TRABAJO FINAL DE MÁSTER

CÓMO UTILIZAR LA IA EN EL SCOUTING DE JUGADORES DE FÚTBOL

MÁSTER EN INTELIGENCIA ARTIFICAL AVANZADA Y APLICADA

UNIVERSITAT DE VALÈNCIA

AUTOR: LOSA BRITO, ALEJANDRO

TUTOR: LAPARRA PÉREZ-MUELAS, VALERO

CURSO: 2023/2024

Agradecimientos

A mi familia, quienes han valorado en primera línea cada paso de este camino. A mi padre, Alfonso, que me introdujo en el apasionante mundo del fútbol y me enseñó los valores del deporte. A mi madre, Yohanka, que me inculcó la importancia de la educación y los estudios en el desarrollo personal. A mi hermana Andrea, compañera de vida y ejemplo de superación ante cualquier adversidad. También, me gustaría dedicarle este trabajo a mi tío Víctor, una persona admirable que me motiva a parecerme a él en un gran número de facetas.

Quiero expresar mi más profundo agradecimiento a los analistas (Javi Navarro, Álvaro Máñez, Pablo Trenado y Fran Moragón) y al analista de datos (José Gila) del Levante Unión Deportiva S.A.D, quienes han compartido conmigo sus conocimientos y experiencias en el análisis y la comprensión del fútbol desde una perspectiva analítica. Sus valiosos aportes y la validación de mis resultados han sido cruciales para el éxito de este proyecto.

Por último, pero no menos importante, agradezco a Valero Laparra por su orientación y apoyo en el desarrollo de este trabajo.

Alejandro Losa Brito

Palabras clave

Scouting, fútbol, sistema recomendador, jugadores, estadísticas individuales, fichajes y predicción del rendimiento.

Key words

Scouting, soccer, recommender system, player, individual statistics, signings and performance prediction.

Resumen

El *Scouting* en el fútbol es un método para encontrar jugadores que se adapten y mejoren el rendimiento de un equipo. Además, cuando se realiza de manera adecuada, se convierte en una de las fuentes fundamentales de ingresos para cualquier club. El objetivo clave es obtener indicadores a través de datos que ayuden a resaltar talentos futbolísticos y diferencien entre jugadores.

La importancia del *Scouting* no sólo se ve reflejada en la construcción de un equipo competitivo capaz de mejorar el rendimiento dentro del campo. Económicamente, supone la generación de ingresos a través de la venta de jugadores, por lo que la compra de un jugador que posteriormente incremente su valor en el mercado generará unos beneficios que ayudarán a mejorar las opciones de adquisición de jugadores en la siguiente temporada, y, por tanto, tener mejores jugadores que aumenten el nivel de competitividad. Además, un buen trabajo de *Scouting* ayuda a conseguir una estabilidad financiera y, por consecuencia, será más atractivo para inversores y patrocinadores.

En este estudio se ha desarrollado una métrica denominada *Performance Evaluation Score* que evalúa el rendimiento de los jugadores en cada partido en una escala del 1 al 100. Además, al combinar datos futbolísticos con datos económicos, enriquece la base de datos y le aporta más valor. En concreto, el precio del jugador en el mercado y la fecha de finalización de su contrato son variables introducidas que serán utilizadas para alcanzar los objetivos propuestos. Esto puede ayudar a los clubs de fútbol a ajustarse al presupuesto de cada equipo y a filtrar jugadores que quedarán libres próximamente.

Por lo tanto, este trabajo se ha centrado en la creación de una base de datos que responde a la pregunta clave de negocio: "¿Qué jugadores tienen un buen rendimiento y deberían ficharse en función de las necesidades del equipo y presupuesto del club?". Además, también permite identificar jugadores con alta probabilidad de que su valor en el mercado aumente con el tiempo.

El *Performance Evaluation Score* y su predicción, no solo facilitarán la toma de decisiones en el *Scouting*, sino que también despertarán el interés de los equipos por tener la capacidad de

optimizar sus beneficios gracias a la introducción de la Inteligencia Artificial en el mundo del fútbol de un modo sencillo y entendible.

Abstract

Scouting in football is essential for identifying players who can enhance a team's performance and potentially become key revenue generators for the club. The primary goal is to gather data-driven insights that spotlight football talents and facilitate differentiation between players.

The significance of scouting extends beyond assembling a competitive team; it also plays a crucial role economically. Effectively scouting and purchasing players who later increase in market value can lead to substantial profits. These profits not only improve a club's ability to acquire better players in subsequent seasons but also enhance its overall competitiveness. Moreover, proficient scouting contributes to financial stability, making the club more appealing to investors and sponsors.

In this study, it's developed a metric called the Performance Evaluation Score, which assesses player performance in each match on a scale from 1 to 100. By integrating football data with economic data, such as a player's market price and the expiration date of their contract, the database becomes a richer resource. This integration allows football clubs to align with their budget constraints and prioritize players who will soon be free agents, optimizing recruitment strategies.

This project is centered around creating a database that addresses the crucial business question: "Which players perform well and should be signed based on the team's needs and the club's budget?" It also aims to identify players likely to increase in market value, providing an opportunity for clubs to capitalize on their scouting investments.

The Performance Evaluation Score and its predictive capabilities not only streamline decision-making but also pique the interest of clubs looking to leverage Artificial Intelligence to maximize their profits in the competitive world of football.

ÍNDICE

1. Introducción al problema

El fútbol, reconocido como una disciplina deportiva de carácter colectivo, implica la participación de dos conjuntos, cada uno integrado por once jugadores. La finalidad primordial radica en superar al adversario en términos de puntuación. Para alcanzar este cometido, los participantes se valen de estrategias con el fin de insertar la esfera en la meta contraria. Este objetivo demanda que los jugadores efectúen avances a través del campo de juego, intercambiando la posesión del balón y procurando eludir la interceptación por parte del equipo oponente.

En este escenario, se destaca la búsqueda de jugadores que posean atributos particulares, propicios para el fomento de situaciones que permitan la anotación o prevención de goles. Además, el fútbol es un deporte que ha captado el interés de millones de personas a nivel global. Dado su amplio reconocimiento, numerosos investigadores, académicos y especialistas han emprendido análisis y estudios para descifrar múltiples facetas vinculadas al desempeño de los jugadores, las métricas de evaluación y la valoración de mercado, entre otros aspectos.

En este contexto académico, es factible hallar una diversidad de investigaciones y estudios que abordan desde la generación de modelos pronósticos hasta la implementación de metodologías de análisis de datos y técnicas de compresión de información.

En el año 1950 se efectuaron las primeras evaluaciones en el ámbito del fútbol, consistiendo en anotaciones manuales sobre los sucesos en el terreno de juego. Posteriormente, hacia 1970, se inició la recolección de información respecto al rendimiento físico de los jugadores, despertando interés en métricas tales como la distancia transitada. Casi dos décadas más adelante, el enfoque de los análisis migró hacia aspectos técnicos específicos, como la cantidad y calidad de tiros o pases efectuados. En 1988, se marcó un hito con la introducción de CAMAS, siendo este el primer sistema computarizado para el análisis de encuentros futbolísticos. Ocho años después a este avance, se fundó AMISCO, convirtiéndose en la pionera entidad dedicada a la generación y compilación de datos relativos a los 22 jugadores participantes en un partido[1].

En la actualidad, se experimenta una recopilación creciente de datos que abarcan dimensiones técnicas, físicas, económicas y sanitarias de los jugadores. El Big Data y la Inteligencia Artificial han hallado aplicabilidad significativa en el ámbito futbolístico. En el presente trabajo el *Scouting*, sector dedicado a la identificación de jugadores con el objetivo de ser contratados, será el foco principal.

Además, con el propósito de identificar jugadores con perfiles similares que puedan integrarse de manera óptima al equipo y potencialmente reemplazar a los miembros actuales, se recurre primordialmente al análisis de componentes principales (*Principal Component Analysis*,

PCA). Este enfoque se destaca por su capacidad para reducir la dimensionalidad de grandes conjuntos de datos, preservando la variabilidad de la información[2].

Adicionalmente, para discernir entre distintos perfiles de jugadores o agrupar atributos similares, se suele implementar *clustering*[3]. Esta metodología facilita la segmentación de los datos en grupos, basándose en la similitud de las características de los jugadores, lo que permite una interpretación más afinada de los datos y contribuye a la toma de decisiones estratégicas en el contexto deportivo.

Además, el Trabajo de Fin de Máster (TFM) de M. del Pilar Selma en 2019[4]. Destaca dentro del amplio cuerpo de literatura sobre análisis deportivo, prestando una considerable atención al estudio de la similitud entre atletas, lo que representa una desviación notable de la norma en la ciencia deportiva.

El TFM mencionado aplica el PCA para identificar las variables más relevantes asociadas a cada posición en el campo. Además, utiliza gráficos de carga Bi-Plot para informar sobre el estilo de juego de los atletas, agrupándolos según sus características distintivas y estableciendo en qué sobresalen.

La investigación también resalta la utilidad de los gráficos de contribuciones para comparar a un deportista específico con el promedio para su posición. Este enfoque ha permitido un análisis detallado sobre en qué aspectos concretos destaca cada jugador, proporcionando así un marco valioso para una mejor comprensión de las habilidades y atributos distintivos de un atleta en el contexto deportivo.

Por otro lado, se ha identificado a Transfermarkt como la web con más información sobre el valor de mercado de los jugadores, convirtiéndose en la fuente primaria para investigadores dedicados a la estimación del valor actual de los atletas. Gracias a la disponibilidad de un punto de referencia específico, se ha observado una mayor diversidad en los modelos predictivos utilizados por investigadores que trataron de predecir el valor económico de los jugadores. Entre las metodologías aplicadas se encuentran los árboles de decisión, bosques aleatorios, k-nearest neighbors (vecinos más cercanos), regresión lineal y redes neuronales[5].

Los proyectos contemplaron la creación de un modelo predictivo capaz de definir el valor financiero de un deportista a partir de sus estadísticas, con el objetivo adicional de discernir cuáles factores ejercen una influencia más significativa en la valorización de dicho atleta.

Paralelamente, se está avanzando en la creación de herramientas diseñadas para evaluar cómo la incorporación de un jugador específico afecta los resultados deportivos del equipo, empleando las cadenas de Markov como herramienta analítica para lograr este propósito[6].

Una vez recolectada información sobre el estado del arte y las metodologías utilizadas en otras investigaciones se ha destacado el PCA por su eficacia en la reducción de la dimensionalidad de los datos, resaltando las relaciones significativas entre sujetos y variables, aspecto esencial cuando los sujetos son atletas y las variables representan sus competencias en

el juego. Este enfoque se convierte en un pilar dentro de este Trabajo de Fin de Máster (TFM), debido a su habilidad para enfatizar la relevancia de ciertas variables en una componente. En este caso, la meta es asignar un valor ponderado a la importancia de cada característica, buscando construir un modelo capaz de calcular el rendimiento de un jugador

El análisis no incorporará el valor económico para la creación del valor que mida el rendimiento, pero sí que será importante en pasos posteriores, ya que facilita que equipos con presupuestos limitados puedan descubrir jugadores de alta rentabilidad y rendimiento deportivo. Esta estrategia no solo es una herramienta para optimizar el talento dentro del campo, sino que también se presenta como una vía para incrementar los ingresos mediante la adquisición y venta estratégica de estos talentos.

La escasez de literatura específica en el área de estudio destaca que para alcanzar los objetivos planteados se necesitará originalidad y un enfoque innovador lo que añade un valor significativo a este campo de investigación. Un ejemplo de esto es la creación de una métrica de rendimiento que no existe y que para obtenerla se necesitará darle una vuelta a los métodos no supervisados de los que disponemos.

No obstante, se ha observado un sistema de evaluación de jugadores implementado por FIFA en sus videojuegos, donde se asigna a cada atleta una puntuación que refleja su desempeño deportivo junto con otros factores subjetivos[7]. A pesar de su popularidad, este método ha enfrentado críticas debido a su tendencia a subestimar a jugadores de alto rendimiento por su menor grado de reconocimiento. Adicionalmente, la aplicación uniforme de esta fórmula a todas las posiciones ignora la diversidad de habilidades cruciales para cada rol específico en el campo, sugiriendo la necesidad de un sistema más adaptativo y preciso que valore adecuadamente las competencias clave según la posición del jugador.

Este escenario ejemplifica cómo un jugador con habilidades superiores puede ser valorado por debajo de otro más mediático, evidenciando una discrepancia en el sistema de calificación de FIFA Ultimate Team de EA Sports. La fama de un jugador fuera del campo de juego parece influir en su calificación dentro del juego, lo que destaca una valoración subjetiva por parte de EA Sports[8].

En el contexto de este Trabajo de Fin de Máster (TFM), estos casos ilustran la urgencia de abordar las deficiencias y sesgos en la evaluación de jugadores. Resalta la necesidad de desarrollar criterios objetivos y justos que se centren únicamente en el rendimiento deportivo, eliminando la influencia de factores externos no relacionados con las capacidades reales en el campo. Esto subraya la importancia de crear un sistema de evaluación más equitativo que pueda servir como un modelo fiable para la identificación y valoración del talento en el deporte.

El empleo del PCA se destaca como un recurso crucial para discernir las variables fundamentales que diferencian a los jugadores, asignando a su vez pesos específicos a estas variables dentro de cada componente. Esto permite una apreciación profunda de qué aspectos son verdaderamente significativos en la evaluación del rendimiento de un jugador. Este enfoque metodológico abre camino para identificar las variables más relevantes dentro de un conjunto de datos, ofreciendo una perspectiva única sobre la valoración del talento deportivo.

La utilización estratégica de la estructura de la base de datos para ponderar las variables según su relevancia es una innovación que facilita comprender las razones detrás del éxito de un jugador sobre otro, basándose en factores estrictamente relacionados con el desempeño en el campo. Al derivar estas variables directamente de los datos, se posibilita la creación de un valor representativo de la habilidad del jugador, libre de distorsiones por su fama o presencia en medios, contrarrestando así las limitaciones observadas en las evaluaciones convencionales como las realizadas por FIFA.

Este enfoque no solo podría mejorar la precisión en la identificación de talento, sino que también ofrece una herramienta valiosa para equipos que buscan optimizar su rendimiento deportivo y obtener ventajas económicas al descubrir jugadores de alta calidad y bajo costo. La importancia de este estudio se subraya aún más al considerar el impacto del *scouting* en la estrategia económica de los clubes deportivos, especialmente frente a otras fuentes de ingreso como los derechos televisivos. Esto resalta el valor potencial del proyecto en el ámbito del fútbol profesional, donde la eficaz identificación y valoración de jugadores puede traducirse en un significativo retorno económico y competitivo[9].

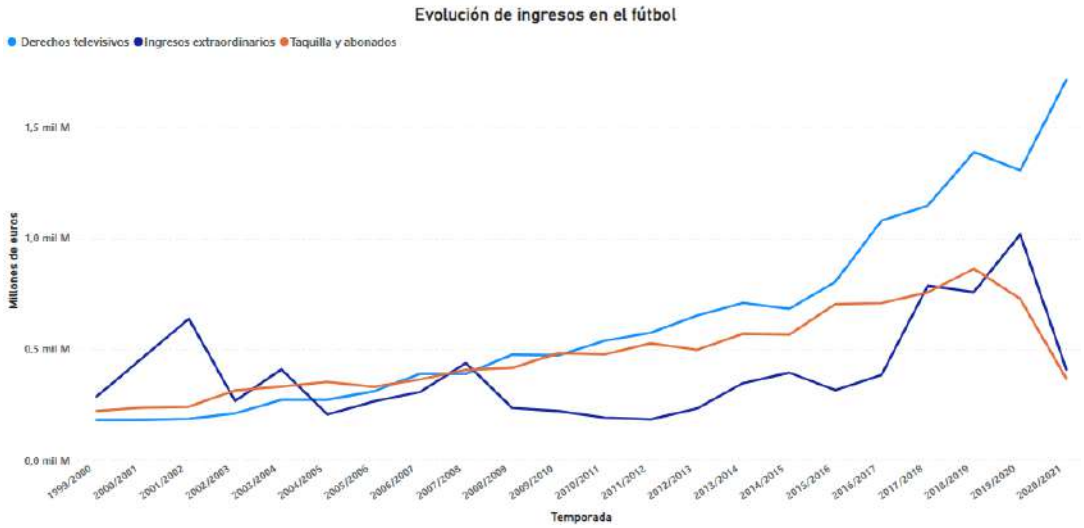

Figura 1. *Evolución de ingresos en la primera división española entre 1999 y 2021.*
Fuente: Elaboración propia a partir de datos del Consejo Superior de Deportes.

La información suministrada por el Consejo Superior de Deportes revela que los derechos televisivos constituyen la fuente primordial de ingresos para los clubes de fútbol, con un incremento destacado a partir de la temporada 2016/2017 y una tendencia ascendente hasta la temporada 2020/2021, donde se alcanzaron ingresos de hasta 1.7 millones de euros[10].

Interesantemente, la figura 1 indica que, durante el período comprendido entre las temporadas 2008/2009 y 2016/2017, los ingresos generados por taquilla y suscripciones de abonados predominaron sobre los ingresos extraordinarios, los cuales se derivan principalmente de la venta de jugadores. No obstante, en temporadas recientes, se observa un equilibrio más pronunciado entre estos dos tipos de ingresos, con una ligera ventaja en los procedentes del mercado de transferencias sobre los obtenidos por la venta de entradas y abonos.

El promedio de ingresos extraordinarios en los últimos cuatro años, cifrado en 743.451.680 euros, resalta la creciente importancia del *scouting* en la estrategia económica de los clubes. La habilidad para identificar y adquirir talento no solo potencia el rendimiento deportivo del equipo, lo cual puede traducirse en un aumento del interés del público y, por ende, en mayores ingresos por venta de entradas, sino que también influye directamente en los beneficios económicos derivados de victorias y posiciones destacadas en competiciones. El caso de La Liga en la temporada 2023/2024, donde el equipo campeón obtuvo 58,4 millones de euros, ejemplifica cómo el éxito deportivo se convierte en un factor determinante para la generación de ingresos significativos.

Este panorama enfatiza la importancia estratégica del *scouting* y la valoración adecuada de jugadores, no solo desde una perspectiva deportiva sino también económica, demostrando cómo una gestión eficaz del talento puede influir positivamente en la solidez financiera y competitiva de un club.

Posición	Fijo	% de variables	Millones de euros por puesto
1º Real Madrid	36,125	17,00%	58,4
2º Barcelona	36,125	15,00%	51,5
3º Girona	36,125	13,00%	44,6
4º Atlético de Madrid	36,125	11,00%	37,8
5º Athletic Club	36,125	9,00%	30,9
6º Real Sociedad	36,125	7,00%	24
7º Real Betis	36,125	5,00%	17,2
8º Villarreal	36,125	3,50%	12
9º Valencia	36,125	3,00%	10,3
10º Alavés	36,125	2,75%	9,4
11º Osasuna	36,125	2,50%	8,6

12º Getafe	36,125	2,25%	7,7
13º Celta de Vigo	36,125	2,00%	6,9
14º Sevilla	36,125	1,75%	6,1
15º Mallorca	36,125	1,50%	5,2
16º Las Palamas	36,125	1,25%	4,3
17º Rayo Vallecano	36,125	1,00%	3,4
18º Cádiz	36,125	0,75%	2,6
19º Almería	36,125	0,50%	1,7
20º Granada	36,125	0,25%	0,90

Tabla 1. *Tabla representativa de ingresos en la primera división española en la temporada 2023/2024 en función de la posición en la clasificación.*
Fuente: Elaboración propia partir de datos extraídos de la siguiente noticia: https://www.marca.com/futbol/primera-division/2024/05/24/664919f846163fdd368b459d.html

En el ámbito del fútbol, se manejan importantes sumas de dinero, donde el rendimiento actual del equipo es crucial, el cual depende directamente de la calidad y desempeño de los jugadores que lo conforman. Este rendimiento se ve influenciado significativamente por las actividades de compra y venta de jugadores que se realizan anualmente. Las elevadas cifras que rodean al mercado de transferencias resaltan la importancia de contar con herramientas eficaces para la identificación de jugadores con un alto rendimiento en el momento actual, facilitando así la toma de decisiones estratégicas en cuanto a fichajes se refiere.

Por otro lado, el rendimiento de un jugador interviene directamente en su precio en el mercado. Siendo importante fichar jugadores cuya progresión se refleje en una revaloración que aumente su precio. Varios estudios sitúan el estado en el que se encuentra la ciencia para predecir el rendimiento de un jugador. El estudio "Artificial neural networks and player recruitment in profesional soccer"[11], utiliza una arquitectura basada en redes neuronales para conseguir clasificar a un jugador en 3 grupos constituidos por jugadores de diferentes niveles de rendimiento (alto, medio, bajo. Los resultados indican una precisión entorno el 78%, afirmando que es posible predecir qué sucederá con un jugador en su carrera.

Este enfoque proporciona una herramienta valiosa para los equipos en su estrategia de reclutamiento, permitiéndoles tomar decisiones más informadas basadas en análisis predictivos. Al anticipar el rendimiento futuro de un jugador, los clubes pueden optimizar sus inversiones en el mercado de transferencias, seleccionando atletas cuyo perfil sugiere una trayectoria ascendente. Así, la aplicación de tecnologías avanzadas como las redes neuronales en el *scouting* y la evaluación de talento representa un avance significativo hacia la maximización del retorno de la inversión en jugadores, combinando la ciencia de datos con la gestión deportiva para potenciar el éxito tanto en el terreno de juego como en el aspecto financiero del club.

El estudio "Player Valuation in European Football"[12], llevado a cabo por Edward Nsolo, Patrick Lambrix y Niklas Carlsson, representa otra contribución significativa al campo de la valoración y predicción del rendimiento de los jugadores en el fútbol. Mediante la aplicación de la técnica de Random Forest, una metodología de aprendizaje automático que construye múltiples árboles de decisión para clasificar a los jugadores según su rendimiento, este estudio arroja luz sobre los atributos que distinguen a los jugadores de élite del resto.

Una de las conclusiones más destacadas del estudio es la variación en la precisión predictiva según la posición del jugador, siendo particularmente alta para los delanteros. Este hallazgo sugiere que las características y estadísticas que se utilizan para evaluar el rendimiento de los jugadores de ataque son más efectivas para predecir su éxito futuro en comparación con las empleadas para otras posiciones. Por lo tanto, para alcanzar niveles similares de precisión en la predicción del rendimiento de jugadores defensivos u ocupantes de otras posiciones en el campo, sería necesario desarrollar y aplicar métricas más sofisticadas o específicas.

Además, la predicción sobre la evolución de un jugador es igualmente crucial para la administración de su valor en el mercado. Los clubes que logran descubrir y nutrir talentos emergentes se posicionan para obtener beneficios significativos a medida que estos jugadores incrementan su cotización en el mercado. Este enfoque predictivo es también vital para el fomento del talento interno, posibilitando la implementación de programas de entrenamiento y desarrollo personalizados que se ajusten al potencial y expectativas de cada jugador.

Por último, para los clubes, en particular aquellos con recursos limitados, invertir en talentos con un margen de mejora notable se presenta como una estrategia sostenible para alcanzar la competitividad y la rentabilidad, permitiéndoles descubrir y desarrollar jugadores que posteriormente podrían ser traspasados por una suma superior.

En resumen, en este proyecto, se pretende obtener una métrica que mida el rendimiento de los jugadores, a partir de casi 200 columnas diferentes con información sobre cada partido de cada jugador. Esto servirá para poder distinguir entre el nivel de los jugadores y poder tener conclusiones sobre posibles fichajes de una manera mucho más rápida que el proceso manual utilizado en la actualidad basado en la visualización de vídeos. Posteriormente, se aplicará el *Machine Learning* y *Deep Learning* para su predicción para conocer el rendimiento futuro y modelar el rendimiento de los distintos jugadores.

2. Objetivos

Este proyecto enfatiza la relevancia de apoyar tanto el aspecto futbolístico como el financiero de los equipos de fútbol profesional. Para lograr esto, se ha definido un objetivo principal, el cual se ha hecho alcanzable a través de la realización de varios objetivos específicos.

2.1 Objetivo Principal

El núcleo de este proyecto radica en emplear datos de rendimiento deportivo para revolucionar los métodos de búsqueda de talentos. La meta es transformar cómo se predicen futuras contrataciones, agilizando el proceso de identificación de los candidatos más prometedores para cada rol, y reduciendo el tiempo que los cazatalentos invierten en la observación presencial de los partidos. Así, se busca concentrar la atención en aquellos jugadores que, de acuerdo con los algoritmos, destacan por sus cualidades deportivas, su compatibilidad con el equipo o la capacidad económica del club para nuevos fichajes. Por tanto, se propone integrar una herramienta para favorecer las estrategias de *scouting* reduciendo el trabajo a la creación de una métrica interpretable que, aprovechando datos deportivos complementados con información adicional fuera del terreno de juego, ayude a los clubes a elegir al jugador perfecto para sus necesidades. Este enfoque no solo tiene el propósito de mejorar el desempeño en el campo, sino también de aumentar los beneficios económicos del club.

Este indicador del rendimiento de un jugador en un partido no solo será fundamental para la toma de decisiones en materia de fichajes, sino también para prever el desempeño futuro del jugador en próximos partidos. De esta manera, se contribuye a reducir la incertidumbre a la que el equipo técnico puede enfrentarse al momento de planificar un encuentro en un contexto determinado. Este método también hace posible detectar posibles debilidades o fortalezas en los equipos rivales mediante el análisis de las actuaciones individuales de sus jugadores.

2.2 Objetivo Particulares

1. Vincular datos tácticos individuales con otras bases de datos que brinden valor económico.

Las bases de datos que agrupan estadísticas de partidos jugados por un futbolista brindan un valor restringido al análisis puro de los eventos en el terreno de juego. No obstante, al combinar estos datos con información adicional sobre el valor de mercado, influencia en redes sociales, nacionalidad, y otros factores, se puede incrementar notablemente su valor. Esta perspectiva más amplia permite valorar el impacto de un jugador más allá de su rendimiento en el campo, introduciendo nuevas facetas para su evaluación.

2. Desarrollo del PES, una métrica diseñada para medir el valor futbolístico de un jugador, independientemente de su valor monetario.

El reto que enfrenta el *Performance Evaluation Score* o PES es la falta de una referencia etiquetada; es decir, no contamos con datos históricos que faciliten estimar el rendimiento futuro de un jugador, lo que dificulta el desarrollo de un modelo supervisado. En este contexto, la habilidad para crear una nueva valoración a través de la creatividad y el dominio de técnicas estadísticas emerge como un factor clave para el éxito del proyecto.

3. Predicción del PES del siguiente partido.

Construir un modelo predictivo que disminuya el margen de error en la estimación del PES se destaca como un elemento fundamental del proyecto, con el propósito de reducir la ambigüedad en las decisiones estratégicas de los cuerpos técnicos. Este avance constituye un paso importante en el campo deportivo, donde prever el rendimiento se considera un desafío complejo debido a la gran cantidad de variables en juego.

4. Presentación clara y directa de los resultados alcanzados.

En el ámbito futbolístico, es habitual encontrar individuos que no están familiarizados con el análisis de datos o la estadística, lo que señala una carencia de formación en estas áreas. Por ello, resulta esencial simplificar la información a compartir para garantizar una comunicación efectiva, asegurando que todas las partes involucradas utilicen un lenguaje común y comprensible.

5. Identificación de jugadores y sus rachas

El resultado final del proyecto deberá concluir en una demostración de su utilidad práctica, por un lado, con su aplicabilidad en el *scouting* mediante un *dashboard* que permita localizar a los mejores jugadores en función del PES, y ajustando los jugadores según la capacidad económica del club que pretende fichar. Por otro lado, gracias al PES predicho se pretende hacer una comparación con el PES real para identificar momento en los que el jugador rinde por encima de lo que se espera y momentos que rinde por debajo.

3. Posibles conjuntos de datos

El siguiente apartado pretende introducir las posibles opciones de datos sobre rendimiento deportivo o valor económico de los jugadores de fútbol. En este punto veremos los motivos principales por los que hemos decidido manejar *Statsbomb* y *Transfermarkt* como las dos fuentes principales del proyecto.

3.1 Opta

La primera base de datos candidata es Opta, una empresa que se especializa en la recopilación, análisis y distribución de datos estadísticos relacionados con el deporte, particularmente conocida por su trabajo en el fútbol, aunque también cubre otros deportes como el rugby, cricket, y más[13].

Opta consta de una página web con datos generales como: asistencias, centros, faltas, intercepciones, pases, tiros, tiros a portería, paradas, fueras de juego, tarjetas, goles y entradas. En total 12 características que nos ayudan a comprender un partido, pero que en un estudio profundo se quedan en escasez de información. Por otro lado, Opta ofrece mejores bases de datos con un mayor número de variables e información, pero mediante su correspondiente pago.

3.2 Fbref

Fbref es una página web que contiene datos sobre rendimiento deportivo de jugadores. Su ventaja principal es que es una fuente de datos abierta, a la cuál podemos acceder mediante la realización de técnicas como *web scraping* que permitirían extraer la información de la web para posteriormente almacenarla y analizarla. Otra de las ventajas es que por cada jugador tenemos hasta 140 características sobre el rendimiento en el campo, y esto se debe a que la gran parte de datos provienen de Opta (versión de pago).

No obstante, tras una extracción de estos datos se puede ver que destaca la presencia de ruido y equivocaciones en su contenido, dificultando su análisis y agregando importancia al tiempo de limpieza, la cual no asegura obtener unos datos finales limpios con calidad. Estos datos pueden contener errores, inconsistencias e incluso mismos jugadores con diferentes nombres[14].

3.3 Transfermarkt

Transfermarkt es un sitio web de fútbol especializado en información relacionada con transferencias de jugadores, valores de mercado y otros datos generales. Se trata de las pocas páginas web que manejan esta información, siendo esta la más conocida y fiable. Los datos que se encuentran dentro pueden estar introducidos o bien de manera automática o bien gracias a la contribución de la comunidad de usuarios de pago de la plataforma. Es por esto, que la

fiabilidad no es del 100%, pero aun así sigue siendo la mejor opción disponible en el mercado, la cual utilizan la gran mayoría de equipos y agentes de la élite del fútbol.

En concreto, y para fines relacionados con *scouting*, esta base de datos nos aportará dos métricas clave: el valor de mercado y la fecha de finalización de contrato. Estas dos enriquecerán los análisis y permitirán ayudar a ajustar los fichajes en base a presupuesto y necesidades[15].

3.4 Skillcorner

Por otro lado, *Skillcorner* no son datos ni de rendimiento deportivo, ni económicos. En este caso, se trata de datos biocondicionales tales como la distancia recorrida, velocidad máxima, esfuerzos de alta intensidad medidos por distancia dada una cierta velocidad, potencia metabólica y posicionamiento de cada jugador en el campo en cada momento del partido lo que ayuda a entender porque zonas se suele mover un jugador.

Toda esta información se encuentra disponible para una gran variedad de competiciones. No obstante, se trata de un *dataset* privado al cual se puede acceder mediante un pago a la compañía. Estos datos servirían como un complemento ideal que podría mejorar la creación del *Performace Evaluation Score* o índice que usaremos para medir el rendimiento deportivo de un jugador basándonos en las características que tengamos a disposición. Es decir, estaríamos teniendo en cuenta no sólo como actúa un jugador cuando tiene el balón, sino también sus condiciones físicas[16].

Finalmente, por motivos económicos no se utilizará esta base de datos en este proyecto.

3.5 Statsbomb

Por último, *Statsbomb* es una base de datos de rendimiento deportivo. Al igual que *Opta* y *Skillcorner*, se trata de un dataset privado. No obstante, gracias a la aportación del Levante UD de este conjunto de datos, podrá ser utilizado para el proyecto. Este contiene más de 180 características sobre el rendimiento de cada jugador en cada partido. La información no contiene errores de formato en el contenido de las diferentes columnas por lo que ahorra mucho esfuerzo en la parte de limpieza. Además, respecto a la cobertura, se tiene información de 30 ligas y unos 17.000 jugadores, lo cual es una opción que admite una gran variedad de datos de diferentes competiciones[17].

Toda la información sobre las características de los datos de Statsbomb se encuentra en su *data layout* adjunto como documento .pdf con nombre ‘Data Layout Player Match Stats.pdf’.

4. Descripción de la base de datos

Comprender en profundidad las variables de nuestra base de datos es esencial, dado que solo deben incluirse en el análisis aquellas con una justificación clara para su uso. La base de datos final integra información de Transfermarkt y de una base de rendimiento deportivo, Statsbomb. Esta fusión ha resultado en un conjunto de datos de **193** columnas y **1.400.000** filas, representando a jugadores en cada uno de los partidos.

El conjunto de datos utilizado en este trabajo de investigación abarca una amplia gama de variables que capturan diferentes aspectos del rendimiento de los jugadores durante un partido. Entre las variables más destacadas se encuentra *player_match_goals*, que cuantifica el número de goles marcados por un jugador en un partido, proporcionando una medida directa de su efectividad ofensiva. Junto a esta, la variable *player_match_assists* refleja el número de asistencias realizadas, es decir, los pases que resultan en un gol, lo que permite evaluar la capacidad del jugador para crear oportunidades de gol para sus compañeros. Por otro lado, *player_match_xG* (Expected Goals) estima la probabilidad de que un disparo se convierta en gol basado en factores como la distancia y el ángulo del tiro, ofreciendo una métrica más avanzada para evaluar la calidad de las oportunidades de gol generadas.

En el ámbito defensivo, la variable *player_match_pressures* cuenta la cantidad de veces que un jugador presiona a un oponente, una medida clave para entender su contribución a la recuperación del balón y la interrupción de las jugadas del rival. Complementariamente, *player_match_ball_recoveries* registra el número de veces que un jugador recupera la posesión del balón, lo que es crucial para analizar su efectividad en tareas defensivas.

También se incluye *player_match_tackles*, que mide la cantidad de entradas exitosas, y *player_match_interceptions*, que contabiliza las veces que un jugador intercepta un pase del oponente. En cuanto al juego aéreo, *player_match_aerials* indica el número de duelos aéreos en los que un jugador participa, mientras que *player_match_successful_aerials* refleja la cantidad de esos duelos que resultan en éxito. Finalmente, la variable *player_match_passes* contabiliza el número total de pases realizados por un jugador, siendo fundamental para evaluar su participación en la construcción del juego.

Para conocer en detalle las columnas y su descripción se presenta adjunto el documento “Data Layout Player Match Stats.pdf” el cual se puede consultar a modo informativo para comprender todas las variables de la base de datos.

Respecto al conjunto de datos ofrecido por Statsbomb, es esencial resaltar su carácter temporal, debido a que cada fila representa el rendimiento de un jugador dividido en 193 características en un partido en concreto situado en un momento del tiempo. Por lo general, es comúnmente conocido el hecho de que el rendimiento del siguiente partido de un jugador puede estar relacionado con su rendimiento pasado, en especial el más reciente.

En cuanto a los datos de Transfermarkt, se han recabado detalles como el nombre del jugador, fecha de nacimiento, edad, posición, pierna dominante, club actual, fecha de finalización del contrato, valor de mercado, agente, redes sociales, y la marca patrocinadora. Aunque esta información es valiosa para análisis que trascienden el *scouting*, este proyecto se enfoca principalmente en dos de ellas: el valor de mercado y la fecha de término del contrato. Estas se emplean como filtros para identificar jugadores relevantes en el mercado de transferencias.

La información sobre la marca patrocinadora, agente y redes sociales se presentan como un elemento de interés para estudios futuros que superen el alcance de este Trabajo de Fin de Máster. En la sección destinada a investigaciones futuras, se explorarán los posibles usos de estos datos.

Para el análisis realizado, se han considerado 17.000 jugadores, partiendo de un total inicial de 10.827 obtenidos mediante el uso un algoritmo capaz de relacionar nombres diferentes basándose en conceptos de similitud como la distancia de Levhenstein. Los 6 mil restantes, se añadieron manualmente mediante la creación de un *Data Quality Control* para la identificación de jugadores nuevos que fueron apareciendo en el tiempo de desarrollo del proyecto. Este proceso se realizó con una frecuencia mensual para distribuir los esfuerzos a lo largo del tiempo, ya que una actualización manual de 400 jugadores cada mes aproximadamente podía llevar alrededor de 3-4 horas de trabajo. Estos ajustes manuales permitieron pasar de una cobertura del 67.2% de los jugadores a un 100%, que, pese a su costo por esfuerzos, su beneficio es una base de datos de referencia de jugadores que ayuda a enriquecer los datos finales con un mayor alcance de jugadores y conclusiones totalmente representativas.

Aplicar el algoritmo basado en distancias no siempre funcionaba, dado que en ambas bases de datos existen muchos jugadores que pueden tener nombres muy distintos. Es por este motivo que para completar la cobertura de jugadores se optó por una actualización manual en la que no sólo aumenta el número de jugadores, sino también la credibilidad en los datos, ya que los enlaces serán al 100% correctos para los 17.000 jugadores.

Las ligas a las cuales se tiene acceso, incluyendo 30 ligas en total, con las cinco grandes ligas europeas: La Liga, Premier League, Serie A, Ligue 1 y Bundesliga. Es importante añadir que se incluyen 4 ligas de futbol femenino: D1 Arkema, Frauen Bundesliga, Liga Femenina, Serie A Women.

En la base de datos, cada fila representa a un jugador y cada columna, a sus atributos. La mayoría de estas columnas muestran el rendimiento del jugador, calculado como el valor absoluto de cada métrica con una granularidad por partido. Es decir, para cada partido de cada jugador se han recogido sus estadísticas exactamente como han ocurrido en el terreno de juego, sin transformaciones adicionales que dificulten su interpretación.

Las variables que no miden cuantitativamente el rendimiento del jugador se utilizan para añadir valor a través de su empleo como filtros después del análisis preliminar. Un ejemplo destacado de estas variables es la fecha de finalización del contrato del jugador, la cual se transforma en un criterio importante para la selección en etapas posteriores del proceso analítico.

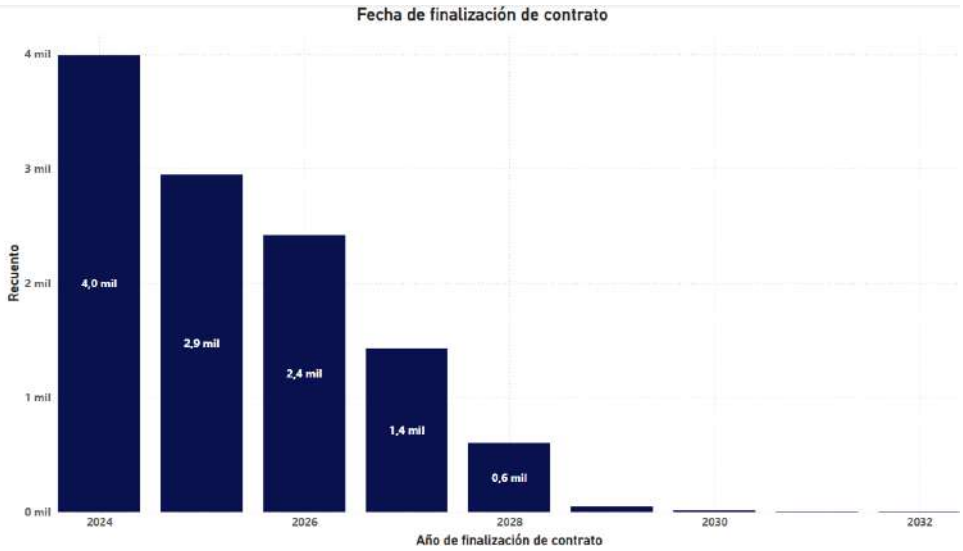

Figura 2. *Distribución de jugadores en función de su fecha de finalización de contrato.*
Fuente: Elaboración propia a partir de datos de Transfermarkt.

En la imagen que estamos analizando, se observa que la mayoría de los futbolistas tienen contratos próximos a expirar en los siguientes dos años, con especial atención a aquellos que concluyen el mismo año. Estos jugadores se convertirán en agentes libres, lo cual significa que cualquier club interesado podría ficharlos sin necesidad de pagar una cláusula de rescisión. Además, al no estar atados a un contrato, estos jugadores tienen la oportunidad de negociar salarios potencialmente más altos en comparación con situaciones donde el nuevo club debe compensar económicamente al anterior. Esto representa un beneficio tanto para el jugador como para el club interesado.

Otro aspecto relevante es el valor de mercado de cada jugador, que ayuda a los clubes a seleccionar y ajustar sus incorporaciones de acuerdo con su presupuesto, permitiéndoles captar los mejores talentos disponibles basándose en su desempeño actual.

Figura 3. *Distribución de jugadores en función de su valor mercado.*
Fuente: Elaboración propia a partir de datos de la base de datos original con proveedor confidencial.

La figura 3 representa una asimétrica donde la mayoría de los jugadores tiene un valor de mercado de entre 0 y 3 millones de euros. La ventaja de esta situación es que hay una gran cantidad de jugadores económicos disponibles para ser contratados. Además, el *Performance Evaluation Score* desarrollado ayudará en la selección, asegurando que los jugadores más prometedores sean considerados.

Además, será crucial tener un conocimiento profundo de cada jugador y realizar un análisis detallado, ya que, en el fútbol, los datos sirven como apoyo en la toma de decisiones, pero el seguimiento y análisis en profundidad son esenciales una vez que un jugador ha captado la atención del equipo.

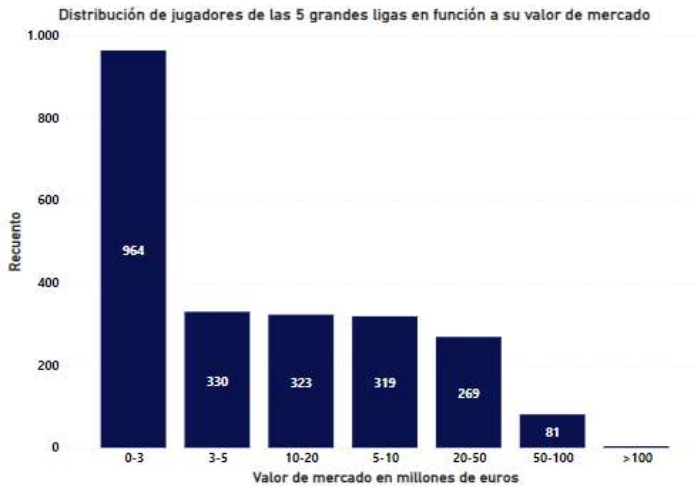

Figura 4. *Distribución de jugadores de las 5 grandes ligas en función de su valor mercado.*
Fuente: Elaboración propia a partir de datos de la base de datos original con proveedor confidencial.

En la figura 4, que revisa las cinco ligas más importantes de fútbol, las diferencias entre los grupos de jugadores no son tan marcadas. Aun así, existen algunas excepciones notables, como

aquellos que se encuentran en la parte derecha, cuya posición representa en su totalidad a jugadores de ataque que sobresalen por su habilidad excepcional para marcar goles o generar jugadas. Entre ellos destacan figuras como Jude Bellingham, Erling Haaland, Kylian Mbappé y Vinicius Junior, cuyos valores de mercado superan los 100 millones de euros. Por otro lado, aquellos jugadores que acumulan menos minutos en el campo suelen ubicarse hacia el extremo izquierdo de la gráfica. Esto punto destaca la importancia de los goles y los minutos en el precio de mercado.

5. Metodología I: Obtención del Performance Evaluation Score

Este segmento explica la metodología que se utilizaron en el proyecto hasta la obtención del PES, detallando el procedimiento para conseguirlo, y cómo esta se vincula con los objetivos planteados. Durante la fase de desarrollo, se usó Python[18] tanto para la extracción de datos como para la automatización de todo el proceso, incluyendo en este transformaciones y enriquecimiento de la BBDD.

Como se ilustra en la figura 5, el proceso inicia con una consulta a la API del proveedor de datos de rendimiento, de donde se obtiene la base de datos X que incluye variables deportivas esenciales para evaluar las acciones en un partido de fútbol.

También se emplearon técnicas de *Web Scraping* a Transfermarkt[19] para recopilar información económica, como el valor de mercado de los jugadores y su situación contractual, almacenándose en la base de datos Y.

Las dificultades para integrar las bases de datos X e Y, especialmente debido a las diferencias en la escritura de los nombres de los jugadores, llevaron a la creación de una base de datos de referencia o *Players Reference Data*. Esta base contiene cinco columnas: una con el nombre del jugador obtenido de X, su identificador en la misma y otra con el identificador correspondiente en Y (Transfermarkt), la fecha de nacimiento del jugador y la fecha en que ese jugador fue añadido a la PRD, *Players Reference Data*. Este trabajo que une una opción automática con una manual garantiza la calidad y exactitud del vínculo entre ambas bases, proporcionando datos confiables y precisos para el análisis.

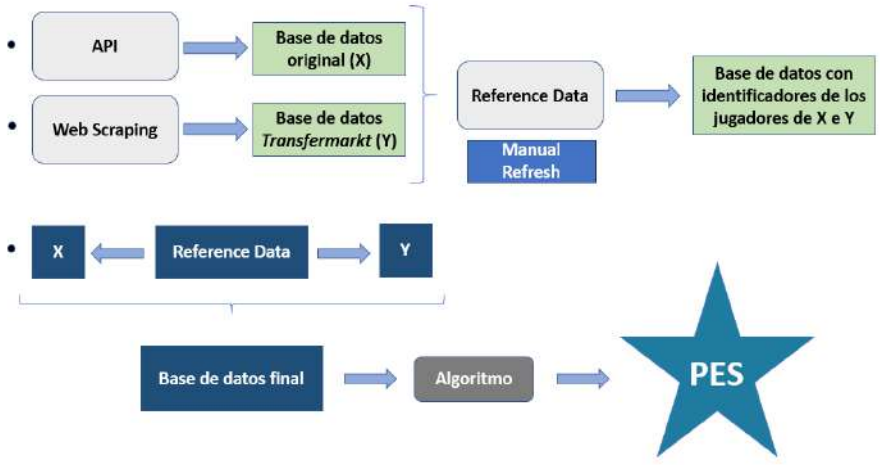

Figura 5. *Mapa conceptual sobre el proceso de extracción del PES.*
Fuente: Elaboración propia.

La base de datos definitiva se crea mediante la integración de las bases X e Y, usando como enlaces los identificadores de cada jugador para cada una de las dos bases de datos, rendimiento deportivo y valor económico. La base de datos de referencia ayuda en este proceso al identificar qué jugador de la base de datos de rendimiento deportivo corresponde en la base de datos extraída mediante *Web Scraping* a Transfermarkt. Los detalles específicos que fundamentan el uso de esta metodología se explicarán más adelante.

Para finalizar, se diseñó un algoritmo utilizando PCA para establecer el PES. Este PES o índice de rendimiento deportivo, que varía entre 1 y 100, refleja el desempeño del jugador en el campo y proporciona una métrica cuantitativa para evaluar su contribución y habilidades deportivas.

5.1 API request

Para recoger los datos, se estableció una conexión con la API de Statsbomb utilizando Python como lenguaje de programación. A través de la función GET, que requería una URL y credenciales específicas para la autenticación del usuario, se accedió inicialmente al listado de ligas disponibles.

No obstante, el interés principal era obtener información sobre los jugadores. Esta se recopiló usando la función GET con una URL adaptada, que seguía un patrón constante y solo variaba según la competición y la temporada específicas. Para este estudio, se eligieron todas las competiciones y temporadas disponibles, utilizando un bucle para recorrer los valores de las diferentes competiciones obtenidas anteriormente. En cada iteración, se añadían las estadísticas de los jugadores para cada partido de cada competición a una base de datos unificada que representase un histórico.

Al finalizar este proceso, se creó un conjunto de datos que incluye aproximadamente 17 mil jugadores de todo el mundo, mostrando diversos aspectos de su rendimiento en el campo, en un total de 1’4 millones de filas.

5.2 Web Scrapping

El fútbol va más allá de ser simplemente un deporte; su alcance e impacto lo han convertido en una importante fuente de ingresos. Los clubes no solo aspiran a sobresalir deportivamente, sino también a aumentar sus beneficios económicos.

Una estrategia fundamental para alcanzar estos objetivos es el mercado de fichajes, donde la compra y venta de jugadores juega un papel crucial. En este contexto, se recomienda que los equipos adquieran jugadores cuyo valor de mercado pueda aumentar a lo largo del tiempo.

Para llevar a cabo análisis efectivos en esta área, es importante contar con información que exceda el desempeño en el campo. Resulta esencial añadir a la base de datos original datos adicionales como el valor de mercado actual del jugador y la fecha en que finaliza su contrato y puede ser adquirido sin costo de transferencia.

El uso de técnicas de *Web Scraping* es fundamental aquí, permitiendo la recolección de datos de fuentes públicas en internet. De manera particular, Transfermarkt, un sitio web alemán especializado en monitorear y analizar las transacciones del mercado de fichajes, se destaca como una fuente destacada. Este portal ofrece información actualizada y detallada sobre el valor de mercado de los jugadores, sus historiales de transferencias y estadísticas de rendimiento.

Así, Transfermarkt se ha consolidado como una herramienta indispensable no solo para profesionales del sector como agentes y directores deportivos, sino también para periodistas y aficionados al fútbol.

En el proyecto actual, se ha utilizado el lenguaje de programación Python para extraer datos de la página web de Transfermarkt. De esta manera, se crearon 10 notebooks para paralelizar el trabajo y reducir en tiempos de extracción. En cada uno de ellos, se utilizan tres URLs correspondientes a tres ligas distintas. Para cada una de estas tres se iterará obteniendo el enlace de cada equipo integrante, y dentro de cada equipo, el enlace correspondiente a cada jugador de ese equipo. Siendo este último, el lugar desde donde se extraen los datos que incluyen tres columnas clave entre el resto de las columnas almacenadas: identificador del jugador, valor de mercado y fecha de finalización de contrato

El resultado final se añade a una base de datos histórica donde cada extracción está marcada con su correspondiente fecha. Esto permitirá coger los datos más actualizados a la hora de realizar el cruce con la base de datos de rendimiento, ya que como se ha dicho la intención es crear una estructura de código reproducible que permita tener todo el proyecto automatizado para aportar el beneficio de que las conclusiones estén basadas en la información más actualizada posible.

Además, en última instancia se aumenta la seguridad de los datos mediante un sistema de copias por si la ejecución fuera errónea tener un *backup* de la versión anterior.

5.3 Reference data

El uso de una base de datos de referencia es fundamental en este proyecto debido a la necesidad de integrar dos bases de datos distintas: una con atributos deportivos de los jugadores y otra con información sobre su valor de mercado obtenida de Transfermarkt. La integración de estas bases enfrenta desafíos, especialmente debido a las diferencias en la forma en que se registran los nombres de los jugadores en cada una, complicando así su fusión directa.

Para resolver esta dificultad, en primer lugar, se normalizan los nombres de ambas bases de datos, eliminando cualquier carácter especial y se utiliza *.lower()* para dejarlos en minúsculas. En segundo lugar, se hace lo mismo para el nombre del equipo. Además, se crean tres columnas nuevas: primera letra del nombre, nombre y última palabra dentro del nombre completo, que podrá variar más, siendo esta el último apellido, o el primero, incluso el nombre si no tiene apellidos registrados. A todo esto, se añade la información de la fecha de nacimiento. Después de conseguir estas variables en la base de datos de rendimiento deportivo y económico, tenemos ya la información necesaria para construir el algoritmo que permitirá encontrar jugadores de una base de datos en la otra.

El primer paso del algoritmo consiste en fijar los valores que coincidan de fechas de nacimiento y primera letra del nombre. Este paso deberá coincidir en la gran mayoría de casos incluyendo la opción correcta. El segundo paso, consiste en emplear la distancia de *Levenshtein*, útil para comparación de cadenas de texto, para el nombre, último apellido y nombre del equipo, de manera que si tienen más de un 60% de similitud entrarán dentro de candidatos potenciales. Este umbral se decidió tras observar que un valor menor proporcionaba enlaces incorrectos, y un valor mayor enlazaba menos jugadores, por lo que el 60% se estableció como un umbral cercano al óptimo en cuanto al beneficio que aporta la relación calidad-cantidad.

En tercer lugar, se mantienen aquellos potenciales jugadores que la similitud sea superior al umbral en como mínimo dos de las tres variables en las que se ha empleado la distancia de *Levenshtein.* El cuarto paso es la creación de un *Score* de similitud basado en la suma de similitud de las dos variables más importantes a coincidir y con menor variabilidad, es decir, nombre del jugador y nombre del equipo. El último paso, consistirá en ordenar a los jugadores potenciales a cruzar de mayor a menor *Score* y quedarse con la primera opción.

Este algoritmo consiguió enlazar 8.900 jugadores de 10.700 iniciales. Los nombres restantes se buscaron y añadieron manualmente. Posteriormente, conforme pasaba el tiempo esta base de datos se fue actualizando para tener la información más reciente posible, añadiendo los nuevos jugadores manualmente debido a su menor volumen.

Estos procesos manuales no solo garantizan la calidad y precisión de los datos, sino que también aportan un valor añadido significativo. Muchas empresas están dispuestas a invertir tiempo y recursos en estas labores manuales para mejorar la credibilidad y fiabilidad de los datos que utilizan, reconociendo que tal inversión puede diferenciarlos notablemente en sus respectivos campos.

Cabe añadir, que se hicieron pruebas directamente cogiendo todo el nombre y buscando similitud, incluso fijando fecha de nacimiento o nombre de equipo. No obstante, en estos tres algoritmos probados o el enlace era de pocos jugadores, para los últimos dos casos, o la calidad del cruce no era buena e incluía información incorrecta de un jugador a otro. Por lo que, la creación de este algoritmo ha requerido de investigación sumado a la prueba y error, dada la

dificultad que supone encontrar el jugador correcto a unir en ambas bases de datos cuyos nombres de jugadores y equipos han podido ser escritos de maneras distintas.

Nombre base de datos original	Nombre de Transfermarkt
Messi	Leo Messi
Lamine Yamal	Lamine
Cristiano	Crisitano Ronaldo
Ferran Torres	F. Torres

Tabla 2. *Ejemplo de la base de datos de referencia.*
Fuente: Elaboración propia a partir de datos de *Transfermarkt* y datos de la base de datos original con proveedor confidencial.

Como se detalla en la tabla 2, los nombres de los jugadores pueden variar desde el número de apellidos utilizados hasta el uso de puntos o incluso apodos de los futbolistas. Es aquí, en la variedad de opciones donde reside la dificultad de encontrar calidad en los enlaces sin perder a muchos jugadores por el camino, es decir, la relación "calidad-cantidad" que se ha mencionado anteriormente.

Una vez aplicado el algoritmo y encontrados los jugadores, se creó una base de datos de referencia llamada *Players Reference Data* que incluye las columnas: nombre del jugador, fecha de nacimiento, identificador del jugador en la BBDD de rendimiento deportivo e identificador de Transfermarkt.

El paso final fue fusionar ambas bases de datos utilizando como puntos de enlace las columnas con los identificadores de los jugadores de las dos bases de datos. Gracias a la *Players Reference Data*, donde antes podíamos hacer análisis por separados sobre rendimiento deportivo y datos económicos, ahora se puede hacer todo desde un mismo sitio o BBDD.

Aunque este proceso fue meticuloso y laborioso, los resultados obtenidos han sido fructíferos, proporcionando dos datos críticos valorados en el ámbito del scouting deportivo: el valor de mercado del jugador y la fecha en que finaliza su contrato. Estos detalles son clave para la toma de decisiones en la gestión de equipos y en las estrategias de mercado, ofreciendo una ventaja competitiva significativa en la evaluación y adquisición de talentos.

Además, desde el inicio de este proyecto con la creación de esta base de datos de referencia hasta su finalización, se programó un *Data Quality Control* para la identificación de jugadores nuevos en la BBDD original que no estaban presentes en la *Players Reference Data*. De esta manera, con un inicio de 10.700 jugadores, se terminó el proyecto con 17.000. Esto proporciona mayor cobertura, que a su vez garantiza que los resultados obtenidos estén actualizados. Los datos llegan en este ámbito, donde los ojos no pueden llegar, ya que no es posible ver a los 17.000 jugadores y mucho menos en cada partido. En cambio, los datos ofrecen esta ventaja que

proporciona un mayor alcance. Este es el motivo por el que la implementación de este control de calidad se propuso como un detalle que podría mejorar tanto el proceso y su infraestructura, como los resultados obtenidos del análisis.

5.4 Aprendizaje no supervisado

El aprendizaje no supervisado es fundamental cuando se trata de analizar conjuntos de datos sin etiquetas predefinidas, permitiendo descubrir patrones y estructuras ocultas sin tener conocimientos previos específicos sobre la organización de los datos. Esto implica que el algoritmo tiene que investigar y analizar los datos de forma independiente para identificar grupos, relaciones o tendencias.

En el contexto de este proyecto, el objetivo principal es desarrollar una métrica de evaluación, y dado que no disponemos de etiquetas preexistentes para guiar un modelo predictivo, el aprendizaje no supervisado se vuelve esencial para explorar y entender los datos. La técnica seleccionada para este propósito es el Análisis de Componentes Principales[20].

El PCA es particularmente útil en este proyecto porque permite reducir la dimensionalidad de los datos mientras se conserva la mayor cantidad de información posible. Esto se logra identificando las direcciones, o componentes principales, en las que los datos varían más. Al aplicar PCA, podemos simplificar la complejidad de los datos sin perder atributos significativos, facilitando así la creación de una métrica robusta que refleje de manera efectiva el rendimiento y el potencial de los jugadores basándose en sus características observadas.

5.5 Proceso de extracción del *PES*

Tras integrar diversas bases de datos en una única BBDD, el próximo paso es cargar esta información consolidada en un *dataframe* para iniciar el análisis. El objetivo es desarrollar un Score que refleje la calidad de cada jugador en una escala del 1 al 100, al que llamaremos PES. Este *Performance Evaluation Score* se determinará considerando aspectos como la liga en la que juega el jugador, los minutos jugados y su posición en el campo, para la formación de grupos de jugadores, y métricas numéricas sobre el rendimiento como pases o tiros para su construcción

Para llevar a cabo este análisis, se empleará Python, un lenguaje de programación destacado por su eficiencia en el manejo de grandes volúmenes de datos y por ofrecer una amplia gama de bibliotecas analíticas y estadísticas. El proceso incluirá la extracción de variables clave que permitan diferenciar significativamente a los jugadores dentro de la base de datos. A partir de estas variables, se desarrollará un modelo que asignará un valor numérico a cada jugador, facilitando así su comparación.

Este valor numérico, que variará entre 1 y 100, servirá para clasificar a los jugadores de mejor a peor basándose en sus atributos deportivos. Además, el uso de diversos filtros permitirá que los equipos personalicen la herramienta según sus necesidades específicas, obteniendo una lista ordenada de los jugadores que mejor se ajusten a sus estrategias y requisitos tácticos.

Es fundamental considerar que, para alcanzar una diferenciación óptima, es necesario combinar variables generales relevantes para todos los equipos con aquellas que sean críticas para estrategias particulares. Por ejemplo, un equipo que priorice el juego aéreo podría valorar más a los delanteros fuertes en duelos aéreos, mientras que uno enfocado en el contraataque podría preferir jugadores con alta velocidad y capacidad de transición rápida.

Aunque inicialmente la métrica se desarrollará sin ajustes específicos para estrategias de equipo particulares, el modelo podría personalizarse más adelante para adaptarse a los requisitos tácticos de cada club. Esto implicaría ajustar el PES con ponderaciones adicionales basadas en variables estratégicamente seleccionadas para reflejar las necesidades tácticas del equipo. Tal personalización podría convertirse en una característica valiosa si se decide comercializar esta herramienta a equipos profesionales, permitiendo una adaptación precisa y efectiva a sus estilos de juego específicos.

En este proyecto, nos enfrentamos al desafío de desarrollar una métrica de evaluación de jugadores sin contar con una variable predeterminada para predecir, lo que nos lleva hacia un enfoque de análisis no supervisado. La ausencia de una variable objetivo complica el desarrollo de un modelo predictivo convencional, ya que no disponemos de un estándar o referente externo con el cual comparar los resultados de manera directa. El reto, entonces, es construir una métrica objetiva y útil sin esos puntos de referencia preestablecidos.

Ante este escenario, hemos optado por emplear técnicas de aprendizaje no supervisado para explorar los datos y descubrir patrones subyacentes que puedan servir como base para la evaluación de jugadores. El PCA, ya mencionado, es clave en este proceso. No solo facilita la reducción de la dimensionalidad de los datos, lo que ayuda a visualizar y entender las estructuras complejas, sino que también permite identificar las características más significativas que influyen en el rendimiento de los jugadores.

Una opción podría haber sido basar el análisis en el valor de mercado del jugador, pero esto habría sesgado los resultados hacia los jugadores más costosos, suponiendo erróneamente que un mayor precio implica una mayor habilidad o rendimiento en el campo. En cambio, el enfoque que hemos adoptado busca evaluar a los jugadores basándose exclusivamente en sus estadísticas de rendimiento, evitando así prejuicios relacionados con su valor de mercado.

La estrategia, por tanto, consiste en desarrollar un modelo que utilice estadísticas de rendimiento para determinar de manera objetiva la calidad de un jugador. Reflexionar sobre cómo emplear PCA de manera innovadora para este fin es crucial, ya que nos permitirá utilizar esta técnica para extraer y ponderar efectivamente los atributos más relevantes de los jugadores,

resultando en una métrica que pueda ser ampliamente aceptada y útil para análisis comparativos y decisiones tácticas en el deporte.

5.6 Tratamiento de los datos

El proceso de preparación de datos en este proyecto está diseñado para garantizar que el conjunto de datos pueda ser aplicado como entrada al algoritmo de PCA. El objetivo de esta preparación es segmentar los datos en subconjuntos que reflejen grupos específicos basados en los minutos jugados, la posición en el campo y la liga en la que compite cada jugador. Esta segmentación es esencial porque permite comparaciones más justas al ajustar por variables que impactan directamente en el rendimiento, pudiendo ser estas diferentes dependiendo del conjunto seleccionado. Por ejemplo, no se deberían seleccionar las mismas variables entre un delantero y un defensa, ya que sus propósitos dentro del campo son distintos.

Debido a la imposibilidad de hacer fichajes de géneros diferentes trataremos a partir de ahora todo el proceso partiendo los datos entre aquellos correspondientes a fútbol femenino y aquellos que se refieren al masculino. Por este motivo, aunque el proceso sea idéntico para ambos casos, generaremos dos bases de datos para aplicarlo de manera separada en cada uno y tratar potencialmente con casos de usos particulares: *Scouting* en fútbol femenino y *Scouting* en fútbol masculino.

Dentro de estos se aplicará una segmentación por conjuntos de jugadores cuyas estadísticas y contexto tienen sentido para poder ser comparables.

Segmentación por subconjuntos:

- **Minutos jugados: Se** agrupará a los jugadores según la cantidad de tiempo que han estado en el campo, partiendo de la premisa de que más minutos en juego ofrecen más oportunidades para demostrar habilidades y acumular estadísticas relevantes.
- **Posición en el campo:** Dado que las estadísticas relevantes varían considerablemente entre las distintas posiciones, es crucial comparar a los jugadores dentro de categorías similares para obtener evaluaciones más equitativas.
- **Liga:** Las diferencias en el nivel de competencia o estilos de juego entre ligas pueden influir significativamente en el rendimiento, por lo que es prudente comparar a los jugadores dentro de contextos competitivos similares.

Como se acaba de destacar, es crucial estructurar los datos en subgrupos bien definidos para un análisis adecuado. Las diferencias en el nivel de juego significan que un logro como anotar 30 goles tiene distinta valoración dependiendo de la liga; lo que es un indicativo de excelencia en una liga superior puede no serlo en otra con un nivel competitivo menor. Por lo tanto, es fundamental comparar a los jugadores con sus pares en ligas de similar competitividad.

Este enfoque permite no solo un análisis más ajustado y equitativo, sino que también ayuda a mantener la coherencia y precisión en la evaluación del rendimiento, asegurando que las conclusiones derivadas sean sólidamente fundamentadas en el contexto competitivo de cada jugador. Este método de segmentación por ligas es esencial para el desarrollo de una métrica de evaluación fiable y aplicable a distintos entornos futbolísticos.

Respecto al tratamiento de variables podemos distinguir dos tipos:

- **Variables cuantitativas:** Se incluirán únicamente aquellas variables que reflejen directamente el rendimiento en el campo, como goles, asistencias, intervenciones para porteros, etc. Se excluirán variables como la edad o el valor de mercado, ya que no aportan directamente a una métrica basada en el rendimiento deportivo en un partido dado.
- **Variables categóricas:** Información como la pierna hábil, la competición y el equipo se considerará valiosa para filtrados en la herramienta interactiva final, pero no se incorporarán en el análisis PCA para prevenir distorsiones en la evaluación del rendimiento basada en habilidades puras.

Dado que una de las segmentaciones por grupo es la posición del jugador, y esta tiene una amplia gama de opciones por sus tecnicismos dependiendo de la formación inicial empleada por cada equipo, tiene sentido aplicar una simplificación de posiciones a aquellas generalmente conocidas con el objetivo facilitar el análisis comparativo y asegura que las evaluaciones sean más precisas y relevantes.

Este enfoque estructurado no solo mejora la calidad y la relevancia de los análisis realizados, sino que también asegura que la métrica final sea robusta y representativa del verdadero rendimiento deportivo de los jugadores, contextualizado según su situación específica. La visualización posterior podría explicar cómo se han categorizado y simplificado las posiciones para facilitar este análisis.

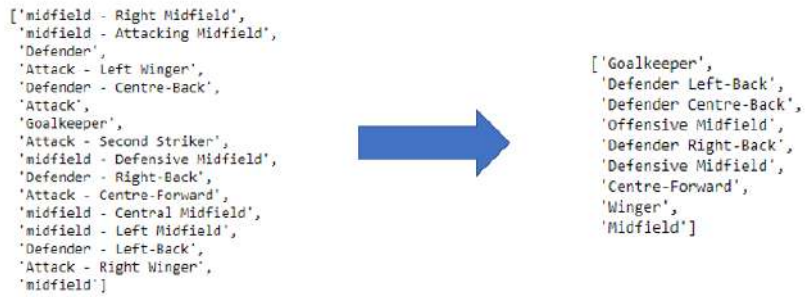

Figure 6. *Reducción del número de posiciones.*
Fuente: Elaboración propia a partir de datos de la base de datos original con proveedor confidencial.

De ahora en adelante, y para facilitar la exposición en las fórmulas y explicaciones, la base de datos final obtenida de la fusión entre la base de datos original y los datos de Transfermarkt será referida como la base de datos ***X***.

$$X = \bigcup \{\{l\} \times \{m\} \times \{p\} : l \in \{GL, PL\}, m \in \{MM, mM\}, p \in \{P, D, C, LI, LD, MC, MCO, E\}\}$$

Siendo:

- GL = grandes ligas, PL = ligas menos populares
- MM = jugadores con más minutos, mM = jugadores con menos minutos
- P = portero, D = delantero, C = central, LI = lateral izquierdo, LD = lateral derecho, MC = mediocentro, MCO = mediocentro ofensivo y E = extremo

Fórmula 1. *Representación de X mediante la unión de subconjuntos formados por todas las combinaciones entre los valores de las variables posición, minutos y competición.*

En la fórmula 1, se puede ver que se han creado tantos conjuntos de datos como posibles combinaciones entre los diferentes valores de las variables posición, competición y minutos. Inicialmente, se estableció un criterio basado en la cantidad de minutos jugados para hacer comparaciones entre jugadores. Se introdujo la variable 'umbral_minutos', definida como la cantidad mínima de minutos que un jugador debe haber jugado para ser considerado en las comparaciones. Comparar a un jugador que ha participado en un partido pocos minutos con otro que ha jugado todo el partido no es justo debido que este último ha podido obtener más valor en las estadísticas. Aunque se ajusten las estadísticas por minuto jugado, un jugador con una actuación excepcional en unos pocos minutos no puede ser considerado superior a otro cuyo rendimiento ha sido consistente a lo largo del partido. Por tanto, deberán compararse estadísticas que pertenezcan a cantidades de minutos similares.

Por esta razón, este umbral se usa para dividir a los jugadores en dos grupos: aquellos que han jugado una cantidad de minutos por encima del umbral y aquellos que no. El valor del umbral se calcula tomando el total de minutos jugados por el jugador más activo y dividiéndolo por tres. Así, los jugadores que no alcanzan este umbral se consideran con menos tiempo de juego, mientras que aquellos que lo superan tienen suficientes minutos acumulados como para hacer una comparación más equitativa y representativa.

$$Umbral\ minutos = \frac{Max\ (X\ player\ season\ minutes)}{3}$$

Fórmula 2. *Creación del umbral que divide separa a jugadores entre aquellos con menos y más minutos disputados.*

Se eligió un tercio del total de minutos jugados por el jugador con más minutos como umbral porque este valor representa en promedio a aquellos jugadores que, si bien no son titulares fijos, tienen participación regular en los partidos. Este criterio permite agrupar por una parte a los

jugadores que han jugado más y por otra aquellos que han jugado menos en el encuentro. Esto será importante, ya que en valor absoluto no tiene sentido comparar entre ambos grupos y estandarizar para que sean comparables elimina coherencia en el contexto futbolístico, donde los que juegan más suelen ser por su mejor rendimiento. Por estos motivos, el umbral quedó fijado en 67 minutos.

Además, como se mencionó anteriormente, se procederá a dividir el conjunto de datos en dos grupos distintos, uno para las ligas de alto prestigio y otro para las ligas de menor categoría. Esta segregación es esencial porque las estadísticas entre diferentes ligas no son directamente comparables, debido a las variaciones en el nivel de competencia. Por ejemplo, marcar 10 goles en la liga española no implica lo mismo que hacerlo en la liga argentina debido a que la primera tiene jugadores de mayor nivel que la segunda en la actualidad. Al subdividir los datos de esta forma, se garantiza que las comparaciones dentro de cada grupo sean válidas y pertinentes, añadiendo robustez y relevancia a las métricas finales que se derivan de estas comparaciones. Por otro lado, también se establece esta separación debido a que dependiendo de la liga puede ser habitual un estilo de juego u otro.

Se ha organizado que las cinco principales ligas europeas —Bundesliga, La Liga, Ligue 1, Premier League y Serie A— se agrupen en una categoría, mientras que las segundas divisiones y ligas de otros países, como la Liga SmartBank, Süper Lig, Primeira Liga, Serie B, Ekstraklasa, Super Liga, Eredivisie, Super League, Bundesliga 2, Jupiler Pro League, Ligue 2, Liga MX y HNL, formen otra. Este enfoque asegura que las comparaciones se realicen entre ligas de niveles de competencia similares, aumentando así la precisión y relevancia de los análisis estadísticos.

Adicionalmente, el análisis se segmentará por posición en el campo, de modo que los jugadores solo sean comparados con otros en la misma posición. Esta distinción es crucial porque el rendimiento y las estadísticas varían significativamente entre posiciones; por ejemplo, no sería adecuado comparar directamente a un delantero con un portero.

Este método permite que la métrica final sea específicamente aplicable y comparativa dentro de grupos homogéneos. Así, por ejemplo, podríamos identificar a los mejores laterales derechos de las principales ligas cogiendo aquellos partidos disputados por el jugador en los que haya jugado más de 67 minutos. Este mismo criterio se aplicará a los diferentes subconjuntos creados por competencia, posición, minutos.

En cuanto al uso del PCA, se ha optado por seleccionar las componentes que explican hasta el 80% de la variabilidad en los datos. Esto se hace para evitar la inclusión de demasiado ruido y asegurar que las componentes seleccionadas proporcionen una representación fiable y precisa del conjunto de datos original. Limitar el número de componentes principales a aquellas que expliquen el 80% de la variabilidad permite mantener un equilibrio óptimo entre la reducción de

la dimensionalidad y la retención de información significativa, lo cual es esencial para generar una métrica útil y representativa.

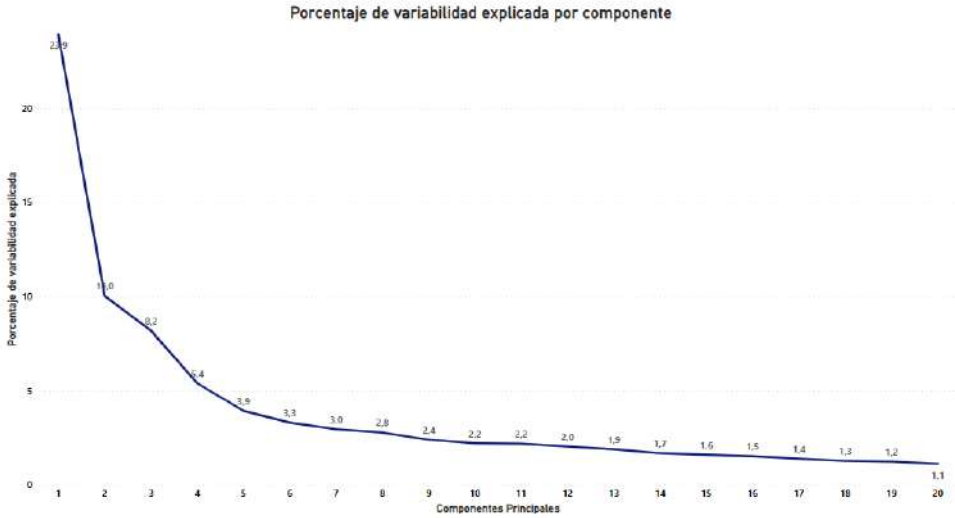

Figura 7. *Porcentaje de variabilidad explicada por las diferentes componentes principales*.
Fuente: Elaboración propia mediante el uso de Power BI.

Como se puede ver en la figura 7, las primeras cuatro componentes principales identificadas a través del PCA son esenciales para comprender la variabilidad en el conjunto de datos. Estas cuatro componentes contribuyen al 47% de la explicación de la variabilidad total. Después de estas, la contribución de las componentes subsecuentes tiende a estabilizarse, y cada una aporta progresivamente menos a la explicación de la variabilidad.

Dado que un 47% de variabilidad explicada podría ser insuficiente para captar adecuadamente la complejidad y la variabilidad inherente de los datos, se ha decidido incluir más componentes en el análisis hasta alcanzar al menos el 80% de la variabilidad explicada.

La elección de expandir la selección de componentes hasta cubrir el 80% de la variabilidad se basa en la necesidad de obtener una representación más completa y representativa del conjunto de datos, asegurando que no se omitan aspectos relevantes que podrían influir en las conclusiones. Este criterio de selección está claramente especificado en el paso 6 del Algoritmo 1, destacando su importancia en el proceso de análisis para garantizar que los resultados sean tanto confiables como útiles para la toma de decisiones y evaluaciones basadas en el *dataset*.

5.7 Algoritmo de creación del PES

El uso del Análisis de Componentes Principales como técnica de aprendizaje no supervisado es habitual en el ámbito deportivo, donde se manejan frecuentemente grandes volúmenes de datos multidimensionales. El PCA juega un papel crucial en la reducción de dimensiones, ayudando a eliminar ruido y simplificar los conjuntos de datos sin perder información esencial, lo cual facilita el manejo y análisis de los datos.

El principal objetivo del algoritmo que se implementará es asignar un peso específico a cada variable dentro del conjunto de datos. Este peso refleja la importancia relativa de cada variable en la explicación de la variabilidad observada. El proceso para determinar estos pesos está detallado en el paso 8 del *Algoritmo 1*. Estos pesos serán proporcionales, sumando un total de 100, y se utilizarán en la ponderación de los valores de cada variable, que estarán normalizados entre 0 y 1 [21]. Utilizando estos pesos en las variables correspondientes, según la *Fórmula 3*, se calcula el PES cuyo valor máximo es 100, como se especifica en la *Fórmula 5*.

El *Algoritmo 1* se explicará en detalle a continuación, desglosando cada paso y explicando el razonamiento detrás de cada decisión. Este enfoque paso a paso no solo aclara cómo funciona el algoritmo, sino también por qué se toman ciertas decisiones metodológicas y cómo estas influyen en los resultados finales.

Es importante subrayar que el *Algoritmo 1* se aplicará a cada uno de los subconjuntos de datos definidos por combinaciones de variables como la posición del jugador, los minutos jugados y la competición en la que participa. Estas divisiones aseguran que las comparaciones y evaluaciones sean justas y relevantes, pues los jugadores solo serán comparados dentro de sus respectivos subgrupos.

Previamente a la aplicación del primer paso, es fundamental estandarizar los datos, dado que, de no hacerlo, se estaría otorgando mayor peso a las columnas con valores absolutos más elevados, y en el análisis se parte del supuesto de que todas las variables deben ser iguales para la computadora. Este proceso es clave aplicarlo con anterioridad a la descomposición en componentes principales. Cabe añadir que este paso es diferente a la normalización de la fórmula 3, dado que en esta última los números quedan entre 0 y 1 para llegar a obtener un PES en un rango concreto, en este caso, el máximo será 100. No obstante, en la estandarización no existe esta restricción en el rango que pueden tomar los datos, sino que se resta el valor con la media y se divide entre la desviación típica, para que no existan variables que expliquen mayor varianza únicamente por su magnitud.

En primer lugar, se aplica PCA para extraer de su descomposición la matriz de pesos P (paso 1) y se coloca el valor absoluto de los pesos para considerar solo la magnitud, sin importar la dirección (paso 2). En términos de valor absoluto, cuanto mayor sean estos pesos dentro de una componente, más relevantes serán las variables dentro de ella. Es cierto que la dirección es importante para entender del tipo de jugador que se trata en un análisis exploratorio de los datos, pero en este caso dentro de cada posición no hay tanta diferencia entre las habilidades de los jugadores y en caso de haberla siempre se podría hacer una clasificación por tipo de jugador mediante un *clustering* y posteriormente comparar el PES creado a partir de sus variables más influyentes teniendo en cuenta para hacerlo tanto la dirección como la magnitud de los pesos en el PCA para ese tipo de jugador, pero esta opción se dejará como análisis futuro por la limitación temporal del proyecto.

Seguidamente, se fuerza a que las variables tengan el signo en función a su significado en el contexto futbolístico. De esta manera, las variables con sentido negativo como el número de pérdidas, tarjetas amarillas o tarjetas rojas, restarán en la *Fórmula 4* y el peso indicará con qué magnitud lo hace, mientras que los goles, asistencias y cualquier otra variable positiva en el fútbol, sumarán en la obtención del PES (paso 3). Por consiguiente, las variables tomarán el signo correspondiente a su naturaleza contextual.

Posteriormente, en la matriz U, para cada componente se almacena la suma de todos los pesos de las diferentes variables (paso 4) con el objetivo de utilizarlo para ser el denominador en el siguiente paso para encontrar la importancia de una variable en una componente.

A continuación, para cada suma de pesos de cada componente Uj, y para cada peso de cada variable en cada componente P"Ti,j, se divide el peso de la variable en la componente (P"Ti,j) por la suma total de los pesos de esa componente (Uj,), de modo que se obtenga una aproximación de la importancia de esa variable en la componente. La suma de estos valores para cada variable que se obtienen en cada componente es 1 (paso 5), siendo esto normal ya que se divide cada peso por la suma de todos los pesos de esa componente. En este punto, se tiene la matriz G de importancia de cada variable en la componente, mientras que lo que se busca es su relevancia en la base de datos.

Sin embargo, dado que no todas las componentes explican la misma variabilidad del dataframe, será necesario ponderar los pesos de cada variable según el porcentaje de variabilidad que explique cada componente, dando, de esta manera, más peso a aquellas que expliquen mayor varianza (paso 5).

Por ello, se obtiene el porcentaje de variabilidad de cada componente. No obstante, como la suma de estos porcentajes será 80 (porcentaje de varianza que se ha decidido explicar del dataset, como se explicó anteriormente) y se quiere que el PES sea entre 1 y 100, primero se relativizarán estos porcentajes manteniendo la proporción, pero la suma pasará a ser 100 (paso 6).

Asimismo, se multiplica el porcentaje de varianza que explica cada componente relativizado a 100, expresado mediante la matriz V, por el peso de cada variable en cada componente, matriz G. Obteniendo así lo que explica esa variable en función de la varianza que explica esa componente en el conjunto de datos, representado por la matriz Z. Con esto, se pasa de tener la importancia de una variable en una componente a tener la importancia de una variable en el conjunto de datos por cada componente (paso 7). Solo será necesario sumar estos valores para cada variable en cada componente y se obtendrá la importancia de cada variable en el conjunto de datos, representado mediante la matriz M (paso 8).

En este punto, dado que se ha cuantificado la importancia de la variable en la BBDD, ya se podría establecer un orden de variables más importantes, o, dicho de otra manera, se podrían

escoger aquellas que más contribuyen a diferenciar entre jugadores y, por consiguiente, altos valores en estas proporcionarán jugadores "buenos".

De esta manera, se ha tomado el concepto de variables con mayor capacidad de separación entre individuos, en este caso jugadores, extraída del PCA y se ha utilizado para, finalmente obtener un peso que represente la importancia de una variable en el *dataset*.

Finalmente, se almacenan las 10 variables más importantes, guardando tanto el nombre como su peso (*paso 9* presente en la *Complementación del algoritmo 1*).

ALGORITMO DE CREACIÓN DE IMPORTANCIA DE VARIABLE EN LA BASE DE DATOS

1) $X = T \cdot P^T + E$
2) P' = |P|. Valores absolutos de P.
3) $P_j'' = P_j' * (1$ si $j \in O, -1$ si $j \in N)$, siendo O el conjunto de variables con sentido futbolístico positivo y N el conjunto de variables con sentido futbolístico negativo. En P, el subíndice j representa las columnas, que, en este caso, sin las variables.
4) $U_{1,j} = \Sigma_{i \in \{1...N\}} (P'')^T_{ij}, \forall j \in \{1\dots m\}$. Al transponer P'', las columnas pasan a ser las componentes principales y para cada una de ellas se suman todos los pesos, almacenando finalmente, la suma total de pesos de variables por componente.
5) $G_{i,j} = ((P'')^T_{i,j} / U_j)*100, \forall i \in \{1\dots n\} \wedge j \in \{1\dots m\}$. La matriz G representa el porcentaje de importancia de cada variable en cada componente.
6) Siendo V una matriz que representa el porcentaje de variabilidad que explica cada componente en la base de datos, V' = (V*100/80)*100. De esta manera, el total de varianza explicada no es 80 (porcentaje de varianza explicada fijada en este caso), sino que la suma pasa a ser 100. Manteniendo la proporción dada la multiplicación por una constante, pero simplificando el proceso de extracción del *Score*. Esta matriz se obtiene de aplicar la función *pca.explained_variance_ratio_* del paquete de Python *sklearn.decomposition*.
7) $Z_{i,j} = G_{i,j} * V'_j, \forall i \in \{1\dots n\} \wedge j \in \{1\dots m\}$. Siendo G_{mxn}.Z el porcentaje de importancia de cada variable en cada componente ponderado por el porcentaje de variabilidad de cada componente relativizado.
8) $M_{1,j} = \Sigma_{i \in \{1...N\}} Z^T_{ij}, \forall j \in \{1\dots m\}$. M representa la importancia de cada variable en el dataset. Siendo M una matriz 1 x m.

Algoritmo 1. *Creación de los pesos de las variables más importantes de la base de datos.*

Fuera del algoritmo de creación de importancia de una variable en el *dataset*, se puede complementar con un noveno paso que permita seleccionar las *k* variables más importantes. En este caso se fija k = 10, para evitar introducir demasiadas variables y que, en lugar de aportar, introduzcan ruido. No obstante, este valor podrá variar y se deja como elección propia. Podría estudiarse un método que minimizase el error del PES en función del número de variables escogidas, pero persiste la misma problemática sobre la inexistencia previa de este valor de rendimiento, por lo que no hay con que comparar para conocer si los resultados son mejores o peores y de manera manual se hace prácticamente inviable. Asimismo, se mantendrá el número de características seleccionadas en 10, dejando abierto su posible cambio en futuras ocasiones.

SELECCIÓN DE LAS K-VARIABLES MÁS IMPORTANTES

9) $B = \{ (v, p) \in M : x \geq f\_k(\{p' : (v', p') \in M\}) \}$. Siendo *f_k* una función que representa el K-ésimo valor más alto de M, en este caso k = 10. B representará las 10 variables más importantes del *dataset*. Además, se aclara que *v* representa el nombre de la variable y *p*, el peso de esta.

Complementación del algoritmo 1. *Paso añadido al algoritmo 1 para la selección de las k variables más importantes.*

Dada la dificultad del algoritmo, se ha decidido plantear el siguiente ejemplo para facilitar su comprensión:

TABLA INICIAL (X)

	Var 1	**Var 2**	**Var 3**	**Var 4**	**Var 5**
F1	1.2	-0.5	0.3	0.8	-1.1
F2	2.3	0.2	-0.7	1.5	0.4
F3	-0.9	1.3	-1.2	0.6	-0.8

1. DESCOMPOSICIÓN EN (X = T * P^T + E):

TABLA T	**Comp 1**	**Comp 2**
F1	0.5	-0.4
F2	1.2	0.8
F3	-0.3	0.6

TABLA P^T	Var 1	Var 2	Var 3	Var 4	Var 5
C1	0.7	-0.3	0.4	0.2	-0.5
C2	0.4	0.5	-0.6	0.7	0.1

2. VALORES ABSOLUTOS DE P (P' = abs(P)):

	Var 1	Var 2	Var 3	Var 4	Var 5
C1	0.7	0.3	0.4	0.2	0.5
C2	0.4	0.5	0.6	0.7	0.1

3. AJUSTE DE SIGNO EN P' SEGÚN EL CONTEXTO:

	Var 1	Var 2	Var 3	Var 4	Var 5
C1	0.7	0.3	-0.4	0.2	-0.5
C2	0.4	0.5	-0.6	0.7	-0.1

4. SUMA DE PESOS ABSOLUTOS POR COMPONENTE (U_j):

	Suma de Pesos
C1	2.1
C2	2.3

5. IMPORTANCIA PORCENTUAL DE CADA VARIABLE EN CADA COMPONENTE (G_ij):

	Var 1	Var 2	Var 3	Var 4	Var 5
C1	0.7/2.1 = 33.33	14.29	-19.05	9.52	-23.81
C2	17.39	21.74	-26.09	30.43	-4.35

6. VARIANZA AJUSTADA “SUMA TOTAL = 100” (V'):

	C1	C2
Varianza explicada	62.5	37.5

7. IMPORTANCIA PONDERADA DE CADA (Z_ij):

	Var 1	Var 2	Var 3	Var 4	Var 5
C1	(33.33x62.5)/100 =20.83	8.93	-11.91	5.95	-14.88
C2	6.52	8.15	-9.79	11.41	-1.63

8. IMPORTANCIA TOTAL POR VARIABLE (M_j):

	Importancia Total
Var 1	20.83+6.52=27.35
Var 2	17.08
Var 3	-21.70
Var 4	17.36
Var 5	-16.51

9. SELECCIÓN DE LAS 10 VARIABLES MÁS (B):

	Importancia Total
Var 1	20.83+6.52=27.35
Var 3	-21.70
Var 4	17.36
Var 2	17.08
Var 5	-16.51

En este caso, como en el ejemplo sólo hay 5 variables se seleccionarían todas. Una vez en el punto 9, ya se han obtenido los pesos de cada variable. El siguiente paso será normalizar los valores de los datos entre 0 y 1, con el objetivo de evitar que variables con mayor magnitud tengan mayor importancia.

Seguidamente, la *Fórmula 3* ha sido aplicada a cada uno de los subconjuntos formados por la combinación de las variables competición, posición y minutos jugados. Esto se ha decidido realizar de esta manera debido a que, si la fórmula fuese aplicada a todo el conjunto de datos, los valores muy altos afectarían a la interpretación de la mayoría de las métricas. Poniendo un ejemplo, si un jugador marca 30 goles en una liga no incluida dentro de las 5 más importantes y un jugador de la Premier League anota 25 goles, el primero tendría un valor superior al segundo. No obstante, si tomamos sólo los jugadores de las grandes ligas, es posible que el segundo jugador sea el máximo anotador y por tanto dentro de este subgrupo tendrá un valor de 1 como

resultado de haber aplicado la fórmula. De esta manera, las estadísticas sólo serán comparables dentro de su subconjunto correspondiente.

Normalización de los datos en un rango entre 0 y 1. Normalización min_max [24].

$$Xi,j = \frac{Xi,j - \min(Xj}{\max(Xj) - \min(Xj)}, \forall\, i \in \{1 \dots n\} \wedge\ j \in \{1 \dots m\}$$

Fórmula 3. *Normalización de los datos en un rango entre 0 y 1*

Posteriormente, al estandarizar los valores del conjunto de datos original entre 0 y 1, se podrá crear un modelo lineal que extraiga un *PES* entre 1 y 100, multiplicando el valor del jugador en cada variable (de 0 a 1) por su peso correspondiente.

El penúltimo paso será aplicar la *Fórmula 4*. Ésta consiste en la combinación de los pesos extraídos anteriormente en el *Algoritmo 2* con la base de datos con valores entre 0 y 1.

Finalmente, dado que hay pesos negativos el resultado de la suma de todas las ponderaciones se relativiza para que siempre sea un número interpretable entre 1 y 100, como se aprecia en la *Fórmula 5*.

FÓRMULA DE OBTENCIÓN DEL PES

$$PES = C_1X_1 + C_2X_2 + C_3X_3 + ... + C_{10}X_{10}$$

Fórmula 4. *Creación del PES.*

FÓRMULA DE OBTENCIÓN DEL PES FINAL (ENTRE 1 Y 100)

$$\text{PES FINAL} = \left(\frac{\text{PES Inicial} - \text{PES Inicial Mínimo}}{\text{PES Inicial Máximo} - \text{PES Inicial Mínimo}}\right) \times 99 + 1$$

Fórmula 5. *Creación del PES FINAL entre 1 y 100.*

El resultado es un *PES* entre el 1 y el 100 que mide el rendimiento de un jugador basado en las características más significativas en función de su posición, liga y minutos, siendo únicamente comparables jugadores de un mismo grupo, como se ha mencionado anteriormente.

Para entender mejor el funcionamiento, veamos dos ejemplos de grupos diferentes:

- Defensas de las grandes ligas con más minutos:

$$PES = C_{Ball\_Recoveries}X_{Ball\_Recoveries} + C_{Agressive\_Actions}X_{Agressive\_Actions} + C_{Crosses}X_{Crosses} + C_{Dribbled\_Past}X_{Dribbled\_Past} + C_{Dispossesions}X_{Dispossessions} + C_{Long\_Balls}X_{Long\_Balls} + C_{Pressures}X_{Pressures} + C_{Tackles}X_{Tackles} + C_{Xgbuildup}X_{Xgbuildup} + C_{Successful\_Aerials}X_{Successful\_Aerials}$$

Donde cada variable tiene su peso asignado y la suma de todos los pesos es 100 como se ha comentado previamente. Los pesos se sustituirán a continuación:

$$PES = 9{,}32X_{Ball\_Recoveries} + 10{,}51X_{Agressive\_Actions} + 9{,}22X_{Crosses} - 10{,}50X_{Dribbled\_Past} - 10{,}36X_{Dispossessions} + 9{,}19X_{Long\_Balls} + 10{,}85X_{Pressures} + 11{,}02X_{Tackles} + 9{,}03X_{Xgbuildup} + 10X_{Successful\_Aerials}$$

Dado que los valores de X han sido normalizados entre 0 y 1, como mucho la suma de las multiplicaciones entre pesos y valores puede llegar a ser 100 en el caso del partido perfecto.

Pongamos de ejemplo a un jugador, Ronal Araújo, perteneciente al FC Barcelona. El PES es creado por el rendimiento en cada partido, así que veremos el funcionamiento para crearlo seleccionando la jornada 22 de La Liga, que enfrentó al Barcelona contra el Villarreal y Araújo obtuvo una puntuación de 39,07 de PES.

$$19{,}86 = 9{,}32 \cdot 0{,}62 + 10{,}51 \cdot 0{,}21 + 9{,}22 \cdot 0{,}38 - 10{,}50 \cdot 0{,}56 - 10{,}36 \cdot 0{,}13 + 9{,}19 \cdot 0{,}27 + 10{,}85 \cdot 0{,}14 + 11{,}02 \cdot 0{,}45 + 9{,}03 \cdot 0{,}57 + 10 \cdot 0{,}15$$

Por último, se aplicaría la *Fórmula 5* para que todos los valores estén entre 1 y 100. Rellenando la fórmula con los respectivos valores:

$$\text{PES FINAL} = \left(\frac{19.87 - 10}{35.67 - 10}\right) \times 99 + 1$$

$$\text{PES FINAL} = 39{,}07$$

- Delanteros de las grandes ligas con más minutos:

Para este grupo el procedimiento es el mismo, el único cambio es que las variables seleccionadas serán diferentes, al igual que sus pesos asociados. En este caso el algoritmo seleccionó las siguientes características como distintivas en los delanteros de las grandes ligas con más minutos de 67.

$$PES = C_{Aerieals}X_{Aerials} + C_{Assists}X_{Assists} + C_{Dribbles\_Faced}X_{Dribbles\_Faced} + C_{Goals}X_{Goals} + C_{Key\_Passes}X_{Key\_Passes} + C_{Pressures}X_{Pressures} + C_{Np\_Shots}X_{Np\_Shots} + C_{Turnovers}X_{Turnovers} + C_{Xg}X_{Xg} + C_{Shot\_Touch\_Ratio}X_{Shot\_Touch\_Ratio}$$

Siguiendo el mismo proceder que en el anterior caso, cada peso Ci sería sustituido por un valor que indicaría la importancia de la variable. Posteriormente se multiplicaría cada peso con su correspondiente valor normalizado entre 0 y 1 y se sumarían para obtener el PES. En último lugar se aplicaría la normalización para obtener el PES entre 1 y 100.

5.7.1 Validación subjetiva del PES

El algoritmo para crear el PES mencionado previamente es una opción que permite automatizar el proceso y asignar un peso a cada variable. Además, cada peso será positivo o negativo dependiendo de su naturaleza. Por ejemplo, a más tarjetas amarillas peor rendimiento del jugador por el contexto negativo de esta característica. En cambio, a mayor número de goles, mejor será el rendimiento. Al final se reduce a que el PES es una suma ponderada de las habilidades, por lo que si tiene mejores valores en las habilidades tendrá mejor rendimiento. Es por esta razón que el PES sirve de indicador para medir el rendimiento.

No obstante, este tiene sus limitaciones ya que estamos cogiendo el valor absoluto de los pesos que obtenemos tras hacer el PCA y asignando variables con mayor importancia en función a la longitud que tiene este peso. De esta manera perdemos el sentido espacial que nos aportan los pesos, solo nos importa su peso en valor absoluto. Por tanto, se ve como alternativa utilizar un estudio más profundo en el que se consideren las variables en el espacio y se observen grupos de jugadores mediante *clustering*. Dentro de cada grupo tendrían más importancia las variables que lo representan en el espacio.

Otra alternativa es la asignación manual de pesos por posición. Es decir, mediante la comunicación con expertos sobre el tema determinaríamos que variables son importantes medir en un central, mediocentro o delantero y asignaríamos un peso manual a cada uno. Esto facilitaría la interpretación de los resultados y proporcionaría una dosis de fiabilidad en los expertos al entregar los resultados, ya que ellos mismos podrían entender por qué motivo un jugador se considera mejor o peor que otro.

No obstante, en el presente proyecto se optó por una opción automática en la que se ha intentado ser creativo a pesar del riesgo que conlleva. De todas maneras, el algoritmo no se trata de una caja negra ya que los pesos y las variables seleccionadas se almacenan para poder ser interpretables. El inconveniente es que las variables seleccionadas serán distintas dependiendo del grupo escogido en la combinación de competición, posición y minutos, ya que como se ha mencionado sólo podrán ser comparables jugadores de un mismo grupo, por ejemplo, delanteros con más minutos de las 5 grandes ligas. Esto genera una amplia combinación de grupos con sus propias variables seleccionadas, lo que dificulta su análisis, pero si fuera necesario entender por qué sale un resultado simplemente deberíamos cargar el archivo guardado con el nombre de las variables y los pesos asociados.

A pesar de estos aspectos negativos, la opción automática sigue siendo una opción válida dado que jugadores con mejores valores en las características tendrán mejor rendimiento y dado que cada posición tiene sus propias características asignando más peso a aquellas que ayudan más a diferenciar entre unos jugadores y otros. Además, observando las puntuaciones asignadas

estas tienen sentido con la realidad y por lo general los mejores jugadores que se encuentran en los mejores equipos se encuentran en las mejores posiciones en el ranking de rendimiento.

Jugador	Equipo	Precio	PES
Kevin De Bruyne	Manchester City	60 M	42’64
Florian Wirtz	Bayern Leverkusen	110 M	40’58
Jamal Musiala	Bayern Munich	110 M	39’58
Cole Palmer	Chelsea F.C	55 M	38’94
Martin Odegaard	Arsenal	95 M	38’12
Xavi Simons	RB Leipzig	80 M	36’54

Tabla 3: *Tabla de mediocentros con más minutos de las 5 grandes ligas ordenados por PES para la temporada 2023/24.*
Fuente: Elaboración propia.

En la tabla 3 vemos como los mejores mediocentros del mundo han aparecido en el top utilizando el algoritmo automático creado. Si bien Kevin De Bruyne, jugador del Manchester City lleva años siendo uno de los mejores mediocentros, aparecen otros nombres que han destacado esta temporada como Florian Wirtz con 21 años, el cual ha destacado esta temporada con un modesto equipo como el Bayern Leverkusen que ha logrado ganar la Bundesliga o liga alemana esta temporada 23/24, siendo esta su primera vez que ganan la competición en toda la historia. Por otro lado, jugadores como Cole Palmer, Musiala u Odegaard no asombran que aparezcan en esta lista dado que son los mediocentros titulares de unos de los mejores clubs del mundo, como es en este caso el Bayern de Munich, Chelsea F.C y Arsenal F.C. Por último, Xavi Simons a su temprana edad, ha dado el salto este año consolidándose como una pieza fundamental en el Leipzig, pasando este mismo año de valer 8 millones a tener un precio de mercado de 80 millones de euros, lo cual refleja la buena temporada del jugador holandés.

Además, si vemos el precio de estos jugadores también se puede ver como tienen un alto valor de mercado, pero esta variable ha estado excluida en la creación del PES ya que ésta debe de ser una métrica basada exclusivamente en el rendimiento deportivo. No obstante, en este caso nos sirve para ver que la percepción global sobre el precio de un jugador corresponde en cierta medida con la ordenación dada del PES, manteniendo como jugadores con alto rendimiento a jugadores considerados con un alto precio en el mercado.

A esta valoración subjetiva se puede añadir que los resultados han sido validados por profesionales del Levante Unión Deportiva, tales como Javi Navarro (Coordinador de metodología), Álvaro Máñez (Analista del Juvenil A) y José Gila (Científico de Datos del Primer Equipo). En la manera de proceder, se les prestó la aplicación interactiva, que mostraremos posteriormente, con los resultados. Después de buscar casos de uso concretos basados en 10 ligas diferentes (España, Inglaterra, Alemania, Italia, Francia, España II, Portugal,

Holanda, Bélgica, Francia II), los profesionales concluyeron que la herramienta era muy útil y sería capaz de cambiar el juego reduciendo mucho los tiempos de búsqueda de jugadores y siendo primer paso para conocer qué jugadores están rindiendo mejor esta temporada. Además, afirman que el PES muestra un punto de vista similar a sus percepciones como trabajadores de un equipo profesional de fútbol.

A pesar de ser un punto de vista subjetivo, el proyecto tiene a ellos mismos como *stakeholders* o personas de interés, lo que de alguna manera deja una sensación positiva de los resultados obtenidos. En el mundo ideal poder haber tenido una métrica de error hubiera facilitado mucho el trabajo, pero en este caso ningún modelo supervisado pudo ser utilizado para la obtención dada la inexistencia de esta variable.

6. Metodología II: Predicción del PES futuro basada en Deep Learning

Hasta este instante, se ha obtenido un valor que nos permite conocer el rendimiento de un jugador en cada partido. Para ello, se ha construido una metodología que permite de manera automatizada ir incluyendo nueva información deportiva como económica, uniendo ambas y creando el PES. Además, la base de datos de referencia, clave para unir *Statsbomb* con *Transfermarket*, tiene unos controles para advertir de nuevos jugadores que deberían de ser añadidos para tener el 100% de cobertura. Durante el desarrollo del proyecto iniciado en diciembre de 2023, se ha ido actualizando la información hasta el día de hoy para aportar información más reciente a los análisis, añadiendo valor de esta manera.

Una vez se tiene toda esta información disponible, el siguiente paso es investigar sobre la posibilidad de utilizar el PES y el resto de las características para predecir el rendimiento de un jugador en el próximo partido.

6.1 Introducción: Parte II

El propósito de obtener la predicción del rendimiento deportivo puede tener casos de uso como entender si un jugador rinde por encima o por debajo de lo esperado. Por otro lado, predecir el partido siguiente podría ser un primer paso para una futura investigación de una predicción más amplia de partidos, como, por ejemplo, diez partidos. Este rango más amplio podría ser usado para ver la tendencia de rendimiento de un jugador y ser un aspecto para considerar a la hora de hacer fichajes.

A pesar de que el PES ha sido construido mediante características como tiros, pases, asistencias, centros, duelos o recuperaciones, estas métricas no pueden ser consideradas en la predicción por un simple problema, no hay manera de saber cuál será el número de éstas en el próximo partido. La única manera sería construir una predicción para cada variable, pero esto implicaría incluir en el modelo final una sucesión de predicción con ciertos errores en cada una.

El método para seguir será utilizar las variables de contexto disponibles para ver si éstas mismas son suficientes para acercarnos a predecir el rendimiento. Como variables de contexto se pueden nombrar las siguientes: jugador, temporada, equipo, competición, fecha del partido, equipo local, equipo visitante, arbitro, edad, precio en el mercado y el mismo PES histórico.

No obstante, el precio de mercado se descartó desde un inicio debido a que se pretende encontrar el rendimiento del jugador por el contexto y no se debe ver influenciado por el precio, ya que de normales jugadores más caros juegan mejor, no obstante añadir esta variable podría incluir un sesgo. Por esta misma razón tampoco se incluyó en la generación del PES.

A partir de aquí, los diferentes modelos construidos se han basado en el método científico de prueba y error. En este, se han ido descartando las opciones que daban peores resultados e investigando de manera más profundas en aquellas que daban mejores. No obstante, se podría considerar que las pruebas han seguido dos caminos, uno en el que cada dato se considera independiente y la predicción se basa en la información de las características de esa fila, y, por otro lado, modelos que consideran que el rendimiento deportivo depende del tiempo y del rendimiento pasado, por lo que se plantearían como series temporales.

6.2 Plan de trabajo justificado

En primer lugar, los datos se han particionado en entrenamiento, validación y test, ya que se utilizarán redes neuronales y de esta manera mientras se entrena el modelo podemos seguir la evolución de la función de pérdida en el entrenamiento y la validación para observar posibles problemas de sobreajuste o falta de entrenamiento. Además, se ha utilizado un *threshold* temporal para partir los datos, de manera que se entrenan con datos hasta ese umbral y se guardan los datos en adelante a esa fecha para testear. De esta manera, en modelos basados en series temporales evitaremos entrenar con datos de todas las épocas y que esto cause que los modelos tienen buenos resultados, aunque engañosos.

Además, se partió de un modelo simple como una Regresión Lineal a modo comparativo con otros modelos. Por otro lado, se añadió la columna 'home_away', que añade la información sobre si el partido lo disputa el jugador como local o visitante.

En cuanto a la función de pérdida, se ha empleado el *mean squared error,* pero como métrica se ha añadido el *mean absolute error* que ayudará a interpretar los resultados en las mismas unidades que el problema, es decir en las del PES, del 1 al 100. También, se ha añadido un *early stopping* para que cuando los resultados en validación no mejoren durante un número de épocas marcado se pare el proceso.

En el punto '8. Resultados Obtenidos' se explicará en mayor detalle cada modelo, tanto la arquitectura como los datos pasados como input si han sido modificados o no. Junto a esta descripción se comentarán los resultados y que aspectos han sido significativos para mejorarlos.

6.2.1 Modelos utilizados

En el contexto del fútbol moderno, donde la ventaja competitiva puede ser mínima, la incorporación de tecnología y análisis avanzado se perfila como un diferenciador clave. Es por esto por lo que en este proyecto que se abordará el problema de la predicción de rendimiento futuro utilizando la inteligencia artificial, en concreto, modelos de aprendizaje profundo, modelos de *machine learning* como el *random forest* y modelos especializados en las series temporales como las redes convolucionales o las *LSTM*.

Además, adoptar la regresión lineal como paso inicial antes de proceder con modelos más complejos ofrece ventajas significativas. Al aplicar inicialmente un modelo simple, es posible detectar patrones, tendencias y posibles anomalías en los datos, elementos cruciales para una primera comprensión de las interacciones entre las variables. En este contexto, la regresión lineal no solo sirve como un punto de partida metodológico, sino también como una herramienta de comparación para evaluar las mejoras en la precisión y eficacia de los modelos predictivos más avanzados.

Por otro lado, en el caso del *Random Forest* en particular, se ha decidido incluir cómo método de comparar con otros modelos más avanzados que la regresión lineal y que no fuesen redes neuronales.

Además, en el contexto de predecir una variable numérica a partir de múltiples variables numéricas, emplear capas densas y *autoencoders* puede ser beneficioso. Las capas densas facilitan la modelización de relaciones complejas y no lineales entre las características, mientras que los *autoencoders* pueden ser empleados para destilar la información más relevante de los datos de entrada, minimizando el ruido y mejorando la calidad de las predicciones. Al combinar ambos enfoques, se optimiza la capacidad del modelo para captar y utilizar las sutilezas de los datos en la predicción de resultados numéricos, mejorando así la precisión y la eficacia del modelo final.

En último lugar, en la modelización de series temporales, como la evolución del rendimiento deportivo de jugadores a lo largo del tiempo, las Redes Neuronales Recurrentes (RNR), y específicamente las *Long Short-Term Memory* (LSTM), se destacan por su capacidad para gestionar secuencias de datos. Las redes LSTM están diseñadas para recordar información durante períodos prolongados, lo cual es fundamental al tratar con secuencias donde el estado anterior tiene influencia en el estado futuro. Este tipo de red es ideal para predecir el rendimiento futuro de los jugadores en partidos venideros, dado que puede capturar patrones a lo largo del tiempo, como tendencias y ciclos de forma y rendimiento, permitiendo una predicción más precisa basada en el historial detallado de cada jugador. Además de que por conocimiento general el jugador suele estar más influido por su rendimiento cercano, pero también se ve afectado, aunque en menor proporción, por lo que ha podido ocurrir anteriormente.

Por otra parte, las Redes Neuronales Convolucionales (CNN) son principalmente reconocidas por su aplicación en el procesamiento de imágenes, pero también pueden ser adaptadas para analizar series temporales. En el contexto del rendimiento deportivo, una CNN puede ser utilizada para examinar segmentos o ventanas temporales del desempeño de los jugadores, identificando patrones significativos que podrían pasar desapercibidos. Al aplicar filtros convolucionales a estos datos, se pueden extraer características importantes de las secuencias de rendimiento que podrían ayudar a predecir el rendimiento futuro.

7. Resultados obtenidos

A continuación, se va a mostrar el proceso seguido, enseñando los resultados obtenidos de cada modelo creado mediante el MAE o Mean Absolute Error, el cual hace más interpretable los resultados, ya que el error se encuentra en las mismas unidades que el problema presentado.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Donde:

- n es el número de observaciones.
- y_i representa el valor real de la i-ésima observación.
- $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima observación.
- $|y_i - \hat{y}_i|$ es la diferencia absoluta entre el valor real y el valor predicho.

Fórmula 6. *Obtención del Mean Absolute Error.*

Además, se mostrarán unas pinceladas de la arquitectura y de los principales cambios añadidos a cada modelo o a los datos de entrada.

En primer lugar, se construyó un modelo simple (M1) a modo comparativo, es decir, una regresión lineal con *One Hot Encoding* aplicado a las variables categóricas. Se obtuvo un MAE de 9’96, siendo este el punto del que se parte.

No obstante, había variables con un número muy grande de categorías, por ejemplo, el nombre del jugador con 17.000 opciones. Esto generaba una cantidad de datos muy grande que podía interferir en peores resultados. Como solución, en el M2, las variables categóricas, se sustituyeron por el promedio del PES por cada temporada por cada valor disponible en cada variable, y también se agregó la desviación estándar, pasando, por ejemplo, de 17.000 columnas a 2, en el caso de los nombres, el promedio y la desviación típica consiguiendo estas métricas mediante la agrupación por cada valor diferente y temporada. De esta manera, y con una regresión lineal, el MAE pasó a ser 7’7, una mejora sustancial que se convierte en el nuevo punto de partida.

Por otro lado, cambiar la función de activación de relu a lineal empeoró ligeramente los resultados, como ocurre en M3. No obstante, en M4, al añadir más capas con más neuronas mejoró ligeramente los resultados, y en M6, donde se continuaron añadiendo varias capas más ocurrió lo mismo, mostrando que mejoró la capacidad de entender las estructuras internas de los datos por parte del modelo, reduciendo el MAE hasta un 7’63.

Hasta el momento los datos de partida comprendían los siguientes aspectos: competición, local o visitante, media del PES por jugador y temporada y su desviación típica, media del PES por equipo y temporada y su desviación típica, edad, media del PES del equipo rival por temporada y su desviación típica y el promedio y desviación típica del árbitro del partido. En M5, se optó por añadir la variable posición, pero los resultados no mejoraron. Esto se debe a que para crear el PES se creó por posición, lo que hace que no por ser una posición u otra el rendimiento va a ser menor o mayor.

Además, se optó por probar una combinación de funciones de activación sigmoid con relu, que no ayudaron a mejorar los resultados, probablemente porque las unidades que se están manejando van del 1 al 100 y la sigmoid hace que estas dimensiones cambien, aunque posteriormente la relu las vuelva a cambiar a semejantes al problema en cuestión.

Modelo	Arquitectura			OPT	Novedades	MAE
	Capa	Neuronas	F. Act.			
M1	1ª	1	Relu	ADAM	Con one hot encoding	9.96
M2	1ª	1	Relu	ADAM	Sin one hot encoding. Siguientes modelos siempre sin one hot encoding.	7.70
M3	1ª	1	Linear	ADAM	F. activación = Linear	7.75
M4	1ª	30	Relu	ADAM	Número de capas Número de neuronas	7.67
	2ª	20	Relu			
	3ª	1	Relu			
M5	1ª	30	Relu	ADAM	INPUT: Añadimos la variable posición	7.67
	2ª	20	Relu			
	3ª	1	Relu			
M6	1ª	80	Relu	ADAM	Número de capas Número de neuronas	7.63
	2ª	50	Relu			
	3ª	30	Relu			
	4ª	20	Relu			
	5ª	1	Relu			
M7	1ª	80	Relu	ADAM	Combinación de sigmoid y relu en la función de activación	7.67
	2ª	50	Relu			
	3ª	30	Sigmoid			
	4ª	20	Sigmoid			
	5ª	1	Relu			

Tabla 4: *Resultados de los modelos de inteligencia artificial junto a una breve descripción de su arquitectura.*

Otra opción fue añadir *dropout*, donde se empezó por valores de 0'2 o 0'3 que resultaron en un MAE muy alto, por lo que finalmente se redujo a 0'1 y se consiguió un MAE de 7'89, empeorando los resultados de partida. Esto se puede deber a que en casos donde el sobreajuste no es un problema significativo, agregar *dropout* podría impedir que la red capte la estructura subyacente de los datos. Además, en tareas donde los datos contienen cierto ruido o la señal es débil, quitar información adicional puede ser contraproducente.

Modelo	Arquitectura			OPT	Novedades	MAE
	Capa	Neuronas	F. Act.			
M8	1ª	80	Relu	ADAM	Dropout. Más epochs en el entrenamiento.	7.89
	2ª Dropout=0'1	50	Relu			
	4ª Dropout=0'1	30	Relu			
	6ª	20	Relu			
	7ª	1	Relu			

Tabla 5: *Resultados de los modelos de inteligencia artificial junto a una breve descripción de su arquitectura.*

Otras pruebas en M9 fueron cambiar el *batch size*, añadiendo más información y pasando de 1000 a 5000 filas por *batch*, las cuales no tuvieron buenos resultados. Con un *batch size* más grande, el gradiente calculado en cada paso de optimización es más representativo del conjunto de datos completo. Esto podría sonar como una ventaja, pero en realidad puede conducir a un aprendizaje que no generaliza bien a datos no vistos, dado que los pequeños *batch sizes*

permiten más ruido en el proceso de entrenamiento, lo cual puede ayudar a escapar de mínimos locales en la función de pérdida. Además, con batch sizes grandes, el descenso de gradiente puede no ser tan efectivo para encontrar áreas prometedoras en el paisaje de la función de pérdida debido a la suavización excesiva del gradiente. También, el proceso de entrenamiento puede volverse más estable (menos variabilidad en la actualización de los parámetros). Esta estabilidad podría mejorar la capacidad de explorar el espacio de parámetros de manera efectiva, aunque esto podría llevar a converger a mínimos locales.

El cambio significativo se produjo en M10, donde se implementó una arquitectura que en primer lugar comprimía y descomprimía la información con un *autoencoder*, y en segundo lugar se añadían dos capas para predecir el rendimiento del siguiente partido. El MAE se bajó a un 6’06. Al aprender a comprimir y luego descomprimir los datos de entrada, los *autoencoders* pueden ser menos propensos al sobreajuste en comparación con redes densas que simplemente aprenden a mapear entradas a salidas sin considerar la estructura interna de los datos. Además, han podido aprender a ignorar el ruido y centrarse en los aspectos más estables y relevantes en los datos.

Modelo	**Arquitectura**			**OPT**	**Novedades**	**MAE**
	Capa	**Neuronas**	**F. Act.**			
M9	1ª	80	Relu	ADAM	Mayor batch size: De 1000 a 5000	7.65
	2ª	50	Relu			
	3ª	30	Relu			
	4ª	20	Relu			
	5ª	1	Relu			
M10	1ª	32	Relu	ADAM	Batch size: 1000 Nueva arquitectura con una primera parte que comprime y descomprime la información	6.06
	2ª	16	Relu			
	3ª	8	Relu			
	4ª	16	Relu			
	5ª	32	Relu			
	6ª	Input dim	Relu			
	7ª	1	Relu			

Tabla 6: *Resultados de los modelos de inteligencia artificial junto a una breve descripción de su arquitectura.*

Dado que los mejores resultados han sido con la arquitectura de *autoencoder* más capas densas, se investigó en profundidad en esta línea. Es por esto por lo que del modelo 11 al 20 se muestran diferentes pruebas, algunas con mejores resultados y otras que los empeoraban drásticamente como M15 (optimizador= SGD) y M20 (haciendo SMOTE de filas con PES mayor a 50 por su escasez), con un MAE de 9’63 y 7’35 respectivamente. En la imagen de la izquierda se puede ver previamente al SMOTE, donde el modelo no predice un PES (eje Y: PES predicho, eje X: PES real) superior a 50. El añadir datos superiores a este PES se hace para que el modelo sea capaz de captar mejor las razones que hacen que un jugador tenga un rendimiento extraordinario. No obstante, se observa que la predicción en rendimientos inferiores a 50 empeora y que aquellos superiores a 50 no es capaz de diferenciarlos bien, prediciendo como

máximo ahora alrededor de un 60 de PES. Esto demuestra que los rendimientos extraordinarios son raros de ver y difíciles de predecir. A pesar de esto y por la distribución de los datos, vemos que un jugador se considera que hace un buen partido (comparado con el resto de información) cuando su PES supera un 40 aproximadamente.

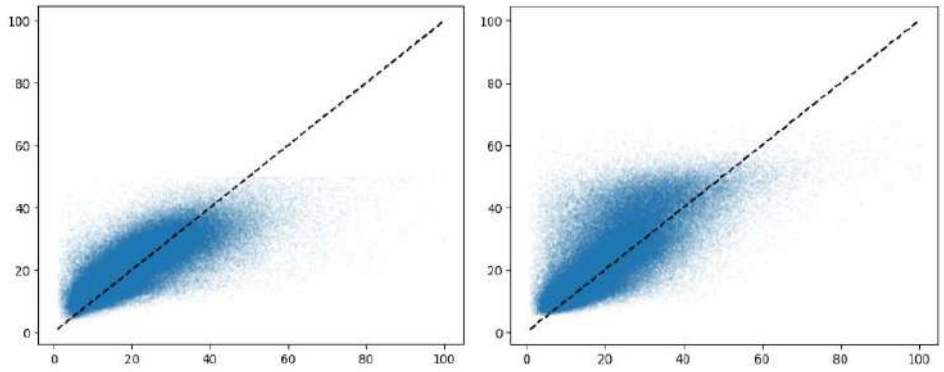

Figure 8: *PES predicho vs PES real antes (izquierda) y después (derecha) de hacer SMOTE.*
Fuente: Elaboración propia.

Otras pruebas como diferente número de neuronas, de capas, cambios de *learning rate* y regularización L1 y L2, obtienen resultados similares al 6'06 conseguido con M10. Cabe añadir que en M18 y M19 los resultados se mejoran a causa de una agregación de columnas que muestran información del rendimiento en partidos previos, siendo el mejor resultado un 6'02 de MAE cuando se añaden cinco columnas con la información del PES de cada jugador en los últimos cinco partidos. Esto muestra que puede haber una relación entre el rendimiento pasado con el del siguiente partido, y es por esto por lo que los siguientes modelos estarán relacionados con series temporales, como, por ejemplo, LSTM o redes convolucionales.

Modelo	Novedad	MAE
M11	Diferente número de neuronas	6.06
M12	Número de capas Número de neuronas	6.09
M13	Cambio de learning rate = 0'1	9.63
M14	Cambio de learning rate = 0'001	6.10
M15	Optimizador = SGD	9.65
M16	Regularzación L1	6.08
M17	Regularización L2	6.09
M18	Variable partido anterior añadida	6.04
M19	5 variables de los últimos 5 partidos añadidas No tiene orden temporal	6.02
M20	Oversampling por encima de 50 (SMOTE)	7.35

Tabla 7: *Resultados de los modelos de inteligencia artificial junto a una breve descripción de su arquitectura.*

Para construir los modelos de series temporales se utilizó una parte densa capaz de entender el resto de las variables como ser local o visitante o la competición entre otras, y, por otro lado,

una parte o bien convolucional o bien LSTM por tal de capturar el rendimiento en el tiempo de los jugadores.

En estos modelos también se han realizado diferentes pruebas para encontrar la manera de mejorar el error. A continuación, se mostrarán únicamente las pruebas más relevantes. Por ejemplo, en el modelo M21 se puede observar esta primera parte de *autoencoder* utilizada en modelos previos y otra parte formada por capas convolucionales capaces de captar patrones en series temporales. Además, los resultados son parecidos, con un MAE de 6’08.

En los modelos M21 y M22, se ha implementado una arquitectura híbrida con el objetivo de integrar tanto la información temporal como no temporal en un solo sistema de predicción. La estructura del modelo se divide en dos componentes principales: una red densa y una red LSTM, cada una encargada de procesar diferentes tipos de datos. Las capas densas están diseñadas para manejar características no secuenciales, tales como la localía o el tipo de rival, extrayendo patrones relevantes a través de una serie de transformaciones no lineales. Por otro lado, la red LSTM se encarga de procesar la serie temporal, capturando la dependencia secuencial de variables como la evolución del PES. Posteriormente, las salidas de ambas redes se concatenan y se introducen en un conjunto final de capas densas para realizar la predicción definitiva. Este diseño permite que el modelo considere tanto la evolución temporal de las variables como las características estáticas, integrando de esta manera ambas fuentes de información.

La elección de esta arquitectura se justifica por la necesidad de combinar diferentes tipos de datos que, por naturaleza, requieren tratamientos específicos. Mientras que las capas densas abordan las relaciones entre variables que no siguen un orden temporal, las LSTM son adecuadas para modelar secuencias temporales, manteniendo intacta la estructura de los datos a lo largo del tiempo. Al concatenar las salidas de ambos submodelos, se logra una representación conjunta que enriquece la capacidad predictiva del modelo, permitiendo que este considere tanto los aspectos secuenciales como los no secuenciales en su análisis.

Modelo	Parte	Capa	Neuronas	MAE
M21	Densa	1	64	6’08
	Densa	2	32	
	Densa	3	16	
	Densa	4	32	
	Densa	5	64	
	Densa	6	1	
	Convolucional	1	10	
	Convolucional	2	15	
	Convolucional	3	20	
	-	Flatten	-	
	-	Concatenate(Densa(6),Flatten)	-	
	Final	Densa	1	

Tabla 8: *Resultados de los modelos de inteligencia artificial junto a una breve descripción de su arquitectura.*

De la misma manera, se ha probado una estructura similar cambiando las capas convolucionales por capas *Long-Short Term Memory*. Los resultados no muestran cambios significativos, aunque haya mejorada unas décimas el error, siendo 6’06 el MAE. Por otro lado, en el modelo 23, con la misma estructura del M22 se añadió el rendimiento de los 20 partidos previos al que se pretende predecir, de manera que se obtiene una serie temporal más amplia. En este caso, los resultados mejoran hasta un MAE de 6’03. En el contexto futbolístico el rendimiento actual de los últimos cinco partidos suele ser más significativo, pero tener una visión amplia del rendimiento en un período más extenso puede ayudar a comprender mejor el nivel deportivo del jugador. Además, las LSTM pueden ser capaces de captar esta importancia a corto plazo, pero sin dejar de lado el rendimiento del jugador en partidos más lejanos.

Modelo	Parte	Capa	Neuronas	MAE
M22	Densa	1	64	6’06
	Densa	2	32	
	Densa	3	16	
	Densa	4	32	
	Densa	5	64	
	Densa	6	1	
	LSTM	1	10	
	LSTM	2	15	
	LSTM	3	20	
	-	Flatten	-	
	-	Concatenate (densa(6), flatten)	-	
	Final	Densa	1	

Tabla 9: *Resultados de los modelos de inteligencia artificial junto a una breve descripción de su arquitectura.*

Por otro lado, dado que la función de activación hasta el momento ha sido la RELU el modelo podría predecir valores más grandes que 100, y el PES se encuentra entre 1 y 100, el siguiente modelo (M24) tomará como variable respuesta un PES transformado a valores entre el rango 0 y 1. A partir de aquí, la estructura es la misma que M23 pero con una función de activación *sigmoid* para la salida del modelo. El MAE en este caso es de 0’061, lo cual indica que el modelo funciona de forma parecida a M23 pero dividiendo los resultados entre 100. De todas maneras, sí que es cierto que de esta manera se aseguraría que el PES predicho no supere el valor máximo, aunque debería de añadirse un step adicional para convertir las predicciones a números en el rango de 1 a 100. Por estas razones y, porque en imágenes anteriores se ha visto que el modelo no predice prácticamente valores por encima de 60, se seguirá trabajando con la misma magnitud del problema original. Además, como se observa en la *figure 9* tiene sentido ya que el PES se distribuye normalmente con una media cerca del 25, con un mínimo de 1 y un máximo de 60, el resto de los datos superiores a 60 parecen ser partidos extraordinarios de ciertos jugadores, lo cual ocurre en una proporción mínima y ninguno llega al 100.

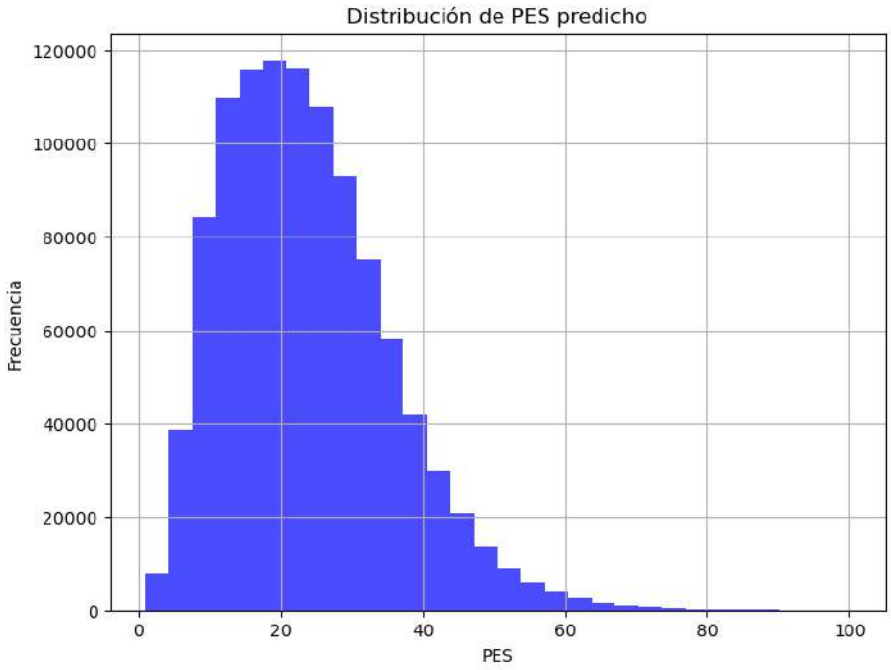

Figure 9: *Distribución del PES predicho.*
Fuente: Elaboración propia.

Además, con el propósito de probar un modelo de *machine learning* que no sean redes neuronales, se decidió trabajar con un *Random Forest*. En este caso (M25), ha obtenido un MAE de 6’31. Esto demuestra que no sólo los modelos de *Deep Learning* pueden ser útiles en este problema. No obstante, dado que el mejor rendimiento se ha obtenido con M19 y M23, pero que M23 añade la componente de captar las series temporales y en el contexto futbolístico tiene sentido conocer la evolución del rendimiento de un jugador como serie temporal, se ha decidido que el modelo 23 será la base para construir el modelo final.

8. Modelo Final

Para la construcción del modelo final se ha utilizado como referencia M23 y se han añadido varias modificaciones. En primer lugar, en la parte densa se ha añadido una capa adicional, en la parte de capas LSTM se ha añadido otra y en la parte final que toma las salidas de ambas partes se han añadido dos capas densas más con el objetivo de aumentar la capacidad del modelo a ajustar los datos. Además, se ha añadido un *learning rate* dinámico como el *reduce on plateou* por tal de ir disminuyéndolo poco a poco, ya que en las etapas iniciales aprende rápido pero llega un momento que se queda en un valor cercano a 6'05, y en este punto tendría sentido reducir el *learning rate*.

Modelo	Parte	Capa	Neuronas	MAE
M22	Densa	1	64	6'01
	Densa	2	32	
	Densa	3	16	
	Densa	4	32	
	Densa	5	64	
	Densa	6	44	
	Densa	7	1	
	LSTM	1	50	
	LSTM	2	25	
	LSTM	3	10	
	LSTM	4	1	
	-	Flatten	-	
	-	Concatenate(densa(6), flatten)	-	
	Final	1	32	
	Final	2	16	
	Final	3	1	

Tabla 10: *Resultados de los modelos de inteligencia artificial junto a una breve descripción de su arquitectura.*

Los resultados muestran el mejor modelo hasta la fecha con un MAE de 6'01. Dado que el PES se encuentra entre 1 y 100, un error alrededor a 6 unidades se podría considerar como un resultado positivo. No obstante, para entender mejor cómo funciona el modelo final se procede a realizar un análisis exploratorio de los resultados.

En la *figure 10* se puede ver cómo la nube de puntos, donde cada uno representa el PES real y predicho para cada jugador en cada partido, sigue una relación positiva, lo que indica que a mayor PES real, mayor PES predicho. No obstante, también se observa algo de variabilidad en los datos

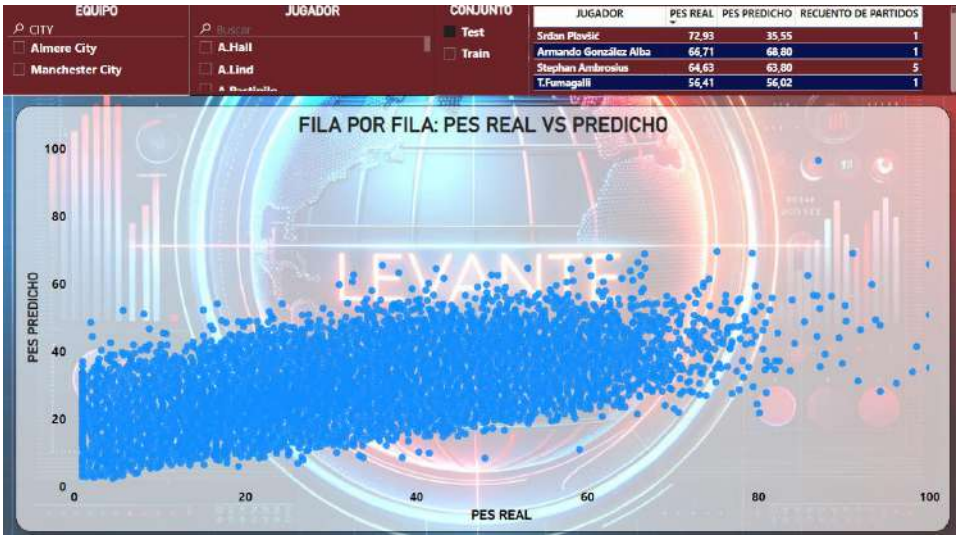

Figure 10: *Gráfico de dispersión, PES real vs PES predicho, cada punto es un partido de un jugador.*
Fuente: Elaboración propia.

No obstante, toda mejora cuando hablamos del rendimiento de los jugadores en promedio. En este caso, la relación para los resultados en el test es prácticamente lineal positiva. Esto significa que el modelo es capaz de predecir muy bien el nivel del jugador, distinguiendo generalmente bien entre jugadores de diferentes niveles por lo que a su rendimiento se refiere.

Figure 11: *Gráfico de dispersión, PES real vs PES predicho, cada punto es el promedio del PES por jugador.*
Fuente: Elaboración propia.

Por tanto, el modelo gana utilidad para poder identificar rachas buenas y malas de los jugadores, comparando el valor predicho (como se supone que debería de jugar un determinado jugador) con su rendimiento real, que como hemos visto es variable. En la imagen inferior se puede ver un ejemplo, con Adrián De La Fuente, central del Levante UD, cuyo rendimiento de la presente temporada fue a menos a mitad de temporada y en el último tramo dio un salto en su

PES. De esta manera, gráficamente se podrían identificar las rachas del jugador, mientras que se podría ver si rinde por encima o por debajo de lo esperado con una simple resta entre el valor predicho y el real.

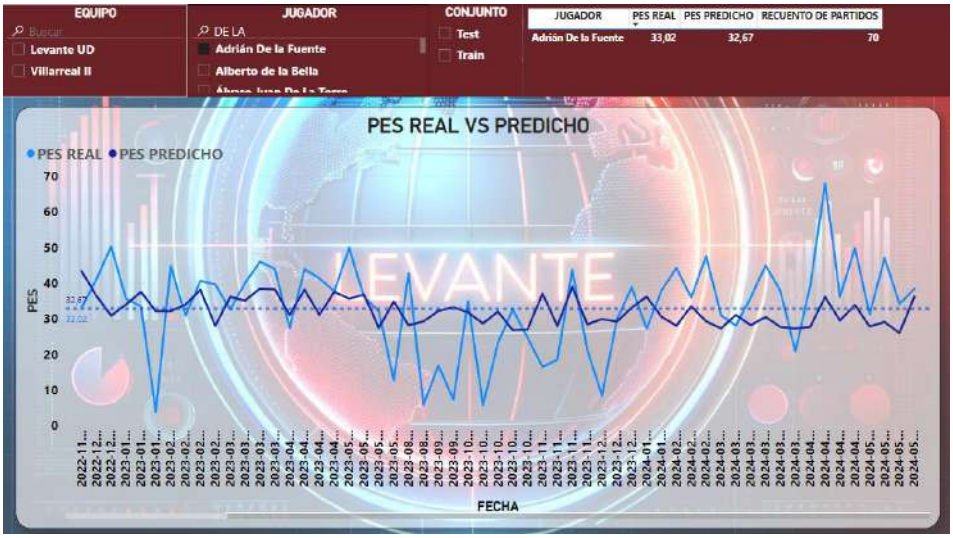

Figure 12: *Evolución del PES real vs PES predicho para Adrián De la Fuente.*
Fuente: Elaboración propia.

Además, no sólo es capaz de predecir bien el nivel esperado del jugador, sino que es capaz de predecirlo en el tiempo. Por ejemplo, en el caso de Ansu Fati, se observa cómo después de una bajada del rendimiento (relacionado con la lesión de ligamentos cruzados) el modelo capta esta nueva tendencia y se adapta a su nueva normalidad, es decir, un rendimiento algo menor que el que tenía previamente.

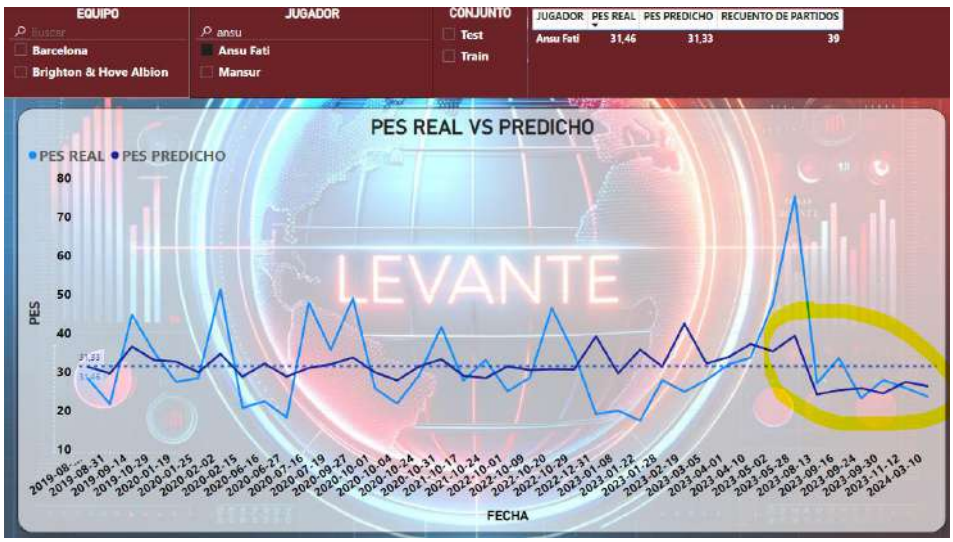

Figure 13: *Evolución del PES real vs PES predicho Ansu Fati.*
Fuente: Elaboración propia.

Por último, si se selecciona el conjunto de test, y se toma el PES promedio de todos los jugadores en el tiempo, se puede observar como el modelo, para fechas y partidos que no ha

visto (partidos con fecha superior a 2024-01-01), predice un rendimiento muy cercano al real, captando bien las bajadas y los picos.

Figure 14: *Evolución del PES real vs PES predicho para todos los jugadores en promedio.*
Fuente: Elaboración propia.

9. Aplicación interactiva con Power BI

Además, se ha creado una aplicación interactiva mediante Power BI con el objetivo de facilitar las tareas de *scouting*. Para ello, se ha importado la base de datos con todas las características de juego de cada jugador y el PES construido, y, por otro lado, la base de datos diseñada para conocer el rendimiento de cada jugador de cada equipo en el siguiente partido. Es por esto por lo que se han construido 3 principales funciones:

1. Identificación de talentos (PES y filtros)
2. Contextualización estilo de juego del jugador por posición (fortalezas y debilidades del jugador)
3. PES del próximo encuentro

El primer punto, es la identificación de talentos, que como vemos en la *figure 15*, en la parte superior disponemos de todos los filtros que permiten a los equipos adaptarse a sus necesidades y presupuesto. De esta manera, se podrá elegir la posición que quieres encontrar nuevos jugadores, y fijar la temporada que quieres ver su rendimiento, que normalmente será la más cercana. Posteriormente, se podrán añadir filtros por competición, edad, precio del jugador en el mercado para ajustar al presupuesto del club y la fecha de finalización de contrato que permitirá conocer a aquellos jugadores que se pueden conseguir a coste cero.

Una vez marcados los filtros, ordenaremos a los jugadores disponibles por su rendimiento deportivo o PES promedio agregado por temporada o el filtro de tiempo añadido en la selección. De esta manera, el resultado será una lista de jugadores de interés de mayor a menor rendimiento en el campo.

Por último, si seleccionamos a un jugador seremos capaces de ver su evolución del PES en el tiempo, pudiendo desagregar el PES promedio en cada uno de los partidos que ha disputado, entendiendo en el tiempo posibles rachas positivas o negativas de cada jugador.

Figure 15: *Página 1 de la aplicación interactiva, PES EVOLUTION.*
Fuente: Elaboración propia.

El segundo punto es la contextualización de cada jugador por cada posición, entendiendo su estilo de juego y sus principales habilidades, así como debilidades. Para ello se ha dividido en tres pantallas que reúnen las principales características que debería tener un defensa, un mediocentro, y, por último, un delantero.

En la *figure 16*, vemos las habilidades representativas en la faceta defensiva, tales como duelos aéreos, presión tras pérdida, entradas, recuperaciones, despejes, faltas y veces que es regateado, entre otras.

Para ver como se utilizaría se han seleccionado tres centrales a comparar: Rúben Días, Ronald Araújo y Pau Torres. En este veríamos que el perfil más completo de jugador defensivo sería Ronald Araújo que es superior en todas las características excepto en centros que le supera Rúben Días. Por el contrario, Pau Torres, a pesar de estar cerca en muchas habilidades, sólo destaca en dos negativas, la pérdida de balones y veces que es regateado. Indicando que el central del Villarreal (Pau Torres) tiene un rendimiento inferior que el central del Barcelona (Araújo) y del Manchester City (Rúben Dias).

No obstante, Rúben Dias tiene mayor PES que Ronald Araújo, lo que se puede deber a ciertas habilidades ofensivas con balón que ayudan a un equipo a generar ocasiones. Para entenderlo mejor, podríamos utilizar otro *dashboard* con un mayor número de características ofensivas.

Además, como bien se presentó en las alternativas de algoritmos de creación del PES, si hubiéramos fijado un peso manual para cada posición, a la hora de construir estos gráficos pondríamos las características que exclusivamente intervinieran en la elaboración del PES, teniendo así una relación directa entre la ordenación del PES y la interpretación de porqué ocurre. No obstante, con el modelo presentado se podría hacer igual, pero las variables seleccionadas cambian en función de la combinación entre competición, posición y minutos, lo que genera muchas diferentes posibilidades de características, y por eso, se ha decidido para contextualizar a cada jugador, entender las variables más representativas en el contexto de su posición. Así pues, se propone en futuras investigaciones probar alternativas de creación de PES.

Figure 16: *Página 2 de la aplicación interactiva, RADAR DEFENSIVO.*
Fuente: Elaboración propia.

Por otro lado, en la *figure 17*, se observan una combinación de características ofensivas y defensivas que representan las habilidades más representativas de un mediocentro, haciendo énfasis en la parte de creación y elaboración de jugadas. Aquí se pueden ver variables como centros, asistencias, tipos de pases, tiros, regates, recuperaciones y duelos aéreos.

En este caso, se han escogido Kevin De Bruyne, Florian Wirtz y Joey Veerman para su comparación. Tres jugadores con un alto PES esta temporada. Aquí se pueden ver tres perfiles de jugador totalmente distintos. Mientras que Florian Wirtz, por su número de pérdidas, regates y goles se muestra como un jugador más atrevido y que toma mayores riesgos en el ataque, los otros dos optan por un perfil con balón centrado en la elaboración del juego minimizando pérdidas. No obstante, Joey Veerman destaca en su capacidad de dar pases de todos los tipos, siendo en su gran parte exitosos, mientras que De Bruyne se centra en aquellos tipos de pase que más influencia tienen en el gol, como asistencias y centros.

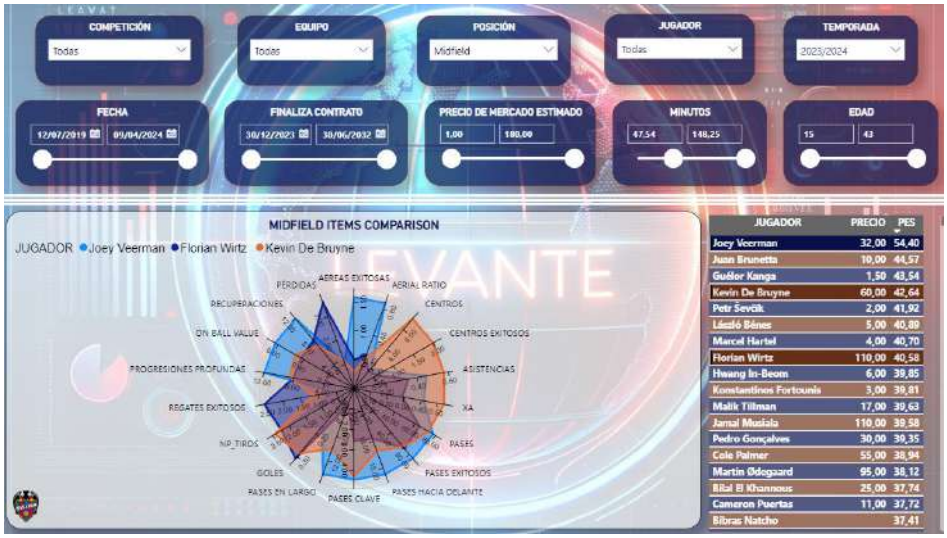

Figure 17: *Página 3 de la aplicación interactiva, RADAR DE MEDIOCENTROS.*
Fuente: Elaboración propia.

En tercer lugar, para los delanteros, se destacan habilidades ofensivas como los tiros, goles, regates, asistencias y capacidad de ganar duelos aéreos. En la *figure 18*, vemos dos perfiles de delanteros. Por un lado, Robert Lewandowski y Harry Kane, focalizados en la finalización. Esto se puede ver en su número de disparos, goles y número de disparos que hacen dividido las veces que ha tocado la pelota, indicando que cada dos veces que tocan el balón disparan a portería una de ellas. Además, este tipo de delanteros referencia tienden a tener buenas capacidades aéreas que les hace ganar duelos, no solo en fase de creación, sino también dentro del area donde un duelo aéreo ganado puede significar un gol. Por otro lado, Iago Aspas destaca en centros, asistencias y regates, por lo que se consideraría un perfil de delantero más asociativo.

Este proceso, nos hace ver la necesidad de en futuras investigaciones utilizar *clustering* por posición para tener perfiles de estilo de juego y crear un PES que dé prioridad a las características representativas de cada clúster.

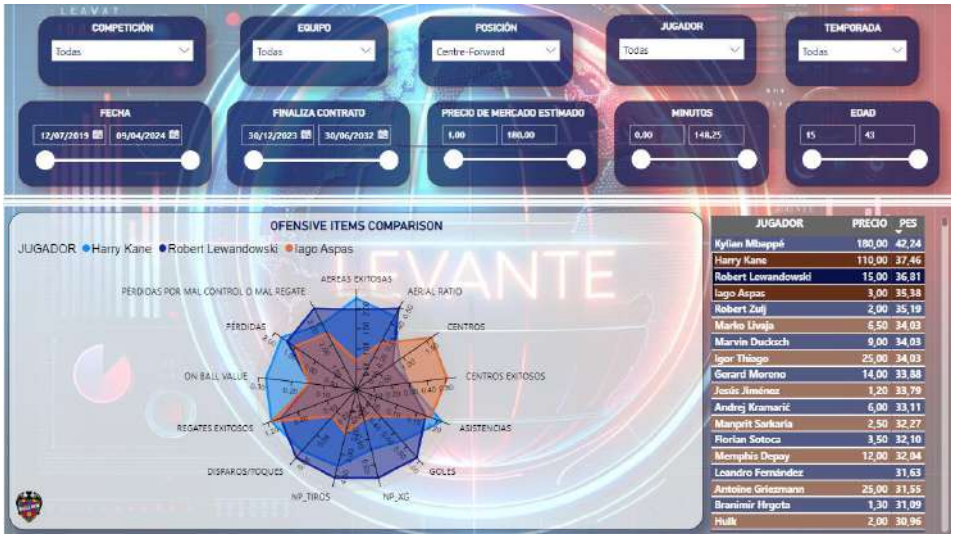

Figure 18: *Página 4 de la aplicación interactiva, RADAR DE DELANTEROS.*
Fuente: Elaboración propia.

A esta aplicación interactiva se le añade la figure 14 mostrada anteriormente, la cual incluye la métrica del PES predicho y la compara con el real para la identificación de rachas positivas y negativas de los diferentes jugadores.

10. Conclusión

En el fútbol se presenta la necesidad de cada temporada competir de la mejor manera posible y, con este objetivo, se intenta formar un equipo que sea capaz de optimizar los resultados. Cada año, jugadores terminan contrato o se retiran, y cada temporada se necesitan reforzar posiciones que han mostrado un nivel menor al esperado en la anterior. Es aquí donde cada equipo debe de hacer un esfuerzo para encontrar al jugador idóneo dadas condiciones y recursos disponibles en el momento.

Con tal de conseguir esto un nuevo reto aparece, el de cuantificar el rendimiento de un jugador en cada uno de sus partidos en función de sus estadísticas. La dificultad de crear una nueva métrica que previamente no se tiene en el conjunto de datos ha supuesto un desafío a la hora de utilizar métodos no supervisados y complementarlos con una serie de pasos para que el proceso genere el *output* esperado, es decir, un valor que mida el rendimiento.

Esta manera de proceder mediante el uso de PCA junto a un algoritmo que obtiene el peso de las variables en el modelo, donde posteriormente se seleccionan las 10 características más importantes por cada posición, minutos y distinción entre grandes ligas y el resto, supone un nuevo proceso en el cual se ha arriesgado en cuanto a creatividad. No obstante, a pesar de no haber manera tangible de mostrar el error que se obtiene, los resultados han sido mostrados a profesionales del Levante Unión Deportiva, los cuáles han confirmado que lo obtenido es muy similar a sus ideas preestablecidas siendo muy útil ya que ahorra trabajo de mirar miles de partidos en vídeo para obtener una visión completa de los jugadores disponibles en el mercado. Además, se añade que se puede conocer el rendimiento de los 17.000 jugadores disponibles, una variedad que nunca se podría llegar a obtener mediante scouting tradicional basado en ver partidos.

Traducido en costes económicos, fichar a un jugador a un buen precio, incluso gratis, y posteriormente venderlo a un precio superior puede proporcionar beneficios al club desde miles hasta millones de euros. Hasta este punto, un equipo podría ser capaz de encontrar al mejor o mejores jugadores dados los recursos económicos disponibles con una simple selección de filtros sobre características y posición sumado a una ordenación del PES. Un proceso sencillo y rápido con un gran impacto.

Por otro lado, el proyecto aporta un valor añadido mediante la creación de la *Players Reference Data*, capaz de permitir el enlace entre una fuente de datos técnicos y deportivos y un conjunto de datos económicos. Esto favorece a un caso de uso en concreto, filtrar por precio de mercado o por la fecha de finalización de contrato, lo que puede llegar a suponer ahorrar millones de euros en un fichaje que pueda aportar un rendimiento similar o incluso superior. Además, esta base de datos ha ido siendo actualizada de 10.700 jugadores hasta 17.000, de manera que se mantiene actualizada con una cobertura del 100% de los jugadores presentes

tomando como referencia el conjunto de datos de rendimiento para la detección de nuevos jugadores. Esto supone que la herramienta creada pueda ser capaz de conseguir nuevos jóvenes talentos o jugadores que fichan de otras ligas que no estaban presentes en las 30 disponibles y que por esa razón no apareciesen.

Otro aspecto que ha permitido llegar a los objetivos con los datos más recientes posibles es que se ha buscado crear el código y los archivos de manera reproducible. De esta forma, cada mes no sólo se han añadido jugadores a la *Players Reference Data*, sino que también se han añadido al conjunto de datos final con el PES los partidos más recientes que se iban disputando en todas las competiciones disponibles. Esta ha resultado ser una buena forma de trabajar que ha añadido valor a los resultados finales, ya que los análisis hechos están actualizados.

En la construcción del modelo se vio un mejor funcionamiento en la utilización de *autoencoders* como primer paso ya que al aprender a comprimir y luego descomprimir los datos de entrada, el modelo puede ser menos propensos al sobreajuste en comparación con redes densas que simplemente aprenden a mapear entradas a salidas sin considerar la estructura interna de los datos. Además, han podido aprender a ignorar el ruido y centrarse en los aspectos más estables y relevantes en los datos. Esto sumado a la incorporación de redes LSTM que tomaban como entrada una serie temporal del PES del jugador en los últimos 20 partidos ha contribuido en el buen funcionamiento del modelo, ya que por un lado es capaz de entender una evolución del rendimiento en el tiempo, y por otro, contemplar las circunstancias a las que se enfrenta el equipo, como jugar en casa o fuera, o el de rival.

Este modelo ha resultado ser útil para cumplir con el último objetivo particular, la identificación de rachas positivas y negativas de los jugadores. Esto se debe a que el modelo predice como se espera que rinda un jugador basándose en su rendimiento en los últimos 20 partidos, lo que puede equivaler a un rango entre 5 a 6 meses por lo general, y otras características como el rival, localidad, árbitro y edad, entre otras mencionadas con más detalle en el apartado 8 sobre los resultados obtenidos. De esta manera, el modelo predice bien lo que se espera que ocurra y cuando un jugador rinde por encima o por debajo de lo esperado durante un tiempo se puede interpretar como rachas positivas o negativas.

Con esto, se concluye que predecir el rendimiento de un jugador en cada partido es muy difícil y variable, debido a que este puede depender de un gran número de variables que difícilmente se van a poder recoger o transformar en datos, como, por ejemplo, un problema que le ha podido ocurrir la noche previa al partido. No obstante, predecir el rendimiento general de los jugadores sí que es posible, y gracias a que cada predicción se basa en rendimientos pasados hasta la fecha, por lo que será más fácil acercarse a lo que pueda ocurrir en el partido siguiente. En caso de que el jugador pueda cambiar la dinámica de rendimiento la predicción podría ser peor, pero la herramienta seguiría siendo útil, ya que su objetivo principal es que sirva para detectar momentos positivos o negativos prolongados en el tiempo y que esto permita entender

mejor al jugador propuesto para ser fichado. Otra opción, sería emplearlo con jugadores del mismo equipo para reconocer malas rachas lo antes posible y ayudar al deportista a revertirlas.

Otro aspecto positivo es que el modelo es capaz de adaptarse a nuevos comportamientos del jugador. Es decir, un jugador que suela rendir entorno al 50, si empieza a tener un rendimiento menor, el modelo lo detectará y se adaptará a la nueva normalidad. Esto se debe a que el modelo es una LSTM, capaz de captar el rendimiento a corto y largo plazo, y ha aprendido de los datos a adaptarse a cambios de rendimiento en los jugadores, probablemente debidos a lesiones o periodos en los que el jugador ha estado inactivo.

En último lugar y haciendo un breve resumen, se puede decir que se han cumplido todos los objetivos vinculando los datos de rendimiento con datos económicos gracias a las *Players Reference Data* (Objetivo 1), que, sumado al desarrollo del PES (Objetivo 2), ha permitido que cualquier equipo pueda encontrar a los mejores jugadores dados los recursos disponibles por el club. Con la predicción del PES en el siguiente partido (Objetivo 3) se rompe con la clásica pregunta de que en el fútbol es imposible predecir el rendimiento que tendrá un jugador, pudiendo concluir que esto es difícil por el gran número de variables a tener en cuenta que son difíciles de conseguir, no obstante si que es posible predecir cómo se esperaría que un jugador rinda dado un contexto y su PES pasado, lo que permite ser utilizado para la identificación de rachas (Objetivo 5), como bien se ha mencionado anteriormente. Estos resultados se ven concluidos en una aplicación interactiva con Power BI, con una presentación clara y un caso de uso en concreto, encontrar fácil y rápidamente opciones a ser consideradas como posibles fichajes para la formación de un equipo (Objetivo 4).

11. Futuras Investigaciones

Dada la longitud de este trabajo, existen ciertos apartados mejorables en futuras investigaciones. En cuanto a la creación del PES, para cada posición se debería probar a realizar un PCA y *clustering* que permitiese identificar grupos que por su estilo de juego. Por ejemplo, un extremo veloz o un extremo que basa su juego en acciones combinativas, o, por otro lado, un delantero que es menos móvil o un delantero que aprovecha su velocidad jugando a las espaldas de los defensas. Una vez se tuvieran los grupos por posición y perfil, se seleccionarían las variables más representativas de cada uno. Este paso o bien podría automatizarse, o bien, mediante un análisis exploratorio se podrían identificar las características que representan a los jugadores de cada *clúster*.

Por otro lado, la asignación de los pesos a las variables podría continuar siendo automático, pero debería de considerarse la opción de fijar manualmente los pesos según el criterio de profesionales. En este último, los entrenadores y directivos podrían influir en las habilidades que les gustaría que su jugador fichado tuviera para cubrir las necesidades del club. De esta manera, serían capaces de elegir tanto las habilidades como la importancia que le proporcionan a cada una de ellas. Con esto se crearía un PES manual, qué junto al PES automático, y, asignando una ponderación a cada uno, se podría construir un PES final más enfocado en las habilidades que se requieran en el deportista.

Además, para la creación del PES se podría mejorar integrando nuevas fuentes de datos, como, por ejemplo, información biocondicional. Dentro de esta, se encuentra la velocidad, distancia recorrida, esfuerzos a altas intensidades y potencia metabólica. Esto enriquecería el rendimiento de un jugador, basándose no sólo en sus registros técnicos, sino que se incorporaría información sobre el estado físico también.

Para aportar más valor en la labor de *scouting*, se propone investigar sobre la posibilidad de predecir un mayor rango temporal, por ejemplo 10 partidos. De esta manera se podría ver tanto el nivel de adaptación a otro equipo y competición. Como una tendencia sobre su evolución a futuro.

Otros aspectos para investigar son la integración de información sobre medios informativos que podrían ser otra manera de tener información sobre el rendimiento de un jugador desde un punto de vista subjetivo que podría llegar a convertirse en cuantitativo mediante algún modelo. Además, otras líneas de investigación incluirían conocer cómo el rendimiento deportivo afecta en el impacto mediático en las redes sociales y cómo se relaciona con las marcas deportivas o agencias de representación.

Por último, el valor aportado por todo este trabajo que ayuda en el sector del *scouting* a encontrar nuevos jugadores adaptados a las necesidades de cada club, podría ser traducido en un producto final que permita su propia monetización.

12. Bibliografía

1. Ayora MJM. Diseño y aplicación de técnicas de machine learning para optimizar el Scouting en clubes de fútbol.
2. Selma M, Pilar M del. Machine Learning en el mundo del fútbol. 1 de octubre de 2019 [citado 20 de julio de 2024]; Disponible en: https://riunet.upv.es/handle/10251/129491
3. Ruiz JMM, Cabrera LV. La ciencia de datos como herramienta diferenciadora en el scouting deportivo en el fútbol de élite [citado 1 de agosto de 2024].
4. Galaz P, Mena S, Saure D. Inferencia Bayesiana de un Modelo Markoviano de Fútbol con aplicación en Scouting. 2021 [citado 23 de julio de 2024].
5. Arts E. Videojuegos de FIFA - Sitio oficial de EA [Internet]. Electronic Arts Inc. 2016 [citado 11 de julio de 2024]. Disponible en: https://www.ea.com/es-es/games/fifa.
6. ¡Así es cómo calcula EA Sports las medias en FIFA Ultimate Team! [Internet]. [citado 13 de julio de 2024]. Disponible en: https://es.besoccer.com/noticia/asi-es-como-calcula-ea-sportslas-medias-en-fifa-ultimate-team-1043599
7. Molina Escondrillas CM. Análisis económico del fútbol profesional. Economic analysis of professional soccer [Internet]. 2016 [citado 11 de julio de 2024]; Disponible en: https://repositorio.upct.es/handle/10317/5755
8. Aguiriano B, Manuel V. La situación económica del fútbol profesional en España [Internet] [Proyecto/Trabajo fin de carrera/grado]. Universitat Politècnica de València; 2022 [citado 2 de agosto de 2024]. Disponible en: https://riunet.upv.es/handle/10251/185544
9. Livescores - Soccer - Betting Showcase [Internet]. [citado 3 de agosto de 2024]. Disponible en: https://optaplayerstats.statsperform.com/en_GB/soccer
10. Estadísticas e Historia del Fútbol | FBref.com [Internet]. [citado 3 de agosto de 2024]. Disponible en: https://fbref.com/es/
11. Fichajes de fútbol, rumores, valores de mercado y noticias [Internet]. [citado 3 de agosto de 2024]. Disponible en: https://www.transfermarkt.es/
12. Home [Internet]. [citado 3 de agosto de 2024]. Disponible en: https://skillcorner.com/es/
13. Livescores - Soccer - Betting Showcase [Internet]. [citado 3 de agosto de 2024]. Disponible en: https://optaplayerstats.statsperform.com/en_GB/soccer
14. Welcome to Python.org [Internet]. Python.org. 2023 [citado 11 de julio de 2024]. Disponible en: https://www.python.org/
15. (Fichajes de fútbol, rumores, valores de mercado y noticias [Internet]. [citado 12 de julio de 2024]. Disponible en: https://www.transfermarkt.es/)

16. 6.5.17. Some properties of PCA models — Process Improvement using Data [Internet]. [citado 14 de julio de 2024]. Disponible en: https://learnche.org/pid/latent-variablemodelling/principal-component-analysis/some-properties-of-pca-model
17. Amesquita D. ¿Cómo normalizar datos entre 0 y 1? [Internet]. Statologos. 2021 [citado 3 de agosto de 2024]. Disponible en: https://statologos.com/normalizar-datos-entre-0-y-1/
18. scikit-learn [Internet]. [citado 3 de agosto de 2024]. RandomForestClassifier. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
19. Team K. Keras documentation: LSTM layer [Internet]. [citado 3 de agosto de 2024]. Disponible en: https://keras.io/api/layers/recurrent_layers/lstm/
20. Codificando Bits [Internet]. [citado 3 de agosto de 2024]. Autoencoders: explicación y tutorial en Python. Disponible en: https://www.codificandobits.com/blog/autoencoders-explicacion-y-tutorial-python/
21. admin. Redes Neuronales en la Predicción de Series Temporales: Estrategias Efectivas [Internet]. 2023 [citado 3 de agosto de 2024]. Disponible en: https://data-universe.org/redes-neuronales-en-la-prediccion-de-series-temporales-estrategias-efectivas/