FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2024/2025

Trabajo de Fin de Grado

TÍTULO: SISTEMA DE RECOMENDACIÓN DE JUGADORES DE FUTBOL CON R

***Alumno*: Alejandro Naranjo Expósito**

***Tutor*: Fernando Pérez Contreras**

Junio de 2025

UNIVERSIDAD COMPLUTENSE
MADRID

Índice

Resumen.

Este trabajo consiste en realizar una machine learnig que sea capaz de recomendar jugadores de fútbol con características similares.

El fin de este trabajo es poder crear un modelo que pueda ayudar a las direcciones deportivas de equipos más modestos a mejorar sus opciones a la hora de fichar jugadores, ya sea para mejorar posiciones que necesiten, o para adelantarse a cubrir las bajas de sus jugadores en plantilla.

Para su estudio se ha utilizado una base de datos con estadísticas de los jugadores que han participado al menos un partido en cualquier club de las cinco grandes ligas europeas según el ranking de la UEFA (UEFA, 2025), las cuales son:

- LaLiga.
- Premier League.
- Bundesliga.
- Serie A.
- Ligue 1.

Los datos han sido tomados de la web de estadísticas de fútbol fbref (Fbref, 2025) a través del uso de la biblioteca worldfootballR (JaseZiv, 2020) en RStudio.

Para su realización se han usado distintas técnicas de depuración de datos, análisis cluster, modelos random forests y medidas de distancias euclídeas. Estás técnicas han sido empleadas en R, debido a que es un sistema de libre acceso, y que tiene un interfaz muy fácil para su entendimiento lo que podría ayudar a su implementación en clubes con menos margen presupuestario.

Abstract

This project involves the development of a machine learning model capable of recommending football players with similar statistical profiles. The main goal is to create a tool that can support the scouting departments of local or regional clubs in improving their recruitment strategies- either by identifying suitable players to strengthen specific positions or by anticipating potential departures.

To achieve this, a dataset was compiled containing statistics of players who have participated in at least one match in any club within the top European leagues, as ranked by UEFA (UEFA, 2025). These leagues are:

- LaLiga.
- Premier League.
- Bundesliga.
- Serie A.
- Ligue 1.

The data was sourced from the football statistics website fbref (Fbref, 2025) using the worldfootballR package (JaseZiv, 2020) in RStudio. The methodology involved data cleaning techniques, cluster analysis, random forest models, and Euclidean distance metrics. These methods were applied using R, an open-source programming environment with a user-friendly interface, making it suitable for implementation by clubs with limited budgets.

Introducción.

La innovación y el deporte siempre han ido de la mano, gran parte de las mejoras que se han realizado en los últimos años viene relacionadas en términos de salud, rendimiento físico, control de situaciones financieras o búsqueda de nuevos activos se han visto desarrolladas por departamentos de investigación de clubes deportivos. Unos de los elementos que más se ha empezado a ver es el uso del análisis de datos para la mejora del rendimiento de los jugadores en el campo o para tomar decisiones de que perfiles de jugadores fichar.

En esto último el mundo del fútbol ha sufrido unas grandes mejoras y se ha visto que ciertos modelos de equipos de la elite europea han utilizado técnicas de análisis de datos para reforzar sus plantillas, sobre todo empezamos a ver estos cambios a partir de salida a cartelera de Moneyball película que habla de cómo Billy Beane, general manager de los Oakland Athletics equipo de la MLB usa el análisis de datos para fichar jugadores y conseguir ganar el título de liga, este modelo implementado en la ficción se empezó a ejecutar en algunos equipos de Europa como el Brentford o el Liverpool desde la llegada de Jürgen Klopp (OneFootball, 2022) en Inglaterra y en España tenemos el caso del CD Leganés desde la llegada de Blue Crow (Martín, 2022) que incluso no solo usan la información del Big Data para tomar decisiones corporativas sobre que entrenadores o directores deportivos deben poner al cargo, todos estos equipos a día de hoy han conseguido resultados positivos como pueden ser los ascensos a las primeras divisiones nacionales del Brentford o el Leganés y en el caso del Liverpool este modelo le llevo a ser uno de los dominadores del fútbol europeo consiguiendo una Premier League y una Champions League.

A parte de departamentos integrados en las estructuras de ciertos equipos existen ciertas páginas web y ciertos programas que pueden ayudar a realizar estas funciones como pueden ser las plataformas de Wyscout (Wyscout, 2025) u OptaAnalyst (OptaAnalyst, 2025) . Estás programas utilizan las estadísticas para realizar informes completos sobre los jugadores de fútbol que contienen en sus bases de datos, pero el gran problema que existe es que estas páginas web solo se pueden usar como opciones de pagó y muchos equipos con presupuestos más bajos no pueden permitirse pagar estas suscripciones mensuales y por lo tanto no se pueden ver beneficiados del uso de los datos como recurso para apoyar a sus departamentos de dirección deportivo.

Objetivos.

El objetivo general de este TFG es desarrollar un sistema de recomendación de jugadores de fútbol mediante el uso de técnicas de machine learning, las cuales han sido entrenadas con datos estadísticos de futbolistas de las cinco grandes ligas europeas (LaLiga, Premier League, Serie A, Bundesliga y Ligue 1) (UEFA, 2025). Este sistema tendrá como objetivo focalizarse en la identificación de jugadores con perfiles de rendimiento similares en términos estadísticos. El fin de todo será el apoyar a clubes con presupuestos limitados o sin departamento de análisis a poder sustituir jugadores clave y traer refuerzos en posiciones especificas dentro de sus sistemas tácticos.

Como objetivos más específicos tenemos los siguientes elementos:

- Realizar una clasificación funcional de los jugadores mediante la aplicación de técnicas de clustering, encontrando los grupos en los que se agrupan los futbolistas con rendimientos estadísticos parecidos.

- Poder implementar un algoritmo de recomendación, el cuál reciba un jugador de referencia, y luego pueda devolver otros futbolistas con perfiles similares al jugador recomendado.
- Demostrar la utilidad práctica del modelo utilizado en contextos reales, usando caso de sustitución de futbolistas clave y su uso para el scouting de jugadores.

Metodología

Para la explicación de la metodología que se ha llevado a cabo durante el trabajo vamos a dividir este apartado en dos partes una parte teórica en la cual se verán reflejados los métodos que hemos utilizado desde un marco teórico y explicativo que sirva para comprender su uso. Y luego se realizará una explicación del porqué del uso de esa metodología y de cómo se ha usado.

Marco teórico

En este marco teórico explicaremos el análisis clúster, los modelos random forest y las medidas de distancias euclídeas.

- Análisis Clúster.

  El análisis clúster es una técnica estadística multivariante que se utiliza para agrupar una muestra en distintos grupos. Estos grupos deben contener unas observaciones homogéneas entre sí, pero heterogéneas respecto al resto de grupos, en consecuencia, de las características que presentan. A diferencia de otros métodos supervisados, como la regresión logística o los árboles de decisión, el análisis clúster es una técnica no supervisada. Esto lo convierte en una técnica que resulta especialmente útil cuando se

requiere clasificar individuos sin tener información previa sobre la pertenencia a un grupo, permitiendo identificar patrones o características comunes dentro de cada uno. Dentro del análisis clúster utilizaremos la técnica de clúster jerárquico con un enfoque aglomerativo o ascendente. Este enfoque fusiona de forma repetitiva cada clúster que se va generando hasta formar un único clúster.

Para vincular existen distintos métodos. En el caso de este trabajo se va a usar el método Ward, este método fue introducido por Joe H. Ward en la década de 1960 en el método en principio se conoció como el método del aumento mínimo de la suma de cuadrados (MISSQ). El cuál consistía en que cada punto de datos comienza con su propio clúster, lo que significa que la suma de cuadrados entre los puntos de los datos comienza en cero y luego aumenta a medida que vamos fusionando los clústeres entre sí.

El método Ward se caracteriza en minimizar la suma de las distancias al cuadrado de los puntos desde los centros de los clústeres a medida que estos se fusionan.

Por lo tanto, es una buena opción a la hora de usar variables cuantitativas, y se ve menos afectado por el ruido de los valores atípicos. (Noble, 2024). El criterio de Ward equivale a considerar la distancia euclídea al cuadrado entre centroides, ajustada por el tamaño de los clústeres como medida de disimilitud. De hecho, se puede demostrar que este método es equivalente a realizar un algoritmo jerárquico cuyo objetivo final aproxima al del algoritmo de k-means, ya que ambos minimizan la variabilidad interna de los grupos en cierta media.

Su utilidad en el contexto del trabajo es permitir agrupar jugadores con perfiles similares gracias a que mide la similitud de jugadores a través de sus características y aplicar el

criterio de mínima varianza, permite descubrir conglomerados naturales de jugadores que comparten patrones comunes. La creación de estos grupos homogéneos sirve luego para poder recomendar jugadores comparables entre sí. Además, a diferencia de métodos de clustering no jerárquicos, el enfoque jerárquico obtiene un dendograma multinivel que puede ser recortado para adaptarlo de la mejor manera al uso que se necesita darle.

En resumen, el análisis de clúster jerárquico con el método de Ward proporciona una herramienta de agrupamiento de perfiles similares de una manera eficiente, que garantiza que cada unión de grupos esté estadísticamente justificada por mínimos incrementos de heterogeneidad. Esto resulta ventajoso en aplicaciones como la del trabajo, donde se busca descubrir segmentos comparables de forma objetiva, maximizando la homogeneidad dentro de cada segmento.

- Análisis de componentes principales (PCA).

  El Análisis de Componentes Principales es una técnica que es utilizada para reducir las dimensiones de un conjunto de datos con múltiples variables, conservando el mayor porcentaje posible de varianza explicada. Con esta técnica se transforma las variables que se tienen en origen por unas nuevas variables que no son correlacionadas, a las cuales se les denomina componentes principales.

  Formalmente, el análisis de componentes principales encuentra una nueva base de ejes ortogonales en el espacio de datos, tales que cada eje coincide con la dirección de máxima varianza de los datos proyectados. Los ejes vienen dados por lo autovectores de la matriz de covarianza de los datos, y la varianza es explicada por cada componente

correspondiente a lo autovalores asociados. El primer componente principal corresponde a la combinación lineal normalizada de las variables originales con mayor varianza explicada, el segundo componente corresponde a la combinación lineal ortogonal del primer componente principal que explica la mayor varianza restante, y así sucesivamente. Esta técnica ofrece diferentes ventajas como son la reducción de la complejidad del modelo, eliminar el ruido y la multicolinealidad. El PCA suele usarse antes de hacer sistemas de clustering. En el contexto del trabajo no es necesario ya que interesa separar a los grupos con todas las variables presentes para poder separar mejor los clústeres, aunque si eso uso el PCA para poder visualizar el conjunto de los clústeres.

- Modelos Random Forest.

  Los Random Forest son modelos de aprendizaje supervisado basados en la combinación de múltiples arboles de decisión. Consiste en modelo de ensamble que fue introducido por Leo Breiman y Adele Cutler a finales de los años 90, y publicado en 2001 uno como un artículo denominado Random Forest para la revista machine learning (Interactive Chaos , 2025). La idea principal es construir una colección de árboles de decisión que trabajen de forma coordinada: cada uno de los árboles aprende a clasificar o predecir de manera ligeramente distinta, y luego sus resultados se agregan para obtener una predicción conjunta más robusta.

  Para comprender la construcción del modelo se dará una explicación teórica de su construcción.

  Estos modelos de Random Forest están compuestos por N árboles de decisión T_1, T_2, …, T_N entrenados de manera ligeramente diferente. La construcción de cada árbol incorpora

dos fuentes principales de aleatoriedad, las cuales son Muestreo aleatorio de instancias (Bootstrap) y la selección aleatoria de atributos en cada división.

El Bootstrap consiste en para cada árbol $T_{i,}$ seleccionar una muestra aleatoria con reemplazo del conjunto de entrenamiento original que es conocido como bagging, creando un subconjunto de datos con un tamaño igual al conjunto de datos original seleccionando de forma aleatoria observaciones con repetición. En media, cada árbol observará aproximadamente un 63% de las instancias únicas, y el resto se verán como out-of-bag para validación interna, gracias a este muestreo se produce diversidad entre los árboles.

El otro método de muestreo que se utiliza es la selección aleatoria de atributos en cada división, esto surge cuando un árbol está creciendo, en cada nada de decisión no se consideran todas las variables predictoras para encontrar la mejor división. En su lugar,el algoritmo Random Forest escoge aleatoriamente un subconjunto m variables de las p disponibles, y solo entre esas ponen los cortes posibles (Interactive Chaos , 2025). Esta técnica fue introducida en 1995 por Tin Kan Ho y adoptada por Breiman, para asegurar que los árboles individuales se diferenciaran no solo por los datos que ven sino también por las características que consideran, consiguiendo que se redujese la correlación entre árboles.

Una vez explicada la aleatoriedad introducida en cada árbol de decisión, se tiene en cuenta que cada árbol se expande completamente hasta que ya no pueda seguir dividiéndose o alcance un criterio de parada propuesto, una vez el árbol ha alcanzado este punto se juntan todos los árboles en el bosque y se decide el valor predicho comprobando

la predicción de cada uno de los árboles y escogiendo como predicción final la predicción que más ‘votos’ haya recibido entre el cómputo global de los árboles. En los problemas de predicción de variables numéricas cada árbol produce un valor numérico y la predicción final será el promedio de las predicciones de todos los árboles, con esto se consigue que si la mayoría de los árboles coinciden mejora la probabilidad de que salga una buena predicción y los valores atípicos tienden a cancelarse. Esto hace que los Random Forest tengan una serie de ventajas y que han hecho que su uso se vuelva extremadamente popular en estadística y machine learning. Estás propiedas son:

- **Alta precisión predictiva y menor sobreajuste**
- **Robustez frente a datos rigurosos y outliers.**
- **Manejo eficaz de datos de alta dimensión y características heterogéneas**
- **Versatilidad e información adicional**

Por todo esto los Random Forest representan un balance entre simplicidad, precisión y robustez. Añadido a su capacidad de manejar datos ruidosos y complexos, junto a su alto desempeño predictivo, los convierte en una elección acertada para su uso como predictora en distintos problemas como el que se plantea en este trabajo para encontrar similitudes entre los jugadores o encontrar la posición en la que mejor se acomodan en el campo.

- Medidas de distancias Euclídeas.

  En una gran cantidad de algoritmos de aprendizaje automático y análisis de datos, especialmente en técnicas de agrupamiento y recomendación basada en vecinos más cercanos como es el sistema de recomendación que se quiere implementar, es

fundamental que exista una métrica que cuantifique la diferencia o similitud entre dos o varios objetos. La distancia euclídea es la mediad de disimilitud más común ese es el hecho por el cual se decidió usarla para medir las diferencias entre los jugadores que pertenecían a un mismo clúster en nuestro universo. Formalmente, la distancia euclídea entre dos puntos se define como la raíz cuadrada del sumatorio de las diferencias entre las variables de los dos puntos al cuadrado, lo que se define matemáticamente en la siguiente formula.

$$d_E(P,Q) = \sqrt{\sum_{i=0}^{n}(p_i - q_i)^2}$$

La utilización de la distancia euclídea es muy común en algoritmos de clustering como es el clustering jerárquico empleado en el trabajo y también se suele usar en sistemas de recomendación basados en cercanía de perfiles que es el uso principal que se le da en este trabajo.

A pesar de su simplicidad e interpretabilidad, el uso de esta distancia tiene una limitación importante y es que depende de la escala de las variables consideradas, por eso en el trabajo se normalizan los datos para poder escalar todas las variables en un mismo rango. Este viene dado por que suma diferencias cuadráticas en cada dimensión, una variable con un rango de variación amplio puede llegar a dominar la métrica de distancia, eclipsando el efecto de otras variables de rango más pequeño. Por ejemplo, en este trabajo variables como partidos o minutos jugados pueden tener un rango muy amplio y desestabilizaría las similitudes entre los jugadores, o variables como los pases que dependiendo de estilos de juegos o posiciones pueden oscilar en cientos de datos de diferencias puede hacer que la diferencia estadística entre distintos jugadores sea mucho

más amplia que la realidad. Además, este problema se puede ver agravado por variables heterogéneas como pueden ser partidos jugados, porcentaje de tiros o kilómetros recorridos por un jugador donde los números no pueden ser ni comparables, ya que usan distintas métricas.

Como se ha comentado con anterioridad la solución clásica a este problema es la normalización de las variables en N(0,1) para que los datos se encuentren dentro de intervalos de [0,1]o [-1,1]. Con esta técnica se busca tipificar los datos y que todas las unidades queden expresadas en unidades que puedan ser comparadas entre sí.

En conclusión, la distancia euclídea es una herramienta fundamental para medir similitudes en espacios multidimensionales y resulta intuitiva para su interpretación geométrica. No obstante, su aplicación directa requiere asegurar la comparabilidad de las escalas de los datos mediante técnicas de normalización, garantizando que el análisis de proximidad releje verdaderamente las diferencias de perfil entre objetos y no artefactos de unidades de medida o dispersión distinta de las variables involucradas. Esto asegura que, por ejemplo, en el objetivo principal del trabajo que es la recomendación de jugadores, la métrica de distancia capture diferencias de rendimiento o estilo de juego, y no se vea sesgada por el hecho de que cierta característica numérica tenga valores más grandes que otra.

Metodología.

- Recolección y depuración de los datos.

Para la posible investigación de este trabajo se utilizan datos extraídos de la librería worldfootballR (JaseZiv, 2020), la cual permite la recolección de información al detalle y más actualizada en lo que se refiere a los jugadores y equipos de fútbol, mediante el uso de la base de datos de FBref.com. (Fbref, 2025) Se recogieron datos pertenecientes a la temporada 2024/2025 de las cinco grandes ligas según el ranking de federaciones de la UEFA (UEFA, 2025), el cual dictamina que las 5 federaciones más grandes son Inglaterra, Italia, España, Alemania y Francia. Lo que dictamina que la primera división de cada una de las federaciones mencionadas con anterioridad conforma las cinco grandes ligas europeas.

De esta base de datos mencionada con anterioridad se extrajeron distintos tipos de estadísticas avanzadas, las cuales se agrupan en las siguientes categorías:

- Estadísticas estándar(standard)
- Disparo (shooting)
- Pases (passing)
- Tipos de pase (passing_types)
- Creación de ocasiones (gca)
- Defensa (defense)
- Posesión (possession)
- Tiempo de juego (playing_time)
- Misceláneo (misc)

Estas estadísticas posibilitan tener una visión global del rendimiento de los distintos jugadores en las diferentes facetas del juego.

Una vez se han recopilado estos datos, se realizó un proceso de limpieza y transformación de los datos, en el que se eliminaron variables que resultaban redundantes como URLs, identificadores de liga, país de nacimiento y otras variables que no aportaban información al trabajo.

Luego se procedió con un filtro para obtener únicamente a los jugadores que han jugado durante la temporada a estudiar, ya que en la base de datos se recogía información de todos los jugadores convocados durante la temporada.

Una vez limpiada la base de datos de variables redundantes y quedándose solo los jugadores que habían jugado al menos un partido a lo largo de la temporada se decide imputar los valores faltantes de las variables numéricas, para ello se utilizó la media de cada variable, mientras que las variables categóricas se complementaron con la moda.

Una vez imputada la base de datos, se agrupa a los futbolistas dependiendo su posición en el campo en defensas ‘DF’, mediocentros ‘MF’ y atacantes ‘FW’, creando la variable ‘Pos_Simplificada’ y además se decide descartar a los jugadores que ocupaban la posición de portero ‘GK’ debido a que su perfil estadístico es diferente al de los jugadores de campo, y además se generó un identificador para cada jugador, combinado su nombre y su equipo, para así poder tener en cuenta por separado las estadísticas de los jugadores que han jugado partidos en dos equipos diferentes durante la presente temporada.

Para finalizar, se decide normalizar todas las variables numéricas mediante un escalado estándar (media cero y desviación típica 1) para que todas las variables tengan la misma influencia durante el modelado.

- Análisis Clúster.

  Una vez han sido preparados los datos, se realizó un análisis clustering para obtener los grupos naturales de jugadores con estilos de juegos parecidos. Para ello se empleó un algoritmo de clústering jerárquico aglomerativo sobre los jugadores utilizando como medida de semejanza la distancia euclídea en el espacio de variables estandarizadas. Para su agrupamiento se usó el método Ward el cual minimiza la suma de las distancias de los puntos al centro de los clústeres a medida que estás se van agrupando hasta formar un único conglomerado con todos los jugadores, lo que genera una jerarquía que es representada por un dendrograma. Para obtener el número de clúster optimo (k), se han utilizado dos criterios:

  - **Método del codo:** Esté método evalúa la inercia intra-clúster o la suma de distancias al centroide dentro de cada uno de los clústeres para diferentes particiones (k=2, 3, 4, …, 15). En el gráfico del codo, se obtiene el valor de K en el que si se añade un clúster adicional se produce una reducción marginal cada vez menor en la variable intragrupo (Dataiku., 2025). Esté punto de inflexión informa del número apropiado de grupos, ya que representa un equilibrio entre la compacidad y la granularidad de los clústeres.

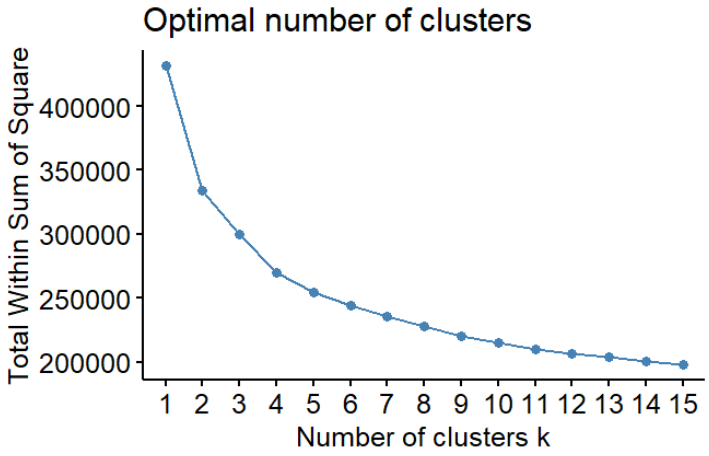

Figura 1

- **Índice de silueta:** Además se calculó el índice de silueta para diferentes configuraciones de k. Con este índice se cuantifica cuanto de separadas se encuentra cada una de las agrupaciones entre sí, combinando la proximidad dentro del mismo clúster y la lejanía con respecto al resto de clústeres. Un valor medio de este índice alto refleja clústeres mejor separados y más densos.

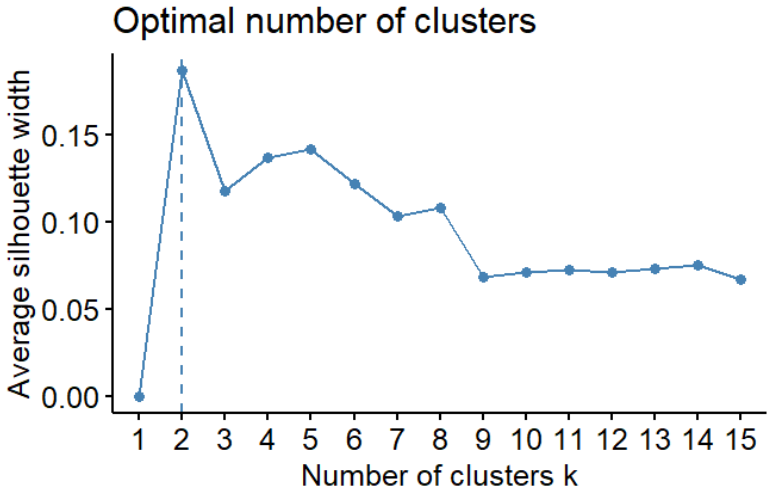

Figura 2

Observando las Figuras 1 y 2 se determina el número de clústeres que mejor se adapta al conjunto de jugadores que disponemos. Teniendo esto en cuenta, se determinó que el número de óptimo de clústeres sería k=5 y se decide por el método del codo ya que el valor optimo del índice de silueta que es k=2 se determina como un valor muy pequeño y que no capturaría de forma correcta los diferentes grupos de futbolistas. Una vez determinado el número de clústeres se procedió a cortar el dendrograma resultante del clúster jerárquico para obtener los 5 clústeres finales, asignando a cada uno un ID del 1 al 5 según al clúster al que pertenece cada jugador. Los perfiles de los jugadores de cada clúster fueron luego analizados y serán mostrados en los resultados.

- Análisis de componentes principales (PCA).

  En el contexto del estudio se ha hecho uso del análisis de componentes principales para la representación en un espacio bidimensional de los jugadores de fútbol facilitando así su visualización para la agrupación de los clústeres como se ve reflejado en la siguiente Figura.

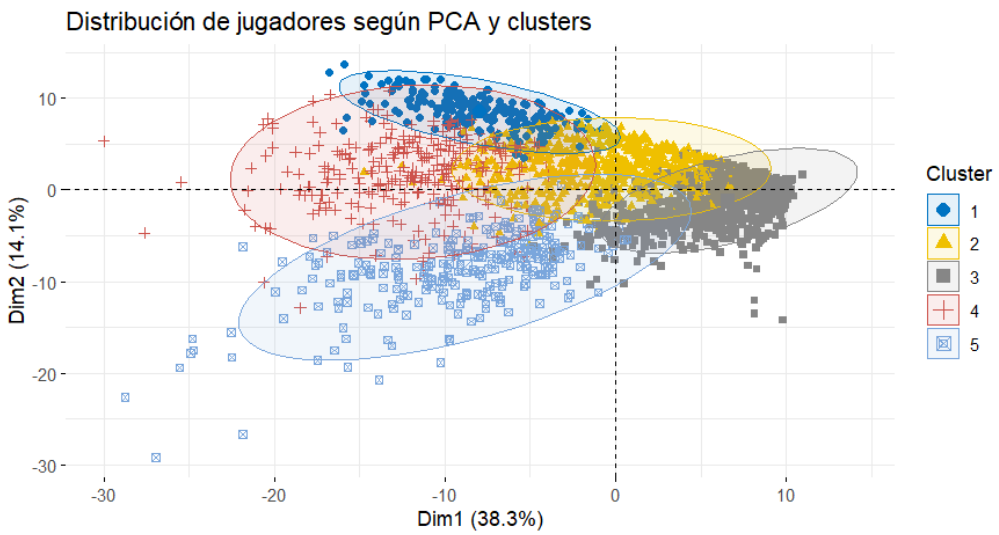

Figura 3

Como se muestra en la Figura 1, gracias al PCA se pueden proyectar al conjunto de jugadores en dos dimesiones (Dim1 y Dim2), ayudando a la visualización clara de la distribución de los clusters generados. La inclusión de elipses de densidad en la representación posibilita observar la cohesión dentro de los distintos grupos y el grado de solapamiento presente entre ellos, lo que aporta una validación visual a los resultados del agrupamiento.

- Modelos Random Forest.

  Con el fin de complementar el análisis clúster y poder predecir con exactitud el grupo al que pertenece cada jugador y la posición a la que mejor se adapta se han implementado dos modelos de clasificación supervisada empleando el algoritmo de Random Forest. El cuál consiste en ensamblar múltiples árboles de decisión entrenados con distintas muestras del conjunto de datos, y que realiza la predicción final por votación, ofreciendo

una buena capacidad predictiva y medidas de importancia de variables. Ahora se comenzará con la explicación de los dos modelos implementados en este trabajo:

- **Clasificación por posición:** En primera instancia se entrenó un modelo capaz de predecir la posición en el campo que ocupaba cada jugador. Las posiciones como se comentaron con anterioridad son las siguientes (DF, MF o FW), a partir de las métricas de rendimiento del jugador. El objetivo de este modelo era comprobar según las estadísticas como se clasificaba cada perfil de jugador y poder también utilizarlo para comprobar la distribución de los jugadores en cada uno de los clústeres. Para la validación del modelo se partió el conjunto de datos en una muestra de entrenamiento y en otra muestra de prueba, para que posteriormente en los resultados se pueda analizar la validez del modelo a través del uso de la matriz de confusión.
- **Clasificación por clúster:** Por otro lado, se realizó un modelo de aprendizaje supervisado para comprobar la clasificación en los clústeres de los distintos jugadores, y que luego se implementará para determinar el clúster al cual pertenece un jugador o un perfil, y poder sacar sus similitudes, como en el caso anterior también se realizó una partición de los datos en entrenamiento y prueba que será analizada en los resultados posteriores

- Sistema de recomendación del jugador

  Para concluir con los métodos utilizados en el trabajo, se explicará el funcionamiento detallado del sistema de recomendación de jugadores, el cual incluye distintos puntos a tratar:

- **Selección del jugador:** En primer lugar se realizara un procedimiento para la selección del jugador, el cuál será pasado por formato texto de la siguiente manera (Nombre Apellido-Equipo), y del que se cogerán las estadísticas pertenecientes a la base de datos, en caso de no existir dentro de la base de datos se debería introducir las estadísticas a analizar y formar un dataset con los datos a estudiar, una vez se escogen los datos, se predice el clúster al cual pertenece el jugador y se determina su posición en el campo para el posterior análisis deportivo. Una vez seleccionado el jugador y determinado el clúster al cual pertenece se pasa al siguiente punto.
- **Cálculo de las distancias en el espacio de características:** Dentro del clúster al que pertenece el jugador seleccionado, el sistema calcula la distancia euclídea entre el jugador seleccionado y el resto de los jugadores pertenecientes a dicho clúster. Debido al de que los datos están normalizados, la distancia euclídea proporciona una medida equilibrada de similitud global en el perfil estadístico. Lo que implica que cuanto menor sea la distancia de un jugador con respecto al jugador objetivo más similares van a ser entre sí.
- **Ranking de jugadores similares:** Una vez calculadas las distancias euclídeas, se genera un ranking entre los 5 jugadores con menor distancia euclídea con respecto al jugador de referencia, dicho resultado implicaría que los jugadores pertenecientes a este ranking serían los que tienen un perfil estadístico más similar al jugador elegido, y por lo tanto serían los jugadores que se podrían usar para suplirle en caso de necesitarlo.

- **Presentación de resultado y visualizaciones:**

  Para la presentación de los datos se usaron, gráficos capaces de mostrar las similitudes entre los jugadores recomendados en el ranking y el jugador de referencia, de esta forma visual el usuario tiene más fácil la visualización de los jugadores pertenecientes al ranking.

Para el análisis de toda la metodología descrita con anterioridad se codificó un algoritmo en R para su posible aplicación práctica en la gestión deportiva de los clubes, y comprobar el funcionamiento en un entorno real que pueda ser comparado con el objetivo general del trabajo.

Resultados.

Una vez explicada la metodología que se ha llevado a cabo, se presentara un análisis de los resultados que han salido en R al utilizar los métodos de estadísticos propuestos, se comprobará si el modelo es eficiente para su uso con jugadores que no estén recogidos en la base de datos y con jugadores que estén recogidos también.

Para comenzar empezaremos a analizar la distribución de los distintos clústeres y que perfiles de futbolistas los conforman.

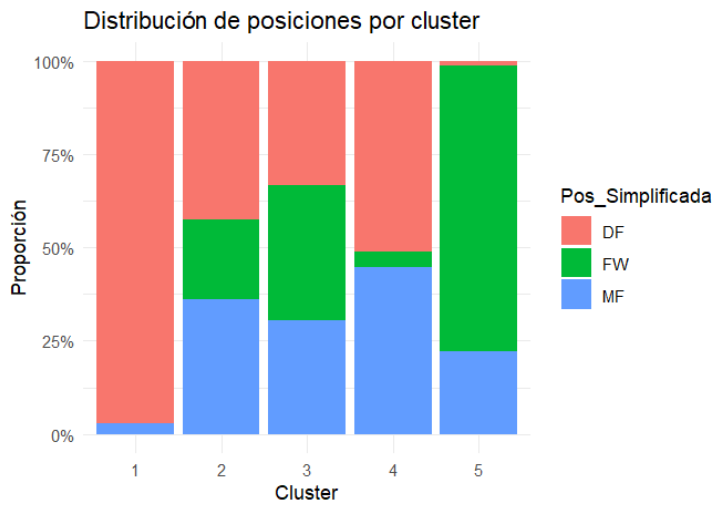

Figura 4

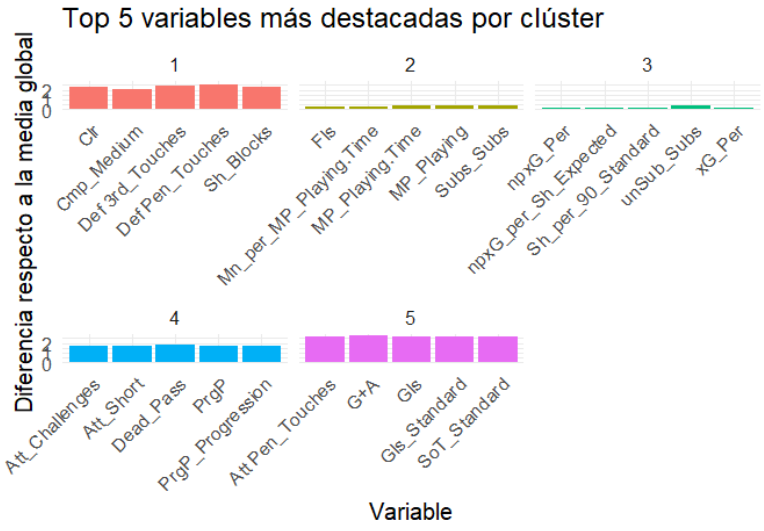

Figura 5

Observando las Figuras 4 y 5 se puede determinar el perfil de los jugadores que compondrán los clústeres, los cuáles serán explicados con mayor detalle a continuación.

- **Clúster 1:**

En esté clúster percibimos que se agrupan jugadores con características principalmente defensivas, ya sea por su posición en el campo, ya que la mayoría de los jugadores que se han agrupado aquí pertenecen a la posición de DF. Como por la variable que más destacan en el perfil de los jugadores que los componen, que como podemos observar en la Figura 5 corresponde a Clr (despejes defensivos realizados), Cmp_Medium (Pases de distancia media completados), Def 3rd_Touches (Toques de balón en el tercio defensivo del campo), Def Pen_Touches (Toques de balón dentro del área de penalti propia) y Sh_Blocks (Tiros bloqueados). Por lo tanto, se concluye que el perfil principal de este clúster serán jugadores que ocupen la posición de defensa central, o centrocampistas defensivos en equipos que defiendan en un bloque bajo.

- **Clúster 2:**

  En este clúster se observa que existe una amplia variedad de perfiles posicionales, por lo que la Figura 4 no es indicativa del perfil de jugadores que se encuentran incluidos, sin embargo, analizando la Figura 5 se observo que las variables principales eran Fls (Faltas Cometidas), Mn_per_MP_Playing_Time (Minutos por partido jugado)**,** MP_Playing_Time (Minutos jugados), MP_Playing (Partidos jugados) y Subs_Subs (Ingresos al campo como suplente), lo que nos lleva a determinar que son jugadores que suelen salir como suplentes y que por lo tanto no suelen jugar demasiados minutos, por lo tanto son jugadores de rol, este clúster puede ser explorado por equipos con bajo presupuesto para encontrar jugadores que no estén cómodos en sus equipos actuales y puedan querer cambiar de equipo.

- **Clúster 3:**

En este clúster se observa que existe una amplia variedad de perfiles posicionales, aunque aquí si se puede ver que predominan los jugadores que ocupan posiciones ofensivas, por lo tanto, la Figura 4 como en el ejemplo anterior no es del todo indicativa del perfil de jugadores, sin embargo, analizando la Figura 5 se observó que las variables principales son npxG_Per(), npxG_per_Sh_Expected(), Sh_per_90_Standard, unSub_Subs() y xG_Per(), analizando estás variables se llegó a la conclusión que en esté clúster encontramos jugadores de carácter defensivo que han sido ineficaces y con pocas contribuciones pero que han sumado una cantidad de minutos.

- **Clúster 4:**

  Analizando la información extraída de las Figuras 4 y 5 se puede determinar que las posiciones de los jugadores pertenecientes a este clúster son MC y DF, y observando la Figura 5 se encuentra que las variables que más destacan en este clúster son Att_Challenges (), Att_Short (), Dead_Pass(), PrgP() y PrgP_Progression(), con estó se determina que el perfil de los jugadores que se encuentran en este grupos son jugadores con grandes capacidades creativas, y que predominan en la construcción del juego, lo que indica que las posiciones que se cubren aquí son laterales ofensivos, mediocentros, centrales en equipos con mucho posesión y con un bloque, y atacantes que contribuyan mucho en la construcción del juego.

- **Clúster 5:**

  Por último, se estudiaron los perfiles que se encuentran dentro de clúster 5, en el cuál encontramos a jugadores con un perfil más ofensivo, observando que las posiciones predominantes son la de jugadores que se sitúan como FW y como MC. Además,

observando las variables que más destacan las cuales son Att Pen_Touches (), G+A (), Gls(), Gls_Standard() y SoT_Standard(), se puede determinar que los jugadores que corresponden a esté clúster son atacantes con una gran contribución de cara al gol, lo que se suelen conocer comúnmente como ‘Killers’, por lo tanto se encontraran los jugadores que pelearan por títulos individuales como la Bota de Oro o por estar entre los máximos goleadores o asistentes de sus respectivas ligas, aquí también podremos encontrar a una gran parte de los candidatos a ganar el premio más relevante a nivel individual como es el Balón de Oro.

Usando la Figura 3 para concluir con el análisis de los distintos clústeres se pueden sacar algunas conclusiones luego de encontrar los perfiles de los jugadores pertenecientes a cada clúster, como se puede observar los jugadores que pertenecen al clúster 4 y al 3 son totalmente opuestos igual que los jugadores del clúster 3 con los del 1, estás visualizaciones pueden quedar más claras al entender el perfil de los jugadores presentes en estos clústeres donde el clúster 3 pertenece a delanteros con bajo rendimiento goleador, el clúster 4 se compone de mediocentros más creativos y por último el clúster 1 se compone de jugadores con un perfil de corte defensivo, por lo tanto podemos entender que son los tres clústeres claros por cada una de las posiciones. Por el contrario, los clústeres 2 y 5 son los dos que más se parecen a todos, aunque luego los valores de 5 son más heterogéneos y se van esparciendo, alejándose del centro, esto puede ser debido a los perfiles de los jugadores que encontramos en estos dos clústeres, en uno nos encontramos con los jugadores que son suplentes o que cuentan con menos minutos en sus equipos también nos podemos encontrar con jugadores jóvenes que hayan entrado hace poco en dinámica del primer equipo. En contraposición en el clúster 5 claramente nos encontramos con los jugadores que

suelen ser referencias en sus equipos, lo que explica que nos encontremos una gran cantidad de outliers ya que probablemente pertenezca a esos jugadores denominados fueras de serie.

Una vez se han determinado los perfiles de los clústeres se procedió a analizar los distintos modelos de clasificación que se han creado.

Para el análisis de los modelos de clasificación se obtuvieron las matrices de confusión

Predicción \ Real	**DF**	**FW**	**MF**
DF	313	5	24
FW	7	168	47
MF	14	43	167

Tabla 1

En primer lugar, se analizó la Tabla 1 correspondiente a la matriz de confusión del modelo de clasificación de las posiciones. Se puede observar que el modelo predice con bastante acierto las posiciones de los jugadores, cabe destacar que la confusión más relevante se produce entre las posiciones de FW y MF. Ya que como se observa en la tabla el modelo tiende a intercambiar ambas categorías, con 43 observaciones de FW (reales) mal clasificadas como MF, y 47 observaciones de MF reales clasificadas como FW, esto sugiere que ambas posiciones pueden compartir ciertas características similares. Por el contrario, para la posición de DF encontramos solo 21 valores mal clasificados lo que indica que la categoría que mejor es clasificada por el modelo es la posición DF.

Obteniendo un Precisión general del 82,23%, y con un Índice Kappa de 0,7279. Estos valores indican un desempeño general bueno y también indican que existe una concordancia entre los valores predichos y los valores reales, como se ve reflejado en la tabla 1, En el contexto del trabajo un Kappa de aproximadamente 0,73 sugiere que el modelo supera con creces a una clasificación realizada por azar, aunque aún existe margen de mejora en la consistencia.

Métrica	DF	FW	MF
Sensibilidad	0.9371	0.7778	0.7017
Especificidad	0.9361	0.9056	0.8964
Valor Predictivo Pos.	0.9152	0.7568	0.7455
Valor Predictivo Neg.	0.9529	0.9152	0.8741
Prevalencia	0.4239	0.2741	0.3020
Tasa de Detección	0.3972	0.2132	0.2119
Prevalencia Detectada	0.4340	0.2817	0.2843
Precisión Balanceada	0.9366	0.8417	0.7990

Tabla 2

En cuanto a la Tabla 2 la cuál presenta los datos de comportamiento por clase, se perciben diferencias relevantes en las métricas de rendimiento. La posición de DF resultó ser la mejor clasificada, con la sensibilidad más alta (93,7%), lo que implica que los jugadores que pertenecían a la posición de DF fueron identificados de forma correcta por el modelo. Analizando el resto de los valores como son especificidad, valor predictivo positivo (VPP) y

valor predictivo negativo (VPN), se saca la conclusión de que el modelo predice con gran exactitud a los jugadores que ocupan la posición de defensa, por encima de los jugadores que corresponden a las otras dos posiciones. Para la posición de FW el desempeño del modelo fue más intermedio, presentado una sensibilidad aproximadamente del 77,8%, lo implica que solo un tercio de los jugadores que realmente juegan como FW fueron mal clasificados. En contra su especificidad fue de 90,6% demostrando que la mayoría de los futbolistas que no juegan de FW se clasificaron correctamente como no pertenecientes a dicha categoría, el VPP fue del 75,7% lo que implica que una cuarta parte de las predicciones que el modelo marco como FW en realidad pertenecían a otra posición. Por otro lado, el VPN de FW fue muy alto 91,5% lo que implica que cuando el modelo predijo que no ocupan la posición de FW acertó en la gran mayoría de casos. Por último, se analizó la predicción para la posición de MF, que resulto ser la más problemática, obtuvo una sensibilidad más baja, alrededor de 70,2% evidenciando que casi un 30% de las instancias MF reales no fueron reconocidas correctamente. Lo que sugiere una dificultad del modelo a la hora de identificar todos los ejemplos de MF. Por coherencia, el VPP fue el más bajo (74,6%), indicando que una proporción significativa de las predicciones resultaron ser incorrectas. Su especificidad (89,6%) no se distanció tanto del resto de categorías, lo que implica que el modelo confunde ocasionalmente las categorías, el VPN de aproximadamente 87,4% fue el más pequeño de todos, aun así, sigue siendo aceptablemente alto; esto significa que, si bien la mayoría de los casos predichos como lo contrario a MF efectivamente no eran MF.

Predicción \ Real	**1**	**2**	**3**	**4**	**5**
1	63	2	1	0	0

Predicción \ Real	1	2	3	4	5
2	12	236	15	13	4
3	0	21	300	0	0
4	2	8	0	68	3
5	0	2	0	3	35

Tabla 3

A continuación, se procedió con el análisis del modelo de clasificación de los clústeres, entrenado para distinguir cinco categorías. La matriz de confusión representada en la Tabla 3, refleja un nivel de confusión relativamente bajo entre los clústeres, no se encontraron patrones de error extremadamente concentrados en una sola pareja de categorías. La confusión más destable ocurre entre los clústeres 2 y 3, encontrando 21 instancias que pertenecían al clúster 2 fueron clasificadas dentro del clúster 3, y a la inversa 15 instancias del clúster 3 se clasificaron de forma errónea como 2. En el resto de las categorías no encontramos una gran cantidad de casos mal clasificados lo que sugiere que las confusiones encontradas no comprometen gravemente el desempeño del modelo para la clase dada, y encontrando solo pequeños errores aislados.

Esto queda relejado al comprobar que el modelo obtuvo una precisión global del 89.09%, con un coeficiente Kappa de 0,8426. Este desempeño es muy elevado, reflejando que 9 de cada 10 observaciones fueron clasificadas de manera correcta dentro de cada uno de los clústeres. Además, un Kappa de aproximadamente 0,84 sugiere un acuerdo casi perfecto entre las predicciones del modelo y los valores reales. Por lo general, el modelo demuestra ser robusto y

bastante fiable para este conjunto de datos, superando con creces al azar y reduciendo al mínimo los errores sistemáticos.

Métrica	**Clase 1**	**Clase 2**	**Clase 3**	**Clase 4**	**Clase 5**
Sensibilidad	0.8182	0.8773	0.9494	0.8095	0.8333
Especificidad	0.9958	0.9152	0.9555	0.9815	0.9933
Valor Predictivo Pos.	0.9546	0.8429	0.9346	0.8395	0.8750
Valor Predictivo Neg.	0.9806	0.9350	0.9657	0.9774	0.9906
Prevalencia	0.0977	0.3414	0.4010	0.1066	0.0533
Tasa de Detección	0.0800	0.2995	0.3807	0.0863	0.0444
Prevalencia Detectada	0.0838	0.3553	0.4074	0.1028	0.0508
Precisión Balanceada	0.9070	0.8963	0.9524	0.8955	0.9133

Tabla 4

Una vez analizadas la matriz de confusiones y las estadísticas generales del modelo se procedió a analizar las estadísticas por clúster las cuales fueron reflejadas en la Tabla 4, lo primero que se puede deducir viendo los datos reflejados con anterioridad es que el modelo tiene un comportamiento bastante equilibrado, con todas las sensibilidades por encima del 80%, lo que implica un alto acierto en las clasificaciones de cada categoría. El clúster que obtuvo la mayor sensibilidad fue el 3 con una sensibilidad del 94,9% lo indica que prácticamente un 95% de los jugadores pertenecientes a dicho clúster fueron identificados correctamente. Sin embargo, pese a tener un VPP de aproximadamente un 93,5% no obtuvo el valor más alto en esta estadística, el

cual pertenece al clúster 1 con un 95,5% lo que significa que todos los jugadores que fueron predichos en el clúster 1 efectivamente pertenecían a este clúster, este dato contrasta con su sensibilidad que es de 81,8% que pese a ser elevada es bastante más baja que la reflejada en el clúster 3. Del resto de clústeres cabe destacar los datos del clúster 5 que obtuvo un buen rendimiento con una sensibilidad de 83,3% y un VPP de 87,5% pese a ser el clúster con menos representaciones del conjunto total con solo 42 jugadores reales de los cuales 35 fueron clasificados de forma correcta.

Por el contrario, los clústeres 2 y 4 obtuvieron un rendimiento algo inferior a los anteriores, aunque igualmente presentan métricas positivas. El clúster 2 obtuvo una sensibilidad de 87,7% lo que implica un gran acierto, aunque se perdiese alrededor del 12,3% de los jugadores. Sin embargo, lo que más destaca es su alta especificidad de 91,5% siendo la más alta de las 5, lo que indica que en algunos casos el modelo predijo que algunos jugadores pertenecían al clúster 2 cuando realmente pertenecían a otra clase, como dato positivo nos sale su VPN de 93,5% el más alto de todos, lo cual indica que cuando el modelo descarta la clase 2 en una instancia, suele acertar. En cuanto al clúster 4, presentó la sensibilidad más baja del modelo (80,9%), lo que sugiere que cerca del 19% de los ejemplos de jugadores que pertenecen al clúster 4 fueron pasados por altos, el porcentaje de omisión más alto entre los 5 clústeres. Su VPP fue de 83,9% algo similar a la clase 2 lo que implica que 1 de cada 6 predicciones de dicho clúster fueron incorrectas. Cabe destacar que, aun siendo la clase con menor sensibilidad, la clase 4 mantuvo una especificidad elevada (98,2%), indicando que raramente se confunde con el resto de los clústeres. Además, el VPN fue aproximadamente de 97,7% evidenciando que casi todos los

casos donde el modelo predijo la no pertenencia al clúster efectivamente no pertenecían a dicho clúster.

Una vez analizados todos los resultados obtenidos tanto para la formación de los clústeres como para comprobar el rendimiento de los distintos modelos se procedió a analizar el objetivo principal del trabajo el cual era el sistema de recomendación. Para ello se comprobó su rendimiento con distintos ejemplos. Se escogieron tres casuísticas distintas, la primera se decidió comprobar en una recomendación de un jugador que se encuentre dentro de la base de datos, para ello se escogió un caso realista como es la búsqueda de un sustituto para el jugador del FC Barcelona Robert Lewandowski, en el segundo caso se propuso la idea de comparar a un jugador que ya no esté en activo con los jugadores que han disputado está temporada, para ello se escogió el caso de Toni Kroos exjugador del Real Madrid y que dejo de jugar en la temporada 2023/2024 y por último se comprobó si se podía usar el sistema para mejorar el rendimiento de un jugador en un equipo, para ello se buscó un sustituto para la delantera del Liverpool debido al descontento del equipo con el rendimiento de Darwin Núñez.

Se comenzó analizando el problema de Lewandowski, para entender la elección del jugador, se escogió un contexto que sería un caso real que tendrá que afrontar la dirección deportiva del FC Barcelona el verano de 2025 y es la búsqueda de un sustituto de garantías a Robert Lewandowski, ya que se plantan en una situación en la cual el jugador arranca la temporada 2026 con 37 años y el club no tiene ningún futbolista de garantías que pueda cubrir una posible retirada o que sirva de sustituto en caso del lesión del jugador, pese al buen rendimiento de Ferran Torres como su suplente durante la temporada de 2025 en el periodo en el que Lewandowski estuvo lesionado.

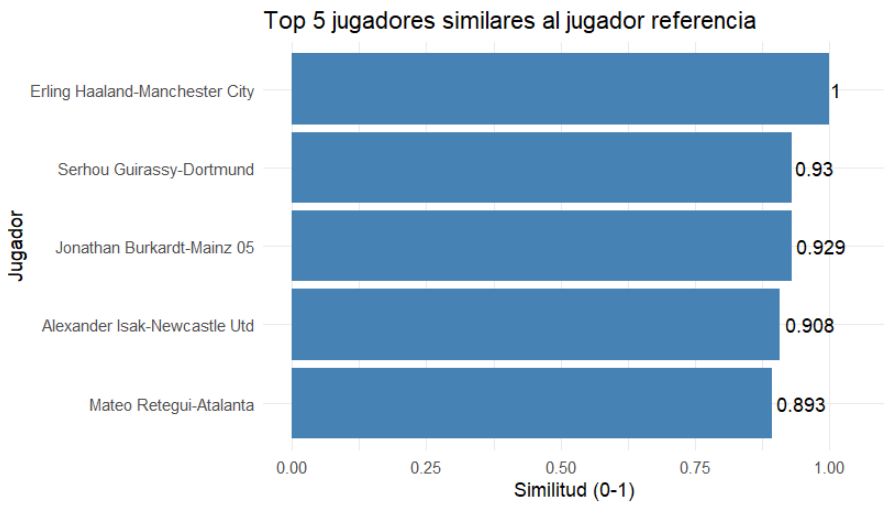

Figura 6

Una vez ejecutado el sistema de recomendación se obtuvo la Figura 6 en la cual observamos a todos los jugadores con mayor similitud con respecto a Robert Lewandowski, el jugador con el que se obtuvo una mayor similitud fue con Erling Haaland, lo que cuadra conociendo a ambos futbolistas ya que ambos corresponden a un perfil claro de delantero con una gran cantidad de goles anotados, que generan juego con balón fuera del área pero que generalmente viven de sus contribuciones dentro del área, este caso permite comprobar que el modelo se puede acercar bastante a la realidad de las direcciones deportivas ya que Haaland es el fichaje soñado por el actual presidente del FC Barcelona Joan Laporta, aunque en sus últimas declaraciones para mundo deportivo descartase su fichaje por el FC Barcelona en el verano de 2025 (Gascón, 2025), por lo tanto aunque la recomendación parece acertada, se descarta presentar el fichaje de Halaand como propuesta para la directiva, el siguiente jugador en salir es Serhou Guirassy del Dortmund, este jugador Guineano se destapo en la temporada 2023/2024 como uno de los mejores goleadores en la Bundesliga con el Sttutgart FC, lo que le valió su fichaje en el verano por el Borussia Dortmund donde se ha confirmado como uno de los mejores delanteros del

mundo a sus 29 años anotando 21 goles en Bundesliga y repartiendo 4 asistencias en 30 partidos, estás son las cifras que se cuentan dentro de la base de datos que no tiene en cuenta las competiciones europeas como la Champions, factor que juega a favor de Guirassy donde ha sido uno de los máximos anotadores de la competición con 13 goles, pese a tener estas cifras espectaculares, su fichaje por el FC Barcelona se comprendería como un movimiento cortoplacista, ya que el buen rendimiento en la elite de Guirassy puede no ir más lejos de los 34 años, es decir sería un fichaje para transición entre Lewandowski y el próximo gran delantero de la entidad blaugrana. El siguiente jugador en aparecer con una gran similitud con Robert Lewandowski es Jonathan Burkardt, este jugador de 24 años sería el jugador más por debajo del radar de los cinco que se encuentran en la Figura 6 como los más similares. Se trata de un jugador alemán con gran proyección que en la temporada 2025 anota 18 goles con el Mainz05 equipo de la Bundesliga, y que ha sido relacionado con equipos como el Bayern de Munich (Vargas, 2025), el fichaje del jugador germano por el Barça sería extraño que ocurriese ya que no ha salido ninguna información que le relacione con el equipo blaugrana pero podría ser un fichaje interesante por edad, precio y capacidad de adaptarse a un rol de suplente mientras Lewandowski sigue en activo y luego ocupar ese puesto de titular en la delantera del blaugrana. Los dos últimos jugadores en aparecer son Alexander Isak y Mateo Retegui, son dos perfiles también interesantes para la directiva azulgrana, Isak se ha convertido en una estrella en el Newcastle United de la Premier League y suena con fuerza para salir de las urracas en el verano de 2025, el gran impedimento para su fichaje por la disciplina blaugrana sería el elevado precio que pide el Newcastle por el jugador que es cercano a los 140 Millones € algo que en la situación económica actual del Barça sería casi imposible su fichaje, además hay que añadir que es un

delantero cotizado en el mercado y que varios equipos con gran potencia económica están pujando por hacerse con el jugador, pese a todo esto algunos medios lo ponen como el principal nombre junto a Haaland para ficharlo en el verano de 2026 en el cual Lewandoski se plantara ya en los 38 años (S.Montero, 2025). Por último, analizaremos el nombre de Mateo Retegui, delantero italo-argentino que juega en el Atalanta de Bergamo, como curiosidad con respecto al resto de jugadores con perfiles similares a Lewandowski este jugador es el único que no ha jugado nunca en ningún equipo de la Bundesliga, por lo que también podemos detectar que existe un patrón entre Lewandowski y los goleadores que se han formado en el fútbol alemán, yendo al nombre de Retegui puede ser una opción interesante, ya que de los jugadores que encontramos es joven todavía con solo 26 años, dato que juega a favor de su comparación con Guirassy, tiene partidos en competiciones europeas factor que sale a su favor en las comparaciones con Burkardt, y se le presupone un valor de mercado menor que Halaand o Isak, que podría ser algo interesante a tener en cuenta para la directiva del Barça. Una vez presentado todos los perfiles de los jugadores recomendados por el programa se pueden sacar varias conclusiones, el modelo con jugadores que se encuentran dentro de la base de datos consigue sacar unas similitudes bastantes realistas, como podemos ver recogiendo noticias de medios de información que confirman el interés del FC Barcelona por algunos de los jugadores presentes en la lista. Como resolución al mercado del Barça la mejor opción sería Haaland o Isak pero debido a su precio tan alto y a los problemas que tiene el Barça financieros actualmente sería casi imposible su fichaje por el conjunto blaugrana este verano, como alternativa propondría a Retegui pese a ser el jugador que menos se parece de la lista, sigue siendo un jugador joven cuyo

valor de mercado no está disparado y que ha jugado ya en Champions demostrando nivel de sobra para ser un delantero con un gran rendimiento en la competición europea.

El siguiente perfil en ser analizado sería el de Toni Kroos, jugador que en la temporada 2024/2025 no disputo ningún partido debido a su retiro la temporada anterior y del que se obtienen sus datos de la última temporada que jugó que fue la temporada 2023/2024, se propuso este ejemplo ya que uno de los problemas que se encontraron en la decepcionante temporada del Real Madrid en la temporada 2024/2025 fue la falta de un sustituto con un perfil parecido al del jugador germano.

Para obtener sus datos se usó la mismas técnicas de depuración que en la base de datos de la temporada 2024/2025, y una vez obtenido su perfil estadístico se usó el modelo de clasificación de clúster para poder predecir en que grupo se encontraría el jugador, que determinaría que el jugador se encontraría en el clúster 1 debido a su perfil de mediocentro de corte más defensivo que desarrolló en su última temporada en el Real Madrid, luego se procedió con el uso del sistema de recomendación que permitiese encontrar los perfiles de los jugadores.

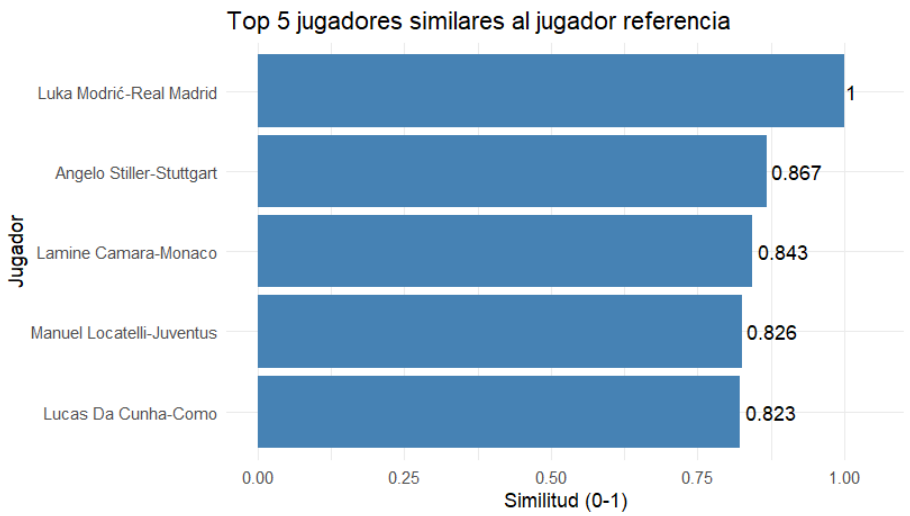

Figura 7

Como se observa en la Figura 7 los jugadores con perfiles similares a la última temporada de Toni Kroos, serían los siguientes, Luka Modric como el jugador más similar, la explicación más lógica sería que en esta campaña hubiese ocupado un rol parecido al que ocupo Toni Kroos en su última temporada en el conjunto merengue, por lo tanto no se tendría en cuenta para sustituir a Toni Kroos además se suma que la siguiente temporada el jugador balcanico anuncio en sus redes sociales que no seguirá en la entidad blanca. Por lo tanto pasamos a analizar a los siguientes jugadores en la lista, el siguiente nombre que aparece es el jugador alemán del Sttugart Angelo Stiller el cual ha sido relacionado con el Real Madrid como opción para ficharle como sustituto de Toni Kroos, y al cual se le ha comparado con el exjugador alemán del Real Madrid (Mx, 2025), utilizando está información se puede concluir que el sistema ha conseguido acertar en gran medida con la similitud sacada entre ambos jugadores. Los siguientes en las listas ya son nombres de distintos jugadores que ocupan una posición de jugador defensivo con buen toque de balón en sus equipos como son Lamine Camara en el Monaco, Manuel Locatelli en la Juventus y Lucas da Cunha en el Como, salvo Locatelli que si es un jugador más reconocido en el panorama

del fútbol mundial el resto son jugadores más jóvenes y con menos recorrido, también menos conocidos por lo tanto pueden ser jugadores interesantes para ser monitoreados. Con estos resultados se sacaron conclusiones de la exactitud de la clasificación de los jugadores en los clústeres mediante el uso del sistema supervisado de clasificación de clúster, usando de un jugador que no pertenece a la base de datos y del cual se pueden tener sus resultados de un rendimiento más cercano.

Para concluir con las pruebas del sistema de recomendación se decidió probar una situación en la cual se mejorase a un jugador con el que están a disgusto en su club como es Darwin Núñez en el Liverpool para ello se decidió probar con unas mejoras porcentuales en ciertas estadísticas en la base de datos ya normalizada del jugador.

Estadística	**Descripción**	**Mejora aplicada**
Gls	Goles totales	+30%
xG_Per	xG por 90 minutos	+20%
Gls_Per	Goles por 90 minutos	+40%
SoT_Standard	Tiros a puerta totales	+50%

Tabla 5

Dichas mejoras quedan reflejadas en la Tabla 5, para su simulación se procedió a actualizar primero estás variables en la línea estadística de Darwin Núñez, una vez se realizaron estás mejoras en la línea estadística se procedió con la clasificación del jugador para comprobar si cambiaba de grupo o se mantenía, en este caso el jugador se mantenía en el mismo clúster el cual

era el grupo 5, una vez realizado esto se procedió a sacar a los jugadores más similares a Darwin Núñez y luego a su nueva línea estadística mejorada.

Jugador	**Distancia**	**Similitud**
Shavy Babicka-Toulouse	6.62	1.000
Justin Njinmah-Werder Bremen	6.77	0.998
Mama Samba Baldé-Brest	6.79	0.997
Ado Onaiwu-Auxerre	7.18	0.991
Eldor Shomurodov-Roma	7.20	0.991

Tabla 6

Jugador	**Distancia**	**Similitud**
Shavy Babicka-Toulouse	7.87	1.000
Diogo Jota-Liverpool	8.03	0.998
Ado Onaiwu-Auxerre	8.04	0.997
Justin Njinmah-Werder Bremen	8.09	0.997
Eldor Shomurodov-Roma	8.11	0.996

Tabla 7

En este caso a diferencia de los anteriores se determinó enseñar los datos con la similitud y con los cálculos de las distancias euclídeas entre jugadores. La Tabla 6 corresponde a los resultados obtenidos con la línea estadística original de Darwin y la Tabla 7 representa los datos de la nueva línea estadística mejorada del jugador del Liverpol, como se pudo observar en ambas tablas aunque el nombre de los jugadores en ambas tablas no cambiaba en exceso, solo cabe destacar el cambio la entrada de Diogo Jota jugador que se disputa el puesto de titular en Liverpool con Darwin y que parte con cierta ventaja sobre el jugador uruguayo, y la salida Mama Samba Baldé del Brest. Este primer cambio nos da detalles que las mejoras se han implementado bien ya que como se ha comentado Diogo Jota es el delantero centro titular por delante de Darwin Núñez en el Liverpool, también se encuentra que existe un distanciamiento entre los jugadores que más se asimilan a Darwin ajustándose más al perfil mejorado del jugador Charrúa, lo que se podría interpretar que cualquiera de los jugadores podría ser un refuerzo de garantías para el Liverpool ya que mejorarían principalmente las contribuciones de gol y los disparos a puerta del Darwin.

Una vez realizadas las distintas pruebas se puede concluir que el sistema de recomendación es bastante acertado y que representa con bastante fiabilidad la realidad.

Conclusiones.

Para concluir, se darán una serie de conclusiones que se han obtenido al realizar todo el análisis del trabajo y de cómo se podría hacer evolucionar el sistema para que pueda ser usado con mayor eficacia dentro de las direcciones deportiva de los clubes.

Como conclusión principal del sistema se ha podido observar que el sistema implementado consigue acercarse bastante a la realidad del mundo del fútbol profesional, encontrando perfiles que puedan encajar en las peticiones que necesita realizar los usuarios del sistema, además la implementación de este sistema permite la mejora en la toma de las decisiones dentro de los fichajes o salidas de los clubes de fútbol, además añade un visión más objetiva a un mundo que solía ser más subjetivo como son los ojeadores, que al final basaban sus informes en las sensaciones que les transmitía un jugador sobre el terreno de juego y no sobre sus datos reales. Aunque el sistema ha resultado tener un buen rendimiento todavía existen muchas mejoras que se podrían implementar para su uso en equipos más modestos o incluso en equipos con fuerte conjunto de analistas, por ejemplo, una de las variables que no se tienen en cuenta en el modelo es el precio del jugador, algo que suele ser fundamental a la hora de realizar fichajes en los equipos de fútbol, siendo un mundo que las diferencias económicas marcan la diferencias a la hora de conseguir objetivos, para su implementación se pensó en el uso de modelos predictivos capaces de simular un valor lo más cercano posible a la realidad, y que luego a la hora de tomar la decisión sobre la elección del jugador a elegir se tenga en cuenta si entra o no en el presupuesto planificado por los directivas. Otro punto importante es el uso del modelo de clasificación por posiciones el cuál a la hora de recomendar jugadores no se ha llegado a implementar pero se podría usar para encontrar jugadores que puedan jugar en una posición distinta a la que ocupan en sus clubes actuales.Un ejemplo claro de casos de jugadores que un cambio en su posición supuso una mejora enorme en su carrera pueden ser jugadores como Joelinton que paso de ser un delantero con poco rendimiento a convertirse en uno de los mejores

mediocampistas del mundo gracias a su capacidad de recorrer gran parte del campo, ser fuerte en los duelos y aportar una gran llegada al área.

Otro de los puntos a mejorar es la implementación de la recomendación de los jugadores adaptada a un cambio de sistema de juego en el equipo, aunque su implementación sería más sencilla gracias a poder ajustar las variables del jugador y la búsqueda de perfiles y no solo de sustitutos de jugadores.

Por último para su implementación en equipos con menor presupuesto se tendría que hacer un trabajo previo en el que se recojan los datos con las estadísticas utilizadas en está base de datos y cuyos cálculos están disponibles para cualquier usuario en una búsqueda en internet, consiguiendo una base datos correcta está herramienta puede ser usada por una gran cantidad de clubes para encontrar perfiles de jugadores que puedan hacer avanzar a estos clubes una mayor dimensión ya que podrían ayudarse del buen rendimiento comprobado en los jugadores de elite para facilitar los pasos a seguir en sus planificaciones deportivas.

Por lo tanto pese a tener estos pequeños debes el sistema de recomendación creado resulta ser de gran ayuda y puede aportar a una innovación en el mundo del deporte sin precedentes, dando una mayor importancia a los datos objetivos, aunque sin perder de vista la subjetividad que muchas veces antes a acertado, estás mejoras tienen que ser herramientas que complementen la toma de decisiones y no sustitutivos a las técnicas que se usaban con anterioridad, ya que los datos no pueden recoger cierta información como el estado anímico de un jugador, su ética de trabajo o su sentimiento de pertenencia al club de propiedad y que en una gran cantidad de casos suponen una importante parte del rendimiento de los jugadores, tampoco con estos sistemas se pueden

prevenir situaciones que supongan un fracaso en el fichaje de un jugador como pueden ser las lesiones.

Bibliografía

Dataiku. (22 de 05 de 2025). Clustering: la fórmula secreta para llegar mejor a tus clientes. Recuperado el 22 de 05 de 2025, de Keyrus: https://keyrus.com/sp/es/insights/clustering-la-formula-secreta-para-llegar-mejor-a-tus-clientes

Fbref. (27 de 05 de 2025). Fbref. Obtenido de Fbref: https://fbref.com/en/

Gascón, J. (20 de 05 de 2025). Laporta revela qué piensa ahora el Barça de Haaland y Nico Williams. Obtenido de Mundo Deportivo: https://www.mundodeportivo.com/futbol/fc-barcelona/20250520/1002468026/laporta-revela-que-piensa-barca-haaland-nico-williams.html

Interactive Chaos . (24 de 05 de 2025). Random Forest. Obtenido de Interactive Chaos : https://interactivechaos.com/es/manual/tutorial-de-machine-learning/random-forest#:~:text=El%20algoritmo%20Random%20Forest%20,autom%C3%A1tico%20m%C3%A1s%20populares%20y%20ampliamente

JaseZiv. (01 de 01 de 2020). worlfootballR_data. Obtenido de github: https://github.com/JaseZiv/worldfootballR_data

Martín, K. (07 de 06 de 2022). Ramineni y Rudd, los 'gurús' a la sombra del Leganés americano. Obtenido de Diario As: https://as.com/futbol/segunda/ramineni-y-ruud-los-gurus-a-la-sombra-del-leganes-americano-n/

Mx, D. T. (14 de 05 de 2025). Angelo Stiller es seguido por el Real Madrid. Obtenido de Onefootball: https://onefootball.com/de/news/angelo-stiller-es-seguido-por-el-real-madrid-41111028

Noble, J. (05 de Agosto de 2024). ¿Qué es el clustering jerárquico? Obtenido de IBM: https://www.ibm.com/es-es/think/topics/hierarchical-clustering

OneFootball. (20 de 04 de 2022). How moneyball has changed football transfers forever. Obtenido de OneFootball: https://onefootball.com/en/news/how-moneyball-has-changed-football-transfers-forever-34949231

OptaAnalyst. (27 de 05 de 2025). OptaAnalyst. Obtenido de OptaAnalyst: https://theanalyst.com/eu

S.Montero. (10 de 05 de 2025). Alexander Isak se pone a tiro del Barça y Flick se frota las manos. Obtenido de El Nacional.cat: https://www.elnacional.cat/es/deportes/alexander-isak-se-pone-tiro-barca-flick-se-frota-manos_1413256_102.html

UEFA. (27 de 05 de 2025). Ranking de federaciones de la UEFA. Obtenido de UEFA.com: https://es.uefa.com/nationalassociations/uefarankings/country/?year=2025

Vargas, I. (24 de 03 de 2025). El Bayern sonríe: el factor que acerca a Jonathan Burkardt. Obtenido de Fichajes.com: https://www.fichajes.com/a3761880895366025073-el-bayern-sonrie-el-factor-que-acerca-a-jonathan-burkardt

Wyscout. (27 de 05 de 2025). Wyscout. Obtenido de Wyscout: https://es.hudl.com/es-xl/products/wyscout

Anexo.

A_minus_xAG_Expected: Diferencia entre las asistencias reales y las asistencias esperadas (xAG).

Ast_Per: Asistencias por cada 90 minutos de juego.

Att_Att_3rd_Tackles: Entradas realizadas en el tercio ofensivo del campo.

Att_3rd_Touches: Toques de balón en el tercio ofensivo del campo.

Att_Challenges: Duelos ofensivos disputados.

Att_Long: Pases largos intentados en ataque.

Att_Medium: Pases de distancia media intentados en ataque.

Att_Pen_Touches: Toques de balón dentro del área penal del oponente.

Att_Short: Pases cortos intentados en ataque.

Att_Take: Intentos de regate para superar a un oponente.

Att_Total: Total de acciones ofensivas intentadas.

Blocks_Blocks: Total de bloqueos defensivos realizados.

Blocks_Outcomes: Resultados de los bloqueos defensivos (por ejemplo, recuperación de balón, despeje).

CK_Pass: Pases realizados desde tiros de esquina.

CPA_Carries: Conducciones que avanzan el balón al área penal contraria.

Carries_Carries: Número total de conducciones de balón.

Clr: Despejes defensivos realizados.

Cmp_Long: Pases largos completados.

Cmp_Medium: Pases de distancia media completados.

Cmp_Outcomes: Resultados de los pases completados.

Cmp_Short: Pases cortos completados.

Cmp_Total: Total de pases completados.

Cmp_percent_Long: Porcentaje de pases largos completados.

Cmp_percent_Medium: Porcentaje de pases de distancia media completados.

Cmp_percent_Short: Porcentaje de pases cortos completados.

Cmp_percent_Total: Porcentaje total de pases completados.

Crs: Centros al área realizados.

CrsPA: Centros al área penal realizados.

Crs_Pass: Pases que son centros al área.

Dead_Pass: Pases realizados desde jugadas a balón parado.

Def_3rd_Tackles: Entradas realizadas en el tercio defensivo del campo.

Def_3rd_Touches: Toques de balón en el tercio defensivo del campo.

Def_GCA: Acciones defensivas que conducen a una oportunidad de gol.

Def_Pen_Touches: Toques de balón dentro del área penal propia.

Def_SCA: Acciones defensivas que conducen a una oportunidad de disparo.

Dis_Carries: Conducciones de balón que resultan en pérdida de posesión.

Dist_Standard: Distancia recorrida en promedio durante un partido.

Err: Errores que conducen a oportunidades de gol para el oponente.

FK_Pass: Pases realizados desde tiros libres.

FK_Standard: Tiros libres directos ejecutados.

Final_Third: Acciones realizadas en el tercio final del campo.

Final_Third_Carries: Conducciones de balón en el tercio final del campo.

Fld: Faltas recibidas.

Fld_GCA: Faltas recibidas que conducen a una oportunidad de gol.

Fld_SCA: Faltas recibidas que conducen a una oportunidad de disparo.

Fls: Faltas cometidas.

GCA90_GCA: Contribuciones a oportunidades de gol por cada 90 minutos.

GCA_GCA: Total de contribuciones a oportunidades de gol.

G_A: Goles y asistencias combinados.

G_A_Per: Goles y asistencias combinados por cada 90 minutos.

G_A_minus_PK_Per: Goles y asistencias sin contar penaltis por cada 90 minutos.

G_minus_PK: Goles anotados excluyendo penaltis.

G_minus_PK_Per: Goles anotados excluyendo penaltis por cada 90 minutos.

G_minus_xG_Expected: Diferencia entre goles reales y goles esperados (xG).

G_per_Sh_Standard: Goles por cada tiro realizado (efectividad de disparo).

G_per_SoT_Standard: Goles por cada tiro a puerta (efectividad de tiros al arco).

Gls: Goles totales marcados.

Gls_Per: Goles por cada 90 minutos.

Gls_Standard: Goles anotados en jugadas (excluye penales).

In_Corner: Saques de esquina recibidos en contra.

KP: Pases clave (pases que conducen a un tiro, aunque no sea asistencia).

Live_Pass: Pases realizados en juego activo (no a balón parado).

Live_Touches: Toques realizados durante el juego activo.

Lost_Aerial: Duelos aéreos perdidos.

Lost_Challenges: Duelos 1v1 perdidos (regates o disputas).

MP_Playing: Partidos jugados.

MP_Playing_Time: Tiempo total en cancha (minutos jugados).

Mid_3rd_Tackles: Entradas realizadas en el tercio medio del campo.

Mid_3rd_Touches: Toques en el tercio medio del campo.

Min_Playing: Minutos totales jugados.

Min_Playing_Time: Igual que el anterior (puede variar según fuente).

Min_percent_Playing_Time: Porcentaje de minutos jugados respecto al total posible.

Mins_Per_90: Normalización de estadísticas por 90 minutos.

Mins_Per_90_Playing: Estadísticas por 90 minutos jugados reales.

Mins_Per_90_Playing_Time: Igual que el anterior, diferente nombre.

Mis_Carries: Conducciones mal ejecutadas (pérdida de posesión por mal control).

Mn_per_MP_Playing_Time: Minutos por partido jugado.

Mn_per_Start_Starts: Minutos por titularidad.

Mn_per_Sub_Subs: Minutos por cada entrada como suplente.

OG: Autogoles.

Off: Veces en fuera de juego.

Off_Outcomes: Resultados de jugadas en las que hubo fuera de juego.

On_minus_Off_Team_Success: Diferencia en éxito del equipo cuando el jugador está en

cancha vs fuera.

On_minus_Off_Team_Success_G: Ídem, pero aplicado a goles.

Out_Corner: Saques de esquina ejecutados.

PK: Goles de penalti.

PK_Standard: Tiros penales convertidos.

PKatt: Penaltis intentados.

PKatt_Standard: Total de penaltis ejecutados.

PKcon: Penaltis encajados (usualmente para porteros).

PKwon: Penaltis ganados (provocados).

PPA: Pases al área penal (Penalty Area Passes).

PPM_Team_Success: Puntos por partido cuando el jugador juega.

PassDead_GCA: Contribuciones a gol derivadas de pases a balón parado.

PassDead_SCA: Contribuciones a tiros derivadas de pases a balón parado.

PassLive_GCA: Contribuciones a gol con pases en juego.

PassLive_SCA: Contribuciones a tiros con pases en juego.

Pass_Blocks: Pases bloqueados por un defensor.

PrgC_Carries: Conducciones progresivas (hacia adelante significativamente).

PrgC_Progression: Metros ganados con conducciones progresivas.

PrgDist_Carries: Distancia total recorrida con conducciones.

PrgDist_Total: Distancia total progresada (pases + conducciones).

PrgP: Pases progresivos (avanzan significativamente el balón hacia la portería rival).

PrgP_Progression: Distancia progresada mediante pases.

PrgR_Progression: Progresión recibida (recepciones que avanzan el balón).

Rec_Receiving: Recepciones de pases.

Recov: Recuperaciones de balón.

SCA90_SCA: Contribuciones a tiros por 90 minutos.

SCA_SCA: Contribuciones totales a tiros.

Sh_Blocks: Tiros bloqueados.

Sh_GCA: Disparos que conducen a goles.

Sh_SCA: Disparos que conducen a tiros posteriores.

Sh_Standard: Tiros realizados.

Sh_per_90_Standard: Tiros por cada 90 minutos.

SoT_Standard: Tiros a puerta.

SoT_per_90_Standard: Tiros a puerta por 90 minutos.

SoT_percent_Standard: Porcentaje de tiros que fueron a puerta.

Starts_Playing: Partidos iniciados como titular.

Starts_Starts: Total de titularidades.

Str_Corner: Saques de esquina ejecutados desde el lado fuerte (derecho para diestros, izquierdo para zurdos).

Subs_Subs: Ingresos como suplente.

Succ_Take: Regates exitosos.

Succ_percent_Take: Porcentaje de regates exitosos.

Sw_Pass: Cambios de juego (pases largos de banda a banda).

TB_Pass: Pases entre líneas (through balls).

TI_Pass: Saques de banda ejecutados.

TO_GCA: Conducciones que conducen a una oportunidad de gol.

TO_SCA: Conducciones que conducen a un tiro.

TklW: Entradas ganadas (recuperando el balón).

TklW_Tackles: Porcentaje de entradas ganadas.

Tkl_Challenges: Duelos defensivos intentados.

Tkl_Int: Entradas + Intercepciones (suma de ambas).

Tkl_Tackles: Entradas realizadas.

Tkl_percent_Challenges: Porcentaje de duelos defensivos ganados.

Tkld_Take: Veces que fue desposeído intentando regate.

Tkld_percent_Take: Porcentaje de regates en los que fue desposeído.

TotDist_Carries: Distancia total conducida con el balón.

TotDist_Total: Distancia total recorrida (con o sin balón).

Touches_Touches: Toques totales de balón.

Won_Aerial: Duelos aéreos ganados.

Won_percent_Aerial: Porcentaje de duelos aéreos ganados.

_2CrdY: Doble amarilla (expulsión por acumulación).

np_G_minus_xG_Expected: Goles no penales menos xG no penal.

npxG_Per: xG no penal por 90 minutos.

npxG_per_Sh_Expected: xG no penal por tiro.

npxG_xAG_Expected: xG + xAG combinados (no penales).

npxG_xAG_Per: xG + xAG combinados por 90 minutos.

onGA_Team_Success: Goles en contra cuando el jugador está en cancha.

onG_Team_Success: Goles a favor cuando está en cancha.

onxGA_Team_Success_G: xG contra mientras el jugador está en cancha.

onxG_Team_Success_G_: xG a favor mientras el jugador está en cancha.

plus_per__minus_90_Team_Success: Diferencia de goles por 90 minutos con el jugador en cancha.

plus_per__minus__Team_Success: Diferencia total de goles con el jugador en cancha.

unSub_Subs: Veces en que fue suplente pero no ingresó.

xAG: Asistencias esperadas (expected assisted goals).

xAG_Expected: Sinónimo de xAG.

xAG_Per: xAG por cada 90 minutos.

xA_Expected: Asistencias esperadas (a veces con nombre alterno).

xG_Per: xG por cada 90 minutos.

xG_xAG_Per: xG + xAG por 90 minutos.

xGplus_per__minus_90_Team_Succe: Diferencia xG +/- por 90 min.

xGplus_per__minus__Team_Success: Diferencia total xG +/- con el jugador.