VNiVERSiDAD D SALAMANCA

GRADO EN ESTADÍSTICA

EL USO DE MÉTODOS ESTADÍSTICOS Y DE APRENDIZAJE AUTOMÁTICO EN FÚTBOL MODERNO

Trabajo de Fin de Grado

Autor: Iván Torre Crespo

Tutor: José Luis Vicente Villardón

Salamanca, 2024

VNiVERSiDAD D SALAMANCA

GRADO EN ESTADÍSTICA

EL USO DE MÉTODOS ESTADÍSTICOS Y DE APRENDIZAJE AUTOMÁTICO EN FÚTBOL MODERNO

Trabajo de Fin de Grado

Autor: Iván Torre Crespo

TORRE CRESPO IVAN - 72103371G

Firmado digitalmente por TORRE CRESPO IVAN - 72103371G
Fecha: 2024.07.02 12:55:25 +02'00'

Tutor: José Luis Vicente Villardón

Salamanca, 2024

INTRODUCTION:

Over the last few decades, thanks to the progress of globalization society has exponentially grown suffering substantial changes. One of the main factors that has promoted this is the introduction and daily use of the internet in most of the actual population.

As could not be otherwise it has also been reflected in football. But, in which way?

Thanks to technology progresses focused on football, it has been possible to optimized equipment such as better pitches or boots preventing injuries to players.

In terms of different kinds of football, tactically speaking, it has undergone a huge evolution, especially in the last decade. The key factor has been the gathering and use of data associated with players and the team in general.

Nowadays, when it comes to physique, players are no longer just footballers, but athletes. The data and its subsequent analysis have enabled the team´s medical staff to carry out individualized follow-ups, thus getting the most out of each player.

Apart from this, tactical information is collected for every single match allowing, for example, to know the percentage of possession of each team and in which zones. In addition, there is also the possibility of knowing the individual performance of one player thanks to GPS technology.

All this data collection and analysis is called descriptive statistics, whose main function is to summarize and describe the main characteristics of the reports. However, experts have developed techniques that allow a deeper analysis by obtaining predictions and conclusions, these techniques are inferential statistics.

Throughout this work, both descriptive and inferential analysis will be carried out in the RStudio programming language.

Daily use of technology, mainly in the young population, could lead one to believe that the implementation of statistics in the world of football would bring this sport even closer to this generation. However, significant voices in the football atmosphere such as Gerard Piqué, former FC Barcelona and Spanish national team player, Javier Tebas, president of La Liga, and Florentino Pérez, president of Real Madrid, have the opposite opinion, stating that football is no longer as attractive to the fans as it used to be. In defense of this claim, they cite, for example, the drop in viewership of La Liga. Nevertheless, this is due to a socio-economic consequence and the rise of piracy, rather than an alleged lack of interest. Each time football is broadcasted on free-to-air television the numbers reach peak viewing figures, being the most watched of the week.

In order to respond to all of the above, this work is going to be performed with a database extracted from a football simulator developed by EA Sports, called FIFA. This simulator is a video game which main users are teenagers, this paper will try to contribute so that young people continue to be fascinated by the magic of this sport, but from a statistical perspective.

The main objective of the project will be, using principal component analysis, to compare which are the ten best strikers in terms of: the attributes given by FIFA and the Overall attribute (rating from 1 to 100 given to each player).

Similarity measures will also be utilized to look at players with as close to similar characteristics as possible, and to test the myth of whether being right- or left-footed affects a possible correlation in terms of ball striking variables.

The work will consist of the following sections:

1. SOCCER
2. STATISTICAL THEORY
3. STATISTICAL PROGRAMME
4. DATABASE
5. RESULTS
6. CONCLUSIONS
7. BIBLIOGRAPHY

ÍNDICE

INTRODUCCIÓN

Desde hace unas décadas, con el avance de la globalización, la sociedad ha evolucionado exponencialmente sufriendo grandes cambios. Uno de los factores que lo ha potenciado es la introducción y el uso diario de internet en la mayor parte de las poblaciones actuales.

Como no podía ser de otra manera también ha afectado al fútbol. ¿De qué manera? Gracias a los avances tecnológicos enfocados a este campo, se ha podido desarrollar equipamiento optimizado como mejores campos y botas de futbol para prevenir lesiones en jugadores.

En cuanto al tipo de futbol, tácticamente hablando, ha sufrido una gran evolución, sobre todo en la última década. El factor clave ha sido la recogida y el uso de datos acerca de los jugadores y el equipo en general.

Físicamente, los jugadores además de ser futbolistas son atletas. Los datos y su posterior análisis han permitido al personal médico de los equipos llevar a cabo seguimientos individualizados sacando así el máximo provecho de los jugadores.

Además, también se recoge información táctica de los partidos permitiendo analizarlos permitiendo saber, por ejemplo, cuenta posesión de balón ha tenido cada equipo y en qué zonas o, gracias a GPS, saber el rendimiento individual de un jugador.

Toda esta recopilación de datos y su análisis se denominan estadísticos descriptivos, cuya principal función es resumir y describir las principales características de los datos. Sin embargo, los expertos han desarrollado técnicas que permiten un análisis más profundo obteniendo predicciones y conclusiones, estas técnicas son los estadísticos inferenciales.

En este trabajo tanto el análisis descriptivo como el inferencial se va a llevar a cabo en el lenguaje de programación RStudio.

El uso cotidiano de la tecnología, en los jóvenes en mayor medida, podría hacer pensar que la implementación de la estadística en el mundo del futbol acercaría aún más el deporte a esta generación. No obstante, grandes voces autorizadas en el mundo del futbol como lo son Gerard Piqué, exjugador del FC Barcelona y de la selección española, Javier Tebas, presidente de La Liga, y Florentino Pérez, presidente del Real Madrid, tienen la opinión contraria afirmando que el futbol, ya no es tan atractivo para el aficionado como lo era antes. Se amparan en la bajada de audiencia que está sufriendo La Liga, sin embargo, esto responde a una consecuencia socioeconómica y el auge de la piratería, más que de un supuesto poco interés en el fútbol. Cada vez que se emite futbol en la televisión abierta tiene cifras que arrasan, siendo lo más visto de la semana.

En respuesta a todo lo anterior, este proyecto se va a realizar con una base de datos extraída de un simulador de fútbol desarrollada por EA Sports, llamada FIFA. Este simulador no deja de ser un videojuego cuyos principales usuarios son adolescentes, este trabajo intentará aportar para que los jóvenes se sigan viendo atraídos por la magia de este deporte, pero desde una perspectiva estadística.

El principal objetivo del trabajo va a ser, usando un análisis de componentes principales, comparar cuales son los diez mejores delanteros en función de: los atributos dados por el FIFA y el atributo Overall (calificación del 1 al 100 que se le da a cada jugador).

También se usarán medidas de similaridad para ver jugadores con características lo más parecidas posibles y comprobar el mito de si ser diestro o zurdo afecta a una posible correlación en cuanto a variables de golpeo de balón.

El trabajo constará de los siguientes apartados:

1. EL FÚTBOL
2. TEORÍA ESTADÍSTICA
3. PROGRAMA ESTADÍSTICO
4. LA BASE DE DATOS
5. RESULTADOS
6. CONCLUSIONES
7. BIBLIOGRAFÍA

CAPÍTULO 1: EL FÚTBOL

El fútbol, conocido como el deporte rey en buena parte del mundo, tiene una historia rica y variada que se remonta siglos e incluso milenios antes del fútbol tal y como se le conoce hoy en día. Por ello, para poner en contexto se va a resumir su historia brevemente.

La fascinación por los juegos de pelota ha estado presente en multitud de civilizaciones a lo largo de la historia, desde civilizaciones precolombinas (mayas y aztecas), pasando por la dinastía Han en china y las culturas mediterráneas (egipcia, romana, griega, …), hasta llegar al siglo XIX a Inglaterra, donde empezó a ser el inicio de un deporte organizado y reglado.

En la época precolombina, las sociedades allí presentes practicaban el tlachtli, un juego por equipos en el cual los jugadores trataban de pasar una pelota echa de caucho a través de unos aros de piedra situados a bastante altura que se encontraban en los laterales del campo. Este juego era una actividad muy importante dentro de estas sociedades, tanto cultural, social y religiosamente hablando, estaban bastante asociados este juego con diversos rituales religiosos.

ILUSTRACIÓN 1.RESTOS DEL DEPORTE TLACHTLI [FUENTE: BRITANNICA]

En la dinastía Han (206 a.C. – 220 d.C.), en china se practicaba el cuju, consistía en chutar una pelota rellena de plumas a través de un aro con una red, se trataba de mantener en el aire la pelota el mayor tiempo posible uso cualquier parte del cuerpo menos las manos. Fue un deporte muy popular en esta sociedad que incluso fue financiada y apoyada por la realeza.

Las culturas antiguas mediterráneas también hacían uso de la pelota para realizar deporte, en este caso, sobre todo la civilización romana y griega, mezclaron el manejo de la pelota con la lucha y la estrategia. En la antigua Grecia se llamaba episkyros, los romanos lo adaptaron a su manera siendo un deporte muy violento que consistía en mantener a toda costa la pelota en el campo de equipo contrario, se llamaba harpastum.

Cientos de años más tarde, en el siglo XVI, nos encontramos con el calcio fiorentino en el auge de la Florencia renacentista. Se considera por muchos el precursor directo del futbol moderno, era un evento tanto social como deportivo muy importante enfrentándose entre sí familias y nobles jugándose el respeto y la admiración del pueblo.

Sin embargo, no fue hasta el siglo XIX cuando de verdad se asentaron las bases estructurales del juego. Pasó de ser un juego sin reglas y violento jugado en suburbios, a un deporte organizado jugado en campos reglados. Fue en 1863 cuando se creó la FA (football Association) en Londres cuando se establecieron las primeras reglas formales. Con la

estandarización de las normas de juego, el deporte se expandió jugando en escuelas y universidades. Este fue el primer paso para la globalización del futbol tal y como lo conocemos.

Como se puede ver, cada civilización que se ha descrito ha aportado al fútbol moderno. Pasando de ser un ritual religioso en América, a un juego de gran relevancia social durante la dinastía Han y durante la época renacentista en Florencia, hasta la unificación de criterios y normas hasta reglar el deporte en 1863 en Inglaterra gracias a la FA.

Normas y Reglas Básicas del Fútbol

Una vez puesto en contexto sobre la historia y la evolución a lo largo de los siglos de este deporte, se van a definir los conceptos básicos del futbol para una correcta compresión para aquellos que no estén familiarizados con este deporte.

El campo en que se juega debe cumplir ciertas medidas establecidas por la FIFA (Federation Internationale de Football Association), entre 90 y 120 metros de longitud y entre 45 y 90 metros de ancho, sin embargo, para partidos oficiales de competiciones FIFA, los mínimos son de 64 metros de ancho y 100 metros de largo y los máximos 75 metros de ancho y 110 metros de largo.

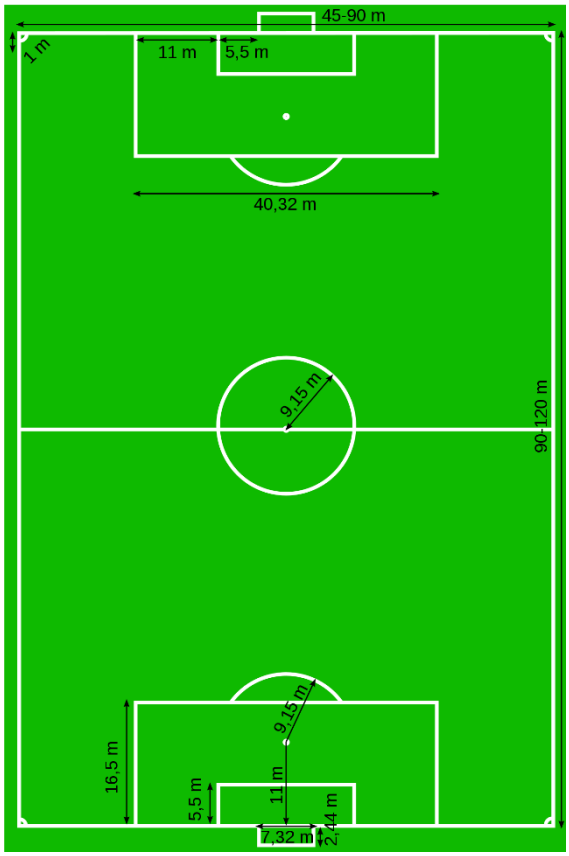


ILUSTRACIÓN 2. DIMENSIONES CAMPO DE FUTBOLE [FUENTE: CONTINUEMOSESTUDIANDO]

El partido se divide en dos tiempos de 45 minutos cada parte, entre ambos partes hay un descanso de 15 minutos. Si el árbitro considera que se ha perdido tiempo puede prolongar el partido más allá de los 45 minutos reglamentarios. Si el partido finaliza con empate en una competición que necesita un ganador se realizaran dos partes más, pero de 15 minutos cada

parte, con un breve descanso entre medias. Cuando termina esta prórroga con empate se resuelve el ganador con penaltis, más adelante se definirá lo que es un penalti.

El partido comienza con 11 jugadores de cada equipo, de estos jugadores el entrenador puede realizar cinco cambios. Sin embargo, solo pueden parar el partido tres veces para hacer cambios, es decir, para poder realizar todos los cambios vas a tener sustituir a varios jugadores en algunas de estos parones. El descanso no cuenta como parón.

Los jugadores se distribuyen por el campo en función de sus características, lo que hace que haya distintos tipos de jugadores en función de su posición.

Estas posiciones son:

- Portero: es el único jugador que dentro del rectángulo que envuelve la portería puede coger el balón con las manos. Solo puede haber uno por equipo. Estos jugadores destacan por su habilidad para leer el juego aéreo para poder anticiparse, sus reflejos, su capacidad de salto y su capacidad de liderazgo. Actualmente algunos porteros de talla mundial son Courtois, Ter Stegen o Neuer. Sin embargo, hay porteros históricos que trascienden el deporte y se impregnan en el subconsciente de la sociedad como Casillas o Buffon.

- Defensa: suelen ser los jugadores van contundentes y fuertes del equipo, es la última línea defensiva antes de llegar a portería, según el criterio del entrenador puede haber entre 3 y 5 defensa en el equipo, siendo los que están en banda (laterales) los más rápidos y habilidosos de los defensas, con el balón en el pie son capaces de poner buenos centros. En el centro de la defensa han destacado en el futbol español en los últimos años Sergio Ramos, Piqué y Puyol. En el lateral por la derecha Carvajal y Jordi Alba por la izquierda han sido los más laureados en el siglo XXI.

- Centrocampista: se pueden dividir en centrocampistas ofensivos y defensivos. Los defensivos se suelen situar muy cerca a los defensas centrales tanto en defensa como en ataque, en defensa ejerce en la práctica como si de un defensa central se tratase solo de que algo más adelantado, suelen ser jugadores fuertes con un gran físico, cuando el equipo tiene la posesión de pelota, se meten entre los centrales para así tener superioridad numérica en esa zona y poder tocar cómodamente para elaborar las jugadas.
  Respecto a los centrocampistas ofensivos, puede ser más posicionales o punzantes. Los posicionales tienen como función enlazar la defensa con los atacantes y dar ritmo al juego, sin embargo, los más punzantes ocupan posiciones más adelantados situándose entre los centrocampistas y los defensas rivales para filtrar pases a los delanteros o disparar a puertos ellos mismos. Estos jugadores suelen ser los más talentosos de la plantilla.
  En el futbol español tenemos grandes ejemplos de estos tipos de jugadores, de perfil defensivo el mejor ha sido Busquets, en una zona más posicional Xavi Hernández y más ofensivo Andrés Iniesta. Este fue el centro del campo que llevo al FC Barcelona y a la Selección Española a la gloria la pasa década.

- Delantero: los encargados por excelencia de marcar goles, se dice de estos jugadores que tienen un “olfato” goleador especial. Puede haber varios tipos de delantero, pero normalmente o son muy rápidos con una gran capacidad para romper la defensa, o son corpulentos permitiéndoles tanto aguantar el balón para dejar de cara a los

centrocampistas u otro delantero, como para rematar los balones centrados de la gente de banda como los laterales.

Además, en función de su posición en el campo pueden ser extremo, delantero centro o segundo delantero. Los extremos son los jugadores encargados de desbordar por banda a los defensas a través del regate, los delanteros centro es el jugador más adelantado del equipo cuya función principal es rematar los balones que le llegan, por último, el segundo delantero es un híbrido entre delantero centro y centrocampista ofensivo.

Por seguir la tendencia de ejemplos del futbol español de los últimos años tenemos que Raúl González es un gran ejemplo de segundo delantero siendo capaz de pese a situarse por detrás delantero centro llegar al área y conseguir marcar infinidad de goles, Fernando Torres en su mejor nivel fue considerado el mejor delantero del mundo, en 2008 metió el gol que dio a España su segunda Eurocopa, finalmente David Villa fue un claro ejemplo de todo lo que tiene que tener un extremo en el futbol, siendo capaz de regatear, marcar goles y asociarse, fue el líder de la selección española en el único mundial que ganó España en el 2010.

La manera de distribuir a los jugadores sobre el campo la decide el entrenador, las más en la actualidad es el 4-4-2 y el 4-3-3, aunque en la liga italiana también es muy común usar la 5-3-2. La manera de interpretar estas alineaciones es el primer número es la cantidad de defensas, el segundo cuantos centrocampistas hay y el ultimo son los delanteros.

Sin embargo, la suma de estos jugadores da diez, en vez de once jugadores, esto sucede debido a que la manera correcta de decir la alineación del equipo sería 1-4-4-2 por ejemplo, el 1 equivale al portero y como en todas las alineaciones va a haber uno, se omite.

El entrenador decidirá la alineación en función de la filosofía que tenga y las cualidades de los jugadores que disponga.

ILUSTRACIÓN 3. IZD: ALINEACIÓN 4-4-2 DCH: 4-3-3 [FUENTE: MAS10.AR]

Teniendo un conocimiento táctico del juego, se va a tratar de resumir las posibles faltas e infracciones que suceden en el trascurso de un partido.

Una de las reglas más complicadas de entender en el futbol es el fuera de juego. El jugador comete dicha infracción si está más cerca de la línea de gol contraria que el balón y el penúltimo defensor (penúltimo porque la mayoría de las veces el ultimo defensor es el portero). El instante en el que se valora la posición del atacante es el momento de dar el pase, además esta norma es válida a partir de la línea del centro del campo en dirección a la portería rival.

Otra consideración para tener en cuenta es que el jugador a de participar de manera activa en la jugada para señalar la infracción, es decir, que el atacante implicado reciba el balón, siga la jugada u obstaculice la visión del portero. Un ejemplo de ello:

ILUSTRACIÓN 4. COMO ES EL FUERA DE JUEGO [FUENTE: ELSUPERHINCHA]

Pese a que el fútbol es un deporte de contacto, el árbitro considerará falta cuando un jugador cargue, golpee o empuje de manera brusca o peligrosa.

Si esa falta se realiza dentro del área grande, se le concede al equipo que le han hecho falta un tiro libre desde el punto de penalti (11 metros) sin más oposición que la del portero. El tiro de penalti es una habilidad más dentro de todas las que tienen los futbolistas, cobrando una especial importancia en los últimos años con el incremento de faltas señaladas y por ende de penaltis concedidos. Los mejores lanzadores de penaltis no son solo aquellos que tienen una gran técnica y habilidad para colocar el balón donde quieren, sino que requiere una gran capacidad de soportar la presión para poder ejecutarlo de la manera correcta, algunos de estos lanzadores son Cristiano Ronaldo, Totti o Messi.

Pese a que existan reglas escritas y regladas sobre las faltas, la última palabra siempre la tiene el árbitro principal del partido. El colegiado puede considerar una falta especialmente peligrosa puede mostrar al jugador infractor tarjeta amarilla. Esta tarjeta por sí solo no acarrea sanción, pero si un jugador acumula dos tarjetas de este color será expulsado del partido y no podrá disputar el próximo encuentro. Algunos ejemplos de cuando un árbitro suele mostrar tarjeta amarilla son en acciones que considera prometedoras de ataque peligroso como agarrar al delantero en un contragolpe o pese a ir a disputar un balón, el defensor llega tarde y golpea al atacante.

El otro tipo de tarjeta es la roja. Este sí que conlleva una sanción inmediata expulsando al jugador automáticamente, además dependiendo de la gravedad del motivo de su expulsión puede ser sancionado sin disputar los próximos encuentros. Alguno de los motivos para mostrar esta tarjeta de manera directa son conductas antideportivas como agredir o escupir a algún jugador o colegiado en el terreno de juego o impedir con la mano una oportunidad manifiesta de gol.

A la hora del desarrollo del partido, las tarjetas tienen un peso muy importante ya que los entrenadores tienden a sustituir a los jugadores que ya han recibido una amonestación, para así no arriesgarse a que les saquen otra tarjeta amarilla y quedarse con un hombre menos lo que resta de partido.

En conclusión, el árbitro pese a seguir su propio criterio tiene que mantener el juego lo más seguro y justo. De esta manera habrá un mejor desarrollo del deporte y tanto futbolistas como aficionados lo disfrutarán.

CAPÍTULO 2: TEORÍA ESTADÍSTICA

En el siguiente apartado se va a describir la influencia del uso de datos en el deporte y más concretamente en el futbol, tras ello se definirán los métodos empleados para el desarrollo de este trabajo tanto descriptivos como analíticos.

Big Data y Estadística en el Deporte

Estando en plena era de la información, el uso de los datos ha crecido exponencialmente, la inferencia en dichos datos se denomina Big Data.

El empleo de Big Data usando la estadística ha revolucionado la sociedad tanto económicamente, por ejemplo, en la logística, como socialmente. El deporte está siendo un gran protagonista en estos avances, debido al gran seguimiento que tiene las sociedades actuales, es un gran introductor del conocimiento e interpretación del Big Data y su estadística a la población.

Conceptualmente, el Big Data es el procesamiento y posterior análisis de datos potencialmente masivos que han sido recopilados por fuentes de datos modernas, como pueden ser GPS o sensores. Haciendo inferencia en el deporte, estos datos pueden ser recopilados con chalecos que miden el rendimiento físico de los deportistas o con sensores que midan más concretamente sus progresos. La posibilidad de poder controlar estos datos ha proporcionado un salto en la calidad de entrenamiento en la mayoría de los deportes.

Una vez tenemos esos datos en posesión, necesitan ser procesados e interpretados. Para ello se han desarrollado diferentes técnicas como la minería de datos, la inteligencia artificial y el aprendizaje automático (más conocido como machine learning). Los expertos en la materia, analistas, son los encargados de a través de los métodos anteriormente descritos, de extraer información y sacar conclusiones, permitiendo encontrar tendencias, predecir resultados u optimizar los recursos de los que se dispone.

Big Data y Estadística en el Fútbol

Como no podía ser de otra manera, al tratarse del deporte rey, el futbol ha sido uno de los deportes que más se está implementando estas nuevas tecnologías. Otros deportes que hacen un gran uso del Big Data son el baloncesto, el béisbol, el tenis y el atletismo entre otros. Concretamente en el fútbol se utiliza en gran medida, algunas de ellas son:

- Análisis del rendimiento individual y colectivo: A través de sensores y de chalecos que llevan los jugadores, los analistas del equipo son capaces de identificar las fortalezas y debilidades tanto del equipo, como de cada jugador. Esto permite al entrenador y sus ayudantes tener más información a su disposición para tomar decisiones sobre que jugador está más forma o que modelo de juego le viene mejor al equipo en ese momento de la temporada.
- Scouting: Este ha sido el área donde más ha cambiado con la entrada del Big Data y los modelos estadísticos. Antes, a la hora de realizar fichajes o encontrar jóvenes talentos el equipo tenía que disponer de un auténtico séquito de ojeadores por todo el mundo y enviar informes a criterio de ellos, con el riesgo que eso conllevaba. Sin embargo, con la implementación del Big Data, los equipos son capaces de recopilar y almacenar datos de jugadores de todo el mundo, permitiendo predecir su potencial futuro de una manera mucho más precisa sin necesidad de tener que enviar ojeadores.

Este es un tema muy interesante que probablemente daría para un trabajo enfocado solo en él.

- Lesiones: Ha permitido crear modelos de predicción de lesiones usando la estadística, de esta manera el equipo técnico tiene más información acerca la tendencia de cada jugador para lesionarse y como evitarlas. Pese a esto, en el futbol sigue siendo un deporte muy lesivo, en gran medida por la cantidad de partidos que tiene que jugar un futbolista y más aún si compite en competiciones internacionales con su país, como Mundiales o Eurocopas.
- Táctica: Actualmente, dentro del *staff* técnico del equipo ha surgido una nueva figura, el analista táctico. Este ayudante posee tanto conocimientos tácticos de futbol elevados como una gran capacidad de análisis de datos, su principal función es mediante el estudio del rival identificar sus patrones de juego, debilidades en su táctica o crear jugadas a balón parado.
- Aficionados: La implementación en el futbol de estadísticas tanto básicas como avanzadas, ha permitido que, a través de redes sociales y aplicaciones móviles, los aficionados sean capaces de seguir en tiempo real las estadísticas del partido que deseen.

Pese a los grandes avances que ha aportado el Big Data y en general la implantación del uso de la estadística en el futbol, también hay que enfrentarse a nuevos desafíos, como el origen y la fiabilidad de los datos con los que se trabaja o la privacidad y seguridad de estos.

En definitiva, el uso de la inferencia estadística en el futbol está haciendo un gran trabajo de divulgación científica en la sociedad, ya que buena parte de los aficionados al futbol ya sean jóvenes o ancianos se empiezan a familiarizar con la ciencia de datos, el Big Data.

Métodos descriptivos

Se ha trabajado este aspecto del trabajo mediante gráficos y diagramas de dispersión. Los gráficos usados han sido los agrupados en barras según la frecuencia y un gráfico radial por sectores. A su vez, un gráfico box-plot acompaña al diagrama de dispersión.

- Gráficos:
  - Agrupados en barras según la frecuencia: Representación visual de datos categóricos. Este gráfico permite ver frecuencias de diferentes subgrupos y compararlas. En nuestro caso pueden ser diferentes temporadas, el valor de mercado según la liga. Un ejemplo:

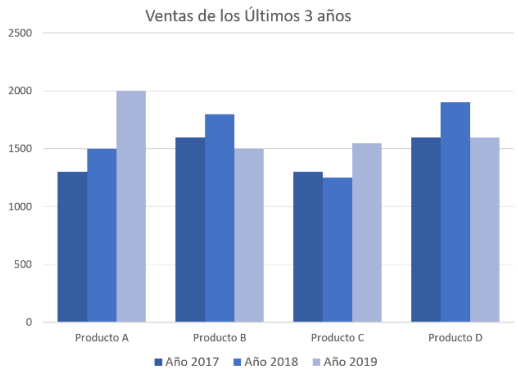


ILUSTRACIÓN 5. GRÁFICO DE BARRAS [FUENTE: PLANDEMEJORA]

  - Radial por sectores: Con una representación gráfica (circular) permite comparar visualmente distintas variables de un conjunto de datos. Cuanto más

se aleja cada variable del eje mayor será su valor. Actualmente está siendo muy usado por analistas deportivos. Un ejemplo:

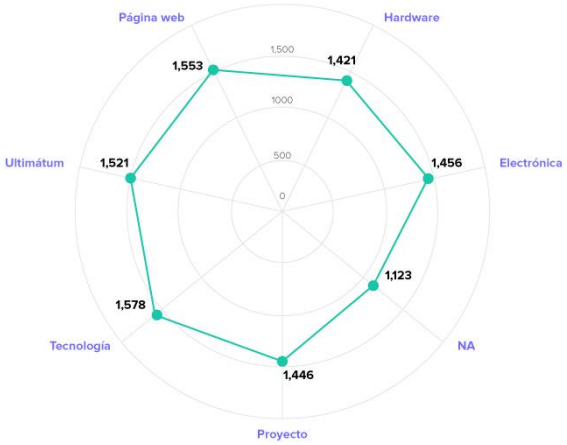


ILUSTRACIÓN 6. GRÁFICO DE RADAR [FUENTE: TUDASHBOARD]

- Box-plot: Representación gráfica de la distribución de los datos numéricos de una variable, se basa en el valor mínimo, en el cuartil uno, dos y tres (en el siguiente punto se definirán los cuartiles) y en punto máximo.

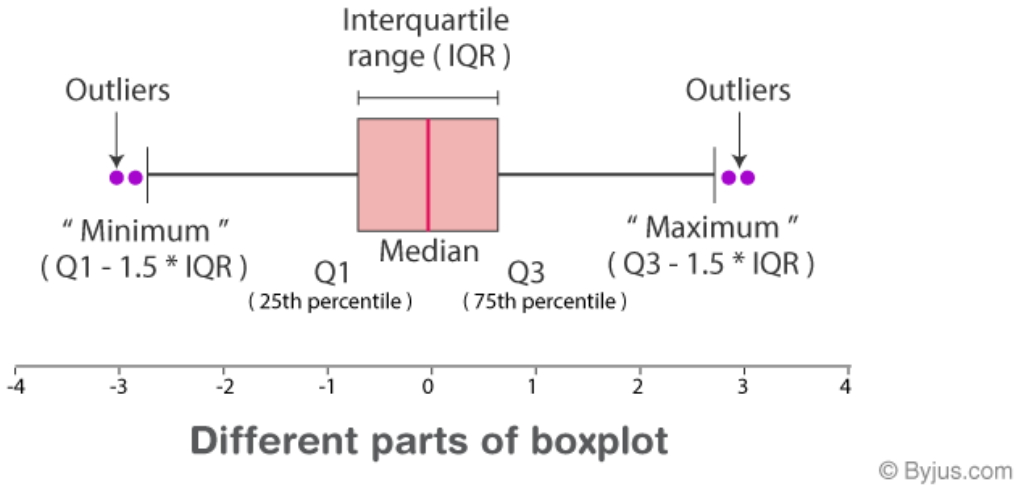


ILUSTRACIÓN 7. GRÁFICO DE BIGOTES(BOXPLOT) [FUENTE: BYJUS]

El rango intercuartílico, entre el primer cuartil y el tercero, equivale al 50% de los datos llamándose caja, los "bigotes" están en los extremos con el mínimo y el máximo, más allá de ellos están los valores atípicos, comúnmente llamados outliers.

- Diagrama de dispersión: Representación gráfica empleada para observar la relación entre dos variables numéricas. Cada punto es representado en el grafo con (x,y)

siendo x e y las dos variables a estudiar. Ese grafico permite visualizar la tendencia, correlación y dispersión de los datos. Por ejemplo:

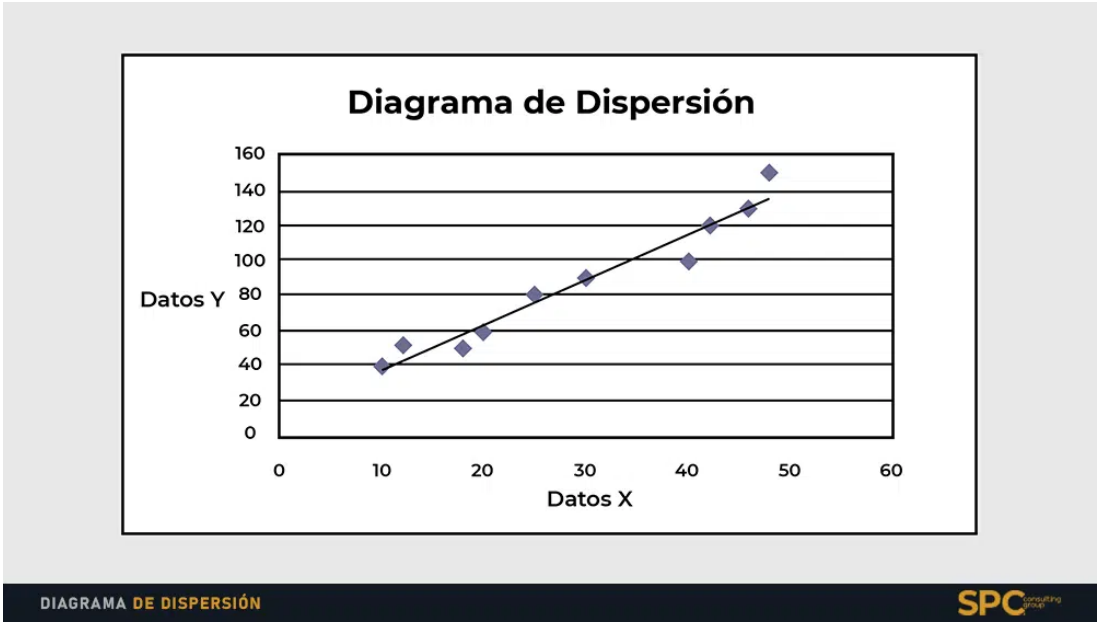


ILUSTRACIÓN 8. GRÁFICO DE DISPERSIÓN [FUENTE: SPCGROUP]

Una vez expuestos los elementos gráficos que se van a emplear en este trabajo, se debería definir aquellos términos que se consideren para una correcta lectura e interpretación de los gráficos, como los cuartiles.

Un cuartil es un tipo de medida estadística que divide el conjunto de datos en cuatro partes iguales, es decir cada cuartil representa el 25% del total de la muestra. Estos cuartiles se sitúan de la siguiente manera:

- Primer cuartil (Q1): Es el valor por el cual por debajo del mismo está el 25% de los datos.
- Segundo cuartil (Q2): Es la mediana de la muestra, divide a partes iguales los datos.
- Tercer cuartil (Q3): Es el valor por el cual el 75% de los datos están bajo el.

La fórmula para calcular el cuartil es la siguiente:

$$Q_a = L_i + \frac{\frac{aN}{4} - F_{i-1}}{f_i} A_i$$

Donde Li es el límite inferior donde se encuentra el cuartil, N es la suma de frecuencias absolutas y Ai es la amplitud del cuartil.

Métodos analíticos

Análisis de clústeres

El análisis de clústeres es una técnica de agrupamiento utilizada para dividir un conjunto de datos en subgrupos homogéneos, de manera que los objetos en el mismo grupo sean más similares entre sí que con los de otros grupos. La medida de disimilaridad (o similitud) entre los objetos es crucial para el éxito del análisis de clústeres. Las distancias y similitudes son herramientas fundamentales para evaluar las relaciones entre los objetos en el espacio multidimensional.

Criterios Basados en Distancias como Indicadores de Disimilaridad

- **Distancia Euclídea**: La distancia euclídea es la medida más común de disimilaridad entre dos puntos en un espacio euclídeo. Se calcula como la raíz cuadrada de la suma de los cuadrados de las diferencias entre las coordenadas correspondientes de los dos puntos.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Distancia Euclídea Normalizada**: Para evitar que las diferentes escalas de las variables influyan en la medida de disimilaridad, se puede utilizar la distancia euclídea normalizada. Esto se logra estandarizando las variables antes de calcular la distancia euclídea.

$$d_{normalizada}(x,y) = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - y_i}{s_i}\right)^2}$$

donde s_i es la desviación estándar de la i -ésima variable.

- **Distancia de Mahalanobis**: La distancia de Mahalanobis considera la correlación entre variables y es útil cuando las variables tienen diferentes varianzas o están correlacionadas. Se define como:

$$d_M(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)}$$

donde S es la matriz de covarianza de las variables.

- **Otras Distancias**: Además de las distancias mencionadas, existen otras métricas que se pueden utilizar en el análisis de clústeres, como:
  - Distancia Manhattan: La suma de las diferencias absolutas entre las coordenadas.
  - Distancia de Minkowski: Generalización de la distancia euclídea y Manhattan.

- Distancia Coseno: Utilizada para medir la similitud entre dos vectores en términos de su ángulo coseno.

Criterios Basados en Similaridades. Medidas de Similaridad

Las medidas de similaridad se utilizan para evaluar cuán parecidos son dos objetos. Algunas medidas comunes incluyen:

- **Coeficiente de correlación de Pearson**: Mide la correlación lineal entre dos variables.
- **Coeficiente de Jaccard**: Utilizado para comparar la similitud de conjuntos.
- **Coeficiente de similitud coseno**: Mide la similitud entre dos vectores de un espacio interno.

Medidas de Similitud y Distancia entre Grupos

- **Distancia Mínima** (Nearest Neighbour Distance): Este criterio, también conocido como enlace sencillo, define la distancia entre dos clústeres como la distancia mínima entre cualquier par de puntos, uno de cada clúster.

$$d(A,B) = min_{i \in A \in B} d(i,j)$$

- **Distancia Máxima** (Furthest Neighbour Distance): Este criterio, conocido como enlace completo, define la distancia entre dos clústeres como la distancia máxima entre cualquier par de puntos, uno de cada clúster.

$$d(A,B) = max_{i \in A, j \in B} d(i,j)$$

- **Distancia entre Centroides:** La distancia entre centroides define la distancia entre dos clústeres como la distancia entre los centroides de los clústeres.

$$d(A,B) = d(\bar{x}_A, \bar{x}_B)$$

Donde $\bar{x}_A$, y $\bar{x}_B$ son los centroides de los clústeres A y B.

Métodos de Análisis Clúster

- **Métodos Jerárquicos**

  Los métodos jerárquicos construyen una jerarquía de clústeres. Pueden ser aglomerativos (enlace ascendente) o divisivos (enlace descendente).

- **Método de la Distancia Mínima (Nearest Neighbour o Single Linkage)**

  Este método define la distancia entre dos clústeres como la distancia mínima entre cualquier par de puntos, uno de cada clúster. Se utiliza para identificar clústeres alargados.

- **Método de la Distancia Máxima (Furthest Neighbour o Complete Linkage)**

  Este método define la distancia entre dos clústeres como la distancia máxima entre cualquier par de puntos, uno de cada clúster. Se utiliza para identificar clústeres compactos.

- **Método de la Media (U.P.G.M.A.)**

  El método de la media de pares no ponderada (U.P.G.M.A.) define la distancia entre dos clústeres como la media aritmética de todas las distancias entre pares de puntos, uno de cada clúster.

- **Método del Centroide**

  Este método define la distancia entre dos clústeres como la distancia entre los centroides de los clústeres.

- **Método de la Mediana**

  Este método es una variante del método del centroide, pero utiliza la mediana en lugar del promedio para calcular el centro del clúster.

- **Método de Ward**

  El método de Ward minimiza el aumento de la suma de las varianzas dentro de cada clúster al fusionar dos clústeres. Es adecuado para identificar clústeres esféricos de tamaño similar.

- **Método Flexible de Lance y Williams**

  El método de Lance y Williams es una generalización de varios métodos de clustering jerárquico. Define una fórmula recursiva para actualizar las distancias entre clústeres a medida que se combinan.

En definitiva, el análisis de clústeres es técnica cuya principal función es segmentar los datos para predecir o averiguar patrones. Debido a la cantidad de métodos y diferentes opciones para escoger la medida que se va a emplear, va a depender del criterio del autor. La distancia va a cobrar importancia en este trabajo ya que es la escogida para realizar los análisis de clúster correspondientes.

Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística multivariante utilizada para reducir la dimensionalidad de un conjunto de datos, mientras se retiene la mayor parte de la variabilidad presente en el mismo. Este método transforma el conjunto de datos original en un nuevo conjunto de variables no correlacionadas denominadas componentes principales. Cada componente principal es una combinación lineal de las variables originales, orientada de tal manera que captura la mayor varianza posible.

El objetivo principal del PCA es simplificar el conjunto de datos original, eliminando la redundancia causada por la correlación entre variables. Para entender cómo funciona el PCA, es fundamental conocer algunos conceptos clave:

- **Varianza**: Es una medida de la dispersión de un conjunto de datos. En el contexto del PCA, la varianza de las variables es importante porque los componentes principales se seleccionan de modo que capturen la mayor varianza posible de los datos originales.
- **Covarianza**: La covarianza mide cómo dos variables cambian juntas. Una covarianza positiva indica que las variables tienden a aumentar juntas, mientras que una negativa indica que una variable tiende a aumentar cuando la otra disminuye. La matriz de covarianza se utiliza en el PCA para identificar las relaciones entre variables.

- **Autovalores y Autovectores**: Los autovalores y autovectores son fundamentales en el PCA. Los autovectores de la matriz de covarianza representan las direcciones de los componentes principales, mientras que los autovalores indican la cantidad de varianza explicada por cada componente principal.

Los pasos por seguir para realizar un correcto PCA son:

1. **Estandarización de los Datos**: Antes de aplicar PCA, es común estandarizar los datos para que todas las variables contribuyan equitativamente al análisis. Esto implica restar la media y dividir por la desviación estándar de cada variable.
2. **Cálculo de la Matriz de Covarianza**: Una vez estandarizados los datos, se calcula la matriz de covarianza para entender cómo las variables varían juntas. Para un conjunto de datos X con n observaciones y p variables, la matriz de covarianza S se calcula como:

$$S = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$$

   donde $\bar{X}$ el vector de medias de las variables.

3. **Descomposición en Autovalores y Autovectores**: A continuación, se realiza la descomposición en autovalores y autovectores de la matriz de covarianza. Los autovalores indican la cantidad de varianza capturada por cada componente principal, y los autovectores definen la dirección de estos componentes. Matemáticamente, si S es la matriz de covarianza, se resuelve la ecuación:

$$Sv = \lambda v$$

   donde λ es el autovalor y v es el autovector correspondiente.

4. **Selección de Componentes Principales**: Los componentes principales se ordenan de acuerdo con la magnitud de los autovalores, de mayor a menor. Se seleccionan los primeros k componentes que explican una proporción significativa de la varianza total del conjunto de datos. La elección de k se puede basar en un criterio predefinido, como retener el 95% de la varianza total.
5. **Transformación de los Datos**: Finalmente, los datos originales se proyectan en el espacio de los componentes principales seleccionados. Esto se realiza multiplicando los datos originales estandarizados por la matriz de autovectores correspondientes a los componentes seleccionados.

$$Z = XV_k$$

donde Z son los datos transformados, X son los datos originales estandarizados, y V_k es la matriz de autovectores correspondientes a los primeros k componentes principales.

Los resultados del PCA se interpretan examinando las cargas de los componentes principales (los coeficientes de los autovectores) y los autovalores asociados. Las cargas indican la contribución de cada variable original a los componentes principales, mientras que los autovalores proporcionan una medida de la varianza explicada por cada componente.

Componentes Principales: El primer componente principal (PC1) es la dirección en el espacio de datos que captura la mayor varianza posible. El segundo componente principal (PC2) es ortogonal al primero y captura la siguiente mayor cantidad de varianza, y así sucesivamente.

Cargas del Componente: Las cargas (o loadings) son los coeficientes que muestran la contribución de cada variable original al componente principal. Estas cargas se pueden interpretar para entender qué variables tienen mayor influencia en cada componente.

Varianza Explicada: Los autovalores asociados a cada componente principal indican la cantidad de varianza capturada por ese componente. La proporción de varianza explicada se utiliza para determinar cuántos componentes retener en el análisis.

Las principales ventajas del PCA:

Reducción de Dimensionalidad: El PCA reduce el número de variables necesarias para describir un conjunto de datos, facilitando su interpretación y análisis.

Eliminación de Multicolinealidad: Al transformar las variables originales en componentes no correlacionados, el PCA elimina la multicolinealidad, lo que puede mejorar el rendimiento de los modelos de machine learning.

Simplicidad y Eficiencia: El PCA es una técnica relativamente simple y eficiente computacionalmente, adecuada para grandes conjuntos de datos.

En cuanto a las desventajas del PCA destacan:

Pérdida de Interpretabilidad: Los componentes principales son combinaciones lineales de las variables originales, lo que puede dificultar la interpretación directa de los resultados.

Suposición de Linealidad: El PCA asume que las relaciones entre las variables son lineales, lo que puede no ser válido en todos los conjuntos de datos.

Sensibilidad a la Escala: El PCA es sensible a la escala de las variables, por lo que es crucial estandarizar los datos antes de aplicar la técnica.

El PCA se utiliza en una amplia variedad de campos y aplicaciones debido a su capacidad para simplificar conjuntos de datos complejos sin perder una cantidad significativa de información.

1. **Análisis Exploratorio de Datos**: El PCA ayuda a identificar patrones ocultos en los datos, como agrupaciones de observaciones similares o la detección de outliers.
2. **Visualización de Datos**: Para datos de alta dimensionalidad, el PCA permite la proyección en 2D o 3D, facilitando la interpretación visual de los datos.

3. **Preprocesamiento de Datos para Machine Learning**: Antes de aplicar técnicas de machine learning, el PCA se usa para eliminar la multicolinealidad y reducir el ruido en los datos, mejorando así el rendimiento de los modelos.
4. **Compresión de Datos**: En el procesamiento de imágenes y señales, el PCA se emplea para comprimir datos sin una pérdida significativa de calidad, lo cual es crucial para el almacenamiento y la transmisión eficiente.
5. **Identificación de Características Importantes**: El PCA puede ayudar a identificar las variables más importantes en un conjunto de datos, facilitando la interpretación y el desarrollo de modelos predictivos.

Se puede concluir con que el Análisis de Componentes Principales es una herramienta estadística poderosa y versátil para reducir la dimensionalidad de los datos mientras se preserva la mayor cantidad de información posible. Al transformar las variables originales en componentes principales no correlacionados, el PCA simplifica el análisis y la visualización de datos complejos. Aunque tiene algunas limitaciones, sus ventajas lo hacen invaluable para el análisis de datos en una amplia variedad de campos, desde la exploración y visualización de datos hasta el preprocesamiento de datos para machine learning. La elección de la cantidad de componentes principales a retener y la interpretación adecuada de los resultados son cruciales para aprovechar al máximo esta técnica.

Análisis de Correlación

Técnica estadística utilizada para medir y analizar la relación entre dos o más variables. Esta técnica evalúa la dirección y la fuerza de la asociación entre las variables, proporcionando una medida cuantitativa que permite entender cómo varían conjuntamente. La correlación no implica causalidad, sino que indica la medida en la que las variables están relacionadas.

Los tres principales coeficientes de correlación son los siguientes:

Coeficiente de Correlación de Pearson

El coeficiente de correlación de Pearson (r) es la medida más comúnmente utilizada para cuantificar la relación lineal entre dos variables. Este coeficiente varía entre -1 y 1, donde:

$r = 1r$ indica una correlación positiva perfecta.

$r = -1$ indica una correlación negativa perfecta.

$r = 0$ indica que no hay correlación lineal.

El coeficiente de correlación de Pearson se calcula utilizando la fórmula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

donde x_i y y_i son los valores de las variables X y Y, y $\bar{x}$ y $\bar{y}$ son las medias de X y Y.

Correlación de Spearman

La correlación de Spearman (p) es una medida no paramétrica de la asociación entre dos variables, basada en los rangos de los datos en lugar de sus valores absolutos. Se utiliza cuando las variables no siguen una distribución normal o cuando la relación entre ellas no es lineal. Se calcula utilizando la fórmula:

$$p = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia entre los rangos de las dos variables para el i-ésimo par de observaciones, y n es el número de observaciones.

Correlación de Kendall

La correlación de Kendall (k) mide la asociación ordinal entre dos variables. Se basa en la concordancia y discordancia de los pares de observaciones. Su fórmula es:

$$k = \frac{(C - D)}{\sqrt{(C + D + T)(C + D + U)}}$$

donde C es el número de pares concordantes, D es el número de pares discordantes, T es el número de pares con la misma ordenación en X pero no en Y, y U es el número de pares con la misma ordenación en Y pero no en X.

Pasos para Realizar un Análisis de Correlación

1. **Recopilación de Datos**: El primer paso en cualquier análisis de correlación es la recopilación de datos. Es importante asegurarse de que los datos sean precisos y relevantes para el análisis.
2. **Visualización de Datos**: Antes de calcular los coeficientes de correlación, es útil visualizar los datos mediante gráficos de dispersión para tener una idea inicial de la relación entre las variables.
3. **Cálculo del Coeficiente de Correlación**: Dependiendo de la naturaleza de los datos, se elige y se calcula el coeficiente de correlación apropiado (Pearson, Spearman, Kendall).
4. **Interpretación de Resultados**: Los coeficientes de correlación se interpretan en términos de dirección (positiva o negativa) y fuerza (débil, moderada, fuerte) de la relación. También se considera la significancia estadística del coeficiente, que indica si la correlación observada es probablemente debida al azar.

El análisis de correlación se utiliza en una amplia variedad de campos y aplicaciones debido a su capacidad para identificar y cuantificar relaciones entre variables.

Investigación Científica: En estudios científicos, la correlación se utiliza para explorar y establecer relaciones entre variables. Por ejemplo, en biología se puede usar para investigar la relación entre distintas características de organismos.

Economía y Finanzas: En economía, la correlación se utiliza para analizar la relación entre variables económicas, como la inflación y el desempleo. En finanzas, se utiliza para medir la relación entre los rendimientos de diferentes activos financieros.

Psicología y Ciencias Sociales: En psicología, el análisis de correlación ayuda a investigar las relaciones entre diferentes variables psicológicas, como la autoestima y el rendimiento académico. En ciencias sociales, se usa para analizar datos de encuestas y estudios sociales.

Marketing y Negocios: En marketing, se utiliza para entender la relación entre diferentes variables de mercado, como la publicidad y las ventas. En negocios, se aplica para analizar datos operativos y de desempeño.

Sus principales ventajas:

Simplicidad y Facilidad de Interpretación: El coeficiente de correlación es fácil de calcular e interpretar, proporcionando una medida clara de la relación entre dos variables.

Versatilidad: El análisis de correlación se puede aplicar a diferentes tipos de datos y en una variedad de campos.

Identificación de Relaciones Iniciales: Ayuda a identificar relaciones iniciales entre variables, lo cual puede ser útil para estudios exploratorios.

Desventajas:

No Implica Causalidad: La correlación no implica causalidad. Una alta correlación entre dos variables no significa que una causa la otra.

Sensibilidad a Valores Atípicos: El coeficiente de correlación de Pearson es sensible a los valores atípicos, que pueden distorsionar la medida de la relación.

Limitación a Relaciones Lineales: El coeficiente de correlación de Pearson mide solo relaciones lineales, por lo que puede no capturar relaciones más complejas entre variables.

Interpretación de Resultados

Correlación Positiva Fuerte: Un coeficiente de correlación de Pearson cercano a 1 indica una relación positiva fuerte entre las variables. Por ejemplo, si $r = 0{,}9$ para las variables de altura y peso, sugiere que a medida que la altura aumenta, el peso también tiende a aumentar de manera significativa.

Correlación Negativa Moderada: Un coeficiente de correlación de Pearson de $r = -0{,}5$ indica una relación negativa moderada entre las variables. Por ejemplo, si $r = -0{,}5$ para las variables de consumo de alcohol y rendimiento académico, sugiere que a medida que el consumo de alcohol aumenta, el rendimiento académico tiende a disminuir, aunque no de manera muy fuerte.

Correlación Débil o Nula: Un coeficiente de correlación cercano a 0 indica una relación débil o nula entre las variables. Por ejemplo, si $r = 0.$ para las variables de horas de sueño y rendimiento en un examen, sugiere que no hay una relación significativa entre estas variables.

Un ejemplo gráfico de los distintos tipos de correlación:

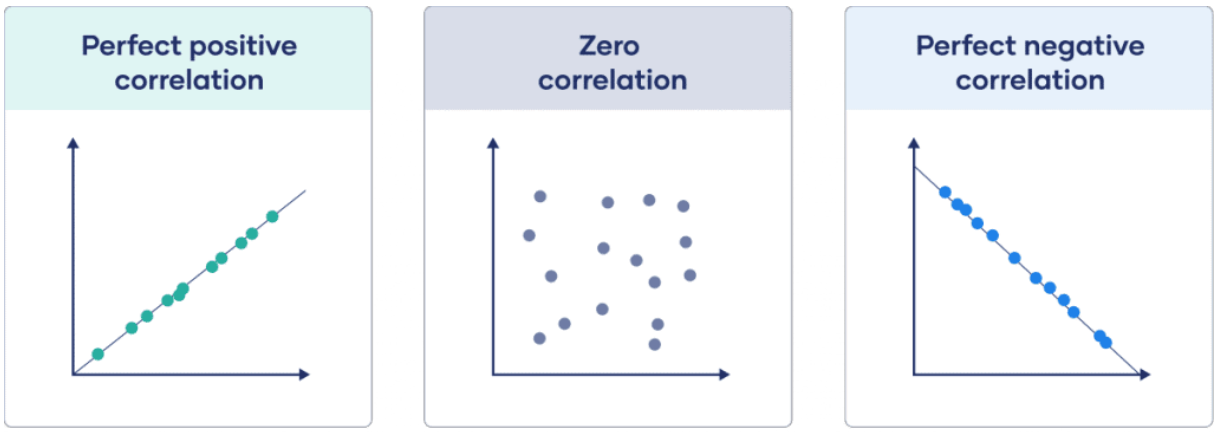


ILUSTRACIÓN 9. TIPOS DE CORRELACIONES [FUENTE: NEXTSCENARIO]

El análisis de correlación es una herramienta estadística esencial para explorar y cuantificar las relaciones entre variables. A través del uso de diferentes coeficientes de correlación, como Pearson, Spearman y Kendall, es posible medir la dirección y la fuerza de estas relaciones en una variedad de contextos y campos. Aunque el análisis de correlación tiene algunas limitaciones, como su incapacidad para inferir causalidad y su sensibilidad a valores atípicos, sus ventajas lo hacen indispensable para el análisis de datos. Entender y aplicar correctamente el análisis de correlación puede proporcionar valiosas percepciones y guiar investigaciones y decisiones basadas en datos.

CAPÍTULO 3: PROGRAMA ESTADÍSTICO

Las personas están muy limitadas si quieren trabajar con grandes bases de datos y muestras a mano, por eso, impulsada por la Segunda Guerra Mundial, se empezaron a desarrollar los primeros almacenes de datos computacional.

La primera fue la Z3 de Konrad Zuse (1941), es considerada la primera computadora programable, pese a que no era capaz de analizar datos, su sucesora Harvand Mark I fue usada para cálculos científicos y militares.

En 1948, el Proyecto RAND impulso el uso de la estadística en la informática, dándole así un enfoque distinto y logrando grandes avances.

Como consecuencia, la empresa IBM construyo el software IBM 704, siendo una de las pioneras en los cálculos estadísticos complejos. A raíz de la popularización en el mundo científico del potencial que tenía la estadística para resolver graves problemas a los que enfrentaba el mundo en aquellos momentos, como por ejemplo como gestionar logísticamente la globalización.

En la actualidad se siguen desarrollando lenguajes de programación, en función del área en la que se está trabajando es más recomendable el uso de uno o de otro.

Los lenguajes de programación más usados en el ámbito del análisis estadístico son:

- Python: El lenguaje más usado en el mundo de la estadística. Muy potente a la par de versátil, popular para trabajar con Inteligencia Artificial y aprendizaje automático.
- R: Programa de software libre especializado en el análisis estadístico. Su popular reside en masiva cantidad de paquetes que tiene capaz de realizar cualquier función estadística o grafica.
- SQL (Structured Query Language): Pese a que no sea un lenguaje de programación en sí mismo, se usa como consultor y gestor de granes bases de datos siendo el paso anterior para realizar el análisis estadístico correspondiente.
- SPSS: El lenguaje más sencillo de todos, aunque sea menos potente y esté más anticuado que el resto, suele ser un perfecto lenguaje para las personas que están aprendiendo.

En este trabajo solo se va a usar el lenguaje de programación RStudio.

Es un entorno de desarrollo integrado (IDE). IDE significa que RStudio está diseñado con el lenguaje de programación R, aun así, también es compatible con los lenguajes de Python y SQL.

Fue desarrollado por la empresa PBC, anteriormente conocida como RStudio, en 2009. El cofundador de la empresa, JJ Allaire, fue el primero desarrollar el lenguaje, tenía experiencia en el sector ya que había sido uno de los integrantes de la creación de R. Su objetivo era optimizar, facilitar y modernizar al usuario el lenguaje que habían creado anteriormente. De esa manera salió al mercado RStudio en febrero de 2011.

Algunas ventajas por las cuales se ha decidido usar este lenguaje en el trabajo son que proporciona una consola interactiva que permite acceder rápidamente a todas las funciones, comandos y scripts que se ha ido obtenido ejecutando el código, permitiendo al usuario tener en todo momento controlado que parte del código se está ejecutando, además otro aspecto a tener en cuenta es la facilidad con que se manipular y exportar los gráficos obtenidos, facilitando de esa manera la elaboración de este trabajo.

El inconveniente que se ha encontrado realizando este trabajo, ha sido en el análisis descriptivo de la base de datos. Mientras que con RStudio realizas los gráficos que se consideren y se plasman en el trabajo sin que el lector pueda realizar ningún cambio, con la extensión de R, RShiny el lector podría haber usado los gráficos descriptivos y manipularlos a su antojo, de esta manera hubiera dado la flexibilidad al usuario de poder realizar sobre la base de datos las consultas que precise. Esta, es una buena forma de mejorar el trabajo en un futuro.

CAPÍTULO 4: LA BASE DE DATOS

En este apartado, se va a relatar como ha sido la obtención y los cambios que se han considerado pertinentes en la base de datos con la que se ha trabajado.

Obtención de datos

En primer lugar, se ha obtenido nuestra base de datos de Kaggle. Kaggle es una empresa subsidiaria de Google, en la cual ofrece una plataforma comercial de datos públicos y basada en la nube, permite además publicar conjuntos de datos, compartir y participar en concursos para resolver desafíos de ciencia de datos. En resumen, es una plataforma donde los científicos de datos pueden desarrollar su aprendizaje en el análisis de datos obteniendo bases de datos fiables y poniéndolas en común con la comunidad científica.

La base de datos empleada para llevar a cabo la investigación son los registros que usa el videojuego FIFA (Federation Internationale Football Associaton) 2019.

El videojuego FIFA es una serie de videojuegos de simulación de futbol bajo el desarrollo de EA Sports. Nació en diciembre de 1993, siendo pionero en la experiencia de juego en 3D, distanciándose de otros competidores que usaban gráficos en 2D. Es una franquicia que saca al mercado un nuevo juego todos los años, en el año 1996 se hizo con la licencia FIFPro lo que le permitió usar los nombres legales de jugadores, equipos, estadios…

Con el paso de los años la franquicia ha tenido un gran impacto en el fútbol moderno, los aficionados, sobre todo los más jóvenes, reciben gracias al FIFA una experiencia inmersiva que se complementa con la realidad del fútbol tanto amateur como profesional.

En qué consisten los datos

En FIFA, hay 28 atributos si eres jugador de campo y 33 si eres portero. Estos atributos tienen como valor máximo 99 y como mínimo 1, reflejan todas las habilidades y características reales de cada jugador aplicándolo al videojuego para que así sea lo más fiel a la realidad.

La calificación general (overall rating) de cada jugador es la evaluación global de su habilidad y rendimiento, pero resumido en un solo número. Para realizar esto, sabemos que EA Sports usa algún tipo de algoritmo (no es público) y depende de la posición en el campo que desempeña y la ponderación que tiene en los atributos claves para la posición que juega.

Algunos de esos atributos claves son:

- Físicos: aceleración y sprint
- Tiro: finalización, potencia de tiro, …
- Pase: visión, centros, …
- Regate: agilidad, control de balón, …
- Defensa: marcaje, intercepciones, …
- Físicos: fuerza, salto, …
- Portero: reflejos, estirada, …

Cambios en la base de datos original

Tras observar la base de datos original te das cuenta de que no existe una variable que divida los clubs por los distintos países. Para ello, se ha empleado la variable Club para agruparlas, de manera que ahora tenemos una nueva variable (League).

Seguimos buscando posibles problemas en las variables de la base de datos original para poder ejecutar correctamente nuestros métodos. En las variables Value y Wage, el valor del

jugador y su sueldo respectivamente, vemos que estas definidas como variables discretas, por lo que las transformamos a variables continuas.

Además, para facilitar la comprensión y los cálculos con estas variables, se ha creado una nueva variable a partir cada una, de tal manera que ahora un jugador que valga €100M pasará a verse 100000000. De esta manera los gráficos y en general los cálculos se verán más claros.

Otra variable por modificar va a ser Position, ya que para la gente que no está familiarizada con términos tácticos futbolísticamente hablando, son algo confusos:

```
> unique(df$Position)
 [1] "RF"  "ST"  "LW"  "GK"  "RCM" "LF"  "RS"  "RCB" "LCM" "CB"  "LDM" "CAM" "CDM" "LS"  "LCB" "RM"  "LAM" "LM"  "LB"  "RDM" "RW"  "CM"  "RB"  "RAM"
[25] "CF"  "RWB" "LWB" ""
```

Se va a simplificar todas estas posiciones clasificándolas en: portero, defensa, centrocampista (midfielder) y delantero (forward). A su vez, al distribuir de esta manera los va a permitir un mejor manejo de los datos.

```
unique(df$Position)
defence <- c("CB", "RB", "LB", "LWB", "RWB", "LCB", "RCB")
midfielder <- c("CM", "CDM","CAM","LM","RM", "LAM", "RAM", "LCM", "RCM", "LDM", "RDM")

df %<>% mutate(Class = if_else(Position %in% "GK", "Goal Keeper",
                               if_else(Position %in% defence, "Defender",
                                       if_else(Position %in% midfielder, "Midfielder", "Forward"))))

rm(defence, midfielder)
```

Al ser FIFA de propiedad anglosajona, las medidas de peso y altura están en libras y pies, algo que para el resto nos resulta difícil de comprender, por lo tanto, se cambiara a kilogramos y centímetros:

```
df %<>%
  mutate(Height = round((as.numeric(str_sub(Height, start=1,end = 1))*30.48)
                        + (as.numeric(str_sub(Height, start = 3, end = 5))* 2.54)),
         Weight = round(as.numeric(str_sub(Weight, start = 1, end = 3)) / 2.204623))
```

Ajustamos el tipo de variable que es “Preferred.Foot”, convirtiéndola en factor para poder trabajar con ella.

Leyendo el nombre de las variables nos damos cuenta de que al ser la mayor parte de ellas compuestas por dos palabras, resulta tedioso, por lo cual se han renombrado algunas de ellas:

```
df %<>%
  rename(
    "Heading.Accuracy"= HeadingAccuracy,
    "Short.Passing"= ShortPassing,
    "FK.Accuracy" = FKAccuracy,
    "Long.Passing"= LongPassing,
    "Ball.Control"= BallControl,
    "Sprint.Speed"= SprintSpeed,
    "Shot.Power"= ShotPower,
    "Long.Shots"= LongShots,
    "Standing.Tackle"= StandingTackle,
    "Sliding.Tackle"= SlidingTackle,
    "GK.Diving"= GKDiving,
    "GK.Handling"= GKHandling,
    "GK.Kicking"= GKKicking,
    "GK.Positioning"= GKPositioning,
    "GK.Reflexes"= GKReflexes
  )
```

Por último, seleccionamos las variables con los que no vamos a trabajar y procedemos a su eliminación. Estas variables son: ID, Body.Type, Real.Face, Joined, Loaned.From, Release.Clause, Photo, Flag, Special y Work.Rate.

La mayoría de estas variables corresponden a su identidad, algo que no tiene relevancia estadística, como la foto de su cara, su bandera o cuando se unió al club.

CAPÍTULO 5: RESULTADOS

Al tratarse de un trabajo estadístico, se ha trabajado primero empleando métodos estadísticos descriptivos y finalmente aplicando métodos inferenciales en nuestra base de datos.

En este trabajo en concreto, se va a incidir de una manera exhaustiva con los descriptivos ya que el objetivo de este trabajo es acercar al usuario estándar del fútbol con escasos conocimientos a la estadística. Por ello, se ha trabajado en una base de datos con la que seamos capaces de ver de una manera sencilla información interesante, alguna de esa información es la que se ha obtenido.

Para los usuarios que hayan querido ir más allá de los descriptivos, se han desarrollado unos métodos estadísticos de inferencia: la similitud entre jugadores usando las distancias como indicadores de disimilaridad, siendo esta el principal objetivo del estudio y la correlación de dos variables, la potencia de tiro y la finalización, si hubiera dicha correlación estudiar si hay diferencias significativas estadísticamente hablando si separamos los jugadores según su pierna buena. De esta manera se podrá averiguar si el mito de que los zurdos tienen tan buen golpeo de balón es verdad, o solo es un mito.

Por último, se ha realizado un análisis de componentes principales sobre los delanteros de nuestra base de datos, resumiendo sus habilidades en una puntación (PCA Score) y comparándola con la valoración dado por el propio FIFA (Overall).

Descriptivos

A continuación, la base de datos va a empezar a ser manipulada. Debido a los usuarios objetivo que tiene este trabajo, se va a mostrar como de una manera sencilla podemos sacar información y obtener conclusiones de gráfico descriptivos a través de RStudio. Los gráficos son:

- gráfico para ver la distribución de los jugadores según su edad
- gráfico para ver visualmente las diferencias entre el potencial de los jugadores según su edad y liga en las que están
- gráfico del valor de mercado de cada liga
- comparación visual de dos jugadores
- estadísticas medias de los jugadores por posición en la liga española

Gráfico para ver la distribución de los jugadores según su edad

Los gráficos usados son gráficos de barras simples. Siendo el eje de las X los años que tiene cada jugador y el eje de las Y las veces de que un jugador tiene esa edad, todo ello dividiendo los jugadores en función de la liga en la que juegan.

Además, se ha añadido visualmente la edad media de los jugadores en cada liga.

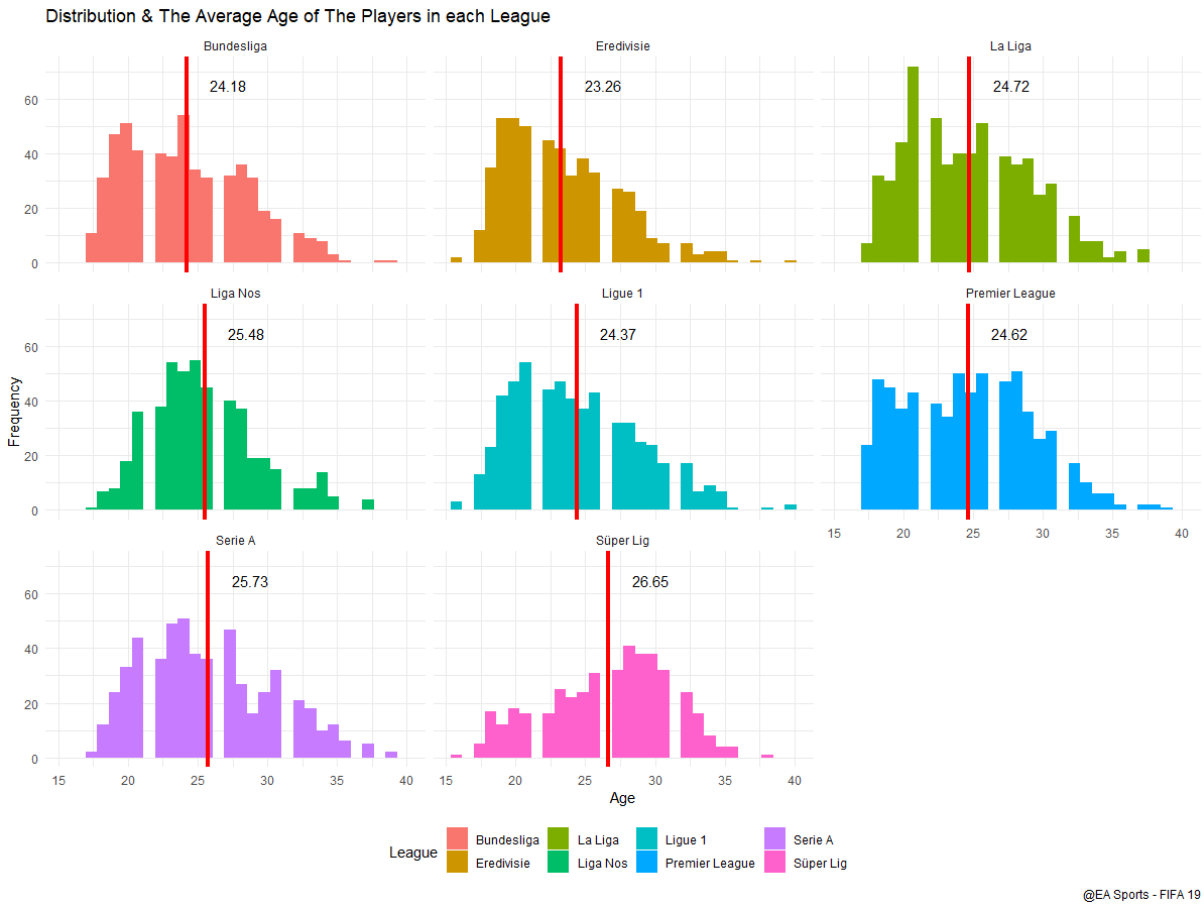


ILUSTRACIÓN 10. PLOT 1

Se puede observar como la diferencia entre la liga con más edad media y la que menos es de algo menos de tres puntos y medio (Süper Lig = 26,65 / Eredivisie = 23,26). Esto corresponde a que en general en la alta élite de los deportes y más concretamente en el fútbol, la franja de edad que se considera óptima para un pleno rendimiento es de entre los 20 y los 30 años.

Aun así, con los avances médicos y en el ámbito de salud deportiva, vemos cada vez más adolescentes de menos de 20 años en la élite, a su vez los futbolistas también alargan más su vida deportiva gracias a estos avances. Dando lugar a que esa horquilla de 20-30 años se esté quedando anticuada.

Pese a que esto que estamos viendo es una norma general en las ocho ligas, hay grandes diferencias entre ellas. Cogiendo de ejemplo los extremos (la que más edad media y la que menos), podemos intuir que la liga holandesa (Eredivisie) está plagada de jóvenes talentos buscando un lugar en el complejo mundo del futbol, ya que a medida que nos movemos en el eje de las X (mayor edad), la frecuencia de jugadores disminuye. Esto pasa también, en la liga portuguesa (Liga Nos).

Respecto a la liga turca (Súper Lig), tenemos un gráfico totalmente distinto, nos encontramos mayor número de jugadores a medida que aumentamos la edad. Debido a que la liga turca no es una de las grandes ligas del continente vemos como muchos jugadores de “avanzada” edad deciden fichar por equipos turcos.

Con estos ejemplos claros, podemos ver como las ligas europeas competitivas pero que no entran al BIG 5 (La Liga, Premier League, Bundesliga, Ligue 1, Serie A) se dividen en dos tipos de ligas, las que tienen un músculo financiero importante y por lo tanto atraen a los jugadores veteranos para ofrecerles un último gran contrato en su carrera; y los que en lugar de tener ese músculo financiero, apuestan por jóvenes talentos para desarrollarlos y poder sacar un buen rédito de ellos vendiéndolos a equipos del BIG 5 normalmente.

Gráfico para ver visualmente las diferencias entre el potencial de los jugadores según su edad y liga en las que están

Como se ha mencionado anteriormente, en el FIFA, cada jugador obtiene una calificación final (overall). Pero también se les otorga una puntación potencial, la cual se define como la calificación máxima que puede conseguir un jugador a lo largo de carrera.

Esta puntación potencial es una proyección que se basa en:

- Edad: A medida que aumenta la edad del jugador disminuye su potencial.
- Calificación actual: Cuanta más elevada sea la calificación actual del jugador más elevado puede ser su potencial.
- Posición: Por norma general tanto los defensas como los porteros suelen tener más competencia, por lo tanto, menos minutos para desarrollarse. Mientras que centrocampistas o delanteros suelen rotar más dentro de un equipo y se les da más oportunidades.
- Rendimiento: El desempeño de un jugador tanto en partidos como en entrenamientos es la parte más determinante, si un jugador es capaz de encadenar buenas actuaciones sobre el terreno de juego, se acercará a su potencial máximo.

En este ejemplo podemos ver un jugador que ya ha llegado al culmen de su carrera y su calificación general ya ha alcanzado su potencial máximo:

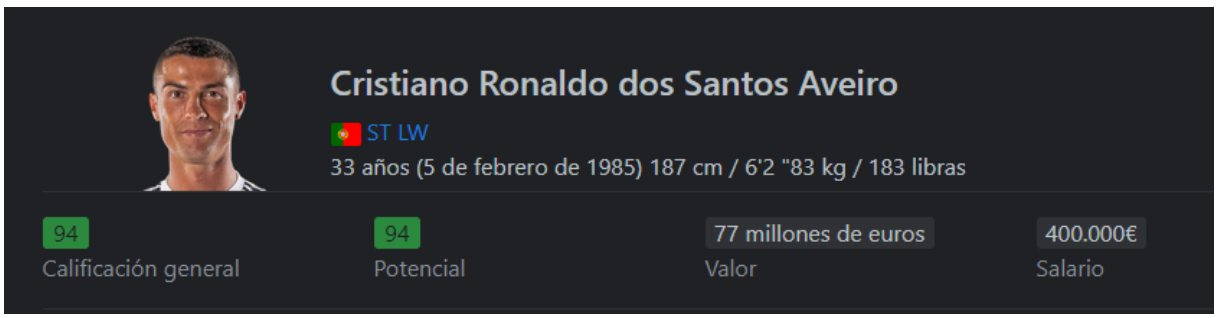


ILUSTRACIÓN 11. RONALDO FIFA 2019 [FUENTE: SOFIFA]

https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/190075/

Por el contrario, en este ejemplo vemos un jugador con un potencial enorme que está por ver si algún día consigue alcanzar.

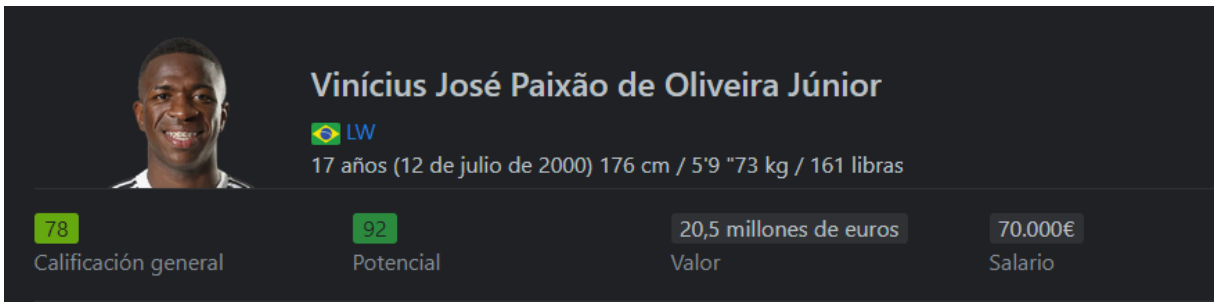


ILUSTRACIÓN 12. VINICIUS FIFA 2019 [FUENTE: SOFIFA]

https://sofifa.com/player/238794/vinicius-jose-de-oliveira-junior/190075/

Una vez definido con lo que vamos a trabajar se extraen los siguientes gráficos:

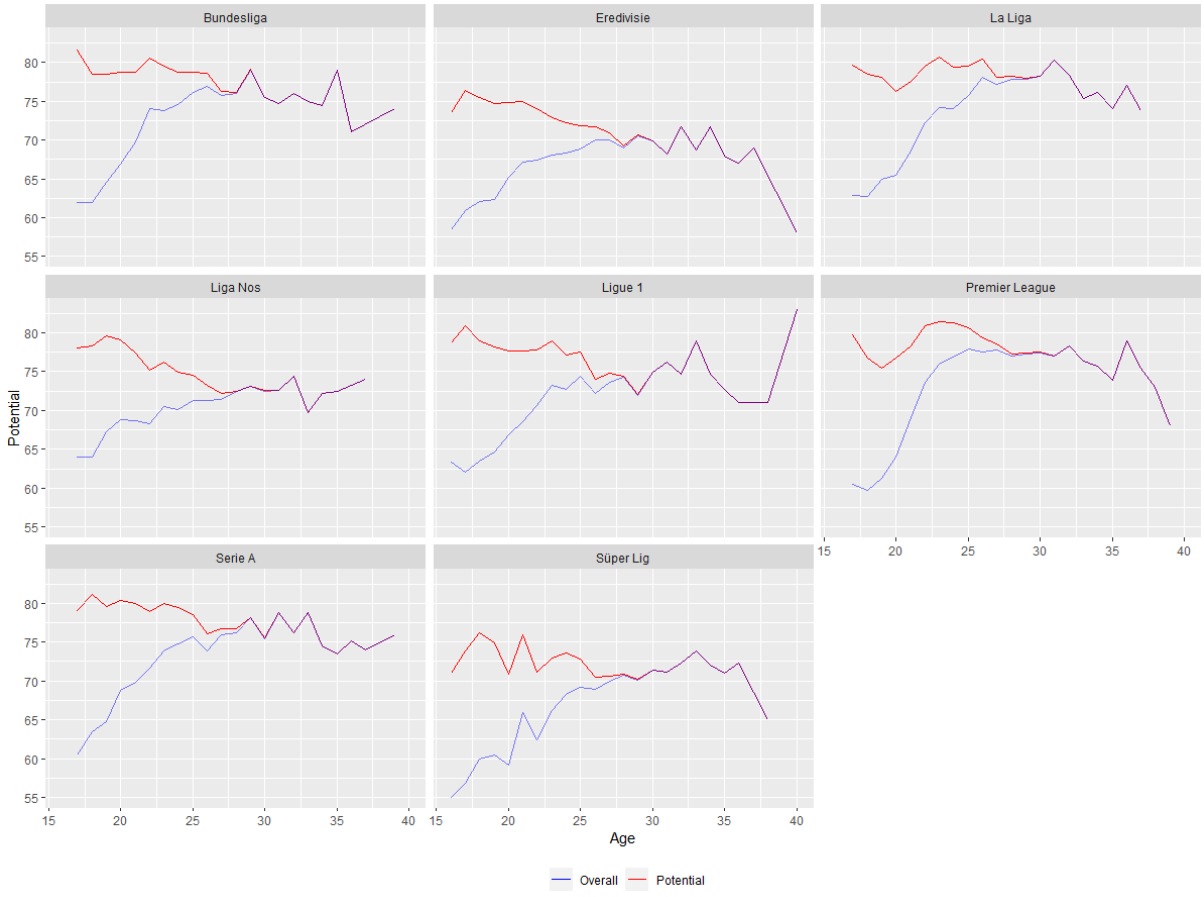


ILUSTRACIÓN 13. PLOT 2

Siendo la línea azul la puntuación actual de los jugadores y la roja la potencial, podemos ver como a medida que los jugadores cumplen años van acercándose a su potencial máximo.

Todos los gráficos coinciden en que torno a los 28 años el potencial y su puntuación se solapan. Por lo tanto, vemos como se cumple lo comentado en el anterior gráfico sobre la horquilla de edad donde se explota el máximo rendimiento del jugador (20 – 30 años).

A partir de la llegada a la etapa de máxima madurez de un futbolista, se mantienen estables en ese nivel unos años hasta que paulatinamente el rendimiento y por tanto su puntuación disminuyen.

Gráfico del valor de mercado de cada liga

Que mejor manera de comprobar el dominio europeo del BIG 5 y el papel que desempeñan las ligas secundarias como la portuguesa (Nos), la holandesa (Eredivisie) y la turca (Süper Lig), que el valor total de cada liga. Es decir, la suma del valor de cada jugador correspondiente a cada liga.

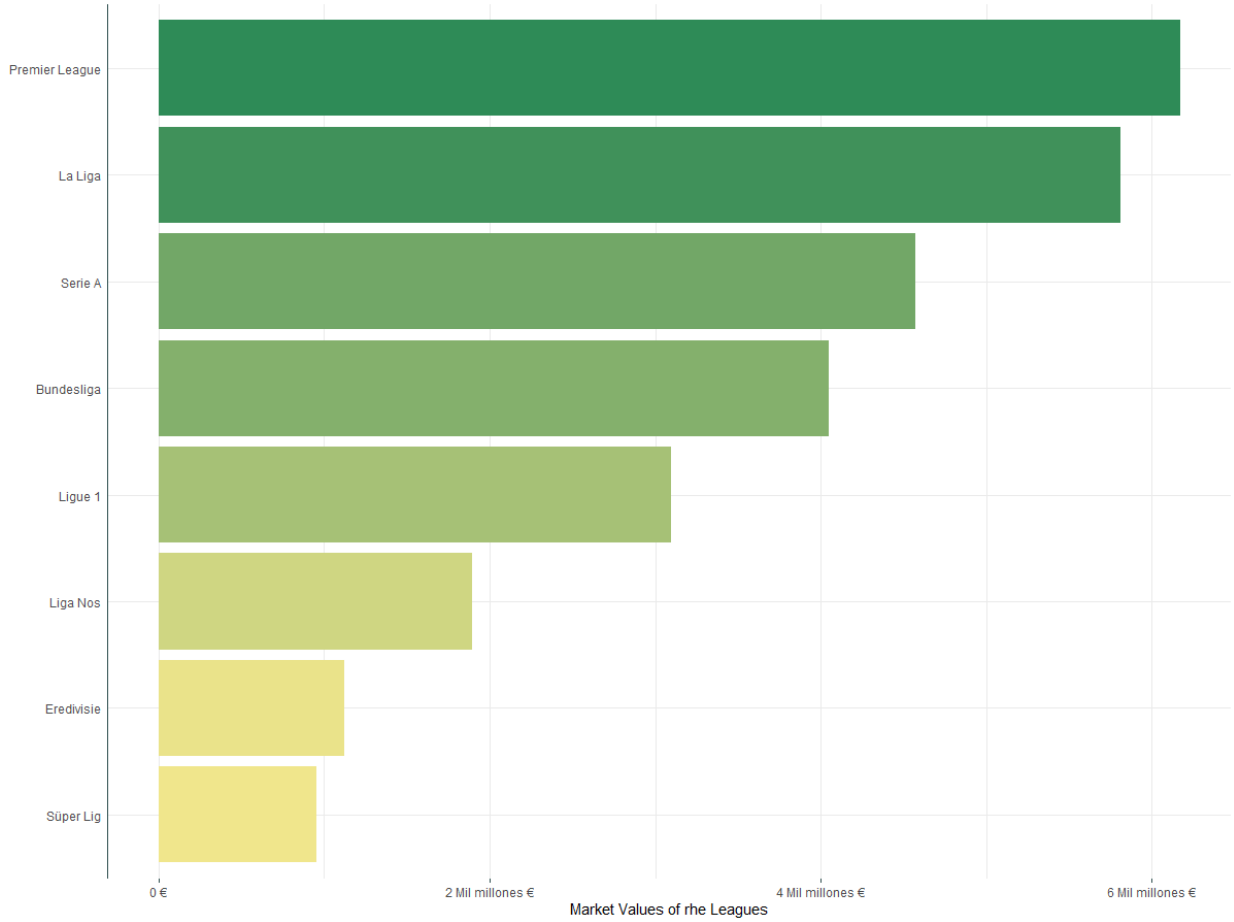


ILUSTRACIÓN 14. PLOT 3

Con el tono verde más oscuro destacan la Premier League y La Liga, con un valor superior a los 6 mil millones de euros y un valor cercano a los 6 mil millones de euros respectivamente. Por lo tanto, las dos ligas con más valor.

Un escalón por debajo se encuentra la Serie A y la Bundesliga, la primera con un valor de 4,5 mil millones de euros y la segunda con un valor ligeramente superior a 4 mil millones de euros.

Cerrando el BIG 5 tenemos la Ligue 1 con un valor que ronda los 3 mil millones de euros.

La Liga Nos destaca siendo el puente entre las grandes ligas europeas y el resto, sobresaliendo del resto de ligas “menores” con un valor cercano a los 2 mil millones de euros.

Vemos en los últimos puestos tenemos a la Eredivisie y la Súper Lig respectivamente, podemos ver como pese a tener modelos de liga totalmente contrarios (la primera centrada en jóvenes promesas a desarrollar para vender y la segunda en dar la oportunidad a jugadores veteranos que “poco” valor de mercado, pero con calidad aún), ambos tienen un valor de

mercado muy similar siendo la Eredivisie con algo más de mil millones de euros la que más tiene.

Comparación visual de dos jugadores

Se ha considerado importante que el lector pueda tener acceso a una herramienta visual como la que se a mostrar a continuación, que permita ver una comparación intuitiva sobre los dos jugadores que considere. En este caso se ha optado por Leo Messi y Cristiano Ronaldo ya que han sido los dos mejores jugadores y más mediáticos de los últimos años en el fútbol.

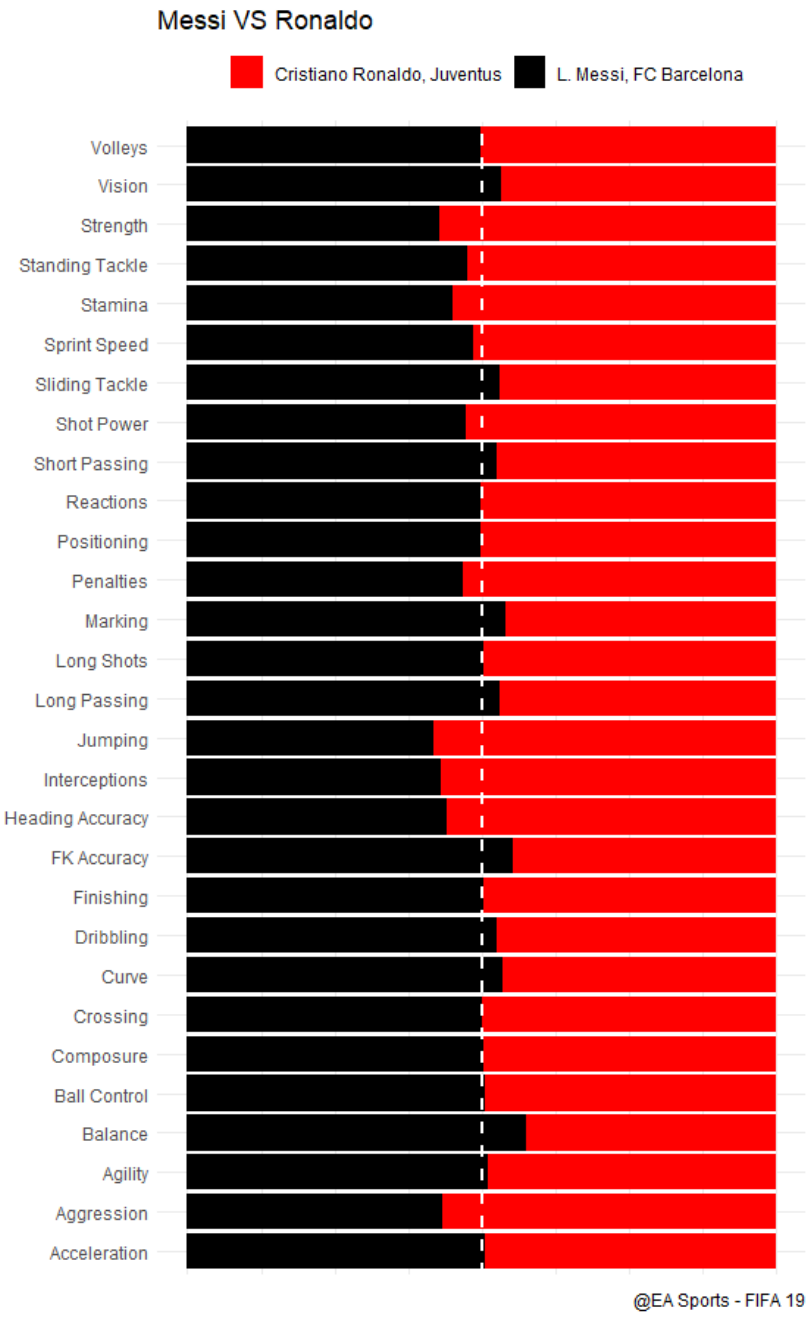


ILUSTRACIÓN 15. PLOT 4

Salta a la vista la enorme igualdad que tienen estos dos astros del fútbol mundial. Las ligeras diferencias son que en los atributos físicos como el salto (Jumping), la resistencia (stamina) o

la fuerza (strength) tiene ventaja Ronaldo; mientras que Messi tiene una mayor puntación en los atributos relacionados con la técnica de pase, como pueden ser: los pases en largo (Long Passing), visión (visión) o pases en corto (short passing).

Estadísticas medias de los jugadores por posición en la liga española

Con el gráfico que vamos a ver a continuación, se le va a dar un enfoque distinto visualizando una serie de atributos (Sprint.Speed, Balance, Dribbling, Finishing y Short.Passing), en función de la posición de los jugadores de La Liga.

La manera correcta de interpretar el gráfico es siguiendo el vector con origen en el centro del círculo y final en el atributo, se podrán apreciar dos posiciones diferentes a cada lado de dicho vector.

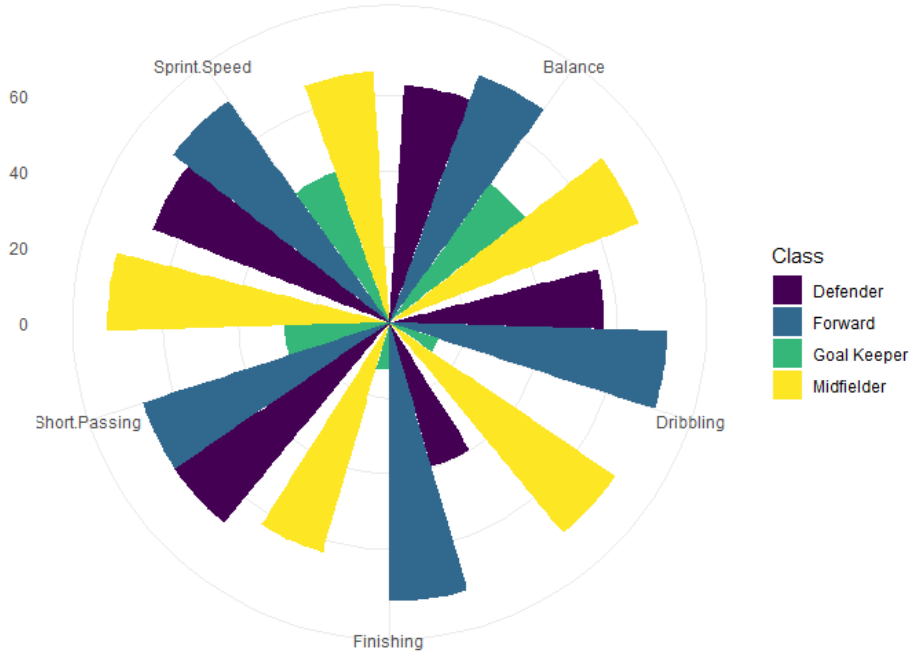


ILUSTRACIÓN 16. PLOT 5

Respecto a la velocidad punta (Sprint.Speed), destacan los delanteros (Forward) de manera obvia, ya que es una posición cuya función principal es penetrar la línea defensiva para marcar. Caso similar ocurre con la finalización (Finishing), pese a que los centrocampistas (midfielder) suelen tener una buena finalización ya que es común que realicen disparos de media distancia a portería, los delanteros es la posición que mejor finaliza por razones obvias.

En cuanto a la calidad para realizar pases en corto (Short.Passing), tanto los defensas (defender) como los centrocampistas tienen una elevada puntación. Los defensas han evolucionado con el tiempo, hasta hace unos años bastaba con ser capaces de hacer cambios de orientación sin excesivas complicaciones (pases en largo), sin embargo, ahora se les exige sacar el balón jugado desde su campo desarrollando esta faceta del juego. Como consecuencia de ellos, tenemos que esta calidad es similar a la de los centrocampistas.

La capacidad de regatear (Dribbling) y balance (balance) de cada jugador van muy ligadas, esto resulta que se muestren resultados muy parecidos en ambas variables, siendo los

delanteros y los centrocampistas los más destacados. Como norma general, los jugadores que estén en banda independientemente de si son una posición u otra, suelen tener una capacidad mayor de regate, y por lo tanto de balance, ya que son los jugadores del equipo cuya finalidad es el desborde.

Analíticos

Los métodos analíticos anteriormente mencionados se van a desarrollar en este apartado.

La similitud entre jugadores

Para estudiar la distancia que hay entre jugadores, es decir, buscar los jugadores con características más similares se ha dado la opción al lector de emplear el método que desee. Esos métodos son los siguientes: Eucledian, Maximum, Manhattan, Canberra, Binary, Minkowski, Pearson, Spearman y Kendall.

```
similarity <- function(df, player, selectLeague, fill_variable, fill_strip,
                       input = c("Eucledian", "Maximum", "Manhattan", "Canberra", "Binary",
                                 "Minkowski", "Pearson", "Spearman", "Kendall")
){
```

ILUSTRACIÓN 17. FUNCION SIMILARIDAD

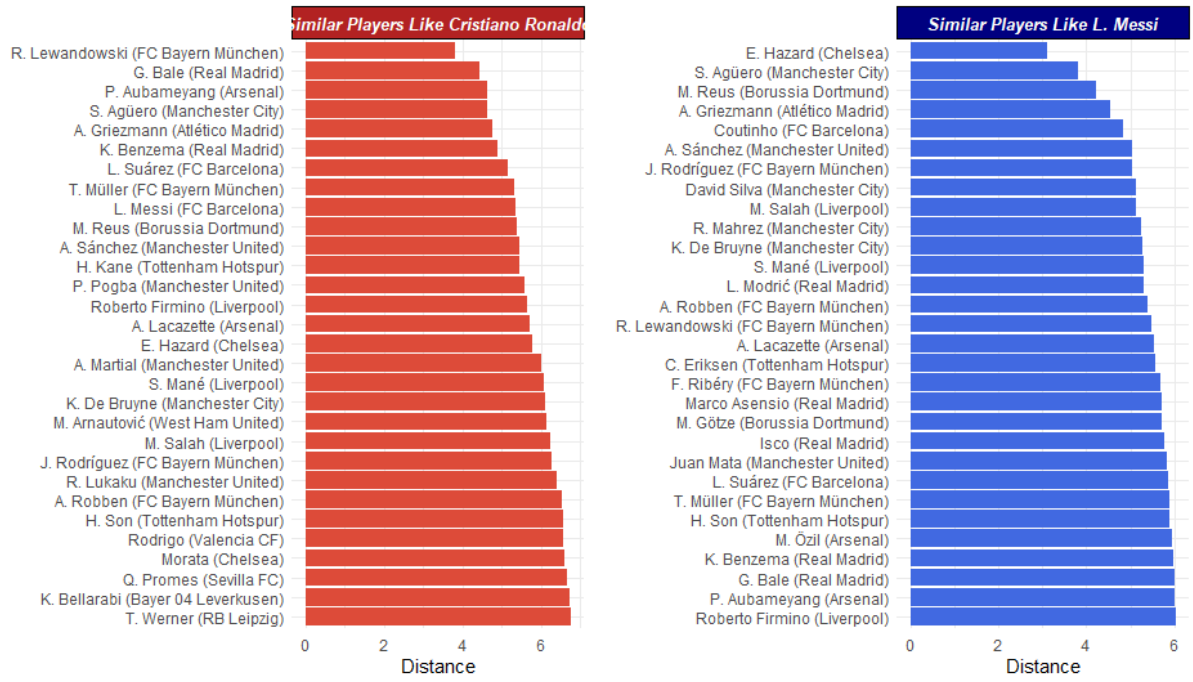
El método que se ha considerado más sólido ha sido la distancia euclidiana, las principales desventajas de usar este método es que es no robusta bien los outliers ni la diferencia de escala en las variables, sin embargo, en esta base de datos las puntaciones están perfectamente escaladas evitando el riesgo de outliers.

Como se ha hablado en el anterior apartado, Messi y Ronaldo son los dos jugadores más dominantes en el futbol actual, por lo tanto, se han considerado ambos para estudiar que jugadores podemos encontrar con las características más parecidas posibles.

```
similarity(df, input = "Euclidean", selectLeague = c("Bundesliga", "Premier League", "La Liga"),
           player = "Cristiano Ronaldo", fill_variable = "#dd4b39", fill_strip = "firebrick"),
similarity(df, input = "Euclidean", selectLeague = c("Bundesliga", "Premier League", "La Liga"),
           player = "L. Messi", fill_variable = "royalblue", fill_strip = "navy")
```

ILUSTRACIÓN 18. FUNCION SIMILARIDAD 2

Este ha sido el resultado:



En ambos jugadores tenemos que hay un jugador que sobresale como el jugador más similar. En el caso de Ronaldo ese jugador es el jugador del Real Madrid Gareth Bale, mientras que el de Messi es Eden Hazard propiedad del Chelsea.

Correlación

Para poder estudiar la correlación entre la finalización y la potencia de tiro hay que averiguar la normalidad de ambas variables. La hipótesis nula será que dicha variable sigue una distribución normal y no se podrá seguir adelante con la correlación.

```
        Shapiro-Wilk normality test

data:  kor$Finishing
W = 0.9751, p-value = 0.03053


        Shapiro-Wilk normality test

data:  kor$Shot.Power
W = 0.9425, p-value = 9.062e-05
```

ILUSTRACIÓN 19. RESULTADOS TEST DE NORMALIDAD

El resultado de ambos p-valores es inferior a 0,05 por lo que podemos afirmar que se rechaza la hipótesis nula, es decir ambas variables no son normales y podemos aplicar un método de correlación que no requieran normalidad como lo es Spearman.

```
        Spearman's rank correlation rho

data:  kor$Shot.Power and kor$Finishing
S = 64431, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7457925
```

ILUSTRACIÓN 20. PUNTUACIÓN DE CORRELACIÓN SPEARMAN

Una correlación de Spearman de 0,7457 indica una fuerte relación positiva entre ambas variables ya que es un valor cercano a 1.

Una vez tenemos estos resultados, se va a comprobar si afecta ser zurdo o diestro a la correlación anterior. Con el siguiente gráfico nos da un vistazo de la correlación coloreando de distinta manera según si pie dominante.

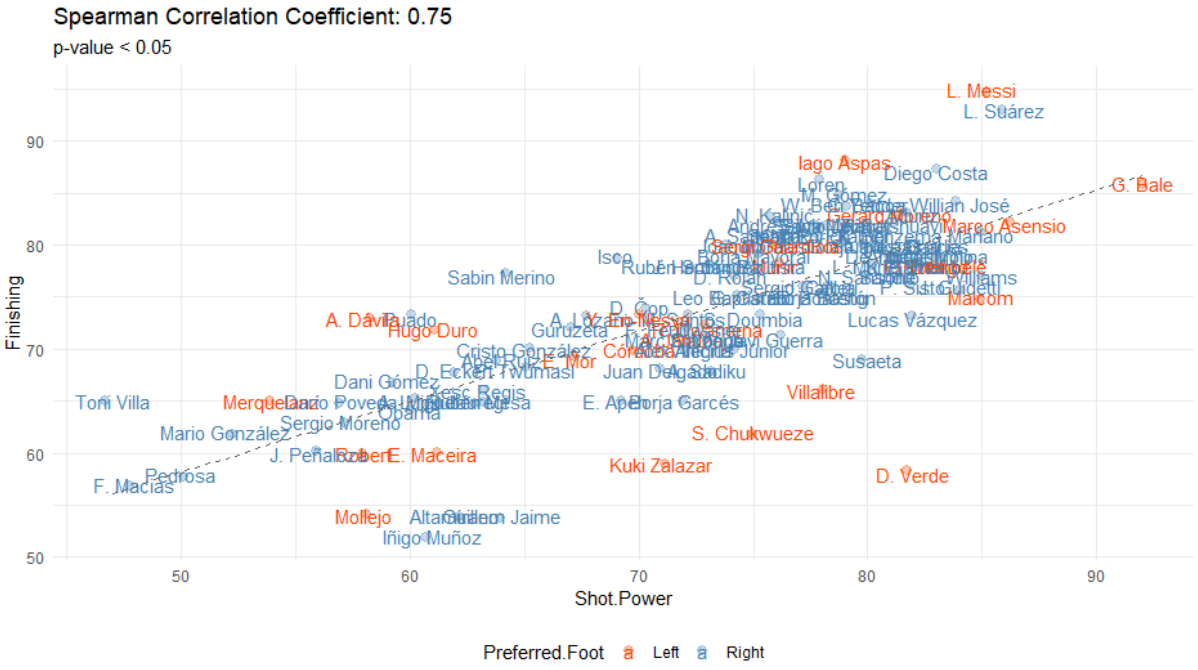


ILUSTRACIÓN 21. PLOT 6. CORRELACIÓN FINISHING-SHOT.POWER SEGUN PREFERRED FOOT

Se ha considerado el test de Wilcoxon adecuado para el comprobar si hay diferencias significativas entre zurdos o diestros a la hora de disparar a puerta (finalización y potencia).

```
        Wilcoxon rank sum test with continuity correction

data:  xt1 and xt2
W = 1160, p-value = 0.8149
alternative hypothesis: true location shift is not equal to 0


        Wilcoxon rank sum test with continuity correction

data:  yt1 and yt2
W = 974.5, p-value = 0.3086
alternative hypothesis: true location shift is not equal to 0
```

ILUSTRACIÓN 22. TEST DE WILCOXON

Ambos valores son superiores a 0,05 por lo que no hay evidencias significativas entre zurdos o diestros.

ACP

Lo primero va a ser preparar los datos, filtrando por “Forward” en primer lugar y seleccionando las variables que se consideren relevantes. Una vez tenemos los datos limpios, normalizamos las variables y establecemos los nombres de cada jugador como nombre de las filas.

```
pca_df <- df %>%
  filter(Class == "Forward") %>%
  select(Name, Position, Club, Class, Age, Overall, Potential, Values, Wages, Crossing:Sliding.Tackle)

pca_df2 <- pca_df %>%
  select(-Club, -Class, -Position, -Overall) %>%
  mutate(Name = paste0(1:nrow(pca_df), "-", Name)) %>%
  mutate_at(vars(Age:Sliding.Tackle), funs(scale)) %>%
  column_to_rownames("Name")
```

ILUSTRACIÓN 23. FUNCIÓN ACP

A continuación, gracias a la función “prcomp” realizamos el ACP, tras ello ajustamos los signos de las rotaciones y sus coordenadas de los componentes principales para hacer más sencilla la interpretación.

```
pca_df2 <- prcomp(pca_df2, scale = FALSE)

pca_df2$rotation <- -pca_df2$rotation
pca_df2$x <- -pca_df2$x
```

ILUSTRACIÓN 24. FUNCIÓN ACP 2

Se calcula la varianza explicada, resultando que los 6 primeros componentes explican el 78,1% de la varianza total. Usando esos 6 primeros componentes permite calcular PCA Score, dándole a cada jugador su valor correspondiente.

```
pl_avo <- pca_df2$sdev^2 / sum(pca_df2$sdev^2)
sum(pl_avo[1:6])

pca_df$`PCA Score` <- rowSums(pca_df2$x[,1:6] * t(matrix(pl_avo[1:6], 6, nrow(pca_df))))
```

```
> sum(pl_avo[1:6])
[1] 0.781055
```

ILUSTRACIÓN 25. CÁLCULO DE VARIANZA EXPLICADA Y SU RESULTADO

Finalmente, con el cálculo del PCA Score correspondiente a cada jugador obtenido, solo queda redondear los resultados, ordenarlos de mayor a menor y empleando dos gráficos de barras a cada lado (usando grid.arrange) mostrar los 10 valores más altos tanto de PCA Score como de Overall.

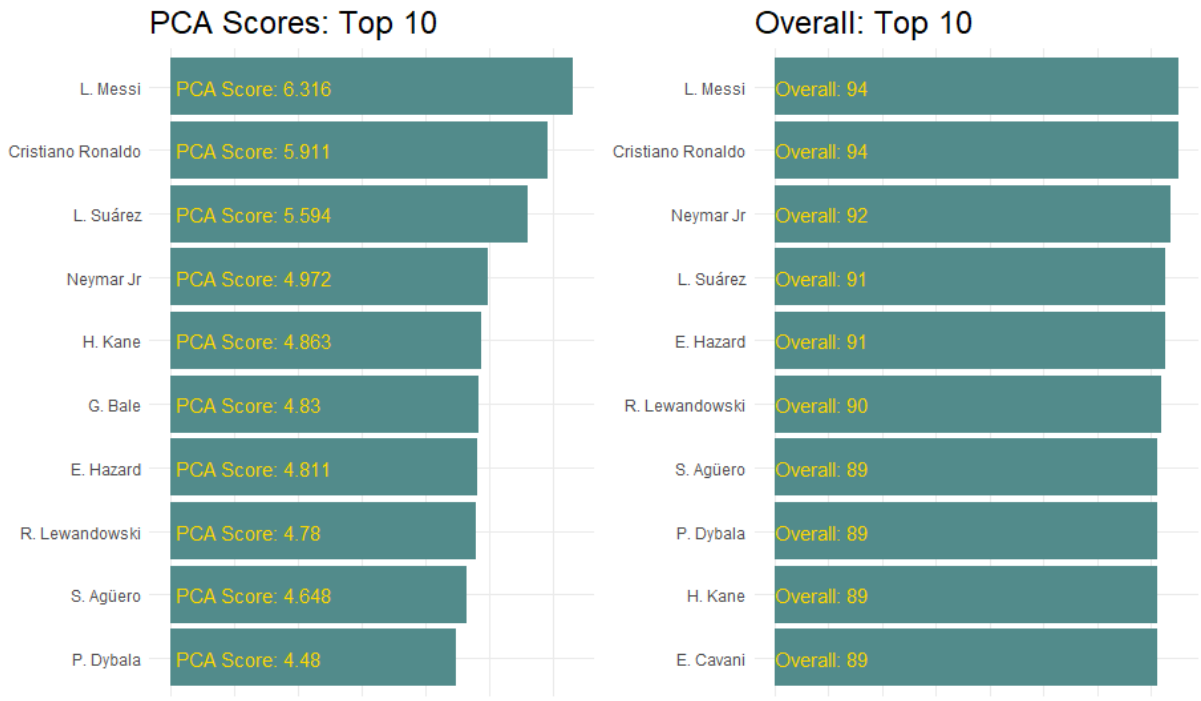


ILUSTRACIÓN 26. PLOT 7. GRÁFICO DE BARRAS PCA SCORES - OVERALL

Los resultados obtenidos son que según el PCA realizado los mejores delanteros en la base de datos se asemejan muchos a las valoraciones Overall. Mientras que para FIFA Cavani con una valoración de 89 puntos entra en el top 10, en el estudio de PCA realizado no aparece dicho top, en su lugar se encuentra Bale.

Conclusiones

Con la visualización de los datos, se ha podido ver la gran diferencia económica que hay entre las ligas del BIG 5 y el resto. Ofreciendo una mayor perspectiva respecto al enfoque que tienen estas ligas no tan potentes, algunas nutriéndose de jóvenes talentos para sacarles un buen rédito económico y otras, fichando a jugadores veteranos ofreciéndoles un último gran contrato, a cambio estos clubs reciben publicidad y visibilidad.

Respecto al objetivo principal del trabajo, podemos concluir que basándonos en esta base de datos y aplicando el análisis de componentes principales, los mejores tres mejores delanteros son Cristiano Ronaldo, Leo Messi y Luis Suarez. Los siete siguientes delanteros tiene una diferencia de "PCA Score" de apenas medio punto. Si lo comparamos con el parámetro "Overall" vemos como nueve de los diez mejores delanteros coinciden, estos parámetros que da EA Sports, pueden verse influidos por el marketing de los jugadores, sin embargo, se aprecia que tiene una gran coherencia la puntación final que se la da a los jugadores con sus atributos.

Con el uso de las distancias como indicadores de disimilaridad, se ha calculado una función que va a permitir al usuario encontrar los jugadores más parecidos al futbolista que deseen. Este tipo de herramientas se llevan empleando en diversos equipos, ya que por presupuesto u otras variables, no siempre los equipos van a poder contratar el jugador que quieren, con esto pueden plantearse fichar jugadores con características similares que se amolden a su idea de juego. Este trabajo ha puesto como ejemplo a los dos mejores delanteros de la base de datos: Messi y Cristiano.

Basándonos en los datos de esta base, queda desmitificada la afirmación popular que atribuye a los futbolistas zurdos mejor disparo a puerta. Si bien es cierto que existe una fuerte correlación entre la variable finalización (Finishing) y potencia (Shot.Power), se concluye que el pie dominante no afecta a la correlación anterior.

Como se mencionó en la introducción de este trabajo, el apartado de estadísticos descriptivos es donde más se puede mejorar, como por ejemplo usando RShiny para dotar al lector la posibilidad de realizar las consultas que desee.

Bibliografía

[1] Historia del Tlachtli. https://www.britannica.com/sports/tlachtli

[2] Dimensiones del campo de juego. https://continuemosestudiando.abc.gob.ar/contenido/6-dimensiones-del-campo-de-juego/

[3] Formaciones de futbol: Estrategias de ataque. https://mas10.ar/2023/10/02/formaciones-de-futbol-estrategias-de-ataque/

[4]Qué es fuera de juego en fútbol y cuando se sanciona. https://elsuperhincha.com/cuando-es-fuera-de-juego-en-futbol/

[5]InStat-La plataforma de Scouting que tu equipo necesita .https://objetivoanalista.com/instat-plataforma-de-scouting/

[6]iq-soccer. https://statsbomb.com/es/que-hacemos/iq-soccer/

[7] Que es un diagrama de barras. https://www.plandemejora.com/que-es-grafica-de-barras/

[8]Grafica de radar. https://tudashboard.com/grafica-de-radar/

[9]Box-plot. https://byjus.com/maths/box-plot/

[10]Diagrama de dispersión. https://spcgroup.com.mx/diagrama-de-dispersion/

[11]Cuartil: Qué es. https://economipedia.com/definiciones/cuartil.html

[12]Introducción al análisis clúster. https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm

[13] Wikipedia. https://es.wikipedia.org/

[14] Análisis de componentes principales. https://rpubs.com/Cristina_Gil/PCA

[15] Que es el análisis de componentes principales y como reducir el tamaño de una base de datos. https://www.hiberus.com/crecemos-contigo/analisis-de-componentes-principales/

[16]Que es un análisis de correlación. Características y ejemplos .https://mexico.unir.net/noticias/economia/analisis-correlacion/

[17] Análisis de correlación. https://datatab.es/tutorial/correlation

[18] Que necesitas y como llevarlo a cabo. https://nextscenario.com/es/analisis-de-correlacion/

[19] Historia de R y RStudio. https://bookdown.org/jboscomendoza/r-principiantes4/un-poco-de-historia.html

[20]La apasionante historia del primer FIFA. https://www.vidaextra.com/deportes/la-apasionante-historia-del-primer-fifa-asi-nacio-la-leyenda