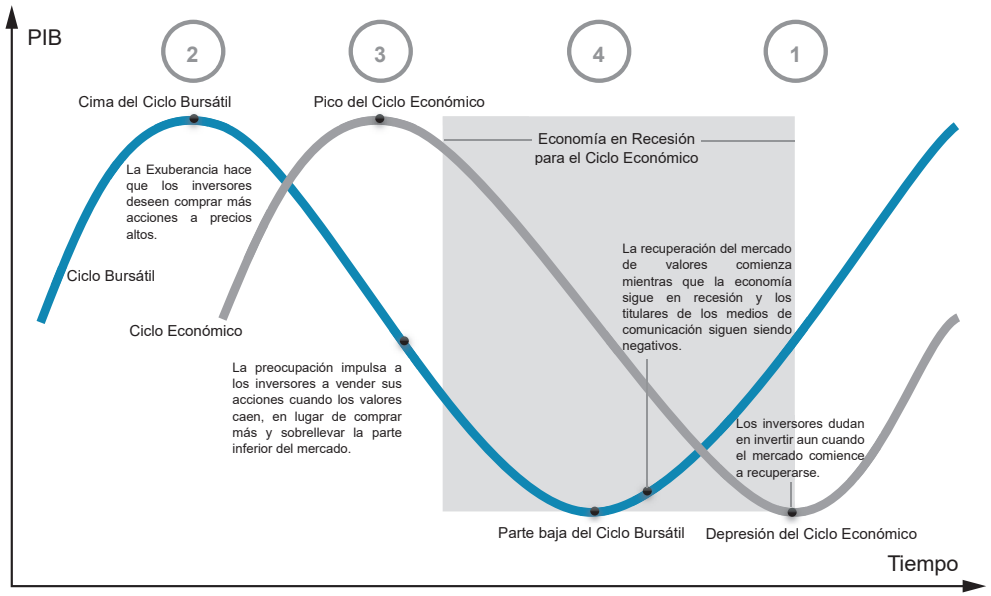


Figura 4.2: Ciclos bursátil y económico.
El ciclo bursátil (azul) como indicador adelantado del ciclo económico (gris) permite predecir el régimen de mercado de la economía dentro de su ciclo económico real. Este fenómeno es especialmente importante durante las fases de recesión y recuperación del ciclo bursátil. La característica más destacada en ambos ciclos son los puntos de inflexión (máximos y mínimos). (**4**) precede a (**1**) finaliza la fase de desaceleración e inicia la fase de recuperación de la economía real. (**2**) precede a (**3**) termina la fase de crecimiento e inicia la fase de recesión real.

La Figura 4.2 describe teóricamente los co-movimientos entre series temporales. Las series azul y gris indican que el índice bursátil tiende a moverse junto con el producto interior bruto de una economía a lo largo del tiempo. Aunque estas regu-

laridades empíricas llevan a pensar en la posible existencia de leyes o restricciones en el comportamiento de unas series temporales que influyen en el movimiento de otras, lo cierto es que la identificación y definición de las etapas de los ciclos económicos ayuda implícitamente a reconocer el tipo de régimen de mercado subyacente en el que se realizará la inversión.

En el caso de los tipos de cambio, los ciclos son patrones que registran los periodos en los que el tipo de cambio negociado se aprecia o deprecia a lo largo del tiempo. Así, los puntos más altos y más bajos de estos ciclos se reconocen como *"picos"* y *"valles"*. Es precisamente en estos puntos de inflexión en los que el régimen del mercado cambia con respecto a su tendencia anterior. Aunque en estas trayectorias la moneda base se aprecia o deprecia con respecto a la moneda contraparte, la cronología del ciclo y otras características como la duración, la amplitud y la asimetría pueden carecer hasta cierto punto de importancia. Sin embargo, es esencial identificar y validar las pautas en los precios, que confirman a través de estos puntos de inflexión, el cambio de dirección de la tendencia precedente. En este sentido, el reconocimiento del régimen de mercado se convierte en un aspecto crítico en la gestión y ejecución de las estrategias de negociación. Esto implica que el tipo de posiciones a desarrollar deben estar a favor de la tendencia dominante del mercado. Lo que supone que el analista debe identificar y validar tal condición no sólo desde el punto de vista técnico sino también considerando los principales índices macroeconómicos.

Entre los indicadores clave, que ayudan a identificar las distintas fases del ciclo económico, están el PIB, el nivel de empleo, la producción industrial, las ventas al por mayor y por menor y el ingreso real. Además de medir el nivel de actividad económica de un país, estos indicadores ayudan a determinar la fortaleza de su moneda en comparación con otras divisas. Por otra parte, como se observo en la Figura 4.2, los mercados financieros también experimentan un comportamiento similar al de los ciclos económicos y los indicadores fundamentales, como los tipos de interés, los tipos de cambio y la inflación, pueden influir en su comportamiento.

Este enfoque integral considera una de las muchas variables que deben tenerse en cuenta para evaluar el peso que el nivel de actividad económica tiene en el valor y el movimiento de las divisas. Sin embargo, el análisis de los ciclos económicos y bursátiles de las economías representadas por el tipo de cambio a negociar es un enfoque que queda fuera del alcance de este estudio, dado su carácter puramente fundamental. En cambio, la definición del régimen de mercado a partir del análisis de los precios históricos ha demostrado ser una herramienta analítica eficaz en la gestión de las estrategias de

negociación, especialmente en la ejecución de operaciones de mercado a corto plazo.

La Figura 4.3, representa el comportamiento histórico del tipo de cambio euro-dólar, según los precios de mercado medidos en sesiones mensuales de negociación, desde enero de 2006 hasta marzo de 2023. El análisis técnico del gráfico permite detectar, tras la apreciación del euro frente al dólar estadounidense experimentada entre 2006 y 2007, un claro cambio de régimen durante 2008. Desde ese año hasta la actualidad, se observa un régimen de mercado enmarcado en un canal bajista. Cabe destacar que la fluctuación del tipo de cambio ha descrito un movimiento en un rango bajista. Los puntos de inflexión situados cerca de los niveles que definen el canal han respetado estas zonas de soporte y resistencia.

Figura 4.3: Régimen de mercado tipo de cambio Eur/Usd.
Régimen de mercado en canal bajista del tipo de cambio Eur/Usd. Datos medidos en un marco de tiempo mensual, muestreados en un horizonte de tiempo de 17 años, desde enero del 2006 hasta marzo del 2023. Puntos de inflexión próximos a las zonas de soporte y resistencia definidas por el canal descendente.

En esencia, la detección y confirmación del régimen de mercado ayuda a aprovechar su situación evitando, por ejemplo, la negociación de instrumentos de alto riesgo, especialmente en condiciones de mercado desfavorables, al tiempo que se intenta identificar las mejores oportunidades de inversión. Así, la solidez de estas evaluaciones permite, a la vista de las señales proporcionadas por el mercado, definir el marco a considerar para llevar a cabo o no las inversiones.

Aunque la estrategia de las economías de mercado también puede contemplar la devaluación competitiva de su moneda para fomentar la inversión extranjera y aumen-

tar la demanda de instrumentos de capital, especialmente de las empresas locales, el objetivo último es impulsar los mercados financieros nacionales y su economía real. En este contexto, el régimen de mercado representado en la Figura 4.3, dado el comportamiento histórico de los precios durante el periodo de tiempo indicado, sugiere que el modelo de negociación debe ser lo suficientemente flexible y versátil como para adaptarse a los rápidos cambios en la dirección de los precios.

4.3.2 Niveles de soporte y resistencia

Los precios de los activos generalmente se mueven entre rangos de precios. Así, los niveles de soporte y resistencia pueden ser entendidos como patrones históricos de congestion en los que el precio actual del activo se encuentra con una zona de soporte (nivel de precios inferior) o resistencia (nivel de precios superior) que le impide avanzar. De manera que cuando el precio del activo se mueve en una tendencia bajista/alcista y cruza de arriba/abajo hacia abajo/arriba dicho nivel de soporte/resistencia, esta zona pasa a convertirse en una nueva zona de resistencia/soporte para el precio.

Los niveles de soporte y resistencia son indicadores de análisis técnico utilizados para identificar los niveles de precios históricos mínimos y máximos, respectivamente, que el valor del activo no ha podido superar en la actualidad. Estas zonas en las que se establecen rangos históricos, dentro de los cuales se mueven los precios de los activos, se han utilizado tradicionalmente como mecanismos de activación para la emisión de ordenes de compra o venta. El desempeño a obtener con la ejecución de este tipo de operaciones viene determinado en parte por el contexto en el que se realiza la negociación. Esto significa que, una vez identificado y validado el régimen de mercado dominante en el que se mueve el activo, las posiciones ejecutadas deben estar en línea con la tendencia dominante. En consecuencia, después de que el valor del tipo de cambio se haya movido en un rango dentro de las zonas históricas de soporte y resistencia, se puede activar una señal de compra cuando el precio del activo rompe el nivel histórico de soporte desde abajo hacia arriba, mientras que se activa una orden de venta cuando el precio del activo rompe el nivel histórico de resistencia desde arriba hacia abajo. Sin embargo, dependiendo del tipo de estrategia de negociación ejecutada, algunos operadores de mercado, tras abrir y mantener abierta una posición de venta o de compra, optan por cerrar dichas posiciones tomando beneficios cuando el valor del activo, tras acercarse a dichas zonas, no cruza los niveles históricos de soporte o resistencia.

Encontrar aquellos niveles en los que el precio cambia de dirección con respecto a la tendencia precedente para anticipar la dirección futura de los precios y obtener beneficios se ha convertido en una de las tareas más desafiantes. Así, el uso de información histórica se ha convertido en una valiosa herramienta para confirmar la emisión de órdenes y evitar falsas señales de mercado. Una de las estrategias de negociación más utilizadas por los agentes del mercado consiste en aprovechar la ruptura de los niveles de soporte y resistencia, históricamente respetados por el movimiento de los tipos de cambio, para generar operaciones de venta y compra, según el caso.

La Figura 4.4 muestra cómo el movimiento del precio del tipo de cambio euro-dólar, según sesiones de negociación de 15 minutos, conduce a la formación de zonas de soporte S_i y resistencia R_i. Una verdadera zona de soporte y resistencia, históricamente, se forma en la media en que el precio del instrumento negociado las respeta. Esto significa que cada vez que el precio del activo intenta cruzar estas zonas y no lo consigue, la fuerza de estos niveles de precios queda perfectamente demostrada.

Figura 4.4: Soporte y resistencia del tipo de cambio euro-dólar.
Precios de mercado del tipo de cambio eur/usd. Datos medidos en sesiones de negociación de 15 minutos, desde el 31 de enero de 2023 hasta el 10 de febrero de 2023. Niveles de resistencia R_i y soporte S_i en colores verde y rojo respectivamente. Nótese que el nivel de resistencia R_1 después de ser superado de abajo hacia arriba por el tipo de cambio, se convierte en un nuevo nivel de soporte S_1. El cruce de estos niveles produce señales de compra o venta que pueden capitalizarse en función del régimen de mercado.

Por otro lado, cuando la fuerza con la que se mueve el precio del activo rompe estos niveles, las zonas de resistencia/soporte, tras ser cruzadas, se convierten en nuevas

zonas de soporte/resistencia. Este comportamiento ha demostrado históricamente que a medida que el tipo de cambio se aprecia o deprecia con el paso del tiempo, los máximos o mínimos globales históricos registrados por el precio del activo tienden a convertirse en niveles de resistencia y soporte con valores máximos y mínimos locales.

También hay que tener en cuenta que la ruptura de estos niveles de soporte y resistencia, y actuando de acuerdo con el régimen de mercado dominante, ofrecen oportunidades de entrada y salida del mercado que deben evaluarse a lo largo del tiempo. La fiabilidad y consistencia de estas pautas confieren a las estrategias de negociación que las utilizan fiabilidad y consistencia en los rendimientos generados.

4.3.3 Puntos de inflexión del mercado

Tradicionalmente, la negociación de tipos de cambio en el mercado de divisas siempre ha estado expuesta a la volatilidad e incertidumbre del entorno. La fluctuación de los tipos de cambio viene determinada en parte por múltiples factores económicos y políticos. Teniendo en cuenta estas consideraciones, los puntos de inflexión también pueden interpretarse como momentos decisivos en los que puede producirse un cambio de régimen en la dirección del mercado. En este sentido, los puntos de inflexión del mercado son técnicamente puntos de reversión en los que cambia la dirección de la tendencia dominante del mercado. Desde el punto de vista del análisis técnico, estos puntos de inflexión pueden reconocerse en los gráficos de precios históricos en aquellos momentos en los que cambia la dirección de la tendencia precedente.

La Figura 4.5 muestra la evolución del tipo de cambio euro-dólar, basada en los precios de mercado $OHLC$ medidos en marcos temporales de 15 minutos. Los datos de la muestra corresponden al periodo comprendido entre el 2 de febrero y el 15 de febrero de 2023. Los niveles de soporte históricos S_i en colores verde y rojo fueron superados por el impulso del tipo de cambio euro-dólar. Durante el periodo analizado, el tipo de cambio ha evolucionado en general en un régimen de mercado bajista.

Las señales sobre el tipo de cambio definidas por las puntas de flecha amarilla y roja son los momentos en los que cambia la dirección de las micro tendencias y pueden entenderse como puntos de giro. Al examinar el indicador RSI en la parte inferior del gráfico, los puntos de giro del tipo de cambio coinciden con la señal RSI en los niveles 30 y 60, respectivamente. Esto significa que cuando la señal RSI alcanza estos niveles, el tipo de cambio está sobrevendido y sobrecomprado, respectivamente. Por lo tanto, es probable que la evolución del tipo de cambio se agote y que las micro tendencias

anteriores cambien de dirección.

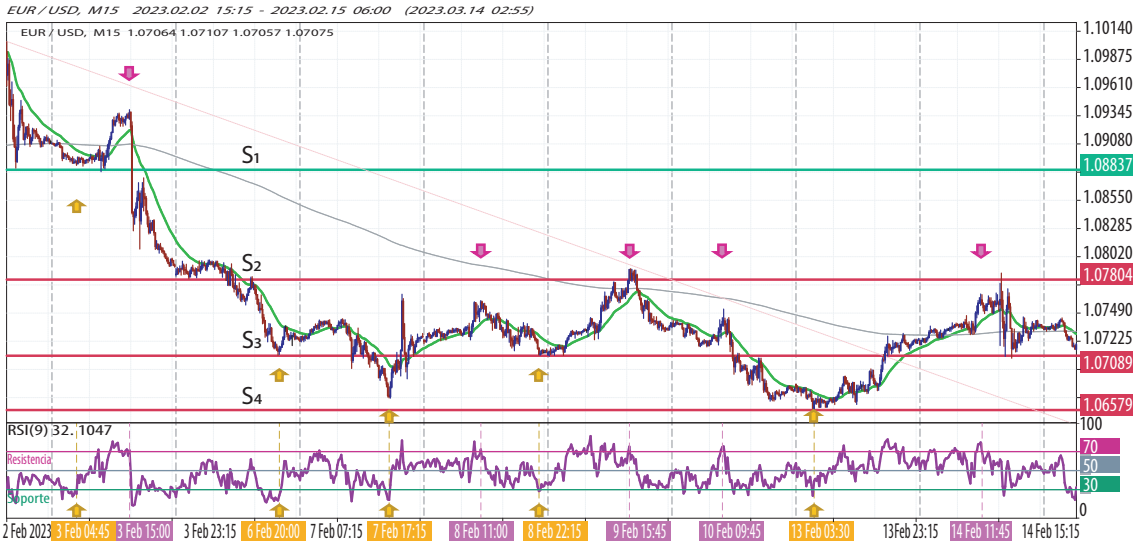


Figura 4.5: Puntos de inflexión.
Precios de mercado del tipo de cambio eur/usd. Datos medidos en sesiones de negociación de 15 minutos, del 2 al 15 de febrero de 2023. Niveles de soporte S_i en verde y rojo. La ruptura de los niveles de soporte S_i en un mercado bajista con movimientos en rango produce señales de compra y venta capitalizables. La señal *RSI* de 9 periodos confirma en los niveles 30 y 70 un tipo de cambio sobrevendido y sobrecomprado, lo que sugiere un agotamiento de la tendencia y un probable cambio de dirección.

El cambio de dirección de la tendencia precedente en el movimiento del tipo de cambio puede deberse a fluctuaciones aleatorias. Esto hace que la definición y validación de tales puntos de inflexión sea una tarea difícil. Lo que significa que, desde el punto de vista del análisis técnico, los estudios deben orientarse a identificar las señales del mercado en las que el cambio de tendencia se debe a cambios de régimen impulsados por acontecimientos de carácter fundamental ocurridos en la economía.

Anticipar la ocurrencia de puntos de inflexión en el movimiento de los tipos de cambio, si bien es una tarea compleja, desde un punto de vista fundamental existe información valiosa que puede ser utilizada como indicadores adelantados para reconocer su posible ocurrencia. Entre los indicadores adelantados *"Leading indicators"* más utilizados para anticipar el cambio de dirección en los precios de mercado se encuentran los índices basados en el análisis de sentimiento, el índice de confianza del consumidor, las ventas al por menor, la producción industrial y los índices de precios al consumo y al productor. Entre los indicadores rezagados *"Lagging indicators"* derivados de los datos históricos del mercado figuran los basados en la media de los datos, que comúnmente

suelen utilizarse para identificar la tendencia de largo plazo.

La definición de estrategias de negociación basadas en la detección y capitalización de puntos de inflexión puede ser determinante en la consecución de beneficios fiables y constantes para los inversores. En consecuencia, el uso de estos puntos de inflexión como señales de entrada o salida del mercado, en la ejecución de posiciones de compra y venta, permite capitalizar los cambios de dirección en el régimen de mercado precedente.Al validar la eficacia del proceso utilizado en la identificación y capitalización de los puntos de inflexión, el proceso puede utilizarse en la definición de los niveles de soporte y resistencia para ayudar al inversor a decidir cuándo vender y cuándo comprar con la información adecuada.

La identificación de puntos de inflexión y la definición de zonas de soporte y resistencia, basadas en datos históricos de mercado, están estrechamente relacionadas y son útiles en la toma de decisiones de inversión. Sin embargo, teniendo en cuenta que las zonas de soporte y resistencia, al estar por debajo y por encima de los precios, su delimitación presupone teóricamente que el tipo de cambio no siga depreciándose y apreciándose, respectivamente. Esto sugiere que los puntos de inflexión proporcionan una valiosa información que ayuda a establecer estos niveles históricos de soporte y resistencia. Esto hace que esta herramienta sea útil para la toma de decisiones de inversión informadas.

En cuanto a la rentabilidad que pueden generar las estrategias de negociación que basan sus operaciones en la capitalización de los puntos de inflexión, cabe señalar que la identificación de estos puntos de inflexión no es necesariamente exacta. Aunque pueden estar condicionados por diversos factores fundamentales, técnicos y de comportamiento, se requiere el uso complementario de indicadores adicionales que ayuden a confirmar las señales del mercado antes de tomar posiciones de compra y venta.

4.3.4 Índice de fuerza relativa RSI

En la actualidad existen numerosas metodologías y reglas de decisión aplicables a la negociación de divisas y al desarrollo de posiciones de compra y venta. Algunos de los enfoques más comunes se basan en el uso de indicadores técnicos, para ayudar a identificar los niveles de soporte y resistencia, así como para confirmar cuándo el activo está sobrecomprado y sobrevendido, antes de abrir posiciones en el mercado.

El Índice de Fuerza Relativa RSI es un indicador de análisis técnico utilizado tradicionalmente para el seguimiento de precios con el fin de identificar cuándo un

instrumento financiero está sobrecomprado o sobrevendido. *J. Welles Wilder* introdujo el cálculo de este índice basado en 14 periodos históricos, aunque las configuraciones basadas en 9 y 25 periodos también se han hecho muy populares hasta la fecha. Este índice actúa como un oscilador de movimiento con valores que van de 0 a 100, lo que facilita el reconocimiento de las zonas de soporte y resistencia en las que el activo está sobrevendido y sobrecomprado.

La Figura 4.6, muestra el comportamiento histórico de los precios de mercado $OHLC$ del tipo de cambio euro-dólar medido en marcos temporales de 15 minutos. La parte inferior de la Figura 4.6 muestra también el índice de fuerza relativa RSI calculado a partir de 9 periodos históricos. Los datos muestreados van del 10 de marzo al 14 de marzo de 2023. Los niveles de resistencia R_i en rojo fueron superados de abajo a arriba por el movimiento del tipo de cambio, lo que supone un régimen de mercado alcista. Esta condición del mercado también la confirma la pendiente descrita por la media móvil exponencial de 27 periodos (línea verde) situada cerca del tipo de cambio.

Figura 4.6: Índice de Fuerza Relativa.
Precios de mercado $OHLC$ del tipo de cambio eur/usd. Datos medidos en marcos temporales de 15 minutos, del 10 al 14 de marzo de 2023. La ruptura de los niveles de resistencia R_i en un mercado alcista con movimientos en rango produce señales de compra capitalizables. La señal RSI de 9 periodos confirma en los niveles 30 y 70 un tipo de cambio sobrevendido y sobrecomprado. Este agotamiento de la tendencia sugiere un cambio de dirección.

Observe que la señal del RSI alcanza máximos y mínimos por encima y por debajo de los niveles 70 y 30 respectivamente. Esto facilita la distinción en los precios de mercado cuando el tipo de cambio forma estos máximos y mínimos. Además, dentro

del análisis técnico de señales, el indicador RSI ayuda a identificar patrones que no son fáciles de distinguir en los gráficos de precios. Entre los patrones *"Chartistas"* más comunes se encuentran los patrones de reversión de tendencia de *"cabeza y hombros"* y los patrones bilaterales de *"triángulo"*. Para más detalles, véase el apartado 4.3.5.

Entre los métodos de negociación más utilizados se encuentra el basado en la identificación y capitalización de divergencias entre el precio del activo negociado y el valor del RSI. Una divergencia bajista/alcista se produce cuando el precio después de haber registrado un valor máximo/mínimo, el RSI no ha superado su máximo/mínimo anterior. Esta señal de mercado es un indicio probable de que el movimiento de los precios está perdiendo fuerza y, por lo tanto, luego de que el tipo de cambio se encuentra sobrecomprado/sobrevendido, es inminente un cambio de dirección de la tendencia precedente.

El análisis de divergencias, como ya se ha mencionado, permite identificar los momentos previos que conducirán a un cambio de dirección en la tendencia de los precios. La Figura 4.7, muestra el proceso que conduce a la formación de una divergencia bajista D_1. Obsérvese que la posterior caída del precio del activo confirma la señal producida por dicha divergencia en la señal del RSI.

Figura 4.7: Divergencia bajista.
Precios de mercado $OHLC$ del tipo de cambio eur/usd. Datos medidos en sesiones de negociación de 15 minutos, del 10 al 14 de marzo de 2023. La divergencia bajista D_1 formada en el gráfico de precios y el RSI confirma en el nivel 70 un agotamiento de la tendencia alcista. En ese nivel de precios, dado que el tipo de cambio está sobrecomprado, es inminente un cambio de tendencia.

En general, las divergencias bajistas pueden aparecen en los gráficos, una vez que la fuerza alcista del precio se ha agotado. Por ejemplo, estas se forman cuando el tipo de cambio después de alcanzar un máximo **b** que supera el máximo anterior **a**, en el RSI ocurre lo contrario, la señal del RSI en $\mathbf{b}'$ no ha superado su máximo anterior $\mathbf{a}'$.

Así, el agotamiento de la tendencia alcista se confirma con una divergencia bajista, lo que lleva al inicio de un proceso de depreciación. Del mismo modo, una divergencia alcista se forma en el gráfico, después de que la fuerza bajista del precio se ha agotado.

En concreto, cuando el precio del activo después de caer a un mínimo por debajo del mínimo anterior, en el RSI sucede lo contrario, después de haber alcanzado un máximo, este valor supera el máximo anterior.

En consecuencia, el agotamiento de la tendencia bajista se confirma con una divergencia alcista, lo que conduce al inicio de un proceso de apreciación. Asimismo, el RSI tiene otra característica importante, y es que ayuda a reconocer más fácilmente los niveles de soporte y resistencia en comparación con los datos de mercado que proporciona el gráfico de precios.

El cálculo del indicador RSI tiene numerosas variantes en la literatura. Sin embargo, el método tradicional introducido por J. Welles Wilder, se obtiene mediante la ecuación 4.5. Este índice RSI mide la fuerza del movimiento del precio, lo que permite reconocer si el tipo de cambio negociado está sobrecomprado o sobrevendido.

Esta categorización supone un agotamiento de la fuerza con la que se ha movido el precio del activo. Esto da lugar a la toma de beneficios o a la apertura de nuevas posiciones de mercado.

$$RSI_{(9p)} = 100 - \frac{100}{(1 + P/G)} \tag{4.5}$$

Donde, P es la media de las variaciones positivas del precio de cierre cp de los últimos i periodos. G es la media de las variaciones negativas del precio de cierre cp de los últimos i periodos. Ambos criterios calculados para 9 periodos históricos.

Obsérvese que el ratio P/G mide la relación entre la media de las variaciones al alza y la media de las variaciones a la baja. Así, un valor alto de este ratio sugiere que el tipo de cambio tiende más a apreciarse que a depreciarse, mientras que un valor bajo sugiere que el tipo de cambio tiende a depreciarse.

Estos resultados varían en función del número de periodos históricos considerados en el cálculo del indicador, lo que también hace difícil su elección.

La ecuación 4.6, una variante del modelo anterior, modifica los valores de P y G por los nuevos valores de $\bar{U}$ y $\bar{D}$, respectivamente. Estos valores, que calculan la media de las variaciones al alza y a la baja, dan mayor peso a la variación más actual entre cp_i y cp_{i-1}.

$$RSI_{(i=n+j)} = 100 - \frac{100}{\left(1 + \bar{U}_i/\bar{D}_i\right)}, \quad j, n \in \mathbb{Z}^+, \quad j \geq 1, \quad 2 \leq n < m \tag{4.6}$$

Esta expresión mide la fuerza del movimiento del precio en la formación de la tendencia, así como su posible cambio de dirección.

Obsérvese que el valor $RSI_{(i=n+j)}$ representa el valor del RSI suavizado y estimado para el periodo i correspondiente a $(n + j)$ periodos en el futuro. j además de ser un entero positivo es mayor o igual a 1. n y m son también enteros positivos, donde n es el número de periodos históricos utilizados en el cálculo del RSI y m es el número total de datos utilizados en el estudio, tal que $2 \leq n < m$.

Asimismo, $\bar{U}, \bar{D} \in \overline{\Delta cp}$, donde, la media de las variaciones alcistas $\bar{U}$ y la media de las variaciones bajistas $\bar{D}$ forman parte del conjunto de variaciones de precio Δcp. La media de las variaciones entre los precios de cierre $\overline{\Delta cp}$ tiene dos procedimientos de cálculo diferentes. En la ecuación 4.7 se estiman los valores de $\bar{U}$ y $\bar{D}$ para el primer periodo futuro $i = (n + j)$, donde $j = 1$.

$$\overline{\Delta cp}_{(j=1)} = \begin{cases} \bar{U}_{i=n+j} = \frac{1}{n} \sum_{i=1+j}^{n+j} u_i, & \text{si } j = 1 \\ \bar{D}_{i=n+j} = \frac{1}{n} \sum_{i=1+j}^{n+j} d_i, & \text{si } j = 1 \end{cases} \tag{4.7}$$

La fórmula señala para el caso $j = 1$ que $\overline{\Delta cp}$, la variación media sobre el precio de cierre, se obtiene calculando dos expresiones. $\bar{U}_{(i=n+j)}$ cuando se trata de movimientos alcistas u_i (variaciones entre precios de cierre, donde el precio de cierre actual cp_i es mayor que el precio de cierre anterior cp_{i-1}) o bien $\bar{D}_{(i=n+j)}$ si se trata de movimientos bajista d_i (variaciones entre precios de cierre, donde el precio de cierre actual cp_i es

menor que el precio de cierre anterior cp_{i-1}). En ambos casos, se realiza el cálculo de la media de las n variaciones entre precios consecutivos históricos u_i y d_i, desde $i = (1 + j)$ hasta $i = (n + j)$

La ecuación 4.8 resume los procedimientos para calcular los valores de $\bar{U}$ y $\bar{D}$ para el segundo y sucesivos periodos futuros $i = (n + j)$, donde $j \geq 2$.

$$\overline{\Delta cp}_{(j\geq 2)} = \begin{cases} \bar{U}_{i=n+j} = \frac{\bar{U}_{i-1}(n-1)+u_i}{n}, & \text{si } j \geq 2 \\ \bar{D}_{i=n+j} = \frac{\bar{D}_{i-1}(n-1)+d_i}{n}, & \text{si } j \geq 2 \end{cases} \tag{4.8}$$

La expresión $\overline{\Delta cp}_{(j\geq 2)}$ indica que para los casos $j \geq 2$, la media móvil suavizada de las variaciones al alza $\bar{U}$ y a la baja $\bar{D}$, se obtiene calculando dos expresiones. El valor de $\bar{U}_{i=n+j}$ utiliza en su cálculo la media de las variaciones al alza del periodo anterior $\bar{U}_{i-1}$ y la variación al alza del periodo actual u_i.

Del mismo modo, el cálculo de $\bar{D}_{i=n+j}$ se basa en la media de la variación a la baja del periodo anterior $\bar{D}_{i-1}$ y la variación a la baja del periodo actual d_i. En ambos casos, nótese que las medias móviles suavizadas de las variaciones al alza y a la baja asignan mayor peso en el cálculo de $\bar{U}_i$ y $\bar{D}_i$ a las variaciones actuales u_i y d_i que a las variaciones medias históricas $\bar{U}_{i-1}$ y $\bar{D}_{i-1}$.

La variación entre precios de cierre Δ_{cp} para dos periodos consecutivos i y $i - 1$, a partir del periodo $i = 2$ viene determinada por las expresiones referidas en la ecuación 4.9:

$$\Delta cp = \begin{cases} u_i = cp_i - cp_{i-1}, & \text{para } i \geq 2 \quad \text{si } cp_i > cp_{i-1} \\ d_i = cp_{i-1} - cp_i, & \text{para } i \geq 2 \quad \text{si } cp_{i-1} > cp_i \\ 0, & \text{para } i \geq 2 \quad \text{si } cp_{i-1} = cp_i \end{cases} \tag{4.9}$$

Δcp corresponde a la diferencia entre dos precios de cierre consecutivos cp, cuyo cálculo está sujeto al signo de las variaciones. Esto sugiere tres tendencias, al alza, a la baja y lateral.

El primer caso, cuando $cp_i > cp_{i-1}$, corresponde a un movimiento al alza u_i. En el segundo caso, cuando $cp_{i-1} > cp_i$, se refiere a un movimiento a la baja d_i. Por último, cuando $cp_{i-1} = cp_i$ dado que $\Delta cp = 0$, se trata de un movimiento lateral.

Es importante señalar que el cálculo de estas variaciones supone tres posibles movimientos en el comportamiento del precio. Sin embargo, la fórmula del índice RSI sólo considera el cálculo de la fuerza de su movimiento a partir de las medias de las variaciones al alza y a la baja.

4.3.5 Patrones de precios

Los patrones en los precios son formaciones gráficas que se han producido a través de interpretaciones realizadas sobre la acción de los precios. Estas formaciones pueden proporcionar señales tempranas de posibles cambios en la dirección de los precios de los instrumentos financieros.

Aunque no todos los patrones son fiables, deben utilizarse como herramienta de confirmación en conjunción con otras técnicas de análisis técnico y análisis fundamental para tomar decisiones fundamentadas en el movimiento de los mercados.

Tradicionalmente, el análisis de patrones de precios se ha utilizado como herramienta de análisis técnico para identificar y confirmar cambios en la dirección de los precios, especialmente cuando se negocia con tipos de cambio.

Estos análisis se han realizado para confirmar los momentos en los que deben tomarse posiciones de compra y venta. Los tipos más comunes de formaciones incluyen patrones de reversión, de continuación y bilaterales. Estos patrones tienen características especiales que hacen necesario reconocerlos para aprovechar y capitalizar las anomalías del mercado.

La Figura 4.8 resume estos tres grandes grupos de pautas en el comportamiento de los precios. (**a**) Patrones de reversión donde hay un cambio de dirección de la tendencia previa. (**b**) Patrones de continuidad que confirman la continuación de la tendencia precedente. (**c**) Patrones bilaterales donde el precio, dada la confirmación de las señales proporcionadas por el mercado, puede cambiar o continuar con la dirección de la tendencia anterior.

Hay que tener en cuenta que en cada pauta se ha establecido el momento en el que hay que entrar en el mercado **ET**, con una operación de compra o venta, según el caso.

El límite de pérdida máxima admisible fijado como **SL** si el precio se mueve en la dirección equivocada. Por último, el límite en el que debe producirse la recogida de beneficios, donde se establece el precio objetivo **XT**.

Los patrones de reversión son formaciones gráficas que ayudan a identificar y confirmar si la tendencia actual está a punto de cambiar de dirección. Entre este grupo de patrones hay básicamente dos efectos predecibles, los patrones de reversión en los que la dirección de la tendencia cambia de alcista a bajista y de bajista a alcista.

En general, en cualquiera de estos casos, después de que la fuerza en el movimiento del tipo de cambio haya sufrido un proceso de agotamiento, aparecen formaciones gráficas que confirman ese agotamiento y que es inminente un cambio de dirección de la tendencia precedente.

Entre los tipos más comunes de formaciones de patrones de reversión se encuentran el doble techo y el doble fondo; el patrón de cabeza y hombros y el patrón de cabeza y hombros invertidos; y la cuña ascendente y descendente.

Los patrones de continuidad son pautas en los precios con formaciones gráficas que indican que la tendencia del mercado continuará en la misma dirección. Este grupo de patrones precede a las indicaciones de continuación en la tendencia alcista o bajista de los precios. Una vez confirmada la formación de estos patrones, la fuerza del movimiento del precio tras alcanzar estos niveles de soporte o resistencia, según el caso, los rompe para continuar la misma tendencia.

Entre los patrones de continuidad de tendencia más comunes se encuentran la cuña descendente y ascendente; el rectángulo alcista y bajista; y el banderín alcista y bajista.

La Figura 4.8 también muestra los patrones bilaterales, a diferencia de los anteriores, estas pautas se caracterizan porque el precio del activo se mueve en un rango lateral en el que no hay una tendencia clara al alza o a la baja.

En cuanto el precio se aproxime al vértice de estas formaciones triangulares, se prevé una ruptura en cualquiera de las dos direcciones. Así, cuando el volumen aumente y el precio rompa la línea de soporte o resistencia, el precio continuará la trayectoria a la baja o al alza. Los patrones bilaterales más comunes son el triángulo ascendente, el triángulo descendente y el triángulo simétrico.

En general, cabe señalar que estas pautas se forman después de que la fuerza del precio se haya agotado al alcanzar un nivel de soporte o resistencia. Así, si la fuerza del precio es débil, confirma estos niveles generando un retroceso en su movimiento tras alcanzarlos. Por el contrario, si la fuerza del precio rompe estos niveles, la tendencia del precio continuará o cambiará de dirección según sea el caso. En cualquiera de estos comportamientos, la confirmación de estos patrones es valiosa para aumentar la probabilidad de éxito en la ejecución de posiciones de compra y venta en el mercado.

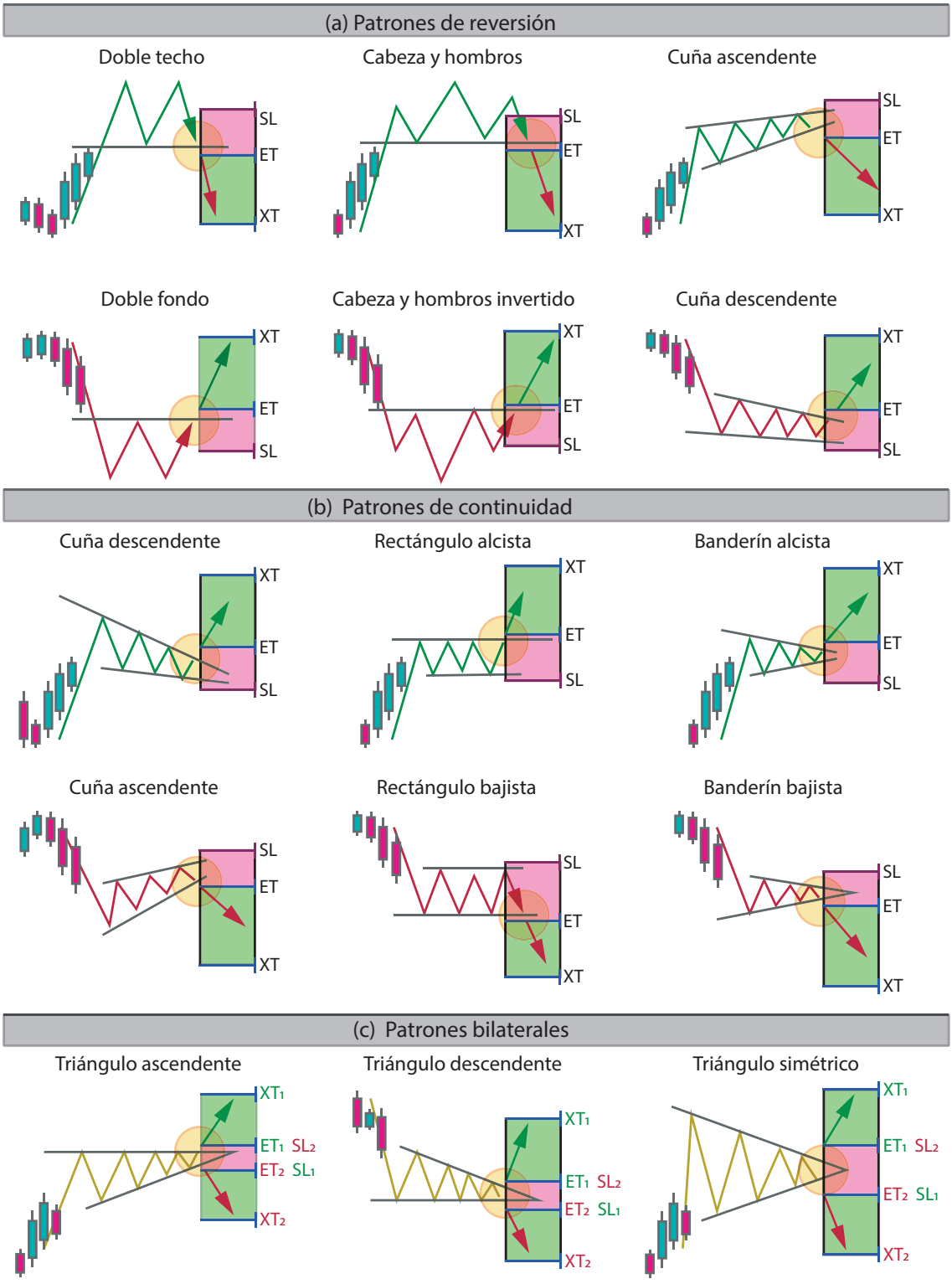


Figura 4.8: Patrones en el movimiento de los precios.
Patrones que señalan una entrada **ET** o una salida anticipada del mercado **XT** debido a una posible reversión o continuación de la tendencia del precio con un nivel de pérdida tolerable **SL**. (**a**) Patrones de cambio de dirección. (**b**) Patrones de continuidad de la tendencia. (**c**) Patrones bilaterales con movimiento del precio en dirección alcista o bajista.

4.4 Enfoque para la solución de problemas

El acto de elegir entre dos o más cursos de acción debe conducir a una toma de decisiones eficaz. Los problemas en los que se consideran múltiples criterios de decisión suelen constar de varias alternativas, en algunos casos contradictorias. Teniendo en cuenta estas observaciones, a continuación se describen brevemente las bases teóricas a considerar para la solución de problemas que pueden tener múltiples criterios y que deben ser considerados en su solución para una efectiva toma de decisiones.

Este apartado describe dos procedimientos útiles para resolver problemas multicriterio: la sección 4.4.1 trata de la programación por objetivos y la sección 4.4.2 se ocupa de la programación no lineal.

4.4.1 Programación por metas

La programación por metas (*Goal programming*) es una técnica de resolución de problemas multi-criterio, una variante de los modelos de programación lineal, que considera varias metas a resolver con la definición de la función objetivo. La estructura general de estos modelos sigue el mismo formato que los modelos de programación lineal. Estos métodos se caracterizan porque parten de la definición de una función objetivo con restricciones, en la que su solución es similar a la obtenida por métodos de programación lineal. En la ecuación 4.10 se formula un problema convencional de programación lineal, la función objetivo **Z** especifica lo que se quiere conseguir, en este caso maximizar la función **Z**. c_1 y c_2 son los coeficientes de las variables de decision x_1 y x_2. Las variables de decisión son, en este caso, los valores a obtener.

$$\begin{aligned} \text{Maximizar:} \quad & \mathbf{Z} = c_1 x_1 + c_2 x_2 \\ \text{sujeto a:} \quad & a_{11} x_1 + a_{12} x_2 \leq b_1 \\ & a_{21} x_1 + a_{22} x_2 \leq b_2 \\ & x_1, x_2 \geq 0 \\ \text{donde:} \quad & \\ & x_1, x_2 = \text{variables de decisión} \end{aligned} \tag{4.10}$$

Nótese que para alcanzar el objetivo de la Función **Z** se han especificado tres limitaciones o condiciones que deben cumplirse en la solución del problema. Estas restriccio-

nes pueden expresarse como una ecuación o bien como una desigualdad que incluya las variables de decisión. Las dos primeras condiciones limitan el uso de las restricciones b_1 y b_2, esto significa que las combinaciones lineales definidas por las variables de decisión x_1 y x_2 deben ser menores o iguales que las restricciones b_1 y b_2.

El modelo anterior forma parte de los modelos de programación lineal clásicos, caracterizados por una única función objetivo. Este mismo modelo, sin embargo, puede transformarse en uno basado en múltiples objetivos que deben alcanzarse según su nivel de importancia. Asumiendo los siguientes objetivos mas adelante se describe el proceso que llevaría a su consecución:

P_1: No utilizar menos de la restricción b_1.
P_2: Alcanzar un nivel de la función objetivo $\mathbf{Z}$ en el nivel k.
P_3: Evitar mantener más de la restricción b_2.
P_4: Reducir la sobre-utilización de la restricción b_1.

Estos objetivos deben formularse en términos de funciones, lo que implica que el analista debe transformar las restricciones del modelo de programación lineal de la ecuación 4.10 en objetivos. Esto implica transformar el modelo original en uno que incluya múltiples objetivos. Cuando esto ha sido posible, el analista debe conseguir un nivel de solución que se acerque lo mejor posible a la consecución de cada uno de estos objetivos. A continuación se describe el procedimiento utilizado en la transformación de las restricciones en restricciones objetivo.

- Convirtiendo la restricción b_1 en los objetivos P_1 y P_4.

El primer objetivo consiste en evitar la sub-utilización de la restricción b_1, lo que supone emplear menos de lo que permite la restricción b_1. En consecuencia, la desigualdad $a_{11}x_1 + a_{12}x_2 \leq b_1$ necesita ser transformada en una *restricción objetivo* como se observa en la ecuación 4.11:

$$a_{11}x_1 + a_{12}x_2 + s_1^- - s_1^+ = b_1 \tag{4.11}$$

En la ecuación anterior nótese la incorporación de dos nuevas variables s_1^- y s_1^+, denominadas variables de desviación. Donde s_1^- es la variable de desviación que representa un valor de utilización inferior a b_1, es decir, una sub-utilización de la restricción b_1, mientras que s_1^+ representa una sobre-utilización de b_1. Así, en la búsqueda de la solución objetivo, al menos una de las variables de decisión será igual a cero. En

caso contrario, cuando $s_1^+ = 0$ y $s_1^- = 0$ significa que el valor de la restricción b_1 se utiliza exactamente. Hay que tener en cuenta que para conseguir este objetivo P_1, de no utilizar menos de la restricción b_1, es necesario crear una nueva función objetivo definida en la ecuación 4.12:

$$\text{Minimizar:} \quad P_1 s_1^- \tag{4.12}$$

Obsérvese que el propósito de la función objetivo es minimizar la variable de desviación s_1^-, la sub-utilización del valor de la restricción b_1. Además, es preciso destacar que el cuarto objetivo P_4, que busca minimizar la sobre-utilización, también está asociado a la utilización de la restricción b_1. Por tanto, al incorporar este objetivo, la nueva función objetivo que toma la ecuación 4.12 queda expresada de la siguiente forma:

$$\text{Minimizar:} \quad P_1 s_1^-, \quad P_4 s_1^+ \tag{4.13}$$

Hay que tener en cuenta que la variable exceso de desviación s_1^+ se minimiza. Además, se debe considerar el orden de prioridad en el cumplimiento de los objetivos en la solución del problema. Esto significa que antes de buscar el cumplimiento del objetivo P_4 se deben perseguir los tres primeros objetivos.

- Ajustando la función objetivo $\mathbf{Z}$ a un nivel deseado k.

En la formulación del modelo de programación por objetivos, el segundo objetivo a conseguir P_2 busca alcanzar un nivel k de solución. Por tanto, reformulando la función objetivo $\mathbf{Z} = c_1x_1 + c_2x_2$ transformándola en una nueva restricción objetivo que incluye un nivel de beneficio deseado k, la expresión 4.10 se transforma en la nueva ecuación 4.14:

$$c_1x_1 + c_2x_2 + s_2^- - s_2^+ = k \tag{4.14}$$

Según la expresión anterior, las variables de desviación s_2^- y s_2^+ representan el nivel de la función objetivo a obtener, por debajo y por encima del nivel k respectivamente. Así, el objetivo P_2, que especifica alcanzar un nivel de solución k, una vez incorporado a la función objetivo se expresa de la siguiente forma:

$$\text{Minimizar:} \quad P_1 s_1^-, P_2 s_2^-, P_4 s_1^+ \tag{4.15}$$

Al observar la nueva función objetivo, cabe destacar que como segunda prioridad, el interés del segundo objetivo (P_2) consiste en minimizar la variable de desviación s_2^-, es decir, que la búsqueda de la mejor solución supere el nivel k a través de la variable de desviación s_2^+.

- Transformado la restricción b_2 en la restricción objetivo P_3.

Considerando la restricción $a_{21}x_1 + a_{22}x_2 \leq b_2$ de la ecuación 4.10, ésta pasa a ser la nueva restricción objetivo:

$$a_{21}x_1 + a_{22}x_2 + s_3^- - s_3^+ = b_2 \tag{4.16}$$

La variable de desviación s_3^- representa el valor de la restricción por debajo de b_2, mientras que la variable de desviación s_3^+ representa el valor de la restricción por encima de b_2. Al añadir este objetivo a la función objetivo, la expresión 4.15 se convierte en la nueva función objetivo:

$$\text{Minimizar:} \quad P_1 s_1^-, \quad P_2 s_2^-, \quad P_3 s_3^+, \quad P_4 s_1^+ \tag{4.17}$$

Obsérvese que la variable de desviación s_3^+ vinculada al objetivo P_3 pretende minimizar el exceso de uso de la restricción b_2. Así, el término $P_3 s_3^+$, según el orden de prioridad, constituye el tercer objetivo que se pretende alcanzar.

- Definición del modelo de programación por metas.

Teniendo en cuenta los resultados obtenidos tras transformar las restricciones en restricciones objetivo, a partir de los objetivos P_i previamente establecidos, se presenta a continuación el modelo de programación por objetivos final:

$$\begin{aligned}
\text{Minimizar:} \quad & P_1 s_1^-, \quad P_2 s_2^-, \quad P_3 s_3^+, \quad P_4 s_1^+ \\
\text{sujeto a:} \quad & a_{11}x_1 + a_{12}x_2 + s_1^- - s_1^+ = b_1 \\
& c_1 x_1 + c_2 x_2 + s_2^- - s_2^+ = k \\
& a_{21}x_1 + a_{22}x_2 + s_3^- - s_3^+ = b_2 \\
& x_1, x_2, s_1^-, s_1^+, s_2^-, s_2^+, s_3^-, s_3^+ \geq 0 \\
\text{donde:} \quad & \\
& x_1, x_2 = \text{variables de decisión} \\
& s_1^-, s_1^+, s_2^-, s_2^+, s_3^-, s_3^+ = \text{variables de desviación}
\end{aligned} \tag{4.18}$$

Al comparar el nuevo modelo de programación por metas con el modelo de programación lineal, mencionado en 4.10, se observa que los términos de la nueva función objetivo no suman en función de $\mathbf{Z}$. Esto se debe a que las variables de desviación s definidas en la nueva función objetivo poseen unidades de medida diferentes. En consecuencia, las funciones objetivo del modelo de programación por metas señalan que las variables de desviación s se minimizarán de forma individual considerando el orden de prioridad de los objetivos P_i definidos en dicha función.

- Ponderación de restricciones objetivo.

Suponiendo que las variables de decisión x_1 y x_2 sólo pueden alcanzar el nivel de restricción b_3 y b_4 respectivamente. Al considerar la inclusión de este objetivo P_5 en el modelo 4.18, además de que es más importante para el analista que la variable de decisión x_2 alcance el nivel de restricción b_4 que la variable x_1. Este hecho implica la adopción de dos nuevas restricciones objetivo, que se describen a continuación:

$$\begin{aligned} a_{31}x_1 + s_5^- &= b_3 \\ a_{41}x_2 + s_6^- &= b_4 \end{aligned} \tag{4.19}$$

En las expresiones anteriores se puede observar que las variables de desviación positivas s_5^+ y s_6^+ se descartaron de la definición de las restricciones objetivo. Esto se debe al propósito del objetivo P_5, que las variables de decisión x_1 y x_2 sólo pueden alcanzar el nivel de restricción b_3 y b_4 respectivamente.

Ahora bien, dado que el propósito adicional del analista, en la búsqueda de la solución del objetivo P_5, es conferir mayor importancia a la variable de decisión x_2 en relación con su nivel de restricción b_4, este nivel de interés puede reflejarse en la definición de esta restricción objetivo asignando como coeficientes los pesos w_1 y w_2 tal y como se muestra a continuación:

$$\text{Minimizar:} \quad P_1 s_1^-, \quad P_2 s_2^-, \quad P_3 s_3^+, \quad P_4 s_1^+, \quad w_1 P_5 s_5^- + w_2 P_5 s_6^- \tag{4.20}$$

De acuerdo con la expresión anterior, la restricción objetivo P_5 está formada por los coeficientes w_1 y w_2, donde $w_1 < w_2$, nótese que se han asignado estos *pesos*, de forma que al minimizar las variables de desviación s_5^- y s_6^- se observa un mayor interés en conseguir una mayor utilización de la restricción b_4 para la variable de decisión x_2 que para la variable x_1. Además, cabe destacar que estos objetivos ponderados, además de sumar, se encuentran en el mismo nivel de prioridad P_5. Así, a continuación se

presenta el nuevo modelo de programación por objetivos en el que se incluye un quinto objetivo restrictivo:

$$
\begin{aligned}
\text{Minimizar:} \quad & P_1 s_1^-, \quad P_2 s_2^-, \quad P_3 s_3^+, \quad P_4 s_1^+, \quad w_1 P_5 s_5^- + w_2 P_5 s_6^- \\
\text{sujeto a:} \quad & a_{11}x_1 + a_{12}x_2 + s_1^- - s_1^+ = b_1 \\
& c_1 x_1 + c_2 x_2 + s_2^- - s_2^+ = k \\
& a_{21}x_1 + a_{22}x_2 + s_3^- - s_3^+ = b_2 \\
& a_{31}x_1 + s_5^- = b_3 \\
& a_{41}x_2 + s_6^- = b_4 \\
& x_1, x_2, s_1^-, s_1^+, s_2^-, s_2^+, s_3^-, s_3^+, s_5^-, s_6^- \geq 0 \\
\text{donde:} \quad & \\
& x_1, x_2 = \text{variables de decisión} \\
& s_1^-, s_1^+, s_2^-, s_2^+, s_3^-, s_3^+, s_5^-, s_6^- = \text{variables de desviación}
\end{aligned}
\tag{4.21}
$$

4.4.2 Programación no lineal

La programación no lineal es una técnica versátil que se aplica ampliamente a una multiplicidad de problemas. A diferencia de los modelos de programación lineal, este enfoque se caracteriza por el hecho de que las funciones objetivo y las restricciones son funciones no lineales. De hecho, muchos problemas del mundo real tienen relaciones no lineales, por lo que es necesaria su modelización no lineal. Esto significa que, aunque dichos problemas se ajustan a la estructura general de la programación lineal, no están formados necesariamente por funciones lineales y, por lo tanto, su solución es extremadamente compleja y difícil, especialmente cuando se trata de encontrar una solución óptima.

En este tipo de enfoques, el espacio de soluciones puede considerar superficies no lineales en las que existe prácticamente un número infinito de puntos o soluciones. Así, una solución óptima (máximo o mínimo) podría coincidir con un pico o un valle lo que dificulta establecer si esta solución corresponde a un óptimo local o si realmente se trata de una solución óptima global.

En este sentido, las técnicas de resolución de problemas de programación no lineal se centran simplemente en encontrar picos o valles en la superficie de la solución. Por tanto, la principal limitación de estos métodos de búsqueda es determinar si la solución

encontrada corresponde a un óptimo local o global. Esto puede llevar a procedimientos de calculo demasiado complejos, lo que hace que este tipo de procedimientos queden por fuera del ámbito de este estudio.

La selección de carteras de inversión es uno de los modelos de programación no lineal más conocidos en el ámbito del mercado de capitales. En los años 50, Harry Markowitz demostró que los inversores basan sus decisiones de inversión en dos criterios particulares: el riesgo s_i y el rendimiento r_i que se obtendrá sobre el valor de la inversión i. El objetivo de este modelo es minimizar el riesgo de la cartera de inversión obteniendo al mismo tiempo una rentabilidad. De hecho, el riesgo está asociado a la variabilidad de los rendimientos obtenidos con la cartera de inversión, lo que le llevó a proponer la diversificación de la cartera de activos como una alternativa para reducir el riesgo.

$$
\begin{aligned}
\text{Minimizar:} \quad & \mathbf{Z} = \sum_{i=1}^{n} x_i^2 s_i^2 + \sum_{i \neq j} x_i x_j r_{ij} s_i s_j \\
\text{sujeto a:} \quad & r_1 x_1 + r_2 x_2 + r_3 x_3 + r_4 x_4 \geq \frac{1}{n} \sum_{i=1}^{n} r_i \\
& x_1 + x_2 + x_3 + x_4 = 1.0 \\
& x_i \geq 0 \\
\text{donde:} \quad & \\
& Z = \text{S} = \text{varianza del retorno anual del portafolio} \\
& x_i, x_j = \text{proporción de dinero invertido en inversiones i y j} \\
& s_i^2 = \text{varianza de la inversión i} \\
& r_{ij} = \text{correlación entre los retornos de las inversiones i y j} \\
& s_i, s_j = \text{desviación típica de los retornos de las inversiones i y j} \\
& r_i = \text{retorno anual esperado de la inversión i}
\end{aligned}
\tag{4.22}
$$

En la actualidad, algunos modelos de aprendizaje automático se basan en el uso de métodos de programación no lineal. Estos enfoques aprovechan las ventajas de estas técnicas para encontrar la solución óptima a cualquier tipo de problema. Entre las aplicaciones más destacadas se encuentran la definición de hiper-parámetros o la selección del mejor subconjunto de características para maximizar la precisión en tareas de clasificación.

4.5 Negociación algorítmica

El desarrollo de los mercados financieros ha propiciado un gran cambio en la forma en que los inversores ejecutan sus órdenes en la negociación de activos. La negociación algorítmica se trata de un método de negociación automatizado que sigue un conjunto de reglas de decisión basadas en una *"Estrategia de Negociación"* que ha demostrado ser viable y coherente en el tiempo. Estos modelos de negociación abren y cierran posiciones, teniendo en cuenta datos de mercado, fundamentales, información heterogénea o una combinación de todos ellos. Así, la emisión de órdenes de mercado para comprar o vender un activo a un precio determinado se basa en la definición de órdenes pendientes de compra o venta. Estas órdenes se definen por debajo/por encima del precio de mercado actual y se activan cuando el precio de mercado alcanza el nivel de precio de la orden. El cierre de posiciones, para obtener beneficios o limitar el nivel de pérdidas, se basa en la emisión de órdenes limitadas. Una vez fijadas, se ejecutan cuando los precios actuales alcanzan el nivel definido por estas órdenes. Todos estos aspectos técnicos, incluida la liquidación automática de una posición, se definen en función de la parametrización del modelo, que entre otros aspectos incluye la gestión del riesgo y del capital.

4.5.1 Estrategia de negociación

La negociación algorítmica de un determinado activo financiero, además de eliminar el error humano en la ejecución, mejora significativamente la eficacia de la negociación. Esto sólo ocurre cuando la estrategia de negociación, además de ser verificable en el tiempo, es cuantificable, consistente, objetiva y extensible. Aunque el proceso que conduce a la construcción de una estrategia de negociación es muy complejo, según la Figura 4.9, deben tenerse en cuenta al menos los siguientes pasos interrelacionados e interdependientes.

Figura 4.9: Desarrollo de estrategias de negociación.
Etapas interrelacionadas e interdependientes utilizadas en la construcción de estrategias de negociación. La primera parte, en términos generales, comprende la formulación, evaluación y mejora de la estrategia. La fase final, de implementación, incluye la negociación, el seguimiento en tiempo real y el refinamiento de la estrategia de negociación.

La Figura 4.9 resumen en ocho pasos el proceso general utilizado en la creación de estrategias de negociación. Sin embargo, a fin de diseñar estrategias eficaces basadas en una metodología sistémica, es preciso utilizar el método científico descrito en la Figura 4.10. Esta metodología contempla tres grandes etapas, la definición de reglas de decisión, la construcción de la estrategia de negociación y finalmente la puesta en marcha y mejora. En la primera fase, hay que tener en cuenta que la definición de reglas de decisión, instrucciones de negociación que se activan en función de las condiciones y señales del mercado, además de traducirse a un lenguaje de programación, debe verificarse y evaluar su eficacia en el tiempo. Este proceso conduce a la formulación y comprobación de pruebas de hipótesis. En la segunda fase, tras validar las reglas de decisión, la construcción de la estrategia de negociación implica no sólo su evaluación preliminar, sino también su optimización y validación en el tiempo. Este proceso incluye la identificación de un instrumento financiero, el análisis retrospectivo de la estrategia a partir de precios históricos de mercado, la formulación y comprobación de hipótesis y la gestión de riesgo y capital. Finalmente, en la última fase, una vez confirmada la viabilidad técnica y financiera de la estrategia de negociación, se utiliza en tiempo real en la negociación del activo en cuestión. Además de su seguimiento y mejora, se evalúa su uso extensivo en diferentes ventanas temporales y con otros tipos de instrumentos financieros.

Es preciso señalar que la formulación de modelos de negociación bajo el enfoque científico ha demostrado ser una metodología valiosa en la construcción de estrategias de negociación eficaces. Los modelos de negociación que surgen bajo este enfoque suplen las siguientes características esenciales:

- Capacidad para prever resultados de negociación coherentes basados en un funcionamiento correcto de las reglas de decisión y la estrategia de negociación.
- Conocimiento exhaustivo y pormenorizado de los principios y fundamentos sobre los que se sustenta la estrategia.
- Formulación basada en un sólido fundamento teórico y en pruebas empíricas que avalan su eficacia.
- Optimización integral de la estrategia diseñada para superar retos específicos y aumentar sus posibilidades de éxito, mitigando el sobreajuste.
- Capacidad de aplicación y uso extensivo hacia otros instrumentos, en contextos diferentes, especialmente bajo condiciones adversas.

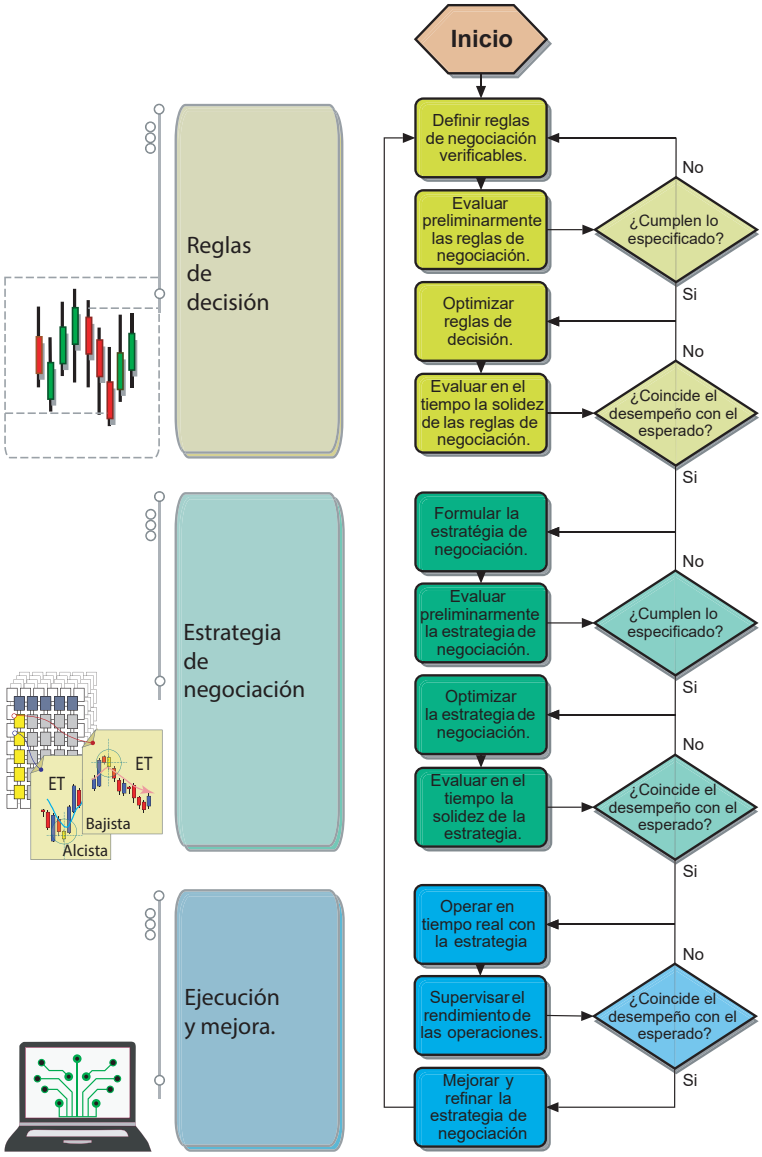


Figura 4.10: Enfoque científico para el desarrollo de estrategias de negociación. Enfoque sistemático utilizado en la construcción de estrategias de negociación. Este proceso de tres fases incluye la definición y mejora de las reglas de decisión utilizadas en la negociación de activos. La formulación, evaluación y mejora de la estrategia de negociación. Por último, la fase de ejecución se ocupa de la negociación, la supervisión en tiempo real y la mejora del modelo de negociación.

4.5.2 Perfil de la estrategia

El perfil de rendimiento es la síntesis estadística del desempeño obtenido con la estrategia de negociación y proporciona una visión detallada de su salud financiera. El análisis exhaustivo de estos resultados estadísticos ayuda a comprender mejor cómo

funcionan las reglas de decisión, la estrategia y el modelo de negociación.

El perfil estadístico de rendimiento es una medida del valor de la estrategia, inherente a las características medidas. En el mejor de los casos, conocer el perfil de la estrategia contribuye a generar confianza en el uso de la misma y a transmitir la sensación de seguridad del analista a la hora de operar con el activo en los mercados. En el caso contrario, proporciona información valiosa que ayuda a diagnosticar y mejorar el rendimiento de la estrategia. Además, conocer el perfil de la estrategia ayuda a identificar las condiciones del mercado que afectan al rendimiento normal de la estrategia y, por tanto, ayuda a comprender mejor las causas que contribuyen a esas desviaciones. Esta información no sólo proporciona una referencia para evaluar y mejorar el rendimiento de la estrategia, sino que también permite comparar el rendimiento obtenido durante las fases de diseño e implementación con la negociación en tiempo real. Una medida precisa, estadísticamente fiable y coherente de los resultados obtenidos revela una gestión adecuada del riesgo y del capital. Estas medidas ayudan a detectar si existe unicidad y coherencia entre los resultados obtenidos en la fase de implementación con respecto a los conseguidos durante la fase de concepción.

Nótese que este proceso de análisis y mejora es mucho más factible en los procesos de negociación algorítmica frente a la negociación racional, donde se hace aún más difícil, dada la dificultad de obtener un mayor número de posiciones que permitan evaluar la viabilidad de la estrategia. Ahora bien, sin importar el tipo de enfoque empleado en la negociación la gestión del riesgo es el principal factor que determina la salud financiera de la estrategia y, por tanto, su viabilidad depende de la gestión del capital.

Por otra parte, el perfil de rendimiento ofrece una imagen clara del desempeño de la estrategia en diferentes condiciones de mercado, volatilidad e incertidumbre. Los estudios empíricos han demostrado un mejor rendimiento de las estrategias de seguimiento de tendencias en mercados con fuertes tendencias alcistas o bajistas, mientras que dichos perfiles pueden verse afectados en mercados laterales o durante cambios de régimen.

Las Tablas 4.1 y 4.2 muestran el perfil de rendimiento de una estrategia de negociación de seguimiento de tendencia que se basa en el tipo de cambio eur/usd. La operativa se realiza en un marco temporal de 15 minutos y el periodo muestreado es de 18 días, del 13 de febrero de 2023 al 3 de marzo del mismo año. La negociación se realiza en una cuenta *"mini"* con un depósito inicial de 1000 dólares americanos, un nivel de apalancamiento de 1:100, un tamaño de contrato de 100000 unidades de la divisa base, el tamaño del lote negociado es 0.070 veces el tamaño de contrato estándar

y un nivel de riesgo de pérdida de dinero por posición del 0.84%.

Tras la realización de cada operación, delimitada por una orden de entrada y otra de salida, se cobra una comisión de 45/1000000 USD por orden. En caso de mantener una posición abierta durante la noche, después de las 17:00 (hora de Nueva York), se cobra una comisión que se calcula en función del tipo de posición, swap largo $-4.96 y swap corto $-0.96.

Tabla 4.1: Perfil de rendimiento de la estrategia

[a] Especificaciones			
Símbolo:	EUR/USD	Moneda base:	USD
Barras:	1471	Tamaño de contrato:	100000
Deposito inicial ($):	1000	Apalancamiento:	1:100
Valor del Pip ($):	0.70	Swaps ($):	-9.40
Saldo final ($):	1151.98	Volumen negociado en lotes:	0.070
Posición cerrada G/P ($):	228.47	Comisiones ($):	-67.09
Nivel de margen mínimo (%):	7939	Tasa libre de riesgo (%):	0.00
Riesgo de mercado (%):	-		
Beneficio Neto Total ($):	151.98	Beneficio bruto ($):	788.52
Pérdida bruta ($):	-636.54	Riesgo por operación (%):	0.84
Factor de beneficio:	1.24	Beneficio esperado:	1.55
Factor de recuperación	0.92	Ratio de Sharpe:	0.08
AHPR:	1.0016	GHPR:	1.0014
Probabilidad de éxito:	0.57	Probabilidad de fracaso:	0.43
Operaciones totales:	98	Operaciones cortas ganadoras:	46
Operaciones largas ganadoras:	52	Operaciones perdedoras:	42
Operaciones nulas:	0	Operaciones ganadoras:	56
ROI:	0.152	Operación mayor beneficio ($):	69.32
Operación mayor pérdida ($):	-61.64	Pérdida media / operación ($):	-14.47
Periodo de análisis (Años):	0.06	Beneficio medio / operación ($):	14.60
Operaciones rentables (%):	55.10	Máx de victorias consecutivas:	6
Máx pérdidas consecutivas:	7	Máx pérdida consecutiva($):	-165.85
Operaciones con pérdidas (%):	44.90	Máx beneficio consecutivo ($):	200.21
Ratio (Win / Loss):	1.23:1	Media victorias consecutivas:	3
Media pérdidas consecutivas:	3		

[a]Barras = Número de velas analizadas en un marco de tiempo de 15 minutos. Posición cerrada G/P = Ganancias/Pérdidas totales de todas las operaciones cerradas (Comisiones y swaps no descontados). Comisiones ($) = Comisión cobrada tras la apertura y cierre de una operación ($ 45/1000000 USD por operación). Swaps ($) = Comisión cobrada cuando la posición se mantiene abierta durante la noche (Swap largo = -4,96, Swap corto = -0,96). Riesgo de mercado (%) = Porcentaje de dinero en riesgo de pérdida sobre el saldo disponible. Beneficio neto total ($) = Resultados financieros de operaciones cerradas (Beneficios + comisiones + swaps). Beneficio bruto ($) = Suma de beneficios de operaciones rentables (Comisiones y swaps descontados). Pérdida bruta ($) = Suma de perdidas de operaciones perdedoras (Comisiones y swaps descontados). Factor de beneficio = Relación entre el beneficio bruto y la pérdida bruta. Beneficio esperado = Retorno medio de una operación. Factor de recuperación = Relación entre el beneficio neto total y la reducción máxima. Ratio de Sharpe = (Rentabilidad - Tasa libre de riesgo) / Desviación estándar de los retornos. AHPR = Media aritmética de las variaciones de capital por operación cerrada. GHPR = Media geométrica de las variaciones de capital por operación cerrada.

Tabla 4.2: Perfil de rendimiento de la estrategia

[a] Especificaciones			
Esperanza matemática:	$ 2.33	Ratio (Beneficio / Pérdida):	1.01 : 1
Ventaja operativa:	0.10	Prueba de corridas Z-Score:	-3.2817
Desviación Estándar Beneficios:	$ 22	Límite de confianza Z-Score:	0.9973
Coeficiente de variación:	9.44	LR Correlación:	0.847
LR Error Estándar:	38.95		
Reducción Absoluta de Saldo:	$ 55.79	Reducción Máxima de Saldo:	$ 165.85
Reducción Relativa de Saldo:	12.95 %		

[a]Riesgo por operación (%) = Cantidad de capital en riesgo por operación. Operaciones ganadoras = Número de operaciones ganadoras (Comisiones y swaps sin descontar). Operaciones perdedoras = Número de operaciones perdedoras (Comisiones y swaps sin descontar). ROI = Beneficio neto total dividido por el depósito inicial. Operación mayor beneficio ($) = Beneficio máximo de todas las operaciones rentables (Comisiones y swaps descontados). Operación mayor pérdida ($) = Pérdida máxima de todas las operaciones perdedoras (Comisiones y swaps descontados). Beneficio medio por operación ($) = Beneficio bruto dividido por el número de operaciones rentables. Pérdida media por operación($) = Pérdida bruta dividida por el número de operaciones no rentables. Operaciones rentables (%) = Número de operaciones rentables (Comisiones y swaps descontados) dividido por el número de operaciones totales. Máx de victorias consecutivas (Recuento) = Serie más larga de operaciones ganadoras. Máx pérdidas consecutivas (Recuento) = Serie más larga de operaciones perdedoras. Operaciones con pérdidas (%) = Número de operaciones con pérdidas (comisiones y swaps descontados) dividido por el número de operaciones totales. Máx Beneficio consecutivo ($) = Beneficio máximo acumulado de una serie de operaciones rentables. Máx Pérdida consecutiva ($) = Máxima pérdida acumulada de una serie de operaciones perdedoras. Ratio (Win / Loss) = Operaciones con beneficios (%) dividido por Operaciones con pérdidas (%) Media victorias consecutivas = Promedio de operaciones ganadoras en series rentables. Media pérdidas consecutivas = Promedio de operaciones perdedoras en series perdedoras. Ratio (Beneficio / Pérdida) = Beneficio medio por operación ($) dividido por la pérdida media por operación ($). Ventaja comercial = Operaciones con beneficios (%) - Operaciones con pérdidas (%). Correlación LR = Correlación entre la línea de balance de la cuenta y la línea de regresión. Z-Score = Detección de periodos consecutivos de beneficios/pérdidas. LR Error Estándar = Error estándar de la desviación entre la línea de saldo de la cuenta y la línea de regresión. Reducción (Drawdown) Absoluta de Saldo ($) = La mayor caída de saldo por debajo del depósito inicial. Reducción máxima del saldo ($) = La mayor caída del saldo en términos monetarios. Reducción relativa del saldo (%) = La mayor caída del saldo en términos porcentuales.

En general, las posiciones de mercado generadas con el modelo de negociación tienen una probabilidad de producir beneficios del 57%, lo que se traduce en un resultado neto de $151.98 en el balance de la cuenta. Las posiciones ganadoras representan el 57.14% de las 98 posiciones efectuadas. El ratio Beneficio/Pérdida, con un valor de 1.23, revela un modesto rendimiento de la estrategia, dando lugar a una posición ganadora por cada posición perdedora. Esto se confirma al analizar el factor de beneficio por posición, la estrategia consigue una ganancia de $1.24 por cada dólar perdido. Hay que señalar que todos los criterios anteriores determinan el perfil de rendimiento de una estrategia de negociación. Sin embargo, hay algunos criterios que describen mejor dicho perfil de rendimiento. Por ello, la Tabla 4.3 resume los criterios más relevantes.

Tabla 4.3: Determinantes del perfil de rendimiento.

[a]**Criterios relevantes**	**Medida**
Beneficio Neto Total ($):	151.98
Retorno sobre la inversión ROI (%):	15.20
Beneficio esperado:	1.55
Probabilidad de éxito (%):	57.14
Ratio de Sharpe:	0.08
Riesgo de perdida de dinero por operación (%):	0.84
Prueba de corridas Z-Score:	-3.2817
Límite de confianza Z-Score (%):	99.73
Factor de beneficio:	1.24

[a]Beneficio neto total ($) = Resultados financieros de operaciones cerradas (Beneficios + comisiones + swaps). ROI = Beneficio neto total dividido por el depósito inicial. Beneficio esperado = Retorno medio de una operación. Probabilidad de éxito = operaciones rentables respecto al total de operaciones. Ratio de Sharpe = (Rentabilidad - Tasa libre de riesgo) / Desviación estándar de los retornos. Riesgo por operación (%) = Cantidad de capital en riesgo por operación. Corridas Z-Score = Detección de periodos consecutivos de beneficios o de pérdidas. Factor de beneficio = Relación entre el beneficio bruto y la pérdida bruta.

Los resultados presentados en la Tabla 4.3, describen un perfil moderado de la estrategia de seguimiento de tendencia. Teniendo en cuenta el número de días de negociación analizados, el beneficio neto total y el rendimiento obtenido por la estrategia sobre el valor de la inversión inicial (ROI = 15.20%) es relativamente modesto. Sin embargo, dada esta condición, la estrategia ha generado beneficios moderados de $ 151.98. En cuanto al beneficio esperado, el valor de 1.55 indica que la estrategia de negociación tiende a generar un mayor número de posiciones ganadoras que perdedoras. La probabilidad de éxito del sistema de negociación, con un índice del 57%, también confirma que el número de posiciones ganadoras es superior al de las posiciones perdedoras.

Al evaluar el Ratio de Sharpe (0.08) en comparación con el riesgo de pérdida de dinero por operación (0.84%), estos valores sugieren que la rentabilidad por operación, derivada de la estrategia de negociación, es relativamente baja en comparación con el riesgo asumido. La prueba de corridas Z-Score (-3.2817) indica que los resultados obtenidos con la estrategia, dada la estructura de sus reglas de decisión de negociación, no son fruto del azar y, por tanto, dichos resultados están determinados por la eficacia de la estrategia. La puntuación Z-Score, por su valor negativo, indica que la estrategia tiene una dependencia positiva, es decir, que el modelo de negociación es propenso a

experimentar largas rachas de victorias y de derrotas. Esto significa que las operaciones ganadoras van seguidas de más posiciones ganadoras, lo que significa que el tamaño del lote puede aumentarse/disminuirse cuando comienza la racha ganadora/perdedora. El límite de confianza Z-Score del 99.3% sugiere que los resultados obtenidos con la estrategia son estadísticamente significativos. No obstante, dado el modesto perfil de la estrategia de negociación, los resultados sugieren mejorar la gestión de exposición al riesgo por posición y mejorar el factor de beneficio antes del cierre de cada posición.

En consecuencia, las operaciones tienen una probabilidad de ganar del 72% y generan unos beneficios netos de 13162.85 USD, durante el periodo de la muestra, de enero de 2006 a diciembre de 2020. Así pues, las operaciones ganadoras representan el 75.34% de las 20894 operaciones realizadas y tardan una media de 7 horas. El 75% de estas operaciones obtuvieron rendimientos brutos en los dos primeros tercios del periodo de tenencia. El Ratio Win/Loss, con un valor de 2.37, muestra que el número de posiciones ganadoras es 2 veces mayor que el número de posiciones perdedoras. El Factor de Beneficio de 1.68 muestra que se gana 1.68 dólares por cada dólar perdido.

Estadísticamente, la prueba de las rachas y el Z-score del sistema de negociación determina la existencia de dependencia entre los resultados de las operaciones. Así, una puntuación Z de -47.68 confirma una dependencia positiva entre las operaciones dentro de un límite de confianza del 99.7%. Esto significa que el número de rachas es menor de lo que implicaría la función de probabilidad normal, de modo que las operaciones ganadoras generan más ganadores y las perdedoras más perdedores. Sin duda, es posible explotar las relaciones de dependencia que existen entre las posiciones y mejorar el rendimiento del sistema, pero el uso de medidas de gestión del capital y de señales de confirmación queda fuera del ámbito de este trabajo.

4.5.3 Estrategia de negociación y modelo de predicción

La Figura 4.11 resume el procedimiento a seguir en la construcción de modelos de predicción de la tendencia de los precios. El enfoque presentado supone, de antemano, un conocimiento del mercado y del instrumento financiero para el que se desea estimar la dirección de su movimiento futuro. Los procedimientos de la Figura 4.11 suponen una clara integración entre la construcción de estrategias de negociación y los modelos de predicción. Desde el punto de vista del análisis técnico, las variables que miden la acción del precio del activo negociado se utilizan en la definición de reglas de decisión. Los patrones en el comportamiento de los precios, dadas las reglas de decisión, son los

que identifican las señales para entrar y salir del mercado. En este sentido, lo que hacen los modelos de previsión es anticipar la aparición de dichas señales, que finalmente son confirmadas entre el modelo de previsión y las reglas de decisión antes de emitir las órdenes que abrirán y cerrarán las posiciones en el mercado.

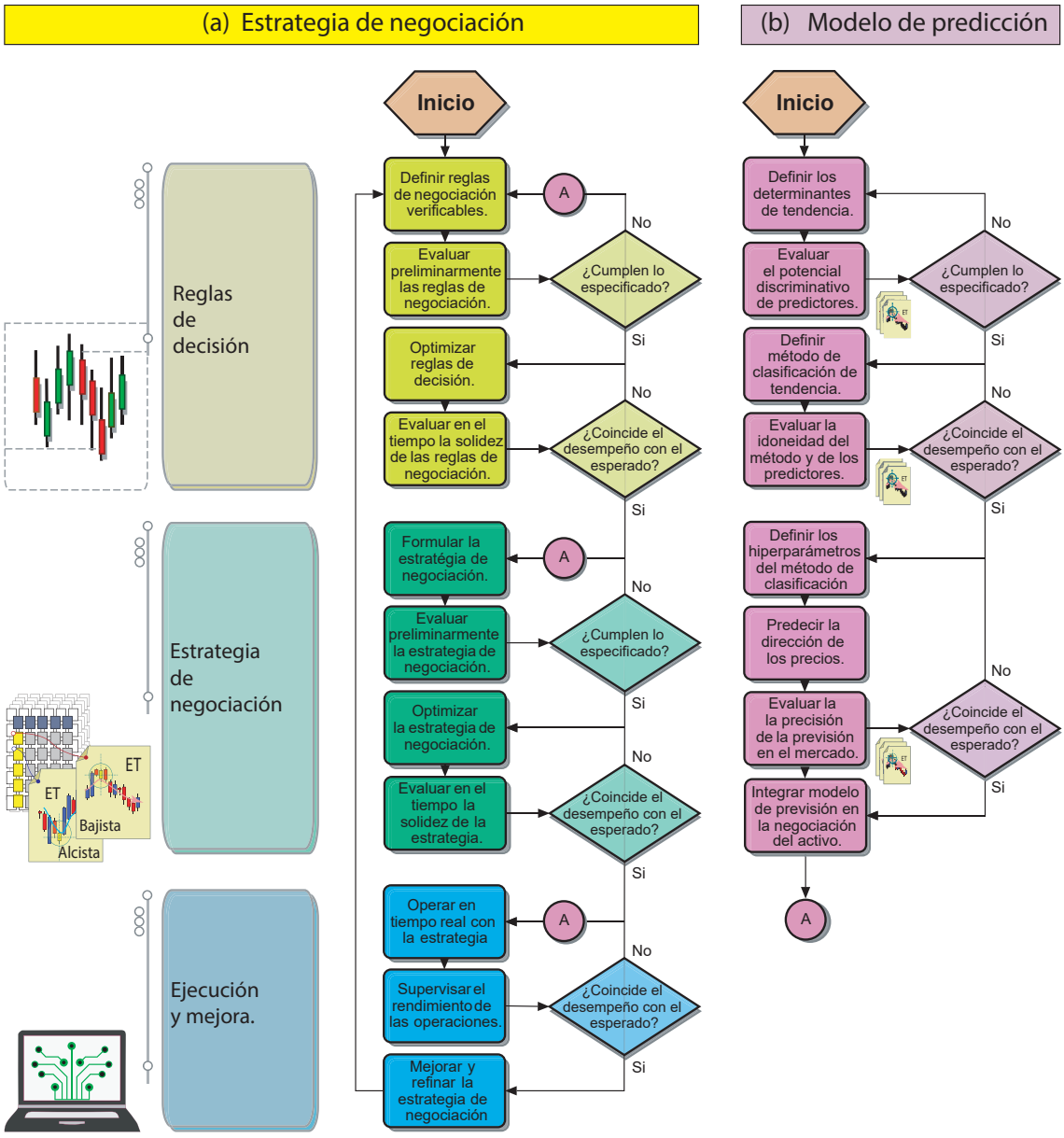


Figura 4.11: Estrategia de negociación y modelo de predicción.
Desarrollo de (**a**) estrategias de negociación y (**b**) modelos de previsión. Estos procedimientos integran elementos comunes, patrones o pautas en los precios, medidos por variables que componen la estrategia de negociación y el modelo de previsión. Los resultados obtenidos con el modelo de previsión se incorporan a la definición de las reglas de decisión, la estrategia de negociación y la negociación de activos en tiempo real.

Como se muestra en la Figura (4.11**b**), cabe señalar que la construcción de modelos de previsión se resume prácticamente en tres etapas: (*i*) identificación y selección de las variables predictoras que mejor influyen en la evolución de los precios y que, dado su poder discriminatorio, permiten diferenciar entre movimientos al alza y a la baja; (*ii*) identificación y selección del mejor método de clasificación, dado el poder discriminatorio de las variables predictoras a utilizar. Además de evaluar la idoneidad de las variables predictoras y de la técnica de clasificación a utilizar; (*iii*) definición de los hiperparámetros y selección de los criterios de evaluación que mejor ayuden a evaluar el rendimiento de la clasificación. Hay que tener en cuenta que, en cada una de estas subetapas, se evalúan los resultados obtenidos. Esto implica que cuando no se cumplen los resultados esperados, es necesario volver a la etapa anterior para realizar las correcciones oportunas. Una vez que el modelo predictivo cumple las expectativas en términos de legibilidad, interpretabilidad y precisión, puede integrarse en el modelo de negociación, bien como herramienta analítica para emitir juicios de valor antes de tomar decisiones de inversión, o como mecanismo de confirmación de las señales del mercado antes de emitir órdenes de mercado.

4.6 Selección de características

La selección de variables mediante el método de inclusión por pasos es una extensión del análisis discriminante que permite detectar el mejor subconjunto de variables discriminantes. Las variables elegibles son aquellas con valores de inclusión superiores a los criterios de entrada definidos en el análisis y al valor de tolerancia especificado ($\mathrm{T} = 0,001$). Antes de incluir una nueva variable en el análisis, las variables incluidas previamente se evalúan para su remoción. Una variable se descarta si su valor F de remoción es menor que el nivel de supresión especificado en el análisis. Si las variables a remover son demasiadas, se elimina la variable que produce el menor valor del estadístico para el resto de las variables. Este proceso finaliza cuando no quedan variables candidatas por remover. Teniendo en cuenta este contexto, a continuación se describe brevemente la fundamentación teórica utilizada en el proceso de selección de características con alto poder discriminativo. Este enfoque es el que se utilizará en la definición del proceso de selección de variables predictoras para la construcción del modelo de clasificación.

En la selección de características, la importancia de cada variable se evalúa individ-

ualmente con respecto al conjunto de datos. El operador de barrido simétrico aplicado por [550] se utiliza para reemplazar, en cada paso, la matriz $\mathbf{W}$ por una nueva matriz $\mathbf{W}^*$. La matriz $\mathbf{W}$, según la expresión 4.23, una vez incluidas las primeras q variables en el análisis, se puede expresar de la siguiente manera:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}_{q \times q} \tag{4.23}$$

Los nuevos elementos de la matriz $\mathbf{W}^*$ se obtienen utilizando la matriz $\mathbf{W}11^{-1}$, como se muestra en la expresión 4.24:

$$\mathbf{W}^* = \begin{bmatrix} -\mathbf{W}_{11}^{-1} & \mathbf{W}_{11}^{-1}\mathbf{W}_{12} \\ \mathbf{W}_{21}\mathbf{W}_{11}^{-1} & \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12} \end{bmatrix} \tag{4.24}$$

de modo que los nuevos elementos de la matriz $\mathbf{W}^*$ se concretan en la expresión 4.25:

$$\mathbf{W}^* = \begin{bmatrix} \mathbf{W}_{11}^* & \mathbf{W}_{12}^* \\ \mathbf{W}_{21}^* & \mathbf{W}_{22}^* \end{bmatrix} \tag{4.25}$$

Simultáneamente, la matriz $\mathbf{T}$ también sigue el mismo procedimiento que la matriz $\mathbf{W}$ y debe sustituirse por la matriz $\mathbf{T}^*$. Estos cálculos producidos durante el método de selección por pasos son los que permiten detectar las características más relevantes respecto al conjunto de datos. Además, estos análisis incluyen el cálculo de algunas medidas estadísticas como la tolerancia, los valores F de remoción e inclusión, las pruebas de igualdad de medias de grupo, entre otras medidas, como se describe a continuación:

4.6.1 Tolerancia

La tolerancia ψ_i, para un predictor potencial i, es la proporción de varianza no explicada por los predictores ya incluidos en el modelo. Es una medida de importancia relativa que determina la calidad del predictor potencial con respecto a otras variables. La tolerancia ψ_i de acuerdo con la expresión 4.26, se calcula como una relación entre la nueva matriz $\mathbf{W}^*$ y la matriz $\mathbf{W}$.

El cálculo de ψ_i, depende de si la variable i está incluida o no en el análisis y si

el valor wii es igual a cero o no. En el caso de que $w_{ii} = 0$, ψ_i es igual a cero. En el caso de que la variable i no esté incluida en el análisis y $wii \neq 0$, ψ_i se calcula como la relación entre la diagonal de $\mathbf{W}^*$ y la diagonal de $\mathbf{W}$. Finalmente, en el caso de que la variable i esté incluida en el análisis y $wii \neq 0$, ψ_i se calcula como el inverso de la relación entre la diagonal de $\mathbf{W}^*$ y la diagonal de $\mathbf{W}$.

$$\psi_i = \begin{cases} 0, & \text{si } w_{ii} = 0 \\ w_{ii}^* \left(w_{ii}\right)^{-1}, & \text{si la variable } i \text{ no está en el análisis y } w_{ii} \neq 0 \\ -w_{ii}^{-1} \left(w_{ii}^*\right)^{-1}, & \text{si la variable } i \text{ está en el análisis y } w_{ii} \neq 0. \end{cases} \tag{4.26}$$

En el método de selección por pasos, las variables independientes se introducen en el análisis en el orden en que disminuye la tolerancia.

Los estadísticos a los que se hace referencia a continuación se calculan para un predictor potencial cuando se cumplen los siguientes supuestos. (i) Cuando su tolerancia es mayor o igual que el límite de tolerancia establecido en el análisis. (ii) La inclusión de un nuevo predictor en el análisis no reduzca la tolerancia, de ninguna de las variables ya incluidas, por debajo o hasta el límite establecido. Obsérvese que la definición de un nivel de tolerancia protege la construcción del modelo de clasificación de la multicolinealidad y la singularidad. Además, evita la inclusión de variables que no sean relevantes o que puedan afectar la estabilidad y precisión del modelo.

4.6.2 Valor F de entrada

El valor F de entrada es una medida estadística que se calcula para cada predictor potencial con el fin de determinar su importancia relativa con respecto a las demás variables. El valor del estadístico F_i que se calcula para cada variable i y que permite determinar si dicha variable es elegible se presenta en la ecuación 4.27. Si el valor "F de entrada" (F_i) calculado para cada variable i supera el límite de entrada definido en el análisis, la variable i entra en el modelo.

$$F_i = \frac{(t_{ii}^* - w_{ii}^*)}{w_{ii}^*} \cdot \frac{(n - q - g)}{(g - 1)} \tag{4.27}$$

con $(g - 1)$, y $(n - q - g)$ grados de libertad.

Obsérvese que el Valor F de entrada calculado para la variable i mide la relación entre la variabilidad entre grupos y la variabilidad dentro de los grupos. Así, el valor F_i de entrada mide la contribución de la variable i a la variabilidad explicada por el modelo después de incluir la variable i en el modelo.

4.6.3 Valor F de remoción

El valor F de remoción es un estadístico calculado para cada variable predictiva potencial a fin de averiguar su relevancia con respecto al resto de variables. En la ecuación 4.28 se presenta la forma de obtener el valor del estadístico F_i. Este valor se calcula para cada variable i y permite determinar si esa variable es elegible o no. Si el valor "F de remoción" (F_i) que se calcula para cada variable i es menor al límite de remoción definido por el análisis, la variable i sale del modelo.

$$F_i = \frac{(w_{ii}^* - t_{ii}^*)}{t_{ii}^*} \cdot \frac{(n-q-g+1)}{(g-1)} \tag{4.28}$$

con $(g-1)$ y $(n-q-g+1)$ grados de libertad.

En el proceso de selección de características por pasos, los valores definidos para la tolerancia ($\psi_i = 0.001$), el nivel de entrada ($F_\varepsilon = 3.84$) y el nivel de eliminación ($F_r = 2.71$) son valores por defecto que han demostrado un mejor rendimiento en la selección de características. Sin embargo, los autores en [551] sugieren en términos de probabilidad un criterio de entrada entre 0.15 y 0.20 para favorecer la entrada de variables importantes.

4.6.4 Lambda de Wilks

La ecuación 4.29 detalla cómo se calcula el estadístico Lambda de Wilks Λ. Esta medida de referencia permite evaluar el poder discriminativo de las variables utilizadas en la construcción del modelo de clasificación. Este criterio se utiliza para probar la hipótesis nula de que no existen diferencias significativas entre las medias de los grupos.

$$\Lambda = \left| \frac{\mathbf{W}_{11}}{\mathbf{T}_{11}} \right| \tag{4.29}$$

con q, $g-1$, and $n-g$ grados de libertad.

Lambda de Wilks Λ según la ecuación 4.29 se define como el cociente entre el determinante de la matriz de varianza-covarianza dentro de grupos ($\mathbf{W}_{11}$) y el determinante de la matriz de varianza-covarianza total ($\mathbf{T}_{11}$). El valor del estadístico Lambda de Wilks Λ se encuentra en el intervalo [0, 1], cuanto más se acerque a cero indica una mayor separación entre las medias de los grupos y, por tanto, el subconjunto de variables seleccionadas son altamente discriminantes.

4.6.5 Prueba F aproximada para Lambda

Siempre que se cumplan los supuestos por defecto del análisis discriminante y se trate de un problema con tres o menos grupos, la transformación de la lambda en un estadístico F producirá un valor con una distribución F exacta. En el caso de problemas de clasificación con cuatro o más grupos, ocurrirá lo mismo con la primera y la segunda variables de entrada. Fuera de este contexto, la selección de predictores utilizando lambda de wilks Λ como criterio de selección producirá un valor F aproximado F_a. La ecuación 4.30 hace referencia a la prueba F aproximada F_a para lambda.

$$F_a = \frac{1 - \Lambda^{\mathfrak{S}}}{\Lambda^{\mathfrak{S}}} \cdot \frac{(r/\mathfrak{S} + 1 - qh/2)}{qh} \tag{4.30}$$

La prueba aproximada F_a para Lambda, también conocida como R de Rao [552], requiere la variable $\mathfrak{S}$, r y h cuyo cálculo se describe en las ecuaciones 4.31, 4.32 y 4.33:

$$\mathfrak{S}^2 = \begin{cases} (q^2 + h^2 - 5) \cdot (q^2 h^2 - 4)^{-1}, & \text{si } h^2 + q^2 \neq 5 \\ 1, & \text{en caso contrario} \end{cases} \tag{4.31}$$

$$r = (n - 1) - \frac{q + g}{2} \tag{4.32}$$

$$h = (g - 1) \tag{4.33}$$

donde qh y $(r/\mathfrak{S} + 1 - qh/2)$ son los grados de libertad. El valor del estadístico F_a es exacto cuando q o h es 1 o 2.

4.6.6 Distancia de Mahalanobis

La distancia de Mahalanobis es una medida útil introducida por Prasanta Chandra Mahalanobis en 1936 para determinar la similitud entre dos o más conjuntos de observaciones. La ecuación 4.34 calcula la distancia de Mahalanobis al cuadrado entre dos conjuntos $M^2_{\mathtt{a}\ \mathtt{b}}$. Esta expresión definida por [553] se utiliza como criterio de selección en la detección de variables que resulten ser más eficaces en la diferenciación entre grupos.

$$M^2_{\mathtt{a}\ \mathtt{b}} = -(n-g)\left(\sum_{i=1}^{q}\sum_{l=1}^{q} w^*_{il}\left(\bar{X}_{i\mathtt{a}} - \bar{X}_{i\mathtt{b}}\right)\left(\bar{X}_{l\mathtt{a}} - \bar{X}_{l\mathtt{b}}\right)\right) \tag{4.34}$$

El Valor F del estadístico $F_{\mathtt{a}\ \mathtt{b}}$ utilizado para evaluar la igualdad de medias entre los grupos $\mathtt{a}$ y $\mathtt{b}$ se basa en el cuadrado de la distancia de Mahalanobis $M^2_{\mathtt{a}\ \mathtt{b}}$ y se calcula a partir de la ecuación 4.35:

$$F_{\mathtt{a}\ \mathtt{b}} = M^2_{\mathtt{a}\ \mathtt{b}} \cdot \frac{n_{\mathtt{a}} n_{\mathtt{b}}}{(n_{\mathtt{a}} + n_{\mathtt{b}})} \cdot \frac{(n-q-g+1)}{q(n-g)} \tag{4.35}$$

La suma de las variaciones no explicadas por el modelo $\mathfrak{V}_r$ se calculan mediante la ecuación 4.36.

$$\mathfrak{V}_r = \sum_{\mathtt{a}=1}^{g-1}\sum_{\mathtt{b}=\mathtt{a}+1}^{g} \frac{4}{(4 + M^2_{\mathtt{a}\ \mathtt{b}})} \tag{4.36}$$

Esta medida ayuda a evaluar la calidad del modelo de clasificación en función del cuadrado de la distancia de Mahalanobis $M^2_{\mathtt{a}\ \mathtt{b}}$ y el número de grupos g. Mientras mas pequeña sea la suma de dichas variaciones no explicadas por las variables utilizadas en el modelo mejor sera el modelo. En caso contrario, cuando la suma de dichas variaciones es mayor significa que muchas de esas variaciones no se explican con las variables seleccionadas y utilizadas en el modelo y, por lo tanto, el modelo es poco preciso.

4.7 Análisis Biplot

Los métodos Biplot han demostrado ser una potente herramienta para la visualización e inspección de grandes matrices de datos multivariantes en un espacio de dimensión reducida. Estos métodos permiten representar y analizar en un mismo plano cartesiano

los marcadores fila (observaciones) y columna (variables) de la matriz de datos, donde estos últimos corresponden a las características medidas sobre las observaciones. La eficacia del análisis de las interrelaciones entre observaciones y variables ha probado ser efectiva con este tipo de representaciones. Concretamente, con el análisis de patrones de similitud entre observaciones y patrones de covariación entre variables. A continuación se hace una breve descripción sobre los fundamentos teóricos, reglas de interpretación y otros aspectos relacionados con los métodos biplot.

4.7.1 Fundamentos teóricos

Sea $\mathbf{X}\,\langle I \times J\rangle$ una matriz de datos multivariantes de tamaño $I \times J$ formada por las características de J variables continuas medidas en I observaciones. La matriz de datos $\mathbf{X}$ puede representarse gráficamente en un biplot de dimensión S mediante marcadores $\mathbf{a}_i$, $i = \{1, 2, \ldots, I\}$ para sus filas y marcadores $\mathbf{b}_j$, $j = \{1, 2, \ldots, J\}$ para sus columnas. Los marcadores $\mathbf{a}_i$ y $\mathbf{b}_j$ se eligen de forma que el producto interno de $\mathbf{a}_i^\top \mathbf{b}_j$ esté lo más cerca posible del elemento x_{ij} de la matriz $\mathbf{X}$.

$$x_{ij} \approx \mathbf{a}_i{}' \mathbf{b}_j \tag{4.37}$$

El ajuste de una matriz de datos de dos vías a una matriz de rango dos precisa de la descomposición de valores singulares (SVD). De este modo, las dos primeras componentes $\mathbf{A}$ y $\mathbf{B}$, siempre y cuando expliquen la mayor parte de la variabilidad total de los datos, pueden representarse y mostrar la información de $\mathbf{X}$ en el biplot. Así, la matriz $\mathbf{X}$ se puede aproximar mediante la ecuación 4.38, disponiendo los marcadores adecuados en dos matrices $\mathbf{A}$ y $\mathbf{B}$. Por tanto, la representación en un mismo gráfico de las I filas de la matriz $\mathbf{A}$ y las J columnas de la matriz $\mathbf{B}$ recibe el nombre de biplot.

$$\mathbf{X} \approx \mathbf{A}\mathbf{B}^\top \tag{4.38}$$

Mediante la descomposición en valores singulares (SVD), la matriz $\mathbf{X}$ puede expresarse como se muestra en la ecuación 4.39.

$$\mathbf{X} \cong \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{4.39}$$

donde $\mathbf{U}$ y $\mathbf{V}$ son matrices con vectores columna ortonormales obtenidos de los vectores propios de los productos matriciales de $\mathbf{X}\mathbf{X}^\top$ y $\mathbf{X}^\top\mathbf{X}$. De modo que $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$. La matriz diagonal $\mathbf{D}$ está formada por los valores singulares de la matriz

$\mathbf{X}$. Como ya se ha mencionado, la mejor aproximación de la matriz de datos $\mathbf{X}$ a una matriz de rango S viene determinada por los primeros valores y vectores singulares S. Por consiguiente, la expresión 4.39 puede definirse tal como aparece en 4.40.

$$\mathbf{X} \cong \sum_{s=1}^{S} \lambda_s \mathbf{u}_s \mathbf{v}_s^{\top} \tag{4.40}$$

De acuerdo con la ecuación 4.40 la matriz $\mathbf{X}$ se puede aproximar por la suma de un número $\mathbf{s}$ de términos de la forma $\lambda_s \mathbf{u}_s \mathbf{v}_s^{\top}$. Nótese que λ_s es el s-ésimo valor singular, mientras que $\mathbf{u}_s$ y $\mathbf{v}_s$ son el s-ésimo vector singular izquierdo y derecho, respectivamente. El número de componentes necesarios para conseguir una aproximación precisa de $\mathbf{X}$ depende del rango de la matriz $\mathbf{X}$ y de sus valores singulares.

Al operar con los componentes adecuados, definidos en la expresión 4.39, se obtienen las coordenadas de los marcadores fila y columna en las matrices $\mathbf{A}$ y $\mathbf{B}$.

$$\mathbf{A} = \mathbf{U}\mathbf{D}^{\omega}, \quad \mathbf{B} = \mathbf{V}\mathbf{D}^{1-\omega}, \quad \omega = \{0, 1\} \tag{4.41}$$

La representación simultánea de las matrices $\mathbf{A}$ y $\mathbf{B}$ en el mismo espacio produce distintos tipos de biplots, como se muestra en 4.42.

$$\textbf{Biplot} \leftarrow \begin{cases} \mathbf{A} = \mathbf{U}\mathbf{D}, \quad \mathbf{B} = \mathbf{V}, & \text{si } \omega = 1 \\ \mathbf{A} = \mathbf{U}, \quad \mathbf{B} = \mathbf{V}\mathbf{D}, & \text{en caso contrario} \end{cases} \tag{4.42}$$

Según la ecuación 4.42 en el primer caso se observa que al sustituir el valor de $\omega = 1$ en la expresión 4.41, la matriz $\mathbf{A}$ de marcadores fila y con mejor información contribuye a una representación $\mathbf{JK - Biplot}$. En caso contrario, cuando $\omega = 0$, la matriz $\mathbf{B}$ de marcadores columna y con mejor información contribuye a una representación $\mathbf{GH - Biplot}$. Obsérvese que tomando de cada caso las matrices $\mathbf{A}$ y $\mathbf{B}$ con mejor información, su representación simultánea en el mismo espacio contribuye a un análisis $\mathbf{HJ - Biplot}$.

4.7.2 Reglas de interpretación

Una representación Biplot, como ya se ha mencionado en la ecuación 4.37, está formada por marcadores fila $\mathbf{a}_i$ que aparecen como puntos y marcadores columna $\mathbf{b}_j$ en forma de vectores. El estudio de los ejes factoriales, las contribuciones y la detección de interacciones entre variables y observaciones ayuda a descubrir patrones y relaciones

ocultas en la estructura de los datos. Este nivel de detalle se alcanza cuando se adoptan algunas reglas de interpretación que ayudan a este tipo de análisis.

Teniendo en cuenta estos aspectos, a continuación se esbozan brevemente las reglas de interpretación que deben considerarse en el análisis biplot.

A. Relaciones entre observaciones

Al explorar las relaciones entre observaciones se pueden identificar tanto grupos de observaciones similares como relaciones entre variables que caracterizan a estos grupos. Este proceso se lleva a cabo teniendo en cuenta las siguientes consideraciones.

- El valor de x_{ij} se aproxima mediante el producto escalar de la longitud del vector $\mathbf{b}_j$ por la proyección del marcadores fila $\mathbf{a}_i$ sobre el vector $\mathbf{b}_j$.
- El valor de x_{ij} se aproxima mediante el producto escalar de la longitud del vector $\mathbf{b}_j$ por la proyección del marcadores fila $\mathbf{a}_i$ sobre el vector $\mathbf{b}_j$.
- La proyección de los marcadores fila $\mathbf{a}_i$ sobre los vectores $\mathbf{b}_j$ se utiliza para analizar las relaciones entre observaciones y variables.
- La distancia entre observaciones $\mathbf{a}_i$ puede entenderse como una medida de disimilaridad.
- La proyección de las observaciones $\mathbf{a}_i$ al vector $\mathbf{b}_j$ define el orden que siguen dichas observaciones en esa variable.
- Una observación $\mathbf{a}_i$ cuanto más se aleja de la punta del vector $\mathbf{b}_j$ más se aleja de la media de esa variable.

B. Relaciones entre variables

El análisis de las relaciones entre variables, además de ayudar a descubrir las variables que más influyen en la distribución de las observaciones, proporciona información valiosa sobre la estructura de las variables en la matriz de datos. La lectura de estas interrelaciones se facilita teniendo en cuenta las reglas de interpretación que se mencionan a continuación.

- El origen del biplot puede entenderse como el valor medio de las diferencias entre observaciones $\mathbf{a}_i$ con respecto al valor de las variables $\mathbf{b}_j$.

- La longitud del vector $\mathbf{b}_j$ es una medida de variabilidad de las mediciones en esa variable.

- La punta del vector $\mathbf{b}_j$ señala la dirección en la que se incrementan los valores de esa variable.

- El coseno de los ángulos que se forman entre vectores $\mathbf{b}_j$ es una medida de covariación entre esas variables. Así, ángulos aproximados de 90° sugieren independencia entre variables, ángulos agudos cercanos a 0° sugieren variables altamente correlacionadas, ángulos aproximados de 180° indican variables inversamente correlacionadas.

C. Plano factorial y contribuciones

Los ejes factoriales constituyen una dimensión en la que se comparan variables y observaciones. Los planos factoriales se eligen de forma que sean capaces de explicar la mayor parte de la variabilidad de los datos. Además, el análisis de las contribuciones entre ejes y variables puede mostrar qué variables son más importantes en cada eje y cómo contribuyen a la variabilidad en ese eje. Teniendo en cuenta estas condiciones, a continuación se exponen las reglas de interpretación más importantes que deben tenerse en cuenta durante el análisis.

- La independencia entre ejes factoriales está confirmada por el ángulo de 90° que se forma entre ellos.

- La relación de los ejes factoriales con las variables $\mathbf{b}_j$ y observaciones $\mathbf{a}_i$ es interpretable a través de las contribuciones relativas.

- Las contribuciones relativas son proporciones de variabilidad explicadas en dos direcciones, desde los ejes factoriales del biplot a los elementos fila y columna, y desde los elementos fila y columna a los ejes factoriales.

- Las relaciones entre ejes factoriales y variables $\mathbf{b}_j$ o ejes factoriales y observaciones $\mathbf{a}_i$ se denominan contribuciones relativas del factor a los elementos columna o fila, respectivamente.

- Las contribuciones relativas permiten identificar qué variables $\mathbf{b}_j$ están mejor representadas en un plano factorial y qué variables $\mathbf{b}_j$ están mejor relacionadas con cada eje factorial.

4.7.3 Análisis GH-Biplot

Esta técnica de análisis multivariante es útil para el análisis de grandes matrices de datos continuos. El método se destaca por su calidad de representación que lo hace especialmente útil en el análisis de relaciones entre variables. En efecto, para las variables, la calidad de la representación es superior a la de las observaciones. A continuación se describen brevemente las propiedades más importantes de esta técnica para el análisis de datos.

El producto escalar de las variables de la matriz $\mathbf{X}$ coincide, según la ecuación 4.43, con el producto escalar de sus marcadores columna $\mathbf{B}$.

$$\mathbf{X}'\mathbf{X} = \mathbf{B}\mathbf{B}' \tag{4.43}$$

La propiedad definida por la ecuación anterior, donde $\mathbf{X}$ es la matriz de datos centrada y $\mathbf{B}$ es la matriz de coordenadas de la variable en el subespacio de dimension reducida permite destacar tres aspectos importantes.

(**i**). El cuadrado de la longitud del marcador columna $\mathbf{b}_j$ se aproxima a la varianza de las medidas en la variable de la matriz $\mathbf{X}$. Por consiguiente, la longitud de los vectores definidos por los marcadores columna $\mathbf{b}_j$ se acerca al valor de la desviación típica de dichas variables.

$$\parallel \ b_j^2 \ \parallel = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \tag{4.44}$$

La expresión anterior corresponde a la norma del vector columna b_j de la matriz de datos $\mathbf{X}$ estandarizada. Lo que significa que la norma del marcador columna b_j corresponde a la desviación típica de la variable en cuestión x_j. Esta expresión es la raíz cuadrada de la varianza de la columna j de la matriz $\mathbf{X}$. Donde x_{ij} es el valor de la variable x_j para la observación i.

Si la matriz de datos $\mathbf{X}$ está centrada, esto implica que la media de cada columna es cero y su desviación típica es 1. Esto resulta útil cuando se compara la importancia relativa de cada variable j de $\mathbf{X}$. Cuanto mayor sea la norma, mayor será la variabilidad de las medidas de esta variable respecto de otras en la matriz de datos.

(**ii**). El valor del coseno del ángulo formado entre dos vectores columna b_j se aproxima

a la correlación existente entre esas variables.

$$\cos\left(b_i, b_j\right) = \frac{b_i^\top b_j}{\mid b_i \mid\mid b_j \mid} \cong \operatorname{corr}\left(x_i, x_j\right) = \frac{\operatorname{cov}(x_i, x_j)}{\sigma_i \sigma_j} \tag{4.45}$$

El coseno del ángulo entre dos marcadores columna b_i y b_j se puede calcular usando su producto punto y sus normas. $b_i^\top b_j$ representa el producto punto entre los dos vectores, y $\mid b_i \mid$ y $\mid b_j \mid$ son las normas de los vectores. Como los marcadores columna b_i y b_j están unidos por su origen, el coseno de su ángulo puede interpretarse como una medida de la similitud entre los vectores. Así, valores cercanos a 1 sugerirían similitud (altamente correlacionados) y valores cercanos a -1 indicarían que están inversamente correlacionados.

Por otra parte, si $r_{ij} = \operatorname{cov}(x_i, x_j)/\sigma_i \sigma_j$ es el coeficiente de correlación de Pearson que mide la fuerza y la dirección de la relación lineal entre las variables x_i y x_j, $\operatorname{cov}(x_i, x_j)$ es la covarianza entre x_i y x_j, σ_i y σ_j son las desviaciones estándar de x_i y x_j, respectivamente. Por lo tanto, si $r_{ij} > 0$, existe una relación lineal positiva entre x_i y x_j. Si $r_{ij} < 0$, existe una relación lineal negativa. Si $r_{ij} = 0$, sugiere que no hay relación lineal entre x_i y x_j.

(**iii**). La distancia Euclídea d^2 entre dos variables x_i y x_j equivale a la distancia existente entre los marcadores columna b_i y b_j de dichas variables.

$$d^2\left(x_i, x_j\right) = d^2\left(b_i, b_j\right) \tag{4.46}$$

Según la expresión anterior, la distancia Euclídea al cuadrado entre las variables x_i y x_j se define como la diferencia entre dichas variables multiplicada por su transpuesta. De modo que $d^2\left(x_i, x_j\right) = \left(x_i - x_j\right)^\top \left(x_i - x_j\right)$.

Al expandir esta expresión y aplicando la propiedad del producto punto para separar la norma al cuadrado de cada variable y el producto punto de ambas variables se tiene que $\left(x_i - x_j\right)^\top \left(x_i - x_j\right) = \|x_i\|^2 + \|x_j\|^2 - 2\left(x_i^\top x_j\right)$.

Al sustituir los marcadores columna b_i y b_j, de los vectores x_i y x_j se tiene la suma de los cuadrados de los marcadores columna b_i y b_j, menos el doble del producto punto entre ellos $\|x_i\|^2 + \|x_j\|^2 - 2\left(x_i^\top x_j\right) = |b_i|^2 + |b_j|^2 - 2\left(b_i^\top b_j\right)$.

De manera que $d^2\left(b_i, b_j\right)$ es la distancia Euclídea al cuadrado entre los puntos en el subespacio de dimension reducida definido por los marcadores columna b_i y b_j, obtenidos de la expresión anterior. Esta distancia mide la similitud entre las variables x_i y x_j en términos de sus correlaciones y covarianzas en el subespacio del análisis

GH – Biplot.

(**iv**). La distancia de Mahalanobis entre dos observaciones x_i y x_j de la matriz $\mathbf{X}$ se aproxima a la distancia Euclídea entre los marcadores fila a_i y a_j, donde S es la matriz de varianzas-covarianzas.

$$(x_i - x_j)^\top S^{-1} (x_i - x_j) \cong (a_i - a_j)^\top (a_i - a_j) \tag{4.47}$$

Teniendo en cuenta que, las coordenadas en $\mathbf{B}$ son las contribuciones de las variables en los ejes factoriales, $x_i = \mathbf{B}a_i$ y $x_j = \mathbf{B}a_j$ se obtiene la siguiente equivalencia: $(x_i - x_j)^\top S^{-1} (x_i - x_j) = (\mathbf{B}a_i - \mathbf{B}a_j)^\top S^{-1} (\mathbf{B}a_i - \mathbf{B}a_j)$.

De la anterior expresión se factoriza $\mathbf{B}(a_i - a_j)$ del producto de matrices, de forma que la ecuación en términos de las coordenadas a_i y a_j queda definida como: $(x_i - x_j)^\top S^{-1} (x_i - x_j) = (\mathbf{B}(a_i - a_j))^\top S^{-1} (\mathbf{B}(a_i - a_j))$.

Haciendo uso de la propiedad de simetría de la matriz S y de la inversa de S, es decir, $S^\top = S$ y $S^{-1} = S$, se obtiene que $\mathbf{B}^\top S^{-1}\mathbf{B} = \left(\mathbf{B}^\top S^{-1}\mathbf{B}\right)^\top = \mathbf{B}^\top S^{-\top}\mathbf{B} = \mathbf{B}^\top S^{-1}\mathbf{B}$. De este modo, la expresión anterior se define como $(x_i - x_j)^\top S^{-1} (x_i - x_j) = (a_i - a_j)^\top \left(\mathbf{B}^\top S^{-1}\mathbf{B}\right) (a_i - a_j)$.

Por lo tanto, la expresión final $(a_i - a_j)^\top \left(\mathbf{B}^\top S^{-1}\mathbf{B}\right) (a_i - a_j) = (a_i - a_j)^\top (a_i - a_j)$ es equivalente a la del enunciado 4.47. En consecuencia, estas expresiones establecen una relación entre la distancia de las observaciones en el espacio de las variables originales y el subespacio de dimensión reducida con sus respectivos marcadores de fila.

(**v**). El producto escalar de los marcadores fila es igual al producto escalar de las filas de la matriz $\mathbf{X}$ con métrica $(\mathbf{X}^\top\mathbf{X})^{-1}$ dentro del espacio columnas. Teniendo en cuenta que $\mathbf{X} = \mathbf{AB}'$ se obtiene la siguiente ecuación:

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{AA}' \tag{4.48}$$

Sustituyendo $\mathbf{X} = \mathbf{AB}'$ en la primera parte de la ecuación anterior se obtiene $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{AB}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{BA}'$. Nótese que $\mathbf{A}$ es la matriz de coordenadas de los marcadores fila y $(\mathbf{B}'\mathbf{B})^{-1}$ es también la métrica en el espacio de las columnas. Lo que da lugar al hecho de que $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{AA}'$.

En este sentido, esta propiedad es útil en el análisis **GH – Biplot** porque permite visualizar la relación entre los marcadores fila y los datos en el espacio de las columnas y comprender la contribución de cada variable en la estructura de los datos.

4.7.4 Medidas de bondad del ajuste

En el análisis estadístico, es esencial evaluar la calidad con la que los modelos utilizados se ajustan a la estructura de los datos. Las medidas de bondad de ajuste se utilizan para determinar la calidad del ajuste y permiten la comparación entre distintos modelos. En el contexto del análisis GH-Biplot, a continuación se describen brevemente las principales medidas de bondad de ajuste utilizadas para evaluar el desempeño de este método.

A. Variabilidad explicada

La bondad de ajuste proporcionada por el método $\mathbf{GH-Biplot}$ en un subespacio de dimensión reducida R puede evaluarse conociendo la proporción de variabilidad de la matriz $\mathbf{X}$ que está siendo explicada mediante su matriz $\tilde{\mathbf{X}}$ aproximada. La "variabilidad total" o también denominada "suma de cuadrados total" definida por $\mathfrak{Vt}_X$ en la ecuación 4.49, hace referencia a la suma de cuadrados de los elementos de la matriz $\mathbf{X}$.

$$\mathfrak{Vt}_X = \| \ \mathbf{X} \ \|^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(x_{ij}^2\right) \tag{4.49}$$

La expresión anterior se utiliza en el contexto del análisis de datos para evaluar la cantidad total de información presente en un conjunto de variables dispuestas en la matriz de datos $\mathbf{X}$ antes de realizar una representación $\mathbf{GH-Biplot}$.

Ahora, la variabilidad total $\mathfrak{Vt}_X$ de la matriz de datos $\mathbf{X}$ puede dividirse en dos partes. La variabilidad explicada por el modelo o su respectiva aproximación, y la variabilidad residual que no puede ser explicada por el modelo o la aproximación. La ecuación en 4.50 se denomina identidad de suma de cuadrados. Esta expresión define que la suma total de cuadrados de la matriz de datos $\mathbf{X}$ se desagrega en la suma de cuadrados explicada por el modelo o su aproximación $\tilde{\mathbf{X}}$, y la suma de cuadrados no explicada por el modelo, que se mide por la distancia al cuadrado entre $\mathbf{X}$ y $\tilde{\mathbf{X}}$.

$$\left\| \ \mathbf{X} \ \right\|^2 \quad = \quad \left\| \ \tilde{\mathbf{X}} \ \right\|^2 + \left\| \ \mathbf{X} - \tilde{\mathbf{X}} \ \right\|^2 \tag{4.50}$$

Desde un punto de vista estadístico, esta identidad es útil en el análisis de datos, ya que permite evaluar el grado de ajuste del modelo o de su aproximación. Además, su uso resulta práctico a la hora de explicar la variabilidad del conjunto de datos, así

como al realizar análisis de varianza y formular pruebas de hipótesis.

La ecuación de identidad también puede expresarse a partir del cuadrado de los valores singulares, como ocurre en la ecuación 4.51. Los componentes λ_r corresponden a valores singulares de la matriz $\mathbf{X}$. Esta expresión relaciona la varianza total de los datos (en la primera parte de la igualdad) con la varianza explicada y la variación residual. Estos valores se definen como una función de los cuadrados de los valores singulares.

$$\sum_{(r)=1}^{\tilde{R}} \lambda_{(r)}^2 \quad = \quad \sum_{(r)=1}^{R} \lambda_{(r)}^2 + \sum_{(r)=R+1}^{\tilde{R}} \lambda_{(r)}^2 \tag{4.51}$$

Los primeros R valores singulares suelen explicar la mayor parte de la variabilidad de los datos y tienden a ser los más importantes en el análisis. En líneas generales, los valores singulares restantes, de $R+1$ a $\tilde{R}$, explican la varianza residual que los primeros valores singulares R no pueden explicar. Por lo tanto, la ecuación, además de que permite determinar la cantidad de varianza explicada por cada conjunto de valores singulares, se utiliza para seleccionar el número adecuado de ejes factoriales para llevar a cabo el análisis.

Por otro lado, otra forma útil de medir la bondad de ajuste del modelo se resume en la función $f(R)$ de la ecuación 4.52. Esta expresión mide la proporción de variabilidad total explicada por los primeros R valores singulares al cuadrado, con respecto a la suma total de los valores singulares al cuadrado.

$$f(R) = \frac{\sum_{r=1}^{R} \lambda_{(r)}^2}{\sum_{r=1}^{\tilde{S}} \lambda_{(r)}^2}, \tag{4.52}$$

donde $\tilde{S}$ es el rango de la matriz X y $\lambda_{(r)}^2$ representa el cuadrado del r-ésimo valor singular. La proporción $f(R)$ es una medida de la bondad de ajuste del modelo y se suele utilizar en el análisis de componentes principales y otros métodos que implican descomposiciones de valores singulares (SVD). Debe tenerse en cuenta que cuanto mayor sea la proporción de variabilidad explicada por el modelo $f(R)$, utilizando los primeros R valores singulares, mejor será la bondad de ajuste del método, lo que significa que una pequeña proporción de la variabilidad restante puede ser explicada por los valores singulares restantes. De lo contrario, si sólo hay una pequeña proporción de variabilidad explicada por los R primeros valores singulares, significa que hay un alto porcentaje de variabilidad explicada en otras dimensiones superiores.

B. Calidad de representación

La fiabilidad de las interpretaciones realizadas sobre el análisis GH-Biplot, en el que se representan simultáneamente variables y observaciones, viene determinada por la bondad de ajuste del método y la calidad de representación de los datos. En esta sección se describen brevemente las medidas más utilizadas para evaluar la calidad de representación de variables y observaciones.

La calidad de representación de las variables viene determinada, en parte, por la matriz de varianzas - covarianzas S. Según la ecuación 4.53, la matriz de varianzas - covarianzas S, está definida por las matrices de vectores propios $\mathbf{U}$ y $\mathbf{V}$ y la matriz diagonal $\mathbf{D}$ que contiene los valores singulares de $\mathbf{X}$.

$$\mathrm{S} = \mathbf{V}\mathbf{D}\mathbf{U}^{\top}\mathbf{U}\mathbf{D}\mathbf{V}^{\top} \tag{4.53}$$

Dado que $\mathbf{A}'\mathbf{A} = \mathbf{U}'\mathbf{U} = \mathbf{I}$, en la ecuación original se puede sustituir $\mathbf{U}^{\top}\mathbf{U}$ por $\mathbf{I}$, La matriz S, se expresa como se define en 4.54.

$$\mathrm{S} = \mathbf{V}\mathbf{D}^{2}\mathbf{V}^{\top} \tag{4.54}$$

Esta expresión indica que la matriz de varianzas - covarianzas S se compone de la matriz de vectores propios ortonormales $\mathbf{V}$ (que refleja la calidad de representación de las variables de $\mathbf{X}$) y los cuadrados de los valores singulares $\mathbf{D}^2$ (que reflejan la calidad de representación de las observaciones de $\mathbf{X}$).

A partir de la expresión anterior, la suma del cuadrado de los valores singulares de S es igual a la norma de Frobenius $\|\cdot\|_F$ y a la traza de la matriz $\mathrm{tr}(\cdot)$, en la forma referida en la ecuación 4.55.

$$\sum_{s=1}^{\tilde{S}} \lambda^4 = \quad \left\| \mathbf{V}\mathbf{D}^{2}\mathbf{V}^{\top} \right\|_F^2 = \quad \mathrm{tr}(\mathrm{S}^2) \tag{4.55}$$

A esta igualdad se la puede denominar "la relación entre la suma de los cuadrados de los valores singulares, la norma de Frobenius de la matriz de vectores propios al cuadrado y la traza de la matriz de varianzas-covarianzas al cuadrado".

La expresión $\sum_{s=1}^{\tilde{S}} \lambda^4$ es la suma de los cuadrados de los valores singulares a la cuarta de la matriz $\mathbf{X}$, donde $\tilde{S}$ es el rango de la matriz.

Esta medida permite evaluar la calidad de representación de la matriz $\mathbf{X}$ a partir de su descomposición en valores singulares (SVD). De manera que, una mayor suma de

cuadrados de los valores singulares a la cuarta parte indica una mejor representación de la matriz $\mathbf{X}$ en cuanto a su estructura latente. Por tanto, esta medida también resulta útil para determinar el número óptimo de dimensiones que deben incluirse en un modelo de análisis de valores singulares o de análisis factorial.

Considerando $\sum_{s=1}^{\tilde{S}} \lambda^4$ como punto de referencia sobre el que se evalúa la calidad de representación de los marcadores columna, la aproximación matricial $\mathbf{X}$ determinada para un subespacio de dimensión R se divide sobre dicho punto de referencia como se indica en la ecuación 4.56.

$$\mathrm{C_r} = \frac{\sum_{r=1}^{R} \lambda^4}{\sum_{r=1}^{\tilde{S}} \lambda^4} \tag{4.56}$$

Es de notar que el índice $\mathrm{C_r}$ mide la proporción del modelo aproximado representado en el subespacio de dimensión R con respecto al espacio de las variables en la matriz $\mathbf{X}$. Cuanto mayor sea el valor de $\mathrm{C_r}$, mejor será la calidad de representación de las variables.

A diferencia de los marcadores columna, que son los mejor representados en el subespacio de dimensión reducida R, los marcadores fila no gozan de esta propiedad. Sin embargo, la calidad de representación de los marcadores fila viene determinada por la relación $\mathrm{R}/\tilde{S}$. Obsérvese que el valor R corresponde a la aproximación del modelo, mientras que $\tilde{S}$ es la suma de los cuadrados de los valores singulares de $\mathbf{X}\mathrm{S}^{-1}\mathbf{X}^{\top}$ sobre el espacio de filas de la matriz $\mathbf{X}$. Por tanto, cuanto mayor sea la proporción $\mathrm{R}/\tilde{S}$, mejor será la calidad de representación de las observaciones.

4.7.5 Contribuciones

En el análisis e interpretación de las relaciones de interacción entre ejes factoriales, variables y observaciones, no sólo es esencial medir y evaluar el ajuste del modelo, dada su aproximación a la estructura de los datos, sino también evaluar el ajuste de cada una de las variables y observaciones. En esta sección se analiza brevemente el concepto de contribución como medida de ajuste individual. Esta descripción sólo tiene en cuenta el papel desempeñado por las variables, dada su calidad de representación definida en el GH-Biplot.

A. Contribución relativa del elemento al factor

Esta medida de desempeño, definida por sus siglas como $CR.E_jF_l$ y que se detalla en la ecuación 4.57, permite determinar la contribución que un elemento columna E_j, dado su poder explicativo, tiene en el comportamiento subyacente del factor F_l. Así, esta contribución mide esa parte de variabilidad del factor l que está siendo explicada por el elemento variable j.

$$CR.E_jF_l = \frac{h_{jl}^2}{\lambda_l^2} \tag{4.57}$$

B. Contribución relativa del factor al elemento

Esta medida de desempeño, definida por sus siglas como $CR.F_lE_j$ y que se detalla en la ecuación 4.58, permite medir la contribución que un factor F_l, dado su poder explicativo, tiene sobre el elemento columna E_j. Por tanto, este índice mide la contribución, es decir, la parte de la variabilidad de cada uno de los elementos variables j que está siendo explicada por el factor l.

$$CR.F_lE_j = \frac{h_{jl}^2}{\sum_{k=1}^{r} j_{jk}^2} \tag{4.58}$$

Implícitamente, esta métrica permite reconocer, en la forma en que se distribuyen las observaciones, las variables responsables de esa distribución en los ejes factoriales.

El análisis de las contribuciones individuales puede hacerse en dos direcciones, desde los factores hacia los elementos variables o en dirección opuesta. Este nivel de detalle puede ser necesario en algunos estudios cuando se desea analizar las relaciones e interacciones entre factores y variables específicas. En consecuencia, estos enfoques pueden ser útiles cuando se desea analizar la influencia que estas contribuciones ejercen sobre el valor o el significado tanto de variables como de factores.

4.7.6 Diagnóstico de modelos

El diagnóstico de modelos mediante métodos de visualización de datos continuos, en particular a partir de métodos biplot, se basa en las aportaciones de Gollob [554] y Mandel [555]. Los primeros estudios desarrollados por Gollob mostraron que el producto escalar de dos vectores puede conducir a la definición de modelos bajo aproximaciones bilineales. Más tarde, Mandel descubrió que las interacciones de los modelos bilin-

eales, además de aproximarse mediante términos multiplicativos, pueden representarse en un biplot. Teniendo en cuenta estos aportes, a continuación se ofrece una breve descripción sobre este tipo de procedimientos para el diagnóstico de modelos.

Las contribuciones de Bradu y Gabriel mostraron que el análisis biplot, además de ser una herramienta útil para el análisis descriptivo de datos multivariantes, también ayuda en el diagnóstico de modelos, especialmente para datos procedentes de tablas de dos vías. Los resultados obtenidos con sus estudios mostraron que, inspeccionando la posición geométrica de los marcadores columna $\mathbf{b}_j$ y fila $\mathbf{a}_i$, es posible deducir el modelo bilineal que mejor se ajusta a la estructura de los datos.

En general, un *modelo aditivo* es ajustable a una matriz de dos vias $\mathbf{X}$ de rango 2 cuando se cumplen dos condiciones. Además de que los marcadores columna $\mathbf{b}_j$ y los marcadores fila $\mathbf{a}_i$ sean colineales, las dos líneas $\mathbf{L}_b$ y $\mathbf{L}_a$ que unen los subconjuntos de marcadores columna y fila deben ser ortogonales. En cambio, los *modelos Tukey*, a diferencia de los anteriores, se caracterizan por el hecho de que existe interacción entre las rectas $\mathbf{L}_b$ y $\mathbf{L}_a$ definidas por los marcadores columna $\mathbf{b}_j$ y fila $\mathbf{a}_i$, lo que hace posible el ajuste de un modelo basado en la interacción.

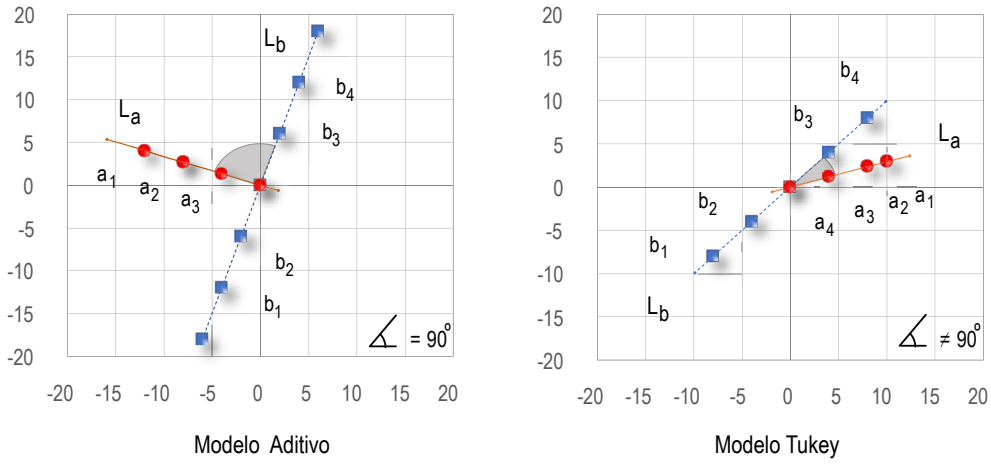


Figura 4.12: Diagnóstico de modelos.
Modelos bilineales que explican la variable dependiente $\mathbf{y}_{ij}$ a partir de los efectos de los factores fila $\mathbf{a}_i$, columna $\mathbf{b}_j$ e interacción $(\mathbf{a}_i)(\mathbf{b}_j)$ fila-columna. (**a**) Modelo Aditivo de la forma: $\mathbf{y}_{ij} = \mu + \gamma\mathbf{a}_i + \delta\mathbf{b}_j$, donde $a_i \in L_a$ y $b_j \in L_b$. Rectas L_a y L_b ($\measuredangle = 90°$) perpendiculares. (**b**) Modelo Tukey de la forma: $\mathbf{y}_{ij} = \mu + \gamma\mathbf{a}_i + \delta\mathbf{b}_j + \lambda(\gamma\mathbf{a}_i)(\delta\mathbf{b}_j)$, donde $a_i \in L_a$ y $b_j \in L_b$. Rectas L_a y L_b ($\measuredangle \neq 90°$) no perpendiculares.

En algunos casos no se sabe de antemano si hay interacción o no cuando se analiza la influencia que una variable puede tener en el comportamiento de otra mediante

tablas bidireccionales. Las aportaciones de Bradu y Gabriel muestran que, inspeccionando mediante una analisis biplot la disposición geométrica de las líneas que unen los respectivos marcadores columna y marcadores fila, es posible identificar el modelo bilineal que mejor describe la estructura de los datos.

El modelo aditivo de la Figura (4.12a) queda expresado en la ecuación 4.59. Este modelo explica las relaciones lineales entre la matriz de datos $\mathbf{y}_{ij}$ y los dos vectores de puntos a_i y b_j. Según este modelo, su media global μ es diferente de cero y sus vectores de marcas a_i y b_j contribuyen de forma aditiva a la variable respuesta y_{ij}.

$$\mathbf{y}_{ij} = \mu + \gamma\mathbf{a}_i + \delta\mathbf{b}_j \tag{4.59}$$

Según la Figura (4.12a), las rectas $\mathbf{L}_a$ y $\mathbf{L}_b$ son perpendiculares ($\measuredangle = 90°$). Obsérvese que los marcadores fila a_i son colineales a la recta L_a si y sólo si a_i puede expresarse como una combinación lineal de los vectores que definen la recta L_a. Es decir, si existen μ, γ y δ tales que, $a_i = \mu + \gamma v_a + \delta w_a$, donde v_a y w_a son los vectores que definen la recta L_a. Análogamente los marcadores columna b_j también son colineales a la recta $\mathbf{L}_b$ ya que $b_j \in \mathbf{L}_b$.

Este modelo es útil principalmente en el análisis de relaciones entre las variables de un conjunto de datos. Diagnosticar, mediante análisis biplot, si un modelo aditivo es apropiado para los datos es un proceso que consiste en comprobar si las observaciones a_i y los marcadores columna b_j se alinean con las líneas L_a y L_b. En caso afirmativo, esto indica que estos componentes tienen una contribución aditiva significativa en la variable respuesta y_{ij}.

Por otra parte, el modelo Tukey de la Figura (4.12b) expresado en la ecuación 4.59 explica las relaciones lineales entre la matriz de datos $\mathbf{y}_{ij}$ y los dos vectores de puntos a_i y b_j. Según este modelo, su media global μ es diferente de cero y sus vectores de marcas a_i y b_j contribuyen de forma aditiva a la variable respuesta y_{ij}. A diferencia de la ecuación 4.59, tiene un termino adicional en el que se mide el efecto interacción entre el factor fila γa_i y el factor columna δb_j. Este efecto interacción denominado $\lambda(\gamma a_i)(\delta b_j)$ es debido a factores externos. Lo que significa que el efecto de la variable a sobre la variable respuesta $\mathbf{y}$ está en función del valor de la variable b, y viceversa.

$$\mathbf{y}_{ij} = \mu + \gamma\mathbf{a}_i + \delta\mathbf{b}_j + \lambda(\gamma\mathbf{a}_i)(\delta\mathbf{b}_j) \tag{4.60}$$

Según la Figura (4.12b), las rectas $\mathbf{L}_a$ y $\mathbf{L}_b$ no son perpendiculares ($\measuredangle \neq 90°$).

Obsérvese que los marcadores fila a_i y b_j son colineales a las rectas L_a y L_b, respectivamente, si y sólo si a_i y b_i pueden expresarse como una combinación lineal de los vectores que definen las rectas L_a y L_b, respectivamente.

Este modelo suele utilizarse en situaciones en las que existe una interacción significativa entre dos o más variables predictoras. Asimismo, se pueden aprovechar los valores de los coeficientes γ, δ y λ a fin de cuantificar la magnitud de los efectos de las variables predictoras y su interacción sobre la variable respuesta.

Estos enfoques han alcanzado un desarrollo destacado en el campo de la agricultura, en particular con la formulación y evaluación de modelos en los que se estudian las interacciones genotipo-ambiente [556]. Otros autores en [557] han explorado los puntos fuertes del análisis biplot en el estudio de modelos de tres vías, especialmente con el análisis de la interacción de segundo orden en un único eje factorial. Investigaciones más especializados han empleado el análisis biplot también en el estudio de las interacciones en modelos de dos y tres vías [558].

4.8 Análisis discriminante

El análisis discriminante es una técnica de análisis multivariante utilizada en el estudio de las diferencias entre grupos. Este método estadístico paramétrico ha demostrado ser una potente herramienta analítica para identificar, a partir de un subconjunto de variables, las características que mejor discriminan a fin de construir modelos de clasificación más precisos. En términos generales, el análisis discriminante extrae combinaciones lineales, denominadas funciones discriminantes, de un subconjunto de variables independientes. Estas funciones ayudan a clasificar las nuevas observaciones en las categorías de clase definidas en la variable dependiente. Por lo tanto, si las variables predictoras son buenas para discriminar entre grupos, también es importante saber en qué se diferencian los grupos con respecto a estas características y la media de los datos del grupo es el mejor estimador de tendencia central útil para detectar si existen o no diferencias estadísticamente significativas entre grupos. El procedimiento descrito a continuación aplica para dos grupos cuyo número de casos o suma de pesos supere el número de grupos no vacíos.

4.8.1 Media de los datos

La media de los datos de una muestra de estudio es el mejor estimador de tendencia central de una población. La ecuación 4.61 a través de la variable, $\bar{X}_{ij}$ calcula la media de una variable i en un grupo j.

$$\bar{X}_{ij} = \frac{1}{n_j} \cdot \sum_{k=1}^{m_j} f_{jk} X_{ijk} \tag{4.61}$$

donde f_{jk} puede interpretarse como un factor que puede utilizarse para dar peso a cada caso k que forma parte del grupo j. X_{ijk} representa el valor de la variable i en el caso k del grupo j. m_j es el número de casos k en el grupo j, n_j es el número total de pesos de los casos en el grupo j.

Otro estadístico básico es el referido en la ecuación 4.62. Este calcula la media de la variable i en todos los grupos.

$$\bar{X}_i = \frac{1}{n} \cdot \sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} X_{ijk} \tag{4.62}$$

La idea que subyace a las ecuaciones 4.61 y 4.62 es que la media de los valores de la variable i puede ser calculada en función de cada uno de los grupos j o del número total de casos o ponderaciones. Debe notarse que el uso de ponderaciones permite diferenciar la importancia asignada a los casos.

4.8.2 Varianza

La variabilidad de los datos respecto a su media se expresa mediante las ecuaciones 4.63 y 4.64. La primera expresión calcula la varianza S_{ij}^2 de la variable i para cada uno de los grupos j a partir de la suma de los cuadrados de los residuos dividida por el número de observaciones en cada grupo j.

$$S_{ij}^2 = \frac{1}{(n_j - 1)} \cdot \sum_{k=1}^{m_j} f_{jk} X_{ijk}^2 - n_j \bar{X}_{ij}^2 \tag{4.63}$$

La ecuación 4.64 a diferencia de la anterior mide la varianza S_i^2 de la variable i.

$$S_i^2 = \frac{1}{(n - 1)} \cdot \sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} X_{ijk}^2 - n \bar{X}_i^2 \tag{4.64}$$

4.8.3 Productos cruzados intragrupos

A fin de realizar el análisis discriminante, las sumas de cuadrados intragrupos y matriz de productos cruzados **W** referida en la ecuación 4.65 se utiliza para calcular la matriz de covarianzas intragrupos. La expresión w_{il}, además de representar la matriz de productos cruzados de las variables, i y l, se utiliza para calcular la covarianza entre dichas variables.

$$w_{il} = \sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}X_{ljk} - \frac{1}{n_j}\cdot\sum_{j=1}^{g}\left(\sum_{k=1}^{m_j} f_{jk}X_{ijk}\right)\left(\sum_{k=1}^{m_j} f_{jk}X_{ljk}\right) \quad i,l = 1,\ldots,p \tag{4.65}$$

donde p es el número de variables.

Obsérvese que la primera parte de la expresión w_{il} representa la suma de los productos cruzados de las variables i y l para cada caso k en cada grupo j. La segunda parte de esta ecuación representa la suma de las medias ponderadas de las variables i y l para cada grupo j, dividida por el número de casos ponderados en el grupo j.

La matriz de productos cruzados es el punto de partida en el cálculo de las combinaciones lineales que producirán la máxima separación entre grupos. La matriz de covarianza intragrupo proporciona información sobre las relaciones entre las variables y cómo varían juntas en cada grupo.

4.8.4 Productos cruzados Totales

Las sumas totales de cuadrados y la matriz de productos cruzados **T** son una medida de la variabilidad total de los datos. La expresión 4.66 es una forma de calcular la matriz de productos cruzados **T** para cada par de variables (i y l). La matriz de productos cruzados es una medida de correlación entre dichas variables.

$$t_{il} = \sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}X_{ljk} - \frac{1}{n}\cdot\left(\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}\right)\left(\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ljk}\right) \tag{4.66}$$

Al desagregar la ecuación 4.66 en dos partes, la primera parte definida por la ex-

presión 4.67 es una medida de la variabilidad de las variables i y l entre los grupos.

$$\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}X_{ljk} \tag{4.67}$$

La segunda parte de la expresión 4.66 definida en la ecuación 4.68 es una medida de la variabilidad total en las variables i y l.

$$-\frac{1}{n}\cdot\left(\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}\right)\left(\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ljk}\right) \tag{4.68}$$

Por lo tanto, al restar la expresión 4.68 de la expresión 4.67, se obtiene una medida de la variabilidad entre las variables i y l que no es explicada por la variabilidad total de los datos. Se debe tener en cuenta que el calculo de la matriz $\mathbf{T}$ de productos cruzados es esencial para estimar las diferencias en las medias de las variables entre grupos y seleccionar las variables con mayor poder discriminante.

4.8.5 Covarianza intragrupos

La matriz de covarianza intragrupos, $\mathbf{C}v$, definida en la ecuación 4.69 es una medida de la variabilidad entre las variables de un mismo grupo. Esta matriz se obtiene a partir de la matriz de productos cruzados intragrupos, $\mathbf{W}$, y un factor de normalización $(n-g)$, donde n es la suma total de pesos y g es el número de grupos.

$$\mathbf{C}v = \frac{\mathbf{W}}{(n-g)}, \quad donde \quad n > g \tag{4.69}$$

La matriz de covarianza intragrupo $\mathbf{C}v$ es útil para el cálculo de la matriz de covarianza total y la matriz de correlación. Además de ser importante en el análisis de componentes principales (PCA), se utiliza para identificar las relaciones lineales entre las variables dentro de cada grupo.

4.8.6 Covarianza de grupos

La matriz de covarianza individual $\mathbf{C}v^{(j)}$ para cada grupo j se presenta en la ecuación 4.70. Esta expresión calcula la covarianza entre los valores de las variables i y l para los casos de estudio k en un determinado grupo j. Note que el término $\bar{X}_{ij}$ calcula la media de los valores de la variable i en el grupo j, en tanto que $\bar{X}_{lj}$ calcula la media

de los valores de la variable l en el grupo j.

$$Cv_{il}^{(j)} = \frac{1}{(n_j - 1)} \cdot \sum_{k=1}^{m_j} f_{jk} X_{ijk} X_{ljk} - \bar{X}_{ij} \bar{X}_{lj} n_j, \quad donde \quad i, l = 1, \ldots, p \tag{4.70}$$

En consecuencia, la matriz de covarianza individual $\mathbf{C}v^{(j)}$ para cada grupo j se calcula restando el producto de los promedios y multiplicando el resultado por un factor de normalización $(n_j - 1)^{-1}$. Este factor de normalización se utiliza para ajustar el tamaño de la muestra y asegurarse de que la matriz de covarianza refleje adecuadamente las relaciones entre variables dentro de cada grupo.

4.8.7 Correlación dentro de grupos

La matriz de correlación intragrupos $\mathbf{R}$ mide la relación lineal entre dos variables dentro de un grupo determinado. La matriz $\mathbf{R}$ se obtiene de la matriz de sumas de cuadrados y productos cruzados $\mathbf{W}$. De acuerdo con la ecuación 4.71, la correlación entre las variables i y l se obtiene dividiendo la entrada w_{il} de la matriz $\mathbf{W}$ por la raíz cuadrada del producto de las entradas w_{ii} y w_{ll} de la misma matriz.

$$r_{il} = \frac{w_{il}}{\sqrt{w_{ii} \cdot w_{ll}}} \tag{4.71}$$

donde $\quad i, l = \{1, \ldots, p\}$ y p es el número de variables.

De modo que la medida de correlación lineal entre las variables i y l dentro de un grupo determinado esta determinado en función de los siguientes casos:

$$r_{il} \leftarrow \begin{cases} r_{il} & \text{si } w_{ii} \cdot w_{ll} > 0 \\ 0 & \text{en caso contrario} \end{cases}$$

El valor de la correlación lineal r_{il} se calcula en la medida en que el producto de los valores de entrada w_{ii} y w_{ll} sea positivo. En caso contrario, se asigna un valor cero, lo que indica que no existe correlación lineal entre estas variables.

4.8.8 Covarianza total

La matriz $\mathbf{T}'$ de la ecuación 4.72 calcula la covarianza total de un conjunto de datos. Esta se obtiene dividiendo la matriz $\mathbf{T}$ por $n-1$, donde n es la suma total de los pesos.

$$\mathbf{T}' = \mathbf{T}(n-1)^{-1} \tag{4.72}$$

La matriz de covarianzas aporta información muy útil para comprender la estructura de los datos y la relación entre las variables.

4.8.9 Valores F y Λ

Medir la importancia relativa de cada variable con respecto a un subconjunto de variables predictoras es crucial con el análisis de la varianza. Los estadísticos univariantes F de Snedecor y Λ definidos en las ecuaciones 4.73 y 4.74 permiten analizar la importancia relativa de la variable I con respecto a las demás variables.

$$F_i = \frac{(t_{ii} - w_{ii})}{w_{ii}} \cdot \frac{n-g}{(g-1)} \tag{4.73}$$

con grados de libertad: n - g, y g - 1.

Cabe destacar que el estadístico F de Snedecor de la variable i se calcula como la relación entre la variabilidad entre grupos y la variabilidad dentro de los grupos para la variable i.

$$\Lambda_i = w_{ii} \cdot (t_{ii})^{-1} \tag{4.74}$$

con grados de libertad: 1, g - 1, y n - g.

Por otro lado, el estadístico Λ para la variable i se calcula como la relación entre la variabilidad dentro del grupo, para la variable i, y la variabilidad total.

Ambos estadísticos son influenciados por sus respectivos grados de libertad.

4.8.10 Funciones lineales discriminantes

Una vez identificadas las variables que producen las mejores diferencias entre grupos, el objetivo del análisis discriminante es obtener, a partir de las variables predictoras X_i una combinación lineal $y(x)$ que maximice la separación entre j grupos. La ecuación

4.75 presenta la función discriminante lineal $y(x)$.

$$y(x) = b_{ij}^{\circ} X_{ij} + \ldots + b_{qj}^{\circ} X_{qj} + a_j^{\circ} \tag{4.75}$$

Cabe señalar que las variables predictoras X_i miden las características $i = \{1, 2, \ldots, q\}$ que mejores diferencias producen entre un número de grupos $j = \{1, 2, \ldots, g\}$. El número de m funciones discriminantes y_i, $i = \{1, 2, \ldots, m\}$, tales que $m = \text{mín}(g - 1, q)$ son las que producen el mejor desempeño en la clasificación.

Los coeficientes b_{ij}° y las constantes a_j° para la construcción de las funciones lineales de Fisher se obtienen según las expresiones 4.76 y 4.77.

4.8.11 Funciones de clasificación

Las funciones lineales discriminantes de Fisher son funciones de clasificación que se construyen a partir de un subconjunto de q variables predictoras independientes. Los coeficientes de dichas funciones se obtienen a partir de la ecuación 4.76.

$$b_{ij}^{\circ} = (n - g) \sum_{l=1}^{q} \left(w_{il}^{*} \cdot \bar{X}_{lj} \right) \tag{4.76}$$

donde la variable independiente $i = \{1, 2, \ldots, q\}$ y el número de grupos $j = \{1, 2, \ldots, g\}$. El coeficiente b_{ij}° calculado con la ecuación 4.76 representa el peso de la variable i en la función lineal discriminante para el grupo j.

La ecuación se calcula sumando, para cada variable i en cada grupo j, el producto de la media de la variable l en el grupo j y el peso w_{il}^{*}. Este peso corresponde a la nueva matriz de sumas de cuadrados y productos cruzados intragrupo W calculados entre las variables i y l según ecuación 4.23. El término $(n - g)$ es prácticamente una constante de normalización utilizada para que los coeficientes tengan una escala común.

Las constantes de las funciones discriminantes lineales pueden calcularse mediante la ecuación 4.77.

$$a_j^{\circ} = \log \Pi_j - \sum_{i=1}^{q} \left(b_{ij}^{\circ} \cdot \bar{X}_{ij} \right) / 2 \tag{4.77}$$

donde $j \in \{1, 2, \ldots, q\}$ y la probabilidad a priori del grupo j se denota como Π_j. Los valores constantes a_j° de las funciones discriminantes lineales para cada grupo j maximizan la separación entre los grupos.

4.8.12 Clasificación

Existen muchos procedimientos válidos para realizar tareas de clasificación. Uno de los métodos más tradicionales se basa en la utilización de las funciones lineales de Fisher $y(x_j)$ obtenidas de la ecuación principal (4.75) y que se describe en la sección 4.8.11. La regla de clasificación descrita en la expresión 4.78 funciona cuando $j = 2$ grupos.

$$y(x_j) = \begin{cases} y(x_1), & \text{si } y(x_1) \geqslant y(x_2) \\ y(x_2), & \text{en caso contrario} \end{cases} \tag{4.78}$$

La clasificación de un nuevo caso k se realiza en función de la puntuación discriminante $y(x_j)$. El modelo asigna el caso k al grupo j que produce la puntuación discriminante más alta.

4.8.13 Bondad de ajuste del modelo

La capacidad de las funciones discriminantes lineales para ajustarse a la estructura de los datos puede valorarse con medidas de bondad de ajuste. Entre las más utilizadas se encuentran la varianza explicada por el modelo $\varpi(y_i)$, la correlación canónica de la función discriminante $C\,(y_i)$ y la prueba de funciones basada en el estadístico Lambda de Wilks Λ. Estas medidas, en su orden, se expresan en las ecuaciones 4.79, 4.80 y 4.81.

$$\varpi(y_i) = \lambda_i / \sum_{i=1}^{m} \lambda_i, \quad m = min\,(g-1, q) \tag{4.79}$$

donde $i = 1, 2, \ldots, m$ funciones discriminantes, g es el número de grupos y q el número de variables utilizadas en el modelo. Nótese que la varianza explicada por la función $\varpi(y_i)$ es una proporción de la varianza total explicable por todas las funciones extraíbles y_m. Por tanto, cuanto más se acerque el valor a uno, mejor explica el modelo la variabilidad total de los datos.

La correlación canónica calculada mediante la expresión 4.80 es una medida de la relación global entre grupos. Este valor permite evaluar la eficacia con la que la combinación lineal consigue separar los grupos. Cuanto mayor sea la correlación canónica $C\,(y_i)$, mayor será la capacidad de la función discriminante lineal para separar los

grupos.

$$C\left(y_i\right) = \sqrt{\lambda_i / \left(1 + \lambda_i\right)} \tag{4.80}$$

donde λ_i es el autovalor vinculado a cada función lineal discriminante y_i.

Por último, la prueba de funciones definida por la ecuación 4.81 mide en la función discriminante lineal la capacidad de diferenciar entre grupos.

$$\Lambda = 1 / \left(1 + \lambda_i\right) \tag{4.81}$$

Conviene señalar que cuanto más cercana a cero sea la lambda de Wilks (Λ), mejor será el rendimiento de la función discriminante.

Parte III

Materiales y métodos

5. Datos

La fiabilidad de los procesos de clasificación y predicción de la dirección de los precios a corto plazo depende de la calidad de los datos y de la información crítica utilizada. Este estudio utiliza puntos de inflexión del mercado definidos por un modelo algorítmico de seguimiento de tendencias, junto con datos de mercado (OHLCV) del tipo de cambio euro-dólar, para realizar la previsión. En la sección 5.1 se abordan los detalles de los datos de mercado utilizados en la medición de características discriminativas. En la sección 5.2 se presenta un resumen de los resultados financieros de la estrategia de negociación basada en los puntos de inflexión detectados en el mercado. En la sección 5.3 se hace referencia a las operaciones de mercado ejecutadas en los puntos de inflexión y las medidas utilizadas para su caracterización. En el apartado 5.3.1 se presentan los índices y variables utilizados para medir y caracterizar dichas operaciones. En el apartado 5.3.2 se introducen las expresiones matemáticas utilizadas en la medición de los indicadores y variables. Finalmente, en el apartado 5.3.3 se resume las principales estadísticas sobre estas mediciones.

5.1 Datos de mercado

El estudio utiliza dos conjuntos de datos con información de mercado sobre el tipo de cambio euro-dólar. Las cotizaciones del par euro-dólar se obtienen de la plataforma de negociación *Alpari International*. Estas cotizaciones corresponden a marcos temporales de 15 minutos y 1 semana. La primera muestra de análisis abarca 15 años, desde el 1 de enero de 2006 hasta el 31 de diciembre de 2020. La segunda muestra abarca 18 años, del 4 de enero de 2004 al 27 de junio de 2021. El primer conjunto de datos se compila en una matriz $\mathbf{Q}\,\langle 375000 \times 7\rangle$ que consta de 375000 observaciones y 7 métricas. El segundo conjunto se compila en una matriz $\mathbf{W}\,\langle 913 \times 7\rangle$ que consta de 913 observaciones y 7 métricas. Cada observación, en ambos conjuntos de datos, consta de número de registro (id), fecha y hora (dt), precio de apertura (op_i), precio máximo (hp_i), precio mínimo (lp_i), precio de cierre (cp_i) y volumen negociado en el mercado (Vo_i).

La Figura 5.1 muestra el historial de precios (a) y los rendimientos logarítmicos (b) del tipo de cambio euro-dólar. Las series de precios y rendimientos corresponden a marcos temporales de 15 minutos y 1 semana. Estos valores proceden de las matrices de datos **Q** y **W**. La Figura (5.1a) muestra que el tipo de cambio, aunque presenta un movimiento de rango, tiende a mantener un régimen de mercado bajista durante el periodo de análisis. El efecto tendencia de las series históricas de precios queda suprimido por los rendimientos logarítmicos. Las series de retornos se obtienen a partir de la siguiente expresión $\vartheta_i = ln\,(cp_i/cp_{i-1})$, donde ϑ_i es el retorno entre precios calculado para cada marco temporal y cp_i es el precio de cierre asociado al instante de tiempo i.

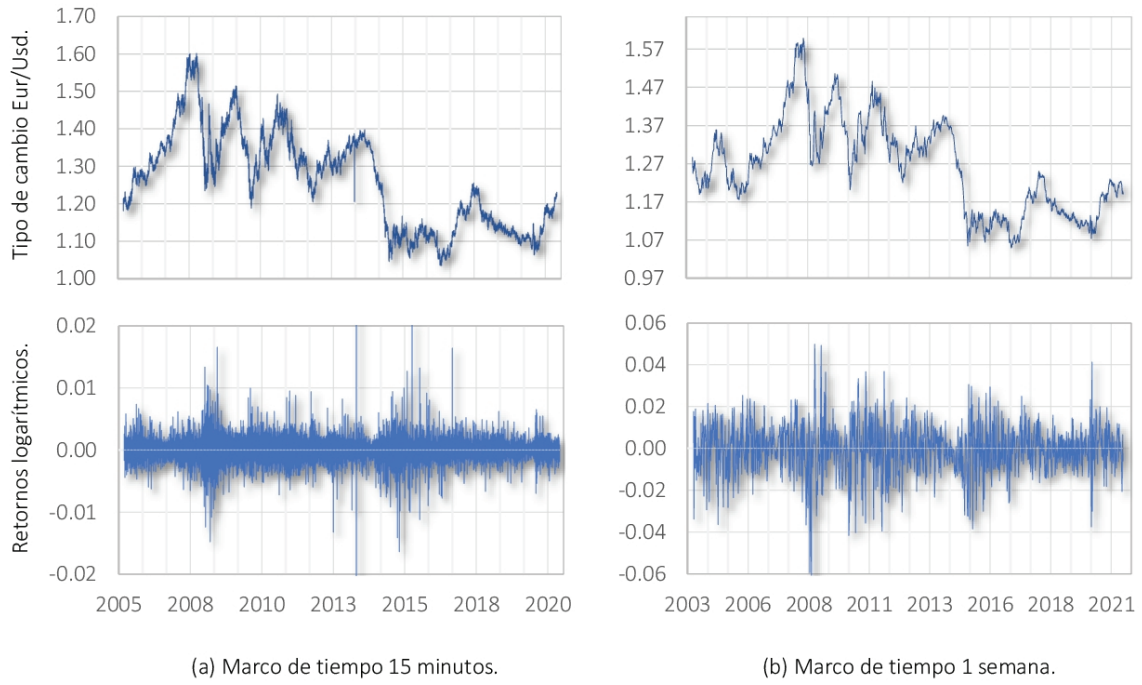


Figura 5.1: Datos de mercado del tipo de cambio euro-dólar.
(**a**) Precios de cierre históricos. Datos medidos en un marco de tiempo de 15 minutos, muestreados en un horizonte de tiempo de 15 años, va desde el 1 de enero del 2006 hasta el 31 de diciembre del 2020. (**b**) retornos logarítmicos. Datos semanales muestreados en un periodo de 18 años, desde enero 4 del 2004 hasta junio 27 del 2021.

La media $\bar{\vartheta}$ de estos rendimientos es mayor en los retornos de 15 minutos que en los de 1 semana $\left(\bar{\vartheta}_{15m} = 8.42 \times 10^{-08}, \bar{\vartheta}_{1w} = -6.56 \times 10^{-05}\right)$. El análisis de la Figura (5.1b) revela que la volatilidad, en ambos marcos de tiempo, es mayor a finales del 2008 y principios del 2009. Posteriormente, la volatilidad en el marco de tiempo

de 15 minutos se hace fuerte a principios del 2014 y finales del 2015, mientras que en el marco de tiempo semanal se hace fuerte a mediados del 2010 y principios del 2020. Los coeficientes de asimetría (γ) de estos rendimientos son negativos. El sesgo negativo es más fuerte en los retornos de 15 minutos ($\gamma_{15m} = -1.02, \gamma_{1w} = -0.30$). El coeficiente de curtosis (κ) es mayor en la distribución de retornos de 15-minutos ($\kappa_{15m} = 7570.83, \kappa_{1w} = 1.60$). Así, las distribuciones de rendimientos logarítmicos, del tipo de cambio euro-dólar para ambos marcos de tiempo, son asimétricas y leptocúrticas. Además, esto sugiere preparar los datos para viabilizar el uso de técnicas de predicción multivariantes.

Por otro lado, el análisis de los rendimientos del tipo de cambio euro-dólar revela resultados significativos. Los rendimientos de 15 minutos muestran una media superior a los rendimientos de 1 semana, lo que indica un potencial de ganancias a corto plazo. La presencia de periodos de mayor volatilidad en ambos marcos temporales pone de relieve la necesidad de tener en cuenta la volatilidad a la hora de tomar decisiones de inversión, especialmente en los marcos temporales más pequeños. Además, los elevados valores de curtosis en la distribución de los rendimientos de 15 minutos sugieren capitalizar y cubrirse frente a las volatilidades, especialmente en los valores extremos, lo que puede ser útil a la hora de tomar decisiones financieras.

5.2 Puntos de inflexión

El uso de información crítica es relevante para mejorar la eficacia de la toma de decisiones de inversión, especialmente cuando se utiliza en la construcción de modelos de clasificación interpretables y precisos. Este estudio se basa en una colección de operaciones intradía exitosas generadas por un modelo de negociación algorítmica. Los datos se obtienen a partir de una estrategia de seguimiento de tendencias que utiliza un enfoque de negociación por impulso (Momentum Trading Strategy). El modelo captura los puntos de inflexión del mercado, donde comienza la dirección del movimiento de los precios, y los denomina momentos de entrada (`ET`). Estos puntos de referencia son útiles para identificar las variables que mejor predicen el inicio de la dirección del movimiento futuro del tipo de cambio euro-dólar. Además, las operaciones realizadas en los puntos de entrada `ET`, a favor de la tendencia, producen márgenes de beneficios positivos con una exposición al riesgo reducida.

Las Tablas, 5.1 y 5.2, muestran el rendimiento de la estrategia de negociación. Cabe señalar que las posiciones con rendimiento financiero negativo se excluyen de la

preparación y el análisis de los datos, ya que afectan negativamente a los resultados del estudio. La apertura y cierre de cada contrato se realiza sobre los precios de cierre (*cp*) en un marco temporal de 15 minutos. Este marco temporal se selecciona en función de los mejores resultados financieros y la menor exposición al riesgo obtenidos con la estrategia de negociación. Cada contrato se ejecuta en el instante `ET` en que se produce el cambio de tendencia. Los costos de comisión de $45/1000000 USD se cobran después de cada transacción. La comisión cobrada por mantener una posición abierta durante la noche se aplica según el tipo de posición, Swap Largo = -$4.96 y Swap Corto = -$0.96. El deslizamiento no se aplica a las operaciones porque los retrasos en la ejecución afectan negativamente a la definición de los momentos de estudio `ET` y a la definición de las microtendencias.

Tabla 5.1: Perfil de rendimiento de la estrategia

[a] Especificaciones			
Símbolo:	EUR/USD	Moneda base:	USD
Barras:	370903	Tamaño de contrato:	100000
Deposito inicial ($):	1000	Apalancamiento:	1:100
Valor del Pip ($):	0.10	Swaps ($):	-267.30
Saldo final ($):	14162.85	Volumen negociado en lotes:	0.010
Posición cerrada G/P ($):	15800.68	Comisiones ($):	-2370.53
Nivel de margen mínimo (%):	6570.47	Tasa libre de riesgo (%):	0.00
Riesgo de mercado (%):	0.063		
Beneficio Neto Total ($):	13162.85	Beneficio bruto ($):	32496.94
Pérdida bruta ($):	-19334.09	Riesgo por operación (%):	0.14
Factor de beneficio:	1.68	Beneficio esperado:	0.63
Factor de recuperación	38.39	Ratio de Sharpe:	0.13
AHPR:	1.0003	GHPR:	1.0003
Probabilidad de éxito:	0.72	Probabilidad de fracaso:	0.28
Operaciones totales:	20894	Operaciones cortas ganadoras:	10502
Operaciones largas ganadoras:	10392	Operaciones perdedoras:	5780
Operaciones nulas:	88	Operaciones ganadoras:	15026
ROI:	13.16	Operación mayor beneficio ($):	74.36
Operación mayor pérdida ($):	-39.79	Pérdida media / operación ($):	-3.12
Periodo de análisis (Años):	15.21	Beneficio medio / operación ($):	2.21
Operaciones rentables (%):	70.32	Máx de victorias consecutivas:	33
Máx pérdidas consecutivas:	21	Máx pérdida consecutiva($):	-342.01
Operaciones con pérdidas (%):	29.68	Máx beneficio consecutivo ($):	278.76
Ratio (Win / Loss):	2.37:1	Media victorias consecutivas:	5
Media pérdidas consecutivas:	2		

[a]Barras = Número de velas analizadas en un marco de tiempo de 15 minutos. Posición cerrada G/P = Ganancias/Pérdidas totales de todas las operaciones cerradas (Comisiones y swaps no descontados). Comisiones ($) = Comisión cobrada tras la apertura y cierre de una operación ($ 45/1000000 USD por operación). Swaps ($) = Comisión cobrada cuando la posición se mantiene abierta durante la noche (Swap largo = -$4.96, Swap corto = -$0.96). Riesgo de mercado (%) = Porcentaje de dinero en riesgo de pérdida sobre el saldo disponible.

Tabla 5.2: Perfil de rendimiento de la estrategia

[a] Especificaciones			
Esperanza matemática:	$ 0.76	Ratio (Beneficio / Pérdida):	0.71:1
Ventaja operativa:	0.41	Prueba de corridas Z-Score:	-47.68
Desviación Estándar Beneficios:	$ 4.10	Límite de confianza Z-Score:	0.997
Coeficiente de variación:	5.43	LR Correlación:	0.99
LR Error Estándar:	397.52		
Reducción Absoluta de Saldo:	$ 34.06	Reducción Máxima de Saldo:	$ 342.84
Reducción Relativa de Saldo:	2.42 %		

[a]Beneficio neto total ($) = Resultados financieros de operaciones cerradas (Beneficios + comisiones + swaps). Beneficio bruto ($) = Suma de beneficios de operaciones rentables (Comisiones y swaps descontados). Pérdida bruta ($) = Suma de perdidas de operaciones perdedoras (Comisiones y swaps descontados). Factor de beneficio = Relación entre el beneficio bruto y la pérdida bruta. Beneficio esperado = Retorno medio de una operación. Factor de recuperación = Relación entre el beneficio neto total y la reducción máxima. Ratio de Sharpe = (Rentabilidad - Tasa libre de riesgo) / Desviación estándar de los retornos. AHPR = Media aritmética de las variaciones de capital por operación cerrada. GHPR = Media geométrica de las variaciones de capital por operación cerrada. Riesgo por operación (%) = Cantidad de capital en riesgo por operación. Operaciones ganadoras = Número de operaciones ganadoras (Comisiones y swaps sin descontar). Operaciones perdedoras = Número de operaciones perdedoras (Comisiones y swaps sin descontar). ROI = Beneficio neto total dividido por el depósito inicial. Operación mayor beneficio ($) = Beneficio máximo de todas las operaciones rentables (Comisiones y swaps descontados). Operación mayor pérdida ($) = Pérdida máxima de todas las operaciones perdedoras (Comisiones y swaps descontados). Beneficio medio por operación ($) = Beneficio bruto dividido por el número de operaciones rentables. Pérdida media por operación($) = Pérdida bruta dividida por el número de operaciones no rentables. Operaciones rentables (%) = Número de operaciones rentables (Comisiones y swaps descontados) dividido por el número de operaciones totales. Máx de victorias consecutivas (Recuento) = Serie más larga de operaciones ganadoras. Máx pérdidas consecutivas (Recuento) = Serie más larga de operaciones perdedoras. Operaciones con pérdidas (%) = Número de operaciones con pérdidas (comisiones y swaps descontados) dividido por el número de operaciones totales. Máx Beneficio consecutivo ($) = Beneficio máximo acumulado de una serie de operaciones rentables. Máx Pérdida consecutiva ($) = Máxima pérdida acumulada de una serie de operaciones perdedoras. Ratio (Win / Loss) = Operaciones con beneficios (%) dividido por Operaciones con pérdidas (%) Media victorias consecutivas = Promedio de operaciones ganadoras en series rentables. Media pérdidas consecutivas = Promedio de operaciones perdedoras en series perdedoras. Ratio (Beneficio / Pérdida) = Beneficio medio por operación ($) dividido por la pérdida media por operación ($). Ventaja comercial = Operaciones con beneficios (%) - Operaciones con pérdidas (%). Correlación LR = Correlación entre la línea de balance de la cuenta y la línea de regresión. Z-Score = Detección de periodos consecutivos de beneficios/pérdidas. LR Error Estándar = Error estándar de la desviación entre la línea de saldo de la cuenta y la línea de regresión. Reducción (Drawdown) Absoluta de Saldo ($) = La mayor caída de saldo por debajo del depósito inicial. Reducción máxima del saldo ($) = La mayor caída del saldo en términos monetarios. Reducción relativa del saldo (%) = La mayor caída del saldo en términos porcentuales.

Por consiguiente, las operaciones tienen una probabilidad de ganar del 72% y generan un beneficio neto de 13162.85 USD durante el periodo de la muestra, de enero de 2006 a diciembre de 2020. Las operaciones ganadoras representan el 75.34% de las 20894 operaciones realizadas, con una duración media de 7 horas. Del total de operaciones, el 75.34% obtuvieron rendimientos brutos en los dos primeros tercios del periodo

de tenencia. El Ratio Win/Loss, con un valor de 2.37, muestra que el número de posiciones ganadoras es 2 veces superior al número de posiciones perdedoras. Además, el Profit Factor de 1.68 indica que se ganan 1.68 USD por cada 1 USD que se pierde.

Estadísticamente, la prueba de rachas y la puntuación Z-score del sistema de negociación determinan la existencia de dependencia entre los resultados de las operaciones. Así, una puntuación Z de -47.68 confirma una dependencia positiva entre las operaciones dentro de un límite de confianza del 99.7%. Esto indica que el número de rachas es menor de lo esperado según la función de probabilidad normal, lo que implica que las operaciones ganadoras generan más ganadores y las perdedoras más perdedores.

Los resultados obtenidos con la estrategia comercial sugieren que existen oportunidades para explotar estas relaciones de dependencia y mejorar el rendimiento del sistema, pero el uso de medidas de gestión del capital y señales confirmatorias queda fuera del alcance de este documento.

En general, los puntos de inflexión del mercado pueden ofrecer oportunidades de inversión o indicar cambios en la oferta y la demanda de un activo negociado. La detección y capitalización de los puntos de inflexión `ET` en la negociación del tipo de cambio euro-dólar proporciona información crítica para lograr rendimientos con una exposición reducida al riesgo.

Los sólidos resultados de la estrategia de negociación demuestran la importancia de estos puntos de inflexión y respaldan su uso en modelos de clasificación para predecir la evolución futura de los precios.

5.3 Operaciones de mercado

Las operaciones ganadoras intradía se obtienen a partir de un modelo de seguimiento de tendencias que aprovecha los puntos de inflexión del mercado (`ET`). El conjunto de datos $\mathbf{A}\,\langle 14770 \times 18\rangle$ constituye la unidad de partida para la detección de micro-tendencias en favor de los movimientos tendenciales de largo plazo que se van a pronosticar. Esta colección de datos consta de 14770 contratos con rendimientos positivos y 18 métricas que miden el comportamiento del mercado antes y después de su apertura en el instante `ET`. Las unidades de análisis constan de 7380 operaciones largas y 7390 operaciones cortas intradía. El periodo de análisis es de 15 años y va desde enero de 2006 hasta diciembre del 2020.

5.3.1 Índices y variables

Las operaciones de compra y venta se miden mediante 18 indicadores. Los quince primeros índices, mencionados en la Tabla 5.3, miden el comportamiento del mercado 18 periodos previos al momento ET (en un marco temporal de 15 minutos). Estos indicadores se calculan con un número de periodos definido por el mejor rendimiento alcanzado con el modelo de clasificación.

Tabla 5.3: Variables e Índices

Nemotécnico [a]	**Denominación**	**Medida**
Acción del Precio:		
s.V.PT	Dispersión de retornos precios de cierre	Numérica continua
x.V*PT	Rendimiento medio	Numérica continua
Sp.CP.PT	Pendiente precios de cierre	Numérica continua
Sp.Vr*CP.PT	Pendiente variaciones precios de cierre	Numérica continua
Sp.Vr.HL.PT	Pendiente variaciones precios alto y bajo	Numérica continua
Sp.Vr.CO.PT	Pendiente variaciones precios cierre y apertura	Numérica continua
s.Vr.RSI.PT	Variación RSI	Numérica continua
Sp.Vr.RSI.PT	Pendiente variaciones RSI	Numérica continua
**Volumen:		
s.Vr.Vo.PT	Variación del volumen	Numérica continua
x.Vr.Vo.PT	Variación media del volumen	Numérica continua
Sp.Vr.Vo.PT	Pendiente de las variaciones de volumen	Numérica continua
Sp.Vo.PT	Pendiente del volumen	Numérica continua
Divergencias:		
Dv.CP-RSI.PT	Divergencia Precio de Cierre y RSI	Numérica continua
Dv.Vo-CP.PT	Divergencia Precio de Cierre y Volumen	Numérica continua
Dv.Vo-RSI.PT	Divergencia RSI y Volumen	Numérica continua
Tras mantener abierta una operación:		
* GPL_$	Beneficio o Pérdida bruta	Numérica continua
* RP% M[1]	Dinero en riesgo por operación	Numérica continua
* Sp.XP-EP	Pendiente precios de apertura y cierre	Numérica continua
Tendencia a largo plazo (W1):		
M.Sp.P.ET	Pendiente de mercado antes de *ET*	Numérica continua
* LTT	Pendiente de mercado luego de *ET*	Numérica continua
* M.Sp	Pendiente de mercado incluye *ET*	Numérica continua
$\text{RSI}(9p)_{Smooth}$	Índice de fuerza relativa	Numérica continua
Tendencia	0 = Alcista, 1 = Bajista	Variable categórica

[a] Todas las variables se miden en marcos de tiempo de 15 minutos, excepto (W1) que se miden en marcos de tiempo de 1 semana. **Volumen = Número de veces que cambia el precio en cada periodo en un marco de tiempo de 15 minutos. (W1) = Formulas utilizadas para definir los movimientos de tendencia a largo plazo. ET = Punto de inflexión del mercado utilizado como momento de estudio. * = Mediciones posteriores al momento de estudio ET (no participan en el proceso de selección de características).

Estos indicadores se agrupan en tres áreas de análisis, precio, volumen negociado y divergencia entre las variables de las dimensiones precio y volumen. Además, estos

indicadores se utilizan como variables candidatas en el proceso de selección de características. Los mejores índices se utilizan para predecir el comienzo del movimiento direccional del tipo de cambio euro-dólar.

Los indicadores (GPL_$, RP% M[1], y Sp.XP-EP) miden el rendimiento obtenido por la operación una vez finalizado el tiempo de mantenimiento del contrato. Estos indicadores se utilizan en la detección de estructuras subyacentes y en la identificación de los determinantes de las diferencias entre los movimientos tendenciales.

Los indicadores (M.Sp.P.ET, LTT y, M.Sp) ayudan a integrar las micro-tendencias de las operaciones ganadoras intradía (**A**) en un marco temporal más amplio (**W**). En consecuencia, las instancias que están a favor de los movimientos de tendencia de largo plazo se extraen y forman la población de estudio **O** (véase la sección 6.3 sobre preparación de datos).

Por último, la variable categórica *Tendencia* etiqueta en el momento `ET` la dirección de la tendencia del tipo de cambio. Las etiquetas de clase utilizadas son *Movimiento tendencial alcista* y *Movimiento tendencial bajista*. Los movimientos de precios lateralizados no se consideran en el presente trabajo.

5.3.2 Cálculo de índices y variables

En la Tabla 5.4 se muestran las fórmulas introducidas en este documento para calcular las variables e índices a utilizar en el análisis técnico de los precios del tipo de cambio.

La tabla está dividida en diferentes secciones que se centran en aspectos específicos del análisis técnico. Estas secciones incluyen: *Acción del precio*, que mide la variación y dispersión del precio del activo negociado; *Volumen*, que mide el volumen del activo negociado o el número de variaciones del precio del tipo de cambio negociado en un intervalo de 15 minutos; *Divergencias*, que examina las discrepancias en el movimiento de dos variables a lo largo del tiempo; *Después de mantener abierta una operación*, que analiza los beneficios, las pérdidas y los riesgos asociados tras mantener abierta una operación; y *Tendencia a largo plazo (W1)*, que se centra en el análisis de tendencias a largo plazo superponiendo las microtendencias de las posiciones ganadoras con los movimientos de tendencia a largo plazo.

Estos indicadores miden el comportamiento de los precios de mercado antes, durante y después de los puntos de inflexión, en los que también se producen operaciones intradía. El objetivo de utilizar esta información crítica es detectar y medir las características discriminantes que mejor distinguen entre tendencias alcistas y bajistas.

Tabla 5.4: Fórmulas de cálculo[1]

Nemotécnico		Fórmulas	
Acción del Precio:			
s.V.PT	$\sqrt{\frac{\sum(x_i-\bar{x})^2}{(n-1)}}$,	$x_i = \frac{cp_i - cp_{i-1}}{cp_{i-1}}$	$i = 1, ..., n;\ n = 18$
x.V*PT	$\frac{\sum_{i=1}^{n} x_i}{n}$,	$x_i = ln\frac{cp_i}{cp_{i-1}}$,	$i = 1, ..., n$
Sp.CP.PT	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = cp_i$,	$x_i = \{x_1, ..., x_n\}$
Sp.Vr*CP.PT	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = ln\frac{cp_i}{cp_{i-1}}$,	$x_i = \{x_1, ..., x_n\}$
Sp.Vr.HL.PT	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{(hp_i - lp_i)}{(hp_{i-1} - lp_{i-1})} - 1$,	$x_i = \{x_1, ..., x_n\}$
Sp.Vr.CO.PT	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{(cp_i - op_i)}{(cp_{i-1} - op_{i-1})} - 1$,	$x_i = \{x_1, ..., x_n\}$
s.Vr.RSI.PT	$\sqrt{\frac{\sum(x_i-\bar{x})^2}{(n-1)}}$,	$x_i = \frac{RSI(9p)_i}{RSI(9p)_{i-1}} - 1$,	$i = 1, ..., n$
Sp.Vr.RSI.PT	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{RSI(9p)_i}{RSI(9p)_{i-1}} - 1$,	$x_i = \{x_1, ..., x_n\}$
Volumen:			
s.Vr.Vo.PT	$\sqrt{\frac{\sum(x_i-\bar{x})^2}{(n-1)}}$,	$x_i = \frac{Vo_i - Vo_{i-1}}{Vo_{i-1}}$	$i = 1, ..., n$
x.Vr.Vo.PT	$\frac{\sum_{i=1}^{n} x_i}{n}$,	$x_i = \frac{Vo_i}{Vo_{i-1}} - 1$,	$i = 1, ..., n$
Sp.Vr.Vo.PT	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{Vo_i - Vo_{i-1}}{Vo_{i-1}}$,	$x_i = \{x_1, ..., x_n\}$
Sp.Vo.PT	$\frac{\sum(x_i-\bar{x})\left(Vo_i-\bar{Vo}\right)}{\sum(x_i-\bar{x})^2}$,	$x_i = \{x_1, ..., x_n\}$	
Divergencias:			
Dv.CP-RSI.PT	$\frac{\sum(cp_i-\bar{cp})(y_i-\bar{y})}{\sqrt{\sum(cp_i-\bar{cp})^2\sum(y_i-\bar{y})^2}}$,	$y_i = RSI(9p)_i$	
Dv.Vo-CP.PT	$\frac{\sum\left(Vo_i-\bar{Vo}\right)(cp_i-\bar{cp})}{\sqrt{\sum\left(Vo_i-\bar{Vo}\right)^2\sum(cp_i-\bar{cp})^2}}$		
Dv.Vo-RSI.PT	$\frac{\sum\left(Vo_i-\bar{Vo}\right)(y_i-\bar{y})}{\sqrt{\sum\left(Vo_i-\bar{Vo}\right)^2\sum(y_i-\bar{y})^2}}$,	$y_i = RSI(9p)_i$	

Tabla 5.4 Fórmulas de cálculo

Nemotécnico	Fórmulas		
Tras mantener abierta una operación:			
⋆ GPL_$	$Vo \cdot cs\,(xp - ep)$,	$Vo \cdot cs\,(ep - xp)$	
⋆ RP% M[1]	$\frac{Vo \cdot cs(lp-ep)}{sb}$,	$\frac{Vo \cdot cs(ep-hp)}{sb}$	
⋆ Sp.XP-EP	Slope $(\Delta p/\Delta t)$,	Long $= \frac{xp-ep}{xt-et}$,	Short $= \frac{ep-xp}{xt-et}$
Tendencia a largo plazo (W1):			
M.Sp.P.ET	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \{cp_{i-96}, ..., cp_{i=et}\}$,	$x_i = \{x_1, ..., x_n\}$
⋆ LTT	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \{cp_{i=et}, ..., cp_{i=96}\}$,	$x_i = \{x_1, ..., x_n\}$
⋆ M.Sp.P.ET	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \{cp_{i-96}, ..., cp_{i=et}, ..., cp_{i=96}\}$,	$x_i = \{x_1, ..., x_n\}$
$\text{RSI}(9p)_S$	$100 - \left(100/\left(1 + \bar{U}/\bar{D}\right)\right)$		
Tendencias:	0 = Alcista $(GPL\_\$ > 0, \quad Slope\,(\Delta p/\Delta t) > 0, \quad Long\ Trade)$ 1 = Bajista $(GPL\_\$ > 0, \quad Slope\,(\Delta p/\Delta t) < 0, \quad Short\ Trade)$		

Entre otra información relevante, la Tabla 5.4, contiene las relaciones matemáticas y estadísticas que se miden entre las distintas variables e indicadores calculados. Algunas fórmulas utilizan la media y la desviación típica para calcular la variación y dispersión de los datos. Estas relaciones pueden revelar patrones y tendencias en los datos que no son evidentes a simple vista.Además, la tabla proporciona fórmulas para calcular las pendientes de las líneas de tendencia, lo que permite analizar y prever la dirección y la fuerza de las tendencias de los precios y el volumen de veces que cambia el precio

[1]Las variables se miden en marcos de tiempo de 15 minutos, excepto (W1) que se miden en marcos de tiempo de 1 semana. (W1) = Formulas utilizadas para definir los movimientos de tendencia de largo plazo. ⋆ = Mediciones posteriores al punto de inflexión ET (No participan en el proceso de selección de características). ν_i = Volumen registrado en el periodo i. cp_i = Precio de cierre del periodo i. hp_i = Precio máximo del periodo i. lp_i = Precio mínimo del periodo i. op_i = Precio de apertura en el periodo i. $RSI(9p)_i$ = Índice de fuerza relativa suavizado del período i calculado a partir de 9 períodos históricos. $x_i = \{x_1, ..., x_n\}$ Serie de tiempo de n observaciones. et = Punto de inflexión o momento de estudio. Vo = Volumen. cs = Tamaño de contrato 100000 unidades de la divisa base (Euro). sb = Saldo inicial de la cuenta en dólares americanos. xp = Precio de salida. ep = Precio de entrada. xt = Momento de cierre. $\bar{U}$ = Media de las diferencias de los precios al alza. $\bar{D}$ = Media de las diferencias de los precios a la baja.

del activo negociado.

En términos generales, las fórmulas utilizadas en el cálculo de indicadores para el análisis técnico del tipo de cambio euro-dólar se centran en evaluar el comportamiento de los precios e identificar patrones para aprovechar las anomalías del mercado, mejorando así la toma de decisiones de inversión.

5.3.3 Estadísticas descriptivas de índices

La tabla 5.5 resume las estadísticas descriptivas de la matriz de datos multivariantes **A**, incluidos los índices relacionados con el análisis técnico y el cálculo de indicadores en un marco temporal de 15 minutos. Estas medidas proporcionan información importante sobre el comportamiento de los precios y el volumen de veces que cambia el precio del tipo de cambio negociado en un periodo determinado.

Tabla 5.5: Estadísticas descriptivas de los índices.

[a] Medida	Media	*DS	Mediana	Mínimo	Máximo	Asimetría	Curtosis
Acción del Precio:							
s.V.PT	0.00051011	0.00034042	0.00042776	0.000044579	0.00439635	2.3730	11.7879
x.V*PT	-0.0000002	0.00019849	0.00000143	-0.00209697	0.00164601	0.0050	5.15348
Sp.CP.PT	-0.0000008	0.00025970	0.00000289	-0.00233251	0.00192982	-0.028	6.19084
Sp.Vr*CP.PT	-0.0000004	0.00003325	-0.0000002	-0.00040831	0.00033698	-0.055	6.48955
Sp.Vr.HL.PT	0.00630912	0.03272341	0.00538741	-0.58322522	0.72498519	0.6652	38.0406
Sp.Vr.CO.PT	0.00800454	0.46597394	0.01403601	-6.46007878	8.77988644	0.0522	31.9152
s.Vr.RSI.PT	0.16739490	0.07751317	0.15280451	0.031895221	1.33360205	2.6943	18.7865
Sp.Vr.RSI.PT	-0.0005816	0.00824177	-0.0004332	-0.04381362	0.08385499	0.2604	3.14648
*Volumen:							
s.Vr.Vo.PT	0.47597319	1.24018626	0.35617189	0.047287739	69.8530494	40.416	1937.19
x.Vr.Vo.PT	0.10736958	0.31392351	0.07387271	-0.12090153	16.9855561	35.890	1639.09
Sp.Vr.Vo.PT	0.00231944	0.05945863	0.00230310	-3.32210627	2.84634278	-1.024	1542.24
Sp.Vo.PT	10.7184210	41.9964534	5.06553148	-295.573787	357.643962	0.5038	6.78630
Divergencias:							
Dv.CP-RSI.PT	0.85151545	0.20628068	0.93089873	-0.65551962	0.99928244	-2.903	9.93598
Dv.Vo-CP.PT	-0.01375776	0.51348876	-0.0276017	-0.97874454	0.9763092	0.0360	-1.1499
Dv.Vo-RSI.PT	-0.01580351	0.49230421	-0.0234383	-0.97742819	0.96606119	0.0382	-1.1572
Tras mantener abierta una operación:							
GPL_$	2.174395396	2.51882690	1.47000000	0.010000000	44.3000000	4.4641	34.7811
RP% M[1]	-0.00079658	0.01564524	-0.0017000	-0.22980000	0.17700000	0.8083	14.9109
Sp.XP-EP	0.000009683	0.03751680	-0.0000302	-0.54240000	1.03680000	1.1255	57.2277

[a]*DS = Desviación Estándar. *Volumen = Número de cambios de precios en cada período en un marco de tiempo de 15 minutos.

La media de los datos en los momentos `ET` (en los que se produce el cambio de dirección de la tendencia precedente) es mayor para las variables Sp.Vo.PT, GPL_$, Dv.CP-RSI.PT y s.Vr.RSI.PT. Además, en este grupo de indicadores, los dos primeros índices (Sp.Vo.PT y GPL_$) muestran un efecto volátil, mientras que los dos últimos tienen una de las desviaciones estándar más bajas. La media y la mediana de los datos son diferentes. Los estadísticos de asimetría y curtosis están lejos de cero. Los indicadores s.Vr.Vo.PT, x.Vr.Vo.PT y Sp.Vr.Vo.PT evidencian valores desproporcionados en las colas de las distribuciones. Dado que ninguna de estas medidas se aproxima a la media, se trata de una prueba contundente de que el conjunto de datos, para algunos indicadores, presenta un comportamiento asimétrico y leptocúrtico.

En términos generales, La matriz de datos **A**, formada por operaciones ganadoras intradía, muestra un comportamiento asimétrico y leptocúrtico en ciertos indicadores, especialmente en los instantes de tiempo (`ET`) en los que se realizan las operaciones de compra o venta. Este desafío, inherente a los datos, requiere una preparación previa antes de construir modelos de clasificación, especialmente cuando se utilizan técnicas paramétricas que se basan en supuestos estadísticos. Además, es crucial disponer de características interpretables que ayuden a explicar y predecir la tendencia, teniendo en cuenta la acción de los precios, para construir modelos precisos e interpretables que mejoren la toma de decisiones de inversión y reduzcan el riesgo.

6. Metodología propuesta

6.1 Introducción

Este capítulo presenta una propuesta metodológica que aborda las oportunidades y limitaciones de los actuales modelos de clasificación. El objetivo es proporcionar una valiosa herramienta metodológica para construir modelos de clasificación interpretables, parsimoniosos y precisos. Este enfoque pretende mejorar la toma de decisiones de inversión mediante la predicción de la tendencia a corto plazo del tipo de cambio euro-dólar sin comprometer la legibilidad de las interpretaciones. En este contexto, la metodología propuesta se basa en el uso sistemático de un conjunto de procedimientos bien definidos. La sección 6.3 trata la preparación de datos, la sección 6.4 describe el proceso de selección de características discriminantes e independientes, la sección 6.5 trata la detección y el análisis de estructuras subyacentes, la sección 6.6 resume la construcción de modelos de clasificación mediante análisis discriminante y la sección 6.7 trata la evaluación del rendimiento del modelo de clasificación con datos fuera de muestra.

6.2 Generalidades

La metodología propuesta permite predecir el inicio del movimiento futuro del tipo de cambio euro-dólar. El diseño metodológico consta de cinco etapas y se fundamenta en cuatro colecciones de datos $\boldsymbol{\Xi}\,\langle i \times j \rangle$ de tamaño i filas por j atributos. Donde i es el número de momentos de estudio `ET` en los que se produce un cambio de dirección de la tendencia precedente, y j son las variables numéricas observables que miden el comportamiento de los precios de cierre antes y después de esos momentos `ET`.

La Figura 6.1 presenta un procedimiento de cinco fases para construir un modelo de clasificación mediante análisis discriminante. Cada fase (representada en amarillo) se divide en acciones (azul) y sus resultados correspondientes (blanco). Estas fases son:

preparación de datos, selección de características, análisis de estructuras, construcción del modelo de clasificación mediante análisis discriminante y evaluación del rendimiento del modelo de clasificación.

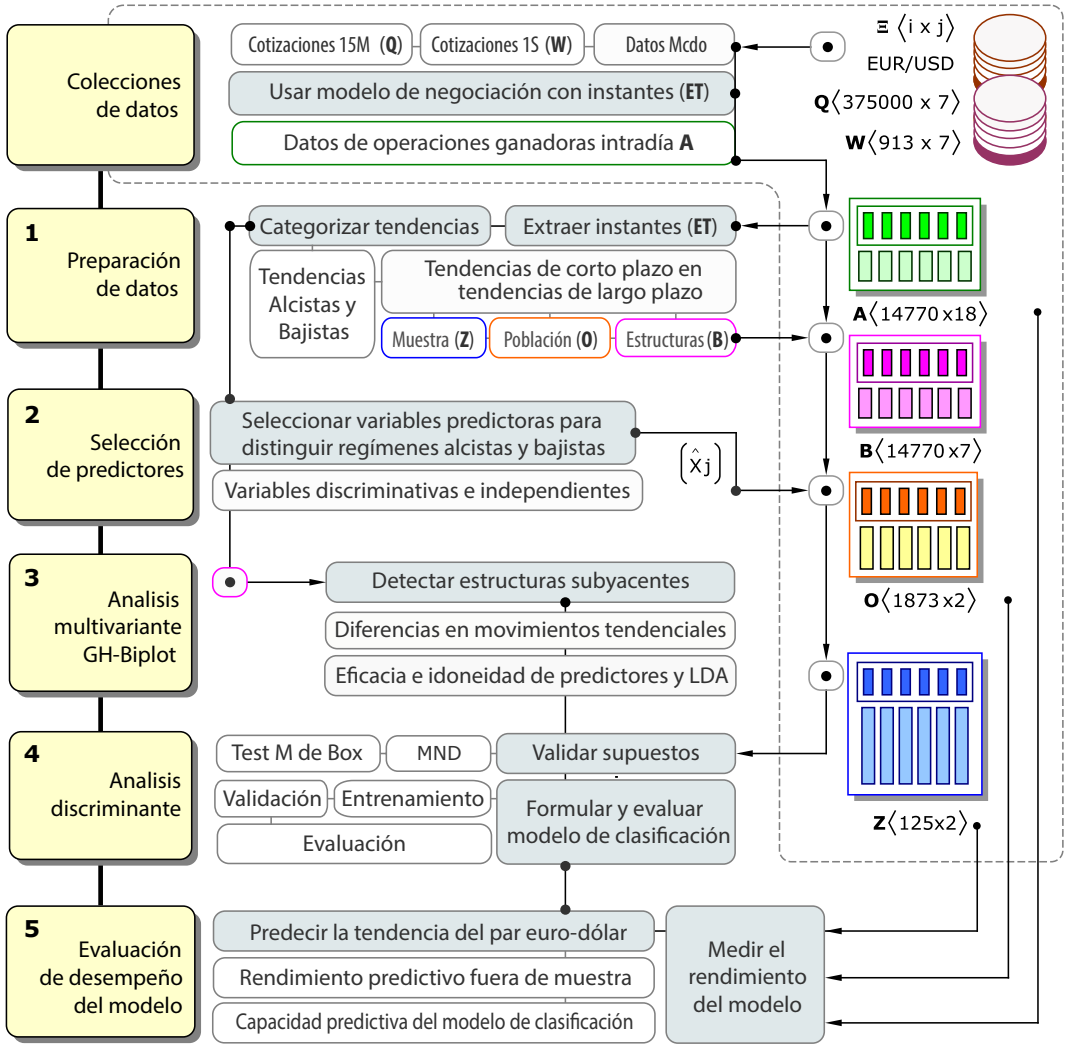


Figura 6.1: Enfoque para la construcción de modelos de clasificación. Metodología propuesta para predecir, en el corto plazo, el comienzo del movimiento futuro del tipo de cambio euro-dólar. Esta metodología se resume en los procedimientos de: (1) Preparación de datos. (2) Selección de características con poder discriminante. (3) Análisis multidimensional de las diferencias entre movimientos de tendencia. (4) Análisis discriminante y (5) Evaluación del poder predictivo del modelo de clasificación.

En primer lugar, se preparan las colecciones de datos para hacerlas analizables y relevantes según los propósitos establecidos en cada fase propuesta. En segundo lugar, se seleccionan los predictores independientes con mayor poder discriminatorio basándose

en su mérito estadístico. Tercero, se detectan y analizan las estructuras subyacentes para identificar los determinantes de las diferencias entre movimientos tendenciales. La eficacia de los predictores seleccionados y la idoneidad del análisis discriminante se validan con los resultados obtenidos de esta fase. En cuarto lugar, estas soluciones parciales se integran en la construcción de un modelo de clasificación interpretable y preciso. En esta fase se clasifica la dirección del movimiento del tipo de cambio mediante análisis discriminante. Por último, se evalúa el poder predictivo del modelo de clasificación utilizando datos fuera de muestra. La eficacia de la metodología propuesta y la generalización del modelo se evalúan con datos en diferentes condiciones de mercado y horizontes temporales.

En términos generales, el enfoque propuesto tiene aplicaciones prácticas en escenarios reales, especialmente en la negociación intradía de instrumentos financieros como los tipos de cambio. En este contexto, es crucial tomar decisiones informadas en condiciones de incertidumbre. Esta metodología permite a los analistas aprovechar los puntos de inflexión del mercado al utilizar la preparación de datos, la selección de características y la detección de estructuras para desarrollar modelos de clasificación más eficaces. Estos modelos de clasificación facilitan la interpretación del comportamiento de los precios y proporcionan predicciones precisas de la tendencia a corto plazo del activo negociado. Además, la incorporación de estas predicciones a los marcos de negociación intradía mejora la eficacia de los juicios de valor, la precisión de la toma de decisiones de inversión y reduce la exposición al riesgo.

6.3 Preparación de datos

La preparación de datos es esencial en los procesos de selección de características, detección de estructuras y formulación de modelos. Los datos que se utilizaran en estas fases se organizan en colecciones $\boldsymbol{\Xi}\langle i \times j \rangle$, donde i representa los instantes de estudio $\texttt{ET}$, y j son las variables utilizadas en los análisis.

La Figura 6.2 muestra las cuatro fases simplificadas del proceso de preparación de datos. En primer lugar, los puntos de inflexión ($\texttt{ET}$) se etiquetan como micro-tendencias alcistas y bajistas para identificar los predictores más discriminativos mediante la selección de características. Segundo, la población de estudio ($\mathbf{O}$) se extrae teniendo en cuenta que las micro-tendencias ($\texttt{ET}$) coinciden con las tendencias a largo

plazo. Los casos que cumplen este requisito se utilizan para predecir, a corto plazo, la dirección del movimiento del tipo de cambio euro-dólar. Tercero, se elige el conjunto óptimo de variables observables (**B**) para la detección de estructuras. La eficacia discriminativa de los predictores y el análisis discriminante se verifican con las variables representadas. Finalmente, una muestra de datos (**Z**), con distribución normal multivariante (*MND*), se extrae de la población de estudio. Este muestreo ayuda a superar la naturaleza asimétrica y leptocúrtica de las observaciones y a cumplir los supuestos estadísticos del análisis discriminante.

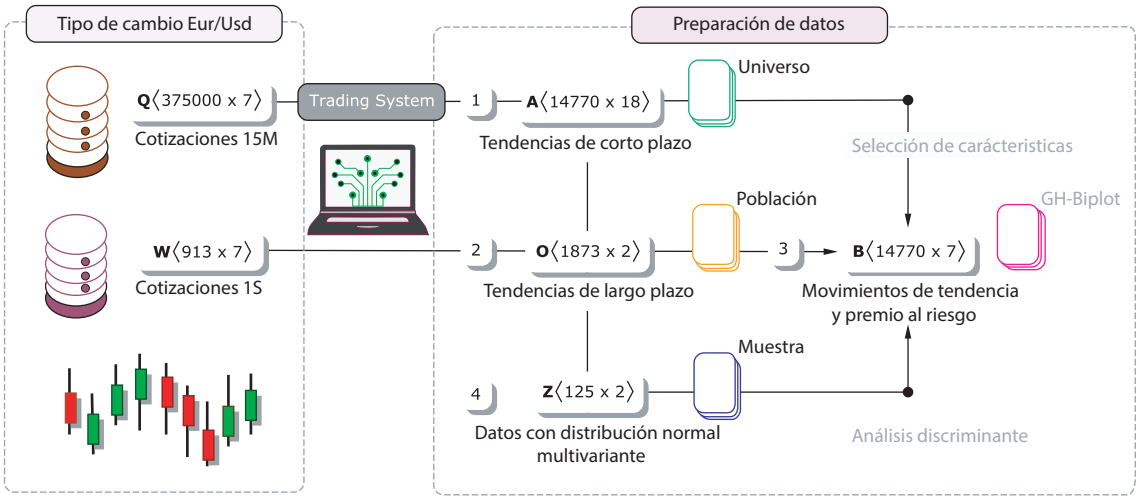


Figura 6.2: Estructura del proceso de preparación de datos.
Proceso de preparación para predecir a corto plazo el comienzo de la dirección del movimiento futuro del tipo de cambio euro-dólar. (1) Definición de micro-tendencias. (2) Validación de puntos de inflexión (ET) en función de los movimientos de tendencia de largo plazo. (3) Preparación de datos para la detección de estructuras subyacentes. (4) Extracción de micro-tendencias con distribución normal multivariante.

Teniendo en cuenta la metodología propuesta, a continuación se aborda cada una de las fases de preparación y disposición de las distintas colecciones de datos para garantizar el cumplimiento de los objetivos propuestos.

6.3.1 Preparación de datos: Selección características.

La preparación de datos para la selección de características es crucial para identificar y destacar atributos relevantes, discriminantes e independientes en el análisis posterior. En este proceso de preparación, las unidades de estudio (ET) se etiquetan según su comportamiento alcista o bajista, y se realizan mediciones utilizando las variables presentadas en la Tabla 5.4. A continuación, se describen los procedimientos asociados.

A. Categorización de puntos de inflexión

En esta fase, las unidades de estudio (ET) se preparan para la discretización en micro-tendencias alcistas y bajistas. Las etiquetas de clase asignadas se utilizan como factor de agrupación en todas las matrices de datos preparadas. Las instancias de la matriz de datos multivariante $\mathbf{A}\ \langle 14770 \times 18 \rangle$, que consisten en operaciones ganadoras intradía de una estrategia de negociación de seguimiento de tendencias, se clasifican en dos grupos: *micro-tendencias alcistas* y *micro-tendencias bajistas*. La información utilizada para asignar estas etiquetas de clase se muestra en la Tabla 6.1.

Tabla 6.1: Categorización de micro-tendencias.

Posición	[a]**Pendiente**	**Tipo**	**Categoría**
Ganadora	Slope $(\Delta p/\Delta t) > 0$	Posición larga	Micro-tendencia alcista
Ganadora	Slope $(\Delta p/\Delta t) < 0$	Posición corta	Micro-tendencia bajista

[a]Pendiente de la operación: Sp.XP-EP = Slope $(\Delta p/\Delta t)$. Las fórmulas para calcular pendientes, para posiciones largas y cortas, se mencionan en la Tabla 5.4.

Este proceso de discretización asigna una etiqueta de clase (micro-tendencia alcista o micro-tendencia bajista) en función del beneficio bruto declarado por la operación realizada en ese instante (ET). Esta asignación de etiqueta de clase se confirma mediante una variable categórica real (tipo de operación). Así, una unidad de estudio (ET) se etiqueta como *micro-tendencia alcista* si la operación larga (*buy*) generada en ese instante produce un beneficio bruto positivo (GPL_$ > 0) con pendiente positiva (Slope $(\Delta p/\Delta t) > 0$). Una unidad de estudio (ET) se etiqueta como *micro-tendencia bajista* cuando la operación corta (*sell*) generada en ese instante produce un beneficio bruto positivo (GPL_$ > 0) con pendiente de mercado negativa (Slope $(\Delta p/\Delta t) < 0$).

B. Medición de variables observables

La precisión de la predicción se fundamenta en el uso de información crítica sobre el comportamiento histórico de los precios. La medición de los puntos de inflexión (ET) y de las características observables en los últimos (n) periodos refleja la importancia del análisis histórico en la predicción de la tendencia futura. En este contexto, la definición del número de periodos históricos en la medición de las variables observables de la Tabla 5.4 es decisiva en la precisión de los modelos de clasificación. Estas mediciones basadas en una secuencia de periodos históricos son las que se utilizan posteriormente en la selección de características.

La tabla 6.2 muestra información sobre las contribuciones y variaciones analizadas en relación con la medición de las características utilizando un número diferente de periodos de tiempo. Los pesos asignados a cada periodo aumentan a medida que se consideran más periodos históricos en el estudio, en relación con el horizonte de análisis intradiario de 96 sesiones diarias de 15 minutos. Esto indica que, a medida que aumenta el número de periodos considerados en el análisis histórico de precios, se da más peso a los datos más antiguos en el proceso de medición de características.

Tabla 6.2: Contribuciones y pesos en diferentes periodos observados.

[a]**Tipo de datos**	**Periodos (n)**	**Pesos (w)**	**Contribuciones**	
			Mediciones (c)	**Variaciones ($c_{\Delta(v)}$)**
Datos más recientes	3	0.031	0.969	0.979
Datos recientes	6	0.063	0.937	0.948
	9	0.094	0.906	0.917
	12	0.125	0.875	0.885
Datos intermedios	15	0.156	0.844	0.854
	18	**0.188**	**0.812**	**0.823**
	21	0.219	0.781	0.792
	24	0.250	0.750	0.760
Datos antiguos intermedios	27	0.281	0.719	0.729
	30	0.313	0.687	0.698
	36	0.375	0.625	0.635
	45	0.469	0.531	0.542
Datos antiguos	72	0.750	0.250	0.260
	96	1.000	0.000	0.010

[a]Un día se compone de 96 periodos o intervalos de 15 minutos. El peso de la medición (w) se calcula dividiendo el número de periodos utilizados en la medición por el número total de periodos disponibles, w = (N° Períodos) / 96. La contribución de la medición (c) se obtiene restando el peso de la medición (w) de la contribución total de la característica medida, $c = (1 - w)$. Esto significa que los valores más antiguos de la característica medida también se reflejan en los (n) períodos más recientes. Las variaciones (Δ_v) entre las mediciones de las características observadas se obtienen dividiendo la diferencia entre dos mediciones consecutivas por el valor de la medición anterior, $\Delta_v = (v_i - v_{i-1})/v_{i-1}$. El número de variaciones ($n_{\Delta(v)}$) de una variable observable medida en 96 períodos se obtiene restando 1 al número de períodos, $n_{\Delta(v)} = (n - 1)$. El peso de las variaciones ($w_{\Delta(v)}$) se calcula como el cociente entre el número de variaciones ($n_{\Delta(v)}$) y el número total de variaciones disponibles, $w_{\Delta(v)} = \big(n_{\Delta(v)}/95\big)$. La contribución de las variaciones ($c_{\Delta(v)}$) se obtiene restando el peso de las variaciones ($w_{\Delta(v)}$) de uno, $c_{\Delta(v)} = \big(1 - w_{\Delta(v)}\big)$. Esto significa que la volatilidad actual de las mediciones viene determinada a su vez por la acción actual de los precios y no por el comportamiento histórico de sus variaciones más antiguas.

Por otro lado, respecto a la columna *Contribuciones*, a medida que aumenta el número de periodos históricos analizados, se observa una tendencia decreciente en las contribuciones. Esto indica que los datos más recientes tienen un impacto más significativo en las contribuciones globales. Este patrón sugiere que los movimientos

y variaciones recientes son más relevantes en el análisis de precios para la toma de decisiones informadas.

Al medir las características con datos de mercado de los últimos 18 periodos según la *Ley de Pareto*, se observa que el 18.8% de los datos más actualizados captura el 81.2% del comportamiento reciente de la acción del precio. Estos resultados ponen de relieve la importancia de considerar un horizonte temporal adecuado para medir la acción del precio a la hora de predecir la tendencia futura y tomar decisiones de inversión mejor informadas. Además, la tabla revela que el uso de datos más antiguos, con 72 y 96 periodos históricos, aunque abarcan un horizonte temporal más largo, muestra contribuciones bajas, del 25% y el 0% respectivamente. Esto sugiere que el uso de datos más antiguos tiene poca relevancia para el análisis de los precios actuales y puede no reflejar las condiciones actuales del mercado de forma representativa.

En términos generales, al utilizar el 18.8% de los datos más actuales para medir las características del momento previo a los puntos de inflexión (`ET`), se captura teóricamente el 81.2% del comportamiento dinámico de la acción del precio que precede al cambio de tendencia. Sin embargo, al medir las características utilizando el 75% de las observaciones intradía disponibles, solo se captura el 25% de la información actual sobre la acción de los precios. Estas consideraciones revelan tres alternativas clave basadas en la medición de características con 12, 18 y 36 periodos históricos. Cada opción captura información relevante que precede al cambio de dirección de la tendencia en los puntos de inflexión (`ET`). Aunque se exploraron 14 alternativas en la medición de características, este estudio introduce la medición de características discriminativas basada en los 18 periodos más recientes. Esto se debe al alto rendimiento obtenido en términos de diferenciación entre movimientos tendenciales y a la mayor precisión alcanzada por el modelo de clasificación.

6.3.2 Preparación de datos: Muestra de estudio

En este apartado se abordan dos aspectos esenciales que garantizan la calidad y representatividad de los datos en la construcción de modelos de clasificación. En primer lugar, la población de estudio se define a partir de los movimientos tendenciales a largo plazo, incluyendo las micro-tendencias y los puntos de inflexión (`ET`). Estos componentes son integrados para establecer la base de la población analizada. En segundo lugar, se extrae de la población de estudio una muestra de observaciones con dis-

tribución normal multivariante para superar la naturaleza asimétrica y leptocúrtica de los datos. Así, el enfoque propuesto garantiza la fiabilidad de los resultados aplicando el análisis discriminante como método de clasificación en la construcción de modelos interpretables y precisos.

A. Definición de la población de estudio

En esta fase, se identifica y compila la población de estudio en la matriz de datos **O**. Estos datos se utilizan para predecir el inicio de la dirección futura del tipo de cambio euro-dólar a corto plazo. La matriz $\mathbf{O}\,\langle 1873 \times 2\rangle$ contiene $\mathbf{i}$ momentos de análisis ET, en los que ocurre el cambios de dirección de la tendencia precedente, y en los que se miden $\mathbf{p}$ variables predictoras seleccionadas $\hat{\mathbf{x}}_j$.

Las micro-tendencias generadas en los instantes (ET) de la matriz **A** se validan con los movimientos tendenciales a largo plazo según los precios de mercado de la matriz de datos $\mathbf{W}\,\langle 913 \times 7\rangle$. Además, la autenticidad de las micro-tendencias generadas en los intervalos de 15 minutos se valida mediante un análisis de 96 periodos semanales antes y después del momento (ET). Los resultados muestran que 96 periodos son suficientes para confirmar la tendencia dominante en un marco temporal semanal.

En este contexto, el instante (ET) es válido cuando el cambio de dirección de la tendencia precedente en las sesiones de negociación de 15 minutos coincide con la tendencia dominante en las sesiones de negociación semanales; en caso contrario, se descarta la observación. Es importante señalar que los cambios de tendencia a corto plazo forman parte de los movimientos de tendencia a largo plazo, pero no determinan su formación. Además, este estudio se limita a predecir el inicio de la dirección de la tendencia en sesiones de negociación de 15 minutos y, por tanto, no significa necesariamente que dicha predicción sea determinante del comienzo del movimiento de tendencia de largo plazo.

La Figura 6.3 muestra los resultados del proceso de selección de unidades de análisis (ET) descrito en la Tabla 6.3. Los regímenes de mercado alcistas y bajistas a largo plazo, con los momentos de estudio (ET), incluyen movimientos de precios con continuidad y cambio de dirección de la tendencia precedente. Es relevante señalar que los momentos de estudio (ET), en los que se producen cambios en la dirección de la tendencia a corto plazo en sesiones de negociación de 15 minutos, están vinculados a la tendencia dominante del mercado en sesiones de negociación semanales.

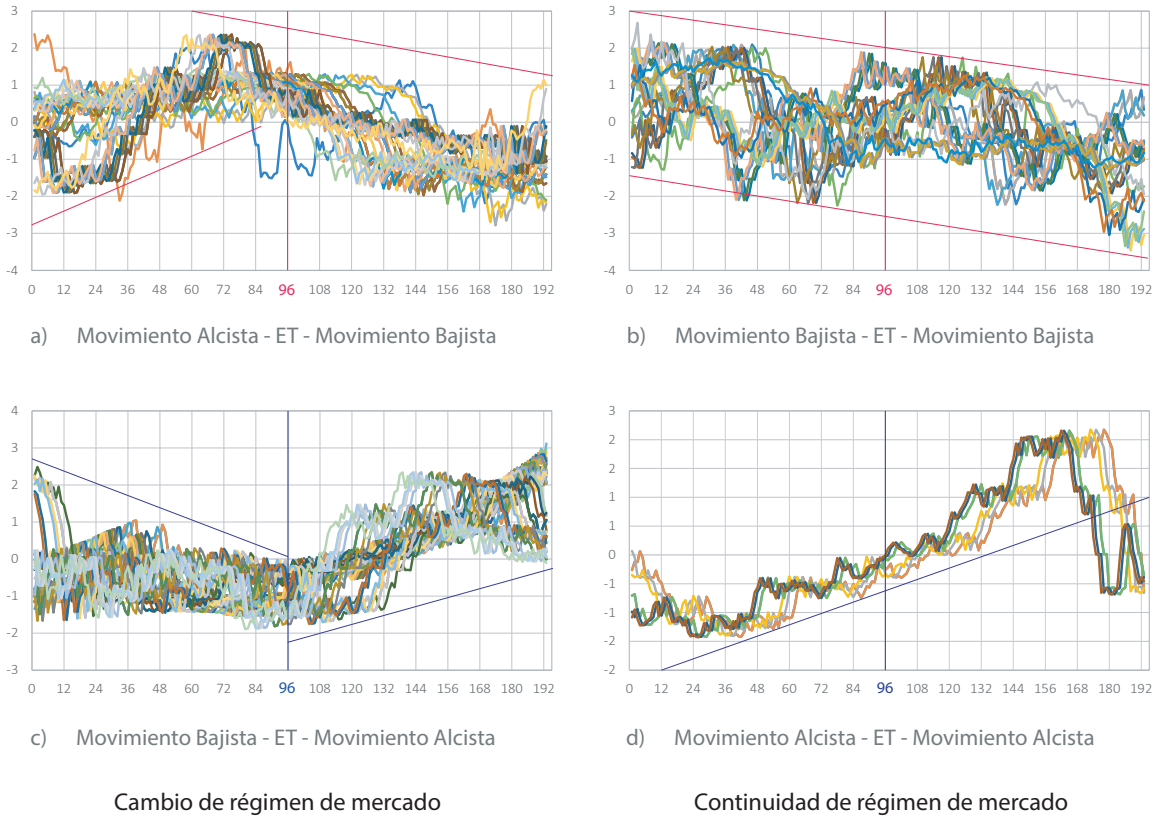


Figura 6.3: Movimientos de mercado del tipo de cambio euro-dólar.
Movimientos de mercado según precios de cierre semanales con integración de datos intradía en momentos (ET). Datos normalizados por filas. (**a**) Cambio de régimen, movimiento direccional alcista a bajista. (**b**) Continuidad de régimen, Movimiento direccional bajista. (**c**) Cambio de régimen, movimiento direccional bajista a alcista. (**d**) Continuidad de régimen, movimiento direccional alcista.

La Tabla 6.3 detalla en su primera columna el proceso de formación de los movimientos tendenciales a largo plazo para la selección de las unidades de estudio (ET). La superposición de puntos de inflexión (ET) y micro-tendencias en los movimientos tendenciales a largo plazo es una referencia que confirma la continuidad (FT) o el cambio de dirección (CT) de la tendencia precedente en las sesiones de negociación semanales. Así, dos formaciones tendenciales se reconocen, *movimientos alcistas* y *movimientos bajistas*, con dos formas de origen, *cambio de tendencia* o *continuidad en la tendencia*. En la segunda y tercer columna se definen las premisas utilizadas para identificar el tipo de régimen previo y posterior al momento (ET). La cuarta columna confirma la pendiente dominante del mercado. Finalmente, la última columna muestra el número de observaciones obtenidas para el estudio.

Tabla 6.3: Micro-tendencias integradas en tendencias de largo plazo

Régimen	[a]**Pendiente línea de regresión precios de cierre Eur/Usd**			**Población**
de Mercado	**Hasta** ET	**Luego de** ET	**Incluye** ET	**O**
$[-96p, ET, 96p]$	$[-96p, ..., ET]$	$[ET, ..., 96p]$	$[-96p, ..., ET, ..., 96p]$	1873
Bajista-CT-Alcista	Slope RL (p/t) < 0	Slope RL (p/t) > 0	Slope RL (p/t) > 0	414
Alcista-FT-Alcista	Slope RL (p/t) > 0	Slope RL (p/t) > 0	Slope RL (p/t) > 0	228
Alcista-CT-Bajista	Slope RL (p/t) > 0	Slope RL (p/t) < 0	Slope RL (p/t) < 0	556
Bajista-FT-Bajista	Slope RL (p/t) < 0	Slope RL (p/t) < 0	Slope RL (p/t) < 0	675

[a]ET = Instante de Entrada y punto de inflexión (ET). CT = Cambio de tendencia, FT = Siguiendo la tendencia. Hasta ET = M.Sp.P.ET = Pendiente del mercado antes del punto de inflexión (ET). Luego de ET = *LTT = Línea de tendencia a largo plazo. Incluye ET = *M.Sp = Pendiente del mercado. Las ecuaciones para su computo están referenciadas en la Tabla 5.4.

Obsérvese que este procedimiento ayuda a identificar y validar los momentos de análisis (ET) en los que las micro-tendencias se mantienen o cambian en favor de la dirección de la tendencia a largo plazo. Además, el desconocimiento de las causas que originan los movimientos de tendencia requiere la identificación de diversos momentos de estudio (ET) que reflejan la relación entre el corto y largo plazo.

En términos generales, la población de estudio **O** consta de 1873 micro-tendencias en favor de la dirección de la tendencia dominante, pero que no son determinantes de la formación inicial de la tendencia de largo plazo. Por consiguiente, la población de estudio está compuesta por 414 y 228 micro-tendencias alcistas. Las primeras con cambio de tendencia a favor de la tendencia alcista dominante, y las segundas respaldan la tendencia alcista de largo plazo. Además, hay 556 micro-tendencias bajistas con cambio de dirección de la tendencia precedente en función de la tendencia bajista dominante, y 675 micro-tendencias bajistas que respaldan la continuidad del movimiento bajista de largo plazo.

Por otra parte, la población de estudio **O** se compone principalmente de dos categorías: alcistas y bajistas, con un mayor número de tendencias bajistas (1231) que alcistas (642). Es esencial tener en cuenta este desequilibrio a la hora de construir modelos de clasificación precisos e interpretables. El muestreo balanceado ofrece una solución útil para equilibrar la representación de las clases sin afectar la selección de características relevantes e interpretables. Este valioso procedimiento ayuda a mejorar la precisión de la clasificación durante el entrenamiento del modelo. Además, la validación del rendimiento del modelo con datos independientes y el uso de métricas

adecuadas son esenciales para garantizar la precisión y la generalización del modelo para cada clase.

B. Definición de la muestra de estudio

En esta fase se prepara la muestra de estudio $\mathbf{Z}$, que se utilizará para formular, validar y evaluar el rendimiento del modelo de clasificación. Para superar el efecto asimétrico y leptocúrtico de la población de estudio $\mathbf{O}\,\langle 1873 \times 2\rangle$ y garantizar el cumplimiento de los supuestos que validan la aplicación del análisis discriminante, se extrae una muestra pequeña con distribución normal multivariante. Este procedimiento garantiza el cumplimiento de los supuestos de normalidad multivariante e igualdad de las matrices de varianza y covarianza dentro de los grupos, tanto para los predictores como para los grupos. A continuación se describe el procedimiento utilizado mediante la formulación de una función objetivo para extraer una muestra de datos multivariantes con distribución normal.

1. Consideraciones generales. La muestra de datos se extrae teniendo en cuenta la estructura de la población $\mathbf{O}$. En este caso, la población corresponde a los puntos de inflexión (`ET`) del tipo de cambio euro-dólar detectados y validados en marcos temporales de 15 minutos y una semana, respectivamente. Cabe destacar que esta población sigue una distribución asimétrica y leptocúrtica, lo que implica una distribución no simétrica y una mayor concentración en torno a determinados valores.

Esta concentración de información recurrente en valores concretos tiene un efecto acumulativo a lo largo del tiempo, alterando la distribución normal inicial y generando una distribución asimétrica y leptocúrtica. Esta característica es relevante en el análisis de datos con métodos de clasificación paramétricos, ya que puede afectar la fiabilidad de los resultados e interpretaciones.

Para determinar y extraer las instancias que preservan la distribución normal multivariante inicial de la población de datos $\mathbf{O}$, se formula una función objetivo. El propósito de esta función es seleccionar las instancias x_i que mejor se ajustan a una distribución normal multivariante, permitiendo realizar inferencias e interpretaciones relevantes en el análisis.

- Normalidad univariante

La búsqueda exhaustiva de casos con distribución normal univariante se realiza mediante la prueba de *Shapiro-Wilk*. El ajuste del modelo base, con distribución normal,

a los datos analizados se evalúa mediante el estadístico de prueba SW. Este valor se calcula mediante la ecuación 6.1 y aplica solo a las colas derechas para tamaños de muestra entre 3 y 50 observaciones. Así, un valor positivo alto de SW indica una clara proximidad de los datos a una distribución normal.

$$SW = \left(\sum_{i=1}^{n} a_i x_i \right)^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 , \tag{6.1}$$

donde, x_i son los valores aleatorios ordenados de la muestra de análisis, a_i son constantes generadas a partir de las covarianzas, varianzas y medias muestrales de una muestra distribuida normalmente de tamaño n. En consecuencia, el estadístico SW es el cociente entre la suma de los productos de los coeficientes y los valores observados, al cuadrado, y la suma de los cuadrados de las diferencias entre los valores observados y la media muestral.

El $p-valor$ de la prueba de normalidad univariante, denominada SW_{Test}, se calcula mediante distintos procedimientos, uno de ellos basado en tablas. Utilizando una relación lineal, el $p - valor$ correspondiente se estima entre los límites p_1 y p_2, que corresponden a los valores SW_1 y SW_2 del estadístico SW. La ecuación 6.2 expresa esta relación lineal.

$$SW_{Test} = p_1 + \frac{(SW - SW_1)}{(SW_2 - SW_1)} (p_2 - p_1) , \tag{6.2}$$

donde, SW_{Test} es el estadístico de prueba o $p - valor$ de la prueba de normalidad univariante. Los valores p_1 y p_2 representan los límites inferiores y superiores respectivamente del rango de valores p que se corresponden con los valores SW_1 y SW_2 del estadístico SW. De este modo, el $p-valor$ proporciona una medida de la significación estadística de la prueba de $Shapiro-Wilk$, determinando si la muestra de datos sigue o no una distribución normal. Esto es especialmente relevante, sobre todo en este caso, en el que se busca preservar la normalidad univariante y multivariante de la muestra extraída de la población **O**.

- Normalidad multivariante

El estadístico de prueba de normalidad multivariante *Henze-Zirkler* mide la distancia funcional no negativa entre dos funciones de distribución. En caso de que los datos sigan una distribución normal multivariante, este estadístico sigue una distribución aproximadamente log-normal. El cálculo de este estadístico comienza calculando la media, la

varianza y un parámetro de suavidad. A continuación, se normalizan logarítmicamente la media y la varianza y, por último, se calcula el $p-valor$. La ecuación 6.3 presenta el cálculo del estadístico de prueba Henze-Zirkler.

$$\text{HZ} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\alpha^2}{2}M_{ij}} - 2\xi^{-\frac{p}{2}}\sum_{i=1}^{n} e^{-\frac{\alpha^2}{2\xi}M_i} + n\left(1+2\alpha^2\right)^{-\frac{p}{2}},$$

donde:

$$\begin{aligned}
\xi &= \left(1+\alpha^2\right) \\
\alpha &= \frac{1}{\sqrt{2}}\left(\frac{n\left(2p+1\right)}{4}\right)^{\frac{1}{p+4}} \\
M_{ij} &= \left(x_i - x_j\right)' S^{-1}\left(x_i - x_j\right) \\
M_i &= m_{ii} = \left(x_i - \bar{x}\right)' S^{-1}\left(x_i - \bar{x}\right) \\
p &= \text{número de variables}
\end{aligned} \tag{6.3}$$

En la ecuación 6.3, M_i es la distancia de $Mahalanobis$ al cuadrado entre la i-ésima observación y su respectivo centroide, M_{ij} es la distancia de $Mahalanobis$ entre la i-ésima y la j-ésima observación. Cuando los datos analizados siguen una distribución normal multivariante, el estadístico de prueba (HZ) se distribuye aproximadamente con una distribución log-normal con media μ y varianza σ^2, como se muestra en las ecuaciones 6.4 y 6.5

$$\mu = 1 - \frac{b^{-\frac{p}{2}}\left(1 + p\alpha^{\frac{2}{b}} + \left(p\left(p+2\right)\alpha^4\right)\right)}{2b^2} \tag{6.4}$$

$$\sigma^2 = 2\left(1+4\alpha^2\right)^{-\frac{p}{2}} + \frac{2b^{-p}\left(1+2p\alpha^4\right)}{b^2} + \frac{3p\left(p+2\right)\alpha^8}{4b^4} \tag{6.5}$$

$$-4\varpi_\alpha^{-\frac{p}{2}}\left(1 + \frac{3p\alpha^4}{2\varpi_\alpha} + \frac{p\left(p+2\right)\alpha^8}{2\varpi_\alpha^2}\right)$$

Según las ecuaciones 6.4 y 6.5, $b = \left(1+2\alpha^2\right)$, $\varpi_\alpha = \left(1+\alpha^2\right)\left(1+3\alpha^2\right)$. En este sentido, la media y la varianza logarítmicas normalizadas del estadístico Henze-Zirkler se obtienen mediante las expresiones 6.6 y 6.7, respectivamente.

$$log\left(\mu\right) = \frac{1}{2}\log\left(\frac{\mu^4}{\sigma^2+\mu^2}\right) \tag{6.6}$$

$$log\left(\sigma^2\right) = log\left(\left(\sigma^2 + \mu^2\right)/\sigma^2\right) \tag{6.7}$$

Utilizando los parámetros dados, la significación de normalidad multivariante se evalúa utilizando el estadístico de prueba HZ_{Test} definido en la ecuación 6.8.

$$\mathrm{HZ}_{Test} = \left(\frac{log\left(\mathrm{HZ}\right) - log\left(\mu\right)}{log\left(\sigma\right)}\right) \tag{6.8}$$

La distribución normal multivariante en los datos del análisis se puede verificar mediante la prueba de *Royston*. Esta prueba se basa en el estadístico de prueba de $Shapiro - Wilk/Shapiro - Francia$ (WF_j). De modo que para cada variable $j = 1, 2, \ldots, p$, se calcula el estadístico WF_j, y se obtienen los valores R mediante la transformación de normalidad de *Royston*.

$$\omega f_j = log\left(1 - WF_j\right), \quad x = log\left(n\right), \quad 12 \leq n \leq 2000 \tag{6.9}$$

$$R = \frac{\left(\omega f_j - \mu\right)}{\sigma} \tag{6.10}$$

Los valores de μ y σ se obtienen mediante la aproximación polinomial definida en la ecuación 6.11. Estos coeficientes son proporcionados por Royston para diferentes tamaños de muestra.

$$\begin{aligned} \mu = &\quad \beta_{0\mu} + \beta_{1\mu}x + \beta_{2\mu}x^2 + \ldots + \beta_{m\mu}x^m \\ log\left(\sigma\right) = &\quad \beta_{0\sigma} + \beta_{1\sigma}x + \beta_{2\sigma}x^2 + \ldots + \beta_{m\sigma}x^m \end{aligned} \tag{6.11}$$

El estadístico de prueba de Royston (R_{Test}) se utiliza para evaluar la normalidad multivariante de los datos. Este estadístico se obtiene aplicando la ecuación 6.12. Obsérvese que el estadístico R_{Test} sigue una distribución chi-cuadrado con k grados de libertad ($R_{Test} \sim \chi^2_k$). Esto significa que la distribución chi-cuadrado proporciona una

base sólida para evaluar la significación estadística de los resultados obtenidos.

$$
\begin{aligned}
\mathrm{R}_{Test} = &\quad \frac{k\sum_{j=1}^{p}\gamma_j}{p} \sim \chi^2_k \\
\text{donde:} & \\
k = &\quad \frac{p}{1+(p-1)\,\bar{s}} \\
\gamma_j = &\quad \left(\phi^{-1}\left(\frac{\phi\,(-R)}{2}\right)\right)^2, \quad j=\{1,2,\ldots,p\}. \\
\bar{s} = &\quad \sum_i\sum_j s_{ij}/p(p-1)', \quad \bar{s}_{ij} = \begin{cases} f\,(t_{ij},n)\,, & \text{if} \quad i\neq j \\ 1 & \text{if} \quad i=j. \end{cases} \\
f\,(t,n) = &\quad t^{\theta}\left(1-\frac{\mu}{g}\,(1-t)^{\mu}\right)
\end{aligned}
\tag{6.12}
$$

El estadístico R_{Test} se obtiene dividiendo la suma de los términos γ_j por el número de variables p consideradas en el análisis. El parámetro k se calcula a partir de p variables y un término adicional $\bar{s}$ que involucra la media de los elementos de la matriz de covarianzas. El término γ_j, calculado para cada variable j, se obtiene a partir de la función inversa normal estándar (ϕ^{-1}). Obsérvese que $\bar{s}$ es la media de los elementos de la matriz de covarianzas que se calcula sumando los elementos individuales y dividiéndolos por $p(p-1)$. La función $f(t,n)$ calcula los elementos de la matriz de covarianzas, mostrando una relación no lineal con los parámetros t y n. Los parámetros μ y θ, se obtienen a partir del modelo de simulación propuesto por [559], asignando los valores $\mu = 0.715$ y $\theta = 5$. Por otro lado, el valor de g viene determinado por la expresión 6.13, con un tamaño de muestra de $10 \leq n \leq 2000$.

$$
g\,(n) = 0.21364 + 0.015124\left(y^2\right) - 0.0018034\left(y^3\right), \quad y = log\,(n)\,. \tag{6.13}
$$

En términos generales, según el teorema del límite central multivariante, la media muestral de una variable aleatoria multivariante se aproxima a una distribución normal multivariante. En este sentido, para obtener resultados válidos en el análisis discriminante cuando una muestra de instancias con características específicas se extrae de una población asimétrica y leptocúrtica, deben cumplirse los supuestos de normalidad multivariante y univariante. Estos criterios deben medirse con los estadísticos de

prueba analizados hasta ahora. Además, también es importante medir la igualdad de las matrices de varianza y covarianza entre los grupos, que se analiza a continuación.

- Igualdad de matrices de varianzas y covarianzas dentro de grupos.

Además de garantizar la normalidad en la estructura de los datos en la muestra de análisis, es importante garantizar la igualdad de las matrices de varianza y covarianza entre los grupos. En este contexto, se utiliza la prueba *M de Box*, representada por la ecuación 6.14 para evaluar el cumplimiento de este criterio.

La prueba *M de Box* es una medida que evalúa tanto la variabilidad entre grupos como la variabilidad dentro de los grupos en el análisis multivariante de la varianza. Su finalidad es determinar si existen diferencias significativas entre los grupos en relación con las variables dependientes consideradas en el estudio.

Utilizando un nivel de significación del 5% se examina el $p-valor$ obtenido mediante la prueba *M de Box*. Si el $p-valor$ es superior a 0.05, se concluye que las matrices de varianza y covarianza son homogéneas entre los grupos. Es decir, no existen diferencias significativas en la estructura de la variabilidad entre los grupos analizados.

$$M-Box = log\left|V'\right|^{(n-g)} - \sum_{j=1}^{n}\left(n_j-1\right) log\left|V^{(j)}\right|, \tag{6.14}$$

donde, el primer termino, $log\left|V'\right|^{(n-g)}$, se refiere a la covarianza total entre grupos y refleja la variabilidad total en todas las variables dependientes utilizadas en el análisis. El segundo término, $\sum_{j=1}^{n}\left(n_j-1\right) log\left|V^{(j)}\right|$, tiene en cuenta las covarianzas dentro de cada grupo y refleja la variabilidad dentro de cada grupo individual. n es el tamaño de la muestra de análisis. g es el número de grupos. V' es la matriz de covarianza total que captura las relaciones entre las variables dependientes en todos los grupos combinados. n_j es el tamaño de cada grupo j. Por último, $V^{(j)}$ es la matriz de covarianza dentro del grupo j, que captura las relaciones entre las variables dependientes dentro de cada grupo.

2. Función objetivo: Extracción de instancias La selección de una muestra de observaciones bajo restricciones de normalidad multivariante se aborda mediante la formulación de un modelo de optimización multicriterio. La figura 6.4, resume en términos generales la selección de instancias considerando dos objetivos esenciales, preservar la normalidad multivariante en la muestra de datos seleccionada y garantizar la igualdad de las matrices de varianzas y covarianzas entre grupos.

Figura 6.4: Modelo de selección de instancias.
La selección de instancias, en la conformación de muestras de estudio, se basa en cinco criterios fundamentales. Estos requisitos incluyen la evaluación de la normalidad univariante y multivariante de los grupos y las variables, y la verificación de la homogeneidad de las matrices de covarianza. Estos criterios garantizan una selección adecuada de las instancias que cumplen los requisitos de normalidad multivariante.

Nótese que para alcanzar el objetivo del modelo de selección de instancias se han especificado cinco condiciones que deben satisfacerse en la solución del problema. Estas restricciones se pueden expresar como una desigualdad en la que se busca el cumplimiento de las condiciones establecidas para las variables de decisión. Además, estas restricciones controlan, en el conjunto de instancias seleccionadas, la distribución y estructura de los datos.

Las dos primeras condiciones (1 y 2) limitan la búsqueda a instancias que en su conjunto asuman una distribución normal. Las condiciones 3 y 4 restringen aún más la selección de instancias, buscando que el conjunto seleccionado asuma una distribución normal multivariante. Por último, la quinta condición restringe aún más la búsqueda de instancias que en conjunto garanticen la homogeneidad de las matrices de varianza y covarianza. Una vez finalizado el proceso de selección, la muestra de estudio extraída cumple todos los supuestos estadísticos establecidos. Esto hace que la muestra del estudio sea adecuada para la aplicación del análisis discriminante.

La ecuación 6.15 presenta la función objetivo que maximiza el número de instancias seleccionadas x_i en el subconjunto de instancias X_j. El modelo de selección multicriterio tiene en cuenta el equilibrio de clases y las restricciones de normalidad e igual-

dad de las matrices de covarianza de las instancias seleccionadas.

$$
\begin{aligned}
\text{Maximizar:} \quad & \mathbf{Z} = f(X_j, x_i) = \sum_{i=1}^{n}(x_i) \\
\text{sujeto a:} \quad & \text{1. Restricciones de normalidad univariante para } X_j \text{ (en grupos):} \\
& SW_{Test} \quad xV(T_1) \geq 0.05 \\
& SW_{Test} \quad Sp(T_1) \geq 0.05 \\
& SW_{Test} \quad xV(T_2) \geq 0.05 \\
& SW_{Test} \quad Sp(T_2) \geq 0.05 \\
& \text{2. Restricciones de normalidad univariante para } X_j \text{ (en variables):} \\
& SW_{Test} \quad xV \geq 0.05 \\
& SW_{Test} \quad Sp \geq 0.05 \\
& \text{3. Restricciones de normalidad multivariante para } X_j \text{ (en grupos):} \\
& HZ_{Test} \quad xV, Sp(T_1) \geq 0.05 \\
& R_{Test} \quad xV, Sp(T_1) \geq 0.05 \\
& HZ_{Test} \quad xV, Sp(T_2) \geq 0.05 \\
& R_{Test} \quad xV, Sp(T_2) \geq 0.05 \\
& \text{4. Restricciones de normalidad multivariante para } X_j \text{ (en variables):} \\
& HZ_{Test} \quad xV, Sp \geq 0.05 \\
& R_{Test} \quad xV, Sp \geq 0.05 \\
& \text{5. Restricciones de homogeneidad de matrices de covarianzas para } X_j\text{:} \\
& M - Box_{Test} \quad xV, Sp \geq 0.05 \\
& \text{6. Restricciones de equilibrio de clases para } X_j\text{:} \\
& \left| \sum_{i=1}^{n}(x_i) \cdot \mathbf{1}_{\{T_1\}} - \sum_{i=1}^{n}(x_i) \cdot \mathbf{1}_{\{T_2\}} \right| \leq \omega \cdot n \\
\text{donde:} \quad & \\
& n = \text{número total de instancias } x_i \text{ seleccionadas} \\
& x_i = \text{variable binaria de selección} \\
& x_i = 1 \quad \text{instancia } x_i \text{ seleccionada} \\
& x_i = 0 \quad \text{instancia } x_i \text{ no seleccionada}
\end{aligned}
\tag{6.15}
$$

donde: T_1 = subconjunto de instancias alcistas seleccionadas

T_2 = subconjunto de instancias bajistas seleccionadas

X_j = conjunto de instancias seleccionadas

$\mathbf{1}_{\{T_1\}}$ = función indicadora, toma valor 1 si $x_i \in T_1$

en caso contrario, toma valor 0 si $x_i \notin T_1$

$\mathbf{1}_{\{T_2\}}$ = función indicadora, toma valor 1 si $x_i \in T_2$

en caso contrario, toma valor 0 si $x_i \notin T_2$

ω = parámetro que controla el equilibrio de clases T_1 y T_2

$\omega = 0.5$ si el número total de instancias seleccionadas n es impar

$\omega = 0.0$ si el número total de instancias seleccionadas n es par

xV, Sp = predictores seleccionados: xV = x.V*PT y Sp = Sp.Vr*CP.PT

SW_{Test} = p-valor de la prueba de Shapiro-Wilk

HZ_{Test} = p-valor de la prueba de Henze-Zirkler

R_{Test} = p-valor de la prueba de Royston

$M - Box_{Test}$ = p-valor de la prueba M de Box

El modelo de optimización multi-criterio (**Z**) maximiza el número de instancias x_i en el subconjunto de instancias seleccionadas X_j, considerando la normalidad, la igualdad de las matrices de covarianza y un equilibrio de clases entre T_1 y T_2. Cada instancia x_i a seleccionar se evalúa en el subconjunto acumulativo de instancias X_j ya seleccionadas, de forma que, si el subconjunto X_j satisface las restricciones establecidas, dicha instancia x_i asume el valor de 1 y, por tanto, se incorpora al subconjunto X_j. En caso contrario, dicha instancia x_i asume el valor de cero y se descarta su inclusión en el subconjunto de instancias seleccionadas X_j. Además, el número de instancias seleccionadas n debe mantener un equilibrio de clases independientemente de que n sea un número entero positivo par o impar. Las instancias x_i con etiqueta de clase T_1 deben ser aproximadamente el 50% del número total n de instancias seleccionadas.

La restricción de equilibrio entre clases establece que la diferencia absoluta entre las sumas de instancias seleccionadas en T_1 y T_2 debe ser menor o igual que ω. El valor de ω se define en función de la disponibilidad de datos y determina el equilibrio entre clases. Si el número total de instancias seleccionadas n es impar, ω debe ser 0.5

para permitir una diferencia de 0.5 entre las sumas de instancias seleccionadas en T_1 y T_2. Si n es par, ω puede ser 0, lo que implica un equilibrio exacto entre las sumas de instancias seleccionadas en ambos grupos.

La figura 6.5 presenta los resultados del proceso de validación estadística de la muestra de instancias seleccionadas mediante el modelo. La eficacia del modelo se evalúa seleccionando un subconjunto (X_j) de instancias de la población de estudio **O**. Este subconjunto cumple las condiciones requeridas y se confirma con pruebas estadísticas rigurosas, como Shapiro-Wilk (SW), Henze-Zirkler (HZ), Royston (R) y M de Box. Los resultados de estas pruebas, con $p - values$ superiores al nivel de significación del 5%, confirman la normalidad multivariante y la homogeneidad de las matrices de varianzas y covarianzas en los casos seleccionados.

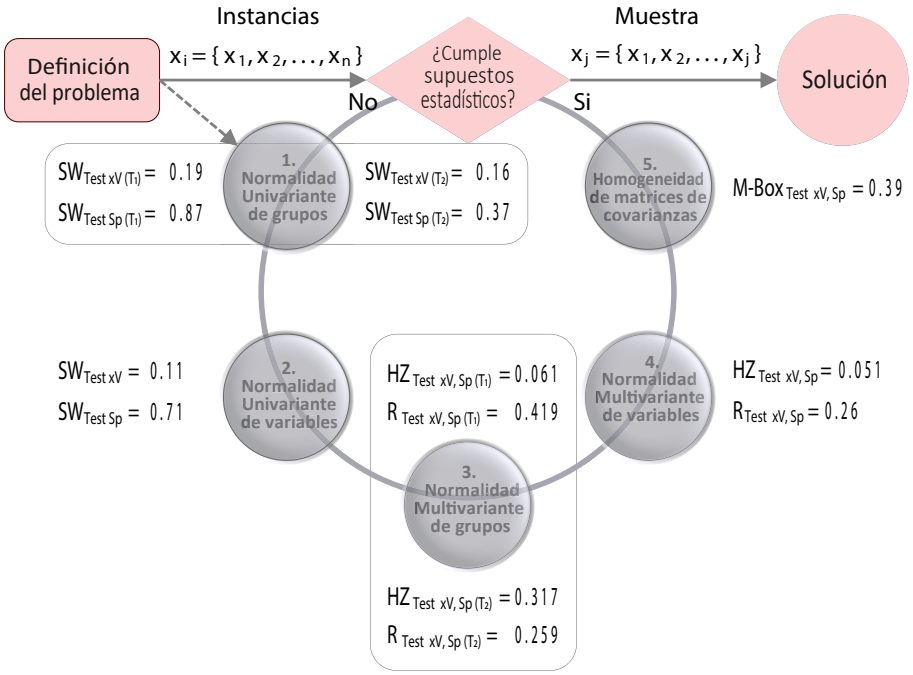


Figura 6.5: Validación estadística de la selección de instancias.
Evaluación de la eficacia del Modelo de Selección de Muestras. La selección de un subconjunto (X_j) de instancias cumplen con los tests estadísticos de Shapiro-Wilk (SW), Henze-Zirkler (HZ), Royston (R) y M de Box. Los $p - values$ obtenidos en estas pruebas, a un nivel de significación superior a 0.05, validan la normalidad multivariante y la homogeneidad de las matrices de varianzas y covarianzas en las instancias seleccionadas.

Como ya se ha mencionado, la población de estudio **O** consta de 1873 microtendencias, a favor de la dirección de la tendencia dominante, con un mayor número

de tendencias a la baja (1231) que al alza (642). El modelo de selección de instancias aborda este desequilibrio de clases sin comprometer la selección de características relevantes e interpretables. De las 1873 observaciones, se seleccionan 125 instancias que cumplen las restricciones de normalidad, igualdad de matrices de covarianza y equilibrio de clases entre $T_1 = 62$ y $T_2 = 63$ movimientos al alza y a la baja, respectivamente. El cumplimiento de estos supuestos es esencial para aplicar técnicas estadísticas paramétricas y obtener resultados válidos y significativos. Por lo tanto, la colección de datos $\mathbf{Z}\,\langle 125 \times 2\rangle$ consta de 125 puntos de inflexión (ET) sobre los que se miden 2 variables predictoras elegidas y garantiza el cumplimiento de los supuestos estadísticos señalados. La figura 6.6 pone de manifiesto la relevancia de que la muestra de instancias extraídas, cumpliendo los supuestos de normalidad e igualdad de matrices de covarianza entre grupos, sea representativa de la población $\mathbf{O}$.

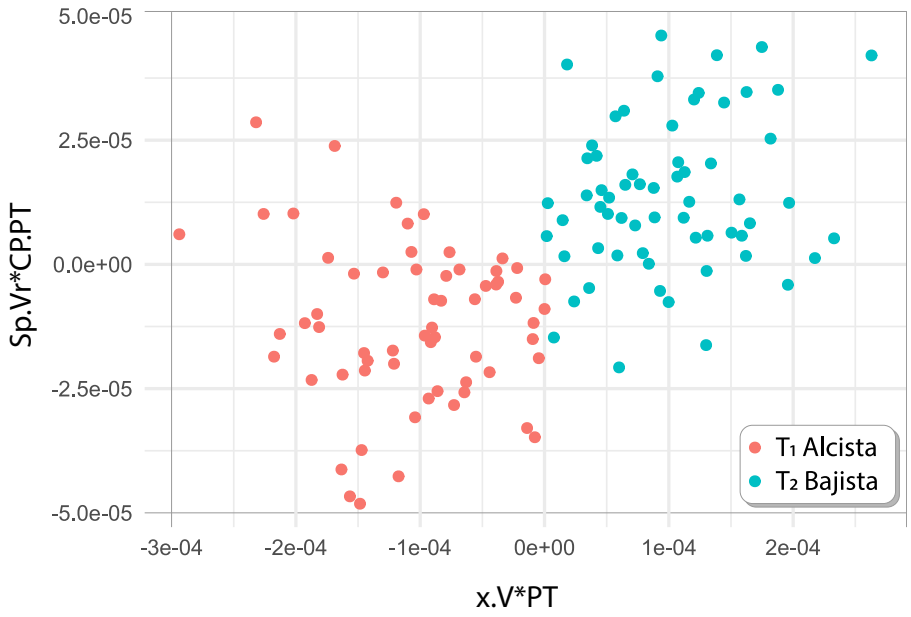


Figura 6.6: Dispersión multivariante de instancias extraídas.
El análisis de la muestra de datos $\mathbf{Z}$ con distribución normal multivariante revela relaciones y patrones significativos entre los mejores predictores. La disposición de las instancias muestra tanto independencia entre ellas como una clara separación entre clases. Estos puntos de inflexión revelan patrones distintivos en los precios, lo que respalda la construcción de modelos de clasificación para favorecer la toma de decisiones de inversión informadas.

Los datos muestran una clara diferenciación entre instancias alcistas y bajistas, especialmente en el eje x con la variable x.V*PT. Esta capacidad discriminativa se mantiene en la muestra extraída, diferenciando los puntos de inflexión que contribuyen

al inicio de los movimientos alcistas y bajistas en el cruce de las variables x.V*PT y Sp.Vr*CP.PT. Estos patrones son útiles en la construcción de modelos de clasificación para predecir nuevas tendencias. El potencial predictivo de estos patrones, conservado en la muestra de instancias extraídas, puede contribuir potencialmente a mejorar la toma de decisiones de inversión basada en datos. En consecuencia, la normalidad multivariante de las instancias extraídas es esencial en el uso de métodos de clasificación paramétricos, especialmente cuando se desea garantizar la fiabilidad de los resultados y las interpretaciones.

La figura 6.7 se divide en dos partes: (a) Gráficos de densidad para variables y movimientos de tendencia, y (b) Gráficos QQ normales para variables, ambos representados por las instancias extraídas.

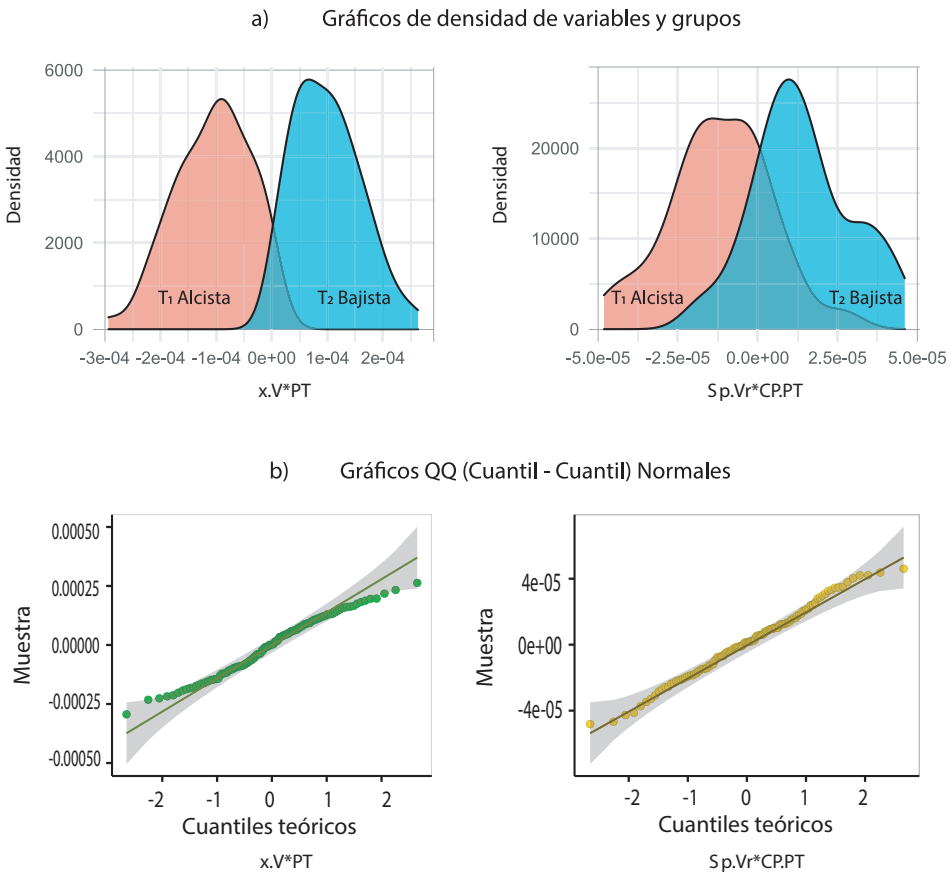


Figura 6.7: Distribución y normalidad multivariante en instancias extraídas. Distribución y normalidad multivariante y univariante de las instancias extraídas. (**a**) Gráficos de densidad para variables y movimientos de tendencia; (**b**) Gráficos QQ Normales para variables. Los gráficos de densidad demuestran la aproximación de las variables y grupos representados a una distribución normal. Los gráficos QQ revelan una marcada linealidad entre cuantiles, lo que indica una aproximación a una distribución normal.

Los gráficos de densidad de la figura muestran que las variables y los grupos representados siguen una distribución normal. Los gráficos QQ revelan una clara linealidad entre los cuantiles de la muestra y los cuantiles esperados, lo que sugiere una aproximación a una distribución normal en los casos seleccionados.

Los gráficos de densidad respaldan la presunción de normalidad multivariante en las instancias extraídas, mostrando que los datos siguen una distribución normal entre variables y grupos. Los gráficos QQ revelan una correspondencia entre cuantiles, lo que indica una distribución normal entre instancias. Estos hallazgos son relevantes para la construcción de modelos de clasificación basados en el análisis discriminante, ya que garantizan resultados confiables.

En términos generales, el modelo de extracción de instancias permite construir muestras de datos multivariantes con distribución normal, útiles en la formulación de modelos de clasificación paramétricos. Los resultados estadísticos y visuales confirman la normalidad multivariante en los datos, lo que hace que este modelo de extracción de instancias sea relevante para formar muestras representativas para la formulación de modelos de clasificación. Además, la relevancia de estos resultados hace que este tipo de muestreo sea pertinente para la predicción de tendencias en la negociación de otro tipo de instrumentos financieros.

6.3.3 Preparación de datos: Detección de estructuras

Comprender el comportamiento y la covariación de las variables observables y su vínculo con los factores subyacentes es esencial para obtener una imagen completa y precisa del fenómeno estudiado. Aunque la selección de variables predictoras, en la construcción de modelos de clasificación sigue un proceso exhaustivo, es necesario mediante el análisis de estructuras validar su poder discriminativo para obtener modelos de clasificación interpretables y precisos.

Esta sección presenta una metodología novedosa para seleccionar un subconjunto óptimo de variables con el fin de mejorar la eficacia de las estructuras detectadas y validar la capacidad discriminativa de los predictores seleccionados. La metodología propuesta utiliza una función objetivo basada en un enfoque de programación lineal entera mixta (MILP) para seleccionar un conjunto óptimo de variables.

El núcleo central de esta metodología se basa en evaluar la adecuación de los datos, el enfoque de solución para la selección de variables y la visualización de datos me-

diante análisis GH-Biplot. Los resultados y las interpretaciones se validan mediante los estadísticos de prueba KMO y MSA para evaluar la adecuación de los datos y la prueba de esfericidad de Bartlett para verificar la estructura de correlación entre las variables y garantizar la solidez del análisis. Los experimentos se basan en las mediciones hechas sobre los puntos de inflexión (ET) del tipo de cambio euro-dólar. La metodología propuesta logra una elevada adecuación muestral (KMO) y una fuerte correlación anti-imagen (MSA) de hasta 0.72, lo que favorece la formación de estructuras para la validación de patrones.

A. Generalidades sobre la detección de estructuras

La eficacia y validez de un modelo de selección de variables, que contribuye a reforzar las estructuras subyacentes en la validación de patrones, se basa en la consideración de aspectos fundamentales. A continuación se examinan los elementos clave que sientan las bases para la formulación de la función objetivo, incluida la evaluación de la adecuación de los datos, la singularidad de las matrices de datos, las medidas de adecuación del muestreo y la prueba de esfericidad de Bartlett.

- Adecuación de los datos.

La evaluación de la adecuación de los datos es crucial antes de confirmar y validar el poder discriminante de los predictores en el análisis y la detección de estructuras. Estudios notables sugieren tamaños de muestra grandes y relaciones de covariación entre variables con coeficientes de correlación ($r > 0.3$) para la detección fiable de estructuras [560].

Aunque la presencia de multicolinealidad entre variables puede detectarse en el análisis de estructuras, su presencia en variables predictoras puede afectar negativamente la estabilidad de los modelos de clasificación. Por tanto, la eficacia del subconjunto de variables seleccionadas puede tener un impacto positivo o negativo en la construcción de estructuras, dependiendo de su capacidad para capturar adecuadamente la variabilidad y las relaciones subyacentes entre las variables observadas.

En este sentido, La selección de un subconjunto óptimo de variables y la evaluación previa de la adecuación de los datos proporciona una base sólida para la detección, análisis, confirmación y validación del poder discriminante de los predictores en las estructuras subyacentes.

- Singularidad

La singularidad se manifiesta en las matrices de datos cuando no se puede calcular su matriz inversa y, por tanto, no tienen una solución única al sistema de ecuaciones lineales. Esta condición dificulta la interpretación de las relaciones entre variables y puede generar correlaciones espurias o infladas en los análisis de correlación. Así,la detección de matrices de datos singulares es crucial para obtener resultados significativos y fiables en el análisis de correlación y la detección de estructuras.

El determinante ($\det(\mathbf{X})$) es una medida comúnmente utilizada para determinar si una matriz de datos $\mathbf{X}$ es singular o no singular. Su cálculo varía en complejidad y eficiencia en función del tamaño y la estructura de la matriz. Normalmente, se utiliza un umbral de tolerancia cercano a cero ($u = 10^{-10}$) para determinar la singularidad, como se indica en la expresión 6.16. Sin embargo, este valor puede ajustarse en función de la precisión numérica requerida y de las características de los datos, tomando valores más pequeños cercanos a $u = 10^{-12}$ en los casos que lo requieran.

$$|\det(\mathbf{X})| \begin{cases} < u, & \text{si la matriz es singular} \\ \geq u, & \text{si la matriz es no singular} \end{cases} \tag{6.16}$$

En la ecuación presentada, u representa el umbral de tolerancia utilizado para determinar la singularidad de la matriz de datos $\mathbf{X}$. El valor de u puede ajustarse en función de la precisión requerida en el análisis. Es importante señalar que cuando el determinante de la matriz es inferior a u, es indicativo de la presencia de una alta multicolinealidad entre las variables de la matriz de datos. En casos extremos, cuando existe una multicolinealidad perfecta (cuando las variables predictoras son combinaciones lineales exactas), el determinante de la matriz de covarianzas (o matriz de correlaciones) de las variables predictoras es igual a cero. En estas situaciones, es pertinente considerar que la eliminación de variables linealmente dependientes o la aplicación de técnicas de regularización como ridge o lasso no sólo reducen la multicolinealidad, sino que también mejoran la singularidad de una matriz de datos.

- Medida de adecuación de los datos de Kaiser-Meyer-Olkin (KMO)

El estadístico KMO es una medida de adecuación del muestreo utilizada para evaluar y determinar qué tan adecuada es una matriz de datos para el análisis y la detección

de estructuras factoriales. Esta medida se calcula utilizando la ecuación 6.17.

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} V_{ij}^{2*}} \tag{6.17}$$

Obsérvese que la matriz de correlaciones r_{ij} representa las correlaciones entre todas las variables implicadas en el análisis, incluida la variable j. Por otra parte, la matriz de covarianza parcial V_{ij} representa las covarianzas parciales entre todas las variables distintas de la variable j. La covarianza parcial mide la relación única entre dos variables después de eliminar la influencia de las otras variables. Los coeficientes de correlación de la anti-imagen están determinados por v_{ij}^*. Por lo tanto, la medida de adecuación KMO proporciona un valor numérico que va de 0 a 1 e indica lo bien que se ajustan los datos a una estructura latente. Aunque no existe una escala con rangos de categorías estándar, se sabe que los valores de KMO superiores a 0.5 indican una idoneidad razonable de los datos para el análisis estructural. Sin embargo, es importante tener en cuenta que la interpretación exacta del valor KMO puede depender del contexto específico y de las características de los datos objeto de estudio. También deben tenerse en cuenta otros factores, como la finalidad del análisis y la complejidad del modelo, a la hora de evaluar la idoneidad de los datos.

- Matriz de correlaciones anti-imagen (MSA)

La Medida de Adecuación de la Muestra (MSA) es una medida clave en el análisis y la detección de estructuras. Estos valores se utilizan para evaluar la adecuación de cada variable en el conjunto de datos. Estas medidas individuales complementan la medida de adecuación global KMO, y proporcionan información detallada sobre la contribución que cada variable hace a la estructura factorial. Los valores MSA se basan en las correlaciones parciales de cada variable con las correlaciones parciales totales, lo que refleja su capacidad para representar la estructura subyacente de los datos. El uso combinado de valores MSA y KMO proporciona una evaluación completa de la idoneidad de los datos para el análisis de estructuras. Estos valores ayudan a formarse una idea más precisa de la estructura de los datos, especialmente para la toma de decisiones relacionadas con el análisis de estructuras.

Las matrices de correlación y de covarianza parcial contienen los coeficientes de correlación y las covarianzas parciales, respectivamente. La matriz de correlaciones refleja las relaciones lineales entre las variables, mientras que la matriz de covarianzas parciales muestra las relaciones únicas tras controlar las influencias de las demás

variables. Específicamente, los elementos diagonales de la matriz de correlaciones anti-imagen representan la medida de adecuación MSA de la muestra para cada variable en particular. La matriz de covarianza anti-imagen de la matriz $\mathbf{V} = (v_{ij})$ viene determinada por la ecuación 6.18.

$$v_{ij} = \frac{r^{(ij)}}{r^{(ii)} r^{(jj)}} \tag{6.18}$$

Los elementos de la matriz de covarianza anti-imagen $\mathbf{V}$ se calculan utilizando los coeficientes de correlación parcial de la matriz de correlación $\mathbf{R}$. Cada elemento v_{ij} en la posición (i, j) de $\mathbf{V}$ representa la covariación específica entre las variables i y j, después de eliminar las influencias de las otras variables.

- Prueba de esfericidad de Bartlett

La prueba estadística de esfericidad de Bartlett es una herramienta crucial que permite evaluar previamente la adecuación de los datos antes de proceder a la detección de estructuras subyacentes. Además, el valor obtenido con esta prueba respalda los resultados obtenidos con la medida KMO. En la ecuación 6.19 se resume el procedimiento utilizado para su calculo.

$$\chi^2 = -\left[(n-1) - \frac{1}{6}(2p+5)\right] \cdot ln\,|R| \tag{6.19}$$

El valor de χ^2 en la prueba de esfericidad de Bartlett, a partir de la ecuación anterior, permite evaluar si la matriz de correlaciones R es una matriz de identidad. Donde n es el tamaño de la muestra y p el número de variables implicadas en el estudio.

En este contexto, cuando el valor χ^2 es mayor que el valor crítico y el $p - valor$ es menor que el nivel de significancia de 5% se rechaza la hipótesis nula H_0. Esto significa que las variables están correlacionadas y, por tanto, el análisis y la detección de estructuras son adecuados para el conjunto de datos. En caso contrario, cuando el valor χ^2 es inferior al valor crítico o el $p-valor$ es mayor al nivel de significancia del 5%, no se rechaza la hipótesis nula. Esto sugiere que las variables no están correlacionadas y, por lo tanto, las variables evaluadas no son adecuadas para el análisis y la detección de estructuras.

B. Función objetivo: Selección de variables para fortalecer estructuras

En el contexto de la selección óptima de variables para la construcción de estructuras y la detección de patrones discriminativos, se utiliza un modelo de programación lineal

entera mixta (MILP) con una función objetivo que busca maximizar la medida de adecuación muestral KMO. La figura 6.8 resume, en la función objetivo, la inclusión de las restricciones más importantes relacionadas con el tamaño de los subconjuntos de variables observables fijas, disponibles y de elección. Estas restricciones aseguran una selección precisa y eficaz de las variables y garantizan resultados fiables en la construcción de estructuras.

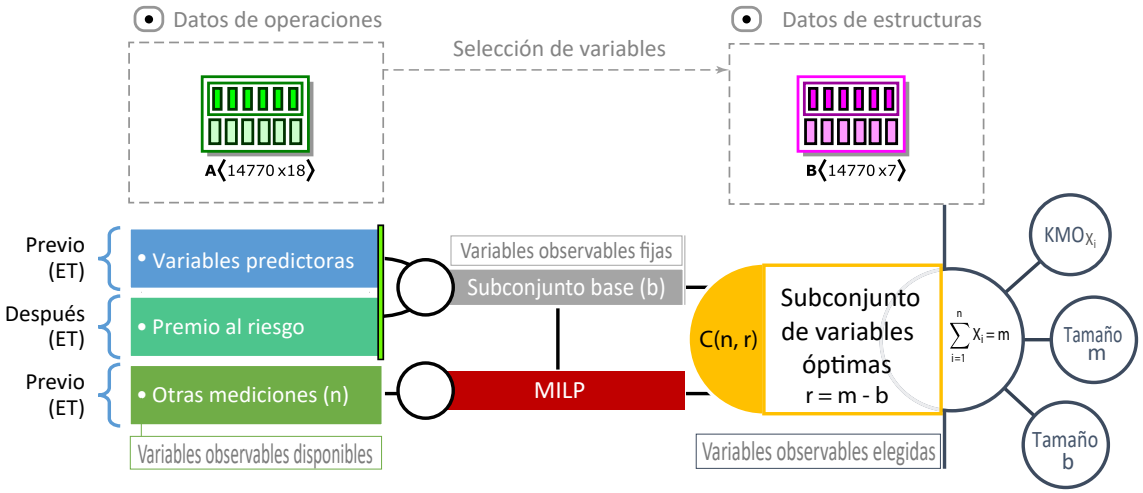


Figura 6.8: Selección óptima de variables para máximo KMO con modelo MILP. Optimización de variables en un modelo MILP para maximizar la medida de adecuación muestral KMO. Incluye variables de entrada, función objetivo y restricciones. La salida es el subconjunto óptimo de variables seleccionadas. Las variables de entrada son predictores, prima de riesgo y otras medidas. El modelo MILP selecciona el mejor subconjunto dentro del espacio de combinaciones $\binom{n}{r}$, maximizando KMO para los tamaños m y b.

La ecuación 6.20 representa la función objetivo que maximiza el valor KMO al seleccionar el subconjunto óptimo de variables observables X_i según restricciones de tamaño en los subconjuntos de variables fijas (b), disponibles(n) y de elección (m).

Obsérvese que la función objetivo pretende maximizar el valor KMO seleccionando el subconjunto óptimo de características. Para garantizar que la maximización se realiza con el mínimo número de variables, se incorpora la restricción adicional de que la suma de las variables seleccionadas sea igual a m, según la restricción $\sum_{i=1}^{n} x_i = m$. Esto garantiza que el subconjunto evaluado tenga el tamaño deseado. Además, al considerar un número fijo de variables (b) el modelo incorpora y selecciona sobre esta base las variables que mejor contribuyen a la construcción de estructuras. Esto demuestra que las variables complementarias, seleccionadas por el modelo, se eligen en función de su contribución a la construcción de estructuras sobre la base de las

variables fijas.

$$
\begin{aligned}
\text{Maximizar:} \quad & \mathbf{KMO}_{X_i} = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} V_{ij}^{2*}} \\
\text{sujeto a:} \quad & X_i = b \cup r \\
& \sum_{i=1}^{n} x_i = m \\
& m \in C(n, r) \\
& C(n, r) = n!/(r! \cdot (n - r)!) \\
& r = m - b, \quad r = \{1, 2, 3, \ldots, n\} \\
& m \geq b \\
& m \leq M, \quad M = b + n \\
& \det(\mathbf{X}_i) \geq u \\
& u = 10^{-10} \\
& x_i \in \{0, 1\}, \forall_i \\
\text{donde:} \quad & \\
& KMO_{X_i} = \text{valor KMO del subconjunto de variables } x_i \text{ evaluadas} \\
& b \cup r = \text{conjunto } X_i \text{ que contiene los elementos de } b \text{ y } r \\
& V_{ij}^{*} = \text{coeficientes de correlación anti-imagen (MSA)} \\
& r_{ij} = \text{correlaciones originales entre variables } i \text{ y } j \\
& KMO = \text{medida de adecuación muestral de grupo:} \quad 0 \leq KMO \leq 1 \\
& MSA = \text{medida de adecuación muestral de cada variable:} \quad 0 \leq MSA \leq 1 \\
& x_i = \text{variable binaria de selección} \\
& i = 1 \quad \text{variable } x_i \text{ incluida en el subconjunto de evaluación} \\
& i = 0 \quad \text{variable } x_i \text{ no incluida en el subconjunto de evaluación} \\
& m = \text{tamaño del subconjunto de variables a evaluar} \\
& C(n, r) = \text{subespacio combinatorio de los conjuntos de variables a evaluar} \\
& M = \text{tamaño máximo permitido del subconjunto de evaluación} \\
& n = \text{número de variables disponibles} \\
& r = \text{cantidad de variables adicionales para alcanzar el tamaño } m \\
& b = \text{número de variables fijas en el subconjunto base}
\end{aligned}
\tag{6.20}
$$

De acuerdo con la ecuación 6.20 el modelo MILP esta formado por una función objetivo y restricciones específicas que se definen de acuerdo a los objetivos del problema. Este modelo busca optimizar la estructura subyacente de un grupo de variables X_i conformado por un subconjunto de variables fijas (b) y variables adicionales (r) necesarias para formar un grupo de evaluación de tamaño m, donde $r = m - b$. Además, el número total de variables disponibles para la formación de grupos es n.

- Restricciones de cumplimiento

El modelo de optimización esta sujeto al cumplimiento de las siguientes restricciones:
El subconjunto de variables seleccionadas (X_i) está compuesto por las variables fijas (b) y las variables adicionales (r), tal como lo confirma la expresión $X_i = b \cup r$.
La suma de todas las variables seleccionadas (x_i) es igual al tamaño deseado del subconjunto (m), donde $\sum_{i=1}^{n} x_i = m$ y $b < m \leq n$.
El tamaño del subconjunto (m), dada su estructura combinatoria, pertenece al espacio de los conjuntos de variables a evaluar ($C(n, r)$), de modo que $m \in C(n, r)$.
La cantidad de variables adicionales (r) necesarias para alcanzar el tamaño deseado del subconjunto (m), considerando las variables fijas (b) esta determinado por $r = m - b$.
El tamaño del subconjunto (m) debe ser mayor o igual al número de variables fijas (b) de modo que $m \geq b$. Además, el tamaño del subconjunto (m) esta condicionado al tamaño máximo permitido (M) de forma que $m \leq M$, y $M = b + n$.
El umbral de tolerancia fijado para que la matriz de datos $(X)_i$ se considere no singular es (u), de forma que $\det(\mathbf{X}_i) \geq u$, donde $u = 10^{-10}$
Las variables de selección (x_i) son binarias, es decir, pueden tomar valores de 1 o 0 cuando son y no son elegidas para la evaluación, respectivamente. De forma que $x_i \in \{0, 1\}, \forall_i$
Las variables de selección (x_i) son binarias, es decir, pueden tomar valores de 0 cuando no son elegidas y 1 cuando son elegidas para la evaluación. De forma que $x_i \in \{0, 1\}, \forall_i$.

- Formación de grupos en el subespacio de combinaciones

Las estructuras latentes se construyen en grupos de tamaño m, compuestos por b variables fijas y r variables observables seleccionadas de un conjunto de n variables disponibles. Las variables fijas (b) se dividen en dos grupos. El primer grupo incluye las variables predictoras, como el rendimiento medio del precio de cierre (x.V*PT) y la pendiente de la linea de regresión de las variaciones de los precios de cierre (Sp.Vr*CP.PT),

tomadas 18 periodos antes del punto de inflexión (ET). El segundo grupo incluye medidas de tendencia (Sp.XP-EP) y prima de riesgo (GPL_$ y RP% M[1]), tomadas después del punto de inflexión (ET). El conjunto de variables disponibles (n) consta de 13 medidas tomadas antes del punto de inflexión (ET). Todas estas mediciones tomadas antes y después de los puntos de inflexion (ET) se compilan en la matriz de datos $\mathbf{A}\ \langle 14770 \times 18 \rangle$ (véase la tabla 5.3).

Los grupos de tamaño m se forman con variables observables fijas b y disponibles n. Para cada grupo, se seleccionan r variables adicionales de entre las disponibles para alcanzar el tamaño m. El número total de combinaciones posibles de r variables seleccionadas de un conjunto de n variables disponibles se calcula mediante la fórmula de combinaciones $C(n, r)$ como se ve en la ecuación 6.21. En consecuencia, para construir grupos de tamaño m (de 6 a 18 variables observables), se itera sobre los valores de m y se calcula r como la diferencia entre m y el número de variables fijas b. Este procedimiento iterativo crea los grupos correspondientes según el espacio de combinaciones posibles $C(n, r)$. Así, cada grupo de variables esta formado por las variables fijas b y una combinación de r variables disponibles.

$$\binom{n}{r} = \frac{n!}{r! \cdot (n-r)!} \tag{6.21}$$

Tabla 6.4: Variables e Indices

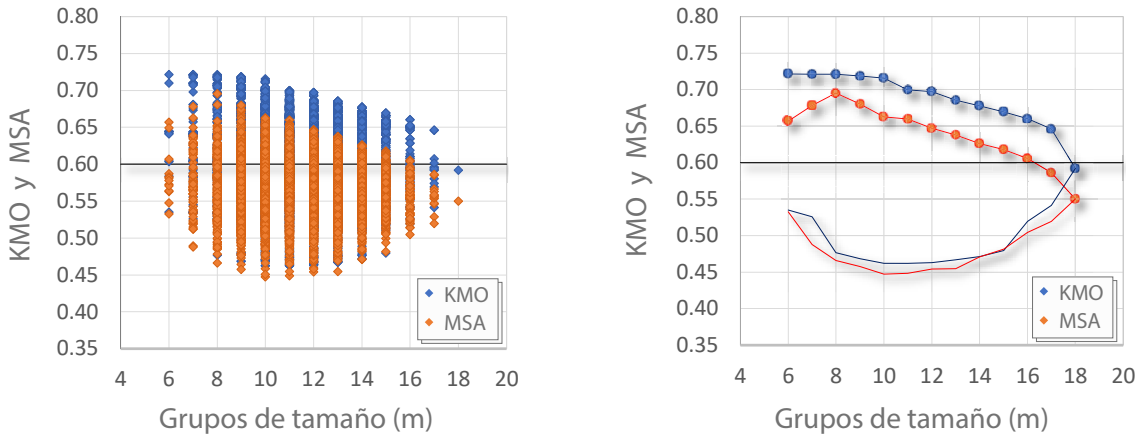
Tamaño de grupos (m)	Variables fijas (b)	Variables disponibles (n)	Variables adicionables (r)	C(n, r)	Combinaciones
6	5	13	1	C(13, 1)	13
7	5	13	2	C(13, 2)	78
8	5	13	3	C(13, 3)	286
9	5	13	4	C(13, 4)	715
10	5	13	5	C(13, 5)	1287
11	5	13	6	C(13, 6)	1716
12	5	13	7	C(13, 7)	1716
13	5	13	8	C(13, 8)	1287
14	5	13	9	C(13, 9)	715
15	5	13	10	C(13, 10)	286
16	5	13	11	C(13, 11)	78
17	5	13	12	C(13, 12)	13
18	5	13	13	C(13, 13)	1
Total					8568

La tabla 6.4 muestra las posibilidades de formar grupos de variables en función del tamaño $b < m \leq n$ y del número de variables adicionales (r). Resumiendo los datos del subespacio, la tabla representa todas las combinaciones posibles de variables para

grupos de 6 a 18 variables. Cinco variables fijas (b) se mantienen en todos los grupos, mientras que 13 variables (n) están disponibles para la selección. El valor de (r) varía de 1 a 13, indicando el número de variables que pueden añadirse al subconjunto base para alcanzar el tamaño del grupo. C(n, r) representa el número de combinaciones posibles al seleccionar (r) variables de un conjunto de (n) variables disponibles.

Los datos de la tabla anterior revelan que el número de combinaciones posibles en el subespacio de combinaciones para grupos de 6 a 18 variables es de 8568 subconjuntos a evaluar. Obsérvese que a medida que aumenta el tamaño del grupo (m), también lo hace el número de combinaciones posibles, que alcanza su máximo cuando (m) es igual al número total de variables disponibles (n = 13). La combinación con el menor número de variables adicionales (r = 1) tiene 13 posibilidades, mientras que la combinación con el mayor número de variables adicionales (r = 13) sólo tiene una posibilidad. Estos resultados proporcionan información valiosa sobre las opciones de formación de grupos en función del tamaño y el número de variables adicionales.

La Figura 6.9 presenta los valores de KMO y MSA para diferentes tamaños de grupo de variables, lo que permite evaluar el ajuste de los datos a diferentes estructuras subyacentes en función del tamaño de los grupos.



(a) Subespacio de combinaciones (b) Valores máximos y mínimos

Figura 6.9: Efecto del tamaño de grupo en medidas de KMO y MSA.
Valores KMO y MSA para distintos tamaños de grupo. (**a**) Subespacio de combinaciones de grupos de variables. (**b**) Valores máximos y mínimos de adecuación muestral. Grupos más pequeños (6, 7 y 8 variables) con mejor ajuste y mayor capacidad para capturar la estructura subyacente de los datos. Grupos más grandes requieren un análisis más detallado para comprender su idoneidad.

El análisis del subespacio de combinaciones de la figura (6.9a) y la evaluación de

los valores máximo y mínimo de KMO y MSA para cada tamaño de grupo de la figura (6.9b) revelan resultados importantes. Puede observarse que, a medida que aumenta el tamaño del grupo de variables, el valor máximo de KMO disminuye gradualmente. Los grupos más pequeños (6, 7 y 8 variables) muestran un mejor ajuste del modelo en comparación con los grupos más grandes. Esto puede indicar una estructura subyacente más robusta o una mayor coherencia entre las variables seleccionadas para detectar y validar los patrones discriminativos de las variables fijas.

Al analizar los valores medios de MSA, la Figura 6.9 muestra que son más altos para los grupos más pequeños de 6, 7 y 8 variables. Esto indica que dichos grupos son capaces de capturar mejor la estructura subyacente de los datos que los grupos más grandes. En estas condiciones, es más probable que los grupos más pequeños revelen patrones y relaciones relevantes entre variables en comparación con los grupos de mayor tamaño.

Por otra parte, es importante señalar que los tamaños de grupo de 16, 17 y 18 variables muestran valores de KMO y MSA más bajos. Sin embargo, esto no implica necesariamente que deban evitarse por completo, ya que cada tamaño de grupo puede proporcionar información única. En estas condiciones, también sería deseable un análisis más detallado de estos grupos para comprender mejor la estructura subyacente de los datos y determinar su idoneidad.

- Enfoques de selección de estructuras.

El subespacio combinatorio de variables ofrece 8568 combinaciones o grupos de diferentes tamaños, que son útiles para detectar estructuras y validar patrones. El modelo MILP mapea desde el subespacio combinatorio cada grupo de variables y evalúa la adecuación muestral de los datos utilizando la medida KMO. El modelo selecciona dentro de cada tamaño de grupo la estructura subyacente que produce los valores más altos de KMO y MSA. La elección de los mejores grupos de variables, que representan las mejores estructuras, son los más adecuados para confirmar los patrones discriminantes mediante el análisis GH-Biplot.

En este contexto, al considerar el subespacio combinatorio de variables para diferentes tamaños de grupo, se pueden identificar tres tipos diferentes de soluciones en relación con la elección de la estructura subyacente que se va a estudiar.

El primer enfoque de solución se centra en la selección y el análisis de los grupos con los mejores resultados, concretamente los de menor tamaño (6, 7 y 8 variables) con los valores más altos de KMO y MSA. Estos grupos representan una combinación

óptima de variables que se ajusta mejor a los datos. Por lo tanto, son especialmente adecuados para evaluar las estructuras subyacentes, explorar y visualizar las relaciones entre variables y desvelar patrones discriminantes mediante el análisis GH-Biplot.

El segundo enfoque de solución es aplicable cuando se conocen de antemano el tamaño del grupo, las variables fijas y las posibles variables que se incluirán en el análisis. En este caso, tomando como referencia el subespacio de combinaciones con los tamaños de grupo que han mostrado los mejores resultados (por ejemplo, grupos de 6, 7 y 8 variables), se configura el grupo con el conjunto específico de variables fijas y las variables de interés. A continuación, se evalúa la estructura subyacente y se compara con los mejores conjuntos. Este enfoque proporciona información valiosa sobre la estructura analizada en comparación con las configuraciones que producen mejores resultados.

El tercer enfoque de solución se basa en seleccionar una estructura que haya experimentado mejoras significativas en comparación con una estructura base inicial. La estructura base (b) se define mediante la ecuación 6.22, que determina el tamaño del conjunto de variables (m). En este enfoque, la estructura base se mejora añadiendo un subconjunto de variables (r) que optimicen la estructura. Si la adición de variables no consigue la mejora requerida (evaluada por el valor KMO), se mantiene la estructura base sin incluir variables adicionales.

$$\sum_{i=1}^{n} x_i = m \begin{cases} m_{(b+r)}, & \text{si } KMO_{(b+r)} > KMO_{(b)} \\ m_{(b)}, & \text{si } KMO_{(b+r)} \leq KMO_{(b)} \end{cases} \tag{6.22}$$

Esta ecuación establece que la suma de las variables (x_i) determina el tamaño del grupo (m). Si la mejora de la estructura, medida por el valor KMO, es significativa al añadir el subconjunto de variables (r) a la estructura base (b), se selecciona el tamaño de grupo $m_{(b+r)}$. En caso contrario, se mantiene el tamaño de grupo $m_{(b)}$. Este enfoque permite obtener una estructura mejorada considerando cuidadosamente la adición de variables.

En términos generales, el primer enfoque de solución se basa en la selección de estructuras óptimas utilizando criterios específicos y bien definidos. El segundo enfoque consiste en elegir una estructura basándose en una selección previa de variables de interés. Por último, el tercer enfoque se centra en la mejora de la estructura existente mediante la inclusión de variables adicionales o el mantenimiento de la estructura base utilizando criterios de mejora establecidos.

6.4 Selección de características

Desde un enfoque estadístico exploratorio - confirmatorio, en esta sección se describe el procedimiento utilizado para identificar y seleccionar los determinantes que explican mejor la acción del precio y ayudan a clasificar y predecir el movimiento direccional del tipo de cambio euro-dólar. La sección consta de tres apartados. El apartado 6.4.1 ofrece una descripción general del proceso de selección de características. El apartado 6.4.2 detalla el procedimiento utilizado para seleccionar variables predictoras independientes con alto poder discriminante. El apartado 6.4.3 describe en detalle el procedimiento utilizado para determinar el potencial discriminante de las variables seleccionadas.

6.4.1 Generalidades del proceso de selección

El poder predictivo de un modelo de clasificación depende de las variables predictoras seleccionadas y del proceso de selección. La inclusión de variables irrelevantes afecta negativamente el rendimiento del modelo. Para abordar este problema, se utiliza un enfoque estadístico exploratorio-confirmatorio basado en las diferencias multidimensionales entre movimientos alcistas y bajistas. Este enfoque ayuda a identificar y confirmar la efectividad de las variables responsables de la separación de estas estructuras de agrupación.

La metodología propuesta se utiliza para identificar y seleccionar los determinantes del movimiento del tipo de cambio euro-dólar a partir de una matriz de datos multivariante $\mathbf{A}\,\langle i \times j \rangle$. En esta matriz, $\mathbf{i}$ representa los movimientos de tendencia basados en puntos de inflexión (ET) a los que $\mathbf{p}$ variables numéricas observables $\mathbf{x}j = (\mathrm{x1}, ..., \mathrm{x}_p)$ se miden. La distinción entre grupos se realiza mediante una variable categórica definida por el vector de datos $\dot{\mathbf{y}} = \{0, 1\}$, donde 0 indica "Movimiento Tendencial Alcista" y 1 indica "Movimiento Tendencial Bajista". El proceso de selección busca identificar el subconjunto de predictores $\hat{\mathbf{x}}_j$, formado por $\mathbf{p}$ variables independientes, que es mejor que $\mathbf{x}_j$, siguiendo la condición $(\hat{\mathbf{x}}_\mathbf{j} \subset \mathbf{x}_\mathbf{j})$ y $(\mathbf{x}_j > \hat{\mathbf{x}}_\mathbf{j})$. El resultado es una nueva matriz $\mathbf{A}'$ con dimensiones $\mathbb{R}^{i \times \hat{j}}$, donde $\hat{j} < j$.

La función de maximización de la ecuación 6.23 evalúa las variables $\mathbf{x}_j$ de la matriz de datos $\mathbf{A}$. Los vectores β_j representan los coeficientes, y las matrices L y T corresponden a las matrices de covarianza.

$$\text{Maximizar:} \quad Z = \frac{\beta_j^\top L \beta_j}{\beta_j^\top T \beta_j} \tag{6.23}$$

La función $\mathbf{Z}$ evalúa las variables $\mathbf{x}_j$ utilizando la matriz de datos $\mathbf{A}$ e intenta maximizar la variabilidad entre grupos en relación con la variabilidad total. Esto se consigue seleccionando un subconjunto de variables $\mathbf{x}_j$. La covarianza total $\mathbf{T} = \mathbf{K}+\mathbf{L}$, se compone de las covarianzas dentro del grupo y entre grupos. β_j es el vector de coeficientes de la función discriminante que evalúa las variables observables $\mathbf{x}_j$. De este modo, el subconjunto de predictores elegibles $\hat{\mathbf{x}}_\mathbf{j}$ es el que mejor contribuye a la diferenciación entre grupos, tiene el mayor poder discriminante y ofrece el mejor rendimiento predictivo. Además, la selección de las características más fehacientes se valida realizando pruebas de significación estadística. Estos predictores se utilizan en la construcción del modelo de clasificación.

6.4.2 Proceso de selección de variables

Este proceso consta de tres etapas: cribado de variables observables, evaluación del potencial discriminante y selección de variables predictoras independientes, como se ilustra en la Figura 6.10.

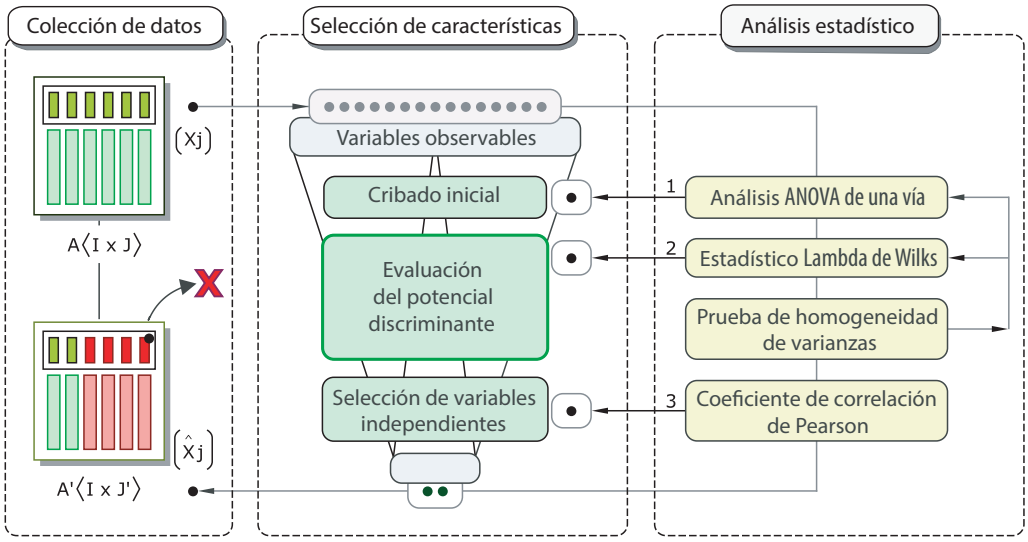


Figura 6.10: Procedimiento para la selección de características.
Procedimiento propuesto para identificar y seleccionar variables predictoras independientes con mayor poder discriminante. Método basado en un enfoque estadístico de tipo exploratorio - confirmatorio. Este procedimiento se resume en tres etapas: (1) Cribado inicial de variables observables. (2) Evaluación del potencial discriminante y (3) Selección de variables predictoras independientes.

En primer lugar, se mide la capacidad que tiene una variable para diferenciar entre grupos. Este criterio es lo que hace que una variable candidata sea potencialmente

elegible. Este criterio se detecta mediante un análisis ANOVA de una vía. El número de variables candidatas se reduce con esta selección inicial. En segundo lugar, se identifican los predictores que mejor contribuyen a la separación entre grupos mediante el estadístico Lambda de Wilks. El potencial discriminante de cada variable candidata se mide mediante este criterio. La prueba de homogeneidad de varianzas de Levene valida la prueba de igualdad de medias de grupo y reduce el número de variables elegibles detectadas con el estadístico Lambda de Wilks. Por último, la asociación entre variables elegibles se mide con el coeficiente de correlación de Pearson. Los predictores que pueden causar problemas de multicolinealidad se descartan de la formulación del modelo. Las variables que se dejan fuera del análisis, no mejoran la separación de los grupos, no aportan información al modelo y, por lo tanto, se excluyen de los procedimientos posteriores.

6.4.3 Potencial discriminante de variables elegibles

En relación con el proceso que se muestra en la Figura 6.10, el segundo momento se basa en el método de selección por pasos. Esto ayuda a identificar, estadísticamente, las variables explicativas que tienen un fuerte poder discriminativo y proporcionan el mejor rendimiento en la construcción de modelos de clasificación. Este procedimiento estadístico es un método iterativo en el que las variables de estudio se analizan individualmente hasta obtener un subconjunto de variables elegibles $\hat{\mathbf{x}}_j$. El estadístico $LambdadeWilks$ y el valor del estadístico asociado, F de $Snedecor$, se utilizan para determinar el potencial discriminante de cada variable y el subconjunto de variables elegibles $\hat{\mathbf{x}}_j$. Los atributos con valores F de $Snedecor$ grandes contribuyen a la separación de las medias de los grupos y, por tanto, discriminan mejor. Por el contrario, los valores más bajos no discriminan debido a atributos con grupos muy espaciados y datos muy dispersos. El procedimiento comienza calculando, para cada variable de análisis, los valores F de $Snedecor$ de inclusión F_ε y de eliminación F_r dados por las ecuaciones (6.24) y (6.25), como sigue:

$$F_\varepsilon = \frac{\left(\check{f} - \check{e}\right)(n - q - g)}{\check{e}\,(g - 1)} \tag{6.24}$$

$$F_r = \frac{\left(\check{e} - \check{f}\right)(n - q - g + 1)}{\check{f}\,(g - 1)} \tag{6.25}$$

Donde $\check{e}$ es la matriz de sumas de cuadrados y productos cruzados dentro de los

grupos, $\check{f}$ es la matriz de sumas de cuadrados y productos cruzados totales, q corresponde al grupo de variables incorporadas en el análisis, g es el número de grupos, n es el número de observaciones.

Las variables con niveles de inclusión superiores al valor F para entrar ($F_\varepsilon = 3.84$) entran en el subconjunto, de variables elegibles $\hat{\mathbf{x}}_j$, y las variables con valores de remoción inferiores al valor F de eliminación ($F_r = 2.71$) salen del análisis. El proceso se repite hasta que no haya variables candidatas para eliminar. Si la tolerancia de la variable de análisis cae por debajo de la tolerancia especificada (0.001), la variable no es elegible. La tolerancia ψ_i puede escribirse como $\psi_i = (1 - R^2)$. La tolerancia se entiende como el porcentaje de varianza no explicada entre la variable analizada y las variables incluidas. El valor disminuye drásticamente si las variables están correlacionadas, mientras que producirá pequeñas variaciones cuando las variables son independientes.

El modelado del poder discriminante de la función lineal se realiza en función de algunas o todas las $\mathbf{p}$ variables elegibles $\hat{\mathbf{x}}_j$. El estadístico Lambda de Wilks es el criterio de referencia utilizado para evaluar el poder discriminante de las variables utilizadas en el modelo. Este criterio contrasta la dispersión de las instancias dentro de los grupos y la dispersión de los datos sin distinguir entre grupos. Su cálculo se da en la ecuación (6.26). Cuanto más se aproxime el valor a cero, mayor será el poder discriminante de las variables analizadas, mientras que si el valor se aproxima a uno, menor será el poder discriminante.

$$\Lambda = \frac{|\mathbf{K}|}{|\mathbf{T}|} \tag{6.26}$$

El estadístico Lambda de Wilks ($\mathbf{\Lambda}$) se define como la razón entre la determinante de la covarianza dentro del grupo ($|\mathbf{K}|$) y la determinante de la covarianza total ($|\mathbf{T}|$), donde $\mathbf{T} = \mathbf{K} + \mathbf{L}$. Mediante pruebas de significación estadística, se identifican las variables elegibles ($\hat{\mathbf{x}}_j$) que, desde una perspectiva estadística y de análisis técnico, explican el origen de una nueva tendencia basada en el proceso de culminación del movimiento anterior. Estos predictores elegibles proporcionan valor explicativo, mejoran la interpretación de los resultados y contribuyen a la formulación de modelos de predicción más precisos e interpretables

En notación matricial $\mathbf{T} = \mathbf{K} + \mathbf{L}$ es la covarianza total, con $\mathbf{K}$ y $\mathbf{L}$ como las covarianzas dentro del grupo y entre grupos. $\mathbf{\Lambda}$ es el estadístico Lambda de Wilks. Finalmente, el uso de pruebas de significación estadística facilita la identificación del

subconjunto de variables elegibles $\hat{\mathbf{x}}_j$ que, por su validez estadística y desde el punto de vista del análisis técnico, ayudan a explicar el origen de una nueva tendencia en función del proceso de culminación del movimiento anterior. Además de esta connotación, los predictores elegibles $\hat{\mathbf{x}}_j$ son más útiles porque aportan valor explicativo a los resultados obtenidos y contribuyen a la formulación de modelos de predicción interpretables y más precisos.

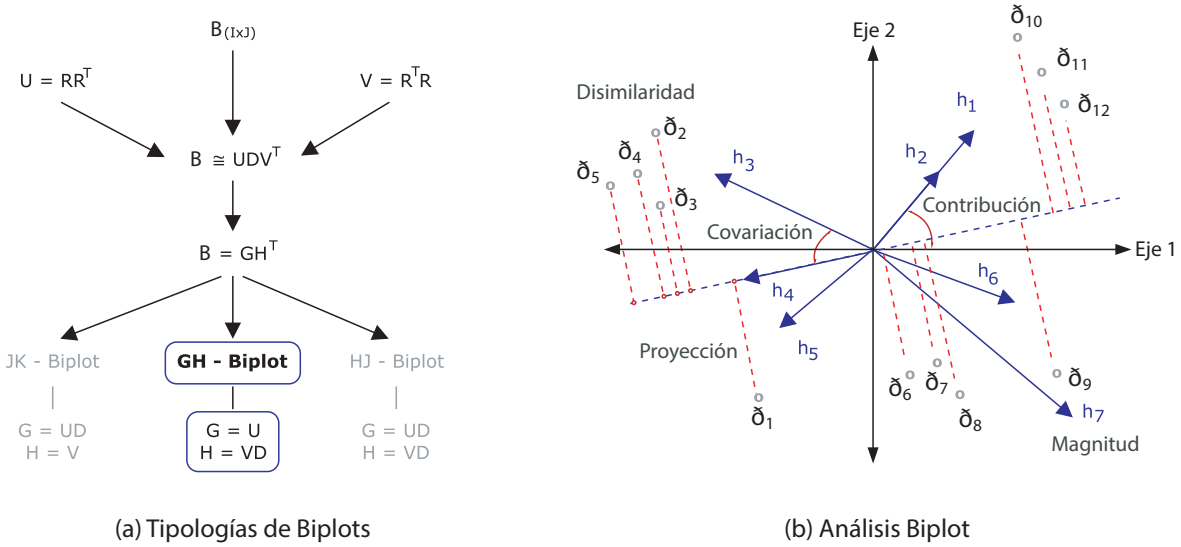
6.5 Análisis multivariante de estructuras

GH – Biplot es un método de análisis multivariante útil para representar e inspeccionar grandes matrices de datos [561]. Esta técnica de análisis es un enfoque alternativo al método original introducido por Gabriel [562] que mejora la calidad de la representación de las observaciones (puntos) y variables (vectores) en el mismo sistema de referencia de dimensiones reducidas. En este trabajo, el análisis **GH – Biplot** se utiliza para detectar estructuras y describir las relaciones subyacentes entre las variables observables seleccionadas $\hat{\mathbf{x}}_j$ y los factores que las explican. Así, el análisis permite sugerir, en función de la estructura de los datos y de la capacidad discriminante de las variables predictoras seleccionadas, la conveniencia de formular un modelo discriminante.

En la Figura (6.11a), la matriz de datos multivariante $\mathbf{B}\,\langle i \times j \rangle$ después de la descomposición de valor singular (SVD) se aproxima a la matriz $\mathbf{B} \cong \mathbf{UDV}^T$, donde, $\mathbf{U}$ y $\mathbf{V}$ son matrices ortogonales, y $\mathbf{D} = (\lambda_1, ..., \lambda_J)$ contiene los valores singulares. En consecuencia, $\mathbf{G}$ es la matriz $\mathbf{U}$, y $\mathbf{H}$ la matriz de las dos primeras columnas del producto $\mathbf{VD}$. Así, esta técnica con la adecuada selección de marcadores, $\eth_i = (\eth_i, ..., \eth_n)$ para las unidades de análisis(ET) y $\mathrm{h_j} = (h_j, ..., h_n)$ para las variables, permite representar simultáneamente en el mismo espacio observaciones (puntos) y variables (vectores).

La Figura (6.11b) muestra la representación gráfica y las reglas de interpretación de un análisis biplot. Aunque el análisis **GH – Biplot** se utiliza específicamente para la detección y estudio de patrones y relaciones de covariación entre variables, este análisis se sirve de las siguientes reglas de interpretación: la longitud de cada vector representa la variabilidad de las medidas; los ángulos formados entre variables determinan el nivel de covariación o correlación; y los ángulos formados entre variables y factores se entienden como la relación o contribución de cada variable al factor. Sin embargo, aunque este tipo de análisis no se centra en el estudio de las relaciones entre observaciones y variables, la distancia entre unidades de análisis (ET) se interpreta como una medida

de similaridad y la proyección ortogonal de cada observación sobre el cuerpo del vector determina el orden de la medida en la matriz original.



(a) Tipologías de Biplots (b) Análisis Biplot

Figura 6.11: Estructura del análisis GH – Biplot.
(a) Tipologías del análisis biplot. (b) Representación gráfica y reglas de interpretación de un análisis biplot. El análisis GH – Biplot es apropiado para la detección de relaciones de covariación entre variables. Los ejes factoriales comparten la variabilidad común de las variables observadas. Los puntos de inflexión (ET) se etiquetan como puntos $\eth_i$, las variables y los índices son los vectores etiquetados como h_j.

6.6 Análisis discriminante

El análisis discriminante es una técnica estadística multivariante introducida por Fisher [458]. Permite modelar y predecir, sobre una variable dependiente categórica, la pertenencia a un grupo en función de un conjunto de variables cuantitativas independientes. La violación de supuestos y la limitación en el cumplimiento de requisitos pueden distorsionar los resultados y las interpretaciones. Para superar estas dificultades, se construye, valida y evalúa la función discriminante lineal a partir de la matriz de datos $\mathbf{Z}\langle i \times j\rangle$ formada por **i** momentos de análisis (en los que cambia la dirección de la tendencia precedente) sobre los que se miden **p** variables predictoras elegidas $\hat{\mathbf{x}}_j = (\hat{x}_1, ..., \hat{x}_p)$. El grupo de pertenencia corresponde al inicio de la dirección del movimiento futuro y se define mediante el vector de etiquetas $\dot{\mathbf{y}} = \{0, 1\}$, donde 0 = Movimiento Tendencial Alcista C_U, y 1 = Movimiento Tendencial Bajista C_D. El vector de datos $\dot{\mathbf{y}}$ se divide en $\breve{y}_g$ grupos con $g = \{1, ...G\}$, donde, $\breve{y}_g$ denota el g^{th} grupo

de tamaño n_g, tal que $\check{y}_g = \{\check{y}_1, \check{y}_2\}$. $\pi_g = \pi_1 = \pi_2 = 1/2$ es la probabilidad a priori de pertenencia de cada caso al grupo $\check{y}_g$. En consecuencia, la función lineal expresada en la ecuación (6.27) es la que mejor discrimina entre movimientos alcistas y bajistas.

$$\mathrm{y}(\hat{X}) = \beta_j^{\top} \hat{\mathrm{X}}_{\mathrm{j}} + \beta_0, \tag{6.27}$$

donde $\mathrm{y}(\hat{x})$ es la puntuación discriminante. β_j es el vector de pesos o coeficientes discriminantes. $\hat{\mathrm{x}}_j$ son las variables independientes seleccionadas y β_0 es una constante. La función (6.27) se puede escribir como la ecuación (6.28)

$$\mathrm{y}(\hat{x}) = \beta_0 + \beta_1 \hat{\mathrm{x}}_1 + \beta_2 \hat{\mathrm{x}}_2. \tag{6.28}$$

Los criterios de clasificación disponibles para la asignación de casos son múltiples, sin embargo, en este trabajo se emplean las funciones de clasificación de Fisher $\mathrm{y}(\hat{x_g})$ derivadas de la ecuación principal (6.28). Para cada grupo g se extrae una función, así $\mathrm{y}(\hat{x}_U)$ y $\mathrm{y}(\hat{x}_D)$ discriminan entre movimientos alcistas $\mathrm{C_U}$ y bajistas $\mathrm{C_D}$, respectivamente. Los coeficientes de estas funciones se obtienen suponiendo normalidad bivariante para el subconjunto de predictores $\hat{\mathbf{x}}_j$, máxima verosimilitud e iguales probabilidades a priori para $\check{y}_g$. El vector de coeficientes de cada función de clasificación viene determinado por la ecuación (6.29).

$$y(\hat{x_g}) = \bar{\hat{x}}_{jg}^{\top} S^{-1} x - \frac{1}{2} \bar{\hat{x}}_{jg}^{\top} S^{-1} \bar{\hat{x}}_{jg} + ln\left(\pi_g\right), \tag{6.29}$$

donde $\bar{\hat{x}}_{jg}^{\top}$ es el vector transpuesto de las medias de las variables predictoras $\hat{\mathbf{x}}_j$ del grupo g. S es la matriz de covarianza intra-grupo combinada, x es el nombre de las variables independientes elegidas, π_g es la probabilidad a priori de pertenencia al grupo g. La clasificación de cada observación i se hace en función de la puntuación discriminante obtenida $\mathrm{y}(\hat{x})$. El modelo asigna el caso al grupo que obtiene la mayor puntuación discriminante. La regla que asigna los casos al grupo de pertenencia se resume a continuación:

Si $\mathrm{y}(\hat{x}_U) \geq \mathrm{y}(\hat{x}_D)$ entonces $\mathrm{y}(\hat{x}_U)$ pertenece a $\mathrm{C_U}$.
En caso contrario, si $\mathrm{y}(\hat{x}_U) < \mathrm{y}(\hat{x}_D)$ entonces $\mathrm{y}(\hat{x}_D)$ pertenece a $\mathrm{C_D}$,

donde $\mathrm{y}(\hat{x}_U)$ es la función lineal alcista, $\mathrm{y}(\hat{x}_D)$ es la función lineal bajista, $\mathrm{C_U}$ y $\mathrm{C_D}$ son las etiquetas de identificación de grupo según movimientos alcistas y bajistas.

Las medidas de bondad de ajuste indican la capacidad del modelo para ajustarse al conjunto de datos. El análisis de estos resultados puede revelar discrepancias entre

valores observados y esperados por el modelo. La varianza explicada, la correlación canónica y la prueba de funciones son las medidas de bondad de ajuste más utilizadas. La varianza total explicada se debe a los primeros valores propios asociados a cada una de las funciones discriminantes extraídas. Así, las m funciones discriminantes $y_i = (y_1, ..., y_m)$, donde $m = Min\,(g-1, \hat{x_j})$ son linealmente independientes y llevan asociado un valor propio λ que indica la proporción de la varianza total explicada por esa función. Para una única función discriminante $m = 1$, sólo hay un valor propio distinto de cero λ_1. La proporción de varianza explicada por la función y_i con respecto a la varianza total explicada por las funciones extraíbles y_m esta determinada por la expresión $\varpi(y_i)$ de la ecuación (6.30)

$$\varpi(y_i) = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i}, \tag{6.30}$$

obsérvese que en el denominador de la ecuación (6.30) está la suma de todos los valores propios λ_i, correspondientes a la varianza total explicada por todas las funciones discriminantes y_m. En el numerador se encuentra el valor propio λ_i asociado a la función discriminante de análisis y_i.

La correlación canónica es una medida de la eficacia del poder discriminante de una función y proporciona información valiosa cuando la diferenciación se hace entre dos grupos. Una vez más, la efectividad de la función se conoce a partir de los valores propios extraídos. La correlación canónica mide el porcentaje de la varianza total que en la i-ésima función y_i está siendo explicada por la diferencia entre grupos. En consecuencia, la i-ésima función y_i es más discriminante cuanto más cercana a uno sea la correlación canónica. Este valor se obtiene a partir de (6.31)

$$C\,(y_i) = \sqrt{\frac{\lambda_i}{1+\lambda_i}}, \tag{6.31}$$

donde, $C\,(y_i)$ es la correlación canónica de la función discriminante y_i, λ_i es el autovalor asociado a la función discriminante y_i.

Finalmente, la prueba de funciones evalúa en el modelo construido la capacidad de diferenciación entre grupos. La significación global de la función discriminante se obtiene mediante el estadístico de contraste Lambda de Wilks de la ecuación (6.32). Donde λ_i es el autovalor de la función discriminante lineal. La función lineal discrimina

más entre grupos cuando el estadístico Lambda de Wilks está más próximo a cero.

$$\Lambda = \frac{1}{1 + \lambda_i}. \tag{6.32}$$

6.7 Evaluación de desempeño del modelo

En esta fase se evalúa el poder predictivo del modelo de clasificación. Luego de entrenar la función discriminante lineal, el modelo se utiliza para clasificar la dirección del movimiento futuro del tipo de cambio euro-dólar. Sustituyendo las medidas de las variables predictoras $\hat{\mathbf{x}}_j$ en la función discriminante $y(\hat{\mathbf{x}})$, se predice, a corto plazo, la dirección que tomará el movimiento del tipo de cambio euro-dólar. El desempeño del modelo se evalúa utilizando cuatro muestras adicionales, OOS_1, OOS_2, OOS_3 y OOS_4, que capturan información de diferentes condiciones de mercado y horizontes temporales en un marco temporal de 15 minutos. Los índices de rendimiento utilizados para medir el poder predictivo del modelo se describen con más detalle en la sección (7.2). El uso de medidas de rendimiento y de datos adicionales permiten una evaluación objetiva y comparativa del modelo con otros enfoques en términos de generalización, precisión e interpretabilidad.

7. Configuración de experimentos

Teniendo en cuenta la naturaleza de los datos y su disposición asimétrica y leptocúrtica, la metodología propuesta está diseñada para reducir el sesgo natural de los datos y dar fiabilidad a las interpretaciones. Para ello, se extrae y prepara un subconjunto de los datos. Los casos son independientes, los predictores y los grupos asumen una distribución normal multivariante, las matrices de varianza-covarianza intragrupo son iguales y la pertenencia de cada caso a un grupo determinado es excluyente.

La identificación de los determinantes del movimiento direccional del tipo de cambio euro-dólar debido al cambio de "régimen de mercado" se aborda como un problema de selección de características observables [563] y la distinción entre grupos se realiza con una única variable categórica compuesta por dos clases (tendencia alcista y tendencia bajista).

La predicción de la dirección del tipo de cambio se aborda como un problema de clasificación de aprendizaje supervisado. La formulación del modelo de clasificación se realiza a partir de los datos muestreados y las variables seleccionadas. La evaluación del rendimiento del modelo de clasificación se realiza con cuatro muestras de datos adicionales. Por lo tanto, para validar la eficacia de los resultados obtenidos en todos los experimentos, la metodología propuesta utiliza un enfoque interpretativo basado en la potencia estadística de las pruebas paramétricas y la fiabilidad de los resultados.

Durante la fase experimental, seis consideraciones técnicas clave influyen significativamente en la calidad de los resultados obtenidos con la metodología propuesta: (i) Garantizar la calidad de los datos manteniendo la precisión y la coherencia, especialmente en las actividades críticas durante la fase de preparación. (ii) Utilizar variables discriminantes que diferencien eficazmente entre observaciones. (iii) Las categorías de la variable dependiente deben ser relevantes en número, independientes y significativamente discriminantes. (iv) La eficacia del proceso de selección de variables predictoras debe ser demostrable y observable mediante el análisis de estructura GH-Biplot. (v) La integración de estas soluciones parciales en un modelo de clasificación debe reflejarse con valores de alto rendimiento. Por último, (vi) Las pruebas de significación

estadística deben respaldar la generalización, la validez de las interpretaciones y las conclusiones realizadas sobre el modelo.

La biblioteca PyCaret [564] de aprendizaje automático y de código abierto se utilizó en el entorno notebook de Python (3.9.5) [565] para la preparación de datos y la implementación de modelos. Los programas IBM SPSS Statistics for Windows, Versión 27.0. (IBM Corp. Released, 2019) [566] y RStudio Team (2021) [567] se utilizaron para el análisis estadístico. El software MultBiplot [568] se utilizó para el análisis GH-Biplot y la visualización de datos multivariantes. El paquete MVN [569] en código R se utilizó para evaluar los supuestos de normalidad multivariante y univariante y proporcionar la evidencia gráfica. Los experimentos se realizaron en un equipo Intel(R) Xeon(R) Silver 4110 CPU @ 2.10Ghz (2 Procesadores) con 32 Gb en RAM a 64-Bits.

7.1 Entrenamiento, validación y evaluación

La muestra de análisis consta de 125 observaciones divididas en dos subconjuntos, uno de entrenamiento y otro de evaluación. Es importante señalar que el tamaño de estos subconjuntos debe ser suficientemente grande y representativo para que el modelo aprenda, generalice y evalúe con precisión su rendimiento. Sin embargo, no existe una solución estándar única que defina su tamaño específico, ya que puede variar en función del contexto. Para realizar el análisis, los casos se seleccionan aleatoriamente, aproximadamente el 70% de los movimientos alcistas y bajistas se utilizan para crear el modelo (conjunto de entrenamiento) y con los mismos casos se hace la validación cruzada del mismo. Los casos restantes, aproximadamente el 30% de los movimientos tendenciales, se reservan para validar el rendimiento del modelo, garantizando así una evaluación imparcial.

En la evaluación del poder predictivo del modelo se utilizan cuatro conjuntos de datos adicionales. El conjunto OOS_3 consta de 1748 movimientos de tendencia confirmados en un marco temporal semanal a largo plazo, mientras que los conjuntos OOS_1, OOS_2 y OOS_4 constan de 6581, 1645 y 2336 movimientos de tendencia, respectivamente, medidos en un marco temporal de 15 minutos. Estos conjuntos abarcan un horizonte temporal de 24 años desde la introducción del euro, lo que permite evaluar exhaustivamente la capacidad de generalización y la precisión del modelo.

La Tabla 7.1, proporciona los valores de los parámetros utilizados en la realización de los experimentos, incluyendo las especificaciones técnicas de los criterios utilizados en el proceso de selección de características, el análisis discriminante, así como el

tamaño de los conjuntos de datos utilizados en el análisis. En el proceso de selección de características, los valores por defecto asignados para Tolerancia (ψ_i), los niveles de entrada ($F_\varepsilon = 3.84$) y remoción ($F_r = 2.71$), en la evaluación de atributos, son los que producen el mejor rendimiento.

Tabla 7.1: Valores de parámetros

[a]**Parámetros**	**Criterios**	**Descripción**
Selección de características:		
F-Snedecor (*inclusión*)	$F_\varepsilon = 3.84$	Umbral para variable de entrada
F-Snedecor (*remoción*)	$F_\varepsilon = 2.71$	Umbral para remoción de variables
Tolerancia mínima	$\psi = 0.001$	Valor para entrar en el análisis
Análisis discriminante:		
Probabilidades iguales	$\pi_g = 0.50$	Definición coeficientes funciones
Matriz de covarianza	Intra-grupos iguales	Clasificación de casos
Validación cruzada	Dejando uno fuera	Clasificación validación cruzada
Conjuntos ET para el análisis:		
Entrenamiento:	(70%) 87/125 instancias	Casos usados para crear el modelo
Validación cruzada:	(70%) 87/125 instancias	Casos validados de forma cruzada
Evaluación:	(30%) 38/125 instancias	Casos usados para evaluar el modelo
Fuera de muestra (OOS):		
OOS_1: 1999-2005	6581 instancias	Movimientos alcistas y bajistas
OOS_2: 2006-2020	14645 instancias	Movimientos en canal bajista
OOS_3: 2006-2020	1748 instancias	Movimientos en canal bajista
OOS_4: 2021-2023	2336 instancias	Movimientos alcistas y bajistas

[a]Valores de los parámetros utilizados en todos los experimentos. Estas asignaciones han mostrado el mejor rendimiento en términos de precisión durante el entrenamiento, la validación y la evaluación del rendimiento del modelo de clasificación.

7.2 Medidas de desempeño

El desempeño del modelo de clasificación se mide en función de los resultados dados en la Tabla 7.2. La matriz confusión está formada por cuatro categorías. En la diagonal principal se ubican las observaciones clasificadas correctamente por el modelo. Están los Verdaderos Positivos T_P que es el número de observaciones clasificadas correctamente como movimientos alcistas y los Verdaderos negativos T_N que es el número de casos clasificados correctamente como movimientos bajistas. En la diagonal opuesta están los casos clasificados incorrectamente por el modelo. Los Falsos Positivos F_P es el número de movimientos bajistas incorrectamente clasificados como alcistas y los Falsos Negativos F_N es el número de movimientos alcistas que se clasifican erróneamente

como bajistas.

Tabla 7.2: Matriz confusión.

Predicciones / Actuales	**Positivos Actuales**	**Negativos Actuales**	**Total**
Predicción Positiva (PP)	Verdaderos Positivos (T_P)	Falsos Positivos (F_P)	PP = T_P + F_P
Predicción Negativa (PN)	Falsos Negativos (F_N)	Verdaderos Negativos (T_N)	PN = F_N + T_N
Total	PA = T_P + F_N	NA = F_P + T_N	PA + NA = PP + PN

Si bien no existe un consenso general sobre el uso eficaz de una medida de rendimiento concreta, *Exactitud* es una de las métricas más utilizadas para evaluar el rendimiento de los modelos a la hora de clasificar nuevas observaciones. La ecuación (7.1) mide la proporción de observaciones que el modelo clasifica correctamente:

$$Exactitud = \frac{T_P + T_N}{F_P + T_P + T_N + F_N}. \tag{7.1}$$

La media geométrica $G-mean$ es una medida de desempeño que se emplea cuando los grupos están desequilibrados [570]. La ecuación (7.4) mide la media geométrica del ratio de aciertos positivos ($Sensibilidad$) y el ratio de aciertos negativos ($Especificidad$). Sensibilidad y Especificidad se miden a partir de las ecuaciones (7.2) y (7.3) respectivamente.

$$Sensibilidad = \frac{T_P}{T_P + F_N}, \tag{7.2}$$

$$Especificidad = \frac{T_N}{F_P + T_N}, \tag{7.3}$$

$$Exactitud \quad G - mean = \sqrt{Sensibilidad \cdot Especificidad}. \tag{7.4}$$

El valor predictivo positivo, también denominado Precisión, es el ratio que se computa entre el número de casos clasificados como verdaderos positivos y el número total de casos positivos. La ecuación (7.5) define su cálculo.

$$\text{Precisión} = \frac{T_P}{F_P + T_P}. \tag{7.5}$$

Puntuación F_1 es la media armónica de Sensibilidad y Precisión. Esta medida es útil para comparar la calidad de la predicción entre modelos y puede ser interpretada como una función de recuentos de verdaderos positivos, falsos negativos y falsos positivos [571]. La ecuación (7.6) muestra la forma de cálculo.

$$F_1 = \frac{2 \cdot \mathit{Sensibilidad} \cdot \text{Precisión}}{\mathit{Sensibilidad} + \text{Precisión}}. \tag{7.6}$$

El estadístico *Kappa de Cohen* se calcula a partir de la ecuación (7.7). Es útil en la construcción de modelos y sirve para saber si el modelo formulado es mejor, igual o peor que un modelo aleatorio [572]. Note que el estadístico Kappa se computa en función de una tasa denominada Línea base. Este índice define un punto de referencia para la comparación de modelos y se calcula con la expresión (7.8).

$$\mathit{Kappa} = \frac{\mathit{Exactitud} - \text{Línea base}}{1 - \text{Línea base}}, \tag{7.7}$$

donde:

$$\text{Línea base} = \frac{(PP \cdot AP)}{(PP + PN)^2} + \frac{(PN \cdot AN)}{(AP + AN)^2}. \tag{7.8}$$

Coeficiente de correlación de Matthews (MCC) es un índice estadístico robusto que proporciona una medida más confiable del rendimiento del modelo de clasificación. Produce puntuaciones más altas sólo cuando las predicciones son correctas en las cuatro categorías de la matriz confusión [544]. Se calcula como se muestra en la ecuación (7.9):

$$MCC = \frac{T_P \cdot T_N - F_P \cdot F_N}{\sqrt{(T_P + F_P) \cdot (T_P + F_N) \cdot (T_N + F_P) \cdot (T_N + F_N)}} \tag{7.9}$$

Parte IV

Discusión de resultados

8. Resultados y discusión

En este capítulo se presentan y analizan los resultados de la metodología descrita en la Figura 6.1. Se abordan cuatro áreas concretas: selección de atributos con alto poder discriminante e independientes (Sección 8.1), análisis multidimensional de las diferencias entre los movimientos de tendencia (Sección 8.2), construcción, validación y evaluación de un modelo de clasificación para predecir a corto plazo la tendencia del par euro-dólar (Sección 8.3), y evaluación del rendimiento del modelo utilizando datos fuera de muestra y comparación con otros enfoques de previsión (Sección 8.4).

8.1 Selección de características

El proceso de selección de características desempeña un papel crucial en la construcción de modelos de clasificación interpretables y precisos. En la Tabla 5.3 se resumen las variables e índices que capturan la información previa al cambio de régimen de mercado en los puntos de inflexión (`ET`). En cuanto a las 15 primeras métricas, que son objeto de análisis en esta sección, no se conocen a priori las variables más eficaces para la construcción de modelos de clasificación. No todas son útiles y significativas en el proceso de discriminación de la dirección de la tendencia del tipo de cambio euro-dólar. La importancia de las variables no solo radica en la significación estadística que cada una aporta en la diferenciación entre tendencias sino en su precisión y capacidad explicativa para la interpretación del problema de estudio. Teniendo en cuenta este contexto, la identificación y selección de las variables predictoras que mejor contribuyen a la diferenciación entre movimientos tendenciales alcistas y bajistas se realiza en tres pasos, utilizando la matriz de datos multivariantes $\mathbf{A}$.

8.1.1 Evaluación de la contribución de cada predictor

La prueba de igualdad de medias de grupo define el potencial que cada variable predictora tiene en la diferenciación de grupos antes de participar en la formulación del modelo. La evaluación de la contribución que cada variable aporta en la diferenciación

de grupos permite reconocer las variables predictoras que deben ser consideradas en el estudio. La prueba de igualdad de medias entre grupos referida en la Tabla 8.1 contrasta la H_0 sobre igualdad de medias entre grupos. Según análisis ANOVA de una vía y con un valor de significación superior al 5%, se acepta la H_0 sobre igualdad de medias entre grupos. Las variables resaltadas en gris no tienen el potencial para discriminar entre movimientos tendenciales alcistas y bajistas, y por tanto, no deben ser consideradas en el estudio.

Tabla 8.1: Análisis ANOVA de una vía

Medida	Suma de Cuadrados	[a]Df		Media Cuadrática	Estadístico F	P-Valor
		Df_1	Df_2			
x.V*PT	0.00	1	14768	0.00	16856.79	0.000
Sp.Vr*CP.PT	0.00	1	14768	0.00	5001.01	0.000
s.Vr.Vo.PT	1.80	1	14768	1.80	1.17	**0.279**
x.Vr.Vo.PT	0.07	1	14768	0.07	0.67	**0.412**
Sp.Vr.Vo.PT	0.00	1	14768	0.00	0.05	**0.830**
Sp.Vo.PT	24.89	1	14768	24.89	0.01	**0.905**
s.V.PT	0.00	1	14768	0.00	5.50	0.019
Sp.CP.PT	0.00	1	14768	0.00	11645.00	0.000
Sp.Vr.HL.PT	0.00	1	14768	0.00	0.16	**0.688**
Sp.Vr.CO.PT	0.04	1	14768	0.04	0.19	**0.667**
s.Vr.RSI.PT	10.41	1	14768	10.41	1962.97	0.000
Sp.Vr.RSI.PT	0.09	1	14768	0.09	1500.26	0.000
Dv.CP-RSI.PT	0.06	1	14768	0.06	1.38	**0.241**
Dv.Vo-CP.PT	872.86	1	14768	872.86	4266.50	0.000
Dv.Vo-RSI.PT	1129.54	1	14768	1129.54	6808.76	0.000

[a]**Df**= Grados de libertad.

8.1.2 Evaluación del poder discriminante de cada predictor

La búsqueda de un modelo de predicción preciso, interpretable y parsimonioso conduce a la identificación y selección de los predictores que mejor explican la acción de los precios y, en consecuencia, mejor contribuyen a la separación entre grupos. El método de selección de variables por pasos es útil para lograr este propósito y garantiza así la disponibilidad de las mejores variables candidatas que se utilizarán en la formulación del modelo.

El estadístico Lambda de Wilks es una herramienta fundamental en la identificación de variables con alto poder discriminante. Este estadístico permite detectar aquellas características que presentan diferencias significativas entre grupos, siendo muy útil en el proceso de selección de variables en la construcción de modelos de clasificación. La Tabla 8.2 resume las estadísticas para el subconjunto de variables candidatas significa-

tivas que potencialmente podrían incorporarse en un único modelo y, que en conjunto, son las variables predictoras que mejor discriminan entre movimientos alcistas y bajistas.

En la tabla 8.2, se sigue un proceso de selección de variables por pasos. En cada paso, se añade al modelo el predictor con el mayor valor F de entrada que supera el criterio de entrada (por defecto, 3.84). Así, las variables con los valores más pequeños del estadístico Lambda de Wilks obtenidos en cada paso indican una mayor contribución al proceso de separación entre grupos, lo que sugiere que forman parte del subconjunto de predictores que mejor contribuye a la diferenciación entre grupos. En el último paso, las variables que quedan fuera del análisis tienen todas valores F de entrada inferiores a 3.84, lo que indica que no cumplen con el criterio de inclusión en el subconjunto.

Tabla 8.2: Poder discriminante de variables predictoras

Paso	Atributo Elegible	Lambda de Wilks	[a] Df			Valor F	Df		P-Valor
			Df_1	Df_2	Df_3		Df_1	Df_2	
1	x.V*PT	0.467	1	1	14768	16856.787	1	14768	0.000
2	Sp.Vr.RSI.PT	0.365	2	1	14768	12851.92	2	14767	0.000
3	Dv.Vo-RSI.PT	0.337	3	1	14768	9666.854	3	14766	0.000
4	s.Vr.RSI.PT	0.327	4	1	14768	7590.954	4	14765	0.000
5	Sp.Vr*CP.PT	0.325	5	1	14768	6137.861	5	14764	0.000
6	Sp.CP.PT	0.324	6	1	14768	5142.587	6	14763	0.000
7	Dv.Vo-CP.PT	**0.323**	7	1	14768	4410.364	7	14762	0.000
8	s.V.PT	**0.323**	8	1	14768	3861.226	8	14761	0.000

[a] **Df**= Grados de libertad.

Según resultados de la Tabla 8.2, s.V.PT no minimiza el valor del estadístico Lambda de Wilks obtenido en el paso previo (7), en consecuencia esta variable (8) no aporta información al modelo y por tanto se excluye de los análisis posteriores. El potencial que cada predictor tiene en la diferenciación de clases y la capacidad para lograr un alto nivel de desempeño en la clasificación se miden a través del estadístico Lambda de Wilks. La Figura 8.1 muestra en orden de importancia el poder discriminante de las mejores variables predictoras.

Las variables x.V*PT, Sp.Vr*CP.PT y Sp.CP.PT son variables que capturan información relevante sobre la acción del precio antes del cambio de régimen de mercado. La segunda, tercera y cuarta variable miden el comportamiento del mercado en función del indicador técnico $RSI(9P)$.

Por otro lado, la variable Dv.Vo-CP.PT mide la divergencia entre el precio de cierre y el volumen, que representa la frecuencia de los cambios en el precio del par euro-dólar

por cada sesión de 15 minutos.

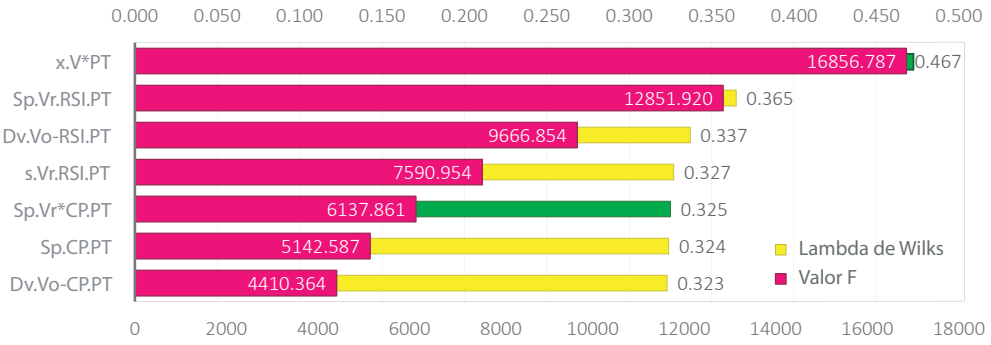


Figura 8.1: Poder discriminante de variables predictoras.
Orden de importancia de las variables predictoras según su capacidad discriminante. El estadístico Lambda de Wilks determina el potencial discriminante que cada predictor tiene en la diferenciación entre movimientos alcistas y bajistas. Los predictores con los valores más altos del Estadístico F de entrada superan el criterio de entrada (por defecto, 3.84) y son adecuados en la diferenciación entre medias de clases.

La prueba de homogeneidad de varianzas de Levene se utiliza para determinar si los grupos de estudio tienen varianzas iguales. Esta prueba es un requisito previo para asegurar la igualdad de varianzas antes de realizar la prueba de igualdad de medias (ANOVA de una vía).

En la Tabla 8.3 se comprueba el cumplimiento del supuesto de homogeneidad de varianzas mediante el estadístico de Levene. Al rechazar la hipótesis nula Ho de igualdad de medias para los predictores 1, 2, y 6 se confirma que las medias de los grupos difieren. Por lo tanto, ahora que se sabe que las medias de los grupos difieren en el caso de los predictores resaltados en gris, es posible, mediante un análisis multivariante, acercarse a la comprensión de las diferencias en la acción de los precios generadas por los cambios en el régimen de mercado.

Tabla 8.3: Prueba de homogeneidad de varianzas

ID	Medida	Estadístico de Levene	Grados de libertad		P-Valor
			Df_1	Df_2	
1	x.V*PT	0.578	1	14768	**0.446**
2	Sp.Vr*CP.PT	1.820	1	14768	**0.177**
3	Sp.Vr.RSI.PT	439.873	1	14768	0.000
4	Dv.Vo-RSI.PT	16.092	1	14768	0.000
5	s.Vr.RSI.PT	188.975	1	14768	0.000
6	Sp.CP.PT	3.165	1	14768	**0.075**
7	Dv.Vo-CP.PT	9.311	1	14768	0.002

8.1.3 Análisis de independencia entre variables predictoras

El análisis de independencia en la selección de variables predictoras es crucial en la construcción de modelos de clasificación. El uso de variables no independientes puede tener un efecto negativo en la estabilidad de los coeficientes de los modelos y por tanto puede afectar los resultados y la interpretación de los mismos. Según los resultados presentados en la Tabla 8.4, existe una alta correlación entre las variables Sp.CP.PT y x.V*PT. La correlación del 95.6% es significativa en el nivel del 1% bilateral. Si todas las variables se utilizan en la formulación del modelo, la inestabilidad en los signos de los coeficientes de las funciones discriminantes lineales de Fisher podrían afectar negativamente los resultados de la clasificación. Por lo tanto, para evitar el problema de multicolinealidad, se excluye de la formulación del modelo la variable Sp.CP.PT.

Tabla 8.4: Prueba de independencia de variables predictoras

Medidas	Sp.Vr*CP.PT	x.V*PT	Sp.CP.PT
Sp.Vr*CP.PT	1	0.362	0.312
x.V*PT		1	**0.956**
Sp.CP.PT			1

En resumen, según metodología propuesta, de 15 variables candidatas, las variables predictoras e independientes seleccionadas son dos: Rendimiento medio de los precios de cierre x.V*PT y Pendiente de la linea de regresión de las variaciones entre precios de cierre Sp.Vr*CP.PT, ambas calculadas sobre la base de los últimos 18 periodos antes del cambio de régimen de mercado en el momento `ET`. Según mérito estadístico, el subconjunto de variables elegidas $\hat{\mathbf{x}}_j$ son las que mejor contribuyen a la separación de grupos y tienen el mayor poder discriminante y, desde el punto de vista del análisis técnico, son las que mejor explican la acción del precio antes de que ocurra un cambio de régimen de mercado en el instante `ET`.

8.2 Análisis Multivariante

El propósito del análisis **GH − Biplot** es detectar y describir, en las estructuras subyacentes de los datos, la naturaleza multidimensional de las diferencias entre el comienzo de un movimiento direccional alcista y bajista en función de las variables predictoras elegidas $\hat{\mathbf{x}}_j$. Los resultados del análisis permiten confirmar la idoneidad de las vari-

ables predictoras seleccionadas y la pertinencia del análisis discriminante en las tareas de clasificación.

El análisis se realiza sobre la colección de datos $\mathbf{B}\langle 14770 \times 7\rangle$ estructurada según el proceso de preparación de datos abordado en la sección (6.3.3). Esta matriz consta de 14770 movimientos tendenciales sobre los que se miden 7 variables numéricas observables y una variable categórica que etiqueta la dirección de la tendencia del tipo de cambio euro-dólar en dos grupos $\tilde{y}_g$. Los grupos, según el vector de etiquetas $\dot{\mathsf{y}} = \{0, 1\}$, están identificados con las categorías 0 = "Movimiento Tendencial Alcista" y 1 = "Movimiento Tendencial Bajista".

La efectividad del análisis $\mathbf{GH - Biplot}$, en la detección de estructuras subyacentes, esta determinada por la suficiencia de los datos. La pertinencia de la matriz de datos multivariantes $\mathbf{B}\langle 14770 \times 7\rangle$ para la detección de estructuras se verifica con las medidas de adecuación del muestreo de Kaiser-Meyer-Olkin KMO y la prueba de esfericidad de Bartlett. Los resultados de estas pruebas estadísticas sugieren que el análisis es procedente. El valor KMO de 0.639 cercano a 1, indica que la proporción de variación de las variables utilizadas en el análisis son causadas por factores subyacentes. El valor Chi-cuadrado $\tilde{\chi}^2 = 47208.373$ asociado al estadístico de prueba KMO y un P-Valor inferior al 5% sugiere el rechazo de la hipótesis nula H_0 de que la matriz de correlación es una matriz de identidad ($\tilde{\chi}^2 = 47208.373$, Df = 21, P-Value = 0.000). De ahí que, la covariación entre variables es adecuada para la detección de estructuras subyacentes.

La interpretación de la solución propuesta se hace sobre los dos primeros factores que explican el 55.012% de la variabilidad acumulada de las variables analizadas. Esto sugiere dos influencias latentes asociadas con el cambio de dirección de las tendencias precedentes, con un importante margen de variación no explicado. Por consiguiente, la proporción de varianza absorbida es suficiente para explicar la relación subyacente entre el comportamiento histórico del tipo de cambio y el beneficio bruto realizable, cuando se conoce a priori la dirección de la tendencia futura.

Las variables manifiestas que mejor explican a los ejes factoriales se resumen en la Tabla 8.5. Las contribuciones relativas del factor a los elementos columna pueden interpretarse como una medida de bondad de ajuste con la que las variables observadas mejor explican a los ejes factoriales. La fiabilidad del análisis y del modelo de clasificación sugerido se verifica con la precisión lograda en el proceso de clasificación.

En la Tabla 8.5 se detalla, por grupos, las variables implicadas en el análisis. El primer grupo mide el comportamiento de los precios de cierre que preceden al cambio

de tendencia en el momento ET. Este grupo de variables 1, 2, 4 y 5 son las que mejor discriminan entre direcciones de movimientos alcistas y bajistas. La solución propuesta, por el método de selección de características, la conforman las variables 1 y 4 resaltadas en color gris. El segundo grupo está formado por las variables 3, 6 y 7 que validan la dirección de la tendencia predicha, luego del momento ET. Otros parámetros disponibles en la Tabla 5.4 no participan en el análisis porque no ayudan a explicar la naturaleza multidimensional de las diferencias entre el comienzo de un movimiento alcista y bajista.

Tabla 8.5: Contribuciones relativas del factor a los elementos columna

ID	Variables	Grupo 1	Grupo 2	Eje 1	Eje 2
1	x.V*PT	1		**875**	2
2	Sp.CP.PT	1		**831**	5
3	Sp.XP-EP		2	411	27
4	Sp.Vr*CP.PT	1		301	75
5	s.Vr.RSI.PT	1		221	119
6	RP%M[1]		2	0	**461**
7	GPL$		2	0	**523**

La Figura 8.2 muestra la representación **GH – Biplot** de la colección de datos **B**. Los marcadores columna o vectores representan las medidas de las variables (parámetros técnicos previos y posteriores al comienzo de cada micro tendencia luego del momento ET). Los marcadores fila representados por puntos definen, en términos de valores, el momento previo al cambio de tendencia ET. La escala de medida de los vectores define la magnitud de la varianza de las mediciones. Los ángulos formados entre vectores representan las correlaciones entre variables. El centro del **GH – Biplot** es el punto medio de las diferencias entre movimientos alcistas y bajistas.

En consecuencia, dos factores latentes se identifican y están asociados con la naturaleza multivariante de las diferencias entre grupos. El primer eje factorial (Eje 1) describe el comportamiento histórico del tipo de cambio que precede al cambio de tendencia. Este eje está altamente correlacionado con las variables: retorno medio x.V*PT y tendencia de los precios de cierre Sp.CP.PT. En el primer factor, existe una mayor variabilidad en las mediciones del lado izquierdo que en el origen del **GH – Biplot**. Estas variaciones negativas, entre precios de cierre, deprecian al tipo de cambio. Así, la acumulación sostenida de rendimientos medios negativos es un indicador que precede al comienzo de un nuevo movimiento al alza. Este efecto experimentado por la depreciación del euro respecto al dólar americano se destaca durante el periodo 2006 - 2020.

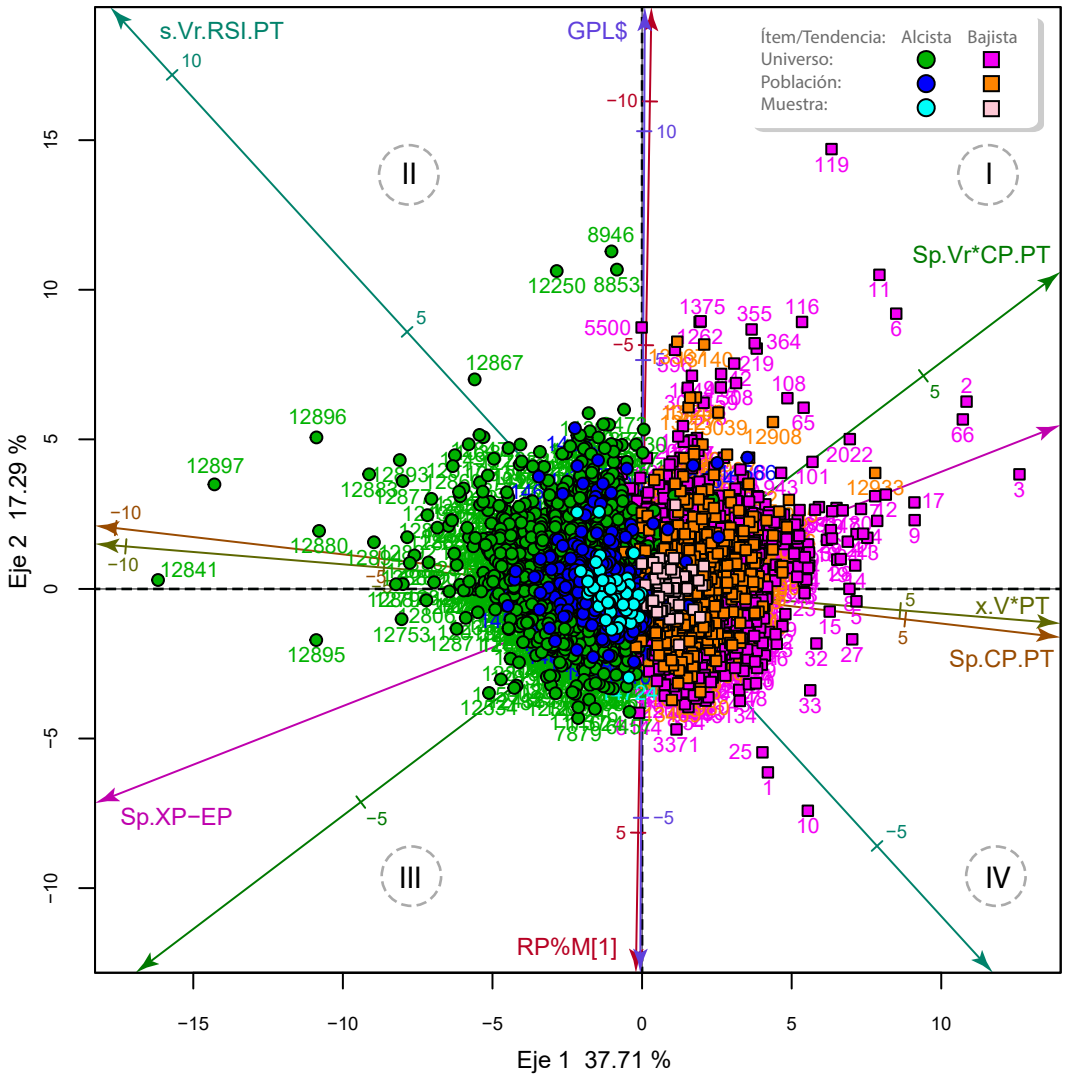


Figura 8.2: GH – Biplot del movimiento direccional del tipo de cambio euro-dólar. Datos normalizados estandarizados por columna y particionados según *SVD*. Plano factorial 1-2. Varianza explicada del 55.012 % (37.71 % del eje 1 y 17.29 % del eje 2). Análisis de la estructura de los datos y perfiles de los puntos de inflexión `ET`. Detección de relaciones subyacentes entre variables (vectores) y diferencias entre movimientos de tendencia. Poder discriminante de variables predictoras en la construcción de modelos de clasificación.

Hacia el lado derecho del primer factor la variabilidad de las mediciones es menor. La apreciación sostenida del tipo de cambio es un efecto de las variaciones positivas, entre precios de cierre y, es un indicador que precede al comienzo de un nuevo movimiento bajista. En consecuencia, el primer eje factorial es el que mejor discrimina entre las medias de los grupos. De acuerdo con los ángulos de covariación formados entre el

factor y las variables que lo explican, el rendimiento medio x.V*PT es mejor predictor. Teniendo en cuenta la alta correlación entre el primer eje factorial y la tendencia de los precios de cierre Sp.CP.PT, la inclusión de esta variable en la construcción de modelos de clasificación puede generar problemas de multicolinealidad.

El segundo eje factorial (Eje 2) describe el premio al riesgo. Este eje es explicado por las variables: Riesgo de mercado RP%M[1] y Ganancia o Perdida Bruta GPL$. Ambos vectores puntúan en sentido contrario. Teóricamente, los beneficios brutos sujetos a riesgo se consiguen, con la apertura y cierre de comercios, en función del comienzo y fin de cada micro-tendencia. Así, en el segundo eje, la variabilidad del beneficio obtenido en función de la exposición al riesgo es mayor por encima del punto medio del biplot que por debajo.

Las variables de plano: Pendiente de las variaciones entre precios de cierre definida por Sp.Vr*CP.PT y medida antes del momento `ET`, y la pendiente de la micro-tendencia Sp.XP-EP medida posterior al evento `ET`, entre ambos vectores, registran una co-variación inversa. De manera que, las pendientes negativas de las variaciones entre precios de cierre preceden a la formación de micro-tendencias alcistas (Puntos en cuadrantes II y III) y las pendientes positivas de las variaciones entre precios de cierre preceden a la formación de micro-tendencias bajistas (Puntos en cuadrantes I y IV). La independencia entre Sp.Vr*CP.PT y s.Vr.RSI.PT se confirma por el ángulo de 90° que forman entre ellas, note que esta característica no es suficiente para discriminar entre movimientos alcistas y bajistas.

En general, tomados en conjunto, estos resultados sugieren la formación de cuatro cuadrantes que explican la naturaleza multivariante de las diferencias entre movimientos alcistas y bajistas. Los cuadrante II y III se pueden interpretar como la zona de sobre-venta, en donde cada marcador fila representa, respecto al primer eje factorial, la depreciación mas baja, experimentada por el tipo de cambio, previo al cambio de tendencia. Los cuadrantes I y IV pueden entenderse como la zona de sobre compra, en donde las observaciones representan, respecto al primer eje factorial, la apreciación mas alta alcanzada, previa al cambio de tendencia. Los cuadrantes I y II se pueden interpretar como zona de exposición de alto riesgo, en donde cada marcador fila corresponde, respecto al segundo eje factorial, a Micro tendencias generadoras de beneficios brutos positivos con exposición a perdidas flotantes antes de alcanzar el precio de cierre objetivo. Los cuadrantes III y IV pueden entenderse como zona de exposición de bajo riesgo, en donde cada marcador fila corresponde, respecto al segundo eje factorial, a micro tendencias generadoras de beneficios brutos positivos con exposición a ganancias

flotantes antes de alcanzar el precio de cierre objetivo. La evidencia de este estudio sugiere que el primer eje factorial es el que mejor explica las diferencias entre las medias de los movimientos al alza y a la baja. Dado el patrón de correlación con las variables observables, la variable predictora x.V*PT es la que mejor contribuye en la diferenciación entre el comienzo de un nuevo movimiento tendencial alcista y bajista. Este predictor es el mas influyente para considerar en la construcción de un modelo de clasificación. Por lo tanto, una función discriminante lineal es suficiente para interpretar, clasificar y predecir con precisión el inicio de un movimiento tendencial, como demuestran los elevados valores de precisión obtenidos en las tareas de clasificación (véase la sección 8.3.4).

8.3 Análisis discriminante lineal (LDA)

Para explorar las ventajas de las soluciones aportadas por la metodología propuesta, en términos de datos, predictores y estructuras, este trabajo evalúa el rendimiento del modelo de clasificación basándose en tres medidas establecidas: (**i**) muestra con normalidad multivariante para superar la naturaleza asimétrica y leptocúrtica de los datos; (**ii**) variables predictoras elegidas según mérito técnico y estadístico; (**iii**) idoneidad del análisis discriminante respecto a la naturaleza multidimensional de las diferencias entre movimientos alcistas y bajistas. Para ello, esta sección presenta resultados en cinco áreas específicas. 1) Estadística descriptiva para el subconjunto de estudio (matriz de datos $\mathbf{Z}$). 2) Verificación de los supuestos y requisitos del análisis discriminante. 3) Evaluación del ajuste del modelo. 4) Clasificación de la dirección de los movimientos del tipo de cambio euro-dólar y 5) Validación del modelo. El análisis discriminante se realiza sobre el conjunto de datos $\mathbf{Z}\,\langle 125 \times 2\rangle$ utilizando dos variables independientes $\hat{\mathbf{x}}_j$ que recogen información sobre la acción del precio. Estos predictores se utilizan para clasificar las observaciones $\mathbf{i}$ en dos grupos $\check{y}_g$. Las variables predictoras $\hat{\mathbf{x}}_j$ son: Rendimiento medio de los precios de cierre x.V*PT y pendiente de la línea de regresión de las variaciones entre precios de cierre Sp.Vr*CP.PT, ambos predictores se calculan 18 periodos previos al momento de estudio (`ET`), en el que ocurre el cambio de dirección de la tendencia precedente. Los grupos, según el vector de etiquetas $\dot{\mathbf{y}} = \{0, 1\}$, son movimientos tendenciales descritos por el tipo de cambio euro-dólar, donde 0 = "Movimiento Tendencial Alcista" y 1 = "Movimiento Tendencial Bajista". Estas etiquetas de clase identifican el comienzo de un nuevo movimiento direccional de corto plazo (medido en un marco de tiempo de 15 minutos) confirmado en un marco de

tiempo de largo plazo (1 semana).

8.3.1 Estadísticas descriptivas

Las estadísticas descriptivas previas al comienzo del evento (ET), en donde ocurre el cambio de dirección en relación con el inicio de un nuevo movimiento direccional, se informan en la Tabla 8.6. En promedio, el rendimiento medio x.V*PT registra, durante el periodo de análisis (2006-2020), una ligera depreciación del euro respecto al dólar americano con una alta variación en los rendimientos medios. La pendiente de las variaciones entre precios de cierre Sp.Vr*CP.PT que preceden a la formación de un nuevo régimen de mercado (alcista/bajista) reporta en promedio una ligera tendencia positiva con mucha variación en las mediciones.

Tabla 8.6: Estadísticas descriptivas de variables y grupos. Valores previos al momento ET

Medida	n	Media	[a]DS	Mediana	Mínimo	Máximo	Asimetría	Curtosis
Variables predictoras:								
x.V*PT	125	-0.0000024	0.0001211	0.0000017	-0.0002938	0.000263	-0.0880643	-0.88373
Sp.Vr*CP.PT	125	0.0000011	0.0000204	0.0000013	-0.0000482	0.0000461	-0.021371	-0.3078016
Movimiento alcista:								
x.V*PT	62	-0.0001056	0.0000676	-0.0000967	-0.0002938	0.0000006	-0.3798269	-0.4650322
Sp.Vr*CP.PT	62	-0.000012	0.000016	-0.0000122	-0.0000482	0.0000286	-0.0526824	-0.0757937
Movimiento bajista:								
x.V*PT	63	0.0000992	0.0000607	0.0000929	0.0000017	0.000263	0.4674263	-0.4223248
Sp.Vr*CP.PT	63	0.0000139	0.0000154	0.0000124	-0.0000207	0.0000461	0.1728039	-0.5063708

[a]DS = Desviación Estándar.

Al analizar los valores medios de los grupos, la media de los rendimientos medios x.V*PT que preceden a la formación de una tendencia alcista reporta en promedio una perdida de valor (depreciación) ligeramente mayor con relación a su proceso de apreciación, y una alta variación entre precios de cierre *cp*. La pendiente de las variaciones entre precios de cierre Sp.Vr*CP.PT que preceden al comienzo de un movimiento direccional alcista/bajista reporta en promedio una ligera tendencia negativa/positiva con una alta variación en las mediciones. Aunque los rendimientos medios x.V*PT que preceden a la formación de movimientos alcistas/bajistas presentan una ligera asimetría negativa/positiva, este efecto asimétrico y leptocúrtico es común en los rendimientos generados por activos financieros [573]. Sin embargo, como se describe en la sección B., durante la preparación de la muestra de estudio, el subconjunto de datos **Z** tiene una

distribución normal multivariante, superando los problemas de asimetría y curtosis. Además, los resultados de las pruebas estadísticas exhaustivas confirman el cumplimiento de los supuestos estadísticos exigidos por el análisis discriminante, como se explica en la siguiente sección.

8.3.2 Validación de supuestos

El supuesto de normalidad multivariante es uno de los requisitos mas importantes que requieren algunos procedimientos estadísticos paramétricos multivariantes para garantizar la fiabilidad de las interpretaciones. Si los contrastes de significación y la evaluación de la bondad de ajuste no se satisfacen, afectara negativamente la confiabilidad de las interpretaciones basadas en los resultados de esos procedimientos [429, 574]. A continuación se presenta los resultados de las pruebas que validan el cumplimiento de estos requisitos.

A. Pruebas de normalidad multivariante para predictores

Existen numerosas pruebas que evalúan los supuestos de normalidad multivariante pero no existe un único procedimiento estándar. Sin embargo, en este estudio se utilizan las pruebas de Henze-Zirkler y Royston sugeridas en [575] debido a su buen control del error tipo I y potencia. La Tabla 8.7 muestra los resultados de la prueba Henze-Zirkler para muestras mayores de 100 observaciones y su resultado se valida con la prueba de Royston.

Los predictores x.VPT y Sp.VrCP.PT, con un nivel de significación superior a 0.05, se ajustan a una distribución normal multivariante. Esto implica que, en el contexto de este análisis, los datos relacionados con estas variables y grupos específicos tienen una mayor probabilidad de seguir una distribución normal univariante. Este hallazgo se ve corroborado por el uso del estadístico de prueba de Henze-Zirkler, que tiene una distribución aproximadamente log-normal. Estos resultados son coherentes con la literatura existente [429] y apoyan la hipótesis de que los conjuntos de datos multivariantes, en el análisis de estos predictores, también asumen una distribución normal univariante.

Tabla 8.7: Pruebas de normalidad multivariante para predictores

Prueba		**Medidas**	**Estadístico**	**P-Valor**
Henze-Zirkler	x.V*PT	Sp.Vr*CP.PT	0.987	0.051
Royston	x.V*PT	Sp.Vr*CP.PT	2.708	0.260

Los métodos gráficos también validan los resultados obtenidos. En la Figura 8.3 se resume dicho análisis. El primer gráfico (**a**) muestra la clara concordancia entre los cuantiles de las distribuciones de probabilidad hipotetizada y observada. Los valores en Y tienden a ser iguales en X. Es decir, los datos evaluados no se desvían de la normalidad multivariante.

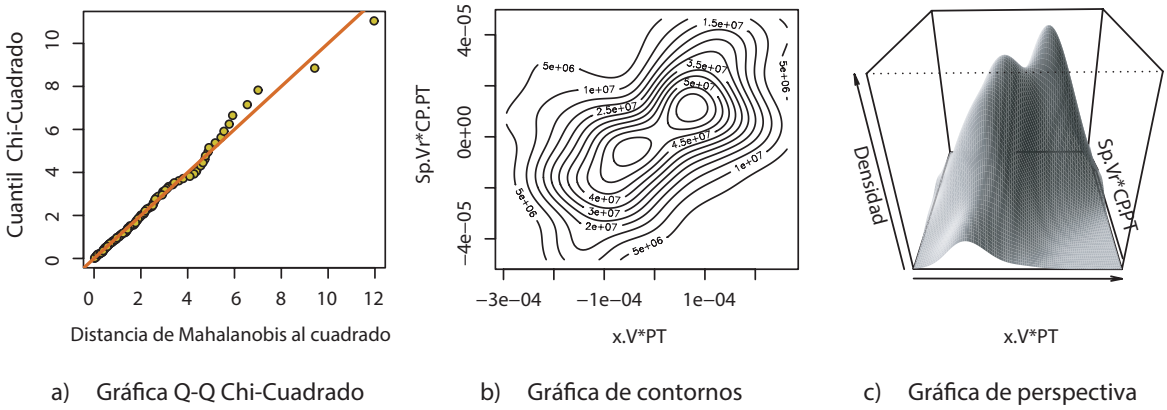


Figura 8.3: Gráficas multivariantes de variables predictoras.
La estructura de los datos, de las variables predictoras, se aproxima a una distribución normal multivariante. Las representaciones se han obtenido a partir de la matriz de datos $\mathbf{Z}\,\langle 125 \times 2\rangle$.(**a**) Gráfica Q-Q Chi-Cuadrado. (**b**) Gráfica de contornos de x.V*PT y Sp.Vr*CP.PT. (**c**) Gráfica de perspectiva de variables predictoras.

La representación de las líneas de contorno elipsoidal (**b**) es útil para verificar la normalidad multivariante de los predictores. La vista superior del gráfico de contorno de x.V*PT y Sp.Vr*CP.PT insinúa una ligera figura de campana de gauss con una separación y diferenciación evidentes entre los rendimientos medios negativos y positivos generados por la depreciación / apreciación del tipo de cambio, antes del cambio de tendencia.

Al observar los histogramas de la Figura 8.4, esta diferencia es más evidente en x.V*PT que en Sp.Vr*CP.PT y esto se debe a que los rendimientos logarítmicos son leptocúrticos y asimétricos. Además, las líneas de contorno confirman la correlación positiva entre grupos, como se detectó en el análisis **GH – Biplot** de la Figura 8.2.

El gráfico de perspectiva (**c**) es una representación bivariada en una superficie de distribución de probabilidad tridimensional. Esta figura proporciona información sobre dónde tienden a concentrarse los datos y cómo se correlacionan las variables. La forma de esta representación se aproxima a una distribución gaussiana.

B. Pruebas de normalidad univariante para predictores

Los resultados obtenidos en las pruebas de la Tabla 8.8 apoyan la afirmación de que el subconjunto de datos cumple el supuesto de normalidad univariante al nivel de significación del 5%. Esto implica que las variables objeto de estudio presentan una distribución que se asemeja a la distribución normal, fundamental para muchos análisis estadísticos.

El valor del estadístico de prueba Anderson-Darling, denominado Estadístico AD, se utiliza para probar la hipótesis nula H_0 de que la muestra procede de una población con distribución normal. Su uso permite evaluar cuantitativamente la adecuación de la distribución de los datos a la normalidad y respalda la validez de los resultados obtenidos en las pruebas. Estos resultados son importantes para garantizar la solidez y fiabilidad de los análisis posteriores basados en el supuesto de normalidad.

Tabla 8.8: Pruebas de normalidad univariante para predictores

Prueba	**Medida**	**Casos**	[a]**Estadístico AD**	**P-Valor**
Anderson-Darling	x.V*PT	125	0.724	0.058
Anderson-Darling	Sp.Vr*CP.PT	125	0.182	0.912

[a]Estadístico de prueba Anderson-Darling.

En la Figura 8.4, se examina la estructura de los datos mediante gráficos de normalidad. Se observa que los predictores x.VPT y Sp.VrCP.PT presentan una distribución normal univariante al nivel de significación del 5%. Estos resultados apoyan el supuesto de normalidad en el análisis de estos predictores y proporcionan una base sólida para su uso en modelos posteriores.

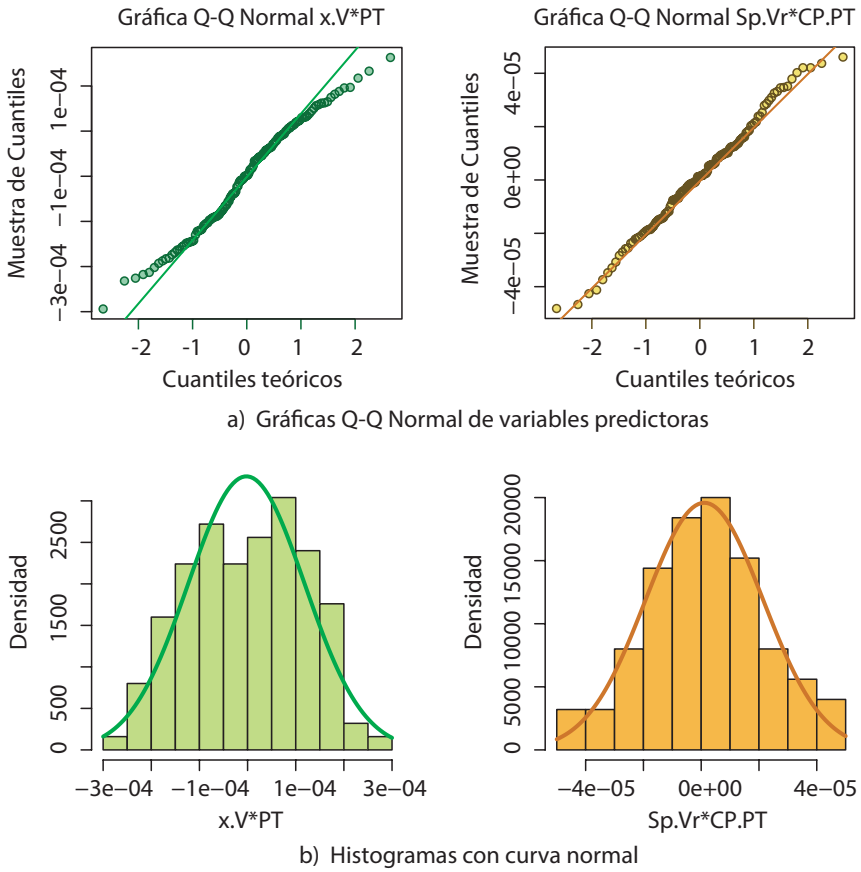


Figura 8.4: Gráficas univariantes de variables predictoras.
La distribución de los datos, de las variables predictoras, se aproxima a una distribución normal. Las representaciones se basan en la matriz de datos $\mathbf{Z}\,\langle 125 \times 2\rangle$. (**a**) Gráficas Q-Q Normal para variables predictoras x.V*PT y Sp.Vr*CP.PT. (**b**) Histogramas con curva normal.

C. Pruebas de normalidad multivariante para grupos

La asunción de una distribución normal multivariante en grupos de dos o más variables es de vital importancia en el análisis discriminante, puesto que garantiza el cumplimiento de los supuestos de normalidad multivariante. Esta condición es fundamental para garantizar la validez y eficacia de los resultados obtenidos en el análisis discriminante. Además, permite una correcta interpretación de las relaciones entre las variables predictoras y las categorías de clasificación.

Los resultados obtenidos de las pruebas de Henze-Zirkler y Royston, que se muestran en la Tabla 8.9, indican que el conjunto de datos de las direcciones de los movimientos al alza y a la baja del tipo de cambio euro-dólar se ajusta adecuadamente a una distribución normal multivariante. Así lo demuestran los valores de significación

obtenidos, que son estadísticamente superiores a 0.05. Estos resultados apoyan la validez del supuesto de normalidad multivariante en el contexto del análisis de grupos.

Tabla 8.9: Pruebas de normalidad multivariante para grupos

Prueba	Tendencia	Medidas		n	Estadístico	P-Valor
Henze-Zirkler	Alcista	x.V*PT	Sp.Vr*CP.PT	125	0.874	0.061
	Bajista	x.V*PT	Sp.Vr*CP.PT	125	0.594	0.317
Royston	Alcista	x.V*PT	Sp.Vr*CP.PT	125	1.737	0.419
	Bajista	x.V*PT	Sp.Vr*CP.PT	125	2.705	0.259

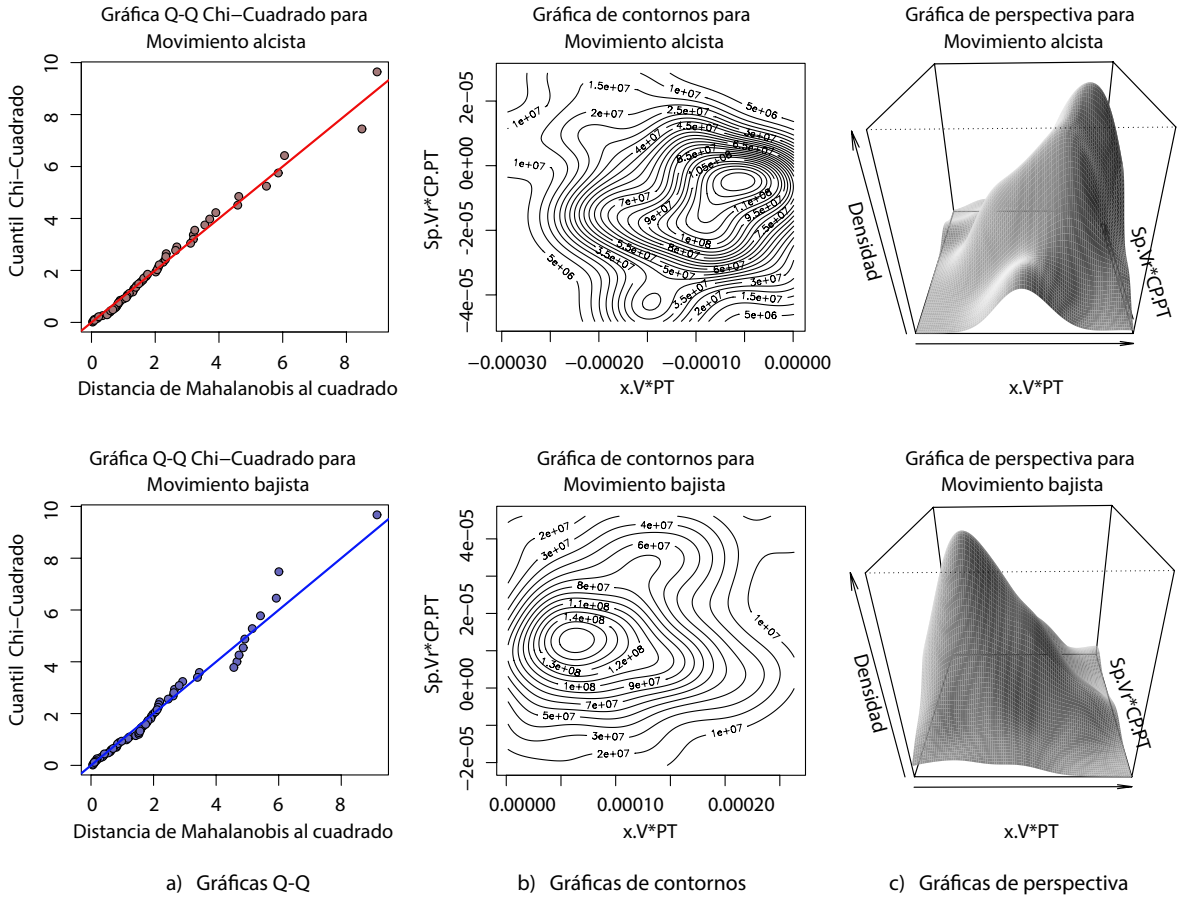


Figura 8.5: Gráficas multivariantes de grupos.
Las estructuras de los grupos, movimientos de tendencia alcistas y bajistas, se aproximan a una distribución normal multivariante. Las representaciones se obtienen a partir de la matriz de datos $\mathbf{Z}\langle 125 \times 2\rangle$. (**a**) Gráficas Q-Q normal para grupos. (**b**) Gráficas de contorno para grupos. (**c**) Gráficas de perspectiva para grupos.

El análisis gráfico complementa y refuerza la evidencia obtenida a partir de las

pruebas estadísticas, proporcionando una mayor confianza en el supuesto de normalidad multivariante en el análisis realizado. Aunque los subconjuntos de observaciones de precios al alza y a la baja se ajustan a una distribución normal multivariante, el análisis gráfico de la Figura 8.5 revela una ligera distribución asimétrica negativa y positiva en los grupos respectivos. Los gráficos de contorno (**b**) y de perspectiva (**c**) muestran esta asimetría, que se atribuye al comportamiento asimétrico y leptocúrtico de los rendimientos de los precios de cierre antes del cambio de tendencia. A pesar de esta observación, es importante señalar que las muestras siguen una distribución normal multivariante, lo que respalda la validez de los resultados obtenidos en el análisis.

Si los resultados de la Tabla 8.9 indican que el subconjunto de datos satisface los supuestos de normalidad multivariante a un nivel de significación del 5% y el análisis gráfico de la Figura 8.5 corrobora los resultados de las pruebas, entonces es valido afirmar, tal como se confirma en el apartado siguiente, que cada predictor y por ende cada grupo dentro de cada predictor tiene una distribución normal univariante [429].

D. Pruebas de normalidad univariante para grupos

Es crucial que los datos de grupo de cada uno de los predictores utilizados en la construcción de modelos de clasificación paramétricos asuman una distribución normal. La normalidad de los datos es un supuesto fundamental en muchos métodos estadísticos y contribuye a obtener resultados más fiables y sólidos. Esto significa que al garantizar la normalidad de cada categoría dentro de cada variable individual es esencial para realizar un análisis estadístico adecuado y obtener resultados válidos.

Tabla 8.10: Pruebas de normalidad univariante para grupos

Prueba	**Medida**	**Tendencia**	**Casos**	[a]**Estadístico AD**	**P-Valor**
Anderson-Darling	x.V*PT	Alcista	62	0.343	0.480
		Bajista	63	0.402	0.348
Anderson-Darling	Sp.Vr*CP.PT	Alcista	62	0.212	0.851
		Bajista	63	0.462	0.251

[a]Estadístico de prueba Anderson-Darling.

Según resultados de la Tabla 8.10, en todos los análisis, el P-Valor del estadístico de prueba Anderson-Darling es mayor a 0.05, en consecuencia, los subconjuntos de observaciones de cada uno de los grupos tendenciales dentro de cada uno de los predictores

x.V*PT y Sp.Vr*CP.PT se ajustan a una población con distribución normal.

El análisis gráfico de los grupos dentro de cada variable complementa las pruebas estadísticas de normalidad multivariante y univariante al proporcionar una visualización de la distribución de los datos y las posibles desviaciones de normalidad. En la Figura **??**, se observa que los grupos de precios en los movimientos al alza y a la baja presentan distribuciones normales en los predictores x.V*PT y Sp.Vr*CP.PT. Estos resultados respaldan la validez de los supuestos de normalidad multivariante y univariante en la muestra de selección, y se confirman mediante la inspección visual de los gráficos correspondientes.

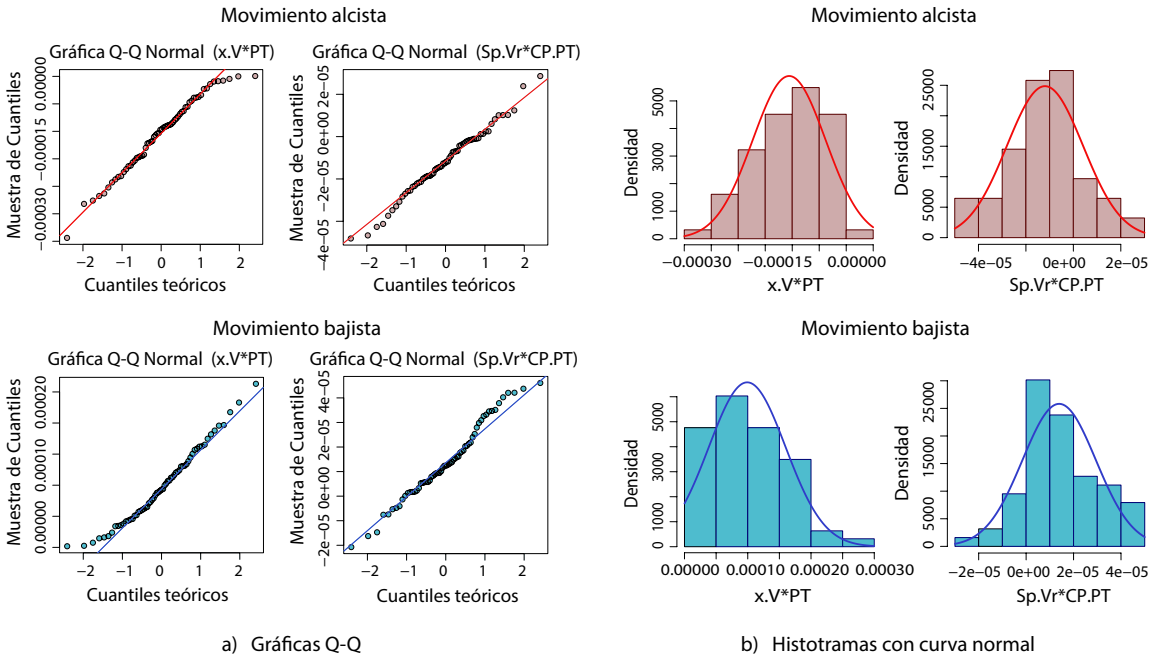


Figura 8.6: Gráficas univariantes de grupos.
Las estructuras de los grupos, movimientos de tendencia alcistas y bajistas, se aproximan a una distribución normal. Las representaciones se obtienen de la matriz de datos $\mathbf{Z}\langle 125 \times 2\rangle$. (**a**) Gráficas Q-Q normal para grupos. (**b**) Histogramas de grupos con curva normal.

En términos generales, los resultados obtenidos corroboran que la muestra de selección ($\mathbf{Z}$) cumple con los supuestos de normalidad tanto multivariante como univariante para los predictores y los grupos. Además, la validez de los resultados obtenidos mediante los estadísticos de prueba se confirma al observar los correspondientes gráficos de normalidad. Esto proporciona una base sólida para análisis posteriores y aumenta la confianza en la interpretación de los resultados.

E. Evaluación de la contribución de cada predictor

La evaluación de la contribución de cada predictor en la construcción de modelos de clasificación es esencial para obtener modelos más eficaces. Esta evaluación permite comprender la importancia relativa que tiene cada variable en la predicción del resultado y ayuda a validar la selección de los predictores más relevantes. Ademas, la prueba de igualdad de medias de grupo se utiliza para evaluar el potencial discriminante de cada variable predictora antes de su inclusión en el modelo.

La Tabla 8.11 muestra los resultados de la prueba, utilizando el estadístico Lambda de Wilks. Los valores más pequeños de Lambda de Wilks indican una mejor capacidad para discriminar entre grupos. Los resultados de la prueba confirman que el rendimiento medio medido 18 periodos antes del cambio de régimen x.V*PT presenta un valor Lambda de Wilks más bajo en comparación con la pendiente de las variaciones medidas 18 momentos antes del cambio de tendencia Sp.Vr*CP.18P, lo que sugiere que el primero es un mejor discriminante entre grupos. Además, los $p - valores$ asociados a ambos predictores son inferiores al 5%, lo que indica significación estadística en la discriminación entre medias de grupos.

Tabla 8.11: Prueba de igualdad de medias de grupos

ID	Medida	Lambda de Wilks	F-Valor	[a]Df$_1$	*Df$_2$	P-Valor
1	x.V*PT	0.293	205.529	1	85	0.000
2	Sp.Vr*CP.PT	0.638	48.238	1	85	0.000

[a]**Df**= Grados de libertad.

F. Evaluación de la colinealidad de los predictores

Es fundamental evaluar la multicolinealidad en las variables utilizadas como predictores en la construcción de modelos de clasificación basados en el análisis discriminante. Esto se debe a que la multicolinealidad puede afectar la estabilidad de los coeficientes y la interpretación de los resultados, comprometiendo la precisión y fiabilidad del modelo. Al identificar y abordar la multicolinealidad, se puede garantizar que los predictores seleccionados sean independientes entre sí, lo que refuerza la capacidad del modelo para discriminar correctamente entre grupos y mejora su capacidad predictiva.

En este contexto, la matriz de correlaciones dentro de grupos, presentada en la Tabla 8.12, revela la correlación entre los predictores x.V*PT y Sp.Vr*CP.PT. El valor obtenido sugiere una clara independencia entre los predictores, lo que implica

que no existen problemas significativos de multicolinealidad. Además, el coeficiente de correlación no es lo suficientemente alto como para generar inestabilidad en los signos de los coeficientes del modelo de clasificación. En consecuencia, la metodología propuesta de selección de predictores es eficaz para superar el problema de multicolinealidad, garantizando así la estabilidad y fiabilidad de los coeficientes en la construcción de modelos de clasificación.

Tabla 8.12: Matriz de correlación intragrupos

Medidas	**x.V*PT**	**Sp.Vr*CP.PT**
x.V*PT	1	-0.026
Sp.Vr*CP.PT	-0.026	1

En la Tabla 8.13, la matriz de estructuras esta definida por las correlaciones dentro de grupos combinados entre los predictores y la función canónica discriminante estandarizada. El orden de las variables esta en función del valor absoluto de esa correlación y de los coeficientes de las funciones de clasificación. Además, dicho orden es idéntico al que se muestra en la prueba de igualdad de medias de grupos de la Tabla 8.11. Esta concordancia confirma la ausencia de multicolinealidad entre las variables independientes seleccionadas.

Los coeficientes estandarizados facilitan la comparación de las variables independientes medidas en diferentes escalas. El coeficiente con el valor absoluto mas alto corresponde a la variable predictora con mayor capacidad discriminante. De manera que, el orden de los predictores, según sus coeficientes estandarizados, muestra la importancia que tiene el rendimiento medio x.V*PT en la clasificación y predicción de la dirección futura del tipo de cambio euro-dólar.

En consecuencia, como la función discriminante no esta afectada por la multicolinealidad, es seguro mencionar que la dirección futura del movimiento del precio esta determinada, principalmente, por el rendimiento medio x.V*PT del tipo de cambio euro-dólar y, por tanto, el rendimiento medio x.V*PT discrimina mejor entre las direcciones de los movimientos tendenciales alcistas y bajistas.

Tabla 8.13: Matriz de estructuras y coeficientes de función estandarizados

ID	**Medidas**	**Función discriminante**	**Coeficientes de función estandarizados**
1	x.V*PT	0.891	0.902
2	Sp.Vr*CP.PT	0.432	0.455

G. Evaluación de homogeneidad de matrices de covarianza

La medición y evaluación de la homogeneidad de las matrices de covarianza dentro de grupos es crucial en la construcción de modelos de clasificación basados en el análisis discriminante. En la Tabla 8.14 se presenta los valores de los rangos y los logaritmos de los determinantes de las matrices de covarianzas de los grupos. Los logaritmos de los determinantes son una medida de la variabilidad dentro de cada grupos. Las pequeñas diferencias en los logaritmos de los determinantes indican grupos con matrices de covarianzas más similares y compactas. El estadístico M de Box de la Tabla 8.15 evalúa el supuesto de igualdad de matrices de covarianzas entre grupos.

Tabla 8.14: Medida de variabilidad de los grupos

Tendencia	**Rango de matriz**	**Logaritmo de determinantes**
Movimiento alcista	2	-41.261
Movimiento bajista	2	-41.395
Agrupados dentro de grupos	2	-41.283

Según resultados de la Tabla 8.15, el P-Valor del estadístico de prueba M de Box es mayor a 0.05, en consecuencia, los subconjuntos de observaciones de los grupos en los predictores x.V*PT y Sp.Vr*CP.PT provienen de poblaciones con matrices de varianzas-covarianzas iguales y, por tanto, implícitamente se puede inferir que la formulación de las funciones de clasificación no requieren de matrices de covarianzas de grupos separados. El cumplimiento de este supuesto le otorga elegibilidad a los resultados generados por el uso del análisis discriminante.

Tabla 8.15: Prueba M de Box de igualdad de matrices de covarianzas entre grupos

M de Box	**F-Valor**	**[a]Df$_1$**	**Df$_2$**	**P-Valor**
3.092	1.004	3	1122087.83	0.390

[a]**Df**= Grados de libertad.

8.3.3 Evaluación del ajuste del modelo

Teóricamente el tamaño de los grupos de la muestra de análisis determina la probabilidad a priori de pertenencia a cada grupo. La definición de los coeficientes de las

funciones de clasificación y el desempeño del proceso de clasificación están determinados por las probabilidades previas utilizadas en la formulación del modelo. Según la Tabla 8.6, la muestra de estudio **Z** esta formada por grupos de igual tamaño. En consecuencia, para evitar cualquier influencia del desequilibrio de los grupos en el cálculo de los coeficientes de las funciones discriminantes, se establecen probabilidades a priori iguales para ambos grupos ($\pi_g = 0.50$). Esta igualdad en las probabilidades a priori garantiza la igualdad de trato de los grupos y ayuda a evitar cualquier sesgo en los resultados del análisis discriminante.

La Tabla 8.16 presenta información sobre la medida de la eficacia de la función discriminante lineal mediante la correlación canónica. En este contexto, la correlación canónica evalúa la relación entre los valores reales de las observaciones de cada grupo y las puntuaciones discriminantes predichas. Esta medida proporciona información sobre la capacidad discriminante de la función lineal para clasificar los grupos. En consecuencia, con un apreciable eigenvalor de **3.048** las variables predictoras que intervienen en la formulación del modelo explican el 100% de la varianza total acumulada, y la correlación canónica del **87**% entre las variables del modelo y las puntuaciones discriminantes pronosticadas deja vislumbrar un modelo que se ajusta al comportamiento de los datos y que ofrece, en términos de eficacia, una alta correlación entre los valores observados y los datos pronosticados.

Tabla 8.16: Eficacia de la función lineal discriminante

Eigenvalor	**% Varianza**	**% Varianza acumulada**	**Correlación canónica**
3.048	100	100	0.868

En la Tabla 8.17 el estadístico Lambda de Wilks se utiliza como una medida que verifica que tan bien la función lineal discriminante separa los casos en los grupos. Valores pequeños, cercanos a cero, del estadístico Lambda de Wilks son indicadores de la capacidad discriminante de la función. Chi-cuadrado es el valor del estadístico asociado utilizado para contrastar la hipótesis H_0. De manera que con un P-Valor inferior al 5% se rechaza la hipótesis H_0 de que las medias de las funciones lineales de Fisher son iguales entre grupos y, por tanto, el modelo lineal discriminante contribuye a la diferenciación de los casos entre las medias de los grupos.

Tabla 8.17: Prueba de funciones

Lambda de Wilks	**Chi-Cuadrado**	[a]**Df$_1$**	**P-Valor**
0.247	117.443	2	0.000

[a]**Df**=Grados de libertad.

8.3.4 Clasificación de la dirección del movimiento

Las funciones lineales de Fisher ayudan a asignar los casos a un grupo de pertenencia en particular. Para cada caso se calcula una puntuación de clasificación y el modelo de análisis discriminante asigna el caso al grupo cuya función de clasificación obtuvo la puntuación discriminante mas alta. Las funciones lineales de Fisher que mejor discriminan entre movimientos tendenciales alcistas y bajistas se resumen en la ecuación (8.1).

$$\mathrm{y}(\hat{X}) = \beta_0 + \beta_1 \hat{\mathrm{X}}_1 + \beta_2 \hat{\mathrm{X}}_2 \tag{8.1}$$

Donde, $\mathrm{y}(\hat{X})$ es la puntuación discriminante, $\hat{\mathrm{X}}_1$ es el rendimiento medio (x.V*PT), $\hat{\mathrm{X}}_2$ es la pendiente de la linea de regresión de las variaciones entre precios de cierre (Sp.Vr*CP.PT), ambos criterios calculados en los últimos 18 periodos. Los coeficientes de las funciones de clasificación β_0, β_1 y β_2 están referidos en la Tabla 8.18.

Ahora bien, partiendo de la función principal (8.1), el modelo de análisis discriminante trabaja con dos funciones lineales denotadas como $\mathrm{y}(\hat{X}_U)$ and $\mathrm{y}(\hat{X}_D)$ para discriminar entre movimientos tendenciales alcistas $\mathrm{C_U}$ y movimientos tendenciales bajistas $\mathrm{C_D}$. La regla que asigna los casos al grupo de pertenencia es:

$$\text{Si } \mathrm{y}(\hat{X}_U) \geq \mathrm{y}(\hat{X}_D) \text{ entonces } \mathrm{y}(\hat{X}_U) \text{ pertenece a } \mathrm{C_U}$$
$$\text{En caso contrario, si } \mathrm{y}(\hat{X}_U) < \mathrm{y}(\hat{X}_D) \text{ entonces } \mathrm{y}(\hat{X}_D) \text{ pertenece a } \mathrm{C_D}$$

La estructura de las puntuaciones discriminantes se afecta con los signos y la magnitud de los coeficientes de las variables que participan en el modelo. La función de clasificación de movimientos al alza, según la Tabla 8.18, se caracteriza porque los coeficientes de los predictores son menores a cero. Esto significa, según la contribución que cada variable hace en el modelo, que cuanto menor sea el rendimiento medio x.V*PT respecto a la pendiente de estas variaciones Sp.Vr*CP.PT, es menos probable que la cotización del euro frente al dólar americano continué depreciándose. En consecuencia, luego de que la caída del tipo de cambio se desacelera y entra en una fase de

agotamiento y corrección, en este nivel, el comportamiento negativo de la acción del precio y que se produce antes del cambio de régimen de mercado, es el valor utilizado por la función discriminante para predecir el comienzo de un movimiento alcista.

Asimismo en la Tabla 8.18, en la función de clasificación del movimiento bajista, el cambio de signo en los coeficientes indica que los casos con rendimientos medios positivos x.V*PT mayores que las pendientes de dichas variaciones Sp.Vr*CP.PT, tienen mas probabilidades de haber llegado al final de un proceso de apreciación, anunciando así el inicio de un régimen de mercado bajista.

Tabla 8.18: Coeficientes de las funciones de clasificación

Medida	**Símbolo**	**Dirección del movimiento**	
		Alcista y($\hat{X}_U$)	**Bajista** y($\hat{X}_D$)
x.V*PT	β_1	-24598.42	22755.91
Sp.Vr*CP.PT	β_2	-44485.91	52112.42
Constante	β_0	-2.26	-2.16

En contraste con lo que se dijo, el diagrama de cajas de la Figura 8.7 muestra la distribución de las medidas de los predictores por grupos. Las variaciones esperadas de los datos se disponen simétricamente en dos regiones diferentes delimitadas por un nivel central cero. Los rendimientos medios negativos x.V*PT indican que el tipo de cambio se ha depreciado y se mantiene en la zona de sobre-venta. En cambio, cuando los rendimientos medios x.V*PT son positivos, el tipo de cambio esta apreciado y se sitúa en zona de sobre-compra. Teóricamente, los cuartiles Q_1, Q_2 y Q_3, de los diagramas de cajas, definen los niveles de soporte y resistencia en los que históricamente ha cambiado la dirección de la tendencia e implícitamente indica el número de movimientos generados en cada nivel. Los resultados muestran que para cada grupo existe una explícita concordancia entre las medidas de los predictores. Obsérvese el 75% de similitud entre los signos de los valores de Sp.Vr*CP.PT y x.V*PT.

los hallazgos obtenidos parecen sugerir que el cambio de dirección en el movimiento de los precios es mas probable cuando se producen las siguientes señales. Un movimiento alcista es mas probable que ocurra cuando el rendimiento medio negativo x.V*PT es mayor a la media de los rendimientos medios negativos x.V*PT y la pendiente de las variaciones Sp.Vr*CP.PT sea negativa. En caso contrario, un movimiento bajista es más probable que ocurra cuando el rendimiento promedio positivo x.V*PT es menor que la media de los rendimientos medios positivos x.V*PT y la pendiente de las varia-

ciones Sp.Vr*CP.PT sea positiva.

En términos generales, debido a que el rendimiento medio tiene la mayor contribución en la función discriminante, una reducción o aumento sostenido en el rendimiento medio del tipo de cambio por encima o por debajo de su valor promedio histórico hace que sea mas probable que cada caso, según la puntuación discriminante mas alta, se clasifique como el comienzo de un movimiento tendencial alcista o bajista respectivamente.

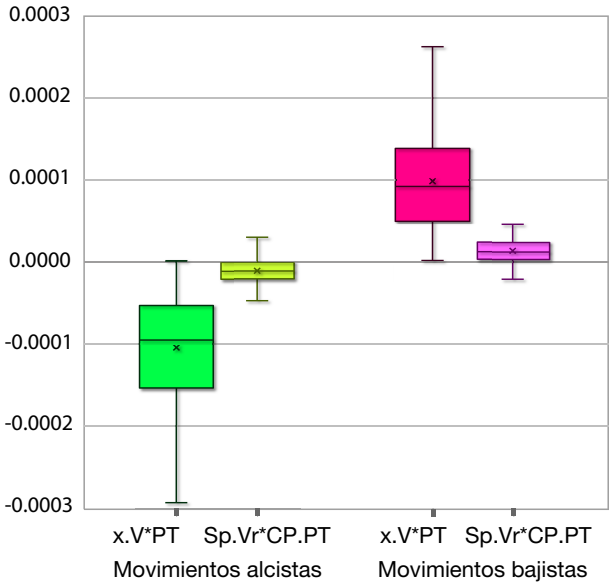


Figura 8.7: Diagrama de cajas para predictores y grupos.
Valores medios de rendimiento x.V*PT y pendientes de las variaciones Sp.Vr*CP.PT en los puntos de inflexión del mercado `ET` en donde cambia la dirección de la tendencia precedente. Valores obtenidos de la muestra de estudio $\mathbf{Z}\,\langle 125 \times 2\rangle$. Puntos de inflexión `ET` que dan lugar a: (**a**) movimientos al alza y (**b**) movimientos a la baja.

8.3.5 Validación del modelo

La validación del modelo de clasificación es esencial para garantizar su fiabilidad, precisión y capacidad de generalización. La Tabla 8.19 muestra el desempeño conseguido con la aplicación del modelo, con una tasa de clasificación correcta del **98.9**%. Según datos del apartado Muestra de Entrenamiento, de 87 casos utilizados para crear la función discriminante lineal, 37 de 38 movimientos bajistas y 49 de 49 trayectorias de precios alcistas se clasifican satisfactoriamente.

En el apartado de Validación cruzada, se evalúa el desempeño del modelo excluyendo los casos utilizados en su formulación. 86 de 87 casos son clasificados correc-

tamente, es decir, que el **98.9**% de los casos seleccionados para crear el modelo y que son validados de forma cruzada son clasificados correctamente.

Finalmente, en la sección de Evaluación, se valida la eficacia del modelo de clasificación. Utilizando el 30% de los movimientos de precios históricos restantes de la muestra **Z** y que no participaron en la formulación del modelo, el **100**% de las observaciones utilizadas para evaluar el rendimiento del modelo se clasifican correctamente.

Tabla 8.19: Resultados de clasificación

Casos de análisis		**Dirección del**	**Predicción**		**Total**
Muestra	**n = 125**	**Precio**	**Alcista**	**Bajista**	
Entrenamiento		Movimiento alcista (n)	49	0	49
		Movimiento bajista (n)	1	37	38
		Movimiento alcista (%)	100%	0%	100%
		Movimiento bajista (%)	2.60%	97.40%	100%
Validación cruzada		Movimiento alcista (n)	49	0	49
		Movimiento bajista (n)	1	37	38
	87	Movimiento alcista (%)	100%	0%	100%
	70%	Movimiento bajista (%)	2.60%	97.40%	100%
Evaluación		Movimiento alcista (n)	13	0	13
		Movimiento bajista(n)	0	25	25
	38	Movimiento alcista (%)	100%	0%	100%
	30%	Movimiento bajista (%)	0%	100%	100%

Este resultado indica que el modelo es eficaz y capaz de detectar con precisión las diferencias entre el inicio de un movimiento alcista y uno bajista. Es importante destacar que el desequilibrio de clases dentro de los subconjuntos de datos no tiene un impacto perjudicial en la precisión de la clasificación. Además, el proceso de preparación y extracción de instancias en la muestra de estudio, la representatividad y equilibrio de grupos en la muestra de datos, el poder discriminante de las variables predictoras seleccionadas y la idoneidad del análisis discriminante son factores determinantes en el rendimiento alcanzado.

Los valores de precisión obtenidos respaldan la utilidad y fiabilidad del modelo en tareas de previsión, lo que lo convierte en una valiosa herramienta para analistas e inversores en la toma de decisiones de inversion.

8.4 Desempeño del modelo de predicción

En esta sección se presentan los resultados de la evaluación del poder predictivo a corto plazo del modelo de clasificación de la dirección del movimiento del tipo de cambio euro-dólar. Se han añadido otros resultados experimentales para evaluar la

eficacia de la metodología propuesta. Estos resultados constituyen una aportación valiosa y concluyente que apoya, complementa y amplía el rendimiento obtenido. Estos experimentos adicionales se basan en datos fuera de muestra de distintas condiciones de mercado y abarcan horizontes temporales diferentes. Los datos utilizados son puntos de inflexión (ET) extraídos de los datos de mercado. El horizonte temporal utilizado para predecir el inicio de un nuevo régimen de mercado corresponde a un periodo futuro medido en un marco temporal de 15 minutos. En cada caso, el inicio de un nuevo movimiento direccional se predice a partir de un evento (ET) en el que históricamente se ha producido el cambio de dirección del movimiento de los precios.

La Tabla 8.20 presenta los resultados de clasificación utilizando datos fuera de muestra a partir de las funciones lineales de Fisher. Los datos se dividen en dos secciones: La primera sección corresponde a los datos de entrenamiento y validación cruzada, y la segunda a los datos de prueba: OOS_1, OOS_2, OOS_3 y OOS_4. Cada conjunto de datos consta de una variable dependiente formada por dos categorías: movimientos alcistas y movimientos bajistas y las dos variables predictoras seleccionadas.

En la primera sección, el conjunto de datos de entrenamiento y validación cruzada muestra una clasificación perfecta del 100% con los movimientos alcistas y del 98.40% con los datos de los movimientos bajistas. Estos resultados indican que el modelo clasifica excepcionalmente bien los movimientos bajistas e incluso es perfecto con los movimientos alcistas.

En la segunda sección, sobre los conjuntos de datos OOS_1, OOS_2, OOS_3 y OOS_4, los resultados de clasificación son coherentes con los obtenidos con el conjunto de entrenamiento y validación cruzada. En general, el método clasifica excepcionalmente bien los datos alcistas y bajistas, con valores del 98.10% al 99.0% y del 98.8% al 99.0%, respectivamente. En este caso, el rendimiento de la clasificación con datos bajistas es incluso mejor.

Estos resultados sugieren que las funciones lineales de Fisher se ajustan a la estructura de los datos y discriminan muy bien entre tendencias alcistas y bajistas. Esto demuestra que el análisis discriminante es una herramienta eficaz para clasificar los datos de tendencias sin sacrificar la legibilidad de las interpretaciones. Además, la relevancia de estos resultados y su relación con el objetivo de la metodología propuesta demuestran una clara eficacia en la predicción de tendencias a corto plazo.

En cuanto a los experimentos adicionales, realizados con los conjuntos OOS_1 y OOS_4, muestran un rendimiento de clasificación coherente con los resultados ya reportados. El modelo clasifica correctamente los movimientos alcistas y bajistas con valores

que oscilan entre el 98.7% y el 99.0% y entre el 98.8% y el 99.9%, respectivamente.

Tabla 8.20: Resultados de clasificación fuera de muestra

Casos de análisis		**Dirección del precio**	**Predicción**		**Total**
Conjuntos	**n**		**Alcista**	**Bajista**	
Entrenamiento y	125	Movimiento alcista (n)	62	0	62
validación		Movimiento bajista (n)	1	62	63
cruzada		Movimiento alcista (%)	100%	0%	100%
		Movimiento bajista (%)	1.60%	98.40%	100%
Evaluación:					
OOS_1: 1999-2005	6581	Movimiento alcista (n)	3306	35	3341
		Movimiento bajista (n)	40	3200	3240
		Movimiento alcista (%)	99.00%	1.00%	100%
		Movimiento bajista (%)	1.20%	98.80%	100%
OOS_2: 2006-2020	14645	Movimiento alcista (n)	7227	91	7318
		Movimiento bajista(n)	84	7243	7327
		Movimiento alcista (%)	98.80%	1.20%	100%
		Movimiento bajista (%)	1.10%	98.90%	100%
OOS_3: 2006-2020	1748	Movimiento alcista (n)	569	11	580
		Movimiento bajista(n)	14	1154	1168
		Movimiento alcista (%)	98.10%	1.90%	100%
		Movimiento bajista (%)	1.20%	98.80%	100%
OOS_4: 2021-2023	2336	Movimiento alcista (n)	1172	16	1188
		Movimiento bajista (n)	11	1137	1148
		Movimiento alcista (%)	98.70%	1.30%	100%
		Movimiento bajista (%)	1.00%	99.00%	100%

Los resultados experimentales demuestran la bondad de ajuste del modelo y su capacidad de generalización y precisión en diferentes condiciones de mercado y horizontes temporales. Sin embargo, existen ciertas limitaciones que deben cumplirse para mantener la coherencia de los resultados obtenidos. Entre las más importantes se encuentran: (i) Garantizar la calidad de los resultados producidos en la preparación y selección de características y su validación con el análisis de estructuras. (ii) Evitar el uso de puntos de inflexión, cuando están influenciados por la publicación de datos fundamentales, dado que la precisión en la predicción puede verse afectada por la volatilidad experimentada. (iv) Utilizar características discriminativas e independientes que proporcionen diferencias significativas entre tendencias para asegurar una clasificación precisa. (v) Garantizar el cumplimiento de los supuestos de normalidad multivariante en la muestra de entrenamiento y validación, cuando se decida utilizar como método de clasificación el análisis discriminante. Si esta condición no se cumple, la fiabilidad de las interpretaciones puede verse comprometida y los resultados obtenidos pueden

ser meramente descriptivos en lugar de inferibles.

Finalmente, además de cumplir con los requerimientos anteriores, es esencial realizar pruebas de significación estadística en cada etapa del proceso para garantizar la fiabilidad y coherencia de los resultados obtenidos y asegurar el éxito de la aplicación del enfoque propuesto a otros instrumentos financieros.

8.4.1 Poder predictivo del modelo de clasificación

La evaluación del poder predictivo de un modelo de clasificación es primordial para garantizar su eficacia y fiabilidad en la toma de decisiones informadas. La Tabla 8.21 presenta la evaluación del poder predictivo del modelo de clasificación obtenido a partir de la metodología propuesta. En la tabla se han incluido tres secciones que contienen diferentes medidas de rendimiento. En la primera sección, se presentan las medidas de rendimiento para el conjunto de datos de entrenamiento y validación. En la segunda sección, se recogen las medidas de rendimiento para los conjuntos de datos de prueba. Finalmente, en la tercera sección se presentan los estadísticos descriptivos de los valores medios de rendimiento, con un intervalo de confianza del 95%.

Tabla 8.21: Poder predictivo del modelo de clasificación

[a]**Datos**	**n**	**ACC**	**SEN**	**ESP**	**AUC**	**PRE**	**F1**	**LB**	**K**	**MCC**
Entrenamiento	125	0.992	0.984	1.000	0.992	1.000	0.992	0.500	0.984	0.984
Validación cruzada	125	0.992	0.984	1.000	0.992	1.000	0.992	0.500	0.984	0.984
Evaluación:										
OOS_1: 1999-2005	6581	**0.989**	0.988	0.989	**0.989**	**0.990**	**0.989**	0.500	**0.977**	**0.977**
OOS_2: 2006-2020	14645	0.988	0.989	0.988	0.988	0.988	0.988	0.500	0.976	0.976
OOS_3: 2006-2020	1748	0.986	0.976	**0.991**	0.983	0.981	0.979	**0.556**	0.968	0.968
OOS_4: 2021-2023	2336	0.988	**0.991**	0.986	0.988	0.987	**0.989**	0.500	**0.977**	**0.977**
Estadísticos descriptivos:										
Rendimiento medio		0.9877	0.9860	0.9885	0.9870	0.9865	0.9862	0.5140	0.9745	0.9745
ICM (95%), LS		0.9897	0.9967	0.9918	0.9913	0.9926	0.9939	0.5585	0.9814	0.9814
ICM (95%), LI		0.9857	0.9752	0.9851	0.9826	0.9803	0.9785	0.4594	0.9675	0.9675
Error estándar		0.0006	0.0033	0.0010	0.0013	0.0019	0.0024	0.0140	0.0021	0.0021
Desviación estándar		0.0012	0.0067	0.0020	0.0027	0.0038	0.0048	0.0280	0.0043	0.0043

[a]Medidas de rendimiento: ACC = Exactitud, SEN = Sensibilidad, ESP = Especificidad, AUC = Área bajo la curva, PRE = Precisión, F1 = Puntuación F1, LB = Línea base, K = Kappa, MCC = Coeficiente de correlación de Matthews, ICM (95%) = Intervalo de confianza para la media (95%), LS = Límite superior, LI = Límite inferior.

En términos generales, el modelo de clasificación se entrenó y validó con un conjunto

de datos de entrenamiento y se evaluó con cuatro conjuntos de datos de prueba (OOS_1, OOS_2, OOS_3 y OOS_4). En cada conjunto de datos se presentan varias métricas de rendimiento, entre las que se incluyen la exactitud (ACC), la sensibilidad (SEN), la especificidad (ESP), el área bajo la curva ROC (AUC), la precisión (PRE), la puntuación F1-score (F1), la línea base (LB), el coeficiente Kappa (K) y el coeficiente de correlación de Matthews (MCC).

Los resultados indican que el modelo posee una capacidad predictiva excepcionalmente alta en todos los conjuntos de datos de evaluación, con una exactitud superior al 98.57%. Además, el modelo presenta valores AUC elevados, superiores al 98.26% en todos los conjuntos de datos de prueba, lo que sugiere que tiene una gran capacidad discriminatoria para diferenciar entre movimientos alcistas y bajistas.

Respecto a las medidas de sensibilidad y especificidad, el modelo presenta valores elevados, superiores a 97.52% y 98.51%, respectivamente, en todos los conjuntos de datos de prueba. Estos resultados indican que el modelo tiene capacidad para detectar tanto el inicio de un movimiento al alza como a la baja, lo que lo convierte en una herramienta valiosa para predecir las tendencias del mercado.

En cuanto a las medidas de precisión y F1-score, el modelo ha arrojado resultados notables, con valores superiores a 98.03% y 97.85%, respectivamente, en todos los conjuntos evaluados. Esto indica que el modelo tiene una capacidad extraordinaria para predecir movimientos alcistas. Además, en términos de F1-score, el modelo presenta un excelente equilibrio entre precisión y sensibilidad, lo que resalta su eficacia para identificar ambos tipos de movimientos.

Los coeficientes Kappa y MCC registran valores elevados, superiores al 96.75%, en todos los conjuntos de datos de prueba, lo que indica que el modelo tiene una capacidad excepcional para predecir la clase correcta. Estos resultados confirman que la muestra de entrenamiento recoge, a través de las variables predictoras, los patrones que mejor diferencian las tendencias. En consecuencia, el modelo tiene una gran capacidad de generalización y produce predicciones muy precisas sin comprometer la legibilidad e interpretabilidad de los resultados.

En cuanto al valor linea base, al sustituirlo por el mejor rendimiento obtenido por el modelo linea base del estado del arte (véase Tabla 8.22) de 0.8955, la metodología propuesta supera este valor en todas las métricas de evaluación de rendimiento.

Finalmente, es importante recalcar que los límites de confianza del 95% para todas las medidas de rendimiento están muy próximos a su valor medio. Esto indica una gran precisión de las medidas de rendimiento y, por tanto, una mayor fiabilidad en los

resultados obtenidos. Además, tanto el error estándar como la desviación estándar son muy pequeños, lo que sugiere una baja variabilidad en el rendimiento del modelo en los distintos conjuntos de datos. Esto, a su vez, apunta a una alta generalización, consistencia y precisión del modelo frente a diferentes condiciones de mercado, evidenciadas por el uso de 24 años de datos de mercado en el análisis.

8.4.2 Comparación con el estado del arte

En el ámbito de la inversión, resulta esencial contar con modelos de clasificación precisos e interpretables para predecir el comportamiento del mercado. La evaluación de estas características es esencial para determinar la eficacia del enfoque utilizado en la toma de decisiones de inversión. En este sentido, la Tabla 8.22 compara la metodología propuesta (LCC) con el estado del arte, y resumen las principales características de las metodología evaluadas. Estas metodología incluyen enfoques ensemble [329] e híbridos [576], redes neuronales [577, 578] y predicción de tendencias a partir del análisis de sentimiento [579]. Estos enfoques se han elegido para la comparación no solo por su eficacia, sino también porque son capaces de predecir la dirección del tipo de cambio euro-dólar.

Las metodología evaluadas en el estudio utilizan datos de mercado e indicadores técnicos como insumos de entrada. Sin embargo, las muestras de estudio utilizadas son limitadas y sólo consideran horizontes temporales concretos. Esto hace que estos modelos sean poco generalizables a largo plazo. Por otra parte, la extracción de características es muy común en estos enfoques, lo que compromete la legibilidad de las interpretaciones a cambio de mejoras en la precisión. Aunque los enfoques evaluados han demostrado su poder predictivo tanto con datos de entrenamiento como con datos fuera de muestra (OOS), se debe tener en cuenta que algunos estudios, como [329] y [578], solo se evaluaron con datos dentro de muestra (TVT). Esto puede llevar a una sobreestimación de su capacidad predictiva y a una falta de generalización a nuevos datos.

Es importante señalar que la definición de los hiperparámetros de los métodos anteriores requiere una asignación cuidadosa para garantizar la precisión indicada. Sin embargo, esto puede generar problemas de sobreajuste y reducir la generalización del modelo en determinadas condiciones de mercado y horizontes temporales. En cambio, la metodología propuesta que utiliza el análisis discriminante no se ve afectada por la inestabilidad de los hiperparámetros debido al uso de variables predictoras discrimi-

nantes e independientes[57].

La Tabla 8.22 muestra que la metodología propuesta (LCC FE y LCC FS) obtiene la mayor precisión (ACC) con datos *OOS*, con precisiones de 0.975 y 0.987 respectivamente. El modelo predice con gran exactitud si la tendencia a corto plazo del tipo de cambio será alcista o bajista. La metodología propuesta es más competitiva que los enfoques del estado de la técnica, que utilizan redes profundas [577] con una precisión de 0.8955. La ventaja competitiva más importante de la metodología propuesta es la mejora en la precisión de la clasificación sin sacrificar la legibilidad de las interpretaciones (FS). Esto se debe al uso de puntos de inflexión para realizar la predicción, la introducción de nuevos atributos, el proceso de selección de características discriminantes e independientes y su validación multivariante previa a la construcción de una función discriminante lineal.

Tabla 8.22: Evaluación de la precisión de los métodos de predicción

[a] **Metodología**	**TC**	**Datos**	**FPM**	**Muestra**	**Predicción**	**PP**	**ACC**
Ensemble bagging SVM [329]	E/U	IM, IT	FE	2014 - 2019	D, A / B / L	TVT	0.808
Hybrid model [576]	E/U	IM, IT	FS, FE	2010 - 2015	D, A / B	OOS	0.8865
Market sentiment ANN [579]	U/E	IM, RS	FS, FSp	2013	D, A / B	OOS	0.6346
Deep networks CNN [577]	E/U	IM	FEng	2010 - 2015	D, A / B 1-30 P	OOS	0.8955
ANN y DTW [578]	E/U	IM	FS, FE	2011 - 2013	D, A / B	TVT	0.72
LCC (FE)[580]	E/U	IM	FE (15)	**1999 - 2023**	I, A / B	OOS	**0.975** ‡
LCC (FS)[580]	E/U	IM	FS (2)	**1999 - 2023**	I, A / B	OOS	**0.987** ‡

[a]**Convenciones**: **Metodología**: LCC = Enfoque de predicción propuesto, Configurador clasificador lineal para predecir el movimiento direccional del tipo de cambio euro-dólar. **TC**: Tipo de cambio negociado. E/U = (EUR/USD) Euro-dólar americano, U/E = (USD/EUR) dólar-euro. IM = Datos de mercado (OHLCV). IT = Indicadores técnicos. RS = Redes sociales (Stock Twits posts). **FPM**: Métodos de preprocesamiento de características. FE = Extracción de características, FS = Selección de características, FSp = Espacio de características, FEng = Ingeniería de características. **Predicción**: D = Predicción diaria, A = Predicción alcista, B = Predicción bajista, L = Predicción lateralizada. I = Predicción intradía. 1-30 P = Predicción de 1 a 30 periodos en el futuro. **PP**: Capacidad del modelo para predecir nuevos datos fuera del conjunto de entrenamiento. TVT = Entrenamiento, validación y evaluación, OOS = Rendimiento fuera de muestra. **ACC**: Índice de Exactitud. FE(15) = Extracción de características de las 15 primeras variables de la Tabla (5.4). FS(2) = Selección de características discriminantes e independientes (x.V*PT y Sp.Vr*CP.PT) que miden la acción del precio (Ver Tabla 5.4). ‡ Pruebas HSD Tukey, con un nivel de significación inferior al 1%, revelan una diferencia significativa en la medida de exactitud a favor del grupo control.

Por otro lado, La metodología propuesta utiliza un conjunto de datos más amplio y representativo, de 1999 a 2023, en comparación con otros enfoques. Esto implica mayor robustez y capacidad de generalización, en la construcción de modelos de clasificación, evaluada ante diferentes condiciones de mercado. El modelo de clasificación resultante

es más preciso y fiable, lo que aumenta la confianza en la validez de los resultados obtenidos. Al utilizar un conjunto de datos más amplio, se reduce la posibilidad de que los resultados sean un artificio de la muestra específica utilizada en el análisis, lo que aumenta la fiabilidad de la metodología propuesta.

En términos generales, los modelos de clasificación basados en enfoques de inteligencia artificial y aprendizaje automático tienen un impacto determinante en la precisión de la clasificación, así como en el uso de recursos computacionales suficientes para llevar a cabo las fases de entrenamiento, validación y despliegue. Sin embargo, la sencillez del modelo hace que sea fácilmente interpretable (Ver clasificación de la dirección del movimiento 8.3.4), estable (Ver evaluación de la colinealidad de los predictores (F.)), generalizable, escalable y preciso (Ver poder predictivo del modelo de clasificación (8.4.1)).

Parte V

Comentarios finales

9. Conclusiones y trabajo futuro

Desde el punto de vista del análisis técnico y estadístico, los resultados empíricos obtenidos con este estudio demuestran que el problema a resolver: *La construcción de modelos de clasificación parsimoniosos, sencillos y altamente precisos, que permitan explicar y predecir a corto plazo la dirección de la tendencia de los precios como ayuda para la toma de decisiones informadas* se puede lograr a través de cinco fases (preparación de datos, selección de variables, detección de estructuras subyacentes, formulación del modelo de clasificación y evaluación del poder predictivo del modelo). Los resultados demuestran que la construcción de un modelo de clasificación robusto se debe a la consecución de cinco aspectos esenciales:

Primero, a una nueva forma de entender el mercado para predecir su dirección futura. Esto se consigue detectando e identificando los puntos de inflexión del mercado (`ET`). Este elemento critico en el que se origina la tendencia del precio y a partir del cual evoluciona, se utiliza como punto de referencia para medir, clasificar y predecir la dirección del movimiento futuro.

Segundo, a la naturaleza y la estructura multidimensional de las diferencias individuales entre tendencias. Esta visión multivariante del problema resulta ser una alternativa adecuada para detectar el potencial discriminante y predictivo de las variables que mejor explican la acción del precio y que mejor contribuyen a la diferenciación entre movimientos de tendencia. En consecuencia, la precisión y consistencia en la clasificación se debe al carácter discriminante y a la independencia de las variables predictoras seleccionadas según el enfoque estadístico exploratorio-confirmatorio del proceso de selección.

Tercero, a la detección y el análisis de las estructuras subyacentes, que comprueba ser una alternativa adecuada para validar la capacidad discriminante de las variables predictoras seleccionadas y resulta ser una estrategia novedosa para confirmar la idoneidad de la técnica a utilizar en la construcción del modelo de clasificación.

Cuarto, a la significación de la muestra de estudio y a la estructura de los datos, de modo que las variables predictoras y los grupos se ajusten a una distribución normal multivariante. Esta solución innovadora, además de eludir la disposición asimétrica

y leptocúrtica de los datos (retornos logarítmicos), garantiza el cumplimiento de los supuestos estadísticos para la aplicación de los métodos paramétricos multivariantes y contribuye a la construcción de modelos de clasificación versátiles con resultados fiables y consistentes. Incluso se adaptan a conjuntos de datos con un mayor número de frecuencias en los que los casos son diferentes y menos iguales.

Quinto, a la evaluación del poder predictivo del modelo de clasificación con los indicadores de desempeño más reportados en la literatura científica. En términos generales, los resultados obtenidos con datos dentro y fuera de la muestra sugieren que el desempeño alcanzado en la clasificación es consistente y preciso. Esta constante se consigue en condiciones normales, es decir, fuera de la influencia de los eventos macroeconómicos o de la publicación de datos fundamentales que generen impulso en los movimientos de los precios.

En consecuencia, esta tesis proporciona una potente herramienta metodológica para predecir la dirección de los precios de los activos financieros, a partir de la formulación de modelos de clasificación parsimoniosos, sencillos y de gran precisión, como ayuda para la toma de decisiones informadas en situaciones de incertidumbre. El enfoque propuesto aprovecha los puntos fuertes de los métodos estadísticos multivariantes a los que se hace referencia en este trabajo e introduce un enfoque de intervención modular que puede resultar atractivo para abordar problemas que requieran algunos o todos los procedimientos aquí definidos. La tasa de precisión en la clasificación del 98.77% lograda con el enfoque propuesto supera los umbrales de precisión publicados, incluso con los métodos más sofisticados descritos actualmente en la literatura científica.

9.1 Conclusiones

- En un esfuerzo por comprender el comportamiento futuro y la tendencia de los mercados, la previsión de los precios de los instrumentos financieros se ha convertido hoy en día en uno de los problemas más desafiantes a los que se enfrentan la industria financiera y los agentes implicados en la negociación. Aunque se han utilizado muchas técnicas eficaces de inteligencia artificial en el ámbito de las series de tiempo y el aprendizaje supervisado, los resultados alcanzados hasta ahora han sido muy prometedores. Sin embargo, los resultados obtenidos con este estudio permiten concluir que estos trabajos tienen claramente algunas limitaciones, en términos de interpretabilidad (el problema de la caja negra), sobreajuste, selección de variables y definición de hiperparámetros para la construcción de los

modelos. En efecto, el problema de la predicción de la dirección del movimiento de los tipos de cambio en el mercado Forex es un tema de investigación abierto que necesita ser considerado y, por tanto, todavía requiere soluciones parsimoniosas, descifrables y más precisas.

- Considerando el creciente interés por el valor de los datos y su influencia en la toma de decisiones de inversión, este trabajo ha llevado a concluir que dos de los criterios en los que se soporta la construcción de modelos de clasificación y predicción más precisos son la calidad de los datos y las actividades críticas definidas en el proceso de preparación y disposición de los mismos. Al explorar este tema, por un lado, la evidencia de este estudio apunta a que, debido al crecimiento exponencial del volumen, la variedad y la velocidad de generación de datos en los mercados financieros, la creación de valor a partir de la preparación y el procesamiento de datos se ha convertido en un problema de creciente interés para la industria financiera y los participantes del mercado. Por otro lado, los resultados de este estudio sugieren que, debido al profundo desconocimiento del funcionamiento de los mercados, se ignoran los datos relevantes desde el punto de vista del análisis técnico y se subestima su valor, utilidad y potencial para generar valor, extraer información y producir conocimiento.

  En este contexto, aunque existen muchos procedimientos que contribuyen a la obtención de datos de calidad para extraer información de ellos, los hallazgos de este estudio indican que no existe un único enfoque estándar. Por lo tanto, en este trabajo, se ha ideado una solución novedosa en la que la preparación y disposición de datos además de incluir las actividades críticas para producir, extraer, compilar, verificar, limpiar, discretizar y ordenar en colecciones de datos, datos precisos, coherentes y de calidad, estratégicamente esté pensada para garantizar el cumplimiento de los objetivos del enfoque propuesto y de las fases de selección de atributos, detección de estructuras y formulación de modelos. Así, con el diseño del proceso de preparación y disposición de datos, se han obtenido resultados precisos en los que se ha contemplado la intervención de tres problemas centrales:

  En primer lugar, el proceso de preparación de datos que conduce a la identificación efectiva de los determinantes del cambio de régimen de mercado, además de validar los puntos de inflexión del mercado (`ET`), contribuye a la definición de la población y la muestra de estudio mediante la selección de instancias. La inter-

vención de este problema sugiere que la captura e identificación de los puntos de inflexión (ET) mediante la construcción e implementación de un modelo de negociación algorítmico de seguimiento de tendencias es una solución novedosa porque proporciona un nuevo marco para comprender y predecir la dirección del mercado. Los puntos de referencia (ET) en los que se produce el cambio de dirección de la tendencia precedente se categorizan en dos clases (movimiento direccional alcista y bajista). Así, la población de estudio se define con las instancias que están a favor de la tendencia dominante. Los hallazgos indican que, las variables observables que miden la acción del precio antes del instante (ET) son las variables candidatas legibles e interpretables que proporcionan la mejor información sobre el cambio de dirección de la tendencia precedente y están disponibles para el proceso de selección de características.

En segundo lugar, los resultados obtenidos con la preparación de datos han revelado que la selección de un subconjunto de variables observadas, además de ser un tema importante por resolver, para la detección de estructuras, es una estrategia novedosa que para este trabajo y futuros estudios sobre el tema ayuda a construir factores comunes, y en este caso, a explicar la naturaleza multidimensional de las diferencias entre los movimientos de tendencia. Los hallazgos han demostrado que, al intervenir este problema mediante la formulación de un modelo de optimización para la selección de un subconjunto de variables cuantitativas, que se incorporan a las variables predictoras seleccionadas y, que mejor explican la prima de riesgo, la acción del precio y validan la dirección de las tendencias definidas, empíricamente ha demostrado ser una solución novedosa. Así, el subconjunto de variables observadas con medidas de adecuación muestral KMO por variables y grupos mayores a 0.5 y medidas de adecuación muestral individual MSA por variable y grupos mayores a 0.5, deben estar presentes al menos en una proporción de 4 por cada 7 variables elegibles. En consecuencia, este enfoque, además de que puede ayudar a mejorar el conocimiento en la búsqueda de variables para la detección de estructuras subyacentes, tiene el potencial de aplicarse a una gran variedad de casos en los que se requiere.

En tercer lugar, considerando la naturaleza de los datos y el problema de su disposición asimétrica y leptocúrtica, se ha ideado una solución innovadora desde la perspectiva de la preparación de los datos, diseñando un modelo de optimización multiobjetivo para la identificación y selección de instancias. Los hallazgos de este

estudio indican que se han obtenido resultados precisos, mostrando que los casos seleccionados son independientes, las variables predictoras seleccionadas y los grupos asumen una distribución normal multivariante, las matrices de varianzas-covarianzas intra-grupos son iguales y la pertenencia de cada caso a un grupo concreto es excluyente. En consecuencia, el modelo de extracción de instancias proporciona una poderosa herramienta metodológica para la conformación de una muestra de estudio representativa, con clases balanceadas y datos con distribución normal multivariante. La importancia de este aporte radica en que, además, de asegurar el cumplimiento de supuestos estadísticos, confiere fiabilidad a los resultados obtenidos en la formulación de modelos de clasificación basados en el uso de métodos estadísticos paramétricos multivariantes. Aunque existen limitaciones en cuanto al número de variables y grupos que participan en el modelo de extracción, en el presente estudio solo se ha examinado la extracción de instancias, que siguen una distribución normal multivariante, en casos en los que sólo intervienen dos variables predictoras y una variable de agrupación compuesta por dos clases. Por tanto, este estudio es el primer paso, y los hallazgos pueden ayudar a resolver esta dificultad presente en un mayor número de variables independientes con un mayor número de clases en la variable de agrupación. Al mismo tiempo, se necesitan más estudios experimentales con otros tipos de instrumentos financieros y en otras aplicaciones para estimar la fiabilidad del modelo de selección de instancias en la extracción de la muestra de estudio y su uso en la formulación de modelos de clasificación.

- En los últimos años, la selección de características ha suscitado un gran interés, especialmente para la construcción de modelos de clasificación que sirvan de apoyo a la toma de decisiones de inversión. Uno de los factores clave para crear modelos más precisos reside en la eficiencia del proceso de selección de características. Al explorar este problema, hay que tener en cuenta al menos dos aspectos. En primer lugar, dada la variedad y el volumen de información potencialmente disponible y utilizable, el desconocimiento de los determinantes del movimiento direccional de los precios de los instrumentos financieros es claramente evidente y puede abordarse como un problema de selección de características. En segundo lugar, el uso de variables irrelevantes y redundantes, además de que no contribuyen a la diferenciación entre movimientos al alza y a la baja afecta negativamente la precisión de los modelos de clasificación. En cualquiera de estos escenarios,

se tiene un problema abierto que, desde el punto de vista del análisis técnico y estadístico, aún debe ser investigado.

Este estudio, que utiliza un enfoque estadístico exploratorio-confirmatorio, aborda este tema desafiante con una solución novedosa y eficaz que, además de mejorar el conocimiento sobre los inductores de movimiento de los precios, permite identificar y seleccionar las variables independientes candidatas que, por su mérito estadístico y su interpretabilidad, explican mejor la acción de los precios, tienen el mayor poder discriminante, producen el mejor desempeño en la clasificación y permiten capitalizar las ineficiencias del mercado a partir de la predicción de la dirección del movimiento futuro.

El procedimiento propuesto puede ser transferible a otros tipos de instrumentos financieros y tiene el potencial de ser utilizado en otros tipos de aplicaciones en las que se requiere detectar el mejor subconjunto de variables predictoras discriminantes e independientes. Además, el enfoque propuesto podría ser una herramienta valiosa para los responsables de la toma de decisiones en el proceso de selección de características.

- El análisis Biplot ha demostrado ser una poderosa herramienta para la inspección de relaciones y patrones de covariación entre variables en grandes matrices de datos multivariantes. Sin embargo, dada la dificultad que implica, desde una perspectiva gráfica multivariante, la detección de estructuras en las que los grupos difieren para sugerir el modelo que mejor se ajusta a la estructura de los datos y validar así el poder discriminante de las variables predictoras a considerar en la construcción de modelos de clasificación, este trabajo, aprovechando la alta calidad de representación que proporciona el análisis $\mathbf{GH-Biplot}$, introduce dos nuevos enfoques exploratorios para el análisis de estructuras. El primer enfoque ayuda a validar, desde una perspectiva exploratoria multivariante, la capacidad discriminante y predictiva de las variables predictoras seleccionadas mediante el estudio de la naturaleza y la estructura de las diferencias individuales de los movimientos de tendencia. En términos generales, esto implica el análisis detallado de las interacciones entre variables, la detección de estructuras significativas e interpretables y el reconocimiento de las diferencias de perfil entre grupos. El segundo enfoque ayuda a sugerir y/o determinar la idoneidad del método de clasificación a utilizar en la predicción. El procedimiento propuesto para el diagnóstico de modelos incluye la inspección y confirmación de relaciones

y patrones modelables en las estructuras de datos, la identificación del modelo más apropiado y la evaluación del ajuste del modelo a la estructura de datos.

Al abordar este problema, los resultados obtenidos con este estudio permiten concluir que el diagnóstico de variables predictoras y de modelos es un proceso de verificación a posteriori. Se lleva a cabo después de conocer las variables y la técnica a utilizar en la construcción del modelo de clasificación y se realiza sobre el conjunto de variables que incluyen las que mejor contribuyen a la construcción de estructuras, según el enfoque propuesto en la preparación de datos. Así, el análisis de los factores subyacentes, además de ayudar a identificar y definir los perfiles de las instancias (ET), según rasgos intrínsecos poco comunes, ayuda implícitamente a reconocer el poder discriminativo de las variables predictoras y a determinar su idoneidad en la construcción de los modelos de clasificación.

Empíricamente, la importancia de los enfoques propuestos radica en que, además de demostrar la idoneidad de los datos para la detección y el análisis de estructuras, las variables predictoras seleccionadas y las funciones lineales de Fisher contribuyen a la diferenciación y la separación de los casos entre grupos. La medida KMO de Kaiser-Meyer-Olkin de adecuación del muestreo, las pruebas de igualdad de medias de grupos (Análisis ANOVA de una vía) y de las diferencias entre las medias de las funciones lineales de Fisher (Estadístico Lambda de Wilks) lo confirman respectivamente.

Desde el punto de vista del análisis de la estructura de los datos, la evidencia sugiere que el primer eje factorial (comparte el 37.71% de la variabilidad de las variables observadas) y la variable predictora x.V*PT son los que mejor explican y contribuyen a la diferenciación entre los movimientos de tendencia. La media de los datos de este predictor es el mejor estimador de las diferencias entre grupos y es el más influyente en la construcción de un modelo de clasificación. En general, los resultados obtenidos con el análisis **GH − Biplot** explican de forma novedosa en cuatro cuadrantes la naturaleza multivariante de las diferencias entre los movimientos al alza y a la baja, en el nivel vertical se explica el valor del tipo de cambio en los puntos de inflexión del mercado (ET), los niveles mínimos/máximos en los que el tipo de cambio se deprecia/aprecia se explican con el primer eje factorial. Así, los cuadrantes II y III definen la zona de sobreventa y los cuadrantes I y IV la zona de sobrecompra. Por otro lado, el nivel horizontal explica la prima de riesgo a la que está expuesto el tipo de cambio negociado y está dominado por

el segundo eje factorial. Así, los cuadrantes I y II son la zona de alta exposición al riesgo, y los cuadrantes III y IV son la zona de baja exposición al riesgo.

- En cuanto a la formulación de modelos de clasificación, si se tienen en cuenta los diversos esfuerzos y la relevancia que ha alcanzado el análisis de las series de tiempo y el uso de técnicas eficaces de aprendizaje automático en el campo de la inteligencia artificial para predecir la dirección de los precios de los activos financieros, este trabajo ha permitido concluir que la construcción de modelos de clasificación parsimoniosos, descifrables y más precisos a partir de datos de mercado es una tarea difícil y compleja que requiere un conocimiento profundo del mercado y debe apoyarse en estrategias de negociación fiables diseñadas a partir de un adecuado análisis técnico y fundamental.

  En este contexto, la evidencia de este trabajo sugiere el desarrollo de un nuevo enfoque para construir modelos de clasificación y predecir en el corto plazo la dirección del movimiento de los precios de los activos financieros. La metodología propuesta aborda esta difícil tarea como un problema de clasificación supervisado utilizando un enfoque estadístico multivariante y paramétrico. Los resultados de este estudio indican que, además de proporcionar nuevos métodos de intervención modular basados en la preparación de los datos, la selección de características y la detección y el análisis de estructuras subyacentes, cuyas soluciones se integran en la construcción de modelos de clasificación más precisos, Indirectamente la solución propuesta proporciona un marco para comprender, desde el punto de vista del análisis técnico y el modelado financiero, el proceso que conduce a los precios de los instrumentos financieros a la formación de niveles históricos de soporte y resistencia en los que la dirección del movimiento futuro también puede ser potencialmente predecible.

  En consecuencia, en este documento se ha destacado que el enfoque propuesto simplifica la complejidad del mercado en dos variables predictoras independientes con alto poder discriminante. El rendimiento medio de los precios de cierre x.V*PT y la pendiente de la recta de regresión de las variaciones entre los precios de cierre Sp.Vr*CP.PT, ambos valores se calculan 18 períodos antes del punto de referencia (`ET`) (Este número de periodos se elige para el cálculo de las métricas aquí descritas porque produce el mejor desempeño en las tareas de clasificación). Por lo tanto, se ha demostrado que la función lineal discriminante explica, de forma inteligible en los puntos de inflexión del mercado (`ET`), la influencia que

tienen las variables predictoras en la acción del precio en la formación de los movimientos de tendencia, y predice la dirección del movimiento futuro.

En general, se han aportado importantes evidencias que confirman que la dirección del movimiento después del instante (ET) está determinada por la dirección y la fuerza del impulso generado antes de los puntos de inflexión (ET), siendo de dos tipos: (i)movimientos bajistas/alcistas con impulso, en donde el tipo de cambio refleja valores de sobreventa/sobrecompra y (ii)movimientos bajistas/alcistas sin impulso. Los hallazgos sugieren que, cuando los movimientos bajistas/alcistas antes del instante ET vienen con impulso, en la función lineal de Fisher que predice la dirección del movimiento alcista/bajista, los coeficientes de los predictores son menores/mayores que cero, la contribución del rendimiento medio x.V*PT es mayor que la contribución de Sp.Vr*CP.PT. Así, cuanto más cerca esté el rendimiento medio del valor mínimo/máximo y el valor de la pendiente de estas variaciones Sp.Vr*CP.PT sea también negativo/positivo, menos probable será que el valor del euro frente al dólar americano siga depreciándose/apreciándose con el tiempo.

Por otro lado, las evidencias de este trabajo indican que, cuando los movimientos bajistas/alcistas anteriores al instante (ET) vienen sin impulso, según la función lineal de Fisher el inicio de un nuevo movimiento alcista o bajista posterior al evento (ET) es más probable que se produzca cuando el rendimiento medio negativo o positivo del tipo de cambio disminuye o aumenta por encima o por debajo del primer cuartil de los valores de rendimiento históricos y el valor de la pendiente de las variaciones entre los precios de cierre sea positivo o negativo.

- El uso de medidas de bondad de ajuste e indicadores de desempeño utilizando datos dentro y fuera de la muestra se ha convertido en un tema de creciente interés en la construcción de modelos de clasificación debido a la necesidad de verificar y mejorar el poder predictivo de dichos modelos. Los hallazgos de este estudio sugieren que, aunque existen muchos criterios para evaluar el desempeño de los modelos de clasificación, no existe un procedimiento estándar único. Sin embargo, las medidas de desempeño obtenidas utilizando las métricas más comúnmente reportadas en la literatura científica (Accuracy, G-mean accuracy, Recall, Specificity, precision, F1, Kappa, Baseline y MCC) superan el desempeño de la clasificación de los métodos más sofisticados publicados hasta la fecha. La evaluación del proceso de clasificación indica que se han obtenido índices de precisión más exactos,

lo que demuestra que el enfoque propuesto proporciona una poderosa herramienta metodológica para construir modelos de clasificación más precisos como ayuda para tomar mejores decisiones de inversión.

La evaluación del poder predictivo con datos dentro y fuera de muestra en un horizonte de aproximadamente 24 años, demuestra que la combinación lineal de las variables predictoras seleccionadas, según el análisis de estructura, y los resultados de las pruebas estadísticas paramétricas, además de proporcionar evidencia estadísticamente significativa de las diferencias entre los movimientos de tendencia, maximiza la precisión de la clasificación. Por otra parte, al mismo tiempo que los supuestos estadísticos se apoyan en pruebas de significación y respaldan la eficacia del modelo de clasificación a partir de pruebas de bondad de ajuste, estos resultados confieren fiabilidad a los niveles de soporte y resistencia definidos a partir de los precios históricos de mercado. Así, el uso de las zonas de soporte y resistencia en las que el activo alcanza niveles de sobreventa y sobrecompra, además de ser fiables, son útiles para gestionar la relación riesgo-beneficio en el desarrollo de posiciones de compra y venta.

En cuanto al número de instrumentos financieros analizados, aunque el alcance de este estudio es limitado, es el primer paso para mejorar la comprensión del funcionamiento del mercado y mejorar la precisión en la predicción de la dirección del par euro-dólar. Además de contribuir con un nuevo enfoque a la construcción de modelos de clasificación con fines predictivos para la toma de decisiones de inversión, la metodología propuesta puede ser transferible. Sin embargo, se necesitan más estudios experimentales con otros tipos de activos y en otras aplicaciones para estimar la fiabilidad del enfoque propuesto en la construcción de modelos de clasificación parsimoniosos, descifrables y precisos. Además, esta investigación podría ser una ayuda útil para los responsables de la toma de decisiones de inversión en la ejecución de operaciones en el mercado. Los resultados del enfoque propuesto permiten integrar las previsiones del modelo de clasificación y el análisis de las zonas de sobrecompra y sobreventa del activo negociado para confirmar y aumentar la eficacia de las órdenes de compra y venta antes de su ejecución.

9.2 Trabajo futuro

Este trabajo es el primer paso para mejorar la comprensión del funcionamiento del mercado en lo que respecta al proceso seguido por los precios de los instrumentos financieros en la definición de los puntos de inflexión y los niveles históricos de soporte y resistencia alcanzados en la negociación. Los resultados obtenidos con el enfoque propuesto tienen muchas implicaciones, especialmente en la definición de los determinantes del movimiento direccional y la predicción del movimiento futuro. En este sentido, el trabajo futuro se centra en utilizar la metodología propuesta para evaluar su eficacia con otros tipos de instrumentos financieros, especialmente sobre la base de estrategias de negociación fiables y consistentes.

Los resultados obtenidos en esta investigación pueden ayudar a mejorar el desempeño financiero de las estrategias de negociación de instrumentos financieros. La integración de los resultados proporcionados por los niveles de sobreventa y sobrecompra y las previsiones del movimiento direccional de los precios de los instrumentos financieros es una potente herramienta de ayuda a la toma de decisiones de inversión. Esta información es útil para confirmar las órdenes de compra y venta antes de su emisión. En este contexto, la implementación de estas herramientas de apoyo a las decisiones de inversión en el sistema de negociación algorítmico es un futuro trabajo en curso. El estudio se centra en analizar y evaluar sistemáticamente, en tiempo real, el desempeño del sistema de negociación. La evaluación permitirá mejorar ampliamente la eficacia del sistema de negociación y del modelo de previsión ante la volatilidad del tipo de cambio euro-dólar.

El enfoque propuesto en la construcción de modelos de clasificación para predecir la dirección de los precios a corto plazo proporciona una precisión de clasificación excepcionalmente alta. Sin embargo, se necesitan más estudios experimentales para estimar la idoneidad de la metodología propuesta y de las variables predictoras seleccionadas en la construcción de otros modelos de clasificación. En este sentido, esta investigación ha planteado la exploración de otros métodos de aprendizaje supervisado. El trabajo futuro previsto se centra también en la evaluación comparativa del desempeño proporcionado por otras técnicas de clasificación.

En conjunto, los resultados de este estudio también son aplicables en la construcción de un indicador de análisis técnico. Para avanzar en esta investigación, el trabajo futuro

se centra en la construcción de un asesor experto o sistema de recomendación adaptable a cualquier tipo de instrumento financiero negociado en cualquier marco temporal. La dirección estimada del precio se basa en las puntuaciones discriminantes. Estos valores se obtendrán a partir del rendimiento medio y la pendiente de la recta de regresión de las variaciones entre los precios de cierre, ambos indicadores calculados n períodos históricos. El indicador actuaría como una herramienta de apoyo a la decisión de inversión basada en la identificación de niveles de soporte y resistencia en los que el valor del activo está infravalorado o sobrevalorado.

Los procedimientos propuestos en este trabajo constituyen una valiosa ayuda para la solución de los problemas aquí planteados. Para facilitar la aplicación de estos procedimientos en otro tipo de estudios, el trabajo futuro se centrará en desarrollar estos procedimientos en una interfaz gráfica en el entorno R. Las aplicaciones más importantes son: (1) Modelo para la identificación y definición de los puntos de inflexión de los precios de los activos financieros negociados en el mercado para diferentes marcos de tiempo. (2) modelo de extracción de un subconjunto de instancias con medidas que siguen una distribución normal multivariante. Este primer enfoque se aplica a las medidas cuantitativas dispuestas en dos variables independientes y una variable de agrupación con dos clases. (3) Modelo de selección de atributos para la construcción de factores comunes. El subconjunto de variables dispuestas para la detección y el análisis de estructuras subyacentes ayuda a validar la idoneidad de las variables predictoras seleccionadas y del modelo de clasificación que se utilizará con fines predictivos. (4) Modelo de selección de variables predictoras independientes, según mérito estadístico, con alto poder discriminante y capacidad predictiva. Este método es adecuado para definir las variables a considerar en la construcción de modelos de clasificación con fines predictivos. (5) Modelo general para la construcción de una función discriminante lineal, que integra la preparación de datos, la selección de variables predictoras y la validación de la eficacia del método y los predictores seleccionados.

Finalmente, una cuestión que aún debe investigarse, desde el punto de vista del análisis técnico, es la relación de causalidad o de covariación entre las variables predictoras seleccionadas y los tiempos de duración de los movimientos tendenciales de los precios. En este contexto, un problema importante que se debe abordar en futuros estudios es la predicción de la duración de los movimientos tendenciales de los precios.

Parte VI

Apéndices

10. Producción académica

Durante el exigente proceso de investigación que ha conducido al desarrollo de esta tesis doctoral, se ha logrado una importante producción académica que abarca diversos aspectos de relevancia científica. Entre los logros más destacados, se ha incluido la participación activa en congresos nacionales e internacionales, donde se han compartido y discutido, con la comunidad científica, los hallazgos y avances de esta investigación. Esta valiosa interacción ha permitido mejorar y enriquecer la calidad del presente trabajo, contribuyendo al fortalecimiento de este estudio.

Entre las participaciones mas significativas esta la asistencia como ponente a tres relevantes congresos internacionales en Almería, España; Covilã, Portugal; y San Juan de Pasto, Colombia. Entre los principales resultados de investigación en estadística multivariante aplicada a los mercados de capitales, destacan la identificación de los determinantes de los rendimientos positivos en la negociación del tipo de cambio euro-dólar y el innovador enfoque "Biplot logístico" aplicado a una estrategia de negociación del par de divisas euro-dólar. También se han realizado estudios de Análisis Multivariante de Sistemas de Especulación centrados en la negociación del tipo de cambio euro-dólar. La divulgación de estos resultados de investigación se resume a continuación:

- **Ponencia Resultado de investigación:** Determinantes de las posiciones de mercado con retornos positivos: Un primer enfoque a partir del análisis discriminante. XXI International Congress of the Word Economy Society / Sociedad de Economía Mundial. University of Beira Interior - Covilã - Portugal. 2019.

- **Conferencia Magistral Resultado de investigación:** Análisis Multivariante de Sistemas de Especulación. 2do Congreso Internacional de Ciencia, Tecnología, Innovación y Desarrollo Territorial. Universidad Autónoma de Nariño – San Juan de Pasto – Nariño – Colombia. 2018.

- **Ponencia Resultado de Investigación:** Biplot Logístico aplicado a una estrategia de negociación que utiliza el par de divisas Euro/Dólar. XX Reunión

de Economía Mundial. XX World Economy Meeting. Universidad de Almería – España. 2018.

La participación en estos congresos internacionales refleja un alto compromiso con la investigación y la generación de conocimiento en el ámbito financiero. El intercambio de ideas ha enriquecido este trabajo de tesis y ha generado oportunidades de colaboración y desarrollo. Además, la relevancia y originalidad de los temas presentados han sido reconocidos por la comunidad científica, reafirmando su valor y potencial impacto en el ámbito financiero.

Por otra parte, la vinculación generada por estos procesos de investigación con grupos de investigación e instituciones brinda la oportunidad de realizar estancias para mejorar la calidad de los resultados de investigación publicados en esta tesis. Durante una pasantía, se llevaron a cabo las siguientes actividades en la Escuela de Ciencias Matemáticas y Computacionales de la Universidad de Investigación de Tecnología Experimental Yachay Tech:

- Exploración del problema de la predicción de la tendencia de los precios de los activos financieros mediante la definición e integración de estrategias de negociación y modelos de previsión.

- Definición y selección de indicadores y medidas determinantes del movimiento direccional de los precios para la construcción de modelos de clasificación, utilizando técnicas estadísticas, de aprendizaje automático y de inteligencia artificial.

Estas actividades de investigación han fortalecido significativamente los conocimientos en el ámbito financiero y han impulsado la búsqueda de soluciones eficaces y precisas. La colaboración con expertos y la inmersión en un entorno académico y de investigación han demostrado ser esenciales para garantizar que los resultados obtenidos tengan un impacto potencial en el ámbito académico.

La publicación de estos resultados de investigación en revistas científicas de alto impacto no solo asegura la revisión y validación por parte de expertos a través del proceso de revisión por pares, sino que también fortalece la solidez y confiabilidad de los hallazgos al difundirlos ampliamente dentro de la comunidad científica y académica.

En este contexto, los resultados de la investigación de esta tesis doctoral han sido publicados en *IEEE Access*, una prestigiosa revista de acceso abierto que publica artículos de resultados de investigación en el campo de la tecnología y la ingeniería,

especialmente en la categoría de "Ciencias de la Computación y Sistemas de Información".

La revista, a julio de 2022, tiene un cuartil Q1 en SJR, lo que indica que se encuentra en el primer cuartil en términos de visibilidad y prestigio dentro de su campo a nivel mundial. Por otro lado, en JCR tiene un cuartil Q2, lo que sugiere que se encuentra en el segundo cuartil en términos de su Factor de Impacto en comparación con otras revistas en su categoría específica.

El artículo publicado lleva por título: "A Novel Linear-Model-Based Methodology for Predicting the Directional Movement of the Euro-Dollar Exchange Rate". En este artículo, que se adjunta a continuación, se presenta una innovadora metodología basada en modelos lineales para predecir el movimiento direccional del tipo de cambio euro-dólar, destacando la relevancia y el valor de los resultados obtenidos.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 May 2023, accepted 2 June 2023, date of publication 9 June 2023, date of current version 10 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3285082

RESEARCH ARTICLE

A Novel Linear-Model-Based Methodology for Predicting the Directional Movement of the Euro-Dollar Exchange Rate

MAURICIO ARGOTTY-ERAZO[1,2], ANTONIO BLÁZQUEZ-ZABALLOS[1], CARLOS A. ARGOTY-ERASO[3], LEANDRO L. LORENTE-LEYVA[2,4], NADIA N. SÁNCHEZ-POZO[5], AND DIEGO H. PELUFFO-ORDÓÑEZ[2,6]

[1]Department of Statistics, University of Salamanca, Campus Miguel de Unamuno, 37007 Salamanca, Spain
[2]Smart Data Analysis Systems Group (SDAS) Research Group, Ben Guerir 43150, Morocco
[3]Facultad de Ciencias Económicas y Administrativas, Universidad de Nariño, Torobajo, San Juan de Pasto 52001, Colombia
[4]Facultad de Derecho, Ciencias Administrativas y Sociales, Universidad UTE, Quito 170147, Ecuador
[5]Centro de Posgrado, Universidad Politécnica Estatal del Carchi, Tulcán 040101, Ecuador
[6]College of Computing, Mohammed VI Polytechnic University, Hay Moulay Rachid, Ben Guerir 43150, Morocco

Corresponding author: Mauricio Argotty-Erazo (mauricio.argotti@usal.es)

This work was supported in part by the College of Computing, Mohammed VI Polytechnic University; and in part by the Smart Data Analysis Systems Group (SDAS) Research Group (https://sdas-group.com/).

ABSTRACT Predicting the price and trends of financial instruments is a major challenge in the financial industry, impacting investment decision-making efficiency for various stakeholders. Although numerous and effective artificial intelligence techniques have been applied to time series analysis, the prediction of exchange rate movements in the Forex market still necessitates parsimonious, interpretable, and accurate solutions. This paper presents a novel methodology for predicting the short-term directional movement of the euro-dollar exchange rate using market data, specifically by measuring price action. The proposed methodology prioritizes using market inflection points and the multidimensional nature of the differences between uptrends and downtrends to construct a linear discriminant function (LDA). The core of our methodology is our novel Linear Classifier Configurator (LCC) which includes stages for data preparation, feature selection, and detection of underlying structures. We validate the results and interpretations using the statistical power of parametric tests. The experiments use market data of the euro-dollar exchange rate in 15-minute and 1-week time frames. Additionally, we incorporate a collection of intraday winning trades provided by an algorithmic trading model applied between January 1999 and April 2023. The proposed LCC methodology achieves an out-of-sample classification accuracy of 98.77%, outperforming other methodologies based on sophisticated approaches such as Long Short-Term Memory (LSTM), Deep reinforcement learning (DRL), Wavelet analysis (WA), Sentiment analysis of textual content, Support Vector Machines (SVM), and Genetic Algorithms (GA). Furthermore, our methodology improves financial performance and reduces risk exposure in trading strategies, as well as it is useful in selecting variables and transferable to other financial assets.

INDEX TERMS Linear discriminant analysis (LDA), foreign exchange market (FOREX), machine learning (ML), supervised learning (SL), time series forecasting (TSF), trading systems.

I. INTRODUCTION

The FOREX or FX market, also known as the "Over the Counter" (OTC) market, is the world's largest financial market with a wide range of players, from financial institutions to individual investors, participating in it. In this market, currency values are defined and currency exchange is encouraged. Recently, the market has experienced rapid growth, making it easier to access currency trading, and introducing new technological challenges in the current trading

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

and regulatory models. This has awakened a growing interest among investors, practitioners, and researchers to predict future price behavior. The euro-US dollar cross (EUR/USD) is one of the most traded financial instruments, and is particularly interesting to investors because the strength of currencies is determined by the level of economic activity in the economies they represent [1].

Currently, exchange rate volatility and the effects generated on the value of currencies are strongly influenced by market liquidity and the interaction of economic, political, psychological, and non-fundamental variables [2]. The 2008 financial crisis and the Covid-19 pandemic have emphasized the importance of monitoring market dynamics to identify, anticipate, and mitigate adverse movements, reinforcing both financial and operational resiliencies. Market participants now require the ability to detect and monitor inflection points and the directional movement of exchange rates to manage their utilization effectively. Thus, informed decision-making is crucial in enhancing trading, minimizing risk exposure, and enabling the timely development of hedging operations to safeguard investment value.

Traditionally, asset value prediction revolves around two main approaches: technical analysis and fundamental analysis [3]. The former relies on studying historical market data and visually identifying patterns in price charts to predict their trends [4]. On the other hand, from a fundamental analysis perspective, proponents of the efficient markets hypothesis [5] argue that the price of a financial asset reflects all market information, asserting that price changes are random and, consequently, future behavior is unpredictable. Despite initial research supporting these postulates, subsequent studies by [6], [7], [8], [9], [10] demonstrate that historical data can predict returns and prices. As a result, the predictability of asset prices and trends remains a debated research topic.

In artificial intelligence, researchers have tackled price forecasting as both a supervised and unsupervised classification problem. A literature review reveals that this complex issue has been examined through statistical, machine learning, and deep learning techniques [11], as well as hybrid models [11], [12]. Additionally, alternative approaches have utilized sentiment analysis of news and text mining in social networks [13], [14], [15]. In recent years, the authors in [16] have observed a strong preference for deep learning techniques, especially in economic and financial applications. These techniques are increasingly used due to their high accuracy in classification and prediction tasks [17], [18]. Although these approaches can identify patterns or trends in historical data, they cannot directly explain the influence of predictor variables on price developments. Due to their black-box nature, in addition to model complexity and lack of readability, they pose problems of overfitting, generalization, and hyperparameter definition [19], [20], [21], [22], the solution of which entails significant efforts. Unlike the methods above, Discriminant analysis does not suffer from the same issue. It is less susceptible to parameter instability and offers competitive interpretability and precision when prior knowledge exists regarding the discriminant variables that effectively distinguish between groups [23].

In particular, linear discriminant analysis (LDA) is the most widely used classification method since it was proposed by Fisher [24], [25], [26]. Although new variants appear in the specialized literature, among the most diverse approaches, the following methods stand out: Quadratic (QDA) [27], Heteroscedastic (HDA) [28], Regularized (RDA) [29], Sparse Linear Discriminant Analysis (S-LDA) [30] and High-Dimensional Discriminant Analysis (HDDA) [31]. LDA is a multivariate parametric statistical method particularly relevant in the financial field for classification and prediction tasks. Besides enhancing the discriminative power of the predictor variables when they are useful for generating significant differences between classes, the main advantage of LDA over other techniques is its ability to incorporate multiple quantitative variables in the analysis. Thus, classification can be approached as a dimensionality reduction problem or, ideally, by selecting the best subset of predictor variables to ensure model interpretability while achieving accurate classification between groups.

Empirical evidence has shown that discriminant analysis is effective in classification when predictor variables produce significant differences between groups. Several noteworthy studies use discriminant analysis and novel information to predict value-at-risk (VaR) and financial instrument prices. These studies use data from investor information [32], descriptive statistics [33], financial ratios [34], news headlines [35], [36], and fundamental data extracted from financial news (twits) [37]. The classification results show that using valuable information with high discriminant power improves the models' predictive power and classification accuracy. Discriminant analysis, as shown in [23], [37], [38], and [39], outperforms certain machine learning algorithms. This supports the notion that LDA is well-suited for forecasting asset price trends. By effectively capturing patterns in extensive datasets and leveraging the multidimensional distinctions between groups, LDA improves accuracy while maintaining interpretability in classification. However, discriminant analysis is less common than other artificial intelligence techniques in predicting price trends and logarithmic asset returns. One contributing factor to its limited usage is meeting the statistical assumptions necessary for its application. This challenge arises when dealing with asymmetrically and leptokurtically distributed data, preventing the variables from assuming a multivariate normal distribution [40], [41]. Although other econometric approaches have also explored this issue [42], [43], [44], [45], [46], [47], [48], [49], [50], the problem of non-normal data distribution in parametric models continues to be a topic of ongoing research and debate. This limitation hampers the use of specific techniques due to the data structure.

Given this context, there is no definitive solution to the problem of price trend prediction. While various approaches have been proposed, current research has not yielded a conclusive answer, leaving room for further investigation and

potential breakthroughs. Consequently, there is a significant demand for the development of parsimonious, simple, and highly accurate forecasting models that can effectively identify patterns or trends in historical data for predicting future price movements. Despite the complexity of the relationship between predictor variables and price developments, formulating classification models that elucidate their influence remains a critical challenge, as it enables more precise, well-informed, and value-driven investment decisions [51].

Research papers thus far have predominantly explored this phenomenon with an emphasis on technique. These studies use a wide variety of data and variables (market data, technical and fundamental indicators) to discover valuable insights that improve forecasting accuracy. However, most of these studies overlook the context and the moments leading up to trend changes. Taking into account these observations, our approach departs from the methodologies employed in the cited articles.

The main contribution of this work, from the technical and statistical analysis field, is to provide a valuable methodological tool to build interpretable, parsimonious, and accurate classification models.

Our approach aims to improve investment decision-making by predicting short-term trends in the euro-dollar exchange rate while maintaining interpretability.

This study introduces a novel framework for understanding and predicting price direction by incorporating market turning points. It also introduces two new variables that effectively discriminate between upward (bullish) and downward (bearish) trends. Additionally, the study applies discriminant analysis to multivariate normal data drawn from an asymmetric and leptokurtic distribution, contributing to its originality.

The novelty of the proposed methodology lies in the systematic use of a well-defined set of procedures, as shown in Figure 1. These processes include data preparation, selection of discriminant and independent features, structure detection to validate the discriminant power of the selected variables, and integration of these partial solutions to construct a linear discriminant function. This function provides a more precise depiction of price action to predict future trends. Moreover, statistical significance tests duly support the quality and reliability of the results and interpretations in each stage of the proposed methodology.

Our proposed methodology achieves a remarkable out-of-sample classification accuracy of 98.77% using 15-minute trading sessions of the euro-dollar exchange rate market data from January 1999 to April 2023. This performance surpasses even the most sophisticated methods discussed in recent works, including [52], [53], [54], [55], [56], [57], and [58].

Consequently, the findings suggest that the change in direction is influenced by the average yield of closing prices and the slope of the regression line of these variations. Thus, at inflection points **ET** (where the change in trend takes place), an bullish/bearish movement is more likely to occur when the average yield decreases/increases above/below its historical average value and the slope of the regression line of the variations between closing prices is negative/positive. In addition, predictions can confirm buy or sell orders before placing them, which increases the probability of successful execution of orders. In predictor variables, our approach helps to identify support and resistance levels at which the price changes its trend. In addition, this method is suitable for trading strategies that work in different time frames.

This paper is structured in the following sections: Section II provides a synthesis of the state of the art of the field of study. Section III describes the structures of the data collections and the proposed methodology. Section IV presents the experiment setup and performance measures used. Section V provides the results and discussion of the executed experiments. Finally, section VI summarizes the study's conclusions.

II. RELATED WORKS

Accurate prediction of price movements in the foreign exchange market is crucial for making informed investment decisions. However, the lack of interpretability in predictive models can hinder the formulation of judgments before making investment decisions. While model development has traditionally emphasized technique and accuracy, there is a growing need to develop predictive approaches that strike a balance between accuracy and interpretability [59].

Ge et al. [60] and Aziz et al. [61] highlight numerous studies in the scientific literature that use statistical techniques, machine learning, and artificial intelligence algorithms to address the problem of price trend prediction. Commonly used approaches include artificial neural networks (ANNs), support vector machines (SVMs), wavelet analysis, textual content sentiment analysis, and genetic algorithms [61]. Deep learning techniques, such as LSTM and deep reinforcement learning, have been utilized to enhance prediction accuracy [62]. Additionally, several studies use SVM and genetic algorithms to reduce dimensionality and improve prediction accuracy [63]. Moreover, textual content sentiment analysis has been leveraged to understand the impact of news on market behavior and achieve a higher prediction accuracy [64].

Although these studies have improved the accuracy of predictions, the importance of the interpretability of the results has often been underestimated. In finance, the explanatory power of predictive models is essential in making investment decisions [65]. There remains a need to develop predictive approaches that balance accuracy with interpretability, generating more compelling results and supporting investment decisions with clear and understandable information [59].

In this context, discriminant analysis is a useful statistical technique that has proven its effectiveness in terms of accuracy and explanatory power in different financial applications [66]. Although less widely used in price prediction than machine learning algorithms, several studies use discriminant analysis to predict the direction of price movements in different financial assets, such as cryptocurrencies, exchange rates, stocks, investment portfolios, and stock market indices.

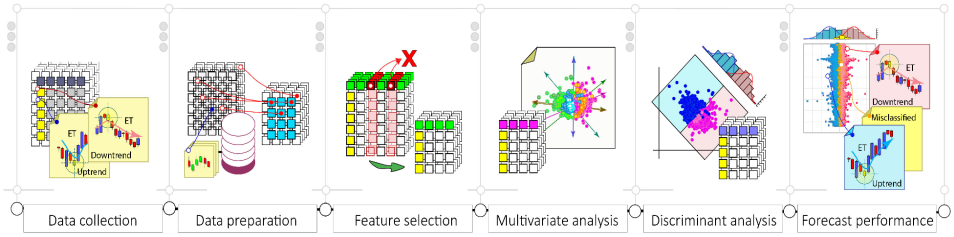


FIGURE 1. High-level diagram of the proposed methodology, Outlines the stages of data preparation, feature selection, and structure detection to predict the directional movement of the euro-dollar exchange rate. Additionally, it highlights the benefits of using the multidimensional nature of the differences between trend movements to construct a linear discriminant function.

A. CRYPTOCURRENCIES

Several studies have demonstrated the effectiveness of discriminant analysis in predicting digital assets. Grobys and Sapkota [32] achieve an 87% accuracy in classifying 146 digital assets using LDA and discriminant variables. The results show that reducing data dimensionality and selecting only the most informative variables improves model performance and allows investors to make informed decisions on high-default-risk assets. Gurrib and Kamalov [35] use sentiment analysis and LDA to predict the direction of bitcoin (BTC). With an accuracy of 58.5%, lower than that achieved by Grobys and Sapkota [32], they demonstrated the potential of Natural Language Processing (NLP) in extracting valuable information from heterogeneous data to design classification models for decision-making. Chen et al. [38] demonstrate the superiority of LDA and LogR over various machine learning algorithms (Quadratic Discriminant Analysis, Support Vector Machine, Random Forest, XGBoost, and Long Short Term Memory) in daily Bitcoin prediction, with an accuracy of 66%. These studies suggest that prediction accuracy does not necessarily depend on the complexity of the technique employed. However, using linear combinations of the original data as input variables may result in a loss of interpretability in predictions. In contrast, our proposed methodology overcomes this limitation, improving the transparency of predictions and facilitating informed decision-making.

B. EXCHANGE RATES

Several authors stress the importance of selecting the appropriate model and transforming the data to improve the accuracy of predictions. Zlicar et al. [67] achieved a 53.25% accuracy in predicting the direction of the most traded exchange rates using a Parzen sequential windowing (PW) algorithm for mapping price data. The results of this approach suggest that proper selection of the learning algorithm (among several methods evaluated, such as SVM, PW, and FDA) is critical to improving classification accuracy. On the other hand, Steurer et al. [68] focus on predicting daily exchange rate movements using machine learning techniques such as ANN, LDA, and LR. Their study also underscores the importance of selecting a suitable model and transforming the data to improve the accuracy of the predictions. Although the appropriate modeling technique choice depends on the problem and the data, it is worth noting that the most complex and attractive methods are not always the most effective. Therefore, it is essential to establish a coherent connection between the objective, the method, and the data used to obtain robust and reliable results. In the proposed methodology, this connection ensures the results' robustness and applicability in informed decision-making.

C. EQUITY INSTRUMENTS

Recent stock market studies have demonstrated the effectiveness of discriminant analysis in building predictive models and improving forecast accuracy. Mndawe et al. [37] use sentiment analysis to predict the stock price trend of South African companies. Their experiments reveal that LDA outperforms Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) with 94% and 96% accuracy. The results highlight the relevance of using fundamental data from news headlines and social media (tweets). Meanwhile, Iworiso and Vrontos [23] propose an approach to predict the direction of the U.S. equity risk premium as a function of 14 financial performance variables. The experiments demonstrate the superiority of high-dimensional discriminant analysis (HDDA), achieving a 67% classification accuracy, while quadratic discriminant analysis (QDA) generates the highest cumulative returns. Bernardi et al. [36] use discriminant analysis to create sentiment indicators and reduce expected VaR violations. Their study focused on examining the influence of news on financial return prediction. The authors also demonstrated that incorporating exogenous quantitative variables from market news improves prediction accuracy. However, despite the accurate results obtained from these studies using fundamental data and discriminative exogenous information, the interpretability and transparency of the models are constrained due to feature extraction. In the proposed methodology, we handle these limitations through appropriate feature selection techniques.

Dao and Ahn [69] use a support vector machine (SVM) fuzzy logic model to predict stock market movements. Among other statistical methods, the authors use LDA to compare machine learning models (multiple discriminant analysis MDA, LogR logistic regression, CART regression

tree, ANN artificial neural network, SVM, and fuzzy SVM models). Gorenc and Dejan [70] and Sharma et al. [71] focus on the use of statistical classifiers to improve the accuracy of predicting the movements of S&P 500 and ICICI Bank stock prices, respectively. Although these approaches use discriminant analysis as a benchmark to compare the effectiveness of other classifiers, feature extraction compromises the interpretability of the predictions of the best approaches. The proposed methodology ensures model transparency and interpretability by selecting readable features, which facilitates, facilitating decision-making processes.

D. INVESTMENT PORTFOLIOS

The use of discriminant analysis has proven effective in structuring investment portfolios with potentially profitable, low-risk assets.

Weiss [33] studies a portfolio classification approach that predicts the VaR and expected loss of 1500 bivariate portfolios containing stock, commodity, and currency futures data. The approach uses LDA to generate a parametric VaR and expected shortfall (ES) risk model. This model uses the descriptive statistics of the bivariate distribution of the portfolios as attributes, categorizing them according to the VaR and ES categories of the dependent variable. The results show a classification accuracy of 67.53% for VaR and 53.80% for ES. Although the interpretability of predictions and the model transparency are compromised, using descriptive statistics as predictor variables is remarkable. These statistics capture key aspects of the data distribution, leading to improved classification accuracy. By adopting this approach, the proposed methodology introduces two discriminant variables based on measuring average closing price returns and the slope of these returns at inflection points.

Okicic et al. [72] present an approach to stock selection and analysis based on LDA. This method seeks to facilitate the configuration and diversification of investment portfolios. The results emphasize the importance of identifying variables that can effectively predict the relevant categories for selecting suitable stocks. For their part, Zopounidis et al. [73] propose an approach using LDA to select financial ratios and compare its discriminant ability with that of other approaches. Studies show that using critical and relevant information improves classification accuracy by capturing distinctive characteristics of the dependent variable categories. The proposed methodology capitalizes on these findings by using discriminative measures to differentiate the categories of the dependent variable.

On the other hand, combining multiple techniques leads to substantial improvements in classification performance. Hwang et al. [74] develop a linear discriminant function to classify stocks into superior and inferior according to their financial performance. They confirm that the combination of data envelopment analysis (DEA) and discriminant analysis (LDA) significantly improves classification accuracy, reaching 85%. This outcome indicates that employing more robust approaches can benefit from incorporating pre-modeling procedures and complementary techniques.

Kwag et al. [34] introduced an approach that uses a logistic regression model to predict stock price direction. The model's prediction parameters are determined using LDA. Furthermore, their study reveals that financial ratios have a greater impact than macroeconomic indicators in explaining and predicting stock price variations.

Ogut et al. [75] propose a model for detecting price manipulation in the Turkish stock market. The study compares several machine learning algorithms, such as discriminant analysis, logistic regression, artificial neural network, and support vector machine. The results show that discriminant analysis helps detect price manipulation in the stock market to prevent fraudulent practices. These studies demonstrate the versatility and effectiveness of discriminant analysis in classification when combined with complementary approaches. The results improve with data preparation, feature extraction, or feature selection. The proposed methodology considers other complementary approaches to improve classification accuracy, such as underlying structure analysis.

E. STOCK MARKET INDEXES

Discriminant analysis has been used in notable studies to predict the direction of stock market indexes.

Leung et al. [39] evaluate the effectiveness of several classification methods (LDA, logit, probit, and probabilistic neural network) in predicting the direction of stock indexes. The study highlights the performance of the discriminant function, which achieved the highest accuracy of 68% when applied to the NIKKEI 225 index. These results indicate that using methods that capture and exploit discriminant relationships and patterns in the data can increase prediction accuracy.

Huang et al. [76] propose a combined approach that integrates multiple classification methods to improve prediction accuracy. Specifically, the model combines support vector machine (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and Elman backpropagation neural networks. By leveraging this combined approach, they achieved improved accuracy in predicting the movement of the NIKKEI 225 index, with a hit rate of 75%. These results suggest that combining different approaches capitalizes on their strengths and compensates for their individual limitations, resulting in more robust and accurate prediction models.

Azis et al. [61] presents a remarkable study summarizing the machine learning approaches used in the investigation of financial phenomena. This article details the most commonly used machine learning techniques in financial asset price prediction. It is worth highlighting that harnessing discriminant information leads to a significant improvement in prediction accuracy.

This brief literature review provides a clear and detailed rationale for why the proposed methodology can achieve

more compelling results in predicting the evolution of the euro-dollar exchange rate. We also identify the issues and challenges these approaches face that remain to be solved.

The literature review supports the selection of discriminant analysis and the criteria defined in the proposed methodology. It offers insights into the strengths and limitations of the cited approaches and identifies best practices that have demonstrated high predictive ability in their respective fields.

We develop the proposed methodology with a comprehensive approach that addresses common challenges in current predictive approaches. We also carefully consider relevant factors such as market inflection points and the introduction of new variables that influence the prediction of financial asset prices. In addition, we address the need for interpretability, stability, generalization, scalability, accuracy, complexity, and transparency of the proposed model.

An exceptional aspect of this methodology is its emphasis on preserving the interpretability of the input variables. This feature allows investors to gain a deeper understanding of how investment decisions are made and instills confidence in the model's outcomes. As a result, the main contribution of this methodology is to offer a transparent and informed understanding of the price action of the traded exchange rate. By leveraging such information, this approach generates more compelling and valuable predictions, thereby empowering investors to make well-informed investment decisions.

III. MATERIALS AND METHODS

A. DATABASE

This research work uses the collection of intraday winning trades generated by an algorithmic trading model. These data are obtained from a trend-following trading strategy based on a momentum trading approach (Momentum Trading Strategy). The model captures the beginning of the direction of price movement and identifies it as an **ET** entry momentum. These reference points are useful for identifying the variables that best predict the onset of the direction of the future movement of the euro-dollar exchange rate. Thus, trades made on **ET** entry points are made in favor of the trend and produce positive profit margins.

Table 1, shows the performance of the trading strategy. It should be noted that positions with negative financial performance are excluded from the data preparation and analysis as they negatively impact the results of the study. The opening and closing of each contract is done on CP closing prices on a 15-minute time frame. This time frame is selected based on the trading strategy's financial performance and risk exposure. Each contract is executed at the **ET** instant at which the trend change occurs. Commission costs of $45/1000000 USD trade are charged after each transaction. The commission charged for holding a position open overnight is applied according to the type of position, Swap long $= -4.96$ and Swap short $= -0.96$. Slippage is not deducted from trades because delays in execution negatively affect the definition of **ET** study times and the definition of microtrends.

Consequently, the trades have a winning probability of 72% and generate net profits of USD 13162.85, during the sample period, from January 2006 to December 2020. Thus, the winning trades represent 75.34% of the 20894 trades made and take an average of 7 hours. Seventy-five percent of these trades achieved gross returns in the first two-thirds of the holding period. The Win/Loss Ratio with a value of 2.37 shows that the number of winning positions is 2 times higher than the number of losing positions. Profit Factor of 1.68 shows that $ 1.68 is earned for every dollar lost.

Statistically, the test of spurts and the Z-score of the trading system determines the existence of dependence between trades' results. Thus, a Z-score of -47.68 confirms a positive dependence between trades within a 99.7% confidence limit. This means, a smaller number of spurts than the normal probability function would imply, so that winning trades generate more winners and losing trades more losers. It is certainly possible to exploit the dependence relationships that exist between positions and improve the performance of the system, but the use of capital management measures and confirmation signals is beyond the scope of this paper.

1) MARKET DATA

The study uses two data sets with market information on the euro-dollar exchange rate. The quotes for the euro-dollar pair are obtained from the *Alpari International* trading platform. These quotes come from 15-minute and 1-week time frames. The first analysis sample spans 15 years, running from January 1, 2006 to December 31, 2020. The second sample is 18 years, from January 4, 2004 to June 27, 2021. The first data set is compiled into a $\mathbf{Q}\,\langle 375000 \times 7\rangle$ matrix consisting of 375000 observations and 7 metrics. The second set is compiled into a matrix $\mathbf{W}\,\langle 913 \times 7\rangle$ consisting of 913 observations and 7 metrics. Each observation, in both data sets, consists of Record Number (id), Date and Time, Opening Price (op_i), High Price (hp_i), Low Price (lp_i), Closing Price (cp_i), and Volume (Vo_i).

Figure 2 shows the price history (a) and log returns (b) of the euro-dollar exchange rate. The price and return series correspond to 15-minute and 1-week time frames. These values come from the data matrices **Q** and **W**. Figure 2a shows that the exchange rate, while exhibiting a range movement, tends to maintain a bear market regime during the analysis period. The trend effect of the historical price series is suppressed by the logarithmic returns. The series of returns are obtained from the following expression $\vartheta_i = ln\,(cp_i/cp_{i-1})$, where ϑ_i is the inter-price return calculated for each time frame and cp_i is the closing price associated to the time instant i.

The mean $\bar{\vartheta}$ of these returns is higher in the 15-minute returns than in the 1-week returns ($\bar{\vartheta}_{15m} = 8.42 \times 10^{-08}$, $\bar{\vartheta}_{1w} = -6.56 \times 10^{-05}$). Analysis of Figure 2b reveals that volatility, in both time frames, is highest in late 2008 and early 2009. Subsequently, volatility in the 15-minute time frame becomes strong in early 2014 and late 2015, while in the weekly time frame it becomes strong in mid-2010 and early 2020. The skewness coefficients (γ) of these

TABLE 1. Trading strategy performance report.

Specifications					
Symbol:	EUR/USD	Base Currency:	USD	Bars:	370903
Initial Deposit ($):	1000	Leverage:	1:100	Contract Size:	100000
Ending Balance ($):	14162.85	Traded Volume in Lots:	0.010	Pip Value ($):	0.10
Closed Trade P/L ($):	15800.68	Commissions ($):	-2370.53	Swaps ($):	-267.30
Minimum Margin Level (%):	6570.47	Risk Free Rate (%):	0.00	Market Risk (%):	0.063
Total Net Profit ($):	13162.85	Gross Profit ($):	32496.94	Gross Loss ($):	-19334.09
Profit Factor:	1.68	Expected Payoff:	0.63	Probability of Success:	0.72
Recovery Factor:	38.39	Sharpe Ratio:	0.13	Probability of Failure:	0.28
AHPR:	1.0003	GHPR:	1.0003	Risk per Trade (%):	0.14
Total Trades:	20894	Short Trades Won:	10502	Long Trades Won:	10392
Null Trades:	88	Winning Trades:	15026	Losing Trades:	5780
ROI:	13.16	Largest profit trade ($):	74.36	Largest loss trade ($):	-39.79
Analysis period (Years):	15.21	Average profit trade ($):	2.21	Average loss trade ($):	-3.12
Profit Trades (%):	70.32	Max consecutive wins:	33	Max consecutive losses:	21
Loss Trades (%):	29.68	Max consecutive profit ($):	278.76	Max consecutive loss ($):	-342.01
(Win / Loss) Ratio:	2.37:1	Average consecutive wins:	5	Average consecutive losses:	2
Mathematical Expectation:	$ 0.76	(Profit / Loss) Ratio:	0.71:1	Trading Advantage:	0.41
Standard Deviation Profits:	$ 4.10	LR Correlation:	0.99	The Runs Test Z-Score:	-47.68
Coefficient of Variation:	5.43	LR Standard Error:	397.52	Z Score Confidence Limit:	0.997
Balance Drawdown Absolute ($):	34.06	Balance Drawdown Maximal ($):	342.84	Balance Drawdown Relative (%):	2.42

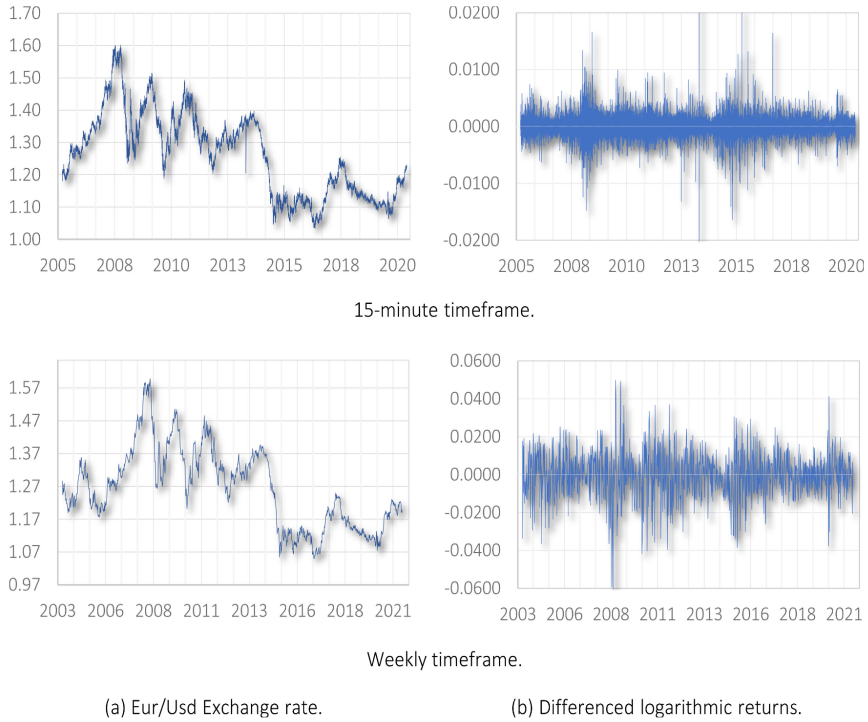


FIGURE 2. Historical closing prices (a) and logarithmic returns (b) of the euro-dollar exchange rate. Data is measured on a 15-minute time frame sampled over a 15-year time horizon, ranging from January 1, 2006, to December 31, 2020. Weekly data is sampled over an 18-year period, from January 4, 2004, through June 27, 2021.

returns are negative. The negative skewness is strongest in the 15-minute ($\gamma_{15m} = -1.02$, $\gamma_{1w} = -0.30$). The kurtosis coefficient (κ) is larger in the 15-minute returns distribution ($\kappa_{15m} = 7570.83$, $\kappa_{1w} = 1.60$). Thus, the distributions of log returns, of the euro-dollar exchange rate for both time frames, are asymmetric and leptokurtic. Furthermore, this suggests

preparing the data to make the use of multivariate forecasting techniques feasible.

2) INTRADAY WINNING TRADE DATA

Intraday winning trades are obtained from a trend-following trading model. This dataset $\mathbf{A}\,\langle 14770 \times 18\rangle$ constitutes the starting unit for the detection of micro-trends in favor of the long-term trend movements to be forecasted. This collection consists of 14770 contracts and 18 metrics that measure the market behavior before and after the opening of each contract at the **ET** time. The units of analysis consist of 7380 long trades and 7390 intraday short trades. The analysis period is 15 years and runs from January 2006 to December 2020.

- *Indexes and variables.*

Trades are measured by 18 indicators. The first fifteen indices, referenced in the table 2, measure market performance 18 periods prior to the **ET** moment (on a 15-minute time frame). These indicators are calculated with a number of periods (18) defined by the best performance achieved with the ranking model. These indicators are grouped into three areas of analysis, price, volume and divergence between price and volume dimension variables. In addition, these indicators are used as candidate variables in the feature selection process. The best indices are used to predict the onset of the directional movement of the euro-dollar exchange rate.

The following three metrics (GPL_$, RP% M[1], and Sp.XP-EP) measure the performance obtained by the trade after the end of the contract's holding time. These indicators are used in the detection of underlying structures and the identification of the determinants of the differences between trend movements. The indicators (M.Sp.P.ET, LTT and, M.Sp) help to integrate the micro-trends of intraday winning trades (**A**) into a larger time frame (**W**). Consequently, the instances that are in favor of long-term trend movements are extracted and form the study population **O** (See Section III-B1).

Finally, the categorical variable "Trend" labels at time **ET** the direction of the exchange rate trend. The class labels used are "Upward Trend Movement" and "Downward Trend Movement". Sideways price movements are not considered in the present work.

- *Descriptive statistics of the indexes.*

Table 3 summarizes the descriptive statistics of the multivariate data matrix **A**. The mean of the data at the **ET** moments (where the change in direction of the preceding trend takes place) is higher for the variables Sp.Vo.18P, GPL_$, Dv.CP.RSI.18P, and s.Vr.RSI.18P. In addition, the first two indices (Sp.Vo.18P, GPL_$) report a volatile effect, while the last two have one of the lowest standard deviations. The mean and median of the data are different. The skewness and kurtosis statistics are far from zero. The indicators s.Vr.Vo.18P, x.Vr.Vo.18P, and Sp.Vr.Vo.18P provide evidence of disproportionate values in the tails of the distributions. Since none of these measures are close to the mean, this is strong evidence that the data set, for some indicators, exhibits asymmetric and leptokurtic behavior.

B. PROPOSED METHODOLOGY

This approach has practical applications in real-world scenarios, particularly in intraday trading of financial instruments like foreign exchange rates. Making informed decisions under uncertainty is crucial in this context. Our methodology allows analysts to leverage market inflection points and use data preparation, feature selection, and structure detection to develop effective classification models. These models facilitate the interpretation of price behavior and offer accurate short-term trend predictions for the traded asset. Incorporating these predictions into intraday trading frameworks improves the effectiveness of value judgments in investment decision-making and minimizes risk exposure. As the core of our methodology is the configuration and validation of the model so that it becomes suitable for linear classification, we name it Linear Classifier Configurator - LCC in short.

The proposed methodology allows predicting, in the short term, the beginning of the direction of the future movement of the euro-dollar exchange rate. The methodological design is based on five stages and four collections of data $\Xi\,\langle i \times j\rangle$ of size i rows by j attributes. Where i is the number of study moments **ET**, at which the change in direction of the preceding trend occurs and j are the numerical observable variables that measure the behavior of the closing prices before and after the moment **ET**. Figure 3 summarizes the procedure followed at each stage. First, the data collections are prepared, in order to make them analyzable and relevant, in fulfillment of the purposes established in each of the proposed phases. Secondly, from the technical analysis point of view, the predictors with the highest discriminatory power are selected on the basis of their statistical merit. Thirdly, the underlying structures are detected and analyzed to identify the determinants of the differences between trend movements. The usefulness of the selected predictors and the adequacy of the discriminant analysis is validated with the results of this stage. Fourth, the classification of the direction of the exchange rate movement is performed using discriminant analysis. Finally, the predictive power of the classification model is evaluated out-of-sample. The effectiveness of the proposed methodology is confirmed by the results of this stage. It is important to consider a suggestion when setting up new data sets before applying the proposed methodology (Figure 3). The benchmarks used for prediction should be measured using the variables specified in Table 2. This recommendation ensures that the features are defined based on measurements that effectively distinguish one observation from another. These benchmarks can be represented by turning points or the opening moments of successful buy or sell trades. These market entry signals must have been previously validated technically and financially through a trading strategy.

1) DATA PREPARATION

To maximize the usefulness of the data, make them analyzable and increase their potential in the processes of feature

TABLE 2. Indexes and variables.

Dimension	Name in short	Features		Formulas
Price action	s.V.PT	Dispersion of returns from closing prices.	$\sqrt{\frac{\sum(\ddot{x}_i-\bar{\ddot{x}})^2}{(n-1)}}$,	$\ddot{x}_i = \frac{cp_i - cp_{i-1}}{cp_{i-1}}$
	x.V*P'T	Average Return.	$\frac{\sum_{i=1}^{n}\check{x}_i}{n}$,	$\check{x}_i = ln\frac{cp_i}{cp_{i-1}}$
	Sp.CP.PT	Close-Price Slope.	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = cp_i$
	Sp.Vr*CP.PT	Slope of Close-Price Variation.	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = ln\frac{cp_i}{cp_{i-1}}$
	Sp.Vr.HL.PT	Slope of High-Low Price variation.	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{(hp_i - lp_i)}{(hp_{i-1} - lp_{i-1})} - 1$
	Sp.Vr.CO.PT	Slope of Close-Open Price variation.	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{(cp_i - op_i)}{(cp_{i-1} - op_{i-1})} - 1$
	s.Vr.RSI.PT	RSI variation.	$\sqrt{\frac{\sum(\dot{x}_i-\bar{\dot{x}})^2}{(n-1)}}$,	$\dot{x}_i = \frac{RSI(9p)_i}{RSI(9p)_{i-1}} - 1$
	Sp.Vr.RSI.PT	RSI variation Slope.	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{RSI(9p)_i}{RSI(9p)_{i-1}} - 1$
Volume	s.Vr.Vo.PT	Variation in Volume.	$\sqrt{\frac{\sum(x_i^{\circ}-\bar{x^{\circ}})^2}{(n-1)}}$,	$x_i^{\circ} = \frac{Vo_i - Vo_{i-1}}{Vo_{i-1}}$
	x.Vr.Vo.PT	Average Variation in Volume.	$\frac{\sum_{i=1}^{n}\hat{x}_i}{n}$,	$\hat{x}_i = \frac{Vo_i}{Vo_{i-1}} - 1$
	Sp.Vr.Vo.PT	Variation slope in Volume.	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \frac{Vo_i - Vo_{i-1}}{Vo_{i-1}}$
	Sp.Vo.PT	Market Volume Slope.	$\frac{\sum(x_i-\bar{x})(Vo_i-\bar{Vo})}{\sum(x_i-\bar{x})^2}$	
Divergences	Dv.CP-RSI.PT	Close-Price and RSI divergence.	$\frac{\sum(cp_i-\bar{cp})(y_i-\bar{y})}{\sqrt{\sum(cp_i-\bar{cp})^2\sum(y_i-\bar{y})^2}}$,	$y_i = RSI(9p)_i$
	Dv.Vo-CP.PT	Close-Price and Volume divergence.	$\frac{\sum(Vo_i-\bar{Vo})(cp_i-\bar{cp})}{\sqrt{\sum(Vo_i-\bar{Vo})^2\sum(cp_i-\bar{cp})^2}}$	
	Dv.Vo-RSI.PT	RSI and Volume divergence.	$\frac{\sum(Vo_i-\bar{Vo})(y_i-\bar{y})}{\sqrt{\sum(Vo_i-\bar{Vo})^2\sum(y_i-\bar{y})^2}}$,	$y_i = RSI(9p)_i$
After holding time	* GPL_$	Gross profit or loss for Long and Short trade.	$V_\varphi \cdot cs\,(xp - ep)$,	$V_\varphi \cdot cs\,(ep - xp)$
	* RP% M[1]	Money Risk by long or short trade.	$\frac{V_\varphi \cdot cs(lp-ep)}{sb}$,	$\frac{V_\varphi \cdot cs(ep-hp)}{sb}$
	* Sp.XP-EP	Open and close price trend line slope.	Slope $(\Delta p/\Delta t)$,	Long = $\frac{xp-sp}{xt-st}$, Short = $\frac{sp-xp}{xt-st}$
Long-term trend	M.Sp.P.ET	Market slope prior to ET moment (W1).	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \{cp_{i-96}, ..., cp_{i=\mathrm{ET}}\}$
	* LTT	Long-term trendline (W1).	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \{cp_{i=\mathrm{ET}}, ..., cp_{i=96}\}$
	* M.Sp	Market slope (W1).	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$,	$y_i = \{cp_{i-96}, \ldots, cp_{i=\mathrm{ET}}, \ldots, cp_{i=96}\}$
N/A	$\mathrm{RSI}(9p)_{Smooth}$	Relative Strength Index.	$100 - \left(100/\left(1 + \bar{U}/\bar{D}\right)\right)$	
∇Trend	Trend	0 = Uptrend: $(GPL\_\$ > 0,$ $Slope\,(\Delta p/\Delta t) > 0,$ $Long\ \ Trade)$ 1 = Downtrend: $(GPL\_\$ > 0,$ $Slope\,(\Delta p/\Delta t) < 0,$ $Short\ \ Trade)$		

*Notes: All variables are measured in 15-minute time frames, except (W1), which is measured in a 1-week time frame. (W1) = Formulas used for the definition of long-term trend movements. * = Measurements after the **ET** turning point (not involved in the feature selection process). Vo_i = Market volume in the period i or Number of times the price changes in each period i, in a 15-minute time frame. i = Analysis period, where $i \in \{1, \ldots, n\}$. cp_i = Closing price for the period i. hp_i = Maximum price of the period i. lp_i = Minimum price of the period i. op_i = Opening price in the period i. $RSI(9p)_i$ = Relative strength index for the period i, measured from the last 9 historical closing prices. $x_i = \{x_1, \ldots, x_n\}$ Time series, $x_i, i \in \{1, \ldots, n\}$. **ET** = Turning point in study time et. V_φ = Volume traded in lots. cs = Contract size 100000 units of base currency (Euro). sb = Initial account balance in US dollars. xp = Exit Price. sp = Entry Price. xt = Exit time. st = Start time. $\bar{U}$ = Average of upward price differences. $\bar{D}$ = Average of falling price differences. ∇Trend = Categorical variable.

selection, structure detection and model formulation, they are prepared and arranged in four data collections $\Xi\,\langle i \times j\rangle$. Where i is the number of **ET** times at which the change in direction of the preceding trend occurs, in 15-minute trading sessions, and j are the variables that measure their behavior before and after the trend shift. According to Figure 4, data preparation is summarized in four phases. First, the bullish and bearish micro-trends are categorized and arranged for the feature selection process. Second, the study population is extracted and prepared based on the micro-trends that are in favor of the long-term trend movements. The instances that meet this requirement are used to predict, in the short run, the

TABLE 3. Descriptive statistics of indices.

Measure	Mean	Std.Dev	Median	Min	Max	Skew	Kurtosis
Price action:							
s.V.18P	0.00051011	0.00034042	0.00042776	0.000044579	0.00439635	2.3730	11.7879
x.V*18P	-0.0000002	0.00019849	0.00000143	-0.00209697	0.00164601	0.0050	5.15348
Sp.CP.18P	-0.0000008	0.00025970	0.00000289	-0.00233251	0.00192982	-0.028	6.19084
Sp.Vr*CP.18P	-0.0000004	0.00003325	-0.0000002	-0.00040831	0.00033698	-0.055	6.48955
Sp.Vr.HL.18P	0.00630912	0.03272341	0.00538741	-0.58322522	0.72498519	0.6652	38.0406
Sp.Vr.CO.18P	0.00800454	0.46597394	0.01403601	-6.46007878	8.77988644	0.0522	31.9152
s.Vr.RSI.18P	**0.16739490**	0.07751317	0.15280451	0.031895221	1.33360205	**2.6943**	18.7865
Sp.Vr.RSI.18P	-0.0005816	0.00824177	-0.0004332	-0.04381362	0.08385499	0.2604	3.14648
Volume:							
s.Vr.Vo.18P	0.47597319	1.24018626	0.35617189	0.047287739	69.8530494	**40.416**	**1937.19**
x.Vr.Vo.18P	0.10736958	0.31392351	0.07387271	-0.12090153	16.9855561	**35.890**	**1639.09**
Sp.Vr.Vo.18P	0.00231944	0.05945863	0.00230310	-3.32210627	2.84634278	-1.024	**1542.24**
Sp.Vo.18P	**10.7184210**	**41.9964534**	5.06553148	-295.573787	357.643962	0.5038	6.78630
Divergences:							
Dv.CP-RSI.18P	**0.85151545**	**0.20628068**	0.93089873	-0.65551962	0.99928244	-2.903	9.93598
Dv.Vo-CP.18P	-0.01375776	0.51348876	-0.0276017	-0.97874454	0.9763092	0.0360	-1.1499
Dv.Vo-RSI.18P	-0.01580351	0.49230421	-0.0234383	-0.97742819	0.96606119	0.0382	-1.1572
After holding time:							
GPL_$	**2.174395396**	**2.51882690**	1.47000000	0.010000000	44.3000000	4.4641	34.7811
RP% M[1]	-0.00079658	0.01564524	-0.0017000	-0.22980000	0.17700000	0.8083	14.9109
Sp.XP-EP	0.000009683	0.03751680	-0.0000302	-0.54240000	1.03680000	1.1255	57.2277

direction of the future movement of the euro-dollar exchange rate. Third, the best set of observable variables is prepared and arranged for the detection of underlying structures. The relevance of the selected discriminant analysis and predictors is determined by the adequacy of the data. Fourth, a multivariate normally distributed data sample (*MND*) is drawn from the study population to overcome the asymmetric and leptokurtic nature of the observations and meet the requirements of the discriminant analysis.

- *Short-term trend movements.*

In this phase, the units of analysis are prepared for discretization into bullish and bearish micro-trends. All prepared data matrices employ these class labels as a factor or grouping variable. In addition, measurements performed on these instances are used for feature selection. Thus, the multivariate data collection $\mathbf{A}\,\langle 14770 \times 18\rangle$, consisting of intraday winning trades, from a trend-following trading strategy, are categorized into two class groups: micro uptrend and micro downtrend. The information used to perform this procedure is reported in the table 4. The discretization process assigns a class label based on two quantitative variables (Gross profit or loss and Trade slope) and confirms this assignment based on a true categorical variable (Trade type). Accordingly, an upward micro-movement occurs if a long trade generates a positive gross profit on a positive slope. A bearish micro-trend is defined by a short trade with positive gross profit on a negative market slope.

- *Long-term trend movements.*

At this stage the study population is identified and compiled into the $\mathbf{O}$ data matrix. These data are used to predict, in the short term, the onset of the future direction of the euro-dollar exchange rate. The matrix $\mathbf{O}\,\langle 1873 \times 2\rangle$ is formed by $\mathbf{i}$ moments of analysis $\mathtt{ET}$, at which the change of the preceding trend occurs, on which $\mathbf{p}$ chosen predictor variables $\hat{\mathbf{x}}_j$ are measured. This collection is the result of confirming each micro-trend, from matrix $\mathbf{A}$, with a long-term trend movement, according to market information from data matrix $\mathbf{W}\,\langle 913 \times 7\rangle$. Consequently each change of direction of the preceding trend is confirmed 96 weekly periods before and after the time $\mathtt{ET}$. The results obtained during data preparation show that this number of periods is sufficient to confirm the dominance of the newly formed trend on a weekly time frame. Thus, the study time $\mathtt{ET}$, is selected as a valid instance, when the change of direction of the preceding trend, in 15-minute trading sessions, is confirmed with a continuation or a change of trend in favor of the dominant trend, in weekly trading sessions, otherwise the observation is discarded. It should be noted that short-term trend changes are part of long-term trend movements, but are not determinative of their formation. Note that the purpose of this study is limited to predicting the beginning of the trend direction in 15-minute trading sessions and, therefore, does not necessarily mean that such prediction is determinant of the beginning of the long-term trend movement.

Figure 5 shows a graphical explanation of the analysis unit selection process described in the table 5. The various long-term bullish and bearish market regimes, according to the study moments $\mathtt{ET}$, are formed by movements with continuity in the trend and movement with change in the direction of the preceding trend. It should be noted that the study

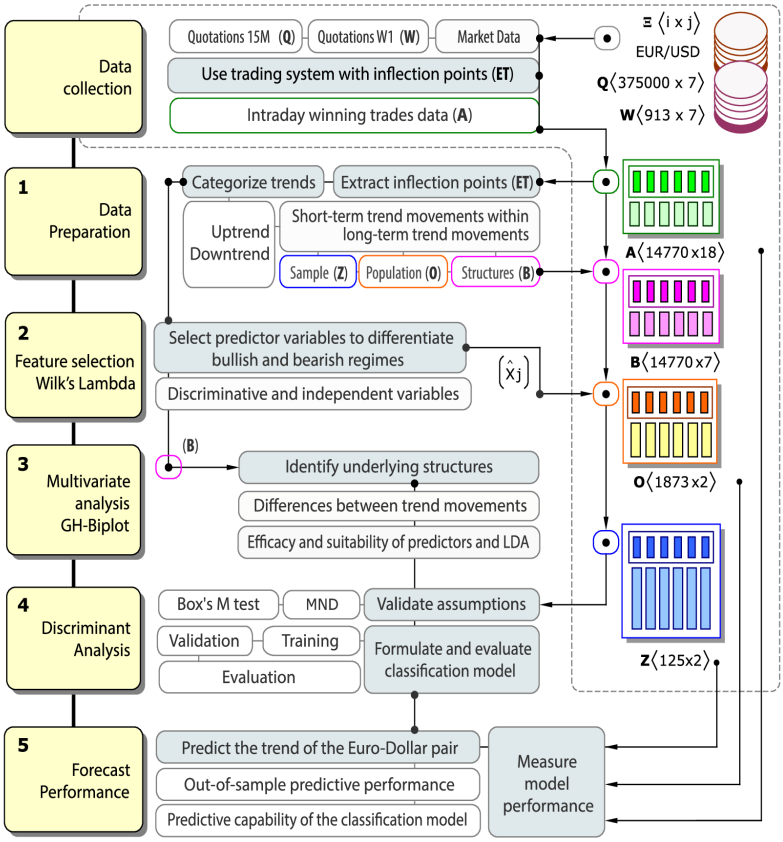


FIGURE 3. Proposed methodology to predict, in the short term, the onset of the future movement of the euro-dollar exchange rate. Each phase (in yellow) is divided into actions (in blue) and their corresponding results (in white). This methodology is summarized in data preparation procedures, selection of features with discriminant power, multidimensional analysis of differences between trend movements, discriminant analysis, and evaluation of the predictive power of the classification model.

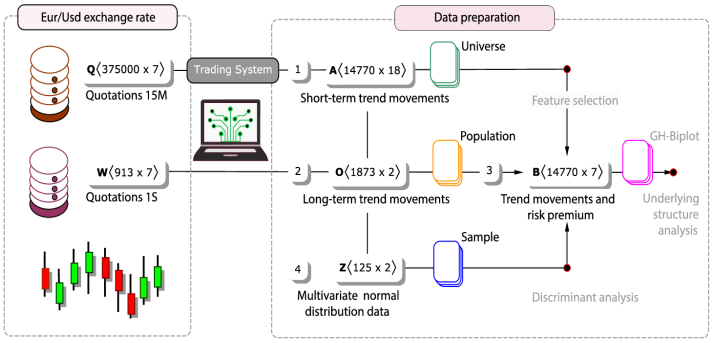


FIGURE 4. Data preparation for short-term prediction of the beginning of the direction of the future movement of the euro-dollar exchange rate. (1) Definition of micro-trends. (2) Definition of micro-trends as a function of long-term trend movements. (3) Preparation of data for underlying structure analysis. (4) Extraction of micro-trends with multivariate normal distribution.

moments **ET** in which the change of direction of the short-term trend occurs, according to the 15-minute trading session, are a function of the dominant market trend, according to the weekly trading session.

TABLE 4. Categorization of micro movements.

Gross profit or loss	* Trade slope	Trade type	Micro movement label
Winning trade	Slope $(\Delta p/\Delta t) > 0$	Long trade (Buy)	Upward Micro movement
Winning trade	Slope $(\Delta p/\Delta t) < 0$	Short trade (Sell)	Downward Micro movement

Winning trade = GPL_$ > 0. *Sp.XP-EP = Slope $(\Delta p/\Delta t)$. The formulas for the calculation of slopes, for long and short positions, are referenced in Table 2.

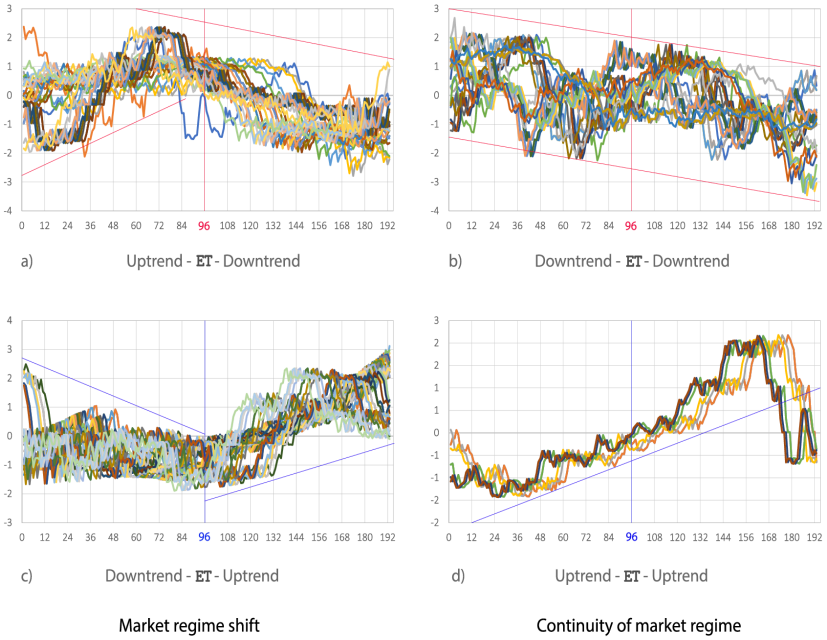


FIGURE 5. Market movements of the euro-dollar exchange rate based on weekly closing prices with the integration of intraday data at time ET. Row-normalized data. (a) Regime shift, upward to downward directional movement. (b) Regime continuity, downward directional movement. (c) Regime shift, downward to upward directional movement. (d) Regime continuity, upward directional movement.

For purposes of selection of the **ET** study units, Table 5 details in the first column the process of formation of the long-term trend movements. The study moments **ET** of micro-trends and long-term trend movements overlap. Thus, the moment of study **ET** is the reference point used to confirm, in weekly trading sessions, the continuity of the preceding trend (FT - Following the trend) or to identify the change in the direction of the preceding trend (CT - Shift trend). Thus, two trend formations are recognized, upward movements and downward movements, with two forms of origin, change of trend or continuity in the trend. The second and third columns define the assumptions used to identify the type of regime before and after the **ET** moment. The fourth column confirms the dominant slope of the market. Finally, the last column details the number of observations obtained for the study.

In summary, the study population **O** consists of 1873 micro-trends that are in favor of the direction of the dominant trend, but are not determinants of the initial long-term trend formation. Consequently, the study population consists of 414 and 228 bullish micro-trends, the former with trend reversal in favor of the dominant uptrend and the latter supporting the long-term uptrend. 556 and 675 bearish micro-trends, the former with change of direction of the preceding trend in favor of the dominant downtrend and the latter supporting the continuation of the long-term downward movement.

- *Trend movements and Risk-award.*

At this stage, the data collection **B** is prepared. According to Figure 4, the best observable variables are evaluated and selected for structure detection. The data are prepared in such a way that the analysis of the underlying structures reveals the multidimensional nature of the differences between trend movements. Consequently, the results of this study help to determine the relevance of the discriminant analysis and the appropriateness of the selected predictors. The data collection $\mathbf{B}\langle 14770 \times 7\rangle$ is made up of i analysis moments **ET** and

TABLE 5. Micro movements into long-term trend movements.

Regime Market Before - ET - After	Market slope prior to ET SRL $[cp_{i-96}, ..., cp_{i=\mathtt{ET}}]$	Long-term trendline SRL $[cp_{i=\mathtt{ET}}, ..., cp_{i=96}]$	Market Slope SRL $[cp_{i-96}, ..., cp_{i=\mathtt{ET}}, ..., cp_{i=96}]$	O 1873
Downtrend - CT - Uptrend	Slope RL $(y_i/x_i) < 0$	Slope RL $(y_i/x_i) > 0$	Slope RL $(y_i/x_i) > 0$	414
Uptrend - FT - Uptrend	Slope RL $(y_i/x_i) > 0$	Slope RL $(y_i/x_i) > 0$	Slope RL $(y_i/x_i) > 0$	228
Uptrend - CT - Downtrend	Slope RL $(y_i/x_i) > 0$	Slope RL $(y_i/x_i) < 0$	Slope RL $(y_i/x_i) < 0$	556
Downtrend - FT - Downtrend	Slope RL $(y_i/x_i) < 0$	Slope RL $(y_i/x_i) < 0$	Slope RL $(y_i/x_i) < 0$	675

ET = Turning point and Entry Time. CT = Trend shift, FT = Following trend. *M.Sp.P.ET = Market slope prior to **ET**. SRL = Slope RL = Slope of the time-series regression line. cp_i = Closing price of period i. *LTT = Long-term trendline. *M.Sp = Market Slope. Market Slope is the slope of the regression line of 96 closing prices before and after the **ET** turning point. The trend formed after the **ET** turning point, which is detected on a 15-minute time frame, is validated on a 1-week time frame with Market Slope. The equations for their computation are referenced in Table 2.

j variables measuring price action before and after the **ET** moment. There are two groups of noteworthy variables, the first group corresponding to the selected predictor variables and the second group consisting of the risk premium analysis (Sp.XP-EP, RP%M[1], GPL$). The latter group validates the direction of microtrends after earning positive gross profits following trading at **ET** inflection points. Variables that do not explain the differences between the onset of an upward and downward movement are excluded from the data structure.

- *Normally distributed multivariate data.*

Finally, in this phase, the study sample $\mathbf{Z}$ is prepared. These data are used to formulate, validate and evaluate the performance of the classification model. To overcome the asymmetric and leptokurtic effect of the study population and to ensure that the assumptions validating the application of discriminant analysis are met, a small sample size with multivariate normal distribution is used. To solve this problem, we draw from the study population $\mathbf{O}\langle 1873 \times 2\rangle$ a sample of observations that follows a multivariate normal distribution. Outliers are excluded from the study sample. This procedure ensures that the assumptions of multivariate normality and equality of variance and covariance matrices within groups are met for predictors and groups. Consequently, the data collection $\mathbf{Z}\langle 125 \times 2\rangle$ consists of 125 **ET** observations on which 2 chosen predictor variables $\hat{\mathbf{x}}_j$ are measured.

2) FEATURE SELECTION

The predictive power of the classification model is determined by the quality of the selected predictor variables and the adequacy of the selection process. Furthermore, the use of irrelevant variables, which do not contribute to the differentiation between classes, negatively affects the performance of the classification model. To address this problem, based on the multidimensional nature of the differences between bullish and bearish movements, the feature selection process is conceived under an exploratory-confirmatory statistical approach. This procedure helps to identify and confirm the effectiveness of the variables responsible for the separation of these clustering structures.

In this way, the proposed methodology helps to identify and select, from a multivariate data matrix $\mathbf{A}\langle i \times j\rangle$ formed by $\mathbf{i}$ trend movements on which p observable numerical variables $\mathbf{x}_j = (\mathrm{x}_1, \ldots, \mathrm{x}_p)$, the determinants $\hat{\mathbf{x}}_j$ of the movement of the euro-dollar exchange rate. The distinction between groups is made with a single categorical variable defined by the data vector $\dot{\mathbf{y}} = \{0, 1\}$, where $0 =$ "Upward Trend Movement" and $1 =$ "Downward Trend Movement". Thus, the vector $\dot{\mathbf{y}}$ is divided into two groups $\check{y}_g = \{\check{y}_1, \check{y}_2\}$. Consequently, the selection process yields the subset of eligible predictors $\hat{\mathbf{x}}_j$ which is formed by $\mathbf{p}$ independent variables $\hat{\mathbf{x}}_j = (\hat{\mathrm{x}}_1, \ldots, \hat{\mathrm{x}}_p)$, and is better than $\mathbf{x}_j$, if $(\hat{\mathbf{x}}_j \subset \mathbf{x_j})$ and $(\mathbf{x}_j > \hat{\mathbf{x}}_j)$, so that, $\mathbf{A}' \in \mathbb{R}^{i \times \hat{j}}$, with $\hat{j} < j$.

The function $\max\left(\beta_j^\top L \beta_j\right) / \left(\beta_j^\top T \beta_j\right)$ evaluates the variables $\mathbf{x}_j$ using the data matrix $\mathbf{A}$. So, it maximizes the between-group variability as a function of the total variability from a subset of $\mathbf{x}_j$ variables. The total covariance $\mathbf{T} = \mathbf{K} + \mathbf{L}$, is composed of the within-group and between-group covariances. β_j is the vector of coefficients of the discriminant function that evaluates the observable variables $\mathbf{x}_j$. Thus, the subset of eligible predictors $\hat{\mathbf{x}}_j$ is the one that best contributes to the differentiation between groups, has the highest discriminant power, and offers the best predictive performance. Additionally, the selection of the most plausible features is validated by performing statistical significance tests. These predictors are used in the construction of the classification model.

- *Variable selection process.*

It is composed of three stages, as shown in Figure 6. First, the ability to differentiate between groups is the potential that an eligible candidate variable has. This criterion is detected by one-way Anova analysis. The number of candidate variables is reduced with this initial screening. Second, the predictors that best contribute to the separation between groups are identified using the Wilks' Lambda statistic. The discriminant potential of each candidate variable is measured by this criterion. Levene's test of homogeneity of variances validates the test of equality of group means and reduces the number of eligible variables detected with the Wilks' Lambda statistic. Finally, the association between eligible variables is measured with Pearson's correlation coefficient. Predictors that may cause multicollinearity problems are discarded from the model formulation. Variables that are left out of the analysis, do not improve the separation of the groups, do not contribute information to

the model and are therefore excluded from the subsequent procedures.

- *Discriminant potential of eligible variables.*

Referring to the process shown in Figure 6, the second step is based on the stepwise selection method. This helps to identify, statistically, explanatory variables that have a strong discriminative power and provide the best performance in the construction of classification models. This statistical procedure is an iterative method in which the study variables are analyzed individually until a subset of eligible variables $\hat{\mathbf{x}}_j$ is obtained. The Wilks' Lambda statistic and the value of the associated statistic, Snedecor's F, are used to determine the discriminant potential of each variable and of the subset of eligible variables $\hat{\mathbf{x}}_j$. Attributes with large Snedecor F-values contribute to the separation of group means and therefore discriminate better. In contrast, lower values do not discriminate due to attributes with closely spaced groups and widely scattered data. The procedure starts by calculating, for each analysis variable, the Snedecor's F-Values of inclusion F_ε and removal F_r given by equations (1) and (2), as follows:

$$F_\varepsilon = \frac{\left(\check{f} - \check{e}\right)(n - q - g)}{\check{e}\,(g-1)} \tag{1}$$

$$F_r = \frac{\left(\check{e} - \check{f}\right)(n - q - g + 1)}{\check{f}\,(g-1)} \tag{2}$$

where $\check{e}$ is Within-groups Sums of Squares and Cross-product Matrix, $\check{f}$ is Total Sums of Squares and Cross-product Matrix, q corresponds to the group of variables incorporated in the analysis, g is the number of groups, n is the number of observations.

Variables with inclusion levels higher than the F-value to enter ($F_\varepsilon = 3.84$) enter the subset, of eligible variables $\hat{\mathbf{x}}_j$, and variables with removal values lower than the elimination F-value ($F_r = 2.71$) exit the analysis. The process is repeated until there are no candidate variables to remove. If the tolerance of the analysis variable falls below the specified tolerance (0.001), the variable is ineligible. The tolerance ψ_i can be written as $\psi_i = \left(1 - R^2\right)$. The tolerance is understood as the percentage of unexplained variance between the analyzed variable and the included variables. The value decreases drastically if the variables are correlated, while it will produce small variations when the variables are independent.

The modeling of the discriminant power of the linear function is performed as a function of some or all of the $\mathbf{p}$ eligible variables $\hat{\mathbf{x}}_j$. The Wilks' Lambda statistic is the benchmark used to assess the discriminant power of the variables used in the model. This criterion contrasts the dispersion of the instances within groups and the dispersion of the data without distinguishing between groups. Its computation is given in the equation (3). The closer the value is to zero, the greater the discriminant power of the variables analyzed, whereas if the value is close to one, the lower the discriminant power.

$$\Lambda = \frac{|\mathbf{K}|}{|\mathbf{T}|} \tag{3}$$

In matrix notation $\mathbf{T} = \mathbf{K} + \mathbf{L}$ is the total covariance, with $\mathbf{K}$ and $\mathbf{L}$ as the within-group and between-group covariances. **Lambda** is the Wilks' Lambda statistic. Finally, the use of statistical significance tests facilitates the identification of the subset of eligible variables $\hat{\mathbf{x}}_j$, which, because of their statistical validity and from a technical analysis point of view, help to explain the emergence of a new trend as a function of the culmination process of the previous movement. In addition to this connotation, the eligible predictors $\hat{\mathbf{x}}_j$ are more useful because they provide explanatory value to the results obtained and contribute to the formulation of more accurate prediction models.

3) MULTIVARIATE ANALYSIS

GH − Biplot is a multivariate analysis method useful for representing and inspecting big data matrices [77]. This analysis technique is an alternative approach to the original method introduced by Gabriel [78] that improves the quality of representation of observations (points) and variables (vectors) in the same reduced-dimensional reference system. In this work, the **GH − Biplot** analysis is used to detect structures and to describe the underlying relationships between the selected observable variables $\hat{\mathbf{x}}_j$ and the factors that explain them. Thus, the analysis allows us to suggest, according to the structure of the data and the discriminant capacity of the selected predictor variables, the appropriateness of formulating a discriminant model.

In Figure 7a, the multivariate data matrix $\mathbf{B}\,\langle i \times j\rangle$ after singular value decomposition (*SVD*) approximates the matrix $\mathbf{B} \cong \mathbf{UDV}^T$, where, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{D} = (\lambda_1, \ldots, \lambda_J)$ contains the singular values. Consequently, $\mathbf{G}$ is the matrix $\mathbf{U}$, and $\mathbf{H}$ the matrix of the first two columns of the product $\mathbf{VD}$. Thus, this technique with the appropriate selection of markers, $\eth_i = (\eth_i, \ldots, \eth_n)$ for the units of analysis and $\mathrm{h_j} = \left(h_j, \ldots, h_n\right)$ for the variables, allows to represent simultaneously in the same space observations (points) and variables (vectors).

The graphical representation and interpretation rules of a biplot analysis are shown in Figure 7b. Although the **GH − Biplot** analysis is specifically used for the detection and study of patterns and covariation relationships between variables, this analysis uses the following interpretation rules: the length of each vector represents the variability of the measurements; the angles formed between variables determine the level of covariation or correlation; and the angles formed between variables and factors are understood as the relationship or contribution of each variable to the factor. However, although this type of analysis does not focus on the study of the relationships between observations and variables, the distance between units of analysis is interpreted as a measure of similarity and the orthogonal projection of each observation on the body of the vector determines the order of the measurement in the original matrix.

4) DISCRIMINANT ANALYSIS

Discriminant analysis is a multivariate statistical technique introduced by Fisher [79]. It allows modeling and predicting,

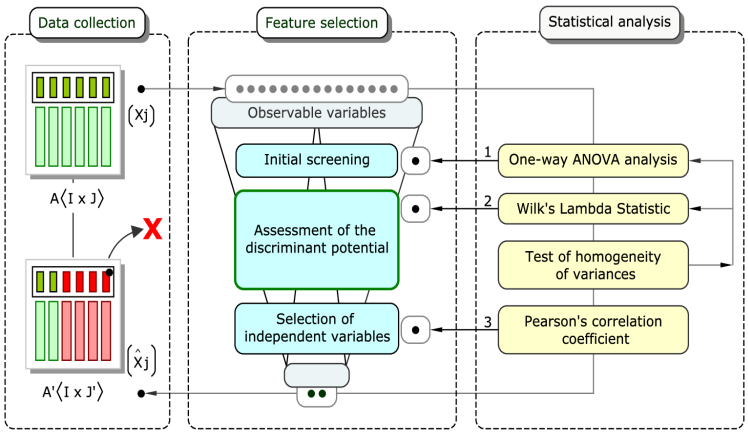
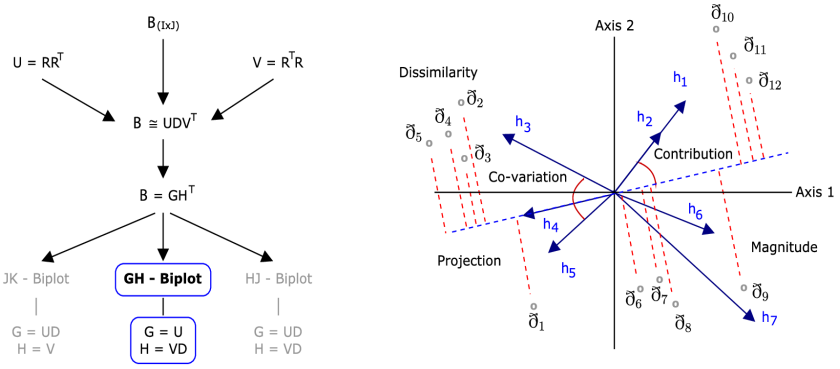


FIGURE 6. Proposed procedure to identify and select the predictor variables with the greatest discriminant power. This procedure is summarized in three stages, initial screening of observable variables, evaluation of the discriminant potential, and selection of independent predictor variables.



(a) Biplot typologies. (b) Biplot analysis.

FIGURE 7. GH – Biplot analysis: A new tool to diagnose the discriminatory power of selected predictor variables and the suitability of classification models in the underlying structures of the variables. (a) Typologies of biplot analysis. (b) Graphical representation and rules of interpretation of a biplot analysis.

on a categorical dependent variable, group membership as a function of a set of independent quantitative variables. Violation of assumptions and limitation in the fulfillment of requirements can distort results and interpretations. To overcome these difficulties, the linear discriminant function is constructed, validated and evaluated from the data matrix $\mathbf{Z}\langle i \times j\rangle$ formed by $\mathbf{i}$ study moments $\mathbf{ET}$ (at which the direction of the preceding trend changes) on which $\mathbf{p}$ chosen predictor variables $\hat{\mathbf{x}}_j = (\hat{x}_1, \ldots, \hat{x}_p)$ are measured. The membership group corresponds to the beginning of the future movement direction and is defined by the label vector $\dot{\mathbf{y}} = \{0, 1\}$, where 0 = Upward Trend Movement "C_U" and 1 = Downward Trend Movement "C_D". The data vector $\dot{\mathbf{y}}$ is divided into $\check{y}_g$ groups with $g = \{1, \ldots G\}$, where, $\check{y}_g$ denotes the g^{th} group of size n_g, such that $\check{y}_g = \{\check{y}_1, \check{y}_2\}$ is the a priori probability of membership of each case to the group $\check{y}_g$. Consequently, the linear function expressed in equation (4) is the one that best discriminates between bullish and bearish movements.

$$y(\hat{\mathbf{X}}) = \beta_j^{\top} \hat{\mathbf{X}}_j + \beta_0, \tag{4}$$

where $y(\hat{\mathbf{X}})$ is the discriminant score. β_j is the vector of discriminant weights or coefficients. $\hat{\mathbf{X}}_j$ are the selected independent variables and β_0 is a constant. The function (4) can be written as the equation (5):

$$y(\hat{\mathbf{x}}) = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2. \tag{5}$$

The classification criteria available for case assignment are multiple, however, in this work the Fisher classification functions $y(\hat{x}_g)$ derived from the main equation (5) are employed.

For each group g a function is extracted, thus $y(\hat{x}_U)$ and $y(\hat{x}_D)$ discriminate between upward "C_U" and downward "C_D" movements respectively. The coefficients of these functions are obtained assuming bivariate normality for the subset of predictors $\hat{\mathbf{x}}_j$, maximum likelihood and equal a priori probabilities for $\check{y}_g$. The vector of coefficients of each ranking function is determined by the equation (6).

$$y(\hat{x_g}) = \bar{\hat{\mathbf{x}}}_{jg}^{\top} S^{-1} x - \frac{1}{2} \bar{\hat{\mathbf{x}}}_{jg}^{\top} S^{-1} \bar{\hat{\mathbf{x}}}_{jg} + \ln\left(\pi_g\right), \tag{6}$$

where $\bar{\hat{\mathbf{x}}}_{jg}^{\top}$ is the transposed vector of the means of the predictor variables $\hat{\mathbf{x}}_j$ of the group g. S is the combined intra-group covariance matrix, x is the name of the chosen independent variables, π_g is the a priori probability of belonging to the group g. The classification of each observation i (turning point **ET**) is made according to the obtained discriminant score $y(\hat{\mathbf{x}})$. The model assigns the case to the group that achieves the highest discriminant score. The rule that assigns the cases to the membership group is summarized below:

$$\begin{gathered}\text{If } y(\hat{x}_U) \geq y(\hat{x}_D) \text{ then } y(\hat{x}_U) \text{ belongs in } C_U. \\ \text{Else if } y(\hat{x}_U) < y(\hat{x}_D) \text{ then } y(\hat{x}_D) \text{ belongs in } C_D,\end{gathered}$$

where $y(\hat{x}_U)$ is the bullish linear function, $y(\hat{x}_D)$ is the bearish linear function, C_U y C_D are the group identification labels according to bullish and bearish movements.

Goodness-of-fit measures indicate the model's ability to fit the data set. Analysis of these results can reveal discrepancies between observed and expected values of the model. The explained variance, the canonical correlation and the function test are the most commonly used goodness-of-fit measures. The total variance explained is due to the first eigenvalues that are associated with each of the extracted discriminant functions. Thus, the m discriminant functions $y_i = (y_1, \ldots, y_m)$, where $m = Min\left(g - 1, \hat{x_j}\right)$ are linearly independent and have an associated eigenvalue λ indicating the proportion of total variance explained by that function. For a single discriminant function $m = 1$, there is only one nonzero eigenvalue λ_1. The ratio of variance explained by the function y_i to the total variance explained by the extractable functions y_m is determined by the expression $\varpi(y_i)$ of the equation (7):

$$\varpi(\mathrm{y}_i) = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i}. \tag{7}$$

Note that in the denominator of the equation (7) is the sum of all the eigenvalues λ_i, corresponding to the total variance explained by all the discriminant functions y_m. In the numerator is the eigenvalue λ_i associated to the discriminant function of analysis y_i.

The canonical correlation is a measure of the effectiveness of the discriminant power of a function and provides valuable information when differentiation is made between two groups. Again, the effectiveness of the function is known from the eigenvalues extracted. The canonical correlation measures the percentage of the total variance that in the $i-th$ function y_i is being explained by the difference between groups. Consequently, the $i-th$ function y_i is more discriminant the closer the canonical correlation is to one. This value is obtained from (8):

$$C(\mathrm{y}_i) = \sqrt{\frac{\lambda_i}{1 + \lambda_i}}, \tag{8}$$

where $C(y_i)$ is the canonical correlation of the discriminant function y_i, λ_i is the eigenvalue associated to the discriminant function y_i.

Finally, the function test evaluates in the constructed model the ability to differentiate between groups. The overall significance of the discriminant function is obtained using the Wilks' Lambda contrast statistic of equation (9):

$$\Lambda = \frac{1}{1 + \lambda_i}, \tag{9}$$

where λ_i is the eigenvalue of the linear discriminant function. The linear function discriminates more between groups when the Wilks' Lambda statistic is closer to zero.

5) FORECAST PERFORMANCE

At this stage, the predictive power of the classification model is evaluated. After training the linear discriminant function, the model is used to classify the direction of the future movement of the euro-dollar exchange rate. By replacing the measures of the predictor variables $\hat{\mathbf{x}}_j$ in the discriminant function $y(\hat{\mathbf{x}})$, it predicts, in the short term, the direction of the movement of the euro-dollar exchange rate. The performance of the model is evaluated using four additional samples, OOS_1, OOS_2, OOS_3 and OOS_4, which capture information from different market conditions and time horizons over a 15-minute time frame. The performance indices used to measure the predictive power of the model are described in more detail in the section (IV-B). The use of performance measures and additional data allows an objective and comparative evaluation of the model with other approaches in terms of generalization, accuracy, and interpretability.

6) LCC ALGORITHM

Algorithm 1 is a pseudocode that gathers all the steps of the proposed LCC methodology.

IV. EXPERIMENTAL SETUP

Considering the nature of the data and their asymmetric and leptokurtic arrangement, the proposed methodology is designed to reduce the natural bias of the data and to give reliability to the interpretations. To this end, a subset of the data is extracted and prepared. The cases are independent. Predictors and groups assume a multivariate normal distribution. Intra-group variance-covariance matrices are equal and the membership of each case to a particular group is exclusive.

The identification of the determinants of the directional movement of the euro-dollar exchange rate due to the change of "market regime" is approached as a problem of selection of observable characteristics [80] and the distinction between

Algorithm 1 Linear Classifier Configurator (LCC) for Predicting the Directional Movement of the Euro-Dollar Exchange Rate

Input: Market Data (OHLCV): $\mathbf{Q}\ \langle 375000 \times 7\rangle$ 15-minutes, $\mathbf{W}\ \langle 913 \times 7\rangle$ 1-week, Trading system, IWT data: $\mathbf{A}\ \langle 14770 \times 18\rangle$.

1: Data preparation:
- Extract **ET** inflection points in $\mathbf{A}\ \langle 14770 \times 18\rangle$ from $\mathbf{Q}\ \langle 375000 \times 7\rangle$ using Trading system.
- Measure **ET** inflection points using Table 2 variables.
- Assign labels to trends using Table 4.
- Extract population data $\mathbf{O}\ \langle 1873 \times 2\rangle$ from $\mathbf{A}\ \langle 14770 \times 18\rangle$ based on weekly market data $\mathbf{W}\ \langle 913 \times 7\rangle$ using Table 5.
- Extract the normally distributed sample $\mathbf{Z}\ \langle 125 \times 2\rangle$ from the population $\mathbf{O}\ \langle 1873 \times 2\rangle$.
- Extract selected subset features $\mathbf{B}\ \langle 14770 \times 7\rangle$ with underlying structures from $\mathbf{A}\ \langle 14770 \times 18\rangle$.

2: Feature selection:
- Use candidate variables from $\mathbf{A}\ \langle 14770 \times 18\rangle$ with specified parameters from Table 6.
- Conduct one-way ANOVA analysis to assess group differences among candidate variables.
- Evaluate candidate variables' discriminant potential via stepwise variable selection (Equations 1, 2, and 3).
- Validate the equality of means among groups using the Levene's test.
- Measure the linear relationship between variables with Pearson's correlation coefficient.
- Extract the subset $\hat{\mathbf{X}}$ of discriminative and independent variables $\mathbf{A}'\ \langle 14770 \times 2\rangle$.

3: Multivariate analysis:
- Use the subset of features $\mathbf{B}\ \langle 14770 \times 7\rangle$ for structural analysis with GH-Biplot (Section III-B3)
- Assess data adequacy for structure detection with the KMO measure and Barlett's test of sphericity.
- Validate:
  - Individual differences in trend movements from a multivariate perspective.
  - The discriminative efficacy and independence of the selected predictors.
  - The effectiveness of discriminant analysis.

4: Discriminant analysis:
- Use the data matrix $\mathbf{Z}\ \langle 125 \times 2\rangle$ and the parameters in Table 6.
- Analyze descriptive statistics of $\mathbf{Z}\ \langle 125 \times 2\rangle$ as described in Table 13.
- Verify the fulfillment of statistical assumptions (Section V-C2).
- Formulate, validate and evaluate the classification model ($y(\hat{\mathbf{X}}) = \beta_0 + \beta_1\hat{X}_1 + \beta_2\hat{X}_2$) following Equations (4), (5) and (6).
- Assess the model fit to the data structure using goodness-of-fit tests based on Equations (7), (8), and (9).

Output: Linear discriminant model $y(\hat{\mathbf{X}}) = \beta_0 + \beta_1\hat{X}_1 + \beta_2\hat{X}_2$.

groups is made with a single categorical variable consisting of two classes (uptrend and downtrend).

The prediction of the direction of the exchange rate is approached as a supervised learning classification problem. The formulation of the classification model is done based on the sampled data and the selected variables. The performance evaluation of the classification model is done with four additional data samples. Therefore, to validate the effectiveness of the results obtained in all experiments the proposed methodology uses an interpretive approach based on the statistical power of parametric tests and on the reliability of the results.

During the experimental phase, six key technical considerations significantly impact the quality of results achieved with the proposed methodology: (i) Ensure data quality by maintaining accuracy and consistency, particularly in critical activities during the preparation phase. (ii) Use discriminative variables that effectively differentiate between observations. (iii) The categories of the dependent variable must be relevant in number, independent, and significantly discriminative. (iv) The effectiveness of the selection process of predictor variables must be observable through GH-Biplot structure analysis. (v) The integration of these partial solutions in a classification model must be reflected with high-performance values. Finally, (vi) statistical significance tests must support the generalization and validity of interpretations and conclusions.

The open source, machine learning PyCaret library was used in Python (3.9.5) notebook environment for data preparation and model implementation. IBM SPSS Statistics for Windows, Version 27.0. (IBM Corp. Released, 2019) [81] and RStudio Team (2021) software were used for statistical analysis. MultBiplot software [82] was used for GH-Biplot analysis and multivariate data visualization. The MVN [83] package in R code was used to evaluate multivariate and univariate normality assumptions and provide the graphical evidence. The experiments were performed on an Intel(R) Xeon(R) Silver 4110 CPU @ 2.10Ghz (2 Processors) with 32 Gb in 64-Bit RAM.

A. TRAINING, VALIDATION AND TESTING

The analysis sample consists of 125 observations divided into two subsets, one for training and one for evaluation. The size of these subsets should be large and representative enough for the model to accurately learn, generalize, and assess its performance. However, there is no standard rule defining their specific size. For the analysis, the cases are randomly selected, approximately 70% of the upward and downward movements are used to create the model (training set) and

the same cases are used to cross-validate the model. The remaining trend movements, approximately 30%, are used to validate the performance of the model. Four additional data samples are used to evaluate the predictive power of the model. Set OOS_3 consists of 1748 confirmed trend movements on a long-term weekly time frame and sets OOS_1, OOS_2 and OOS_4 consist of 6581, 14645 and 2336 trend movements measured on a 15-minute time frame. The table, 6, provides the parameter values used to conduct all experiments, including technical specifications of the criteria used in the feature selection process, the discriminant analysis and the analysis sample sizes. In the feature selection process, the default values assigned for Tolerance (ψ_i), input levels ($F_\varepsilon = 3.84$) and removal ($F_r = 2.71$), in the attribute evaluation, are the ones that produce the best performance.

B. PERFORMANCE MEASUREMENTS

The performance of the classification model is measured based on the results given in the table 7. The confusion matrix consists of four categories. In the main diagonal are placed the observations correctly classified by the model. There are the True Positives T_P which is the number of observations correctly classified as bullish movements and the True Negatives T_N which is the number of cases correctly classified as bearish movements. On the opposite diagonal are the cases incorrectly classified by the model. The False Positive F_P is the number of bearish moves incorrectly classified as bullish and the False Negative F_N is the number of bullish moves that are misclassified as bearish.

Although there is no general consensus on the effective use of a particular performance measure, *Accuracy* is one of the most commonly used metrics to evaluate the performance of models in classifying new observations. The equation (10) measures the proportion of observations that the model classifies correctly:

$$Accuracy = \frac{T_P + T_N}{F_P + T_P + T_N + F_N} \tag{10}$$

The geometric mean $G - mean$ is a performance measure that is used when groups are unbalanced [84]. Equation (13) measures the geometric mean of the positive hit ratio (Recall or Sensitivity) and the negative hit ratio (Specificity). *Recall* and *Specificity* are measured from equations (11) and (12) respectively.

$$Recall = \frac{T_P}{T_P + F_N} \tag{11}$$

$$Specificity = \frac{T_N}{F_P + T_N} \tag{12}$$

$$G - mean \quad Accuracy = \sqrt{Recall \cdot Specificity} \tag{13}$$

The positive predictive value, also called *Precision*, is the ratio computed between the number of cases classified as true positives and the total number of positive cases. The equation (14) defines its calculation.

$$Precision = \frac{T_P}{F_P + T_P} \tag{14}$$

$F_1 - Score$ is the harmonic mean of *Recall* and *Precision*. This measure is useful for comparing prediction quality between models and can be interpreted as a function of counts of true positives, false negatives, and false positives [85]. Equation (15) shows how it is calculated.

$$F_1 = \frac{2 \cdot Recall \cdot Precisin}{Recall + Precision} \tag{15}$$

The *Kappa* statistic is calculated from the equation (16). It is useful in model building and is used to find out whether the formulated model is better, equal or worse than a random model [86]. Note that the *Kappa* statistic is computed as a function of a prime rate called *Baseline*. This rate defines the baseline or benchmark for model comparison and is computed with the expression (17):

$$Kappa = \frac{Accuracy - Baseline}{1 - Baseline}, \tag{16}$$

where

$$Baseline = \frac{(PP \cdot AP)}{(PP + PN)^2} + \frac{(PN \cdot AN)}{(AP + AN)^2}. \tag{17}$$

Matthews correlation coefficient *MCC* is a robust rate statistic that provides a more reliable measure of the performance of the classification model. It produces higher scores only when the predictions are correct in all four categories of the confusion matrix [87]. It is calculated as shown in equation (18):

$$MCC = \frac{T_P \cdot T_N - F_P \cdot F_N}{\sqrt{(T_P + F_P) \cdot (T_P + F_N) \cdot (T_N + F_P) \cdot (T_N + F_N)}}. \tag{18}$$

V. RESULTS AND DISCUSSION

This section presents the results of the methodology outlined in Figure 3. The results are discussed in four distinct areas, highlighting key findings and insights. These areas include: A. Selection of attributes with high discriminant power that best explain the price action; B. Multidimensional analysis of differences between trend movements; C. Construction, validation, and evaluation of a classification model for predicting the short-term trend of the euro-dollar pair; and D. Evaluation of the model's performance using out-of-sample data and comparison with other forecasting approaches.

A. FEATURE SELECTION

According to the table 2, of the first 15 metrics that capture information prior to the market regime change, at the **ET** instant, the variables that are relevant to the study are not known a priori. Not all of them are useful and significant in the process of discriminating the direction of the future movement of the euro-dollar exchange rate. The importance of the variables lies not only in the statistical and individual significance that each one contributes to the differentiation between groups, but also in their explanatory capacity for the interpretation of the study problem. The identification and selection of the predictor variables that best contribute

TABLE 6. Parameter values.

Parameters	Criteria	Description
Feature selection		
F-Snedecor to enter	$F_e = 3.84$	Threshold for input variable.
F-Snedecor to remove	$F_e = 2.71$	Threshold for variable removal.
Minimum tolerance	$\psi = 0.001$	Minimum value to enter in the analysis.
Discriminant analysis		
Probabilities of equal groups	$\pi_g = 0.50$	Definition of function coefficients.
Covariance matrix	Within-groups equal	Classification of cases.
Cross-validation	Leave-one-out classification	Classification for cross validation.
ET Inflection point sets for analysis		
Training set:	(70%) 87/125 observations	Selected cases to create the model.
Cross-validation set:	(70%) 87/125 observations	Cross-validated cases.
Testing set:	(30%) 38/125 observations	Cases selected to evaluate the model.
Out-of-sample sets:		
OOS_1: 1999-2005	6581 observations	Price movements in bear and bull markets.
OOS_2: 2006-2020	14645 observations	Price movements in a bearish market channel.
OOS_3: 2006-2020	1748 observations	Price movements in a bearish market channel.
OOS_4: 2021-2023	2336 observations	Price movements in bear and bull markets.

TABLE 7. Confusion matrix.

Predictions / Actual	Actual Positives	Actual Negatives	Total
Predicted Positive PP	True Positives T_P	False Positives F_P	PP = T_P + F_P
Predicted Negative PN	False Negatives F_N	True Negatives T_N	PN = F_N + T_N
Total	AP = T_P + F_N	AN = F_P + T_N	AP+AN = PP+PN

to the differentiation between upward and downward trend movements is done in three steps, using the multivariate data matrix **A**.

1) EVALUATION OF THE CONTRIBUTION OF EACH PREDICTOR

The test for equality of group means, from an exploratory point of view, helps to identify the potential that each predictor variable has in group differentiation before participating in the formulation of the model. The evaluation of the contribution that each variable makes in the differentiation of groups allows to recognize the predictor variables that should be considered in the study. The test of equality of group means referred to in the table 8 contrasts the H_0 on equality of means between groups. According to one-way ANOVA analysis and with a significance value higher than 5%, the H_0 on equality of means between groups is accepted, so that the variables highlighted in gray do not have the potential to discriminate between upward and downward trend movements and, therefore, should not be considered in the study.

2) EVALUATION OF THE DISCRIMINANT POWER OF EACH PREDICTOR

The search for a simple, parsimonious and easy to explain prediction model leads to the identification and selection of predictors that best explain the price action and, consequently, best contribute to the separation between groups. The stepwise variable selection method is useful in achieving this purpose and thus ensures the identification and arrangement of the best candidate variables to be used in the model formulation.

Table 9 summarizes the statistics for the subset of significant candidate variables that could be incorporated into a single model and, taken together, are the predictor variables that best discriminate between upward and downward trend movement.

At each step, the predictor with the highest F-value for entry that exceeds the entry criterion (default, 3.84) is added to the model. Thus, the small values of the Wilks' Lambda statistic obtained at each step indicate that the subset of predictors chosen are the ones that best contribute to the separation between groups. In the last step the variables that are left out of the analysis all have F-values for entry smaller than 3.84, so no more are added to the subset.

According to the results of the table 9, s.V.18P does not minimize the value of the Wilks' Lambda statistic obtained in the previous step (7), consequently this variable (8) does not contribute information to the model and therefore is excluded from further analysis.

The potential that each predictor has in class differentiation and the ability to achieve a high level of classification performance are measured through the Wilk's Lambda statistic. Figure 8 shows in order of importance the discriminant power of the best predictor variables. x.V*18P, Sp.Vr*CP.18P and Sp.CP.18P are the variables that best explain the price action prior to the market regime change. The second, third and

TABLE 8. One-Way ANOVA.

Measure	Sum of squares	Df_1	Df_2	Mean square	F-Statistic	P-Value
x.V*18P	0.00	1	14768	0.00	16856.79	0.000
Sp.Vr*CP.18P	0.00	1	14768	0.00	5001.01	0.000
s.Vr.Vo.18P	1.80	1	14768	1.80	1.17	**0.279**
x.Vr.Vo.18P	0.07	1	14768	0.07	0.67	**0.412**
Sp.Vr.Vo.18P	0.00	1	14768	0.00	0.05	**0.830**
Sp.Vo.18P	24.89	1	14768	24.89	0.01	**0.905**
s.V.18P	0.00	1	14768	0.00	5.50	0.019
Sp.CP.18P	0.00	1	14768	0.00	11645.00	0.000
Sp.Vr.HL.18P	0.00	1	14768	0.00	0.16	**0.688**
Sp.Vr.CO.18P	0.04	1	14768	0.04	0.19	**0.667**
s.Vr.RSI.18P	10.41	1	14768	10.41	1962.97	0.000
Sp.Vr.RSI.18P	0.09	1	14768	0.09	1500.26	0.000
Dv.CP-RSI.18P	0.06	1	14768	0.06	1.38	**0.241**
Dv.Vo-CP.18P	872.86	1	14768	872.86	4266.50	0.000
Dv.Vo-RSI.18P	1129.54	1	14768	1129.54	6808.76	0.000

TABLE 9. Discriminant power of predictor variables.

Step	Selected Features	Wilks' Lambda	Df_1	Df_2	Df_3	F-Value	Df_1	Df_2	P-Value
1	x.V*18P	0.467	1	1	14768	16856.787	1	14768	0.000
2	Sp.Vr.RSI.18P	0.365	2	1	14768	12851.92	2	14767	0.000
3	Dv.Vo-RSI.18P	0.337	3	1	14768	9666.854	3	14766	0.000
4	s.Vr.RSI.18P	0.327	4	1	14768	7590.954	4	14765	0.000
5	Sp.Vr*CP.18P	0.325	5	1	14768	6137.861	5	14764	0.000
6	Sp.CP.18P	0.324	6	1	14768	5142.587	6	14763	0.000
7	**Dv.Vo-CP.18P**	**0.323**	7	1	14768	4410.364	7	14762	0.000
8	**s.V.18P**	**0.323**	8	1	14768	3861.226	8	14761	0.000

fourth variables measure market behavior as a function of the technical indicator $RSI(9P)$. Finally, the variable Dv.Vo-CP.18P measures the divergence between the closing price and the volume in the market.

Levene's test for homogeneity of variances allows us to establish whether the study groups come from populations with equal variances. The reliability of the test of equality of means, one-way ANOVA, is validated by compliance with the test of homogeneity of variances. Consequently, by verifying in the table 10 the fulfillment of the assumption of homogeneity of variances, with Levene's statistic, the rejection of the hypothesis Ho that the means of the groups of predictors 1, 2 and 6 are equal is reaffirmed. Therefore, now that it is known for the predictors highlighted in gray that the group means do indeed differ, it is possible through multivariate analysis to approach the knowledge of the differences that changes in the market regime generate in the price action.

3) INDEPENDENCE ANALYSIS BETWEEN PREDICTOR VARIABLES

According to the results presented in the table 11, there is a high correlation between the variables Sp.CP.18P and x.V*18P. The 95.6% correlation is significant at the 1% bilateral level. If all variables are used in the model formulation. The instability in the signs of the coefficients of the Fisher linear discriminant functions could negatively affect the classification results. Therefore, to avoid the multicollinearity problem, the variable Sp.CP.18P is excluded from the model formulation.

In summary, according to the proposed methodology, out of 15 candidate variables, the predictor and independent variables selected according to the exploratory confirmatory approach are two. Average yield of closing prices x.V*18P and Slope of the regression line of the variations between closing prices Sp.Vr*CP.18P, both calculated on the basis of the last 18 periods before the change of market regime at the **ET** time. According to statistical merit, the chosen subset of variables $\hat{\mathbf{x}}_j$ are the ones that best contribute to group separation and have the highest discriminant power and, from a technical analysis point of view, are the ones that best explain the price action before a market regime shift occurs at the **ET** instant.

B. MULTIVARIATE ANALYSIS

The purpose of the **GH − Biplot** analysis is to detect and describe, in the underlying structures of the data, the multidimensional nature of the differences between the onset of an upward and downward directional movement as a function of the chosen predictor variables $\hat{\mathbf{x}}_j$. The results of the analysis allow us to confirm the suitability of the selected predictor variables and the relevance of the discriminant analysis in the classification tasks.

The analysis is performed on the data collection $\mathbf{B}\langle 14770 \times 7\rangle$ consisting of 14770 trend movements on which 7 numerical observable variables and a categorical

FIGURE 8. Discriminant ability of the predictor variables. Order of importance of predictors according to discriminant capacity between groups. The Wilks' Lambda statistic measures the discriminant potential of each predictor to differentiate between bullish and bearish movements. Predictors with F-value greater than the entry criterion (default, 3.84) are suitable for differentiating between class averages.

TABLE 10. Variance Homogeneity Test.

ID	Measure	Levene Statistic	Df_1	Df_2	P-Value
1	x.V*18P	0.578	1	14768	**0.446**
2	Sp.Vr*CP.18P	1.820	1	14768	**0.177**
3	Sp.Vr.RSI.18P	439.873	1	14768	0.000
4	Dv.Vo-RSI.18P	16.092	1	14768	0.000
5	s.Vr.RSI.18P	188.975	1	14768	0.000
6	Sp.CP.18P	3.165	1	14768	**0.075**
7	Dv.Vo-CP.18P	9.311	1	14768	0.002

TABLE 11. Independence test of predictor variables.

Measure	Sp.Vr*CP.18P	x.V*18P	Sp.CP.18P
Sp.Vr*CP.18P	1	0.362	0.312
x.V*18P		1	**0.956**
Sp.CP.18P			1

variable labeling the direction of the trend of the euro-dollar exchange rate are measured in two groups $\check{y}_g$. The groups, according to the label vector $\dot{\mathbf{y}} = \{0, 1\}$, are identified with the categories $0 =$ "Upward Trend Movement" and $1 =$ "Downward Trend Movement".

The effectiveness of the $\mathbf{GH} - \mathbf{Biplot}$ analysis in detecting underlying structures is determined by the adequacy of the data. The relevance of the multivariate data matrix $\mathbf{B}\langle 14770 \times 7\rangle$ for structure detection is verified by the Kaiser-Meyer-Olkin KMO sampling adequacy measures and Bartlett's test of sphericity. The results of these statistical tests suggest that the analysis is appropriate. The KMO value of 0.639 close to 1 indicates that the proportion of variation in the variables used in the analysis are caused by underlying factors. The Chi-square value $\tilde{\chi}^2 = 47208.373$ associated with the KMO test statistic and a P-Value of less than 5% suggests rejection of the hypothesis H_0 that the correlation matrix is an identity matrix ($\tilde{\chi}^2 = 47208.373$, Df $= 21$, P-Value $= 0.000$). Hence, covariation between variables is suitable for the detection of underlying structures.

The manifest variables that best explain the factor axes are summarized in the table 12. The relative factor contributions to the column items can be interpreted as a measure of goodness-of-fit with which the observed variables best explain the factor axes. The reliability of the analysis and of the suggested classification model is verified by the accuracy achieved in the classification process.

The table 12 details, by groups, the variables involved in the analysis. The first group measures the behavior of the closing prices preceding the trend change at the time **ET**. This group of variables 1, 2, 4 and 5 are the ones that best discriminate between upward and downward movement directions. The proposed solution, by the feature selection method, is formed by variables 1 and 4 highlighted in gray color. The second group is formed by variables 3, 6 and 7 that financially validate the expected trend direction with the realization of profits after the **ET** moment. Other parameters available in the table 2 do not participate in the analysis because they do not help to explain the multidimensional nature of the differences between the beginning of a bullish and bearish movement.

Figure 9 shows the $\mathbf{GH} - \mathbf{Biplot}$ representation of the $\mathbf{B}$ data collection. The interpretation of the proposed solution is made on the first two factors that explain 55.012% of the accumulated variability of the analyzed variables. This

TABLE 12. Relative contributions of the factor to the column elements.

ID	Variables	Group 1	Group 2	Axis 1	Axis 2
1	x.V*18P	1		**875**	2
2	Sp.CP.18P	1		**831**	5
3	Sp.XP-EP		2	411	27
4	Sp.Vr*CP.18P	1		301	75
5	s.Vr.RSI.18P	1		221	119
6	RP%M[1]		2	0	**461**
7	GPL$		2	0	**523**

suggests two latent influences associated with the change in direction of the preceding trends, with a significant unexplained margin of variation. Therefore, the proportion of variance absorbed is sufficient to explain the underlying relationship between the historical behavior of the exchange rate and gross realizable profit, when the direction of the future trend is known a priori.

The column markers or vectors represent the variable measurements (technical parameters before and after the onset of each microtrend after the **ET** moment). The row markers represented by dots define, in terms of values, the time prior to the trend change **ET**. The scale of measurement of the vectors defines the magnitude of the variance of the measurements. The angles formed between vectors represent the correlations between variables. The center of the GH-Biplot is the midpoint of the differences between upward and downward movements.

Consequently, two latent factors are identified and are associated with the multivariate nature of the differences between groups. The first factor axis (Axis 1) describes the historical behavior of the exchange rate preceding the trend change. This axis is highly correlated with the variables: average return x.V*18P and closing price trend Sp.CP.18P. In the first factor, there is greater variability in the left-hand side measurements than in the origin of the GH-Biplot. These negative variations, between closing prices, depreciate the exchange rate. Thus, the sustained accumulation of negative average returns is an indicator that precedes the beginning of a new upward movement. This effect experienced by the depreciation of the euro against the US dollar is highlighted during the period 2006 - 2020.

Towards the right side of the first factor, the variability of the measurements is lower. The sustained appreciation of the exchange rate is an effect of positive variations between closing prices and is an indicator that precedes the beginning of a new downward movement. Consequently, the first factorial axis is the one that best discriminates between group means. According to the covariation angles formed between the factor and the explaining variables, the mean return x.V*18P is a better predictor. Note that the high correlation with the closing price trend Sp.CP.18P may generate multicollinearity problems, in case it is also decided to include this variable in the construction of the model.

The second factorial axis (Axis 2) describes the risk premium. This axis is explained by the variables: Market Risk RP%M[1] and Gross Profit or Loss GPL$. Both vectors score in the opposite direction. Theoretically, gross profits subject to risk are achieved, with the opening and closing of trades, depending on the start and end of each micro-trend. Thus, on the second axis, the variability of the profit obtained as a function of risk exposure is greater above the midpoint of the biplot than below.

The plane variables: slope of the variations between closing prices Sp.Vr*CP.18P measured before the moment **ET** and slope of the micro-trend Sp.XP-EP measured after the event **ET**, between both vectors, register an inverse covariation. Thus, negative slopes of the variations between closing prices precede the formation of upward micro-trends (Points in quadrants II and III) and positive slopes of the variations between closing prices precede the formation of downward micro-trends (Points in quadrants I and IV). The independence between Sp.Vr*CP.18P ands.Vr.RSI.18P is confirmed by the 90° angle they form between them, note that this feature is not sufficient to discriminate between bullish and bearish movements.

Overall, taken together, these results suggest the formation of four quadrants that explain the multivariate nature of the differences between bullish and bearish moves. Quadrants II and III can be interpreted as the oversold zone, where each row marker represents, with respect to the first factor axis, the lowest depreciation experienced by the exchange rate prior to the trend change. Quadrants I and IV can be understood as the overbought zone, where the observations represent, with respect to the first factorial axis, the highest appreciation reached prior to the change in trend. Quadrants I and II can be interpreted as high risk exposure zone, where each row marker corresponds, with respect to the second factorial axis, to Micro trends generating positive gross profits with exposure to floating losses before reaching the target closing price. Quadrants III and IV can be understood as low risk exposure zone, where each row marker corresponds, with respect to the second factorial axis, to positive gross profit generating micro trends with exposure to floating gains before reaching the target closing price.

The evidence from this study suggests that the first factor axis best explains the differences between the means of the

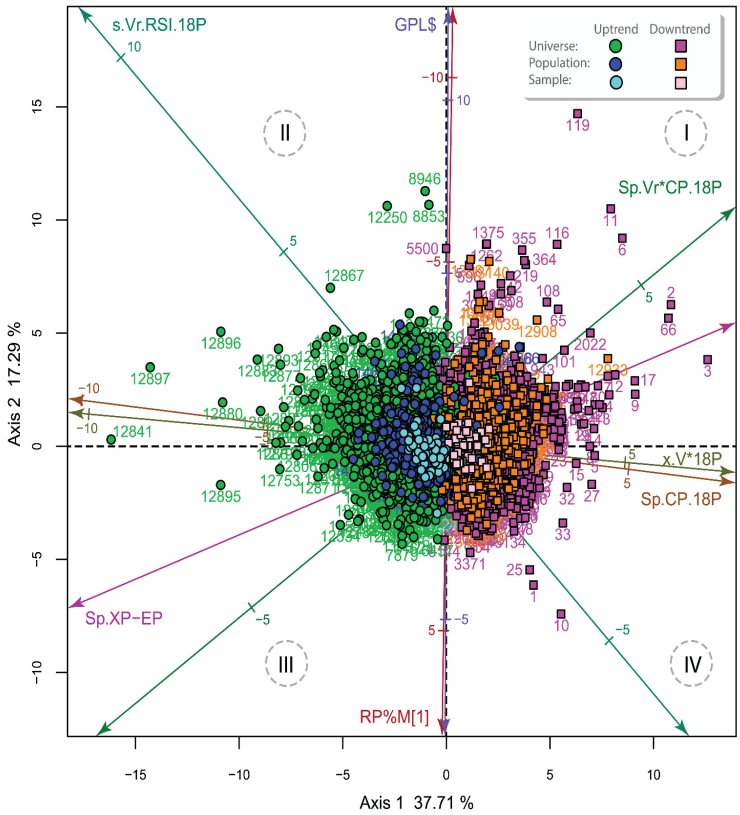


FIGURE 9. **GH – Biplot of euro-dollar exchange rate movement determinants.Normalized data is standardized by column and partitioned according to *SVD*. Factorial plane 1-2. Explained variance of 55.012% (37.71% of axis 1 and 17.29% of axis 2). Detection of underlying relationships between variables (vectors) and identification of predictors that best discriminate between bullish and bearish micro-trends (row markers in quadrants II-III and quadrants I-IV, respectively).**

upward and downward movements. Given the pattern of correlation with the observable variables, the predictor variable x.V*18P is the best contributor in differentiating between the onset of a new upward and downward trend movement. This predictor is the most influential to consider in building a classification model. Therefore, a discriminant analysis model is sufficient to distinguish, classify and predict the beginning of a trend movement.

C. LINEAR DISCRIMINANT ANALYSIS MODEL

To explore the advantages of the solutions provided by the proposed methodology, in terms of data, predictors and context, this paper evaluates the performance of the classification model based on three established measures: (**i**) sample with multivariate normality to overcome the asymmetric and leptokurtic nature of the data; (**ii**) predictor variables chosen according to technical and statistical merit; (**iii**) adequacy of the discriminant analysis with respect to the multidimensional nature of the differences between bullish and bearish movements. To this end, this section presents the results in five specific areas. 1) Descriptive statistics for the study subset ($\mathbf{Z}$ data matrix). 2) Verification of assumptions and requirements of the discriminant analysis. 3) Evaluation of model fit. 4) Classification of the direction of euro-dollar exchange rate movements and 5) Validation of the model.

Discriminant analysis is performed on the data set $\mathbf{Z}\langle 125 \times 2\rangle$ using two independent variables $\hat{\mathbf{x}}_j$ that collect information about the price action. These predictors are used to classify the observations $\mathbf{i}$ into two groups $\check{y}_g$. The $\hat{\mathbf{x}}_j$ predictor variables are: Average closing price returns x.V*18P and slope of the regression line of the inter-closing price changes Sp.Vr*CP.18P, both predictors are calculated 18 periods prior to the **ET** study time, at which the change in direction of the preceding trend occurs. The clusters, according to the vector of labels $\dot{\mathbf{y}} = \{0, 1\}$, are trend movements described by the euro-dollar exchange rate,

where 0 = "Upward Trend Movement" and 1 = "Downward Trend Movement". These class labels identify the beginning of a new short-term directional movement (measured on a 15-minute time frame) confirmed on a long-term (1-week) time frame.

1) DESCRIPTIVE STATISTICS

Descriptive statistics prior to the onset of the **ET** event, where the change in direction occurs relative to the onset of new directional motion, are reported in the table 13. On average, the mean return x.V*18P records, over the analysis period (2006-2020), a slight depreciation of the euro against the US dollar with a high variance in mean returns.

The slope of the variations between closing prices Sp.Vr*CP.18P preceding the formation of a new market regime (bullish/bearish) reports on average a slight positive trend with high variation in the measurements.

When analyzing the average values of the groups, the average of the mean returns x.V*18P that precede the formation of an uptrend reports on average a slightly higher loss of value (depreciation) relative to its appreciation process, and a high variation between closing prices.

The slope of the closing price-to-closing price variations Sp.Vr*CP.18P preceding the onset of a directional up/down move reports on average a slight negative/positive trend with high variance in the measurements.

While, the average x.V*18P returns preceding the formation of bullish/bearish moves report a slight negative/positive skewness, this asymmetric and leptokurtic effect is common in the returns generated by financial assets. However, the subset of data overcomes this problem, as confirmed in the next section.

2) VALIDATION OF ASSUMPTIONS

The multivariate normality assumption is one of the most important requirements of some multivariate parametric statistical procedures to ensure the reliability of interpretations. If significance tests and goodness-of-fit assessment are not satisfied, it will negatively affect the reliability of interpretations based on the results of these procedures [88], [89]. The results of tests validating compliance with these requirements are presented below.

- Multivariate normality tests for predictors.

There are numerous tests that evaluate multivariate normality assumptions but there is no single standard procedure. However, the Henze-Zirkler and Royston tests suggested in [90] are used in this study because of their good control of type I error and power. Table 14 shows the results of the Henze-Zirkler test for samples larger than 100 observations and its result is validated with the Royston test.

The predictors x.V*18P and Sp.Vr*CP.18P with a significance level greater than 0.05 assume a multivariate normal distribution. In this context, the Henze-Zirkler test statistic follows an approximately log-normal distribution and, therefore, it is more likely that the data, as a function of the variables and groups, also assume a univariate normal distribution [88].

Graphical methods also validate the results obtained. Figure 10 summarizes this analysis. The first graph (**a**) shows the clear agreement between the quantiles of the hypothesized and observed probability distributions. Values in Y tend to be equal in X. That is, the evaluated data do not deviate from multivariate normality.

The ellipsoidal contour line plot (**b**) is useful to verify the multivariate normality of the predictors. The top view of the contour plot of x.V*18P and Sp.Vr*CP.18P hints at a slight Gaussian bell figure with an obvious separation and differentiation between the negative and positive mean returns generated by the Depreciation / Appreciation of the exchange rate, prior to the change in trend. Looking at the histograms in Figure 11, this difference is more evident in x.V*18P than in Sp.Vr*CP.18P and this is because the logarithmic returns are leptokurtic and asymmetric. In addition, the contour lines confirm the positive correlation between groups, as detected in the **GH − Biplot** de la figura 9.

The perspective plot (**c**) is a bivariate representation on a three-dimensional probability distribution surface. It gives information about where the data tend to be concentrated and since the variables are correlated, its shape approximates a Gaussian distribution.

- Univariate normality tests for predictors.

The results of the tests in the table 15 indicate that the subset of data satisfies the assumptions of univariate normality at a significance level of 5%. $AD - Statistic$ is the value of the Anderson-Darling test statistic used to test the hypothesis H_0 that the sample is from a normally distributed population.

Figure 11 verifies through normality plots the structure of the analyzed data. The predictors x.V*18P and Sp.Vr*CP.18P have a univariate normal distribution at 5% of significance level.

- Multivariate normality tests for groups.

Since the significance value derived from the Henze-Zirkler and Royston tests are mathematically greater than 0.05 the results of the 16 table allow us to conclude that the data set of the directions of the upward and downward movements of the euro-dollar exchange rate conform to a multivariate normal distribution.

While the subsets of observations on uptrending and downtrending price directions fit a multivariate normal distribution, in Figure 12 in the contour (**b**) and outlook (**c**) plots it is observed that the distribution of data for the uptrending and downtrending groups have a slight negative and positive skewed distribution. As mentioned, this is due to the fact that the closing price returns, prior to the trend change, have an asymmetric and leptokurtic behavior. However, it is emphasized that these samples follow a multivariate normal distribution.

If the results of the table 16 indicate that the subset of data satisfies the assumptions of multivariate normality at a significance level of 5% and the graphical analysis of the

TABLE 13. Descriptive statistics of predictors and groups. Values prior to time ET.

Measure	n	Mean	Std.Dev	Median	Min	Max	Skew	Kurtosis
Predictor variables								
x.V*18P	125	-0.0000024	0.0001211	0.0000017	-0.0002938	0.000263	-0.0880643	-0.883730
Sp.Vr*CP.18P	125	0.0000011	0.0000204	0.0000013	-0.0000482	0.0000461	-0.021371	-0.3078016
Upward movement								
x.V*18P	62	-0.0001056	0.0000676	-0.0000967	-0.0002938	0.0000006	-0.3798269	-0.4650322
Sp.Vr*CP.18P	62	-0.000012	0.000016	-0.0000122	-0.0000482	0.0000286	-0.0526824	-0.0757937
Downward movement								
x.V*18P	63	0.0000992	0.0000607	0.0000929	0.0000017	0.000263	0.4674263	-0.4223248
Sp.Vr*CP.18P	63	0.0000139	0.0000154	0.0000124	-0.0000207	0.0000461	0.1728039	-0.5063708

TABLE 14. Multivariate normality tests for predictors.

Test	Measures		Statistic	P-Value
Henze-Zirkler	x.V*18P	Sp.Vr*CP.18P	0.987	0.051
Royston	x.V*18P	Sp.Vr*CP.18P	2.708	0.260

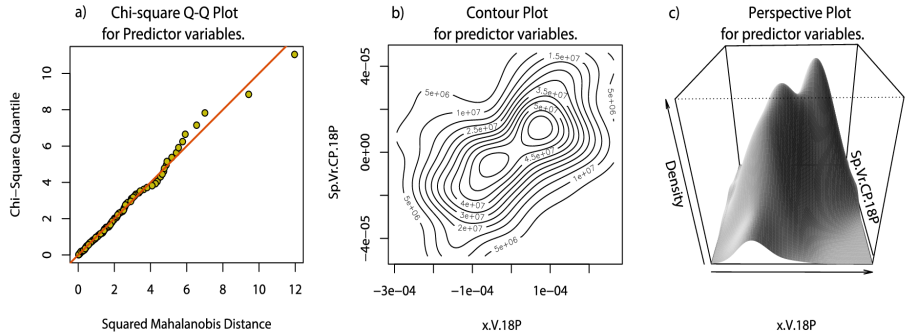


FIGURE 10. Multivariate charts of predictor variables. The structure of the data matrix $\mathbf{Z}\langle 125 \times 2\rangle$ approximates a multivariate normal distribution. (a) Chi-Square Q-Q Plot of long-term movements. (b) Contour Plot from x.V*18P and Sp.Vr*CP.18P. (c) Perspective plot from predictor variables.

TABLE 15. Univariate normality tests for predictors.

Test	Measure	Cases	AD-Statistic	P-Value
Anderson-Darling	x.V*18P	125	0.724	0.058
Anderson-Darling	Sp.Vr*CP.18P	125	0.182	0.912

TABLE 16. Multivariate normality tests for groups.

Test	Trend	Measures		n	Statistic	P-Value
Henze-Zirkler	Upward movement	x.V*18P	Sp.Vr*CP.18P	125	0.874	0.061
	Downward movement	x.V*18P	Sp.Vr*CP.18P	125	0.594	0.317
Royston	Upward movement	x.V*18P	Sp.Vr*CP.18P	125	1.737	0.419
	Downward movement	x.V*18P	Sp.Vr*CP.18P	125	2.705	0.259

figure 12 corroborates the results of the tests, then it is valid to state, as confirmed in the following section, that each predictor and hence each group within each predictor has a univariate normal distribution [88].

- Unvariate normality tests for groups.

According to results in table 17, in all analyses, the P-Value of the Anderson-Darling test statistic is greater than 0.05, consequently, the subsets of observations from each of

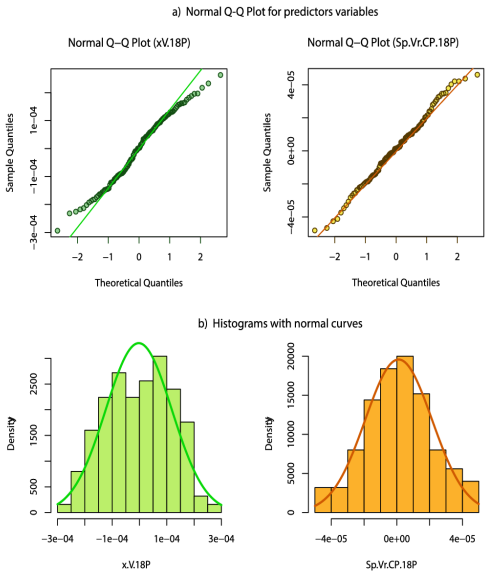


FIGURE 11. Univariate charts of predictor variables. The structure of the data matrix $\mathbf{Z}(125 \times 2)$ approximates a normal distribution. (a) Normal Q-Q Plot for predictors. (b) Histograms with normal curves.

the trend groups within each of the predictors x.V*18P y Sp.Vr*CP.18P come from a normally distributed population.

As seen in 13, the distributions of the data for the upward and downward trending groups follow a normal distribution within each of the predictors x.V*18P and Sp.Vr*CP.18P. It is concluded that the selection sample satisfies the multivariate and univariate normality assumptions, and the validity of the results obtained with the test statistics is confirmed by inspection of the multivariate and univariate normality plots.

- Assessment of the contribution of each predictor.

The test for equality of group means measures the potential that each independent variable has before becoming part of the model. The table 18 shows the results of the one-way ANOVA and at a significance value of less than 5% the predictor variables discriminate between group means.

The Wilks' Lambda statistic is another criterion that measures the discriminant potential of the variables. Small values indicate that the selected variable is better at discriminating between groups. The table 18 suggests that the mean return measured 18 periods prior to the regime shift x.V*18P is a better discriminant than the slope of the variances measured 18 moments prior to the trend change Sp.Vr*CP.18P.

- Assessment of the collinearity of the predictors.

The matrix of correlations within groups, from the table 19, shows the correlation between predictors x.V*18P and Sp.Vr*CP.18P. The value obtained suggests a clear independence between predictors. The correlation coefficient is not large enough for instability problems to occur in the signs of the coefficients of the model variables. Consequently, the proposed methodology through the selection of predictors helps to overcome the multicollinearity problem.

In the table 20, the structure matrix is defined by the correlations within groups combined between the predictors and the standardized canonical discriminant function. The order of the variables is a function of the absolute value of this correlation and the coefficients of the ranking functions. Moreover, this order is identical to that shown in the test for equality of group means in the 18 table. This concordance confirms the absence of multicollinearity between the selected independent variables.

The standardized coefficients facilitate the comparison of the independent variables measured on different scales. The coefficient with the highest absolute value corresponds to the predictor variable with the highest discriminant capacity. Thus, the order of the predictors, according to their standardized coefficients, shows the importance of the average return x.V*18P in predicting the future direction of the euro-dollar exchange rate.

Consequently, as the discriminant function is not affected by multicollinearity, it is safe to mention that the future direction of the price movement is mainly determined by the average return x.V*18P of the euro-dollar exchange rate and, therefore, the average return x.V*18P discriminates better between the directions of the upward and downward trend movements.

- Assessment of homogeneity of covariance matrices.

The table 21 presents the values of the ranks and logarithms of the determinants of the covariance matrices of the groups. The logarithms of the determinants are a measure of the variability of the groups. Small differences in the logarithms of the determinants indicate groups with equal covariance matrices. The *Box's M* statistic of the table 22 evaluates the assumption of equality of covariance matrices between groups.

According to results from the table 22, the P-Value of the *Box's M* test statistic is greater than 0.05, consequently, the subsets of observations of the groups in the predictors x.V*18P and Sp.Vr*CP.18P come from populations with equal variance-covariance matrices and, therefore, it can be implicitly inferred that the formulation of the classification functions do not require separate group covariance matrices.

3) ASSESSMENT OF MODEL FIT

Theoretically, the size of the groups in the analysis sample determines the a priori probability of belonging to each group. The definition of the coefficients of the classification functions and the performance of the classification process are determined by the prior probabilities used in the formulation of the models. According to the table 13, the study sample (Z) is formed by groups of equal size. Consequently, so that the calculation of the coefficients of the functions is not affected, equal prior probabilities are used for both groups ($\pi_g = 0.50$).

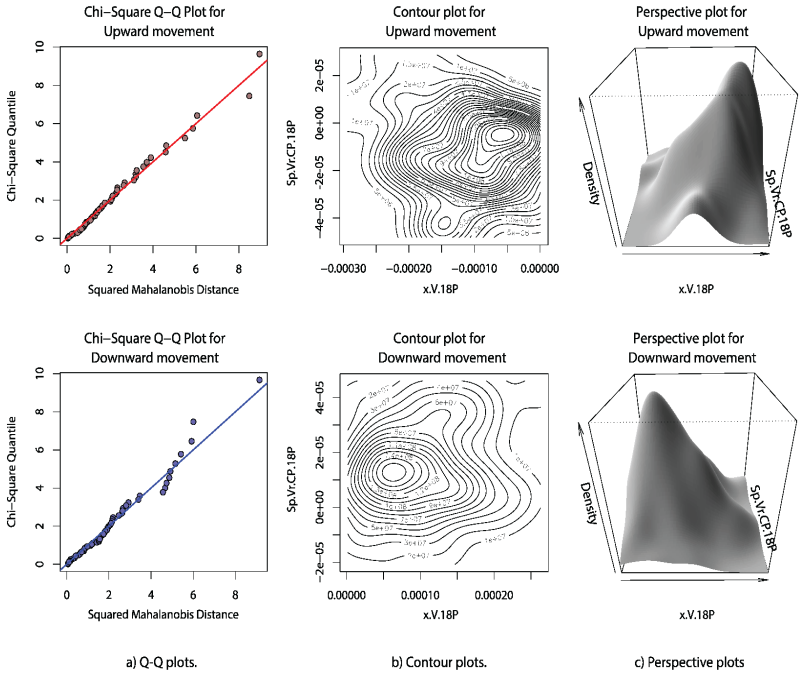


FIGURE 12. Multivariate charts of groups. The group structures of the data matrix $\mathbf{Z}(125 \times 2)$, upward and downward trending movements, approximate a multivariate normal distribution.(a) Normal Q-Q Plot for groups. (b) Contour plots for groups. (c) Perspective plots for groups.

TABLE 17. Univariate normality tests for groups.

Test	Measure	Trend	Cases	AD-Statistic	P-Value
Anderson-Darling	x.V*18P	Upward movement	62	0.343	0.480
		Downward movement	63	0.402	0.348
Anderson-Darling	Sp.Vr*CP.18P	Upward movement	62	0.212	0.851
		Downward movement	63	0.462	0.251

TABLE 18. Tests of equality of group means.

ID	Measure	Wilks' Lambda	F-Value	Df_1	Df_2	P-Value
1	x.V*18P	0.293	205.529	1	85	0.000
2	Sp.Vr*CP.18P	0.638	48.238	1	85	0.000

TABLE 19. Within-groups correlation matrix.

Measure	x.V*18P	Sp.Vr*CP.18P
x.V*18P	1	-0.026
Sp.Vr*CP.18P	-0.026	1

The table 23 provides information related to the efficacy measure of the linear discriminant function. For the case in question, since there are two groups to be discriminated, the canonical correlation is used as a measure of the function's effectiveness. This measure is similar to Pearson's correlation between the actual values of the observations in each group and the predicted discriminant scores.

As a result, with an appreciable eigenvalue of **3.048** the predictor variables involved in the model formulation explain 100% of the total cumulative variance, and the **87**% canonical

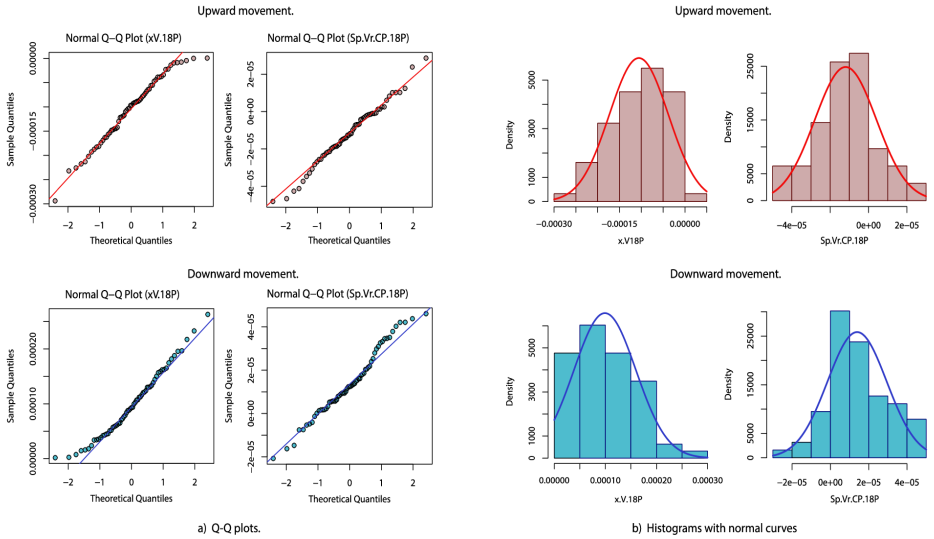


FIGURE 13. Univariate charts of groups. The group structures of the data matrix $\mathbf{Z}(125 \times 2)$, upward and downward trending movements, approximate a normal distribution.(a) Normal Q-Q Plot for groups. (b) Histograms with normal curves.

TABLE 20. Structure matrix and standardized function coefficients.

ID	Measure	Discriminant function	Standardized function coefficients
1	x.V*18P	0.891	0.902
2	Sp.Vr*CP.18P	0.432	0.455

TABLE 21. Measure of variability of the groups.

Trend	Matrix rank	Log Determinant
Upward movement	2	-41.261
Downward movement	2	-41.395
Pooled within-groups	2	-41.283

TABLE 22. Box's M-Test of equality of covariances across groups.

Box's M	F-Value	Df_1	Df_2	P-Value
3.092	1.004	3	1122087.83	0.390

correlation between the observations in the groups and the predicted discriminant scores gives a glimpse of a model that fits the behavior of the data and provides, in terms of efficiency, a high correlation between the observed values and the predicted data.

In the table 24 the Wilks' Lambda statistic is used as a measure that checks how well the linear discriminant function separates cases into groups. Small values, close to zero, of the Wilks' Lambda statistic are indicators of the discriminant ability of the function. Chi-square is the value of the associated statistic used to test the H_0 hypothesis. So with a P-value of less than 5%, the hypothesis H_0 that the means of Fisher linear functions are equal between groups is rejected and, therefore, the linear discriminant model contributes to the differentiation of cases between group means.

4) CLASSIFYING THE DIRECTION OF EXCHANGE RATE MOVEMENT

The Fisher linear functions help to assign the cases to a particular membership group. For each case a classification score is calculated and the discriminant analysis model assigns the case to the group whose classification function obtained the highest discriminant score. The linear Fisher functions that best discriminate between upward and downward trending

TABLE 23. Efficiency of the linear discriminant function.

Eigenvalue	% Variance	% Cumulative Variance	Canonical correlation
3.048	100	100	0.868

TABLE 24. Test of functions.

Wilks' Lambda	Chi-Square	Df_1	P-Value
0.247	117.443	2	0.000

movements are summarized in equation (19).

$$y(\hat{\mathbf{X}}) = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 \hat{X}_2, \tag{19}$$

where, $y(\hat{X})$ is the discriminant score, $\hat{X}_1$ is the average return (x.V*18P), $\hat{X}_2$ is the slope of the regression line of the variations between closing prices (Sp.Vr*CP.18P), both criteria calculated over the last 18 periods. The coefficients of the ranking functions β_0, β_1 y β_2 are referenced in the table 25.

Now, starting from the main function (19), the discriminant analysis model works with two linear functions denoted as $y(\hat{X}_U)$ and $y(\hat{X}_D)$ to discriminate between upward trending movements C_U and downward trending movements C_D. The rule that assigns the cases to the membership group is:

$$\text{If } y(\hat{X}_U) \geq y(\hat{X}_D) \text{ then } y(\hat{X}_U) \text{ belongs in } C_U$$
$$\text{Else if } y(\hat{X}_U) < y(\hat{X}_D) \text{ then } y(\hat{X}_D) \text{ belongs in } C_D$$

The structure of the discriminant scores is affected by the signs and magnitude of the coefficients of the variables participating in the model. The upward movement ranking function, according to the table 25, is characterized by the fact that the coefficients of the predictors are less than zero. This means, according to the contribution that each variable makes in the model, that the smaller the average return x.V*18P with respect to the slope of these variations Sp.Vr*CP.18P, the less likely it is that the price of the euro against the US dollar will continue to depreciate. Consequently, after the fall of the exchange rate slows down and enters a phase of exhaustion and correction, at this level, the negative behavior of the price action and occurring before the change of market regime, is the value used by the discriminant function to predict the beginning of an upward movement.

Also in the table 25, in the bearish movement classification function, the change of sign in the coefficients indicates that the cases with positive mean returns x.V*18P greater than the slopes of these variations Sp.Vr*CP.18P, are more likely to have reached the end of an appreciation process, thus announcing the beginning of a bearish market regime.

In contrast to what was said, the box plot in Figure 14 shows the distribution of predictor measures by groups. The expected variances of the data are symmetrically arranged in two different regions bounded by a central zero level. Negative mean returns x.V*18P indicate that the exchange rate has depreciated and remains in the oversold zone. Conversely, when the mean returns x.V*18P are positive, the exchange rate is appreciating and is in the overbought zone. Theoretically, the Q_1, Q_2 and Q_3 quartiles of the box plots define the levels of support and resistance at which the direction of the trend has historically changed and implicitly indicate the number of moves generated at each level. The results show that for each group there is explicit agreement between the predictor measures. Note the 75% similarity between the signs of the values of Sp.Vr*CP.18P and x.V*18P.

The findings obtained seem to suggest that a change of direction in price movement is more likely when the following signals occur. An upward movement is more likely to occur when the negative mean return x.V*18P is greater than the mean of the negative mean returns x.V*18P and the slope of the variances Sp.Vr*CP.18P is negative. Otherwise, a downward movement is more likely to occur when the positive average return x.V*18P is less than the mean of the positive average returns x.V*18P and the slope of the variances Sp.Vr*CP.18P is positive.

In summary, because the mean return has the largest contribution in the discriminant function, a sustained decrease or increase in the mean return of the exchange rate above or below its historical average value makes it more likely that each case, according to the highest discriminant score, will be classified as the beginning of an upward or downward trend movement respectively.

5) MODEL VALIDATION

Table 26 presents the results of applying the model. The initial analysis achieved an impressive classification accuracy of **98.9**%. Out of 38 bearish moves, 37 were correctly classified, and all 49 bullish moves were accurately identified. These cases train and cross-validate the discriminant function, showcasing the model's effectiveness and ability to generalize during cross-validation. In the subsequent phase, a perfect classification accuracy of **100**% was attained by employing observations excluded during the model formulation. This outcome indicates that the model is efficient and capable of accurately detecting the differences between the onset of a bullish move and a bearish one. Importantly, it should be highlighted that the class imbalance within the data subsets does not have a detrimental impact on classification accuracy. The exceptional performance can be attributed to the representative nature of the data sample,

TABLE 25. Coefficients of the classification functions.

Measure	Symbol	Predicted trend	
		Upward movement $y(\hat{X}_U)$	Downward movement $y(\hat{X}_D)$
x.V*18P	β_1	-24598.42	22755.91
Sp.Vr*CP.18P	β_2	-44485.91	52112.42
Constant	β_0	-2.26	-2.16

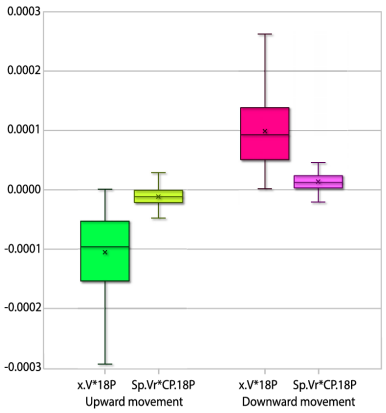


FIGURE 14. Box plots of predictors and groups. Values of x.V*18P and Sp.Vr*CP.18P at the inflection points ET before the trend shift, from the data matrix Z(125 × 2). (a) Upward movement. (b) Downward movement.

the discriminative power of the chosen predictor variables, and the suitableness of the employed discriminant analysis methodology.

The obtained accuracy values support the model's usefulness and reliability in forecasting tasks, making it a valuable tool for market analysts and investors, empowering them to make well-informed decisions.

D. FORECAST PERFORMANCE

This section presents the evaluation results of the short-term predictive power of the model for predicting the direction of movement of the euro-dollar exchange rate. Additional experimental results are presented to assess the effectiveness of the proposed methodology. These results are based on out-of-sample data encompassing diverse market conditions and spanning various time horizons. In each instance, the model predicts the initiation of a new directional movement from an inflection point **ET** when the change in the direction of price movements has historically occurred.

Table 27 presents the classification results using out-of-sample data from Fisher's linear functions. The data are divided into two sections: The first corresponds to the training and cross-validation data, and the second to the test data: OOS_1, OOS_2, OOS_3, and OOS_4. Overall, the method ranks the bullish and bearish data exceptionally well, with values from 98.10% to 99.0% and 98.8% to 99.0%, respectively. The results suggest Fisher's linear functions fit the data structure and accurately discriminate between bullish and bearish trends. This finding demonstrates that discriminant analysis is an effective tool for classifying trend data. Furthermore, the relevance of these results and their relation to the objective of the proposed methodology demonstrate effectiveness in predicting short-term trends.

Additional experiments conducted with the OOS_1 and OOS_4 ensembles show classification performance consistent with earlier results. The model correctly classifies bullish and bearish movements with values ranging from 98.7% to 99.0%, and from 98.8% to 99.9%, respectively. It is worth noting that the class imbalance observed in OOS_3 does not have a detrimental impact on classification accuracy. This exceptional performance can be attributed to the representative nature of the training sample, the discriminative quality of the selected predictor variables, and the suitableness of the discriminant analysis methodology.

The experimental results demonstrate the model's goodness of fit and its generalizability and accuracy under different market conditions and time horizons. However, specific requirements must be met to maintain the consistency of the obtained results: (i) Ensure the quality of the results during feature preparation, selection, and validation through structural analysis. (ii) Avoid using inflection points when they are influenced by the release of fundamental data, as prediction accuracy may be affected by the resulting volatility. (iii) Use discriminant and independent characteristics that provide significant differences between trends to ensure accurate classification. (iv) Ensure compliance with the assumptions of multivariate normality in the training and validation samples when opting for discriminant analysis as a classification method. Failure to meet these conditions may compromise the reliability of interpretations, rendering the results merely descriptive rather than inferable. Additionally, it is crucial to perform statistical significance tests at each stage of the process to ensure the reliability and consistency of the obtained results. This strategy ensures the successful application of the proposed approach to other financial instruments.

- Predictive capability of the classification model

Table 28 provides an overview of the predictive power of the classification model obtained through the proposed methodology. Three sections with different performance measures have been included: the first details the performance measures for the training and validation data set; the second, the test data sets; and the third, the descriptive statistics of the mean performance values. The classification model undergoes training and validation on a single training dataset,

TABLE 26. Classification results.

Sample analysis cases Subset	125	Trend	Predicted group membership Upward movement	Downward movement	Total
Training set		Upward Movement (n)	49	0	49
		Downward Movement (n)	1	37	38
		Upward Movement (%)	100%	0%	100%
		Downward Movement (%)	2.60%	97.40%	100%
Cross-Validation		Upward Movement (n)	49	0	49
Training set		Downward Movement (n)	1	37	38
	87	Upward Movement (%)	100%	0%	100%
	70%	Downward Movement (%)	2.60%	97.40%	100%
Testing set		Upward Movement (n)	13	0	13
		Downward Movement(n)	0	25	25
	38	Upward Movement (%)	100%	0%	100%
	30%	Downward Movement (%)	0%	100%	100%

TABLE 27. Out-of-sample classification results.

Sample analysis cases Subset		Trend	Predicted group membership Upward movement	Downward movement	Total
Training and Cross-Validation Set	125	Upward Movement (n)	62	0	62
		Downward Movement (n)	1	62	63
		Upward Movement (%)	100%	0%	100%
		Downward Movement (%)	1.60%	98.40%	100%
Testing sets:					
OOS_1: 1999-2005	6581	Upward Movement (n)	3306	35	3341
		Downward Movement(n)	40	3200	3240
		Upward Movement (%)	99.00%	1.00%	100%
		Downward Movement (%)	1.20%	98.80%	100%
OOS_2: 2006-2020	14645	Upward Movement (n)	7227	91	7318
		Downward Movement(n)	84	7243	7327
		Upward Movement (%)	98.80%	1.20%	100%
		Downward Movement (%)	1.10%	98.90%	100%
OOS_3: 2006-2020	1748	Upward Movement (n)	569	11	580
		Downward Movement(n)	14	1154	1168
		Upward Movement (%)	98.10%	1.90%	100%
		Downward Movement (%)	1.20%	98.80%	100%
OOS_4: 2021-2023	2336	Upward Movement (n)	1172	16	1188
		Downward Movement(n)	11	1137	1148
		Upward Movement (%)	98.70%	1.30%	100%
		Downward Movement (%)	1.00%	99.00%	100%

and its performance is evaluated on four distinct test datasets: OOS_1, OOS_2, OOS_3, and OOS_4. Several performance metrics are measured on each dataset, including accuracy, sensitivity, specificity, AUC, precision, F1-score, baseline, Kappa coefficient, and Matthews correlation coefficient.

The results indicate that the model has an exceptionally high predictive ability on all test data sets, with an accuracy of over 98.57%. Additionally, the high values of the area under the curve (AUC) indicate a strong discriminatory ability to differentiate between bullish and bearish movements. The model demonstrates impressive sensitivity and specificity, exceeding 97.52% and 98.51%, respectively, across all test datasets. These figures indicate the model's capability to accurately detect the onset of bullish and bearish movements. Furthermore, the model achieves remarkable accuracy and F1-score values, surpassing 98.03% and 97.85%, respectively. These results highlight the model's exceptional ability to predict bullish movements and maintain a strong balance between precision and sensitivity, confirming its effectiveness in identifying both types of trends. Moreover, the Kappa and Matthews correlation coefficients (MCC) exhibit high values exceeding 96.75% across all test datasets. Therefore, the model possesses an exceptional ability to predict the correct class, further reinforcing its overall performance and reliability.

The outstanding performance achieved in this study confirms that the training sample effectively picks up the patterns that best differentiate the trends through the predictor variables. Consequently, the model has a high generalization capacity and produces very accurate predictions without compromising the readability and interpretability of the predictions.

TABLE 28. Model's predictive power.

Data	n	Accuracy	Recall	Specificity	AUC	Precision	F1	Baseline	Kappa	MCC
Training set	125	0.992	0.984	1.000	0.992	1.000	0.992	0.500	0.984	0.984
Cross-validation	125	0.992	0.984	1.000	0.992	1.000	0.992	0.500	0.984	0.984
Testing sets:										
OOS_1: 1999-2005	6581	**0.989**	0.988	0.989	**0.989**	**0.990**	**0.989**	0.500	**0.977**	**0.977**
OOS_2: 2006-2020	14645	0.988	0.989	0.988	0.988	0.988	0.988	0.500	0.976	0.976
OOS_3: 2006-2020	1748	0.986	0.976	**0.991**	0.983	0.981	0.979	**0.556**	0.968	0.968
OOS_4: 2021-2023	2336	0.988	**0.991**	0.986	0.988	0.987	**0.989**	0.500	**0.977**	**0.977**
Descriptive statistics:										
Mean performance		0.9877	0.9860	0.9885	0.9870	0.9865	0.9862	0.5140	0.9745	0.9745
95% confidence	UL	0.9897	0.9967	0.9918	0.9913	0.9926	0.9939	0.5585	0.9814	0.9814
interval for the mean	LL	0.9857	0.9752	0.9851	0.9826	0.9803	0.9785	0.4594	0.9675	0.9675
Standard error		0.0006	0.0033	0.0010	0.0013	0.0019	0.0024	0.0140	0.0021	0.0021
Standard deviation		0.0012	0.0067	0.0020	0.0027	0.0038	0.0048	0.0280	0.0043	0.0043

Finally, the proposed methodology outperforms the reference value (Baseline) across all performance evaluation metrics. Even if the best performance of the state of the art is considered as the benchmark (0.8955, see Table 29), the proposed methodology is superior in all performance evaluation metrics. Moreover, the 95% confidence limits for all performance metrics are closely aligned with their mean values, indicating high accuracy in the results. Additionally, both the standard error and standard deviation are minimal, suggesting low variability in model performance across different datasets. This consistency implies a high level of generalization, accuracy, and reliability of the model over various time horizons and market conditions.

- Comparison with the state of the art

Accurate and interpretable classification models are essential for predicting market behavior in the investment field. Evaluating these characteristics is crucial to determine the effectiveness of the chosen approach. In this regard, Table 29 comprehensively compares the proposed methodology with state-of-the-art techniques, summarizing the key characteristics of the evaluated methodologies. These methodologies include ensemble approaches [91], hybrid approaches [57], neural networks [92], [93], and trend prediction based on sentiment analysis [58]. These approaches are chosen not only for their effectiveness but also for their ability to predict the direction of the euro-dollar exchange rate.

The methodologies evaluated in the study rely on market data (MI) and technical indicators (TI) as inputs. However, these models have limitations due to the restricted nature of the study samples, which only consider specific time horizons. Consequently, their generalizability in the long term may be compromised. Moreover, feature extraction (FE) is widespread in these approaches, which enhances accuracy but often hinders interpretability. While the assessed approaches have demonstrated predictive power using both training and out-of-sample (OOS) data, particular studies, such as [91] and [93], were solely evaluated using in-sample data (TVT). This reliance on in-sample data may result in an overestimation of their predictive ability and a lack of generalizability to new data.

The definition of the hyperparameters in the referred methods is one problem that requires careful assignment to guarantee the reported accuracy. The incorrect assignment of hyperparameters can lead to overfitting issues and limit the model's generalization capacity to specific market conditions and time horizons. In contrast, the proposed methodology, which uses discriminant analysis, avoids the instability of hyperparameters by employing discriminant and independent predictor variables [23].

Table 29 presents the results of the accuracy (Acc) comparison, highlighting the exceptional performance of the proposed methodology (FE and FS) with OOS data, achieving accuracies of 0.975 and 0.987, respectively. The statistical analysis using the HSD Tukey test, with a significance level of less than 1%, confirms the superiority of the proposed methodology in terms of accuracy. Compared to the state-of-the-art approaches, which use deep neural networks [92] with a 0.8955 accuracy, the proposed methodology (FS) demonstrates remarkable improvement in classification accuracy while maintaining the interpretability of the results. This improvement is attributed to various factors, including the incorporation of inflection points, the introduction of new attributes, the discriminant and independent feature selection process, and the multivariate validation prior to constructing a linear discriminant function.

The proposed methodology leverages a comprehensive and representative data set spanning from 1999 to 2023, surpassing the scope of other approaches. This substantial data coverage enhances robustness and generalization when constructing classification models evaluated under different market conditions. Consequently, the resulting classification model is more accurate and reliable, instilling greater confidence in the results' validity. Using a larger data set reduces the risk of the results being influenced by specific sample artifacts, increasing the reliability of the proposed methodology.

Overall, classification models based on artificial intelligence and machine learning techniques significantly impact classification accuracy, requiring sufficient computational resources for training, validation, and deployment phases. However, the simplicity of the formulated model makes it

TABLE 29. Accuracy assessment of prediction approaches.

Methodology	Instrument	Data	FPM	Sample	Prediction	PP	Acc
Sadeghi, 2021 [91]	EUR/USD	MI, TI	FE	2014 - 2019	Daily, Up / Down / Sideway	TVT	0.808
Hybrid model [57]	EUR/USD	MI, TI	FS, FE	2010 - 2015	Daily, Up / Down	OOS	0.8865
Market sentiment ANN [58]	USD/EUR	MI, SN	FS, FSp	2013	Daily, Up / Down	OOS	0.6346
Deep networks CNN [92]	EUR/USD	MI	FEng	2010 - 2015	Daily, Up / Down 1-30 Periods	OOS	0.8955
ANN and DTW [93]	EUR/USD	MI	FS, FE	2011 - 2013	Daily, Up / Down	TVT	0.72
Proposed methodology – LCC (FE)	EUR/USD	MI	FE (15)	**1999 - 2023**	Intraday, Up / Down	OOS	**0.975** ‡
Proposed methodology – LCC (FS)	EUR/USD	MI	FS (2)	**1999 - 2023**	Intraday, Up / Down	OOS	**0.987** ‡

MI = Market Information (OHLCV). **TI** = Technical Indicators. **SN** = Social Networks (StockTwits posts). **FPM** = Feature Pre-processing Method. FE = Feature Extraction, FS = Feature Selection, FSp = Feature Space, FEng = Feature Engineering. **PP** = Model's ability to predict new data outside the training set. TVT = Training, Validation and Testing, OOS = Out-of-Sample Performance. **Acc** = Accuracy. FE(15) = Feature extraction from the first 15 variables in Table 2. FS(2) = Feature selection using discriminative and independent predictors x.V*PT and Sp.Vr*CP.PT measuring price action (See Table 2).
‡ HSD Tukey tests, at less than 1% significance level, reveal a significant difference in accuracy in favor of the control group.

easily interpretable (see Linear discriminant model), stable (see Assessment of the collinearity of the predictors), generalizable (see Out-of-sample results), scalable (see Extra-out-of-sample results) and accurate (see Model's predictive power).

VI. CONCLUSION

This paper demonstrates the effectiveness of the proposed methodology for the formulation of a simple, accurate, parsimonious and easy-to-explain prediction model [69]. Specifically, the here-introduced Linear Classifier Configurator (LCC) integrates data preparation for discriminant analysis, feature selection, and the formulation of a classification model that predicts, in the short term, the onset of the direction of the future movement of the euro-dollar exchange rate. This ingenious approach begins with the exploratory statistical analysis of candidate variables. The effectiveness of the selected variables is validated with the results of the ranking process. The prediction model is formulated from a representative sample following a multivariate normal distribution. Finally, the model is evaluated using out-of-sample data. On average, the proposed methodology provides an exceptionally satisfactory out-of-sample classification accuracy of 98.77%.

The suitability of the LCC methodology makes it particularly useful for identifying the best determinants of the direction of the euro-dollar exchange rate movement. The selected independent predictor variables, according to statistical merit, are the ones that best explain the price action before the change of market regime (object of prediction), have the highest discriminant capacity, the highest predictive power and are the variables that best contribute to the formulation of the classification model. A significant advantage of the proposed method is its ability to deal with and overcome the asymmetric and leptokurtic arrangement of the studied data. Thus, the preparation of data and the extraction of a representative study sample guarantees the requirements of multivariate normality between predictors and groups, overcomes the influence of imbalance in the groups and ensures the fulfillment of statistical assumptions that validate the interpretations made on the results obtained with the discriminant analysis.

The proposed methodology is applicable, from the point of view of technical analysis, to the study of the price behavior of financial instruments, especially those traded on the basis of a reliable and consistent trading strategy over time. In this context, the proposed method is appropriate to identify the determinants of the price movement of the traded asset. The quality of the selected variables is validated with the performance of the constructed classification model.

Consequently, the obtained results allow to conclude, that on the basis of a reliable and consistent trading strategy in time, it is predictable in the short term (a period in the future in 15-minute trading session) the future direction of the euro-dollar exchange rate movement (as a consequence of the change of the market regime). In these cases the results of the study can help to improve the financial performance of trading strategies. In addition, the prediction model can act as a decision support tool. Thus, the proposed model can confirm a buy or sell order prior to its issuance.

The obtained findings suggest that the change of direction is mainly influenced by the average return x.V*18P and the slope of the variations between closing prices Sp.Vr*CP.18P. Consequently, an upward or downward movement is more likely to occur when the average negative or positive return of the exchange rate decreases or increases above or below its historical average value and the slope of the inter-closing price changes is negative or positive.

The implementation of the predictive model in the algorithmic trading system is a future work in progress. The study focuses on systematically analyzing and evaluating, in real time, the performance of the trading system. The evaluation will comprehensively improve the effectiveness of the trading system and the forecasting model in the face of euro-dollar exchange rate volatility. In addition to the exploration of other supervised learning methods, future work planned by the authors also focuses on benchmarking the performance provided by other classification techniques.

An important problem to be addressed in future studies is the prediction of the duration of trending price movements. Although, the prediction model achieves, out of the analysis sample, **98.77**% classification accuracy, a problem that still needs to be investigated, from a technical analysis point of

view, is the causality or covariation relationships between the identified metrics and trend price movement duration times.

ACKNOWLEDGMENT

The authors would like to thank the proofreading and valuable feedback by Juan Sebastián Mejía-Ordóñez.

REFERENCES

[1] E. Ilzetzki, C. M. Reinhart, and K. S. Rogoff, "Why is the euro punching below its weight?" *Econ. Policy*, vol. 35, no. 103, pp. 405–460, Apr. 2021.
[2] M. A. Ayadi, W. Ben Omrane, J. Wang, and R. Welch, "Senior official speech attributes and foreign exchange risk around business cycles," *Int. Rev. Financial Anal.*, vol. 80, Mar. 2022, Art. no. 102011.
[3] N. Eiamkanitchat, T. Moontuy, and S. Ramingwong, "Fundamental analysis and technical analysis integrated system for stock filtration," *Cluster Comput.*, vol. 20, no. 1, pp. 883–894, Mar. 2017.
[4] R. D. Edwards, J. Magee, and W. C. Bassetti, *Technical Analysis of Stock Trends*. Boca Raton, FL, USA: CRC Press, 1967.
[5] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *J. Finance*, vol. 25, no. 2, pp. 383–417, May 1970.
[6] K.-Y. Woo, C. Mai, M. McAleer, and W.-K. Wong, "Review on efficiency and anomalies in stock markets," *Economies*, vol. 8, no. 1, p. 20, Mar. 2020.
[7] J. M. Griffin, P. J. Kelly, and F. Nardari, "Do market efficiency measures yield correct inferences? A comparison of developed and emerging markets," *Rev. Financial Stud.*, vol. 23, no. 8, pp. 3225–3277, Aug. 2010.
[8] L. Menkhoff and M. P. Taylor, "The obstinate passion of foreign exchange professionals: Technical analysis," *J. Econ. Literature*, vol. 45, no. 4, pp. 936–972, Nov. 2007.
[9] A. W. Lo and A. C. MacKinlay, "Stock market prices do not follow random walks: Evidence from a simple specification test," *Rev. Financial Stud.*, vol. 1, no. 1, pp. 41–66, Jan. 1988.
[10] J. Conrad and G. Kaul, "Mean reversion in short-horizon expected returns," *Rev. Financial Stud.*, vol. 2, no. 2, pp. 225–240, Apr. 1989.
[11] D. Pradeepkumar and V. Ravi, "Soft computing hybrids for FOREX rate prediction: A comprehensive review," *Comput. Oper. Res.*, vol. 99, pp. 262–284, Nov. 2018.
[12] Z. Hajirahimi and M. Khashei, "Hybridization of hybrid structures for time series forecasting: A review," *Artif. Intell. Rev.*, vol. 56, pp. 1201–1261, May 2022.
[13] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong," *Inf. Process. Manage.*, vol. 57, no. 5, Sep. 2020, Art. no. 102212.
[14] G. Zhou, "Measuring investor sentiment," *Annu. Rev. Financial Econ.*, vol. 10, no. 1, pp. 239–259, Nov. 2018.
[15] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Exp. Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, Nov. 2014.
[16] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning : A systematic literature review: 2005–2019," *Appl. Soft Comput.*, vol. 90, May 2020, Art. no. 106181.
[17] Z. Hu, Y. Zhao, and M. Khushi, "A survey of forex and stock price prediction using deep learning," *Appl. Syst. Innov.*, vol. 4, no. 1, p. 9, Feb. 2021.
[18] S. Nosratabadi, A. Mosavi, P. Duan, P. Ghamisi, F. Filip, S. Band, U. Reuter, J. Gama, and A. Gandomi, "Data science in economics: Comprehensive review of advanced machine learning and deep learning methods," *Mathematics*, vol. 8, no. 10, p. 1799, Oct. 2020.
[19] A. Thakkar and K. Chaudhari, "A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions," *Exp. Syst. Appl.*, vol. 177, Sep. 2021, Art. no. 114800.
[20] M.-L. Shen, C.-F. Lee, H.-H. Liu, P.-Y. Chang, and C.-H. Yang, "An effective hybrid approach for forecasting currency exchange rates," *Sustainability*, vol. 13, no. 5, p. 2761, Mar. 2021.
[21] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[22] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.
[23] J. Iworiso and S. Vrontos, "On the directional predictability of equity premium using machine learning techniques," *J. Forecasting*, vol. 39, no. 3, pp. 449–469, Apr. 2020.
[24] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis," *J. Behav. Experim. Finance*, vol. 32, Dec. 2021, Art. no. 100577.
[25] G. P. Zhang, "Neural networks for classification: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 30, no. 4, pp. 451–462, Nov. 2000.
[26] L. A. Berrueta, R. M. Alonso-Salces, and K. Héberger, "Supervised pattern recognition in food analysis," *J. Chromatography A*, vol. 1158, nos. 1–2, pp. 196–214, Jul. 2007.
[27] S. Bose, A. Pal, R. SahaRay, and J. Nayak, "Generalized quadratic discriminant analysis," *Pattern Recognit.*, vol. 48, no. 8, pp. 2676–2684, Aug. 2015.
[28] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Commun.*, vol. 26, no. 4, pp. 283–297, Dec. 1998.
[29] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
[30] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.
[31] C. Bouveyron, S. Girard, and C. Schmid, "High dimensional discriminant analysis," *Commun. Statistics-Theory Methods*, vol. 36, no. 14, pp. 2607–2623, 2007.
[32] K. Grobys and N. Sapkota, "Predicting cryptocurrency defaults," *Appl. Econ.*, vol. 52, no. 46, pp. 5060–5076, Oct. 2020.
[33] G. N. F. Weiß, "Copula-GARCH versus dynamic conditional correlation: An empirical study on VaR and ES forecasting accuracy," *Rev. Quant. Finance Accounting*, vol. 41, no. 2, pp. 179–202, Aug. 2013.
[34] S. W. Kwag and Y. S. Kim, "Stock price predictability of financial ratios and macroeconomic variables: A regulatory perspective," *Ind. Eng. Manage. Syst.*, vol. 12, no. 4, pp. 406–415, Dec. 2013.
[35] I. Gurrib and F. Kamalov, "Predicting Bitcoin price movements using sentiment analysis: A machine learning approach," *Stud. Econ. Finance*, vol. 39, no. 3, pp. 347–364, Apr. 2022.
[36] M. Bernardi, L. Catania, and L. Petrella, "Are news important to predict the value-at-risk?" *Eur. J. Finance*, vol. 23, no. 6, pp. 535–572, May 2017.
[37] S. T. Mndawe, B. S. Paul, and W. Doorsamy, "Development of a stock price prediction framework for intelligent media and technical analysis," *Appl. Sci.*, vol. 12, no. 2, p. 719, Jan. 2022.
[38] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," *J. Comput. Appl. Math.*, vol. 365, Feb. 2020, Art. no. 112395.
[39] M. T. Leung, H. Daouk, and A.-S. Chen, "Forecasting stock indices: A comparison of classification and level estimation models," *Int. J. Forecasting*, vol. 16, no. 2, pp. 173–190, Apr. 2000.
[40] J. Pekár and M. Pcolár, "Empirical distribution of daily stock returns of selected developing and emerging markets with application to financial risk management," *Central Eur. J. Operations Res.*, vol. 30, no. 2, pp. 699–731, Jun. 2022.
[41] R. A. Eisenbeis, "Pitfalls in the application of discriminant analysis in business, finance, and economics," *J. Finance*, vol. 32, no. 3, pp. 875–900, Jun. 1977.
[42] S. Patra, "Revisiting value-at-risk and expected shortfall in oil markets under structural breaks: The role of fat-tailed distributions," *Energy Econ.*, vol. 101, Sep. 2021, Art. no. 105452.
[43] J. B. Horpestad, Š. Lyócsa, P. Molnár, and T. B. Olsen, "Asymmetric volatility in equity markets around the world," *North Amer. J. Econ. Finance*, vol. 48, pp. 540–554, Apr. 2019.
[44] C. G. Corlu and A. Corlu, "Modelling exchange rate returns: Which flexible distribution to use?" *Quant. Finance*, vol. 15, no. 11, pp. 1851–1864, Nov. 2015.
[45] H.-C. Liu and J.-C. Hung, "Forecasting S&P-100 stock index volatility: The role of volatility asymmetry and distributional assumption in GARCH models," *Exp. Syst. Appl.*, vol. 37, no. 7, pp. 4928–4934, Jul. 2010.
[46] P. Choi and K. Nam, "Asymmetric and leptokurtic distribution for heteroscedastic asset returns: The SU-normal distribution," *J. Empirical Finance*, vol. 15, no. 1, pp. 41–63, Jan. 2008.

[47] J.-C. Hung, M.-C. Lee, and H.-C. Liu, "Estimation of value-at-risk for energy commodities via fat-tailed GARCH models," *Energy Econ.*, vol. 30, no. 3, pp. 1173–1191, May 2008.

[48] D. N. Politis, "A heavy-tailed distribution for arch residuals with application to volatility prediction," Dept. Econ., Univ. California San Diego, San Diego, CA, USA, Tech. Rep., 2004.

[49] P. Theodossiou, "Financial data and the skewed generalized T distribution," *Manage. Sci.*, vol. 44, no. 12, pp. 1650–1661, Dec. 1998.

[50] T. Bollerslev, "A conditionally heteroskedastic time series model for speculative prices and rates of return," *Rev. Econ. Statist.*, vol. 69, no. 3, pp. 542–547, 1987.

[51] R. P. Buckley, D. A. Zetzsche, D. W. Arner, and B. W. Tang, "Regulating artificial intelligence in finance: Putting the human in the loop," *Sydney Law Rev.*, vol. 43, no. 1, pp. 43–81, 2021.

[52] E. Y. Matsumoto, E. Del-Moral-Hernandez, C. E. Yoshinaga, and A. de Campos Pinto, "Forecasting US dollar exchange rate movement with computational models and human behavior," *Exp. Syst. Appl.*, vol. 194, May 2022, Art. no. 116521.

[53] N. Gabdrakhmanova, V. Fedin, B. Matsuta, and M. Pilgun, "The modeling of forecasting new situations in the dynamics of the economic system on the example of several financial indicators," *Proc. Comput. Sci.*, vol. 186, pp. 512–520, 2021.

[54] D. C. Yıldırım, I. H. Toroslu, and U. Fiore, "Forecasting directional movement of forex data using LSTM with technical and macroeconomic indicators," *Financial Innov.*, vol. 7, no. 1, pp. 1–36, Dec. 2021.

[55] S. A. S. Alzaeemi and S. Sathasivam, "Examining the forecasting movement of palm oil price using RBFNN-2SATRA metaheuristic algorithms for logic mining," *IEEE Access*, vol. 9, pp. 22542–22557, 2021.

[56] C. Lu, "Analysis of early warning of RMB exchange rate fluctuation and value at risk measurement based on deep learning," *Comput. Econ.*, vol. 59, pp. 1501–1524, Aug. 2021.

[57] P. P. Das, R. Bisoi, and P. K. Dash, "Data decomposition based fast reduced kernel extreme learning machine for currency exchange rate forecasting and trend analysis," *Exp. Syst. Appl.*, vol. 96, pp. 427–449, Apr. 2018.

[58] V. Plakandaras, T. Papadimitriou, P. Gogas, and K. Diamantaras, "Market sentiment and exchange rate directional forecasting," *Algorithmic Finance*, vol. 4, nos. 1–2, pp. 69–79, Apr. 2015.

[59] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[60] W. Ge, P. Lalbakhsh, L. Isai, A. Lenskiy, and H. Suominen, "Neural network–based financial volatility forecasting: A systematic review," *ACM Comput. Surveys*, vol. 55, no. 1, pp. 1–30, Jan. 2022.

[61] S. Aziz, M. Dowling, H. Hammami, and A. Piepenbrink, "Machine learning in finance: A topic modeling approach," *Eur. Financial Manage.*, vol. 28, no. 3, pp. 744–770, Jun. 2022.

[62] Y. Qiu, Y. Qiu, Y. Yuan, Z. Chen, and R. Lee, "QF-TraderNet: Intraday trading via deep reinforcement with quantum price levels based profit-and-loss control," *Frontiers Artif. Intell.*, vol. 4, pp. 1–12, Oct. 2021.

[63] M. O. Özorhan, İ. H. Toroslu, and O. T. Şehitoğlu, "A strength-biased prediction model for forecasting exchange rates using support vector machines and genetic algorithms," *Soft Comput.*, vol. 21, no. 22, pp. 6653–6671, Nov. 2017.

[64] Y. Qiu, Z. Song, and Z. Chen, "Short-term stock trends prediction based on sentiment analysis and machine learning," *Soft Comput.*, vol. 26, pp. 2209–2224, Jan. 2022.

[65] S. Ahmed, M. M. Alshater, A. E. Ammari, and H. Hammami, "Artificial intelligence and machine learning in finance: A bibliometric review," *Res. Int. Bus. Finance*, vol. 61, Oct. 2022, Art. no. 101646.

[66] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Appl. Soft Comput.*, vol. 90, May 2020, Art. no. 106181.

[67] B. Žlicar and S. Cousins, "Discrete representation strategies for foreign exchange prediction," *J. Intell. Inf. Syst.*, vol. 50, no. 1, pp. 129–164, Feb. 2018.

[68] E. Steurer, M. Rothenhausler, and Y. Yeo, "Forecasting discretized daily USD/DEM exchange rate movements with quantitative models," in *Proc. 22nd Annu. Conf. German-Classification-Soc.*, 1999, pp. 467–472.

[69] T. Dao and H. Ahn, "An optimized combination of π-fuzzy logic and support vector machine for stock market prediction," *J. Intell. Inf. Syst.*, vol. 20, no. 4, pp. 43–58, Dec. 2014.

[70] M. G. Novak and D. Velušcek, "Prediction of stock price movement based on daily high prices," *Quant. Finance*, vol. 16, no. 5, pp. 793–826, May 2016.

[71] M. Sharma, S. Sharma, and G. Singh, "Performance analysis of statistical and supervised learning techniques in stock data mining," *Data*, vol. 3, no. 4, p. 54, Nov. 2018.

[72] J. Okicic, S. Remetic-Horvath, and B. Buyukdemir, "Stock selection based on discriminant analysis: Case of capital market of Bosnia and Herzegovina," *J. Econ. Social Stud.*, vol. 4, no. 2, pp. 5–30, 2014.

[73] C. Zopounidis, M. Doumpos, and S. Zanakis, "Stock evaluation using a preference disaggregation methodology," *Decis. Sci.*, vol. 30, no. 2, pp. 313–336, Mar. 1999.

[74] S.-N. Hwang, C.-T. Lin, and W.-C. Chuang, "Stock selection using data envelopment analysis-discriminant analysis," *J. Inf. Optim. Sci.*, vol. 28, no. 1, pp. 33–50, Jan. 2007.

[75] H. Ögüt, M. M. Doğanay, and R. Aktas, "Detecting stock-price manipulation in an emerging market: The case of Turkey," *Exp. Syst. Appl.*, vol. 36, no. 9, pp. 11944–11949, Nov. 2009.

[76] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Comput. Operations Res.*, vol. 32, no. 10, pp. 2513–2522, Oct. 2005.

[77] M. P. G. Villardón, "Una alternativa de representacion simultanea: HJ-BIPLOT," *Qüestiió, Quaderns d'estadística I Investigació Operativa*, vol. 10, pp. 13–23, Apr. 1986.

[78] K. R. Gabriel, "The biplot graphic display of matrices with application to principal component analysis," *Biometrika*, vol. 58, no. 3, pp. 453–467, 1971.

[79] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[80] Z. Bousbaa, O. Bencharef, and A. Nabaji, "Stock market speculation system development based on technico temporal indicators and data mining tools," in *Heuristics for Optimization and Learning*. Cham, Switzerland: Springer, 2021, pp. 239–251.

[81] *IBM SPSS Statistics for Windows, Version 26.0*, IBM Corp., Armonk, NY, USA, 2019.

[82] J. L. Vicente-Villardón, "MULTBIPLOT: A package for multivariate analysis using biplots," Mathlab Softw., Departamento de Estadística, Universidad de Salamanca, Salamanca, Spain, Tech. Rep., 2015.

[83] S. Korkmaz, D. Goksuluk, and G. Zararsiz, "MVN: An R package for assessing multivariate normality," *R J.*, vol. 6, no. 2, p. 151, 2014.

[84] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 4th Int. Conf. Mach. Learn.* Princeton, NJ, USA: Citeseer, vol. 97, 1997, pp. 179–186.

[85] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize F1 measure," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2014, pp. 225–239.

[86] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[87] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Dec. 2020.

[88] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman, *Using Multivariate Statistics*. Boston, MA, USA: Pearson Education, 2007.

[89] H. Finch, "Identification of variables associated with group separation in descriptive discriminant analysis: Comparison of methods for interpreting structure coefficients," *J. Experim. Educ.*, vol. 78, no. 1, pp. 26–52, Sep. 2009.

[90] T. Svantesson and J. W. Wallace, "Tests for assessing multivariate normality and the covariance structure of MIMO data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, Apr. 2003, p. 656.

[91] A. Sadeghi, A. Daneshvar, and M. M. Zaj, "Combined ensemble multiclass SVM and fuzzy NSGA-II for trend forecasting and trading in forex markets," *Exp. Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115566.

[92] S. Galeshchuk and S. Mukherjee, "Deep networks for predicting direction of change in foreign exchange rates," *Intell. Syst. Accounting, Finance Manage.*, vol. 24, no. 4, pp. 100–110, Oct. 2017.

[93] L. C. Tiong, D. C. Ngo, and Y. Lee, "Forex prediction engine: Framework, modelling techniques and implementations," *Int. J. Comput. Sci. Eng.*, vol. 13, no. 4, pp. 364–377, 2016.

MAURICIO ARGOTTY-ERAZO was born in La Unión, Nariño, Colombia, in 1976. He received the degree (Hons.) in industrial engineering and the Specialist degree in administration and management from Universidad Cooperativa de Colombia, Pasto, in 2000 and 2009, respectively, the master's degree (Hons.) in financial management and consulting from Universidad Mariana, in 2015, and the master's degree (Hons.) in advanced multivariate data analysis and big data from Universidad de Salamanca, Spain, in 2018, where he is currently pursuing the Ph.D. degree in applied multivariate statistics.

From 2002 to 2016, he was a Research Professor and a Coordinator of the Engineering Laboratories, Industrial Engineering Program, Universidad Cooperativa de Colombia, Pasto. In 2016, he was a Postgraduate Professor with the Senior Management, Accounting, and International Auditing Specialization Programs and the Administration and Competitiveness Master's Program, Universidad Mariana, Pasto, Colombia. Currently, he is a Master's Thesis Advisor with the Marketing Master's Program, Universidad de Nariño, Pasto. His research interests include quantitative analysis, simulation, financial modeling, multicriteria optimization, advanced multivariate data analysis and big data, statistical analysis, data visualization, machine learning, and artificial intelligence applied to financial markets.

ANTONIO BLÁZQUEZ-ZABALLOS was born in Macotera, Salamanca, Spain, in 1969. He received the B.S. degree in mathematics and the Ph.D. degree in mathematical sciences from Universidad de Salamanca, Spain, in 1992 and 1998, respectively.

He was a Research Professor with the Department of Statistics, Universidad de Salamanca, in the following positions: from 1994 to 1999, he was an Assistant University Professor. From 1999 to 2004, he was a full-time Associate Professor. From 2004 to 2009, he was an Assistant Professor (Ph.D.). Since 2009, he has been an Associate Professor (Ph.D.). His research interests include multivariate statistical methods applied to the analysis of economic and biological data and applied mathematics.

Dr. Blázquez-Zaballos awards and distinctions include the Extraordinary Doctorate Award (1997–1998).

CARLOS A. ARGOTY-ERASO was born in Pasto, Nariño, Colombia, in 1977. He received the degree in financial administration from Universidad Mariana, Pasto, in 2002. He is currently pursuing the master's degree in marketing with Universidad de Nariño, Colombia.

From 1998 to 2005, he was the Assistant Manager in the private sector. He promoted the launching of new products and the brand positioning with Lincoln Inversiones, Pasto. From 2002 to 2004, he worked in private banking, promoting access to new financial products and services. From 2005 to 2007, he was the Business Manager of GMAC Financiera de Colombia S.A., Pasto. Since 2008, he has been working for different local and national companies, both public and private, in researching market behavior and consumer needs. His research interests include the measurement, modeling, the management of financial risk, the formulation of credit scoring models, digital marketing and consumer psychology, predictive models of consumer behavior based on market data, and artificial intelligence methods.

LEANDRO L. LORENTE-LEYVA was born in Manzanillo, Cuba, in 1985. He received the degree in mechanical engineering and the master's degree in CAD/CAM from Universidad de Holguín, Cuba, in 2009 and 2011, respectively. He is currently pursuing the Ph.D. degree in engineering and industrial production with Universitat Politécnica de Valéncia (UPV), Spain. From 2009 to 2013, he was a Specialist with Astilleros de Oriente, Manzanillo, Cuba. From 2014 to 2015, he was a Project Advisor with Ministerio de Industrias, Cuba—PDVSA Industrial, Caracas—Venezuela. In 2015, he was a Senior Engineer with Empresa de Ingeniería del Reciclaje, Ministerio de Industrias, Cuba. From 2015 to 2020, he was a Researcher/Professor with Universidad Técnica del Norte, Ecuador. Since 2021, he has been a Program Coordinator with the Postgraduate Center, Universidad Politécnica Estatal del Carchi, Ecuador. He is currently a Senior Researcher with the SDAS Research Group. His professional and research interests include discrete and computational optimization and artificial intelligence applications, manufacturing engineering, computer-aided design, planning and manufacturing (CAD/CAPP/CAM), and application of numerical methods in engineering. In addition, time series forecasting, data analysis, and the applications of heuristic methods as solutions to engineering problems.

NADIA N. SÁNCHEZ-POZO was born in Tulcán, Ecuador, in 1992. She received the degree in electronics and communication networks engineering from Universidad Técnica del Norte, Ibarra, Ecuador, in 2018, and the master's degree in data science from Universitat Oberta de Catalunya, Spain, in 2021. From 2018 to 2022, she was a DL/ML Engineer with the SDAS Research Group. Currently, she is a Master's Thesis Advisor and a Lecturer with the Applied Statistics Master's Program, Universidad Politécnica Estatal del Carchi (UPEC), Ecuador. Her main research interests include big data, machine learning, and natural language processing.

DIEGO H. PELUFFO-ORDÓÑEZ was born in Pasto, Colombia, in 1986. He received the degree in electronic engineering, the master's degree in industrial automation, and the Ph.D. degree in engineering from Universidad Nacional de Colombia, Manizales, Colombia, in 2008, 2010, and 2013, respectively. In 2012, he undertook his doctoral internship with KU Leuven, Belgium. From 2013 to 2014, he was a Postdoctoral Researcher with Université Catholique de Louvain, Louvain la-Neuve, Belgium. From 2014 to 2015, he was a Lecturer with Universidad Cooperativa de Colombia, Pasto. From 2015 to 2017, he was a Researcher/Professor with Universidad Técnica del Norte, Ecuador. From 2017 to 2020, he was a Researcher/Professor awith Yachay Tech. Ecuador. From 2020 to 2022, he was a Consultant/Curriculum Author with deeplearning.ai. Currently, he is an Assistant Professor with the College of Computing, Mohammed VI Polytechnic University, Morocco. He is also the Founder and the Head of the SDAS Research Group. He is also an invited Lecturer and an External Researcher with Corporación Universitaria Autónoma de Nariño, Pasto, and a member of the SEDMATEC Research Group. Also, he is an External Collaborator with the Writing Laboratory, Tecnológico de Monterrey, Mexico. He is also an External Supervisor of Ph.D. Programs with Universidad de Granada, Spain, Universitat Politécnica de Valéncia, Spain, and Universidad Nacional de La Plata, Argentina. His main research interests include kernel-based and spectral methods for data clustering and dimensionality reduction, complex high-dimensional data, signal, image, and video analysis for medical and industry applications. He has served as an organizing committee member (the general chair, the session chair, and the competitions chair) and a keynote speaker at several conferences. Also, he has served as a Guest Editor for the *Computers and Electrical Engineering Journal*.

Bibliografía

[1] M. A. Ayadi, W. Ben Omrane, J. Y. Wang, and R. Welch. Senior official speech attributes and foreign exchange risk around business cycles. *International Review of Financial Analysis*, 80, 2022.

[2] Afees A. Salisu and Xuan Vinh Vo. Predicting stock returns in the presence of covid-19 pandemic: The role of health news. *International Review of Financial Analysis*, 71:101546, 2020.

[3] Walid Ben Omrane and Tanseli Savaser. Exchange rate volatility response to macroeconomic news during the global financial crisis. *International Review of Financial Analysis*, 52:130–143, 2017.

[4] N. Karnaukh, A. Ranaldo, and P. Soderlind. Understanding fx liquidity. *Review of Financial Studies*, 28(11):3073–3108, 2015.

[5] Y. W. Cheung and M. D. Chinn. Currency traders and exchange rate dynamics: a survey of the us market. *Journal of International Money and Finance*, 20(4):439–471, 2001.

[6] Muhammad Naeem, Hammad Ejaz Khan, and Raza Ali. Bibliometric literature review on exchange rate: A future research agenda. *Etikonomi*, 21(1):41–54, 2022.

[7] Martha Flores-Sosa, Ezequiel Aviles-Ochoa, and Jose M. Merigo. Exchange rate and volatility: A bibliometric review. *International Journal of Finance & Economics*, 27(1):1419–1442, 2022.

[8] M. Fratzscher. What explains global exchange rate movements during the financial crisis? *Journal of International Money and Finance*, 28(8):1390–1407, 2009.

[9] Michael T. Belongia and Peter N. Ireland. Strengthening the second pillar: a greater role for money in the ecb's strategy. *Applied Economics*, 54(1):99–114, 2022.

[10] Marwadi and Rofikoh Rokhim. Jakarta interbank spot dollar rate (jisdor) as the reference rate: Is it effective? *Indonesian Capital Market Review*, 8(2), 2016.

[11] Michael Broll. The skewness risk premium in currency markets. *Economic Modelling*, 58:494–511, 2016.

[12] Pasquale Della Corte, Tarun Ramadorai, and Lucio Sarno. Volatility risk premia and exchange rate predictability. *Journal of Financial Economics*, 120(1):21–40, 2016.

[13] Lukas Menkhoff. The use of technical analysis by fund managers: International evidence. *Journal of Banking & Finance*, 34(11):2573–2586, 2010.

[14] G. Allayannis and E. Ofek. Exchange rate exposure, hedging, and the use of foreign currency derivatives. *Journal of International Money and Finance*, 20(2):273–296, 2001.

[15] N. Eiamkanitchat, T. Moontuy, and S. Ramingwong. Fundamental analysis and technical analysis integrated system for stock filtration. *Cluster Computing-the Journal of Networks Software Tools and Applications*, 20(1):883–894, 2017.

[16] John J Murphy. Technical analysis of the futures markets. *A Comprehensive Guide to Trading Methods and Applications*, 1986.

[17] Thomas Gehrig and Lukas Menkhoff. Extended evidence on the use of technical analysis in foreign exchange. *International Journal of Finance & Economics*, 11(4):327–338, 2006.

[18] Robert D Edwards, John Magee, and WH Charles Bassetti. *Technical analysis of stock trends*. CRC press, 1967.

[19] William F Sharpe, Gordon J Alexander, and Jeffrey W Bailey. Investments. 1985.

[20] Burton G Malkiel. *A random walk down Wall Street. W. W.* Norton & Co, New York, 1973.

[21] Michael C Jensen. Random walks: reality or myth comment. *Financial Analysts Journal*, 23(6):77–85, 1967.

[22] Paul A Samuelson. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6:41–49, 1965.

[23] Holdbrook Working. New concepts concerning futures markets and prices. *The American Economic Review*, 51(2):160–163, 1961.

[24] Holbrook Working. A random-difference series for use in the analysis of time series. *Journal of the American Statistical Association*, 29(185):11–24, 1934.

[25] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[26] Kai-Yin Woo, Chulin Mai, Michael McAleer, and Wing-Keung Wong. Review on efficiency and anomalies in stock markets. *Economies*, 8(1), 2020.

[27] Lukas Menkhoff and Mark P Taylor. The obstinate passion of foreign exchange professionals: technical analysis. *Journal of Economic Literature*, 45(4):936–972, 2007.

[28] Andrew W. Lo and A. Craig MacKinlay. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies*, 1(1):41–66, 1988.

[29] Weiwei Jiang. Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, 184:115537, 2021.

[30] Gourav Kumar, Sanjeev Jain, and Uday Pratap Singh. Stock market forecasting using computational intelligence: A survey. *Archives of Computational Methods in Engineering*, 28(3):1069–1101, 2021.

[31] M. Watorek, S. Drozdz, J. Kwapien, L. Minati, P. Oswiecimka, and M. Stanuszek. Multiscale characteristics of the emerging global cryptocurrency market. *Physics Reports-Review Section of Physics Letters*, 901:1–82, 2021.

[32] Zahra Hajirahimi and Mehdi Khashei. Hybridization of hybrid structures for time series forecasting: a review. *Artificial Intelligence Review*, pages 1–61, 2022.

[33] D. Pradeepkumar and V. Ravi. Soft computing hybrids for forex rate prediction: A comprehensive review. *Computers & Operations Research*, 99:262–284, 2018.

[34] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005-2019. *Applied Soft Computing*, 90, 2020.

[35] Xiaodong Li, Pangjing Wu, and Wenpeng Wang. Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong. *Information Processing & Management*, 57(5), 2020.

[36] Guofu Zhou. Measuring investor sentiment. *Annual Review of Financial Economics*, 10:239–259, 2018.

[37] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670, 2014.

[38] Z. X. Hu, Y. Q. Zhao, and M. Khushi. A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, 4(1), 2021.

[39] Saeed Nosratabadi, Amirhosein Mosavi, Duan Puhong, Pedram Ghamisi, Ferdinand Filip, Shahab S. Band, Uwe Reuter, Joao Gama, and Amir H. Gandomi. Data science in economics: Comprehensive review of advanced machine learning and deep learning methods. *Mathematics*, 8(10), 2020.

[40] Salim Lahmiri. Modeling and predicting historical volatility in exchange rate markets. *Physica A: Statistical Mechanics and its Applications*, 471:387–395, 2017.

[41] Werner Kristjanpoller and Marcel C. Minutolo. Forecasting volatility of oil price using an artificial neural network-garch model. *Expert Systems with Applications*, 65:233–241, 2016.

[42] John W. Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32:100577, 2021.

[43] Guoqiang Peter Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462, 2000.

[44] Luis A. Berrueta, Rosa M. Alonso-Salces, and Károly Héberger. Supervised pattern recognition in food analysis. *Journal of Chromatography A*, 1158(1):196–214, 2007.

[45] S. Usmani and J. A. Shamsi. News sensitive stock market prediction: literature review and suggestions. *Peerj Computer Science*, 2021.

[46] Nagendra Kumar and Andreas G Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech communication*, 26(4):283–297, 1998.

[47] Smarajit Bose, Amita Pal, Rita SahaRay, and Jitadeepa Nayak. Generalized quadratic discriminant analysis. *Pattern Recognition*, 48(8):2676–2684, 2015.

[48] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[49] Charles Bouveyron, Stéphane Girard, and Cordelia Schmid. High-dimensional discriminant analysis. *Communications in Statistics Theory and Methods*, 36(14):2607–2623, 2007.

[50] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

[51] I. Gurrib and F. Kamalov. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Studies in Economics and Finance*, 39(3):347–364, 2022.

[52] Mauro Bernardi, Leopoldo Catania, and Lea Petrella. Are news important to predict the value-at-risk? *European Journal of Finance*, 23(6):535–572, 2017.

[53] K. Grobys and N. Sapkota. Predicting cryptocurrency defaults. *Applied Economics*, 52(46):5060–5076, 2020.

[54] Seung Woog Kwag and Yong Seog Kim. Stock price predictability of financial ratios and macroeconomic variables: A regulatory perspective. *Industrial Engineering and Management Systems*, 12(4):406–415, 2013.

[55] Sibusiso T. Mndawe, Babu Sena Paul, and Wesley Doorsamy. Development of a stock price prediction framework for intelligent media and technical analysis. *Applied Sciences-Basel*, 12(2), 2022.

[56] Gregor Weiss. Copula-garch versus dynamic conditional correlation: an empirical study on var and es forecasting accuracy. *Review of Quantitative Finance and Accounting*, 41(2):179–202, 2013.

[57] J. Iworiso and S. Vrontos. On the directional predictability of equity premium using machine learning techniques. *Journal of Forecasting*, 39(3):449–469, 2020.

[58] Zheshi Chen, Chunhong Li, and Wenjun Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365, 2020.

[59] M. T. Leung, H. Daouk, and A. S. Chen. Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of Forecasting*, 16(2):173–190, 2000.

[60] Ross P Buckley, Dirk A Zetzsche, Douglas W Arner, and Brian W Tang. Regulating artificial intelligence in finance: Putting the human in the loop. *Sydney Law Review, The*, 43(1):43–81, 2021.

[61] Jeffrey A Frankel and Kenneth A Froot. Chartists, fundamentalists, and trading in the foreign exchange market. *The American Economic Review*, 80(2):181–185, 1990.

[62] Stephen J Taylor. Rewards available to currency futures speculators: compensation for risk or evidence of inefficient pricing? *Economic Record*, 68:105–116, 1992.

[63] Lukas Menkhoff. Examining the use of technical currency analysis. *International Journal of Finance & Economics*, 2(4):307–318, 1997.

[64] Yu-Hon Lui and David Mole. The use of fundamental and technical analyses by foreign exchange dealers: Hong kong evidence. *Journal of International money and Finance*, 17(3):535–545, 1998.

[65] Thomas Oberlechner. Importance of technical and fundamental analysis in the european foreign exchange market. *International Journal of Finance & Economics*, 6(1):81–93, 2001.

[66] Yin-Wong Cheung, Menzie D Chinn, and Ian W Marsh. How do uk-based foreign exchange dealers think their market operates? *International Journal of Finance & Economics*, 9(4):289–306, 2004.

[67] Michael C Jensen. Some anomalous evidence regarding market efficiency. *Journal of financial economics*, 6(2/3):95–101, 1978.

[68] Michael C Jensen and George A Benington. Random walks and technical theories: Some additional evidence. *The Journal of finance*, 25(2):469–482, 1970.

[69] Benoit Mandelbrot. Forecasts of future prices, unbiased markets, and "martingale" models. *The Journal of Business*, 39(1):242–255, 1966.

[70] Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419, 1963.

[71] Holbrook Working. The investigation of economic expectations. *The American Economic Review*, 39(3):150–166, 1949.

[72] Louis Bachelier. Théorie de la spéculation. In *Annales scientifiques de l'École normale supérieure*, volume 17, pages 21–86, 1900.

[73] Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, and Philip S. Yu. Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems*, 143:236–247, 2018.

[74] Mahinda Mailagaha Kumbure, Christoph Lohrmann, Pasi Luukka, and Jari Porras. Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, 197:116659, 2022.

[75] S. Aziz, M. Dowling, H. Hammami, and A. Piepenbrink. Machine learning in finance: A topic modeling approach. *European Financial Management*, 2021.

[76] Ahmed M Khedr, Ifra Arif, Magdi El-Bannany, Saadat M Alhashmi, and Meenu Sreedharan. Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *Intelligent Systems in Accounting, Finance and Management*, 28(1):3–34, 2021.

[77] Trilok Nath Pandey, Alok Kumar Jagadev, Satchidananda Dehuri, and Sung-Bae Cho. A review and empirical analysis of neural networks based exchange rate prediction. *Intelligent Decision Technologies*, 12(4):423–439, 2018.

[78] Irfan Ramzan Parray, Surinder Singh Khurana, Munish Kumar, and Ali A Altalbe. Time series data analysis of stock price movement using machine learning techniques. *Soft Computing*, 24(21):16509–16517, 2020.

[79] Luca Cagliero, Paolo Garza, Giuseppe Attanasio, and Elena Baralis. Training ensembles of faceted classification models for quantitative stock trading. *Computing*, 102(5):1213–1225, 2020.

[80] Francesco Rundo. Deep lstm with reinforcement learning layer for financial trend prediction in fx high frequency trading systems. *Applied Sciences*, 9(20):4460, 2019.

[81] Krzysztof Drachal and Michał Pawłowski. A review of the applications of genetic algorithms to forecasting prices of commodities. *Economies*, 9(1):6, 2021.

[82] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[83] Ankit Thakkar and Kinjal Chaudhari. A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Systems with Applications*, 177:114800, 2021.

[84] Mei-Li Shen, Cheng-Feng Lee, Hsiou-Hsiang Liu, Po-Yin Chang, and Cheng-Hong Yang. An effective hybrid approach for forecasting currency exchange rates. *Sustainability*, 13(5), 2021.

[85] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[86] Shian-Chang Huang and Cheng-Feng Wu. Energy commodity price forecasting with deep multiple kernel learning. *Energies*, 11(11):3029, 2018.

[87] J. Pekar and M. Pcolar. Empirical distribution of daily stock returns of selected developing and emerging markets with application to financial risk management. *Central European Journal of Operations Research*, 30(2):699–731, 2022.

[88] Robert A. Eisenbeis. Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, 32(3):875–900, 1977.

[89] Saswat Patra. Revisiting value-at-risk and expected shortfall in oil markets under structural breaks: The role of fat-tailed distributions. *Energy Economics*, 101:105452, 2021.

[90] Jone B Horpestad, Štefan Lyócsa, Peter Molnár, and Torbjørn B Olsen. Asymmetric volatility in equity markets around the world. *The North American Journal of Economics and Finance*, 48:540–554, 2019.

[91] Canan G Corlu and Alper Corlu. Modelling exchange rate returns: which flexible distribution to use? *Quantitative Finance*, 15(11):1851–1864, 2015.

[92] Hung-Chun Liu and Jui-Cheng Hung. Forecasting s&p-100 stock index volatility: The role of volatility asymmetry and distributional assumption in garch models. *Expert Systems with Applications*, 37(7):4928–4934, 2010.

[93] Pilsun Choi and Kiseok Nam. Asymmetric and leptokurtic distribution for heteroscedastic asset returns: the su-normal distribution. *Journal of Empirical finance*, 15(1):41–63, 2008.

[94] Jui-Cheng Hung, Ming-Chih Lee, and Hung-Chun Liu. Estimation of value-at-risk for energy commodities via fat-tailed garch models. *Energy Economics*, 30(3):1173–1191, 2008.

[95] Dimitris N Politis. A heavy-tailed distribution for arch residuals with application to volatility prediction. 2004.

[96] Panayiotis Theodossiou. Financial data and the skewed generalized t distribution. *Management Science*, 44(12-part-1):1650–1661, 1998.

[97] Tim Bollerslev. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, pages 542–547, 1987.

[98] Thomas D Willett. New developments in financial economics. *Journal of Financial Economic Policy*, 2021.

[99] Paul Goodwin and Robert Fildes. *Forecasting with Judgment*, pages 541–572. Springer International Publishing, Cham, 2022.

[100] John Maynard Keynes. The general theory of employment. *The quarterly journal of economics*, 51(2):209–223, 1937.

[101] Maury Klein. The stock market crash of 1929: A review article. *Business History Review*, 75(2):325–351, 2001.

[102] J.K. Galbraith and P. Chemla. *Brève histoire de l'euphorie financière*. Le grand livre du mois. Editions du Seuil, 1992.

[103] Olivier Blanchard, Changyong Rhee, and Lawrence Summers. The stock market, profit, and investment. *The Quarterly Journal of Economics*, 108(1):115–136, 1993.

[104] J Bradford De Long, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann. Noise trader risk in financial markets. *Journal of political Economy*, 98(4):703–738, 1990.

[105] Yasuhiro Sakai. Jm keynes on probability versus fh knight on uncertainty: reflections on the miracle year of 1921. In *JM Keynes Versus FH Knight*, pages 39–60. Springer, 2019.

[106] IV Rozmainskiy. The common theory of jm keynes: lessons 75 years later. *Terra Economicus*, 10(1):46–52, 2012.

[107] Bruno S Frey and Marcel Kucher. History as reflected in capital markets: the case of world war ii. *The Journal of Economic History*, 60(2):468–496, 2000.

[108] William O Brown, Jr and Richard CK Burdekin. German debt traded in london during the second world war: a british perspective on hitler. *Economica*, 69(276):655–669, 2002.

[109] Barry Smart. Another great transformation or common ruin? *Theory, Culture & Society*, 28(2):131–151, 2011.

[110] Victor Troster, Elie Bouri, and David Roubaud. A quantile regression analysis of flights-to-safety with implied volatilities. *Resources Policy*, 62:482–495, 2019.

[111] Taufiq Choudhry, Syed S Hassan, and Sarosh Shabi. Relationship between gold and stock markets during the global financial crisis: Evidence from nonlinear causality tests. *International Review of Financial Analysis*, 41:247–256, 2015.

[112] Robert J Shiller. The subprime solution. In *The Subprime Solution*. Princeton University Press, 2012.

[113] Ray Barrell, Ian Hurst, and Simon Kirby. Financial crises, regulation and growth. *National Institute Economic Review*, 206(1):56–65, 2008.

[114] Richard H Thaler and Cass R Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[115] Maria Jose Roa Garcia. Financial education and behavioral finance: new insights into the role of information in financial decisions. *Journal of economic surveys*, 27(2):297–315, 2013.

[116] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[117] Bruce G Carruthers. From uncertainty toward risk: The case of credit ratings. *Socio-Economic Review*, 11(3):525–551, 2013.

[118] Bruce G Carruthers and Laura Ariovich. *Money and credit: A sociological approach*, volume 6. Polity, 2010.

[119] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*, 2020.

[120] International Monetary Fund. Monetary and Capital Markets Department. *Annual Report on Exchange Arrangements and Exchange Restrictions 2021*. International Monetary Fund, USA, 2022.

[121] Shaista Jabeen, Muhammad Farhan, Muhammad Ahmad Zaka, Muhammad Fiaz, and Mobina Farasat. Covid and world stock markets: A comprehensive discussion. *Frontiers in Psychology*, page 4837, 2022.

[122] Satish Kumar, Sandeep Rao, Kirti Goyal, and Nisha Goyal. Journal of behavioral and experimental finance: A bibliometric overview. *Journal of Behavioral and Experimental Finance*, page 100652, 2022.

[123] Mohsin Ali, Nafis Alam, and Syed Aun R Rizvi. Coronavirus (covid-19): An epidemic or pandemic for financial markets. *Journal of Behavioral and Experimental Finance*, 27:100341, 2020.

[124] Omair Haroon and Syed Aun R Rizvi. Covid-19: Media coverage and financial markets behavior a sectoral inquiry. *Journal of Behavioral and Experimental Finance*, 27:100343, 2020.

[125] Kristi Raik. The ukraine crisis as a conflict over europe's political, economic and security order. *Geopolitics*, 24(1):51–70, 2019.

[126] Richard Connolly. *Russia's Response to Sanctions*, pages i–ii. Cambridge University Press, 2018.

[127] Qiang Ji, Dayong Zhang, and Yuqian Zhao. Searching for safe-haven assets during the covid-19 pandemic. *International Review of Financial Analysis*, 71:101526, 2020.

[128] Walid Mensi, Shawkat Hammoudeh, Duc Khuong Nguyen, and Sang Hoon Kang. Global financial crisis and spillover effects among the us and brics stock markets. *International Review of Economics & Finance*, 42:257–276, 2016.

[129] Marie Brière, Ariane Chapelle, and Ariane Szafarz. No contagion, only globalization and flight to quality. *Journal of international Money and Finance*, 31(6):1729–1744, 2012.

[130] Joe Brocato and Kenneth L Smith. Sudden equity price declines and the flight-to-safety phenomenon: additional evidence using daily data. *Journal of Economics and Finance*, 36(3):712–727, 2012.

[131] Sang Hoon Kang and Seong-Min Yoon. Financial crises and dynamic spillovers among chinese stock and commodity futures markets. *Physica A: Statistical Mechanics and its Applications*, 531:121776, 2019.

[132] Alexander Engel. Futures and risk: the rise and demise of the hedger-speculator dichotomy. *Socio-Economic Review*, 11(3):553–576, 2013.

[133] Chris Starmer. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature*, 38(2):332–382, 2000.

[134] Wing-Keung Wong. Review on behavioral economics and behavioral finance. *Studies in Economics and Finance*, 2020.

[135] Aditya Sharma and Arya Kumar. A review paper on behavioral finance: study of emerging trends. *Qualitative Research in Financial Markets*, 2019.

[136] Michael M Pompian. *Behavioral finance and wealth management: how to build investment strategies that account for investor biases*. John Wiley & Sons, 2011.

[137] John J Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[138] Robert A Olsen. Behavioral finance and its implications for stock-price volatility. *Financial analysts journal*, 54(2):10–18, 1998.

[139] Cars Hommes and Florian Wagener. Complex evolutionary systems in behavioral finance. In *Handbook of financial markets: Dynamics and evolution*, pages 217–276. Elsevier, 2009.

[140] Soosung Hwang and Alexandre Rubesam. A behavioral explanation of the value anomaly based on time-varying return reversals. *Journal of banking & finance*, 37(7):2367–2377, 2013.

[141] Collin Read. The efficient market hypothesis, 2012.

[142] Paul A Samuelson, M Davis, and A Etheridge. *Louis Bachelier's theory of speculation: the origins of modern finance*. JSTOR, 2006.

[143] Harry M Markowitz. Portfolio theory: as i still see it. *Annu. Rev. Financ. Econ.*, 2(1):1–23, 2010.

[144] William F Sharpe, Gordon J Alexander, and Jeffery V Bailey. *Investment*. Prentice Hall Incorporated, 1999.

[145] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

[146] William F Sharpe. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, pages 169–185, 1998.

[147] Don Galagedera. A review of capital asset pricing models. *Managerial Finance*, 33(10):821–832, 2007. Galagedera, Don/ABB-2772-2021 Galagedera, Don/0000-0003-2990-4997 Si 13.

[148] Fischer Black and Myron Scholes. The valuation of option contracts and a test of market efficiency. *The journal of finance*, 27(2):399–417, 1972.

[149] Paul A Samuelson. Proof that properly anticipated prices fluctuate randomly. In *The world scientific handbook of futures markets*, pages 25–38. World Scientific, 2016.

[150] Eric B Lindenberg and Stephen A Ross. Tobin's q ratio and industrial organization. *Journal of business*, pages 1–32, 1981.

[151] Nai-Fu Chen, Richard Roll, and Stephen A Ross. Economic forces and the stock market. *Journal of business*, pages 383–403, 1986.

[152] Haim Levy. The capm is alive and well: A review and synthesis. *European Financial Management*, 16(1):43–71, 2010.

[153] Francesco Rocciolo, Andrea Gheno, and Chris Brooks. Explaining abnormal returns in stock markets: An alpha-neutral version of the capm. *International Review of Financial Analysis*, 82:102143, 2022.

[154] Keith Cuthbertson, Dirk Nitzsche, and Stuart Hyde. Monetary policy and behavioural finance. *Journal of Economic Surveys*, 21(5):935–969, 2007.

[155] Hersh Shefrin and Meir Statman. Behavioral portfolio theory. *Journal of financial and quantitative analysis*, 35(2):127–151, 2000.

[156] John Y Campbell and John H Cochrane. Explaining the poor performance of consumption-based asset pricing models. *The Journal of Finance*, 55(6):2863–2878, 2000.

[157] Constantinos Antoniou, John A Doukas, and Avanidhar Subrahmanyam. Investor sentiment, beta, and the cost of equity capital. *Management Science*, 62(2):347–367, 2016.

[158] Frankie Chau, Rataporn Deesomsak, and Dimitrios Koutmos. Does investor sentiment really matter? *International Review of Financial Analysis*, 48:221–232, 2016.

[159] Lu Zhang. The investment capm. *European Financial Management*, 23(4):545–603, 2017.

[160] Meir Statman. Behavioral finance: Finance with normal people. *Borsa Istanbul Review*, 14(2):65–73, 2014.

[161] Andrei Semenov. Departures from rational expectations and asset pricing anomalies. *The Journal of Behavioral Finance*, 10(4):234–241, 2009.

[162] Abhijeet Chandra and M Thenmozhi. Behavioural asset pricing: Review and synthesis. *Journal of Interdisciplinary Economics*, 29(1):1–31, 2017.

[163] Robert J Shiller. *Market volatility*. MIT press, 1992.

[164] Robert J Shiller. Irrational exuberance. In *Irrational exuberance*. Princeton university press, 2015.

[165] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[166] Rongxin Chen, Gabriele M Lepori, Chung-Ching Tai, and Ming-Chien Sung. Explaining cryptocurrency returns: A prospect theory perspective. *Journal of International Financial Markets, Institutions and Money*, 79:101599, 2022.

[167] Gerd Gigerenzer. *Risk savvy: How to make good decisions*. Penguin, 2015.

[168] William Forbes, Robert Hudson, Len Skerratt, and Mona Soufian. Which heuristics can aid financial-decision-making? *International Review of Financial Analysis*, 42:199–210, 2015.

[169] Hammad Siddiqi. Anchoring-adjusted capital asset pricing model. *Journal of Behavioral Finance*, 19(3):249–270, 2018.

[170] Maqsood Ahmad. The role of recognition-based heuristics in investment management activities: are expert investors immune?–a systematic literature review. *Qualitative Research in Financial Markets*, 2021.

[171] Fergus Bolger and Nigel Harvey. Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology Section A*, 46(4):779–811, 1993.

[172] Michael Lawrence, Paul Goodwin, Marcus O'Connor, and Dilek Önkal. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of forecasting*, 22(3):493–518, 2006.

[173] Malcolm Baker and Jeffrey Wurgler. Investor sentiment in the stock market. *Journal of economic perspectives*, 21(2):129–152, 2007.

[174] Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680, 2006.

[175] Nicholas Barberis, Andrei Shleifer, and Robert Vishny. A model of investor sentiment. *Journal of financial economics*, 49(3):307–343, 1998.

[176] Robert D Edwards, John Magee, and WH Charles Bassetti. *Technical analysis of stock trends*. CRC press, 2018.

[177] Maliheh Rezaei Adariani. Evaluation of the profitability of technical analysis for asian currencies in the forex spot market for short-term trading. *AU-GSB e-JOURNAL*, 5(2), 2012.

[178] Cheol-Ho Park and Scott H Irwin. What do we know about the profitability of technical analysis? *Journal of Economic surveys*, 21(4):786–826, 2007.

[179] Martin J Pring. *Technical analysis explained: The successful investor's guide to spotting investment trends and turning points*. McGraw-Hill Professional, 2002.

[180] Anne Walters, Vikash Ramiah, and Imad Moosa. Ecology and finance: A quest for congruency. *Journal of Behavioral and Experimental Finance*, 10:54–62, 2016.

[181] Peter CB Phillips, Yangru Wu, and Jun Yu. Explosive behavior in the 1990s nasdaq: When did exuberance escalate asset values? *International economic review*, 52(1):201–226, 2011.

[182] Harry G Johnson. The case for flexible exchange rates, 1969. *Federal Reserve Bank of St. Louis Review*, (1), 1969.

[183] Jeffrey A Frankel, Kenneth A Froot, et al. Understanding the us dollar in the eighties: the expectations of chartists and fundamentalists. *Economic record*, 62(1):24–38, 1986.

[184] Jeffrey A Frankel and Kenneth Froot. Chartists, fundamentalists and the demand for dollars. *Available at SSRN 228045*, 1991.

[185] Charles Goodhart. The foreign exchange market: a random walk with a dragging anchor. *Economica*, pages 437–460, 1988.

[186] Alan Kirman. Ants, rationality, and recruitment. *The Quarterly Journal of Economics*, 108(1):137–156, 1993.

[187] Robert J Shiller. Investor behavior in the october 1987 stock market crash: Survey evidence, 1987.

[188] Robert Shiller. Speculative booms and crashes. In *Monetary Economics in the 1990s*, pages 58–74. Springer, 1996.

[189] Michael John Artis. International economic policy co-ordination: Theory and practice. *Oxford Review of Economic Policy*, 5(3):83–93, 1989.

[190] Robert J Shiller, Stanley Fischer, and Benjamin M Friedman. Stock prices and social dynamics. *Brookings papers on economic activity*, 1984(2):457–510, 1984.

[191] Fischer Black. Noise. *The journal of finance*, 41(3):528–543, 1986.

[192] J Bradford DeLong, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann. The economic consequences of noise traders, 1987.

[193] John Y Campbell and Albert S Kyle. Smart money, noise trading and stock price behaviour. *The Review of Economic Studies*, 60(1):1–34, 1993.

[194] Mark P Taylor. What do investment managers know? an empirical study of practioners' predictions. *Economica*, pages 185–202, 1988.

[195] Mark P Taylor. Expectations, risk and uncertainty in the foreign exchange market: some results based on survey data. *The Manchester School*, 57(2):142–153, 1989.

[196] John Cochrane. *Asset pricing: Revised edition*. Princeton university press, 2009.

[197] Allan Timmermann and Clive WJ Granger. Efficient market hypothesis and forecasting. *International Journal of forecasting*, 20(1):15–27, 2004.

[198] Thomas Gehring and Lukas Menkhoff. Technical analysis in foreign exchange-the workhorse gains further ground. Technical report, Diskussionsbeitrag, 2003.

[199] Prodromos Tsinaslanidis and Francisco Guijarro. What makes trading strategies based on chart pattern recognition profitable? *Expert Systems*, 38(5):e12596, 2021.

[200] Audeliano Wolian Li and Guilherme Sousa Bastos. Stock market forecasting using deep learning and technical analysis: a systematic review. *IEEE access*, 8:185232–185242, 2020.

[201] Rubén Aguilar-Rivera, Manuel Valenzuela-Rendón, and JJ Rodríguez-Ortiz. Genetic algorithms and darwinian approaches in financial applications: A survey. *Expert Systems with Applications*, 42(21):7684–7697, 2015.

[202] Rodolfo C Cavalcante, Rodrigo C Brasileiro, Victor LF Souza, Jarley P Nobrega, and Adriano LI Oliveira. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55:194–211, 2016.

[203] Gourav Kumar, Sanjeev Jain, and Uday Pratap Singh. Stock market forecasting using computational intelligence: A survey. *Archives of Computational Methods in Engineering*, 28(3):1069–1101, 2021.

[204] Yuhong Li and Weihua Ma. Applications of artificial neural networks in financial economics: a survey. In *2010 International symposium on computational intelligence and design*, volume 1, pages 211–214. IEEE, 2010.

[205] Sneha Soni. Applications of anns in stock market prediction: a survey. *International Journal of Computer Science & Engineering Technology*, 2(3):71–83, 2011.

[206] George S Atsalakis and Kimon P Valavanis. Surveying stock market forecasting techniques–part ii: Soft computing methods. *Expert Systems with applications*, 36(3):5932–5941, 2009.

[207] Azadeh Nikfarjam, Ehsan Emadzadeh, and Saravanan Muthaiyah. Text mining approaches for stock market prediction. In *2010 The 2nd international conference on computer and automation engineering (ICCAE)*, volume 4, pages 256–260. IEEE, 2010.

[208] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3):797–827, 2006.

[209] Christopher Krauss. Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2):513–545, 2017.

[210] John Lehoczky and Mark Schervish. Overview and history of statistics for equity markets. *Annual Review of Statistics and Its Application*, 5:265–288, 2018.

[211] Rodolfo Toríbio Farias Nazário, Jéssica Lima e Silva, Vinicius Amorim Sobreiro, and Herbert Kimura. A literature review of technical analysis on stock markets. *The Quarterly Review of Economics and Finance*, 66:115–126, 2017.

[212] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1):65–91, 1993.

[213] Yong Hu, Kang Liu, Xiangzhou Zhang, Lijun Su, EWT Ngai, and Mei Liu. Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review. *Applied Soft Computing*, 36:534–551, 2015.

[214] Narasimhan Jegadeesh and Sheridan Titman. Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of finance*, 56(2):699–720, 2001.

[215] Thomas J George and Chuan-Yang Hwang. The 52-week high and momentum investing. *The Journal of Finance*, 59(5):2145–2176, 2004.

[216] Doron Avramov, Tarun Chordia, Gergana Jostova, and Alexander Philipov. Momentum and credit rating. *The journal of Finance*, 62(5):2503–2520, 2007.

[217] Yunlin Yang, Bartosz Gebka, and Robert Hudson. Momentum effects in china: A review of the literature and an empirical explanation of prevailing controversies. *Research in International Business and Finance*, 47:78–101, 2019.

[218] Timothy King and Dimitrios Koutmos. Herding and feedback trading in cryptocurrency markets. *Annals of Operations Research*, 300(1):79–96, 2021.

[219] Fotini Economou, Konstantinos Gavriilidis, Bartosz Gebka, and Vasileios Kallinterakis. Feedback trading: a review of theory and empirical evidence. *Review of Behavioral Finance*, 2022.

[220] Werner FM De Bondt and Richard Thaler. Does the stock market overreact? *The Journal of finance*, 40(3):793–805, 1985.

[221] Josef Lakonishok, Andrei Shleifer, and Robert W Vishny. Contrarian investment, extrapolation, and risk. *The journal of finance*, 49(5):1541–1578, 1994.

[222] Yangru Wu. Momentum trading, mean reversal and overreaction in chinese stock market. *Review of Quantitative Finance and Accounting*, 37(3):301–323, 2011.

[223] Joseph Kang, Ming-Hua Liu, and Sophie Xiaoyan Ni. Contrarian and momentum strategies in the china stock market: 1993–2000. *Pacific-Basin Finance Journal*, 10(3):243–265, 2002.

[224] Bruce N Lehmann. Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, 105(1):1–28, 1990.

[225] Tim A Herberger, Matthias Horn, and Andreas Oehler. Are intraday reversal and momentum trading strategies feasible? an analysis for german blue chip stocks. *Financial Markets and Portfolio Management*, 34(2):179–197, 2020.

[226] Jianhua Gang, Zongxin Qian, and Tiange Xu. Investment horizons, cash flow news, and the profitability of momentum and reversal strategies in the chinese stock market. *Economic Modelling*, 83:364–371, 2019.

[227] Michael Cooper. Filter rules based on price and volume in individual security overreaction. *The Review of Financial Studies*, 12(4):901–935, 1999.

[228] Roberto C Gutierrez Jr and Eric K Kelley. The long-lasting momentum in weekly returns. *The Journal of finance*, 63(1):415–447, 2008.

[229] Chunjuan Zhuang. Improving performance of exchange rate momentum strategy using volatility information. *Physica A: Statistical Mechanics and its Applications*, 510:741–753, 2018.

[230] Ahmad Raza, Ben R Marshall, and Nuttawat Visaltanachoti. Is there momentum or reversal in weekly currency returns? *Journal of International Money and Finance*, 45:38–60, 2014.

[231] Clifford Asness, Andrea Frazzini, Ronen Israel, and Tobias Moskowitz. Fact, fiction, and momentum investing. *The Journal of Portfolio Management*, 40(5):75–92, 2014.

[232] Gil Cohen. Algorithmic trading and financial forecasting using advanced artificial intelligence methodologies. *Mathematics*, 10(18):3302, 2022.

[233] Dimitar Bogoev and Arzé Karam. Detection of algorithmic trading. *Physica A: Statistical Mechanics and its Applications*, 484:168–181, 2017.

[234] Chenye Liu, Ying Wu, and Dongming Zhu. Price overreaction to up-limit events and revised momentum strategies in the chinese stock market. *Economic Modelling*, page 105910, 2022.

[235] Emilios C Galariotis. Contrarian and momentum trading: a review of the literature. *Review of Behavioral Finance*, 2014.

[236] Sanjay Sehgal and Vibhuti Vasishth. Past price changes, trading volume and prediction of portfolio returns: Evidence from select emerging markets. *Journal of Advances in Management Research*, 2015.

[237] Yaohu Lin, Shancun Liu, Haijun Yang, and Harris Wu. Stock trend prediction using candlestick charting and ensemble machine learning techniques with a novelty feature engineering scheme. *IEEE Access*, 9:101433–101446, 2021.

[238] Bin Gao and Chunpeng Yang. Forecasting stock index futures returns with mixed-frequency sentiment. *International Review of Economics & Finance*, 49:69–83, 2017.

[239] Michael Melvin, John Prins, and Duncan Shand. Forecasting exchange rates: An investor perspective. In *Handbook of economic forecasting*, volume 2, pages 721–750. Elsevier, 2013.

[240] Kai-Yin Woo, Chulin Mai, Michael McAleer, and Wing-Keung Wong. Review on efficiency and anomalies in stock markets. *Economies*, 8(1):20, 2020.

[241] Chiranjit Dutta, Kara Karpman, Sumanta Basu, and Nalini Ravishanker. Review of statistical approaches for modeling high-frequency trading data. *Sankhya B*, pages 1–48, 2022.

[242] Gianluca Piero Maria Virgilio. High-frequency trading: a literature review. *Financial markets and portfolio management*, 33(2):183–208, 2019.

[243] Maureen O'hara. High frequency market microstructure. *Journal of financial economics*, 116(2):257–270, 2015.

[244] Lucimar Antônio Cabral de Ávila, Alanna Santos de Oliveira, Jéssica Rayse de Melo Silva Ávila, and Rodrigo Fernandes Malaquias. Behavioral biases in investors' decision: studies review from 2006-2015. 2016.

[245] Qingyuan Han and Steve Keen. Aggregate excess demand on wall street. *Heliyon*, 7(11):e08355, 2021.

[246] Marcus O'Connor, William Remus, and Ken Griggs. Judgemental forecasting in times of change. *International Journal of Forecasting*, 9(2):163–172, 1993.

[247] Robert C Blattberg and Stephen J Hoch. Database models and managerial intuition: 50% model+ 50% manager. *Management science*, 36(8):887–899, 1990.

[248] Oliver Schaer, Simon Spavound, et al. Review of forewarned: A sceptic's guide to prediction. *Foresight: The International Journal of Applied Forecasting*, (48):17–18, 2018.

[249] Paul Goodwin. *Forewarned: A Sceptic's Guide to Prediction*. Biteback Publishing, 2017.

[250] Robert Fildes and Fotios Petropoulos. Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting*, 36:5–12, 2015.

[251] Vicki G Morwitz, Joel H Steckel, and Alok Gupta. When do purchase intentions predict sales? *International Journal of Forecasting*, 23(3):347–364, 2007.

[252] Jon Scott Armstrong. *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer, 2001.

[253] Theodoros Evgeniou, Lily Fang, Robin M Hogarth, and Natalia Karelaia. Competitive dynamics in forecasting: The interaction of skill and uncertainty. *Journal of Behavioral Decision Making*, 26(4):375–384, 2013.

[254] Chip Heath and Amos Tversky. Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of risk and uncertainty*, 4(1):5–28, 1991.

[255] John W Atkinson. Motivational determinants of risk-taking behavior. *Psychological review*, 64(6p1):359, 1957.

[256] Roopesh Ranjan and Tilmann Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91, 2010.

[257] Allan H Murphy and Robert L Winkler. Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7(4):435–455, 1992.

[258] Meysam Arvan, Behnam Fahimnia, Mohsen Reisi, and Enno Siemsen. Integrating human judgement into quantitative forecasting methods: A review. *Omega*, 86:237–252, 2019.

[259] J Scott Armstrong and Kesten C Green. Forecasting methods and principles: Evidence-based checklists. *Journal of Global Scholars of Marketing Science*, 28(2):103–159, 2018.

[260] Dick van Dijk and Philip Hans Franses. Combining expert-adjusted forecasts. *Journal of Forecasting*, 38(5):415–421, 2019.

[261] Vera Shanshan Lin. Improving forecasting accuracy by combining statistical and judgmental forecasts in tourism. *Journal of China Tourism Research*, 9(3):325–352, 2013.

[262] Philippe Baecke, Shari De Baets, and Karlien Vanderheyden. Investigating the added value of integrating human judgement into statistical demand forecasting systems. *International Journal of Production Economics*, 191:85–96, 2017.

[263] Teemu Seeve and Eeva Vilkkumaa. Identifying and visualizing a diverse set of plausible scenarios for strategic planning. *European Journal of Operational Research*, 298(2):596–610, 2022.

[264] Paul Goodwin. Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, 30(2):127–135, 2002.

[265] Ahti A Salo and Derek W Bunn. Decomposition in the assessment of judgmental probability forecasts. *Technological Forecasting and Social Change*, 49(1):13–25, 1995.

[266] Trilok Nath Pandey, Alok Kumar Jagadev, Satchidananda Dehuri, and Sung-Bae Cho. A review and empirical analysis of neural networks based exchange rate prediction. *Intelligent Decision Technologies*, 12(4):423–439, 2018.

[267] Wei Huang, Kin Keung Lai, Yoshiteru Nakamori, Shouyang Wang, and Lean Yu. Neural networks in finance and economics forecasting. *International Journal of Information Technology & Decision Making*, 6(01):113–140, 2007.

[268] Po-Hsuan Hsu, Mark P Taylor, and Zigan Wang. Technical trading: Is it still beating the foreign exchange market? *Journal of International Economics*, 102:188–208, 2016.

[269] Lorna Katusiime, Abul Shamsuddin, and Frank W Agbola. Foreign exchange market efficiency and profitability of trading rules: Evidence from a developing country. *International Review of Economics & Finance*, 35:315–332, 2015.

[270] Camillo Lento and Nikola Gradojevic. The profitability of technical analysis during the covid-19 market meltdown. *Journal of Risk and Financial Management*, 15(5):192, 2022.

[271] Mary E Thomson, Andrew C Pollock, Dilek Önkal, and M Sinan Gönül. Combining forecasts: Performance and coherence. *International Journal of Forecasting*, 35(2):474–484, 2019.

[272] Aaron L Bodoh-Creed. Mood, memory, and the evaluation of asset prices. *Review of Finance*, 24(1):227–262, 2020.

[273] Elia Yathie Matsumoto, Emilio Del-Moral-Hernandez, Claudia Emiko Yoshinaga, and Afonso de Campos Pinto. Forecasting us dollar exchange rate movement with computational models and human behavior. *Expert Systems with Applications*, 194:116521, 2022.

[274] Thomas Lux and Michele Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500, 1999.

[275] Ayaz Hussain Bukhari, Muhammad Asif Zahoor Raja, Muhammad Sulaiman, Saeed Islam, Muhammad Shoaib, and Poom Kumam. Fractional neuro-sequential arfima-lstm for financial market forecasting. *Ieee Access*, 8:71326–71338, 2020.

[276] Dennys CA Mallqui and Ricardo AS Fernandes. Predicting the direction, maximum, minimum and closing prices of daily bitcoin exchange rate using machine learning techniques. *Applied Soft Computing*, 75:596–606, 2019.

[277] Adrián Carro, Raúl Toral, and Maxi San Miguel. Markets, herding and response to external information. *PloS one*, 10(7):e0133287, 2015.

[278] Nijolė Maknickienė, Indrė Lapinskaitė, and Algirdas Maknickas. Application of ensemble of recurrent neural networks for forecasting of stock market sentiments. *Equilibrium-quarterly journal of economics and economic policy*, 13(1):7–27, 2018.

[279] Frederik Kunze. Predicting exchange rates in asia: New insights on the accuracy of survey forecasts. *Journal of Forecasting*, 39(2):313–333, 2020.

[280] Patricia C O'brien. Analysts' forecasts as earnings expectations. *Journal of accounting and Economics*, 10(1):53–83, 1988.

[281] Satish K Mittal. Behavior biases and investment decision: theoretical and research framework. *Qualitative Research in Financial Markets*, 2019.

[282] Robert Fildes and Spyros Makridakis. The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review/Revue Internationale de Statistique*, pages 289–308, 1995.

[283] Daniel G Goldstein and Gerd Gigerenzer. Fast and frugal forecasting. *International journal of forecasting*, 25(4):760–772, 2009.

[284] Fotios Petropoulos and Yael Grushka-Cockayne. Fast and frugal time series forecasting. *arXiv preprint arXiv:2102.13209*, 2021.

[285] Daphne Sobolev. The effect of price volatility on judgmental forecasts: The correlated response model. *International Journal of Forecasting*, 33(3):605–617, 2017.

[286] Andrew C Pollock and Mary E Wilkie. Directional judgemental financial forecasting: trends and random walks. In *Modelling reality and personal modelling*, pages 253–271. Springer, 1993.

[287] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

[288] Dale Griffin, Richard Gonzalez, and Carol Varey. The heuristics and biases approach to judgment under uncertainty. *Blackwell handbook of social psychology: Intraindividual processes*, 1:207–235, 2001.

[289] Yuhui Wang, Shenghua Luan, and Gerd Gigerenzer. Modeling fast-and-frugal heuristics. *PsyCh Journal*, 11(4):600–611, 2022.

[290] Sagar P Kothari, Eric So, and Rodrigo Verdi. Analysts' forecasts and asset pricing. *Annual Review of Financial Economics*, 8:197–219, 2016.

[291] Natalia Karelaia and Robin M Hogarth. Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological bulletin*, 134(3):404, 2008.

[292] Rongjun Yu. Stress potentiates decision biases: A stress induced deliberation-to-intuition (sidi) model. *Neurobiology of stress*, 3:83–95, 2016.

[293] Sophie Cockcroft and Mark Russell. Big data opportunities for accounting and finance practice and research. *Australian Accounting Review*, 28(3):323–333, 2018.

[294] Minjian Ye and Guangzhong Li. Internet big data and capital markets: a literature review. *Financial Innovation*, 3(1):1–18, 2017.

[295] Ting-Ting Chen, Bo Zheng, Yan Li, and Xiong-Fei Jiang. New approaches in agent-based modeling of complex financial systems. *Frontiers of Physics*, 12(6):1–12, 2017.

[296] Antonio Díaz and Carlos Esparcia. Assessing risk aversion from the investor's point of view. *Frontiers in psychology*, 10:1490, 2019.

[297] Norbert Marwan. How to avoid potential pitfalls in recurrence plot based data analysis. *International Journal of Bifurcation and Chaos*, 21(04):1003–1017, 2011.

[298] Maqsood Ahmad. The role of cognitive heuristic-driven biases in investment management activities and market efficiency: a research synthesis. *International Journal of Emerging Markets*, (ahead-of-print), 2022.

[299] Eric C So. A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics*, 108(3):615–640, 2013.

[300] Falk Lieder, Thomas L Griffiths, Quentin J M Huys, and Noah D Goodman. The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, 25(1):322–349, 2018.

[301] Oscar Bustos and Alexandra Pomares-Quimbaya. Stock market movement forecast: A systematic review. *Expert Systems with Applications*, 156:113464, 2020.

[302] Ezra W Zuckerman. Structural incoherence and stock market activity. *American Sociological Review*, 69(3):405–432, 2004.

[303] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene Stanley. Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121(2):461–504, 2006.

[304] David H Romer. Rational asset price movements without news, 1992.

[305] Ken T Trotman, Hwee C Tan, and Nicole Ang. Fifty-year overview of judgment and decision-making research in accounting. *Accounting & Finance*, 51(1):278–360, 2011.

[306] Cornelius Casey and Thomas I Selling. The effect of task predictability and prior probability disclosure on judgment quality and confidence. *Accounting Review*, pages 302–317, 1986.

[307] Ray Pfeiffer, Karen Teitel, Susan Wahab, and Mahmoud Wahab. Identifying the news in analysts' earnings forecasts revisions: An alternative to the random walk expectation. *Review of Pacific Basin Financial Markets and Policies*, 24(04):2150032, 2021.

[308] Roy Batchelor and Tai Yeong Kwan. Judgemental bootstrapping of technical traders in the bond market. *International Journal of Forecasting*, 23(3):427–445, 2007.

[309] Mary E Thomson, Andrew C Pollock, Karen B Henriksen, and Alex Macaulay. The influence of the forecast horizon on judgemental probability forecasts of exchange rate movements. *The European Journal of Finance*, 10(4):290–307, 2004.

[310] Bertrand Jacquillat and Pascal Grandin. Performance measurement of analysts' forecasts. *Journal of Portfolio Management*, 21(1):94, 1994.

[311] Sundaresh Ramnath, Steve Rock, and Philip Shane. The financial analyst forecasting literature: A taxonomy with suggestions for further research. *International Journal of Forecasting*, 24(1):34–75, 2008.

[312] Manish Kumar and M Thenmozhi. Forecasting stock index movement: A comparison of support vector machines and random forest. In *Indian institute of capital markets 9th capital markets conference paper*, 2006.

[313] Pedro N Rodriguez and Arnulfo Rodriguez. Predicting stock market indices movements. *WIT Transactions on Modelling and Simulation*, 38, 2004.

[314] Xinjie Di. Stock trend prediction with technical indicators using svm. *Standford: Leland Stanford Junior University*, 2014.

[315] Jinji Hao and Jonathon Skinner. Analyst target price and dividend forecasts and expected stock returns. *Journal of Asset Management*, pages 1–13, 2022.

[316] Zhi Da, Keejae P Hong, and Sangwoo Lee. What drives target price forecasts and their investment value? *Journal of Business Finance & Accounting*, 43(3-4):487–510, 2016.

[317] Fernando GDC Ferreira, Amir H Gandomi, and Rodrigo TN Cardoso. Artificial intelligence applied to stock market trading: a review. *IEEE Access*, 9:30898–30917, 2021.

[318] Bhaskar Tripathi and Rakesh Kumar Sharma. Modeling bitcoin prices using signal processing methods, bayesian optimization, and deep neural networks. *Computational Economics*, pages 1–27, 2022.

[319] Ilia Zaznov, Julian Kunkel, Alfonso Dufour, and Atta Badii. Predicting stock price changes based on the limit order book: a survey. *Mathematics*, 10(8):1234, 2022.

[320] Maciej Wujec. Analysis of the financial information contained in the texts of current reports: A deep learning approach. *Journal of Risk and Financial Management*, 14(12):582, 2021.

[321] Saleh Albahli, Aun Irtaza, Tahira Nazir, Awais Mehmood, Ali Alkhalifah, and Waleed Albattah. A machine learning method for prediction of stock market using real-time twitter data. *Electronics*, 11(20):3414, 2022.

[322] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009.

[323] K Puneeth, Sagar Rudagi, M Namratha, Ranispoorti Patil, and Rohini Wadi. Comparative study: Stock prediction using fundamental and technical analysis. In *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, pages 1–4. IEEE, 2021.

[324] Marco Ortu, Nicola Uras, Claudio Conversano, Silvia Bartolucci, and Giuseppe Destefanis. On technical trading and social media indicators for cryptocurrency price classification through deep learning. *Expert Systems with Applications*, 198:116804, 2022.

[325] Asit Kumar Das, Debahuti Mishra, Kaberi Das, Arup Kumar Mohanty, Mazin Abed Mohammed, Alaa S Al-Waisy, Seifedine Kadry, and Jungeun Kim. A deep network-based trade and trend analysis system to observe entry and exit points in the forex market. *Mathematics*, 10(19):3632, 2022.

[326] Swaty Dash, Pradip Kumar Sahu, Debahuti Mishra, Pradeep Kumar Mallick, Bharti Sharma, Mikhail Zymbler, and Sachin Kumar. A novel algorithmic forex trade and trend analysis framework based on deep predictive coding network optimized with reptile search algorithm. *Axioms*, 11(8):396, 2022.

[327] Dushmanta Kumar Padhi, Neelamadhab Padhy, Akash Kumar Bhoi, Jana Shafi, and Seid Hassen Yesuf. An intelligent fusion model with portfolio selection and machine learning for stock market prediction. *Computational Intelligence and Neuroscience*, 2022, 2022.

[328] Faruk Ozer and C Okan Sakar. An automated cryptocurrency trading system based on the detection of unusual price movements with a time-series clustering-based approach. *Expert Systems with Applications*, 200:117017, 2022.

[329] Alireza Sadeghi, Amir Daneshvar, and Mahdi Madanchi Zaj. Combined ensemble multi-class svm and fuzzy nsga-ii for trend forecasting and trading in forex markets. *Expert Systems with Applications*, 185:115566, 2021.

[330] Nicola Uras and Marco Ortu. Investigation of blockchain cryptocurrencies' price movements through deep learning: A comparative analysis. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 715–722. IEEE, 2021.

[331] Ernest Kwame Ampomah, Zhiguang Qin, and Gabriel Nyame. Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information*, 11(6):332, 2020.

[332] Do-Hyung Kwon, Ju-Bong Kim, Ju-Sung Heo, Chan-Myung Kim, and Youn-Hee Han. Time series classification of cryptocurrency price trend based on a recurrent lstm neural network. *Journal of Information Processing Systems*, 15(3):694–706, 2019.

[333] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2):654–669, 2018.

[334] Hongping Hu, Li Tang, Shuhua Zhang, and Haiyan Wang. Predicting the direction of stock markets using optimized neural networks with google trends. *Neurocomputing*, 285:188–195, 2018.

[335] Oscar Bustos, Alexandra Pomares, and Enrique Gonzalez. A comparison between svm and multilayer perceptron in predicting an emerging financial market: Colombian stock market. In *2017 Congreso Internacional de Innovacion y Tendencias en Ingenieria (CONIITI)*, pages 1–6. IEEE, 2017.

[336] Pranjal Chakraborty, Ummay Sani Pria, Md Rashad Al Hasan Rony, and Mahbub Alam Majumdar. Predicting stock movement using sentiment analysis of twitter feed. In *2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT)*, pages 1–6. IEEE, 2017.

[337] Scott Coyne, Praveen Madiraju, and Joseph Coelho. Forecasting stock prices using social media analysis. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 1031–1038. IEEE, 2017.

[338] Alexiei Dingli and Karl Sant Fournier. Financial time series forecasting-a deep learning approach. *International Journal of Machine Learning and Computing*, 7(5):118–122, 2017.

[339] Chien-Feng Huang and Hsu-Chih Li. An evolutionary method for financial forecasting in microscopic high-speed trading environment. *Computational Intelligence and Neuroscience*, 2017, 2017.

[340] Minh Dang and Duc Duong. Improvement methods for stock market prediction using financial news articles. In *2016 3rd National foundation for science and technology development conference on information and computer science (NICS)*, pages 125–129. IEEE, 2016.

[341] Luca Di Persio and Oleksandr Honchar. Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International journal of circuits, systems and signal processing*, 10(2016):403–413, 2016.

[342] Mojgan Ghanavati, Raymond K Wong, Fang Chen, Yang Wang, and Simon Fong. A generic service framework for stock market prediction. In *2016 IEEE International Conference on Services Computing (SCC)*, pages 283–290. IEEE, 2016.

[343] Jian Chai, Jiangze Du, Kin Keung Lai, and Yan Pui Lee. A hybrid least square support vector machine model with parameters optimization for stock forecasting. *Mathematical Problems in Engineering*, 2015, 2015.

[344] Rajashree Dash and PK Dash. A comparative study of radial basis function network with different basis functions for stock trend prediction. In *2015 IEEE Power, Communication and Information Technology Conference (PCITC)*, pages 430–435. IEEE, 2015.

[345] Stefan Feuerriegel and Ralph Fehrer. Improving decision analytics with deep learning: the case of financial disclosures. *arXiv preprint arXiv:1508.01993*, 2015.

[346] Rafael Thomazi Gonzalez, Carlos Alberto Padilha, and Dante Augusto Couto Barone. Ensemble system based on genetic algorithm for stock market forecasting. In *2015 IEEE congress on evolutionary computation (CEC)*, pages 3102–3108. IEEE, 2015.

[347] Akhil Sethia and Purva Raut. Application of lstm, gru and ica for stock price prediction. In *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 2*, pages 479–487. Springer, 2019.

[348] Mária Bohdalová and Michal Greguš. Fractional brownian motion in ohlc crude oil prices. In *Advances in Time Series Analysis and Forecasting: Selected Contributions from ITISE 2016 3*, pages 77–87. Springer, 2017.

[349] Yauheniya Shynkevich, T Martin McGinnity, Sonya A Coleman, Ammar Belatreche, and Yuhua Li. Forecasting price movements using technical indicators: Investigating the impact of varying input window length. *Neurocomputing*, 264:71–88, 2017.

[350] Seçkin Karasu and Aytaç Altan. Crude oil time series prediction model based on lstm network with chaotic henry gas solubility optimization. *Energy*, 242:122964, 2022.

[351] D Th Vezeris, Christos J Schinas, and Garyfalos Papaschinopoulos. Profitability edge by dynamic back testing optimal period selection for technical parameters optimization, in trading systems with forecasting: The d-backtest ps method. *Computational Economics*, 51:761–807, 2018.

[352] Ranjeeta Bisoi and Pradipta K Dash. A hybrid evolutionary dynamic neural network for stock market trend analysis and prediction using unscented kalman filter. *Applied Soft Computing*, 19:41–56, 2014.

[353] Alexandre Martins Carvalho, Flavio Barboza, and José Augusto Fiorucci. Can an automated trading algorithm based on a graphical analysis present a good result? *Revista Eniac Pesquisa*, 9(1):129–150, 2020.

[354] Seyma Caliskan Cavdar and Alev Dilek Aydin. Hybrid model approach to the complexity of stock trading decisions in turkey. *The Journal of Asian Finance, Economics and Business*, 7(10):9–21, 2020.

[355] Ranjeeta Bisoi, PK Dash, and AK Parida. Hybrid variational mode decomposition and evolutionary robust kernel extreme learning machine for stock price and movement prediction on daily basis. *Applied Soft Computing*, 74:652–678, 2019.

[356] Faheem Aslam, Khurrum S Mughal, Ashiq Ali, and Yasir Tariq Mohmand. Forecasting islamic securities index using artificial neural networks: performance evaluation of technical indicators. *Journal of Economic and Administrative Sciences*, 37(2):253–271, 2021.

[357] Chandrasekar Ravi. Fuzzy crow search algorithm-based deep lstm for bitcoin prediction. *International Journal of Distributed Systems and Technologies (IJDST)*, 11(4):53–71, 2020.

[358] Chang Li, Dongjin Song, and Dacheng Tao. Multi-task recurrent neural networks and higher-order markov random fields for stock price movement prediction: Multi-task rnn and higer-order mrfs for stock price classification. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1141–1151, 2019.

[359] Ikhlaas Gurrib and Elgilani Elshareif. Optimizing the performance of the fractal adaptive moving average strategy: The case of eur/usd. *International Journal of Economics and Finance*, 8(2):171–178, 2016.

[360] Shangkun Deng, Chongyi Xiao, Yingke Zhu, Jingyuan Peng, Jie Li, and Zonghua Liu. High-frequency direction forecasting and simulation trading of the crude oil futures using ichimoku kinkohyo and fuzzy rough set. *Expert Systems with Applications*, 215:119326, 2023.

[361] Rubén Arévalo, Jorge García, Francisco Guijarro, and Alfred Peris. A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. *Expert Systems with Applications*, 81:177–192, 2017.

[362] Murat Ozturk, Ismail Hakki Toroslu, and Guven Fidan. Heuristic based trading system on forex data using technical indicator rules. *Applied Soft Computing*, 43:170–186, 2016.

[363] Aleksandar Rakićević, Radovan Končarević, and Bratislav Petrović. Comparison of moving averages for trading trends: the case of the belgrade stock exchange. In *Proceedings of the 14th International Symposium SymOrg*, pages 6–10, 2014.

[364] João Carapuço, Rui Neves, and Nuno Horta. Reinforcement learning applied to forex trading. *Applied Soft Computing*, 73:783–794, 2018.

[365] Orawan Chantarakasemchit and Siranee Nuchitprasitchai. Enhancing forex rates prediction with machine learning on eur to usd with moving average methods and financial factors. In *Recent Advances in Information and Communication Technology 2021: Proceedings of the 17th International Conference on Computing and Information Technology (IC2IT 2021)*, pages 44–54. Springer, 2021.

[366] Vinicius Amorim Sobreiro, Thiago Raymon Cruz Cacique da Costa, Rodolfo Toríbio Farias Nazário, Jéssica Lima e Silva, Eduardo Alves Moreira, Marcius Correia Lima Filho, Herbert Kimura, and Juan Carlos Arismendi Zambrano. The profitability of moving average trading rules in brics and emerging stock markets. *The North American Journal of Economics and Finance*, 38:86–101, 2016.

[367] Valeriy Zakamulin and Javier Giner. Trend following with momentum versus moving averages: A tale of differences. *Quantitative Finance*, 20(6):985–1007, 2020.

[368] Przemysław Juszczuk and Jan Kozak. Classification and preprocessing in the stock data. In *Business Information Systems Workshops: BIS 2017 International Workshops, Poznań, Poland, June 28-30, 2017, Revised Papers 20*, pages 269–281. Springer, 2017.

[369] Przemysław Juszczuk and Lech Kruś. Selecting the efficient market indicators in the trading system on the forex market. In *Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology–ISAT 2018: Part III*, pages 122–133. Springer, 2019.

[370] Nguyen Trong Khanh and Nguyen Thi Ngoc Anh. Hybrid classifier by integrating sentiment and technical indicator classifiers. *Context-Aware Systems and Applications, and Nature of Computation and Communication*, page 25, 2018.

[371] Watthana Pongsena, Prakaidoy Sitsayabut, Nittaya Kerdprasop, and Kittisak Kerdprasop. Development of a model for predicting the direction of daily price changes

in the forex market using long short-term memory. *International Journal of Machine Learning and Computing*, 11(1):61–67, 2021.

[372] N Nurrimah and Z Rustam. Stock price trend prediction method based on support vector machines with fisher score. In *AIP Conference Proceedings*, page 030014. AIP Publishing LLC, 2020.

[373] Nima Zarrabi, Stuart Snaith, and Jerry Coakley. Exchange rate forecasting using economic models and technical trading rules. *The European Journal of Finance*, 28(10):997–1018, 2022.

[374] Firuz Kamalov. Forecasting significant stock price changes using neural networks. *Neural Computing and Applications*, 32:17655–17667, 2020.

[375] Widya Hardiyanti. Risk value analysis of gold futures trading investment using fundamental analysis, technical analysis, and value at risk. *Journal of Research in Mathematics Trends and Technology*, 3(1):8–19, 2021.

[376] Furong Ye, Liming Zhang, Defu Zhang, Hamido Fujita, and Zhiguo Gong. A novel forecasting method based on multi-order fuzzy time series and technical analysis. *Information Sciences*, 367:41–57, 2016.

[377] SZ Mahfooz, Iftikhar Ali, and Muhammad N Khan. Improving stock trend prediction using lstm neural network trained on a complex trading strategy. *International Journal for Research in Applied Science and Engineering Technology*, 10(7):4361–4371, 2022.

[378] Fagner A De Oliveira, Cristiane N Nobre, and Luis E Zárate. Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index–case study of petr4, petrobras, brazil. *Expert systems with applications*, 40(18):7596–7606, 2013.

[379] Adam Fadlalla and Farzaneh Amani. Predicting next trading day closing price of qatar exchange index using technical indicators and artificial neural networks. *Intelligent Systems in Accounting, Finance and Management*, 21(4):209–223, 2014.

[380] Rajendran Sugumar, Alwar Rengarajan, and Chinnappan Jayakumar. A technique to stock market prediction using fuzzy clustering and artificial neural networks. *Computing and Informatics*, 33(5):992–1024, 2014.

[381] Axel Groß-Klußmann and Nikolaus Hautsch. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, 18(2):321–340, 2011.

[382] Don K Mak and Don K Mak. Analysis of the trading tactics. *Trading Tactics in the Financial Market: Mathematical Methods to Improve Performance*, pages 119–124, 2021.

[383] M Ananthi and K Vijayakumar. Retracted article: stock market analysis using candlestick regression and market trend prediction (ckrm). *Journal of Ambient Intelligence and Humanized Computing*, 12(5):4819–4826, 2021.

[384] Vasu Kalariya, Pushpendra Parmar, Patel Jay, Sudeep Tanwar, Maria Simona Raboaca, Fayez Alqahtani, Amr Tolba, and Bogdan-Constantin Neagu. Stochastic neural networks-based algorithmic trading for the cryptocurrency market. *Mathematics*, 10(9):1456, 2022.

[385] Hong-Yong Wang, Hong Li, and Jin-Ye Shen. A novel hybrid fractal interpolation-svm model for forecasting stock price indexes. *Fractals*, 27(04):1950055, 2019.

[386] Raymond ST Lee. Cosmos trader–chaotic neuro-oscillatory multiagent financial prediction and trading system. *The Journal of Finance and Data Science*, 5(2):61–82, 2019.

[387] Pavan Kumar Illa, Balakesavareddy Parvathala, and Anand Kumar Sharma. Stock price prediction methodology using random forest algorithm and support vector machine. *Materials Today: Proceedings*, 56:1776–1782, 2022.

[388] Maedeh Tajmazinani, Hossein Hassani, Reza Raei, Saeed Rouhani, et al. Modeling stock price movements prediction based on news sentiment analysis and deep learning. *Annals of Financial Economics (AFE)*, 17(01):1–19, 2022.

[389] Lakshmana Phaneendra Maguluri and R Ragupathy. A cluster based non-linear regression framework for periodic multi-stock trend prediction on real time stock market data. *International Journal of Advanced Computer Science and Applications*, 11(9), 2020.

[390] Lewen Wang and Zuoxian Ye. An interpretable framework for stock trend forecasting. In *Journal of Physics: Conference Series*, volume 1634, page 012026. IOP Publishing, 2020.

[391] Xinyi Guo and Jinfeng Li. A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 472–477. IEEE, 2019.

[392] Guilherme A Bileki, Flávio Barboza, Luiz Henrique C Silva, and Vanderlei Bonato. Order book mid-price movement inference by catboost classifier from convolutional feature maps. *Applied Soft Computing*, 116:108274, 2022.

[393] Nusrat Rouf, Majid Bashir Malik, and Tasleem Arif. Predicting the stock market trend: An ensemble approach using impactful exploratory data analysis. In *International Conference on Information, Communication and Computing Technology*, pages 223–234. Springer, 2021.

[394] Munish Khanna, Mohak Kulshrestha, Law K Singh, Shankar Thawkar, and Kapil Shrivastava. Performance evaluation of machine learning algorithms for stock price and stock index movement prediction using trend deterministic data prediction. *International Journal of Applied Metaheuristic Computing (IJAMC)*, 13(1):1–30, 2022.

[395] Dang Lien Minh, Abolghasem Sadeghi-Niaraki, Huynh Duc Huy, Kyungbok Min, and Hyeonjoon Moon. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *Ieee Access*, 6:55392–55404, 2018.

[396] Pengyue Wang, Xuesheng Li, Zhiliang Qin, Yuanyuan Qu, and Zhongkai Zhang. Stock price forecasting based on wavelet filtering and ensembled machine learning model. *Mathematical Problems in Engineering*, 2022, 2022.

[397] Sourav Malakar, Saptarsi Goswami, Amlan Chakrabarti, and Basabi Chakraborty. A hybrid and adaptive approach for classification of indian stock market-related tweets. In *Data Management, Analytics and Innovation*, pages 325–342. Springer, 2020.

[398] Mehdi Khashei and Bahareh Mahdavi Sharif. A kalman filter-based hybridization model of statistical and intelligent approaches for exchange rate forecasting. *Journal of Modelling in Management*, 2020.

[399] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, 2019.

[400] Ernest Kwame Ampomah, Gabriel Nyame, Zhiguang Qin, Prince Clement Addo, Enoch Opanin Gyamfi, and Micheal Gyan. Stock market prediction with gaussian naïve bayes machine learning algorithm. *Informatica*, 45(2), 2021.

[401] Yajie Zhang and Sijie Lu. Multi-model fusion method and its application in prediction of stock index movements. In *2021 6th International Conference on Machine Learning Technologies*, pages 58–64, 2021.

[402] Xiao-dan Zhang, Ang Li, and Ran Pan. Stock trend prediction based on a new status box method and adaboost probabilistic support vector machine. *Applied Soft Computing*, 49:385–398, 2016.

[403] R Halliday. Equity trend prediction with neural networks. 2004.

[404] Dingming Wu, Xiaolong Wang, Jingyong Su, Buzhou Tang, and Shaocong Wu. A labeling method for financial time series prediction based on trends. *Entropy*, 22(10):1162, 2020.

[405] José Antonio Sanz, Dario Bernardo, Francisco Herrera, Humberto Bustince, and Hani Hagras. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *IEEE Transactions on Fuzzy Systems*, 23(4):973–990, 2014.

[406] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239, 2017.

[407] Varun Dogra, Sahil Verma, Kavita Verma, NZ Jhanjhi, Uttam Ghosh, and Dac-Nhuong Le. A comparative analysis of machine learning models for banking news extraction

by multiclass classification with imbalanced datasets of financial news: Challenges and solutions. 2022.

[408] Zhen Shao, Qingru Zheng, Shanlin Yang, Fei Gao, Manli Cheng, Qiang Zhang, and Chen Liu. Modeling and forecasting the electricity clearing price: A novel belm based pattern classification framework and a comparative analytic study on multi-layer belm and lstm. *Energy Economics*, 86:104648, 2020.

[409] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 475–482. Springer, 2009.

[410] Sukarna Barua, Md Monirul Islam, Xin Yao, and Kazuyuki Murase. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, 26(2):405–425, 2012.

[411] Florin Gorunescu. Data mining techniques and models. In *Data Mining*, pages 185–317. Springer, 2011.

[412] Hakan Gunduz and Zehra Cataltepe. Borsa istanbul (bist) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications*, 42(22):9001–9011, 2015.

[413] Jiajia Li and Phayung Meesad. Combining sentiment analysis with socialization bias in social networks for stock market trend prediction. *International Journal of Computational Intelligence and Applications*, 15(01):1650003, 2016.

[414] Raghavendra Reddy and Gopal K Shyam. Market data analysis by using support vector machine learning technique. In *Proceedings of International Conference on Computational Intelligence and Data Engineering*, pages 19–27. Springer, 2019.

[415] Narendhar Gugulothu, Easwar Subramanian, and Sanjay P Bhat. Sparse recurrent mixture density networks for forecasting high variability time series with confidence estimates. In *International Conference on Artificial Neural Networks*, pages 422–433. Springer, 2019.

[416] A Wijitcharoen, B Watanapa, P Padungweang, and W Anantasabkit. Ica-deoda: An independent feature extraction model for stock index forecasting. In *2016 International Computer Science and Engineering Conference (ICSEC)*, pages 1–4. IEEE, 2016.

[417] A.L. Yuille, F.E. Ruiz, P.S. Pérez, and B.I. Bonev. *Information Theory in Computer Vision and Pattern Recognition*. Springer London, 2009.

[418] H. Zhang, Y. Chen, X. Chu, Z. Zhang, T. Hao, Z. Wu, and Y. Yang. *Neural Computing for Advanced Applications: Third International Conference, NCAA 2022, Jinan, China, July 8–10, 2022, Proceedings, Part I*. Communications in Computer and Information Science. Springer Nature Singapore, 2022.

[419] Adamantios Ntakaris, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Mid-price prediction based on machine learning methods with technical and quantitative indicators. *Plos one*, 15(6):e0234107, 2020.

[420] Anwar Ul Haq, Adnan Zeb, Zhenfeng Lei, and Defu Zhang. Forecasting daily stock trend using multi-filter feature selection and deep learning. *Expert Systems with Applications*, 168:114444, 2021.

[421] H. R. Syahputra and Z. A. Husodo. Predicting the trend of indonesian stock price movements using discriminant analysis and support vector machine. *Proceedings of the International Conference of Applied Business and Management (Icabm2020)*, pages 684–707, 2020. Syahputra, Hanandi Rahmad Husodo, Zaafri Ananto Borges, AP Vieira.

[422] Mahinda Mailagaha Kumbure, Christoph Lohrmann, and Pasi Luukka. A study on relevant features for intraday s&p 500 prediction using a hybrid feature selection approach. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 93–104. Springer, 2021.

[423] Yang Liu, Qingguo Zeng, Huanrui Yang, and Adrian Carrio. Stock price movement prediction from financial news with deep learning and knowledge graph embedding. In *Pacific rim knowledge acquisition workshop*, pages 102–113. Springer, 2018.

[424] Lu-Tao Zhao, Shen Miao, Jing aQu, and Xue-Hui Chen. A multi-factor integrated model for carbon price forecasting: market interaction promoting carbon emission reduction. *Science of The Total Environment*, 796:149110, 2021.

[425] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[426] Hakan Gündüz, Zehra Çataltepe, and Yusuf Yaslan. Stock daily return prediction using expanded features and feature selection. *Turkish Journal of Electrical Engineering and Computer Sciences*, 25(6):4829–4840, 2017.

[427] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier, 1992.

[428] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[429] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. *Using multivariate statistics*, volume 5. pearson Boston, MA, 2007.

[430] Junwen Yang, Yunmin Wang, and Xiang Li. Prediction of stock price direction using the lasso-lstm model combines technical indicators and financial sentiment analysis. *PeerJ Computer Science*, 8:e1148, 2022.

[431] Prakash K Aithal, Acharya U Dinesh, and M Geetha. Identifying significant macroeconomic indicators for indian stock markets. *IEEE Access*, 7:143829–143840, 2019.

[432] Yea-Win Wu. Configuring an improved backpropagation network for forecasting study of interest rate in traditional money market and derivative commodity market. In *Soft Computing in Intelligent Systems and Information Processing. Proceedings of the 1996 Asian Fuzzy Systems Symposium*, pages 521–526. IEEE, 1996.

[433] Dong-Hee Cho, Seung-Hyun Moon, and Yong-Hyuk Kim. Genetic feature selection applied to kospi and cryptocurrency price prediction. *Mathematics*, 9(20):2574, 2021.

[434] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.

[435] Dattatray P Gandhmal and K Kumar. Wrapper-enabled feature selection and cplm-based narx model for stock market prediction. *The Computer Journal*, 64(2):169–184, 2021.

[436] Tian Xia, Qibo Sun, Ao Zhou, Shanguang Wang, Shilong Xiong, Siyi Gao, Jinglin Li, and Quan Yuan. Improving the performance of stock trend prediction by applying ga to feature selection. In *2018 IEEE 8th International Symposium on Cloud and Service Computing (SC2)*, pages 122–126. IEEE, 2018.

[437] Chenn-Jung Huang, Dian-Xiu Yang, and Yi-Ta Chuang. Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications*, 34(4):2870–2878, 2008.

[438] Norberto Ritzmann Júnior and Julio Cesar Nievola. A generalized financial time series forecasting model based on automatic feature engineering using genetic algorithms and support vector machine. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[439] F. Saeed, F. Mohammed, and N. Gazem. *Emerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing*. Advances in Intelligent Systems and Computing. Springer International Publishing, 2019.

[440] Z.A. Zhao and H. Liu. *Spectral Feature Selection for Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2011.

[441] M.S. Raza and U. Qamar. *Understanding and Using Rough Set Based Feature Selection: Concepts, Techniques and Applications*. Springer Singapore, 2017.

[442] Hoai Bach Nguyen, Bing Xue, and Peter Andreae. Mutual information estimation for filter based feature selection using particle swarm optimization. In *European Conference on the Applications of Evolutionary Computation*, pages 719–736. Springer, 2016.

[443] Gavin Brown. A new perspective for information theoretic feature selection. In *Artificial intelligence and statistics*, pages 49–56. PMLR, 2009.

[444] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

[445] K Fathima Bibi and M Nazreen Banu. Feature subset selection based on filter technique. In *2015 International Conference on Computing and Communications Technologies (ICCCT)*, pages 1–6. IEEE, 2015.

[446] Artur J Ferreira and Mário AT Figueiredo. Efficient feature selection filters for high-dimensional data. *Pattern recognition letters*, 33(13):1794–1804, 2012.

[447] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, 8(1):43–48, 2010.

[448] Kilho Shin and Seiya Miyazaki. A fast and accurate feature selection algorithm based on binary consistency measure. *Computational Intelligence*, 32(4):646–667, 2016.

[449] Antonio Arauzo-Azofra, Jose Manuel Benitez, and Juan Luis Castro. Consistency measures for feature selection. *Journal of Intelligent Information Systems*, 30(3):273–292, 2008.

[450] Manoranjan Dash, Huan Liu, and Hiroshi Motoda. Consistency based feature selection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 98–109. Springer, 2000.

[451] Subrat Kumar Nayak, Pravat Kumar Rout, Alok Kumar Jagadev, and Tripti Swarnkar. Elitism based multi-objective differential evolution for feature selection: A filter approach with an efficient redundancy measure. *Journal of King Saud University-Computer and Information Sciences*, 32(2):174–187, 2020.

[452] Salimeh Yasaei Sekeh and Alfred O Hero. Feature selection for mutlti-labeled variables via dependency maximization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3127–3131. IEEE, 2019.

[453] Christoph Lohrmann and Pasi Luukka. Nonspecificity, strife and total uncertainty in supervised feature selection. *Engineering Applications of Artificial Intelligence*, 109:104628, 2022.

[454] Lin Sun, Lanying Wang, Jiucheng Xu, and Shiguang Zhang. A neighborhood rough sets-based attribute reduction method using lebesgue and entropy measures. *Entropy*, 21(2):138, 2019.

[455] Sheng-yi Jiang and Lian-xi Wang. Efficient feature selection based on correlation measure between continuous and discrete features. *Information Processing Letters*, 116(2):203–215, 2016.

[456] Fadi Thabtah, Firuz Kamalov, Suhel Hammoud, and Seyed Reza Shahamiri. Least loss: A simplified filter method for feature selection. *Information Sciences*, 534:1–15, 2020.

[457] Lin Sun, Tianxiang Wang, Weiping Ding, Jiucheng Xu, and Yaojin Lin. Feature selection using fisher score and multilabel neighborhood rough sets for multilabel classification. *Information Sciences*, 578:887–912, 2021.

[458] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[459] Serkan Gunal and Rifat Edizkan. Subspace based feature selection for pattern recognition. *Information Sciences*, 178(19):3716–3726, 2008.

[460] Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205. Ieee, 2015.

[461] Kui Zhang, Yuhua Li, Philip Scarf, and Andrew Ball. Feature selection for high-dimensional machinery fault diagnosis data using multiple models and radial basis function networks. *Neurocomputing*, 74(17):2941–2952, 2011.

[462] Hoang TP Thanh and Phayung Meesad. Stock market trend prediction based on text mining of corporate web and time series data. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 18(1):22–31, 2014.

[463] Weerachart Lertyingyod and Nunnapus Benjamas. Stock price trend prediction using artificial neural network techniques: Case study: Thailand stock exchange. In *2016 International Computer Science and Engineering Conference (ICSEC)*, pages 1–6. IEEE, 2016.

[464] Ming-Chi Lee. Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, 36(8):10896–10904, 2009.

[465] Hani AK Ihlayyel, Nurfadhlina Mohd Sharef, Mohd Zakree Ahmed Nazri, et al. An enhanced feature representation based on linear regression model for stock market prediction. *Intelligent Data Analysis*, 22(1):45–76, 2018.

[466] S Foroozan, MA Azmi Murad, NM Sharef, and AR Abdul Latiff. Improving sentiment classification accuracy of financial news using n-gram approach and feature weighting methods. In *2015 2nd International Conference on Information Science and Security (ICISS)*, pages 1–4. IEEE, 2015.

[467] Sepideh Foroozan Yazdani, Masrah Azrifah Azmi Murad, Nurfadhlina Mohd Sharef, Yashwant Prasad Singh, and Ahmed Razman Abdul Latiff. Sentiment classification of financial news using statistical features. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(03):1750006, 2017.

[468] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1):306–324, 2015.

[469] Sui Xue-shen, Qi Zhong-ying, Yu Da-Ren, Hu Qing-hua, and Zhao Hui. A novel feature selection approach using classification complexity for svm of stock market trend prediction. In *2007 International Conference on Management Science and Engineering*, pages 1654–1659. IEEE, 2007.

[470] Masoomeh Rashidpoor Toochaei and Farzad Moeini. Evaluating the performance of ensemble classifiers in stock returns prediction using effective features. *Expert Systems with Applications*, 213:119186, 2023.

[471] Gang Ji, Jingmin Yu, Kai Hu, Jie Xie, and Xunsheng Ji. An adaptive feature selection schema using improved technical indicators for predicting stock price movements. *Expert Systems with Applications*, 200:116941, 2022.

[472] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[473] Isaac Kofi Nti, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. Efficient stock-market prediction using ensemble support vector machine. *Open Computer Science*, 10(1):153–163, 2020.

[474] Armin Mahmoodi, Leila Hashemi, Milad Jasemi, Soroush Mehraban, Jeremy Laliberté, and Richard C Millar. A developed stock price forecasting model using support vector machine combined with metaheuristic algorithms. *OPSEARCH*, pages 1–28, 2022.

[475] Kyung Keun Yun, Sang Won Yoon, and Daehan Won. Prediction of stock price direction using a hybrid ga-xgboost algorithm with a three-stage feature engineering process. *Expert Systems with Applications*, 186:115716, 2021.

[476] Mehreen Rehman, Gul Muhammad Khan, and Sahibzada Ali Mahmud. Foreign currency exchange rates prediction using cgp and recurrent neural network. *IERI Procedia*, 10:239–244, 2014.

[477] Shun Li, Da Shi, and Shaohua Tan. Financial data modeling using a hybrid bayesian network structured learning algorithm. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 6(1):48–71, 2012.

[478] Samant Saurabh and Kushankur Dey. Unraveling the relationship between social moods and the stock market: Evidence from the united kingdom. *Journal of Behavioral and Experimental Finance*, 26:100300, 2020.

[479] Ikhlaas Gurrib and Firuz Kamalov. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Studies in Economics and Finance*, 2021.

[480] Shuai Wang and Wei Shang. Forecasting direction of china security index 300 movement with least squares support vector machine. *Procedia Computer Science*, 31:869–874, 2014.

[481] Vidya Moni, Maheshwari Mattipalli, and Altaf QH Badar. Machine learning classification techniques to predict directional change of energy prices using high dimensionality

reduction. In *2022 International Conference on Computer Science and Software Engineering (CSASE)*, pages 247–252. IEEE, 2022.

[482] Frances B Shin and David H Kil. Classification cramer–rao bounds on stock price prediction. *Journal of Forecasting*, 17(5-6):389–399, 1998.

[483] Cecille Freeman, Dana Kulić, and Otman Basir. An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recognition*, 48(5):1812–1826, 2015.

[484] Nagwan Abdel Samee, Ghada Atteia, Reem Alkanhel, Amel Ali Alhussan, and Hussah Nasser AlEisa. Hybrid feature reduction using pcc-stacked autoencoders for gold/oil prices forecasting under covid-19 pandemic. *Electronics*, 11(7):991, 2022.

[485] Wenbo Ge, Pooia Lalbakhsh, Leigh Isai, Artem Lenskiy, and Hanna Suominen. Neural network–based financial volatility forecasting: A systematic review. *ACM Comput. Surv.*, 55(1), jan 2022.

[486] Ibrahim D Raheem. Global financial cycles and exchange rate forecast: A factor analysis. *Borsa istanbul review*, 20:S81–S92, 2020.

[487] O. Bustos and A. Pomares-Quimbaya. Stock market movement forecast: A systematic review. *Expert Systems with Applications*, 156, 2020.

[488] L. Kavtaradze and M. Mokhtari. Factor models and time-varying parameter framework for forecasting exchange rates and inflation: A survey. *Journal of Economic Surveys*, 32(2):302–334, 2018.

[489] Yifu Qiu, Yitao Qiu, Yicong Yuan, Zheng Chen, and Raymond Lee. Qf-tradernet: Intraday trading via deep reinforcement with quantum price levels based profit-and-loss control. *Frontiers in Artificial Intelligence*, 4, 2021.

[490] Deniz Can Yıldırım, Ismail Hakkı Toroslu, and Ugo Fiore. Forecasting directional movement of forex data using lstm with technical and macroeconomic indicators. *Financial Innovation*, 7(1):1–36, 2021.

[491] Yue Qiu, Zhewei Song, and Zhensong Chen. Short-term stock trends prediction based on sentiment analysis and machine learning. *Soft Computing*, pages 1–16, 2022.

[492] Saeed Seifollahi and Mehdi Shajari. Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to forex market prediction. *Journal of Intelligent Information Systems*, 52(1):57–83, 2019.

[493] Shahrokh Firouzi and Xiangning Wang. A comparative study of exchange rates and order flow based on wavelet transform coherence and cross wavelet transform. *Economic Modelling*, 82:42–56, 2019.

[494] Mustafa Onur Özorhan, İsmail Hakkı Toroslu, and Onur Tolga Şehitoğlu. A strength-biased prediction model for forecasting exchange rates using support vector machines and genetic algorithms. *Soft Computing*, 21(22):6653–6671, 2017.

[495] Arisara Pornwattanavichai, Saranya Maneeroj, and Somjai Boonsiri. Bertforex: Cascading model for forex market forecasting using fundamental and technical indicator data based on bert. *IEEE Access*, 10:23425–23437, 2022.

[496] Ch Sree, M Meghana, R Manjula, D Mohan, et al. Bitcoin price prediction using machine learning's boosting algorithms. In *Proceedings of Second International Conference on Sustainable Expert Systems*, pages 115–125. Springer, 2022.

[497] Adil Gürsel Karaçor and Turan Erman Erkan. Exploiting visual features in financial time series prediction. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 14(2):61–76, 2020.

[498] Shamima Ahmed, Muneer M. Alshater, Anis El Ammari, and Helmi Hammami. Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, 61:101646, 2022.

[499] Yin Shi and Xiaoni Li. A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. *Heliyon*, 5(12), 2019.

[500] Hafiz A. Alaka, Lukumon O. Oyedele, Hakeem A. Owolabi, Vikas Kumar, Saheed O. Ajayi, Olugbenga O. Akinade, and Muhammad Bilal. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94:164–184, 2018.

[501] P. Ravi Kumar and V. Ravi. Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. *European Journal of Operational Research*, 180(1):1–28, 2007.

[502] Daobo Yan, Yi Xiong, Zhihong Zhan, Xiaohong Liao, Fangchao Ke, Hailiang Lu, Yulun Ren, Shuang Liao, Lipin Sun, and Qixin Wang. Research on eigenvalue selection method of power market credit evaluation based on non parametric bayesian discriminant analysis and cluster analysis. *Energy Reports*, 7:990–997, 2021.

[503] Oriol Amat, Raffaele Manini, and Marcos Anton Renart. Credit concession through credit scoring: Analysis and application proposal. *Intangible Capital*, 13(1):51–70, 2017.

[504] A. I. Marques, V. Garcia, and J. S. Sanchez. A literature review on the application of evolutionary computing to credit scoring. *Journal of the Operational Research Society*, 64(9):1384–1399, 2013.

[505] Lyn C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172, 2000.

[506] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181, 2020.

[507] J. Kim. Economic integration of major industrialized areas: An empirical tracking of the continued trend. *Technological Forecasting and Social Change*, 67(2-3):187–202, 2001.

[508] C. Zopounidis, M. Doumpos, and S. Zanakis. Stock evaluation using a preference disaggregation methodology. *Decision Sciences*, 30(2):313–336, 1999.

[509] S. James Press and Sandra Wilson. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364):699–705, 1978.

[510] Blaz Zlicar and Simon Cousins. Discrete representation strategies for foreign exchange prediction. *Journal of Intelligent Information Systems*, 50(1):129–164, 2018.

[511] W. Huang, Y. Nakamori, and S. Y. Wang. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10):2513–2522, 2005.

[512] E. Steurer, M. Rothenhausler, and Y. Yeo. Forecasting discretized daily usd/dem exchange rate movements with quantitative models. In *22nd Annual Conference of the German-Classification-Society*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 467–472, 1999.

[513] Tuanhung Dao and Hyunchul Ahn. An optimized combination of π-fuzzy logic and support vector machine for stock market prediction. *Journal of Intelligence and Information Systems*, 20(4):43–58, 2014.

[514] Marija Gorenc Novak and Dejan Velušček. Prediction of stock price movement based on daily high prices. *Quantitative Finance*, 16(5):793–826, 2016.

[515] Manik Sharma, Samriti Sharma, and Gurvinder Singh. Performance analysis of statistical and supervised learning techniques in stock data mining. *Data*, 3(4), 2018.

[516] Jasmina Okicic, Sonja Remetic-Horvath, and Baris Buyukdemir. Stock selection based on discriminant analysis: Case of capital market of bosnia and herzegovina. *Journal of Economic and Social Studies*, 4(2):5–30, 2014.

[517] Shiuh-Nan Hwang, Chin-Tsai Lin, and Wang-Ching Chuang. Stock selection using data envelopment analysis-discriminant analysis. *Journal of Information & Optimization Sciences*, 28(1):33–50, 2007.

[518] Hulisi Ogut, M. Mete Doganay, and Ramazan Aktas. Detecting stock-price manipulation in an emerging market: The case of turkey. *Expert Systems with Applications*, 36(9):11944–11949, 2009.

[519] D. Plikynas, L. Simanauskas, and S. Buda. Research of neural network methods for compound stock exchange indices analysis. *Informatica*, 13(4):465–484, 2002.

[520] S. Aziz, M. Dowling, H. Hammami, and A. Piepenbrink. Machine learning in finance: A topic modeling approach. *European Financial Management*, 2021.

[521] Dattatray P. Gandhmal and K. Kumar. Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, 34:100190, 2019.

[522] Mehrdad Kaveh and Mohammad Saadi Mesgari. Application of meta-heuristic algorithms for training neural networks and deep learning architectures: A comprehensive review. *Neural Processing Letters*, pages 1–104, 2022.

[523] Jung-Pin Lai, Yu-Ming Chang, Chieh-Huang Chen, and Ping-Feng Pai. A survey of machine learning models in renewable energy predictions. *Applied Sciences*, 10(17):5975, 2020.

[524] MOON Kyoung-Sook, JUN Sookyung, and KIM Hongjoong. Speed up of the majority voting ensemble method for the prediction of stock price directions. *Economic Computation & Economic Cybernetics Studies & Research*, 52(1), 2018.

[525] Vijay Krishna Menon, Nithin Chekravarthi Vasireddy, Sai Aswin Jami, Viswa Teja Naveen Pedamallu, Varsha Sureshkumar, and KP Soman. Bulk price forecasting using spark over nse data set. In *International Conference on Data Mining and Big Data*, pages 137–146. Springer, 2016.

[526] Phayung Meesad and Risul Islam Rasel. Predicting stock market price using support vector regression. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1–6. IEEE, 2013.

[527] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[528] Huai Lin Dong, Juan Juan Huang, Zhu Hua Cai, and Qing Feng Wu. Research on predicted model of least squares support vector machine based on genetic algorithm. In *Advanced Materials Research*, volume 753, pages 2875–2881. Trans Tech Publ, 2013.

[529] Yanyan Zhuo. Research on stock index forecasting based on machine learning. In *2018 6th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2018)*, pages 66–72. Atlantis Press, 2018.

[530] Lydie Myriam Marcelle Amelot, Ushad Subadar Agathee, and Yuvraj Sunecher. Time series modelling, narx neural network and hybrid kpca–svr approach to forecast the foreign exchange market in mauritius. *African Journal of Economic and Management Studies*, 2020.

[531] Piotr Czekalski, Michal Niezabitowski, and Rafal Styblinski. Ann for forex forecasting and trading. In *2015 20th International Conference on Control Systems and Computer Science*, pages 322–328. IEEE, 2015.

[532] Ali H Dhafer, Fauzias Mat Nor, Gamal Alkawsi, Abdulaleem Z Al-Othmani, Nuradli Ridzwan Shah, Huda M Alshanbari, Khairil Faizal Bin Khairi, and Yahia Baashar. Empirical analysis for stock price prediction using narx model with exogenous technical indicators. *Computational Intelligence and Neuroscience*, 2022, 2022.

[533] SM Prabin and MS Thanabal. A repairing artificial neural network model-based stock price prediction. *International Journal of Computational Intelligence Systems*, 14(1):1337–1355, 2021.

[534] A Christy and G MeeraGandhi. Combining bitemporal conceptual datamodel with multiway join relations for forecasting. *Procedia Computer Science*, 57:1104–1114, 2015.

[535] J S Kallimani S Vasanthi A Mateen Buttar et al. J Faritha Banu, S B Rajeshwari. Modeling of hyperparameter tuned hybrid cnn and lstm for prediction model, 2022.

[536] Isaac Kofi Nti, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction. *Journal of Big Data*, 8(1):1–28, 2021.

[537] Priya Chakriswaran, Durai Raj Vincent, Kathiravan Srinivasan, Vishal Sharma, Chuan-Yu Chang, and Daniel Gutiérrez Reina. Emotion ai-driven sentiment analysis: A survey, future research directions, and open issues. *Applied Sciences*, 9(24):5462, 2019.

[538] Brett Lantz. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.

[539] Arsalan Dezhkam, Mohammad Taghi Manzuri, Ahmad Aghapour, Afshin Karimi, Ali Rabiee, and Shervin Manzuri Shalmani. A bayesian-based classification framework for financial time series trend prediction. *The Journal of Supercomputing*, pages 1–38, 2022.

[540] Rogério P Espíndola and Nelson FF Ebecken. On extending f-measure and g-mean metrics to multi-class problems. *WIT Transactions on Information and Communication Technologies*, 35, 2005.

[541] Masoomeh Rashidpoor Toochaei and Farzad Moeini. Evaluating the performance of ensemble classifiers in stock returns prediction using effective features. *Expert Systems with Applications*, page 119186, 2022.

[542] Ted Byrt, Janet Bishop, and John B Carlin. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429, 1993.

[543] Samuel Asante Gyamerah, Philip Ngare, and Dennis Ikpe. On stock market movement prediction via stacking ensemble learning method. In *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 1–8. IEEE, 2019.

[544] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[545] Grzegorz TRATKOWSKI and Krzysztof PIONTEK. Verification of a particular investment strategy on the us stock market–random forest algorithm and the quality of price changes forecasts.

[546] Maria Trigka, Andreas Kanavos, Elias Dritsas, Gerasimos Vonitsanos, and Phivos Mylonas. The predictive power of a twitter user's profile on cryptocurrency popularity. *Big Data and Cognitive Computing*, 6(2):59, 2022.

[547] Jingyi Shen and M Omair Shafiq. Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of big Data*, 7(1):1–33, 2020.

[548] Minqi Jiang, Jiapeng Liu, Lu Zhang, and Chunyu Liu. An improved stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica A: Statistical Mechanics and its Applications*, 541:122272, 2020.

[549] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[550] Arthur Pentland Dempster. Elements of continuous multivariate analysis. Technical report, 1969.

[551] Michael C Costanza and AA Afifi. Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of the American Statistical Association*, 74(368):777–785, 1979.

[552] Maurice M Tatsuoka. 9: Multivariate analysis in educational research. *Review of research in education*, 1(1):273–319, 1973.

[553] Donald F Morrison. Multivariate statistical methods-2. 1976.

[554] Harry F Gollob. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33(1):73–115, 1968.

[555] John Mandel. The partitioning of interaction in analysis. *Journal of Research of the National Bureau of Standards: Physics and chemistry*, 73:309, 1969.

[556] Hugh G Gauch Jr. Statistical analysis of yield trials by ammi and gge. *Crop science*, 46(4):1488–1500, 2006.

[557] JC Gower. Three-dimensional biplots. *Biometrika*, 77(4):773–785, 1990.

[558] K Ruben Gabriel, M Purificación Galindo, and José Luis Vicente-Villardón. Use of biplots to diagnose independence models in three-way contingency tables. In *Visualization of Categorical Data*, pages 391–404. Elsevier, 1998.

[559] GJS Ross, RD Jones, RA Kempton, FB Laukner, RW Payne, D Hawkins, and RB White. Mlp: maximum likelihood program. *NBS SPECIAL PUBLICATION 503*, page 87, 1980.

[560] Noora Shrestha. Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9(1):4–11, 2021.

[561] María Purificación Galindo Villardón. Una alternativa de representacion simultanea: Hj-biplot. *Qüestiió: quaderns d'estadística i investigació operativa*, pages 13–23, 1986.

[562] Karl Ruben Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[563] Zineb Bousbaa, Omar Bencharef, and Abdellah Nabaji. Stock market speculation system development based on technico temporal indicators and data mining tools. In *Heuristics for Optimization and Learning*, pages 239–251. Springer, 2021.

[564] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. PyCaret version 2.3.1.

[565] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[566] IBM Corp. *IBM SPSS Statistics for Windows, Version 26.0*. IBM Corp, Armonk, NY, 2019.

[567] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2021.

[568] J. L. Vicente-Villardón. *MultBiplot: A package for Multivariate Analysis using Biplots. Mathlab Software*. Departamento de Estadística. Universidad de Salamanca, 2015.

[569] Selcuk Korkmaz, Dinçer Göksülük, and GÖKMEN Zararsiz. Mvn: An r package for assessing multivariate normality. *R JOURNAL*, 6(2), 2014.

[570] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186. Citeseer, 1997.

[571] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer, 2014.

[572] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[573] Walter Krämer. Asymmetry in the distribution of daily stock returns. *Empirical Economics*, 60(3):1115–1125, 2021.

[574] Holmes Finch. Identification of variables associated with group separation in descriptive discriminant analysis: Comparison of methods for interpreting structure coefficients. *The Journal of Experimental Education*, 78(1):26–52, 2009.

[575] Thomas Svantesson and Jon W Wallace. Tests for assessing multivariate normality and the covariance structure of mimo data. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 4, pages IV–656. IEEE, 2003.

[576] P.P. Das, R. Bisoi, and P.K. Dash. Data decomposition based fast reduced kernel extreme learning machine for currency exchange rate forecasting and trend analysis. *Expert Systems with Applications*, 96:427–449, 2018.

[577] Svitlana Galeshchuk and Sumitra Mukherjee. Deep networks for predicting direction of change in foreign exchange rates. *Intelligent Systems in Accounting, Finance and Management*, 24(4):100–110, 2017.

[578] Leslie CO Tiong, David CL Ngo, and Yunli Lee. Forex prediction engine: framework, modelling techniques and implementations. *International Journal of Computational Science and Engineering*, 13(4):364–377, 2016.

[579] Vasilios Plakandaras, Theophilos Papadimitriou, Periklis Gogas, and Konstantinos Diamantaras. Market sentiment and exchange rate directional forecasting. *Algorithmic Finance*, 4(1-2):69–79, 2015.

[580] Mauricio Argotty-Erazo, Antonio Blázquez-Zaballos, Carlos A Argoty-Eraso, Leandro L Lorente-Leyva, Nadia N Sánchez-Pozo, and Diego H Peluffo-Ordóñez. A novel linear-model-based methodology for predicting the directional movement of the euro-dollar exchange rate. *IEEE Access*, 2023.