Los algoritmos como herramienta para el éxito en el fútbol moderno

Luis Melián Peña

Grado en Business Analytics

Curso 2023/2024

Convocatoria ordinaria

Mayo 2024

AGRADECIMIENTOS

Quiero mostrar mi agradecimiento a las siguientes personas y entidades. Sin ellas no habría sido posible la realización de este Trabajo de Fin de Grado:

A mis padres, ustedes son mi inspiración y ejemplo a seguir. Gracias por acompañarme en cada paso que doy, por el amor incondicional, el constante apoyo y por ser mi principal fuente de motivación. Ninguna palabra será suficiente para agradecerles todo lo que hacen por mí.

A mi tutora, María Jesús Gómez y a la profesora Ana Lazcano, por su tiempo, paciencia y orientación dedicados a la elaboración de este proyecto. Sus conocimientos y sugerencias fueron fundamentales en cada etapa de este trabajo.

A You First, la empresa donde realicé mis prácticas curriculares, por facilitarme la búsqueda de los datos y permitirme conciliar en todo momento mi trabajo con la asistencia a clases y la realización de este proyecto.

A la Universidad Francisco de Vitoria, por brindarme la oportunidad de crecer académicamente y proporcionarme los recursos necesarios para llevar a cabo este proyecto. También quiero agradecer a mis compañeros de clase por su compañerismo y apoyo durante este proceso.

Y, por último, a todas aquellas personas que han ayudado a que este trabajo sea hoy una realidad.

RESUMEN

Este trabajo explora el uso de algoritmos como herramienta para mejorar el rendimiento y la gestión en el fútbol moderno. Por un lado, se emplea un algoritmo k-means para identificar las canteras de fútbol más productivas en Europa, concluyendo que el FC Barcelona y el Real Madrid son las más exitosas. Por otro lado, se utiliza un análisis de componentes principales acompañado de un modelo de regresión lineal para determinar las variables más influyentes en el valor de mercado de los jugadores. Las variables más relevantes identificadas son: la altura, la puntuación Elo y el potencial del jugador.

Posteriormente, se presenta el modelo de negocio TalentScope, el cual evidencia la capacidad de los algoritmos para proponer soluciones innovadoras y exitosas en el ámbito del fútbol. Esta plataforma está diseñada para evaluar y predecir el potencial de jugadores de fútbol mediante el uso de algoritmos e inteligencia artificial. Su público objetivo serán clubes, agencias de representación y federaciones deportivas. La plataforma obtendrá beneficios a través de suscripciones anuales básicas y premium.

PALABRAS CLAVE

Canteras productivas de fútbol, predicción de talentos, valor de mercado de jugadores, algoritmos, aprendizaje automático.

ABSTRACT

This project explores the use of algorithms as a tool to improve performance and management in modern football. On the one hand, a k-means algorithm is used to identify the most productive football academies in Europe, concluding that FC Barcelona and Real Madrid are the most successful. On the other hand, a principal component analysis and a linear regression model are used to determine the most influential variables in the market value of the players. The most relevant variables identified are height, Elo score, and player potential.

Subsequently, the TalentScope business model is presented, demonstrating the ability of algorithms to propose innovative and successful solutions in football. This platform evaluates and predicts soccer players' potential through algorithms and artificial intelligence. Its target audience will be clubs, representation agencies, and sports federations. The platform will be profitable through basic and premium annual subscriptions.

KEYWORDS

Productive football academies, talent prediction, players' market value, algorithms, machine learning.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

LISTA DE ACRÓNIMOS

ACRÓNIMO	SIGNIFICADO
PCA	Principal Component Analysis
ETL	Extract, Transform and Load
PC1	Componente principal 1
PC2	Componente principal 2
MSE	Mean Squared Error
IA	Inteligencia artificial

1. DESCRIPCIÓN DEL PROBLEMA A RESOLVER

El fútbol en Europa representa no solo un espectáculo deportivo seguido por millones de personas, sino también una industria con un impacto económico significativo. La constante búsqueda de mejoras en el rendimiento y estrategias de mercado por parte de los clubes destaca la necesidad de innovación. No obstante, la brecha económica entre equipos y el acceso desigual a recursos tecnológicos plantean un desafío para mantener un entorno competitivo equitativo. La adopción de algoritmos de aprendizaje automático se perfila como una solución prometedora, ofreciendo una metodología avanzada para la identificación y evaluación de talentos, y abriendo nuevas vías para equilibrar la competencia y gestión de recursos en el fútbol europeo.

2. MOTIVACIÓN Y ORIGEN

Este trabajo se inspira en la necesidad de democratizar las oportunidades en el fútbol europeo, facilitando a todos los clubes, independientemente de su tamaño o presupuesto, la posibilidad de competir en igualdad de condiciones. La motivación principal es explorar cómo las herramientas de análisis de datos y aprendizaje automático pueden revolucionar la identificación y desarrollo de talentos futbolísticos, contribuyendo así a una distribución más equitativa de oportunidades y recursos. La iniciativa para este estudio parte de una observación detallada de las tendencias actuales en el fútbol, donde la tecnología juega un papel cada vez más crucial en la configuración del futuro del deporte.

La capacidad para predecir el potencial de desarrollo de jugadores y el impacto económico de las decisiones deportivas mediante la aplicación de algoritmos abre un nuevo horizonte en la gestión deportiva. Este enfoque innovador no solo tiene el potencial de mejorar el rendimiento en el campo, sino que también promete transformar la estructura financiera y operativa de los clubes, permitiéndoles tomar decisiones más informadas y estratégicas. Al analizar detenidamente las prácticas existentes y proponer soluciones innovadoras, este trabajo aspira a sentar las bases de una transformación en la gestión del talento y la estrategia deportiva, promoviendo un ambiente más competitivo en el fútbol europeo.

3. ESTADO DEL ARTE

El uso de algoritmos en el fútbol moderno ha aumentado en los últimos años para tomar decisiones más informadas en los ámbitos del scouting de jugadores, la predicción de resultados y el análisis de rendimiento. Tanto el clustering como el análisis de componentes principales (PCA) se han convertido en técnicas fundamentales para analizar datos en el fútbol de hoy en día. Estas técnicas ofrecen una amplia serie de ventajas para los clubes y los jugadores, como la capacidad de identificar grupos de

datos similares, la reducción de la dimensionalidad de los datos y la facilidad de interpretación de los resultados.

En el fútbol, el clustering se ha utilizado para una variedad de propósitos, como la identificación de canteras productivas, jugadores con potencial y partidos similares. En un artículo publicado en la revista Expert Systems with Applications, se propuso un enfoque basado en clustering para clasificar a los jugadores de fútbol según su estilo de juego mediante el análisis de datos de rendimiento de jugadores de la Primera División española (*Alonso-Fernández et al., 2018).* Los datos se analizaron utilizando un algoritmo de clustering jerárquico, y los clústeres resultantes se interpretaron en términos de roles de jugador y estilos de juego. Los autores mostraron que el enfoque propuesto era capaz de identificar clústeres significativos de jugadores y que estos podían utilizarse para mejorar la comprensión del rendimiento de los jugadores. Sugerían que el enfoque propuesto podría utilizarse por entrenadores, ojeadores y analistas para identificar jugadores talentosos y para desarrollar estrategias de entrenamiento y tácticas efectivas.

Un estudio más reciente utilizó el clustering para identificar las canteras europeas que producían más jugadores profesionales (*Peeters et al., 2022)*. Se utilizó una base de datos de más de 100.000 jugadores de fútbol profesional de todo el mundo. Las variables utilizadas para el clustering fueron la edad de debut, el número de partidos jugados, el número de goles marcados y el valor de mercado de los jugadores. Los resultados mostraron que las canteras de los clubes de la Premier League inglesa eran las más productivas, seguidas de las canteras de los clubes de la Bundesliga alemana y de La Liga española. Las canteras de los clubes de la Liga portuguesa, la Ligue 1 francesa y la Serie A italiana también se encontraban entre las más productivas de Europa.

Por su parte, el análisis de componentes principales se ha utilizado en el fútbol para reducir el número de variables que se tienen en cuenta en el análisis de rendimiento de los jugadores. El estudio "The evaluation of playing styles integrating with contextual variables in professional soccer" (Pino-Ortega et al., 2021) examinó los estilos de juego en el fútbol profesional, integrando variables contextuales. Para ello, se recopilaron datos de 20 equipos de la Superliga China durante la temporada 2019-2020 y se utilizaron tres variables contextuales: el nivel de calidad del equipo, el factor de campo y la fecha del partido. El PCA se utilizó para identificar los principales componentes de los estilos de juego encontrándose los siguientes: intensidad del juego, posesión del balón y defensa. También se utilizó un análisis de regresión logística para examinar la asociación entre los estilos de juego y el rendimiento del equipo. Los resultados mostraron que los equipos que emplearon estilos de juego basados en la intensidad y la posesión del balón tuvieron un mayor rendimiento, siendo capaces de controlar el ritmo del partido y crear más oportunidades de gol.

Algunos años antes se había presentado el análisis "Using machine learning to identify elite soccer players" (*Fawcett et al., 2014)* en el que los autores utilizaron varios algoritmos de aprendizaje automático, incluidas las máquinas de soporte vectorial, para analizar un conjunto de datos de jugadores de la Premier League. Estos algoritmos tenían como objetivo anticipar si un jugador podría alcanzar los 50 partidos en la Premier League, un punto de referencia habitual para los jugadores de élite. Los modelos de aprendizaje automático lograron una asombrosa precisión del 80% en la predicción de jugadores de élite, superando significativamente a los métodos tradicionales de scouting. Esto demostró el potencial de los enfoques basados en datos para identificar talentos individuales que podrían pasar desapercibidos.

Más recientemente, en una ponencia presentada en la Conferencia MIT Sloan Sports Analytics celebrada en Boston, MA. (*Fawcett, 2020)* se explicó el creciente papel del aprendizaje automático en el fútbol moderno. Fawcett, reconocido experto en el campo del análisis de datos deportivos, destacó cómo los algoritmos de aprendizaje automático estaban siendo utilizados por los clubes de fútbol para mejorar sus procesos de reclutamiento, predicción de resultados, análisis de rendimiento y tácticas. Explicó también cómo los modelos de aprendizaje automático eran útiles para predecir los resultados de los partidos con mayor precisión que los métodos tradicionales, teniendo en cuenta una variedad de factores, como estadísticas de jugadores, resultados de partidos anteriores, y condiciones del campo. Según el autor, el futuro del fútbol estará sin duda ligado al aprendizaje automático y los clubes de fútbol que se adapten más rápido a esta tecnología estarán mejor posicionados para el éxito en la era de los datos.

4. OBJETIVOS DEL TFG

Los objetivos de este trabajo son estudiar el uso de algoritmos en el fútbol europeo con el fin de identificar las canteras que proporcionan más jugadores al fútbol profesional y los atributos más influyentes a la hora de estimar el valor de mercado de los jugadores.

5. METODOLOGÍA Y PLAN DE TRABAJO

Primera fase: se realiza la extracción, transformación y carga de datos del proyecto. El objetivo de esta fase es elaborar la base de datos final que será utilizada en la segunda fase del proyecto. Esta base de datos debe ser completa, precisa y consistente, para que los resultados del análisis sean fiables.

Segunda fase: parte técnica del proyecto donde se llevará a cabo el análisis de los datos. El entregable de esta fase serán los distintos análisis necesarios para realizar la toma de decisiones y de donde se sacarán las conclusiones en la tercera fase.

Tercera fase: es importante comprender los resultados de los estudios realizados para el desarrollo del modelo de negocio. Esto significa analizar los datos detenidamente, buscar patrones y tendencias y dar conclusiones que aporten valor.

6. INGENIERÍA DEL DATO

6.1. EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE DATOS

El proceso de extracción, transformación y carga (ETL) consiste en recopilar datos de múltiples fuentes, limpiarlos y convertirlos en un formato compatible con un almacén de datos. Su principal objetivo es facilitar el análisis de datos y la toma de decisiones. Este tipo de procesos puede automatizar tareas manuales, como la recopilación y la carga de datos; de esta forma se liberan recursos para que las organizaciones se puedan concentrar en tareas más estratégicas y ahorrar tiempo. Para este proyecto, el proceso de ETL ha seguido los siguientes pasos:

1º. **Obtención del número de observaciones, variables y tipos de datos:** la base de datos está compuesta por 1893 registros y 21 atributos distintos, lo que proporciona una cantidad considerable de información susceptible de ser analizada estadísticamente de manera sólida. La calidad general de los datos es elevada; sin embargo, se ha detectado la presencia de valores faltantes en ciertas columnas, las cuales necesitan ser revisadas detenidamente. En particular, las columnas 'valor_actual', 'agencia', 'altura' y 'edad_debut' presentan un número reducido de entradas vacías que podrían tener un impacto significativo en cualquier análisis subsecuente.

   La diversidad en los tipos de datos es notable, abarcando desde tipos objeto, que son generalmente cadenas de texto, hasta tipos numéricos y de fecha/hora. Esto sugiere que el conjunto de datos incluye tanto atributos cualitativos como cuantitativos. Resalta la correcta tipificación de la columna 'fecha_nacimiento', como datetime64, lo que facilita cualquier operación que involucra cálculos de edad o intervalos de tiempo.

   Surge una preocupación en cuanto a la columna 'altura', la cual se identifica como tipo objeto. Esto podría implicar la presencia de datos no numéricos o con un formato inconsistente, lo que requeriría una limpieza de datos para convertir estos valores a un formato numérico estándar. De manera similar, la columna “traspasos” se presenta como tipo objeto, lo cual podría indicar una representación textual de lo que se espera sean valores numéricos, como la cantidad de traspasos. La conversión de estos datos a un tipo numérico permitirá un análisis cuantitativo más preciso.

El uso de memoria es de aproximadamente 310.7 KB, lo que se considera manejable y no debería presentar problemas para su manipulación en la mayoría de los entornos de análisis de datos modernos.

2º. **Estandarización de nombres de equipos:** los nombres en las columnas 'equipo' y 'cantera' se han unificado y corregido para que sean consistentes. Para estandarizar los nombres de los equipos, primero identificamos los nombres únicos en las columnas 'equipo' y 'cantera'. Se han identificado algunas irregularidades, como diferencias en la forma de escribir los nombres de los equipos, por ejemplo, entre "Atlético de Madrid" y "Atletico de Madrid", "FC Barcelona" y "Barcelona", "Real Madrid" y "Madrid", entre otros.

3º. **Eliminación de jugadores duplicados**: se ha procedido a eliminar todas las entradas duplicadas de jugadores identificadas mediante este código. Además, se ha constatado que existen dos jugadores con el nombre de Nuno Santos y dos jugadores con el nombre de Adama Traoré. Por esta razón, se mantendrán ambos registros en la base de datos sin ser eliminados.

	nombre	equipo	liga	posicion	fecha_nacimiento	edad	cantera
205	Ignasi Miquel	Granada	Primera España	Defensa central	1992-09-28	31.00000	Arsenal FC
206	Álvaro Carreras	Granada	Primera España	Lateral izquierdo	2003-03-23	20.00000	Manchester United
209	José Pozo	Rayo Vallecano	Primera España	Mediocentro ofensivo	1996-03-15	27.00000	Manchester City
211	Pablo Maffeo	Mallorca	Primera España	Lateral derecho	1997-07-12	26.00000	Manchester City
218	Sergi Canós	Valencia CF	Primera España	Extremo izquierdo	1997-02-02	26.00000	Liverpool
220	Arnau Puigmal	Almería	Primera España	Mediocentro ofensivo	2001-01-10	23.00000	Manchester United
448	Paulino de la Fuente	Real Oviedo	Segunda España	Extremo derecho	1997-06-27	26.00000	Inter

Tabla 1. Jugadores duplicados.
Fuente: Elaboración propia.

4º. **Revisión de valores nulos:** se dispone de un total de 202 valores nulos en el conjunto de datos, distribuidos de la siguiente manera: 'valor_actual' (3), 'altura' (55), 'edad_debut' (1) y 'agencia' (143). Es importante decidir cómo manejar estos valores nulos. Se podría considerar reemplazarlos con valores promedio o medianos, o dejarlos como están, dependiendo del contexto y la importancia de cada columna. En este caso, al tener acceso a los datos faltantes de la variable altura desde la fuente de datos utilizada para el trabajo, se han introducido manualmente.

5º. **Sustitución de valores nulos en columna "agencia":** se ha llevado a cabo un proceso de limpieza de datos en el cual los valores ausentes en la columna 'agencia' han sido sustituidos por el término

"desconocido". Este procedimiento ha permitido estandarizar la información contenida en dicha columna, garantizando que todos los registros presenten un valor definido. Como resultado de esta intervención, la columna de agencia ahora presenta una total ausencia de valores nulos, es decir, no hay registros vacíos, lo que facilita el análisis posterior de los datos y mejora la integridad de la base de datos al asegurar que cada jugador tiene asignada una agencia, aunque sea desconocida.

6º. **Identificación de la fila donde la edad de debut es nula para su posterior corrección:** se ha efectuado una actualización puntual en la entrada correspondiente al jugador cuyo campo 'edad_debut' estaba vacío. El valor ausente se ha sustituido por 15, que corresponde a la edad de debut del futbolista Mikki van Sas. Posteriormente, se ha llevado a cabo una comprobación para asegurarse de que la modificación se ha aplicado de manera adecuada en el registro específico.

7º. **Identificación de las filas donde el valor actual es nulo para completarlas:** se ha procedido a la identificación y actualización de las entradas en la base de datos que presentaban valores nulos en la columna 'valor_actual'. Para cada jugador con un valor de mercado ausente, se ha consultado la plataforma Transfermarkt con el fin de obtener la información faltante. En el caso de Juanma García del Albacete, Marcos Navarro del Levante y Óscar Clemente también del Levante, se han añadido sus respectivos valores de mercado que ascienden a 0.7 millones, 0.1 millones y 0.8 millones. Tras realizar estas incorporaciones, se ha verificado que los datos se han incorporado correctamente al conjunto de datos.

8º. **Retención de la primera nacionalidad del jugador:** implementación de una función en el conjunto de datos para retener solamente la primera nacionalidad listada para cada jugador. La columna 'nacionalidad' se ha modificado para mostrar solamente la nacionalidad que el jugador utilizará para representar a su selección nacional en competiciones internacionales.

9º. **Selección de la cantera de debut en caso de debut y de la última cantera en la que jugó en caso de no debutar:** se ha empleado una función en la base de datos para determinar y asignar la cantera de origen de cada jugador. La función opera bajo el siguiente criterio: si un jugador no ha hecho su debut y ha estado en múltiples canteras, se elige la segunda cantera como su lugar de formación; en caso de que el jugador ya haya debutado, se selecciona la primera cantera registrada; y si el jugador solo cuenta con una cantera en su historial, se asigna esa como su cantera de debut. El hecho de que se seleccione únicamente la cantera de debut podría suponer un sesgo de infravaloración de canteras que venden jugadores a tempranas edades.

10º. **Transformación de las columnas “internacionalidad" y “comunitario” en numéricas:** las variables cuantitativas desempeñan un rol fundamental en la ejecución de operaciones matemáticas

y análisis estadísticos avanzados. Son particularmente cruciales en el contexto de los modelos estadísticos y de machine learning, ya que permiten una interpretación y evaluación más directas de los resultados obtenidos. Dichas variables numéricas ofrecen la ventaja de poder ser transformadas a través de procesos de escalado o normalización, lo cual es un paso crítico para optimizar el desempeño de ciertos modelos de aprendizaje automático, especialmente aquellos que son susceptibles a la magnitud de los atributos. Además, cuando se pretende aplicar métodos de reducción de la dimensionalidad, como es el caso del PCA, resulta indispensable que las variables involucradas sean numéricas, pues esto permite calcular con exactitud las direcciones que maximizan la varianza en el conjunto de datos.

11º. **Identificación de posiciones mal escritas o repetidas y estandarización de estas:** el gráfico de caja que se presenta revela ciertas irregularidades en la denominación de las posiciones de los jugadores, señalando la necesidad de adoptar un criterio uniforme para la nomenclatura de dichas posiciones. Se observa que algunas etiquetas están abreviadas o incompletas, lo que podría generar confusión al tratar de identificar las posiciones específicas que representan. Asimismo, se identifican ciertas posiciones que podrían ser percibidas como redundantes o duplicadas, situación que podría llevar a equívocos o indicar un potencial error en la clasificación dentro del conjunto de datos.

Para simplificar el análisis y mejorar la claridad visual de la información, se ha decidido consolidar las distintas posiciones del campo medio, tales como pivote, interior izquierdo, interior derecho, mediocentro ofensivo, mediocentro, centrocampista defensivo y mediocentro defensivo, bajo la categoría general de 'centrocampista'.

Además, se han normalizado las denominaciones de las posiciones restantes para que el conjunto de datos distinga claramente entre ocho categorías de posición: portero, lateral izquierdo, lateral derecho, defensa central, centrocampista, extremo derecho, extremo izquierdo y delantero centro. Esto se realiza con el objetivo de facilitar la interpretación de los datos y evitar ambigüedades en el análisis estadístico posterior.

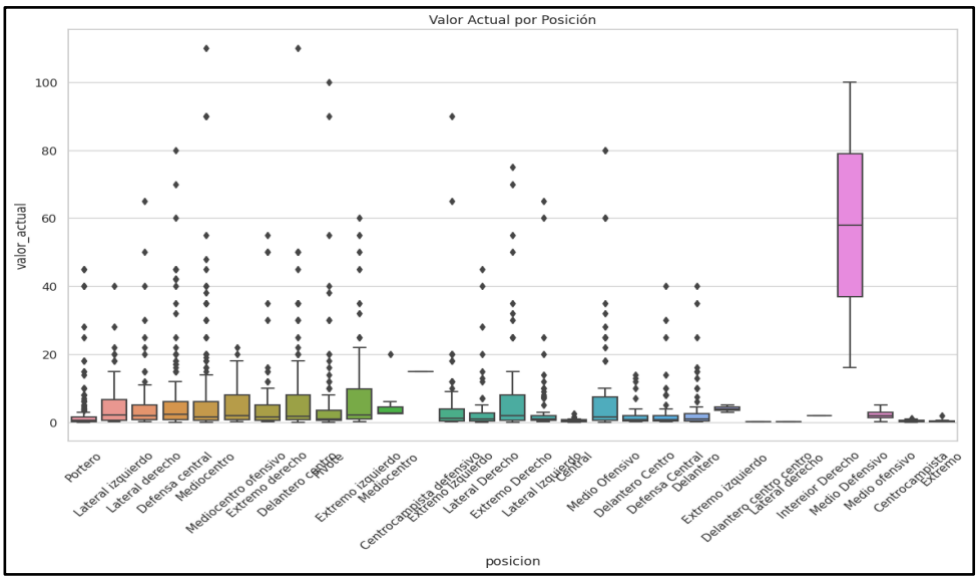

Ilustración 1. Boxplot del valor actual por posición previo a la limpieza.
Fuente: Elaboración propia

12º. **Supresión de columnas irrelevantes y cambios de nombre en variables:** se ha eliminado la columna 'ficha' del conjunto de datos, considerando que contiene enlaces URL a perfiles de jugadores y que estos enlaces son irrelevantes para el análisis en cuestión. Para garantizar la claridad y la funcionalidad de la base de datos, es esencial que los nombres de las variables sean intuitivos y autoexplicativos, permitiendo así a cualquier usuario comprender con facilidad el contenido que representa cada una de las variables. Modificar el nombre de la columna de 'datos' a 'debut' clarifica de manera inmediata que esta sección del conjunto de datos está dedicada a registrar la información relacionada con los partidos disputados con su equipo de debut.

   De manera similar, reemplazar la denominación de 'valor_actual' por 'valor_mercado' mejora la comprensión de los datos, ya que indica con precisión que la variable se refiere al valor de mercado estimado del jugador en cuestión. Estas alteraciones en los nombres de las variables son cruciales para asegurar que la base de datos sea intuitiva y su contenido sea fácilmente comprensible para los usuarios.

13º. **Transformación de la columna “debut” en numérica:** sustitución de la expresión "No debuta" por el número 0 en el conjunto de datos. Este cambio asigna el valor de 0 minutos jugados, lo cual es equivalente a la no participación o debut de un jugador en partidos oficiales. Este ajuste facilita el manejo de la columna, ya que al convertirla en un campo numérico se simplifica el procesamiento y análisis de los datos posteriores.

14º. **Revisión final de valores faltantes y almacenamiento de los datos limpios para su posterior análisis:** finalmente, se ha completado con éxito la limpieza de la base de datos, resultando en una

ausencia total de valores nulos en todo el conjunto de datos. El código ejecutado ha generado un DataFrame de pandas denominado “data”, el cual se ha almacenado en un archivo de Excel titulado 'Europe limpio.xlsx'. Este archivo se sitúa en la raíz del directorio actual del sistema de archivos.

Para proceder con el almacenamiento del archivo, se ha definido la ruta '/Europe limpio.xlsx' en la variable output_file_path. La ruta especificada en esta variable representa el destino y el nombre del archivo en el que se desea guardar el dataset una vez que ha sido depurado y está listo para su uso o análisis posterior.

6.2. ANÁLISIS DESCRIPTIVO DE LOS DATOS

El conjunto de datos actualizado y limpio presenta 1882 filas y 20 columnas, evidenciando una estructura diversificada y una composición meticulosamente curada. Se han depurado los tipos de datos, manteniendo una mezcla equilibrada de textos (objetos) y valores numéricos (enteros y flotantes), garantizando así su relevancia para el análisis. Las columnas como 'nombre', 'equipo', 'liga', 'posición' y 'cantera' destacan por su amplia gama de valores únicos, reflejando la riqueza y la diversidad de la información contenida. Este refinamiento ha enriquecido la integridad de la base de datos, eliminando los valores nulos y consolidando las variables para una interpretación más clara y un análisis estadístico robusto.

6.2.1. DESCRIPCIÓN DE VARIABLES DEL DATASET

VARIABLE	DESCRIPCIÓN	TIPO
nombre	Representa el nombre del jugador.	**object**
equipo	Indica el equipo al que pertenece el jugador.	**object**
liga	Indica la liga en la que compite el jugador.	**object**
posición	Representa la posición donde el jugador se sitúa en el terreno de juego. La posición de juego es una elección del entrenador del equipo. Hay jugadores que pueden desarrollar varias posiciones; en ese caso, hemos anotado su posición principal.	**object**
fecha_nacimiento	Representada en formato dd/mm/yy.	**object**
edad	Edad actualizada del jugador.	**int64**

cantera	Club de formación del jugador. La cantera de un jugador de fútbol se refiere a las divisiones juveniles donde los jóvenes talentos son reclutados, entrenados y desarrollados para convertirse en futuros jugadores profesionales.	**object**
debut	Número de partidos jugados con el equipo de debut de un jugador. El debut es la primera aparición en un partido oficial con un equipo profesional. En caso de no debutar, aparecen 0 partidos.	**int64**
valor_mercado	Mecanismo que determina la valía de un futbolista en términos económicos, tras conjugar factores relativos al rendimiento deportivo, posibilidades de fichaje, ausencia de lesiones y generación de negocio con contratos de patrocinio. Se expresa en millones de euros.	**float64**
nacionalidad	Indica el país de origen del jugador, afectando su elegibilidad para selecciones nacionales y necesidades de permisos de trabajo.	**object**
valor_equipo	Incluye el valor de sus activos, como el estadio, los jugadores, los contratos de patrocinio y otros activos relacionados con el club. Este valor puede fluctuar con el tiempo y depende de muchos factores, como el éxito deportivo, la marca del equipo, la cantidad de seguidores, entre otros. Se expresa en millones de euros.	**float64**
agencia	Agencia de representación del jugador. Una agencia de representación de jugadores es una entidad que se encarga de representar y gestionar los intereses de deportistas, en particular, jugadores de deportes profesionales. Estas agencias trabajan en nombre de los jugadores para ayudarles a obtener contratos con equipos deportivos, patrocinadores y otros acuerdos comerciales.	**object**
altura	Altura del jugador en centímetros.	**int64**
internacional	Si es jugador internacional o no con su selección nacional. La internacionalidad es la participación de un deportista cuando es seleccionado para representar a su país en competiciones oficiales frente a otras naciones. En tal caso, se dice que el deportista es internacional por su país.	**int64**

edad_debut	Edad al debutar profesionalmente. El debut en el primer equipo suele ocurrir en la adolescencia o la temprana adultez, generalmente entre los 17 y los 20 años, pero esto puede variar. Algunos jugadores talentosos pueden debutar a una edad aún más temprana, mientras que otros pueden hacerlo más tarde, dependiendo de su desarrollo y oportunidades.	**float64**
comunitario	Si el jugador es considerado comunitario en su liga de competición o no. Se considera un jugador extracomunitario a cualquiera que no pertenece a un país de la unión europea. Los futbolistas considerados comunitarios pueden serlo bien por haber nacido en los países de la UE o bien por haber obtenido la doble nacionalidad de forma posterior.	**int64**
traspasos	Cantidad de traspasos que ha producido un jugador durante su carrera. Los traspasos de jugadores de fútbol son acuerdos en los que un club de fútbol permite que uno de sus jugadores pase a formar parte de otro club a cambio de una suma de dinero o, en algunos casos, a cambio de otros jugadores u otros tipos de compensaciones.	**int64**
Elo	Puntuación Elo del jugador. Es el valor que determina el nivel de un jugador, teniendo en cuenta variables como la dificultad de los partidos, rivales, competiciones, goles y asistencias decisivas, estados de forma, etc.	**int64**
potencial	Potencial estimado del jugador. Es una estimación del posible Elo máximo que puede alcanzar un futbolista atendiendo a su rendimiento particular en un momento preciso de su carrera y los años de vida activa que el jugador en cuestión todavía tiene por delante.	**int64**
raza	Raza del jugador diferenciada entre blanco o negro.	**object**

Tabla 2. Descripción de variables del dataset.
Fuente: Elaboración propia.

6.2.2. PERIODICIDAD DE LOS DATOS

La periodicidad de los datos se refiere a la frecuencia con la que se recopilan, actualizan o reportan datos en un conjunto de información. Puede variar según el contexto y el tipo de datos que se esté considerando. En el caso del conjunto de datos utilizado para este proyecto, se recopilaron y actualizaron a finales del año 2023.

Algunos aspectos como el valor de mercado, el valor del equipo y los traspasos pueden actualizarse con cierta frecuencia, posiblemente al final de cada temporada, durante las ventanas de transferencias, o incluso basándose en el rendimiento inmediato de los jugadores y otros factores del mercado. Otros aspectos como la fecha de nacimiento, la nacionalidad o la cantera de la que proceden, son estáticos y no cambian con el paso del tiempo.

6.2.3. MEDIDAS DE TENDENCIA CENTRAL, DISPERSIÓN Y FRECUENCIAS

Las medidas de tendencial central, dispersión y frecuencias son términos estadísticos que describen diferentes aspectos de nuestro conjunto de datos.

MEDIDAS DE TENDENCIA CENTRAL

Estas medidas proporcionan un punto central alrededor del cual se distribuyen los datos. Las medidas de tendencia central más utilizadas son la media y la mediana.

Media: la edad media de los jugadores es de aproximadamente 25 años, lo que sugiere una mezcla de juventud y experiencia dentro de la liga. En promedio, los jugadores juegan alrededor de 17 partidos con su equipo de debut, con una edad de debut aproximada de 17.6 años, lo que indica que muchos de ellos comienzan sus carreras profesionales a una edad temprana.

Desde una perspectiva de mercado, el valor promedio de mercado de un jugador es de 5.44 millones de euros, aunque este número puede variar ampliamente dependiendo de factores como la posición en el campo, el rendimiento y la reputación. Además, el valor promedio de los equipos a los que pertenecen estos jugadores es de 139.23 millones de euros, lo que refleja la considerable inversión económica en el talento futbolístico en Europa.

La estatura promedio de los jugadores es de unos 182 cm, lo cual es consistente con las demandas físicas de los atletas de alto nivel en este deporte. Un dato interesante es que aproximadamente el 25% de los jugadores ha tenido experiencia internacional, lo que destaca la presencia de talento con reconocimiento más allá de sus ligas nacionales.

La mayoría de los jugadores, un 90%, son considerados comunitarios, lo que significa que poseen un pasaporte de la Unión Europea, facilitando su traspaso entre equipos dentro de esta zona sin restricciones por cuestiones de nacionalidad. Hablando de traspasos, en promedio, un jugador ha sido transferido casi 5 veces durante su carrera, indicando un mercado de transferencias bastante activo.

Mediana: la mediana de los datos ofrece una perspectiva más precisa de la situación típica de un jugador de fútbol en este conjunto de datos, al ser menos sensible a valores extremos que la media.

La mediana del valor de mercado es de 1.2 millones de euros, mostrando que más de la mitad de los jugadores están valorados por debajo de la media del valor de mercado, lo que señala una distribución donde unos pocos jugadores de alto perfil podrían estar elevando significativamente el promedio. La mediana de la altura se mantiene en 182 cm, lo que es coherente con el perfil físico esperado para el deporte.

La mediana de la edad es de 25 años. Sin embargo, la mediana de la edad de debut es de 17.5 años, lo que refleja la tendencia de los jugadores a comenzar sus carreras profesionales temprano. En cuanto al movimiento entre clubes, la mediana de traspasos es 4, sugiriendo que los jugadores tienden a cambiar de equipo varias veces a lo largo de su carrera. El puntaje Elo mediano es de 62 y el potencial es de 71, lo que sugiere que hay una buena perspectiva de crecimiento y desarrollo en habilidades futbolísticas en la población de jugadores analizada.

MEDIDAS DE DISPERSIÓN

Estas medidas describen la variabilidad o el grado de dispersión de los datos en torno a una medida de tendencia central. Las medidas de dispersión que se utilizarán son: la varianza, la desviación típica y el rango.

Varianza: la varianza de los datos de los jugadores de fútbol muestra patrones interesantes en cuanto a la dispersión de sus características. La edad y la edad de debut de los jugadores tienen una baja varianza, lo que indica una homogeneidad en estas métricas; la mayoría de los jugadores se encuentran cerca de la edad promedio de 25 años y debutando profesionalmente alrededor de los 17 años.

En contraste, hay una alta varianza en el número de partidos jugados antes del debut y en los valores de mercado y de equipo, lo que refleja una amplia gama de experiencias individuales y diferencias económicas significativas entre los jugadores y los clubes a los que pertenecen. La altura muestra una variabilidad moderada, mientras que la experiencia internacional y la posesión de la nacionalidad comunitaria son características con muy baja varianza, sugiriendo que la mayoría de los jugadores no han jugado a nivel internacional y que casi todos tienen pasaporte de la Unión Europea. La cantidad de traspasos y la clasificación Elo tienen varianzas que indican una diversidad moderada en la movilidad de los jugadores entre clubes y en sus respectivas habilidades.

Desviación típica: la edad de los jugadores tiene una desviación típica baja de 4.24, lo que confirma una concentración de edades alrededor de la media de 25 años. Un valor de 53.39 en la desviación típica de debut sugiere que hay una variabilidad considerable en la cantidad de partidos jugados antes de debutar profesionalmente. El valor de mercado con 11.91 y el valor del equipo con 222.75 también

tienen desviaciones típicas altas, lo que indica una amplia gama de valores económicos entre los jugadores y sus clubes.

La altura presenta una desviación menor de 6.79, lo que indica cierta consistencia en las características físicas de los jugadores. Los traspasos con 3.09 y la clasificación Elo con 11.34 muestran una variabilidad moderada, y el potencial con 8.82 muestra una dispersión significativa, lo que indica diferentes percepciones sobre la capacidad de mejora y desarrollo futuro de los jugadores.

Rango: se ha obtenido un rango de 24 años en la edad de los jugadores, lo cual es consistente con una combinación de jugadores emergentes y veteranos. El valor de mercado tiene un rango de casi 110 millones de euros, evidenciando la existencia de tanto talentos emergentes como estrellas consolidadas con valoraciones económicas muy dispares.
En cuanto al valor del equipo, el rango es superior a 1064 millones de euros, reflejando la heterogeneidad económica entre los clubes de fútbol. La altura varía en 37 cm, sugiriendo una diversidad física acorde con las diferentes demandas de las posiciones en el campo. La variable de internacional muestra un rango de 1, lo que significa que está binariamente distribuida entre jugadores que han y no han jugado a nivel internacional. La edad de debut tiene un rango de 9.4 años, indicando que algunos jugadores comienzan su carrera profesional mucho más temprano que otros.

MEDIDAS DE FRECUENCIA

Son herramientas estadísticas utilizadas para describir la cantidad de veces que ocurre un evento o la cantidad de veces que se presenta un valor particular dentro de un conjunto de datos. Las medidas de frecuencia que se van a utilizar son la tabla de frecuencias y el histograma.

Tablas de Frecuencia: la tabla de frecuencias para la altura muestra una distribución en la que las alturas más comunes se sitúan entre los 178 y 183 centímetros, con un máximo de 134 individuos que miden 185 cm, lo que indica que la mayoría de las personas en este conjunto de datos son bastante altas, lo cual es típico en deportes como el baloncesto o el voleibol. Las alturas menores a 168 cm o mayores a 200 cm son mucho menos frecuentes, lo que sugiere que hay pocos individuos extremadamente bajos o altos en este grupo. Este rango predominante de alturas altas, pero no extremas podría ser indicativo de deportistas profesionales donde la altura es una ventaja, pero no necesariamente el único factor determinante del éxito.

Histogramas: el gráfico de barras mostrado representa la distribución de edades de un conjunto de datos, donde cada barra indica la frecuencia de individuos para cada edad específica. Observamos que las edades están distribuidas a lo largo del eje X, y la altura de cada barra refleja la cantidad de individuos

de esa edad en el conjunto de datos. La representación visual ayuda a identificar rápidamente las edades más comunes y a observar la dispersión de edades dentro del conjunto de datos.

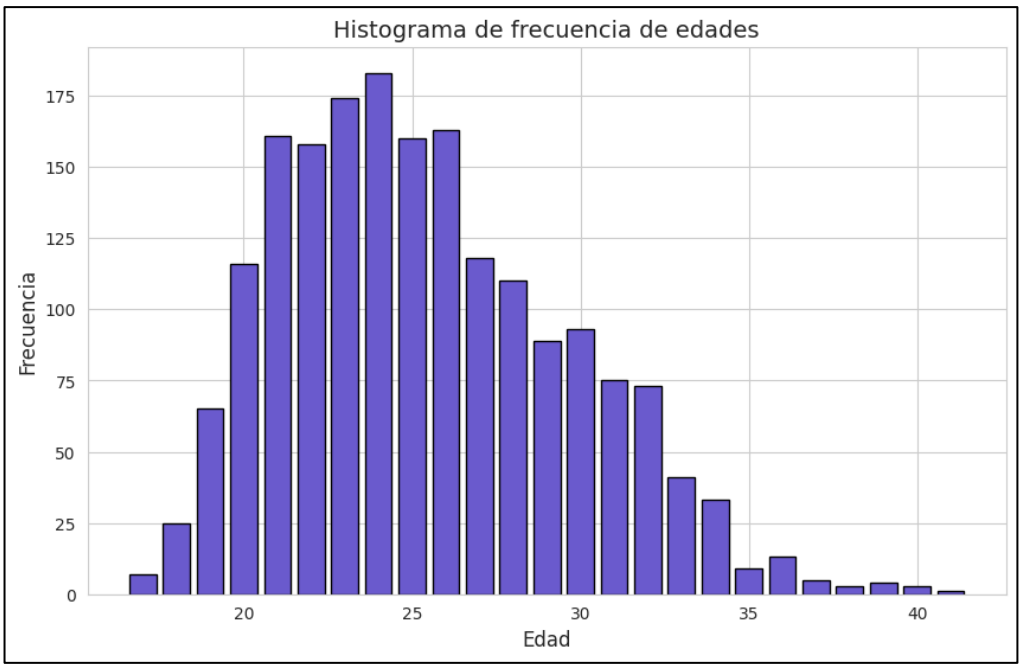

Ilustración 2. Histograma de frecuencia de edades.
Fuente: Elaboración propia

7. GRÁFICOS DESCRIPTIVOS

La sección de gráficos en un análisis descriptivo es fundamental para visualizar y entender mejor los datos con los que estamos trabajando. Esta sección permite presentar de manera clara y eficiente la distribución, tendencias, y patrones subyacentes en los datos a través de diversas formas gráficas como histogramas, boxplots, gráficas de dispersión, diagramas de barras, entre otros. Los gráficos descriptivos no solo facilitan la identificación de características clave de los datos, como la centralidad, dispersión, y la presencia de outliers, sino que también son herramientas esenciales para comunicar hallazgos complejos de manera simple a una audiencia diversa.

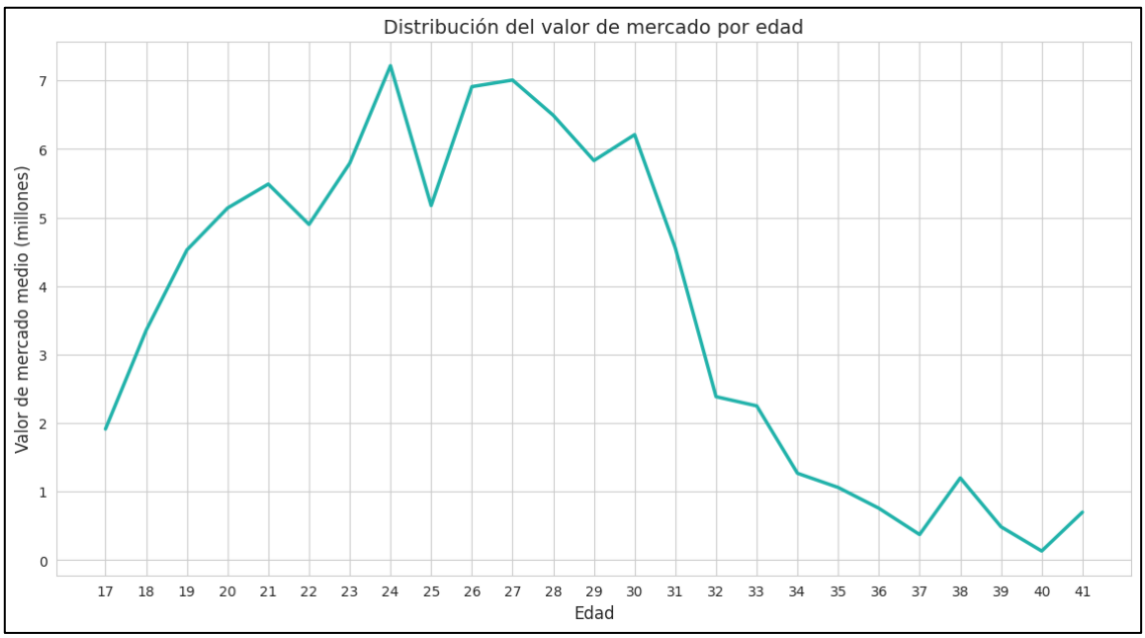

Ilustración 3. Distribución del valor de mercado por edad.
Fuente: Elaboración propia

En este gráfico de líneas se puede observar un aumento en el valor medio de los jugadores hasta los 24 años. Esto podría indicar que los jugadores desarrollan sus habilidades y experiencia hasta este punto, lo que incrementa su valor en el mercado. Este pico es seguido de una disminución pronunciada. Esto puede reflejar que los jugadores alcanzan su máximo valor de mercado aproximadamente en la mitad de sus veinte años, posiblemente debido a que se encuentran en su mejor forma física y han acumulado experiencia significativa sin llegar a la fase donde la edad comienza a afectar su rendimiento.

Después del pico, el valor medio disminuye con la edad, lo que podría sugerir que los clubes y los agentes ven un menor potencial de retorno de inversión en jugadores más mayores debido a la disminución esperada en rendimiento y la menor cantidad de años que les quedan jugando al máximo nivel.

Hacia el final de la carrera de un jugador, el valor medio se estabiliza y hasta muestra ligeros aumentos en ciertas edades. Esto podría deberse a que algunos jugadores mantienen un alto nivel de habilidad y reputación que sostiene su valor a pesar de su avanzada edad.

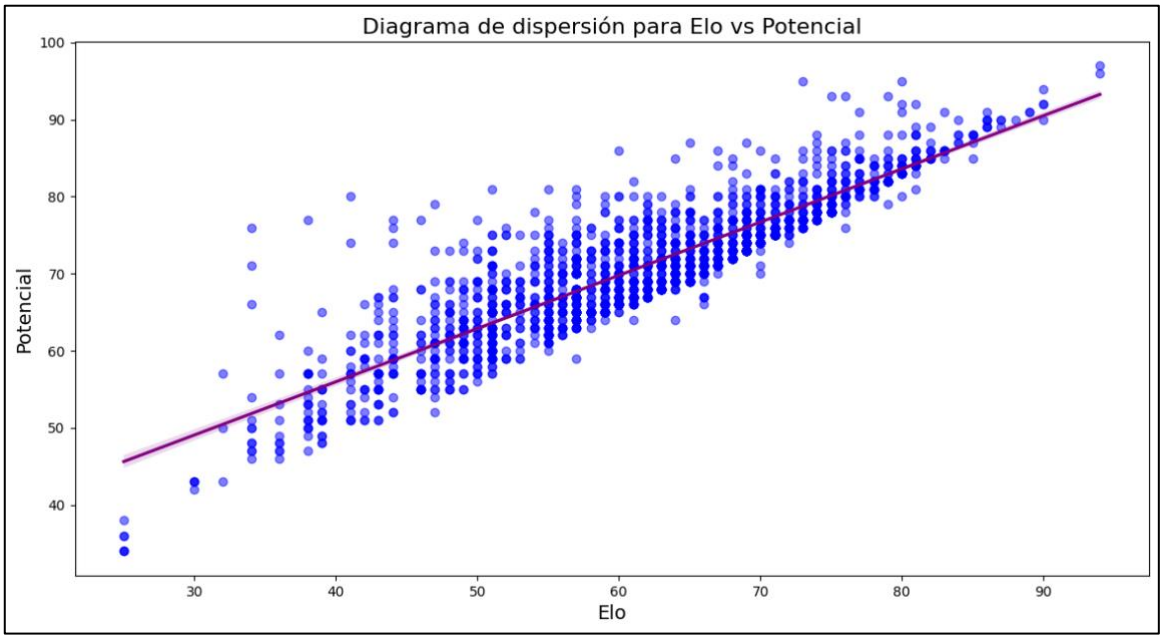

Ilustración 4. Diagrama de dispersión para Elo vs Potencial.
Fuente: Elaboración propia

En esta representación, el Elo varía aproximadamente entre 30 y 90, mientras que el potencial varía entre 40 y 100. Existe una correlación positiva entre las variables Elo y potencial. A medida que aumenta el Elo, también tiende a aumentar el potencial. Se puede apreciar que los datos son más densos en los rangos medios del Elo, aproximadamente entre 50 y 70, donde también hay una mayor variación en el potencial.

La variabilidad del potencial parece disminuir a medida que el Elo aumenta. Esto se puede ver por la forma en que los puntos se concentran en un rango más estrecho de potencial a valores más altos de Elo. Destacan algunos jugadores con un Elo muy bajo y un potencial relativamente alto. Estos podrían ser jóvenes talentos aún no completamente desarrollados o jugadores que no han tenido la oportunidad de demostrar su valía en partidos de alto nivel aún.

La línea de tendencia muestra la relación promedio entre el Elo y el potencial. Su pendiente positiva respalda la correlación positiva entre estas variables. Aunque hay una clara tendencia general, también hay suficiente dispersión de puntos alrededor de la línea.

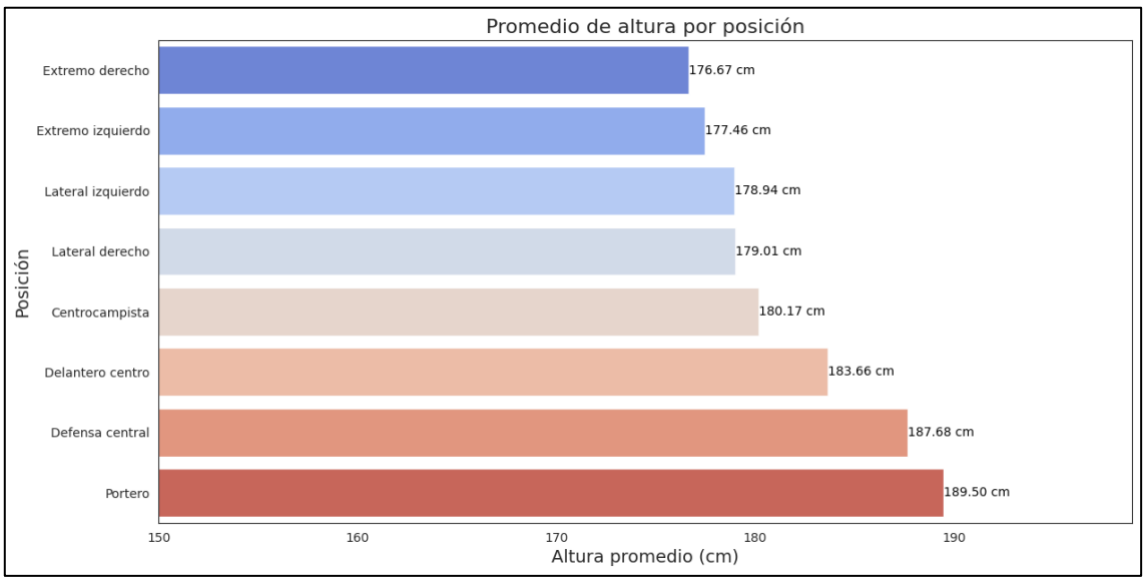

Ilustración 5. Promedio de altura por posición.
Fuente: Elaboración propia

El gráfico de barras horizontales muestra la altura promedio para cada posición. El eje X comienza desde los 150 cm y contiene el valor numérico promedio anotado dentro de cada barra. Los valores son claros y legibles, lo que facilita la comparación directa de las alturas promedio entre las diferentes posiciones.

Las alturas promedio se encuentran en un rango que va desde aproximadamente 176 cm para los extremos hasta casi 190 cm para los porteros, lo que refleja las diferencias físicas típicas asociadas con cada rol en el campo. Los jugadores que juegan pegados a la línea de banda, como pueden ser laterales o extremos, tienden a ser más bajos que los defensas centrales o porteros. Esto se debe a que son posiciones en las que se requiere una gran velocidad. Está demostrado que los jugadores de menor estatura son más veloces que aquellos de mayor estatura.

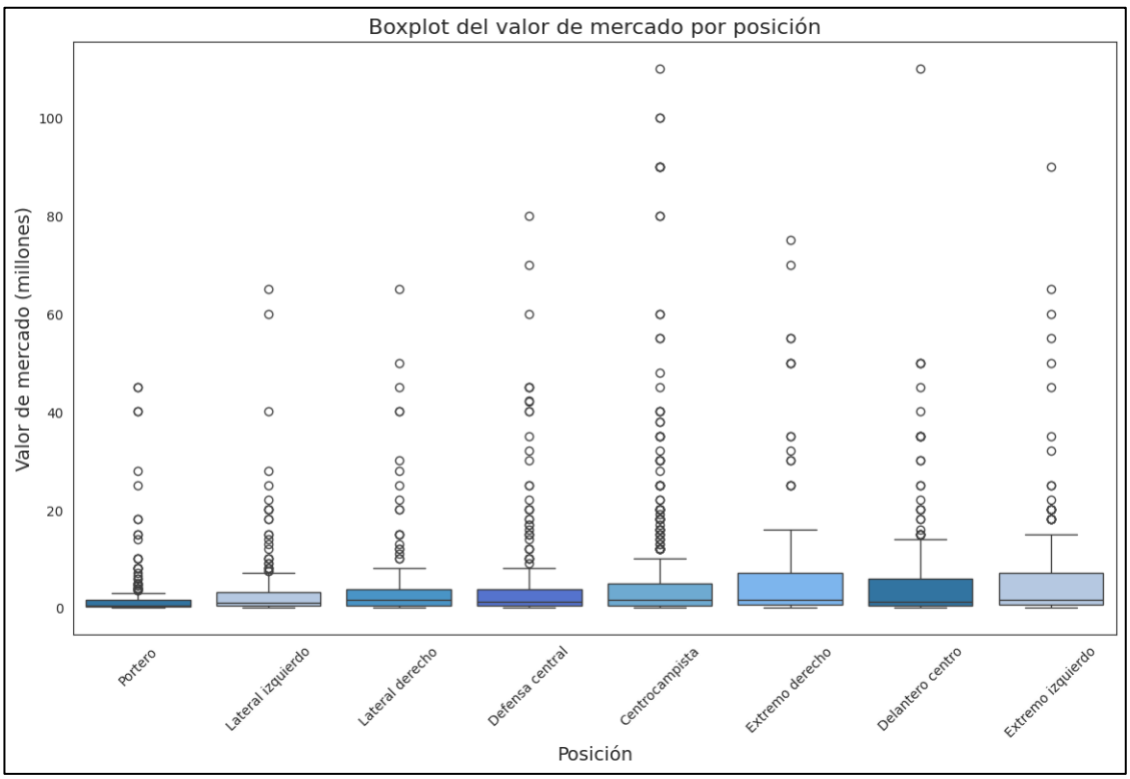

Ilustración 6. Boxplot del valor de mercado por posición.
Fuente: Elaboración propia

En el eje X podemos observar como las posiciones están distribuidas de izquierda a derecha, empezando por los porteros y terminando por los extremos izquierdos. La línea dentro de cada caja indica la mediana de los valores actuales para esa posición. La altura de la caja indica el rango intercuartílico (entre el primer y tercer cuartil), lo que da una idea de la dispersión de los datos alrededor de la mediana.

Los bigotes de los diagramas se extienden hasta los valores máximo y mínimo dentro de un rango que no se considera atípico. Los puntos fuera de los bigotes representan valores atípicos o extremos. Los delanteros centro y extremos derecho e izquierdo parecen tener medianas más altas comparadas con las otras posiciones, lo que podría indicar que, en este contexto específico, estas posiciones están mejor valoradas.

Algunas posiciones como mediocentros y defensa centrales muestran una mayor variabilidad en los valores actuales en comparación con posiciones como los porteros, que tiene una variabilidad más baja y valores más concentrados.

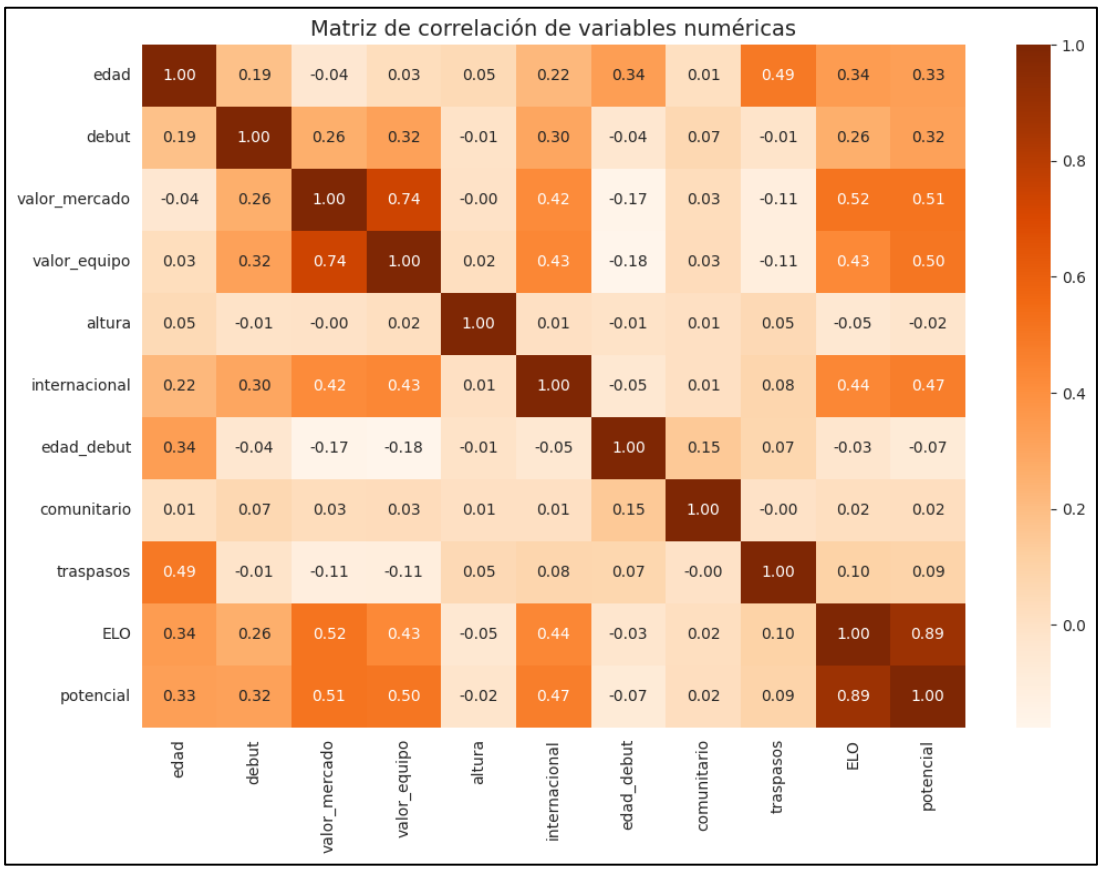

Ilustración 7. Matriz de correlación de variables numéricas.
Fuente: Elaboración propia

En esta matriz, los diferentes tonos de naranja representan la fuerza y la dirección de la correlación entre las variables listadas en las filas y columnas. Los valores de correlación varían de -1 a 1, donde 1 indica una correlación positiva perfecta, 0 indica ninguna correlación y -1 indica una correlación negativa perfecta.

Correlaciones altas: el valor actual del jugador y el valor equipo tienen una correlación de 0.74, lo que sugiere que los valores actuales tienden a ser más altos en equipos de mayor valor. El Elo y el potencial tienen una correlación de 0.89, indicando que un Elo más alto está fuertemente asociado con un mayor potencial.

Correlaciones moderadas positivas: la variable debut y valor de equipo tienen una correlación de 0.32, sugiriendo una relación moderada entre el tiempo de debut y el valor del equipo.

Correlaciones débiles positivas: la edad y el debut tienen una correlación de 0.19, lo que puede sugerir que los jugadores debutan a edades ligeramente variables. La edad del jugador y la edad en el debut tienen una correlación de 0.34, indicando una relación débil.

Correlaciones bajas o insignificantes: la edad y el valor actual del jugador tienen una correlación casi nula de -0.04, lo que significa que no hay una relación lineal aparente entre la edad de los jugadores y su valor actual.

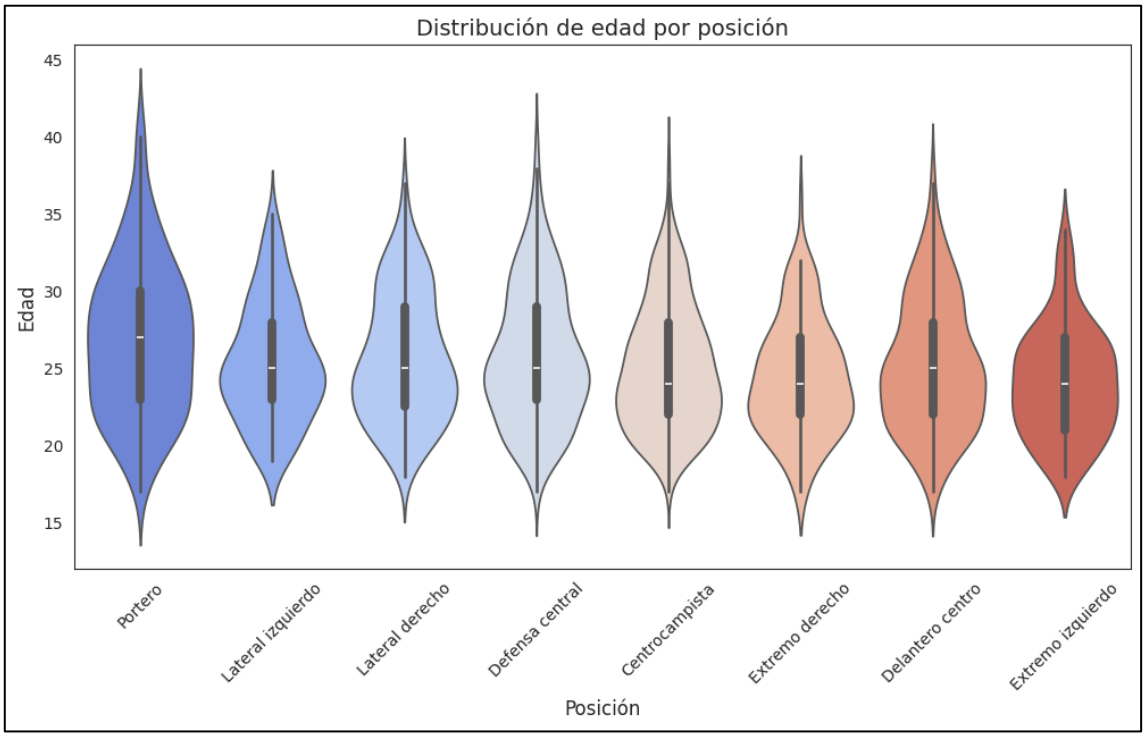

Ilustración 8. Distribución de edad por posición.
Fuente: Elaboración propia

El gráfico muestra un análisis de la edad de jugadores de fútbol en diferentes posiciones en el campo a través de un diagrama de violín combinado con cajas de bigotes. Las posiciones enumeradas son: portero, lateral izquierdo, lateral derecho, defensa central, centrocampista, extremo derecho, delantero centro y extremo izquierdo. Cada violín representa la distribución de la edad para una posición específica, con la parte más ancha indicando la concentración más alta de edades.

Se observa una variabilidad en la distribución por posición, con porteros y defensas centrales mostrando una distribución de edades más amplia y un rango intercuartílico más elevado, lo que sugiere que estas posiciones pueden ser ocupadas efectivamente por jugadores en un rango de edad más extenso. Por otro lado, las posiciones como extremo derecho y extremo izquierdo muestran distribuciones más estrechas y picudas, indicando una menor variabilidad en la edad de los jugadores que ocupan estas posiciones, lo que podría sugerir una preferencia por jugadores más jóvenes debido a la necesidad de velocidad y agilidad en esas áreas del campo. Los bigotes de las cajas, que se extienden desde el cuartil 25 al cuartil 75, y las líneas dentro de las cajas, que representan la mediana, proporcionan un resumen estadístico adicional sobre la distribución de las edades en cada posición.

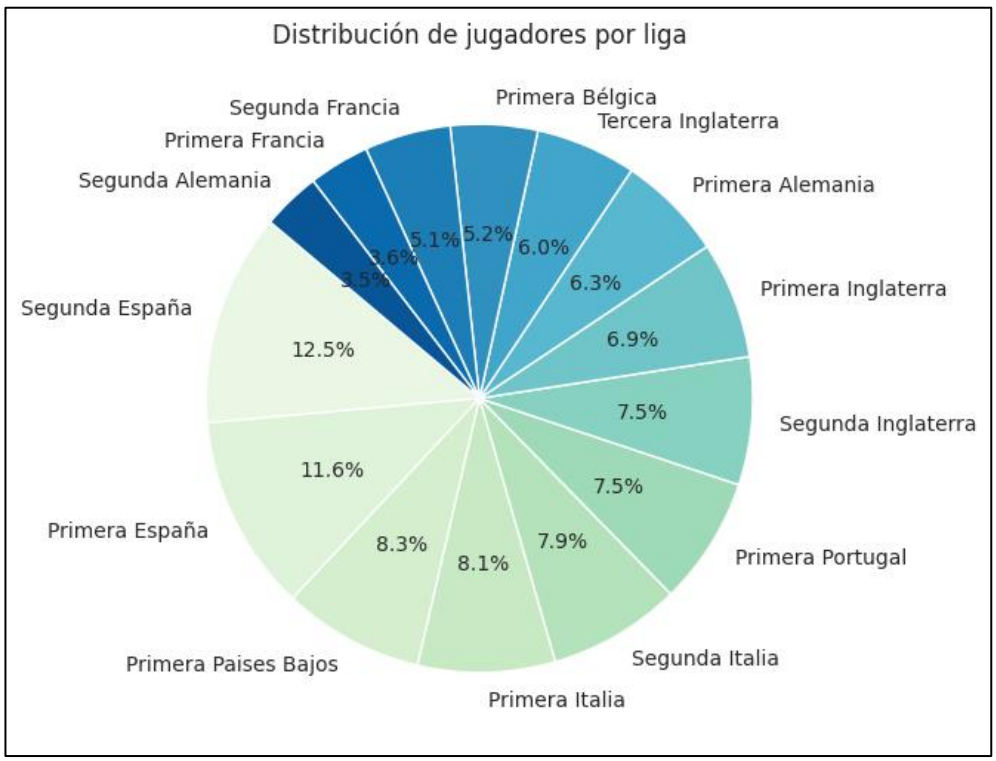

Ilustración 9. Distribución de jugadores por liga.
Fuente: Elaboración propia

La primera y segunda liga española lideran la representación con un 11.6% y 12.5% respectivamente. Esto indica una fuerte presencia de jugadores de estas dos ligas en el conjunto de datos. Existe una buena representación de varias ligas europeas, incluyendo la primera liga de Países Bajos y la primera y segunda liga de Italia, cada una con más del 7% de representación. Por otra parte, ligas de segunda división y otras menos prominentes como la Tercera de Inglaterra y la Primera de Bélgica ocupan segmentos más pequeños, sugiriendo una menor concentración de jugadores en estas categorías. Este tipo de visualización es útil para comparar rápidamente la proporción de jugadores en diferentes niveles competitivos y ligas nacionales.

Estos porcentajes también resaltan la importancia de considerar posibles sesgos en la recopilación y selección de datos. La sobrerrepresentación o subrepresentación de ciertas ligas podría ser un indicativo de sesgos en el proceso de recopilación de datos. Los sesgos en los datos se refieren a distorsiones o desviaciones en los resultados de un análisis debido a la manera en que los datos han sido recogidos, procesados o interpretados. Estos sesgos pueden llevar a conclusiones erróneas y afectar la validez y confiabilidad de un estudio o modelo.

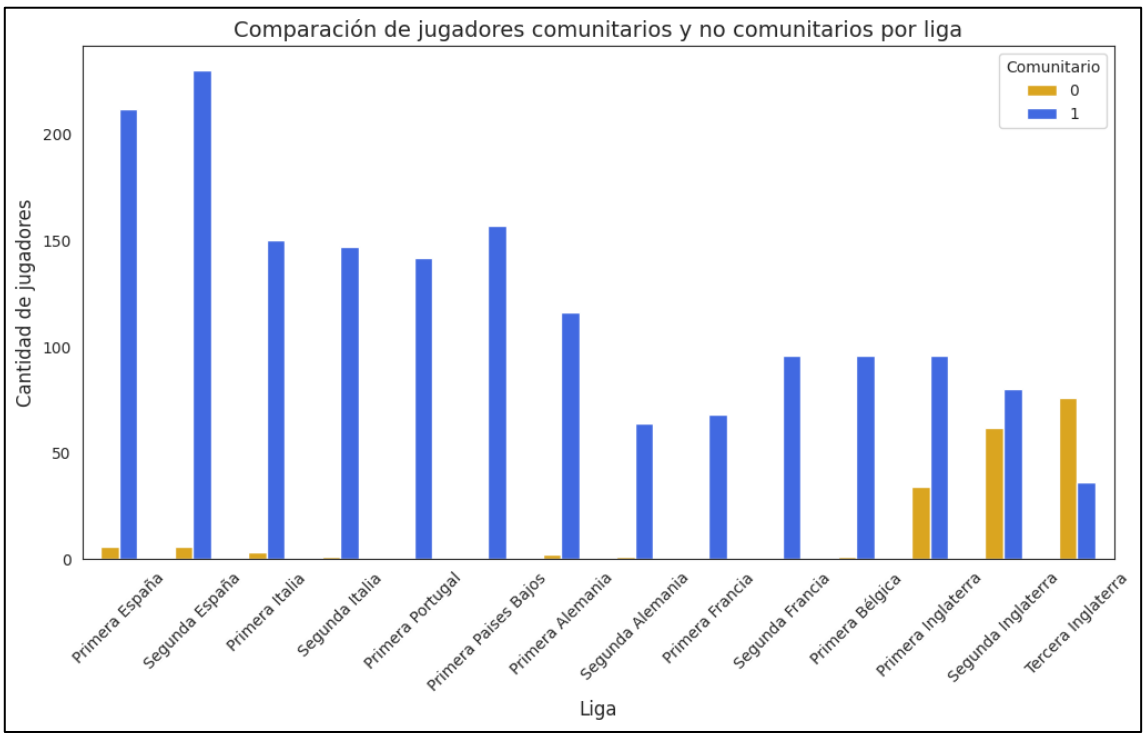

Ilustración 10. Comparación de jugadores comunitarios y no comunitarios por liga.
Fuente: Elaboración propia

En todas las ligas presentadas, la cantidad de jugadores comunitarios es mayor que la de jugadores no comunitarios. La primera y segunda liga española son las ligas con más jugadores comunitarios del set de datos. El término aquí se refiere a jugadores que son parte de la Unión Europea y que tienen ciertas facilidades para jugar en ciertas ligas debido a leyes de trabajo y migración.

La representación gráfica sugiere que las ligas de fútbol en Inglaterra tienen más jugadores no comunitarios en comparación con otras ligas. Esto se debe al impacto del Brexit en el mercado de transferencias y en las regulaciones laborales. Antes del Brexit, las reglas permitían a los clubes de la Premier League y otras divisiones inglesas contratar jugadores europeos con mucha facilidad, lo que les permitía adquirir jóvenes talentos desde los 16 años. Sin embargo, después del Brexit, se han introducido regulaciones más estrictas que afectan la manera en que los clubes ingleses pueden fichar a jugadores de la Unión Europea.

8. MODELOS ANALÍTICOS

Después de realizar el análisis descriptivo de las variables, se procede a emplear los modelos analíticos. Este estudio se enfoca en la exploración de dos modelos analíticos aplicados al fútbol profesional, con el fin de comprender mejor las dinámicas que afectan tanto al desarrollo de talentos en las canteras como a la valoración de los jugadores en el mercado.

El primero de los modelos analizados es el algoritmo k-means, una técnica de agrupamiento que permite clasificar clubes de fútbol según la cantidad de jugadores producidos por sus canteras. El segundo modelo consiste en la aplicación del análisis de componentes principales (PCA) junto con un modelo de regresión lineal para evaluar y comprender las variables que tienen mayor influencia en el valor de mercado de los jugadores.

8.1. ALGORITMO K MEANS

El algoritmo k-means es una técnica de agrupamiento que se utiliza en el aprendizaje automático no supervisado para dividir un conjunto de datos en k grupos o clústeres, basándose en las características de estos. El objetivo es agrupar los datos de tal manera que los puntos dentro de un mismo grupo sean lo más similares posible entre sí, mientras que los puntos en grupos diferentes sean lo más distintos posible. La similitud se suele medir utilizando la distancia entre los puntos, siendo la distancia euclidiana una de las más comunes.

En este estudio, la aplicación del algoritmo k-means tiene el objetivo de identificar las canteras que más jugadores aportan al fútbol profesional. La capacidad de una academia para nutrir y proyectar jugadores hacia el profesionalismo no solo refleja su calidad formativa, sino que también incide directamente en la sostenibilidad y el éxito a largo plazo de los equipos. Ante este escenario, el análisis cuantitativo de las contribuciones de las canteras a nivel profesional se convierte en un campo de estudio de gran relevancia.

Este enfoque no solo facilitará la identificación de las canteras más exitosas, sino que también proporcionará valiosos insights sobre las prácticas y estrategias de desarrollo de jugadores que distinguen a los programas de formación líderes en el mundo. Los resultados de este análisis tienen el potencial de informar decisiones estratégicas tanto para clubes en busca de optimizar sus inversiones en canteras como para agentes y otros actores del fútbol que buscan colaborar con las academias más prometedoras.

8.1.1. MÉTODO DEL CODO

El método del codo es una técnica visual utilizada para determinar el número óptimo de clústeres en un análisis de k-means. Se busca un punto en la gráfica donde el descenso de la curva se aplana significativamente, lo que indica que agregar más clústeres no mejora significativamente la varianza explicada por el modelo. Este punto se denomina “codo”.

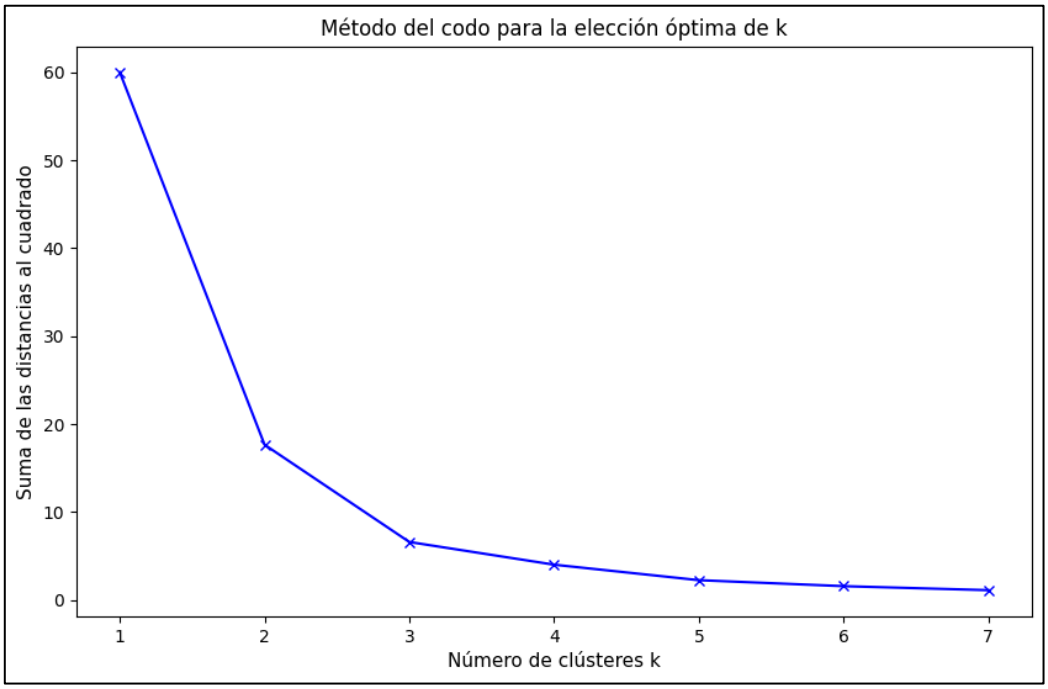

Ilustración 11. Método del codo para la elección óptima de k.
Fuente: Elaboración propia

La línea en el gráfico cae rápidamente a medida que aumentamos k de 1 a 2 y luego se aplana a medida que seguimos aumentando k. El "codo" parece estar alrededor de k=3, lo que sugiere que el número óptimo de clústeres para este conjunto de datos particular es 3. Elegir más de 3 clústeres no mejoraría mucho la compactación de los clústeres.

8.1.2. ANÁLISIS DE SILUETA

La siguiente herramienta presentada es otra técnica utilizada para determinar el número óptimo de clústeres en un conjunto de datos. A través del análisis de silueta, se evalúa la precisión con la que cada elemento ha sido clasificado en su clúster correspondiente, lo cual es un reflejo de la distancia entre los clústeres y la cohesión dentro de ellos.

El valor del coeficiente de silueta varía entre -1 y 1. Valores cercanos a 1 indican que el objeto está bien emparejado con su propio clúster y mal emparejado con clústeres vecinos. Un valor de 0 sugiere que el elemento se sitúa justo en el límite de decisión entre dos clústeres, y los valores negativos indican que esos casos podrían haber sido asignados al clúster incorrecto.

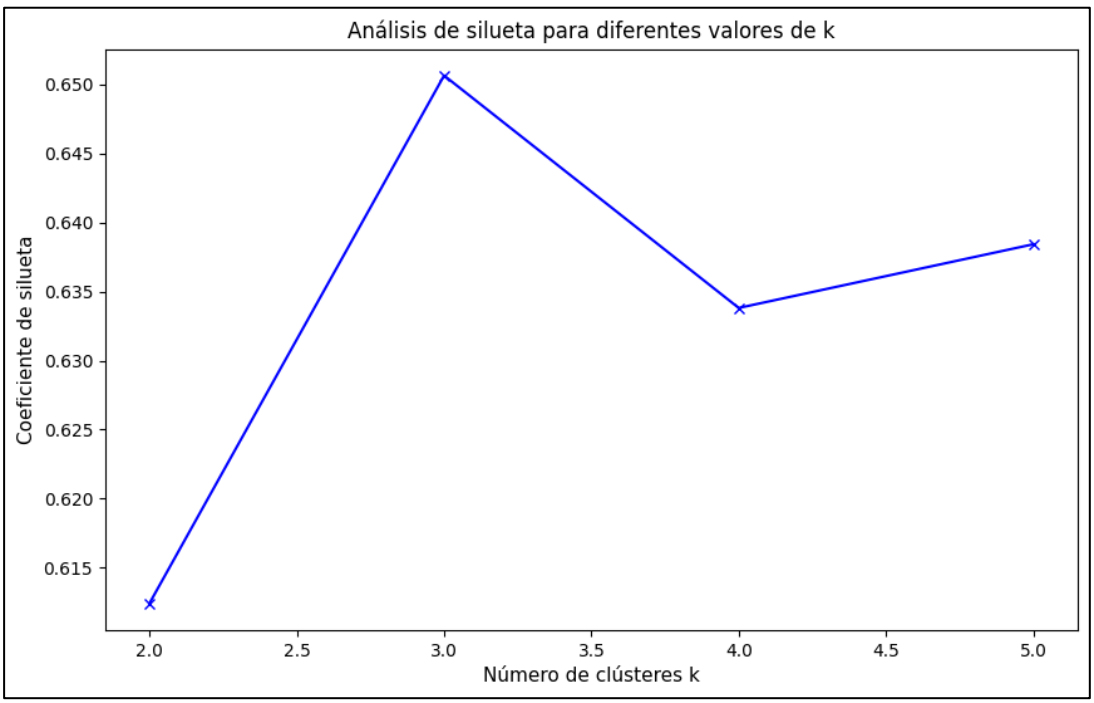

Ilustración 12. Análisis de silueta para diferentes valores de k.
Fuente: Elaboración propia

Se observa un incremento en el coeficiente de silueta al pasar de 2 a 3 clústeres, indicando que un agrupamiento en tres clústeres es más adecuado para estos datos que un agrupamiento en dos. Ya que no se observan mejoras sustanciales en el coeficiente para k=4 o k=5, se puede deducir que un número de clústeres de k=3 es preferente. Esta conclusión es consistente con los resultados obtenidos anteriormente mediante el método del codo.

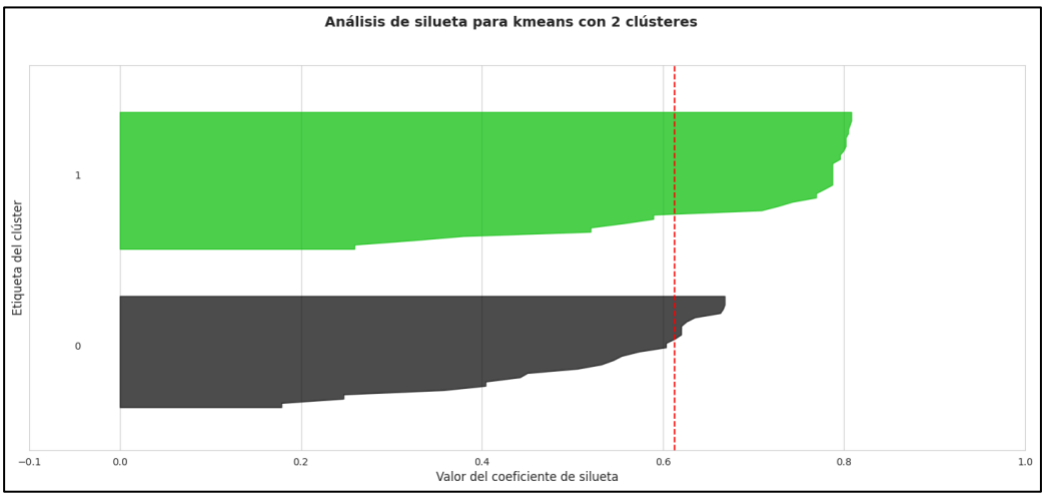

Ilustración 13. Análisis de silueta para kmeans con 2 clústeres
Fuente: Elaboración propia

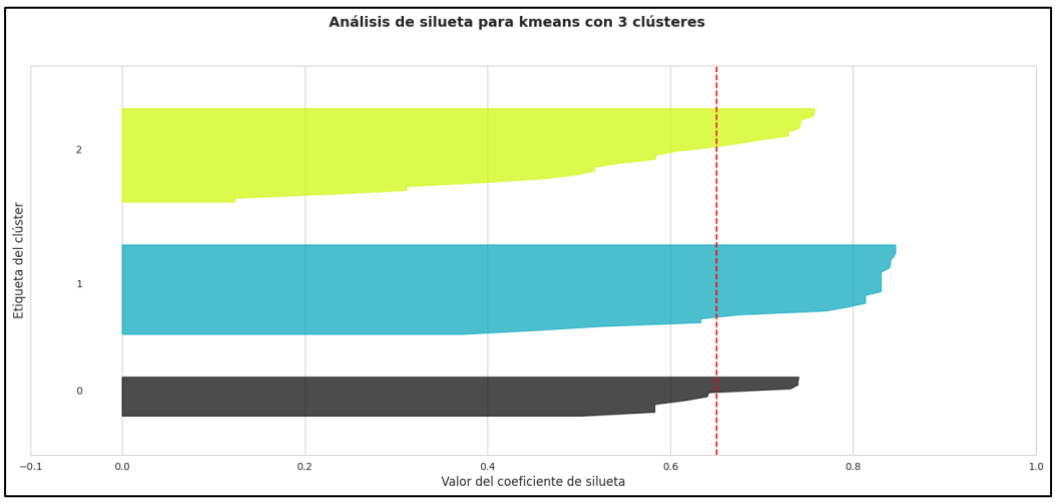

Ilustración 14. Análisis de silueta para kmeans con 3 clústeres
Fuente: Elaboración propia

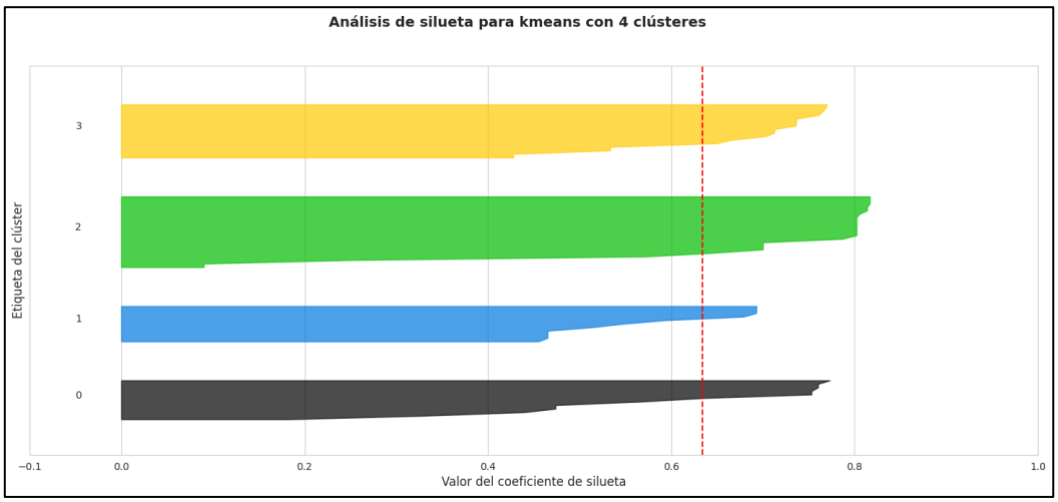

Ilustración 15. Análisis de silueta para kmeans con 4 clústeres
Fuente: Elaboración propia

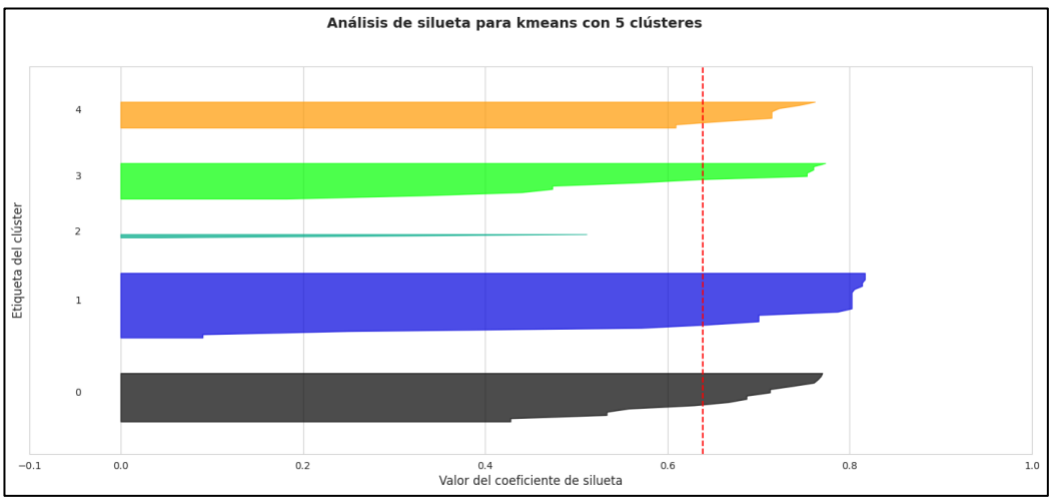

Ilustración 16. Análisis de silueta para kmeans con 5 clústeres
Fuente: Elaboración propia

Los gráficos de silueta generados para diferentes números de clústeres en este modelo de k-means sugieren que, aunque el modelo con dos clústeres tiene el coeficiente de silueta medio más alto, indicando una buena separación de clústeres, existe una gran disparidad en el tamaño de estos clústeres. Al aumentar el número de clústeres a tres y cuatro, la distribución del tamaño de los clústeres se vuelve más equilibrada, aunque con una disminución en el valor medio del coeficiente de silueta, lo cual sugiere que los puntos están menos cohesionados dentro de los clústeres. Con cinco clústeres, la tendencia continúa, y el valor medio del coeficiente de silueta disminuye aún más, lo que puede ser indicativo de una agrupación subóptima.

En conclusión, la configuración de tres clústeres parece la opción más equilibrada, proporcionando una adecuada cohesión dentro de los clústeres y una clara separación entre ellos. Teniendo en cuenta la homogeneidad en las dimensiones de los clústeres y un coeficiente de silueta comparativamente alto, optar por tres clústeres parece ser la alternativa más acertada para este análisis de clústeres.

8.1.3. CLUSTERIZACIÓN CON K=3

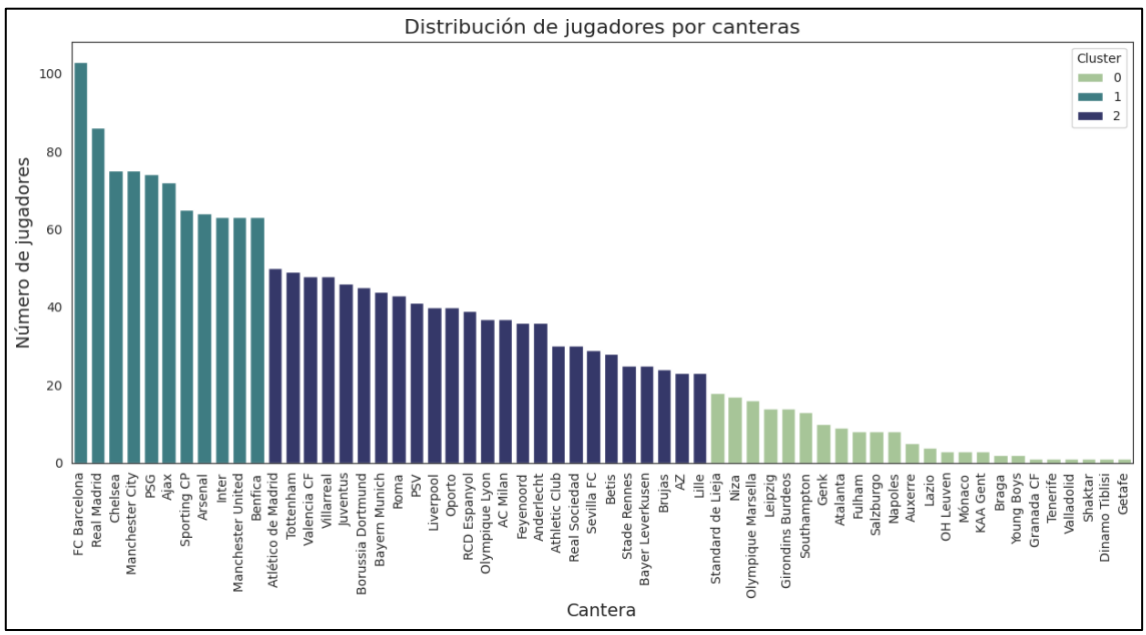

Ilustración 17. Distribución de jugadores por canteras.
Fuente: Elaboración propia

El gráfico muestra un histograma donde se diferencian tres clústeres basados en el número de jugadores provenientes de cada cantera. Observamos una variabilidad significativa en la producción de jugadores, con algunos clubes liderando en números y otros aportando menos. Los clubes del clúster 0 son aquellos con menos jugadores provenientes de sus canteras, lo que indica una selección más exclusiva, mientras que los clubes en el clúster 1 presentan una producción media y aquellos en el clúster 2 la más alta. La distribución sugiere que la mayoría de los clubes tienden a producir una cantidad moderada de jugadores, con unos pocos destacándose en ambos extremos de la escala.

Destaca el FC Barcelona apareciendo en el extremo izquierdo, lo cual indica que su cantera es una de las más eficientes en términos de número de jugadores formados. Esto no solo refleja el énfasis del club en el desarrollo de talento a través de su reconocida academia La Masía, sino que también sugiere un enfoque exitoso en nutrir y promover jugadores hacia el profesionalismo. Otros clubes destacados por su alto número de jugadores formados incluyen al Real Madrid y al Manchester United, ambos conocidos por su compromiso con la formación de jugadores y su impacto en el fútbol de alto nivel.

Por otro lado, clubes como el Valladolid y el Getafe que aparecen hacia el extremo derecho del gráfico en el clúster 2, podrían ser ejemplos de entidades con programas de cantera menos extensos en términos de cantidad, pero que quizás enfaticen la calidad sobre la cantidad o tengan otras estrategias para integrar jugadores en sus equipos principales.

8.1.4. ANÁLISIS DE LOS CENTROIDES

Cada clúster es representado por un centroide, que es calculado como el promedio de los puntos en el clúster. Esta metodología es fundamental en el campo del aprendizaje automático y la minería de datos, especialmente en aplicaciones que requieren la identificación de patrones o tendencias dentro de grandes conjuntos de datos.

- **Clúster 0**: incluye jugadores de alrededor de 25 años en promedio, con 43 partidos en su equipo de debut, con un valor de mercado promedio de aproximadamente 12 millones de euros, con un valor de equipo de alrededor de 310 millones de euros, con una altura media de 182 cm, con un 51.6% de participaciones internacionales, con un debut promedio a los 17 años, con un promedio de transferencias de 4.75 veces, con un Elo promedio de 68.74, con un potencial promedio de 77.57 y con un 89.4% de porcentaje de comunitarios.

- **Clúster 1**: incluye jugadores de alrededor de 25 años en promedio, con 7 partidos en su equipo de debut, con un valor de mercado promedio de aproximadamente 1.6 millones de euros, con un valor de equipo de alrededor de 40 millones de euros, con una altura media de 182 cm, con un 15.7% de participaciones internacionales, con un debut promedio a los 17.8 años, con un promedio de transferencias de 5.12 veces, con un Elo promedio de 58.64, con un potencial promedio de 68.18 y con un 89.8% de porcentaje de comunitarios

- **Clúster 2**: incluye jugadores de alrededor de 25.7 años en promedio, con 63 partidos en su equipo de debut, con un valor de mercado promedio de aproximadamente 36.28 millones de euros, con un valor de equipo de alrededor de 916 millones de euros, con una altura media de 182 cm, con un 74.7% de participaciones internacionales, con un debut promedio a los 17 años, con un promedio

de transferencias de 3.67 veces, con un Elo promedio de 73.62, con un potencial promedio de 82.15 y con un 91.9% de porcentaje de comunitarios.

8.1.5. MAPA DE CALOR DE LOS CENTROIDES

El mapa de calor representa los centroides de los tres clústeres. Se muestra el promedio de jugadores para cada uno de los clústeres. El clúster 1 tiene el valor más alto con un promedio de 73 jugadores, lo que indica que es el más grande en términos de número promedio de jugadores. El clúster 0 tiene un promedio mucho más bajo de jugadores, con un valor de 6.79, lo que sugiere que es el más pequeño o menos poblado de los tres. Finalmente, el clúster 2 tiene un valor intermedio de 36.64. El uso de colores varía de azul a rojo, donde el rojo indica un valor más alto y el azul un valor más bajo.

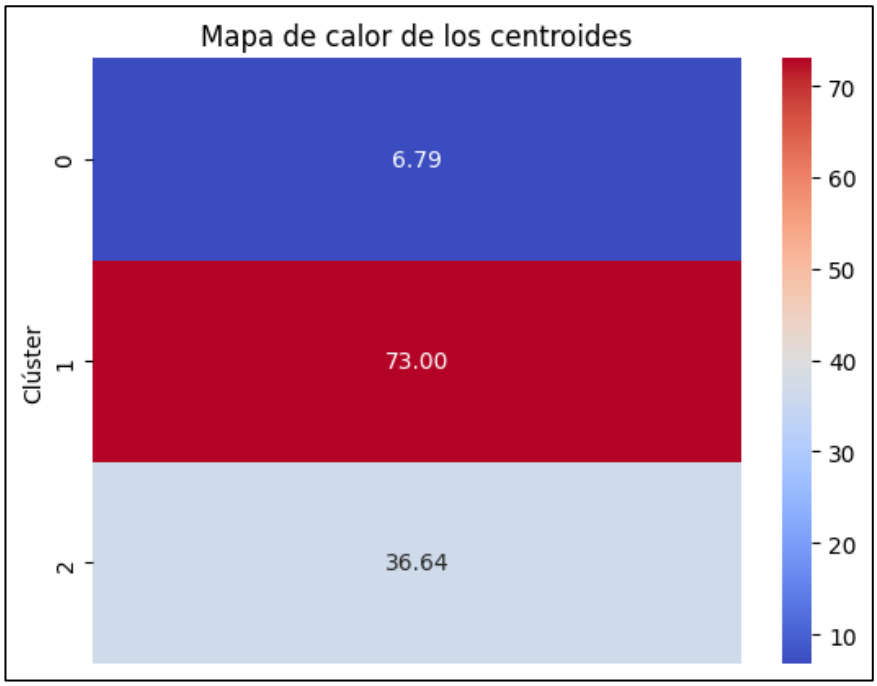

Ilustración 18. Mapa de calor de los centroides
Fuente: Elaboración propia

8.2. ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales es una técnica de reducción de dimensionalidad lineal que puede utilizarse para extraer información de un espacio de alta dimensión proyectándolo en un subespacio de menor dimensión. Intenta preservar las partes esenciales que tienen mayor variación de los datos y eliminar las partes no esenciales con menor variación.

El análisis detallado y la evaluación del valor de mercado de los jugadores de fútbol constituyen una faceta crítica en la gestión deportiva moderna, implicando un desafío considerable debido a la influencia e interdependencia entre diversas variables que caracterizan a los jugadores. En este contexto, el PCA

ofrece una solución elegante y eficiente al permitir la reducción de la dimensionalidad de los datos, facilitando así la identificación de las variables más significativas que influyen en el valor de mercado de los jugadores.

El propósito del algoritmo en este estudio es identificar aquellas variables fundamentales que impactan significativamente en el valor de mercado de los jugadores. El fin es dotar a clubes, representantes y demás actores relevantes en el ámbito del fútbol profesional de un instrumento analítico de primera línea. Mediante este método, se pretende aclarar las relaciones entre las distintas características de los jugadores y su cotización en el mercado, ofreciendo así orientaciones más precisas para la toma de decisiones en cuanto a traspasos, formación de talentos y formulación de estrategias comerciales.

8.2.1. SELECCIÓN DE VARIABLES NUMÉRICAS

Para realizar un análisis detallado, hemos seleccionado cuidadosamente un conjunto de variables numéricas que se consideran clave, excluyendo intencionadamente la variable de interés 'valor_mercado'. Las variables elegidas para este análisis incluyen: 'edad', 'debut', 'valor_equipo', 'altura', 'internacional', 'edad_debut', 'comunitario', 'traspasos', 'Elo', y 'potencial'. La razón detrás de esta selección es identificar aquellos factores que potencialmente influyen en el valor de mercado de los jugadores, sin incorporar el valor de mercado mismo para evitar cualquier influencia directa en el análisis.

Para preparar las variables para un análisis equitativo es crucial estandarizarlas, dado que el PCA es especialmente sensible a las escalas en las que se presentan las variables. Utilizando “StandardScaler” de scikit-learn en Python, se transforman los datos de tal manera que cada variable opere en una escala uniforme, con una media de cero y una desviación estándar de uno. Este proceso de estandarización es fundamental, ya que asegura que ninguna variable domine el análisis debido a diferencias en su magnitud o unidades de medida. Se establece una base nivelada para todas las variables, lo que permite un análisis más equitativo, donde cada una tiene la misma oportunidad de influir en los resultados del estudio.

8.2.2. APLICACIÓN DEL PCA

En este paso, se aplica el PCA con el fin de simplificar la complejidad de nuestros datos. Esta estrategia ayudará a filtrar el ruido y destacar las relaciones subyacentes entre las variables, lo que aclarará interpretaciones y facilitará la toma de decisiones analíticas más informadas. Al transformar los datos estandarizados con PCA, se obtendrán perspectivas más detalladas mediante una representación de datos más manejable y enfocada.

8.2.3. VARIANZA EXPLICADA POR COMPONENTE

Componente	Varianza Explicada(%)	Varianza Acumulada(%)
1	29.725103	29.725103
2	16.468213	46.193316
3	11.037425	57.230741
4	10.160003	67.390744
5	8.279850	75.670594
6	8.167391	83.837985
7	6.498018	90.336003
8	5.052191	95.388194
9	3.561300	98.949493
10	1.050507	100.000000

Tabla 3. Varianza explicada por componente.
Fuente: Elaboración propia

El primer componente explica aproximadamente el 29.72% de la varianza total de los datos. Esto indica que casi un tercio de la información en el conjunto de datos puede ser representada por este único componente. El segundo componente añade otro 16.47% a la varianza explicada, llevando la varianza acumulada al 46.19%. Esto significa que juntos, los dos primeros componentes capturan casi la mitad de toda la variabilidad en los datos.

A medida que se agregan más componentes, la varianza acumulada aumenta, pero cada componente adicional contribuye menos a la varianza explicada que el anterior. Por ejemplo, el tercer componente añade un 11.04% adicional, mientras que el décimo componente solo añade un 1.05%.

8.2.4. VARIANZA EXPLICADA ACUMULADA POR COMPONENTES

El gráfico de la varianza acumulada indica que para representar al menos un 80% de la varianza de los datos debemos elegir 6 componentes principales. Esto se puede observar claramente en el gráfico, donde la línea de varianza acumulada cruza el umbral del 80% justo antes de añadir el séptimo componente. La línea azul vertical que marca el número de componentes necesarios para superar el umbral del 80% de varianza explicada se sitúa en 6 componentes.

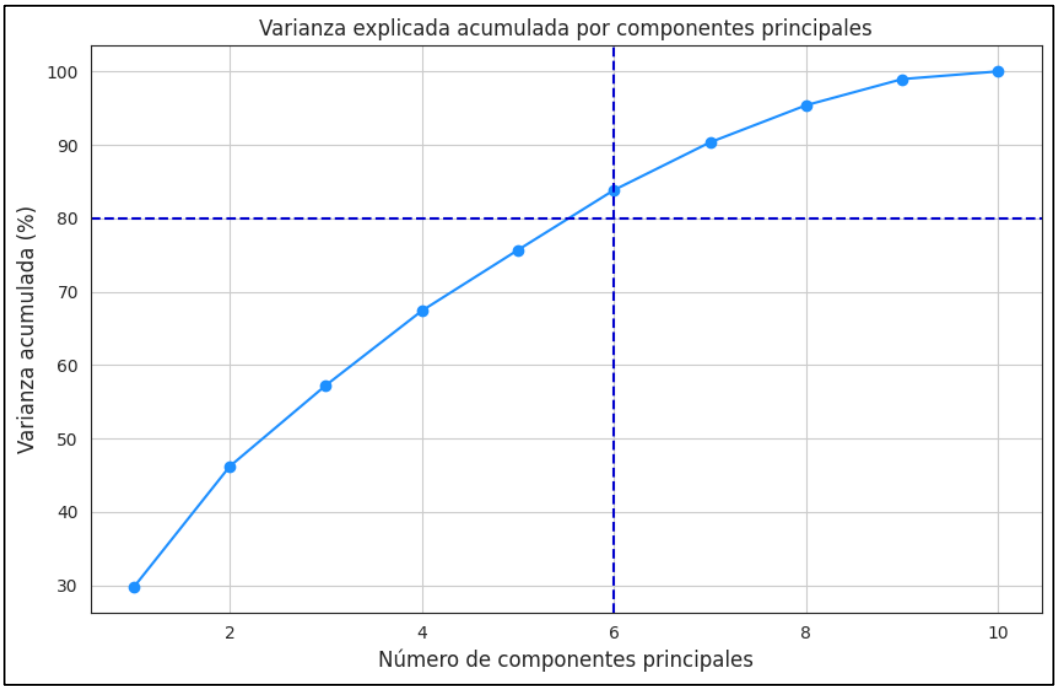

Ilustración 19. Varianza explicada acumulada por componentes principales.
Fuente: Elaboración propia

8.2.5. APLICACIÓN DEL PCA CON 6 COMPONENTES

Se reduce la dimensionalidad de los datos al elegir 6 componentes mientras se conserva la mayor parte de la información original, lo que es útil para simplificar los análisis posteriores y facilitar la visualización de los datos sin perder características esenciales.

8.2.6. ANÁLISIS DE LOS DOS PRIMEROS COMPONENTES PRINCIPALES SEGÚN EL VALOR DE MERCADO

La visualización de las observaciones en el espacio definido por los dos primeros componentes principales muestra una dispersión de puntos donde el color representa el valor de mercado de los jugadores en millones de euros.

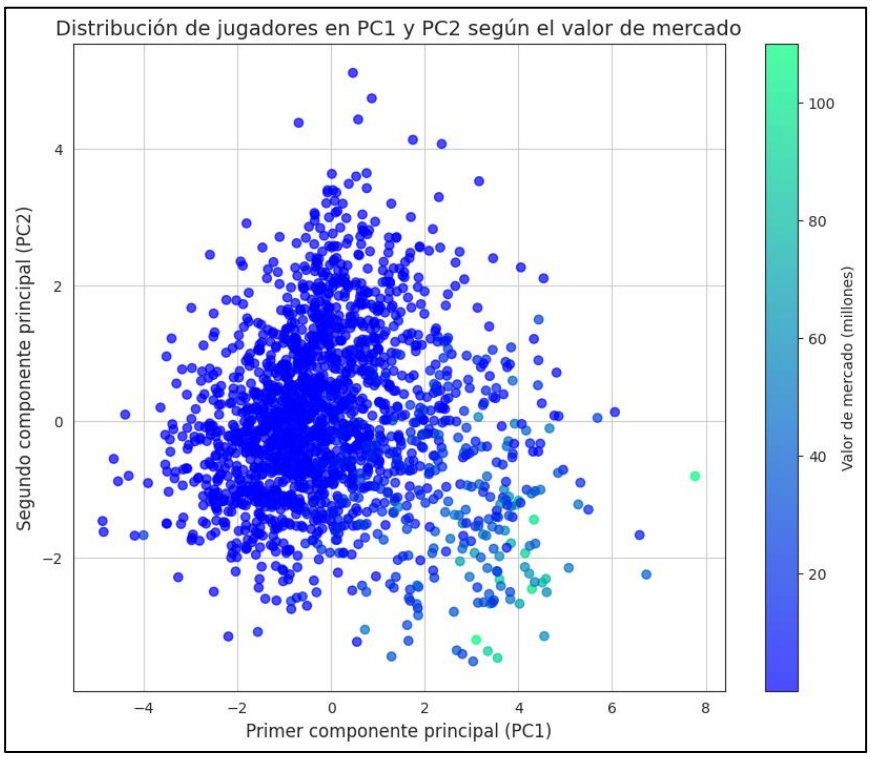

Ilustración 20. Distribución de jugadores en PC1 y PC2 según valor de mercado.
Fuente: Elaboración propia

La mayoría de los puntos se concentran cerca del origen, lo que sugiere que gran parte de los jugadores tienen valores promedio en las características representadas por ambos componentes principales. Sin embargo, hay una dispersión notable hacia la derecha a lo largo del eje PC1, indicando un grupo de jugadores con mayor calidad y potencial. Curiosamente, los jugadores con los valores de mercado más altos no están necesariamente concentrados en el centro del gráfico, sino que tienden a dispersarse hacia valores más altos de PC1.

El valor de mercado parece aumentar a medida que nos movemos hacia la derecha a lo largo de PC1, que podría estar capturando atributos relacionados con la habilidad o potencial que el mercado valora. No hay un patrón tan claro en la dirección de PC2, lo que indica que la experiencia por sí sola no es un factor tan directo en el valor de mercado como lo son las características capturadas por PC1.

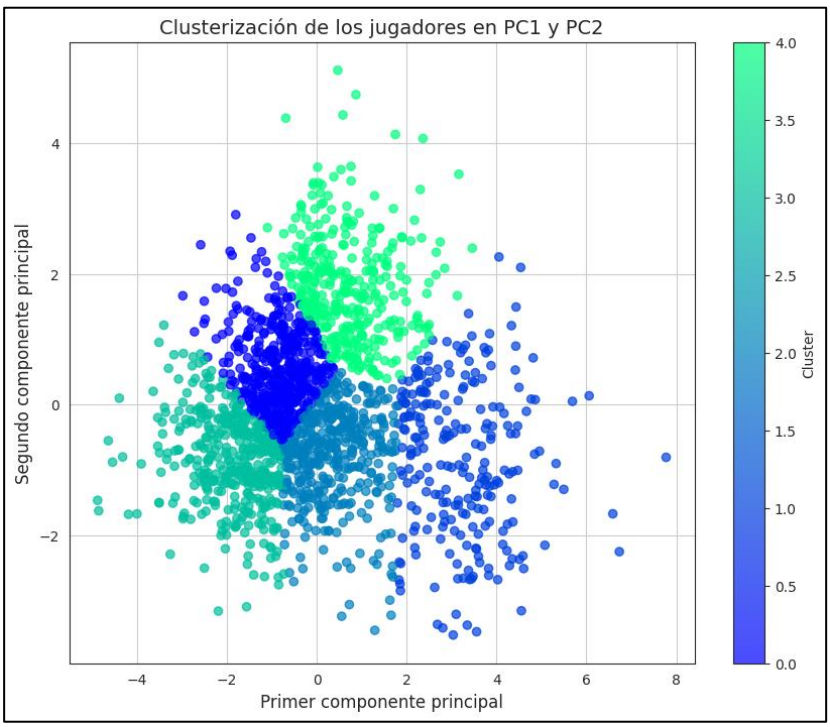

Ilustración 21. Clusterización de los jugadores en PC1 y PC2.
Fuente: Elaboración propia

Una vez realizada la clusterización, se identifican cinco subgrupos identificados por diferentes colores. El clúster más grande, en azul oscuro, se centra alrededor del origen, representando a los jugadores con características medias en las dimensiones capturadas por PC1 y PC2. También se identifican dos clústeres extendidos principalmente a lo largo del eje PC1, pudiendo representar grupos con mayor calidad o potencial. El clúster dispersado a lo largo de PC2 podría diferenciar jugadores por su experiencia o etapas de carrera, mientras que el clúster con puntos dispersos más alejados en ambas dimensiones sugiere un grupo menos definido, pero posiblemente con atributos únicos que los distinguen.

La superposición entre los clústeres indica variabilidad en las características de los jugadores y sugiere que las diferencias entre estos grupos no son completamente discretas, reflejando la complejidad y multidimensionalidad del valor de un jugador de fútbol.

8.2.7. MODELO DE REGRESIÓN LINEAL MÚLTIPLE

En este estudio, el PCA va acompañado de un modelo de regresión lineal múltiple. Se trata de una técnica estadística y de aprendizaje automático que se emplea para modelar la relación entre una variable dependiente y varias variables independientes. Este modelo presupone una relación lineal entre dichas variables, haciéndolo idóneo tanto para la predicción como para la inferencia. El objetivo principal de aplicar la regresión lineal es desarrollar un modelo predictivo capaz de estimar el valor de mercado de

un jugador de fútbol basándose en aspectos cuantificables y observables, tales como características físicas y futbolísticas del jugador, entre otras.

Definición de variables

Variable dependiente (Y): el valor de mercado del jugador, expresado en millones de euros; representa la variable que el modelo intenta predecir.

Variables independientes (X): selección de diversas variables relevantes en el ámbito futbolístico, como la edad del jugador, la altura, el potencial, la cantidad de traspasos realizados, las participaciones en partidos internacionales, entre otras.

Evaluación del modelo

El MSE (Error Cuadrático Medio) es una métrica utilizada para medir la calidad de un estimador o modelo. Esta métrica evalúa el promedio de los cuadrados de los errores, es decir, la diferencia cuadrática entre los valores observados y los predichos por el modelo. El MSE obtenido en este modelo es de 55.36, lo cual indica variabilidad en las predicciones del modelo. Esto puede deberse a que los valores de mercado varían significativamente desde unos pocos millones hasta cientos de millones

El R^2 (Coeficiente de Determinación) es otra métrica usada para evaluar la bondad de ajuste de un modelo de regresión lineal. Expresa la proporción de la variabilidad de los datos que es explicada por el modelo. El valor de R^2 obtenido es de 0.52, indicando que aproximadamente el 52% de la variabilidad en el valor de mercado de los jugadores se explica por las variables independientes seleccionadas. Aunque este valor señala una correlación moderada, también implica que existen otros factores no considerados en el modelo que influyen en el valor de mercado. En el contexto económico, un R^2 de alrededor del 50% se considerarse aceptable, ya que los fenómenos económicos suelen estar influenciados por numerosos factores, muchos de los cuales pueden ser difíciles de cuantificar o incluso se desconocen.

8.2.8. INTERPRETACIÓN DE COEFICIENTES EN MODELO DE REGRESIÓN

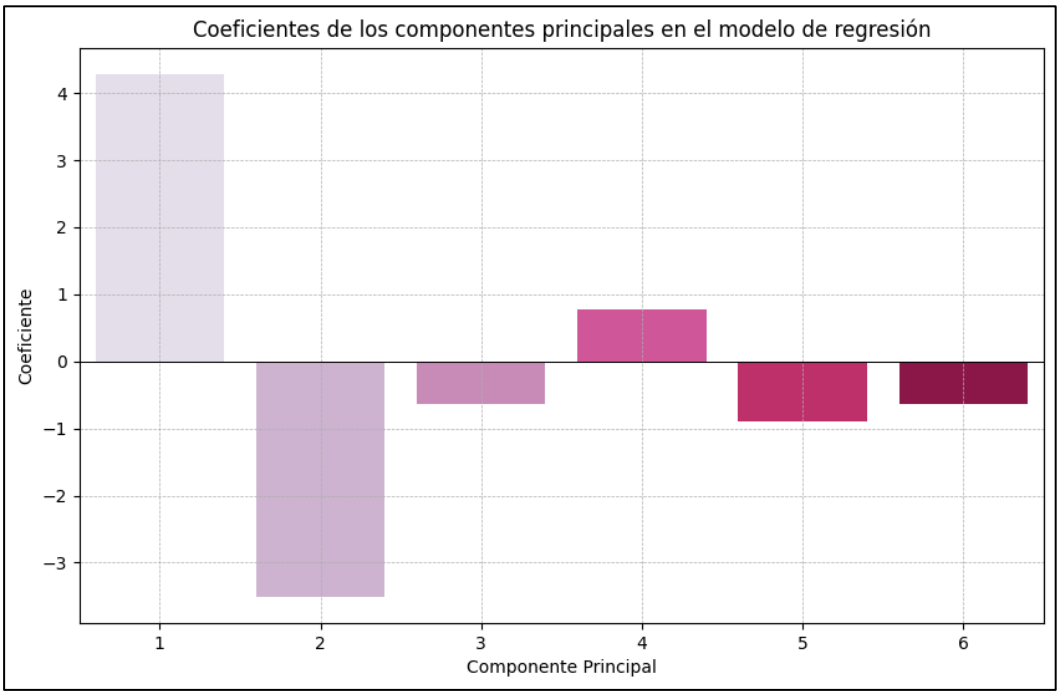

Ilustración 22. Coeficientes de los componentes principales en el modelo de regresión.
Fuente: Elaboración propia

La importancia de los componentes principales en la predicción del valor de mercado de los jugadores se revela a través de los coeficientes obtenidos en el modelo de regresión lineal. Cada componente principal se convierte en una variable independiente en el modelo y su coeficiente indica el grado de impacto que tiene sobre la variable dependiente, en este caso, el valor de mercado del jugador. Un coeficiente alto en valor absoluto sugiere que cambios en ese componente tienen un efecto significativo en el valor de mercado, ya sea incrementándolo (coeficientes positivos) o disminuyéndolo (coeficientes negativos).

8.2.9. REGRESIÓN LINEAL EN EL PRIMER COMPONENTE

La representación gráfica del primer componente en este análisis es una decisión basada en la preponderancia de su coeficiente en el modelo de regresión, como se evidencia en el gráfico anterior. Este primer componente demuestra tener el mayor impacto en la variable dependiente, ya que captura las características más influyentes que determinan el valor de mercado. Por lo tanto, destaca como el factor individual más crítico en el modelo, haciéndolo un candidato ideal para una visualización focalizada.

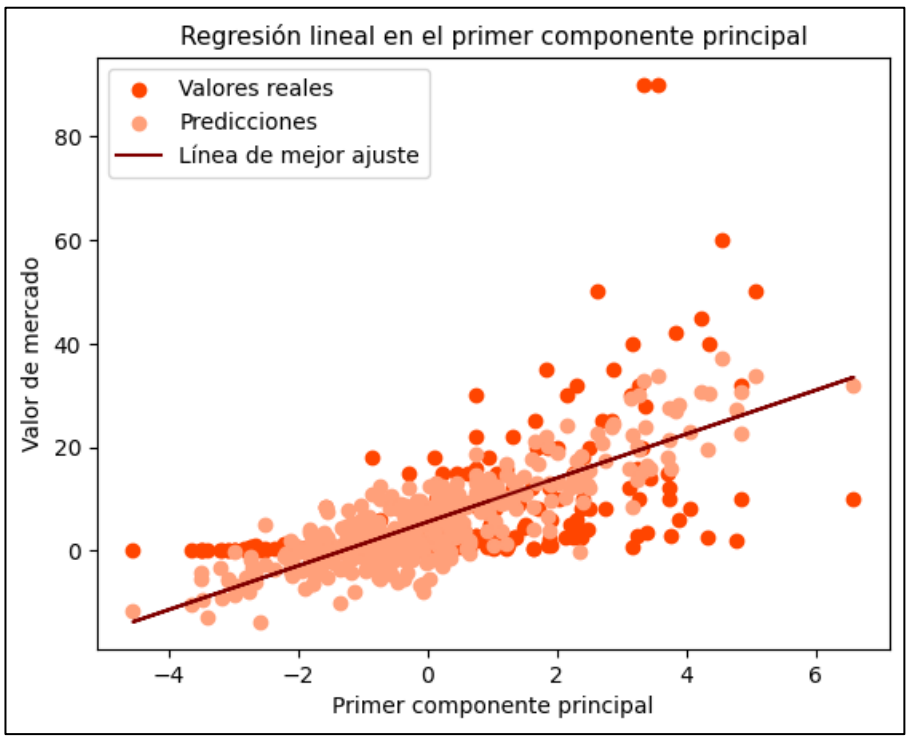

Ilustración 23. Regresión lineal en el primer componente principal.
Fuente: Elaboración propia

En la visualización, los puntos rojos representan los valores reales del valor de mercado, mientras que los puntos de color salmón indican las predicciones del modelo de regresión. La línea de mejor ajuste, trazada a través de los datos, ofrece una representación visual del modelo de regresión lineal aplicado. Esta línea de color marrón contrasta con los puntos y sirve para ilustrar la dirección y la fuerza de la relación entre el componente principal y la variable de mercado, proporcionando una interpretación directa de la tendencia general y la calidad del ajuste del modelo.

8.2.10. REPRESENTACIÓN DE LAS CARGAS DE LAS VARIABLES

Los valores de la siguiente tabla representan las cargas de las variables originales en cada componente principal, que reflejan cómo cada variable contribuye a la componente. Estas cargas también se han representado en forma de matriz, asignando colores oscuros cuando la carga es negativa y colores claros cuando la carga es positiva.

	PC1	PC2	PC3	PC4	PC5	PC6
edad	0.276625	0.577010	0.075844	-0.020149	0.129009	-0.005768
debut	0.300412	-0.090703	-0.218046	0.115069	0.745737	0.316391
valor_equipo	0.377873	-0.329794	-0.071607	0.129492	0.022242	-0.013329
altura	-0.002250	0.079438	0.159045	0.937802	-0.052804	-0.259698
internacional	0.399383	-0.077862	0.011590	0.075939	0.133436	0.036485
edad_debut	-0.022512	0.485909	-0.448776	-0.098421	0.215965	-0.567504
comunitario	0.029479	0.089839	-0.766376	0.211100	-0.423319	0.410044
traspasos	0.110563	0.539614	0.352399	0.036681	-0.122566	0.521307
ELO	0.498185	-0.005974	0.040975	-0.142582	-0.312138	-0.211664
potencial	0.518790	-0.047930	0.039369	-0.091764	-0.260485	-0.155408

Tabla 4. Representación de las cargas de las variables.
Fuente: Elaboración propia

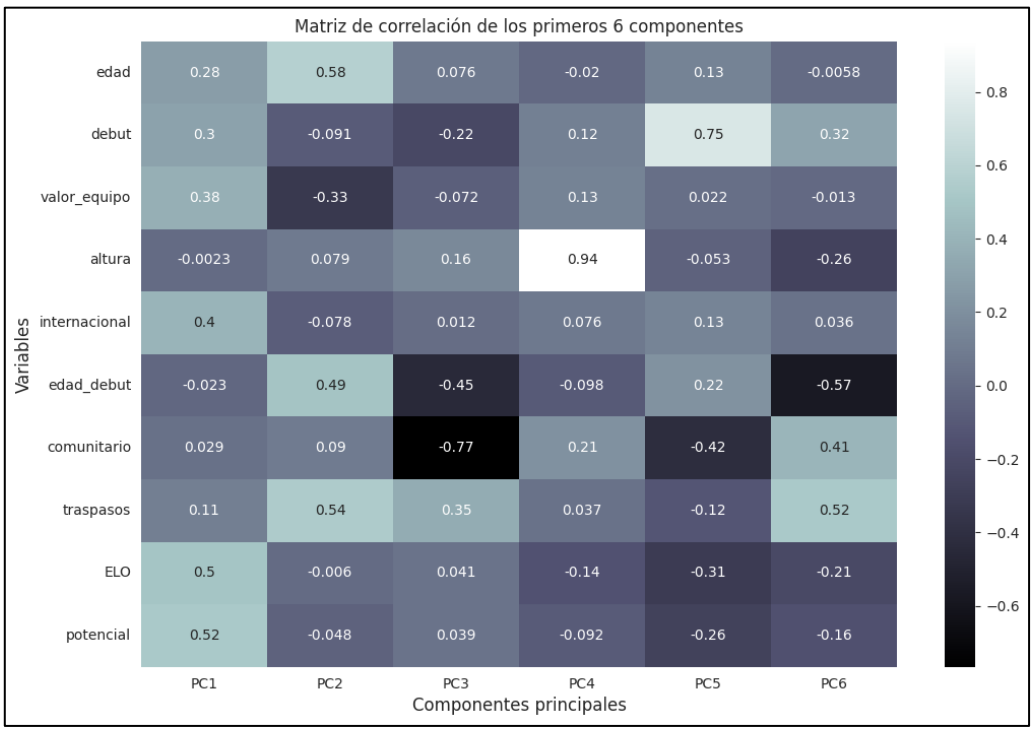

Ilustración 24. Matriz de correlación de los primeros 6 componentes.
Fuente: Elaboración propia

La relación entre los coeficientes de los componentes en el modelo de regresión y las cargas de los componentes principales subraya la complejidad del impacto que tienen las características de los jugadores en su valor de mercado. No se trata solo de la influencia individual de cada característica sino de cómo las interacciones entre diferentes características, capturadas a través de los componentes principales, contribuyen conjuntamente a determinar el valor de mercado.

En este análisis, 'Elo' y 'potencial' emergen como predictores claves del valor de mercado debido a sus altas cargas positivas en el PC1 y el significativo coeficiente positivo de este componente en el modelo de regresión. Esto indica que altos valores en 'Elo' y 'potencial' están asociados con aumentos en el valor de mercado, subrayando la importancia de estas características en la percepción del valor de un jugador.

Este hallazgo es coherente con la intuición de que jugadores con mejor desempeño y mayor potencial son más valorados en el mercado.

Asimismo, la variable 'altura' sobresale en el PC4, el otro componente con coeficiente positivo. Por lo tanto, los jugadores con una estatura notable están siendo cada vez más valorados en el mercado. Esto refleja una creciente tendencia en la que los clubes valoran no solo el talento técnico, sino también las características físicas de los jugadores al momento de realizar inversiones en el mercado de fichajes.

Por otro lado, la 'edad' y los 'traspasos', con su contribución a PC2 y el coeficiente negativo asociado a este componente, sugieren una relación inversa con el valor de mercado. Esto podría interpretarse como que un mayor número de 'traspasos' o una mayor 'edad' podrían percibirse como indicativos de una menor proyección futura o de una carrera más cercana a su declive, lo cual afectaría negativamente el valor de mercado del jugador.

La interpretación combinada de las cargas de los componentes y los coeficientes de regresión ofrece una visión integral de qué características y combinaciones de estas ejercen la mayor influencia sobre el valor de mercado, proporcionando información de valor para clubes, agentes y federaciones del ámbito futbolístico.

9. MODELO DE NEGOCIO: TALENTSCOPE

El análisis de datos realizado previamente evidencia la capacidad de los algoritmos para proponer soluciones innovadoras y exitosas en el ámbito del fútbol. Por ello, este modelo de negocio presenta la aplicación de los algoritmos en el scouting de jugadores a través de una plataforma revolucionaria llamada Talentscope. El objetivo principal de esta plataforma será transformar radicalmente las metodologías tradicionales de scouting implementando tecnologías emergentes. A continuación, se detalla el diseño y funcionamiento de este innovador modelo de negocio.

¿Qué será TalentScope?

TalentScope será una plataforma diseñada para evaluar y predecir el potencial de jugadores de fútbol mediante el uso de algoritmos e inteligencia artificial. Estará diseñada para analizar un registro completo de datos, incluyendo métricas físicas, técnicas, tácticas, psicológicas y de impacto en el juego. Los algoritmos de aprendizaje profundo permitirán identificar patrones y correlaciones que podrían no ser detectados por el ojo humano o los sistemas de evaluación convencionales. Esto habilitará a los entrenadores, reclutadores y clubes deportivos para tomar decisiones más informadas sobre el desarrollo y la gestión de talentos, así como identificar jugadores prometedores para el éxito futuro.

TalentScope destacará por ofrecer una interfaz de usuario sencilla y adaptable que facilitará el acceso a análisis completos y detallados del potencial de los jugadores. La plataforma brindará herramientas avanzadas para el análisis y la visualización de datos, permitiendo una exploración profunda del rendimiento de los jugadores. También, se adaptará a las preferencias individuales de cada equipo al permitir la modificación de los parámetros de evaluación, lo que la hará compatible con diferentes filosofías deportivas y tácticas de juego. De esta manera, garantizará que la información proporcionada no solo sea específica a las necesidades del usuario, sino que también pueda revelar posibles talentos no identificados a primera vista.

Imaginemos que un club de fútbol, que juega de manera muy directa y utiliza frecuentemente centros al área, está buscando reforzar su delantera con un jugador que se ajuste a las necesidades tácticas del equipo. Utilizando TalentScope, el club podrá ingresar una serie de criterios de búsqueda personalizados para encontrar al jugador ideal. Por ejemplo, el club podría especificar a la plataforma que está buscando un delantero alto y rápido, con gran capacidad de pases y que destaque por su remate de cabeza.

Una vez que se hayan ingresado estos criterios, la plataforma realizará una búsqueda exhaustiva en su base de datos y seleccionará a los jugadores que mejor se ajusten a estos requisitos. Sin embargo, TalentScope no se limitará a proporcionar información sobre el rendimiento pasado de los jugadores, sino que también utilizará algoritmos de aprendizaje profundo para predecir su potencial futuro. De esta manera, no solo identificará a los jugadores que mejor se ajustan a las necesidades del club en el presente, sino que también ayudará a predecir qué jugadores tendrán un rendimiento destacado en el futuro.

Posteriormente, el club tendrá a su disposición una lista de jugadores potenciales, sobre los cuales podrá realizar un análisis en detalle del rendimiento pasado y de su futuro potencial. Esto permitirá al club tomar una decisión informada sobre quién sería el mejor fichaje para el equipo, no solo en el presente, sino también a largo plazo.

9.1. PROPUESTA DE VALOR

El mercado donde operará TalentScope se caracteriza por una demanda creciente de tecnologías avanzadas para la identificación y el desarrollo de talentos en el fútbol. Los clubes buscan constantemente ventajas competitivas para identificar e incorporar jugadores prometedores antes que sus competidores. La integración de IA y análisis predictivo en la evaluación de jugadores representa una tendencia emergente dirigida a optimizar el rendimiento y la inversión en talento. TalentScope se distinguirá de otras plataformas del sector por su enfoque único en la predicción del potencial de desarrollo a largo plazo, ofreciendo un análisis más profundo y personalizado que las evaluaciones tradicionales.

9.2. IDENTIFICACIÓN DE CLIENTES POTENCIALES Y TEST DE CONCEPTO

Estas serán las diferentes entidades dentro del ámbito del fútbol que podrán beneficiarse de las innovaciones de nuestra plataforma, abriendo nuevas puertas hacia el éxito y la sostenibilidad en el deporte más popular en Europa:

- **Clubes de fútbol profesional:** desde clubes de élite que compiten en ligas mayores hasta equipos en divisiones inferiores buscando una ventaja competitiva para ascender.
- **Agencias de representación:** interesadas en identificar y representar a los futuros talentos del fútbol.
- **Federaciones deportivas:** organizaciones a nivel nacional o regional interesadas en fortalecer sus selecciones y ligas locales.

Las entrevistas con clubes de fútbol profesional, agencias de representación y federaciones deportivas ofrecerán una excelente oportunidad para recoger esta información directamente de las fuentes más relevantes. Al entender las perspectivas únicas de cada grupo de clientes, se podrá adaptar la tecnología y los servicios para ofrecer soluciones que no solo cumplan, sino que superen sus expectativas.

9.2.1. OBJETIVOS DE LAS ENTREVISTAS

- Comprender las necesidades individuales de los clientes en el contexto del scouting de talentos y la gestión de jugadores, con el fin de adaptar nuestra plataforma a sus necesidades específicas.
- Investigar las expectativas de los clientes respecto a una solución basada en inteligencia artificial, poniendo el foco en aspectos de precisión, facilidad de uso y resultados esperados.
- Recoger impresiones sobre la plataforma, evaluando la recepción inicial y el posible interés de contratación.
- Escuchar propuestas de personalización que permitan a TalentScope alinearse mejor con las estrategias deportivas y tácticas específicas de cada cliente.

9.2.2. POSIBLES PREGUNTAS

- ¿Qué herramientas emplean actualmente en el proceso de scouting y evaluación de jugadores? ¿Existen necesidades o desafíos específicos en su proceso que no estén siendo completamente satisfechos?
- ¿Qué características le gustaría ver en una plataforma como TalentScope para considerarla una inversión valiosa para su organización?
- Al evaluar el potencial de jugadores emergentes o menos conocidos, ¿qué tipo de información o datos consideran indispensables para tomar decisiones informadas?
- En relación con TalentScope, ¿qué dudas o cuestiones se le plantean sobre la plataforma? ¿cuáles son sus impresiones iniciales? ¿Qué aspectos consideran positivos o negativos?

- ¿Se plantean la incorporación de TalentScope en sus procesos de evaluación de talentos? ¿De qué presupuesto disponen para innovación y tecnología?

9.3. DESARROLLO DEL PROTOTIPO

La validación de TalentScope se iniciará con el desarrollo de un conjunto de productos mínimos viables que incorporen las características esenciales de la plataforma. A través de pruebas piloto con clubes de fútbol, agencias de representación y federaciones deportivas, se recogerán datos cruciales sobre la usabilidad, la precisión de los análisis predictivos y el valor real que estos servicios aportan a la gestión de talentos. Este enfoque interactivo permitirá realizar ajustes en función del feedback, asegurando que el producto final responda efectivamente a las necesidades y expectativas del mercado objetivo.

9.4. ESTUDIO DE PATENTABILIDAD

Realizar un estudio de patentabilidad para este modelo de negocio será crucial para proteger la innovación tecnológica y el valor intelectual detrás de sus algoritmos de aprendizaje profundo y análisis predictivo. Este proceso implicará investigar bases de datos de patentes existentes para asegurarse de que las innovaciones son únicas y no infringen derechos de terceros. La obtención de patentes no solo protegerá contra la competencia directa, sino que también aumentará el valor de la empresa ante inversores y socios estratégicos.

9.5. BÚSQUEDA DE FINANCIACIÓN

La búsqueda de financiación se enfocará en inversores que tengan interés específico en tecnología deportiva, inteligencia artificial y big data. Presentaciones a fondos de capital de riesgo, inversores ángeles especializados en tecnología y deportes, y programas de subvenciones para innovación podrán ser caminos viables. Será vital preparar un pitch convincente que destaque el potencial de mercado, la innovación tecnológica y el equipo detrás del proyecto, demostrando no solo la viabilidad económica sino también el impacto transformador en la industria del fútbol.

9.6. FUENTES DE INGRESO

Para TalentScope, el modelo de ingresos se centrará en suscripciones básicas y premium que ofrecerán desde análisis básicos hasta servicios personalizados y consultoría. Se venderán licencias de software a clubes, federaciones y agencias complementadas con talleres y capacitaciones en scouting y análisis de rendimiento. Los paquetes personalizados permitirán una mayor flexibilidad y adaptación a las necesidades específicas de los clientes.

9.7. ESTRATEGIA DE PRECIOS Y TEST CUANTITATIVO

Tras realizar un análisis detallado de posibles competidores como Olocip o Driblab se ha determinado la siguiente estructura de precios:

- **Suscripción anual básica:** se estima que TalentScope podrá establecer una suscripción anual de 65.000 euros para clubes, 41.500 euros para federaciones y de 26.500 euros para agencias de representación.

- **Suscripción anual premium:** Para aquellos que desearan acceder a registros exclusivos y servicios personalizados, TalentScope ofrecerá una suscripción premium. Los precios para esta categoría serán de 80.000 euros para clubes, 52.000 euros para federaciones y 35.000 euros para agencias de representación, subrayando la calidad de la tecnología de IA y los servicios añadidos como talleres y formaciones.

Para estimar la demanda de TalentScope según los precios establecidos, se realizará un estudio de mercado detallado enviando cuestionarios online a federaciones, agencias y clubes de fútbol de alto poder económico que puedan permitirse este tipo de inversión. El enfoque a los clubes estará principalmente en las cinco principales ligas europeas (España, Francia, Inglaterra, Italia y Alemania), aunque también se tomarán en consideración equipos concretos de Holanda, Bélgica, Portugal y algunas segundas divisiones importantes.

Estos cuestionarios incluirán preguntas sobre el tipo de suscripción en que podrían estar interesados los posibles clientes, las necesidades actuales no satisfechas en el proceso de scouting, la importancia de la personalización y adaptabilidad de la plataforma y la disposición a sustituir las herramientas actuales por TalentScope. Para aumentar la tasa de respuesta, se considerará ofrecer incentivos, como accesos a informes exclusivos de la industria o pruebas gratuitas de la plataforma. Los datos recogidos permitirán segmentar el mercado por tipo de cliente y calcular la demanda potencial basándose en el número de clientes en cada segmento, su presupuesto para scouting y tecnología y su interés expresado en las innovaciones que TalentScope propone. Esta información será fundamental para hacer una previsión de ingresos realista y ajustar estrategias de marketing y ventas para maximizar la penetración en el mercado.

9.8. PLAN DE LANZAMIENTO

Una vez se haya estimado la demanda, será importante diseñar cuidadosamente un plan de lanzamiento para generar expectativa y captar la atención de los principales stakeholders en la industria del fútbol. Esto podría incluir una combinación de marketing de contenidos, demostraciones en eventos deportivos clave, pruebas gratuitas para clubes selectos y colaboraciones con influencers y profesionales

reconocidos en el ámbito deportivo. Además, será esencial establecer una fuerte presencia online a través de redes sociales y una página web optimizada para los motores de búsqueda, donde se muestren casos de éxito y testimonios detallados.

9.9. FEEDBACK CONTINUO

El proceso de recopilación de feedback deberá ser continuo y estructurado, utilizando encuestas online para entender experiencias, sugerencias y críticas de los clientes. Este feedback será imprescindible para mejorar la plataforma, adaptándola a las necesidades y desafíos del mercado futbolístico. Además, fomentar una comunidad de usuarios activa a través de foros o grupos de discusión podría proporcionar insights constantes y fomentar la innovación colaborativa.

9.10. POSIBLES RIESGOS

Los posibles riesgos a los que se enfrentará TalentScope serán desafíos tecnológicos, como la constante necesidad de actualización ante los avances en inteligencia artificial, cuestiones de seguridad de datos, que podrían llevar a violaciones de la privacidad, y la aceptación en el mercado, marcada especialmente por el escepticismo de algunos scouts y clubes reacios a sustituir métodos convencionales de evaluación por soluciones basadas en IA. La complejidad de estos riesgos se magnifica por la presión de mantener tanto la precisión predictiva como la personalización en un entorno deportivo diverso y exigente.

Para contrarrestar estos desafíos, se adoptarán políticas robustas de protección de datos acordes con las regulaciones más estrictas y un enfoque proactivo en la demostración del valor tangible que TalentScope aportará al scouting. Esto, combinado con una comunicación efectiva que destaque los casos de éxito y la integración de la herramienta en las prácticas de scouting existentes, ayudará a superar el escepticismo y fomentar una adopción más amplia.

9.11. LINEAS FUTURAS DE INVESTIGACIÓN

La aplicación de la inteligencia artificial y el análisis predictivo, como los implementados en TalentScope, tiene un gran potencial más allá del scouting de jugadores. Estas tecnologías son capaces de transformar otras áreas del fútbol y del deporte, ofreciendo soluciones innovadoras a problemas comunes y abriendo nuevas oportunidades de negocio, como podrían ser la venta de entradas y la prevención de lesiones.

En el contexto de la venta de entradas, la aplicación de algoritmos avanzados permitiría analizar detalladamente los patrones de compra y las preferencias de los aficionados, posibilitando una estrategia de precios y promociones altamente eficiente. Este enfoque no solo buscaría maximizar ingresos y asistencia, sino que también se orientaría hacia la mejora de la experiencia del aficionado, ajustando

precios y ofertas a sus expectativas y posibilidades económicas. La clave radicaría en la integración de variables como el rendimiento del equipo y la relevancia del encuentro, permitiendo así ajustes dinámicos que reflejen la demanda real y anticipada.

La prevención de lesiones representaría otro campo fértil para el análisis predictivo, el cual, mediante el análisis de datos de entrenamientos y partidos, podría identificar patrones que indiquen un riesgo incrementado de lesiones. La personalización de los programas de entrenamiento, tomando en cuenta el historial de lesiones del jugador, su carga de trabajo actual y sus datos biométricos, contribuiría significativamente a minimizar los riesgos de lesión. Este enfoque tendría el potencial de mejorar el rendimiento del equipo al mantener a los jugadores clave en forma a la vez que sería capaz de reducir significativamente los costos asociados a la rehabilitación y al tiempo de inactividad de los jugadores lesionados.

10. CONCLUSIONES

A lo largo de este trabajo se ha evidenciado cómo la aplicación de algoritmos de aprendizaje automático en el ámbito del fútbol proporciona ventajas competitivas significativas en todos los niveles. La implementación de herramientas como TalentScope no solo optimiza el proceso de scouting y selección de talentos, sino que también proporciona insights valiosos que pueden marcar la diferencia en el rendimiento deportivo de un equipo.

La contratación de analistas de datos se presenta como una necesidad para los clubes que buscan mantenerse competitivos en un entorno cada vez más exigente. En un mercado donde la capacidad de identificar y desarrollar talento de manera eficiente marca la diferencia entre el éxito y el fracaso, aquellos clubes que adopten estas tecnologías de manera temprana tendrán una clara ventaja sobre aquellos que no lo hagan. La capacidad de analizar grandes cantidades de datos de manera rápida y precisa, y de obtener insights valiosos a partir de ellos, permitirá a los clubes tomar decisiones más informadas y estratégicas en todas sus áreas, incluyendo el reclutamiento de talentos, la optimización de la venta de entradas, la prevención de lesiones, la toma de decisiones financieras y el estilo de juego del equipo, entre otras muchas.

Además, el modelo de negocio TalentScope no solo representa una herramienta poderosa para mejorar el rendimiento deportivo de un equipo, sino que también es un activo estratégico que puede marcar la diferencia en el éxito a largo plazo de un club. Al centrarse en las necesidades específicas de los clubes de fútbol, agencias de representación y federaciones deportivas, la plataforma se posiciona como una solución integral que cumple y supera las expectativas del cliente. Su flexibilidad y adaptabilidad, junto

con su capacidad para ofrecer análisis personalizados y consultoría especializada, la convierten en una herramienta indispensable para cualquier entidad involucrada en la gestión de talentos en el fútbol.

En resumen, este trabajo ha demostrado que el uso de algoritmos es fundamental para el futuro del fútbol y del deporte en general. TalentScope representa una innovación significativa en este sentido, ofreciendo a los clubes deportivos una herramienta poderosa y efectiva para mejorar su rendimiento deportivo y alcanzar el éxito en el campo. Aquellos que se aprovechen de estas tecnologías de manera temprana tendrán una clara ventaja competitiva en el mercado y estarán mejor posicionados para alcanzar sus objetivos deportivos y financieros.

11. REFERENCIAS BIBLIOGRÁFICAS

- Akhanli, S., & Hennig, C. (2022, 20 abril). Clustering of football players based on performance data and aggregated clustering validity indexes. arXiv.org. https://arxiv.org/abs/2204.09793

- Bauer, P., & Anzer, G. (2021). Data-driven detection of counterpressing in professional football. Data Mining And Knowledge Discovery, 35(5), 2009-2049. https://doi.org/10.1007/s10618-021-00763-7

- Bergkamp, T., Niessen, A. S. M., Hartigh, R. D., Frencken, W., & Meijer, R. R. (2019). Methodological Issues in Soccer Talent Identification Research. Sports Medicine, 49(9), 1317-1335. https://doi.org/10.1007/s40279-019-01113-w

- Figueiredo, A. J., Coelho-e-Silva, M. J., & Malina, R. M. (2011). Predictors of functional capacity and skill in youth soccer players. Scandinavian Journal Of Medicine & Science In Sports, 21(3), 446-454. https://doi.org/10.1111/j.1600-0838.2009.01056.x

- Fujii, K. (2021). Data-Driven Analysis for Understanding Team Sports Behaviors. Journal Of Robotics And Mechatronics, 33(3), 505-514. https://doi.org/10.20965/jrm.2021.p0505

- Gil, S. M., Bidaurrazaga-Letona, I., Martin-Garetxana, I., Lekue, J. A., & Larruskain, J. (2019). Does birth date influence career attainment in professional soccer? Science And Medicine In Football, 4(2), 119-126. https://doi.org/10.1080/24733938.2019.1696471

- Hao, H., & Al-Barakati, A. (2021). Sports intensity and energy consumption based on fractional linear regression equation. Applied Mathematics And Nonlinear Sciences, 7(2), 115-126. https://doi.org/10.2478/amns.2021.2.00137

- Kong, L., Zhang, T., Zhou, C., Gómez, M., Hu, Y., & Zhang, S. (2022). The evaluation of playing styles integrating with contextual variables in professional soccer. Frontiers In Psychology, 13. https://doi.org/10.3389/fpsyg.2022.1002566

- Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zając, A., & Stanula, A. (2014). Application of Neural and Regression Models in Sports Results Prediction. Procedia-Social And Behavioral Sciences, 117, 482-487. https://doi.org/10.1016/j.sbspro.2014.02.249

- McCormack, S., Jones, B., Scantlebury, S., Collins, N., Owen, C., & Till, K. (2021). Using Principal Component Analysis to Compare the Physical Qualities Between Academy and International Youth Rugby League Players. International Journal Of Sports Physiology And Performance, 16(12), 1880-1887. https://doi.org/10.1123/ijspp.2021-0049

- Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal Of Operational Research, 263(2), 611-624. https://doi.org/10.1016/j.ejor.2017.05.005

- Plakias, S., Moustakidis, S., Kokkotis, C., Papalexi, M., Tsatalas, T., Giakas, G., & Tsaopoulos, D. (2023). Identifying soccer Players' playing Styles: A Systematic review. Journal Of Functional Morphology And Kinesiology, 8(3), 104. https://doi.org/10.3390/jfmk8030104

- Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F. M., & Baca, A. (2023). Machine learning application in soccer: a systematic review. Biology Of Sport, 40(1), 249-263. https://doi.org/10.5114/biolsport.2023.112970

- Schlenger, J., Wunderlich, F., Raabe, D., & Memmert, D. (2023). Systematic Analysis of Position-Data-based Key Performance Indicators. International Journal Of Computer Science In Sport, 22(1), 80-101. https://doi.org/10.2478/ijcss-2023-0006

- Sebzalli, Y., & Wang, X. (2001). Knowledge discovery from process operational data using PCA and fuzzy clustering. Engineering Applications Of Artificial Intelligence, 14(5), 607-616. https://doi.org/10.1016/s0952-1976(01)00032-x

- Vidal-Codina, F., Evans, N., Fakir, B. E., & Billingham, J. (2022, 1 febrero). Automatic event detection in football using tracking data. arXiv.org. https://arxiv.org/abs/2202.00804

- Wold, S., Esbensen, K. and Geladi, P. (1987) Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems, 2, 37-52. http://dx.doi.org/10.1016/0169-7439(87)80084-9

12. ANEXOS

12.1. VARIABLES DE LA BASE DE DATOS

nombre	equipo	liga	posicion	fecha_nacimiento	edad	cantera	debut	valor_mercado	nacionalidad	valor_equipo	agencia	altura	internacional	edad_debut	comunitario	traspasos	Elo	potencial	raza
Toni Fuidias	Girona FC	Primera España	Portero	15/4/01	22	Real Madrid	0	0,3	España	219	Desconocido	195	0	18,4	1	1	42	54	Blanco
Paulo Gazzaniga	Girona FC	Primera España	Portero	2/1/92	31	Valencia CF	0	3	Argentina	219	LGT Football	195	1	19,6	1	12	75	78	Blanco
Miguel Gutiérrez	Girona FC	Primera España	Lateral izquierdo	27/7/01	22	Real Madrid	10	15	España	219	Wasserman	180	0	16,3	1	1	75	81	Blanco
Arnau Martínez	Girona FC	Primera España	Lateral derecho	25/4/03	20	FC Barcelona	0	15	España	219	Tripple Match	182	0	17,5	1	5	79	86	Blanco
Eric García	Girona FC	Primera España	Defensa central	9/1/01	22	FC Barcelona	70	12	España	219	DE LA PEÑA & SOSTRES	182	1	16,6	1	3	72	79	Blanco
David López	Girona FC	Primera España	Defensa central	9/10/89	34	RCD Espanyol	237	3	España	219	Sports and Management	185	0	20	1	6	76	84	Blanco
Bernardo Espinosa	Girona FC	Primera España	Defensa central	11/7/89	34	Sevilla FC	2	1,5	Colombia	219	You First	192	0	18,4	1	12	63	80	Blanco
Aleix García	Girona FC	Primera España	Centrocampista	28/6/97	26	Villarreal	1	10	España	219	MagicPlayers	173	1	16,8	1	12	79	82	Blanco
Iván Martín	Girona FC	Primera España	Centrocampista	14/2/99	24	Villarreal	0	4	España	219	Universal Tweenty Two	178	0	18,5	1	3	74	78	Blanco
Cristian Portugués	Girona FC	Primera España	Extremo derecho	21/5/92	31	Valencia CF	2	3,5	España	219	Wasserman	167	0	17,9	1	10	72	81	Blanco
Valery Fernández	Girona FC	Primera España	Lateral izquierdo	23/11/99	24	FC Barcelona	0	2,5	España	219	Desconocido	182	0	18,8	1	3	57	66	Blanco
Dani García	Athletic Club	Primera España	Centrocampista	24/5/90	33	Real Sociedad	0	2	España	290	RGFOOTBALL	179	0	19,8	1	9	70	80	Blanco
Yuri Berchiche	Athletic Club	Primera España	Lateral izquierdo	10/2/90	33	Real Sociedad	88	2	España	290	Media Base Sports	181	0	19,1	1	9	78	83	Blanco
Mario Hermoso	Atlético de Madrid	Primera España	Defensa central	18/6/95	28	Real Madrid	0	20	España	472	Gesport Espizua SL	184	1	18,3	1	3	81	84	Blanco
Marcos Llorente	Atlético de Madrid	Primera España	Centrocampista	30/1/95	28	Real Madrid	39	35	España	472	Desconocido	184	1	18,6	1	3	81	86	Blanco
Pablo Barrios	Atlético de Madrid	Primera España	Centrocampista	15/6/03	20	Real Madrid	0	20	España	472	CAA Stellar	181	0	17,4	1	1	64	76	Blanco
Saúl Ñíguez	Atlético de Madrid	Primera España	Centrocampista	21/11/94	29	Real Madrid	0	12	España	472	CAA Stellar	184	1	15,8	1	1	77	91	Blanco
Antoine Griezman	Atlético de Madrid	Primera España	Delantero centro	21/3/91	32	Real Sociedad	202	25	Francia	472	By and For	176	1	18,5	1	5	94	97	Blanco
Álvaro Morata	Atlético de Madrid	Primera España	Delantero centro	23/10/92	31	Real Madrid	95	20	España	472	Niagara Sports Company	189	1	17,1	1	9	86	90	Blanco
Iñaki Peña	FC Barcelona	Primera España	Portero	2/3/99	24	Villarreal	0	4	España	862,5	INTEGRAL ADVISING FOOTBALL S.L	184	0	16,5	1	2	43	56	Blanco
Ander Astralaga	FC Barcelona	Primera España	Portero	3/3/04	19	Athletic Club	0	0,5	España	862,5	CAA Stellar	190	0	16,7	1	1	48	66	Blanco
Iñigo Martínez	FC Barcelona	Primera España	Defensa central	17/5/91	32	Real Sociedad	239	8	España	862,5	Camelo Sánchez	182	1	19,3	1	2	75	86	Blanco
Marcos Alonso	FC Barcelona	Primera España	Lateral izquierdo	28/12/90	32	Real Madrid	0	3	España	862,5	LIAN Sports Group	188	1	18,2	1	5	73	86	Blanco
Pablo Gavi	FC Barcelona	Primera España	Centrocampista	5/8/04	19	Betis	0	90	España	862,5	DE LA PEÑA & SOSTRES	173	1	16,3	1	1	80	92	Blanco
Fermín López	FC Barcelona	Primera España	Centrocampista	11/5/03	20	Betis	0	8	España	862,5	CAA Stellar	174	1	17,5	1	1	65	77	Blanco
Oriol Romeu	FC Barcelona	Primera España	Centrocampista	24/9/91	32	RCD Espanyol	0	5	España	862,5	MagicPlayers	183	0	16,9	1	7	75	78	Blanco
Ferran Torres	FC Barcelona	Primera España	Extremo derecho	29/2/00	23	Valencia CF	97	35	España	862,5	Leaderbrock	184	1	16,6	1	2	75	82	Blanco
Aitor Fernández	CA Osasuna	Primera España	Portero	3/5/91	32	Athletic Club	0	3	España	116,6	Sports and Management	182	0	19,3	1	6	68	77	Blanco
Nacho Vidal	CA Osasuna	Primera España	Lateral derecho	24/1/95	28	Valencia CF	9	3	España	116,6	Promoesport	180	0	19,2	1	2	67	77	Blanco
Juan Cruz	CA Osasuna	Primera España	Lateral izquierdo	28/7/92	31	Atlético de Madrid	0	2,5	España	116,6	Roalza	182	0	18,3	1	7	74	78	Blanco
Jesús Areso	CA Osasuna	Primera España	Lateral derecho	2/7/99	24	Athletic Club	0	2	España	116,6	Global Ases	182	0	18,2	1	3	66	72	Blanco
Lucas Torró	CA Osasuna	Primera España	Centrocampista	19/7/94	29	Real Madrid	0	8	España	116,6	Toldra Consulting S.L	190	0	17,2	1	6	74	77	Blanco
Moi Gómez	CA Osasuna	Primera España	Extremo izquierdo	23/6/94	29	Villarreal	222	5	España	116,6	InterStarDeporte	176	0	16,7	1	6	75	78	Blanco

Fuente: Elaboración propia