



## Resumen

---

La estimación del valor de mercado de los jugadores es crucial para la planificación de un equipo. Permite a los clubes de fútbol afrontar las negociaciones con la información adecuada, así como calcular correctamente el valor de sus activos.

Aunque ha habido varios enfoques para calcular el valor de los jugadores, muchos de los modelos propuestos son de caja negra y no proporcionan información sobre qué atributos son más importantes a la hora de determinar el valor de los jugadores, que puede cambiar según la posición.

En este proyecto exploraremos diferentes modelos de aprendizaje automático para evaluar el valor de mercado de los jugadores según su posición. El objetivo principal es comprender qué características tienen una mayor importancia a la hora de calcular el valor de mercado de cada jugador por posición.

**Palabras clave:** analítica deportiva, análisis de datos, ciencia de datos, aprendizaje automático, fútbol.

## Resum

---

L'estimació del valor de mercat dels jugadors és crucial per a la planificació d'un equip. Permet als clubs de futbol afrontar les negociacions amb la informació adequada, així com calcular correctament el valor dels seus actius.

Encara que hi ha hagut diversos enfocaments per a calcular el valor dels jugadors, molts dels models proposats són de caixa negra i no proporcionen informació sobre quins atributs són més importants a l'hora de determinar el valor dels jugadors, que pot canviar segons la posició.

En este projecte explorarem diferents models d'aprenentatge automàtic per a avaluar el valor de mercat dels jugadors segons la seua posició. L'objectiu principal és comprendre quines característiques tenen una major importància a l'hora de calcular el valor de mercat de cada jugador per posició.

**Paraules clau:** analítica esportiva, anàlisi de dades, ciència de dades, aprenentatge automàtic, futbol.

## Abstract

---

The estimation of the market value of players is crucial for the planning of a team. It allows football clubs to approach negotiations with the right information as well as to correctly estimate the value of their assets.

Although there have been several approaches to estimate the value of players, many of the proposed models are black box and do not provide information on which attributes are most important in determining the value of players, which can change depending on the position.

In this project we will explore different machine learning models models to assess the market value of players according to their position. The main objective is to understand which characteristics are the most important when calculating the market value of each player by position.

**Keywords:** sports analytics, data analytics, data science, machine learning, football.

# Índice

1.	Introducción .....	11
1.1.	Motivación .....	11
1.2.	Objetivos .....	11
1.3.	Impacto esperado .....	12
1.4.	Estructura de la memoria .....	12
2.	Estado del arte .....	14
2.1.	Analítica deportiva en fútbol .....	14
2.2.	Valoración de futbolistas .....	15
2.3.	Crítica al estado del arte .....	17
3.	Metodología .....	18
4.	Conjuntos de datos .....	19
4.1.	Obtención de datos .....	19
4.1.1.	Conexión entre transfermarkt y fbref .....	19
4.1.2.	Adquisición de datos .....	21
4.2.	Tratamiento de los datos .....	22
4.3.	Descripción de los datos .....	22
4.4.	Análisis exploratorio de datos .....	25
4.4.1	Análisis univariante .....	25
4.4.2	Análisis multivariante .....	39
5.	Modelos propuestos .....	52
6.	Experimentos .....	54
6.1.	Jugadores de campo .....	54
6.1.1	Regresión lineal múltiple .....	54
6.1.2	Elastic Net .....	56
6.1.3	Árboles de decisión .....	56
6.1.4	Random Forest .....	57
6.1.5	GAM .....	58
6.1.6	EBM .....	60
6.1.7	Redes neuronales .....	60
6.2.	Defensas .....	61
6.2.1	Regresión lineal múltiple .....	61
6.2.2	Elastic Net .....	63
6.2.3	Árboles de decisión .....	63
6.2.4	Random Forest .....	64
6.2.5	GAM .....	65
6.2.6	EBM .....	66
6.2.7	Redes neuronales .....	66



6.3.	Centrocampistas .....	67
6.3.1	Regresión lineal múltiple .....	67
6.3.2	Elastic Net .....	69
6.3.3	Árboles de decisión .....	69
6.3.4	Random Forest .....	70
6.3.5	GAM .....	71
6.3.6	EBM .....	72
6.3.7	Redes neuronales .....	72
6.4.	Delanteros .....	73
6.4.1	Regresión lineal múltiple .....	73
6.4.2	Elastic Net .....	75
6.4.3	Árboles de decisión .....	75
6.4.4	Random Forest .....	76
6.4.5	GAM .....	76
6.4.6	EBM .....	77
6.4.7	Redes neuronales .....	78
6.5.	Porteros .....	79
6.5.1	Regresión lineal múltiple .....	79
6.5.2	Elastic Net .....	80
6.5.3	Árboles de decisión y Random Forest .....	81
6.5.4	GAM .....	81
6.5.5	EBM .....	82
6.5.6	Redes neuronales .....	82
6.6.	Análisis de resultados .....	84
7.	Conclusiones .....	86
	Legado .....	87
	Relación con los estudios cursados .....	88
	Referencias .....	89
	Glosario .....	91
	Analítica deportiva .....	91
	Analítica en fútbol .....	91
	Objetivos de desarrollo sostenible .....	93

# Índice de figuras

Figura 1: Estadísticas descriptivas de la edad. Jugadores de campo a la izquierda, porteros a la derecha .....	26
Figura 2: Histograma de edad. Jugadores de campo a la izquierda, porteros a la derecha .....	26
Figura 3: Estadísticas descriptivas de los valores de los fichajes. Jugadores de campo a la izquierda, porteros a la derecha .....	27
Figura 4: Histogramas de costes de fichaje. Jugadores de campo a la izquierda, porteros a la derecha .....	27
Figura 5: Estadísticas descriptivas de la duración restante de contrato (meses). Jugadores de campo a la izquierda, porteros a la derecha .....	28
Figura 6: Histogramas de duración de contrato restante (meses). Jugadores de campo a la izquierda, porteros a la derecha .....	28
Figura 7: Número de fichajes por año. Jugadores de campo a la izquierda, porteros a la derecha .....	29
Figura 8: Número de fichajes por país del equipo de origen. Jugadores de campo a la izquierda, porteros a la derecha .....	29
Figura 9: Número de fichajes por país del equipo de destino. Jugadores de campo a la izquierda, porteros a la derecha .....	30
Figura 10: Número de fichajes por posición .....	30
Figura 11: Gráfico de densidad de partidos jugados para jugadores de campo .....	31
Figura 12: Gráfico de densidad de titularidades de jugadores de campo .....	31
Figura 13: Gráfico de densidad de goles sin contar penaltis .....	32
Figura 14: Gráfico de densidad de goles de penalti .....	32
Figura 15: Gráfico de densidad para asistencias de jugadores de campo .....	33
Figura 16: Gráfico de densidad para tiros y tiros a puerta .....	33
Figura 17: Gráfico de densidad para acciones defensivas .....	34
Figura 18: Ocasiones de gol creadas en grandes competiciones por jugadores de campo .....	34
Figura 19: Gráfico de densidad de tarjetas para jugadores de campo .....	35
Figura 20: Gráfico de densidad de partidos jugados para porteros .....	35
Figura 21: Diagrama de densidad para titularidades en porteros .....	36
Figura 22: Diagrama de densidad de minutos jugados por porteros .....	36
Figura 23: Gráfico de densidad de goles encajados .....	37
Figura 24: Gráfico de densidad de tiros a puerta en contra .....	37
Figura 25: Histograma de porterías a cero .....	38
Figura 26: Penaltis recibidos y parados .....	38
Figura 27: Tarjetas en porteros .....	39
Figura 28: Asistencias en porteros .....	39

Figura 29: Mapa de calor de correlaciones de variables referidas a participación .....	40
Figura 30: VIF en variables referidas a participación .....	40
Figura 31: Correlación y VIF entre variables referidas a minutos jugados.....	41
Figura 32: Correlación entre variables relativas a disparos a portería y goles .....	41
Figura 33: VIF en variables relativas a disparos a portería y goles .....	42
Figura 34: Correlación y VIF entre variables referidas a disparos a portería y goles.....	42
Figura 35: Correlaciones entre variables relacionadas con la creación de ocasiones .....	43
Figura 36: VIF en variables referidas a ocasiones de gol creadas .....	43
Figura 37: Correlaciones entre variables .....	44
Figura 38: VIF en las variables restantes .....	44
Figura 39: Correlaciones de variables relativas a porteros .....	45
Figura 40: VIF de variables relativas a porteros .....	45
Figura 41: Fichajes internos por país y edad .....	46
Figura 42: Fichajes por edad y país de origen del equipo que abandona el jugador .....	46
Figura 43: Fichajes por edad y país de origen del equipo por el que ficha el jugador .....	47
Figura 44: Fichajes internos de porteros por país y edad .....	47
Figura 45: Fichajes de porteros por edad y país de origen del equipo que abandona el futbolista .....	48
Figura 46: Fichajes de porteros por edad y país de origen del equipo por el que ficha el jugador .....	48
Figura 47: Gasto en futbolistas por países .....	49
Figura 48: Diferencia entre gastos e ingresos por fichajes .....	49
Figura 49: Gasto total en fichajes por posición .....	50
Figura 50: Gasto promedio por posición .....	50
Figura 51: Regresión lineal múltiple para jugadores de campo (1) .....	54
Figura 52: Regresión lineal múltiple para jugadores de campo (2) .....	55
Figura 53: Elastic Net para jugadores de campo (1) .....	56
Figura 54: Elastic Net para jugadores de campo (2) .....	56
Figura 55: Resultados de árboles de decisión para jugadores de campo .....	57
Figura 56: Variables influyentes en árbol de decisión para jugadores de campo .....	57
Figura 57: Shapley values de RF para jugadores de campo .....	58
Figura 58: Efectos no lineales de variables con GAM para jugadores de campo .....	59
Figura 59: Importancia de cada variable en GAM para jugadores de campo .....	59
Figura 60: Importancia de variables en EBM para jugadores de campo .....	60
Figura 61: Redes neuronales de dos capas para jugadores de campo .....	60
Figura 62: Redes neuronales de tres capas para jugadores de campo .....	60
Figura 63: Shapley values para redes neuronales en jugadores de campo .....	61
Figura 64: Regresión lineal múltiple para defensas (1) .....	62
Figura 65: Regresión lineal múltiple para defensas (2) .....	62

Figura 66: Elastic Net para defensas (1) .....	63
Figura 67: Elastic Net para defensas (2) .....	63
Figura 68: Árboles de decisión para defensas (1) .....	63
Figura 69: Árboles de decisión para defensas (2) .....	64
Figura 70: Random forest para defensas .....	64
Figura 71: GAM para defensas (1) .....	65
Figura 72: GAM para defensas (2) .....	65
Figura 73: EBM para defensas .....	66
Figura 74: Redes neuronales de dos capas para defensas .....	66
Figura 75: Redes neuronales de tres capas para defensas .....	66
Figura 76: Shapley values en redes neuronales para defensas .....	67
Figura 77: Regresión lineal múltiple para centrocampistas (1) .....	68
Figura 78: Regresión lineal múltiple para centrocampistas (2) .....	68
Figura 79: Elastic Net para centrocampistas (1) .....	69
Figura 80: Elastic Net para centrocampistas (2) .....	69
Figura 81: Árboles de decisión para centrocampistas (1) .....	69
Figura 82: Árboles de decisión para centrocampistas (2) .....	70
Figura 83: Random forest para centrocampistas .....	70
Figura 84: GAM para centrocampistas (1) .....	71
Figura 85: GAM para centrocampistas (2) .....	71
Figura 86: EBM para centrocampistas .....	72
Figura 87: Redes neuronales de dos capas para centrocampistas .....	72
Figura 88: Redes neuronales de tres capas para centrocampistas .....	72
Figura 89: Shapley values en redes neuronales para centrocampistas .....	73
Figura 90: Regresión lineal múltiple para delanteros (1) .....	74
Figura 91: Regresión lineal múltiple para delanteros (2) .....	74
Figura 92: Elastic Net para delanteros (1) .....	75
Figura 93: Elastic Net para delanteros (2) .....	75
Figura 94: Árboles de decisión para delanteros (1) .....	75
Figura 95: Árboles de decisión para delanteros (2) .....	75
Figura 96: Random Forest para delanteros .....	76
Figura 97: GAM para delanteros (1) .....	77
Figura 98: GAM para delanteros (2) .....	77
Figura 99: EBM para delanteros .....	78
Figura 100: Redes neuronales de dos capas para delanteros .....	78
Figura 101: Redes neuronales de tres capas para delanteros .....	78
Figura 102: Shapley values en redes neuronales para delanteros .....	79
Figura 103: Regresión lineal múltiple para porteros (1) .....	80
Figura 104: Regresión lineal múltiple para porteros (2) .....	80

Figura 105: Elastic Net para porteros (1) .....	80
Figura 106: Elastic Net para porteros (2) .....	81
Figura 107: GAM para porteros (1) .....	81
Figura 108: GAM para porteros (2) .....	82
Figura 109: Redes neuronales de dos capas para porteros .....	82
Figura 110: Redes neuronales de tres capas para porteros .....	82
Figura 111: Shapley values de redes neuronales para porteros .....	83

# 1. Introducción

---

## 1.1. Motivación

El mercado de fichajes de futbolistas se ha posicionado durante las últimas décadas como el más importante de la industria deportiva en cuanto a relevancia económica y mediática, llegando a mover cifras desorbitadas de dinero, las cuales pueden llegar a cientos o miles de millones de euros a lo largo de cada temporada [1].

Para un equipo de fútbol es de vital importancia estimar con precisión el valor de mercado de los distintos futbolistas de su equipo y de aquellos que se hayan en su órbita de fichajes, ya que puede marcar la diferencia entre una buena y una mala planificación deportiva, afectando al éxito deportivo del equipo, así como a su situación económica y financiera [2].

Predecir el valor de mercado de estos futbolistas puede ayudar a los clubes a abordar aspectos esenciales relativos a la gestión económica y de plantilla [2], entre los cuales se encuentran:

- Toma de decisiones estratégicas, tales como determinar cuánto dinero ofrecer a la hora de tratar de comprar a un futbolista y cuánto dinero pedir a otros equipos si el objetivo es venderlo.
- Identificación de oportunidades en el mercado, que pueden incluir jugadores infravalorados o con potencial de crecimiento.
- Planificación de la plantilla, contribuyendo a construir un equipo equilibrado y competitivo que permita alcanzar los objetivos deportivos de cada temporada, optimizando el presupuesto disponible, así como sentar las bases de un equipo competitivo de cara a temporadas futuras.

La mayoría de los modelos propuestos para calcular el valor de mercado de los futbolistas han resultado ser de caja negra, por lo que no es posible acceder a información como qué atributos son más importantes a la hora de determinar el valor de cada uno [3].

Un modelo de aprendizaje automático explicable podría proporcionar a los clubes herramientas más precisas para evaluar el valor de los futbolistas, así como dar a entender con claridad cuáles son los factores que influyen en el valor de mercado de los distintos futbolistas (habilidades técnicas, rendimiento reciente, edad, posición, marketing, etc.) y en qué medida contribuye cada uno de ellos.

## 1.2. Objetivos

El objetivo principal de esta obra no es otro que el desarrollo y la posterior implementación de diferentes modelos de aprendizaje automático explicables que permitan estimar con precisión el valor de mercado de distintos futbolistas para la presente temporada (2024-2025), así como dilucidar qué factores han sido relevantes, y en qué medida, para inferir dichas predicciones.

Para ello, implementaremos y compararemos diversos algoritmos de aprendizaje automático para determinar cuáles ofrecen un mejor equilibrio entre precisión y explicabilidad. También evaluaremos el rendimiento de los distintos modelos entrenados para dilucidar cuáles



ofrecen mejores resultados para este propósito.

### 1.3. Impacto esperado

Se espera que este trabajo pueda llegar a ser de ayuda para la toma de decisiones relativas a la gestión de plantilla de distintos clubes de fútbol, que los modelos desarrollados a lo largo del mismo faciliten la evaluación del mercado de fichajes y ayuden a los equipos que hagan uso de ellos a planificarse de cara a alcanzar los objetivos de cada temporada.

Algunos de los principales beneficiarios serían las direcciones deportivas de los equipos de fútbol. Los modelos ofrecerán una base sólida para estimar el valor real de los futbolistas, tanto de la plantilla actual como de los objetivos de fichaje, lo que les permitiría optimizar las inversiones, negociar de forma más informada en traspasos y renovaciones y evitar sobrevaloraciones o infravaloraciones que comprometan la salud económica y el rendimiento deportivo del club.

Los departamentos financieros de los clubes también podrían sacar provecho de esta obra, ya que la precisión en la valoración de jugadores contribuiría a una mejor gestión de recursos y una mayor estabilidad económica para el club.

Otra finalidad de esta obra es la de apoyar y servir como referencia a futuras investigaciones relacionadas, o bien al desarrollo de nuevas metodologías, ya sea dentro del mundo del fútbol o en otros deportes.

Por último, se proyecta poder contribuir a la creación de análisis avanzados por parte de periodistas o aficionados al fútbol mediante modelos que faciliten la comprensión del mercado de fichajes.

### 1.4. Estructura de la memoria

Una vez acabada la introducción pasaremos al apartado 2, donde se llevará a cabo el estado del arte, apartado en el cual realizaremos un recorrido por algunas de las distintas obras de analítica deportiva que se han llevado a cabo hasta el día de hoy, profundizando en aquellas referidas a analítica en fútbol y en concreto a estimar valores de mercado de futbolistas.

A continuación, en el apartado 3, comentaremos la metodología que llevaremos a cabo, desde la extracción de datos hasta el posterior entrenamiento de modelos que traten de predecir el valor de mercado de futbolistas.

En el apartado 4 hablaremos de los conjuntos de datos a analizar, de cómo se han obtenido y tratado antes de empezar a entrenar modelos. Este apartado contendrá también un análisis exploratorio, el cual realizaremos con el objetivo de explicar la distribución de las variables y tratar posibles problemas de multicolinealidad.

Posteriormente, en el apartado 5, comentaremos en detalle cuáles son los modelos que vamos a llevar a cabo para estimar valores de mercado y determinar qué variables influyen más a la hora de valorar un futbolista.

En el apartado 6 llevaremos a cabo distintos experimentos para tratar de estimar el valor de mercado de los futbolistas, obteniendo posteriormente conclusiones sobre cuáles han sido mejores y sobre qué variables han influido en mayor medida a la hora de entrenarlos.

A continuación, tendremos un apartado de conclusiones, en el que comentaremos cuáles han sido los objetivos del trabajo, qué procedimientos hemos llevado a cabo para lograrlos y qué resultados hemos obtenido. Asimismo, se formularán propuestas de trabajo futuro en este mismo apartado.

Para terminar, tendremos un apartado de referencias, un glosario y un capítulo sobre los objetivos de desarrollo sostenible a los que esta obra ha podido contribuir.

## 2. Estado del arte

---

La creciente relevancia económica del mercado de fichajes de fútbol, estimada en 741,45 millones de dólares en 2024 y con pronóstico de alcanzar los 906 millones de dólares en 2029 [4], ha traído consigo multitud de incentivos para desarrollar herramientas y métodos que permitan evaluar de manera precisa el valor de mercado de cada uno de los futbolistas que forman parte del mismo.

El uso de estas herramientas se torna cada vez más imprescindible para los clubes de cara a llevar a cabo una buena planificación de plantilla y mantener su competitividad, especialmente en el más alto nivel.

Asimismo, existe un interés general desde la perspectiva del aficionado por estimar valores de mercado de jugadores de fútbol, de cara a opinar y debatir si un fichaje ha sido justo, si se ha pagado de más o de menos por él, qué futbolistas pueden ser interesantes para su equipo de cara a la siguiente temporada, etc.

Esto se ve reflejado en la gran popularidad de portales web de fútbol como Transfermarkt [5], que supera los 30 millones de visitas mensuales según la herramienta de análisis de tráfico Semrush [6].

Todos estos factores han propiciado la creación de distintas obras sobre el tema en cuestión que han usado diferentes métodos para la estimación de valores de mercado, desde modelos de regresión y econométricos hasta enfoques basados en métricas de rendimiento deportivo.

### 2.1. Analítica deportiva en fútbol

La analítica deportiva ha revolucionado el panorama del fútbol moderno, transformando la forma en que se analizan el rendimiento, la estrategia y la gestión de los equipos. A medida que el deporte ha evolucionado hacia una mayor profesionalización y competitividad, el uso de datos se ha convertido en una herramienta clave para la toma de decisiones tanto dentro como fuera del terreno de juego.

Entre las obras dedicadas a este tópico, podemos encontrar análisis de cómo la analítica de datos impacta las decisiones gerenciales en el fútbol europeo [7], destacando su importancia en la toma de decisiones tácticas y administrativas. Se presentan ejemplos de cómo equipos como el FC Barcelona han implementado tecnologías innovadoras para mejorar su rendimiento y estrategias, tales como WIMU, un sistema que mide el rendimiento físico de los jugadores durante entrenamientos y partidos, proporcionando datos sobre diferentes aspectos del desempeño.

También se destaca la necesidad de adaptarse a nuevas tendencias como los Esports. Pese a todo también se hace hincapié que, aunque los datos son cruciales, el factor suerte también juega un papel en el deporte, lo que hace que un análisis exhaustivo sea esencial para el éxito de las organizaciones deportivas [7].

En este mismo contexto, se destaca la creciente importancia de los sistemas de calificación para medir el rendimiento de los jugadores en el fútbol. A través de un análisis exhaustivo de datos provenientes de diversas fuentes, como OPTA [8] y whoscored.com [9], se abordan desafíos asociados a la recopilación, limpieza e integración de datos en el ámbito de la evaluación del rendimiento [10]. El estudio subraya que, aunque existe una abundancia de información disponible sobre el fútbol, la calidad y el detalle de los datos pueden variar

significativamente, lo que puede afectar la precisión de los modelos analíticos. Por lo tanto, se acentúa la necesidad de desarrollar conjuntos de alta calidad y de aplicar métodos de análisis adecuados para obtener hallazgos fundamentados en el rendimiento de los jugadores, contribuyendo así a una mejor toma de decisiones en el ámbito deportivo.

Encontramos también un análisis integral sobre el impacto de la analítica de datos en el fútbol profesional [11]. En esta obra se detalla cómo la implementación de esta tecnología no solo mejora el rendimiento deportivo de los clubes, sino que también optimiza la toma de decisiones administrativas. A través de casos de estudio, se evidencia la utilidad de la analítica para identificar talentos, analizar rivales y gestionar recursos de manera eficiente, y se comentan casos de éxito de equipos que utilizaron técnicas de este tipo. Uno de los ejemplos que menciona es el caso de éxito del Sevilla CF, equipo que, bajo la administración directiva de Ramón Rodríguez Verdejo, alias "Monchi", utilizó análisis estadístico para identificar talentos, comprarlos a un coste bajo y venderlos a equipos de mayor nivel a precios altos. El impacto económico de estas prácticas fue un superávit de 118 millones de euros en fichajes desde la temporada 2009/10 hasta la 2018/19, además del alzamiento de varios títulos de UEFA Europa League por parte del club deportivo.

En definitiva, vemos que la analítica de datos se ha convertido en una herramienta fundamental para los equipos de fútbol a la hora de tomar decisiones informadas.

## 2.2. Valoración de futbolistas

Mediante la revisión de las obras, observamos que la valoración de los jugadores es un proceso complejo que involucra múltiples variables, tanto cuantitativas como cualitativas, incluyendo la edad, los minutos jugados o la posición entre otras, así como lo difícil que puede llegar a ser el proceso de extraer datos como detalles contractuales y cuantías de fichajes debido a su escasa disponibilidad, lo que dificulta aún más el trabajo de análisis [12].

Una gran cantidad de obras con este propósito utilizan los valores de mercado de Transfermarkt como referencia para entrenar modelos. Estas estimaciones no siempre son fiables, ya que suelen subestimar los valores de mercado, especialmente en el caso de los jugadores más valiosos, por lo que deben usarse con precaución en investigaciones científicas y deportivas, así como en negociaciones contractuales [13]. Parece preferible utilizar precios de fichajes reales, tal y como haremos en este trabajo. La mayoría de los estudios que tratan de estimar valores de mercado de futbolistas se centran en variables clásicas, tales como goles, asistencias, edad y minutos jugados.

Se ha intentado asimismo predecir estos valores de mercado mediante un análisis ANOVA a los jugadores de la liga desde la temporada 18/19 a la temporada 21/22, incluyendo un total de 1720 futbolistas en la investigación, los cuales se dividieron en defensas, centrocampistas y delanteros a la hora de realizar el análisis. Se determinó que la edad, los minutos jugados y la posición final del equipo en la temporada influían de forma notable en el valor de mercado de los futbolistas estudiados. También se concluyó que había variables de rendimiento como goles, asistencias y acciones defensivas que son indicadores clave [14]. Se determinó también que los delanteros suelen tener un mayor valor de mercado, debido a su mayor impacto en el juego, así como que el historial de lesiones no influye a la hora de tasar a un futbolista. En el estudio se usan valores de transfermarkt como referencia para estimar el valor de mercado de los futbolistas, y en las conclusiones se hace hincapié en que sería interesante utilizar modelos más complejos para el propósito de estimar valores de mercado.

Podemos observar el uso de técnicas como regresión LASSO para analizar variables que evalúan el desempeño de los jugadores en función de su posición en el campo, con el objetivo de dilucidar cuáles son las que influyen en el valor de mercado de los futbolistas [15], aunque con una muestra limitada, de tan solo 37 futbolistas, que dificulta extraer conclusiones precisas sobre las variables analizadas, que eran tanto de rendimiento deportivo (tiros a puerta, goles,

asistencias, etc.) como características del jugador (edad, posición, altura, peso, etc.).

Aun con estas limitaciones, se concluyó que las variables con más impacto en el valor de mercado son tiros y goles dentro del área para delanteros, asistencias, pases completados y recuperaciones de balón para centrocampistas, intercepciones y despejes para defensas y paradas y goles encajados para porteros. Se concluyó también que existe una relación positiva entre el rendimiento y el valor de mercado, y por último que los jugadores de ataque son más valorados y tienen un impacto crucial en el rendimiento del equipo [15].

Otros trabajos han utilizado modelos de regresión multinivel para estimar los valores de mercado de los jugadores, técnica que consiste en agrupar los datos en distintas jerarquías. Para llevarla a cabo se anidan los futbolistas dentro de equipos en primer lugar, y luego los equipos en ligas, etc., y se entrena un mismo modelo con todas estas agrupaciones. Para entrenar el modelo se usó un dataset de 4217 individuos, utilizando su valor de mercado de Transfermarkt para entrenar los modelos, y se incluyeron variables relativas a características del jugador (edad, altura y pie dominante), a su rendimiento (goles, asistencias, minutos jugados, etc.) y de popularidad, tales como visualizaciones en wikipedia o menciones en redes sociales [16]. Mediante el uso Akaike Information Criterion (AIC), una métrica que indica si el modelo tiene un buen equilibrio entre precisión y simplicidad para evitar el sobreajuste, se observa que el modelo mejora de forma significativa al añadir un nuevo bloque de variables, permitiéndonos concluir que existen variables relevantes de todo tipo. Como aspecto negativo se utilizan valores de mercado de Transfermarkt, de cuya imprecisión hemos hablado anteriormente.

En otras obras observamos el uso de modelos econométricos como mínimos cuadrados ordinarios (OLS) o mínimos cuadrados generalizados (GLS) para estimar el valor de mercado de 150 de los delanteros más valiosos del momento en abril de 2015 basándose en métricas de rendimiento. Como valor de mercado se utilizó el establecido en Transfermarkt, basados en opiniones de terceros en lugar de en fichajes reales, y se contaba con 14 variables para entrenar los modelos, incluyendo factores como partidos jugados, goles o asistencias durante la temporada en la que los datos fueron recopilados. [17]. El mejor de ellos obtuvo un R-cuadrado ajustado de 0.57, por lo que explicaba un 57% de la variabilidad del modelo. Como contrapartida, además de utilizar solamente una posición hace uso de valores de mercado de Transfermarkt.

Hemos podido encontrar en otros trabajos pruebas como el Kruskal-Wallis y el Mann-Whitney U test, la cuales han sido utilizadas para comparar las diferencias en los valores de transferencia entre diferentes grupos de jugadores [18], divididos por edades, posiciones, etc. Mediante estos métodos se determinó que los delanteros suelen tener un mayor valor de mercado comparados con otros tipos de jugadores, así como que la edad del jugador y su rendimiento son también datos influyentes a la hora de determinar cuánto se pagará por él. Sin embargo, el número de individuos era de 108 únicamente, lo que puede haber sido perjudicial a la hora de extraer conclusiones.

Asimismo, en otras obras, la regresión múltiple ha sido usada como herramienta para el valor de mercado de los futbolistas en posiciones de ataque, analizando factores físicos y de rendimiento [19]. Mediante este método se ha logrado desarrollar un modelo útil que indicaba que factores como la edad, la estatura y la nacionalidad influyen en el valor de mercado, mientras que el número de tarjetas recibidas no tiene un impacto significativo. El reducido número de individuos, de tan solo 100, es un factor que puede haber influido de forma negativa a la hora de captar relaciones entre variables, aunque no es el único inconveniente que presenta esta obra. También vemos que usa valores de mercado de Transfermarkt para estimar el valor de los futbolistas, lo cual no es una buena práctica tal y como hemos visto anteriormente debido a la escasa fiabilidad de estas estimaciones.

## 2.3. Crítica al estado del arte

Vemos que el complejo desafío de determinar el valor de mercado de los futbolistas ha provocado un intenso interés académico, dando lugar a una amplia y diversa colección de trabajos de investigación que abordan esta problemática desde múltiples ángulos, explorando distintas variables y proponiendo una variedad de enfoques metodológicos.

Las técnicas utilizadas hasta el momento para determinar qué variables influyen, y en qué medida, en el valor de mercado de los futbolistas son variadas y, en algunos casos, sofisticadas. No obstante, una revisión crítica de la literatura muestra que muchas de estas investigaciones presentan limitaciones significativas que conviene tener en cuenta. Una de las principales es la dependencia casi exclusiva de los valores estimados por el portal Transfermarkt como proxy del valor de mercado.

Si bien este sitio web es una referencia ampliamente utilizada en la industria y en trabajos académicos debido a su accesibilidad y amplitud de datos, no deja de ser una estimación basada en criterios parcialmente subjetivos y que no necesariamente reflejan el valor real de mercado en las transacciones, tal y como hemos mencionado anteriormente.

Además, se observa que muchos de estos estudios se basan en muestras de tamaño reducido, lo que limita la capacidad de generalizar los resultados. En otros casos, el análisis se restringe a delanteros, lo cual impide tener una visión integral del mercado futbolístico y reduce el alcance del estudio a contextos específicos.

En contraposición a estas limitaciones, en nuestro trabajo se propone una metodología diferente. Nos centraremos en fichajes reales, es decir, operaciones de traspaso que efectivamente se han producido en el mercado, lo que nos permitirá utilizar datos concretos y verificables sobre los valores de transferencia, lejos de basarnos en estimaciones que, tal como se ha visto en la literatura revisada, no siempre son precisas. Además, se incluirán futbolistas de distintas posiciones, abarcando así una gama representativa del espectro profesional.

Asimismo, se utilizará una muestra suficientemente amplia como para permitir la extracción de conclusiones robustas y estadísticamente significativas. Con ello, buscamos mejorar la validez externa de los resultados y ofrecer un análisis más completo y realista del valor de mercado de los jugadores en función de distintas variables.

### 3. Metodología

---

A continuación, resumiremos los pasos a seguir de cara a obtener modelos explicativos que estimen el valor de mercado de futbolistas.

En primer lugar, extraeremos datos sobre traspasos de futbolistas que se hayan llevado a cabo estos últimos años. Para ello, conectaremos los perfiles de transfermarkt de cada uno de ellos con su perfil de fbref. Del primero obtendremos sus traspasos, del segundo su rendimiento en el momento de los fichajes.

También obtendremos el rendimiento de futbolistas con contrato en vigor, de cara a que los modelos entrenados puedan estimar sus valores de mercado. Esto se hará mediante sus perfiles de transfermarkt y fbref conectados anteriormente, extrayendo de transfermarkt la duración restante de contrato y de fbref sus métricas de rendimiento y su edad.

Una vez obtenidos estos datos se hará una limpieza de los mismos, así como un análisis exploratorio, todo ello de cara a tratar datos faltantes o posibles problemas de multicolinealidad.

Tras hacer una selección de modelos explicativos, se entrenarán usando estos datos y se evaluará el rendimiento de cada uno de ellos, obteniendo posteriormente conclusiones de cuáles han funcionado mejor y determinando cuáles de las variables extraídas han sido más importantes en el proceso.

## 4. Conjuntos de datos

---

Este capítulo detalla el proceso que hemos seguido para la obtención, tratamiento y análisis de los datos que servirán para entrenar nuestros modelos. Abordaremos desde la estrategia inicial para conectar diversas fuentes de información futbolística hasta los pasos específicos para extraer la información clave para el propósito de crear modelos explicativos.

Finalmente, realizaremos un análisis exploratorio exhaustivo, tanto a nivel univariante como multivariante, con el objetivo de comprender a fondo las características de nuestros datos y las relaciones existentes entre ellos, lo cual es crucial para la construcción y la posterior interpretación de los modelos de valoración de jugadores.

### 4.1. Obtención de datos

El primer paso para crear los modelos será conseguir el conjunto deseado de datos, el cual consistirá en un excel en el que cada fila contenga las estadísticas de un jugador para una temporada en la que ha sido traspasado, en la cual incluiremos métricas relativas al rendimiento de la propia temporada y totales acumulados entre esta y las anteriores que haya disputado el jugador, así como el coste del fichaje e información sobre la duración restante de su contrato y su edad en el momento del traspaso.

Para obtener el conjunto de datos deseado indagaremos en internet para encontrar bases de datos con estadísticas variadas sobre jugadores de fútbol, desde goles y asistencias hasta entradas o intervenciones exitosas, así como títulos a lo largo de su carrera. En la página fbref [20] encontramos estadísticas de este tipo para todos los jugadores, seccionadas por competición y temporada, tanto datos totales como de media cada 90 minutos.

También incluye métricas como pases progresivos y goles esperados, aunque están disponibles solamente a partir de la temporada 2017-2018, y únicamente en las cinco grandes ligas y en champions league, encontrándose ausentes en otras ligas y torneos, incluyendo la copa del rey, la supercopa de europa, etc. A nivel de selecciones, por su parte, incluye estadísticas como pases progresivos y goles esperados en las principales competiciones, es decir, en la eurocopa, en el mundial y en la copa américa, sin contar las fases previas que determinan qué equipos clasifican a estas competiciones.

Para obtener información de los traspasos de los futbolistas a analizar usaremos transfermarkt [5], plataforma que proporciona información detallada sobre los valores de mercado de los jugadores, fichajes, etc.

#### 4.1.1. Conexión entre transfermarkt y fbref

Usaremos la librería worldfootballR [21] para realizar scrapping en la web de fbref, portal web cuyas condiciones de uso lo permite con limitaciones relativas al número de solicitudes por minuto. Obtendremos la URL y la posición de los jugadores que vayamos a analizar.

Podemos extraer toda esta información mediante la función `fb_big5_advanced_season_stats` de la librería worldfootballR, la cual devuelve información de futbolistas de las cinco grandes ligas para una temporada y un tipo concreto de estadística que se le pasa como parámetro. La ejecutaremos para distintos años con el objetivo de extraer las URLs de todos los futbolistas que han participado en las cinco grandes ligas desde la temporada 2017-2018, las cuales se irán añadiendo junto a la posición de cada jugador a un

array asociativo con el objetivo de comprobar que contengan información.

Una vez que tengamos estas URLs, y con el objetivo de dar importancia a jugadores que no participen en las cinco grandes ligas, incluiremos también a aquellos que hayan jugado esta última temporada en los principales equipos de otros países, tales como Argentina, Portugal, etc. También se incluirán jugadores de equipos que hayan disputado la uefa champions league esta temporada 2024-2025, para lo cual tendremos que usar la función `fb_team_player_stats`, que devuelve información sobre un equipo y un tipo de estadística concreta para la presente temporada introduciendo la URL del equipo en cuestión. Por ahora nos quedaremos únicamente con los valores de las columnas deseadas, que serán la URL y la posición de cada jugador.

Al igual que en la extracción anterior, añadiremos la URL y la posición de cada jugador, con el objetivo de comprobar posteriormente si la URL tiene contenido para analizar. Para la liga argentina será algo más complicado, dado que no aparece la plantilla actual de sus equipos en `fbref` al entrar en la sección de estadísticas de los clubes. Esto puede deberse al hecho de que acaba de terminar la presente temporada de la liga profesional de argentina, y provocará que tengamos que buscar a los jugadores uno por uno en la web para encontrar su URL y su posición.

Al extraer todos los datos anteriormente mencionados y realizar una ligera inspección, nos encontramos con que uno de los jugadores extraídos no tiene posición asignada. Al buscarlo en la página web tampoco nos aparece, por lo que no podemos extraer sus estadísticas, por lo que terminamos optando por eliminarlo de la lista de jugadores a analizar. Una vez realizado el procedimiento anterior obtenemos un listado de un total de 7027 jugadores que se usarán para entrenar y validar el modelo.

El siguiente objetivo consistirá en enlazar de algún modo estos enlaces con enlaces a transfermarkt, con la finalidad de poder enlazar las estadísticas de cada jugador con la cantidad de dinero que se pagó por ellos en cada movimiento de mercado. Para ello ejecutaremos la función `player_dictionary_mapping` perteneciente a la librería `worldfootballR`, que nos devolverá el enlace de transfermarkt y el de `fbref` para cada jugador que haya militado en las cinco grandes ligas desde la temporada 2017-2018:

Al visionar el resultado obtenido, nos encontramos con un resultado diferente al esperado, y es que el dataframe resultante cuenta con 14872 observaciones, cuando nosotros habíamos obtenido únicamente 6571 futbolistas al extraer las plantillas de los equipos que habían participado en las cinco grandes ligas desde la temporada 2017-2018.

Al profundizar en esta cuestión, nos encontramos con que muchos de los futbolistas incluidos en el dataframe no han militado en las cinco grandes ligas en ningún momento. Además, alguno de los futbolistas que han participado en alguna de las cinco grandes ligas, como es el caso de Gabriel Magalhaes o Samu Aghehowa, no están incluidos en este dataframe. Estos futbolistas, unidos a aquellos que queremos analizar fuera de las cinco grandes ligas, hacen un total de 468 futbolistas a analizar que no tenemos enlazados con su perfil de transfermarkt.

El siguiente paso será extraer URLs de transfermarkt correspondientes a jugadores de equipos externos a las cinco grandes ligas para extraer los nombres de las URLs y compararlos con los nombres de los jugadores faltantes. Nos encontramos con que 248 de esos jugadores sí que residen en los datos extraídos anteriormente, pero hay 220 que siguen ausentes.

Tras indagar en las funciones para hacer scrapping, nos encontramos con que ninguna de ellas nos ayudaría a relacionar los jugadores ausentes con su URL de transfermarkt, por lo que tendremos que añadirlos a mano. Una vez terminado este procedimiento, revisamos el dataframe resultante para asegurarnos de que no hay URLs repetidas, así como para renombrar a aquellos futbolistas cuyo nombre aparezca repetido.

Observamos dos jugadores apodados “Jota”, otros dos llamados “Wallace”, tres llamados “Guilherme”, dos conocidos como “Aaron Ramsey” y así hasta alrededor de 50 nombres repetidos. Esto ha sido especialmente problemático con dos jugadores llamados “Alessandro Russo”, ya que para ninguno de los dos aparece un segundo apellido que ayude a distinguirlos en ninguna web.

Tampoco ha sido posible buscar fichas técnicas de partidos disputados por los mismos, por lo que se ha optado por añadir su posición al final de sus nombres, quedando como “Alessandro Russo (GK)” y “Alessandro Russo (CM)”. Algo parecido ocurre en el caso de dos futbolistas llamados “Ibrahim Cissé”, ambos sin segundo apellido y que además juegan en la misma posición (defensa central). Al final se ha decidido diferenciarlos por el código ISO de sus países de origen, quedando como “Ibrahim Cissé (FRA)” el francés e “Ibrahim Cissé (CIV)” el costamarfileño.

También observamos un total de 6 enlaces repetidos en el dataframe resultante, fruto de erratas a la hora de añadir enlaces de transfermarkt de jugadores con nombres parecidos, tales como Frederik Sorensen y Frederik Winther. Se procedió a buscar sus enlaces reales en transfermarkt para corregir el dataframe, así como a cambiar las posiciones de dichos jugadores, ya que también eran erróneas.

También se comprobaron URLs repetidas en el conjunto de datos original para corregir aquellas que no fueran correctas en el nuestro. Vemos también que un jugador llamado William Vick, perteneciente a la plantilla del Sturm Graz, no aparece en transfermarkt, por lo que no podremos contar con él para entrenar el modelo.

Procederemos ahora a visualizar los valores para la columna que indica la posición de los futbolistas, con el objetivo de detectar errores. Vemos jugadores cuya columna de posición contiene valores como “attack” o “NA”. Procedemos a entrar en sus perfiles y corregir estos valores.

A continuación, revisaremos que no existan caracteres extraños, encontrándonos con que varios jugadores tienen cambiados caracteres con tildes y diéresis por otros caracteres extraños. Una vez sustituidos estos caracteres por los correspondientes en cada caso, tenemos listo el dataframe que conecta las URLs de transfermarkt y fbref, por lo que podemos empezar a extraer datos de estas páginas.

#### 4.1.2. Adquisición de datos

Procedemos a extraer datos de fichajes las páginas web anteriormente mencionadas, haciendo distinción entre porteros y jugadores de campo, ya que irán en dataframes distintos, incluyendo únicamente movimientos a partir de la temporada 2019-2020, dado que obtendremos una mayor cantidad de información para estos jugadores.

Para cada traspaso que encontramos en transfermarkt para un jugador, extraeremos estadísticas acumuladas de las tres temporadas anteriores al fichaje del mismo. Adicionalmente, para las estadísticas principales (goles, asistencias, etc.) obtendremos datos de la última temporada y de grandes competiciones, además de los acumulados de las tres últimas. Para porteros, estas estadísticas serán goles y tiros a puerta en contra, y para todo tipo de jugadores incluiremos partidos y minutos jugados, además de titularidades, entre las estadísticas principales a analizar. Nos encontramos con que algunos jugadores cuentan con una gran cantidad de datos faltantes, debido principalmente a que muchos de ellos juegan en ligas con poco reconocimiento.

A continuación, obtendremos información sobre contratos en vigor de los jugadores de nuestra base de datos. De aquellos que tengan uno extraeremos estadísticas de las últimas tres temporadas, de cara a que los distintos modelos que entrenemos estimen sus valores de

mercado teniendo en cuenta estas estadísticas. Una vez hecho esto, aplicamos el mismo procedimiento en jugadores con contrato vigente, con la idea de extraer sus estadísticas de las últimas tres temporadas y poder estimar sus valores de mercado a partir de los modelos entrenados.

## 4.2. Tratamiento de los datos

Una vez hemos extraído los datos requeridos para entrenar el modelo, el siguiente paso será proceder con la limpieza de datos, centrándonos en el tratamiento de datos nulos, faltantes y erróneos, y en la posterior transformación de los datos, convirtiendo en dummy las variables categóricas. Empezamos con el dataset que contiene fichajes de jugadores de campo, y observamos que 116 de esos fichajes no tienen información sobre partidos jugados. Al analizarlos en profundidad, vemos que no tienen información de ninguna variable, por lo que los eliminamos de la base de datos.

Una vez eliminados, observamos que ciertas variables cuentan con una gran cantidad de datos faltantes. Es el caso de bloqueos y ocasiones de gol creadas la última temporada o porcentaje de duelos aéreos ganados, las cuales cuentan con hasta 1166 individuos en los que su valor no se encuentra presente. Estas variables se eliminan del dataset, junto a otras con más de 270 faltantes, son eliminadas del dataset.

También son eliminados un total de 92 individuos con 5 o más datos faltantes. El resto de datos faltantes han podido ser inferidos investigando en otras páginas web de contenido futbolístico, tales como besoccer o footystats. En el caso de los porteros, se eliminaron variables con más de 50 faltantes e individuos con más de 5.

Al realizar comprobaciones de comparación de variables para comprobar que no haya datos erróneos (más tiros a puerta que goles, más asistencias que ocasiones creadas, etc.), comprobamos que el conjunto de datos no presenta problemas de este tipo, por lo que damos la limpieza por finalizada.

Vamos ahora con la transformación, creando variables dummy a partir de nuestras variables categóricas, que son el equipo que abandona el futbolista en el momento del fichaje, el equipo en el que recalca, los países de ambos equipos y la posición del jugador. Dado que las variables relativas a equipos y países cuentan con un gran número de categorías que provocarían un ruido inmenso en caso de ser incluidas en su totalidad, se opta por incluir los principales países y equipos y marcar el resto como "Otros".

Esto nos deja un total de 73 variables que utilizaremos para entrenar los modelos, incluyendo la variable explicada, doce variables relativas al rendimiento, una a la duración de contrato restante y 59 variables dummy, cuyos valores base serán "Otros" en caso de variables relativas al origen y destino del fichaje y "Right-back" en el caso de la posición.

Centraremos y escalaremos las variables para evitar inconvenientes relativos a la diferencia de escalas de las mismas, pero no aplicaremos transformaciones para paliar la asimetría pese a la pérdida de rendimiento que puede causar en los modelos, ya que de hacerlo perderíamos explicabilidad en los mismos.

## 4.3. Descripción de los datos

En el caso de los jugadores de campo contamos con un dataset resultante de 3151 individuos y 73 variables, mientras que en el caso de los porteros tenemos un dataset distinto con un total de 185 individuos y 36 variables a analizar.

Ambos grupos tienen variables en común, y son las siguientes:

- Name: Contiene el nombre del futbolista.
- Exp\_contr: Indica los meses que le quedaban de contrato al futbolista en el momento de su fichaje.
- Team\_from: Equipo del que salió el futbolista en el momento del fichaje.
- Team\_to: Equipo al que fue a parar el futbolista en el fichaje en cuestión.
- Transfer\_value: Cuantía que se pagó por el futbolista.
- Position: Posición del futbolista.
- Year: Año del fichaje.
- Age: Edad del futbolista en el momento del fichaje.
- Matches\_pl: Partidos jugados en total durante las últimas tres temporadas.
- Matches\_pl\_LS: Partidos jugados en total durante la última temporada.
- Matches\_pl\_BC: Partidos jugados en grandes competiciones durante las últimas tres temporadas.
- Matches\_pl\_BC\_LS: Partidos jugados en grandes competiciones durante la última temporada.
- Starts: Partidos en los cuales el futbolista ha sido titular durante las últimas tres temporadas.
- Starts\_LS: Partidos en los cuales el futbolista ha sido titular durante la última temporada.
- Starts\_BC: Partidos en los cuales el futbolista ha sido titular en grandes competiciones durante las últimas tres temporadas.
- Starts\_BC\_LS: Partidos en los cuales el futbolista ha sido titular en grandes competiciones durante la última temporada.
- Minutes\_pl: Minutos jugados durante las últimas tres temporadas.
- Minutes\_pl\_LS: Minutos jugados durante la última temporada.
- Minutes\_pl\_BC: Minutos jugados en grandes competiciones durante las últimas tres temporadas.
- Minutes\_pl\_BC\_LS: Minutos jugados en grandes competiciones durante la última temporada.
- Assists: Asistencias proporcionadas en total durante las últimas tres temporadas.
- Yellow\_cards: Tarjetas amarillas recibidas durante las últimas tres temporadas.
- Red\_cards: Tarjetas rojas recibidas durante las últimas tres temporadas.

También tenemos variables que solamente están presentes en el dataset de jugadores de campo, como pueden ser:

- NP\_goals: Goles totales anotados por un jugador durante las últimas tres temporadas sin recurrir al penalti.
- NP\_goals\_LS: Goles anotados por un jugador durante la última temporada sin recurrir al penalti.
- NP\_goals\_BC: Goles anotados por un jugador durante las últimas tres temporadas sin recurrir al penalti en grandes competiciones.
- NP\_goals\_BC\_LS: Goles anotados por un jugador durante la última temporada y en grandes competiciones sin recurrir al penalti.
- Pen\_goals: Goles totales de penalti anotados por un jugador durante las últimas tres temporadas.
- Pen\_goals\_LS: Goles totales de penalti anotados por un jugador durante la última temporada.
- Pen\_goals\_BC: Goles totales de penalti anotados por un jugador durante las últimas tres temporadas en grandes competiciones.
- Pen\_goals\_BC\_LS: Goles totales de penalti anotados por un jugador durante la última temporada en grandes competiciones.
- Assists\_LS: Asistencias proporcionadas en total durante la última temporada.
- Assists\_BC: Asistencias proporcionadas en total durante las últimas tres temporadas en grandes competiciones.
- Assists\_BC\_LS: Asistencias proporcionadas en total durante la última temporada en grandes competiciones.
- Shoots: Tiros realizados por el jugador durante las últimas tres temporadas.
- Shoots\_on\_target: Tiros con dirección a la portería contraria realizados por el jugador durante las últimas tres temporadas.
- Tackles\_won: Número de entradas en las que el jugador ha impedido con éxito que uno de sus rivales avance conduciendo el balón durante las últimas tres temporadas.
- Interceptions: Anticipaciones en las que un jugador recupera la posesión del balón para su equipo.
- GCA\_BC: Ocasiones de gol creadas en grandes competiciones durante las últimas tres temporadas.
- GCA\_BC\_LS: Ocasiones de gol creadas en grandes competiciones durante la última temporada.

Asimismo, contamos con variables que solamente se encuentran en el dataset de porteros, como pueden ser:

- GA: Goles recibidos en total durante las últimas tres temporadas.

- GA\_LS: Goles recibidos en total durante la última temporada.
- GA\_BC: Goles recibidos durante las últimas tres temporadas en grandes competiciones.
- GA\_BC\_LS: Goles recibidos durante la última temporada en grandes competiciones.
- SoTA: Tiros a puerta recibidos en total durante las últimas tres temporadas.
- SoTA\_LS: Tiros a puerta recibidos en total durante la última temporada.
- SoTA\_BC: Tiros a puerta recibidos durante las últimas tres temporadas en grandes competiciones.
- SoTA\_BC\_LS: Tiros a puerta recibidos durante la última temporada en grandes competiciones.
- CS: Conocidas como “cleansheets” en inglés, se refiere al número de partidos en los cuales un portero no recibió ningún gol durante las últimas tres temporadas.
- PKA: Penaltis recibidos por el portero durante las últimas tres temporadas.
- PKSv: Penaltis parados por el portero durante las últimas tres temporadas.
- Goals: Goles anotados por el portero durante las últimas tres temporadas.

## 4.4. Análisis exploratorio de datos

Desarrollar modelos de predicción del valor de mercado de futbolistas mediante técnicas de aprendizaje automático requiere un conocimiento previo de las distintas variables que vamos a utilizar. Es esencial realizar un análisis exploratorio de datos para comprender la estructura y características del conjunto de datos, identificar patrones, detectar valores atípicos y evaluar la calidad de la información disponible.

### 4.4.1 Análisis univariante

Antes de construir modelos de aprendizaje automático es fundamental comprender la naturaleza de cada variable de forma individual. Por ello debemos explorar sus principales características, tales como su distribución, tendencia central y dispersión, con la idea de detectar patrones que podrían tener un impacto en el rendimiento de los modelos predictivos.

Empezaremos con la edad de los futbolistas a la hora de ser traspasados de un equipo a otro.

count	3151.000000	count	185.000000
mean	24.466519	mean	26.405405
std	3.425869	std	3.867784
min	16.000000	min	18.000000
25%	22.000000	25%	23.000000
50%	24.000000	50%	26.000000
75%	27.000000	75%	29.000000
max	36.000000	max	35.000000
Name: Age, dtype: float64		Name: Age, dtype: float64	

Figura 1: Estadísticas descriptivas de la edad. Jugadores de campo a la izquierda, porteros a la derecha

Observamos en la Figura 1 un promedio de edad de alrededor de 24.46 años para jugadores de campo y 26.4 para porteros, así como una mayor edad en todos los cuartiles para los guardametas, lo que responde al hecho de que los porteros por norma general tienen carreras más longevas que los jugadores de campo.

También observamos una mayor desviación estándar en los porteros, sugiriendo una mayor variabilidad en sus edades.

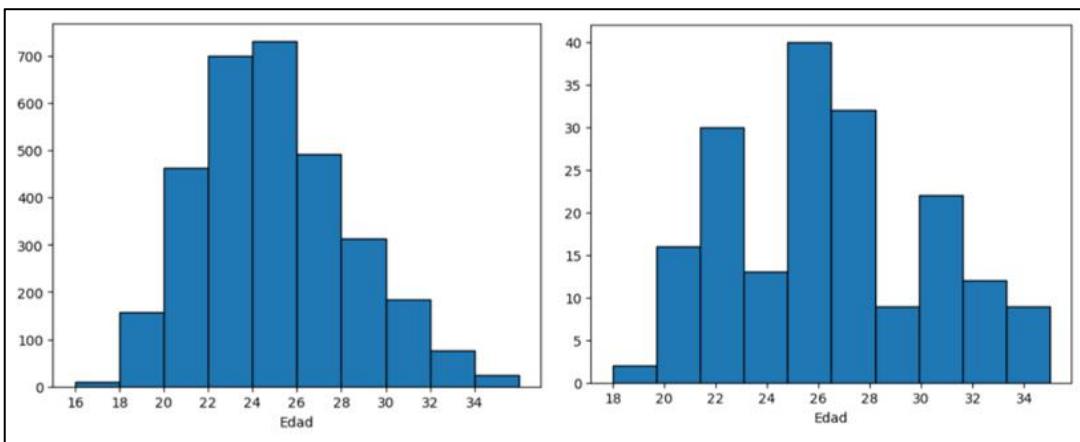


Figura 2: Histograma de edad. Jugadores de campo a la izquierda, porteros a la derecha

La Figura 2 muestra que la distribución de los jugadores de campo es asimétrica positiva, con la mayoría de jugadores concentrados entre los 20 y los 28 años y con una caída significativa después de los 30, lo que indica que los equipos suelen apostar por jugadores con proyección o bien por aquellos que estén en el pico de su carrera. Se observa un pico claro entre los 24 y los 26 años, momento en el que los jugadores suelen alcanzar su máximo rendimiento y llamar la atención de más clubes en consecuencia.

Observando el histograma de porteros vemos que los fichajes están más repartidos en cuanto a edad, con una mayor presencia entre los 26 y los 28 años. Esto nos da a entender que los clubes valoran más la experiencia en porteros que su proyección, a diferencia de lo que ocurre en jugadores de campo. Vamos con los costes de los distintos fichajes a continuación.

count	3,151	count	185
mean	9,231,057	mean	6,238,897
std	13,855,635	std	7,924,571
min	1,000	min	51,000
25%	1,500,000	25%	1,000,000
50%	4,000,000	50%	3,000,000
75%	11,100,000	75%	7,700,000
max	127,200,000	max	50,200,000
Name: Transfer_value, dtype: object		Name: Transfer_value, dtype: object	

Figura 3: Estadísticas descriptivas de los valores de los fichajes. Jugadores de campo a la izquierda, porteros a la derecha

Observamos en la Figura 3 que el coste medio de fichaje de un jugador de campo, así como todos sus cuartiles, superan a los de los porteros, debido a su mayor impacto y protagonismo durante los partidos.

La variabilidad es mayor también en jugadores de campo, así como el rango en el que varían los precios que se pagan por ellos, lo cual podría deberse a que hay más factores que pueden afectar a su precio que en el caso de los porteros.

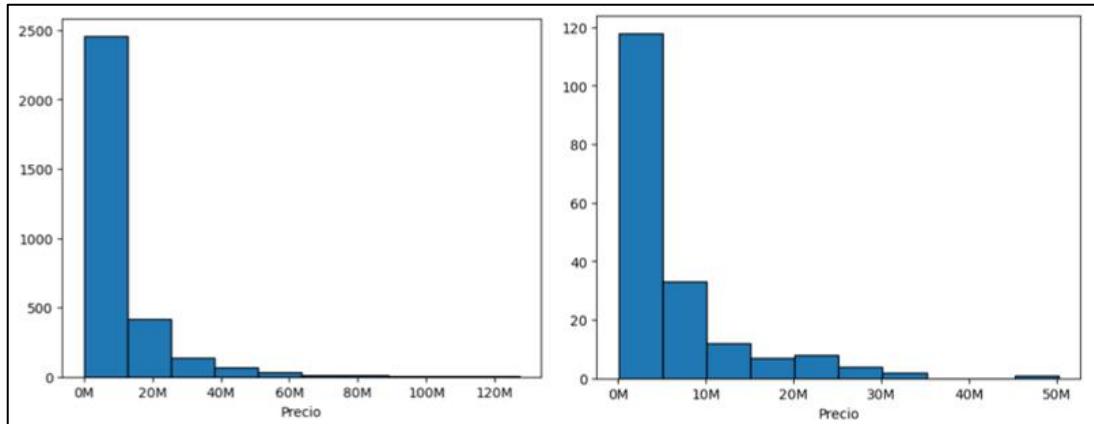


Figura 4: Histogramas de costes de fichaje. Jugadores de campo a la izquierda, porteros a la derecha

En la Figura 4 vemos que tanto en porteros como en jugadores de campo hay una distribución asimétrica positiva, con la mayoría de jugadores concentrados en valores bajos en cuanto a precios de fichaje. En el caso de los jugadores de campo sus fichajes se concentran mayoritariamente en precios menores que 20 millones de euros, mientras que en los porteros en precios inferiores a 10 millones.

Esto nos puede dar una idea de los segmentos que está el mercado de fichajes, con una pequeña porción de fichajes de grandes estrellas y una gran mayoría de jugadores cuyos precios se sitúan en rangos más modestos.

Vamos a continuación con la duración de contrato de los futbolistas.

count	3,151	count	185
mean	22	mean	22
std	12	std	12
min	3	min	9
25%	11	25%	11
50%	22	50%	22
75%	29	75%	24
max	67	max	59
Name: Exp_contr, dtype: object		Name: Exp_contr, dtype: object	

Figura 5: Estadísticas descriptivas de la duración restante de contrato (meses). Jugadores de campo a la izquierda, porteros a la derecha

En la Figura 5 observamos datos similares en la media y en la desviación típica. La única diferencia reside en los cuartiles, los cuales tienen un mayor valor para los jugadores de campo en los dos últimos y un menor valor en el primer cuartil, lo que indica que su rango de duraciones de contrato es mayor y que tendrán una mayor diversidad en esta variable.

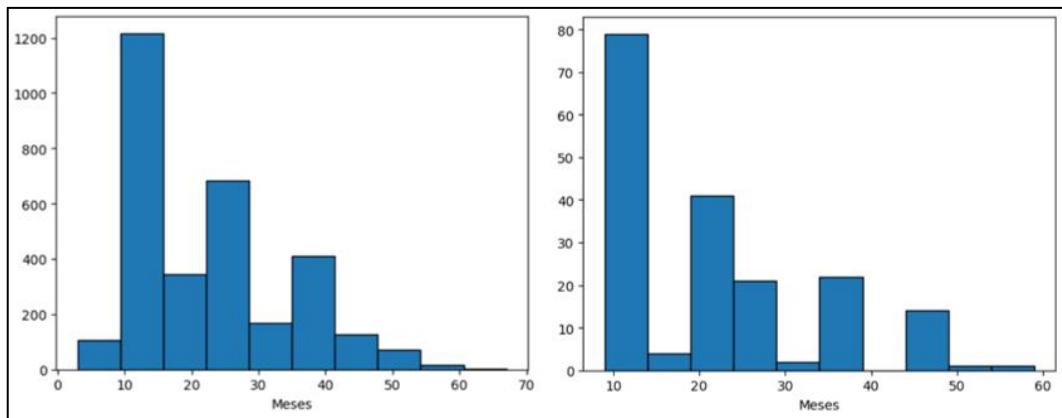


Figura 6: Histogramas de duración de contrato restante (meses). Jugadores de campo a la izquierda, porteros a la derecha

En la figura 6 podemos comprobar un pico pronunciado en duraciones cortas de contrato tanto en porteros como en jugadores de campo, lo que sugiere una posible estrategia de los clubes de aprovechar oportunidades para fichar jugadores cuyo contrato está cerca de expirar, probablemente para obtenerlos a un menor coste de traspaso.

También observamos picos menores en duraciones medias de contrato, especialmente en el caso de los porteros, que pueden deberse que los equipos que han acometido los fichajes lo han hecho por el potencial de los jugadores o bien por necesidad, sin importarles invertir más dinero en un jugador con una mayor duración restante de contrato.

Analicemos a continuación los fichajes realizados por año.

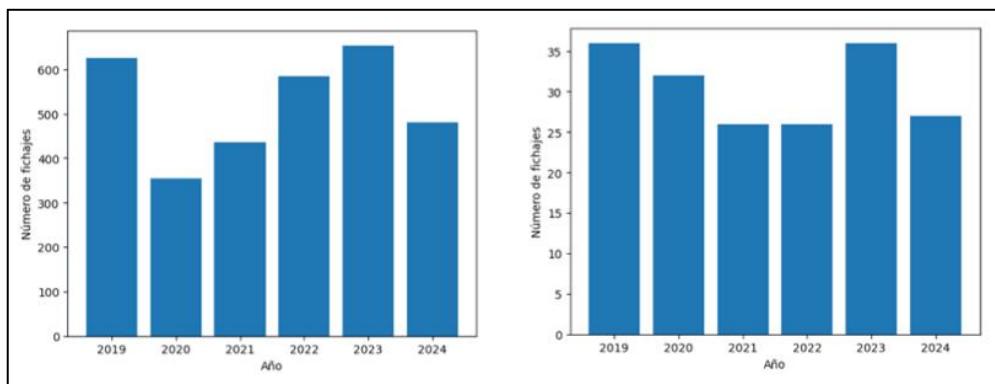


Figura 7: Número de fichajes por año. Jugadores de campo a la izquierda, porteros a la derecha

Si nos fijamos en la Figura 7 vemos un decrecimiento del número de fichajes tras el COVID-19, con un descenso considerable de 2019 a 2020, especialmente en los jugadores de campo, aunque se recupera la tendencia normal en todos los futbolistas conforme se va superando la pandemia.

El número de porteros traspasados se ha mantenido más estable en el tiempo que el de jugadores de campo, sufriendo en menor medida los efectos del COVID que en el caso de los jugadores de campo. Vamos ahora con los países en los que se encuentran jugando los futbolistas a la hora de ser traspasados.

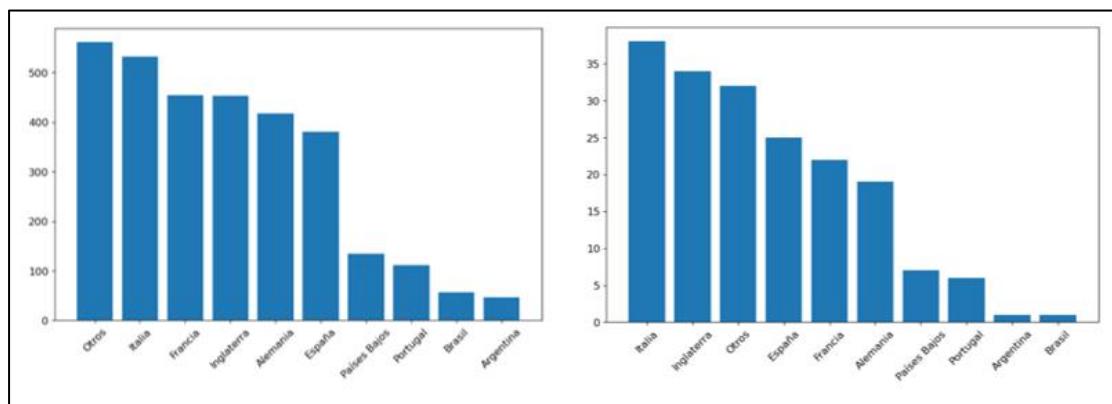


Figura 8: Número de fichajes por país del equipo de origen. Jugadores de campo a la izquierda, porteros a la derecha

Observamos en la Figura 8 que Italia, Francia, Inglaterra y España suelen ser los principales países de los equipos de los que proceden los futbolistas fichados, tanto porteros como jugadores de campo, aunque debemos tener en cuenta que muchos de ellos recalcan en equipos de esos mismos países. También hay una parte importante de futbolistas proveniente de países menos reconocidos por su tradición futbolística.

Analizamos a continuación los países a los que fueron a jugar los futbolistas tras ser traspasados.

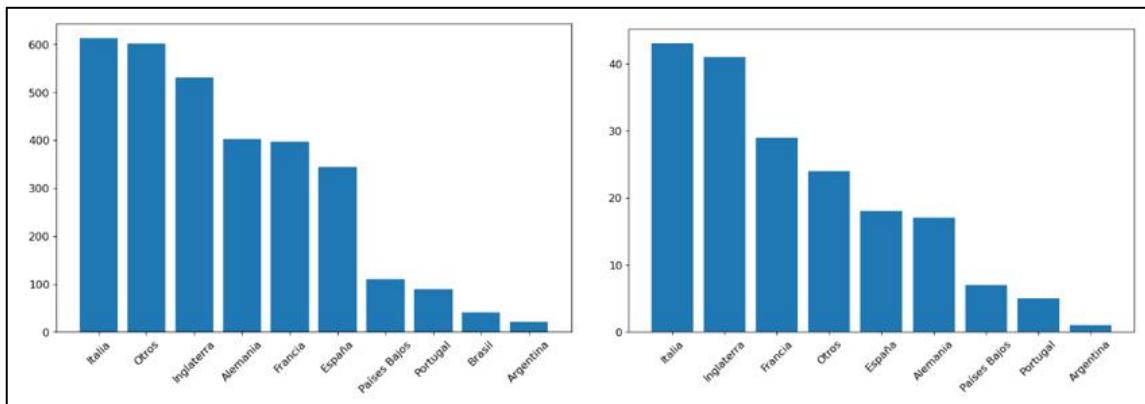


Figura 9: Número de fichajes por país del equipo de destino. Jugadores de campo a la izquierda, porteros a la derecha

Al igual que en los países de origen, los de destino son mayoritariamente Italia, Francia, Inglaterra y España, tal como se observa en la Figura 9. Vamos ahora con la posición de los jugadores.

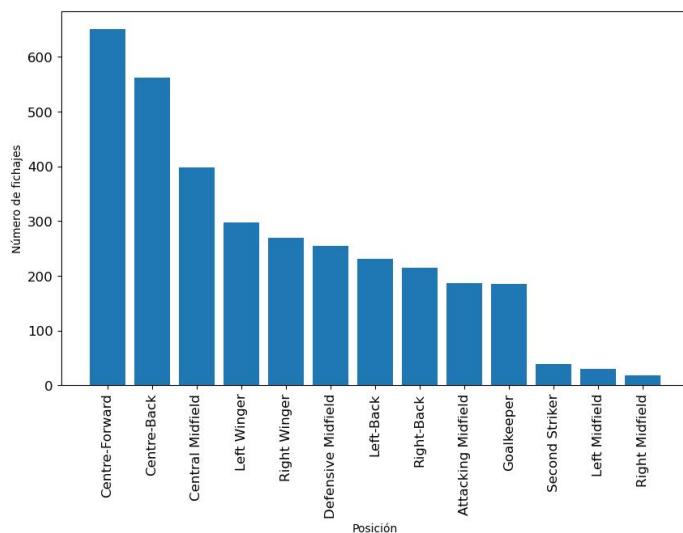


Figura 10: Número de fichajes por posición

En la Figura 10 vemos que, con un total de 650, los delanteros centro se consagran como los futbolistas con los que más operaciones se llevan a cabo, seguidos por los defensas centrales, lo cual podría explicarse debido a la importancia estratégica de estos jugadores y sus roles determinantes en ataque y en defensa, respectivamente. Se observa también un mayor número de traspasos por futbolistas que juegan por la mitad del campo que por aquellos que juegan por las bandas.

Las siguientes variables a analizar tendrán en consideración métricas en la última temporada (sufijo '\_LS'), en grandes competiciones (sufijo '\_BC') y en grandes competiciones durante la última temporada (sufijo 'BC\_LS'). A continuación, analizaremos los partidos jugados en jugadores de campo.

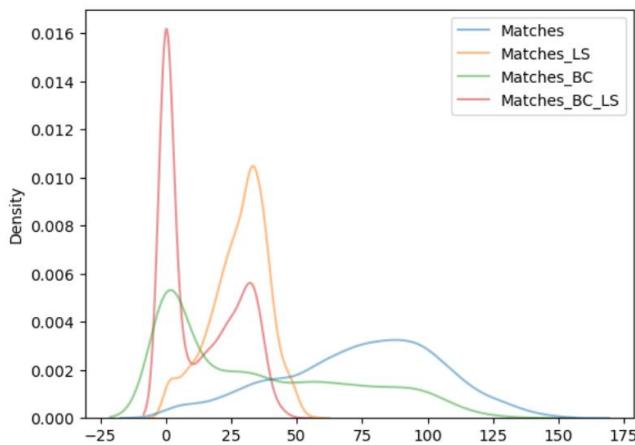


Figura 11: Gráfico de densidad de partidos jugados para jugadores de campo

Se puede vislumbrar en la Figura 11 una estructura más dispersa en los encuentros disputados en total, mientras que en el resto de variables existen asimetrías positivas con picos pronunciados. En las variables referidas a grandes competiciones vemos individuos concentrados cerca del cero, lo que nos indica que gran parte de los fichajes no ha disputado partidos en Champions league ni en las 5 grandes ligas durante las últimas temporadas.

En cuanto a partidos jugados la última temporada, vemos un pico mucho más pronunciado alrededor de los 30-40 partidos jugados. La curva cae rápidamente a ambos lados de este pico, indicando que la mayoría de los futbolistas fichados jugaron un número de partidos relativamente similar en la última temporada.

A continuación, estudiaremos las titularidades de los jugadores de campo.

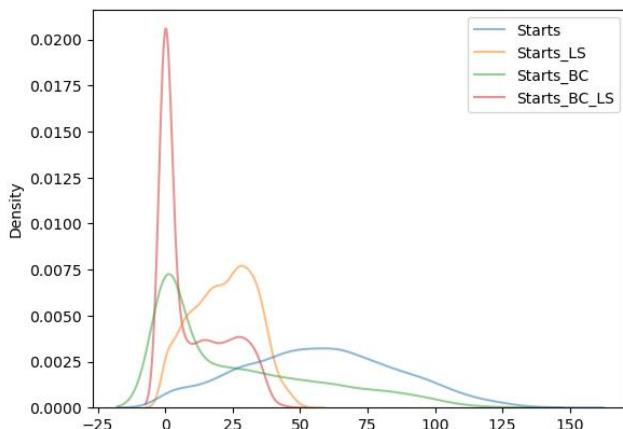


Figura 12: Gráfico de densidad de titularidades de jugadores de campo

En la Figura 12 vemos densidades muy parecidas a las de partidos jugados, con una estructura más dispersa en las titularidades totales, y picos pronunciados en el resto de variables, así como jugadores que no han disputado ningún partido como titulares en grandes competiciones o lo han hecho en escasas ocasiones durante las últimas temporadas.

Vamos a analizar los goles anotados sin recurrir al penalti.

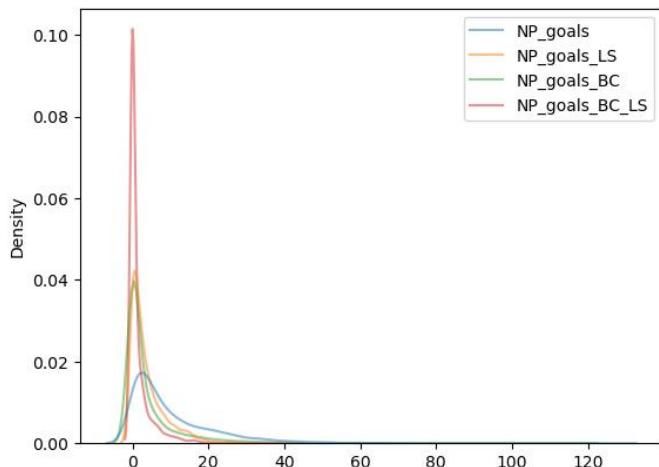


Figura 13: Gráfico de densidad de goles sin contar penaltis

Observamos en la Figura 13, tal como era de esperar, que las densidades de los goles muestran distribuciones asimétricas positivas con picos cercanos al cero, donde se encuentran la mayoría de futbolistas por los factores comentados en la explicación del gráfico anterior.

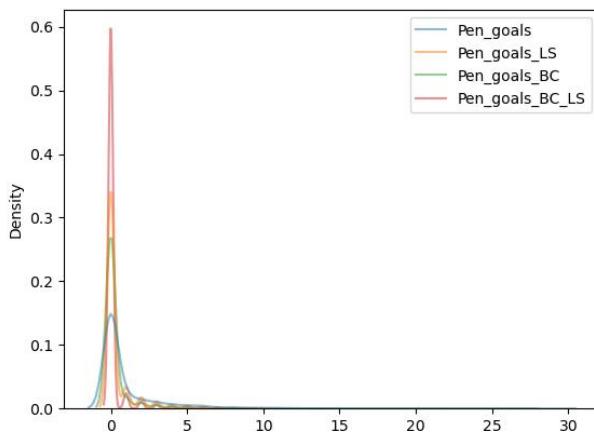


Figura 14: Gráfico de densidad de goles de penalti

Los resultados que vemos en la Figura 14 son los esperados, distribuciones asimétricas positivas con picos alrededor de cero en las cuatro densidades, mostrando una amplia mayoría de jugadores que apenas ha metido goles de penalti (o no lo han hecho directamente) y unos pocos que lo han hecho en más ocasiones. Es una tendencia similar a la anterior pero mucho más pronunciada, ya que en cada equipo habitualmente hay un encargado de tirar penaltis o unos pocos, provocando que el resto de futbolistas, conformando una gran mayoría, no lleguen ni siquiera a tirar un penalti o lo hagan en muy pocas ocasiones.

Esto provoca que la mayoría de jugadores no hayan anotado goles de penalti en las temporadas anteriores a sus traspasos, y que aquellos que lo hayan hecho sean considerados como casos atípicos.

Observemos los datos sobre asistencias.

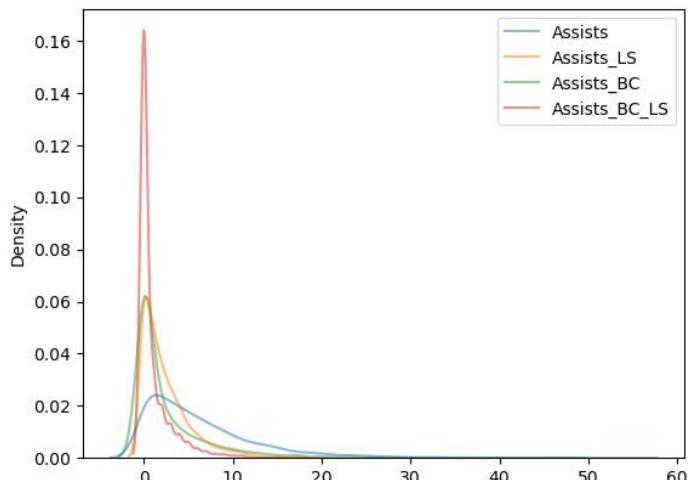


Figura 15: Gráfico de densidad para asistencias de jugadores de campo

En la Figura 15 se ve que, al igual que en las dos variables anteriores, la gran mayoría de fichajes ha repartido escasos pases de gol, o bien no lo ha hecho en ninguna ocasión durante las temporadas anteriores a sus fichajes. Esto puede deberse a una escasa participación en sus equipos o bien a roles más defensivos o puramente goleadores, haciendo que aquellos acostumbrados a hacerlo destaque como datos atípicos con respecto al resto de jugadores.

Analicemos los disparos.

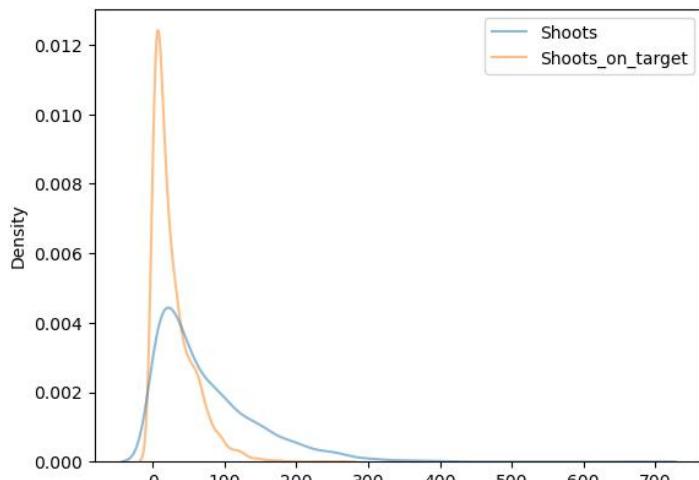


Figura 16: Gráfico de densidad para tiros y tiros a puerta

Observamos en la Figura 16, Tal como en otros casos, vemos que hay una serie de jugadores que se encargan en mayor medida de disparar a portería, bien por su posición o por su rol en el campo, así como una gran mayoría que los hace en pocas ocasiones, haciendo que los primeros sean considerados datos atípicos y provocando una distribución asimétrica positiva con picos alrededor del cero.

Pasamos a visualizar acciones defensivas a continuación.

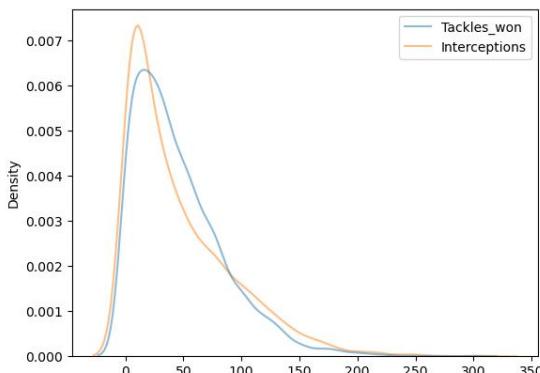


Figura 17: Gráfico de densidad para acciones defensivas

En la Figura 17 vemos distribuciones asimétricas positivas tanto para entradas con éxito (tackles won) como para intercepciones. Al ser parte de los jugadores de corte ofensivo o tener roles que no requieren este tipo de acciones, aquellos que las lleven a cabo con regularidad serán datos atípicos, y las distribuciones de densidad tendrán picos alrededor del cero.

Analicemos las ocasiones de gol creadas en grandes competiciones a continuación.

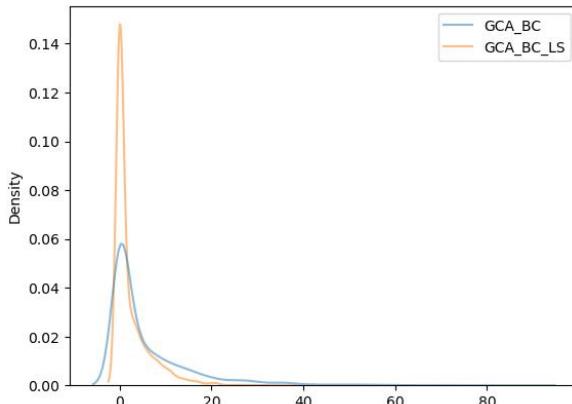


Figura 18: Ocasiones de gol creadas en grandes competiciones por jugadores de campo

Observamos en la Figura 18 algo parecido a muchas variables anteriores, distribuciones asimétricas positivas con picos en el cero, que muestran un gran número de jugadores que no han creado apenas ocasiones de gol, ya sea por posición, rol o escasa participación, y una minoría que se dedican a ello.

Veamos las amonestaciones.

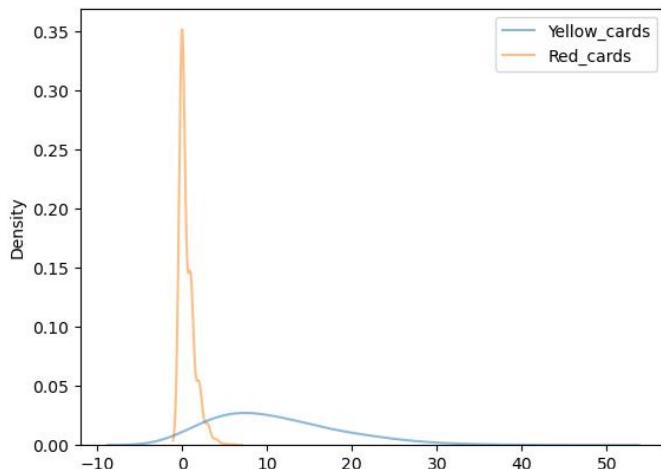


Figura 19: Gráfico de densidad de tarjetas para jugadores de campo

Podemos vislumbrar en la Figura 19 una distribución simétrica para tarjetas amarillas y una más asimétrica positiva con un pico en cero para tarjetas rojas. Este fenómeno se explica teniendo en cuenta que las tarjetas rojas son menos frecuentes, así como que hay unos pocos futbolistas que las reciben con más regularidad que el resto, mientras que las tarjetas amarillas son más usuales en todos los partidos de fútbol.

Vamos ahora con métricas de rendimiento para los porteros, empezando por el número de partidos jugados.

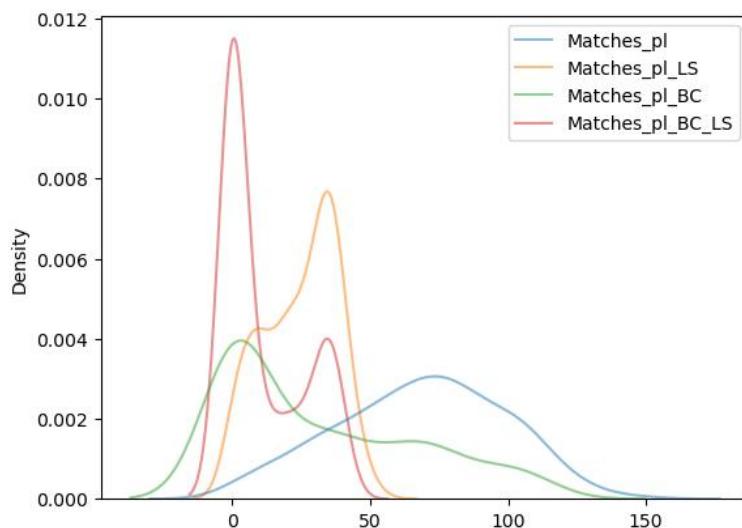


Figura 20: Gráfico de densidad de partidos jugador para porteros

Vemos en la figura anterior una distribución simétrica en partidos totales, a diferencia de en el resto de variables relacionadas, donde encontramos distribuciones asimétricas.

En el caso de partidos jugados la última temporada, la distribución es asimétrica negativa, con el pico más pronunciado alrededor de los 40 partidos, mientras que en las variables referidas a partidos jugados en grandes competiciones observamos una asimetría positiva con picos alrededor de cero, lo que indica que muchos de los jugadores traspasados no disputaron grandes competiciones durante los años previos a sus fichajes.

Vamos ahora con las titularidades.

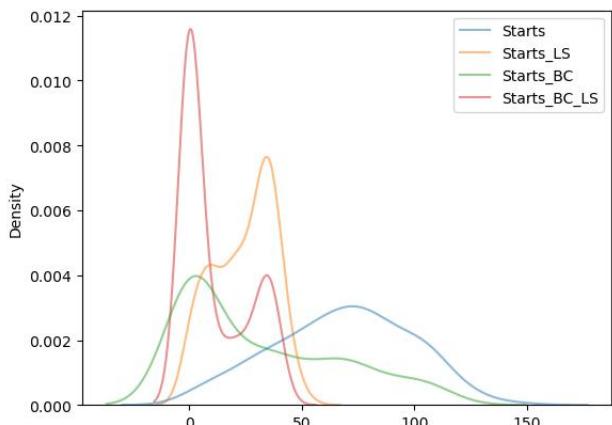


Figura 21: Diagrama de densidad para titularidades en porteros

Vemos en la figura 21 distribuciones similares a las de partidos jugados, indicando una muy posible correlación entre variables, tal como pasaba con los jugadores de campo. En el caso de los porteros, además, es menos probable que salgan desde el banquillo, por lo que los valores de esta variable serán más similares a los de partidos jugados que en el caso de los jugadores de campo.

Analicemos los minutos jugados.

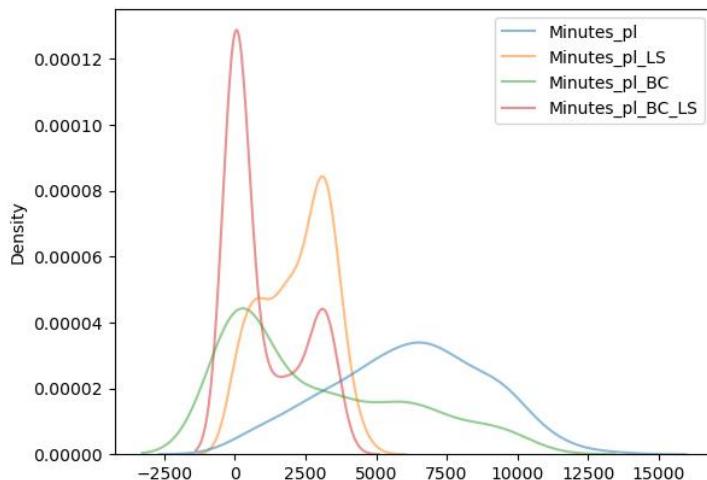


Figura 22: Diagrama de densidad de minutos jugados por porteros

Al ser altamente probable que aquel portero que empieza jugando el partido se encuentre todavía en el terreno de juego al producirse el pitido final, es de esperar que esta variable esté altamente correlacionada con las anteriores, más aún que en el caso de los jugadores de campo.

Las densidades de las variables referidas a minutos jugados son prácticamente idénticas a las de titularidades, tal como se ve en la Figura 22, con una distribución simétrica en minutos totales, asimétrica positiva para los minutos jugados la última temporada y asimétrica positiva en el resto de variables.

Vamos con goles encajados

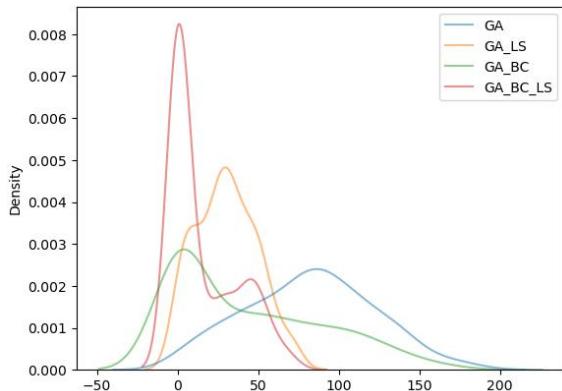


Figura 23: Gráfico de densidad de goles encajados

Si nos fijamos en la figura anterior vemos de nuevo una gran variabilidad, sobre todo en goles encajados durante las tres últimas temporadas, y un gran número de porteros que no ha encajado goles en grandes competiciones durante la última temporada. Esto puede deberse a que muchos de ellos no han participado en grandes competiciones.

Observamos también una asimetría positiva en variables referidas a la última temporada, mientras que el resto tienen una distribución simétrica, con un pico alrededor de 40 en goles en contra durante la última temporada y otro cerca de 90 para las tres últimas.

A continuación, comentaremos los tiros a puerta en contra.

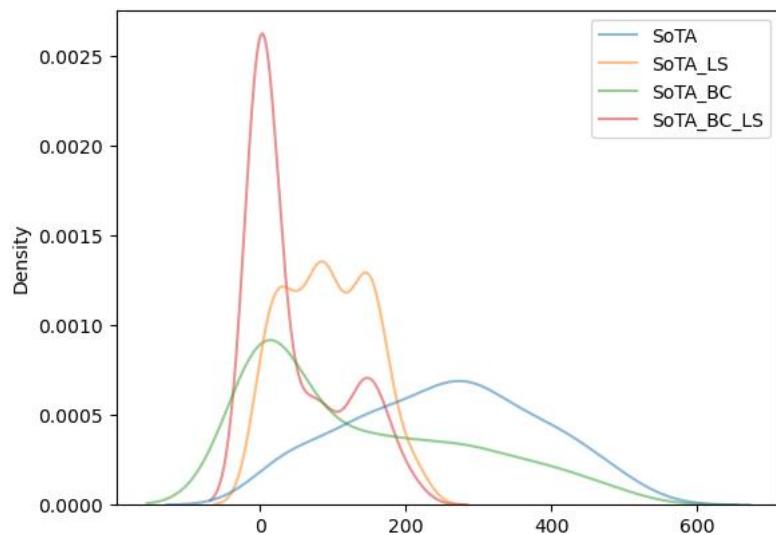


Figura 24: Gráfico de densidad de tiros a puerta en contra

Tal como vemos en la figura 24, las distribuciones vuelven a ser asimétricas positivas con picos en cero en el caso de las variables referidas a grandes competiciones y simétricas en el resto, destacando una vez más que no todos los porteros han disputado grandes competiciones.

A continuación, analizaremos las porterías a cero.

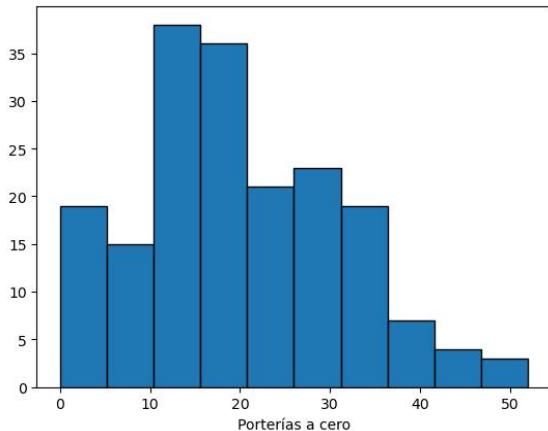


Figura 25: Histograma de porterías a cero

Observamos una ligera asimetría positiva en la figura anterior, así como un pico entre las 10 y las 20 porterías a cero, llegando algunos a tener hasta 50, aunque son casos poco frecuentes. Esto puede darnos una idea de que a la hora de fichar suelen elegirse porteros con buen rendimiento, aunque sin llegar a ser estrellas consolidadas.

Vamos ahora con penaltis en contra.

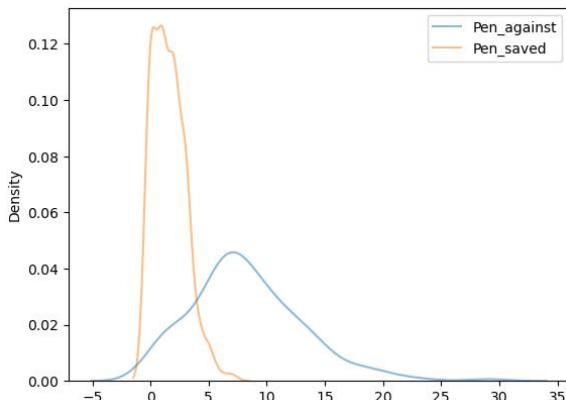


Figura 26: Penaltis recibidos y parados

En la Figura 26 vemos una distribución simétrica en cuanto al número de penaltis recibidos, con un pico cercano a 7, aunque los penaltis parados tienen una asimetría positiva con pico en cero, dándonos una idea de la gran dificultad de detener penaltis, así como de que existen unos pocos porteros especializados en ello.

Analicemos las tarjetas a continuación.

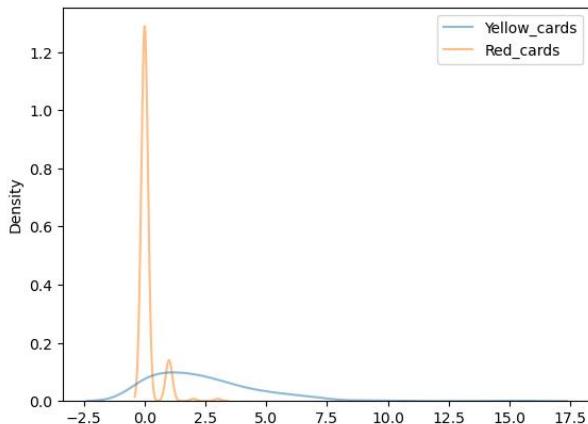


Figura 27: Tarjetas en porteros

En cuanto a amonestaciones, se ve en la Figura 27 las tarjetas amarillas suelen ser más frecuentes en todos los partidos, aunque en el caso de los porteros rara vez las reciben, solamente en casos en los que protestan o pierden tiempo. En cuanto a las expulsiones o tarjetas rojas, son más inusuales todavía, con un pequeño número de porteros que las han recibido, aunque en pocas cantidades por lo general.

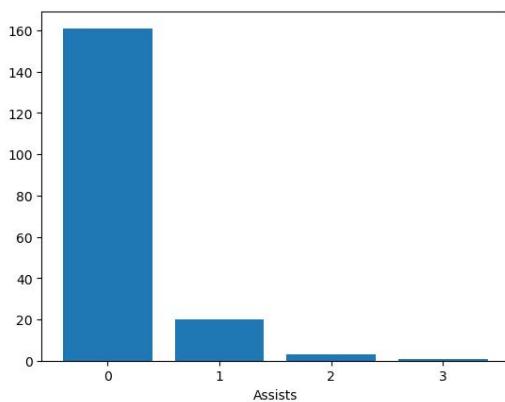


Figura 28: Asistencias en porteros

En cuanto a asistencias, vemos en la Figura 28 que apenas unos pocos porteros han llegado a dar alguna. Vemos 20 porteros que han repartido una solamente, tres que han dado dos y uno que ha repartido 3 asistencias, labor poco usual para un guardameta.

#### 4.4.2 Análisis multivariante

En el contexto del fútbol profesional, el valor de mercado de los futbolistas se ve influenciado por una amplia gama de variables interrelacionadas: desde métricas de rendimiento en el campo hasta características demográficas y contextuales, como la edad, la posición, el club al que pertenecen o su nacionalidad. Comprender estas relaciones no solo permite estimar con mayor precisión el valor de un jugador, sino también identificar los factores más relevantes que lo determinan.

En este apartado se abordarán los métodos multivariantes utilizados para examinar la influencia conjunta de diversas características en el valor de mercado de los futbolistas. Este

análisis servirá como base para el desarrollo posterior de modelos de aprendizaje automático que no solo sean precisos, sino también interpretables para clubes, analistas y otros actores del ecosistema futbolístico.

En primer lugar, buscaremos correlaciones entre variables relativas a jugadores de campo, empezando por aquellas que hemos observado antes que muy probablemente estén correlacionadas, y tratando de no usar pares de variables con más de un 0'75 de coeficiente de correlación.

Usaremos tanto el coeficiente de Pearson como el VIF para cuantificar la correlación y la multicolinealidad entre variables. Esta última métrica, el VIF, responde a las siglas de Variance Inflation Factor, y mide cuánto aumenta la varianza del coeficiente estimado de una variable explicativa debido a su correlación con las demás variables del modelo.

Cuanto más alta sea esta correlación, más sobreestimada estará la varianza del coeficiente, y por lo tanto menos confiables serán sus estimaciones. Valores por encima de cinco indican una correlación fuerte que puede hacer que nos planteemos prescindir de algunas variables.

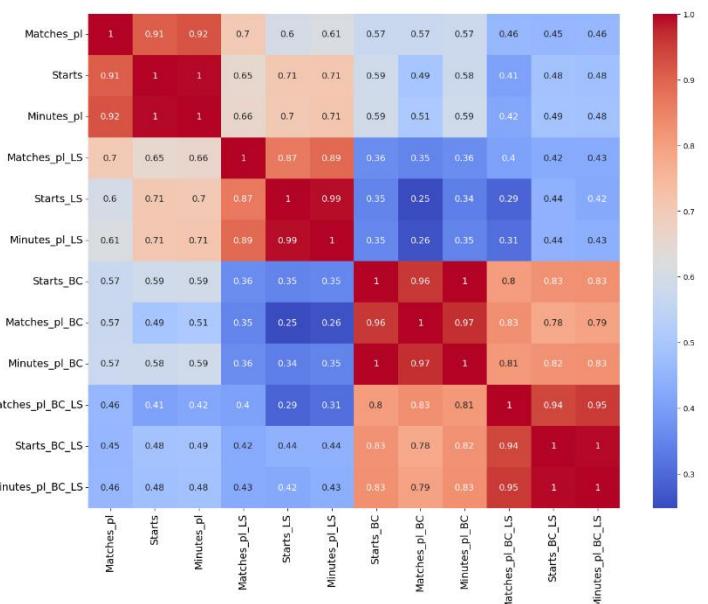


Figura 29: Mapa de calor de correlaciones de variables referidas a participación

Si nos fijamos en la Figura 29 vemos una correlación muy alta en variables referidas a las mismas temporadas y competiciones, por lo que podría ser interesante eliminar algunas de ellas en modelos que se vuelvan inestables y difíciles de interpretar al incluirlas.

	Variable	VIF
0	const	9.070808
1	Matches_pl	33.963559
2	Starts	480.725603
3	Minutes_pl	594.019233
4	Matches_pl_LS	20.901276
5	Starts_LS	266.519597
6	Minutes_pl_LS	331.775167
7	Starts_BC	1154.960340
8	Matches_pl_BC	86.116457
9	Minutes_pl_BC	1483.293649
10	Matches_pl_BC_LS	56.014326
11	Starts_BC_LS	621.149836
12	Minutes_pl_BC_LS	811.155110

Figura 30: VIF en variables referidas a participación

En la figura anterior vemos un VIF superior a 5 en todas estas variables, lo que confirma un alto nivel de colinealidad entre estas variables. Decidimos quedarnos solamente con las referidas a minutos jugados, obteniendo los siguientes resultados en cuanto a correlación.

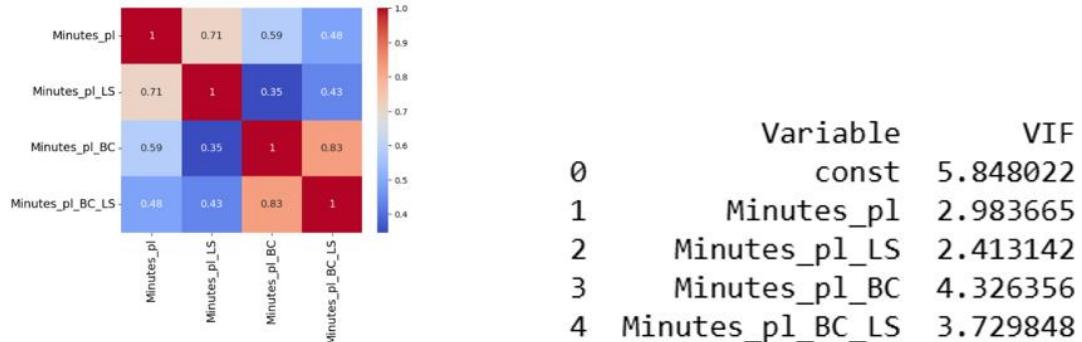


Figura 31: Correlación y VIF entre variables referidas a minutos jugados.

Vemos en la Figura 31 que el VIF es inferior a 5 en todas las variables, aunque tenemos un coeficiente de correlación de 0.83 entre partidos jugados en grandes competiciones las últimas tres temporadas y partidos jugados en grandes competiciones durante la última temporada, por lo que eliminaremos la segunda de cara a evitar inestabilidad en alguno de nuestros modelos.

Vamos ahora con otro bloque de variables relacionadas entre sí, como son las referidas a disparos a portería y goles.

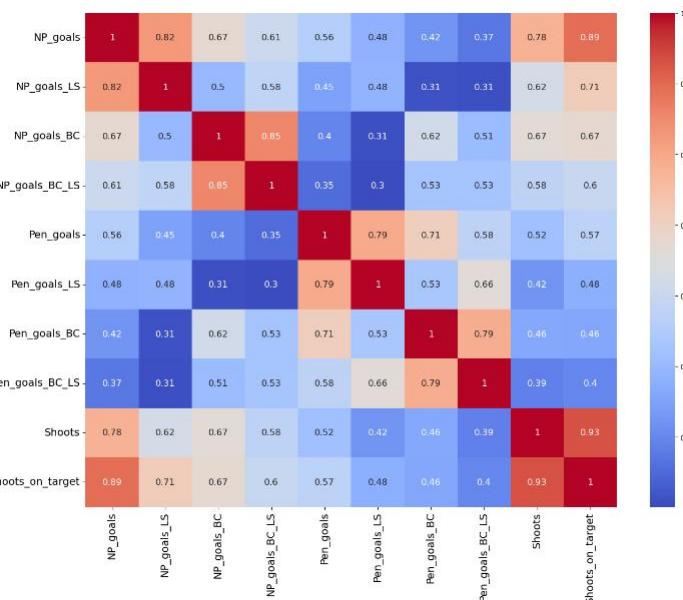


Figura 32: Correlación entre variables relativas a disparos a portería y goles

Si nos fijamos en la Figura 32, vemos una correlación muy fuerte entre algunas de las variables, tales como disparos y disparos a portería o disparos y goles sin contar penaltis.

	Variable	VIF
0	const	2.296536
1	NP_goals	10.128117
2	NP_goals_LS	4.730565
3	NP_goals_BC	7.363418
4	NP_goals_BC_LS	5.577082
5	Pen_goals	6.440626
6	Pen_goals_LS	5.319940
7	Pen_goals_BC	6.577137
8	Pen_goals_BC_LS	5.197397
9	Shoots	8.871579
10	Shoots_on_target	16.305934

Figura 33: VIF en variables relativas a disparos a portería y goles

Los valores del VIF son mayores que cinco en la mayoría de los casos, tal como podemos ver en la figura anterior, lo que indica una fuerte colinealidad entre variables de este tipo. Esto nos lleva a prescindir de cinco de las variables que causan estos inconvenientes, como son NP\_goals\_LS, NP\_goals\_BC\_LS, Pen\_goals\_LS, Pen\_goals\_BC\_LS y Shoots\_on\_target.

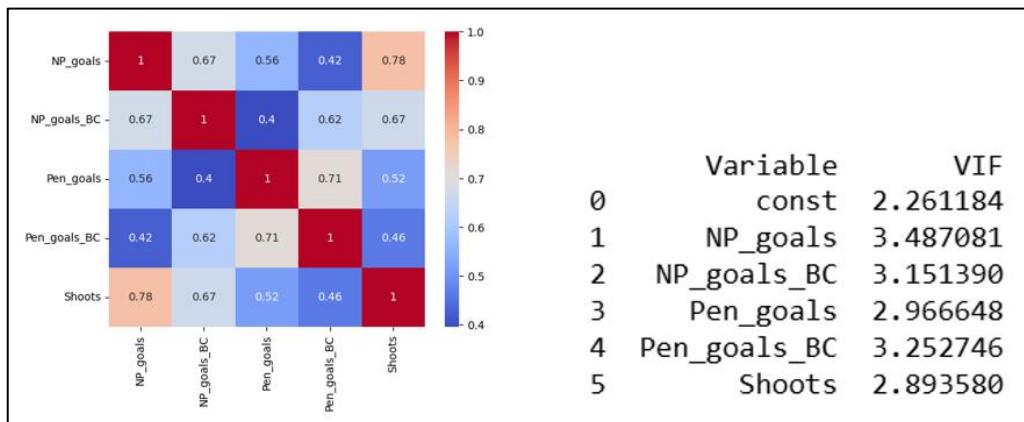


Figura 34: Correlación y VIF entre variables referidas a disparos a portería y goles.

Seguimos teniendo una correlación considerable entre tiros y goles, tal como se ve en la Figura 34, así como entre goles de penalti y goles de penalti en grandes competiciones, a pesar de que el VIF no es preocupante en ninguna de ellas. Deberemos tener esto en cuenta a la hora de entrenar los modelos.

Vamos ahora con el análisis de variables relacionadas con la creación de ocasiones de gol.

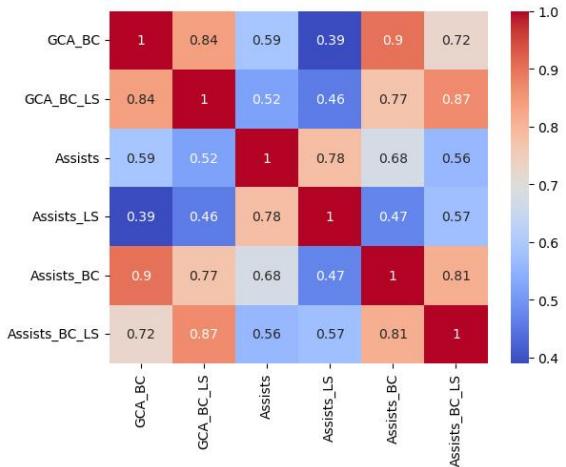


Figura 35: Correlaciones entre variables relacionadas con la creación de ocasiones

Observamos en la figura anterior una fuerte correlación entre muchas de estas variables, como es el caso de asistencias y ocasiones de gol creadas durante la última temporada, lo cual es lógico si atendemos al hecho de que una asistencia cuenta como creación de una ocasión de gol.

	Variable	VIF
0	const	2.094514
1	GCA_BC	10.564567
2	GCA_BC_LS	8.483755
3	Assists	4.462283
4	Assists_LS	3.622569
5	Assists_BC	12.148509
6	Assists_BC_LS	8.964993

Figura 36: VIF en variables referidas a ocasiones de gol creadas

Queda de manifiesto una fuerte colinealidad en este grupo de variables. Nos quedaremos únicamente con las variables referidas a ocasiones de gol creadas en grandes competiciones y asistencias totales durante las últimas tres temporadas, eliminando gran parte de la correlación en los datos.

También se observa una correlación de 0.82 entre intercepciones y entradas con éxito, por lo que se elimina esta última.

Al examinar las variables restantes, vemos que Shoots, NP\_goals\_BC, y Minutes\_pl\_LS están muy correlacionadas con otras, por lo que se eliminan también.

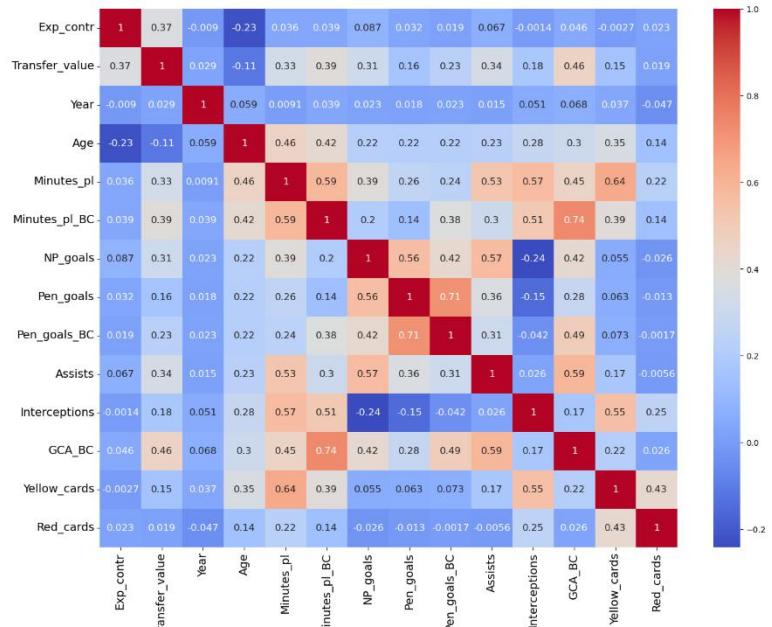


Figura 37: Correlaciones entre variables

Observamos en la Figura 37 que algunos pares de variables tienen un coeficiente de Pearson de entre 0'7 y 0'75, detalle que habrá que tener en cuenta en caso de observar inestabilidad o problemas relacionados a la hora de entrenar los modelos.

	Variable	VIF
0	const	1371199.92
1	Exp_contr	1.22
2	Transfer_value	1.78
3	Year	1.02
4	Age	1.69
5	Minutes_pl	4.26
6	Minutes_pl_BC	4.08
7	NP_goals	2.60
8	Pen_goals	2.70
9	Pen_goals_BC	2.65
10	Assists	2.52
11	Interceptions	2.90
12	GCA_BC	4.12
13	Yellow_cards	2.17
14	Red_cards	1.24

Figura 38: VIF en las variables restantes

Vemos en esta figura que ninguna variable tiene un VIF mayor que 5, por lo que la colinealidad se ha reducido a unos niveles que ya no son preocupantes a la hora de entrenar nuestros modelos. Estas serán las variables numéricas que utilizaremos para entrenar modelos en jugadores de campo.

Replicamos el proceso con porteros, para los cuales además de eliminar variables agrupamos algunas de ellas, siendo el caso de porcentaje de paradas para agrupar variables relativas a tiros a puerta en contra y goles encajados.

Estos son los resultados.

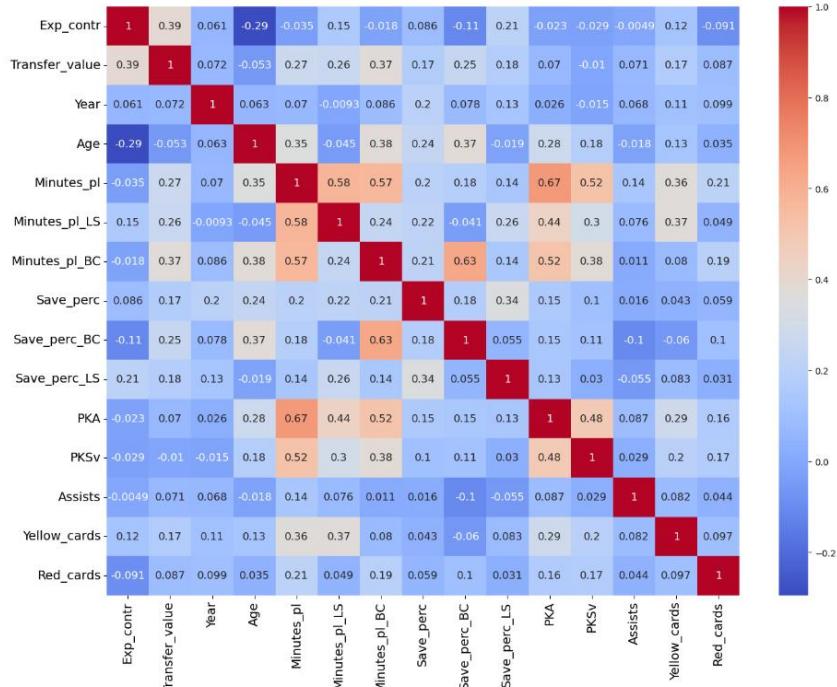


Figura 39: Correlaciones de variables relativas a porteros

	Variable	VIF
0	const	1418303.68
1	Exp_contr	1.42
2	Transfer_value	1.71
3	Year	1.10
4	Age	1.70
5	Minutes_pl	3.36
6	Minutes_pl_LS	1.99
7	Minutes_pl_BC	3.09
8	Save_perc	1.33
9	Save_perc_BC	2.02
10	Save_perc_LS	1.25
11	PKA	2.19
12	PKSv	1.54
13	Assists	1.06
14	Yellow_cards	1.34
15	Red_cards	1.11

Figura 40: VIF de variables relativas a porteros

Observamos en las dos figuras previas que todas las correlaciones están por debajo de 0'7 y todos los VIF por debajo de 4, por lo que hemos paliado en gran medida los problemas de multicolinealidad existentes en las variables.

A continuación, se hará un análisis de costes de fichaje para jugadores de campo según edad, país del equipo saliente y país del equipo en el que recalca.

Empezaremos con fichajes en los que el jugador no cambia de país

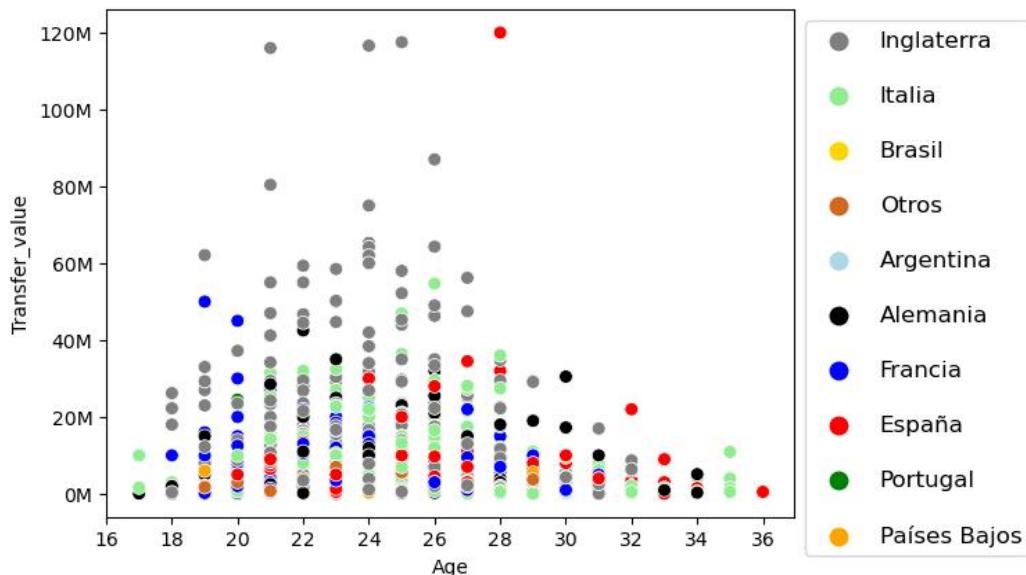


Figura 41: Fichajes internos por país y edad

Queda de manifiesto en la Figura 41 que los grandes traspasos se dan cuando el jugador tiene entre 22 y 28 años, un rango de edades en las que o bien tiene mucha proyección y algo de experiencia o bien está en el punto álgido de su carrera futbolística.

Vemos también una predominancia absoluta de Inglaterra en cuanto a fichajes internos, ya que la mayoría de grandes fichajes en los que el jugador no cambia de país se producen allí. La única excepción marcada es el fichaje de Antoine Griezmann, quien se fue de un equipo español para recalcar en otro por un coste de unos 120 millones de euros.

Una vez visualizados estos fichajes exploraremos aquellos en los que el jugador cambia de país, empezando por visualizar los países de origen de los equipos que el futbolista abandona en estos traspasos.

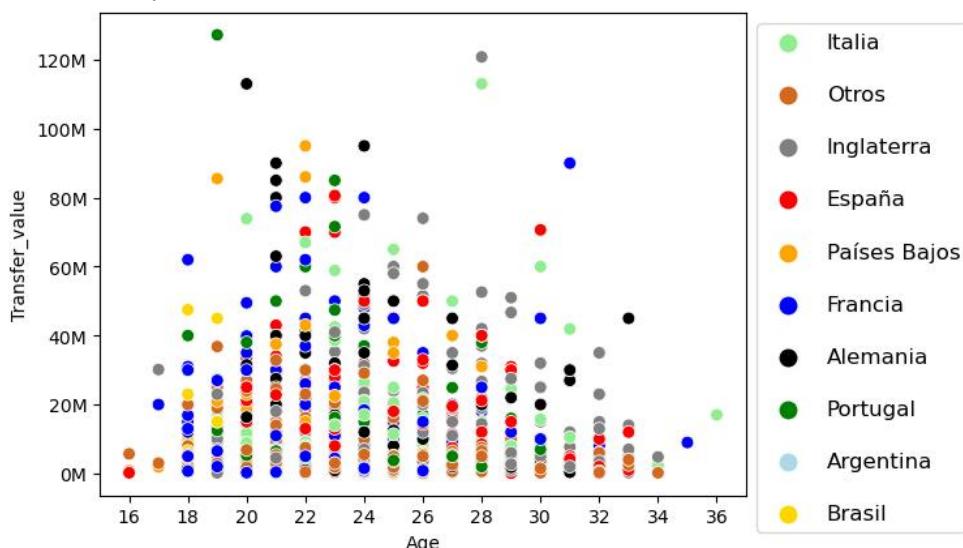


Figura 42: Fichajes por edad y país de origen del equipo que abandona el jugador

Vemos ahora unos rangos de edad más amplios para grandes traspasos, en los que se ha pagado 100 millones de euros por futbolistas de 19 y 20 años. Los países de los equipos de origen son bastante variados en todo el gráfico.

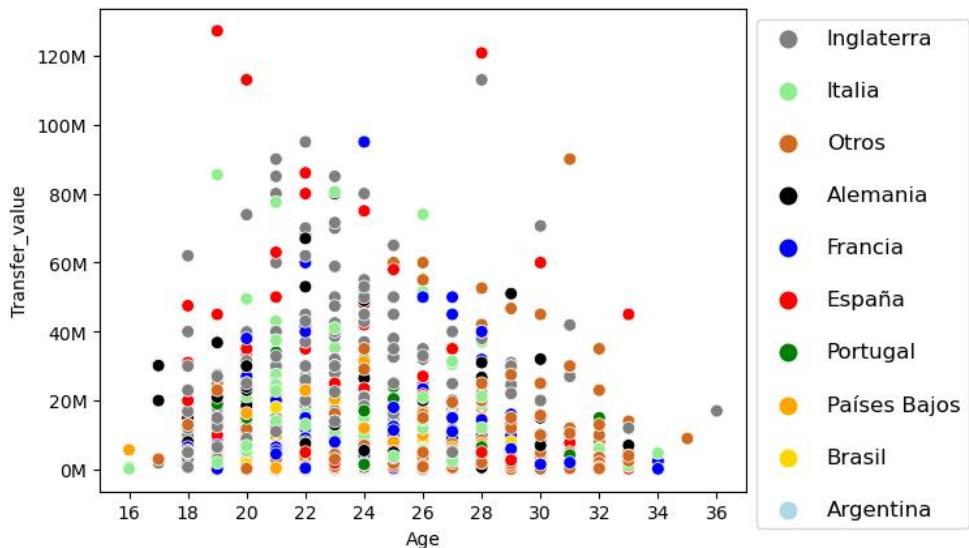


Figura 43: Fichajes por edad y país de origen del equipo por el que ficha el jugador

En esta figura caso se observa una predominancia de España en cuanto a grandes fichajes de jugadores provenientes de equipos extranjeros, apostando en mayor medida por talento exterior, a diferencia de lo que veíamos en Inglaterra, donde se fichan principalmente jugadores de otros equipos de la Premier League.

Examinaremos ahora estos mismos datos para porteros, empezando de nuevo por fichajes entre equipos de un mismo país.

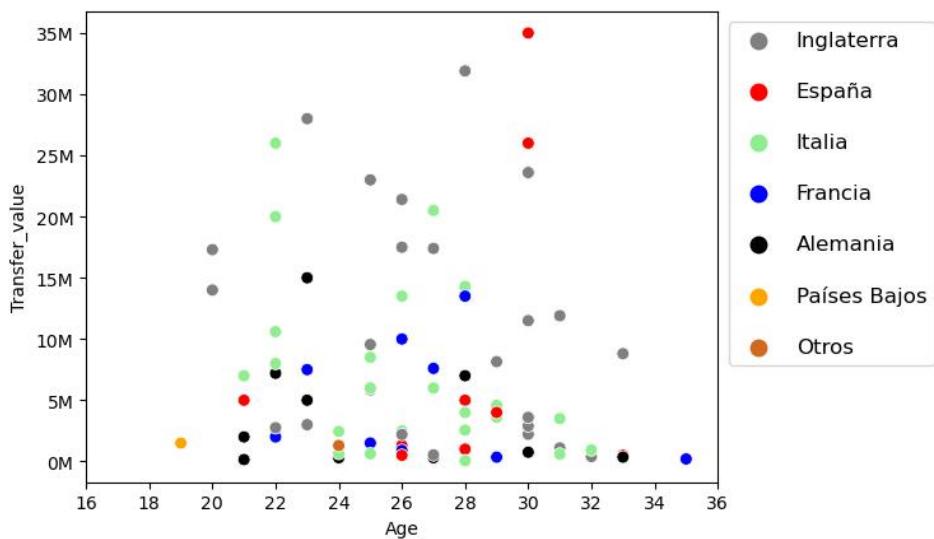


Figura 44: Fichajes internos de porteros por país y edad

Vemos en la Figura 44 un rango más amplio de edad en cuanto a los fichajes de mayor precio con respecto a los jugadores de campo, debido a la mayor duración general de la carrera de los porteros. En cuanto a países, vemos que Inglaterra e Italia predominan en número de fichajes y son, junto a España, los países donde se fichan principalmente porteros de gran valor provenientes de equipos del mismo país.

Ahora exploraremos aquellos fichajes en los que el portero cambia de país, empezando por visualizar los países de origen de los equipos que el futbolista abandona en estos traspasos.

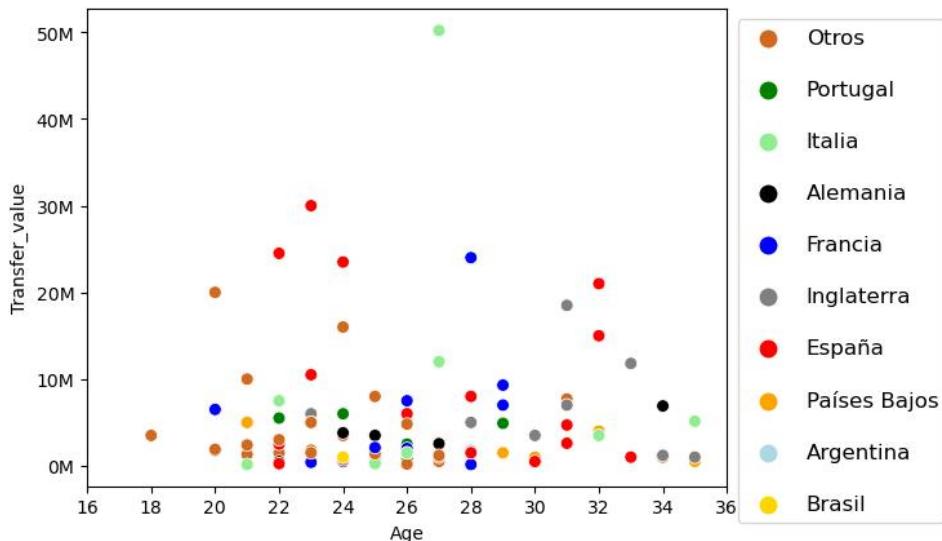


Figura 45: Fichajes de porteros por edad y país de origen del equipo que abandona el futbolista

En este gráfico destaca el fichaje de André Onana por el Manchester United por una cuantía de 50 millones de euros, así como una gran presencia de porteros que provienen de España.

Visualizaremos ahora los países en los que recalcan estos jugadores.

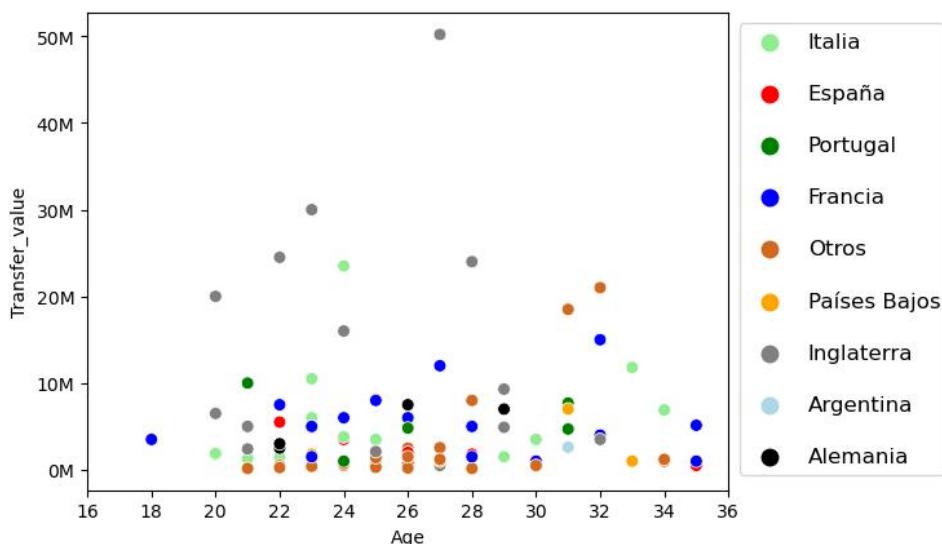


Figura 46: Fichajes de porteros por edad y país de origen del equipo por el que ficha el jugador

Vemos en la Figura 46 un dominio de Inglaterra en cuanto a fichajes más caros de porteros extranjeros.

Procedemos a analizar el gasto por países para todo tipo de futbolistas, tanto porteros como jugadores de campo.

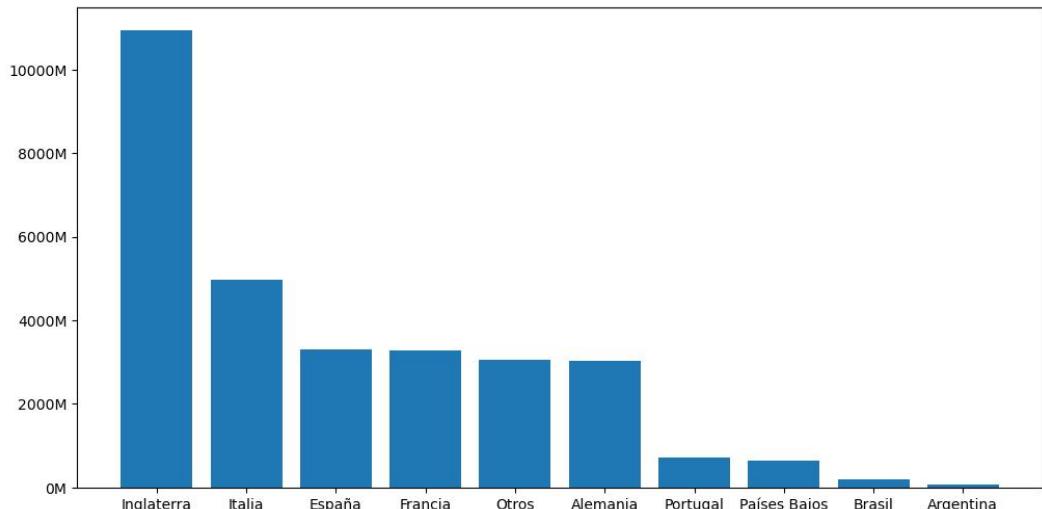


Figura 47: Gasto en futbolistas por países

Observamos en la figura anterior un predominio absoluto de Inglaterra en cuanto a gasto en fichajes se refiere, con más de 10000 millones invertidos en fichajes por parte de equipos ingleses en el mercado de verano desde el año 2019 hasta el 2024. Italia es el segundo país con más gasto, con unos 4700 millones invertidos, mientras que el resto de las cinco grandes ligas rondan los 3000 millones.

Analicemos a continuación la diferencia entre gastos e ingresos por fichajes.

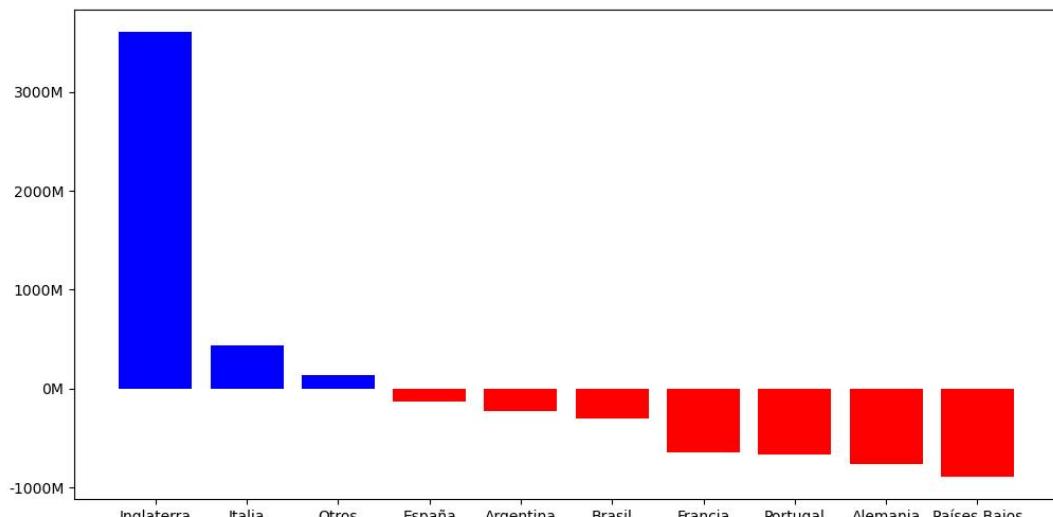


Figura 48: Diferencia entre gastos e ingresos por fichajes

Queda de manifiesto en esta figura que Inglaterra e Italia son los únicos países de los principales a nivel futbolístico cuyo gasto total en fichajes supera al ingreso, siendo en Inglaterra mucho más marcada esta diferencia. El resto de países analizados ingresan más de lo que gastan a la hora de fichar.

Estudiaremos a continuación el gasto en fichajes por posición.

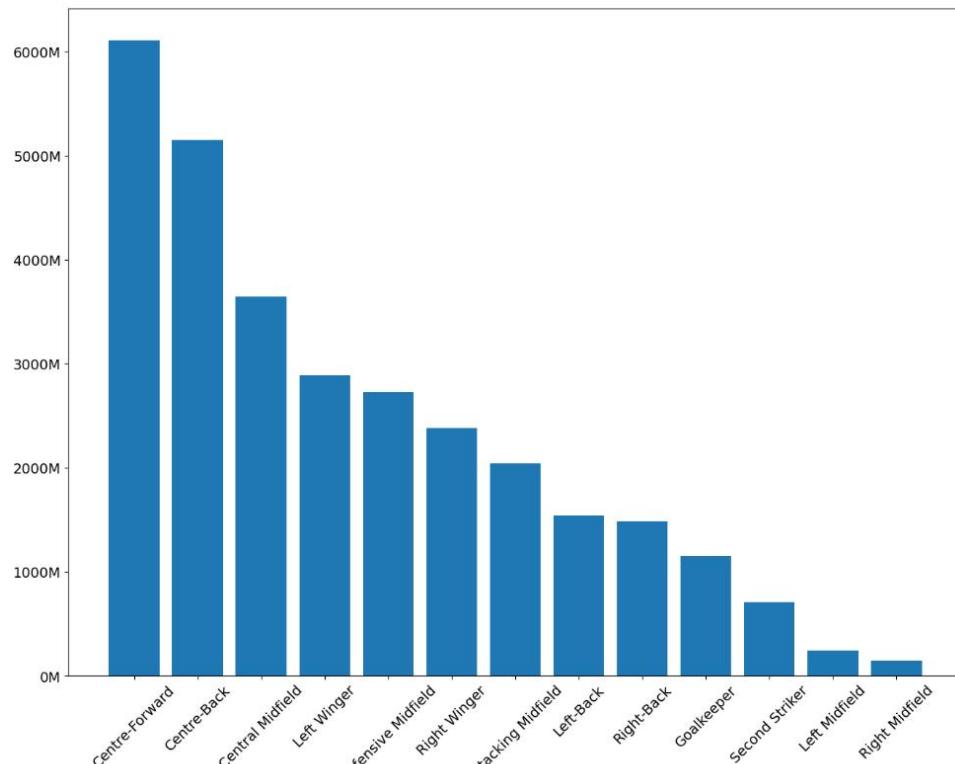


Figura 49: Gasto total en fichajes por posición

Anteriormente vimos que los delanteros centro y los defensas centrales eran los futbolistas que más se fichaban, por lo que no es de extrañar que el gasto total en los mismos sea superior al resto de posiciones.

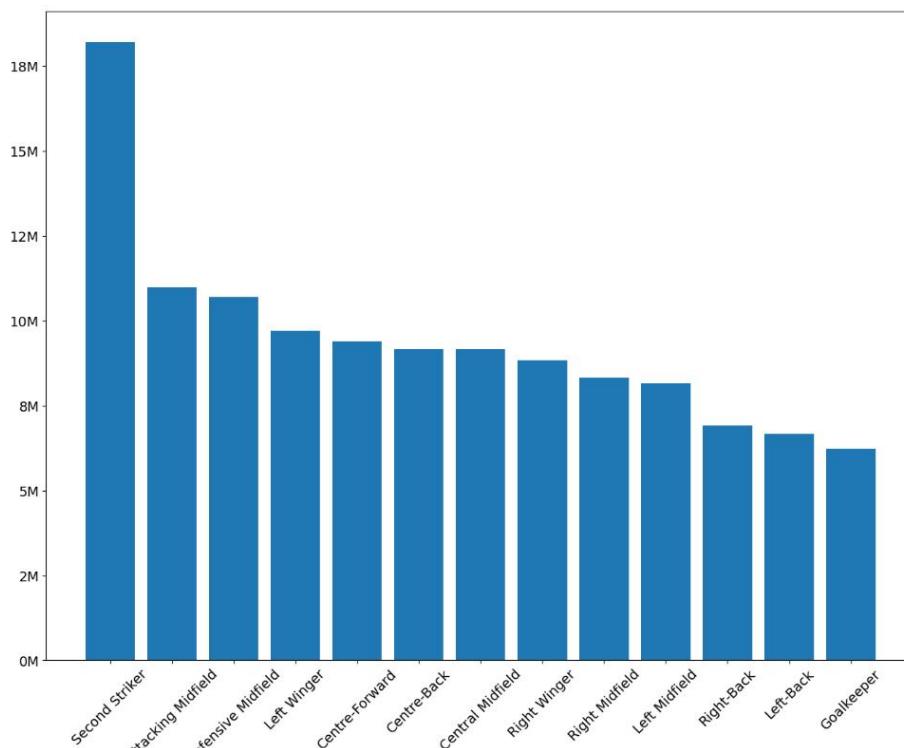


Figura 50: Gasto promedio por posición

En cuanto al gasto promedio, observamos en la Figura 50 que es superior en posiciones menos usuales o con menos movimiento en los mercados de fichajes, así como que los porteros tienen la media de costes de fichaje más baja de entre todas las posiciones.

## 5. Modelos propuestos

---

Una vez realizado el análisis exploratorio de los datos, se definirán los modelos a entrenar, los cuales se describen en detalle a continuación.

Los modelos descritos se ejecutarán, en primer lugar, para todos los jugadores de campo. Una vez hecho esto, dividiremos a los jugadores de campo por demarcación, entrenando por separado modelos para defensas, centrocampistas y delanteros, y por último se entrenarán para porteros. Esto se hará con la idea de dilucidar si es mejor tratar de predecir los valores de mercado de todos los jugadores de campo juntos o por separado.

Estos son los experimentos que llevaremos a cabo.

- Regresión lineal múltiple en dos capas. En la primera de ellas tendremos todas las variables salvo aquellas relativas al equipo de origen del futbolista y al país en el que se encuentra este equipo, variables que se añadirán en la segunda capa.
- Elastic Net, modelo de regresión lineal que combina dos técnicas de regularización para mejorar el rendimiento de la regresión lineal múltiple.
- Árboles de decisión, con la idea de captar interacciones entre variables. Se probarán distintos hiperparámetros con la idea de encontrar la mejor combinación de los mismos, y a continuación se comprobará qué variables han contribuido en mayor medida a las divisiones dentro del árbol.
- Random Forest con distintos hiperparámetros, algoritmo basado en árboles de decisión que entrena varios de ellos y combina sus resultados para crear predicciones robustas. Para medir la contribución de las distintas variables al modelo utilizaremos los shapley values, herramienta de interpretación que sirve para explicar cómo ha afectado cada una a cada predicción del mismo.
- Generalized Additive Model (GAM), modelo con el cual se tratarán de encontrar relaciones no lineales entre las variables independientes y el valor de mercado de los futbolistas.
- Explainable Boosting Machine (EBM), técnica con la que trataremos de captar interacciones entre variables, en la cual se entrena varios modelos y se combinan las predicciones para hacer una estimación final manteniendo la transparencia sobre la contribución individual de cada variable y de las interacciones identificadas en la predicción.
- Redes neuronales, técnica basada en el aprendizaje de patrones complejos en los datos mediante la interconexión de múltiples capas de nodos que procesan y transforman la información, permitiendo modelar relaciones no lineales y realizar predicciones.

Durante estos experimentos, se alternarán distintos valores de hiperparámetros, tales como el número de capas de la red, el número de neuronas en cada capa, la proporción de neuronas que estarán inactivas en cada iteración para evitar el sobreajuste (dropout), el optimizador de la red y el tamaño del subconjunto de los datos de entrenamiento que se utiliza para actualizar los pesos de la red en una sola iteración, conocido como batch size.

Llevaremos a cabo un tuning de hiperparámetros, el cual consiste en probar distintas combinaciones de valores de esos hiperparámetros y evaluar cuál ofrece el mejor desempeño

Se extraerán conclusiones al final para dilucidar cuáles han sido los modelos que mejor han funcionado.





## 6.1.2 Elastic Net

Tras entrenar un modelo con las variables relativas a rendimiento, posición, tiempo de contrato restante y equipo de origen del futbolista, obtenemos los siguientes resultados.

```
Mejor alpha: 0.004641588833612782
Mejor l1_ratio: 0.1
R^2 en los datos de entrenamiento: 0.4930121059036896
MSE en los datos de entrenamiento: 0.5069878940963105
```

Figura 53: Elastic Net para jugadores de campo (1)

Vemos en la figura anterior que el modelo tiene un R-cuadrado de 0.493, por lo que apenas mejora a la regresión lineal múltiple. El MSE tampoco mejora en gran medida a lo que ya teníamos.

Variables con coeficiente distinto de cero:		
	Variabile	Coeficiente
16	Reduced_team_from_Paris SG	0.387440
19	Reduced_team_from_Atlético Madrid	0.373503
27	Reduced_country_from_Inglaterra	0.360514
32	Reduced_country_from_Portugal	0.344343
1	Age	-0.328191
20	Reduced_team_from_Bayern Munich	0.280910
9	GCA_BC	0.268477
14	Reduced_team_from_Juventus	0.266075
39	Position_Centre-Back	0.263434
21	Reduced_team_from_Bor. Dortmund	0.256135
17	Reduced_team_from_Real Madrid	0.243135
0	Exp_contr	0.237680
15	Reduced_team_from_Napoli	0.233014
22	Reduced_team_from_Chelsea	0.215863
42	Position_Defensive Midfield	0.215267
33	Reduced_country_from_Paises Bajos	0.213035
4	NP_goals	0.200909
3	Minutes_p1_BC	0.189994
28	Reduced_country_from_Francia	-0.170690
35	Reduced_country_from_Brasil	0.159074
12	Reduced_team_from_Inter	0.130236
31	Reduced_country_from_Alemania	-0.122605
8	Interceptions	0.103890
45	Position_Left Midfield	-0.090540
7	Assists	0.085122

Figura 54: Elastic Net para jugadores de campo (2)

Ordenando las variables por el valor absoluto de sus coeficientes, se dilucida en la Figura 54 que aparecen algunas que no eran relevantes en regresión lineal múltiple, tales como que el equipo de origen sea el Atlético de Madrid.

Vemos que las variables con mayor relevancia son dummies referentes al equipo o al país del mismo, y la explicación de los coeficientes es idéntica a la de la regresión lineal múltiple, por lo que militar en el PSG aumenta el valor de mercado en 0'387 desviaciones estándar, hacerlo en el Atlético de Madrid en 0'373, etc.

## 6.1.3 Árboles de decisión

Se han entrenado varios modelos variando parámetros para tratar de encontrar la mejor combinación posible, utilizando el 80% de los datos para entrenamiento y el 20% como conjunto de prueba. Estos parámetros han sido el criterio de impureza del árbol, su máxima profundidad, el número de muestras necesarias para dividir un nodo y el número mínimo de muestras que debe tener cada hoja del árbol.

Estos han sido los resultados.

```

Mejores hiperparámetros: {'criterion': 'absolute_error', 'max_depth': 10, 'min_samples_leaf': 15, 'min_samples_split': 2}
R^2 en entrenamiento: 0.48227840393358135
MSE en entrenamiento: 0.5061275607774506
R^2 en prueba: 0.32180478877225915
MSE en prueba: 0.7388443739759898

```

Figura 55: Resultados de árboles de decisión para jugadores de campo.

Vemos en la Figura 55 que tanto el R-cuadrado como el MSE son mejores en el conjunto de entrenamiento que en el de prueba, indicando un cierto sobreajuste, y que los resultados en el conjunto de prueba no superan a los de regresión lineal. El modelo tan solo explica un 32% de la variabilidad en los datos de test. El MSE es incluso mayor que en los modelos de regresión lineal múltiple, por lo que podemos concluir que los resultados son bastante inferiores a los obtenidos con técnicas de regresión lineal.

En cuanto a las variables que han contribuido a las divisiones del árbol, tenemos las siguientes.

GCA_BC	0.280150
Exp_contr	0.222889
Age	0.118206
Minutes_pl_BC	0.094810
Minutes_pl	0.074191
Reduced_country_from_Inglaterra	0.066880
NP_goals	0.041227
Assists	0.032716
Interceptions	0.030212
Reduced_country_from_Países Bajos	0.009134
Yellow_cards	0.008357
Pen_goals_BC	0.006750
Year	0.006428
Reduced_country_from_Portugal	0.004398
Position_Centre-Forward	0.001908
Pen_goals	0.001061
Reduced_country_from_Alemania	0.000682

Figura 56: Variables influyentes en árbol de decisión para jugadores de campo

Observamos en la figura anterior que las ocasiones de gol creadas en grandes competiciones contribuyen en un 28% al proceso de decisión del árbol, siendo la variable que más lo hizo. Otras como la edad o el tiempo restante de contrato también fueron de gran importancia para el proceso.

### 6.1.4 Random Forest

Tras entrenar un modelo con 100 árboles y usando un 20% de las muestras para datos de prueba, obtenemos un R-cuadrado de 0'42 con un MSE de 0.63, equivalente a unos 8'72 millones de euros, resultados peores que los que hemos obtenido mediante regresión lineal.

Estos son los shapley values de las variables con mayor importancia.

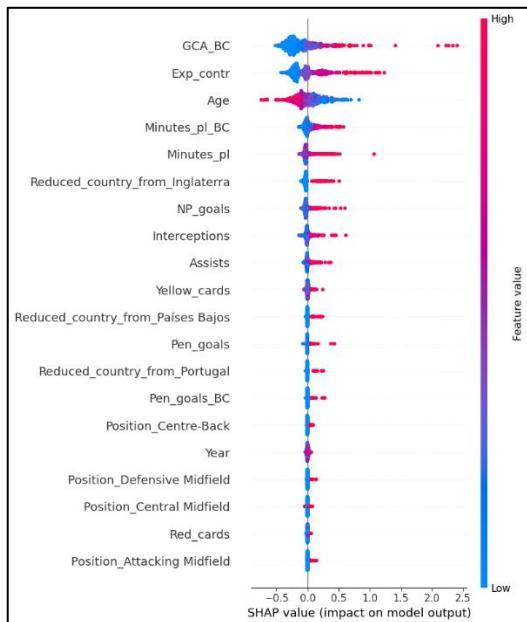


Figura 57: Shapley values de RF para jugadores de campo.

Tal como observábamos en los árboles, las variables que más han contribuido a las decisiones son las ocasiones de gol creadas en grandes competiciones, el tiempo restante de contrato y la edad. Todas las variables suelen tener un impacto positivo en el valor de mercado de los futbolistas salvo la edad, cuyo impacto es negativo en aquellos individuos a los que afecta en mayor medida a la hora de estimar su valor de mercado.

### 6.1.5 GAM

Al entrenar un modelo de este estilo buscamos relaciones no lineales de las variables independientes con la variable objetivo. Con un R-cuadrado de 0'5639 y un MSE, por ahora los mejores valores de estos parámetros, estos son los efectos de las distintas variables no dummy.

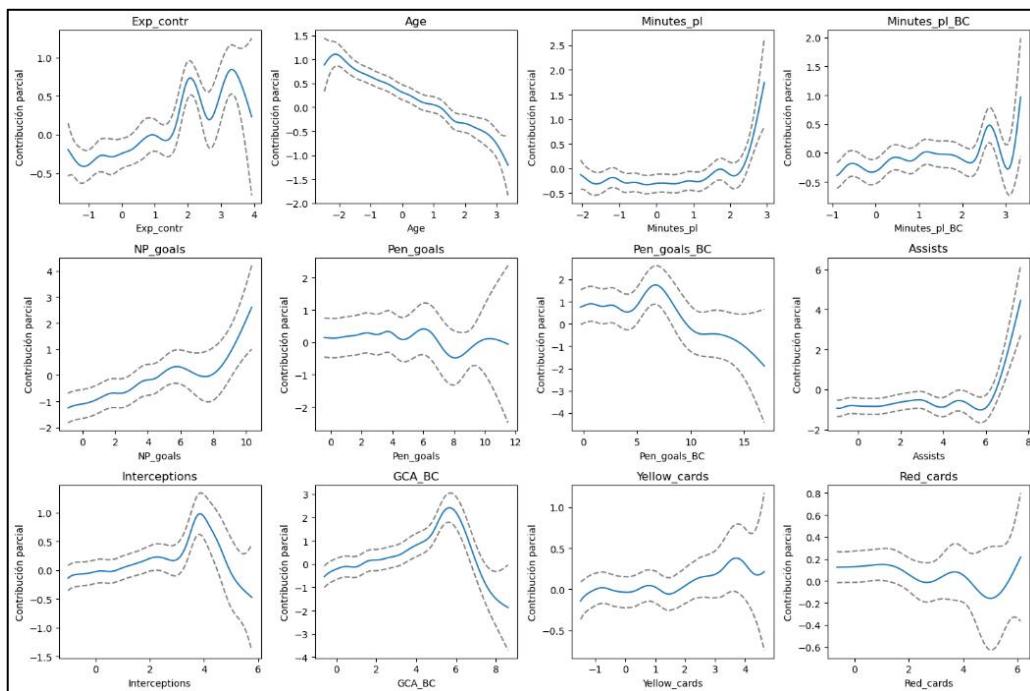


Figura 58: Efectos no lineales de variables con GAM para jugadores de campo.

Observamos en la Figura 58 relaciones no lineales entre el valor de mercado de futbolistas y el resto de variables. Las relaciones suelen ser positivas, salvo en el caso de la edad. Las variables referidas a tarjetas y goles de penalti tienen relaciones más débiles. A continuación, observaremos el impacto de todas las variables en el modelo, habiéndolas filtrado previamente para observar solamente aquellas con un p-value inferior a 0'05.

	Variable	Partial_Effect_Variance
7	Assists	1.266831
9	GCA_BC	1.118040
6	Pen_goals_BC	0.887825
4	NP_goals	0.727323
1	Age	0.350193
0	Exp_contr	0.155947
2	Minutes_pl	0.154817
8	Interceptions	0.110444
5	Pen_goals	0.053363
3	Minutes_pl_BC	0.052368
15	Reduced_team_from_Bayern Munich	0.030457
14	Reduced_team_from_Atlético Madrid	0.028391
12	Reduced_team_from_Paris SG	0.027068
11	Reduced_team_from_Napoli	0.020888
10	Reduced_team_from_Juventus	0.018733
13	Reduced_team_from_Real Madrid	0.017863
16	Reduced_team_from_Bor. Dortmund	0.013804
23	Reduced_country_from_Portugal	0.013328
18	Reduced_country_from_Inglaterra	0.008284
17	Reduced_team_from_Chelsea	0.008117
26	Position_Centre-Back	0.006984
25	Reduced_country_from_Brasil	0.004592
24	Reduced_country_from_Paises Bajos	0.004221
19	Reduced_country_from_Francia	0.003555
27	Position_Defensive_Midfield	0.003281
22	Reduced_country_from_Alemania	0.001874
20	Reduced_country_from_España	0.000963
21	Reduced_country_from_Italia	0.000898

Figura 59: Importancia de cada variable en GAM para jugadores de campo.

Vemos en la figura anterior que las asistencias y las ocasiones de gol creadas son aquellas variables que más contribuyen al modelo.



En la figura anterior se percibe que la mejor combinación de este tipo de redes ha llegado a explicar un 83'5% del valor de mercado de los jugadores de campo, con un MAE de tan solo 0'22, lo que equivale a unos tres millones de euros. Es el mejor modelo de todos los que hemos entrenado, y sus shapley values son los siguientes.

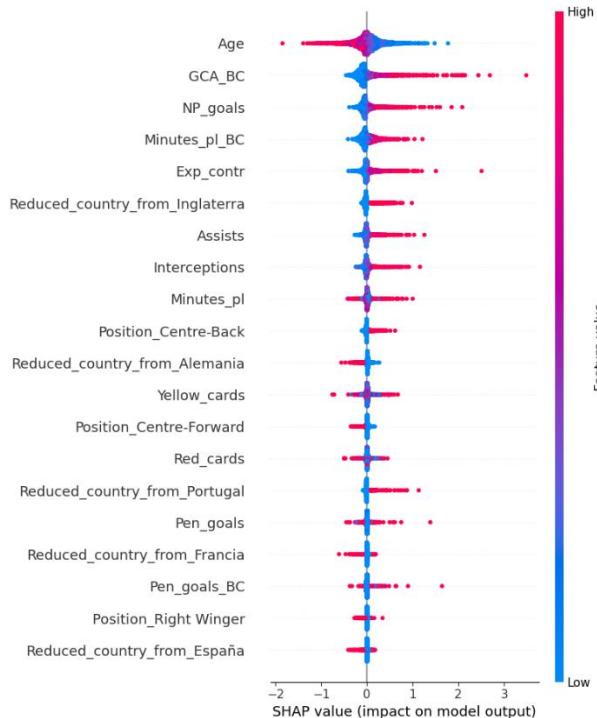


Figura 63: Shapley values para redes neuronales en jugadores de campo

Vemos en la Figura 63 que las variables que más importancia han adquirido son la edad, que afecta de forma negativa al valor de mercado del futbolista, las ocasiones de gol creadas y los goles que no han sido de penalti. Estas dos últimas afectan de manera positiva el valor de mercado de los jugadores.

Podemos concluir que las redes neuronales han sido las técnicas de machine learning que mejor han funcionado en este caso, seguidas por los modelos GAM.

## 6.2. Defensas

Repetiremos los experimentos anteriores utilizando únicamente futbolistas cuya posición sea defensa central o lateral, es decir, de corte defensivo. Con un total de 1008 fichajes de este tipo, trataremos de encontrar el modelo que mejor explique sus valores de mercado.

### 6.2.1 Regresión lineal múltiple

Tal como hicimos anteriormente, entrenaremos dos modelos de regresión lineal múltiple. En el primero de ellos tendremos en cuenta factores relativos al rendimiento y la situación del jugador (edad y tiempo de contrato restante), y en el segundo añadiremos información relativa al equipo de origen del jugador. Estos son los resultados.



## 6.2.2 Elastic Net

Tras llevar a cabo un conjunto de pruebas con modelos de este tipo, estos son los resultados.

```
Mejor alpha: 0.05994842503189409
Mejor l1_ratio: 0.1
R^2 en los datos de entrenamiento: 0.4328431164653581
MSE en los datos de entrenamiento: 0.5671568835346419
```

Figura 66: Elastic Net para defensas (1)

Vemos en la figura anterior que el modelo consigue explicar el 43'3% de la variabilidad del valor de fichaje de los futbolistas, con un MSE de unos 7'8 millones de euros. Estas han sido las variables más significativas para obtener estos resultados.

Variables con coeficiente distinto de cero:		
	Variable	Coeficiente
0	Exp_contr	0.300123
1	Age	-0.287818
3	Minutes_pl_BC	0.219900
27	Reduced_country_from_Inglaterra	0.190811
2	Minutes_pl	0.160687
9	GCA_BC	0.143209
4	NP_goals	0.136956
7	Assists	-0.089100
28	Reduced_country_from_Francia	-0.081288
8	Interceptions	0.056681
33	Reduced_country_from_Paises_Bajos	0.045608
6	Pen_goals_BC	-0.040907
14	Reduced_team_from_Juventus	0.035307
10	Yellow_cards	-0.031594
11	Red_cards	-0.012448
32	Reduced_country_from_Portugal	0.009149
20	Reduced_team_from_Bayern_Munich	0.005052
5	Pen_goals	-0.001703

Figura 67: Elastic Net para defensas (2)

Tal como ocurría en regresión lineal múltiple, la edad (de forma negativa), el tiempo de contrato restante y los minutos jugados en grandes competiciones han sido las variables más importantes a la hora de explicar el valor de mercado.

## 6.2.3 Árboles de decisión

Llevamos a cabo un conjunto de pruebas con distintos hiperparámetros. Estos son los resultados.

```
Mejores hiperparámetros: {'criterion': 'squared_error', 'max_depth': 10, 'min_samples_leaf': 15, 'min_samples_split': 2}
R^2 en entrenamiento: 0.5081981463398161
MSE en entrenamiento: 0.45416962636602
R^2 en prueba: 0.45802959586798175
MSE en prueba: 0.6932410637015808
```

Figura 68: Árboles de decisión para defensas (1)

En la figura 68 vemos un R-cuadrado de 0'458, en los datos de prueba, por lo que el mejor árbol explica casi un 46% de la varianza del valor de mercado, mejorando ligeramente los resultados obtenidos en regresión lineal múltiple, con un MSE de 0'69. Veamos las variables que más han contribuido a las divisiones del árbol.

Exp_contr	0.445336
Interceptions	0.175004
GCA_BC	0.134111
NP_goals	0.091126
Age	0.081297
Minutes_pl_BC	0.031404
Reduced_country_from_Países Bajos	0.020862
Reduced_country_from_Inglaterra	0.012001
Minutes_pl	0.004367
Assists	0.002623
Yellow_cards	0.001869
dtype: float64	

Figura 69: Árboles de decisión para defensas (2)

Percibimos en la Figura 69 que el tiempo restante de contrato ha sido utilizado para tomar el 44% de las decisiones dentro del árbol, siendo la variable con mayor influencia en el mismo, seguido de las intercepciones y las ocasiones de gol creadas en grandes competiciones.

## 6.2.4 Random Forest

Tras haber entrenado un total de 100 árboles, tal como en el caso en el que usábamos a todos los jugadores de campo, obtenemos un modelo que explica el 54% de la variabilidad del valor de mercado de los defensas con un MSE de 0.58 (unos 8 millones de euros). Son los mejores resultados para defensas hasta el momento, y estas han sido las variables más relevantes.

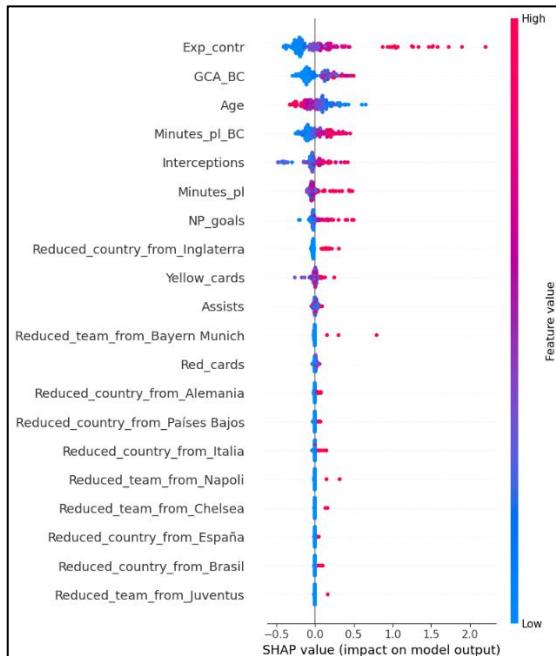


Figura 70: Random forest para defensas

Observamos en este caso que el tiempo restante de contrato es la variable que tiene un impacto importante sobre el mayor número de individuos, seguido por las ocasiones de gol creadas en grandes competiciones y la edad, afectando esta última de manera negativa al valor de mercado, tal como ocurría en los modelos anteriores.

## 6.2.5 GAM

A continuación, intentaremos vislumbrar los efectos no lineales de las variables no dummy sobre el valor de mercado de los defensas. Obtenemos un modelo con un R-cuadrado de 0'6 y un MSE de 0'4, unos 5'5 millones de euros, resultados algo superiores a los anteriores.

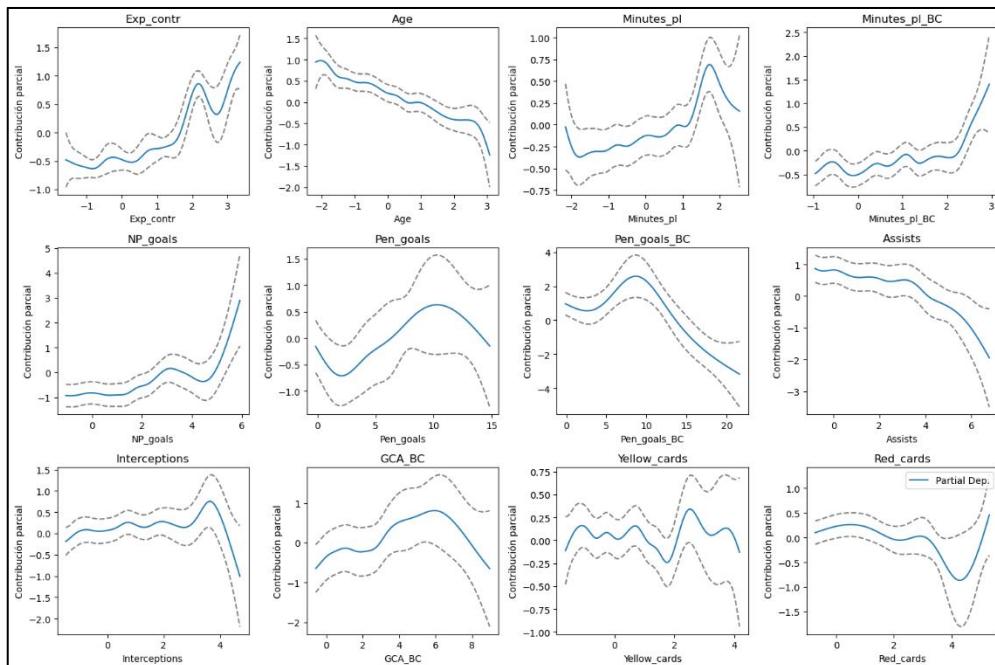


Figura 71: GAM para defensas (1)

Vemos en esta figura un efecto positivo en varias de las variables numéricas, exceptuando la edad y las tarjetas rojas, entre otras. La edad, así como el tiempo restante de contrato y los penaltis anotados en grandes competiciones parecen ser las variables con mayor influencia en el valor de mercado.

Variable	Partial_Effect_Variance
Pen_goals_BC	3.025592
NP_goals	0.681925
Exp_contr	0.304808
Age	0.247525
GCA_BC	0.200128
Minutes_pl_BC	0.185300
Red_cards	0.120671
Reduced_team_from_Napoli	0.091168
Minutes_pl	0.088280
Interceptions	0.083838
Reduced_team_from_Bayern_Munich	0.039942
Reduced_team_from_Juventus	0.038144
Reduced_team_from_Real_Madrid	0.026897
Reduced_country_from_Países_Bajos	0.005505
Reduced_country_from_Inglaterra	0.005019

Figura 72: GAM para defensas (2)

Los goles de penalti en grandes competiciones, así como los goles que no han sido de penalti y el tiempo restante de contrato parecen ser las variables con mayor relevancia en este modelo, tal como se observa en la Figura 72.



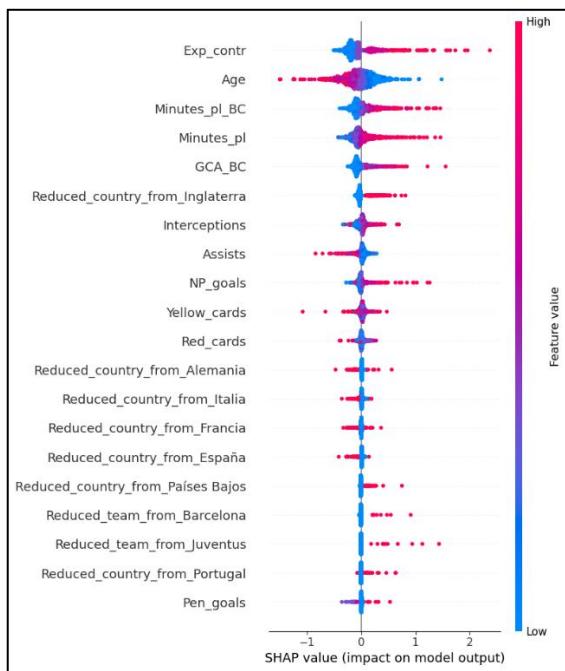


Figura 76: Shapley values en redes neuronales para defensas

Vemos en la figura anterior que las variables que afectan con fuerza a más individuos parecen ser el tiempo restante de contrato, la edad y los minutos jugados, tanto en grandes competiciones como en general. La edad afecta de forma negativa al valor de mercado, tal como ocurría en experimentos anteriores.

Una vez más, el rendimiento de las redes neuronales ha sido superior al del resto de modelos, aunque en este caso han sido las de dos capas las que mejor han funcionado. Los modelos GAM han vuelto a ser los segundos con un mayor rendimiento, y Random Forest también ha dado resultados bastante decentes.

## 6.3. Centrocampistas

Volveremos a entrenar los modelos anteriores enfocándonos exclusivamente en jugadores que se desempeñan como centrocampistas. Con una muestra de 887 traspasos, buscaremos identificar el modelo que ofrezca la mejor explicación de sus valores de mercado.

### 6.3.1 Regresión lineal múltiple

Empezamos entrenando un modelo con solo métricas de rendimiento como variables independientes.



### 6.3.2 Elastic Net

Estos son los resultados obtenidos tras llevar a cabo varias pruebas.

```
Mejor alpha: 0.004641588833612782
Mejor l1_ratio: 0.1
R^2 en los datos de entrenamiento: 0.5123142203269124
MSE en los datos de entrenamiento: 0.4876857796730875
```

Figura 79: Elastic Net para centrocampistas (1)

Observamos en la Figura 79 unos resultados ligeramente mejores que en el modelo de regresión lineal múltiple, con un R-cuadrado de 0'51 y un MSE de 0'48. Por su parte, las variables con mayor importancia son las siguientes.

	Variable	Coeficiente
17	Reduced_team_from_Real Madrid	0.708476
21	Reduced_team_from_Bor. Dortmund	0.515904
27	Reduced_country_from_Inglaterra	0.478166
16	Reduced_team_from_Paris SG	0.426455
14	Reduced_team_from_Juventus	0.368378
1	Age	-0.359167
18	Reduced_team_from_Barcelona	0.354355
19	Reduced_team_from_Atlético Madrid	0.354017
20	Reduced_team_from_Bayern Munich	0.337053
33	Reduced_country_from_Países Bajos	0.318331
32	Reduced_country_from_Portugal	0.311963
9	GCA_BC	0.288982
13	Reduced_team_from_AC Milan	0.274091
0	Exp_contr	0.241621
12	Reduced_team_from_Inter	-0.240424
31	Reduced_country_from_Alemania	-0.223518

Figura 80: Elastic Net para centrocampistas (2)

Vemos en la figura anterior que las variables más importantes a la hora de estimar el valor de mercado de los centrocampistas son aquellas relativas al equipo de origen y al país en el que se encuentra. La única métrica de rendimiento que se encuentra entre las variables de mayor relevancia es el número de ocasiones de gol creadas en grandes competiciones.

### 6.3.3 Árboles de decisión

Tras múltiples experimentos, este es el mejor árbol entrenado.

```
Mejores hiperparámetros: {'criterion': 'squared_error', 'max_depth': 5, 'min_samples_leaf': 5, 'min_samples_split': 2}
R^2 en entrenamiento: 0.5678357524149055
MSE en entrenamiento: 0.4372797123853997
R^2 en prueba: 0.24832211071636323
MSE en prueba: 0.7160650216738692
```

Figura 81: Árboles de decisión para centrocampistas (1)

Observamos que el R-cuadrado es mucho mejor en los datos de entrenamiento que en los de test, lo que indica un sobreentrenamiento del modelo. Las métricas para los datos de test muestran que el modelo no es muy bueno. Veamos las variables más importantes en este caso.

GCA_BC	0.304392
Exp_contr	0.199811
Interceptions	0.155624
Minutes_pl_BC	0.093223
Age	0.070162
Minutes_pl	0.058738
Assists	0.044973
NP_goals	0.029581
Yellow_cards	0.029345
Reduced_country_from_Inglaterra	0.014151
dtype: float64	

Figura 82: Árboles de decisión para centrocampistas (2)

Las ocasiones de gol creadas en grandes competiciones vuelven a tener un papel importante, tal como se percibe en la figura 82, y en este caso se utilizan para el 30% de las divisiones. Las intercepciones y el tiempo restante de contrato completan el podio de las métricas más decisivas a la hora de entrenar este modelo.

### 6.3.4 Random Forest

Con un R-cuadrado de 0'52 y un MSE de 0'45, los resultados mejoran ligeramente a los modelos que teníamos anteriormente para centrocampistas. Estas han sido las variables más importantes a la hora de entrenar el modelo.

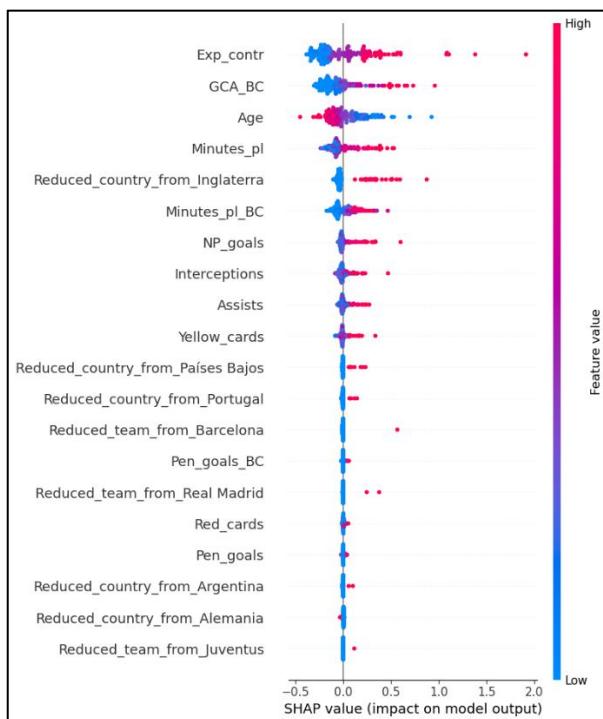


Figura 83: Random forest para centrocampistas

Vemos en la Figura 83 las mismas 3 variables más importantes que había en los árboles de decisión. La variable de procedencia más importante que tenemos es la que indica que el jugador proviene de un equipo inglés, y la única con un efecto negativo sobre el valor de mercado del futbolista vuelve a ser la edad.

### 6.3.5 GAM

A la hora de tratar de captar efectos no lineales de las variables no dummy sobre el valor de mercado de los centrocampistas, nos encontramos con un R-cuadrado de 0'642 y un MSE de 0'36, equivalente a unos 5 millones de euros, superando con creces a los resultados anteriores. En cuanto a los efectos de las variables, tenemos los siguientes.

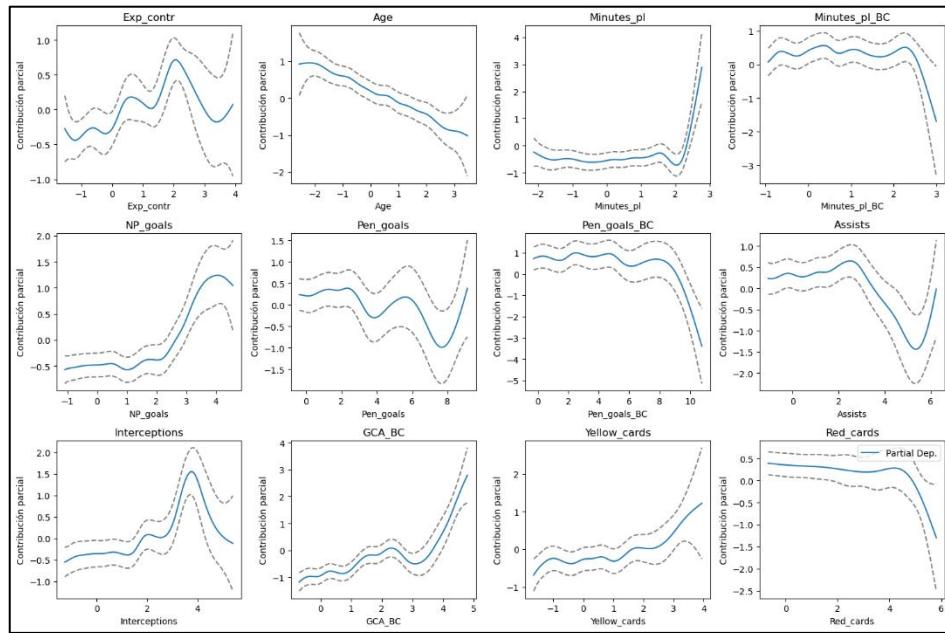


Figura 84: GAM para centrocampistas (1)

Las tarjetas rojas tienen un efecto claramente negativo sobre el valor de mercado, así como la edad. Las ocasiones de gol creadas en grandes competiciones y los minutos jugados parecen ser las variables que más contribuyen al modelo junto a los goles de penalti en grandes competiciones, estos últimos de forma negativa.

Variable	Partial_Effect_Variance
Pen_goals_BC	0.889984
GCA_BC	0.819317
NP_goals	0.454920
Assists	0.410499
Minutes_pl	0.409183
Age	0.385078
Interceptions	0.344810
Minutes_pl_BC	0.182783
Pen_goals	0.165696
Exp_contr	0.101677
Reduced_team_from_Atlético Madrid	0.078043
Reduced_team_from_Real Madrid	0.063722
Reduced_team_from_Barcelona	0.035777
Reduced_team_from_Juventus	0.033137
Reduced_team_from_Paris SG	0.030427
Reduced_team_from_Bor. Dortmund	0.022619
Reduced_country_from_Inglaterra	0.015869
Reduced_country_from_Portugal	0.006173
Reduced_country_from_España	0.005752
Reduced_country_from_Alemania	0.005573
Reduced_country_from_Francia	0.005536
Reduced_country_from_Italia	0.004645

Figura 85: GAM para centrocampistas (2)

Las variables más importantes parecen ser las mencionadas anteriormente como tal, así como las asistencias y los minutos jugados.



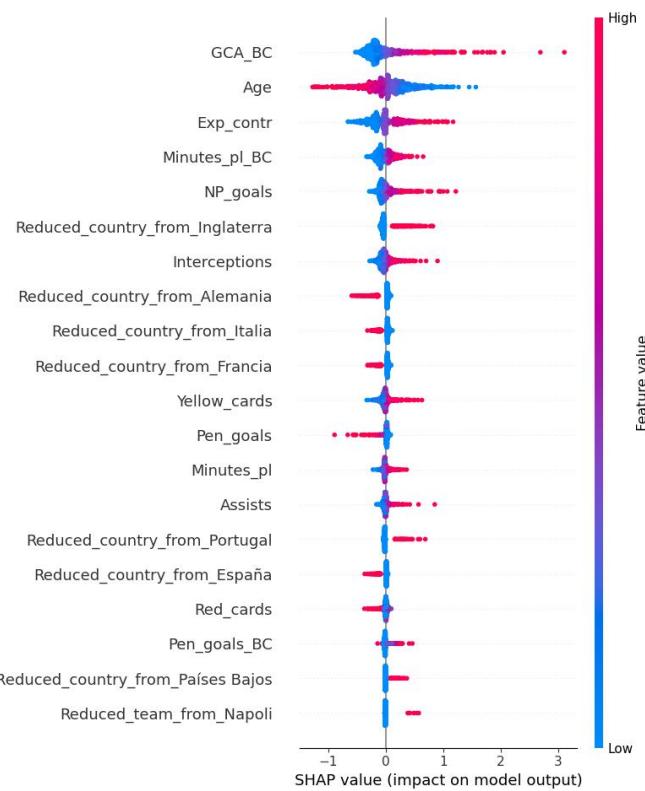


Figura 89: Shapley values en redes neuronales para centrocampistas

Las ocasiones de gol creadas, así como el tiempo restante de contrato y los minutos jugados en grandes competiciones, son métricas que afectan de forma bastante positiva al valor de mercado de los centrocampistas en una gran cantidad de fichajes. La edad, por el contrario, lo hace de forma negativa.

Entre las técnicas aplicadas, las redes neuronales han vuelto a ser las que mejor rendimiento han ofrecido, en especial las de tres capas, mientras que los modelos GAM también muestran un desempeño destacado.

## 6.4. Delanteros

En esta ocasión, llevaremos a cabo los experimentos considerando solo a los jugadores que ocupan la posición de delantero. Con un total de 1256 fichajes en esta posición, el objetivo será encontrar el modelo más adecuado para describir sus valoraciones en el mercado.

### 6.4.1 Regresión lineal múltiple

Empezamos con un modelo que usará como variables independientes aquellas relativas al rendimiento del jugador y a su situación, teniendo en cuenta su edad y el tiempo restante de contrato.



## 6.4.2 Elastic Net

Tras llevar a cabo un conjunto de pruebas, este ha sido el modelo con mejores resultados.

```
Mejor alpha: 0.05994842503189409
Mejor l1_ratio: 0.1
R^2 en los datos de entrenamiento: 0.5187618715565414
MSE en los datos de entrenamiento: 0.4812381284434586
```

Figura 92: Elastic Net para delanteros (1)

El R-cuadrado y el MSE son similares a los del modelo de regresión lineal múltiple entrenado en el apartado anterior. Observemos ahora las variables que más han contribuido al modelo.

Variable	Coeficiente
GCA_BC	0.322604
Age	-0.277152
Exp_contr	0.221130
NP_goals	0.201332
Reduced_country_from_Inglaterra	0.182718
Assists	0.180174
Reduced_country_from_Portugal	0.099837
Minutes_pl_BC	0.080712
Reduced_country_from_Francia	-0.074792
Reduced_country_from_Alemania	-0.066847
Minutes_pl	-0.055017
Pen_goals	0.041158
Pen_goals_BC	0.031889
Interceptions	-0.004870
Red_cards	0.002012

Figura 93: Elastic Net para delanteros (2)

En la Figura 93, tal como observábamos en regresión lineal múltiple, las ocasiones de gol creadas en grandes competiciones, la edad y el tiempo restante de contrato son métricas que explican en gran medida el precio de este tipo de futbolistas.

## 6.4.3 Árboles de decisión

Estos son los resultados tras varios experimentos.

```
Mejores hiperparámetros: {'criterion': 'absolute_error', 'max_depth': 10, 'min_samples_leaf': 10, 'min_samples_split': 2}
R^2 en entrenamiento: 0.4797202470280978
MSE en entrenamiento: 0.5256586067922349
R^2 en prueba: 0.2967468795352599
MSE en prueba: 0.6736128384342772
```

Figura 94: Árboles de decisión para delanteros (1)

Vemos en la figura anterior que el R-cuadrado del modelo no supera a los anteriores en el conjunto de entrenamiento ni en el de prueba, ni tampoco lo hace el MSE.

GCA_BC	0.237574
Age	0.223365
NP_goals	0.214942
Exp_contr	0.121657
Minutes_pl_BC	0.064608
Assists	0.053819
Minutes_pl	0.051194
Interceptions	0.018269
Pen_goals_BC	0.005355
Yellow_cards	0.003338
Pen_goals	0.001971
dtype:	float64

Figura 95: Árboles de decisión para delanteros (2)

Los goles anotados sin contar penaltis parecen haber cobrado importancia en este tipo de modelos, junto a algunas de las variables cuya importancia había quedado patente en modelos anteriores.

#### 6.4.4 Random Forest

Con un R-cuadrado de 0'48 y un MSE de 0'49, métricas que tampoco superan al rendimiento de las regresiones lineales, estas son las variables que más han contribuido a las decisiones.

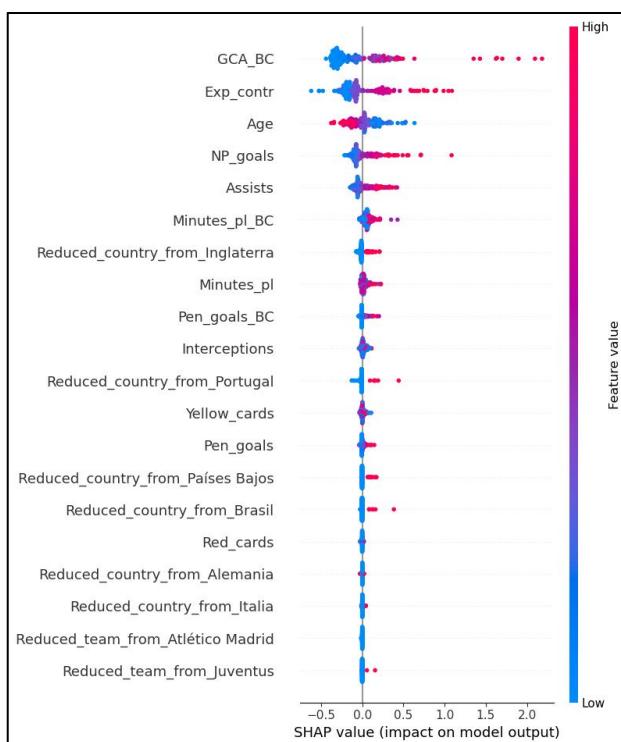


Figura 96: Random Forest para delanteros

Las variables con mayor importancia parecen no haber cambiado con respecto a los modelos anteriores.

#### 6.4.5 GAM

Al tratar de buscar relaciones no lineales entre las variables independientes no dummy y el valor de mercado de los futbolistas, obtenemos un R-cuadrado de 0'65 y un MSE de 0'35, equivalente a unos 4'84 millones de euros, los mejores resultados hasta el momento. Observemos los efectos de cada variable sobre el valor de mercado.

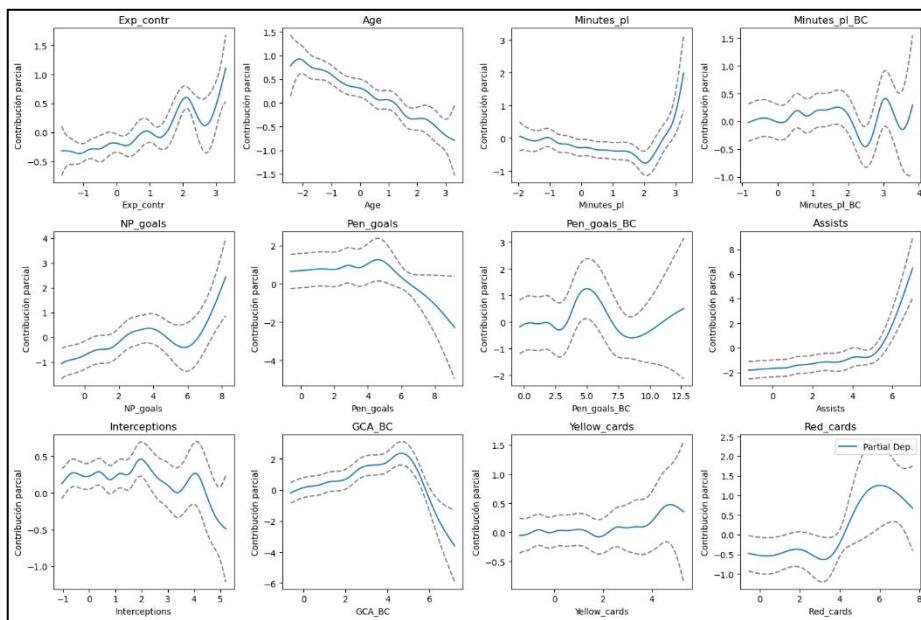


Figura 97: GAM para delanteros (1)

Los efectos son no lineales en su mayoría, tal como se ve en la Figura 97, exceptuando la edad. Las variables con mayor importancia han sido las siguientes.

Variable	Partial_Effect_Variance
Assists	3.782518
GCA_BC	2.083065
Pen_goals	0.849154
NP_goals	0.517235
Pen_goals_BC	0.270618
Age	0.250768
Minutes_pl	0.209425
Exp_contr	0.117307
Minutes_pl_BC	0.036623
Reduced_country_from_Portugal	0.018120
Reduced_team_from_Paris SG	0.014913
Reduced_country_from_Brasil	0.009687
Reduced_country_from_Inglaterra	0.006000
Reduced_country_from_Francia	0.002006
Reduced_country_from_Alemania	0.001694

Figura 98: GAM para delanteros (2)

Observamos en la figura anterior un fuerte impacto proveniente de goles, asistencias y ocasiones de gol creadas a la hora de entrenar el modelo, quedando en un segundo plano las variables dummy relativas al equipo de origen del futbolista.

#### 6.4.6 EBM

Al entrenar un modelo que capta interacciones entre variables hemos obtenido un R-cuadrado de 0'5, con un MSE de 0'4, resultado que no mejora a lo que ya teníamos. Estas han sido las variables y combinaciones de variables con mayor importancia a la hora de entrenar el modelo.



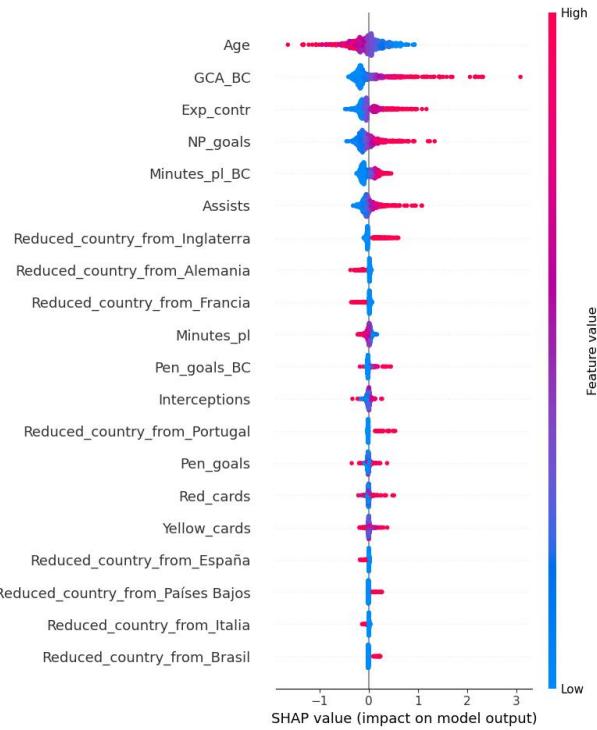


Figura 102: Shapley values en redes neuronales para delanteros

Las variables más relevantes parecen ser las mismas que en el resto de posiciones: edad, ocasiones de gol creadas en grandes competiciones y tiempo restante de contrato.

En este caso, las redes neuronales han demostrado por cuarta vez consecutiva ser los modelos de machine learning con mejor desempeño para predecir el valor de mercado, destacando en este caso las de dos capas, con los modelos GAM en segundo lugar.

## 6.5. Porteros

Por último, trataremos de encontrar el mejor modelo para los porteros, posición en la que solamente contamos con 185 fichajes, lo cual nos limitará a la hora de obtener un modelo explicativo y extraer conclusiones sobre qué variables han sido más relevantes.

### 6.5.1 Regresión lineal múltiple

En primer lugar, llevaremos a cabo una regresión lineal teniendo en cuenta solamente variables relativas a rendimiento, edad y tiempo de contrato restante, obteniendo los siguientes resultados.



El R-cuadrado y el MSE parecen mejorar a los que teníamos en regresión lineal múltiple. Veamos ahora qué variables han tenido un mayor peso a la hora de entrenar el modelo.

	Variable	Coeficiente
0	Exp_contr	0.314746
25	Reduced_country_from_Inglaterra	0.309177
4	Minutes_pl_BC	0.307844
13	Reduced_team_from_Inter	0.255414
18	Reduced_team_from_Barcelona	0.246829
2	Minutes_pl	0.182139
9	PKSv	-0.173609
8	PKA	-0.159076
29	Reduced_country_from_Alemania	-0.145958
1	Age	-0.141112

Figura 106: Elastic Net para porteros (2)

Vemos que la edad tiene una importancia menor que en los casos anteriores, debido posiblemente a la mayor longevidad de la carrera de los porteros. El tiempo de contrato restante, los minutos jugados en grandes competiciones y provenir de un equipo inglés son las variables con mayor peso en este caso.

### 6.5.3 Árboles de decisión y Random Forest

Agrupo ambos en un mismo apartado debido a que tienen resultados similares: el R-cuadrado es negativo en ambos casos. Esto puede deberse a la escasez de individuos de este tipo, lo cual empobrece su aprendizaje y dificulta extraer patrones que expliquen los valores de mercado.

### 6.5.4 GAM

Tratar de encontrar relaciones no lineales entre el valor de mercado de los porteros y las variables numéricas nos ha llevado a obtener un modelo que explica el 73'4% de la varianza del coste de los fichajes, con un MSE de apenas 0'26, lo que mejora en gran medida los resultados que ya teníamos. Estas han sido las relaciones entre variables.

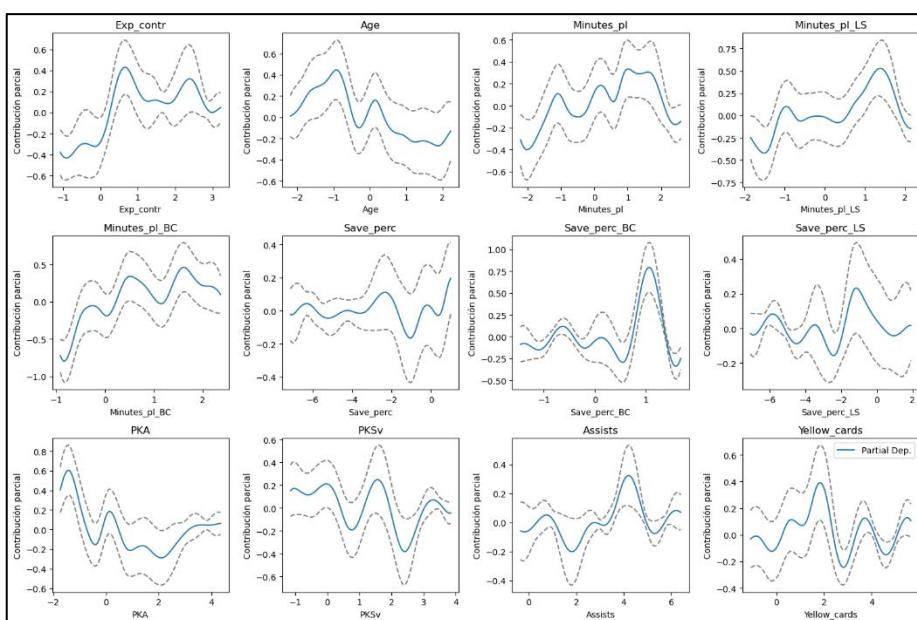


Figura 107: GAM para porteros (1)

Vemos en la Figura 107 que la relación de la edad con el valor de mercado no es lineal, a diferencia de los casos anteriores, y que no parece haber una variable que destaque por



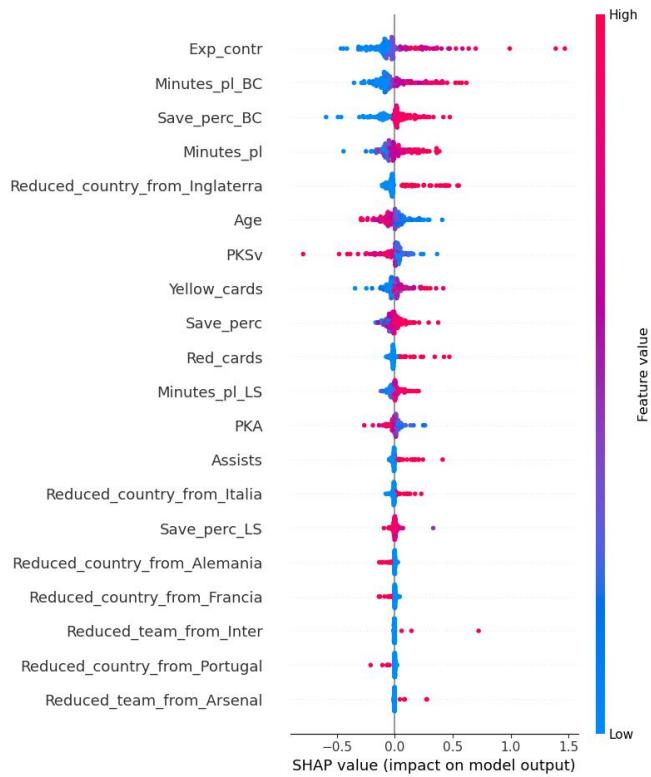


Figura 111: Shapley values de redes neuronales para porteros

Percibimos en la Figura 111 que el tiempo de contrato restante, así como el porcentaje de paradas y las variables relativas a minutos disputados, son las variables que en más fichajes influyen a la hora de estimar sus valores de mercado.

Las redes neuronales han vuelto a destacar por encima del resto de modelos para predecir el valor de mercado, muchos de los cuales han ofrecido resultados muy pobres para el caso de los porteros, debido posiblemente al escaso número de fichajes de los mismos.

## 6.6. Análisis de resultados

Veamos en una tabla los valores de R-cuadrado obtenidos en cada modelo, marcando con un guion los experimentos con resultados negativos o inferiores a 0'05.

	Jugadores de campo	Defensas	Centrocampistas	Delanteros	Porteros
Regresión lineal múltiple	0'49	0'42	0'49	0'52	0'35
Elastic Net	0'49	0'43	0'51	0'52	0'47
Árboles de decisión	0'32	0'45	0'24	0'29	-
Random Forest	0'42	0'54	0'52	0'48	-
GAM	0'56	0'6	0'64	0'65	0'73
EBM	0'43	0'52	0'53	0'5	-
Redes neuronales 2 capas	0'67	<b>0'84</b>	0'73	<b>0'72</b>	0'66
Redes neuronales 3 capas	<b>0'83</b>	0'8	<b>0'79</b>	0'69	<b>0'83</b>

Dilucidamos que las redes neuronales han sido los modelos que mejores resultados nos han ofrecido, seguidas de los modelos GAM. En cuanto al resto de modelos con los que hemos experimentado, los resultados han sido variados dependiendo del conjunto de datos utilizado.

Las variables que más importancia han cobrado a lo largo de todos ellos han sido la edad, que ha afectado de forma negativa al valor de mercado de los futbolistas, los meses de duración de contrato restante y las variables referidas a minutos jugados y ocasiones de gol creadas.

Hemos observado también que las variables de rendimiento han cobrado más valor que aquellas relativas al equipo de origen del futbolista en la mayoría de los casos. Dependiendo de la demarcación del jugador, observamos que hay ciertas variables que han adquirido más importancia que otras en cada caso.

En el caso de los defensas y los centrocampistas, las intercepciones han tenido más importancia que en el resto de jugadores, así como las asistencias en el caso de los segundos, afectando estas variables de forma positiva al valor de mercado de estos jugadores.

En el caso de los delanteros, tal y como era de esperar, los goles que no han sido de penalti también han contribuido a elevar el precio en esta demarcación, así como los goles de penalti en algunos modelos concretos.

En el caso de los porteros, el porcentaje de paradas ha jugado un papel crucial en la mayoría de modelos que hemos utilizado para estimar sus valores de mercado, así como el provenir de un equipo inglés, lo cual también contribuye de forma positiva al precio de los

mismos.

Todo esto es lo que hemos podido conseguir teniendo en cuenta factores de rendimiento y relativos a los equipos de origen de los futbolistas, así como relativos a su situación, tanto en términos de edad como de duración restante de contrato.

De este modo, hemos conseguido modelos que explican en mayor medida la varianza de los valores de mercado de los futbolistas con respecto a los mencionados en el estado del arte, y todo ello usando precios de fichajes reales y sin perder explicabilidad en el proceso.

Hay que tener en cuenta que los factores tenidos en cuenta no lo son todo en el mundo del fútbol, sino que existen más métricas con las que estimar cuánto pagará un equipo por incorporar a un nuevo futbolista a sus filas, como la popularidad en redes sociales, la cual puede aumentar el número de ventas de camisetas del club, el número de entradas vendidas y los ingresos por patrocinios que recibe el equipo en cuestión.

Otros factores que debemos tener en cuenta son la necesidad de un equipo por fichar o vender jugadores, la cual puede influir en el precio del mismo, la situación del jugador dentro de su equipo, su historial de lesiones, su demanda en el mercado, su perfil psicológico o actitud, etc.

Los modelos entrenados en esta obra pueden ser de ayuda a la hora de moverse en el mercado de fichajes, pero es realmente complicado conocer cuánto debe pagar un equipo por un futbolista teniendo en cuenta solamente las variables que hemos utilizado para entrenar modelos, por lo que cada caso debe analizarse de forma individual para estimar una cifra razonable.



## 7. Conclusiones

---

A lo largo de esta obra, hemos abordado el reto de estimar el valor de mercado de los jugadores de fútbol, un elemento esencial para la organización y administración de las plantillas en los distintos equipos. Frecuentemente, los modelos convencionales empleados tienen restricciones importantes, tales como su carácter de modelo de caja negra o el uso de estimaciones poco exactas para calcular los valores de mercado. La incomprensión de los modelos complejos puede obstaculizar la toma de decisiones, lo que acentúa la necesidad de métodos que sean explicables.

Para tratar este problema, hemos puesto en marcha una serie de modelos de aprendizaje automático que no solo facilitan la realización de aproximaciones sobre el valor de los futbolistas, sino también la comprensión de las variables que afectan dichas proyecciones. Hemos empleado métodos como la regresión múltiple lineal, Elastic Net, árboles de decisión, Random Forest y redes neuronales, llevando a cabo múltiples experimentos para evaluar el rendimiento de los modelos. Mediante un estudio detallado de los datos y la utilización de diferentes métodos, hemos obtenido resultados que no solo son robustos en términos de exactitud, sino también en la posibilidad de explicarlos.

Entre los descubrimientos más significativos, podemos aclarar que factores vinculados al desempeño, como los goles, las asistencias y los minutos jugados, sumados a rasgos del jugador como la edad y la duración pendiente del contrato del futbolista, ejercen un efecto considerable en su valor de mercado. Además, hemos comprobado que métodos como las redes neuronales y los modelos GAM proporcionan los mejores resultados en problemas de esta naturaleza.

Respecto al trabajo futuro, este podría centrarse en expandir el conjunto de datos para incorporar más factores, como el impacto de elementos externos como la popularidad en las redes sociales y el historial de lesiones, o bien aspectos relacionados con la relación de un jugador con el club al que pertenece.

## Legado

---

Este trabajo deja como contribución una aproximación a la compleja tarea de estimar el valor de mercado de futbolistas mediante modelos de machine learning. A través del análisis de múltiples algoritmos y conjuntos de datos, tanto de jugadores de campo de todas las posiciones como separados por demarcación, se ha evidenciado que técnicas como las redes neuronales y los modelos GAM pueden ofrecer resultados sólidos y útiles para apoyar la toma de decisiones en contextos reales de mercado, aunque sin llegar a sustituir en ningún momento a la intuición humana.

También se ha logrado identificar un conjunto de variables que han sido especialmente influyentes en la estimación del valor de los jugadores, algunas de ellas diferenciadas por posición en el campo y un conjunto de ellas que afectan a todas las demarcaciones. Este enfoque específico permite avanzar hacia modelos más personalizados y ajustados a las características particulares de cada futbolista, aportando mayor precisión y realismo a las valoraciones.

Esta obra también da a entender que, si bien las variables de rendimiento deportivo, la edad o la duración de contrato son fundamentales, no son las únicas métricas que hay que tener en cuenta para determinar el valor de mercado. Otros factores no directamente cuantificables, como pueden ser la popularidad, el historial de lesiones o la relación del futbolista con su equipo actual, siguen siendo determinantes en la negociación de fichajes.

De esta forma, el presente trabajo no solo ofrece herramientas concretas para el análisis y la valoración de futbolistas, sino que también establece un punto de partida para futuras investigaciones que busquen incorporar dimensiones más amplias en el análisis de este tipo.



## Relación con los estudios cursados

---

Este Trabajo de Fin de Grado está estrechamente vinculado con los contenidos y competencias adquiridas a lo largo del Grado en Ciencia de Datos, abarcando las distintas etapas del ciclo de vida de un proyecto de análisis de datos. Desde el inicio, se ha aplicado el conocimiento adquirido en técnicas de web scraping para la recopilación automatizada de datos desde fuentes online, una habilidad fundamental desarrollada en asignaturas relacionadas con la adquisición de datos.

Posteriormente, se ha llevado a cabo un análisis exploratorio de datos utilizando Python, mediante el cual se han identificado patrones dentro del conjunto de datos, así como correlaciones entre variables que han guiado la preparación del mismo de cara al uso de técnicas de machine learning. Esta fase, esencial en cualquier proyecto de análisis, ha puesto en práctica conocimientos en limpieza, visualización y transformación de datos, competencias clave abordadas en asignaturas como Análisis Exploratorio de Datos, Programación o Visualización de Datos,

El núcleo del trabajo se ha centrado en la aplicación de diversas técnicas de machine learning supervisado, como redes neuronales, modelos GAM y árboles de decisión, para la estimación del valor de mercado de futbolistas. Este enfoque ha permitido no solo poner en práctica los algoritmos estudiados durante la carrera, sino también realizar comparativas de rendimiento, ajustar hiperparámetros y analizar la importancia relativa de las variables en la predicción, habilidades aprendidas en asignaturas como Técnicas escalables en Aprendizaje Automático o Modelos Estadísticos para la Toma de Decisiones.

En conjunto, este trabajo representa una síntesis de los conocimientos adquiridos en el Grado en Ciencia de Datos, demostrando no solo el dominio de las herramientas utilizadas, sino también la capacidad para abordar un problema del mundo real con aplicando una metodología clara y sacando conclusiones fundamentadas en los datos.

## Referencias

---

- [1] Alonso, Á. B., Fernández, P. S., & de Hoyos, I. U. (2007). El mercado de traspaso de futbolistas: un análisis internacional. *Decisión*, (10), 31-55.
- [2] Goddard, J., & Sloane, P. (Eds.). (2014). *Handbook on the economics of professional football*. Edward Elgar Publishing.
- [3] Huang, C., & Zhang, S. (2023). Explainable artificial intelligence model for identifying Market Value in Professional Soccer Players. *arXiv preprint arXiv:2311.04599*.
- [4] <https://www.mordorintelligence.com/es/industry-reports/football-market>
- [5] <https://www.transfermarkt.com/>
- [6] <https://www.semrush.com/website/>
- [7] Gallego, N. (2022). *Analítica de datos y su influencia sobre la gerencia deportiva*.
- [8] [https://optaplayerstats.statsperform.com/en\\_GB/soccer](https://optaplayerstats.statsperform.com/en_GB/soccer)
- [9] <https://es.whoscored.com/>
- [10] Asif, R., Zaheer, M. T., Haque, S. I., & Hasan, M. A. (2016). Football (soccer) analytics: A case study on the availability and limitations of data for football analytics research. *International Journal of Computer Science and Information Security*, 14(11), 516.
- [11] Rodriguez De la Torre, A., & Valencia Cárdenas, J. P. (2024). Impacto de la analítica de datos en las instituciones deportivas: Un enfoque en el fútbol profesional colombiano.
- [12] Franceschi, M., Brocard, J. F., Follert, F., & Gouguet, J. J. (2024). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys*, 38(3), 577-600.
- [13] Coates, D., & Parshakov, P. (2022). The wisdom of crowds and transfer market values. *European Journal of Operational Research*, 301(2), 523–534.
- [14] González Correa, R. (2023). *Determinantes del valor de mercado de un jugador de LaLiga* (Bachelor's thesis).
- [15] He, M., Cachuch, R., & Knobbe, A. J. (2015, June). Football Player's Performance and Market Value. In *Mls@ pkdd/ecml* (pp. 87-95).
- [16] Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Datadriven estimation of market value in association football. *European Journal of Operational Research*, 263(2), 611-624.
- [17] Majewski, S. (2016). Identification of factors determining market value of the most valuable football players. *Central European Management Journal*, 24(3), 91-104.
- [18] Metelski, A. (2021). Factors affecting the value of football players in the transfer market. *Journal of Physical Education and Sport*, 21, 1150-1155.

[19] Kologlu, Y., Birinci, H., Kanalmaz, S. I., & Ozyilmaz, B. (2018). A multiple linear regression approach for estimating the market value of football players in forward position. arXiv preprint arXiv:1807.01104.

[20] <https://fbref.com/>

[21] <https://github.com/JaseZiv/worldfootballR>

# Glosario

---

En el presente apartado introduciremos términos relativos al análisis deportivo con la idea de ofrecer al lector información que le será útil a la hora de comprender el trabajo realizado.

## Analítica deportiva

- Traspaso: proceso mediante el cual un jugador cambia de equipo, ya sea mediante la venta, cesión o intercambio del mismo.
- Mercado de fichajes: periodo del año mediante el cual están permitidos los traspasos de jugadores.
- Fair play financiero: conjunto de normas implementadas para equilibrar los ingresos y gastos de los clubes de fútbol, evitando que estos incurran en deudas excesivas.

## Analítica en fútbol

- Gol: acción en la que un equipo consigue introducir el balón en la portería rival, marcando un tanto a favor de su equipo.
- Asistencia: pase que realiza un jugador y que es inmediatamente anterior a un tiro que acaba en gol por parte de su equipo.
- Tiro: acción de golpear el balón con cualquier parte del cuerpo reglamentaria, generalmente con el pie, con el objetivo de conseguir un gol. Si el balón lleva dirección a portería, se considera un tiro a puerta.
- Entrada: acción en la que un jugador intenta detener el avance del balón por parte de un oponente.
- Regate: movimiento mediante el cual un jugador intenta superar a su oponente con el balón y avanzar hacia la portería contraria.
- Centro: pase dirigido al área rival con la intención de que un compañero remate a portería.
- Parada: acción en la cual un portero detiene un disparo a puerta de un jugador rival, evitando que acabe en gol.
- Duelo aéreo: lance del juego en el que dos o más jugadores de equipos distintos intentan golpear con la cabeza un balón que está en el aire, ya sea para hacerse con el control del mismo, rematar a portería o despejarlo.
- Tarjeta amarilla: amonestación emitida por el árbitro a un jugador por incumplimiento de las reglas del juego. Si se muestran dos a un mismo jugador, este será expulsado del partido.
- Tarjeta roja: sanción mostrada por el árbitro a un jugador para expulsarlo permanentemente del partido debido a faltas graves, como juego violento, comportamiento violento o acumulación de dos tarjetas amarillas.



- Penalti: ocasión de gol de la que dispone un equipo cuando uno de sus integrantes ha sido objeto de falta dentro del área rival, y en la que se cuenta con muchas posibilidades de anotar un gol.
- Portero: conocido como “goalkeeper” en inglés, es el jugador encargado de defender la portería, evitando que el balón entre en la misma.
- Jugador de campo: designación que se da a los futbolistas que no ocupan la posición de portero.
- Defensa central: conocido como “central back” en inglés, es un jugador de corte defensivo que juega en el centro de la defensa, siendo el encargado de evitar que los jugadores rivales consigan rematar balones a portería.
- Lateral: conocido como “left back” o “right back” en inglés, dependiendo de en qué banda se ubique, es un futbolista de corte defensivo que juega por las bandas.

Su principal función es evitar que el equipo rival centre al área o llegue a posiciones de disparo, así como apoyar en ataque y enviar centros al área rival.

- Centrocampista: conocido como “midfielder” en inglés, es un futbolista que juega en el centro del campo. Su función principal es la de conectar el ataque con la defensa, recuperar balones y crear ocasiones de gol.

Dependiendo de su posición específica, podemos clasificarlos en mediocentros defensivos, mediocentros ofensivos, etc.

- Extremo: conocido como “right winger” o “left winger” en inglés, se trata de un jugador de corte ofensivo que se posiciona en las bandas, tratando de desbordar, centrar al área, dar asistencias o disparar a portería.
- Delantero centro: conocido como “striker” en inglés, es un futbolista cuya función principal es la de finalizar jugadas y tratar de marcar goles.

# Objetivos de desarrollo sostenible

---

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. <b>Fin de la pobreza.</b>				x
ODS 2. <b>Hambre cero.</b>				x
ODS 3. <b>Salud y bienestar.</b>				x
ODS 4. <b>Educación de calidad.</b>				x
ODS 5. <b>Igualdad de género.</b>				x
ODS 6. <b>Agua limpia y saneamiento.</b>				x
ODS 7. <b>Energía asequible y no contaminante.</b>				x
ODS 8. <b>Trabajo decente y crecimiento económico.</b>		x		
ODS 9. <b>Industria, innovación e infraestructuras.</b>	x			
ODS 10. <b>Reducción de las desigualdades.</b>	x			
ODS 11. <b>Ciudades y comunidades sostenibles.</b>				x
ODS 12. <b>Producción y consumo responsables.</b>				x
ODS 13. <b>Acción por el clima.</b>				x
ODS 14. <b>Vida submarina.</b>				x
ODS 15. <b>Vida de ecosistemas terrestres.</b>				x
ODS 16. <b>Paz, justicia e instituciones sólidas.</b>		x		
ODS 17. <b>Alianzas para lograr objetivos.</b>				x

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Tal como se ha indicado en la tabla anterior, el trabajo realizado puede llegar a tener distintos grados de relación con un total de cuatro objetivos de desarrollo sostenible.

En el objetivo de trabajo decente y crecimiento económico por la aportación de la inteligencia artificial a la innovación del deporte, concretamente en el plano económico del mismo.

El uso de modelos predictivos basados en inteligencia artificial permite una mayor precisión y objetividad en la valoración de jugadores, optimizando la toma de decisiones de clubes, agentes y analistas deportivos.

Esto, a su vez, genera nuevas oportunidades laborales en áreas como el análisis de datos deportivos, la gestión económica de clubes y la consultoría especializada en valoración de talento.

También está estrechamente relacionado con el objetivo de industria, innovación e infraestructuras, ya que la aplicación de modelos de inteligencia artificial para la estimación del valor de mercado de futbolistas representa un claro ejemplo de cómo la tecnología puede transformar industrias tradicionales, como la deportiva, mediante la innovación y la digitalización.



El uso de machine learning en el análisis de jugadores permite una evaluación más precisa y objetiva de su rendimiento y valor económico, lo que mejora la toma de decisiones dentro de los clubes y otras entidades del ecosistema futbolístico. Esto impulsa la modernización de la industria del deporte, promoviendo una gestión más eficiente y basada en datos.

Además, el desarrollo e implementación de estas tecnologías fomenta la creación de infraestructuras digitales avanzadas en el ámbito deportivo, como bases de datos optimizadas, plataformas de análisis automatizado y herramientas predictivas para la gestión financiera de los clubes.

Esta transformación no solo aumenta la competitividad y sostenibilidad del sector, sino que también incentiva la inversión en innovación y en profesionales especializados en ciencia de datos aplicada al deporte.

Asimismo, cabe mencionar que esta obra contribuye en gran medida al objetivo de reducción de las desigualdades, ya que promueve un enfoque más equitativo y justo en la valoración de futbolistas a través del uso de inteligencia artificial y machine learning.

Tradicionalmente, la evaluación y el valor de mercado de los jugadores han estado influenciados por factores subjetivos como la nacionalidad, la procedencia del club de origen o incluso sesgos inconscientes relacionados con la raza o la reputación mediática.

Esta obra contribuye a que el análisis de futbolistas se lleve a cabo única y exclusivamente a partir de criterios puramente objetivos y relacionados con su rendimiento deportivo, dejando atrás sesgos por raza, nacionalidad u otros motivos.

También guarda cierta relación con el objetivo de paz, justicia e instituciones sólidas, dado que aporta transparencia a la valoración de futbolistas, ayudando de este modo a reducir el fraude en fichajes, ya que contribuye a evitar situaciones en las que un equipo pague un precio excesivo por un jugador para evitar sanciones por fair play financiero a un club con el que mantiene buena relación.

Esto, a su vez, contribuye a la consolidación de instituciones deportivas más sólidas, donde la toma de decisiones se base en criterios verificables y no en acuerdos informales o prácticas poco transparentes.