Trabajo Fin de Grado

Grado en Ingeniería de las Tecnologías Industriales

Diseño y aplicación de técnicas de machine learning para optimizar el Scouting en clubes de fútbol

Autor: Carlos Soria Polo

Tutor: Alicia Robles Velasco

Dpto. Organización Industrial y Gestión de Empresas II
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2021

Trabajo Fin de Grado
Grado en Ingeniería de las Tecnologías Industriales

Diseño y aplicación de técnicas de machine learning para optimizar el Scouting en clubes de fútbol

Autor:
Carlos Soria Polo

Tutor:
Alicia Robles Velasco
Profesor sustituto interino

Dpto. de Organización y Gestión de Empresas II
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla
Sevilla, 2021

Trabajo Fin de Grado: Diseño y aplicación de técnicas de machine learning para optimizar el Scouting en clubes de fútbol

Autor: Carlos Soria Polo
Tutor: Alicia Robles Velasco

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

El Secretario del Tribunal

Fecha:

Agradecimientos

Me gustaría dedicar unas palabras a las personas que han sido partícipes tanto de todos los éxitos durante mi vida en general como de estos últimos cuatro años cursando el Grado en Ingeniería de las Tecnologías Industriales.

A mis padres, a mi hermana y a mi pareja, por la confianza que han depositado en mí. Su apoyo tanto moral como económico, ha resultado imprescindible para poder conseguir mis metas tanto formativas como personales. También, han sabido construir un entorno en el que poder trabajar sin tener mayores preocupaciones que las puramente académicas.

A mis amigos, por servir como desahogo en los momentos de estrés y de apoyo moral para poder seguir adelante.

A mis compañeros durante la carrera, a los cuales conocí cursando la intensificación, que han sido de gran ayuda para la resolución de dudas y para servir como apoyo para afrontar todo lo que nos venía.

A mis profesores durante la carrera y, en especial, a mi tutora de TFG, Alicia Robles Velasco, que me ha ayudado en todo lo que ha podido y más, de forma muy atenta y entusiasta, permitiéndome realizar un trabajo sobre una temática que, personalmente, me apasiona.

Carlos Soria Polo

Sevilla, 2021

Resumen

El objeto del presente trabajo fin de grado es realizar un estudio acerca de los algoritmos de aprendizaje automático, más conocido como 'Machine Learning', y cómo estos pueden optimizar la búsqueda de jugadores objeto de adquisición por parte de clubes de fútbol.

Esto se consigue gracias a la implementación de algoritmos en el lenguaje de programación Python, que permiten agrupaciones y buscar similitudes entre los futbolistas que forman parte de una base de datos previamente elaborada. Concretamente, se ha hecho uso de dos técnicas de Machine Learning: Análisis de las Componentes Principales (PCA, en adelante, por sus siglas en inglés) y Clustering.

El marco teórico de este trabajo se fundamenta en el análisis de un club de fútbol. Para ello, se analiza la organización del departamento que decide la adquisición de futbolistas y, en concreto, las tareas de obtención y procesamiento de datos que sirven de apoyo a dichas decisiones.

Haciendo uso de las planificaciones de la presente temporada de dos equipos y gracias a la implementación de las técnicas de ML utilizadas en este trabajo, se han obtenido planificaciones alternativas a las realizadas por los clubes que se han escogido como ejemplo. Estas planificaciones alternativas se han obtenido a partir de una lista de jugadores, que, según las técnicas aplicadas, tienen características parecidas a los jugadores adquiridos por los clubes en la realidad, siempre teniendo en cuenta que el presupuesto invertido en la compra de los jugadores en la planificación alternativa no puede sobrepasar el presupuesto real.

Abstract

The present end of degree project's object is to conduct a study on automated algorythms, also known as 'machine learning', and how these can optimize the search of players to be acquired by football clubs.

This can be achieved by algorythm implementation in the programming language Python, which allows to group and look for similarities among the players who are part of the database previously elaborated. Specifically, two Machine Learning techniques have been used: Principal Component Analysis (PCA) and Clustering.

This project's theoterical framework is based on the analysis of one football club. This is done by analyzing the organization of the department which decides the footballers' acquisition and, more precisely, the data collection and processing tasks, which support these decisions.

Using the current season's adquisition planning and thanks to the implementation of the two techniques used in this project, alternative plannings have been obtained to those carried out by the clubs which have been chosen as an example for the case studies. These alternative plannings have been constituted from a list of players which, according to the applied techniques, have similar characteristics to the ones acquired by the clubs in reality, always taking into account that the amount invested in the purchase of players in the alternative planning cannot excede the real budget.

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

El deporte, y particularmente el fútbol, que es el que se trata en este trabajo, está sometido, igual que cualquier aspecto de la sociedad, a constante cambio y adaptación a las nuevas tecnologías disponibles tanto para los usuarios de a pie, como para los trabajadores y directivos de cualquier empresa que se precie. Esto ha precipitado, en cierto modo, que las empresas tengan que modernizarse y adaptarse a los tiempos para poder seguir compitiendo en un mercado en el que se necesita estar en constante renovación para estar al frente en materias como la innovación y el desarrollo, que repercutirán directamente en el bienestar de la empresa.

El éxito en el fútbol y en cualquier deporte, aunque no lo parezca, está poco determinado por componentes azarosos. Todo es fruto del trabajo y la dedicación de las personas que forman parte del club: presidente, consejeros, dirección deportiva, futbolistas, cuerpo técnico, trabajadores rasos, etc. Como se ha comentado con anterioridad, el trabajo y la dedicación necesitan también de una innovación, que te hacen estar a la vanguardia y poder, así, elevar tu techo competitivo.

Este trabajo y dedicación tiene que ir de la mano de un conocimiento exhaustivo de la realidad material de la actualidad, en el que las tecnologías son vitales para el desarrollo de la actividad productiva. En el fútbol, particularmente, la digitalización avanza a pasos agigantados: modernización de equipos del staff técnico, estudios de estadísticas de rivales para la preparación de cara a los partidos y, como uno de los avances más importantes y que se va a desgranar en el presente trabajo: el Machine Learning y el Big Data.

Según FundéuRAE, la definición de Machine Learning (anglicismo de aprendizaje automático, que se tratará con detenimiento en un capítulo específico, "es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender" (FundéuRAE, 2018). Al ser tan versátil y tener un campo de aplicación sumamente amplio, se puede aplicar a una gran variedad de aspectos deportivos que toman parte en la dirección de un club de fútbol: gestión de esfuerzos para prevención de lesiones, adecuación de onces iniciales frente a un rival atendiendo a sus debilidades, predicción de partidos, y, el aspecto tratado en este trabajo: scouting y dirección deportiva.

Por su parte, el Big Data es el conjunto de técnicas que permiten analizar, procesar y gestionar conjuntos de datos extremadamente grandes que pueden ser analizados informáticamente para revelar patrones, tendencias y asociaciones, especialmente en relación con la conducta humana y las interacciones de los usuarios (RAE, n.d.). De igual manera que con el Machine Learning, el Big Data puede tener muchísimas aplicaciones: estudio pormenorizado del rival, estudio del rendimiento de los futbolistas de nuestra plantilla, análisis económico de los clubes de nuestro entorno, etc. En este caso, el uso va encaminado a analizar los

datos obtenidos de cada futbolista y, de esa manera, poder tener más argumentos e información para poder ir al mercado a acometer un fichaje.

Una de las áreas que determina el crecimiento de un club, por no decir la que más, es la dirección deportiva, que ayuda a los equipos a encontrar valor añadido gracias a las diferentes herramientas de búsqueda de talento. Esto es vital, debido a que confiere las posibilidades de conformar una plantilla que, en un futuro, puede dar, aparte de buen rendimiento deportivo, un buen rendimiento económico, que también juega un papel vital en el éxito de los clubes de fútbol.

Las direcciones deportivas de los mejores clubes del mundo han encabezado esta modernización de los métodos de trabajo, ya que poseen una cantidad mayor de recursos para destinar a los departamentos de I+D. La inversión en estos departamentos supone una diferenciación con respecto a los competidores, ya que permite encontrar talento y abarcar más territorio de manera más sencilla y eficiente, que, al final, es lo que se persigue como en cualquier empresa de cualquier sector.

La importancia del desarrollo de estas tecnologías es crucial, debido a que hay una grandísima cantidad de ligas, equipos y jugadores a los que seguir. Por ello, se necesitan herramientas informáticas que permitan hacer este seguimiento de manera más sencilla. Teniendo en cuenta la gran cantidad de datos que se ofrecen, se debe aprender a gestionarlos y a hacer de esa buena gestión una ventaja competitiva frente a los competidores, que son los equipos que tienen, aproximadamente, los mismos objetivos y el mismo presupuesto que el del caso de estudio.

Por último, hay que recalcar que el Machine Learning (en adelante, ML) aplicado en el scouting no es más que una herramienta que permite facilitar las tareas a realizar. No obstante, hay factores intangibles que el ML no puede analizar porque no se pueden cuantificar, al menos de manera objetiva, como por ejemplo la adaptabilidad del futbolista a diferentes situaciones y escenarios, la actitud del futbolista con sus compañeros dentro y fuera del campo, etc. Por tanto, esta herramienta es útil para hacer una primera criba de todos los datos que poseemos y otorgar al director deportivo diferentes posibilidades, pero después es trabajo del staff técnico el decidir si el futbolista que el ML destaca es idóneo para lo que se busca o no.

2 OBJETIVOS

A continuación se procederá a definir tanto el objetivo general del presente Trabajo Fin de Grado como la justificación, que trata de situar el objeto de este trabajo en la realidad.

2.1 Objetivo general

El objetivo principal de este proyecto es gestionar de manera eficiente la mayor cantidad de datos posible para hacer un estudio del mercado de cara a acometer un fichaje que se ajuste de mejor manera tanto a la posición y al perfil de futbolista que estamos buscando.

Esto se va a llevar a cabo a partir de algoritmos de Machine Learning que harán uso de bases de datos para establecer aquellos parámetros que son más importantes y que más peso tienen en dicho análisis. Teniendo en cuenta que el objetivo es encontrar jugadores que tengan características en común, los algoritmos que mejor se ajustan son PCA y Clustering. Estos algoritmos permiten agrupar a los jugadores mediante ciertas variables y relaciones lineales y no lineales entre ellas.

2.2 Justificación

El problema que se aborda en este proyecto es, aparte de novedoso, muy útil para los equipos tanto con un gran presupuesto, que ya han puesto el ojo en estas tecnologías de cara a hacer más eficaz la labor de scouting y necesitan de un continuo desarrollo de esta herramienta para poder diferenciarse de sus competidores, como para los equipos con un presupuesto menor que empiezan a dedicar partidas presupuestarias a innovar en este respecto.

Con la grandísima cantidad de datos que se ofrecen desde muchísimos sitios web y aplicaciones, es interesante poder aprender a gestionarlos de manera que sirvan como ayuda para poder cribar futbolistas y, así, poder centralizar el trabajo en la decisión entre los futbolistas que proporcione el ML.

3 ORGANIZACIÓN DE UN CLUB DE FÚTBOL

Para poner en contexto el funcionamiento de una sociedad anónima deportiva, sobre todo de cara a los no seguidores de fútbol, se procede a comentar tanto la estructura de un club de fútbol como el funcionamiento de la dirección deportiva en particular. Además, se tratará cómo se podría encuadrar el Machine Learning en el área de scouting, en función de la filosofía de compra de jugadores que siga cada club.

3.1 Organigrama

La estructura de un club de fútbol, por lo general, está compuesta de:

- **Consejo de Administración (Presidente y consejeros):** es el órgano que toma las decisiones de gestión ordinaria de un club. Es el órgano de poder del equipo.
- **Director General:** es la cúspide de todo el organigrama ejecutivo. Por lo general, es responsable de muchas de las áreas de gestión de un club, como el área deportiva, marketing, presupuestaria, comunicación, instalaciones, etc. Sirve de enlace con el Consejo de Administración.
- **Director Deportivo:** es la cabeza visible de la dirección deportiva del club y es el encargado de, una vez recogida toda la información de la dirección deportiva, tomar la última decisión acerca de la adquisición de un futbolista, y, una vez decidido, el encargado de acometer su fichaje y, consecuentemente, negociar con el club vendedor.
- **Otros directores:** en cada área de relevancia en un club de fútbol, igual que en área deportiva, hay una persona con mayor capacidad de decisión. Como ejemplo de directores que suelen haber en un club: marketing, financiero, jurídico, comunicación, RRHH, instalaciones, etc.
- **Dirección deportiva:** es el área dedicada al estudio y análisis del mercado futbolístico, se dedica a la compraventa de jugadores y a la confección de la plantilla. Posteriormente, se verá que, según el tipo de filosofía en la dirección deportiva de un club, que varía de un equipo a otro, esta área tiene más o menos peso.

A continuación, se mostrarán algunos ejemplos reales de organigramas de dirección de un club de fútbol.

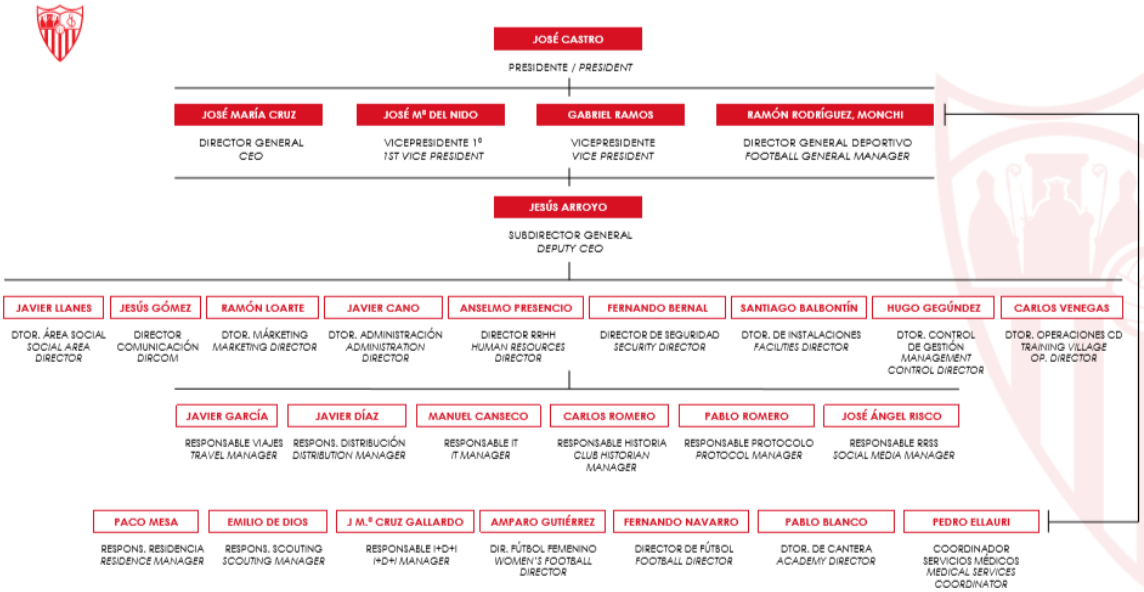

Figura 3-1: Organigrama del Sevilla Fútbol Club (Sevilla FC, 2020a)

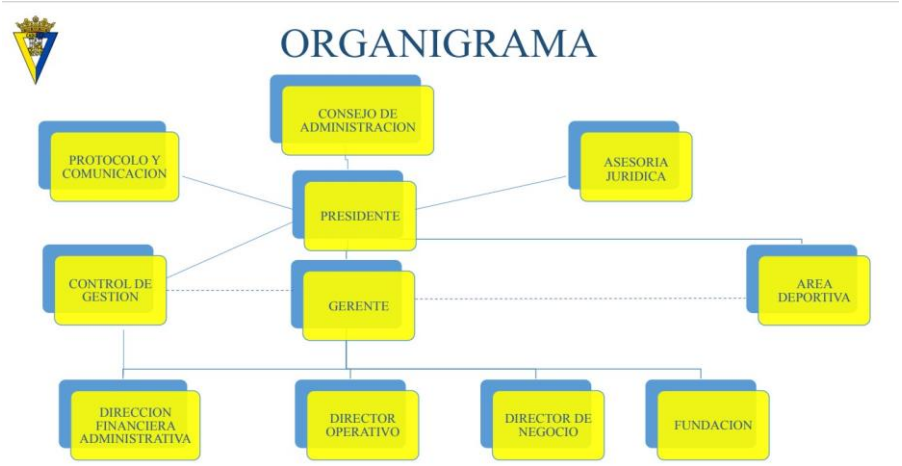

Figura 3-2: Organigrama del Cádiz CF (Cádiz CF, 2020)

Figura 3-3: Organigrama del FC Barcelona (*El Organigrama Ejecutivo Del Barça*, 2017)

Como se observa en las tres figuras dispuestas de la página anterior, al ser clubes pertenecientes a la liga española, su estructura es relativamente parecida. Como luego se verá, todos siguen lo que se conoce como modelo mixto de gestión de la dirección deportiva.

Sí es cierto que la diferencia de presupuesto entre un club y otro se hace notar, ya que se ve cómo en el FC Barcelona (y en menor medida en el Sevilla FC), se tiene un organigrama mucho más completo en el que hay áreas enfocadas exclusivamente a la imagen corporativa del club, así como a la relación con los aficionados. Destaca la potenciación, en los clubes con mayor presupuesto, de las áreas referentes a la tecnología de información, innovación, etc., con menor incidencia en clubes del tamaño del Cádiz CF o inferior, debido a la falta de recursos para poder invertir en estas áreas.

3.2 Dirección deportiva

La dirección deportiva es una de las áreas vitales de un club de fútbol, ya que, como se ha comentado previamente, se encarga de todo lo relacionado con la confección de la plantilla acorde a lo que el club necesita y se puede permitir, y lo que el entrenador requiere para poder llevar a cabo su idea de juego.

Según la filosofía que posea el club, se tienen distintos tipos de roles para la dirección deportiva:

- **Modelo presidencialista:** el presidente es el encargado de la gestión deportiva de la plantilla, por lo tanto, de la compraventa de fichajes, siempre apoyándose en un equipo de trabajo, pero al final la decisión sobre la adquisición o la venta de un futbolista recae en él. Como ejemplo principal, tenemos al Real Madrid CF de Florentino Pérez.
- **Modelo anglosajón:** la figura importante es el entrenador, que decide tanto los perfiles a fichar como los nombres, así como la gestión de la plantilla y las ventas pertinentes. Como su propio nombre indica, es la filosofía predominante de los equipos de la Premier League inglesa, con los ejemplos más representativos del Arsenal de Arséne Wenger, el Manchester United de Sir Alex Ferguson, el Liverpool de Rafa Benítez, entre otros. En España, al no tener tanto peso el entrenador, no tenemos casos de este modelo de gestión.
- **Modelo mixto:** está basado en la coordinación entre 3 equipos de trabajo: el presidente de la entidad junto al equipo de administración, que es el encargado de la parte económica de la planificación, y es el órgano que dictamina si un futbolista es asequible para su adquisición, así como la decisión de venta de un jugador según el valor de la oferta que se reciba por él; el entrenador que decide el perfil de jugador que necesita para cumplir un rol en una determinada posición y; por último, la dirección deportiva, que es la encargada de hacer cumplir las necesidades determinadas por el entrenador en forma de nombres, después de la labor de scouting.

Obviamente, teniendo en cuenta las restricciones de funcionamiento dependiente de la filosofía del club, el scouting tendrá una función más o menos importante y con mayor peso en el modelo mixto que en el modelo presidencialista o el modelo anglosajón, debido a que la labor de rastreo de futbolistas es más necesaria si es la dirección deportiva la que decide el nombre del jugador, mientras que en los otros dos modelos el fichaje de los futbolistas depende más de ofrecimientos de agentes, experiencias previas con algún futbolista en el caso de que hayan compartido equipo previamente con la persona encargada de fichar (ya sea presidente o entrenador), etc. (Sevilla FC, 2020b)

3.3 Scouting

"¿Qué es scouting? Es el análisis científico que realiza un Técnico Deportivo recabando la máxima información posible sobre el elemento que se le asigne, utilizando los medios tecnológicos adecuados para transmitirla a los interesados, minimizando, por tanto, los riesgos que puede causar la competencia". (Botello, 2012)

En otras palabras, el scouting es el área dentro de la dirección deportiva de un club encargada de, a partir de un análisis exhaustivo de datos y visionado de partidos de otros equipos, encontrar talento optimizando los recursos existentes.

3.3.1 Factores de importancia

El scouting alcanza una relevancia importante teniendo en cuenta varios factores:

1. **Digitalización del mundo del fútbol:** el acceso a partidos de fútbol de cualquier rincón del mundo cada vez es mayor, puedes acceder a servidores que tienen almacenados partidos de ligas tan recónditas como la liga tailandesa, la liga panameña o la liga congoleña, todo gracias a la informatización que permite tener acceso a una cantidad ingente de partidos de todo el mundo. En los clubes de fútbol, la plataforma más usada es Wyscout (suscripción de pago), en la cual, aparte de tener una base de datos de miles de futbolistas de cualquier liga, tiene unos 200000 partidos de todas las ligas para poder ver desde casa, cosa que facilita tremendamente la tarea de accesibilidad a todos los partidos.
2. **Diferenciación frente a los rivales:** teniendo en cuenta el nivel de profesionalidad que ha alcanzado el fútbol en aspectos puramente deportivos como la calidad técnica y física de los futbolistas, los clubes necesitan otras vías que le permitan, a partir de un buen trabajo, elevar su techo competitivo. Por ejemplo, un equipo que ahora mismo compita por no descender siempre va a buscar, como

mínimo, avanzar para tener la opción de no tener que sufrir durante la temporada para no descender de categoría.

Una de las áreas que otorga esta posibilidad es la dirección deportiva, y más particularmente, el trabajo de scouting. Si se tiene un departamento potente, organizado y con las ideas claras de lo que se busca, se podrá encontrar talento más rápido y con más acierto que tus rivales. Esto llevará a dicho club a partir en mejor posición para alcanzar los objetivos, tanto deportiva como económicamente.

Ejemplos de clubes que, mediante la labor de scouting y la dirección deportiva, han conseguido elevar su nivel competitivo para poder conseguir objetivos más ambiciosos, se tienen muchos como: Sevilla FC, Mónaco, Lille, Atalanta, Borussia Dortmund, Hoffenheim, etc.

A continuación, se presentan dos tablas de los equipos con mejor y peor balance (ingreso por ventas-gasto en futbolistas) de las grandes ligas europeas desde 2016 hasta 2021:

League -> Big-5

Balance	Club	League	Spent	Earned
+191M€	LOSC Lille	FRA	252M€	443M€
+151M€	Olympique Lyonnais	FRA	348M€	499M€
+133M€	Atalanta BC	ITA	241M€	374M€
+132M€	AS Monaco	FRA	626M€	758M€
+87M€	TSG Hoffenheim	GER	150M€	237M€
+78M€	AS St-Etienne	FRA	68M€	146M€
+77M€	Borussia Dortmund	GER	582M€	659M€
+76M€	Sampdoria UC	ITA	248M€	324M€
+75M€	Genoa CFC	ITA	181M€	256M€
+67M€	Valencia CF	ESP	292M€	359M€
+66M€	Athletic Club	ESP	79M€	145M€
+66M€	Girondins Bordeaux	FRA	90M€	156M€

Figura 3-4: Balance mejor ingreso-gasto de los clubes de las 5 ligas más importantes de Europa (Poli et al., 2020)

Net transfer spending since summer 2016
€ Million

League -> Big-5

Balance	Club	League	Spent	Earned
-631M€	Manchester City	ENG	1006M€	375M€
-586M€	Manchester United	ENG	832M€	246M€
-471M€	FC Barcelona	ESP	1171M€	700M€
-455M€	Paris St-Germain	FRA	854M€	399M€
-386M€	Internazionale	ITA	664M€	278M€
-346M€	Everton FC	ENG	701M€	355M€
-339M€	Aston Villa	ENG	407M€	68M€
-311M€	Milan AC	ITA	577M€	266M€
-308M€	Chelsea FC	ENG	968M€	660M€
-299M€	Arsenal FC	ENG	588M€	289M€

Figura 3-5: Balance peor ingreso-gasto de los clubes de las 5 ligas más importantes de Europa (Poli et al., 2020)

De estas dos figuras (3-4 y 3-5) se pueden sacar varias conclusiones:

- Los clubes ingleses son los que se encuentran, en mayor número, en los peores puestos de este ranking. Esto viene motivado por dos razones:
    - La cantidad ingente de dinero que reciben de la liga inglesa por los derechos televisivos. Esto les permite hacer grandes inversiones en futbolistas sin tener que preocuparse en demasía por buscar talento a bajo precio, ni por vender a futbolistas de su plantilla para que el modelo de negocio sea sostenible. El caso de la Premier League es un caso especial, ya que prácticamente ninguna liga en el mundo recibe los ingresos que ellos tienen por derechos televisivos. Este hecho les permite poder ser tremendamente ineficientes en lo que a optimización de recursos se refiere, pero poder seguir compitiendo al más alto nivel por la calidad de sus plantillas.
    - El modelo de gestión anglosajón, aparte de no tener como meta principal la optimización de los recursos (por lo comentado en el punto anterior), es ineficiente ya que no es la dirección deportiva la que, después de un análisis exhaustivo, decide el jugador a fichar, sino que es el propio manager el que da un nombre, por lo que el estudio del fichaje a acometer es mucho menor. Esto lleva a que, en la Premier más que en cualquier otra liga europea, se produzcan fichajes que requieren una gran inversión que al final no terminen por triunfar.

- Por el contrario, no hay ningún club inglés en los 10 mejores balances expuestos en la figura 3-4, lo que acentúa aún más la idea de la "mala gestión" de estos equipos con respecto a equipos de otros países con menos recursos.
- Los equipos con mejores balances, por lo general, son equipos de segundo o tercer escalón de Europa., entendiendo por escalón al grupo de clubes al que pertenecen según al techo competitivo al que aspiran.
- Los equipos que están en el primer escalón se puede decir que son los equipos que la mayoría de años tienen como objetivo la consecución de un título, en particular, su liga doméstica y, sobre todo, la Liga de Campeones. Los equipos de segundo escalón son los que pelean por entrar en la Liga de Campeones y los de tercer escalón son los que intentan conseguir plaza en la Europa League, etc.

  Esta clasificación es un poco orientativa y subjetiva, pero es relativamente común en los clubes para definir quiénes son tus competidores y quiénes están algún escalón por encima o por debajo.

- Los equipos con peores balances son aquellos que tienen un presupuesto más alto. Esto se observa claramente en la figura 3-5 en la columna 'spent' (gastado en inglés), en la cual muchos de los clubes superan los 800M€ de inversión e incluso los 1000M€ desde 2016, lo cual es una cantidad enorme sólo al alcance de grandes clubes de Europa.

3. **Aumento de ingresos:** ahondando un poco más en el anterior punto, el trabajo de scouting hace, no solo tener un techo competitivo más alto porque el nivel de los futbolistas es mayor, sino que hace que se tenga una mayor cantidad de beneficio por la venta de los futbolistas, ya que lo que se ha fichado, con una labor de scouting por medio, por un precio relativamente asequible y ajustado a unas necesidades concretas, después de un buen rendimiento, se puede obtener una gran plusvalía. Esto permite a los clubes, en los posteriores mercados, trabajar con un mayor presupuesto, que hace aumentar las posibilidades en el mercado de fichar futbolistas más contrastados y con mayor recorrido.

   Además, si los futbolistas adquiridos ofrecen un buen rendimiento deportivo, esto directamente repercute en unos mayores ingresos por obtener mejores resultados en las competiciones que se disputan, por lo que no sólo es dinero obtenido directamente mediante la plusvalía obtenida por la venta de un futbolista, sino que, además, un mejor resultado deportivo te otorga unos ingresos mayores.

3.3.2 Estructuración

Para este punto, y algunos de los que vienen a continuación, nos vamos a centrar en una masterclass del director deportivo del Sevilla FC, que desgrana la estructura y la metodología de trabajo del club sevillano (Sevilla FC, 2020c)

El área de scouting, por lo general, cuenta con los siguientes elementos:

1. **Director deportivo:** es el que da los perfiles de búsqueda de futbolistas a los diferentes scouts y el que, después del proceso de análisis del mercado y la obtención de nombres, decide los nombres que finalmente van a ser objetivos principales de la secretaría técnica.
2. **Coordinador de scouting:** dentro del área de scouting, es el nexo de unión entre todos los scouts y que organiza la información recogida por el equipo para transmitírsela al director deportivo.
3. **Scout:** es el encargado de, a partir de unas órdenes recibidas por el director deportivo y el coordinador de scouting, analizar la(s) competición(es) que se le asigne(n). Su función se basa en ver partidos y rellenar bases de datos, que después se remitirán a su superior.
4. **Área de Big Data:** a partir de las bases de datos de cosecha propia y de portales web como Wyscout, Instat, etc. se realizan análisis de las mismas para la obtención de nombres concretos en base a perfiles definidos por el director deportivo. Por lo general, en esta área, también se desarrollan algoritmos de Machine Learning que sirven como apoyo a la labor de los scouts y del director deportivo.

3.3.3 Metodología de trabajo

En este apartado se explican las cuatro fases más relevantes del proceso de scouting (Figura 3-6). Para ello, se toma como referencia, como se ha comentado previamente, el caso del Sevilla FC. (Sevilla FC, 2020c)

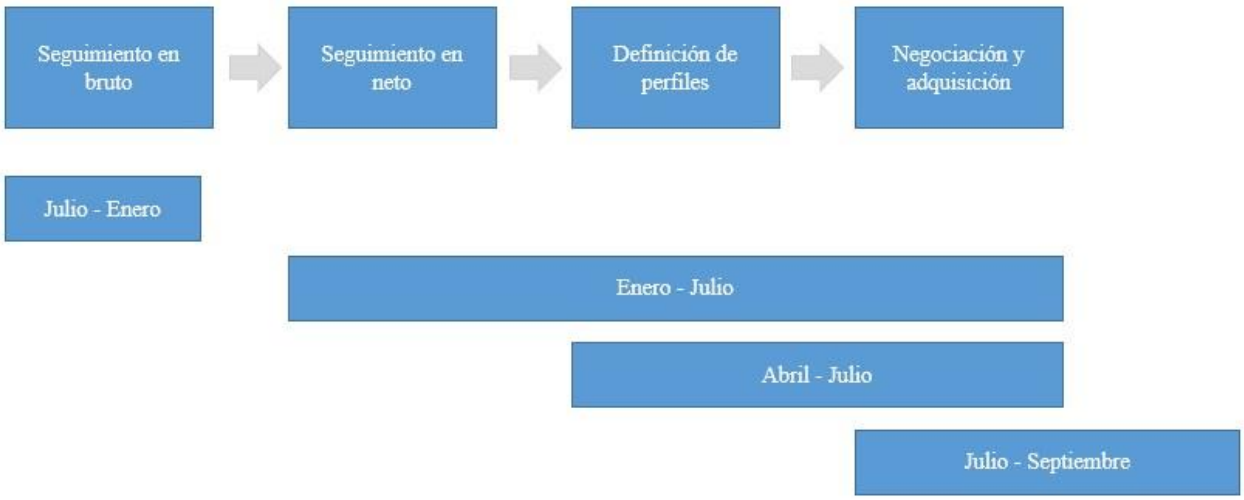

Figura 3-6: Fases relevantes en la planificación del área de scouting

3.3.3.1 Seguimiento en bruto

El seguimiento en bruto es el análisis del fútbol sin un objetivo de perfil y posición concreta de futbolista. Con este seguimiento se busca engordar las bases de datos del club, de cara a posteriores seguimientos en los que sí esté definido ya, por parte del entrenador, el tipo de futbolista que se busca. Este proceso abarca desde julio (que es, aproximadamente, cuando finaliza la temporada) hasta diciembre.

Uno de los criterios a utilizar es la clasificación de las ligas donde se realiza el rastreo según su nivel futbolístico, de tal manera que tenemos:

1. **Ligas tipo A:** son las ligas de primer nivel europeo e internacional que, históricamente, más nutren y han nutrido al Sevilla FC (por su nivel) de futbolistas. Tenemos: Bundesliga (Alemania), Ligue 1 (Francia), Serie A (Italia), Premier League (Inglaterra), Jupiler Pro League (Bélgica), Eredivisie (Holanda), Liga NOS (Portugal), Primera División (Argentina) y Brasileirao (Serie A).
2. **Ligas tipo B:** son las ligas de segundo nivel europeo e internacional con un nivel competitivo más bajo, pero que igualmente hay que seguir, para poder abarcar lo máximo posible. El seguimiento a estas ligas se centra más en las selecciones nacionales de cada uno de los países que en los clubes que pertenecen a cada una de las ligas, haciendo un seguimiento del primer equipo y las selecciones inferiores (U21, U20, U19, etc.). Se aprovecha el trabajo de selección de los seleccionadores de cada uno de los equipos para buscar de manera más eficiente el talento. Tenemos: Bundesliga (Austria), Ekstraklasa (Polonia), Superliga (Suiza), Liga HNL (Croacia), 1. Liga (República Checa), Primera División (Chile), Primera (Colombia), División Profesional (Bolivia), Liga 1 (Perú), Liga MX (México) y MLS (EEUU).
3. **Ligas tipo C:** son las más residuales y se centra más en el seguimiento de competiciones federales. Tenemos: Copa de Oro (países de Norteamérica, Centroamérica y el Caribe), Copa Conmebol (países de Sudamérica), Copa de Asia y Copa África.

Estos tres tipos de ligas se reparten entre los scouts (en este caso, en el área de scouting del Sevilla FC poseen doce scouts), de manera que, a cada uno de ellos, por lo general, se le asigna una liga tipo A y 2/3 ligas tipo B.

Los objetivos principales de este seguimiento son los siguientes:

- **Visionado del máximo número de partidos posibles sin objetivo concreto:** como se ha comentado previamente, tiene como fin la potenciación de la base de datos del club.
- **Realización de onces ideales cada mes:** cada scout, en cada una de las ligas de clase A que tenga asignadas, realiza mensualmente un once ideal. Para el caso de las ligas de clase B, el once ideal es

global entre todas las ligas de ese tipo asignadas a un mismo scout. En cada posición se determina, en ese período de tiempo, el futbolista que más les ha llamado la atención.

- **Selección de jugadores elegidos de cada campeonato por cada posición:** en las ligas de clase C, se hace un XI del campeonato completo.

De este seguimiento se obtiene lo que Ramón Rodríguez Verdejo (alias 'Monchi'), director deportivo del Sevilla FC, trata como 'el primer filtro', que es la obtención de información y consiguiente incorporación a la base de datos de unos 500-550 futbolistas.

3.3.3.2 Seguimiento en neto

El seguimiento en neto es el siguiente paso en el proceso de la planificación deportiva del club. Este paso comprende los posteriores seis meses al seguimiento en bruto, de tal manera que empieza en enero y finaliza en junio, que es, precisamente, cuando finaliza la temporada y el club comienza a moverse en el mercado para acometer los fichajes de cara a realizar la planificación, atendiendo a los perfiles definidos por el entrenador y en base a los datos recogidos por el área de scouting a partir del trabajo previo (Sevilla FC, 2020d).

Esta etapa comprende los siguientes aspectos:

- **Visionado de jugadores seleccionados en la fase seguimiento en bruto:** desaparece la asignación de campeonatos de cada uno de los scouts y se trabaja sobre la información que se recoge en el seguimiento en bruto, por lo tanto, en cada uno de los futbolistas concretos.
- **Seguimiento en distintas situaciones:** en este punto se tiene en cuenta el aspecto mental del futbolista y la atención a los pequeños detalles, centrándose sobre todo en cómo se desenvuelve el futbolista en cada una de las situaciones que aparecen. Por ejemplo: partidos de local, partidos de visitante, partidos contra rivales fuertes, partidos contra rivales asequibles, partidos con su selección, etc. Este punto sirve para recabar información del aspecto mental del futbolista, para complementarlo con la información recogida en el seguimiento en bruto de aspectos más físicos y referidos al mundo futbolístico.
- **Seguimiento por diferentes scouts:** los scouts que han recogido información a partir del visionado de partidos de una liga asignada no vuelven a ver a ese futbolista, sino que será otro scout el que visualice los partidos de ese futbolista para contrastar opiniones. Por lo general, se visiona de cada jugador unos 6-7 partidos en distintas condiciones, por lo comentado en el punto anterior.

Cada uno de los scouts que hace el visionado de cada uno de los futbolistas obtenidos en el seguimiento en neto realiza un informe, centrado en los siguientes seis aspectos (Sevilla FC, 2020e):

- **Perfil físico:** analizar parámetros físicos de los jugadores para establecer si son acordes al perfil de futbolista que demandamos para una posición concreta.
- **Perfil técnico-táctico:** cada scout en cada uno de los informes evalúa en calificaciones de la A hasta la E (siendo A el mejor valor y E el peor) la calidad técnica y táctica del futbolista. Por tanto, es un parámetro completamente subjetivo, ya que no atiende a datos concretos, por lo que es importante que el director deportivo confíe y delegue responsabilidades en el cuerpo de scouting. Además, al haber una cantidad ingente de informes por parte de los scouts (cada futbolista puede tener unos 6-7 informes, siendo el total de futbolistas a analizar unos 500-550), es importante que el director deportivo encuentre personas que crea que tienen el criterio futbolístico suficiente para que dichos informes tengan valor.
- **Perfil psicológico:** es tarea del director deportivo, normalmente, recabar información en este aspecto del círculo cercano del futbolista al que se le está realizando el seguimiento, para así saber cuál es su manera de afrontar los problemas y de solucionarlos. Es un componente subjetivo y atiende completamente al criterio de la persona que te aporta la información, por lo que es tarea de la persona que recaba la información de dilucidar y filtrar cuáles son las ideas útiles para el análisis en este respecto y cuáles no.
- **Condiciones económicas:** también es una tarea a realizar por el director deportivo, que básicamente necesita conocer, a partir del club vendedor y/o del representante del jugador, los siguientes aspectos:
    - **El coste del jugador:** precio que pide el club que posee los derechos federativos del futbolista)
    - **El valor de la cláusula de rescisión**: montante que figura en el contrato que tiene que abonar el futbolista (personalmente la debe abonar él, pero el que hace el pago, realmente, es el club comprador) para rescindir unilateralmente su contrato.
    - **Salario demandado:** es la cantidad que pide el representante del jugador al club comprador que debería ser el sueldo a percibir por el futbolista en caso de querer recalar en su equipo.
- **Tiempo de adaptación estimado:** conocer cómo puede responder el jugador en estudio a un cambio en aspectos como la subida de exigencia si viene de un club de menor entidad, el manejo de la presión, adaptación a la ciudad, etc.
- **Revalorización futura:** en el modelo de negocio del Sevilla FC (y de la mayoría de los clubes deportivos, que no pertenecen al primer escalón competitivo) es vital atender este aspecto, que básicamente trata de analizar si el jugador a analizar nos puede otorgar unas plusvalías después de

un buen rendimiento en el club. El Sevilla, concretamente, basa este aspecto en la filosofía 70%-30%, en la que el 70% de los jugadores pueden dar un rendimiento económico futuro importante, sabiendo que pueden dar un rendimiento deportivo positivo a medio-largo plazo, y un 30% de ellos que estén más enfocados a un rendimiento a corto plazo. Por lo general, el parámetro principal que se tiene en cuenta a la hora de definir si buscamos rendimiento inmediato o futura revalorización en un jugador es la edad, ya que es más fácilmente revalorizable en el futuro un futbolista joven, y, por el contrario, más sencillo que nos otorgue buen rendimiento deportivo a corto plazo un futbolista experimentado.

A partir del estudio, que se ha centrado en estos aspectos, se elabora la lista de futbolistas por posición (unos 20 por cada una, aproximadamente), de los cuales se decidirán los fichajes para la planificación de la temporada.

3.3.3.3 Definición de perfiles

En este punto es el momento de involucrar al entrenador en la hoja de ruta de la dirección deportiva, estando antes centrado completamente en el aspecto deportivo. A partir de abril, aproximadamente, es cuando el director deportivo pregunta al entrenador por los perfiles necesarios de cara a la próxima temporada. Se entiende por perfil de futbolista como las características que definen su rol preferido en la posición en la que se desenvuelve en el terreno de juego.

El entrenador, cuestionado por el director deportivo, le traslada las características que necesita ver cumplidas por las futuras adquisiciones y, el director deportivo busca en cada una de las listas que tiene confeccionadas por posición (a partir del seguimiento en neto), los jugadores que más se ajustan en cada caso al perfil definido por el entrenador.

Se analiza cada una de las listas y, según todos los perfiles que ha definido el entrenador, se definen cuáles son los jugadores que cumplen sus exigencias.

Se vuelve a hablar con el entrenador, ya con los jugadores elegidos, y se le pregunta si conoce a alguno de los futbolistas (de experiencias previas en otros clubes, mayormente). Si el entrenador conoce alguno de los futbolistas que el director deportivo le lleva, el director deportivo, en el caso del Sevilla FC, se posiciona como primera opción de compra del club. Es obvio que, si el director deportivo define que esos futbolistas cumplen el perfil, se ha definido que ese futbolista es acorde a las necesidades del club (a partir del seguimiento en bruto y en neto) y el entrenador lo conoce de experiencias previas, es el idóneo.

En el caso de que el entrenador no conozca ninguno de los futbolistas que el director deportivo le traslada en este punto, que es lo que ocurre normalmente, es el director deportivo el responsable final de la

decisión de adquisición del futbolista. De igual manera, aunque ahora sea el director deportivo el que decide, también es importante mantener al entrenador informado, por lo que la dirección deportiva surte de información (una recopilación de vídeos de partidos donde el jugador tenga jugadas destacadas, por ejemplo) para que el entrenador conozca de una manera sencilla el tipo de jugador que se está analizando (Sevilla FC, 2020e).

3.3.3.4 Negociación y adquisición del futbolista

Una vez definida la primera opción, se acomete la negociación con el club vendedor bajo las siguientes premisas:

- Tener en cuenta coste total del futbolista: pago al club vendedor más el salario a percibir por el futbolista.
- Tener opciones suficientes como para no pagar un sobreprecio por la primera opción.
- Conocer y planificar las pretensiones de pago del club, atendiendo al presupuesto y a la planificación de la dirección deportiva.
- Maximizar beneficios propios sin abusar de la parte contraria y mantener buenas relaciones con todos los clubes.

Una vez tenidos en cuenta todos estos puntos y después de la negociación, lo que era una lista de unos 500-550 jugadores (con 6-7 informes para cada uno de los futbolistas), ha pasado a ser una lista de 20 jugadores definitivos en el seguimiento en neto, 8-10 después de hacer el cribado por los perfiles que definía el entrenador y, finalmente, acaba por materializarse este esfuerzo y trabajo en la adquisición de un futbolista (Sevilla FC, 2020f).

4 MACHINE LEARNING

Según (Hurwitz & Kirsch, 2018), el Machine Learning (en adelante ML) es un tipo de inteligencia artificial que aprende de los datos en vez de utilizar una programación explícita. Esto lo hace a partir de algoritmos que, de forma iterativa, aprenden de los datos para mejorar, y a partir de ellos, predice valores de salida. Como el modelo de Machine Learning aprende a base de datos, es posible producir modelos más precisos a partir de ellos. El hecho de que el algoritmo aprenda en base a datos nos asegura que el resultado que otorgue va a ser objetivo y va a estar actualizado constantemente.

El ML es uno de los temas más importantes en las empresas y organizaciones que están en desarrollo y que buscan formas para innovar de cara a la gestión de los datos. Esto va encaminado a ayudar al negocio a predecir cambios, de manera que se pueda actuar incluso antes de que se produzca el cambio, y así prevenir la improvisación y la falta de previsión.

El origen del Machine Learning no es tan cercano como podría parecer, por lo disruptivo de su tecnología. Se remonta a los años 50, década en la cual un investigador del IBM (International Business Machines) desarrolló un programa de autoaprendizaje para jugar a las damas. Él mismo fue el que utilizó el término de Machine Learning. Posteriormente, en el periódico "IBM Journal of Research and Development" se explicó su aproximación al Machine Learning.

El Machine Learning ha ganado importancia gracias a la cantidad ingente de datos almacenados. El análisis de estos datos se realiza a través del Big Data, fenómeno íntimamente relacionado con el ML.

- **Big Data**

Según la definición de (*Qué Es Big Data | Universidad Complutense de Madrid*, n.d.), el Big Data es el análisis masivo de una base de datos tan grande que las aplicaciones de software de procesamiento utilizadas tradicionalmente no son capaces de capturar, tartar y valorar en un tiempo razonable.

Las fuentes principales de obtención de datos de cara a la gestión del Big Data son múltiples:

- **Producidos por personas:** en cualquier actividad cotidiana del día a día como: responder un mail, escribir un comentario en una red social, introducir datos en una base de datos, navegación en Internet, etc.
- **Entre máquinas:** aplicaciones en los smartphones, redes de comunicación, sensores de cualquier ámbito, máquinas expendedoras, etc.

- **Biométricas:** sensores de huellas dactilares y faciales, escáneres de retina, lectores de ADN, etc.
- **Marketing web:** registro de los movimientos de los usuarios a través de webs de manera que el contenido nuevo a visualizar se ajuste a sus gustos y/o necesidades, así como el uso del propio usuario de la página en la que entra. Por ejemplo, si el usuario está buscando o pasa mucho tiempo en una parte concreta de una página web, el contenido publicitario se ajustará a esa necesidad que el análisis determina que puede tener.
- **Transacciones:** bancarias, reservas de billetes y hoteles, compras, etc.

También podemos dividir los datos en función de su estructura:

- **Datos estructurados:** tipo de datos que tienen un formato y una longitud definida claramente. Ejemplos: números, cadenas de caracteres, etc.
- **Datos no estructurados:** tipo de datos que no se pueden almacenar de forma tradicional, ya que no se puede desglosar la información de una forma definida en longitud y formato, sino que son variables. Ejemplos: emails, presentaciones, archivos PDF, etc.
- **Datos semiestructurados:** tipo de datos que no se encuadran en los estructurados ni en los no estructurados, debido a que no hay un patrón concreto que siga siempre. Ejemplos: HTML.

Por último, en la figura 4-1 se muestran los pasos descritos a continuación, que tratan las fases del ciclo de gestión de la información:

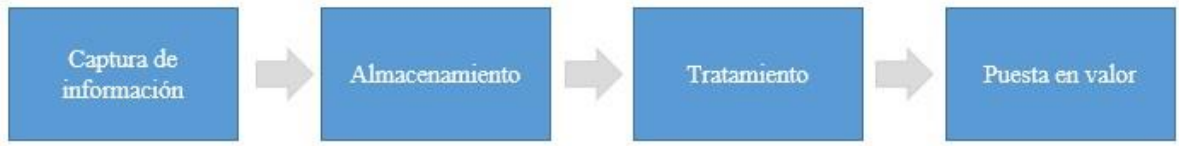

Figura 4-1: Fases del ciclo de gestión de la información del Big Data

1. **Captura de información:** es el punto en el que necesitamos definir cuál es la fuente de información de donde se van a obtener los datos, de manera que sea lo más veraz y ajustada a la realidad posible. El método más conocido para gestionar la captura de los datos es el web scraping, que es una técnica que permite extraer información de sitios web, así como la gestión de información a través de APIs, que permiten la comunicación entre componentes de software.
2. **Almacenamiento:** una vez procesada la captura de los datos, se procede al almacenamiento de estos. Normalmente, se hace uso de una base de datos a partir de sistemas SQL u hojas de cálculo.
3. **Tratamiento:** dependiendo del tipo de información que hayamos recogido, necesitaremos tratar los datos de una manera u otra, desde un tratamiento sencillo a un tratamiento más complejo. Se puede hacer una clasificación de los datos, buscar patrones de comportamiento, etc. En este punto es donde

el Big Data enlaza con el ML, ya que es una técnica de gestión de datos que los utiliza para hacer predicciones futuras.

4. **Puesta en valor:** los datos y las conclusiones obtenidas de estos necesitan de un análisis exhaustivo, de manera que se puedan inferir conclusiones concretas para poder aplicarlos a nuestra metodología de trabajo o para aplicar algún cambio o acometer alguna acción en algún aspecto. Podemos optar por una visualización en una gráfica, recomendación de compra de algún producto sobre otro, etc.

Todo esto no sirve más que para generar una ventaja competitiva frente al resto de empresas que hacen uso de los datos de una manera más rudimentaria y sin un análisis tan concienzudo, lo que puede llevar a conclusiones erróneas y errores en la gestión de esta. Actuar observando los datos no nos asegura no fallar en nuestra predicción, pero sí nos asegura poder actuar con mejor conocimiento de causa y, en la mayoría de los casos, con mayor acierto.

El ML y el uso de la Inteligencia Artificial se han vuelto relevantes en el mundo empresarial por los siguientes factores:

- Los procesadores informáticos se han vuelto mucho más potentes y con capacidad de gestionar, cada vez, mayor cantidad de datos.
- El coste de almacenar y manejar una gran cantidad de datos cada vez es menor, derivado de lo que se comentaba en el primer punto. La potencia de proceso de la gestión de datos ha llevado a que el procesamiento de estos sea más rápido y analítico.
- La distribución de los datos en clusters han mejorado la habilidad de los algoritmos para analizar datos complejos en tiempo récord.
- Se tiene acceso a mayor cantidad de datos a partir de servicios en la nube (APIs, Application Programming Interfaces).
- El desarrollo de los algoritmos se produce a partir de comunidades de código o fuente abierta (lo que se conoce en inglés como open-source communities), lo que hace que sea más accesible para los desarrolladores tanto los esquemas de trabajo como el acceso a las librerías propias de estos algoritmos.
- La interpretación de los resultados es sencilla y accesible incluso para las personas que no son del sector, por lo que cualquier persona en una empresa o corporación debe ser capaz de interpretarlos si están dispuestos de manera sencilla y pedagógica.

Tipos de aprendizaje:

Conocer el objetivo que se persigue mediante ML es básico para saber qué algoritmo o modelo de programación se debe seguir, ya que este hecho condicionará completamente la visión que se va a tener a la hora de analizar el problema.

Según un artículo de (Recuero de los Santos, 2017), planteado el caso de esta compañía (Telefónica) dedicada a la telefonía, se puede llegar a plantear el caso de la retención de los clientes, en cuyo caso se plantea una segmentación de los clientes. Para poder llegar a hacer una segmentación de los clientes, se debe definir la estrategia a seguir, lo cual no es trivial.

Si se busca una segmentación de los clientes, se podría plantear la pregunta de dos maneras distintas:

- "Mis clientes, ¿se agrupan de alguna manera, de forma natural?". En esta pregunta se plantea el objetivo de manera general, cuestionando si los clientes tienen alguna relación entre sí, pero con una visión generalizada de cómo se podrían agrupar estos. Esto es un ejemplo de un objetivo basado en aprendizaje no supervisado.
- ¿Podemos identificar grupos de clientes con una alta probabilidad de solicitar la baja del servicio en cuanto finalice su contrato? ¿Se dará de baja el cliente? En esta pregunta, por el contrario, el objetivo de la retención del cliente y, por consiguiente, el estudio de la solicitación de bajas es el objetivo concreto, por lo que se está ante un ejemplo de aprendizaje supervisado.

En este punto, se tratarán, a partir del ejemplo aportado por esta fuente, los tres tipos de aprendizaje aplicado en ML que se pueden utilizar teniendo en cuenta la naturaleza del problema y la disponibilidad de la variable de salida: sistemas de aprendizaje supervisado, sistemas de aprendizaje no supervisado y sistemas de aprendizaje reforzado. Teniendo en cuenta que el aprendizaje utilizado en este proyecto es el no supervisado, en un apartado posterior se comentará de forma específica este tipo de aprendizaje de forma pormenorizada.

- **Aprendizaje supervisado**

Como se infiere del ejemplo del apartado 4.1, el aprendizaje supervisado es aquel que conoce, en base a datos recogidos previamente, cual es la respuesta correcta y aprende en base a ese conocimiento previo. Por lo tanto, la principal diferencia entre este tipo de aprendizaje y el no supervisado es que el algoritmo, para entrenar, utiliza ejemplos de respuestas correctas a partir de análisis previos ya realizados. Según (Martínez Heras, 2021), el proceso para realizar análisis, explicado de manera superficial y poco pormenorizado, a partir de aprendizaje supervisado es el siguiente (figura :

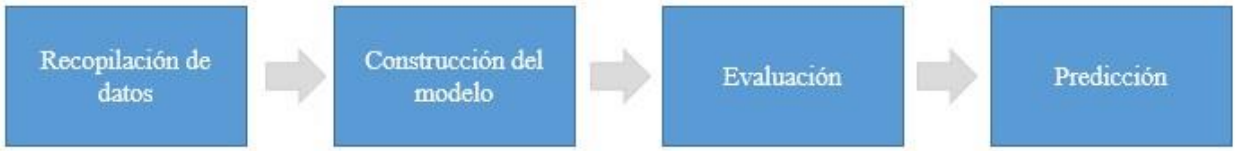

Figura 4-2: Proceso – Aprendizaje supervisado

1. Recopilación de datos históricos en los cuales se incluyan la respuesta correcta al problema.
2. Construcción de un modelo de ML a partir de dichos datos.
3. Evaluación del modelo usando datos nuevos.
4. Predicción del rendimiento que se puede obtener de dicho modelo, a partir de las respuestas correctas de los datos ofrecidos.

- **Aprendizaje no supervisado**

Según (Martínez Heras, 2021), el aprendizaje no supervisado es aquel aprendizaje que no necesita que se le diga la respuesta correcta, sino que se busca que el algoritmo aprenda de los datos que se le introducen. Esto se hace porque los datos pueden tener implicaciones futuras que no resultan obvias a la persona que está analizando el problema, de cara a tomar mejores decisiones.

El artículo aporta un caso concreto en el que un supermercado analiza los productos vendidos de forma anónima (sin conocer quién compra qué productos) y el algoritmo reseña aquellos datos que son relevantes o interesantes para resaltar. De esta manera, por ejemplo, se podrían establecer relaciones entre productos que, a priori, no tienen relación (concretamente se hace alusión a una relación entre compra de cerveza y pañales para bebés). Otros ejemplos o hábitos en los que se podría aplicar este tipo de aprendizaje puede ser la recomendación de las plataformas audiovisuales o de compras de contenidos o artículos recomendados para un cliente, segmentación de pacientes en hospitales sin conocimiento previo del criterio de reparto para dividir a estos, etc.

- **Aprendizaje reforzado**

Según (Martínez Heras, 2021), el aprendizaje por refuerzo es aprendizaje supervisado sin llegar a serlo. La diferencia entre este sistema de aprendizaje y el supervisado es que en el aprendizaje reforzado se ofrecen los datos como ejemplo sin dar la respuesta correcta.

Esta fuente aporta como ejemplo el aprendizaje de una máquina a través de ML para aprender a jugar al ajedrez. Se le enseñan las normas básicas de juego y se le deja aprender por sí mismo, de cara a poder llegar a nuevas soluciones y tácticas no utilizadas ni enseñadas de antemano.

Como la inteligencia artificial no conoce soluciones ejemplo, no puede establecer comparaciones entre acciones que ha realizado con respecto a soluciones ni ejemplos previos, por lo tanto es el aprendizaje más complicado de los descritos.

4.1 Aprendizaje no supervisado

Una vez definido, en la clasificación de aprendizajes, el aprendizaje no supervisado, se realizará una mayor profundización sobre ello, ya que va a ser el empleado en este trabajo.

El proceso que sigue el aprendizaje no supervisado es el mostrado en la figura 4-3:

Figura 4-3: Proceso – Aprendizaje no supervisado

1. Recopilación de datos históricos.
2. Entrenamiento del algoritmo en busca de patrones, grupos, etc. en los que segmentar los datos.
3. Utilizar y analizar la información obtenida de cara a poder tomar mejores decisiones.

4.1.1 Justificación de la aplicación de aprendizaje no supervisado en este estudio de Scouting

Debido a la naturaleza del problema que se está describiendo, relacionado con el scouting, lo más idóneo, a priori, es considerar un sistema de aprendizaje no supervisado, debido a la falta de concreción de nuestro objetivo. Se busca hacer una agrupación en función de las características (atributos) de los futbolistas, de manera que se puedan buscar similitudes entre unos y otros.

La idea del estudio es realizar un análisis que relacione a los jugadores de la base de datos en clusters, para poder establecer similitudes entre ellos mediante los atributos que posean. Esto permitirá conocer, a partir de las estadísticas, qué jugadores se parecen entre sí, cosa que no se sabe a priori, y que resulta muy útil a la hora de hacer estudios de mercado en bases de datos con jugadores relativamente desconocidos.

Relacionándolo con el caso estudiado (el caso del Sevilla Fútbol Club), partimos de una base de datos (que luego se tratará concienzudamente) que correspondería a la base de datos que se trata en el seguimiento

en bruto, ya que en ese período es cuando se tiene una gran cantidad de futbolistas que se necesita cribar. Esos jugadores corresponderían a lo que el club y su área deportiva han definido que son futbolistas interesantes para su estudio, después de un análisis de su rendimiento durante seis meses.

La implementación de los algoritmos utilizados de ML (PCA y Clustering) se podrían encuadrar a caballo entre el seguimiento en neto y la definición de perfiles, ya que a partir de una gran cantidad de datos y atendiendo a unas necesidades concretas definidas por el entrenador, se eligen aquellos jugadores de la base de datos que más se asemejan a lo que se busca.

4.1.2 Clustering

El clustering es un método de ordenación de datos en forma de grupos. A cada uno de los grupos que forman parte de un análisis por clustering se les denomina clusters. Para realizar un análisis por clustering, debemos organizar los datos en función de los atributos para poder, así, establecer relaciones entre ellos.

Se propone el siguiente ejemplo sencillo en el que se tiene una serie de datos representados en dos dimensiones (Martínez Heras, 2021):

Figura 4-4: Conjunto de datos antes de ser agrupados

El análisis de este conjunto de datos (o cualquier conjunto de datos en general) es, en parte, subjetivo. En función de lo que se busca, se pueden definir un número determinado de clusters, por lo que en un análisis por clustering, una persona puede visualizar unos clusters determinados en función de sus intereses, mientras que otra persona puede ver otro número de clusters. En el ejemplo aportado, tienen cabida dos posibles interpretaciones:

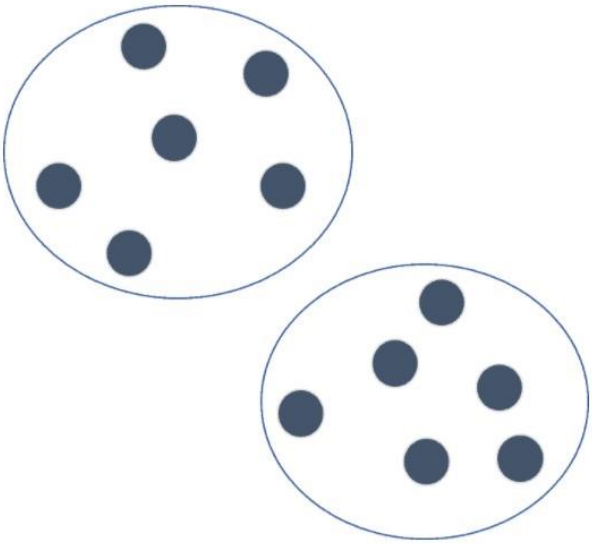

Figura 4-6: Agrupación de datos en dos clusters

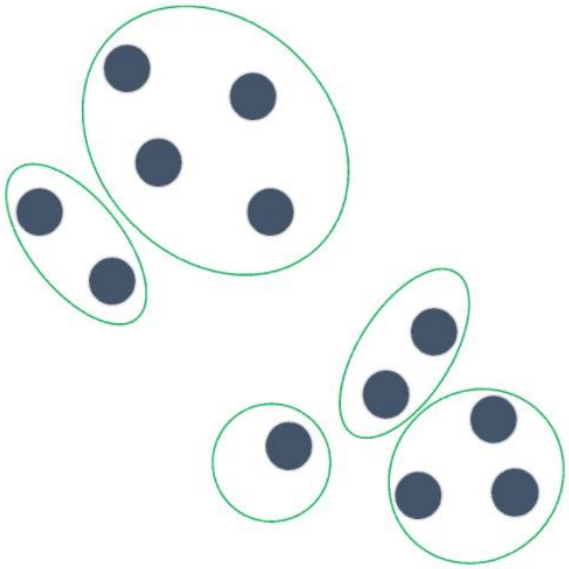

Figura 4-5: Agrupación de datos en cinco clusters

Como se puede observar en este ejemplo, cualquiera de las dos definiciones es válida, debido a la no restricción del número de clusters por el análisis por clustering. Dependiendo de la intencionalidad de la persona a la hora de realizar dicho análisis, se definirá un número de clusters u otro.

Una manera de representar los clusters más sencilla, de forma analítica, es por medio de la definición de centroides (siempre que los atributos sean numéricos). Un centroide, que es el centro geométrico del cluster, está definido con la media de cada uno de los atributos. Por ejemplo, si se identifican cuatro datos con los siguientes atributos: (3,7), (11,10), (4,2), (1,9), el centroide se encontrará en el punto:

- Media aritmética del primer atributo: $\frac{3+11+4+1}{4} = \frac{19}{4} = 4{,}75$
- Media aritmética del segundo atributo: $\frac{7+10+2+9}{4} = \frac{28}{4} = 7$

Por tanto, el centroide en este caso, definiendo un solo cluster para todos los datos, se encuentra en el punto: (4,75;7). Si los atributos no son numéricos, una idea podría ser definir una variable binaria que sea falso o verdadero en función de un criterio definido previamente, para poder obtener una variable numérica a partir de una variable discreta.

Para el caso en el que se tengan más centroides, para poder definir a qué cluster pertenece un determinado dato, la manera más lógica es saber qué cluster está más cerca de cada uno de los datos a analizar. Esto se puede determinar a partir de la distancia euclídea.

Por tanto, para el ejemplo anterior, sabiendo la posición del centroide y la posición de los datos en función de los atributos:

- Distancia del punto 1 (3,7) al centroide: $d_1 = \sqrt{(3-4{,}75)^2 + (7-7)^2} = 1{,}75$
- Distancia del punto 2 (11,10) al centroide: $d_2 = \sqrt{(11-4{,}75)^2 + (10-7)^2} = 6{,}93$
- Distancia del punto 3 (4,2) al centroide: $d_3 = \sqrt{(4-4{,}75)^2 + (2-7)^2} = 5{,}06$
- Distancia del punto 4 (1,9) al centroide: $d_4 = \sqrt{(1-4{,}75)^2 + (9-7)^2} = 16{,}43$

En este caso se tiene un único centroide, pero si la toma de decisión fuese entre varios centroides, la lógica sería calcular la distancia de cada punto a cada uno de los centroides asociados a cada cluster. La menor distancia entre el punto que se considera y el centroide asociado a un cluster es la que define el cluster al que pertenece dicho punto.

Esto ocurre en un análisis en dos dimensiones (para dos atributos), pero si la base de datos de la que se dispone tiene una cantidad de atributos considerable, no se puede realizar una representación gráfica clara de esta. Por tanto, se necesitan algoritmos que definan ese número de clusters de una manera analítica. El principal algoritmo en este respecto es el k-Means.

4.1.2.1 El algoritmo k-Means.

Según (Kubat, 2017), k-Means es el algoritmo más simple para detectar clusters. Concretamente, la "k" del nombre del algoritmo hace referencia al número de clusters necesarios, que es un parámetro que utiliza el usuario.

La estructura de funcionamiento del algoritmo k-Means es la siguiente:

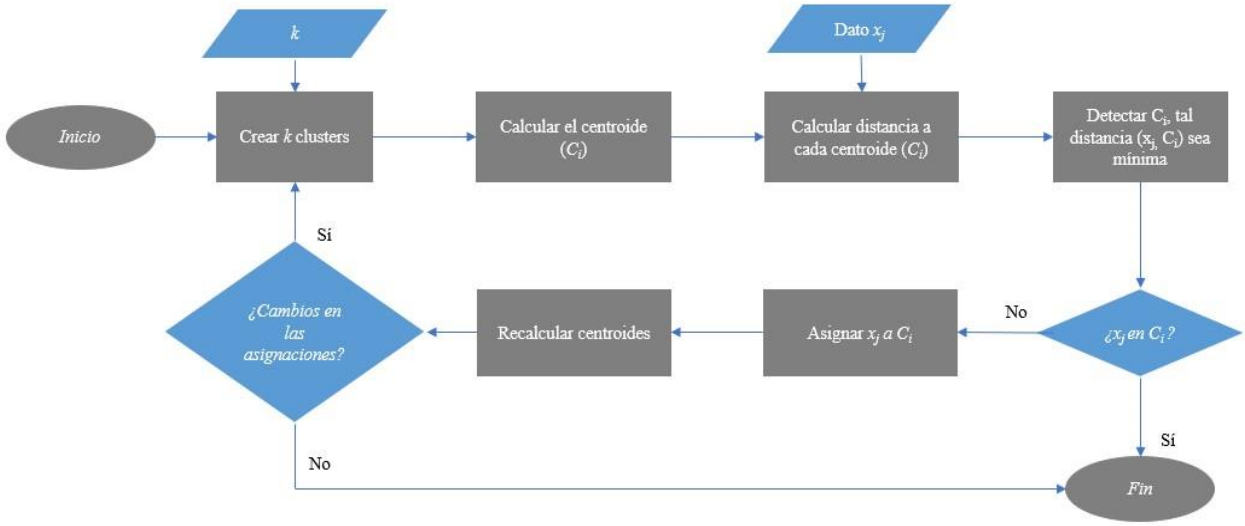

Figura 4-7: Proceso – Algoritmo k-Means

Se aporta un ejemplo gráfico que hace más pedagógica esta explicación del algoritmo:

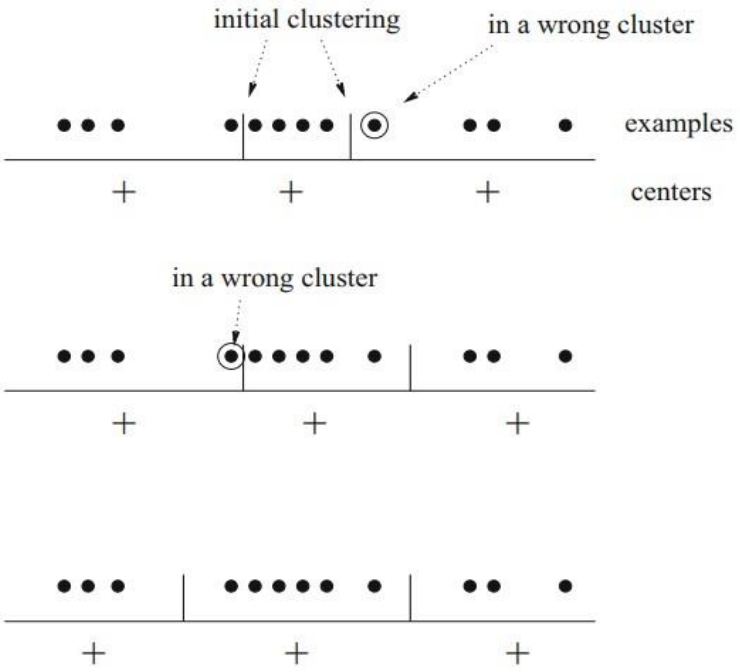

Figura 4-8: Ejemplo gráfico algoritmo k-Means (Kubat, 2017)

En este ejemplo se observa cómo se realiza, primeramente, un reparto de los datos en tres clusters, que es el clustering inicial propuesto por el algoritmo. Seguidamente, se puede ver que uno de los datos no está en el cluster correcto, ya que, por cercanía, debería estar en el contiguo, por lo que el algoritmo en la siguiente iteración corregirá la posición hasta llevarlo al cluster cuyo centroide está más cercano. El algoritmo seguirá realizando iteraciones hasta que todos los datos estén en el cluster que le corresponda.

En el algoritmo k-means es importante:

- Normalización de atributos: una diferencia grande entre atributos hace que se desvirtúe el análisis. Por lo tanto, una opción interesante sería reescalar cada uno de los valores teniendo en cuenta el máximo y mínimo de cada atributo, de la manera siguiente:

$$x_{nuevo} = \frac{x - MAX}{MAX - MIN}$$

- Inicialización: es relevante puesto que el algoritmo, para entrenar y poder llegar a la distribución final de clusters, necesita realizar iteraciones para poder hacer cambios en la disposición inicial de los datos en cada uno de los clusters. Cuanto mejor sea el reparto de datos en la inicialización, más rápido actuará el algoritmo y más rápidamente se encontrará la mejor solución. En el caso de que no sea fácil definir el criterio para realizar la inicialización, porque no haya un atributo claro para

acometerla, se cogen *k* ejemplos y se consideran vectores de código para definir los centroides, de manera que los ejemplos se asocian a cada uno de los clusters definidos a partir de cada centroide.

4.1.2.2 Método del codo

Según (Gonzalo, 2019), el método del codo es un método aplicado para determinar, de una forma relativamente estadística, el número de clusters idóneo para el análisis que se esté realizando.

Con este método, los principales objetivos son dos:

1. Minimizar la varianza dentro de cada uno de los clusters.
2. Maximizar la varianza entre cada uno de los clusters.

Para medir estas varianzas (en términos de distancia), se necesitan definir los centroides de cada uno de los clusters, pues es esta distancia (la de cada uno de los datos al centroide) la que va a definir el número de clusters óptimo para el análisis, de acuerdo con los objetivos definidos previamente. Si el número de clusters aumenta, la varianza dentro de cada cluster tiende a disminuir. Cuanto menor sea esta distancia, mejor será el análisis, ya que el cluster es más compacto.

Aplicando el método del codo para el caso de estudio del presente proyecto, se obtiene este resultado:

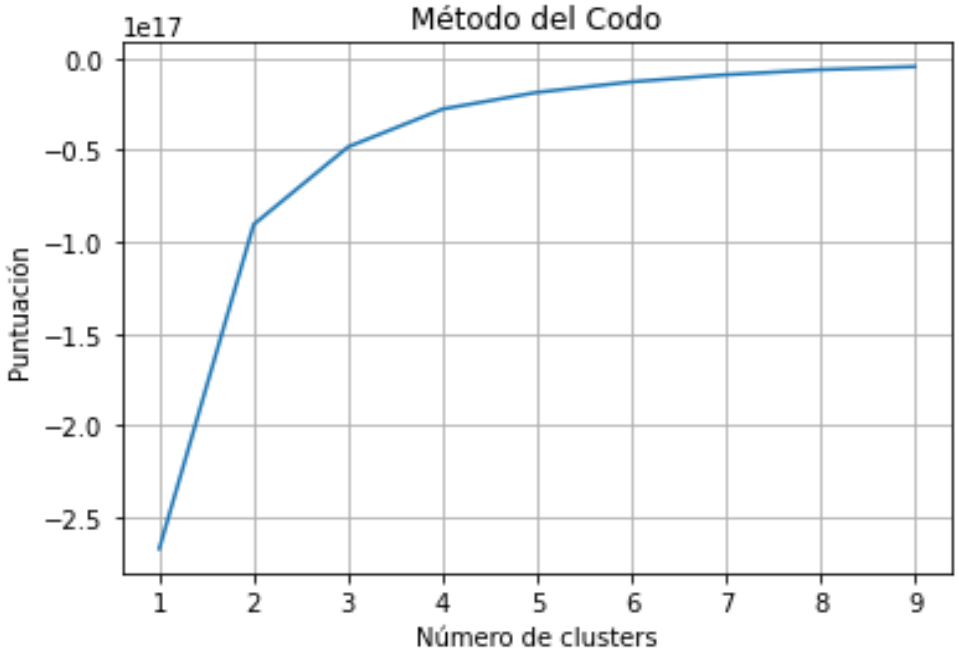

Figura 4-9: Método del codo

La manera óptima de definir el número de clusters idóneo para cualquier estudio mediante este método es ciertamente arbitrario, ya que depende del criterio del usuario el escoger un número de clusters determinado u otro, pues es una cuestión subjetiva.

Representando el rango de 1 cluster a 10 clusters, se obtiene esta gráfica. Como se puede observar, en este caso, el número de clusters óptimo se podría establecer en 5, ya que se observa que un mayor número de clusters no repercute en una disminución significativa de la varianza inter-cluster (dentro del cluster).

Debido a la ambigüedad de este análisis, se opta por utilizar otro método para complementar y objetivar de mayor manera el método del codo. Este método se denomina el análisis de la silueta.

4.1.2.3 Análisis de la silueta

Según (Gonzalo, 2019), el análisis de la silueta realiza una medición de la calidad del clustering a través de la distancia de separación entre clusters (distancia intra-cluster). De esta manera se puede calcular la distancia de cada uno de los puntos a los clusters vecinos. Esta distancia se mide en el rango [-1,1], de forma que:

- Un valor cercano a 1 indica que el punto se encuentra alejado del resto de clusters.
- Un valor cercano a 0 indica que el punto se encuentra cercano o incluso en la frontera entre dos clusters.
- Un valor cercano a -1 indica que el punto está ubicado en un cluster erróneo.

Este método calcula la media de los coeficientes de silueta de todas las observaciones para un número distinto de clusters. El número óptimo de clusters a calcular sería aquel valor que maximice esta media.

El coeficiente de silueta se calcula tal que:

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

Siendo a el valor de la distancia media entre clusters y b la distancia media al cluster más cercano.

4.1.3 Principal Component Analysis (PCA)

De acuerdo con la información extraída de (Uc3m, n.d.), en muchas ocasiones, a la hora de trabajar con bases de datos, se recogen muchas variables, regularmente sin atender a lo relacionada o no que estén algunas de las variables con respecto a otras. De hecho, es bastante común que muchas de las variables recogidas, sin saberlo de antemano, estén correlacionadas. Esto hace que dichas variables, a la hora de realizar un análisis exhaustivo de la base de datos, no representen un hecho diferencial con respecto al resto, ya que no aportan información adicional no conocida previamente, al ser las variables no relacionadas las que nos aportan una mayor información.

Este artículo cita un ejemplo médico de este suceso de correlación de variables, que es la fuerte relación entre la presión sanguínea a la salida del corazón y a la salida de los pulmones. Por tanto, una de las dos variables puede ser obviada en el estudio de los datos por la razón comentada previamente: no representan independencia del resto de variables, por lo que pierden utilidad. Es por ello por lo que es vital realizar un filtrado de datos, y, por consiguiente, una reducción del número de variables, en el estudio de la base de datos.

El análisis de componentes principales (PCA por las siglas en inglés de Principal Component Analysis) es una técnica desarrollada por Pearson a finales del siglo XIX y, unos años después, por Hotteling en los años 30 (del siglo XX). No obstante, el auge de esta herramienta ocurrió con la expansión del uso de ordenadores.

Esta técnica considera p variables interrelacionadas dentro del conjunto total de atributos que tenemos inicialmente en la base de datos. El objetivo de esta técnica es redimensionar ese conjunto en uno nuevo al que se le conoce como conjunto de componentes principales. Este conjunto, y sus variables, se constituyen a partir de combinaciones lineales de las variables que presentan una correlación alta entre sí, así como de las variables que no recojan una gran variabilidad.

El cálculo de las componentes principales sigue esta estructura:

1. Se consideran tres agrupaciones de variables: un conjunto inicial $x = (x_1, x_2, \ldots, x_p)$, un nuevo conjunto de variables $y = (y_1, y_2, \ldots, y_p)$, que no presentan correlación entre sí, siendo cada una de las variables de dicho conjunto combinación lineal de las del conjunto inicial, y el vector de constantes $a_j' = (a_{j1}, a_{j2}, \ldots, a_{jp})$. Teniendo en cuenta estos tres conjuntos de variables se tiene que:

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jp}x_p = a_j'x$$

   Como el hecho de aumentar el valor de las componentes del vector de constantes concurriría en una maximización de la varianza, restringimos el módulo del vector de constantes a 1.

El primer componente se calcula eligiendo a_1 (autovector correspondiente a la primera componente del vector de constantes) de tal forma que la componente y_1 tenga la mayor varianza posible. De igual forma se calcula y_2 (sin tener correlación con y_1) a partir de a_2, y así sucesivamente hasta calcular las y_p variables incorreladas entre sí de mayor a menor varianza.

Para realizar la elección de a_1, el método más común es utilizar el método de los multiplicadores de Lagrange, de tal manera que:

$$Var(y_1) = Var(a_1'x) = a_1' \sum a_1 = a_1'\lambda_1 I a_1 = \lambda_1 a_1' a_1 = \lambda_1$$

Siendo λ: el multiplicador de Lagrange.

Para calcular la segunda componente y_2, el procedimiento es análogo al anterior, simplemente teniendo en cuenta que: $Cov(y_2, y_1) = 0$. El resto de las componentes se calculan de igual manera, obteniéndose, entonces: $Var(y_j) = \lambda_j$.

2. Una vez calculadas todas las varianzas de las *p* variables, se debe tener en cuenta el porcentaje de varianza recogido por cada una de las componentes principales con respecto al total. Calculados todos los porcentajes, se procede a escoger aquellas variables que representen un gran porcentaje de varianza con respecto al total, de manera que, con una dimensión menor del problema, se recoja una varianza considerable.

4.2 Métricas estadísticas para el análisis de datos

A continuación, se definen diferentes elementos que se utilizan para realizar análisis estadísticos de los datos. A la hora de realizar dicha definición, se ha tenido en cuenta que en el caso de estudio que se está analizando, muchas de las variables observadas son discretas, al constituir su naturaleza en un conjunto finito de valores y, además, no siguen una distribución de acuerdo con una función definida (variable aleatoria continua).

- **Esperanza matemática**

Según (Gutiérrez Moya, 2016), teniendo una variable aleatoria real *X* y una función de variable real *g(X)*, se puede definir la esperanza matemática como:

$$E[g(X)] = \sum_{x_i \in \mathbb{R}} g(x_i)p(x_i)$$

La esperanza matemática representa el valor promedio esperable que toma una variable.

- **Media**

Según (Gutiérrez Moya, 2016), un caso particular de la esperanza matemática es la media, que es el valor promedio de una variable aleatoria. Se podría definir a partir de la variable aleatoria discreta *X*, de tal manera que:

$$\mu = E[X] = \sum_{x_i \in \mathbb{R}} x_i p_i$$

- **Varianza y desviación típica**

Según (Gutiérrez Moya, 2016), a partir del concepto de esperanza matemática se puede definir la varianza de la función de variable real *g(X)* como:

$$Var[g(X)] = E[g(X) - E\{g(X)\}]^2 = E[g(X)^2] - [E\{g(X)\}]^2$$

La desviación típica o estándar de *g(X)* es la raíz cuadrada positiva de la varianza:

$$\sigma_{g(X)} = \sqrt{Var\ g(X)}$$

Ambos parámetros son medidas de dispersión que representan la variabilidad de un conjunto de datos con respecto a la media aritmética de los mismos.

- **Cuantiles**

Según (Gutiérrez Moya, 2016), el cuantil *100*p (%)* de la variable aleatoria X es el menor valor x_p que cumple que $F(x_p) \geq p$. Si la variable aleatoria es discreta, se debe cumplir:

$$\begin{cases} F(x_p^-) = P(X < x_p) \leq p \\ F(x_p) = P(X < x_p) \geq p \end{cases} \qquad p \in (0,1)$$

Los cuantiles más utilizados en este estudio son los cuartiles. El cuartil es el valor de una variable que deja el 25% de las muestras a su izquierda, o, lo que es lo mismo, el 25% de los datos toman valores de dicha variable menores a dicho cuartil. Se definen como:

$$\text{Cuartiles } (Q_i) \begin{cases} F(Q_i^-) = P(X < Q_i) \leq \frac{i}{4} \\ F(Q_i^-) = P(X < Q_i) \geq \frac{i}{4} \end{cases} \qquad i = 1,2,3$$

- **Covarianza y coeficiente de correlación**

Según (Gutiérrez Moya, 2016), a partir de una variable aleatoria bidimensional *(X,Y)* discreta se puede definir la covarianza de dicha variable como:

$$\sigma_{XY} = E[(X - \mu_X) - (Y - \mu_Y)] = \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y)p(x_i - y_j)$$

A su vez, se podría definir, de igual manera a partir de una variable aleatoria bidimensional *(X,Y)*, el coeficiente de correlación de X sobre Y como:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - E(X)) - (Y - E(Y))]}{\sqrt{V(X)V(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{V(X)V(Y)}}$$

- Como luego se especificará en el apartado referente a la matriz de correlación, el valor del coeficiente de correlación está acotado, de tal manera que: $\rho_{XY} \in [-1,1]$

- **Matriz de correlación**

Según (*Interpretar Todos Los Estadísticos y Gráficas Para Análisis de Elementos - Minitab*, n.d.), una matriz de correlación es un análisis de las variables del caso de estudio para poder calcular los valores de correlación de Pearson. Estos valores miden el grado de linealidad entre todas las variables, unas con respecto a las otras. Dichos valores pueden ir desde -1 hasta +1, dependiendo del grado de correlación entre dichas variables.

De tal manera que se tienen los siguientes posibles casos:

- Valor de correlación alto y positivo: las variables asociadas a dicho valor miden la misma característica. La condición para considerar una correlación como valor alto es, por lo general,

un valor mayor que 0.7, pero dependiendo del caso de estudio, este valor puede variar hacia arriba o hacia abajo.

- Valor de correlación alto y negativo: las variables asociadas a dicho valor están relacionadas en sentido inverso (el aumento en valor de una variable propicia una disminución de la otra, y viceversa).
- Valor de correlación bajo: las variables están escasamente interrelacionadas.

5 BASE DE DATOS

Ante la falta de disponibilidad de bases de datos ajustadas a lo que se requería, se ha optado por realizar una base de datos manualmente. Esta base de datos se ha confeccionado a partir de los datos de una aplicación y página web de resultados y datos de fútbol llamada Sofascore (*SofaScore: The Fastest Football Scores and Livescore for 2021*, n.d.)

En una situación real, en un club de fútbol, la base de datos se obtendría a partir de alguna aplicación, servidor o página web de pago (Wyscout, es la más conocida y usada en el mundo futbolístico) de la cual esté haciendo uso la entidad. De tal manera, el proceso de exportar los datos de su origen a la base de datos con la que se va a trabajar es un proceso arduo y repetitivo, que no aporta ningún valor añadido.

Con respecto a la base de datos en sí, se ha optado por analizar 600 jugadores de las siguientes ligas europeas:

- Primera División (España)
- Segunda División (España)
- Premier League (Inglaterra)
- Ligue 1 (Francia)
- Serie A (Italia)
- Bundesliga (Alemania)
- Eredivisie (Países Bajos)

Figura 5-2: Ligas europeas utilizadas para la base de datos (ESPN, 2018)

Figura 5-1: Segunda División (*Segunda División de España - Wikipedia, La Enciclopedia Libre*, n.d.)

La base de datos de 600 jugadores cuenta con datos obtenidos a partir de la aplicación Sofascore, como se ha comentado previamente en la descripción. Esta aplicación, además de ser gratuita, cuenta con una gran

cantidad de datos tanto en lo que respecta a atributos de cada uno de los futbolistas, como en lo referente a variedad de ligas y países.

En este trabajo, los atributos se han clasificado en los siguientes grupos:

- Grupo 1: Partidos
- Grupo 2: Estadísticas
- Grupo 3: Ataque
- Grupo 4: Pases
- Grupo 5: Habilidad
- Grupo 6: Defensa
- Grupo 7: Disciplina
- Grupo 8: Valoración
- Grupo 9: Portería

Para cada uno de los atributos grupos se tienen los siguientes atributos:

Grupo 1: PARTIDOS	Grupo 2: ESTADÍSTICAS	Grupo 3: ATAQUE
Liga	Posición	Goles totales
Equipo	Nacionalidad	Frecuencia de goles
Partidos jugados	Altura	Goles por partido
Minutos	Edad	Disparos por partido
Promedio minutos	Polivalencia	Asistencias
Minutos jugados por el equipo	Segunda posición preferida	
Minutos jugados sobre el total del equipo (%)	Pie preferido	

Tabla 1: Definición de atributos relativos a los grupos 1, 2 y 3

Grupo 4: PASES	Grupo 5: HABILIDAD	Grupo 6: DEFENSA
Toques por partido	Regates por partido	Penaltis cometidos
Ocasiones creadas	Regates por partido (%)	Porterías a cero
Pases clave por partido	Duelos ganados por partido	Posesión perdida
Completados (%)	Duelos ganados por partido (%)	Intercepciones por partido
Completados en campo propio		Entradas por partido
Completados en campo propio (%)		Posesión recuperada
Completados en campo contrario		Regateado por partido
Completados en campo contrario (%)		Despejes por partido
Balones largos por partido		
Balones largos completados (%)		
Centros completados		
Centros completados por partido (%)		

Tabla 2: Definición de atributos relativos a los grupos 4, 5 y 6

Grupo 7: DISCIPLINA	Grupo 8: VALORACIÓN	Grupo 9: GOLES RECIBIDOS	Grupo 10: PARADAS
Faltas cometidas por partido	Paradas	Goles concedidos por partido	Paradas realizadas
Faltas recibidas por partido	Anticipación	Penaltis en contra	Goles/Paradas (%)
Tarjetas amarillas	Táctico	Penaltis parados	Paradas por partido
Tarjetas rojas	Distribución de balón	Penaltis parados (%)	Paradas por partido (%)
	Juego aéreo		Salidas por partido
	Ataque		Paradas con balón atrapado
	Técnica		Paradas con despeje
	Defensa		
	Creatividad		
	Valor de mercado		

Tabla 3: Definición de atributos relativos a los grupos 7, 8, 9 y 10

- **Definición de variables**

Para clarificar el sentido de cada atributo, se procede a realizar una breve explicación de cada uno de ellos. Además, para complementar la división que se ha realizado previamente en función de la naturaleza de cada uno de los atributos, se va a añadir una segunda división basada en otro criterio de reparto, para observar si la variable depende del período de estudio en el que se recogen los datos o no. Para esta división tendremos, entonces, dos tipos de variables: estáticas o dinámicas.

Concretamente, la base de datos realizada se ha construido a partir de datos recogidos en el período septiembre 2020 – marzo 2021.

➢ **Variables estáticas:** son aquellas que no dependen del período de estudio, por lo que son características intrínsecas a cada uno de los futbolistas.

- **Posición:** rol que desempeña un jugador en el terreno de juego. Variable categórica. En este análisis (como se puede observar en la siguiente figura) se tienen 10 posiciones posibles:

1. Portero
2. Lateral derecho
3. Lateral izquierdo
4. Central
5. Central
6. Pivote
7. Extremo derecho
8. Mediocentro
9. Delantero
10. Mediapunta
11. Extremo izquierdo

Figura 5-3: Sistema de juego referencia para definir la variable posición en la BBDD (Herráez, n.d.)

- **Nacionalidad:** país del que es originario el futbolista. Variable categórica.
- **Altura:** dimensión vertical de un jugador. Variable numérica (unidad: centímetros).
- **Polivalencia:** número posiciones distintas en las que puede jugar un futbolista aparte de su posición natural (la definida en la variable posición). Cuanto mayor sea esta variable, más polivalente es el futbolista. Variable numérica (unidad: adimensional).
- **Segunda Posición Preferida:** a partir de las posiciones definidas en el atributo 'Polivalencia' en las que el futbolista puede jugar, aquella en la que más cómodo se siente y más está acostumbrado a jugar sin ser su posición natural. Variable categórica.
- **Pie preferido:** pie con el que el jugador tiene una mayor habilidad. Variable categórica.
- **Paradas:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un portero (medidor exclusivamente dedicado a esta demarcación) a la hora de atajar un tiro rival (parada). Variable numérica (unidad: adimensional).
- **Anticipación:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un portero (medidor exclusivamente dedicado a esta demarcación) respecto a su capacidad de predecir hacia dónde va a ir el balón o, incluso, pronosticar cuándo tiene que salir del área para despejar un balón. Variable numérica (unidad: adimensional).
- **Táctico:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un futbolista respecto a su capacidad para posicionarse correctamente en el campo, siendo capaz de adivinar el devenir de la jugada y situarse en consecuencia. Variable numérica (unidad: adimensional).

- **Distribución de balón:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un portero (medidor exclusivamente dedicado a esta demarcación) a la hora de realizar pases en corto y/o en largo y, así, poder ayudar al juego del equipo como si fuese un jugador de campo. Variable numérica (unidad: adimensional).
- **Juego aéreo:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un portero (medidor exclusivamente dedicado a esta demarcación) respecto a su capacidad de intervenir una jugada por arriba. Entiéndase por jugada por arriba una jugada que no transcurre con el balón por el césped, por ejemplo, centros laterales, córners, faltas, etc. Variable numérica (unidad: adimensional).
- **Ataque:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un futbolista (exceptuando porteros) respecto a su capacidad atacante. Variable numérica (unidad: adimensional).
- **Técnica:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un futbolista (exceptuando porteros) respecto a su habilidad con balón e, incluso, sin él. Variable numérica (unidad: adimensional).
- **Defensa:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel en el aspecto defensivo de un futbolista (exceptuando porteros). Variable numérica (unidad: adimensional).
- **Creatividad:** valoración propia de la aplicación Sofascore basada en el rendimiento del jugador durante los dos últimos años que sirve para medir el nivel de un futbolista (exceptuando porteros) de encontrar una solución inesperada para el rival. Variable numérica (unidad: adimensional).
- **Valor de mercado:** valor (estimación) en euros de cada futbolista según la aplicación Sofascore. Variable numérica (unidad: euros, €).

➢ **Variables dinámicas:** son aquellas que dependen del período de estudio.

- **Partidos jugados:** número de partidos en los que el futbolista participa, ya sea como titular o saliendo del banquillo. Variable numérica (unidad: adimensional).
- **Promedio minutos:** número de minutos que juega un futbolista durante un partido. Por ejemplo, si el futbolista en todos los partidos que ha jugado ha sido titular y ha jugado el partido completo, el promedio de minutos de este será 90 minutos. Variable numérica (unidad: minutos).
- **Minutos:** producto entre partidos jugados y promedio minutos. Se obtiene el número total de minutos que juega el futbolista. Variable numérica (unidad: minutos).
- **Minutos jugados por el equipo:** número de partidos jugados por el equipo en el que juega un determinado futbolista multiplicado por 90 minutos que tiene un partido. Nos indica el número

total de minutos que ha jugado el equipo desde el principio de liga hasta el momento en el que se recogen los datos. Variable numérica (unidad: minutos).

- **Minutos jugados sobre el total del equipo:** porcentaje de minutos que juega un futbolista sobre el total de minutos que ha jugado el equipo. El fin de este índice es poder analizar la participación de un futbolista durante la temporada. Variable numérica (unidad: adimensional).
- **Goles totales:** goles que marca un futbolista durante un determinado período de estudio. Variable numérica (unidad: adimensional).
- **Frecuencia de goles:** ratio que relaciona la cantidad de goles anotados (goles totales) y los minutos jugados por el jugador. Variable numérica (unidad: adimensional).
- **Goles por partido:** ratio que relaciona la cantidad de goles anotados (goles totales) y los partidos jugados. Variable numérica (unidad: adimensional).
- **Disparos por partido:** ratio que relaciona la cantidad de disparos y los partidos jugados. Variable numérica (unidad: adimensional).
- **Asistencias:** variable que contabiliza el número de veces que un jugador da un pase que precede a un gol de un jugador de su equipo. Variable numérica (unidad: adimensional).
- **Toques por partido:** número de veces que un futbolista toca el balón de media por partido. Variable numérica (unidad: adimensional).
- **Ocasiones creadas:** variable que recoge la cantidad de pases que un futbolista da que precede a un tiro de un compañero. La diferencia con respecto a la variable 'Asistencias' es que, para que se contabilice una asistencia, se necesita que el pase del jugador propicie un gol, mientras que la ocasión creada recoge un pase que suceda a un tiro. Variable numérica (unidad: adimensional).
- **Pases clave por partido:** ratio que relaciona las ocasiones creadas con el número total de partidos jugados. Variable numérica (unidad: adimensional).
- **Completados (%):** porcentaje de pases que acaban en un jugador del equipo sobre el total de pases que se realiza. Variable numérica (unidad: adimensional).
- **Completados en campo propio:** número de pases acertados que se dan en la mitad del campo donde está el portero de tu equipo. Estos pases se caracterizan por ser de mayor facilidad, ya que el número de jugadores rivales que pueden propiciar un error propio es menor. Variable numérica (unidad: adimensional).
- **Completados en campo propio (%):** porcentaje de pases en campo propio acertados sobre el total de pases completados en campo propio. Variable numérica (unidad: adimensional).
- **Completados en campo contrario:** número de pases acertados que se dan en la mitad del campo donde se encuentra el portero del equipo rival. Estos pases se caracterizan, por el contrario, por ser de una dificultad técnica mayor, debido a la gran cantidad de jugadores rivales que pretenden

realizar un robo de balón cuando la posesión de este la tiene el rival. Variable numérica (unidad: adimensional).

- **Completados en campo contrario (%):** porcentaje de pases en campo contrario acertados sobre el total de pases en campo contrario. Variable numérica (unidad: adimensional).
- **Balones largos por partido:** ratio que relaciona el número de pases a gran distancia que se realizan con respecto al número total de partidos jugados. Variable numérica (unidad: adimensional).
- **Balones largos completados (%):** porcentaje de pases de larga distancia completados sobre el total de intentos. Variable numérica (unidad: adimensional).
- **Centros completados:** pase o lanzamiento acertado del balón desde el lateral del campo, normalmente, hacia el área contraria. Variable numérica (unidad: adimensional).
- **Centros completados por partido (%):** porcentaje de acierto basado en la ratio de centros completados relativo al número de partidos que ha jugado un futbolista. Variable numérica (unidad: adimensional).
- **Regates por partido:** movimiento que realiza un futbolista gracias a su habilidad para sortear a un rival sin que este le robe el balón. Esta variable es una ratio que relaciona el número de regates que se realiza con respecto al total de partidos jugados Variable numérica (unidad: adimensional).
- **Regates por partido (%):** porcentaje de acierto de un futbolista con respecto al total de regates intentados por partido. Variable numérica (unidad: adimensional).
- **Duelos ganados por partido:** un duelo es una disputa de balón con un contrario en la cual no está definido quién tiene la posesión de balón. Por tanto, esta ratio define el número de duelos ganados respecto al número total de partidos disputados por el jugador. Variable numérica (unidad: adimensional).
- **Duelos ganados por partido (%):** porcentaje de balones ganados mediante duelos con respecto al total de duelos en los que un futbolista se ve inmerso. También tiene en cuenta el número total de partidos jugados, de igual forma que la variable anterior. Variable numérica (unidad: adimensional).
- **Penaltis cometidos:** número de infracciones o faltas que comete un futbolista dentro del área, produciendo un penalti en contra para su equipo. Variable numérica (unidad: adimensional).
- **Porterías a cero:** cantidad de veces que un equipo no encaja gol en un partido. Variable numérica (unidad: adimensional).
- **Posesión perdida:** media de ocasiones por partido en las que un futbolista pierde el balón cuando lo tiene en su poder. Variable numérica (unidad: adimensional).

- **Intercepciones por partido:** número de veces sobre el total de partidos jugados en las que un futbolista roba el balón al contrario cuando este ha realizado un pase a un compañero suyo. Variable numérica (unidad: adimensional).
- **Entradas por partido:** número de veces por partido que un futbolista roba el balón al rival cuando lo tiene en su posesión. Variable numérica (unidad: adimensional).
- **Posesión recuperada:** número de veces por partido que un jugador roba el balón en el tercio del campo más alejado de su portería. Variable numérica (unidad: adimensional).
- **Despejes por partido:** un despeje es un lanzamiento del balón por parte de un futbolista con el fin de alejar lo máximo posible el mismo de su portería. Esta ratio relaciona el número de veces que un jugador acomete esta acción sobre el total de partidos jugados. Variable numérica (unidad: adimensional).
- **Faltas cometidas por partido:** número de infracciones o entradas por partido que realiza un futbolista a un rival golpeándole a él en vez de al balón. Variable numérica (unidad: adimensional).
- **Faltas recibidas por partido:** número de infracciones o entradas por partido que realiza el rival sobre el futbolista golpeándole a él en vez de al balón. Variable numérica (unidad: adimensional).
- **Tarjetas amarillas:** número de amonestaciones que ha realizado el árbitro a un futbolista en el período de estudio. Variable numérica (unidad: adimensional).
- **Tarjetas rojas:** número de expulsiones que ha realizado el árbitro a un futbolista en el período de estudio. Variable numérica (unidad: adimensional).
- **Goles totales en contra:** número de goles encajados (variable exclusiva para porteros). Variable numérica (unidad: adimensional).
- **Goles concedidos por partido:** ratio que recoge el número total de goles encajados respecto al número de partidos jugados (variable exclusiva para porteros). Variable numérica (unidad: adimensional).
- **Penaltis en contra:** número de penaltis que ha cometido el equipo en el que juega un determinado portero. Variable numérica (unidad: adimensional).
- **Penaltis parados:** número de penaltis que ha parado el portero. Variable numérica (unidad: adimensional).
- **Penaltis parados (%):** ratio entre las dos variables anteriores, o lo que es lo mismo, entre número de penaltis parados y el total de penaltis cometidos por el equipo. Esta ratio refleja el acierto porcentual de los porteros a la hora de atajar penaltis. Variable numérica (unidad: adimensional).
- **Paradas realizadas:** número de paradas totales que ha realizado un portero durante el período de estudio. Variable numérica (unidad: adimensional).

- **Goles/Paradas (%):** ratio que relaciona el número de goles totales recibidos y el número total de paradas, de manera que se indica el porcentaje de goles que encaja un portero respecto al total de paradas que realiza. Variable numérica (unidad: adimensional).
- **Paradas por partido:** ratio que relaciona el número de paradas realizadas dependiendo del total de partidos. Variable numérica (unidad: adimensional).
- **Paradas por partido (%):** porcentaje que refleja el acierto en el número de paradas con respecto al número de tiros que recibe un portero. Variable numérica (unidad: adimensional).
- **Salidas por partido:** número de veces por partido que un portero sale del área para interceptar el balón. Variable numérica (unidad: adimensional).
- **Paradas con balón atrapado:** número total de paradas que realiza un portero atajando el balón sin dejar que se escape. Variable numérica (unidad: adimensional).
- **Paradas con despeje:** número total de paradas que realiza un portero sin atajar el balón, dejando que se escape. Variable numérica (unidad: adimensional).

5.1 Análisis de los datos mediante gráficas

En este apartado se recogen las gráficas obtenidas a partir del software de programación Python (que se tratará después). Estas gráficas se han realizado para cuatro posiciones distintas: portero, central, mediocentro y delantero. Para la realización de éstas, se han tenido en cuenta las variables más representativas en cada una de las posiciones, para poder acometer un análisis visual y de forma preliminar de cuáles son los valores esperados en cada uno de los casos.

Antes de visualizar las gráficas, debemos tener en cuenta que no hay definido un criterio estadístico de cuál es el valor mínimo aceptable de cada una de las variables para realizar un filtrado sobre todas las posibilidades, sino que depende de la inversión a realizar por el club comprador, por lo que, a mayor gasto, mejores deben ser los datos del jugador a fichar.

En cada una de las posiciones se han utilizado tres tipos de gráficas:

- **Scatterplot (2D):** son gráficas de dispersión bidimensionales cuya función básica es representar dos variables y ver la relación, entre ellas, de manera que se puedan observar relaciones lineales, no lineales, incorrelaciones, etc., así como visualizar tendencias de crecimiento o decrecimiento de los datos.
- **Histograma:** son gráficos de barras representados en intervalos que muestran la frecuencia de los datos para determinados valores. A partir del histograma se puede visualizar una distribución de los datos, de manera que se puede determinar cuáles son los valores más repetidos de una variable concreta.

- **Scatterplot (3D):** son gráficos de dispersión tridimensionales y, al igual que en el caso bidimensional, su principal función es relacionar variables a partir de datos recogidos en una muestra.

5.1.1 Gráficas relativas a la posición portero

En la gráfica de la figura 5-4 se enfrentan las variables: 'Paradas realizadas' (eje x) contra 'Goles totales en contra' (eje y). Temiendo en cuenta la naturaleza de estos atributos, buscaremos aquel portero que se encuentre más cercano a la esquina inferior derecha en la gráfica, que implica un número menor de goles totales en contra para un mayor número de paradas.

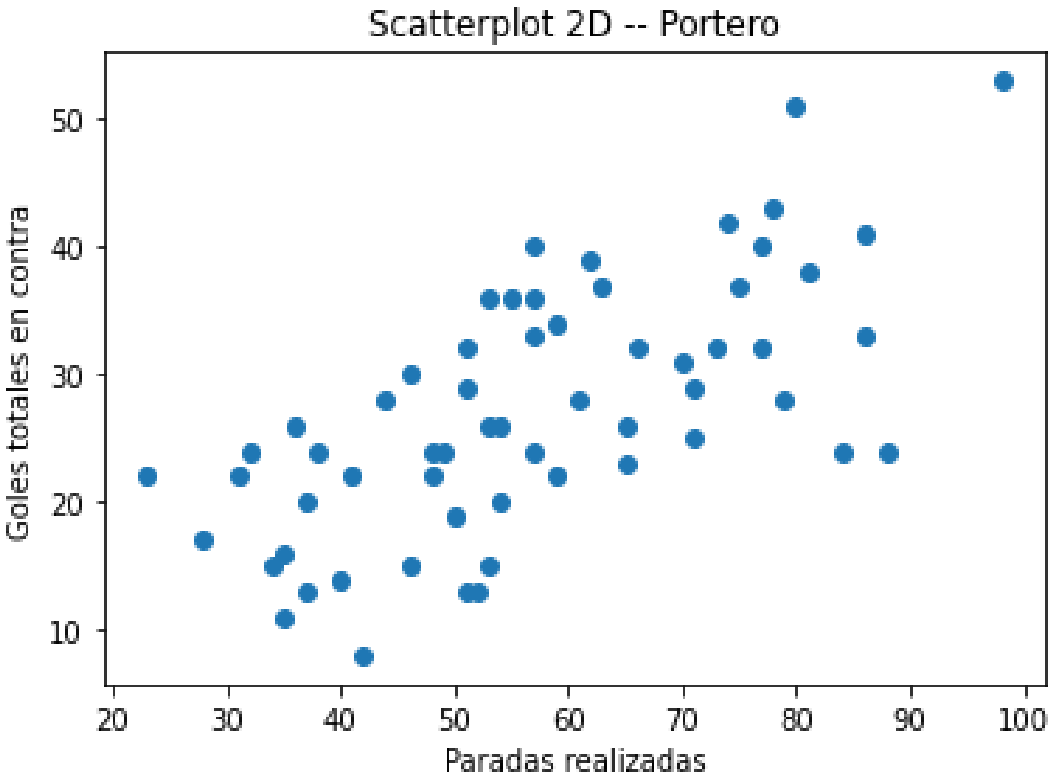

Figura 5-4: Gráfica Scatterplot (2D) de la posición portero

Como se puede observar, la tendencia es que a medida que aumenta el número de paradas realizadas, lo hace también el número total de goles en contra. Esto es consistente con el hecho de que la correlación entre ambas variables, además de ser bastante alta, es positiva (el valor concreto de dicha correlación es 0.671662).

Como valor destacable, comentar el portero que se encuentra en la esquina superior derecha, destacado del resto de porteros de los cuales se han recogido en este análisis. Este portero es Sam Johnstone (West Bromwich Albion, Premier League). Pertenece a un equipo humilde de la liga inglesa, por lo que es natural que tenga un número tanto de paradas como de goles recibidos relativamente alto.

En la figura 5-5 se puede observar que es ciertamente relativa a la anterior, ya que muestra un coeficiente consistente en la división entre el número total de goles en contra entre el número de paradas realizadas (cociente entre las dos variables utilizadas previamente). Por la naturaleza de esta variable, se intuye que el mejor valor para un portero es el menor, ya que, para que este coeficiente descienda de valor, o el portero recibe un número reducido de goles o realiza un gran número de paradas.

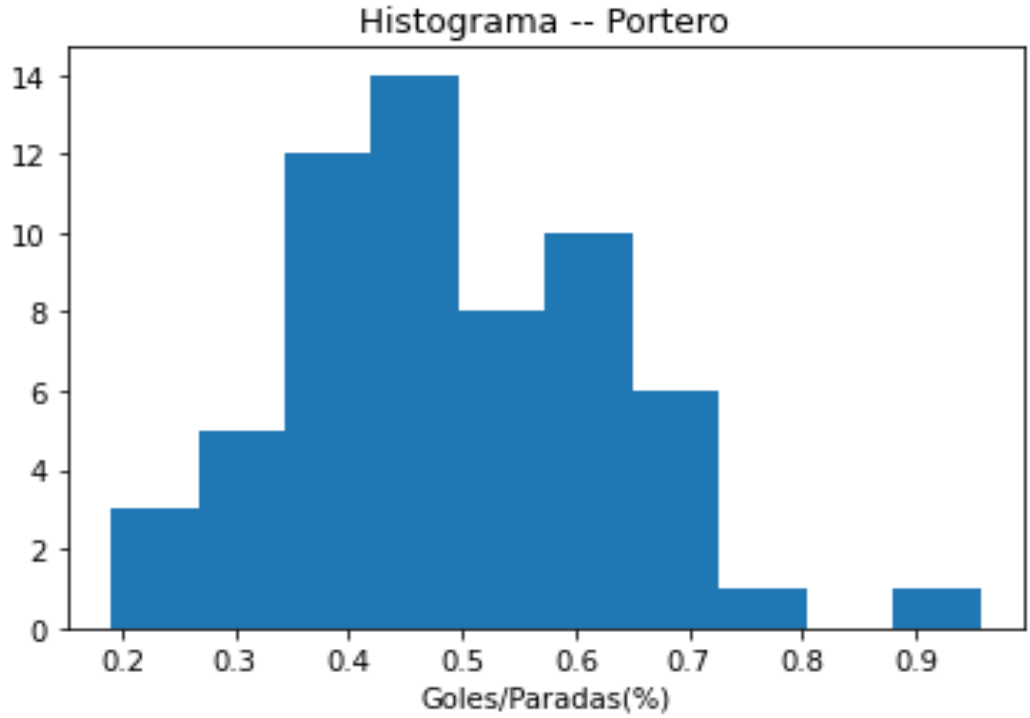

Figura 5-5: Histograma de la posición portero

Como se puede observar, los valores medios de este coeficiente rondan el intervalo [0.35,0.7] aproximadamente, teniendo en cuenta que, en ese intervalo, de la muestra total de 60 porteros, tenemos, aproximadamente, unos 50.

Fuera de ese intervalo, los porteros que se encuentran en el intervalo [0.2,0.35] son los porteros que mejor desempeño tienen en este aspecto. De hecho, realizando un análisis de quiénes son los porteros que tienen un valor del coeficiente dentro de ese intervalo, se encuentran porteros de equipos punteros (con alto valor de mercado) en su gran mayoría, por lo que el análisis de esta variable está recogiendo, de manera correcta, quiénes son los mejores porteros. Obviamente, si el equipo comprador tiene un presupuesto limitado dedicado a la adquisición de un portero, se debería ir a un valor de este coeficiente un poco mayor para encontrar porteros acordes a su inversión. En el intervalo [0.7,1] encontramos únicamente 2 porteros que, a priori, descartaremos, por mal rendimiento, ya que reciben un número considerable de goles para un número de paradas no muy considerable.

En la siguiente gráfica (figura 5-6) se tiene el caso tridimensional de una gráfica de dispersión (scatterplot). Se enfrentan las variables: 'Paradas realizadas' en el eje x (nótese el sentido de crecimiento de dicho parámetro, que aparece al contrario de la variable en el eje y), 'Penaltis parados (% sobre 1)' en el eje y, y, por último, 'Valor de mercado' en el eje z (para su mejor representación, se ha dividido el valor de mercado entre 100 millones, de tal manera que el valor máximo 1.0 representa 100 millones de euros), variable que resulta importante analizar por lo explicado al principio del presente apartado, ya que es el parámetro más importante de todos a la hora de acometer una planificación deportiva.

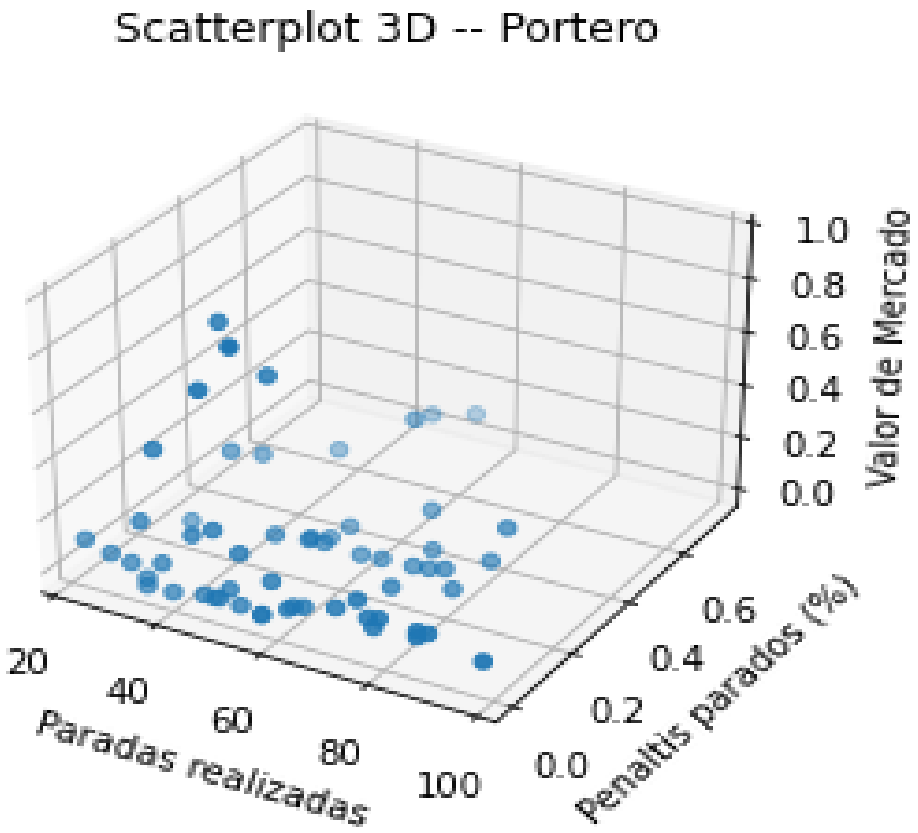

Figura 5-6: Gráfica Scatterplot (3D) de la posición portero

La tendencia general de esta gráfica es que los porteros se agrupen, sobre todo, en un rango [0,0.4] de porcentaje de penaltis parados y un valor de mercado bajo, que es lógico en ambos casos, ya que muchos de los porteros recogidos en la BBDD son miembros de equipos con un presupuesto no muy elevado, y, además, siendo complicado tener un porcentaje de acierto en penaltis parados superior a ese rango.

En este caso, los porteros más interesantes son aquellos que se encuentran con un valor superior a 50-60 paradas realizadas, con un porcentaje de penaltis parados superior a 0,3-0,4 (30-40%) y con un valor de mercado entre [0,0.2] (o lo que es lo mismo, entre 0 y 20 millones de €).

Otro apunte interesante sobre esta gráfica es cómo los porteros caros (con un valor de mercado superior a 50 millones de €), por lo general, no despuntan en valor en ninguno de los dos parámetros. En el caso de la variable 'Paradas realizadas' es lógico, ya que en los equipos de mayor nivel hay mejores jugadores y, generalmente, se defiende mejor, lo que hace que el portero realice un menor número de paradas en cantidad. Para la variable 'Penaltis parados (%)', ésta está sometida a mucha variabilidad, ya que es complicado para un portero parar muchos penaltis porque depende en gran parte del desempeño del lanzador para poder pararlo, por lo que es "normal" que, incluso los grandes porteros no tengan unas estadísticas muy reseñables en este respecto.

5.1.2 Gráficas relativas a la posición central

En la gráfica de la figura 5-7 se enfrentan las variables: 'Balones largos por partido' en el eje x y 'Posesión perdida' en el eje y. Para este caso, se buscará maximizar la variable en el eje horizontal y minimizar la variable en el eje vertical, por lo que el mejor valor se encontraría en la esquina inferior derecha.

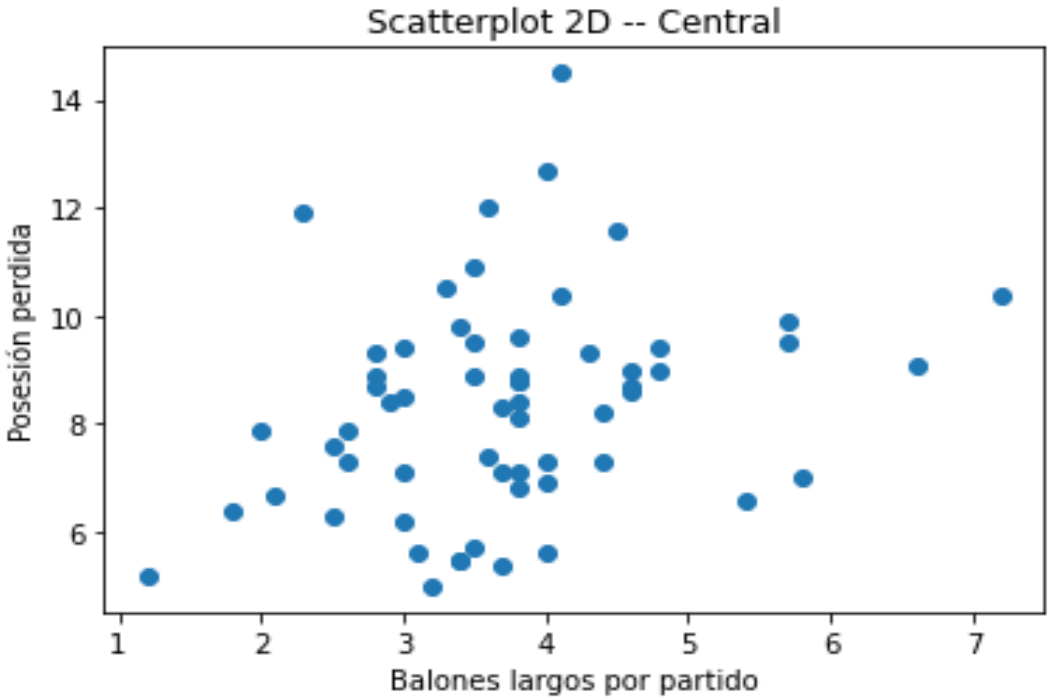

Figura 5-7: Gráfica Scatterplot (2D) de la posición central

En este caso, las dos variables, ateniéndose a los datos recogidos, no parece que recojan una correlación grande, ya que no se observa una tendencia de crecimiento clara como sí ocurría en el caso anterior, aunque sí que se visualiza una cierta tendencia de crecimiento entre las dos variables (con una correlación positiva). El coeficiente de correlación entre estas dos variables es 0.273, lo que ratifica lo comentado previamente, tenemos una correlación positiva con un valor relativamente pequeño.

Parece razonable asegurar que, bajo este criterio, se escojan a los centrales en el rango de 4 o más balones largos por partido, y con una posesión perdida menor de 10 (veces por partido). Los balones largos, como objetivo principal, tienden a otorgar al equipo una mejor salida de balón, que redunda en un mayor número de registros a la hora de tener la posesión de balón y atacar con más facilidad. Este hecho está directamente relacionado con la posesión perdida, ya que una peor salida de balón es consecuencia directa de una pérdida de la posesión. Con respecto a la posesión perdida, muchos equipos priorizan que sus centrales no pierdan el balón con facilidad, ya que una pérdida de balón de un central, al estar cerca de tu propia portería, puede propiciar una situación de peligro para el rival.

Como valores atípicos, destacar aquellos valores que no tienen un gran número de balones largos (en el entorno de 1-4 balones largos por partido) y, sin embargo, tienen unos datos de posesión perdida muy altos (por encima de 12). Estos 4-5 centrales, bajo este análisis, se suprimirían de la lista final de centrales por un rendimiento bajo en este respecto.

En este histograma (figura 5-8) se puede observar la variable: 'Duelos ganados por partido', variable de vital importancia en lo que respecta a cualidades importantes que se buscan en un defensa, pero sobre todo aquellos defensas que juegan en la posición de central.

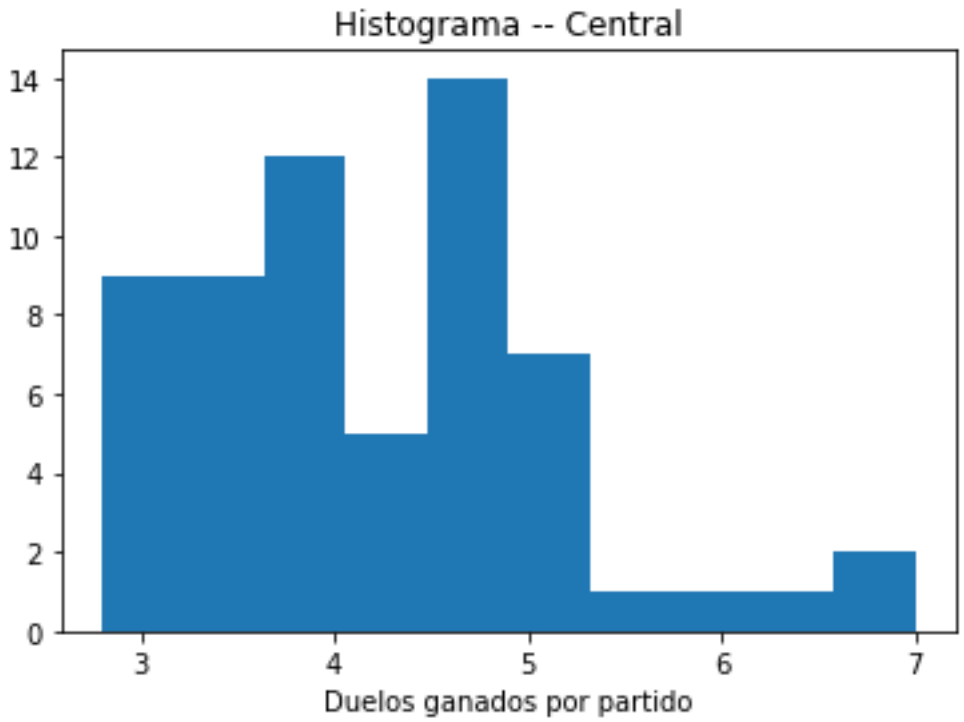

Figura 5-8: Histograma de la posición central

Como se puede observar, la mayoría de los centrales se mueven en el entorno de [2.5,5.25] duelos ganados por partido, reseñando como valores atípicos aquellos por encima de este intervalo.

Observando los valores atípicos para ver quiénes son aquellos centrales que pertenecen a este intervalo, parece que el factor más relevante para que un central tenga un alto valor de duelos ganados por partido no es tanto el valor de mercado (que en la posición de portero sí se vio reflejado una relación entre un mejor desempeño con un mayor valor de mercado), sino que tiene más relación con el estilo de juego del equipo en el que están jugando. Tenemos jugadores como Davíd Garcia (Osasuna, La Liga) con 6.8 duelos ganados por partido y Loïc Badé (Lens, Ligue 1) con 7 duelos ganados por partido, que pertenecen a equipos relativamente humildes (en términos presupuestarios) pero sin embargo tienen un estilo de juego ofensivo y buscando el robo de balón rápido, lo que propicia que en muchas ocasiones se potencien, sobre todo en los centrales, el buscar los duelos individuales con los rivales para poder quitarles la posesión de balón.

En este gráfico tridimensional correspondiente a la figura 5-9 se presentan las siguientes variables: 'Intercepciones por partido' en el eje x, 'Posesión recuperada' en el eje y, y, de igual manera que en todos los gráficos de dispersión tridimensionales de este apartado, la variable 'Valor de mercado' en el eje z.

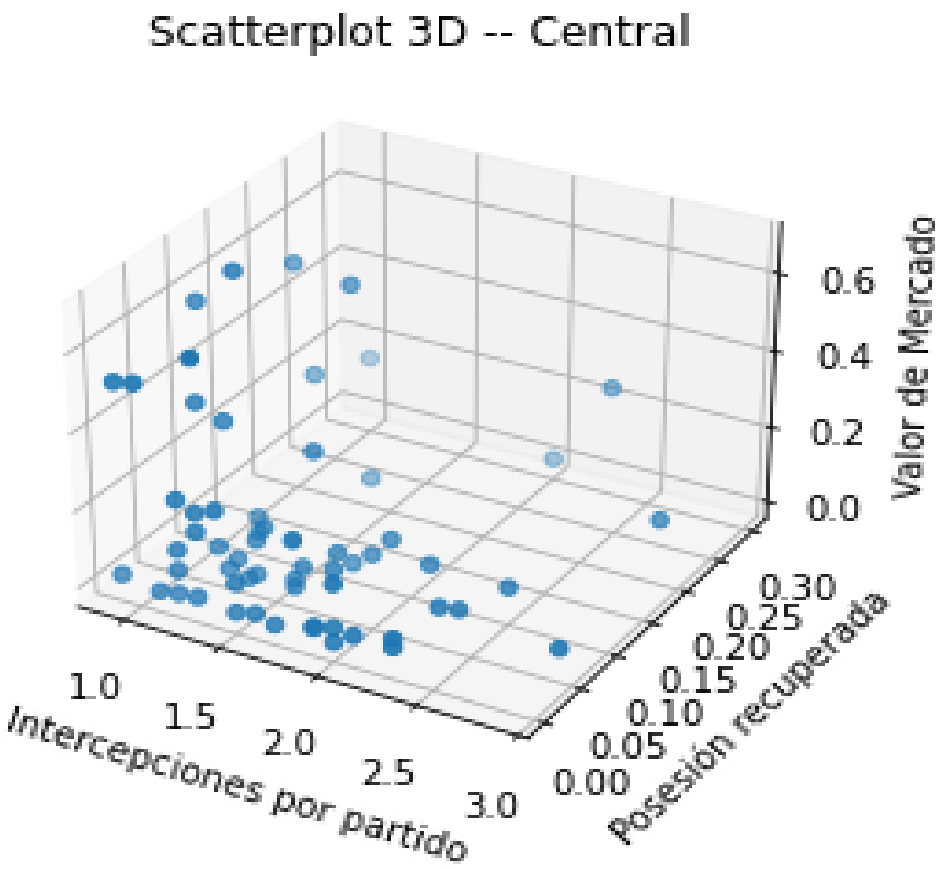

Figura 5-9: Gráfica Scatterplot (3D) de la posición central

El mejor valor en este caso es el valor que, de nuevo teniendo en cuenta que z=0, maximiza las variables en el eje x e y. En esta gráfica se puede observar como la tendencia general es que los datos estén dentro del intervalo [1.0,2.0] para la variable 'Intercepciones por partido' (eje x) y [0.00,0.20] para la variable 'Posesión recuperada' (eje y). Esta tendencia, incluso, se cumple para todos los valores de mercado, ya que hay centrales más caros en los que no se observa una clara mejoría en lo que respecta a estas estadísticas, salvo algún caso aislado.

Como valores anómalos resaltar, sobre todo, el central que llega a estar prácticamente en el punto ideal de la gráfica, maximizando las dos variables de los ejes x e y, y minimizando el valor de mercado (eje z). Este central es Francesco Acerbi (Lazio, Serie A). El hecho de que este futbolista sea el que mejores estadísticas tiene en este aspecto es ciertamente consistente con la filosofía de juego de este equipo, que utiliza a los centrales (al jugar con 3 en vez de con 2) como recuperadores de balón, al tener éstos más libertad en el campo. Esta cuestión es análoga a lo que ocurría en el caso anterior del central francés Loïc Badé, cuyo equipo juega también con 3 centrales, lo que le hace tener más libertad para recuperar balones y participar en más duelos individuales (y consecuentemente, más intercepciones de balón).

5.1.3 Gráficas relativas a la posición mediocentro

En esta gráfica (figura 5-10) se contraponen las variables: 'Pases completados en campo propio' (eje x) y 'Pases completados en campo contrario' (eje y), siendo ambas variables una media de pases por partido.

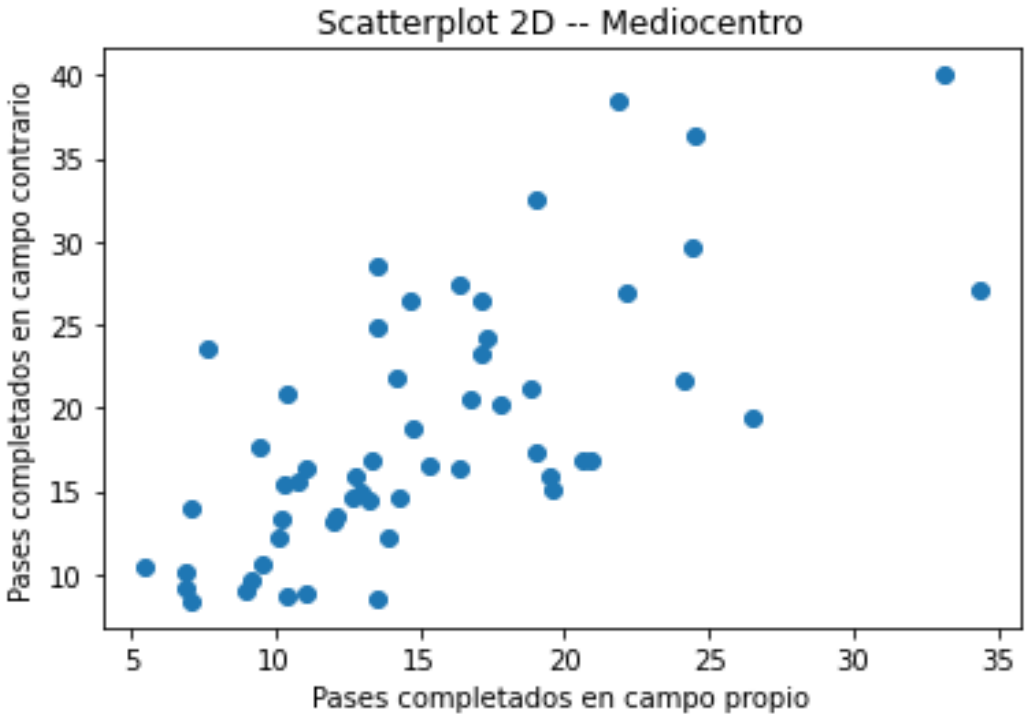

Figura 5-10: Gráfica Scatterplot (2D) de la posición mediocentro

El punto que maximiza ambas variables se encuentra en la esquina superior derecha. Hay que puntualizar, antes de realizar el análisis, que la variable que recoge los pases en campo contrario es más valiosa para los equipos, ya que denota una capacidad técnica del futbolista mayor, debido a la mayor dificultad de realizar pases en el campo rival (donde, lógicamente, hay más jugadores rivales) que en campo propio. Por tanto, a la hora de observar los valores anómalos, se van a priorizar aquellos jugadores que maximicen la variable recogida en el eje vertical.

También hay que tener en cuenta el equipo en el que juega cada futbolista, ya que puede jugar en un buen equipo con una filosofía de juego de posición, en la que la mayoría de pases se den en campo rival (y, además, muchos son pases de seguridad, de fácil concreción), por lo tanto no tiene el mismo valor que un futbolista dé el doble de pases en campo rival que otro si el que menos pases de este estilo realiza es de un equipo netamente inferior que el que más hace.

Estos dos atributos, basándonos en los datos recogidos, mantienen cierta correlación, ya que se observa una tendencia teniendo en cuenta que, a medida que crecen los pases completados en campo propio, lo hacen en más o menos la misma proporción los pases en campo contrario. Esta correlación se refuta matemáticamente a partir del coeficiente de correlación entre ambas variables, que es 0.690917. Además, la tendencia en este caso es que los jugadores den, de media, menos de 20 pases en campo propio y menos de 25 pases en campo contrario.

Como valores anómalos, resaltar los 3 futbolistas que tienen más de 35 pases en campo rival por partido, que son Manuel Locatelli (Sassuolo, Serie A), Joshua Kimmich (Bayern de Múnich, Bundesliga) y Ryan Gravenberch (Ajax, Eredivisie). Estos tres jugadores, aparte de ser piedras angulares en el centro del campo de sus respectivos equipos, juegan en contextos de equipos que utilizan la posesión de balón para someter al rival, por tanto, es coherente con el análisis realizado previamente, son jugadores que van a tener, "forzadamente" un número alto de pases tanto en campo propio como, sobre todo, en campo rival.

Este histograma (figura 5-11) recoge la variable 'Pases clave por partido', que es un atributo principalmente enfocado en los mediocentros que juegan más cerca de área contraria, para analizar su capacidad creativa en lo que respecta a creación de ocasiones claras de gol.

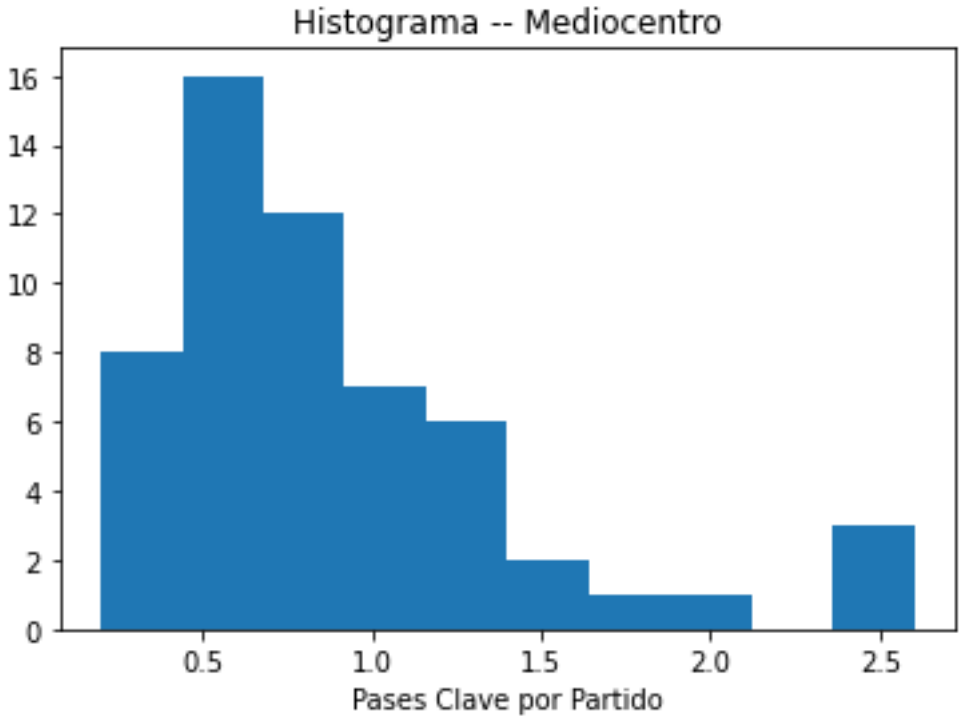

Figura 5-11: Histograma de la posición mediocentro

La tendencia general de número de pases clave por partido se mueve en el intervalo [0.2,1.35]. Este intervalo recoge 49 jugadores, que es un 87.5% del total de mediocentros recogidos en la muestra (56 mediocentros se tienen en la BBDD), por lo tanto, a la hora de buscar un futbolista, muy probablemente se mueva en este entorno de pases clave por partido.

Como valores anómalos, cabe destacar aquellos con un índice superior a 2, que son: Oussama Tannane (Vitesse, Eredivisie), Joey Veerman (Heerenveen, Eredivisie), Joshua Kimmich (Bayern de Múnich, Bundesliga) y Rodrigo de Paul (Udinese, Serie A).

Con respecto a los dos jugadores de la liga holandesa (Eredivisie), es relativamente natural que en una liga en la que se marcan muchos goles y se generan muchas ocasiones de gol, tanto por la filosofía ofensiva del fútbol holandés como por la debilidad de sus defensas, haya jugadores que despunten en esta característica. Por lo tanto, el hecho de que despunten en estas características no es indicativo de una capacidad superior de sendos futbolistas en este respecto, por lo que habría que analizar otros atributos para poder llegar a considerarlos en una lista final de futbolistas de potencial adquisición.

Por otro lado, Joshua Kimmich es un futbolista que ha aparecido como jugador destacable en el anterior análisis, lo que es completamente entendible debido al sometimiento de los equipos alemanes frente al Bayern de Múnich, y además denota un nivel superior en este futbolista (de ahí su elevadísimo valor de mercado: 98 millones de euros). Por último, Rodrigo de Paul es el futbolista más creativo y que más ocasiones genera del

Udinese, por lo que es normal que destaque en este análisis y, posiblemente, sea un futbolista que en muchos de los aspectos relativos a la creatividad, despunte.

En la siguiente gráfica (figura 5-12) se encuentran tres variables: 'Pases completados (%)' en el eje x, 'Pases clave por partido' en el eje y, y 'Valor de mercado' en el eje z. El valor óptimo en este caso, de igual manera que previamente, se sitúa en el menor valor de mercado posible (z=0) y se maximiza para valores cercanos a 1 en la variable del eje x y un valor próximo a 3 en la variable del eje y.

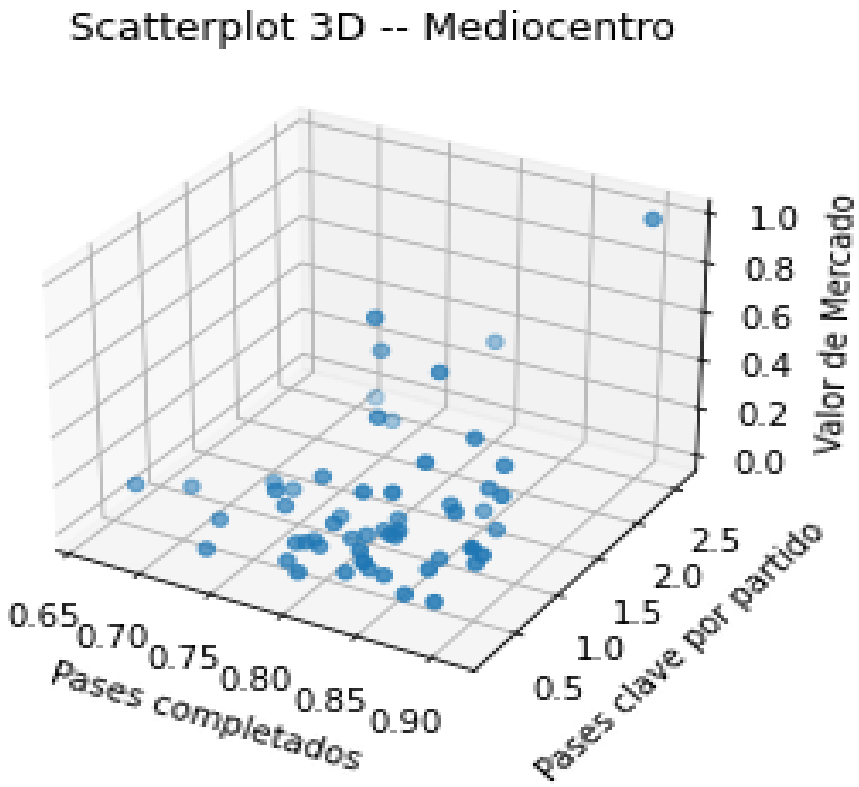

Figura 5-12: Gráfica Scatterplot (3D) de la posición mediocentro

En este caso, podemos observar que los mediocentros relativamente asequibles se encuentran relativamente parejos en las dos variables del caso de estudio, por lo que en este aspecto parece que hay pocos jugadores que despunten sobre el resto.

Vuelve a ser el caso más anómalo Joshua Kimmich, que maximiza el porcentaje de pases acertados y el número de pases clave por partido, por lo que si este análisis de scouting se realizara en un club con un presupuesto grande, definitivamente sería la opción número 1 del equipo de Scouting, ateniéndonos a estos preanálisis gráficos.

5.1.4 Gráficas relativas a la posición delantero

En el gráfico de dispersión de la figura 5-13 se enfrentan las variables: 'Disparos por partido' (eje x) contra 'Regates por partido' (eje y). El punto de mejor valoración en estos dos atributos se encuentra en la esquina superior derecha en esta gráfica.

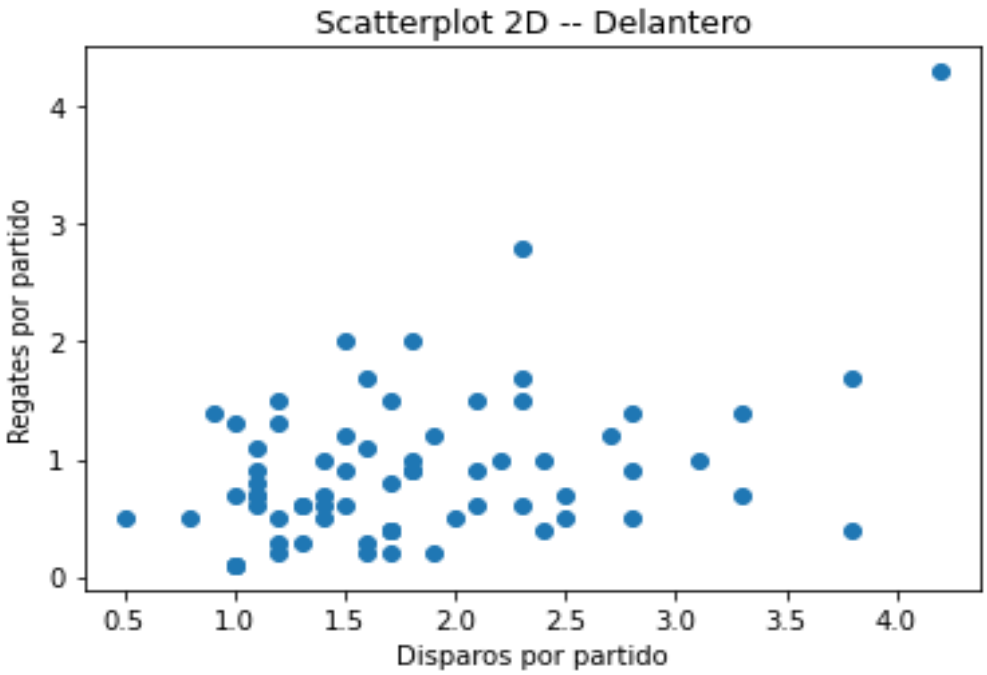

Figura 5-13: Gráfica Scatterplot (2D) de la posición delantero

En este caso, la variable 'Regates por partido' es más valiosa, en lo que se refiere a rendimiento individual de un futbolista, porque la variable 'Disparos por partido', más que denotar una capacidad superior de un delantero sobre los demás, tiene mayor relación con el rendimiento conjunto del equipo. Si el delantero juega en un equipo muy propenso a atacar, posiblemente éste tendrá una estadística de dicha variable mayor, por lo que no es indicativo de si un delantero es mejor o peor.

Estas dos variables parecen, de igual manera que las anteriores, que contienen cierta correlación, pero no con una tendencia tan clara como ocurría en los anteriores gráficos de dispersión bidimensionales, ya que parece que la tendencia es que el aumento de disparos por partido repercute en un aumento de regates por partido, pero en menor grado. Concretamente, estos dos atributos tienen un coeficiente de correlación: 0.410795.

Como valores anómalos, centrándonos en maximizar los regates por partido, destacan dos futbolistas: Lionel Messi (FC Barcelona, La Liga) y Junior Messias (Crotone, Serie A). El primero potencia muchísimo su

regate para sortear rivales y, además, es uno de los mejores futbolistas de la historia, por lo que es natural que aparezca como un futbolista interesante en este análisis. El segundo, aún jugando en un club bastante humilde económicamente hablando, viene de Brasil, que es una liga y un país donde se potencia bastante el uso del regate como arma para superar a los contrarios.

En este gráfico (figura 5-14) se puede ver un histograma en el que la variable de representación es 'Goles por partido', valor clave a la hora de analizar el rendimiento de un delantero.

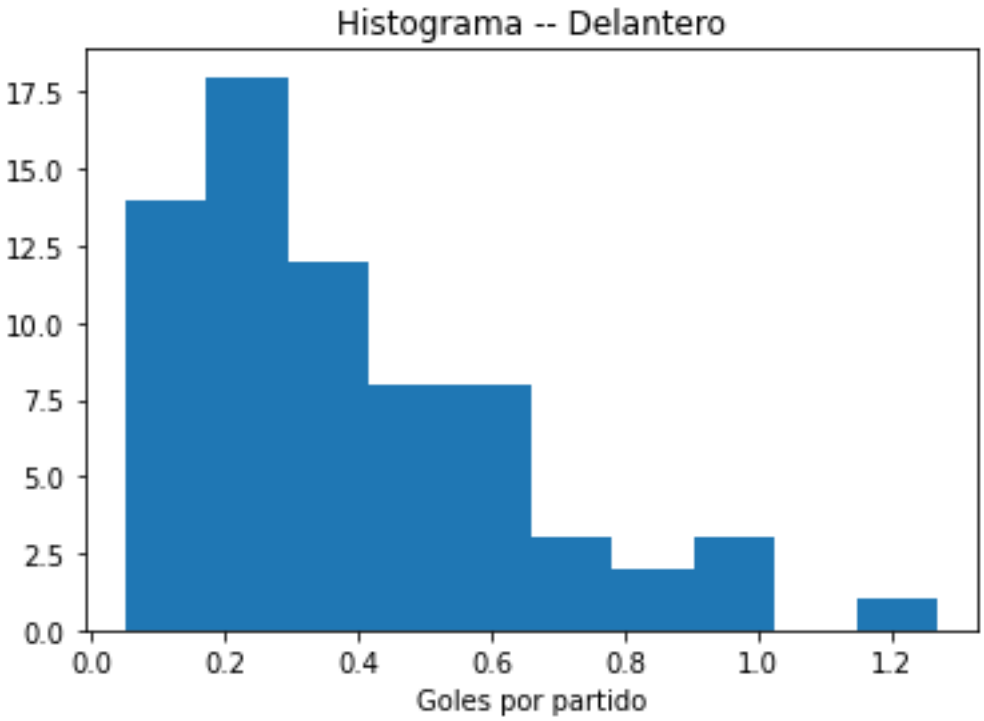

Figura 5-14: Histograma de la posición delantero

La mayoría de los delanteros se encuentran en el intervalo [0.1,0.65], aproximadamente 60 delanteros de los 69 que hay. Por encima de este intervalo, encontramos jugadores de alto nivel en equipos punteros, como, por ejemplo: Lionel Messi (FC Barcelona, La Liga), Benzema (Real Madrid, La Liga), Lewandowski (Bayern Múnich, Bundesliga), etc., por lo tanto, esta variable es interesante para equipos que tienen un presupuesto relativamente alto para la adquisición de un delantero. Para casos en los que el presupuesto es limitado, quizá hay que buscar otras variables interesantes para analizar, o bien buscar en ligas menores futbolistas con unos números parecidos a los de los mejores delanteros de las ligas recogidas en esta BBDD (como por ejemplo la liga austriaca, belga, polaca, rusa, etc.), como hace el director deportivo del Sevilla FC en el análisis del mercado, tal y como se ha expuesto en el punto 3 de este trabajo.

En este gráfico de dispersión tridimensional (figura 5-15) se enfrentan tres variables: ‘Goles por partido’ en el eje x, ‘Asistencias’ en el eje y, y ‘Valor de Mercado’ en el eje z. Nuevamente, el punto que se aspira a conseguir se encuentra en el valor que maximiza las variables de los ejes x e y, para el mínimo valor de mercado posible.

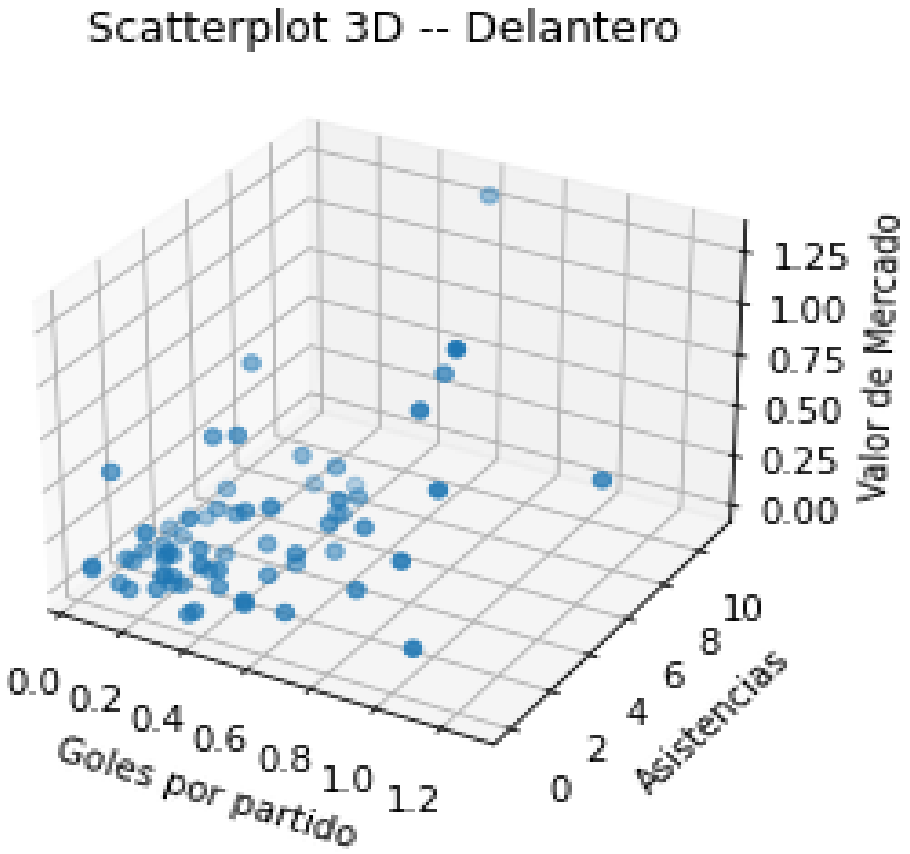

Figura 5-15: Gráfica Scatterplot (3D) de la posición delantero

La tendencia general de este gráfico es que la mayoría de los valores se agrupen en un valor de mercado inferior a 25 millones de euros (0.25) y, además, con un valor inferior a 0.6 goles por partido e inferior a 6 asistencias. Como valores anómalos ocurre lo mismo que previamente, son los jugadores con un valor de mercado alto los que despuntan en estas características.

De hecho, si hacemos la matriz de correlación únicamente teniendo en cuenta los delanteros, dos de las variables que más correlación tienen con ‘Valor de mercado’ son, precisamente, ‘Goles por partido’ y ‘Asistencias’ (0.474 y 0.565, respectivamente), lo que explica la alta cotización de los jugadores que tienen buenas estadísticas en estos atributos.

6 IMPLEMENTACIÓN Y RESULTADOS

Una vez explicada la base de datos configurada para el presente trabajo, así como el preanálisis por medio de gráficas que se ha realizado, se debe tratar el modo de implementación mediante los algoritmos expuestos previamente. La implementación de dichos algoritmos se debe realizar mediante software de programación que tenga las librerías necesarias para ello, como es el caso de Python.

6.1 Lenguaje de programación: Python

A continuación, se va a tratar el lenguaje de programación utilizado para la importación y tratamiento de la base de datos, así como el posterior análisis mediante algoritmos de ML.

Según (González Duque, 2018), Python es un lenguaje de programación creado por Guido van Rossum, ingeniero holandés que trabajaba en el Centro de Investigación de Ciencias de la Computación de Ámsterdam, a principios de los años 90. Debe el nombre al grupo de comedia inglesa "Monty Python". Al tener una sintaxis limpia y un código muy legible, además de ser un entorno de programación de código abierto, lo que hace que se pueda utilizar dicho código en cualquier entorno o problema, favoreciendo su manipulación y adaptación a cada caso. Además, es uno de los mejores lenguajes para realizar un tratamiento de una base de datos de un volumen considerable. Es por ello por lo que, de cara al análisis de Big Data y, por consiguiente, del Machine Learning, es el lenguaje preponderante en las empresas que desempeñan labores en este respecto.

Precisamente esa es la razón del por qué se ha optado por hacer uso de Python en este trabajo, porque es un lenguaje puede hacer uso de librerías dedicadas al estudio y análisis de bases de datos mediante ML o Data Analysis. De cara al entorno laboral es conveniente conocer este lenguaje de programación, debido a su versatilidad y adaptabilidad a cualquier caso de estudio, así como su sencillez a la hora de ser tanto programado como interpretado.

Entre sus principales características se puede destacar:

- **Lenguaje interpretado o de script:** se ejecuta a partir de un programa intermedio llamado intérprete. Esto hace que la compilación del código, en vez de realizarse a lenguaje máquina (lenguaje compilado) se haga a través de dicho intérprete. El lenguaje interpretado es más flexible y portable, mientras que el compilado tiene una ejecución más rápido. Python es un lenguaje de compilación semi interpretado, debido a que, aun siendo lenguaje interpretado, contiene muchas características de los lenguajes compilados.

- **Tipado dinámico:** no es necesario definir el tipo de dato de una determinada variable del código, sino que, según el valor asignado, el lenguaje directamente asocia esa variable a un tipo de dato. Si cambia el valor de la variable y ese valor es de otro tipo distinto al definido previamente, el tipo de la variable cambia.
- **Fuertemente tipado:** no se permite tratar a una variable como si fuese un tipo distinto del determinado por el lenguaje, sino que se tiene que definir explícitamente el nuevo tipo de esa variable.
- **Multiplataforma:** Python está disponible en una gran cantidad de plataformas como pueden ser: Windows, Mac OS, Linux, DOS, etc.
- **Orientado a objetos:** los casos concretos de estudio del mundo real tratados pasan a ser clases y objetos en este lenguaje de programación, y el algoritmo es el que se encarga de establecer la relación entre estos objetos.
- **Lenguaje sencillo:** la sintaxis de programación es cercana al lenguaje natural, por lo que los códigos que se realizan en este lenguaje parecen pseudocódigo. Esto hace que sea muy accesible para todo el mundo y que sea una opción ideal para adentrarse en el mundo de la programación.

A continuación, se van a tratar las librerías utilizadas en Python tanto para el tratamiento de la base de datos como para la implementación de los algoritmos.

- **Librería Pandas**

Según (Castro, n.d.), Pandas es una librería de análisis de BBDD con Python. Entre sus funciones y características destacan:

- Herramienta para lectura y escritura de datos (CSV, SQL, documentos de texto, etc.)
- Estructuras tabulares de datos (DataFrame)
- Hace más intuitivo el uso de la librería Numpy (Numpy es la librería dedicada a la computación científica con Python).
- Facilita el manejo de series temporales.
- Alinea fácilmente los datos y maneja los que faltan en la BBDD aportada.
- En uso para una gran cantidad de campos de aplicación: finanzas, economía, estadística, publicidad, etc.

- **Librería Numpy**

Según (*NumPy*, n.d.), la librería Numpy es una herramienta creada en 2005 que habilita la computación numérica con Python. Además, es un software libre, de manera que los usuarios pueden utilizar, estudiar y mejorarlo de manera completamente libre. Esta librería tiene como principales funciones y características las siguientes:

- Contiene funciones matemáticas, generación de números aleatorios a partir de distribuciones estadísticas, algoritmos de algebra lineal, etc.
- Soporta una gran cantidad de hardware y plataformas computacionales.
- El código que utiliza NumPy está programado en C, de manera que se beneficia tanto de la rapidez del código compilado como de la flexibilidad que otorga Python.
- Sintaxis muy sencilla.

- **Librería Matplotlib**

Según (*Matplotlib: Python Plotting — Matplotlib 3.4.1 Documentation*, n.d.), la librería Matplotlib utilizada para crear gráficas y visualizaciones, tanto en 2D como en 3D, en Python. Entre las posibilidades que ofrece esta librería destacan: gráficos de barras (pudiendo realizar análisis tanto variables numéricas como categóricas), histogramas, gráficos de dispersión, representaciones de funciones tridimensionales, mapas de color, etc. Entre sus principales características se podría destacar:

- Pocas líneas de código necesarias para realizar muchas representaciones gráficas, lo que facilita la codificación.
- Figuras interactivas para poder realizar un análisis más exhaustivo.
- Personalización de ejes, estilos de líneas, etc.
- Tiene un gran número de paquetes asociados a la librería que están en continuo desarrollo para ofrecer un mayor número de funcionalidades al usuario.

- **Librería Scikit-learn**

Según (*About Us — Scikit-Learn 0.24.1 Documentation*, n.d.), scikit-learn es una librería creada en 2007 por Matthieu Brucher como parte de un proyecto de verano sobre codificación organizado por Google. Hasta 2010, no se produjo la primera versión pública de esta librería, aunque desde entonces, cada 3 meses aproximadamente se han producido renovaciones de esta. Concretamente, en este proyecto, se han utilizado los siguientes paquetes de la librería scikit-learn:

- **Sklearn.decomposition (PCA):** realiza el análisis PCA (en el número de PCA definidos previamente por el usuario) de los datos procedentes de la base de datos importada en el código.
- **Sklearn.pipeline(make_pipeline):** línea de código necesaria para realizar la estandarización de los datos.
- **Sklearn.preprocessing(StandardScaler):** estandarización de los datos restando la media de los atributos y dividiendo por su desviación típica.
- **Sklearn.cluster(k-Means):** paquete que contiene la codificación necesaria para poder realizar el algoritmo k-Means a partir de un número de clusters definido previamente.

6.2 Filtrado de variables

Previo a la implementación de los casos de estudio, teniendo en cuenta la naturaleza de alguno de los algoritmos a aplicar, se debe realizar un filtrado de variables de manera que se reduzca la dimensionalidad del problema. Teniendo en cuenta este criterio, se observa lo siguiente:

- **Caso de estudio relativo al clustering:** sí precisa de un filtrado de variables previo, ya que el propio algoritmo no lo realiza, por tanto, para realizar un análisis mediante clustering es muy importante elegir aquellas variables que estén incorreladas y que sean significativas para el análisis del problema.
- **Caso de estudio relativo al PCA:** no precisa de un filtrado de variables previo, debido a que el propio PCA es un algoritmo que recoge aquellas variables que menos correladas están, por lo que realizar un filtrado previo para, después, aplicar este algoritmo, resulta redundante.

El diagrama de flujo que describe el filtrado de variables realizado es el siguiente:

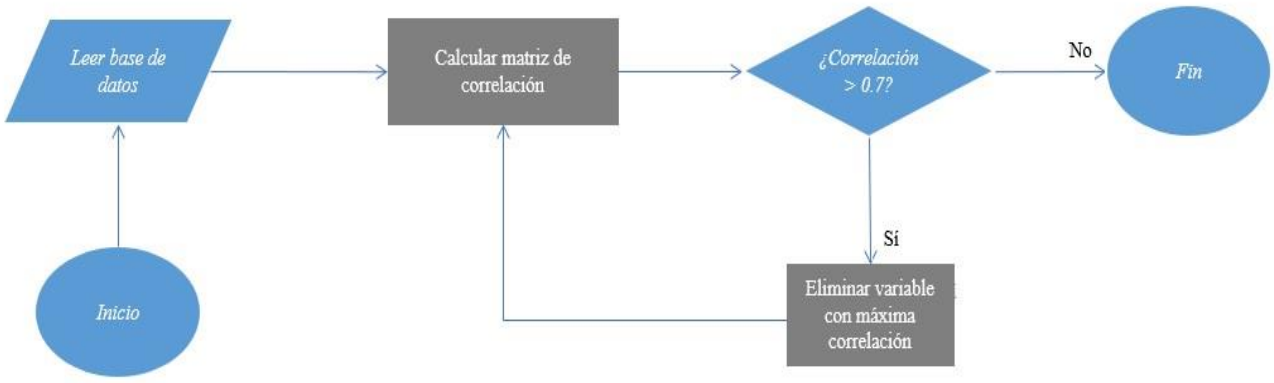

Figura 6-1: Proceso de filtrado de variables

Como se puede observar en este diagrama de flujo, a partir de la matriz de correlación original (la que relaciona todas las variables del problema), se realiza un proceso iterativo de filtrado de variables en el que el criterio de filtración es que se eliminen las variables que tienen mayor correlación.

Este proceso iterativo va calculando matrices de correlación a partir de los datos que, según el algoritmo, deben ser objeto de estudio. En cada iteración se elimina, del conjunto inicial de datos, la variable que presenta una mayor correlación con otra, de manera que de las dos variables que presenten una alta correlación, se elimina una de las dos.

Concretamente, en este ejemplo utilizado para el algoritmo de clustering, se ha utilizado un valor de correlación umbral de 0.7, pudiéndose variar dicho valor a cualquier otro en función de la naturaleza del problema.

Este proceso, una vez concluido, tiene como solución un nuevo conjunto de datos que tiene como característica que ninguna de sus variables presenta una correlación superior al valor 0.7 (considerado umbral, como se ha especificado previamente), por lo que se han eliminado aquellas variables que no aportaban una información de valor al problema.

6.3 Implementación de los algoritmos

En el presente apartado se pretende explicar, de forma esquemática y resumida, la implementación de cada uno de los dos algoritmos utilizados.

6.3.1 Metodología de aplicación del algoritmo de PCA

Para la aplicación de este análisis, se ha realizado el siguiente proceso:

1. Importación de la base de datos en Python (archivo csv), realizado con la librería Pandas.
2. Importación de las funciones necesarias para la aplicación del algoritmo PCA, presentes en la librería scikit-learn.
3. Para la conversión de variables categóricas en numéricas, se ha utilizado, de la librería Pandas, la función ‘dummies’. Esta función habilita al usuario a convertir las variables categóricas en variables binarias. Por ejemplo, si un futbolista es zurdo, la variable ‘Pie Preferido_Izquierdo’ valdría 1, ya que el jugador en concreto cumple que, para la variable ‘Pie Preferido’, su valor es ‘Izquierdo’.

   Esto permite incluir en el algoritmo todas las variables del problema, incluso las categóricas, y así se puede realizar un análisis con mayor certidumbre que sin incluirlas.

4. Escalado de datos del modelo PCA, a partir de las funciones scale y StandardScaler de scikit-learn, que permiten realizar una normalización de los datos.
5. A partir de la base de datos inicial, se definen cuatro subgrupos según la posición a la que pertenecen, de manera que:
    - Portero: contiene a todos los porteros de la base de datos.
    - Defensa: contiene a todos los laterales (derechos e izquierdos) y centrales.
    - Centrocampista: contiene a todos los pivotes, mediocentros y mediapuntas.
    - Atacantes: contiene a todos los extremos (derechos e izquierdos) y delanteros.
6. Aplicación del algoritmo PCA para cada uno de los subgrupos.

6.3.2 Metodología de implementación del modelo de clustering

Los pasos que realiza el algoritmo son los siguientes:

1. Importación de la base de datos en Python (archivo csv), realizado con la librería Pandas.
2. Filtrado de variables acorde a lo explicado en el apartado 6.2
3. Importación de las funciones necesarias para la aplicación del algoritmo de clustering, presentes en la librería scikit-learn.
4. Aplicación del método del codo acorde a lo explicado en el apartado 4.1.2.2, de cara a la definición del número de clusters, realizándose previamente un escalado de datos a partir del conjunto inicial.
5. Aplicación del análisis de la silueta acorde a lo explicado en el apartado 4.1.2.3, de cara a la definición del número de clusters, realizándose previamente un escalado de datos a partir del conjunto inicial.
6. Comparación de los resultados obtenidos en cada uno de los métodos de definición del número de clusters para observar cuál es el valor que mejor se ajusta al conjunto de datos.
7. A partir de la base de datos inicial, se definen cuatro subgrupos según la posición a la que pertenecen, de manera que:
    - **Portero:** contiene a todos los porteros de la base de datos.
    - **Defensa:** contiene a todos los laterales (derechos e izquierdos) y centrales.
    - **Centrocampista:** contiene a todos los pivotes, mediocentros y mediapuntas.
    - **Atacante:** contiene a todos los extremos (derechos e izquierdos) y delanteros.
8. Aplicación del algoritmo de clustering utilizando el algoritmo k-means para cada uno de los subgrupos. Una vez definido el número de clusters óptimo según los métodos del codo y la silueta, se organizan los jugadores de cada subgrupo en el número de clusters determinado.

6.3.2.1 Elección del número de clusters para cada posición

A partir del método del codo y del análisis de la silueta, métodos aplicados en cada una de las posiciones, se han definido el número de clusters óptimo. Como la aplicación de cada método nos proporciona un valor de clusters óptimo según cada caso, lo más productivo es realizar una combinación entre los dos métodos para cada posición y ver cuál es el número de clusters que mejor se adapta a los resultados obtenidos a partir de los dos métodos. Las gráficas y el número de clusters definidos para cada posición son los siguientes:

- Posición 1: Portero

A continuación, se muestran las dos gráficas que muestran cada uno de los métodos aplicados a la portería.

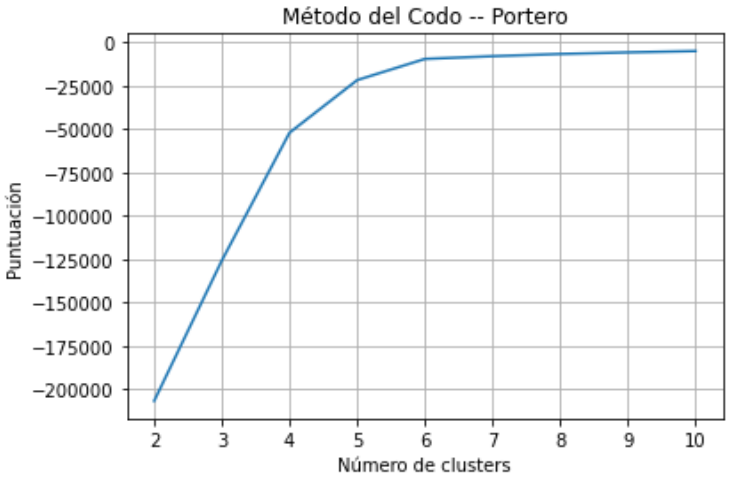

Figura 6-2: Método del codo (Portero)

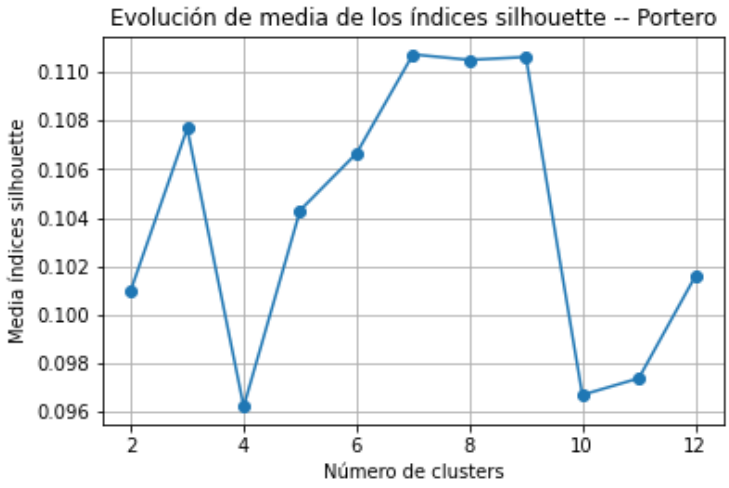

Figura 6-3: Análisis de la silueta (Portero)

En este caso, podemos observar que el número de clusters óptimo para el método del codo es entre 3 y 5, ya que es cuando se observan codos en dicha gráfica. A partir de 5, no existe mucha variación, por lo que se infiere que un aumento en dicho valor no representa una gran variación con respecto a un valor menor de número de clusters. Entre esos tres valores (3, 4 o 5 clusters), el número óptimo de clusters para el análisis de la silueta (valor que maximiza la media de los índices de dicho análisis) es 3, por lo que el valor elegido como óptimo para el caso de los porteros es 3.

- **Posición 2: Defensa**

A continuación, se presentan las gráficas de cada uno de los métodos aplicados en la posición ‘defensa’.

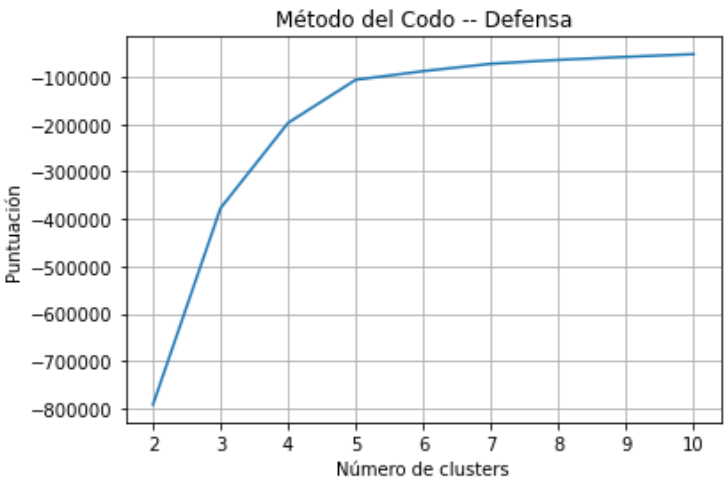

Figura 6-4: Método del codo (Defensa)

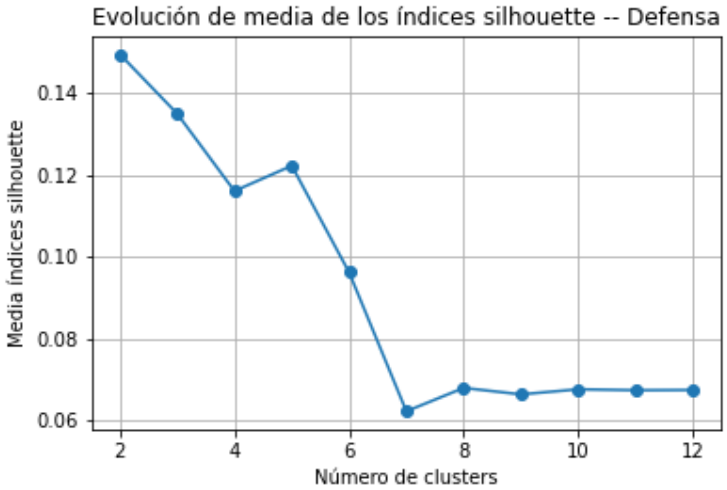

Figura 6-5: Análisis de la silueta (Defensa)

Para los defensas podemos observar que ocurre algo parecido para los porteros en el caso del método del codo, en el que vemos que existe una mayor variación entre los valores 3 y 5, donde se pueden observar codos en la gráfica de este método. Entre estos tres valores, el que maximiza en el análisis de la silueta la media de los índices silueta es, de nuevo, 3, por tanto se define 3 como número óptimo de clusters para esta posición.

- **Posición 3: Centrocampista**

Se procede a presentar las dos gráficas correspondientes a la aplicación del método del codo y del análisis de la silueta para la posición 'centrocampista'.

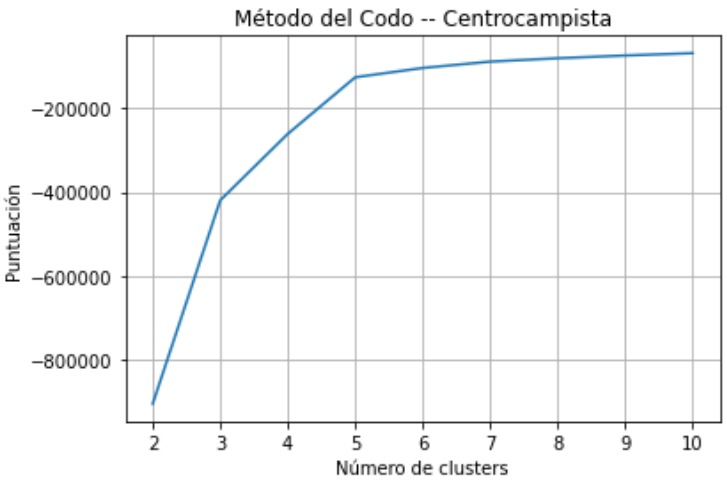

Figura 6-7: Método del codo (Centrocampista)

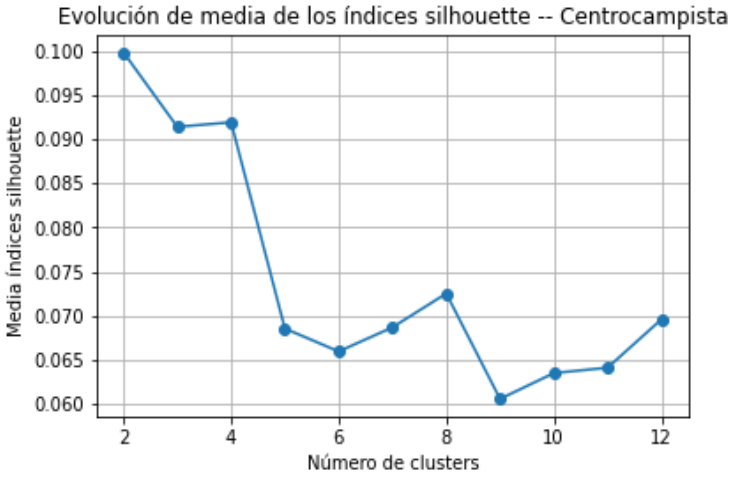

Figura 6-6: Análisis de la silueta (Centrocampista)

Para esta posición se puede observar, en la gráfica perteneciente al método del codo, dos codos que se producen para 3 clusters y 5 clusters. Para esos valores de número de clusters, en el análisis de la silueta, el valor que maximiza la media de los índices silueta es 3, por lo que ese va a ser el valor óptimo de clusters a utilizar para este grupo de futbolistas.

- **Posición 4: Atacante**

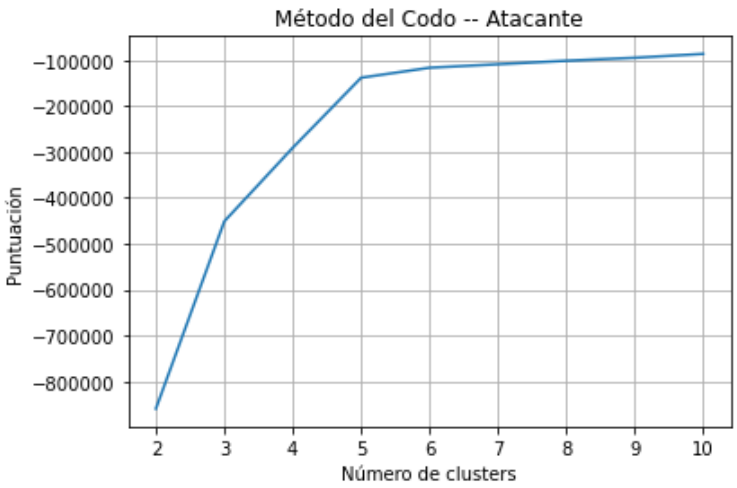

Figura 6-8: Método del codo (Atacante)

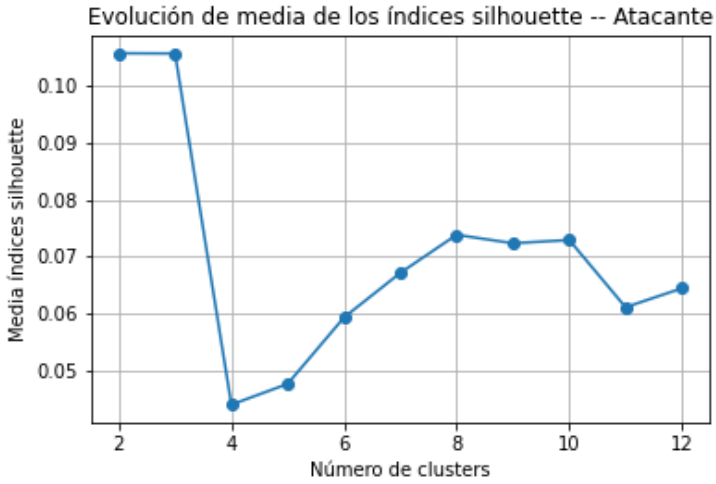

Figura 6-9: Análisis de la silueta (Atacante)

Para los atacantes podemos ver que en la primera gráfica se produce el codo, al igual que en los centrocampistas, en los valores 3 y 5 de número de clusters. Observando el valor de la media de los índices silueta en la segunda gráfica, se puede ver que el valor que maximiza esta media es 3, por lo tanto este será el valor óptimo de número de clusters.

El hecho de que en todos los casos el valor óptimo sea 3 no es casualidad, ya que el método del codo, aplicado a esta base de datos, generalmente tiene los codos en prácticamente el mismo entorno de número de clusters. Por otra parte, en el análisis de silueta ocurre algo parecido, debido a la maximización de media de índices silueta cuanto menor es el número de clusters. Esto propicia que el valor óptimo se encuentre casi siempre en 3, ya que es el menor valor posible de clusters (sin contar 2, que en el método del codo no es un valor en el que se encuentre un codo).

6.4 Análisis de resultados: casos de estudio

En este apartado, a partir de los resultados obtenidos del apartado anterior, se pretenden realizar listas de fichajes acorde a lo pretendido por la dirección deportiva de un club. Se realizarán casos de estudio ficticios basados en planificaciones deportivas reales, adaptadas al presupuesto manejado en cada caso de estudio. Dichas listas servirán para ofrecer a la dirección deportiva distintas alternativas para que la planificación propuesta sea lo más eficiente posible.

El proceso a seguir en el caso de estudio es el propuesto en la siguiente figura:

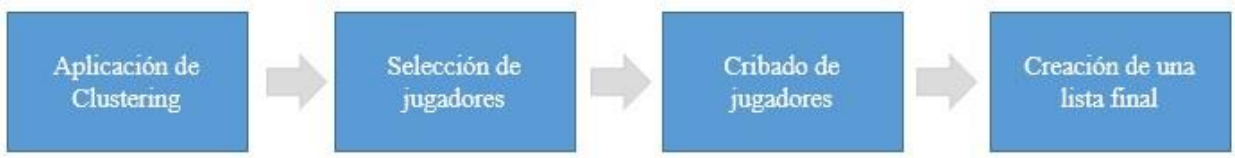

Figura 6-10: Procedimiento de análisis de los casos de estudio

1. **Aplicación de Clustering:** se aplica el algoritmo de Clustering para determinar, según cada uno de los datos contenidos en la base de datos, a qué cluster asocia el algoritmo a cada jugador.
2. **Selección de jugadores:** escoger los jugadores dentro de cada cluster con un valor de mercado parecido al jugador que se está analizando. Si uno de los fichajes de la planificación real pertenece a un cluster determinado, se escogerán, para su posterior análisis, a jugadores con un valor de mercado parecido que pertenezcan al mismo cluster.
3. **Cribado de jugadores:** en el caso de que haya futbolistas escogidos a partir del paso anterior que no hayan jugado nunca en la posición en la que se está buscando un jugador, este no se considerará.
4. **Creación de una lista final:** a partir de los jugadores que se tienen en la lista confeccionada en los pasos anteriores, se observa cuál es el valor del sumatorio de todas las PCA de cada uno de los futbolistas. Se escogerán aquellos futbolistas que tengan el valor más parecido al futbolista que se ha fichado en la planificación real, considerándose estos nuevos futbolistas para la construcción de las planificaciones alternativas.

6.4.1 Caso de estudio I: Club de presupuesto medio-alto

Aprovechando que el análisis del organigrama y la dirección deportiva se han realizado del Sevilla FC, se va a coger como ejemplo su planificación deportiva de la temporada anterior (2020-2021) para intentar realizar una nueva planificación acorde a los resultados obtenidos a partir de la implementación de cada uno de los algoritmos. El Sevilla FC, ha destinado de su presupuesto total (aproximadamente 200 millones de euros), 70 millones de euros a la compra de fichajes. Según (*Sevilla FC - Fichajes 20/21 | Transfermarkt*, 2021), los fichajes realizados por el Sevilla FC en la temporada 2020/2021 fueron (no se consideran futbolistas propiedad del Sevilla FC que vuelven de cesión, solo se consideran futbolistas adquiridos a otros clubes en dicha temporada):

PRESUPUESTO DESTINADO A FICHAJES: 92,8M€		
Posiciones	**Jugador fichado**	**Valor de mercado (M€)**
Portero	Yassine Bounou	17,5
Central	Karim Rekik	2,9
Lateral izquierdo	Marcos Acuña	13,5
Mediocentro	Iván Rakitic	11,0
Mediapunta	Óscar Rodríguez	10,9
	Alejandro Gómez	9,6
Extremo izquierdo	Oussama Idrissi	8,0
Extremo derecho	Suso	19,4

Tabla 4: Fichajes realizados por el Sevilla FC (Temporada 2020/2021)

A continuación, se presentan las tablas (tabla 5 hasta la tabla 9) que recogen, para cada uno de los jugadores de la planificación deportiva del Sevilla FC, las diferentes alternativas a proponer al director deportivo como posibles variantes a los fichados. En la tabla 5, y en las sucesivas tablas tanto en este caso de estudio como en el siguiente, la construcción de la misma es relativamente parecida, de tal manera que:

- **Índice:** es el número asociado al futbolista en la base de datos.
- **Valor PCA combinado:** una vez determinado el número de PCAs que debe tener cada subgrupo, se calcula el valor de cada PCA para cada futbolista y se realiza un sumatorio de todos los PCA para cada uno de los futbolistas. Este índice sirve para ver una similitud entre jugadores, de tal manera que si un jugador tiene un valor de PCA combinado relativamente parecido, tendrá unos datos también parecidos (y, por ende, unas características similares).

TABLA PCA Y CLUSTERING – JUGADOR 1: YASSINE BOUNOU				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
3	Yassine Bounou	Portero	-8,67941	14,5
37	Alphonse Aréola	Portero	-5,70188	12,1
28	Hugo Lloris	Portero	-5,80391	11,6
35	Nick Pope	Portero	4,12172	11,5
15	Fernando Pacheco	Portero	0,980656	14,9
27	Emiliano Martínez	Portero	-3,63878	17,3
4	Álex Remiro	Portero	1,15027	21,0
30	Illan Meslier	Portero	-5,28679	11,2
5	Sergio Asenjo	Portero	1,00832	10,3
16	Marko Dmitrovic	Portero	9,06598	9,2

Tabla 5: Tabla PCA y Clustering para el subgrupo ‘Portero’ con parecido a Yassine Bounou (Sevilla FC)

TABLA PCA Y CLUSTERING – JUGADORES 2-3: KARIM REKIK Y MARCOS ACUÑA				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
600	Karim Rekik	Central	-14,2475	2,9
601	Marcos Acuña	Lateral izquierdo	-3,52804	13,5
129	Rayan Aït-Nouri	Lateral izquierdo	-5,50877	19,4
131	Patrick van Aanholt	Lateral izquierdo	1,89366	7,7
121	Aaron Creswell	Lateral izquierdo	0,441195	6,9
204	Loïc Badé	Central	4,32164	5,8
157	Adriá Pedrosa	Lateral izquierdo	1,9116	5,0
214	Stefan Mitrovic	Central	5,17612	2,4
216	Nicolas Pallois	Central	-8,6567	2,1
210	Brendan Chardonnet	Central	-11,5868	2,1

Tabla 6: Tabla PCA y Clustering para el subgrupo 'Defensa' con parecido a Karim Rekik y Marcos Acuña (Sevilla FC)

TABLA PCA Y CLUSTERING – JUGADOR 4: IVAN RAKITIC				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
358	Ivan Rakitic	Mediocentro	-8,25183	11,0
327	Maximilian Eggestein	Mediocentro	5,688	11,0
328	Diadie Samassekou	Mediocentro	3,26987	13,6
334	Suat Serdar	Mediocentro	7,07238	13,8
299	Ryan Gravenberch	Mediocentro	-8,63236	22,0
319	Djibril Sow	Mediocentro	-5,28382	9,5
326	Wataru Endo	Mediocentro	2,36703	7,3
333	Leandro Barreiro	Mediocentro	-8,37115	7,3
329	Carlos Gruezo	Mediocentro	14,8469	5,7
325	Marcus Ingvartsen	Mediocentro	-6,15517	3,6

Tabla 7: Tabla PCA y Clustering para el subgrupo 'Centrocampista' con parecido a Ivan Rakitic (Sevilla FC)

TABLA PCA Y CLUSTERING – JUGADORES 5-6: ALEJANDRO GÓMEZ Y ÓSCAR RODRÍGUEZ				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
603	Alejandro Gómez	Mediapunta	-13,0147	9,6
602	Óscar Rodríguez	Mediapunta	-7,506	10,9
365	Denis Suárez	Mediapunta	-7,79533	10,8
352	Nahitán Nández	Mediocentro	24,7165	24,0
301	Teun Koopmeiners	Mediocentro	5,70042	15,0
390	Leandro Trossard	Mediapunta	-9,30727	16,1
369	Marc Cucurella	Mediapunta	-5,27514	21,0
300	Pablo Rosario	Mediocentro	0,281267	7,0
307	Joey Veerman	Mediocentro	5,85791	6,0
367	Rubén García	Mediapunta	5,87805	10,5

Tabla 8: Tabla PCA y Clustering para el subgrupo 'Centrocampista' con parecido a Alejandro Gómez y Óscar Rodríguez (Sevilla FC)

TABLA PCA Y CLUSTERING – JUGADORES 7-8: OUSSAMA IDRISSI Y SUSO				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
604	Oussama Idrissi	Extremo izquierdo	3,07888	8,0
605	Suso	Extremo derecho	4,35815	19,4
478	Portu	Extremo derecho	0,205262	22,0
548	Giovanni Simeone	Delantero	9,73289	17,6
515	Calvin Stengs	Extremo derecho	0,916882	16,0
484	Brais Méndez	Extremo derecho	-4,85281	11,6
530	Nikolai Laursen	Extremo derecho	0,521137	6,25
490	Óscar Plano	Extremo derecho	17,0066	3,1
482	Jorge de Frutos	Extremo derecho	0,513649	3,6
522	Benjamin Nygren	Extremo derecho	9,30772	2,5

Tabla 9: Tabla PCA y Clustering para el subgrupo 'Atacante' con parecido a Oussama Idrissi y Suso (Sevilla FC)

Una vez analizados todos los jugadores, se procederá a realizar la tabla final donde se recogen, para cada uno de los futbolistas, los dos que, perteneciendo al mismo cluster, tienen un valor de sumatorio de PCA más cercano, obteniéndose la siguiente tabla (tabla 10):

TABLA FINAL – CASO DE ESTUDIO I (SEVILLA FC)							
Yassine Bounou	Karim Rekik	Marcos Acuña	Ivan Rakitic	Alejandro Gómez	Óscar Rodríguez	Oussama Idrissi	Suso
Hugo Lloris	Brendan Chardonnet	Rayan Aït-Nouri	Ryan Gravenberch	Leandro Trossard	Marc Cucurella	Calvin Stengs	Jorge de Frutos
Alphonse Areola	Nicolas Pallois	Aaron Creswell	Leandro Barreiro	Denis Suárez	Pablo Rosario	Nikolai Laursen	Portu

Tabla 10: Tabla final de jugadores para las planificaciones alternativas (Sevilla FC)

Hay que puntualizar que, tanto para este caso como para el siguiente, se da el caso de que, para dos jugadores cuyas tablas se hayan hecho juntas (aquellos jugadores que, además de jugar en la misma posición,

pertenezcan al mismo cluster), haya un jugador propuesto que tenga un valor de PCA cercano a ambos. Lo que se ha hecho ha sido añadirlo a uno de los jugadores y, para no caer en repetición, añadir otro nuevo futbolista a la lista definitiva.

6.4.2 Caso de estudio II: Club de presupuesto medio-bajo

De igual forma que en el caso de estudio anterior, se va a coger como ejemplo la planificación deportiva de dicho equipo en la temporada anterior para realizar una nueva planificación apoyándonos en los algoritmos utilizados.

Para este caso de estudio se ha escogido al Granada, ejemplo de buena gestión de la dirección deportiva, ya que ha conseguido unos resultados deportivos muy superiores a las aspiraciones del club. Concretamente, el Granada ha invertido el último año alrededor de 20 millones de euros en fichajes (21,7M€), teniendo un presupuesto general aproximado de 70 millones de euros. En la última columna de la siguiente tabla, se añade también el cluster al que pertenece cada futbolista una vez aplicado el algoritmo de clustering en Python.

Según (*Granada CF - Fichajes 20/21 | Transfermarkt*, 2021), los fichajes realizados por el Granada CF en la temporada 2020/2021 fueron (no se considerarán cesiones ni jugadores libres, sólo jugadores que han supuesto una inversión de cara a su adquisición en materia de traspaso) los siguientes (tabla 11):

PRESUPUESTO DESTINADO A FICHAJES: 22,3M€			
Posiciones	**Jugador fichado**	**Valor de mercado (M€)**	**Cluster al que pertenece**
Lateral derecho	Dimitri Foulquier	3,3	3
Pivote	Maxime Gonalons	3,3	2
Mediocentro	Luis Milla	3,6	1
Extremo derecho	Alberto Soro	3,1	3
Delantero	Luis Javier Suárez	9,0	3

Tabla 11: Fichajes realizados por el Granada CF (Temporada 2020/2021)

Una vez definidos los jugadores adquiridos por el Granada CF, su correspondiente valor de mercado y el cluster al que pertenece cada futbolista, se definen listas de jugadores que pueden ser alternativa de adquisición. Estos jugadores deben pertenecer al mismo cluster. Para no desvirtuar el análisis y una posterior combinación, se van a buscar jugadores los cuales su valor de mercado no suponga una inversión mayor al 50% del total presupuestado para la partida de fichajes. A continuación, se definen las tablas (tabla 12 hasta la tabla 15) donde se aplicarán los algoritmos de Clustering y PCA para cada uno de los jugadores, siendo en este caso 5 jugadores los sometidos a análisis.

TABLA PCA Y CLUSTERING – JUGADOR 1: DIMITRI FOULQUIER				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
67	Dimitri Foulquier	Lateral derecho	5,36441	3,3
111	Jens Stryger Larsen	Lateral derecho	11,5463	3,2
128	Ezgjan Alioski	Lateral izquierdo	-0,7876	4,2
112	Gaetano Letizia	Lateral derecho	-4,19586	2,2
73	Damián Suárez	Lateral derecho	18,6206	2,1
106	Darko Lazovic	Lateral derecho	1,48028	5,3
74	Isaac Carcelén	Lateral derecho	-2,03268	1,7
77	Luis Pérez	Lateral derecho	-7,57016	0,985
187	Jeison Murillo	Central	-10,7708	5,8
237	Riccardo Gagliolo	Central	2,40059	3,1

Tabla 12: Tabla PCA y Clustering para el subgrupo 'Defensa' con parecido a Dimitri Foulquier (Granada CF)

TABLA PCA Y CLUSTERING – JUGADOR 2: MAXIME GONALONS				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
267	Maxime Gonalons	Pivote	1,78802	3,3
264	Vicente Iborrra	Pivote	-5,77122	3,2
386	Mateusz Klich	Mediapunta	1,52772	3,4
362	Gonzalo Melero	Mediapunta	11,2745	4,9
343	Adrian Tameze	Mediocentro	-2,57922	5,4
300	Pablo Rosario	Mediocentro	0,281267	7,0
266	Mickaël Malsa	Pivote	15,334	2,5
345	Milan Badelj	Mediocentro	8,984	1,5
351	Sasa Lukic	Mediocentro	1,26966	4,7
347	Giulio Maggiore	Mediocentro	-0,0334	2,3

Tabla 13: Tabla PCA y Clustering para el subgrupo 'Centrocampista' con parecido a Maxime Gonalons (Granada CF)

TABLA PCA Y CLUSTERING – JUGADOR 3: LUIS MILLA				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
606	Luis Milla	Mediocentro	-19,5028	3,6
283	Cheick Doucouré	Pivote	-6,17526	3,7
294	Hicham Boudaoui	Pivote	11,2123	3,8
297	Lamine Fomba	Pivote	-10,4571	4,2
284	Digbo Maiga	Pivote	9,16934	5,3
258	John Lundstram	Pivote	6,33907	8,3
287	Jordan Ferri	Pivote	2,98019	3,3
292	Moreto Cassama	Pivote	13,1991	2,4
290	Paul Lasne	Pivote	7,43797	1,8
295	Pedro Chirivella	Pivote	-8,11	2,7

Tabla 14: Tabla PCA y Clustering para el subgrupo ‘Centrocampista’ con parecido a Luis Milla (Granada CF)

TABLA PCA Y CLUSTERING – JUGADORES 4-5: ALBERTO SORO Y LUIS JAVIER SUÁREZ				
Índice	**Nombre**	**Posición**	**Valor PCA combinado**	**Valor de mercado (M€)**
607	Alberto Soro	Extremo derecho	4,7563	3,1
576	Luis Javier Suárez	Delantero	-4,43561	9,0
488	Salvi	Extremo derecho	-7,23583	2,5
546	M’bala Nzola	Delantero	0,65211	3,9
461	Martin Terrier	Extremo izquierdo	6,98963	11,0
493	David Ferreiro	Extremo derecho	-10,4787	1,0
516	Matús Bero	Extremo derecho	-3,25816	1,6
482	Jorge de Frutos	Extremo derecho	0,513649	3,6
587	Lucas Boyé	Delantero	-15,0474	1,9
580	Jonathan Calleri	Delantero	-4,51537	5,2

Tabla 15: Tabla PCA y Clustering para el subgrupo 'Atacante' con parecido a Alberto Soro y Luis Javier Suárez (Granada CF)

De igual manera que en el caso anterior, una vez estén analizados todos los jugadores de la planificación, se realiza la tabla (tabla 16) que recoge las dos opciones con un valor de PCA más cercano a cada uno de los jugadores.

TABLA FINAL – CASO DE ESTUDIO II (GRANADA CF)				
Dimitri Foulquier	Maxime Gonalons	Luis Milla	Alberto Soro	Luis Javier Suárez
Riccardo Gaggliolo	Mateusz Klich	Lamine Fomba	M'bala Nzola	Jonathan Calleri
Darko Lazovic	Sasa Lukic	Pedro Chirivella	Jorge de Frutos	Matús Bero

Tabla 16: Tabla final de jugadores para las planificaciones alternativas (Granada CF)

Como observación hay que puntualizar que, en muchas de las tablas que se han realizado, la posición del futbolista a analizar no coincide con los futbolistas que se han propuesto como alternativa a este a la hora de realizar la agrupación con clustering. Esto ocurre porque, aunque se haya considerado a un futbolista en una posición concreta, los datos recogidos en la BBDD pueden propiciar que se encuentren más similitudes en otra demarcación que no sea la suya.

Este hecho se puede observar en la tabla 7, en la que Ivan Rakitic es mediapunta, y todas las alternativas propuestas son jugadores que se desenvuelven en la posición mediocentro. Este hecho es consistente con la posición y el rol que Ivan tiene en el campo, ya que es un futbolista que interviene mucho en la jugada, por lo que, si se buscan otros jugadores que puedan parecerse a él, se buscará, por lo general, a un mediocentro antes que a un mediapunta. De hecho, en las posiciones del centro del campo (grupo 'Centrocampista'), no es raro que un jugador en un club juegue en una posición y en otro club juegue en otra, debido a la versatilidad de funciones que un centrocampista puede otorgar al juego del equipo.

Por el contrario, y como se ha definido al principio del presente apartado, con la posición de lateral, independientemente de que el cluster nos indique cierto parecido, no podríamos considerar a un lateral izquierdo para jugar en la posición de lateral derecho, ya que se busca que el futbolista, en esa posición, siempre juegue a pie natural. O lo que es lo mismo, si se busca un lateral derecho, no se puede considerar un lateral izquierdo en el análisis, ya que probablemente este sea zurdo, hecho que hace que no se pueda desenvolver bien en el lateral del lado opuesto.

6.5 Análisis económico de los resultados

El enfoque de este proyecto es netamente deportivo, pero este hecho no es óbice para que se le dé un importante enfoque desde una perspectiva económica, ya que realizar todo este proceso de planificación deportiva, como se ha comprobado, debe estar supeditado al presupuesto de la partida de fichajes.

A partir de las tablas finales de los dos casos de estudio, se procede a dar propuestas de planificaciones alternativas que cumplan con la condición de que el valor total de mercado de todos los jugadores propuestos no supere el montante de dinero invertido en la planificación real de cada uno de los equipos.

Como se comentaba en el marco teórico del presente trabajo (apartado 3), es importante que el director deportivo tenga a su disposición una lista relativamente de opciones para cada una de las posiciones en las que está buscando un futbolista, de manera que no se pague un sobreprecio por la primera opción (que en este caso serían los jugadores que se han fichado en la planificación real).

Concretamente en este trabajo se ha optado por realizar una planificación alternativa completa en la que se combinan jugadores fichados en la realidad con jugadores propuestos por los algoritmos implementados para constatar la viabilidad económica del presente proyecto, que en un club deportivo es de vital importancia independientemente de su capacidad de inversión.

6.5.1 Propuestas de planificación alternativa para el Caso de estudio I

Siendo el presupuesto total dedicado a fichajes 92,8 millones de euros, se proponen tres planificaciones alternativas (tablas 17, 18 y 19). En dichas planificaciones alternativas pueden aparecer algunos de los jugadores fichados en la planificación real del club en la presente temporada.

- **Planificación alternativa 1:**

PLANIFICACIÓN 1		
Jugador	**Posición**	**Valor de Mercado (M€)**
Yassine Bounou	Portero	14,5
Nicolas Pallois	Central	2,1
Aaron Creswell	Lateral izquierdo	6,9
Ivan Rakitic	Mediocentro	11,0
Marc Cucurella	Mediapunta	21,0
Óscar Rodríguez	Mediapunta	10,9
Nikolai Laursen	Extremo derecho	6,25
Suso	Extremo derecho	19,4
	GASTO TOTAL	**92,05 M€**

Tabla 17: Planificación alternativa 1 (Sevilla FC)

- **Planificación alternativa 2:**

PLANIFICACIÓN 2		
Jugador	**Posición**	**Valor de Mercado (M€)**
Hugo Lloris	Portero	11,6
Karim Rekik	Central	2,9
Marcos Acuña	Lateral izquierdo	13,5
Leandro Barreiro	Mediocentro	7,3
Denis Suárez	Mediapunta	10,8
Leandro Trossard	Mediapunta	16,1
Oussama Idrissi	Extremo izquierdo	8,0
Portu	Extremo derecho	22,0
	GASTO TOTAL	**92,20 M€**

Tabla 18: Planificación alternativa 2 (Sevilla FC)

- **Planificación alternativa 3:**

PLANIFICACIÓN 3		
Jugador	**Posición**	**Valor de Mercado (M€)**
Alphonse Aréola	Portero	12,1
Brendan Chardonnet	Central	2,1
Rayan Aït-Nouri	Lateral izquierdo	19,4
Ryan Gravenberch	Mediocentro	22,0
Alejandro Gómez	Mediapunta	9,6
Pablo Rosario	Mediocentro	7,0
Calvin Stengs	Extremo derecho	16,0
Jorge de Frutos	Extremo derecho	3,6
	GASTO TOTAL	**91,80 M€**

Tabla 19: Planificación alternativa 3 (Sevilla FC)

Las tres planificaciones propuestas cumplen con el máximo exigido y, por tanto, a priori, son válidas para su propuesta al director deportivo.

En la tabla 17, resalta el hecho que por Oussama Idrissi (extremo izquierdo) se ha propuesto como alternativa el jugador Nikolai Laursen (extremo derecho), pero el jugador propuesto, como posición alternativa, tiene el extremo izquierdo, por lo que la propuesta del algoritmo es válida al poder jugar este futbolista en la posición que estamos buscando. De hecho, en la realidad, una de las cosas que más se valora, es la polivalencia del jugador, ya que dependiendo de las necesidades concretas del momento, es importante tener a un jugador que puedas disponer en una posición u otra y que ofrezca un rendimiento relativamente parecido en ambas.

En la tabla 19, por el contrario, se tiene que en la planificación real se ficha a un mediapunta (Óscar Rodríguez) y se ficha a un mediocentro (Pablo Rosario) en su lugar. Pasa algo parecido en este caso, ya que el jugador propuesto puede desempeñar tanto las funciones de mediocentro, como de mediapunta (e incluso, de pivote), por lo que es muy valorable que pueda jugar en casi todas las posiciones del centro del campo indistintamente.

6.5.2 Propuestas de planificación alternativa para el Caso de estudio II

Siendo la partida de fichajes de la presente temporada para el Granada CF de 22,3 millones de euros, se ofrecen, igual que en el caso anterior, tres planificaciones alternativas a la realizada en la realidad (tablas 20, 21 y 22). De igual manera que anteriormente, se pueden introducir futbolistas que son objeto de compra en dicha planificación.

Estas planificaciones, al tener una inversión netamente inferior con respecto al caso de estudio anterior, resulta un poco más complicada porque hay menos variedad de jugadores a los que poder acceder por cuestiones económicas, pero es igualmente abordable el problema como se demuestra a continuación.

- **Planificación alternativa 1:**

PLANIFICACIÓN 1		
Jugador	**Posición**	**Valor de Mercado (M€)**
Darko Lazovic	Lateral derecho	5,3
Mateusz Klich	Mediapunta	3,4
Luis Milla	Mediocentro	3,6
M'Bala Nzola	Extremo derecho	3,9
Matús Bero	Extremo derecho	1,6
	GASTO TOTAL	**17,80 M€**

Tabla 20: Planificación alternativa 1 (Granada CF)

- **Planificación alternativa 2:**

PLANIFICACIÓN 2		
Jugador	**Posición**	**Valor de Mercado (M€)**
Riccardo Gaggliolo	Lateral derecho	3,1
Sasa Lukic	Mediocentro	4,7
Lamine Fomba	Pivote	4,2
Jorge de Frutos	Extremo derecho	3,6
Jonathan Calleri	Delantero	5,2
	GASTO TOTAL	**20,80 M€**

Tabla 21: Planificación alternativa 2 (Granada CF)

- **Planificación alternativa 3:**

PLANIFICACIÓN 3		
Jugador	**Posición**	**Valor de Mercado (M€)**
Dimitri Foulquier	Lateral derecho	3,3
Maxime Gonalons	Pivote	3,3
Pedro Chirivella	Pivote	2,7
Alberto Soro	Extremo derecho	3,1
Luis Javier Suárez	Delantero	9,0
	GASTO TOTAL	**21,40 M€**

Tabla 22: Planificación alternativa 3 (Granada CF)

Como se demuestra en las tres tablas anteriores, se cumple con los objetivos de gasto total al no superarse los 22,30M€ de la planificación real, por lo que son planificaciones válidas y ajustadas a la realidad económica del club.

De igual manera que previamente, se ha comprobado que los jugadores que se han añadido a las planificaciones alternativas a partir de la implementación de los algoritmos PCA y Clustering que difieren de lo que se buscaba en la planificación real, pueden jugar, también, en la posición que se buscaba en un principio.

7 CONCLUSIONES

El objetivo del presente trabajo, que era facilitar la labor de búsqueda de jugadores para poder dar alternativas a la dirección deportiva se ha conseguido, ya que se ha podido realizar un tratamiento de los datos de manera que estos nos otorguen unos resultados en forma de agrupación y similitud entre futbolistas.

En este trabajo se ha definido el scouting con el fin de situar la realidad de esta práctica en el mundo del fútbol, así como el caso en particular de un club con una dirección deportiva que marca la diferencia con respecto a las demás. Esto ha ayudado a situar teóricamente el trabajo y enmarcarlo dentro de la realidad futbolística en una materia que, además, está al alza.

Como se ha podido observar, los algoritmos de Machine Learning son muy potentes a la hora de interpretar los datos, por lo que los clubes, dentro de la medida de lo posible, a la hora de realizar adquisiciones, deben tener en cuenta que existen tecnologías relacionadas con la inteligencia artificial que les pueden ser de gran ayuda. Concretamente, en este trabajo se ha hecho uso de PCA y Clustering, que son los más utilizados en análisis relativos a cuestiones deportivas, en la que es vital poder categorizar a los jugadores a partir de datos recogidos de partidos anteriores.

Para poder llevar a cabo este análisis, se ha construido una base de datos amplia, con un gran número de variables, para poder analizarla y poder realizar descartes de aquellas variables que no fueran lo suficientemente significativas. Esto ha permitido que los algoritmos implementados puedan establecer relaciones entre los futbolistas a partir de variables que no tienen mucha correlación entre sí, lo que confiere a la base de datos una capacidad de análisis realmente exhaustiva.

Seguidamente, se ha hecho un análisis preliminar consistente en gráficas para poder realizar un estudio más visual, antes de implementar algoritmos, para observar, según las variables más importantes en cada caso, cuales son los valores más comunes dentro de una posición concreta y quiénes son los jugadores que más despuntan en dichos análisis.

Como resultado final, tras la implementación de los algoritmos de PCA y Clustering, se han aportado planificaciones alternativas como distintas soluciones a aportar al director deportivo cumpliendo, siempre, las restricciones económicas en función de la realidad del club que se está tratando en cada caso de estudio.

En el caso de este trabajo, la mayor limitación ha sido la de no disponer de una base de datos lo suficientemente extensa como para poder realizar un tratamiento de los datos y obtener unas conclusiones

medianamente verídicas, por lo que se optó por realizar una base de datos a mano, con jugadores de todas las posiciones y de distintas ligas para, así, poder tener mayor variedad.

No obstante, más que el tratamiento en sí de los datos mediante los algoritmos de programación, lo importante es pensar cómo modelar la base de datos de manera que los resultados que se obtengan sean lo más realistas y útiles para su análisis.

Como futuras líneas de investigación, se proponen las siguientes:

- Realizar una base de datos únicamente para una posición concreta, y aplicar los algoritmos de Machine Learning para esa posición únicamente.
- Realizar una base de datos solo para una liga o país concreto, restringiendo los jugadores que se van a analizar.
- Realizar una base de datos de un período superior al estudiado, tratando datos de temporadas anteriores.

Independientemente de las distintas soluciones que se puedan aportar, como se ha comentado al principio, el planteamiento, tratamiento y análisis de los distintos casos de estudio que se han planteado son correctos y realistas desde el punto de vista de lo que se realiza en un club de fútbol, como se ha podido ver en el marco teórico, por lo que la viabilidad de este proyecto en la realidad es plena.

8 BIBLIOGRAFÍA

About us — scikit-learn 0.24.1 documentation. (n.d.). Retrieved April 26, 2021, from https://scikit-learn.org/stable/about.html

Botello, J. (2012). *¿Qué es Scouting? ¿Qué es Scouting Deportivo? Accede y conoce este apasionante mundo del #ScoutingDeportivo*. Retrieved March 25, 2021, from https://www.scoutingdeportivo.com/Scouting-Deportivo-Jesus-Botello-Analista-Deportivo/

Cádiz CF. (2020). *Organigrama Cádiz Club de Fútbol S.A.D. | Cádiz CF - Web Oficial*. Retrieved March 11, 2021, from https://www.cadizcf.com/club/organigrama

Castro, N. (n.d.). *Introducción Data Science con Python*.

ESPN. (2018, April 8). *¿Cuándo inician las ligas más importantes en Europa?*. Retrieved March 26, 2021, from https://espndeportes.espn.com/futbol/nota/_/id/4555427/cuando-inician-las-ligas-mas-importantes-en-europa

FundéuRAE. (2018). Retrieved March 09, 2021, from *«aprendizaje automático», mejor que «machine learning» | Fundéu*. https://www.fundeu.es/recomendacion/aprendizaje-automatico-mejor-que-machine-learning/

González Duque, R. (2018). *Python para Todos* (pp. 7–9).

Gonzalo, Á. (2019). *Segmentación utilizando K-means en Python*. Retrieved May 25, 2021, from https://machinelearningparatodos.com/segmentacion-utilizando-k-means-en-python/

Granada CF - Fichajes 20/21 | Transfermarkt. (2021). Retrieved June 14, 2021, from https://www.transfermarkt.es/fc-granada/transfers/verein/16795/saison_id/2020

Gutiérrez Moya, E. (2016). *Lecciones de fiabilidad industrial*. Sección de Publicaciones, Escuela Técnica Superior de Ingeniería.

Herráez, B. (n.d.). *Entrenadores de Futbol - www.entrenadores.info*. Retrieved April 18, 2021, from http://www.escoladefutbol.com/beto/docs/sist_f11/sist_f11.htm

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning IBM Limited Edition*. Retrieved March 22, 2021, from http://www.wiley.com/go/permissions.

Interpretar todos los estadísticos y gráficas para Análisis de elementos - Minitab. (n.d.). Retrieved April 13, 2021, from https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/item-analysis/interpret-the-results/all-statistics-and-graphs/

Kubat, M. (2017). An Introduction to Machine Learning. In *An Introduction to Machine Learning*. Springer International Publishing. Retrieved March 24, 2021, from https://doi.org/10.1007/978-3-319-63913-0

Martínez Heras, J. (2021, January 22). *¿Cómo aprende la Inteligencia Artificial? - IArtificial.net*. Retrieved March 23, 2021, from https://www.iartificial.net/como-aprende-la-inteligencia-artificial/#Aprendizaje_No_Supervisado

Matplotlib: Python plotting — Matplotlib 3.4.1 documentation. (n.d.). Retrieved April 26, 2021, from https://matplotlib.org/

NumPy. (n.d.). Retrieved April 26, 2021, from https://numpy.org/

Poli, R., Besson, R., Ravenel, L., & Gonzalez, T. (2020). *Weekly Post 324*. Retrieved March 26, 2021, from https://football-observatory.com/IMG/sites/b5wp/2020/wp324/en/

Qué es Big Data | Universidad Complutense de Madrid. (n.d.). Retrieved March 22, 2021, from https://www.masterbigdataucm.com/que-es-big-data/

RAE. (n.d.). *Definición de big data - Diccionario panhispánico del español jurídico - RAE*. Retrieved March 9, 2021, from https://dpej.rae.es/lema/big-data

Recuero de los Santos, P. (2017, November 16). *Machine learning: conoce qué es y las diferencias entre sus tipos*. Retrieved March 23, 2021, from https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/

Segunda División de España - Wikipedia, la enciclopedia libre. (n.d.). Retrieved March 26, 2021, from https://es.wikipedia.org/wiki/Segunda_División_de_España

Sevilla FC. (2020a). *Estructura Organizativa | Sevilla FC*. Retrieved March 11, 2021, from https://www.sevillafc.es/el-club/la-entidad/estructura-organizativa

Sevilla FC. (2020b). *MONCHI 13 - Masterclass: el factor Suerte (I) - YouTube*.

https://www.youtube.com/watch?v=y9pmDtZY3PY

Sevilla FC. (2020c, April 27). *MONCHI 13 Masterclass: seguimiento en bruto (III) - YouTube*. Retrieved March 19, 2021, from https://www.youtube.com/watch?v=hjv2Pyeaczo

Sevilla FC. (2020d, May 4). *MONCHI 13 Masterclass: seguimiento en neto (IV) - YouTube*. Retrieved March 19, 2021, from https://www.youtube.com/watch?v=M5uot8QS5Eg

Sevilla FC. (2020e, May 5). *MONCHI 13 masterclass : afinando al máximo (V) - YouTube*. Retrieved March 19, 2021, from https://www.youtube.com/watch?v=fqNsISuT-IQ&t=3s

Sevilla FC. (2020f, May 12). *MONCHI 13 Masterclass: la negociación (VII) - YouTube*. Retrieved March 19, 2021, from https://www.youtube.com/watch?v=qPTfaI7Rmhk&t=620s

Sevilla FC - Fichajes 20/21 | Transfermarkt. (2021). Retrieved June 14, 2021, from https://www.transfermarkt.es/sevilla-fc/transfers/verein/368/plus/?saison_id=2020&pos=&detailpos=&w_s=

SofaScore: The Fastest Football Scores and Livescore for 2021. (n.d.). Retrieved March 30, 2021, from https://www.sofascore.com/

Uc3m. (n.d.). *Tema 3: Análisis de Componentes Principales*.

9 ANEXO

9.1 Código para realizar las gráficas

```
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

DatosCompleto = pd.read_csv('BaseDeDatosDef.csv',encoding='UTF-
8',header=0,sep=';',decimal=',')
# DatosCompleto lee la base de datos CSV realizada de forma completa
Datos = DatosCompleto[['Liga','Equipo','Partidos_jugados',
      'Promedio_minutos','Minutos_jugados_por_el_equipo','Posición',
      'Nacionalidad',   'Altura',   'Edad',     'Polivalencia',
      'Segunda_Posición_Preferida', 'Pie_preferido',  'Goles_totales',
      'Disparos_por_partido', 'Asistencias',    'Toques_por_partido',
      'Ocasiones_creadas',    'Completados_porc'
      ,'Completados_en_campo_propio',     'Completados_en_campo_propio_porc',
      'Completados_en_campo_contrario',
      'Completados_en_campo_contrario_porc',    'Balones_largos_por_partido',
      'Balones_largos_completados_porc', 'Centros_completados',
      'Regates_por_partido',  'Regates_por_partido_porc',
      'Duelos_ganados_por_partido', 'Duelos_ganados_por_partido_porc',
      'Penaltis_cometidos',   'Porterías_a_cero'      ,'Posesión_perdida',
      'Intercepciones_por_partido', 'Entradas_por_partido'
      ,'Posesión_recuperada', 'Regateado_por_partido' ,'Despejes_por_partido',
      'Faltas_cometidas_por_partido'      ,'Faltas_recibidas_por_partido',
      'Amarillas' ,'Rojas',   'Paradas',  'Anticipación',   'Táctico',
      'Distribución_de_balón',      'Juego_aéreo',    'Ataque',   'Técnica',
      'Defensa',  'Creatividad',    'Valor_de_Mercado',
      'Goles_totales_en_contra'     ,     'Penaltis_en_contra',
      'Penaltis_parados'      ,     'Paradas_realizadas',
      'Paradas_por_partido_porc',   'Salidas_por_partido'
      ,'Paradas_con_balón_atrapado','Parada_con_despeje']]
# Datos recoge, a partir de DatosCompleto, únicamente las variables que NO
son combinación lineal de otras


' ---------- Gráficas Portero ---------- '
# Scatterplot 3D
DatosPortero = DatosCompleto[DatosCompleto['Posición']=='Portero']
x11 = DatosPortero.Paradas_realizadas
y11 = DatosPortero.Penaltis_parados_porc
z11 = DatosPortero.Valor_de_Mercado
NombrePortero = DatosPortero.Nombre
fig11=plt.figure()
ax11=fig11.add_subplot(projection='3d')
ax11.scatter(x11,y11,z11)
ax11.set_xlabel('Paradas realizadas',fontsize=10)
ax11.set_ylabel('Penaltis parados (%)',fontsize=10)
ax11.set_zlabel('Valor de Mercado',fontsize=10)
plt.title('Scatterplot 3D -- Portero')
plt.show(fig11)
```

```
# Histograma
x12 = DatosPortero.Paradas_Goles_porc
fig12 = plt.hist(x12)
plt.ylabel('Número')
plt.xlabel('Goles/Paradas(%)')
plt.title('Histograma -- Portero')
plt.show(fig12)

# Scatterplot 2D
x13 = DatosPortero.Paradas_realizadas
y13 = DatosPortero.Goles_totales_en_contra
fig13,ax13=plt.subplots()
ax13.scatter(x13,y13)
plt.xlabel('Paradas realizadas')
plt.ylabel('Goles totales en contra')
plt.title('Scatterplot 2D -- Portero')
plt.show(fig13)


' ---------- Gráficas Central ---------- '
# Scatterplot 3D
DatosCentral = DatosCompleto[DatosCompleto['Posición']=='Central']
NombreCentral = DatosCentral.Nombre
x21 = DatosCentral.Intercepciones_por_partido
y21 = DatosCentral.Posesión_recuperada
z21 = DatosCentral.Valor_de_Mercado/100000000
fig21 = plt.figure()
ax21 = fig21.add_subplot(projection='3d')
ax21.scatter(x21,y21,z21)
ax21.set_xlabel('Intercepciones por partido',fontsize=10)
ax21.set_ylabel('Posesión recuperada',fontsize=10)
ax21.set_zlabel('Valor de Mercado',fontsize=10)
plt.title('Scatterplot 3D -- Central')
plt.show(fig21)

# Histograma
x22 = DatosCentral.Duelos_ganados_por_partido
fig22 = plt.hist(x22)
plt.ylabel('Número')
plt.xlabel('Duelos ganados por partido')
plt.title('Histograma -- Central')
plt.show(fig22)

# Scatterplot 2D
x23 = DatosCentral.Balones_largos_por_partido
y23 = DatosCentral.Posesión_perdida
fig23,ax23=plt.subplots()
ax23.scatter(x23,y23)
plt.xlabel('Balones largos por partido')
plt.ylabel('Posesión perdida')
plt.title('Scatterplot 2D -- Central')
plt.show(fig23)


' ---------- Gráficas Mediocentro ---------- '
# Scatterplot 3D
DatosMediocentro = DatosCompleto[DatosCompleto['Posición']=='Mediocentro']
NombreMediocentro = DatosMediocentro.Nombre
x31 = DatosMediocentro.Completados_porc
```

```
y31 = DatosMediocentro.Pases_clave_por_partido
z31 = DatosMediocentro.Valor_de_Mercado/1e8
fig31 = plt.figure()
ax31 = fig31.add_subplot(projection='3d')
ax31.scatter(x31,y31,z31)
ax31.set_xlabel('Pases completados',fontsize=10)
ax31.set_ylabel('Pases clave por partido',fontsize=10)
ax31.set_zlabel('Valor de Mercado',fontsize=10)
plt.title('Scatterplot 3D -- Mediocentro')
plt.show(fig31)

# Histograma
x32 = DatosMediocentro.Pases_clave_por_partido
fig32 = plt.hist(x32)
plt.ylabel('Número')
plt.xlabel('Pases Clave por Partido')
plt.title('Histograma -- Mediocentro')
plt.show(fig32)

# Scatterplot 2D
x33 = DatosMediocentro.Completados_en_campo_propio
y33 = DatosMediocentro.Completados_en_campo_contrario
fig33,ax33=plt.subplots()
ax33.scatter(x33,y33)
plt.xlabel('Pases completados en campo propio')
plt.ylabel('Pases completados en campo contrario')
plt.title('Scatterplot 2D -- Mediocentro')
plt.show(fig33)



' ---------- Gráficas Delantero ---------- '
# Scatterplot 3D
DatosDelantero = DatosCompleto[DatosCompleto['Posición']=='Delantero']
NombreDelantero = DatosDelantero.Nombre
x41 = DatosDelantero.Goles_por_partido
y41 = DatosDelantero.Asistencias
z41 = DatosDelantero.Valor_de_Mercado/1e8
fig41 = plt.figure()
ax41 = fig41.add_subplot(projection='3d')
ax41.scatter(x41,y41,z41)
ax41.set_xlabel('Goles por partido',fontsize=10)
ax41.set_ylabel('Asistencias',fontsize=10)
ax41.set_zlabel('Valor de Mercado',fontsize=10)
plt.title ('Scatterplot 3D -- Delantero')
plt.show(fig41)

# Histograma
x42 = DatosDelantero.Goles_por_partido
fig42 = plt.hist(x42)
plt.ylabel('Número')
plt.xlabel('Goles por partido')
plt.title('Histograma -- Delantero')
plt.show(fig42)


# Scatterplot 2D
x43 = DatosDelantero.Disparos_por_partido
y43 = DatosDelantero.Regates_por_partido
fig43,ax43=plt.subplots()
```

```
ax43.scatter(x43,y43)
plt.xlabel('Disparos por partido')
plt.ylabel('Regates por partido')
plt.title('Scatterplot 2D -- Delantero')
plt.show(fig43)
```

9.2 Código para implementar PCA

```
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import seaborn as sns
from sklearn.decomposition import PCA
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import scale

DatosCompleto = pd.read_csv('BaseDeDatosDef.csv',encoding='UTF-
8',header=0,sep=';',decimal=',')

Datos = DatosCompleto[['Partidos_jugados',
      'Promedio_minutos','Minutos_jugados_por_el_equipo','Posición','Nacional
idad','Altura',   'Edad',     'Polivalencia',
      'Segunda_Posición_Preferida','Pie_preferido',   'Goles_totales',
      'Disparos_por_partido', 'Asistencias',    'Toques_por_partido',
      'Ocasiones_creadas',    'Completados_porc'
      ,'Completados_en_campo_propio',     'Completados_en_campo_propio_porc',
      'Completados_en_campo_contrario',
      'Completados_en_campo_contrario_porc',    'Balones_largos_por_partido',
      'Balones_largos_completados_porc',  'Centros_completados',
      'Regates_por_partido',  'Regates_por_partido_porc',
      'Duelos_ganados_por_partido', 'Duelos_ganados_por_partido_porc',
      'Penaltis_cometidos',   'Porterías_a_cero'      ,'Posesión_perdida',
      'Intercepciones_por_partido', 'Entradas_por_partido'
      ,'Posesión_recuperada', 'Regateado_por_partido' ,'Despejes_por_partido',
      'Faltas_cometidas_por_partido'      ,'Faltas_recibidas_por_partido',
      'Amarillas' ,'Rojas',   'Paradas',  'Anticipación',   'Táctico',
      'Distribución_de_balón',      'Juego_aéreo',    'Ataque',   'Técnica',
      'Defensa',  'Creatividad',    'Valor_de_Mercado',
      'Goles_totales_en_contra'     ,     'Penaltis_en_contra',
      'Penaltis_parados'      ,     'Paradas_realizadas',
      'Paradas_por_partido_porc',   'Salidas_por_partido'
      ,'Paradas_con_balón_atrapado','Parada_con_despeje']]
Datos = pd.get_dummies(Datos)
X_cols = list(Datos.columns.values)
ColPosicion = DatosCompleto.filter(["Posición"])
DatosPosicion = pd.concat([Datos,ColPosicion], axis=1)



' ---------- ANÁLISIS POSICIÓN: PORTERO ---------- '
# Cogemos de la tabla que acabamos de hacer, solamente los porteros
DP = DatosPosicion[DatosPosicion['Posición']=='Portero']
# De la tabla inicial, cogemos los datos solamente de los porteros
NP = DatosCompleto[DatosCompleto['Posición']=='Portero']
```

```
# De la tabla NP cogemos los nombres de los porteros
NombrePortero = pd.DataFrame(NP.Nombre)
# Borramos la columna posición ya que no es relevante una vez cogidos ya los
jugadores de la posición que buscábamos
DatosPortero = DP.drop(["Posición"], axis=1)


' ---------- PCA: PORTERO ----------'
ss = StandardScaler()
DatosPortero[X_cols] = ss.fit_transform(DatosPortero[X_cols])
PCAP = PCA(n_components=20)
PCA_P = PCAP.fit_transform(DatosPortero[X_cols])
ResP = pd.DataFrame(data=PCA_P,
columns=['PCA1','PCA2','PCA3','PCA4','PCA5','PCA6','PCA7','PCA8','PCA9','PCA1
0','PCA11','PCA12','PCA13','PCA14','PCA15','PCA16','PCA17','PCA18','PCA19','P
CA20'])
SumaP = ResP.sum(axis=1)
VarP = PCAP.explained_variance_ratio_
VarSumP = PCAP.explained_variance_ratio_.sum()
fig1 =
sns.barplot(x=['1','2','3','4','5','6','7','8','9','10','11','12','13','14','
15','16','17','18','19','20'], y=VarP)
plt.xlabel("PCA")
plt.ylabel("Varianza de cada PCA")
plt.title("Varianza explicada por cada PCA - Porteros")
plt.show(fig1)


' ---------- ANÁLISIS POSICIÓN: DEFENSA ---------- '
DLD = DatosPosicion[DatosPosicion['Posición']=='Lateral derecho']
DC = DatosPosicion[DatosPosicion['Posición']=='Central']
DLI = DatosPosicion[DatosPosicion['Posición']=='Lateral izquierdo']
NLD = DatosCompleto[DatosCompleto['Posición']=='Lateral derecho']
NC = DatosCompleto[DatosCompleto['Posición']=='Central']
NLI = DatosCompleto[DatosCompleto['Posición']=='Lateral izquierdo']
NombreLatDerecho = pd.DataFrame(NLD.Nombre)
NombreCentral = pd.DataFrame(NC.Nombre)
NombreLatIzquierdo = pd.DataFrame(NLI.Nombre)
NombreDefensa =
pd.concat([NombreLatDerecho,NombreLatIzquierdo,NombreCentral], axis=0)
DatosLatDerecho = DLD.drop(["Posición"], axis=1)
DatosCentral = DC.drop(["Posición"], axis=1)
DatosLatIzquierdo = DLI.drop(["Posición"], axis=1)
DatosDefensa = pd.concat([DatosLatDerecho,DatosLatIzquierdo,DatosCentral],
axis=0)

' ---------- PCA: DEFENSA ----------'
DatosDefensa[X_cols] = ss.fit_transform(DatosDefensa[X_cols])
PCAD = PCA(n_components=47)
PCA_D = PCAD.fit_transform(DatosDefensa[X_cols])
ResD = pd.DataFrame(data=PCA_D,
columns=['PCA1','PCA2','PCA3','PCA4','PCA5','PCA6','PCA7','PCA8','PCA9','PCA1
0','PCA11','PCA12','PCA13','PCA14','PCA15','PCA16','PCA17','PCA18','PCA19','P
CA20','PCA21','PCA22','PCA23','PCA24','PCA25','PCA26','PCA27','PCA28','PCA29'
,'PCA30','PCA31','PCA32','PCA33','PCA34','PCA35','PCA36','PCA37','PCA38','PCA
39','PCA40','PCA41','PCA42','PCA43','PCA44','PCA45','PCA46','PCA47'])
SumaD = ResD.sum(axis=1)
VarD = PCAD.explained_variance_ratio_
VarSumD = PCAD.explained_variance_ratio_.sum()
```

```
fig2 =
sns.barplot(x=['1','2','3','4','5','6','7','8','9','10','11','12','13','14','
15','16','17','18','19','20','21','22','23','24','25','26','27','28','29','30
','31','32','33','34','35','36','37','38','39','40','41','42','43','44','45',
'46','47'], y=VarD)
plt.xlabel("PCA")
plt.ylabel("Varianza de cada PCA")
plt.title("Varianza explicada por cada PCA - Defensas")
plt.show(fig2)



' ---------- ANÁLISIS POSICIÓN: CENTROCAMPISTA ---------- '
DPI = DatosPosicion[DatosPosicion['Posición']=='Pivote']
DMC = DatosPosicion[DatosPosicion['Posición']=='Mediocentro']
DMP = DatosPosicion[DatosPosicion['Posición']=='Mediapunta']
NPI = DatosCompleto[DatosCompleto['Posición']=='Pivote']
NMC = DatosCompleto[DatosCompleto['Posición']=='Mediocentro']
NMP = DatosCompleto[DatosCompleto['Posición']=='Mediapunta']
NombrePivote = pd.DataFrame(NPI.Nombre)
NombreMediocentro = pd.DataFrame(NMC.Nombre)
NombreMediapunta = pd.DataFrame(NMP.Nombre)
NombreCentrocampista =
pd.concat([NombrePivote,NombreMediocentro,NombreMediapunta], axis=0)
DatosPivote = DPI.drop(["Posición"], axis=1)
DatosMediocentro = DMC.drop(["Posición"], axis=1)
DatosMediapunta = DMP.drop(["Posición"], axis=1)
DatosCentrocampista =
pd.concat([DatosPivote,DatosMediocentro,DatosMediapunta], axis=0)

' ---------- PCA: CENTROCAMPISTA ----------'
DatosCentrocampista[X_cols] = ss.fit_transform(DatosCentrocampista[X_cols])
PCAC = PCA(n_components=46)
PCA_C = PCAC.fit_transform(DatosCentrocampista[X_cols])
ResC = pd.DataFrame(data=PCA_C,
columns=['PCA1','PCA2','PCA3','PCA4','PCA5','PCA6','PCA7','PCA8','PCA9','PCA1
0','PCA11','PCA12','PCA13','PCA14','PCA15','PCA16','PCA17','PCA18','PCA19','P
CA20','PCA21','PCA22','PCA23','PCA24','PCA25','PCA26','PCA27','PCA28','PCA29'
,'PCA30','PCA31','PCA32','PCA33','PCA34','PCA35','PCA36','PCA37','PCA38','PCA
39','PCA40','PCA41','PCA42','PCA43','PCA44','PCA45','PCA46'])
SumaC = ResC.sum(axis=1)
VarC = PCAC.explained_variance_ratio_
VarSumC = PCAC.explained_variance_ratio_.sum()
fig3 =
sns.barplot(x=['1','2','3','4','5','6','7','8','9','10','11','12','13','14','
15','16','17','18','19','20','21','22','23','24','25','26','27','28','29','30
','31','32','33','34','35','36','37','38','39','40','41','42','43','44','45',
'46'], y=VarC)
plt.xlabel("PCA")
plt.ylabel("Varianza de cada PCA")
plt.title("Varianza explicada por cada PCA - Centrocampistas")
plt.show(fig3)


' ---------- ANÁLISIS POSICIÓN: ATACANTES ---------- '
DEI = DatosPosicion[DatosPosicion['Posición']=='Extremo izquierdo']
DED = DatosPosicion[DatosPosicion['Posición']=='Extremo derecho']
DDC = DatosPosicion[DatosPosicion['Posición']=='Delantero']
NEI = DatosCompleto[DatosCompleto['Posición']=='Extremo izquierdo']
NED = DatosCompleto[DatosCompleto['Posición']=='Extremo derecho']
```

```
NDC = DatosCompleto[DatosCompleto['Posición']=='Delantero']
NombreExtIzquierdo = pd.DataFrame(NEI.Nombre)
NombreExtDerecho = pd.DataFrame(NED.Nombre)
NombreDelantero = pd.DataFrame(NDC.Nombre)
NombreAtacante =
pd.concat([NombreExtIzquierdo,NombreExtDerecho,NombreDelantero], axis=0)
DatosExtIzquierdo = DEI.drop(["Posición"], axis=1)
DatosExtDerecho = DED.drop(["Posición"], axis=1)
DatosDelantero = DDC.drop(["Posición"], axis=1)
DatosAtacante = pd.concat([DatosExtIzquierdo,DatosExtDerecho,DatosDelantero],
axis=0)
' ---------- PCA: ATACANTE ----------'
DatosAtacante[X_cols] = ss.fit_transform(DatosAtacante[X_cols])
PCAA = PCA(n_components=52)
PCA_A = PCAA.fit_transform(DatosAtacante[X_cols])
ResA = pd.DataFrame(data=PCA_A,
columns=['PCA1','PCA2','PCA3','PCA4','PCA5','PCA6','PCA7','PCA8','PCA9','PCA1
0','PCA11','PCA12','PCA13','PCA14','PCA15','PCA16','PCA17','PCA18','PCA19','P
CA20','PCA21','PCA22','PCA23','PCA24','PCA25','PCA26','PCA27','PCA28','PCA29'
,'PCA30','PCA31','PCA32','PCA33','PCA34','PCA35','PCA36','PCA37','PCA38','PCA
39','PCA40','PCA41','PCA42','PCA43','PCA44','PCA45','PCA46','PCA47','PCA48','
PCA49','PCA50','PCA51','PCA52'])
SumaA = ResA.sum(axis=1)
VarA = PCAA.explained_variance_ratio_
VarSumA = PCAA.explained_variance_ratio_.sum()
fig4 =
sns.barplot(x=['1','2','3','4','5','6','7','8','9','10','11','12','13','14','
15','16','17','18','19','20','21','22','23','24','25','26','27','28','29','30
','31','32','33','34','35','36','37','38','39','40','41','42','43','44','45',
'46','47','48','49','50','51','52'], y=VarA)
plt.xlabel("PCA")
plt.ylabel("Varianza de cada PCA")
plt.title("Varianza explicada por cada PCA - Atacantes")
plt.show(fig4)
```

9.3 Código para implementar Clustering

```
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import scale
from sklearn.cluster import AgglomerativeClustering


DatosCompleto = pd.read_csv('BaseDeDatosDef.csv',encoding='UTF-
8',header=0,sep=';',decimal=',')
# DatosCompleto lee la base de datos CSV realizada de forma completa
DatosNoComb = DatosCompleto[['Liga','Equipo','Partidos_jugados',
      'Promedio_minutos','Minutos_jugados_por_el_equipo','Posición',
```

```
        'Nacionalidad',   'Altura',   'Edad',     'Polivalencia',
        'Segunda_Posición_Preferida','Pie_preferido',  'Goles_totales',
        'Disparos_por_partido','Asistencias',    'Toques_por_partido',
        'Ocasiones_creadas',    'Completados_porc'
        ,'Completados_en_campo_propio',     'Completados_en_campo_propio_porc',
        'Completados_en_campo_contrario',
        'Completados_en_campo_contrario_porc',    'Balones_largos_por_partido',
        'Balones_largos_completados_porc',  'Centros_completados',
        'Regates_por_partido',  'Regates_por_partido_porc',
        'Duelos_ganados_por_partido','Duelos_ganados_por_partido_porc',
        'Penaltis_cometidos',   'Porterías_a_cero'      ,'Posesión_perdida',
        'Intercepciones_por_partido','Entradas_por_partido'
        ,'Posesión_recuperada','Regateado_por_partido' ,'Despejes_por_partido',
        'Faltas_cometidas_por_partido'      ,'Faltas_recibidas_por_partido',
        'Amarillas' ,'Rojas',   'Paradas',  'Anticipación',   'Táctico',
        'Distribución_de_balón',      'Juego_aéreo',    'Ataque',   'Técnica',
        'Defensa',  'Creatividad', 'Goles_totales_en_contra'  ,
        'Penaltis_en_contra',   'Penaltis_parados'      ,
        'Paradas_realizadas',   'Paradas_por_partido_porc',
        'Salidas_por_partido'
        ,'Paradas_con_balón_atrapado','Parada_con_despeje']]


' ---------- Filtrado de variables ---------- '

Datos =
DatosNoComb[['Partidos_jugados','Promedio_minutos','Minutos_jugados_por_el_eq
uipo','Altura','Edad','Polivalencia','Goles_totales','Disparos_por_partido',
'Asistencias','Toques_por_partido','Ocasiones_creadas',
'Completados_porc','Completados_en_campo_propio',
'Completados_en_campo_propio_porc','Completados_en_campo_contrario',
'Completados_en_campo_contrario_porc','Balones_largos_por_partido',
'Balones_largos_completados_porc','Centros_completados',
'Regates_por_partido','Regates_por_partido_porc',
'Duelos_ganados_por_partido','Duelos_ganados_por_partido_porc',
'Penaltis_cometidos','Porterías_a_cero','Posesión_perdida',
'Intercepciones_por_partido','Entradas_por_partido','Posesión_recuperada',
'Regateado_por_partido','Despejes_por_partido',
'Faltas_cometidas_por_partido','Faltas_recibidas_por_partido',
'Amarillas','Rojas','Paradas','Anticipación','Táctico',
'Distribución_de_balón','Juego_aéreo','Ataque','Técnica',
'Defensa','Creatividad','Goles_totales_en_contra',
'Penaltis_en_contra','Penaltis_parados','Paradas_realizadas',
'Paradas_por_partido_porc','Salidas_por_partido',
'Paradas_con_balón_atrapado','Parada_con_despeje']]
# Datos recoge, a partir de DatosNoComb, sólamente las variables numéricas,
excluyendo las categóricas

Fuerte = list()
while len(Fuerte) != 2:
    MatrizCorrelacion = Datos.corr()
    print(len(MatrizCorrelacion))
    # MatrizCorrelacion recoge, a partir de Datos, sólamente las variables
      númericas, excluyendo las categóricas
    Pares = MatrizCorrelacion.unstack()
    Pares = Pares[abs(Pares)>0.7] # Elimino aquellos pares de variables con
      correlación menor a 0.7 en valor absoluto
    #Elimino los 1, asociados a la diagonal de la matriz de correlación
    df = pd.DataFrame(data=Pares,columns=['c1'])
    Fuerte = df[df.c1 != 1]
```

```
    #Ahora ordeno de mayor a menor valor absoluto
    Fuerte['Abs'] = Fuerte['c1'].abs()
    Orden = Fuerte.sort_values('Abs', ascending=False)
    print(Orden.index[0][0])
    del(Datos[Orden.index[0][0]])

ColPosicion = DatosCompleto.filter(["Posición"])
DatosPosicion = pd.concat([Datos,ColPosicion], axis=1)




' ---------- ANÁLISIS POSICIÓN: PORTERO ---------- '
# Cogemos de la tabla que acabamos de hacer, solamente los porteros
DP = DatosPosicion[DatosPosicion['Posición']=='Portero']
# De la tabla inicial, cogemos los datos solamente de los porteros
NP = DatosCompleto[DatosCompleto['Posición']=='Portero']
# De la tabla NP cogemos los nombres y el valor de mercado de los porteros
NombrePortero = pd.DataFrame(NP.Nombre)
ValorPortero = pd.DataFrame(NP.Valor_de_Mercado)
# Borramos la columna posición ya que no es relevante una vez cogidos ya los
jugadores de la posición que buscábamos
DatosPortero = DP.drop(["Posición"], axis=1)

' ---------- Método del codo ---------- '
Nc = range(2, 11)
kmeans = [KMeans(n_clusters=i) for i in Nc]
score = [kmeans[i].fit(DatosPortero).score(DatosPortero)
         for i in range(len(kmeans))]
fig1 = plt.figure()
fig1 = plt.plot(Nc,score)
plt.xlabel('Número de clusters')
plt.ylabel('Puntuación')
plt.title('Método del Codo -- Portero')
plt.grid()
plt.show()

' ---------- Análisis de la silueta ---------- '
range_n_clusters = range(2, 13)
valores_medios_silhouette = []
Datos_escalado = scale(DatosPortero)

for n_clusters in range_n_clusters:
    modelo = AgglomerativeClustering(
                    affinity   = 'euclidean',
                    linkage    = 'ward',
                    n_clusters = n_clusters
                )

    cluster_labels = modelo.fit_predict(Datos_escalado)
    silhouette_avg = silhouette_score(Datos_escalado, cluster_labels)
    valores_medios_silhouette.append(silhouette_avg)

fig2, ax = plt.subplots(1, 1, figsize=(6, 3.84))
ax.plot(range_n_clusters, valores_medios_silhouette, marker='o')
ax.set_title("Evolución de media de los índices silhouette -- Portero")
ax.set_xlabel('Número de clusters')
ax.set_ylabel('Media índices silhouette')
plt.grid()
```

```
' ---------- Algoritmo k-Means ---------- '
KMeans_Portero = KMeans(n_clusters=3).fit(DatosPortero)
Label_Portero = pd.DataFrame(KMeans_Portero.labels_, columns=['Cluster'])
Centroides_Portero = pd.concat([NombrePortero,Label_Portero+1,ValorPortero],
axis=1)
CP_Ordenado = Centroides_Portero.sort_values('Cluster')



' ---------- ANÁLISIS POSICIÓN: DEFENSA ---------- '
# Cogemos de la tabla que acabamos de hacer, solamente los defensas
DLD = DatosPosicion[DatosPosicion['Posición']=='Lateral derecho']
DC = DatosPosicion[DatosPosicion['Posición']=='Central']
DLI = DatosPosicion[DatosPosicion['Posición']=='Lateral izquierdo']
# De la tabla inicial, cogemos los datos solamente de los defensas
NLD = DatosCompleto[DatosCompleto['Posición']=='Lateral derecho']
NC = DatosCompleto[DatosCompleto['Posición']=='Central']
NLI = DatosCompleto[DatosCompleto['Posición']=='Lateral izquierdo']
# De las tablas NLD,NC,NLI cogemos los nombres y los valores de mercado de
los defensas
NombreLatDerecho = pd.DataFrame(NLD.Nombre)
NombreCentral = pd.DataFrame(NC.Nombre)
NombreLatIzquierdo = pd.DataFrame(NLI.Nombre)
ValorLatDerecho = pd.DataFrame(NLD.Valor_de_Mercado)
ValorCentral = pd.DataFrame(NC.Valor_de_Mercado)
ValorLatIzquierdo = pd.DataFrame(NLI.Valor_de_Mercado)
# Juntamos todos los nombres en una misma tabla
NombreDefensa =
pd.concat([NombreLatDerecho,NombreLatIzquierdo,NombreCentral], axis=0)
ValorDefensa = pd.concat([ValorLatDerecho,ValorLatIzquierdo,ValorCentral],
axis=0)
# Borramos la columna posición ya que no es relevante una vez cogidos ya los
jugadores de la posición que buscábamos
DatosLatDerecho = DLD.drop(["Posición"], axis=1)
DatosCentral = DC.drop(["Posición"], axis=1)
DatosLatIzquierdo = DLI.drop(["Posición"], axis=1)
# Juntamos todos los datos en una misma tabla
DatosDefensa = pd.concat([DatosLatDerecho,DatosLatIzquierdo,DatosCentral],
axis=0)

' ---------- Método del codo ---------- '
Nc = range(2, 11)
kmeans = [KMeans(n_clusters=i) for i in Nc]
score = [kmeans[i].fit(DatosDefensa).score(DatosDefensa)
         for i in range(len(kmeans))]
fig3 = plt.figure()
fig3 = plt.plot(Nc,score)
plt.xlabel('Número de clusters')
plt.ylabel('Puntuación')
plt.title('Método del Codo -- Defensa')
plt.grid()
plt.show()

' ---------- Análisis de la silueta ---------- '
range_n_clusters = range(2, 13)
valores_medios_silhouette = []
```

```
Datos_escalado = scale(DatosDefensa)

for n_clusters in range_n_clusters:
    modelo = AgglomerativeClustering(
                    affinity   = 'euclidean',
                    linkage    = 'ward',
                    n_clusters = n_clusters
                )

    cluster_labels = modelo.fit_predict(Datos_escalado)
    silhouette_avg = silhouette_score(Datos_escalado, cluster_labels)
    valores_medios_silhouette.append(silhouette_avg)
fig4, ax = plt.subplots(1, 1, figsize=(6, 3.84))
ax.plot(range_n_clusters, valores_medios_silhouette, marker='o')
ax.set_title("Evolución de media de los índices silhouette -- Defensa")
ax.set_xlabel('Número de clusters')
ax.set_ylabel('Media índices silhouette')
plt.grid()

' ---------- Algoritmo k-Means ---------- '
KMeans_Defensa = KMeans(n_clusters=3).fit(DatosDefensa)
Label_Defensa = pd.DataFrame(KMeans_Defensa.labels_, columns=['Cluster'])
Centroides_Defensa = pd.concat([NombreDefensa.reset_index(drop=True),
Label_Defensa+1, ValorDefensa.reset_index(drop=True)], axis=1)
CD_Ordenado = Centroides_Defensa.sort_values('Cluster')
```

```
' ---------- ANÁLISIS POSICIÓN: CENTROCAMPISTA ---------- '
# Cogemos de la tabla que acabamos de hacer, solamente los centrocampistas
DPI = DatosPosicion[DatosPosicion['Posición']=='Pivote']
DMC = DatosPosicion[DatosPosicion['Posición']=='Mediocentro']
DMP = DatosPosicion[DatosPosicion['Posición']=='Mediapunta']
# De la tabla inicial, cogemos los datos solamente de los centrocampistas
NPI = DatosCompleto[DatosCompleto['Posición']=='Pivote']
NMC = DatosCompleto[DatosCompleto['Posición']=='Mediocentro']
NMP = DatosCompleto[DatosCompleto['Posición']=='Mediapunta']
# De las tablas NLD,NC,NLI cogemos los nombres y los valores de mercado de
los centrocampistas
NombrePivote = pd.DataFrame(NPI.Nombre)
NombreMediocentro = pd.DataFrame(NMC.Nombre)
NombreMediapunta = pd.DataFrame(NMP.Nombre)
ValorPivote = pd.DataFrame(NPI.Valor_de_Mercado)
ValorMediocentro = pd.DataFrame(NMC.Valor_de_Mercado)
ValorMediapunta = pd.DataFrame(NMP.Valor_de_Mercado)
# Juntamos todos los nombres en una misma tabla
NombreCentrocampista =
pd.concat([NombrePivote,NombreMediocentro,NombreMediapunta], axis=0)
ValorCentrocampista =
pd.concat([ValorPivote,ValorMediocentro,ValorMediapunta], axis=0)
# Borramos la columna posición ya que no es relevante una vez cogidos ya los
jugadores de la posición que buscábamos
DatosPivote = DPI.drop(["Posición"], axis=1)
DatosMediocentro = DMC.drop(["Posición"], axis=1)
```

```
DatosMediapunta = DMP.drop(["Posición"], axis=1)
# Juntamos todos los datos en una misma tabla
DatosCentrocampista =
pd.concat([DatosPivote,DatosMediocentro,DatosMediapunta], axis=0)

' ---------- Método del codo ---------- '
Nc = range(2, 11)
kmeans = [KMeans(n_clusters=i) for i in Nc]
score = [kmeans[i].fit(DatosCentrocampista).score(DatosCentrocampista)
         for i in range(len(kmeans))]
fig5 = plt.figure()
fig5 = plt.plot(Nc,score)
plt.xlabel('Número de clusters')
plt.ylabel('Puntuación')
plt.title('Método del Codo -- Centrocampista')
plt.grid()
plt.show()

' ---------- Análisis de la silueta ---------- '
range_n_clusters = range(2, 13)
valores_medios_silhouette = []
Datos_escalado = scale(DatosCentrocampista)

for n_clusters in range_n_clusters:
    modelo = AgglomerativeClustering(
                    affinity   = 'euclidean',
                    linkage    = 'ward',
                    n_clusters = n_clusters
             )

    cluster_labels = modelo.fit_predict(Datos_escalado)
    silhouette_avg = silhouette_score(Datos_escalado, cluster_labels)
    valores_medios_silhouette.append(silhouette_avg)

fig6, ax = plt.subplots(1, 1, figsize=(6, 3.84))
ax.plot(range_n_clusters, valores_medios_silhouette, marker='o')
ax.set_title("Evolución de media de los índices silhouette --
Centrocampista")
ax.set_xlabel('Número de clusters')
ax.set_ylabel('Media índices silhouette')
plt.grid()

' ---------- Algoritmo k-Means ---------- '
KMeans_Centrocampista = KMeans(n_clusters=3).fit(DatosCentrocampista)
Label_Centrocampista = pd.DataFrame(KMeans_Centrocampista.labels_,
columns=['Cluster'])
Centroides_Centrocampista =
pd.concat([NombreCentrocampista.reset_index(drop=True),
Label_Centrocampista+1, ValorCentrocampista.reset_index(drop=True)], axis=1)
CC_Ordenado = Centroides_Centrocampista.sort_values('Cluster')




' ---------- ANÁLISIS POSICIÓN: ATACANTES ---------- '
```

```
# Cogemos de la tabla que acabamos de hacer, solamente los atacantes
DEI = DatosPosicion[DatosPosicion['Posición']=='Extremo izquierdo']
DED = DatosPosicion[DatosPosicion['Posición']=='Extremo derecho']
DDC = DatosPosicion[DatosPosicion['Posición']=='Delantero']
# De la tabla inicial, cogemos los datos solamente de los atacantes
NEI = DatosCompleto[DatosCompleto['Posición']=='Extremo izquierdo']
NED = DatosCompleto[DatosCompleto['Posición']=='Extremo derecho']
NDC = DatosCompleto[DatosCompleto['Posición']=='Delantero']
# De las tablas NLD,NC,NLI cogemos los nombres y los valores de mercado de
los atacantes
NombreExtIzquierdo = pd.DataFrame(NEI.Nombre)
NombreExtDerecho = pd.DataFrame(NED.Nombre)
NombreDelantero = pd.DataFrame(NDC.Nombre)
ValorExtIzquierdo = pd.DataFrame(NEI.Valor_de_Mercado)
ValorExtDerecho = pd.DataFrame(NED.Valor_de_Mercado)
ValorDelantero = pd.DataFrame(NDC.Valor_de_Mercado)
# Juntamos todos los nombres en una misma tabla
NombreAtacante =
pd.concat([NombreExtIzquierdo,NombreExtDerecho,NombreDelantero], axis=0)
ValorAtacante = pd.concat([ValorExtIzquierdo,ValorExtDerecho,ValorDelantero],
axis=0)
# Borramos la columna posición ya que no es relevante una vez cogidos ya los
jugadores de la posición que buscábamos
DatosExtIzquierdo = DEI.drop(["Posición"], axis=1)
DatosExtDerecho = DED.drop(["Posición"], axis=1)
DatosDelantero = DDC.drop(["Posición"], axis=1)
# Juntamos todos los datos en una misma tabla
DatosAtacante = pd.concat([DatosExtIzquierdo,DatosExtDerecho,DatosDelantero],
axis=0)

' ---------- Método del codo ---------- '
Nc = range(2, 11)
kmeans = [KMeans(n_clusters=i) for i in Nc]
score = [kmeans[i].fit(DatosAtacante).score(DatosAtacante)
         for i in range(len(kmeans))]
fig5 = plt.figure()
fig5 = plt.plot(Nc,score)
plt.xlabel('Número de clusters')
plt.ylabel('Puntuación')
plt.title('Método del Codo -- Atacante')
plt.grid()
plt.show()

' ---------- Análisis de la silueta ---------- '
range_n_clusters = range(2, 13)
valores_medios_silhouette = []
Datos_escalado = scale(DatosAtacante)

for n_clusters in range_n_clusters:
    modelo = AgglomerativeClustering(
                    affinity   = 'euclidean',
                    linkage    = 'ward',
                    n_clusters = n_clusters
             )

    cluster_labels = modelo.fit_predict(Datos_escalado)
    silhouette_avg = silhouette_score(Datos_escalado, cluster_labels)
    valores_medios_silhouette.append(silhouette_avg)

fig6, ax = plt.subplots(1, 1, figsize=(6, 3.84))
```

```
ax.plot(range_n_clusters, valores_medios_silhouette, marker='o')
ax.set_title("Evolución de media de los índices silhouette -- Atacante")
ax.set_xlabel('Número de clusters')
ax.set_ylabel('Media índices silhouette')
plt.grid()

' ---------- Algoritmo k-Means ---------- '
KMeans_Atacante = KMeans(n_clusters=3).fit(DatosAtacante)
Label_Atacante = pd.DataFrame(KMeans_Atacante.labels_, columns=['Cluster'])
Centroides_Atacante = pd.concat([NombreAtacante.reset_index(drop=True),
Label_Atacante+1, ValorAtacante.reset_index(drop=True)], axis=1)
CA_Ordenado = Centroides_Atacante.sort_values('Cluster')
```