VNiVERSiDAD D SALAMANCA

TRABAJO DE FIN DE GRADO

GRADO EN ESTADÍSTICA

Facultad de Ciencias

Curso 2023/2024

APLICACIÓN DE ALGORITMOS MACHINE LEARNING AL DIAGNÓSTICO DE INFARTOS

Autora: Claudia Aranda Santolino

Tutores: Pedro Ignacio Dorado Díaz

José Luis Vicente Villardón

VNiVERSiDAD D SALAMANCA

TRABAJO DE FIN DE GRADO

GRADO EN ESTADÍSTICA

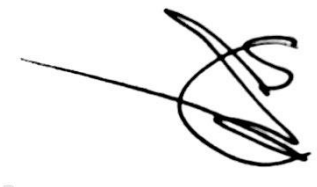
Facultad de Ciencias

Curso 2023/2024

APLICACIÓN DE ALGORITMOS MACHINE LEARNING AL DIAGNÓSTICO DE INFARTOS

Fdo Autora: Claudia Aranda Santolino

01/07/2024



Certificado del/los tutor/es TFG

D./Dña. Pedro Ignacio Dorado Díaz, profesor/a del Departamento de Estadística de la Universidad de Salamanca y José Luis Vicente Villardón, profesor/a del Departamento de Estadística de la Universidad de Salamanca,

HACE/N CONSTAR:

Que el trabajo titulado "APLICACIÓN DE ALGORITMOS MACHINE LEARNING AL DIAGNÓSTICO DE INFARTOS", que se presenta, ha sido realizado por Claudia Aranda Santolino, con DNI 45132019Q y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, de de 20

DORADO DIAZ, PEDRO IGNACIO (FIRMA)
Firmado digitalmente por DORADO DIAZ, PEDRO IGNACIO (FIRMA)
Fecha: 2024.07.01 09:28:40 +02'00'

VICENTE VILLARDON JOSE LUIS - 07839395Y
Firmado digitalmente por VICENTE VILLARDON JOSE LUIS - 07839395Y
Fecha: 2024.07.01 09:47:41 +02'00'

Fdo.: _____

Resumen

A nivel mundial, las enfermedades cardiovasculares (ECV) siguen siendo la principal causa de muerte, lo que requiere métodos de diagnóstico innovadores. En los últimos años, el uso de algoritmos de aprendizaje automático (*machine learning*, ML) en medicina se ha popularizado ampliamente, representando así un avance significativo en la medicina moderna. Siendo el objetivo principal del presente proyecto desarrollar un modelo predictivo en el diagnóstico de infartos de alta precisión mediante la aplicación de estos. Con el fin de alcanzarlo, varios modelos de ML, como árboles de decisión, *random forest*, regresión logística regularizada y *support vector machine*, se desarrollaron, optimizaron, validaron y compararon utilizando la base de datos “Stroke Prediction Dataset" disponible de forma pública en la web Kaggle. Según los resultados de la investigación, el modelo optimizado de *random forest* se estableció como el modelo más efectivo para la predicción de infartos en este contexto, al obtener la mayor puntuación en una medida llamada F3, destacando también el alto valor obtenido en la medida conocida como *accuracy*. Estos hallazgos señalan un gran potencial en las técnicas de aprendizaje automático para el diagnóstico médico, proporcionando una herramienta prometedora que podría ayudar a los profesionales del área de la salud a mejorar el pronóstico de los pacientes mediante diagnósticos precisos y oportunos.

Abstract

Globally, cardiovascular diseases (CVD) remain the leading cause of death, requiring innovative diagnostic methods. In recent years, the use of machine learning (ML) algorithms in medicine has become widely popular, representing a significant advance in modern medicine. The main objective of this project is to develop a predictive model for high-precision infarct diagnosis by applying machine learning algorithms. In order to achieve this, several ML models, such as decision trees, random forest, regularised logistic regression and support vector machine, were developed, optimised, validated and compared using the publicly available "Stroke Prediction Dataset" database on the Kaggle website. According to the results of the research, the optimised random forest model was established as the most effective model for stroke prediction in this context, scoring highest on a measure called F3, and also scoring high on the measure known as accuracy. These findings point to great potential in machine learning techniques for medical diagnosis, providing a promising tool that could help healthcare professionals improve patient prognosis through accurate and timely diagnoses.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. ENFERMEDADES CARDIOVASCULARES. EL INFARTO.

Las enfermedades cardiovasculares (ECV) se posicionan como la principal causa de muerte en todo el mundo, superando las muertes por cualquier otra causa conocida según la Organización Mundial de la Salud [1]. Alrededor de 17,7 millones de personas mueren cada año a causa de estas enfermedades, lo que representa aproximadamente un tercio de todas las muertes registradas a nivel mundial.

Las enfermedades cardiovasculares (ECV) son un conjunto de enfermedades que afectan al corazón y a los vasos sanguíneos. Estas incluyen la enfermedad coronaria (responsable de casi el 42% de las muertes por ECV), una enfermedad que afecta a los vasos sanguíneos que alimentan el músculo cardíaco; enfermedades cerebrovasculares que afectan a los vasos sanguíneos del cerebro derivando en accidentes cerebrovasculares (ACV) (responsables de casi el 40% de las muertes por ECV); enfermedades de las arterias periféricas que limitan el flujo sanguíneo que va dirigido a las extremidades; cardiopatías reumáticas asociadas con daños cardíacos causados por fiebres, cardiopatías congénitas; trombosis venosas profundas y embolias pulmonares, las cuales ocurren cuando los coágulos de sangre bloquean los vasos sanguíneos en el corazón o los pulmones.

Este tipo de eventos cardiovasculares agudos, como ataques cardíacos y accidentes cerebrovasculares (ACV), a menudo son causados por bloqueos en el suministro de sangre al corazón o al cerebro principalmente debidos a una acumulación de depósitos de grasa llamados placas en las paredes de los vasos sanguíneos.

Estos ataques cardíacos y accidentes cerebrovasculares suelen ocurrir como resultado de una combinación de diferentes factores de riesgo. Factores como el tabaquismo, la hipertensión, el colesterol alto, una dieta poco saludable, la obesidad, la falta de actividad física, el consumo excesivo de alcohol, la diabetes… El riesgo de ataque cardíaco aumenta también en presencia de la COVID-19, la cual puede provocar respuestas inflamatorias, así como la coagulación sanguínea, lo que altera la salud cardiovascular.

La contaminación ambiental ha sido reconocida también en la última década como uno de los principales contribuyentes a las enfermedades cardíacas. Se ha demostrado que los contaminantes en la sangre pueden causar inflamación y daño arterial, lo que puede conllevar enfermedades como la cardiopatía isquémica (falta de sangre y oxígeno en el corazón). Un estudio reciente presentado en las Sesiones Científicas de la Asociación Estadounidense del Corazón [2] relaciona directamente la disminución de la contaminación durante el confinamiento (disminución de los niveles de partículas finas (PM2,5)) debido a la pandemia producida por la COVID-19 con la reducción en el número de ataques cardíacos (disminución del 6% de los ataques cardíacos en Estados Unidos). Por todo esto dicho factor es considerado ya una causa tan importante de las enfermedades cardíacas como lo son la hipertensión y el tabaquismo, superando al colesterol alto, el sobrepeso y el sedentarismo. Este

reciente descubrimiento ha conducido al surgimiento de la cardiología ambiental, una nueva subespecialidad en el ámbito de la cardiología.

Respecto a la población más vulnerable a sufrir este tipo de enfermedades y ataques tanto cardiovasculares como cerebrovasculares predomina la población de países de ingresos bajos y medios donde ocurren más del 75% de este tipo de muertes. Destacando América Latina y el Caribe donde podemos apreciar una alta prevalencia de hipertensión en la población que se convierte en un factor preocupante, especialmente cuando una proporción significativa de personas, alrededor del 28% en mujeres y el 43% en hombres no son conscientes de padecer esta afección, según un estudio llevado a cabo por "NCD Risk Factor Collaboration" [3].

Uno de los grupos de edad más afectados y quizás el más preocupante se trate de los adultos jóvenes (personas de edad entre los 25 y 44 años) entre los cuales el número de problemas cardíacos está en aumento destacando que uno de cada cinco pacientes con ataques cardíacos en adultos jóvenes tenía entre 40 y 44 años[4]. Aún a la vista de los mencionados acontecimientos, muchos jóvenes continúan desconociendo que corren el riesgo de desarrollar enfermedades cardíacas. Según una encuesta, casi la mitad de los estadounidenses menores de 45 años siguen sin creer que tengan un riesgo elevado de sufrir enfermedades cardíacas. [4]

El aumento de hábitos como la mala alimentación y la falta de ejercicio está propiciando la aparición anticipada de factores de riesgo en la población, más concretamente en nuestro país, donde el 35% de los menores en España presentan dos o más factores de riesgo cardiovascular según la fundación española del corazón [5]. Como cabe esperar, a pesar de todo esto, las mencionadas siguen siendo situaciones menos habituales que en el grupo de edad correspondiente a las personas mayores (personas de 60 años o más), población en la que sigue siendo más habitual este tipo accidentes cardiovasculares y cerebrovasculares.

Las enfermedades cardíacas, como el infarto de miocardio, tienen un impacto que va más allá de lo físico, afectando a la calidad de vida del paciente, a sus interacciones sociales, a su vida laboral y a su situación socioeconómica entre otras. A la hora de inspeccionar este tipo de consecuencias me gustaría diferenciar entre los países más afectados por este tipo de enfermedades, que como ya he mencionado anteriormente son los países de medianos y bajos ingresos y nuestra propia nación, España.

Un estudio realizado en varios países [6] considerados de medios-bajos ingresos, entre los cuales encontramos Argentina, China, La India y Tanzania, revela que las complicaciones de estas enfermedades reducen drásticamente la capacidad laboral de siete de cada diez pacientes, reduciendo así los ingresos mensuales incluso de personas con altos ingresos. El Dr. Andrés Pichon-Rivière, director del Instituto de Eficiencia Clínica y Sanitaria (IECS) y coautor del estudio mencionado, destaca cómo la pérdida de ingresos, que en los casos analizados oscila entre 403 y 1.860 pesos (21,78 € y 100,52 €), tiene un fuerte impacto en las familias, especialmente en aquellas con recursos limitados. Recordemos que las enfermedades cardiovasculares afectan a la economía no sólo directamente a través de menores ingresos, sino también indirectamente. Pichon-Rivière destaca cómo este efecto se extiende a las esferas familiares, sociales y económicas del paciente, llegando incluso en algunos casos a forzar situaciones tan extremas como la necesidad de que los niños de su

entorno deban abandonar el colegio para poder así trabajar y aumentar los ingresos mensuales o la necesidad de la venta de bienes inmuebles, todo con el fin de cubrir gastos. En definitiva, muchos de los pacientes residentes en dichos países sufren el fuerte impacto económico de este tipo de enfermedades que consumen gran parte de su presupuesto familiar.

Teniendo presente la anterior información y siendo conscientes de las diferencias de recursos entre nuestro país y los mencionados en el estudio previo, veamos como nuestra población se ve también afectada por las consecuencias socioeconómicas de las enfermedades cardiovasculares y los accidentes cerebrovasculares. Para ello me centraré en dos estudios que exploran este tipo de cardiopatías desde dicha perspectiva.

Por un lado, un estudio del Hospital Universitario 12 de Octubre de Madrid, realizado en 1999, basado en 155 pacientes con un primer infarto agudo de miocardio [7], en el cual se realizaron dos evaluaciones, una al ingreso y otra seis meses después, teniendo en cuenta variables clínicas, socioeconómicas, de apoyo social y de calidad de vida. Los resultados mostraron que la mayoría de los pacientes (90,9%) tenían habilidades de lectura y escritura, aunque un porcentaje significativo (42,3%) no tenía educación más allá de la escuela primaria. Aunque el 45,1% estaba empleado al principio, sólo el 14% volvió a trabajar después de seis meses, en su mayoría con educación secundaria o terciaria y trabajadores calificados. Respecto al apoyo social se pudo apreciar que el 78,7% contaba con apoyo instrumental y el 69,9% contaba con apoyo emocional. Además, se encontró una relación significativa entre el apoyo instrumental y emocional y la calidad de vida de los pacientes. Como conclusiones se obtuvieron que la percepción de apoyo está relacionada con el tamaño de la red social, y las personas con menor calidad de vida tienden a tener un menor nivel de educación.

Por otro lado, en un estudio paralelo llevado a cabo el mismo año, en 1999, por el Hospital Universitario Marqués de Valdecilla, Santander, Cantabria [8] se estudió a 584 pacientes menores de 65 años que fueron hospitalizados por infarto agudo de miocardio durante un período de cuatro años. Entre otras variables, se analizaron la edad, el sexo, los antecedentes de enfermedad coronaria, la situación laboral anterior, la categoría laboral, el área económica de trabajo, el tratamiento, las complicaciones, los días de baja y la situación laboral posterior al infarto. Los resultados mostraron que antes del infarto, el 65,3% de los pacientes estaban activos en el trabajo. Después del evento, el promedio de número de días de ausencia laboral fue de 243,9, lo que varía significativamente según la edad y el área de actividad. El 56,6% de las personas anteriormente empleadas lograron volver a trabajar, y se observó una dependencia de este retorno dependiendo de la edad, las categorías laborales (especialmente en trabajos administrativos) y los antecedentes de angina. Las conclusiones extraídas mostraron que la probabilidad de reincorporarse al trabajo después de un infarto de miocardio era menor en pacientes ancianos con una categoría ocupacional inferior, que pertenecían al sector agrícola o industrial y que no tenían enfermedades coronarias conocidas.

Como podemos apreciar ambos estudios muestran que el nivel educativo del paciente juega un papel importante y que la vuelta al trabajo después de un infarto presenta muchos desafíos, especialmente para las personas con menor nivel educativo. Además, se destaca la importancia del apoyo social y emocional para la calidad de vida, destacando el papel de las redes sociales en el proceso de recuperación.

1.2. INTELIGENCIA ARTIFICIAL, MACHINE LEARNING Y SUS APLICACIONES EN LA MEDICINA.

La detección y el tratamiento tempranos son muy importantes para las personas que padecen enfermedades cardiovasculares o tienen un alto riesgo de padecerlas debido a la presencia de uno o más factores de riesgo. La identificación temprana y la intervención oportuna a través de servicios de asesoramiento o medicación adecuada son estrategias clave para reducir el impacto de estas enfermedades cardiovasculares y mejorar la calidad de vida de los pacientes. Veamos cómo se llevan a cabo actualmente estos diagnósticos a los que nos referimos.

A lo largo de miles de años, el cerebro humano ha evolucionado hasta nuestros días lo que nos permite entender lo que vemos, razonar de forma lógica y aprender de vivencias previas de forma que ampliamos nuestros conocimientos personales diariamente, todo esto conlleva a la capacidad de los médicos, gracias a sus estudios y años de experiencia, para poder diagnosticar enfermedades en base a imágenes, datos, información, síntomas... Sin embargo, esto nunca había ocurrido en seres inertes, como los ordenadores, hasta ahora, gracias a la implementación de la inteligencia artificial.

Resulta complicado encontrar una definición uniforme de lo que se conoce como inteligencia artificial. Con frecuencia nos referimos a ella como el campo de las ciencias de la computación que intenta imitar los procesos cognitivos humanos, la capacidad de aprendizaje y el almacenamiento de conocimiento [9].

Estas habilidades son la base del Aprendizaje Automático (AA), también conocido como *Machine Learning* (ML), que actualmente se presenta como una herramienta bioinformática muy útil en el procesamiento de datos. El objetivo del aprendizaje automático es replicar la capacidad de los humanos para aprender de experiencias pasadas. Un ejemplo de cómo se logra esto sería proporcionando al algoritmo un conjunto de tamaño considerable de imágenes, los datos de entrenamiento, cada una de las cuales está etiquetada con el concepto que se desea reconocer. Siendo el objetivo conseguir la capacidad de entrenar un conjunto de hipótesis o técnicas potenciales que identifiquen la apariencia visual de cada idea. Luego se elige el método que proporciona la mejor respuesta, es decir, la capacidad de predecir con la mayor precisión las etiquetas de nuevas imágenes que no se utilizaron en la etapa de entrenamiento (datos de prueba) [10].

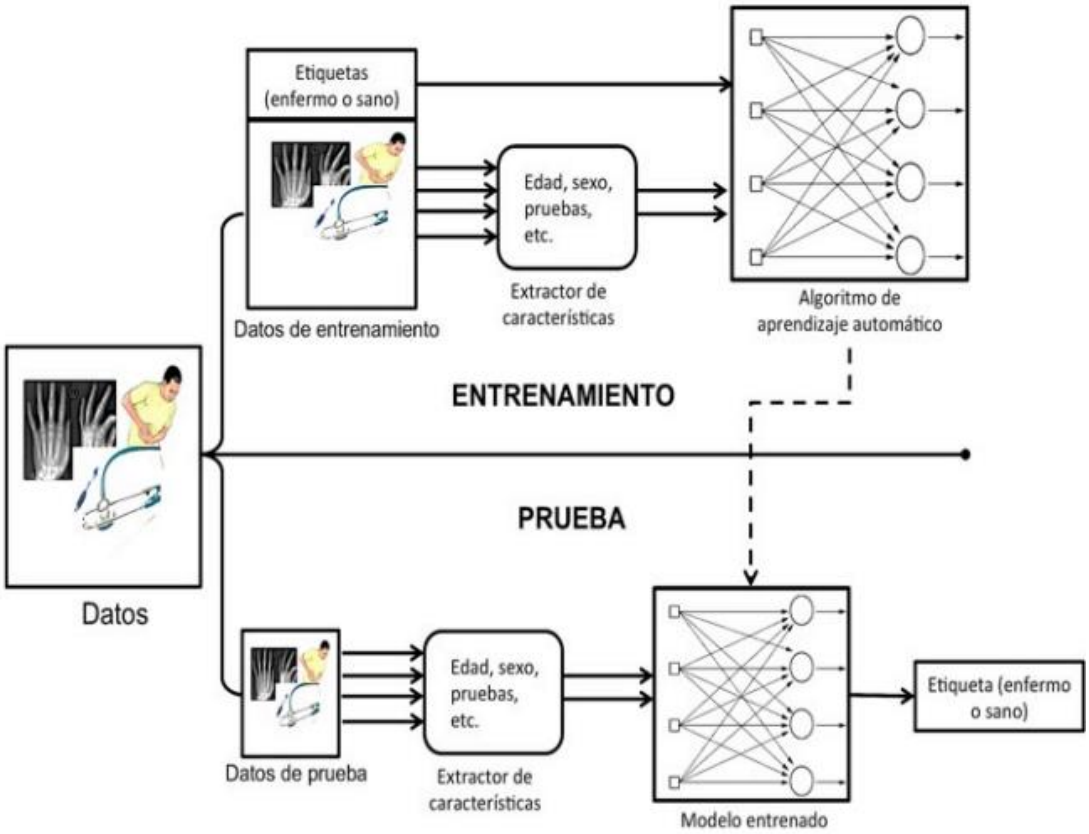


Figura 1. Gráfico explicativo del aprendizaje automático, obtenido de [10]

La mayoría de las técnicas de aprendizaje automático se encuentran en tres categorías según el tipo de técnica de aprendizaje utilizada: aprendizaje por refuerzo, aprendizaje no supervisado y aprendizaje supervisado [11].

En el **aprendizaje supervisado**, comenzamos entrenando el algoritmo al proporcionarle entradas o características (inputs o *features*) vinculadas a un resultado ya conocido o etiqueta (*output* o *label*) definido por personas expertas en el campo que está siendo estudiado. Estos algoritmos tienen como finalidad que la máquina aprenda reglas generales que relacionen las entradas con las salidas. En el ámbito médico, se puede usar un modelo para conectar características de lesiones en la piel, como tamaño, color y forma, con un resultado específico, como clasificar la lesión como benigna o maligna. Durante este proceso basado en multitud de datos proporcionados, el ordenador adquiere experiencia y conocimiento al aprender patrones. Luego, se prueba la capacidad predictiva del modelo al introducir un nuevo registro y evaluar cómo se comporta en la predicción.

En contraste con el método anterior, en el **aprendizaje no supervisado** el sistema no recibe información preexistente. Por el contrario, le proporcionamos una gran cantidad de datos sin etiquetas, y es el sistema el encargado de automáticamente identificar tendencias o patrones ocultos y organizar los datos en grupos. En medicina, por ejemplo, podríamos proporcionar imágenes de tumores cerebrales y permitir que el sistema organice estas imágenes en grupos según los patrones que encuentre.

En el **aprendizaje por refuerzo**, el algoritmo se vale de datos etiquetados y no etiquetados y aprende de forma similar a como lo hacen los niños, interactuando con dichos datos y tomando decisiones recibiendo así recompensas (*feedback*) positivas o negativas según lo que hace. Esto le permite mejorar y desarrollar mejores métodos de clasificación de los datos para obtener así mas *feedbacks* positivos y cada vez menos negativos. De esta forma el algortimo "aprende" sin disponer de ninguna orientación específica.

En el aprendizaje supervisado, algunos algoritmos utilizados son: *decision trees, K-nearest neighbours, support vector machine, naive bayes, linear regression, logistic regression, bagging, random forest, boosting, xgboost, neural networks*, etc. En el aprendizaje no supervisado, los algoritmos utilizados son: *hierarchical clustering, k-means clustering, linear discriminant analysis*, etc. En el aprendizaje por refuerzo (*reinforcement learning*, RL) hay varios algoritmos que se utilizan, como el *Q-learning*, los métodos de gradiente de políticas, los métodos de Montecarlo y el aprendizaje por diferencia temporal.

Dentro del *Machine Learning*, un tipo de algoritmos a destacar es el *Deep Learning* (DL), el cual utiliza redes neuronales artificiales formadas por varias capas con el fin de extraer relaciones sumamente complejas entre características y etiquetas, superando en algunos campos las habilidades humanas, como a la hora de clasificar imágenes. Este tipo de algoritmos se pueden usar también para aprendizaje supervisado, no supervisado y por refuerzo.

Con toda la información previa podemos asumir que la IA puede acelerar y ahorrar dinero en los procesos de diagnósticos médicos. Los especialistas en la salud se podrían valer del aprendizaje automático (ML) para analizar pacientes, clasificar su riesgo y tomar decisiones. De hecho, ya se han creado, entre otros, modelos para identificar enfermedades como autismo [12], cáncer de piel [13], enfermedades cardíacas [14] y cataratas congénitas [15], los cuales podrían funcionar como herramientas capaces de descartar sujetos que no corresponden con el diagnóstico previsto para ellos, dejando a los médicos los casos potencialmente positivos. Cabe la posibilidad incluso de que los propios pacientes pudieran usar también estos modelos para obtener evaluaciones preliminares por ejemplo con la implementación de estos en los dispositivos móviles [16] o ser utilizados por los médicos de familia en zonas poco habitadas en las que no disponen de especialistas para realizar el diagnóstico o el seguimiento de estos pacientes.

2. MOTIVACIÓN Y OBJETIVOS.

2.1. MOTIVACIÓN

La predicción temprana de eventos cardíacos agudos, como un infarto, se presenta como una necesidad en el panorama actual de la atención médica para salvaguardar vidas y mejorar la calidad de atención al paciente. La premisa fundamental es que la anticipación y la intervención proactiva pueden reducir, si no prevenir, los daños irreparables causados por estos eventos críticos.

El infarto ha sido históricamente una de las principales causas de morbilidad y mortalidad en todo el mundo. A pesar de los avances médicos, la capacidad de prever y responder a los episodios

cardíacos sigue siendo un desafío. La importancia de una detección rápida se ve reforzada por la posibilidad de resultados catastróficos, como la pérdida irreversible de tejido cardíaco o, en los casos más extremos, la pérdida de vidas humanas.

En este contexto, los avances tecnológicos en el campo médico, particularmente en la aplicación de técnicas de aprendizaje automático (ML) para la predicción de eventos cardiovasculares, abren nuevas perspectivas, gracias a la capacidad única de estos algoritmos para examinar grandes conjuntos de datos y descubrir patrones en ellos ofrece una oportunidad sin precedentes para innovar la forma en que abordamos la prevención y el tratamiento de enfermedades cardiacas. Al alimentarlos con datos clínicos y biomédicos, los algoritmos de ML tienen el potencial de convertirse en herramientas de vanguardia para la identificación temprana de personas en riesgo de infarto. Estos modelos sofisticados pueden superar incluso las limitaciones de los métodos convencionales, permitiendo así una evaluación más completa de los factores de riesgo y brindando predicciones personalizadas basadas en la complejidad única de cada paciente.

En este caso, la motivación detrás del presente estudio va más allá de la mera exploración de nuevas tecnologías; se trata de un compromiso con el objetivo de transformar la atención cardiovascular utilizando la inteligencia artificial en el mundo real. Finalmente, este trabajo se alinea con la idea de un futuro en el que la predicción del infarto mediante el uso de técnicas de IA como el aprendizaje automático sea una práctica clínica que cambie paradigmas, mejore la eficiencia de la atención médica y, lo más importante, proteja la salud y el bienestar de las comunidades.

2.2. OBJETIVO PRINCIPAL

El objetivo principal (OP) de este trabajo fin de grado es la creación y evaluación de modelos de aprendizaje automático que puedan predecir con precisión la ocurrencia de infartos basándome en la base de datos "*Stroke Prediction Dataset*" disponible de forma pública y gratuita en la plataforma online Kaggle [17].

2.3. OBJETIVOS SECUNDARIOS

Con el fin de poder alcanzar el objetivo principal explicado en el anterior subapartado tendré también que alcanzar otros objetivos secundarios que me guiarán hasta mi objetivo principal.

- (OS1) Conocimiento y comprensión de los métodos supervisados de ML más utilizados en el ámbito de la medicina.

- (OS2) La comparación de varios métodos supervisados de ML, como *decision tree, random forest*, regresión logística regularizada y *support vector machine*. El objetivo de este análisis comparativo será determinar la eficacia relativa de cada método en la predicción de infartos.

- (OS3) Llevaré también a cabo la optimización y validación de los modelos, usando métodos como la validación cruzada basada en *k-folds.* También evaluaré la generalización y la robustez de los modelos.

- (OS4) Por último, llevaré a cabo una comparación de los resultados obtenidos con investigaciones anteriores basadas en la misma base de datos permitiendo así una mayor comprensión de los resultados de mi proyecto.

3. ANTECEDENTES. ESTADO DEL ARTE

Como ya ha sido mencionado anteriormente, este Trabajo de Fin de Grado se basa en una base de datos pública, es por esto, que como cabe esperar no es ni mucho menos el primer trabajo basado en esta ni será el último, por lo que en cierta parte se apoyará en estos trabajos previos, específicamente en los siguientes.

3.1. ESTUDIO REALIZADO POR EL DEPARTAMENTO DE INGENIERÍA INFORMÁTICA E INFORMÁTICA DE LA UNIVERSIDAD DE PATRAS, GRECIA

Los resultados y conclusiones de un estudio realizado por el Departamento de Ingeniería Informática e Informática de la Universidad de Patras, Grecia [18], utilizando la misma base de datos que usaré en el presente proyecto, muestran una evaluación exhaustiva del rendimiento de los modelos de aprendizaje automático (ML) en la predicción temprana de accidentes cerebrovasculares.

Se valieron de diez validaciones cruzadas para evaluar la eficiencia de los modelos en el conjunto de datos para evaluar su rendimiento en el entorno WEKA 3.8.6. El modelo de apilamiento (*stacking*), que combinó clasificadores base como *naive Bayes*, *Random Forest* (RF), *J48* (*Decision Tree*, DT) y *RepTree*, demostró ser el más eficiente si nos referimos a la capacidad de los modelos de aprendizaje automático para predecir con precisión los casos de accidente cerebrovascular en todas las métricas analizadas, superando al resto de clasificadores.

Tanto el modelo de apilamiento (*stacking*), como el modelo RF, si nos centramos en la métrica AUC ROC, tuvieron una capacidad de discriminación similar, con valores de 0.989 y del 0.986, respectivamente, lo que nos indica que ambos son capaces de identificar con éxito los casos de accidente cerebrovascular. El modelo de apilamiento también superó a los clasificadores RF, 3-NN y DT (J48) según la métrica *F1-measure* en un 0.8%, 5.9% y 6.5%.

Los modelos propuestos, especialmente DT y RF, superaron significativamente su rendimiento en términos de *recall* (sensibilidad), *F-measure* y *accuracy* en comparación con un estudio previo realizado por el mismo departamento sobre la misma base de datos. En resumen, la investigación ha demostrado que el método *stacking* sigue siendo el más efectivo y la principal recomendación.

3.2. Estudio realizado por el Departamento de Ingeniería Eléctrica e Informática de la Universidad North South en Bangladesh y el Departamento de Tecnología de la Información de la Universidad Taif en Arabia Saudita

De forma paralela los resultados y conclusiones de un estudio diferente realizado por el Departamento de Ingeniería Eléctrica e Informática de la Universidad North South en Bangladesh y el Departamento de Tecnología de la Información de la Universidad Taif en Arabia Saudita [19], utilizando la misma base de datos que se utilizará en este estudio y usada en el proyecto mencionado previamente, abordan otros temas fundamentales.

La visualización de datos analizó variables importantes como género, edad, hipertensión, enfermedad cardíaca, estado civil, nivel promedio de glucosa e índice de masa corporal (IMC). El conjunto de datos contiene una mayor cantidad de muestras femeninas, y la edad promedio es de la década de los 40, con un límite superior de aproximadamente 60 años. Además, se investigó la relación entre las características y la variable objetivo, que en este caso era el riesgo de accidente cerebrovascular.

La variable objetivo se correlacionó positivamente con la edad, la hipertensión, el nivel promedio de glucosa, la enfermedad cardíaca, el estado civil y el IMC, mientras que el género no.

Se utilizaron *Random Forest* (RF), *Decision Tree* (DT) y *Voting Classifier* para evaluar los modelos. El modelo *Random Forest* obtuvo un valor F1 del 96%, superando a todos los demás métodos. También se destacaron las métricas *recall* y *accuracy* para personas saludables y con antecedentes de accidente cerebrovascular. Los hallazgos se compararon con investigaciones anteriores y se notó que el algoritmo *Random Forest* era más efectivo, con una métrica *precision* del 96 %, en comparación con investigaciones anteriores que alcanzaron un 73 % y un 77 % con el mismo algoritmo.

3.3. Estudio de la Carnegie Mellon University

Un estudio de la Carnegie Mellon University [20] examinó qué factores están relacionados con la probabilidad de que ocurra un accidente cerebrovascular. Tras descubrir que factores como el tipo de vivienda, el tipo de trabajo y el género no tenían una correlación significativa con el riesgo de sufrir un accidente cerebrovascular, optaron por no incluirlos en sus modelos predictivos.

Tras esto, intentaron explicar la variabilidad de los datos mediante el análisis de componentes principales (PCA) con regresión lineal; sin embargo, solo pudieron explicar el 80 % de la variabilidad utilizando los cinco primeros componentes principales. Las relaciones entre las características eran no lineales o de baja dependencia cuando lo intentaron con menos componentes. El resultado insatisfactorio del modelo lineal fue una precisión del 73%. Posteriormente utilizaron métodos de agrupamiento para clasificar. Construyendo un dendrograma

que mostró que la mayoría de los grupos se correlacionaban bien con la presencia o ausencia de accidentes cerebrovasculares.

Finalmente, clasificaron los datos utilizando un *decision tree*. Este modelo creó reglas para categorizar los datos y, al evaluarlo, descubrieron que el clasificador tenía una precisión del 78.6%. En resumen, concluyeron que, el árbol de decisiones era el modelo más adecuado para predecir la ocurrencia de un accidente cerebrovascular.

3.4. Estudio realizado por School of Computer Engineering KIIT, University Bhubaneswar, Odisha

Otro estudio realizado por School of Computer Engineering KIIT, University Bhubaneswar, Odisha [21] utilizando la misma base de datos que los proyectos anteriores, que también usaré, se centró en analizar los resultados de varios clasificadores de aprendizaje automático utilizando la matriz de confusión para evaluar métricas como Sensibilidad, Especificidad y Precisión, entre otras.

Fueron utilizados numerosos algoritmos, incluidos *Random Forests, Support Vector Machines, AdaBoost, XGBoost y Decision Trees*. Se pudo apreciar que *XGBoost* superó a los demás modelos en eficiencia con una precisión del 95.11% mienstras que *Random Forest*, un competidor cercano, tenía una precisión del 92.93%. Aunque *Random Forest* tenía la sensibilidad más alta, *XGBoost* tenía una *accuracy* superior debido a su relación más sólida con el conjunto de datos. *XGBoost* tenía un área bajo la curva (AUC) de 0,943, lo que demostró su capacidad para distinguir entre clases, según la curva ROC.

En resumen, el estudio destaca el papel de los modelos de aprendizaje automático en la predicción temprana de accidentes cerebrovasculares El Clasificador *XGBoost* demostró ser el más preciso, con resultados prometedores al considerar seis o siete características.

3.5. Proyecto llevado a cabo por investigadores de Irlanda, China y Singapur en 2022

Un proyecto llevado a cabo por investigadores de Irlanda, China y Singapur en 2022 [22] proporciona un análisis detallado de varios algoritmos de referencia para la predicción de accidentes cerebrovasculares. Entre ellos utilizaron una técnica de aprendizaje profundo llamada red neuronal convolucional (CNN) con el fin de predecir accidentes cerebrovasculares. Tras realizar experimentos con diferentes configuraciones de variables y componentes principales, se concluyó que la *neural network* ofrecía mejor precisión cuando se utilizaban características específicas de edad, enfermedad cardíaca, nivel promedio de glucosa e hipertensión. Con este conjunto de características, la precisión alcanzaba un 80% y la tasa de error era del 16%.

En resumen, podemos diferenciar este estudio del resto en que demuestra que una combinación de variables acotada por ciertas características puede ser de gran ayuda en la predicción de accidentes cerebrovasculares.

	3.1. Estudio Universidad de Patras, Grecia	3.2. Departamento de Ingeniería Eléctrica e Informática de la Universidad North South en Bangladesh y el Departamento de Tecnología de la Información de la Universidad Taif en Arabia Saudita	3.3. Carnegie Mellon University	3.4. School of Computer Engineering KIIT, University Bhubaneswar, Odisha	3.5. Investigadores de Irlanda, China y Singapur
Base de Datos	Stroke Prediction Dataset, Kaggle	Stroke Prediction Dataset, Kaggle	Stroke Prediction Dataset, Kaggle	Stroke Prediction Dataset, Kaggle	Stroke Prediction Dataset, Kaggle
Algoritmos	Stacking, Random Forest, Naive Bayes, RepTree	Random Forest, Decision Tree, Voting Classifier	Decision Tree	Neural Network, Decision Tree, Random Forest, Support Vector Machiner	Neural Network, Decision Tree, Random Forest
Evaluación	Cross-validation, AUC, F1-measure, precision, recall	F1 Score, Accuracy	Sensitivity, Specificity, F1 Score	Precision, Recall, F-score, Accuracy	Precision, Recall, F-score, Accuracy
AUC ROC /Accuracy	Stacking: 98.9%, Random Forest: 98.6%	Random Forest: 96%	Decision Tree:78.6%.	XGBoost: 95.08%, Decision Tree: 87.31%	Neural Network: 80%
Recomendación	Stacking, recomendado por superar al resto de los algoritmos en *F1-measure.*	Random Forest, recomendado por valor alto AUC ROC	Decision Tree, recomendado por precision.	XGBoost recomendado por precisión	Neural Network recomendado por precisión

Tabla 1. Tabla comparativa estudios previos, estado del arte.

4. METODOLOGÍA

Veamos ahora que pasos debemos seguir a la hora de implementar un modelo de aprendizaje automático.

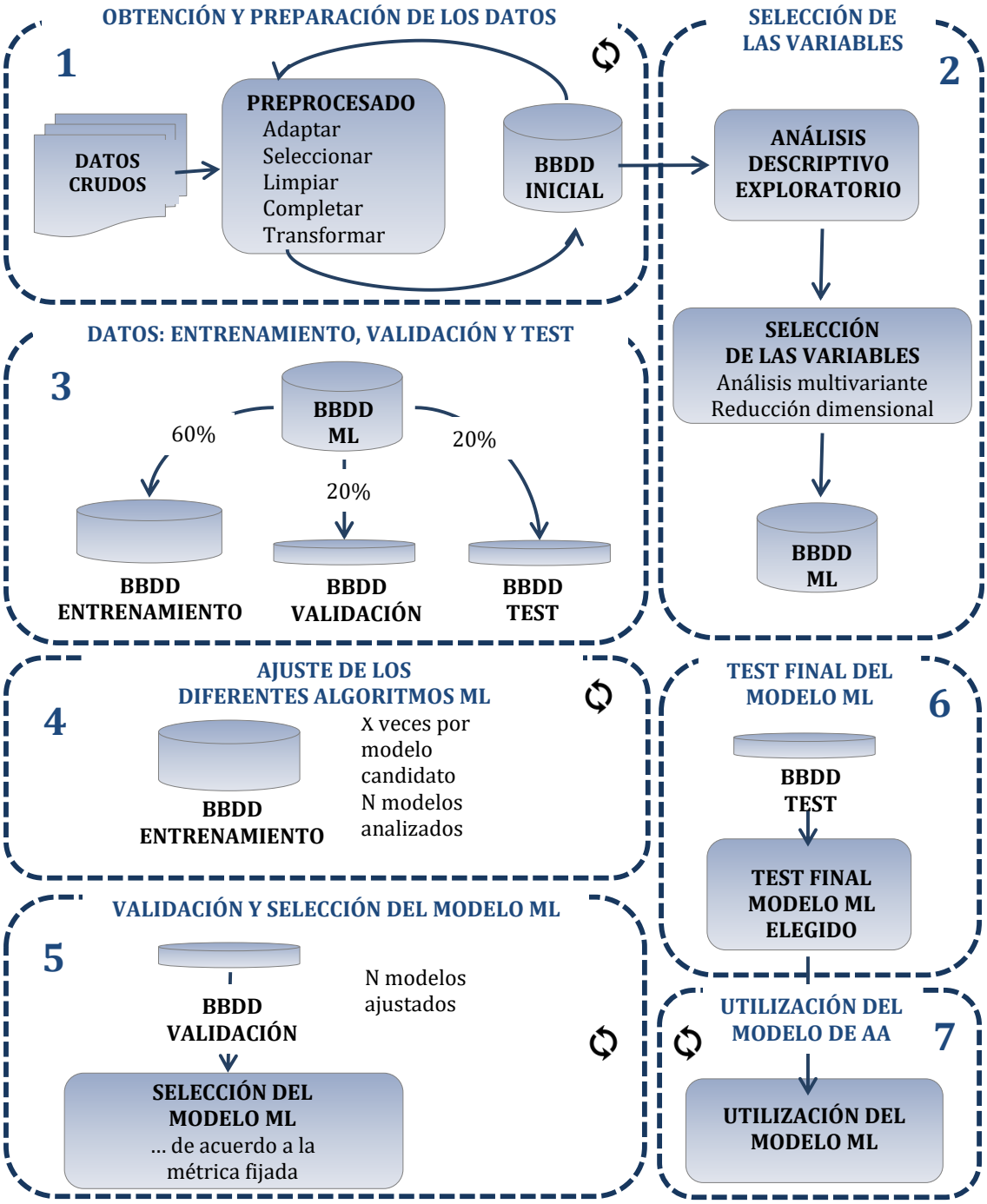


Figura 2. Gráfico explicativo de la metodología para implementar un modelo de aprendizaje automático, obtenido de [9].

El proceso comienza con el preprocesamiento de datos, que consiste en convertir datos brutos en datos estructurados. Este proceso incluye la construcción de una base de datos inicial y la selección de variables pertinentes. Posteriormente, se lleva a cabo un análisis descriptivo y exploratorio para determinar las variables que serán utilizadas en los algoritmos de aprendizaje automático.

Tras esto, el conjunto de datos se divide en tres subconjuntos: entrenamiento, validación y prueba, que normalmente asignan el 60%, 20% y 20% de los datos respectivamente. Los algoritmos de ML seleccionados se ajustan utilizando el subconjunto de entrenamiento.

A continuación, se utiliza el subconjunto de validación para evaluar la calidad del modelo. En la evaluación de modelos de aprendizaje automático, es esencial entender conceptos como verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos, que son elementos clave de la matriz de confusión. La exhaustividad, también llamada *precisión* o valor predictivo positivo, representa la proporción de verdaderos positivos respecto al total predicciones positivas, mientras que el valor predictivo negativo indica la proporción de verdaderos negativos respecto al total de predicciones negativas. La exactitud es la proporción de predicciones correctas, y la especificidad mide la proporción de verdaderos negativos respecto al total de "sanos" (suma de falsos positivos y verdaderos negativos). La sensibilidad es la capacidad del modelo para identificar correctamente los casos positivos, y la curva ROC es una representación gráfica de la sensibilidad frente a la especificidad, utilizada para evaluar la capacidad discriminativa del modelo. El AUC-ROC (Area Under the Receiver Operating Characteristic Curve) representa el área bajo la curva ROC oscila entre 0 y 1, un valor de AUC-ROC cercano a 1 indica un modelo excelente que puede distinguir perfectamente entre las dos clases, mientras que un valor cercano a 0.5 indica un modelo que no tiene capacidad discriminativa mejor que el azar.

He de destacar también una medida de la relación entre una exposición y un resultado, llamada *odds ratio* (OR). La OR es la probabilidad de que se produzca un resultado dada una exposición específica en comparación con la probabilidad de que se produzca un resultado sin esa exposición[23]. Podemos encontrar 3 escenarios, si el OR es igual 1 significa que la exposición y el resultado no están relacionados; si el OR es mayor que 1 sugiere que la exposición aumenta la posibilidad de que se dé el resultado y si es el OR es menor que 1 quiere decir que la exposición disminuye la posibilidad de que se dé el resultado.

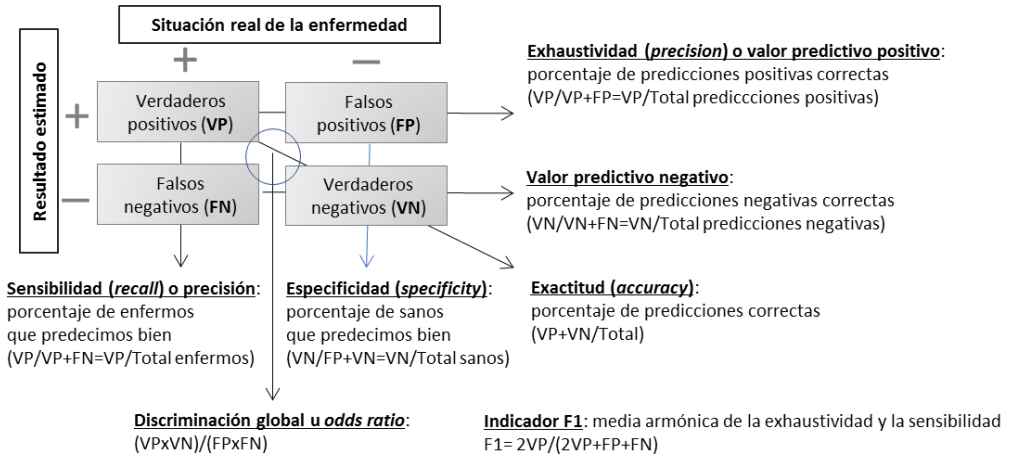


Figura 3. Explicación de la matriz de confusión, obtenido de [9].

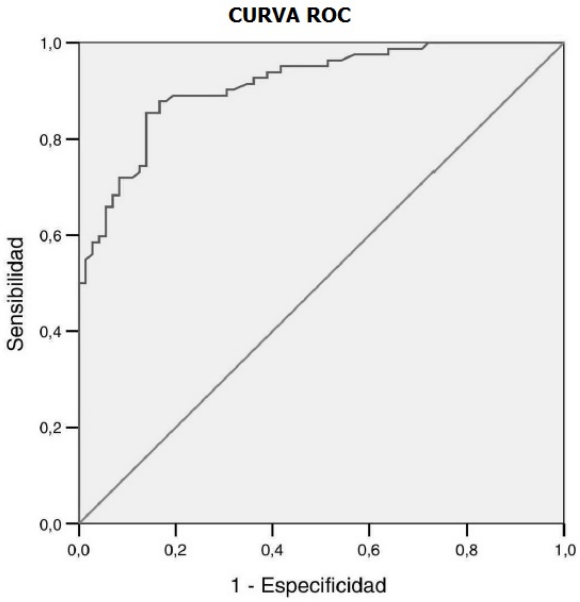


Figura 4. Curva ROC con métrica AUC ROC, obtenido de [9].

El proceso de entrenamiento y validación se repite varias veces, aleatorizando ambos subconjuntos (k-folds). Esto se hace con el objetivo de optimizar los parámetros internos del algoritmo, evaluar la robustez del modelo y evitar el sobreajuste o subajuste.

Una vez construido el modelo final, se prueba con el subconjunto de datos de prueba para verificar su comportamiento con los datos que no se utilizaron en la construcción y validación. Se consideraría también ampliar el conjunto de datos de entrenamiento para obtener un estimador más confiable si se observan discrepancias entre el rendimiento del conjunto de pruebas y el de validación.

4.1. BASE DE DATOS

Como ya se ha mencionado previamente mi estudio se basa en una base de datos pública y gratuita disponible en el sitio web Kaggle, llamada *"Stroke Prediction Dataset"* [17], cuyo autor es un usuario de la web llamado "Fedesoriano"[24]. El cual es Associate Team Leader en Oliver Wyman con experiencia en Data Science en KPMG y Management Solutions. También tiene experiencia en investigación en universidades prestigiosas y una sólida formación académica en Ciencia de Datos y Tecnologías de Big Data.[25]

La base de datos se estructura de la siguiente manera:

Cada fila contiene información relevante sobre un paciente (5110 filas) y cada columna corresponde a una variable (12 variables) siendo estas las siguientes:

Variable	Descripción	Tipo
id	Identificador único del paciente	Numérica
gender	Género del paciente ("Male", "Female" o "Other")	Carácter
age	Edad del paciente	Numérica

hypertension	Indica si el paciente tiene hipertensión (0 si no tiene, 1 si tiene)	Numérica
heart_disease	Indica si el paciente tiene una enfermedad cardíaca (0 si no tiene, 1 si tiene)	Numérica
ever_married	Indica si el paciente ha estado casado alguna vez ("No" o "Yes")	Carácter
work_type	Tipo de trabajo del paciente ("children", "Govt_jov", "Never_worked", "Private" o "Self-employed")	Carácter
Residence_type	Tipo de residencia del paciente ("Rural" o "Urban")	Carácter
nivel_glucosa	Nivel promedio de glucosa en sangre del paciente	Numérica
IMC	Índice de masa corporal del paciente	Carácter
smoking_status	Estado de fumador del paciente ("formerly smoked", "never smoked", "smokes" o "Unknown")	Carácter
stroke	Indica si el paciente ha tenido un derrame cerebral (1 si ha tenido, 0 si no)	Numérica

Tabla 2. Tipo de variables en la base de datos inicial.

4.2. IMPORTAR Y TRATAR LA BASE DE DATOS EN RSTUDIO

Empecemos con el desarrollo de mi propio estudio el cual realizaré usando RStudio, versión 4.3.3, valiéndome de diversas librerías indicadas en el anexo.

En primer lugar, para mejorar la limpieza de salida durante la ejecución, desactivo las advertencias y mensajes en R.

Para inspeccionar las primeras filas y comprender la estructura de los datos, importo mi conjunto de datos relacionado con la predicción de accidentes cerebrovasculares que obtuve de Kaggle desde un archivo CSV y luego lo visualizo en un formato tabular, fijándome también en el resumen estadístico de los datos, lo que permite una visión general y rápida sobre las variables.

Inicialmente, recuento las filas en las que hay valores faltantes en algunas variables, pero primeramente no se detecta ninguna. A continuación, elimino la primera columna, la cual corresponde al id de los pacientes, con la que no trabajaré, para simplificar el conjunto de datos al no considerarla relevante para mi análisis.

Tras esto cambio los nombres de las columnas para que sean más fáciles de leer a: "género","edad","hipertensión","enfermedades_vasculares","casado_alguna_vez", "tipo_trabajo", "tipo_residencia", " nivel_glucosa", "IMC" (Índice de Masa Corporal),"fumador" e "infarto" respectivamente.

Y posteriormente recodifico las variables categóricas:

-variable “género”, 1 <- Male, 2 <- Female, 3<- Other

-variable "casado_alguna_vez", 0<- No, 1<- Yes

-variable “tipo_trabajo”, 1<- children, 2 <- Govt_jov, 3<- Never_worked, 4<- Private, 5<- Self-employed

-variable “tipo_residencia”, 1<- Rural, 2<- Urban

-variable “fumador”, 1<- formerly smoked, 2<- never smoked, 3<- smokes, 4<- Unknown

-variable “hipertensión”, 1 <- el paciente presenta hipertensión, 0<- el paciente no presenta hipertensión

-variable “enfermedades_vasculares”, 1 <- el paciente presenta enfermedades vasculares, 0 <- el paciente no presenta enfermedades vasculares.

Además de transformar las variables “IMC” junto con “nivel_glucosa” y “edad” en numéricas.

Tras esto, convierto la variable cualitativa binaria "infarto" en un factor, ya que es la variable sobre la que trabajaré y me referiré a ella como “variable respuesta” en múltiples ocasiones.

Después de todos estos cambios y visualizando el resumen de variables destacamos que hay una distribución desigual en algunas variables, como el estado civil y el hábito de fumar. Además, llama la atención que solo hay un individuo en la categoría 3 de género, es por esto que decido eliminar esta categoría, así como el individuo que esta contenía no considerándolo significativo para mi estudio.

Recuerdo que previamente no fui capaz de identificar ningún dato faltante, pero tras todas estas modificaciones detecto 201 individuos (lo que supone el 3,93% de las observaciones iniciales) que presentan un valor vacío en la variable IMC. Para tomar una decisión sobre cómo proceder creo un dataframe llamado “filas_na_IMC” el cual solo contiene las filas de “Datastroke” que cumplen la condición de tener valor NA (abreviatura de “*Not Available”)* en la variable IMC. Hago esto con el fin de encontrar algo lo suficientemente destacable entre los individuos que la forman como para conservar dichas observaciones, rápidamente fijándonos en la variable respuesta “infarto” encuentro que 40 individuos de los 249 que presentan haber sufrido un infarto se encuentran entre estas observaciones, por lo que decido proseguir sin eliminar dichas filas, completando los datos faltantes usando un método de aprendizaje automático llamado *K Nearest Neighbors.*

El método de imputación conocido como “*K Nearest Neighbors*" (KNN), propuesto por Troyanskaya en 2001 [26], es una técnica efectiva para rellenar los datos faltantes. Consiste en asignar un valor estimado a cada valor perdido en una persona a partir de la información disponible de los k individuos más similares o cercanos, también conocidos como donantes o vecinos. La similitud se evalúa tomando todas las variables y tomando mediciones completas. KNN ayuda a preservar la estructura original de los datos y evita distorsiones en la distribución de la variable imputada al sustituir los datos faltantes por valores plausibles cercanos al real.

Tras realizar la imputación vuelvo a comprobar que ya no hay ningún dato faltante en ninguna de las variables y efectivamente compruebo que no.

Tras todos estos cambios considero que la base de datos ya está lista para trabajar sobre ella y la guardo como un archivo .rda con el nombre de “Datastroke_preprocesada”.

4.3. EXPLORATORY DATA ANALYSIS

El Análisis Exploratorio de Datos (EDA) es una fase crucial en cualquier proyecto que se base en un base de datos, ya que proporciona una comprensión profunda de la naturaleza y las características de los datos antes de aplicar modelos analíticos más avanzados. Este proceso implica explorar, visualizar y resumir los datos de manera sistemática para identificar patrones, tendencias, relaciones y posibles anomalías. Gracias al EDA puedo comprender la estructura y la distribución de los datos, pudiendo así seleccionar y aplicar las técnicas más apropiadas, para mi caso, lo que conduce a resultados más precisos y significativos.

Empezaré este análisis exploratorio visualizando por medio de histogramas la distribución de las variables numéricas, "Edad", "Nivel de Glucosa" e “IMC".

Una vez visto esto visualizo la relación entre variables numéricas y la variable respuesta en forma de *Boxplots.*

Prosigo visualizando la distribución de las variables categóricas, “género”, “hipertensión”, “enfermedades_vasculares”, “casado_alguna_vez”, “tipo_trabajo”, “tipo_residencia” y “fumador”, en este caso por medio de gráficos de barras.

Creo los gráficos de mosaico, los cuales muestran las relaciones entre las variables categóricas y la respuesta. Para mostrar la proporción de cada combinación de categorías, utilizando rectángulos. Estos me son útiles para encontrar patrones y asociaciones en los datos entre las diferentes categorías de las variables cualitativas con la ocurrencia de infartos.

Posteriormente me dispongo a crear tablas de contingencia y pruebas de asociación (prueba de chi cuadrado de Pearson) para las variables categóricas y la variable respuesta para poder comprender como estas se relacionan, las añado al anexo I en forma tabular para facilitar su lectura y comprensión. La tabla de contingencia muestra cómo se distribuyen las distintas variables cualitativas con respecto a la presencia o ausencia de infarto, lo que nos da una idea de las posibles asociaciones entre ellas. Por otro lado, la prueba de chi cuadrado de Pearson es especialmente adecuada para evaluar la asociación entre variables categóricas independientes y una variable de respuesta nominal como el infarto. Esta prueba compara las frecuencias observadas con las frecuencias esperadas bajo la hipótesis nula de independencia entre las variables, permitiendo así determinar la fuerza y significancia de estas asociaciones. Si por ejemplo estamos buscando si hay una relación entre ser hombre o mujer y tener un infarto. La prueba de chi cuadrado de Pearson nos dice si existe o no una relación significativa entre ambas variables. Esto es crucial en mi caso, ya que me permite identificar qué variables están más estrechamente vinculadas con el infarto y qué factores podrían ser predictores importantes en mi análisis. Así que, al combinar ambas herramientas, obtengo una comprensión más completa de cómo las variables cualitativas se relacionan con el infarto.

Debido a los resultados vistos hasta ahora me gustaría centrarme en visualizar la relación de la variable numérica edad con algunas otras que considero que se pueden estar viendo interferidas por esta. Visualizo así la relación entre las tres variables numéricas valiéndome de una matriz de

correlación entre variables numéricas y obteniendo valores de correlación aproximados. Genero gráficos de dispersión con líneas de ajuste no lineales para cada categoría de infarto visualizando así la relación no lineal entre las variables numéricas, Nivel de Glucosa, Edad y IMC con la variable respuesta, Infarto.

Referido a comprender de qué manera se relaciona o puede estar afectando la variable numérica edad con tipo de trabajo en primer lugar realizo un análisis de varianza (ANOVA) con el fin de encontrar diferencias significativas en la edad media entre las distintas categorías de la variable tipo de trabajo. Escojo la técnica estadística ANOVA ya que me permite comparar las medias de alguna variable continua (en mi caso la edad) de tres o más grupos (la variable tipo de trabajo tiene 5 categorías o grupos) haciéndome saber si al menos la media de edad de uno de estos grupos difiere significativamente de los demás. Tras esto creo un mapa de calor para la distribución de las edades medias de la variable categórica, tipo de trabajo, para entender cuáles son los grupos que se diferencian de los demás.

Me gustaría ver también como se comportan los dos grupos de pacientes definidos según si sufren o no infarto en términos de las variables numéricas, para esto realizo las pruebas t de Student y la prueba de Wilcoxon-Mann-Whitney. Asumiendo ciertos supuestos sobre la distribución de los datos, como la normalidad y la homogeneidad de las variaciones, utilizo la prueba t de Student para determinar si hay una diferencia significativa entre las medias de dos grupos independientes. La prueba de Wilcoxon-Mann-Whitney, por otro lado, es una alternativa no paramétrica a la prueba t; se utiliza cuando los datos no cumplen con estos supuestos. El uso de ambas pruebas me garantiza robustez en el análisis al considerar supuestos paramétricos y no paramétricos.

Tras todo lo visto hasta ahora decido enfocar mi estudio en una población más concreta con el fin de obtener mejores resultados. Más concretamente me centraré en aquellos pacientes que tengan una edad igual superior a los 35 años y que su situación laboral sea distinta a “nunca haber trabajado” o “dedicarse al cuidado de niños” puesto que son ocupaciones que se corresponden como he comprobado con edades más tempranas. Lo hago utilizando la función de indexación de filas para seleccionar las filas que cumplan con esas condiciones y asignarlas a Datastroke_filtrado. Primero, he eliminado las filas donde la edad es menor o igual a 34 y donde el tipo de trabajo es igual a 1 o 3. Luego, he eliminado las categorías 1 y 3 en la columna tipo_trabajo y, por último, he verificado que las categorías 1 y 3 en tipo_trabajo habían sido eliminadas usando la función *unique*.

4.4. DATOS, ENTRENAMIENTO, TEST, BALANCEO Y VALIDACIÓN

4.4.1. División de los datos.

En primer lugar, fijo una semilla (123) para controlar la aleatoridad y asegurar la reproducibilidad de mi estudio. Tras esto, utilizo la función *createDataPartition* para dividir el conjunto de datos (Datastroke_filtrado) en datos de entrenamiento (train_data) y datos de prueba (test_data) en una proporción de 60% y 40% respectivamente y visualizo un resumen de ambos conjuntos de datos para verificar que la división se ha realizado correctamente.

4.4.2. Balanceo de los datos de entrenamiento.

Tras esto para abordar el desbalanceo en la variable infarto en mi conjunto de datos de entrenamiento habiendo un 92.44% de casos sin infarto, decido solucionarlo utilizando una técnica llamada Random Over Sampling (ROS). Esta técnica consiste en aumentar el número de muestras de la clase minoritaria (en este caso, los pacientes que han sufrido un infarto), mediante la generación de nuevas muestras sintéticas a partir de las muestras existentes en esta clase. Solo llevo a cabo este balanceo en el conjunto de datos de entrenamiento con el propósito de usar el conjunto de prueba (test) como una representación de datos nuevos y no vistos. Si balancease los datos antes de esta división, podría introducir un sesgo que no estaría presente en datos reales, haciendo que el conjunto de prueba no fuese representativo de la realidad, lo que podría llevar a una sobreestimación del rendimiento del modelo.

Para hacerlo me valgo del paquete ROSE en R [27]. Gracias al cual uso una función que realiza el sobremuestreo aleatorio de la clase minoritaria, los pacientes que han sufrido un infarto, hasta igualar el número de muestras de la clase mayoritaria, es decir, aquellos que no han sufrido un infarto. Ajusto dicha función con ciertos argumentos definiendo la variable de respuesta (infarto ~ .), los datos (data =train_data), el método de sobremuestreo (method = "over") y una semilla aleatoria para reproducibilidad (seed = 12345). El resultado de esta función es un nuevo conjunto de datos balanceado, el cual guardo bajo el nombre de train_data_balanced. Tras esto compruebo que efectivamente el sobremuestreo se ha llevado a cabo según lo esperado y visualizo un resumen de los datos ya balanceados.

4.4.3. Validación de los modelos.

Para comparar modelos, debo asegurarme de que uso las mismas particiones de *train*/validation al realizar las *cross-validations*, ya que, de lo contrario, podría obtener resultados diferentes entre modelos que sean debidos al azar.

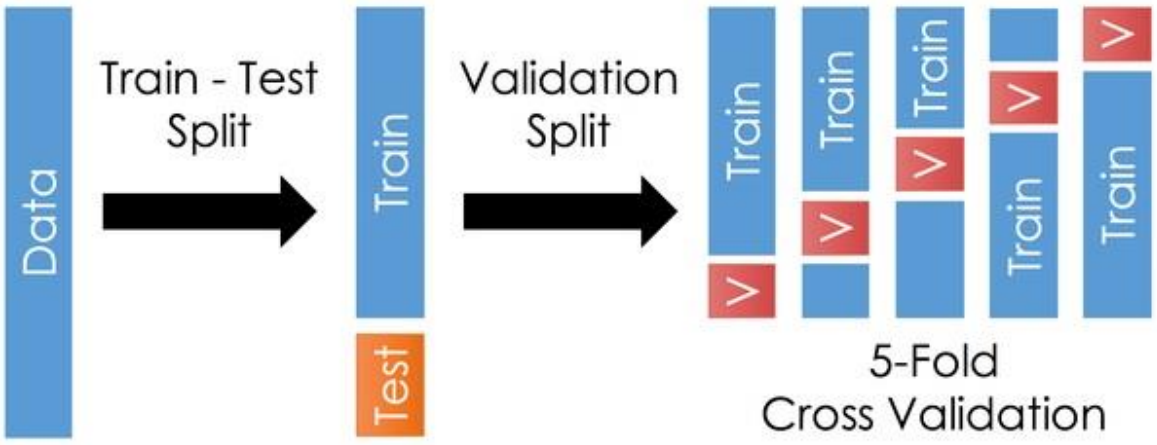


Figura 5. División de *train, test y validation data, 5-fold cross validation* [28]

En el proceso de construcción y evaluación de modelos para predecir infartos, es fundamental contar con una validación robusta que refleje con precisión el rendimiento del modelo en datos no vistos. Para lograr esto, he diseñado una estrategia de validación cruzada personalizada que se centra en minimizar los falsos negativos, es decir, los casos en los que el modelo predice incorrectamente que no habrá un infarto cuando sí ocurre.

Utilizaré la métrica F3, que es una medida que da más peso a la sensibilidad que a la precisión. Considerando que, en el contexto médico de la predicción de infartos, es crucial identificar correctamente la mayoría de los casos de infarto. La sensibilidad indica la capacidad del modelo para identificar correctamente los casos positivos, mientras que la precisión mide la proporción de predicciones positivas que son correctas. El valor F3 me permitirá encontrar un equilibrio óptimo entre estas dos métricas, poniendo tres veces más peso en la sensibilidad que en la precisión.

$$\mathrm{F}\beta = \ = \frac{(1+\beta^2)\cdot(\mathrm{precision}\cdot\mathrm{recall})}{\beta^2\cdot precision + recall}$$

En primer lugar, defino una métrica de evaluación personalizada para la validación cruzada. Calculo la precisión, la sensibilidad y el valor F3 para cada iteración de la validación. Tras esto establezco la validación cruzada con method = "cv" para utilizar validación cruzada estratificada. Se dividirá el conjunto de datos en 5 particiones (number = 5) y se especificará un índice de particiones (index = misFolds) para asegurar la coherencia en las particiones entre las iteraciones. La función de resumen personalizada “f3_summary” se utilizará para calcular las métricas durante la validación cruzada, centrándose en minimizar los falsos negativos y se probarán diferentes valores para los parámetros de cada modelo hasta determinar cuáles optimizan mejor el rendimiento (valor más alto de F3) del modelo que está siendo entrenado. Todo esto lo englobo en lo que llamo un control de entrenamiento (trctrl).

4.5. APLICACIÓN Y COMPARACIÓN DE ALGORTIMOS MACHINE LEARNING.

4.5.1. *Decision Tree*.

Un árbol de decisiones es un mapa que muestra las diferentes opciones y resultados posibles de una serie de decisiones relacionadas [29]. Te permite comparar acciones potenciales teniendo en cuenta sus costos, probabilidades y beneficios. Se usa para desarrollar métodos matemáticos que ayudan a predecir la mejor opción. Comienza con una sola opción que se divide en varios caminos que conducen a más opciones. Este tipo de mapas tienen tres tipos de puntos: los que muestran las probabilidades de resultados específicos, los que representan las decisiones que debes tomar y los que muestran el resultado final de un camino de decisión.

4.5.1.1. Aplicación del algoritmo *decision tree* a los datos de entrenamiento.

Recordemos que busco predecir la variable infarto a partir de todas las demás variables de los datos de entrenamiento.

En este caso uso modelo de árbol de decisión (decision tree) con los datos de entrenamiento balanceados. Este algoritmo me permite crear un modelo que toma decisiones basadas en una serie de reglas derivadas de los datos. Entreno el modelo utilizando el método “rpart”, que es una implementación eficiente de los árboles de decisión, configuro el hiperparámetro de complejidad (cp) en un rango de 0.01 a 0.1 con incrementos de 0.01, mediante el objeto “miGriddt”. Uso el control de entrenamiento previamente definido (trctrl) para asegurar una validación cruzada adecuada, y opto por maximizar el valor F3 (minimizar los falsos negativos).

4.5.1.1.1. Ajuste y optimización de los hiperparámetros del algoritmo *decision tree.*

En la búsqueda de mejorar el rendimiento del modelo de árbol de decisión, realizo ajustes y optimizaciones en los hiperparámetros, específicamente en el parámetro de complejidad (cp). Puesto que inicialmente con los parámetros definidos anteriormente, obtuve una tendencia desfavorable (la medida F3 disminuía drásticamente a medida que el cp aumentaba), decido afinar el rango y los incrementos del parámetro de complejidad. En la segunda iteración, pruebo un rango de cp más pequeño, de 0.001 a 0.01, con incrementos más finos de 0.0001.

4.5.2. *Random forest*.

El bosque aleatorio o *Random forest* es un algoritmo de aprendizaje automático muy popular que combina varios árboles de decisión para hacer una predicción final. Su popularidad se debe a su flexibilidad y capacidad para resolver problemas de clasificación y regresión. En resumen, el *random forest* crea un "bosque" de árboles de decisión, cada uno de los cuales está entrenado en una muestra aleatoria del conjunto de datos que se obtiene mediante bootsrap. El bootstrap en el contexto de un random forest se refiere a una técnica de muestreo utilizada para crear múltiples subconjuntos de datos a partir del conjunto de datos original. En cada iteración, se selecciona aleatoriamente, con reemplazo, una muestra de datos, permitiendo que algunos datos se repitan mientras otros puedan ser omitidos. Cada uno de estos subconjuntos se utiliza para entrenar un árbol de decisión individual en el random forest. [30]. Para obtener una predicción más precisa y confiable, luego combina las predicciones de esos árboles. Su habilidad para evitar el sobreajuste es una de sus principales ventajas porque promedia las predicciones de varios árboles no correlacionados. Además, es simple de usar y permite evaluar el valor de las variables del modelo. Este algoritmo se utiliza en una variedad de sectores, desde finanzas hasta atención médica y comercio electrónico, para ayudar en tareas como evaluar los riesgos crediticios, realizar diagnósticos médicos y recomendar productos.

4.5.2.1. Aplicación del algoritmo *random forest* a los datos de entrenamiento.

En este modelo de predicción de infartos, utilizo el algoritmo de *random forest* con los datos de entrenamiento balanceados. Primero, normalizo los datos centrándolos y escalándolos, asegurando que todas las variables tengan la misma importancia. Luego, utilizo el algoritmo “ranger”, una implementación eficiente del bosque aleatorio, para entrenar el modelo. Durante el entrenamiento, probaré cinco diferentes configuraciones del modelo (tuneLength = 5) para encontrar la mejor combinación de parámetros que maximice el valor de F3 o lo que es lo minimice los falsos negativos en la predicción de infartos y configuro una cuadrícula de hiperparámetros mediante miGridrf, que incluye un rango para mtry (número de predictores) de 6 a 15 y un rango de

min.node.size (número mínimo de observaciones que debe tener un nodo para que se le permita dividirse) de 1 a 5.

Para poder interpretar los resultados correctamente recordemos que "splitrule" es el criterio utilizado para decidir cómo dividir los nodos en un árbol de decisión dentro de un random forest, determinando qué división maximizará la precisión de las predicciones. Destaquemos dos de estas reglas:

Gini impurity (Impureza de Gini): La impureza de Gini informa sobre qué tan mezcladas están las diferentes clases en un nodo del árbol. Si un nodo tiene principalmente una sola clase, la impureza de Gini será baja. Pero si tiene una mezcla de diferentes clases, la impureza de Gini será alta. Entonces, al dividir un nodo en dos, el criterio de Gini busca la división que reducirá más la mezcla de clases en los nodos resultantes, tratando de hacerlos lo más puros posible en términos de la clase que contienen.

Extra-Trees: Los extra-trees son como una versión más rápida y menos exigente de los árboles de decisión normales. En lugar de buscar las mejores formas de dividir los datos, como lo hacen los árboles de decisión típicos, los árboles extra eligen divisiones al azar para cada característica, resultando en árboles más rápidos de entrenar. Aunque los árboles individuales pueden no ser muy precisos, juntar muchos de ellos en un Random Forest ayuda a compensar esto y a hacer predicciones más sólidas.

4.5.2.1.1. Ajuste y optimización de los hiperparámetros del algoritmo *random forest.*

A la vista de los resultados previos (aumento de la medida de F3 a partir de 4 predictores), decido refinar el rango del parámetro mtry, de 2 a 11, manteniendo las mismas reglas de división y tamaños mínimos de nodo.

4.5.3. Regresión logística regularizada.

La regresión logística regularizada es una técnica de machine learning utilizada para predecir resultados binarios, como "sí" o "no", (en nuestro caso sufrirá un infarto o no lo sufrirá). Funciona igual que la regresión logística normal, que calcula la probabilidad de un resultado usando una combinación de las características de entrada. Sin embargo, añade una penalización para mantener los coeficientes (pesos) del modelo pequeños [31]. Esto se hace para evitar que el modelo se ajuste demasiado a los datos de entrenamiento, lo que mejora su capacidad para hacer predicciones precisas en nuevos datos.

4.5.3.1. Aplicación del algoritmo Regresión logística regularizada a los datos de entrenamiento.

Aplico el modelo de regresión logística regularizada a mis datos de entrenamiento usando la librería caret, uso el control del entrenamiento definido previamente (como en los casos anteriores). Luego, entreno el modelo especificando el método glmnet y aplicando preprocesamiento para centrar y escalar los datos. Finalmente, ajusto los hiperparámetros hasta maximizar el valor F3, incluyo un rango para alpha de 0 a 1 y para lambda de 0.0001 a 0.1 y recalco la importancia de las variables basada en permutaciones, finalmente visualizo los resultados para evaluar el desempeño del modelo.

En cuanto a los parámetros alpha (α) y lambda (λ) podemos decir que son esenciales en la regresión logística regularizada. Alpha determina la proporción de las penalizaciones, un valor alto de alpha promueve la regularización Lasso, favoreciendo la selección automática de características al reducir algunos coeficientes a cero, mientras que un valor bajo de alpha prioriza la regularización Ridge, que mantiene todos los predictores en el modelo, pero controla la magnitud de sus coeficientes para evitar el sobreajuste. Por otro lado, lambda controla la fuerza global de la regularización; a medida que lambda aumenta, se incrementa la penalización sobre los coeficientes de los predictores, lo cual simplifica el modelo al reducir la influencia de variables menos relevantes o disminuir el tamaño absoluto de los coeficientes.

4.5.3.1.1. Ajuste y optimización de los hiperparámetros del algoritmo regresión logística regularizada.

A la luz de los resultados obtenidos con los parámetros definidos inicialmente para entrenar el modelo de regresión, decido refinar los rangos de los parámetros para una exploración más detallada, amplió el rango de alpha, manteniéndolo entre 0 y 1 pero dividido en 10 puntos e incremento significativamente la granularidad de lambda, manteniéndolo entre 0.0001 y 0.1 pero dividido en 100 puntos.

4.5.4. *Support Vector Machine*.

El Support Vector Machine (SVM) es un modelo de machine learning utilizado para clasificación y regresión. Funciona encontrando el hiperplano que mejor separa las distintas clases en el espacio de características, maximizando el margen entre las clases [32]. Utiliza funciones kernel para proyectar los datos en un espacio de mayor dimensión donde las clases sean linealmente separables. Los puntos de datos más cercanos al hiperplano, llamados vectores de soporte determinan su posición, lo que permite al modelo generalizar mejor con nuevos datos.

4.5.4.1. Aplicación del algoritmo *support vector machine* a los datos de entrenamiento.

Aplico el modelo SVM a mis datos de entrenamiento usando la librería caret, defino el control del entrenamiento con validación cruzada como ya he definido anteriormente. Luego, entreno el modelo especificando el uso del kernel radial y aplicando preprocesamiento para centrar y escalar los datos. Finalmente, ajusto los hiperparámetros configurando una cuadrícula mediante miGridsvm, que incluye un rango para sigma de 0.001 a 0.1 y para C de 0.1 a 1 y visualizo los resultados para evaluar el desempeño del modelo.

4.5.4.1.1. Ajuste y optimización de los hiperparámetros del algoritmo *support vector machine.*

A la vista de los resultados (al incrementar sigma, el valor de F3 también aumenta ligeramente para todos los valores de C y los valores más altos de C (0.9 y 1) logran los valores más altos de F3), decido refinar los rangos de sigma y C para una exploración más detallada, amplio el rango de sigma (0.5 a 1), dividido en 20 puntos e incremento el rango de C (3 a 8), dividido en 40 puntos.

4.5.5. Comparación y elección del modelo final.

Para seleccionar el mejor modelo entre una lista de modelos optimizados en R, utilizo el paquete caret, que facilita la comparación y evaluación de diferentes modelos a través de técnicas de *cross-validation*. Primero, creo una lista de los modelos optimizados que quiero comparar. En este caso, cuatro modelos: *random forest* optimizado(RF_OPTIMIZADO), *support vector machine* optimizado (SVM_OPTIMIZADO), *decision tree* (DT_OPTIMIZADO) y una regresión logística regularizada optimizada (RLR_OPTIMIZADO).

Utilizo la función resamples de caret para evaluar estos modelos. Esta función aplica técnicas de *cross-validation* para obtener una evaluación robusta del desempeño de cada modelo. Al hacerlo, genero un objeto que contiene los resultados de las métricas de rendimiento de cada modelo. A partir de estos resultados, utilizo la función summary para resumir las métricas de rendimiento, como F3-*score*, *precision* y *recall*.

Para visualizar estas métricas de manera clara, extraigo sus respectivas estadísticas y presento los valores medios (Mean), el valor mínimo, el valor máximo y el valor correspondiente al primer y tercer cuartil de cada métrica para cada modelo.

Al tener las métricas resumidas visualizo comparaciones gráficas utilizando bwplot y dotplot para las métricas F3 y *Recall*.

Finalmente, para seleccionar el mejor modelo, me enfoco en la métrica que he ido optimizando en todos los modelos, el F3-*score*. Extraigo los valores medios de F3 para cada modelo y comparo estos valores para determinar qué modelo tiene el valor más alto. El modelo con el mayor F3-score es seleccionado como el mejor modelo. En nuestro caso, al evaluar los resultados, determino que el modelo RF_OPTIMIZADO tiene el mayor F3-score, por lo que lo identificaré como el mejor modelo.

Después de identificar el mejor modelo entre el resto de modelos optimizados, el siguiente paso es extraer este modelo de la lista de modelos optimizados y también obtener el método utilizado para entrenarlo para poder utilizarlo en futuras predicciones. Esto lo hago asignando el mejor modelo a una nueva variable, llamada final_model. Sin embargo, no basta con tener el modelo en sí; también es importante saber cómo se entrenó. Por lo tanto, necesito conocer el método específico que se utilizó para entrenar este modelo ganador, para ello extraigo el método de entrenamiento utilizado para el mejor modelo y lo asigno a otra variable, final_method.

4.6. MEDICIÓN DEL RENDIMIENTO DEL MODELO FINAL MEDIANTE UNA MATRIZ DE CONFUSIÓN SOBRE LOS DATOS TEST Y SOBRE LOS DATOS ORIGINALES COMPLETOS SIN BALANCEAR MEDIANTE CROSS VALIDATION.

Evaluo en primera instancia el modelo final seleccionado utilizando los datos de prueba (test_data). Primero, creo predicciones sobre estos datos utilizando el modelo elegido previamente como el mejor. Luego, comparo estas predicciones con las etiquetas reales para calcular una matriz de confusión que muestra la distribución de las predicciones correctas e incorrectas junto con métricas clave como precisión y exhaustividad (*recall*). Finalmente, calculo el F3-Score.

Para medir el rendimiento del modelo final mediante una validación cruzada sobre los datos originales completos, utilizo la función train del paquete caret en R. El modelo final, previamente seleccionado, es entrenado nuevamente utilizando la variable objetivo infarto y las características predictivas del conjunto de datos Datastroke_filtrado (no está balanceado ni dividido en *train* y *test*). Mantengo las mismas configuraciones de preprocesamiento, control de entrenamiento y métrica de evaluación (F3) que en la selección del modelo. Además, ajusto los parámetros del modelo final utilizando la cuadrícula óptima previamente determinada mediante final_model$bestTune. Este enfoque me permite evaluar el rendimiento del modelo final sobre los datos test y los datos originales completos, garantizando una evaluación objetiva y precisa de su capacidad predictiva en la detección de infartos.

4.7. DESARROLLO DE APLICACIÓN PREDICTORA DE INFARTOS BASADA EN EL MODELO SELECCIONADO.

Decido crear una aplicación web interactiva utilizando R y el paquete shiny para ayudar a predecir el riesgo de infarto en base a factores de salud específicos que serían las variables. Esta es la última fase de mi apartado de metodología y suponiendo un escenario ideal en el que en España lo pudieran usar los médicos para apoyarse en sus diagnósticos, aunque hay que ser conscientes de que para que esto fuera posible habría primero que plantear un ensayo clínico, que lo apruebe un Comité Ético, ejecutarlo y luego conseguir la aprobación de la agencia española de medicamentos y productos sanitarios (AEMPS).

He diseñado varios campos de entrada en la interfaz de usuario (UI) para que los usuarios ingresen información relevante en dos columnas pudiendo ingresar su respuesta a las variables numéricas de forma manual, como su edad, nivel de glucosa e índice de masa corporal (IMC), y a la hora de ingresar su información sobre las variables categóricas puedan seleccionar la opción más adecuada para ellos por medio de desplegables como en el caso de género, edad, estado civil, tipo de trabajo y tipo de residencia y su estado como fumador, además de un panel donde se mostrará el resultado de la predicción, acompañado de algunas recomendaciones sobre cómo actuar, en cualquier caso. Todo esto traté de hacerlo a través de un diseño que fuera fácil de entender y atractivo.

En lo que respecta a la lógica del servidor, creé una función llamada predict_infarto que recibe los valores del usuario, los transforma en un *dataframe* y luego utiliza el modelo que he desarrollado previamente entrenado (caret.ranger.simple) basado en el modelo final de *random forest* seleccionado previamente para realizar la predicción del riesgo de infarto. Luego programé la aplicación para que reaccionara a los cambios en las entradas del usuario y mostrara el resultado de la predicción en tiempo real.

Finalmente, utilicé la función shinyApp, que combina la lógica del servidor y la interfaz de usuario para crear una aplicación web interactiva. Esta aplicación se puede ejecutar localmente en RStudio o desplegar en un servidor de un servidor web.

5. RESULTADOS

5.1. RESULTADOS DEL *EXPLORATORY DATA ANALYSIS*

5.1.1. Visualización y comparación para cada variable numérica no estandarizada de: su distribución mediante un histograma y su relación con la variable respuesta usando *boxplot*.

5.1.1.1. Distribución Edad y Edad VS Infarto.

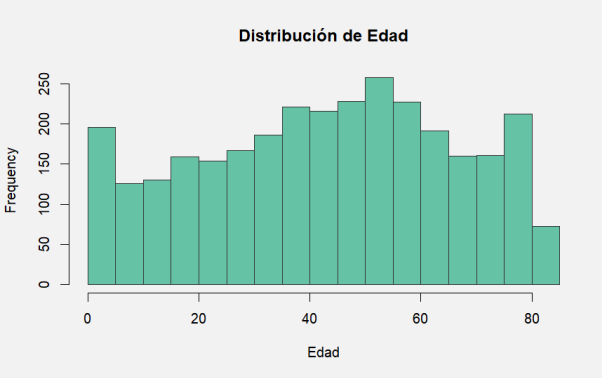


Figura 6. Histograma distribución variable edad.

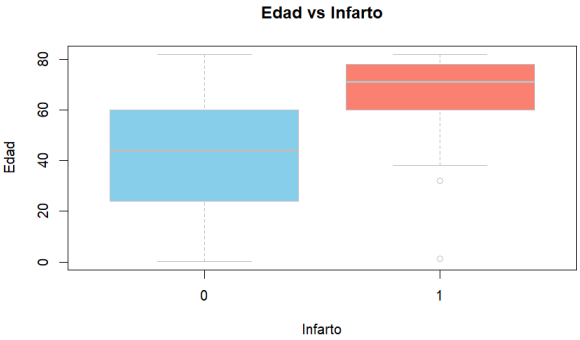


Figura 7. *Boxplot* edad vs. infarto.

Los individuos de nuestra base de datos tienen una amplia gama de edad, desde los bebés hasta los adultos mayores, con una edad mínima de 0,08 años y una edad máxima de 82 años. Sin embargo, los datos de cuartiles muestran que la mayoría de las personas se concentran en la franja de edad entre los 25 y 60 años.

En el caso del *boxplot* de Edad vs Infarto podemos ver como claramente hay una diferencia respecto a la edad de aquellos que sufren o no un infarto, estando el 50% de los individuos que no han sufrido un infarto entre los 25 y casi 60 años mientras que la mitad de los individuos que sí han sufrido un infarto tienen una edad entre 60 y 80 años, vemos un par de valores atípicos señalizados por puntos en edades más bajas, podríamos considerarlos casos aislados.

Estos resultados me conducen a empezar a plantearme si realmente tener una base de datos con un rango de edad tan amplio me beneficiará en mi estudio, sobre todo viendo que la mayor parte de los pacientes que sufren infartos tienen una edad media considerablemente alta y precisamente hay muchos menos individuos que han sufrido infarto en mi base de datos que pacientes que no lo han sufrido.

5.1.1.2. Distribución Nivel de Glucosa y Nivel de glucosa VS Infarto.

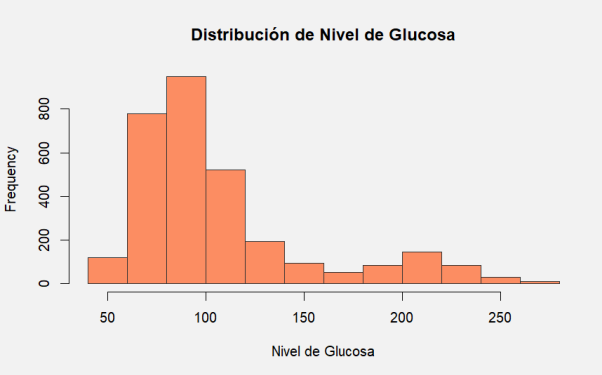


Figura 8. Histograma distribución variable nivel de glucosa

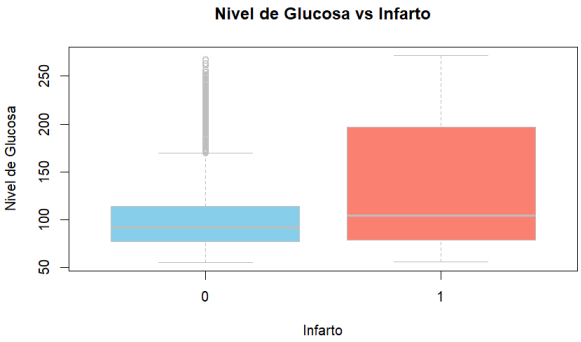


Figura 9. Boxplot nivel de glucosa vs. infarto.

Se observa una variabilidad significativa en los niveles de glucosa en sangre, con valores mínimos de 55.12 hasta valores máximos de 271.74. La mayoría de las personas tienen niveles de glucosa entre el primer y el tercer cuartil, entre 77.07 y 113.50, aunque la distribución tiende a sesgarse hacia valores más altos.

En cuanto a la relación entre el nivel de glucosa y la variable respuesta vemos una gran diferencia en cuanto a la variabilidad, habiendo mucha más variabilidad en los valores de nivel de glucosa de los pacientes que sí han sufrido un infarto frente a los que no, que se agrupan en valores mucho más bajos la mayoría. Sin embargo, debemos destacar la señalización de varios valores atípicos en valores altos de nivel de glucosa dentro del grupo de individuos que no han sufrido un infarto.

5.1.1.3. Distribución IMC y IMC VS Infarto.

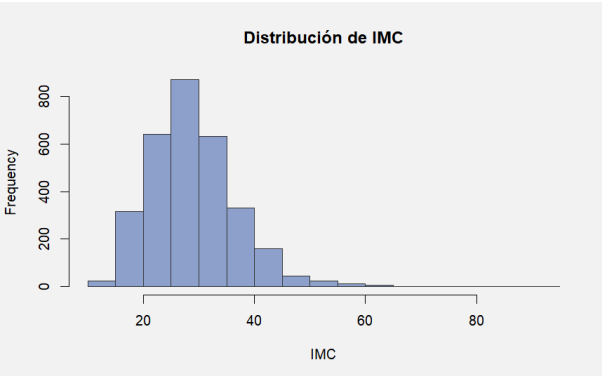


Figura 10. Histograma distribución variable IMC

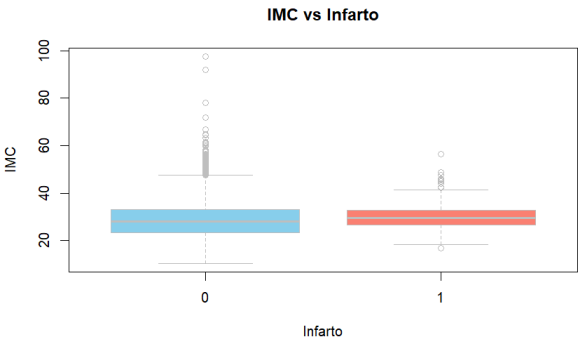


Figura 11. Boxplot IMC vs. infarto.

Por último, pero no menos importante, el índice de masa corporal (IMC), una medida del peso relativo en relación con la altura muestra una distribución más equilibrada en comparación con las variables anteriores. La mayoría de las personas tienen un IMC entre 23.50 y 33.10, aunque los valores varían desde 10.30 hasta 97.60, lo que indica una concentración en el rango de sobrepeso a obesidad, con algunos valores atípicos en el extremo superior.

Visualizando el *boxplot* referente a la relación entre el IMC y la variable respuesta sorprendentemente y en contra de lo esperado vemos que no podemos sacar ninguna conclusión significativa que diferencie a los pacientes que han sufrido infartos de los que no, puesto que la mayoría se engloban en los mismos valores de IMC (valores bajos) habiendo por supuesto excepciones e individuos con valores a típicos más altos representados como círculos.

Esto nos hace otra vez replantearnos la distribución de nuestros datos así como si realmente todas nuestras observaciones son igual de relevantes, si alguna variable está interfiriendo en otra o si quizás debería acotar el rango de ciertas variables.

5.1.2. Visualización y comparación para cada variable categórica de: su distribución mediante un gráfico de barras y visualización de la relación con la variable infarto usando gráficos de mosaico

5.1.2.1. Visualización de la distribución de la variable categórica género mediante gráfico de barras y gráfico de mosaico.

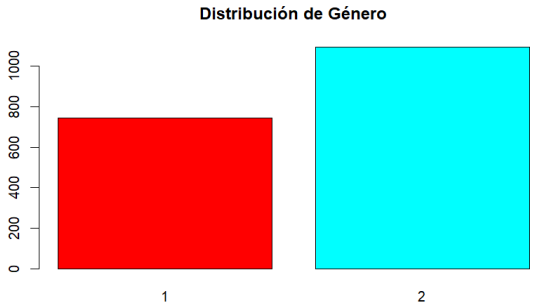


Figura 12. Gráfico de barras distribución variable Género

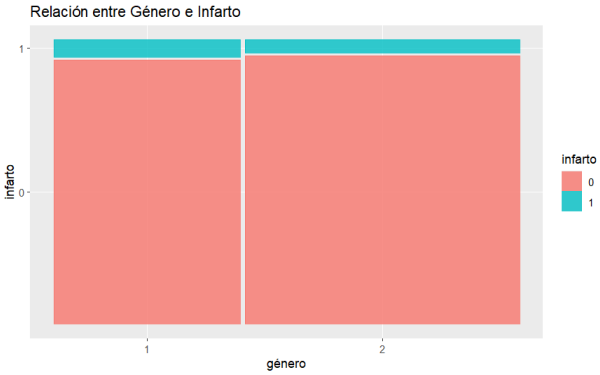


Figura 13. Gráfico de mosaico relación género e infarto.

Con 2115 y 2994 observaciones, el género de las personas de nuestra base de datos muestra una distribución bastante equilibrada entre las categorías masculina y femenina, aunque hay más observaciones de sujetos que sean mujeres. No parece que haya diferencias en cuanto a la proporción de infartos entre hombres y mujeres.

5.1.2.2. Visualización de la distribución de la variable categórica hipertensión mediante gráfico de barras y gráfico de mosaico.

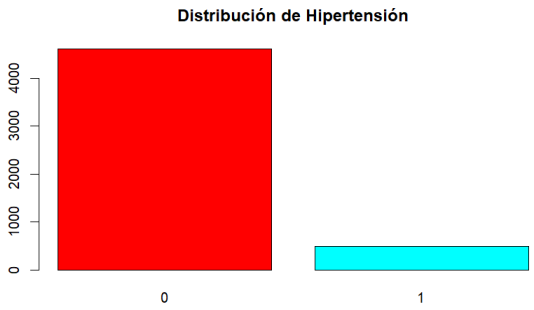


Figura 14. Gráfico de barras distribución variable Hipertensión

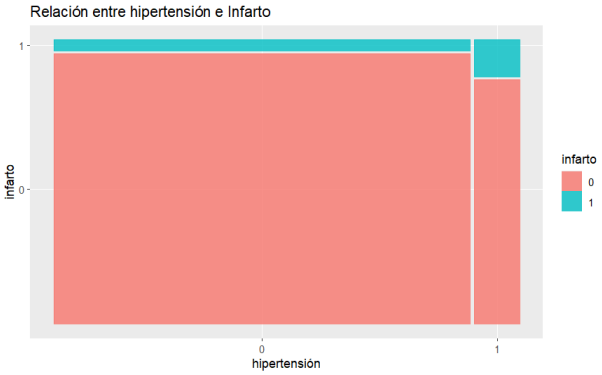


Figura 15. Gráfico de mosaico relación hipertensión e infarto.

La variable de hipertensión muestra si una persona tiene o no hipertensión, con una gran mayoría de personas (4611 observaciones) que no tienen hipertensión en comparación con 498 que sí la tienen. Se observa que hay claramente más infartos (en porcentaje) en los hipertensos que en los que no.

5.1.2.3. Visualización de la distribución de la variable categórica enfermedades cardiovasculares mediante gráfico de barras y gráfico de mosaico.

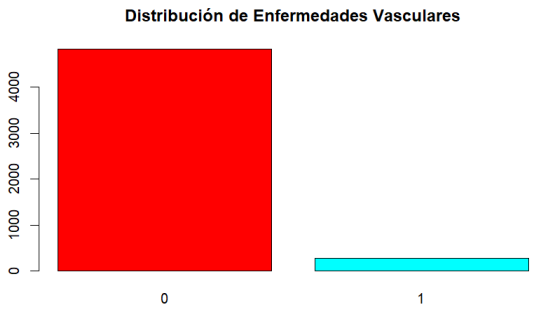


Figura 16. Gráfico de barras distribución variable enfermedades vasculares

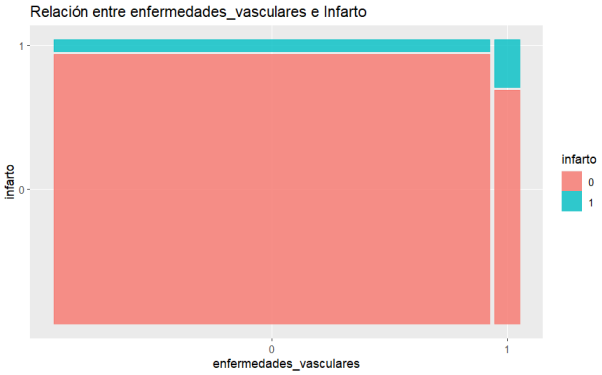


Figura 17. Gráfico de mosaico relación enfermedades vasculares e infarto.

La mayoría de las personas (4833 observaciones) no tienen enfermedades cardiovasculares, mientras que 276 sí. Vemos mayor porcentaje de infartos en las personas con enfermedades vasculares que en las que no las tienen.

5.1.2.4. Visualización de la distribución de la variable categórica casado alguna vez mediante gráfico de barras y gráfico de mosaico.

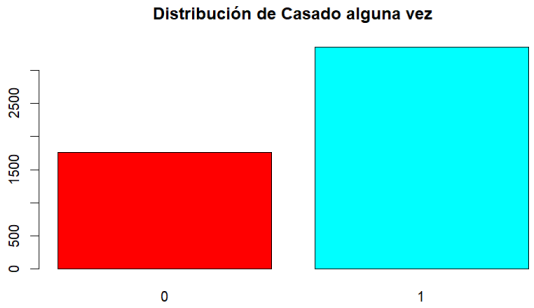


Figura 18. Gráfico de barras distribución variable casado alguna vez

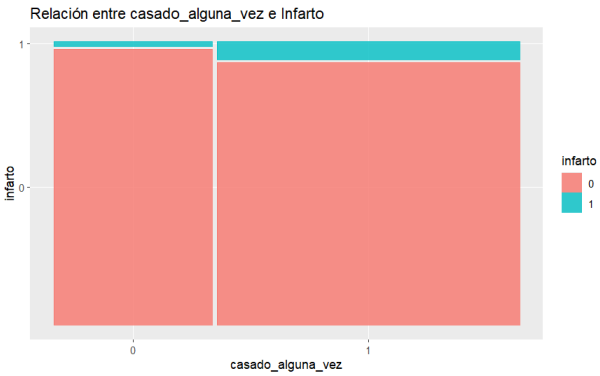


Figura 19. Gráfico de mosaico relación casado alguna vez e infarto.

La variable que indica si las personas han estado casadas alguna vez muestra una proporción mayoritaria de personas casadas en comparación con personas sin casarse, con 3353 observaciones frente a 1756. Apreciamos como los pacientes casados a nivel descriptivo presentan un mayor porcentaje de infartos que los que no.

5.1.2.5. Visualización de la distribución de la variable categórica tipo de trabajo mediante gráfico de barras y gráfico de mosaico.

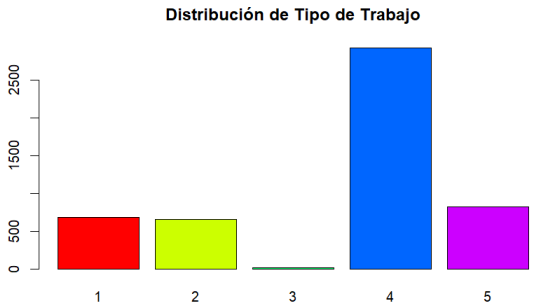


Figura 20. Gráfico de barras distribución variable tipo de trabajo

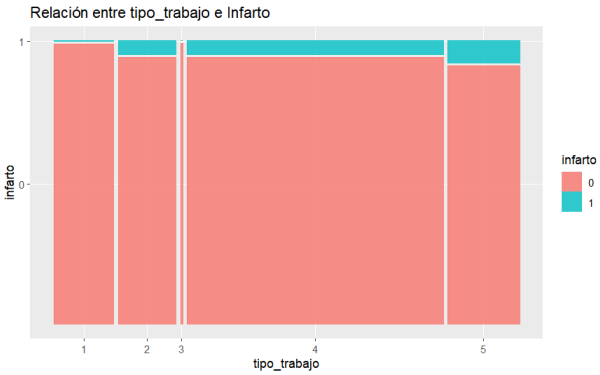


Figura 21. Gráfico de mosaico relación tipo de trabajo e infarto.

En cuanto a la distribución del tipo de trabajo vemos como la mayor parte de nuestra población se dedica al sector privado (2924 individuos) mientras que el caso menos común es el de las personas que nunca han trabajado (22 individuos). Parece que a nivel descriptivo el mayor porcentaje de infartos se da en los de tipo de trabajo 5 (autónomos) mientras que en las categorías 1 y 3, nunca ha trabajado o se dedica al cuidado de los niños apenas hay presencia de infartos.

5.1.2.6. Visualización de la distribución de la variable categórica tipo de residencia mediante gráfico de barras y gráfico de mosaico.

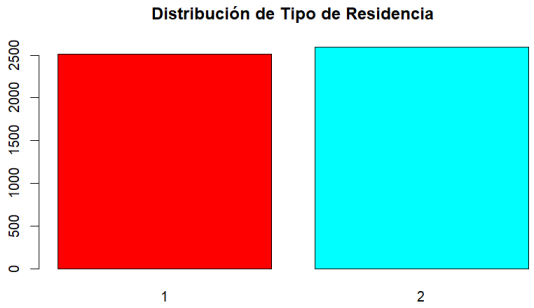


Figura 22. Gráfico de barras distribución variable tipo de residencia

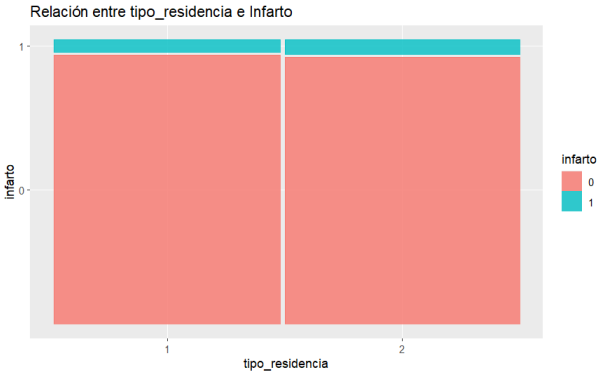


Figura 23. Gráfico de mosaico relación tipo de residencia e infarto.

Podemos ver como las dos categorías de la variable tipo de residencia están prácticamente igualadas en frecuencia, podemos asumir que esta variable actúa como una variable control, se han recogido el mismo número de observaciones en zonas rurales que en zonas urbanas premeditadamente, reduciendo el riesgo de sesgo en el presente análisis. No vemos diferencia en cuanto a infarto entre los que viven en áreas rurales y urbanas.

5.1.2.7. Visualización de la distribución de la variable categórica fumador mediante gráfico de barras y gráfico de mosaico.

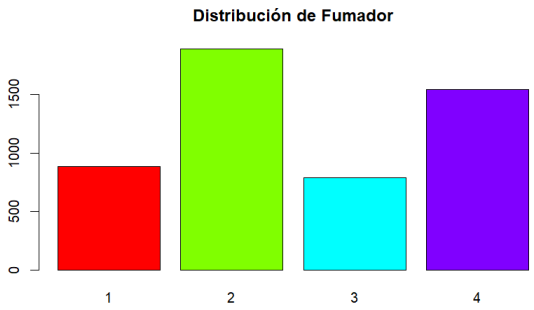


Figura 24. Gráfico de barras distribución variable fumador

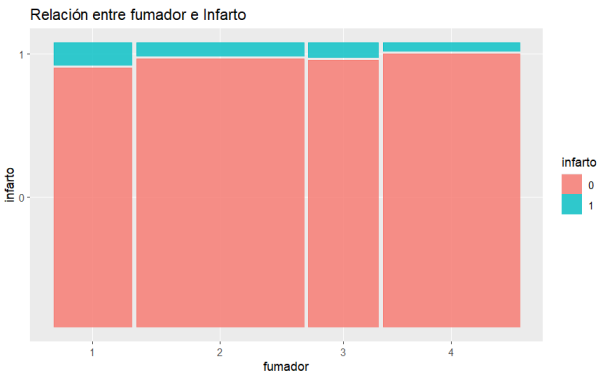


Figura 25. Gráfico de mosaico relación fumador e infarto.

Fijándonos en la variable fumador podemos ver dos categorías muy iguales, sorprendentemente son las que corresponden a fumaba antes y fuma actualmente, sin embargo, la más común entre nuestros individuos es nunca ha fumado. Aparentemente la categoría 1 (antes era fumador) es la que mayor porcentaje de infartos presenta. Recordemos que hay muchos individuos sobre los que desconocemos su estado de fumador (categoría 4).

5.1.3. Tablas de contingencia y pruebas de asociación- prueba de Chi-cuadrado de Pearson para variables categóricas y respuesta:

Recordemos en primer lugar la estructura de la tabla de contingencia, la cual muestra la distribución de los casos de infarto según las distintas categorías de las variables categóricas. Las filas

representan las categorías y las columnas representan la presencia o ausencia de infarto (codificadas como 0 y 1, respectivamente).

Podemos encontrar detalladamente las tablas de contigencia y las pruebas de asociación de Chi-cuadrado en el Anexo I, veamos lo más destacable de estas:

Tanto en la variable hipertensión como en la variable enfermedades cardiovasculares, en la variable fumador y en la variable casado alguna vez podemos ver como todas ellas muestran una asociación significativa con el infarto (p-valor < 0.001), con valores estimados de aproximadamente < 2.2e-16, < 2.2e-16, 2.008e-06 y 1.686e-14 respectivamente y valores relativamente altos para el estadístico X^2 llegando a alcanzar valores de 81.57 y 90.229 para las relaciones de las variables hipertensión y ECV con infarto respectivamente. Estos valores de X^2 indican que las diferencias entre las frecuencias observadas y esperadas bajo la hipótesis nula de independencia son muy grandes, lo que sugiere asociaciones fuertes y significativas entre estas variables y el riesgo de infarto.

Por otro lado, también me gustaría hacer cierto hincapié en la variable tipo de trabajo y cómo se relaciona con la variable respuesta. La prueba chi cuadrado muestra una asociación significativa entre el tipo de trabajo y el infarto (p-valor = 5.409e-10. Sin embargo, la tabla de contingencia revela que las categorías 1 (niños) y 3 (nunca trabajó) tienen números muy bajos de infartos, lo que sugiere que estas categorías pueden no ser representativas o pueden estar asociadas con otras variables, sospecho que esta podría ser la edad puesto que el nunca haber trabajado o el cuidar niños suele ser la situación de muchos jóvenes que aún no han llegado a la edad en la que poder trabajar, o cómo aún son jóvenes dedican sus esfuerzos a los estudios y algunos ganan sus primeros sueldos trabajando como canguro de niños.

Es por todo esto que debido a los resultados encontrados previamente y los recientes cada vez estoy más de acuerdo con la idea de centrarme en pacientes con edades más tardías, para llevar a cabo mi proyecto de forma más precisa y eficiente, pero antes de tomar ninguna decisión precipitada voy a estudiar la relación de la variable Edad con las dos variables que creo que se pueden estar viendo interferidas por ellas, las variables numéricas y tipo de trabajo.

5.1.4. Matriz de correlación entre variables numéricas.

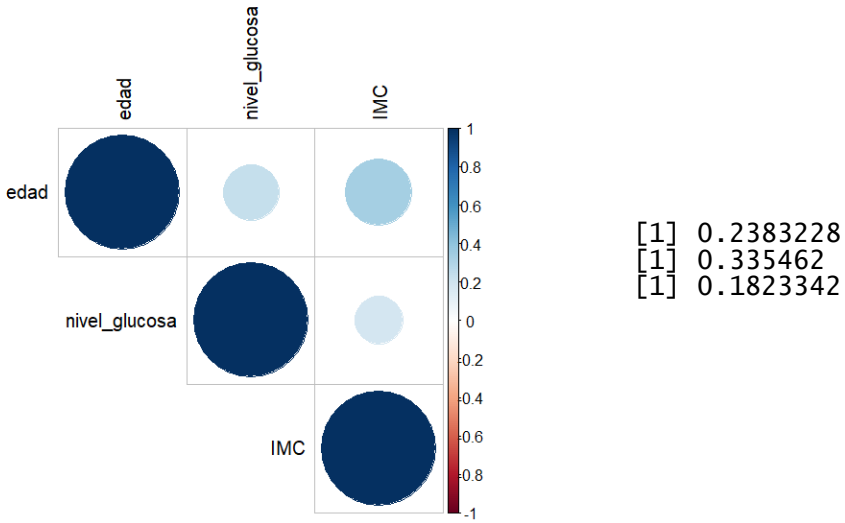


Figura 26. Matriz de correlación y coeficientes de correlación aproximados entre variables numéricas

La correlación entre la edad y el nivel medio de glucosa en sangre (nivel_glucosa) es de 0.238, muestra una correlación positiva débil entre la edad y el nivel medio de glucosa en sangre. Es decir, a medida que aumenta la edad, tiende a aumentar ligeramente el nivel medio de glucosa en sangre, pero la relación no es muy fuerte.

La correlación entre la edad y el índice de masa corporal (IMC) es de 0.335, muestra una correlación positiva entre la edad y el índice de masa corporal, como en el caso anterior, a medida que aumenta la edad, tiende a aumentar el índice de masa corporal, pero la relación no es muy fuerte.

La correlación entre el nivel medio de glucosa en sangre (nivel_glucosa) y el índice de masa corporal (IMC) es de 0.182. Muestra correlación positiva débil entre el nivel medio de glucosa en sangre y el índice de masa corporal. En resumen, hay una ligera tendencia a que los niveles más altos de glucosa en sangre se asocien con un índice de masa corporal más alto, pero la relación no es muy fuerte.

5.1.5. Relación no lineal entre variables numéricas y respuesta mediante gráfico de dispersión con líneas de ajuste no lineales para cada categoría de infarto.

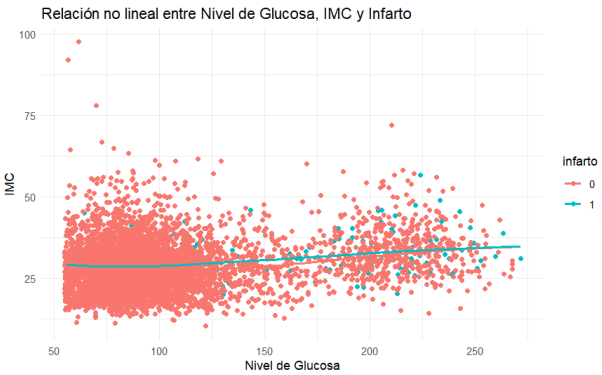


Figura 27. Gráfico de dispersión, Nivel de Glucosa, IMC e Infarto

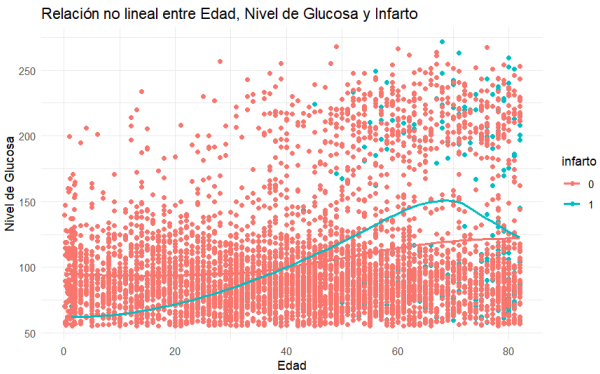


Figura 28. Gráfico de dispersión, Nivel de Glucosa, Edad e Infarto

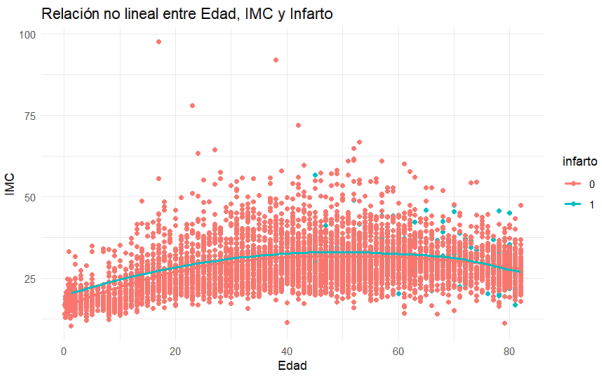


Figura 29. Gráfico de dispersión, Edad, IMC e Infarto

Relación no lineal entre Nivel de Glucosa, IMC e Infarto. En general podemos decir que las líneas se mantienen bastante constantes horizontalmente a lo largo del gráfico lo que nos indica que

la variable infarto no está influenciada por las variables numéricas (nivel de glucosa e IMC) o que la relación entre estas variables es muy débil.

Relación no lineal entre Edad, Nivel de Glucosa e Infarto, podemos ver que la línea azul en el gráfico de la relación no lineal entre Edad, Nivel de Glucosa e Infarto muestra una tendencia ascendente hasta la edad de 70 años y luego comienza a descender. Indicando que, hasta esa edad, el riesgo de infarto aumenta a medida que la edad y los niveles de glucosa en sangre aumentan. A partir de los 70 años, la tendencia descendente podría indicar que otros factores, como el cuidado de la salud, el tratamiento médico o la selección natural, podrían estar influyendo en la disminución del riesgo de infarto. También debemos destacar la presencia de muchos puntos dispersos por encima de las líneas de ajuste, especialmente en los últimos valores de edad, lo que indica una mayor variabilidad en los datos en esa región

Relación no lineal entre Edad, IMC e Infarto. Podemos observar una leve forma cóncava de las líneas, lo podría indicar que, hay una asociación creciente entre la edad y el IMC en relación con la probabilidad de infarto hasta cierto punto, seguida de una asociación decreciente después de ese punto. Esto podría sugerir que, para ciertos grupos de edad y rangos de IMC, el riesgo de infarto disminuye con la edad, pero para otros grupos de edad y rangos de IMC, el riesgo de infarto aumenta con la edad.

5.1.6. Pruebas t de student y Wilcoxon-Mann-Whitney en variables numéricas: edad, imc y nivel de glucosa con la variable infarto.

Los resultados detallados de ambas pruebas se encuentran en el Anexo I. En cuanto a la edad, tanto la prueba t de Student como la de Wilcoxon-Mann-Whitney devolvieron valores de p extraordinariamente bajos (< 2.2e-16), indicando una marcada diferencia en las edades entre los grupos con y sin infarto, con el grupo afectado exhibiendo una media de edad considerablemente mayor (67.73 años) en comparación con el grupo no afectado (41.97 años). Similarmente, para el Índice de Masa Corporal (IMC), las pruebas t y de Wilcoxon-Mann-Whitney mostraron valores de p muy bajos (0.000221 y 4.828e-05 respectivamente), demostrando una diferencia significativa en los IMC entre ambos grupos, con el grupo afectado teniendo un IMC medio notablemente más alto (30.34) en comparación con el grupo no afectado (28.88). En cuanto al nivel de glucosa, ambas pruebas exhibieron valores de p extremadamente bajos (2.373e-11 y 3.583e-09 respectivamente), lo que sugiere una marcada disparidad en los niveles de glucosa entre los dos grupos, con el grupo afectado mostrando un nivel de glucosa medio más elevado (132.54) en comparación con el grupo no afectado (104.79). Estos resultados recalcan la importancia de la edad, el IMC y el nivel de glucosa como factores de riesgo para la ocurrencia de infarto.

5.1.7. ANOVA y mapa de calor para la distribución de las edades medias de la variable categórica tipo de trabajo.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tipo_trabajo	4	1215	328	303832	1110	<2e-16 ***
Residuals	5104	1396	769	274		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tabla 3. Prueba ANOVA edad media y tipo de trabajo.

Los resultados del ANOVA muestran que hay una diferencia significativa en la edad entre las categorías de tipo_trabajo (DF(4, 5104) = 1110, p < 2e-16). Esto indica que al menos una de las categorías de tipo_trabajo tiene una edad media significativamente diferente de las demás. El p-valor extremadamente pequeño (<2e-16) sugiere que esta diferencia no es debida al azar y es muy significativa estadísticamente.

Veamos cuales son aquellas categorías que difieren del resto en términos de edad mediante un mapa de calor:

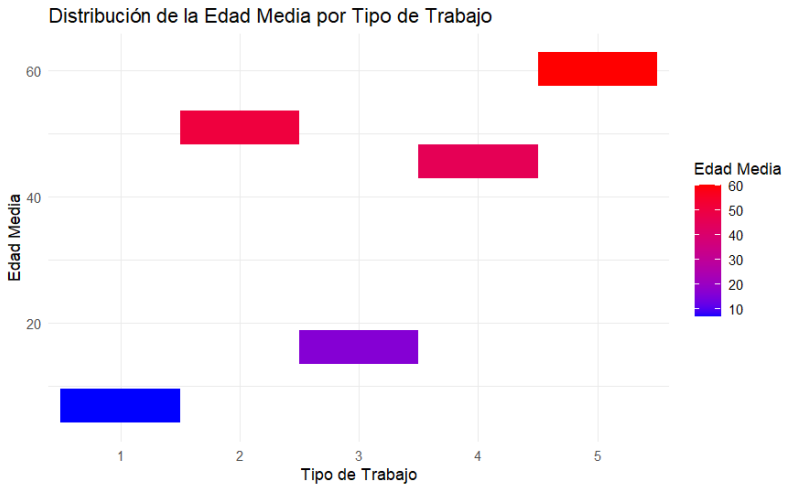


Figura 30. Mapa de calor para la distribución de las edades medias de la variable categórica tipo de trabajo.

Tal y como era de esperar tras los resultados de las tablas de contingencia y la prueba de chi cuadrado de Pearson, las categorías 1 y 3 correspondientes a las situaciones, “trabajar cuidando niños” y “el paciente nunca ha trabajado”, respectivamente, presentan unas edades medias por debajo de los 20 años, pacientes que como hemos visto previamente no suelen presentar infartos.

5.1.8. Filtrado de los datos eliminando observaciones correspondientes a: pacientes menores de los 35 años y pacientes en situación laboral de nunca haber trabajado o dedicarse al cuidado de niños.

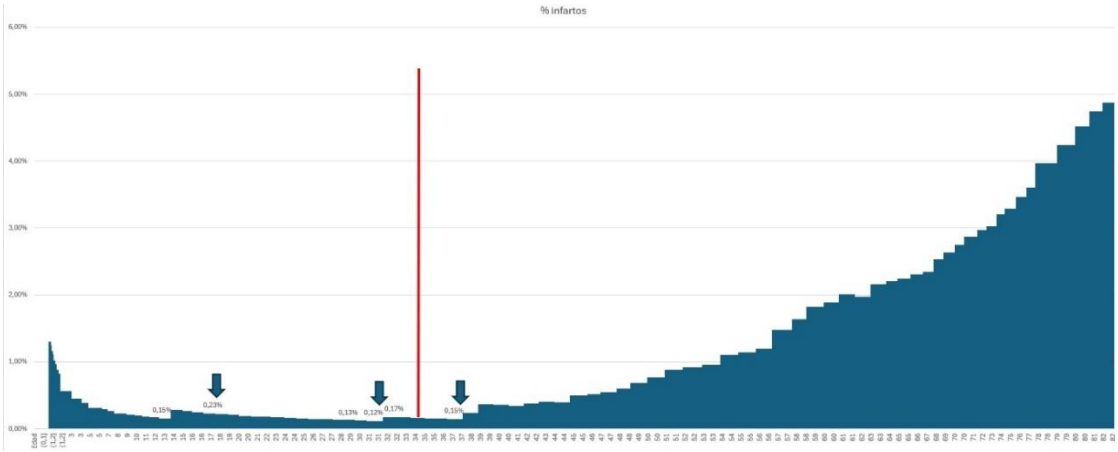


Figura 31. Gráfico de porcentaje acumulado de infartos por edad

La decisión de utilizar únicamente observaciones de pacientes de 35 años o mayores se fundamenta en la necesidad de eliminar situaciones excepcionales que pueden generar ruido en el análisis y distorsionar las relaciones observadas. Al observar el gráfico que muestra el porcentaje acumulado de infartos en función de la edad, se identifican casos anómalos de infarto a edades muy tempranas señaladas con flechas, que no representan la problemática principal de interés del estudio. Esta exclusión permite concentrarme en la población en la que realmente existe un riesgo significativo de infarto, lo que resulta en resultados más relevantes y aplicables para el objetivo del análisis.

La base de datos filtrada por edad y tipo de trabajo tras ser tratada de la misma forma que ha sido tratada la base de datos inicial consta de 3265 observaciones y 11 variables. La mayoría de los pacientes son del género femenino, la edad promedio es de aproximadamente 57 años, la mayoría no tienen hipertensión ni enfermedad cardiovascular, la mayoría han estado casados alguna vez, y la mayoría reside en áreas urbanas. Además, el nivel de glucosa en sangre y el índice de masa corporal (IMC) varían considerablemente entre los pacientes, con valores que van desde mínimos hasta máximos significativos. Por último, la mayoría de los pacientes nunca han fumado. He de recalcar que la proporción de pacientes que no han sufrido un infarto es del 92.46%. Encontrará el análisis exploratorio de los datos filtrados en el Anexo I.

5.2. DATOS, ENTRENAMIENTO, TEST, BALANCEO Y VALIDACIÓN.

Tras realizar la división estratificada en función de la variable infarto de la base de datos en los 2 grupos, entrenamiento y test, estos subgrupos quedan de la siguiente manera:

	TRAIN DATA	**TEST DATA**
CANTIDAD DE FILAS	1960	1305
CANTIDAD DE COLUMNAS	11	11
GÉNERO	1:803 / 2:1157	1:530 / 2:775
EDAD (AÑOS)	Min.: 35.00, Max.: 82.00	Min: 35, Max: 82
HIPERTENSIÓN (0/1)	0:1664 / 1: 296	1122 / 183
ENFERMEDADES CARDIOVASCULARES (0/1)	0:1784 / 1: 176	1208 / 97

CASADO ALGUNA VEZ (0/1)	0:187 / 1:1773	124/ 1181
TIPO DE TRABAJO	1:322 / 2:1176 / 3: 462	1:232 / 2:791 / 3: 282
TIPO DE RESIDENCIA	1:963 / 2:997	1: 639, 2: 666
NIVEL GLUCOSA	Min.: 55.22, Max.: 271.74	Min: 55.28, Max: 263.56
ÍNDICE DE MASA CORPORAL	Min: 14.20, Max: 71.90	Min: 11.30, Max: 92.00
FUMADOR (1/2/3/4)	417 / 797 / 370 / 376	1:329 / 2:519 / 3:204/ 4:253
INFARTO (0/1)	0:1812 / 1:148	0:1207 / 1:98

Tabla 4. Tabla comparativa datos de Entrenamiento, validación y test.

Lo más destacable es que la partición se ha hecho en función de la variable respuesta y los porcentajes de las categorías con infartos y sin infarto en train data y test data son:

TRAIN DATA		TEST DATA	
0	1	0	1
0.9244898	0.0755102	0.92490421	0.07509579

Tabla 5. Tabla de porcentajes de frecuencia de la variable infarto en train y test data.

5.2.1. Balanceo de los datos de entrenamiento.

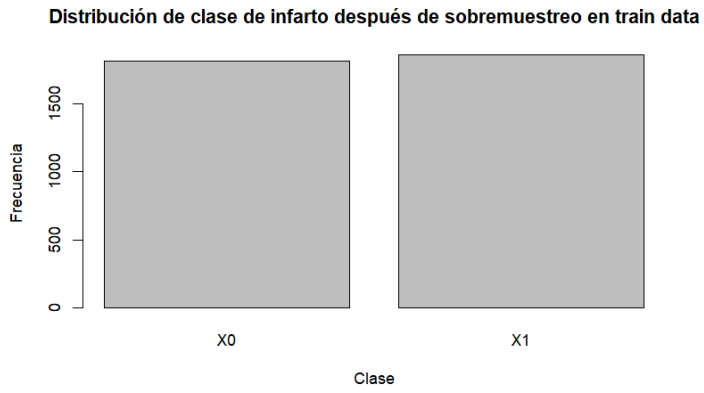


Figura 32. Gráfico de barras frecuencia de la variable infarto de *train data* balanceado.

Lo más destacable de los datos balanceados obtenidos con el método ROSE es que como podemos ver en el gráfico previo ahora hay un equilibrio más parejo entre las 2 clases de infarto y no infarto. Antes del balanceo, la mayoría de las observaciones en el conjunto de datos de entrenamiento estaban en la categoría de no infarto (92.44%), mientras que después del balanceo, la proporción de infartos y no infartos es más equilibrada, con un 50.62% de pacientes que han sufrido un infarto y un 49.37% que no lo han sufrido. Esto indica que el proceso de balanceo ha sido efectivo para corregir el desbalance inicial en los datos y ahora son más adecuados para construir modelos predictivos precisos. Además, podemos observar que las estadísticas descriptivas de las variables, como edad, hipertensión, estado civil, nivel de glucosa en sangre y IMC, han cambiado ligeramente después del balanceo, pero siguen manteniendo rangos y valores similares a los datos originales.

5.3. RESULTADOS APLICACIÓN, OPTIMIZACIÓN Y COMPARACIÓN DE ALGORTIMOS MACHINE LEARNING.

En el contexto de mi análisis de la aplicación y comparación de algoritmos de machine learning, no quiero olvidar el considerar como afecta el infarto a los menores de 35 años antes de centrarme en los pacientes mayores a esta edad que son el grupo en que me enfocaré. Destaquemos algunos valores extraídos de nuestra base de datos, entre los menores de 18 años, que representan 856 individuos, solo se registraron 2 casos de infarto, lo que equivale a un porcentaje muy bajo del 0,23%. Si extendemos este análisis hasta los 32 años, la probabilidad de un infarto sigue siendo excepcionalmente baja, con un mínimo del 0,12%. Esto sugiere que, para los menores de 35 años, la probabilidad de sufrir un infarto en los próximos 3 años es extremadamente baja, con un alto grado de certeza de aproximadamente el 99,85%. Sin embargo, a medida que la edad aumenta, la incidencia de infartos se vuelve más significativa. Por ejemplo, entre los mayores de 38 años, la probabilidad de un infarto es del 7,98%, lo que indica la importancia de considerar la edad como un factor de riesgo significativo.

5.3.1. Resultados de la aplicación del algoritmo *Decision Tree*.

cp	Precision	Recall	F3
0.01	0.6924564	0.7674565	0.7581053
F3 se utilizó para seleccionar el modelo óptimo utilizando el valor más alto. El valor final utilizado para el modelo fue cp = 0,01.			

Tabla 6. Resultados del modelo óptimo de *decision tree*.

La muestra se dividió en cinco partes iguales, y el modelo se entrenó y evaluó cinco veces utilizando una parte diferente como conjunto de prueba en cada iteración (validación cruzada). Se realizaron pruebas de varios valores del parámetro de complejidad (cp), así como evaluaciones del rendimiento del modelo en términos del valor de la medida F3, precision y recall. El valor de F3 más alto es del 75.81%, lo que indica que el modelo tiene un buen equilibrio entre precisión y recall. Entre todas las configuraciones de parámetros probadas, el modelo con el valor de cp = 0,01 fue el seleccionado.

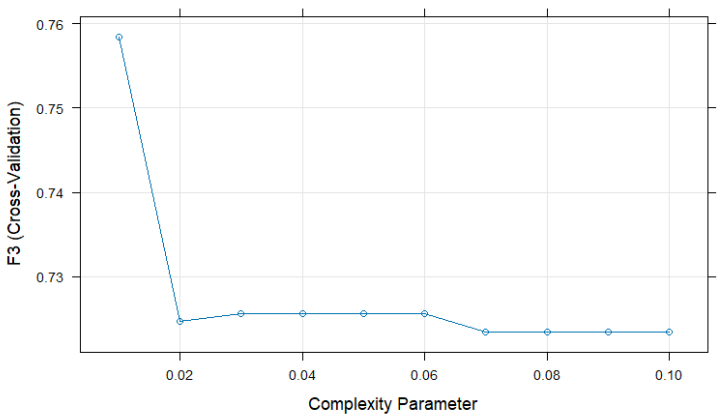


Figura 33. Gráfico de la relación entre el parámetro de complejidad (complexity parameter) y la F3 del modelo *decisión tree*.

Este gráfico muestra la relación entre el parámetro de complejidad y el valor que va tomando la medida F3 en la validación cruzada del modelo. La línea que se extiende a lo largo de todo el eje X, desde valores altos de F3 (casi 0.76), disminuye drásticamente hasta valores más bajos (menores de 0.730), donde parece estabilizarse incluso disminuir levemente, muestra cómo a medida que el parámetro de complejidad aumenta la medida F3 puede sufrir grandes cambios prejudiciales en este caso.

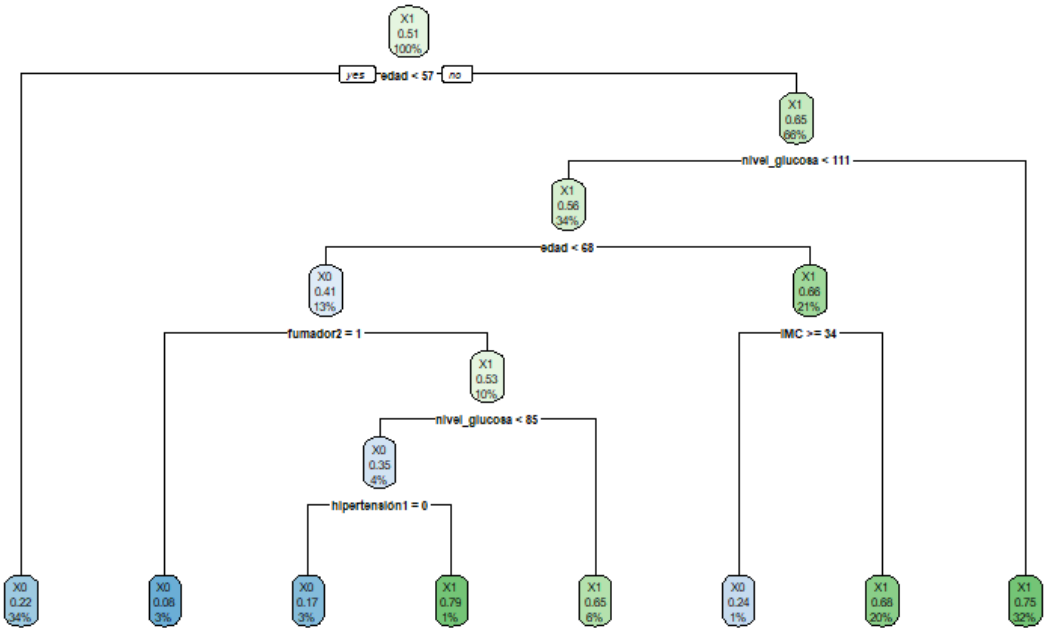


Figura 34. Gráfico de la representación del árbol de decisión.

La división inicial se basa en la edad de los pacientes, un grupo incluye personas menores de 57 años (probabilidad de 0,22 de no sufrir un infarto, 34% de la muestra) , mientras que las personas mayores de 57 años se dividen en dos grupos adicionales: personas con un nivel de glucosa superior a 111 (probabilidad de 0,75 de sufrir un infarto, 32% de la muestra) y personas con nivel de glucosa menor. Mientras tanto, los pacientes con nivel de glucosa menor a 111 se clasifican aún más de nuevo según su edad en menores y mayores de 68 años, clasifica a aquellos pacientes mayores de 68 años según su IMC superior o igual a 34 (el modelo les asigna un 0.24% de probabilidad de no sufrir un infarto, siendo el 1% de la muestra), o con un IMC menor a 34 (0.68% de probabilidad de sufrir un infarto, 20% de la muestra). En cuanto a aquellos pacientes con una edad menor a los 68 años el modelo asigna una probabilidad de 0.06 de no sufrir un infarto a los que nunca han fumado sin embargo para aquellos que si han fumado encontramos otra división, tener un nivel de glucosa superior 85 (0.65% de probabilidad de sufrir un infarto, 6% de la muestra) o tener un nivel menor que este y no sufrir hipertensión (0.17% de probabilidad de no sufrir un infarto, 3% de la muestra) o sufrir de hipertensión (0.79% de probabilidad de sufrir un infarto, 1% de la muestra)

5.3.1.1. Resultados de la optimización de los hiperparámetros del modelo *decision tree*.

cp	Precision	Recall	F3
0.0018	0.7190918	0.7822624	0.7747860
F3 se utilizó para seleccionar el modelo óptimo utilizando el valor más alto. El valor final utilizado para el modelo fue cp = 0,0018.			

Tabla 7. Resultados del modelo óptimo al ajustar los hiperparámetros de *decision tree*.

Tras ajustar lo parámetros a la vista de los resultados previos y entrenar el modelo nuevamente, el mejor valor de F3 (0.7748) fue alcanzado usando un cp de 0.0018, lo que resulta en una *precision* de 0.7191 y un *recall* de 0.7823.

A la vista de los resultados obtenidos queda en evidencia que al ajustar los valores del parámetro de complejidad he conseguido mejorar levemente el valor obtenido de F3 al entrenar el modelo por lo que será este segundo modelo ajustado el que usaré en la comparación con el resto de los modelos.

5.3.2. Resultados de la aplicación del algoritmo *Random Forest*.

mtry	splitrule	min.node.size	Precision	Recall	F3
6	extratrees	2	0.8198145	0.9502538	0.9353391
F3 was used to select the optimal model using the largest value. The final values used for the model were mtry = 6, splitrule = extratrees and min.node.size = 2.					

Tabla 8. Resultados del modelo óptimo de *random forest*.

Estos resultados muestran la evaluación del algoritmo Random Forest mediante validación cruzada con 5 pliegues. Se probaron diferentes configuraciones de hiperparámetros, como el número de variables predictoras a considerar en cada división del árbol (mtry), el criterio de división (splitrule) y el (min.node.size), un parámetro que controla el número mínimo de observaciones que deben estar en un nodo terminal (hoja) del árbol. La métrica principal utilizada para evaluar los modelos fue la F3. A través de este proceso, se generan métricas clave como precisión y recall para cada configuración.

Se observa que el modelo seleccionado, con mtry = 6 y splitrule = extratrees y min.node.sizes=2 obtuvo el mayor valor para F3 (0.9353391).

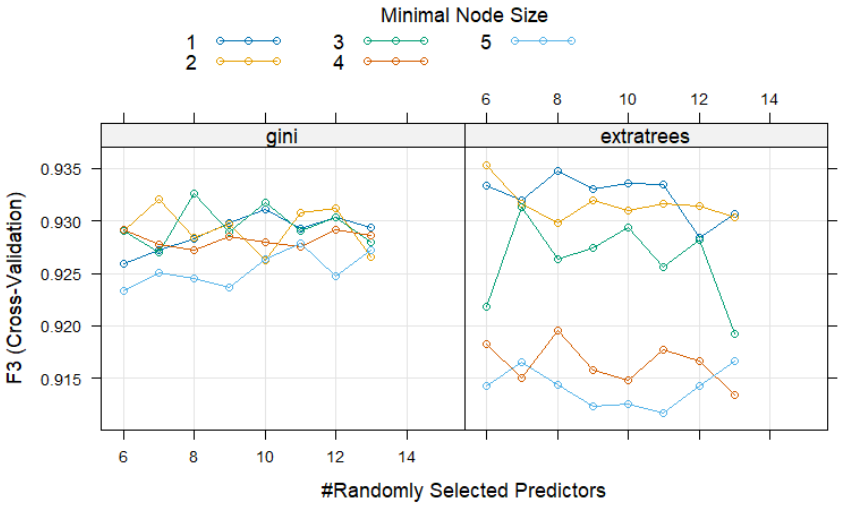


Figura 35. Gráfico splitting rule.

La precisión del modelo utilizando la validación cruzada en función del número de predictores seleccionados al azar para construir los árboles de decisión se muestra en el gráfico "splitting rule". Recordemos que "gini" y "extratrees" son dos reglas de división diferentes, las líneas representadas con distintos colores identifican el tamaño mínimo del nodo, el eje X hace referencia a los predictores elegidos de forma aleatoria y el eje Y a los valores que va tomando la medida F3 a lo largo de la cross validation.

Las líneas correspondientes a la regla de división "gini" parecen mantenerse en valores altos de F3 entre 0.92 y 0.935 independientemente de los predictores seleccionados. Sin embargo, las líneas representadas en la regla de división extratrees si que parecen sufrir numerosos cambios en los valores que toma F3, siendo más bajos para los valores más altos (4 y 5) del parámetro min.node.size.

5.3.2.1. Resultados de la optimización de los hiperparámetros del modelo *random forest*.

mtry	splitrule	min.node.size	Precision	Recall	F3
8	extratrees	1	0.8180225	0.9524494	0.9369993
F3 se utilizó para seleccionar el modelo óptimo utilizando el valor más alto. Los valores finales utilizados para el modelo fueron mtry = 8, splitrule = extratrees y min.node.size = 1.					

Tabla 9. Resultados del modelo óptimo al ajustar los hiperparámetros de *random forest*.

Tras ajustar los parámetros a la vista de los resultados previos y entrenar el modelo nuevamente, el mejor valor de F3 (0.9369993), superior al previo, es alcanzado usando los valores mtry = 8 y splitrule = extratrees y min.node.sizes=1, lo que resulta en una *precision* de 0.8180225 y un *recall* de 0.9524494, será este segundo modelo ajustado el que usaré en la comparación con el resto de los modelos.

5.3.3. Resultados Regresión Logística Regularizada.

alpha	lambda	Precision	Recall	F3
0.5	0.0556	0.6847433	0.7172785	0.7137624
F3 se utilizó para seleccionar el modelo óptimo utilizando el valor más alto. Los valores finales utilizados para el modelo fueron alfa = 0,5 y lambda = 0,0556.				

Tabla 10. Resultados del modelo óptimo de Regresión Logística Regularizada.

Vemos cómo se presentan los resultados de rendimiento del modelo para diferentes combinaciones de dos parámetros de ajuste: "*alpha*" y "*lambda*". Estos parámetros controlan el tipo y la intensidad de la regularización aplicada al modelo. Se registran métricas como *precision*, *recall* y F3-*score* para cada combinación de parámetros de ajuste y como en los casos anteriores es el F3-*score* el usado para seleccionar el modelo óptimo. Los valores finales de los parámetros de ajuste seleccionados fueron "*alpha*" = 0.5 y "*lambda*" = 0.0556, los cuales resultan en un F3-*score* de 0.7137624.

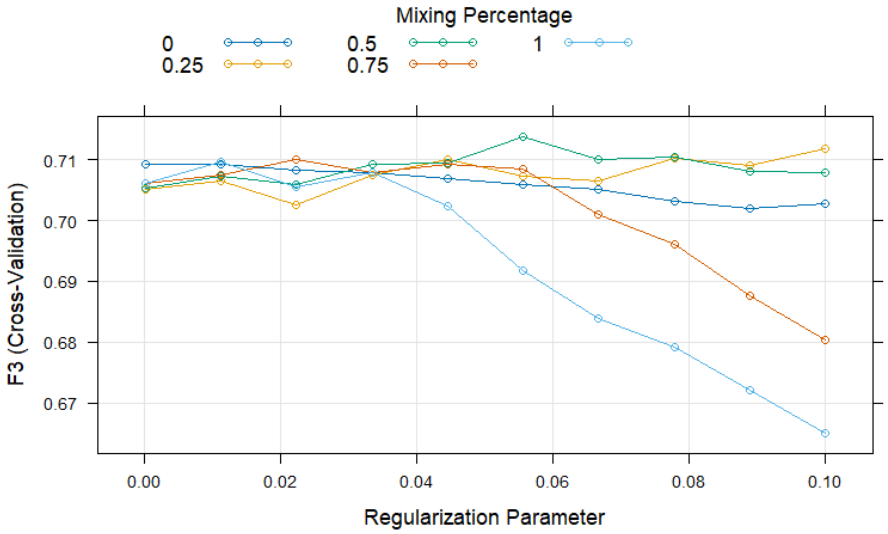


Figura 36. Gráfico F3, Alpha y lambda para el modelo de regresión.

En el gráfico de regresión logística regularizada, tres de las cinco líneas, correspondientes a distintos valores de un parámetro llamado *mixing percentage* (*Alpha*), mantienen un alto y constante valor de F3-score alrededor de 0.71 a lo largo del eje del "*regularization parameter*" (*lambda*).

Sin embargo, las líneas correspondientes a los valores más alto del parámetro Alpha muestran una caída drástica en el F3-score a medida que avanza a lo largo del eje del parámetro lambda, alcanzando valores realmente bajos (menores a 0.67). Tal y como veíamos en los resultados el valor más alto de F3 es alcanzado por la línea correspondiente a "*alpha*" = 0.5 y "*lambda*" = 0.0556.

5.3.3.1. Resultados de la optimización de los hiperparámetros del modelo Regresión Logística Regularizada.

alpha	lambda	Precision	Recall	F3
0.4444444	0.059636364	0.6841918	0.7183468	0.7146649
F3 se utilizó para seleccionar el modelo óptimo utilizando el valor más alto. Los valores finales utilizados para el modelo fueron alfa = 0,4444444 y lambda = 0,05963636.				

Tabla 11. Resultados del modelo óptimo al ajustar los hiperparámetros de Regresión Logística Regularizada.

Tras ajustar los parámetros a la vista de los resultados previos y entrenar el modelo nuevamente, el mejor valor de F3 (0.7146649) fue alcanzado usando un valor para alpha de 0.4444444 y un valor para lambda de 0.059636364, lo que resulta en una *precision* de 0.6841918 y un *recall* de 0.7183468.

A la vista de los resultados obtenidos queda en evidencia que al ajustar los valores de los parámetros Alpha y lambda es posible mejorar levemente el valor obtenido de F3 al entrenar el modelo por lo que será este segundo modelo ajustado el que usaré en la comparación con el resto de los modelos.

5.3.4. Resultados *Support Vector Machine*.

sigma	C	Precision	Recall	F3
0.1	1	0.7466906	0.7704756	0.7678890
F3 se utilizó para seleccionar el modelo óptimo utilizando el valor más alto. Los valores finales utilizados para el modelo fueron sigma = 0,1 y C = 1.				

Tabla 12. Resultados del modelo óptimo de *Support Vector Machine*.

Vemos como se probó una gama de valores para los parámetros de regularización C y sigma (que controla la flexibilidad del modelo), y se encontró que el valor óptimo fue de sigma=0.1 y c=1, con un valor para F3 máximo de aproximadamente 76.78% y el valor de *recall* más alto de 0.77.

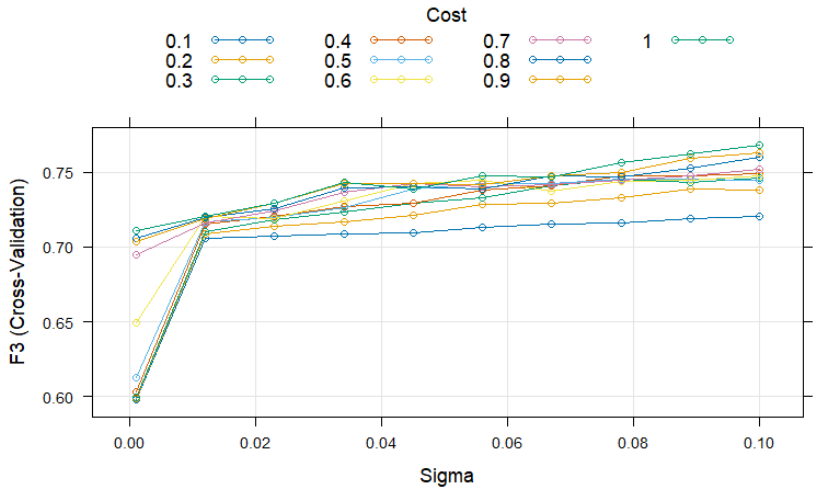


Figura 37. Gráfico F3, sigma y coste para el modelo SVM.

En el gráfico mostrado del modelo definido de SVM todas las líneas, correspondientes a distintos valores de un parámetro llamado coste (*cost*), siguen la misma tendencia en cuanto a los valores que toma la medida F3 según va aumentando un parámetro llamado sigma. Todas ellas parten de valores bajos para F3 (entre 0.60 y 0.71 aproximadamente) cuando sigma toma el valor 0 pero en cuanto este aumenta ligeramente (toma el valor 0.1) todas las lineas independientemente del valor de coste se aúnan en torno a el valor 0.71 para la medida F3, a partir de esa situación a medida que el valor de sigma aumenta el valor de F3 también aumenta más ligeramente para todos los costes destacando que el coste que menor valor para f3 consigue alcanzar es el más bajo 0.1 y los valores de coste más altos 0.9 y 1 son los que obtienen los valores más altos para F3.

5.3.4.1. Resultados de la optimización de los hiperparámetros del modelo *Support Vector Machine*.

sigma	C	Precision	Recall	F3
1	7.487	0.911839	0.9242092	0.9228907
F3 se utilizó para seleccionar el modelo óptimo utilizando el valor más alto. Los valores finales utilizados para el modelo fueron sigma = 1 y C = 7.487				

Tabla 13. Resultados del modelo óptimo de *Support Vector Machine*.

Tras reajustar los valores que toman los parámetros sigma y C a la vista de los resultados anteriores al entrenar nuevamente el modelo SVM presenta un alto rendimiento en la identificación de casos de infarto, con una precisión del 91.18% y una recall del 92.42% al usar los hiperparámetros sigma = 1 y C = 7.487 con todo esto el valor más alto del F3 Score (0.9228907) es alcanzado.

Con todo esto concluyo que en este caso se hace evidencia de la importancia de ajustar los parámetros a la hora de entrenar un modelo puesto que su rendimiento incrementará notablemente (en este caso hemos pasado de un valor de F3 de 0.7678890 a 0.9228907), por esto usaremos el modelo ajustado (resultados tabla 13) en la comparación con el resto de modelos ya optimizados también.

5.3.5. Resultados de la comparación de modelos y elección del modelo final.

```
$F3
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
RF_OPTIMIZADO  0.9287950 0.9333028 0.9342866 0.9369993 0.9412609 0.9473514    0
SVM_OPTIMIZADO 0.9077839 0.9119000 0.9184449 0.9228907 0.9220254 0.9542991    0
DT_OPTIMIZADO  0.7402283 0.7656676 0.7736679 0.7747860 0.7767453 0.8176207    0
RLR_OPTIMIZADO 0.6948744 0.7084850 0.7181136 0.7146649 0.7238067 0.7280449    0

$Precision
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
RF_OPTIMIZADO  0.8020690 0.8103448 0.8144828 0.8180225 0.8191856 0.8440304    0
SVM_OPTIMIZADO 0.8986207 0.9117241 0.9151139 0.9118390 0.9158621 0.9178744    0
DT_OPTIMIZADO  0.6668966 0.6749482 0.7379310 0.7190918 0.7405107 0.7751724    0
RLR_OPTIMIZADO 0.6620690 0.6703448 0.6935818 0.6841918 0.6970324 0.6979310    0

$Recall
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
RF_OPTIMIZADO  0.9428118 0.9444015 0.9517185 0.9524494 0.9578264 0.9654889    0
SVM_OPTIMIZADO 0.9069767 0.9114619 0.9185083 0.9242092 0.9231844 0.9609145    0
DT_OPTIMIZADO  0.7404844 0.7646259 0.7775362 0.7822624 0.7899838 0.8386817    0
RLR_OPTIMIZADO 0.6946355 0.7096774 0.7209469 0.7183468 0.7302780 0.7361963    0
```

Figura 38. Comparación de todos los modelos optimizados según las medidas F3*, Precision* y *Recall*.

Model <chr>	Metric <chr>	Mean <dbl>
RF_OPTIMIZADO	F3	0.9369993
SVM_OPTIMIZADO	F3	0.9228907
DT_OPTIMIZADO	F3	0.7747860
RLR_OPTIMIZADO	F3	0.7146649
RF_OPTIMIZADO	Precision	0.8180225
SVM_OPTIMIZADO	Precision	0.9118390
DT_OPTIMIZADO	Precision	0.7190918
RLR_OPTIMIZADO	Precision	0.6841918
RF_OPTIMIZADO	Recall	0.9524494
SVM_OPTIMIZADO	Recall	0.9242092
DT_OPTIMIZADO	Recall	0.7822624
RLR_OPTIMIZADO	Recall	0.7183468

Figura 39. Comparación de todos los modelos optimizados usando las medias de las medidas F3*, Precision* y *Recall*.

Tanto en la figura 38 como en la figura 39 podemos ver cómo en la métrica de F3, RF y SVM muestran un rango más alto (mínimo de 0.9287950 y 0.9077839 y máximo de 0.9473514 y 0.9542991, respectivamente) en comparación con, RLR (0.6948744 a 0.7280449) y DT (0.7402283 a 0.8176207). Además, RF presenta una mediana y una media ligeramente más altas (0.9342866 y 0.9369993, respectivamente) que el modelo SVM. En cuanto a la medida *precision* vemos como SVM es el que consigue mejores resultados (vemos una media de 0.9118390 y un máximo de 0.9178744. Por otro lado, los valores de *recall* de RF, alcanzan una media superior de 0.9524494 y un máximo de 0.9654889. Todo esto demuestra su capacidad para identificar correctamente las predicciones

positivas, minimizando los falsos positivos (FP) frente al resto de los modelos, lo cuál es nuestro objetivo.

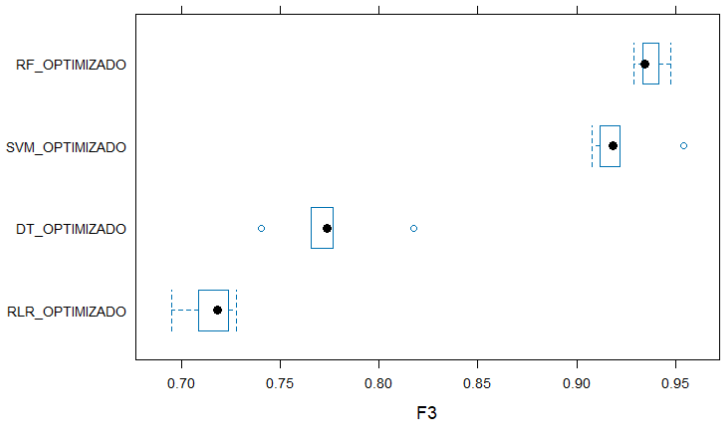


Figura 40. *Boxplot* comparativo de los valores de F3 para todos los modelos optimizados.

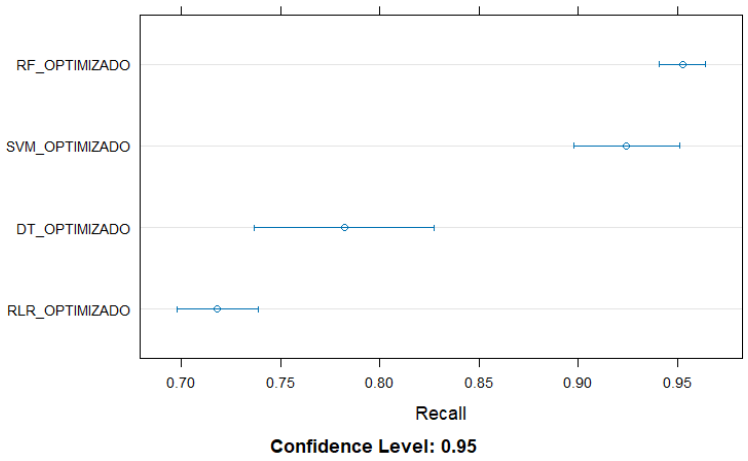


Figura 41. Gráfico comparativo de los intervalos de confianza de los resultados obtenidos para cada modelo de los valores de Recall.

En las figuras 40 y 41 he querido visualizar lo que mencionaba antes para no dejar lugar a dudas, en estos dos gráficos podemos ver cómo tanto en la medida *recall* (figura 41) como en la medida F3 (figura 40) cómo los valores alcanzados por los modelos RF y SVM se encuentran destacablemente desplazados hacia valores muy altos de estas medidas a diferencia del resto de modelos que se agrupan en valores más bajos. Sin embargo, puesto que he entrenado todos los modelos con vistas a aumentar el valor para la medida F3 dando más peso a la medida *recall* que a cualquier otra tomaré como modelo final aquel con la media más alta para F3, en mi caso RF_OPTIMIZADO será el seleccionado como modelo final.

5.4. RESULTADOS DE LA MEDICIÓN DEL RENDIMIENTO DEL MODELO FINAL MEDIANTE UNA MATRIZ DE CONFUSIÓN SOBRE LOS DATOS TEST Y SOBRE LOS DATOS ORIGINALES COMPLETOS SIN BALANCEAR MEDIANTE *CROSS VALIDATION*.

5.4.1. Resultados de la matriz de confusión y los estadísticos del rendimiento del modelo final sobre los datos test.

```
Confusion Matrix and Statistics

          NO_INFARTO INFARTO
NO_INFARTO       1182       1
INFARTO            25      97

               Accuracy : 0.9801
                 95% CI : (0.9709, 0.9869)
    No Information Rate : 0.9249
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8711

 Mcnemar's Test P-Value : 6.462e-06

            Sensitivity : 0.98980
            Specificity : 0.97929
         'Positive' Class : INFARTO


 F3-Score:  0.9661355
```

Figura 42. Matriz de confusión y los estadísticos del rendimiento del modelo final sobre los datos test.

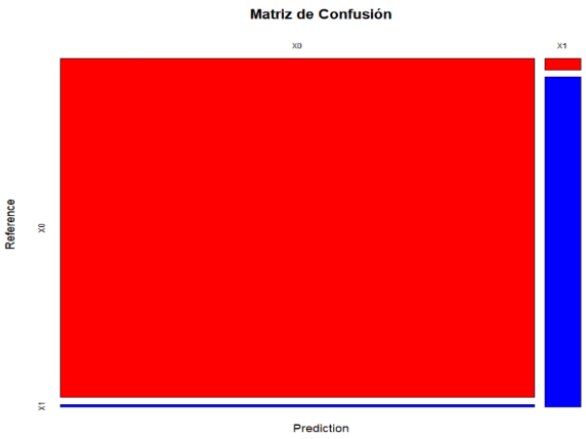


Figura 43. Matriz de confusión del modelo final sobre los datos test.

La evaluación del modelo final (*random forest*) en el conjunto de datos de prueba (test_data, sin balancear) muestra una alta accuracy (0.9801) y un buen equilibrio entre sensibilidad (0.9898) y especificidad (0.9792). La sensibilidad alta indica que el modelo identifica correctamente la mayoría de los casos positivos (infartos), mientras que la especificidad alta sugiere que rara vez clasifica erróneamente los casos negativos. La puntuación F3 en la que nos hemos centrado a lo largo de todo el proyecto es de 0.9661355 la cual indica que el modelo es altamente efectivo en identificar casos de infarto y en minimizar la cantidad de falsos negativos, lo que es esencial en nuestro contexto donde cada caso de infarto no detectado podría tener consecuencias críticas para la salud del paciente.

5.4.2. Resultados de la medición del rendimiento del modelo final con una *cross validation* sobre los datos originales completos sin balancear.

	mtry <int>	splitrule <fctr>	min.node.size <int>	Precision <dbl>	Recall <dbl>	F3 <dbl>	PrecisionSD <dbl>	RecallSD <dbl>	F3SD <dbl>
1	8	extratrees	1	0.9900607	0.9253854	0.9314672	0.007127521	0.001305584	0.001524684

Figura 44. Métricas del rendimiento del modelo final con una cross validation sobre los datos originales completos sin balancear.

En cuanto a la *precision* del modelo medido en todos los datos sin balancear es muy alta (0.9900607), al igual que el valor obtenido en *recall* (0.9253854) y en F3(0.9314672) todo esto indica

como en el caso anterior que el modelo es excelente para identificar correctamente las instancias positivas (verdaderos positivos), con una baja tasa de falsos positivos. El valor alto de F3 indica que el modelo mantiene un buen balance con una ligera preferencia por el *recall*, lo cual es adecuado en nestro contexto donde es más crítico minimizar los falsos negativos. Referente a las desviaciones estándar para *precision* (0.007127521), *recall* (0.001305584) y F3 (0.001524684) son bajas, lo que muestra que el modelo tiene un rendimiento consistente y estable a través de las diferentes particiones de la validación cruzada.

En resumen, el modelo *Random Forest* con los parámetros especificados (mtry=8, *splitrule=extratrees y min.node.size*=1) demuestra un rendimiento excelente y consistente en términos de *precision*, *recall* y F3 sobre los datos test y los datos originales sin balancear, lo que justifica aún más su selección como el modelo final para el análisis de la predicción de infartos.

5.5. APLICACIÓN DE PREDICCIÓN DE INFARTOS BASADA EN EL MODELO FINAL, *RANDOM FOREST*.

Al abrir la aplicación nos encontramos con desplegables para las variables categóricas y celdas rellenables para las numéricas.

Veamos dos ejemplos con los datos de dos pacientes distintos:

Figura 45. Aplicación predictora de infartos, caso de bajo riesgo de infarto.

Tal y como podemos apreciar en la captura según los datos proporcionados por el paciente nuestro algoritmo no detecta riesgo de infarto y da la información mostrada.

Figura 46. Aplicación predictora de infartos, caso de alto riesgo de infarto.

A diferencia del anterior ejemplo los datos de este paciente sí que llevan al algoritmo a detectar un alto riesgo de infarto y es por esto que la aplicación proporciona esa información.

6. DISCUSIÓN Y CONCLUSIONES.

La comparación de los resultados de mi estudio con los hallazgos de investigaciones previas nos proporcionará una visión sobre la utilidad de los algoritmos aplicados en el diagnóstico de infartos. Opté por explorar los modelos DT y RF debido a su importancia en los estudios analizados y ser los únicos que aplicaron y tienen en común los 5 estudios que conforman mi estado del arte, también me pareció interesante aplicar el modelo SVM puesto que se mencionaba en el estudio realizado por University Bhubaneswar, Odisha pero no parecía obtener resultados sobresalientes en este, además quise probar un modelo que no se hubiera aplicado en uno de los estudios previos tenidos en cuenta para apreciar si podía aportar algo novedoso, seleccioné de modelo RLR.

Debo destacar que debido al contexto de mi estudio he preferido centrarme en minimizar los falsos negativos (notificar al paciente que no sufrirá ningún infarto cuando sí que lo sufrirá) más que en optimizar el *accuracy* del modelo a diferencia de los estudios que conforman el estado del arte. Es por esto que también veremos algunas diferencias a sus resultados y aportaciones.

Similarmente al estudio de la Carnegie Mellon University, donde se concluyó que el árbol de decisiones era el modelo más adecuado en la predicción, mis resultados también respaldan la utilidad de este algoritmo en el contexto del diagnóstico de infartos, aunque no concluyo que sea el más adecuado habiendo otros más idóneos.

Lo más destacable y en lo que más diferenciamos mis resultados de los estudios que forman el apartado titulado estado del arte, es el comportamiento y los resultados que nos proporciona el modelo SVM, resulta sorprendente ver como inicialmente resultaba en valores de F3 relativamente bajos, pero al ajustar los parámetros estos se disparan hasta posicionarse en el segundo puesto como mejor predictor, precedido por muy poca diferencia por el que ha sido el modelo final RF.

En particular, el estudio realizado por el Departamento de Ingeniería Eléctrica e Informática de la Universidad North South en Bangladesh y el Departamento de Tecnología de la Información de la Universidad Taif en Arabia Saudita, destacó el desempeño superior del algoritmo *Random Forest*, al igual que hemos visto en el presente trabajo donde hemos obtenido un valor de F3 del 0.9661 midiendo sobre los datos test sin balancear, un valor realmente difícil de superar.

6.1. CONCLUSIONES.

1. El Análisis Exploratorio de Datos (EDA) ha sido una fase fundamental en mi estudio. A través de técnicas de visualización y análisis estadístico, he obtenido una comprensión profunda de la estructura y relación de los datos y variables.

2. Tratar la base de datos previamente a implementar los algoritmos ha sido un reto significativo sobre todo a la hora de imputar los datos faltantes en la variable IMC utilizando el método k-nearest neighbors (KNN).

3. La partición de los datos en entrenamiento (incluyendo validación) y test se realizó en función de la variable respuesta "infarto", asegurando una distribución equilibrada entre el conjunto de entrenamiento y el de prueba. Esto permite una evaluación más precisa y sin sesgos del modelo.

4. El balanceo de los datos de entrenamiento mediante *Random Over Sampling* (ROSE) fue esencial para corregir el desbalance inicial de clases tan dispares de presencia y ausencia de infartos, mejorando así la capacidad predictiva del modelo.

5. Se puede apreciar la influencia significativa de variables como la hipertensión, las enfermedades cardiovasculares y el estado marital en el riesgo de sufrir un infarto. Además, se observa una asociación entre el tipo de trabajo y la incidencia de infartos, con interferencias relacionadas con la edad.

6. La exclusión de pacientes menores de 35 años permite focalizar el análisis en la población con mayor riesgo de infarto, eliminando casos anómalos y garantizando resultados más precisos.

7. Tras la aplicación, optimización de parámetros, validación (mediante *cross-validation*) y comparación de los modelos DT, RF, RLR y SVM (llevados a cabo tal y como se prometía en los objetivos secundarios), selecciono *Random Forest* como el modelo final a usar en mi caso debido a su valor en la medida F3 superior, su alta sensibilidad, y *precision*.

8. La creación de una aplicación web interactiva mediante R y Shiny permite una evaluación del riesgo de infarto basada en factores de salud específicos de manera accesible y visualmente atractiva.

9. Los resultados de mi estudio sugieren al igual que en el caso del análisis de la Carnegie Mellon University, que el modelo de árbol de decisión muestra una capacidad razonable para predecir la ocurrencia de infartos, aunque otros algoritmos pueden ofrecer mejores resultados.

10. Los hallazgos de mi estudio respaldan la eficacia del algoritmo Random Forest, como lo destacó también el trabajo realizado por el Departamento de Ingeniería Eléctrica e Informática de la Universidad North South en Bangladesh y el Departamento de Tecnología de la Información de la Universidad Taif en Arabia Saudita.

11. El objetivo principal de mi trabajo fue la creación y evaluación de modelos de aprendizaje automático capaces de predecir con precisión la ocurrencia de infartos utilizando la base de datos "Stroke Prediction Dataset" de Kaggle. Este objetivo se ha alcanzado satisfactoriamente, desarrollando un modelo que logra valores para las medidas accuracy y sensitivity aproximadamente del 98%, lo que supera ampliamente las expectativas iniciales.

7. BIBLIOGRAFÍA

[1] «Cardiovascular diseases (CVDs)». Accedido: 31 de enero de 2024. [En línea]. Disponible en: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] R. B. Hayes *et al.*, «PM2.5 air pollution and cause-specific cardiovascular disease mortality», *Int. J. Epidemiol.*, vol. 49, n.º 1, pp. 25-35, feb. 2020, doi: 10.1093/ije/dyz114.

[3] «Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants», *Lancet Lond. Engl.*, vol. 398, n.º 10304, pp. 957-980, sep. 2021, doi: 10.1016/S0140-6736(21)01330-1.

[4] J. Yang *et al.*, «Risk Factors and Outcomes of Very Young Adults Who Experience Myocardial Infarction: The Partners YOUNG-MI Registry», *Am. J. Med.*, vol. 133, n.º 5, pp. 605-612.e1, may 2020, doi: 10.1016/j.amjmed.2019.10.020.

[5] I. Alonso, «El 35% de los menores españoles tiene dos o más factores de riesgo cardiovascular», Fundación Española del Corazón. Accedido: 16 de enero de 2024. [En línea]. Disponible en: https://fundaciondelcorazon.com/prensa/notas-de-prensa/3808-el-35-de-los-menores-espanoles-tiene-dos-o-mas-factores-de-riesgo-cardiovascular.html

[6] M. D. Huffman *et al.*, «A Cross-Sectional Study of the Microeconomic Impact of Cardiovascular Disease Hospitalization in Four Low- and Middle-Income Countries», *PLoS ONE*, vol. 6, n.º 6, p. e20821, jun. 2011, doi: 10.1371/journal.pone.0020821.

[7] «Calidad de vida y apoyo social en pacientes con infarto agudo de miocardio no complicado». Accedido: 16 de enero de 2024. [En línea]. Disponible en: https://www.revespcardiol.org/es-pdf-X0300893299001282

[8] «Impacto del infarto de miocardio en la situación laboral de los pacientes». Accedido: 16 de enero de 2024. [En línea]. Disponible en: https://www.revespcardiol.org/es-pdf-X0300893299001489

[9] P. I. Dorado-Díaz, J. Sampedro-Gómez, V. Vicente-Palacios, y P. L. Sánchez, «Aplicaciones de la inteligencia artificial en cardiología: el futuro ya está aquí», *Rev. Esp. Cardiol.*, vol. 72, n.º 12, pp. 1065-1075, dic. 2019, doi: 10.1016/j.recesp.2019.05.016.

[10] V. Arias, J. Salazar, y C. Garicano, «Una introducción a las aplicaciones de la inteligencia artificial en Medicina: Aspectos históricos», *Rev. Latinoam. Hipertens.*, vol. 14, 2019.

[11] M. Á. Vega, L. M. Q. Mora, y M. V. C. Badilla, «Inteligencia artificial y aprendizaje automático en medicina», *Rev. Medica Sinerg.*, vol. 5, n.º 8, Art. n.º 8, ago. 2020, doi: 10.31434/rms.v5i8.557.

[12] D. P. Wall, J. Kosmicki, T. F. DeLuca, E. Harstad, y V. A. Fusaro, «Use of machine learning to shorten observation-based screening and diagnosis of autism», *Transl. Psychiatry*, vol. 2, n.º 4, Art. n.º 4, abr. 2012, doi: 10.1038/tp.2012.10.

[13] A. Esteva *et al.*, «Dermatologist-level classification of skin cancer with deep neural networks», *Nature*, vol. 542, n.º 7639, pp. 115-118, feb. 2017, doi: 10.1038/nature21056.

[14] G. D. Magoulas y A. Prentza, «Machine Learning in Medical Applications», en *Machine Learning and Its Applications: Advanced Lectures*, G. Paliouras, V. Karkaletsis, y C. D. Spyropoulos, Eds., en Lecture Notes in Computer Science. , Berlin, Heidelberg: Springer, 2001, pp. 300-307. doi: 10.1007/3-540-44673-7_19.

[15] E. Long *et al.*, «An artificial intelligence platform for the multihospital collaborative management of congenital cataracts», *Nat. Biomed. Eng.*, vol. 1, n.º 2, Art. n.º 2, ene. 2017, doi: 10.1038/s41551-016-0024.

[16] C. Gui y V. Chan, «Machine learning in medicine», *Univ. West. Ont. Med. J.*, vol. 86, n.º 2, Art. n.º 2, dic. 2017, doi: 10.5206/uwomj.v86i2.2060.

[17] «Stroke Prediction Dataset». Accedido: 18 de enero de 2024. [En línea]. Disponible en: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[18] E. Dritsas y M. Trigka, «Stroke Risk Prediction with Machine Learning Techniques», *Sensors*, vol. 22, n.º 13, Art. n.º 13, ene. 2022, doi: 10.3390/s22134670.

[19] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, y M. Monirujjaman Khan, «Stroke Disease Detection and Prediction Using Robust Learning Approaches», *J. Healthc. Eng.*, vol. 2021, p. e7633381, nov. 2021, doi: 10.1155/2021/7633381.
[20] «Visual Analysis and Prediction of Stroke». Accedido: 26 de febrero de 2024. [En línea]. Disponible en: https://www.stat.cmu.edu/capstoneresearch/spring2021/315files/team16.html
[21] «IEEE Xplore Full-Text PDF»: Accedido: 26 de febrero de 2024. [En línea]. Disponible en: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10088363&casa_token=7R37F4ZVCNsAAAAA:NZ6EV0oFlzC-2txulw0kgTpsNBCshXTe2A4hPRjNGrCquQPBi1K1zP20ex07uPST2M74kLfjTw&tag=1
[22] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, y D. John, «A predictive analytics approach for stroke prediction using machine learning and neural networks», *Healthc. Anal.*, vol. 2, p. 100032, nov. 2022, doi: 10.1016/j.health.2022.100032.
[23] M. Szumilas, «Explaining Odds Ratios», *J. Can. Acad. Child Adolesc. Psychiatry*, vol. 19, n.º 3, p. 227, ago. 2010.
[24] «fedesoriano | Grandmaster». Accedido: 6 de mayo de 2024. [En línea]. Disponible en: https://www.kaggle.com/fedesoriano/competitions
[25] «(1) Federico Soriano Palacios | LinkedIn». Accedido: 6 de mayo de 2024. [En línea]. Disponible en: https://www.linkedin.com/in/federico-soriano-palacios/?originalSubdomain=es
[26] O. Troyanskaya *et al.*, «Missing value estimation methods for DNA microarrays», *Bioinformatics*, vol. 17, n.º 6, pp. 520-525, jun. 2001, doi: 10.1093/bioinformatics/17.6.520.
[27] N. Lunardon, G. Menardi, N. Torelli, y N. Lunardon, «ROSE: Random Over-Sampling Examples».
[28] F. Fallucchi, M. Coladangelo, R. Giuliano, y E. William De Luca, «Predicting Employee Attrition Using Machine Learning Techniques», *Computers*, vol. 9, n.º 4, Art. n.º 4, dic. 2020, doi: 10.3390/computers9040086.
[29] C. Kingsford y S. L. Salzberg, «What are decision trees?», *Nat. Biotechnol.*, vol. 26, n.º 9, pp. 1011-1013, sep. 2008, doi: 10.1038/nbt0908-1011.
[30] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.
[31] «(5) Regresión logística regularizada en R | LinkedIn». Accedido: 14 de mayo de 2024. [En línea]. Disponible en: https://www.linkedin.com/pulse/regresi%C3%B3n-log%C3%ADstica-regularizada-en-r-martin-ehman/
[32] «Support Vector Machine (SVM) Algorithm», GeeksforGeeks. Accedido: 14 de mayo de 2024. [En línea]. Disponible en: https://www.geeksforgeeks.org/support-vector-machine-algorithm/