Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO DE FIN DE GRADO

Grado en Estadística

Estudio de técnicas de clustering aplicadas a una competición profesional de fútbol

Autor:

Víctor Mulero Merino

Tutores:

José Belarmino Pulido Junquera

Miguel Alejandro Fernández Temprano

Curso 2022-2023

Resumen

En la actualidad, el fútbol es el deporte más importante en Europa y ha evolucionado mucho en cuanto al análisis y optimización del rendimiento de los equipos gracias al uso de la estadística y el *Big Data*. Este Trabajo de Fin de Grado tiene como primer objetivo comparar las cinco grandes ligas europeas de fútbol: *LaLiga*, *Premier League*, *Serie A*, *Ligue 1* y *Bundesliga* para observar las diferencias y similitudes que existen entre los jugadores de cada competición. También se pretende conocer si un procedimiento de clasificación no supervisada como el análisis clúster permite clasificar a los jugadores de acuerdo a su posición en el campo a través de las variables disponibles.

Se utilizarán datos entre los años 2017 y 2022 abarcando un total de 5 temporadas. Se dispone de información sobre todos los jugadores de cada liga con estadísticas sobre los partidos en los que participaron. Para el análisis se seleccionarán las variables más importantes de los diferentes aspectos del juego y se utilizarán para identificar los clústeres que serán analizados posteriormente.

Abstract

Nowadays, soccer is the most important sport in Europe and has evolved a lot in terms of analysis and optimization of team performance thanks to the use of statistics and Big Data. The first objective of this work is to compare the five major European soccer leagues: *LaLiga*, *Premier League*, *Serie A*, *Ligue 1* and *Bundesliga* in order to observe the differences and similarities that exist between the players of each competition. It is also intended to know if an unsupervised classification procedure such as cluster analysis allows to classify the players according to their position on the field through the available variables.

Data will be used between the years 2017 and 2022 covering a total of 5 seasons. Information is available for all players in each league with statistics on the games in which they participated. For the analysis, the most important variables of the different aspects of the game will be selected and used to identify the clusters that will be analyzed later.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Las cinco grandes ligas europeas

En el fútbol europeo hay cinco grandes ligas, conocidas como *Big Five*, que son consideradas las principales competiciones de fútbol en Europa. Son las ligas más fuertes y dominan el fútbol mundial con ingresos y recursos superiores a cualquier otra agrupación de ligas [1]. Estas ligas son la *Premier League* en Inglaterra, *LaLiga* en España, *Bundesliga* en Alemania, *Serie A* en Italia y *Ligue 1* en Francia. Es común pensar que cada una de ellas tiene su estilo de juego distintivo con su carácter defensivo u ofensivo.

1.1.1. Sistema de competición

Las cinco ligas siguen el mismo sistema de competición. A lo largo de la temporada los equipos se enfrentan entre sí en dos ocasiones: una vez en su propio campo y otra en el campo del equipo rival. El equipo ganador del partido obtiene 3 puntos, el equipo perdedor no suma ninguno y en caso de empate se otorga un punto a cada uno. Los puntos acumulados a lo largo de la temporada determinan la clasificación final de los equipos y el líder se proclama campeón de la liga.

Lo realmente interesante para este trabajo son las variables que se recogen sobre los jugadores en los partidos. Durante el juego se marcan y se reciben goles, se realizan tiros a puerta, pases, robos de balón, etc. Se utilizarán estadísticas de este tipo para analizar los perfiles de los jugadores.

1.2. Objetivos

En primer lugar, previo a este trabajo se encontraba el TFG realizado por Mario Garrido Tapias [2], donde se analizó la liga española de fútbol (*LaLiga*). Utilizó distintas técnicas de análisis clúster para agrupar a los jugadores según sus diferentes perfiles. El presente

trabajo es una extensión del anterior, donde se utilizan datos de más ligas aparte de la española y se pretende comparar a sus jugadores y su estilo de juego.

En este trabajo se plantea la pregunta de si las cinco grandes ligas europeas son diferentes entre sí en cuanto al juego que se desempeña en ellas. Tradicionalmente, la *Premier League* se conoce por su enfoque ofensivo, con un estilo de juego rápido y dinámico, además de contar con una gran competitividad que hace que se considere la mejor liga del mundo [3]. Otro ejemplo es la liga italiana, conocida históricamente por ser la más defensiva y por utilizar el *Catenaccio*, que es un estilo de juego que consiste en replegarse para defender y atacar de contragolpe [4].

Se utilizarán técnicas de análisis clúster para identificar los perfiles de los futbolistas de manera que queden separados en jugadores defensivos (defensas), de creación de juego (centrocampistas) y ofensivos (delanteros). Se utilizará el *análisis de correspondencias* para comprobar si realmente hay relación entre esos perfiles y las diferentes ligas, observando si hay alguna liga caracterizada por ser más defensiva u ofensiva que el resto. Si se consiguen buenos resultados en la clasificación de los jugadores en sus posiciones, estas técnicas permitirán además identificar qué variables son más relevantes en las características posicionales de los jugadores.

Capítulo 2

Marco teórico

2.1. Análisis de la Varianza (ANOVA)

El ANOVA es una técnica estadística utilizada para comparar las medias de tres o más grupos de un factor y determinar si existen diferencias significativas entre ellos basándose en si las diferencias observadas son lo suficientemente grandes. La *tabla ANOVA* es una tabla que resume los resultados del análisis, incluyendo los factores, suma de cuadrados, grados de libertad, valor del estadístico de contraste F y el p-valor que determina si las diferencias son significativas.

2.1.1. Tipos de ANOVA

Existen varios tipos de ANOVA según el diseño del estudio. En este trabajo se utilizarán los siguientes:

- **ANOVA de un factor (one-way ANOVA):** compara los efectos de un solo factor sobre la variable respuesta. Por ejemplo: *se realiza un estudio que compara el rendimiento académico entre tres métodos de estudio (Método A, Método B y Método C). El objetivo es determinar si hay diferencias significativas en las medias del rendimiento académico entre los métodos.*
- **ANOVA de dos factores (two-way ANOVA):** compara los efectos de dos factores y su interacción sobre la variable respuesta. Por ejemplo: *un estudio que investiga el efecto del método de enseñanza (Método A y Método B), el nivel de motivación (alto y bajo) y su interacción en el rendimiento académico de los estudiantes. El análisis nos permite determinar si hay diferencias significativas en el rendimiento académico entre los grupos definidos por los factores estudiados.*

2.1.2. Hipótesis del ANOVA

Para que el ANOVA sea válido se deben verificar las siguientes condiciones:

- **Normalidad:** Las observaciones se deben distribuir normalmente. Se puede comprobar analizando los residuos: creando un histograma, un *Q-Q plot* o realizando algún test de normalidad como el *test de Shapiro-Wilk*.
- **Homogeneidad de varianzas:** Las varianzas dentro de cada grupo deben ser iguales. Se puede comprobar con el *test de Levene* o realizando un gráfico *residuos vs predichos* y comprobando que no aumente ni disminuya la varianza a lo largo del *eje x*. Este gráfico utiliza los residuos estandarizados y también se puede usar para detectar observaciones atípicas, que tienen residuos estandarizados altos (por encima de 3 en valor absoluto).
- **Independencia:** Las observaciones de cada grupo deben ser independientes entre sí. Se puede asumir independencia si el muestreo se ha realizado correctamente y de manera aleatoria. También se puede realizar un gráfico de residuos frente al orden en el que se recopilaron los datos y comprobar que no hay ningún patrón ni tendencia.

2.1.3. Análisis post-hoc

Cuando se realiza un ANOVA y resulta que un factor es significativo se puede hacer un *análisis post-hoc* para observar los grupos que son significativamente diferentes entre sí. Para esto es muy común usar el *test de Tukey*, que realiza pruebas de comparaciones múltiples y obtiene un intervalo de confianza y p-valor para la hipótesis de igualdad de medias entre cada par de categorías del factor.

2.2. Análisis en Componentes Principales

El *análisis en componentes principales* (PCA) es una técnica estadística ampliamente utilizada para el análisis de datos multidimensionales. Permite reducir la dimensionalidad de un conjunto de variables a la vez que mantiene la mayor cantidad posible de información relevante [5]. Se basa en conceptos clave como la matriz de covarianzas, los autovalores y los autovectores. Los autovalores indican la cantidad de varianza explicada por cada componente y los autovectores representan las direcciones de mayor varianza en los datos.

2.2.1. Procedimiento del PCA

El PCA consiste en buscar ejes ortogonales que capturen la variabilidad no explicada por ejes anteriores. Esto se hace mediante la *descomposición en valores singulares* (SVD). Si el número de variables de las que se dispone es n se pueden obtener un máximo de n componentes principales. En la Figura 2.1 se muestra un diagrama de dispersión en 2 dimensiones junto con los ejes que se forman utilizando PCA. El *Eje* PC_1 recoge la mayor

inercia posible de los datos, mientras que el *Eje* PC_2 es ortogonal al *Eje* PC_1 y captura la mayor variabilidad posible de la no recogida anteriormente.

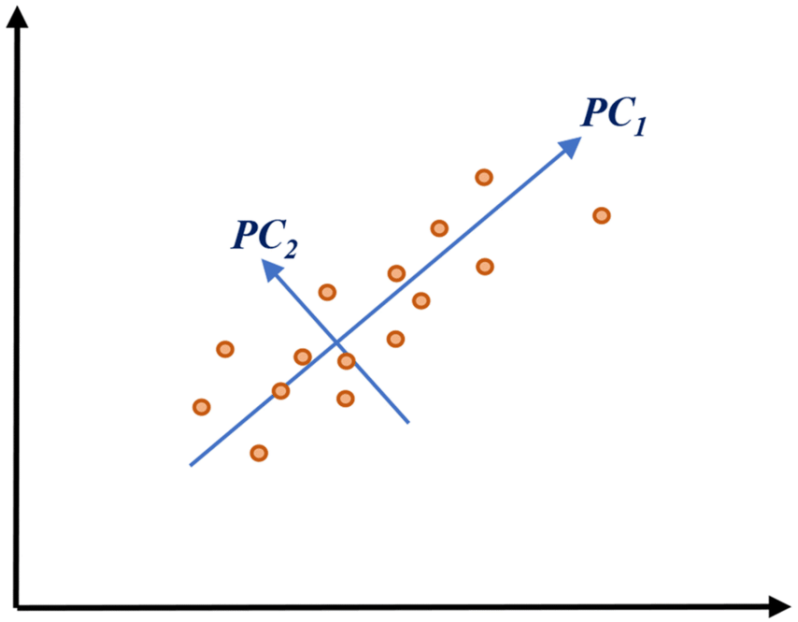

Figura 2.1: *Ejemplo Análisis en Componentes Principales [6]*

El PCA se puede dividir en los siguientes pasos, según Paloma Recuero de los Santos [7]:

1. Estandarizar las variables. Es muy común en la práctica hacer que todas las variables estén centradas ($\mu = 0$) y tengan la misma varianza ($\sigma = 1$). Se hace sobre todo cuando las variables están medidas en unidades diferentes. Si se estandarizan las variables se dice que es un *análisis normado*.

2. Calcular la matriz de covarianzas (matriz de correlación en *análisis normado)* y obtener los autovalores (λ) y autovectores asociados. Se ordenan en función de los autovalores, de mayor a menor. Los autovalores representan la inercia explicada por cada componente y los autovectores conforman las componentes principales.

3. Seleccionar el número óptimo de componentes. Es posible usar diferentes estrategias según Rukshan Pramoditha [8]. Se puede considerar cierto porcentaje de inercia acumulado y retener el número de componentes s necesario para alcanzar dicho porcentaje. En *análisis normado* es habitual extraer las componentes con autovalores mayores que 1. Esto es conocido como la *Regla de Kaiser*. También puede ser de utilidad realizar un *scree plot*, del cual hablaremos en la siguiente sección.

4. Proyectar los datos sobre un espacio de dimensionalidad menor utilizando los autovectores como pesos. Es muy común extraer 2 componentes principales y proyectar las observaciones en 2 dimensiones para visualizar los datos.

2.2.2. Gráficos del PCA

En esta sección se muestran posibles gráficos que se pueden hacer con PCA, junto con un ejemplo de cada uno. Los ejemplos se han obtenido utilizando el lenguaje *R* y el conjunto de datos *mtcars*, que contiene información acerca del rendimiento y las características de 32 modelos de automóviles.

Scree plot

Muestra los autovalores en el *eje y* y el número de componentes principales en el *eje x*. Se busca el codo del gráfico para seleccionar el número de componentes que retener, como se muestra mediante la línea roja punteada en la Figura 2.2.

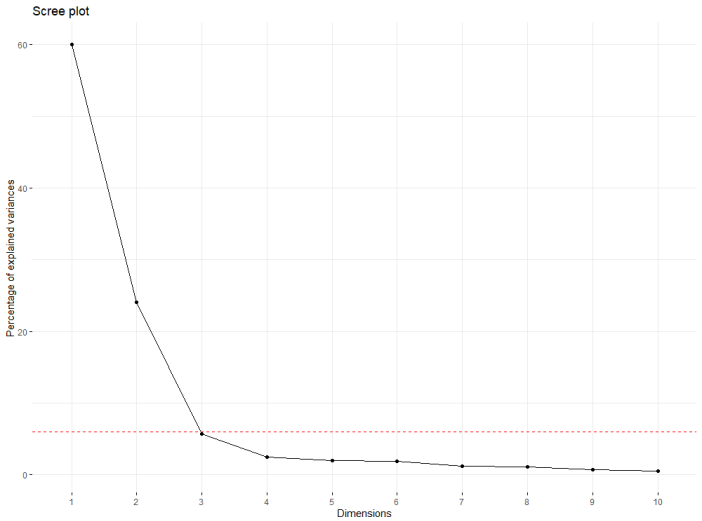

Figura 2.2: *Ejemplo Scree plot de PCA*

Gráfico de cargas

El gráfico de cargas muestra la relación entre las componentes principales y las variables originales. Si el vector correspondiente a una variable forma un ángulo pequeño con respecto al eje de una componente principal, hay una fuerte influencia de la variable en esa componente. Además, cuanto mayor sea la longitud del vector, mayor es la contribución de la variable sobre el eje. Sin embargo, si la componente y la variable son perpendiculares, no hay asociación entre ambas. Se puede ver un ejemplo en la Figura 2.3.

Biplot

Es un gráfico que utiliza las proyecciones de las observaciones del PCA y los vectores que representan las variables. Las observaciones con valores de las variables parecidos tienen proyecciones similares en las componentes principales. Se muestra en la Figura 2.4.

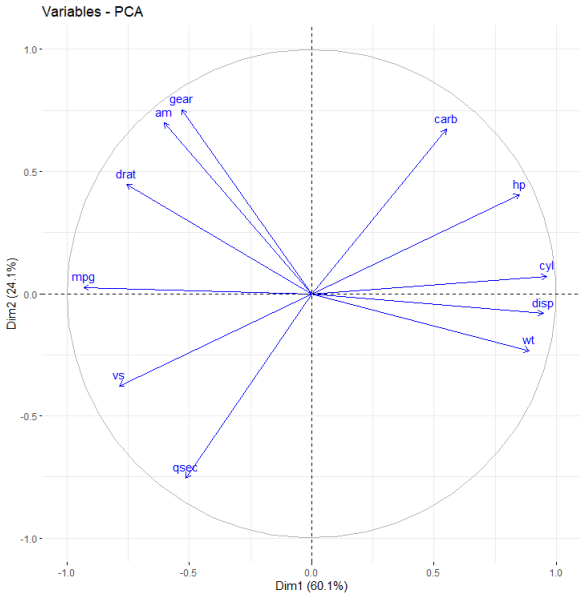

Figura 2.3: *Ejemplo Gráfico de cargas de PCA*

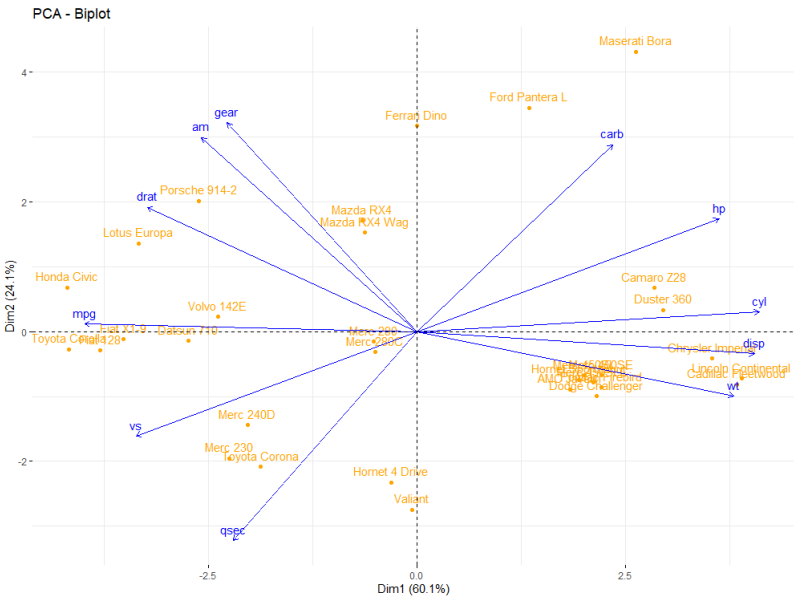

Figura 2.4: *Ejemplo Biplot de PCA*

2.3. Análisis de correspondencias

El análisis de correspondencias (CA) es una técnica estadística que se utiliza para explorar la relación entre dos o más variables categóricas en un conjunto de datos [9]. Es especialmente eficaz para analizar tablas de contingencia con datos de frecuencias numéricas con una representación gráfica que permite una interpretación rápida de los datos [10]. Cuando se dispone de dos variables categóricas para el análisis se trata de un *análisis de correspondencias simples* (SCA). Si hay más de dos variables involucradas se denomina *análisis de correspondencias múltiples* (MCA).

2.3.1. Procedimiento del CA

Al igual que en PCA, el CA consiste en construir dimensiones ortogonales secuencialmente utilizando la *descomposición en valores singulares*. En este trabajo se usa el SCA partiendo de una tabla bidimensional $r \times c$ y se pueden extraer un máximo de $min(r$ - 1, c - 1) ejes. También se utiliza el MCA con 3 variables. El CA comienza con un *test de independencia* χ^2 y continúa obteniendo las proyecciones de las categorías de las variables, de manera que se puedan representar en un número reducido de dimensiones, según Alboukadel Kassambara [11].

Test de independencia χ^2

Para realizar un CA se comienza con un *test de independencia* χ^2 para evaluar si hay una dependencia significante entre las categorías de las filas y las columnas. La hipótesis nula de este test es que las variables son independientes y se rechaza si las frecuencias observadas son muy diferentes de las esperadas.

Proyecciones y calidad

En el análisis de correspondencias se obtiene una *descomposición en valores singulares* a partir de una matriz de residuales **S**, que surge de la tabla de contingencia inicial. A partir de SVD se calculan las proyecciones multiplicando las coordenadas originales de las categorías por las cargas de cada eje. Con el CA también se calcula la calidad de representación de cada categoría en los ejes, de manera que se puede saber cómo de bien está representada la categoría en el número de dimensiones elegido.

2.3.2. Gráficos del CA

Para los gráficos de esta sección se ha utilizado el conjunto de datos *housetasks*, con información sobre las tareas domésticas que realiza una pareja.

Scree plot

Al igual que en PCA, se puede realizar un *scree plot* para visualizar la varianza explicada en función de las componentes que se usan. Se muestra un ejemplo en la Figura 2.5.

Biplot

Representa la proyección de las filas y las columnas simultáneamente. Se puede realizar el gráfico solo para las filas y solo para las columnas, pero lo interesante es visualizar ambas a la vez. Un ejemplo de *Biplot de CA* se muestra en la Figura 2.6.

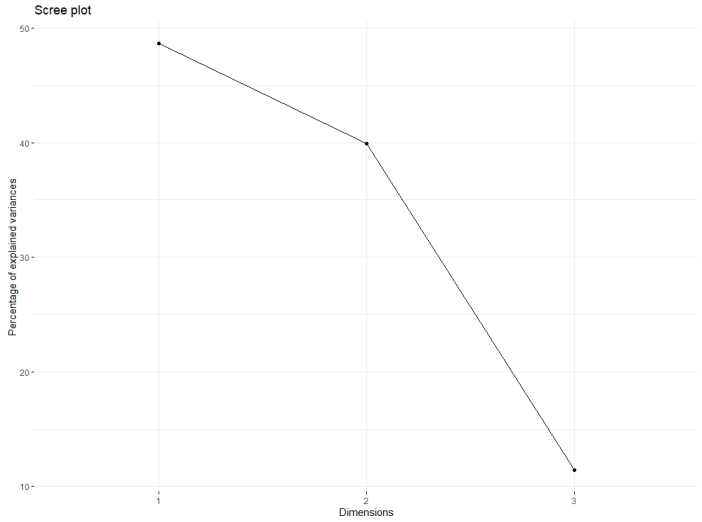

Figura 2.5: *Ejemplo Scree plot de CA*

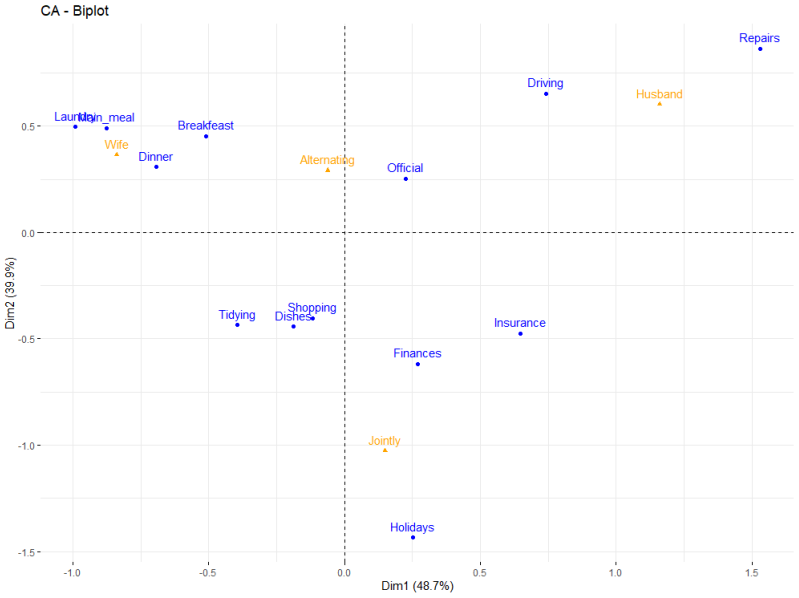

Figura 2.6: *Ejemplo Biplot de CA*

2.4. Procedimientos de clasificación

Existen dos tipos de clasificación cuando se busca categorizar a los individuos según sus características:

- **Clasificación supervisada**: Se tiene un conocimiento previo del número de grupos que hay y las etiquetas que se asignan. Es decir, los grupos están creados previamente. En este tipo de clasificación se encuentran la *regresión logística*, *máquinas de vectores de soporte* (SVM), *árboles de decisión*, *bosques aleatorios*, ...
- **Clasificación no supervisada**: Los grupos no están creados de antemano ni se conoce el número de grupos que hay que formar. El análisis clúster pertenece a este tipo de clasificación. Agrupa los individuos en conjuntos conocidos como clústeres,

de forma que los individuos dentro del mismo grupo son lo más homogéneos posible y lo más diferentes posible al resto de grupos.

2.4.1. Clustering jerárquico

Se crea una clasificación jerárquica en la que los grupos se van dividiendo o formando sucesivamente [12]. Existen dos tipos de *clustering jerárquico*:

- **Disociativo:** También conocido como método descendente. Se comienza con un único clúster en el que se encuentran todos los datos y se va dividiendo en clústeres más pequeños.
- **Aglomerativo:** También conocido como método ascendente. Comienza con tantos grupos como individuos y en cada paso se unen los grupos A y B con *índice de agregación* $\delta(A, B)$ menor, de manera que al final todos los individuos se encuentran en un mismo clúster.

Para entender el *clustering jerárquico aglomerativo* es necesario conocer los conceptos de *índice de disimilaridad* e *índice de agregación*.

Índice de disimilaridad

El *índice de disimilaridad* $d(x, y)$ mide diferencias entre los individuos x e y. Cumplen las propiedades: $d(x, y) > 0$, $d(x, x) = 0$ y $d(x, y) = d(y, x)$. Algunos índices son:

- **Distancia euclídea:** $d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_p - y_p)^2}$
- **Distancia de Manhattan:** $d(x, y) = |x_1 - y_1| + \ldots + |x_p - y_p|$
- **Distancia de Mahalanobis:** $d(x, y) = (x - y)^T S^{-1}(x - y)$, donde S es la matriz de covarianzas.

Índice de agregación

El *índice de agregación* $\delta(A, B)$ mide diferencias entre los grupos A y B. Algunos métodos son los siguientes:

- **Single Linkage:** Considera la distancia entre clústeres como la distancia mínima entre los individuos más próximos. $\delta(A, B) = \min_{x \in A; y \in B} d(x, y)$
- **Complete Linkage:** Considera la distancia entre clústeres como la distancia entre los individuos más alejados. $\delta(A, B) = \max_{x \in A; y \in B} d(x, y)$
- **Método de Ward:** Opta por fusionar los dos grupos que menos incrementen la suma de los cuadrados de las desviaciones al unirse. Cada individuo tiene un peso

$p(x)$ ($p(x) = \frac{1}{n}$ si todos tienen el mismo peso). El *índice de agregación* se calcula de la siguiente forma:

$\delta(A, B) = \frac{n_A \cdot n_B}{n_A + n_B} d(g_A, g_B)^2$, siendo $n_C = \sum_{x \in C} p(x)$, $g_C = \frac{1}{n_C} \sum_{x \in C} xp(x)$

Dendrograma

Para visualizar los clústeres que se han creado con el *clustering jerárquico* se utiliza el *dendrograma*. Los ejes verticales representan los elementos o grupos y las horizontales muestran las uniones entre ellos. En la Figura 2.7 se muestra un ejemplo de dendrograma utilizando el *Método de Ward* con el conjunto de datos *USArrests*, que contiene estadísticas de arrestos en diferentes estados de Estados Unidos en 1973. Se puede cortar el dendrograma por diferentes alturas según el número de clústeres deseado.

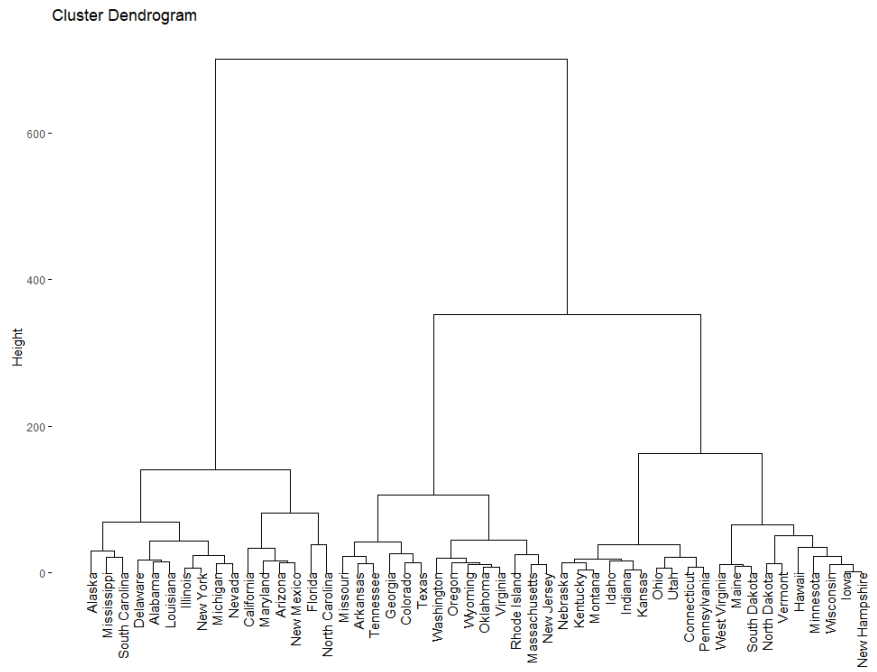

Figura 2.7: *Ejemplo Dendrograma*

2.4.2. Clustering no jerárquico

A diferencia del *clustering jerárquico*, los métodos no jerárquicos clasifican a los individuos en k grupos, donde k es un número que se especifica a priori. Comienzan con una partición inicial de individuos y se van intercambiando de un grupo a otro buscando un óptimo local de la función objetivo, la cual depende del método que se esté utilizando.

Método de las K-Medias

Se trata de un método de análisis clúster que consiste en asignar cada individuo al clúster con el centroide más cercano. El algoritmo consta de los siguientes pasos:

1. Elegir los k centroides iniciales. Una forma de seleccionarlos es el *Método de Forgy*, que consiste en elegir aleatoriamente k observaciones y utilizarlas como centroides. Otra forma es utilizar un método jerárquico para construir una partición inicial idónea, por ejemplo con el *Método de Ward*, para elegir como centroides los centros de la partición.
2. Asignar cada individuo al centroide más cercano.
3. Recalcular los centroides de la nueva partición. Cada centroide se calcula como el punto promedio de todas las observaciones asignadas a cada clúster.
4. Repetir los pasos 2 y 3 hasta alcanzar la convergencia.

Se ha utilizado el conjunto de datos *iris*, que contiene información sobre 150 muestras de flores de 3 especies diferentes, para aplicar el *Método de las K-Medias*, eliminando la variable respuesta y manteniendo las 4 variables explicativas con las que cuenta el conjunto. Se crean $k = 3$ grupos y se utiliza PCA para representarlos en 2 dimensiones.

Figura 2.8: *Ejemplo K-Medias*

2.4.3. Elección del número de clústeres

Algunos métodos para seleccionar el número de clústeres son los siguientes [13]:

Método del codo

Se realiza el *Método de las K-Medias* con diferentes valores del número de clústeres ($k = 2, 3, \ldots, n$) y se hace un gráfico representando la suma de los cuadrados dentro de cada

grupo (*wss*) frente al número de clústeres. Se busca el codo del gráfico para seleccionar el número de clústeres que crear. En la Figura 2.9 se muestra el *Método del codo* para el conjunto de datos *iris*. El codo del gráfico está en $k = 2$ ó $k = 3$.

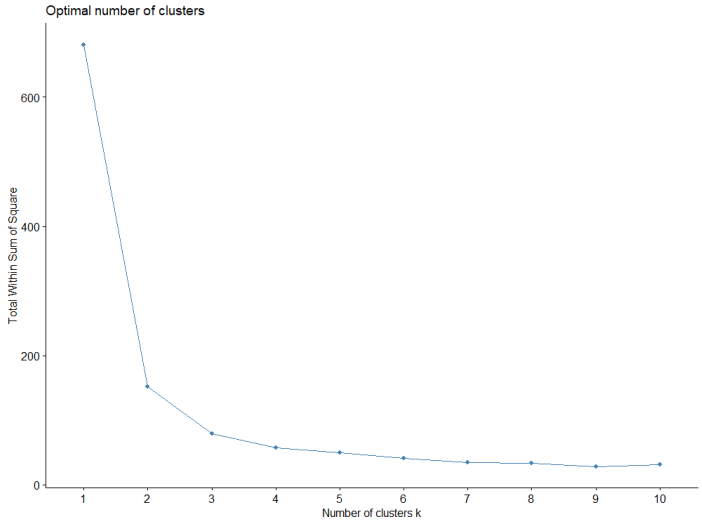

Figura 2.9: *Ejemplo Método del codo*

Método de la silueta

El *Método de la silueta* es una técnica que se utiliza para evaluar la calidad del *clustering*, proporcionando una medida de lo bien que se agrupan los datos, llamada *coeficiente de silueta*. Se realiza el *Método de las K-Medias* con $k = 2, 3, \ldots, n$, se calcula el valor promedio del *coeficiente de silueta* y se busca el número de clústeres que maximiza ese coeficiente. El coeficiente de silueta para una observacion x ($s(x)$) y el coeficiente promedio ($\bar{s}$) se calculan de la siguiente manera:

$$s(x) = \frac{b(x) - a(x)}{\text{máx}\{a(x), b(x)\}}$$

$a(x) =$ distancia promedio entre x y las observaciones del mismo clúster

$b(x) =$ distancia promedio entre x y las observaciones del clúster más cercano

$$\bar{s} = \frac{1}{N} \sum_{x} s(x)$$

Se puede realizar una representación gráfica en la que se muestra el *coeficiente de silueta promedio* frente al número de clústeres. Con el conjunto *iris* se obtiene la Figura 2.10 y el número óptimo de clústeres con este criterio sería 2.

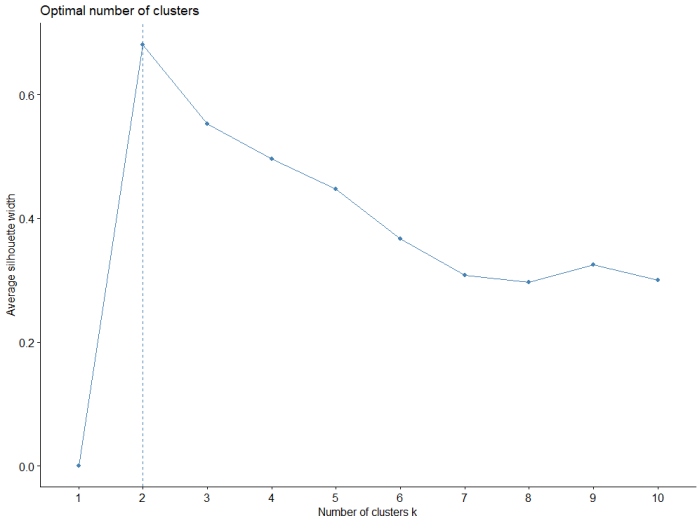

Figura 2.10: *Ejemplo Método de la silueta*

Capítulo 3

Búsqueda de datos

3.1. Bases de datos disponibles

Al comienzo de este trabajo se hizo una búsqueda con el propósito de encontrar una base de datos que ofreciera información sobre los jugadores que participaron en las cinco grandes ligas europeas a lo largo de varios años, siendo necesarias estadísticas como el número de goles, número de asistencias, pases realizados, etc. Tras realizar una búsqueda web se identificaron las páginas más interesantes:

- **football-data.co.uk** [14]

  Es un sitio web que proporciona datos y estadísticas sobre las ligas de fútbol de diversos países, tanto europeos como no europeos. La base de datos abarca información desde la temporada 1993/1994 e incluye, en algunos casos, datos de divisiones inferiores, como la Segunda División en España. Cabe destacar que este sitio web también está relacionado con el ámbito de las apuestas deportivas.

  Sin embargo, tras evaluar los datos descargables en formato CSV se determinó que no eran adecuados para el propósito de este trabajo ya que ofrecía información acerca de cada partido que se disputó, como los nombres de los equipos, fecha y hora y el resultado del encuentro, pero no había estadísticas sobre cada jugador. Por tanto, se descartó esta opción.

- **footystats.org** [15]

  Ofrece información acerca de las ligas de todo el mundo y de competiciones en las que se enfrentan equipos de diferentes países, como la *Champions League*. También cuenta con una gran cantidad de datos acerca de las competiciones femeninas de fútbol.

  Se pueden descargar varios archivos en formato CSV para cada liga y temporada:

  1. *League* - Nombre de la liga, temporada, número de equipos, etc.

2. *Matches* - Información sobre cada partido que se ha jugado en la liga en la temporada correspondiente.
3. *Teams* - Información acerca de cada equipo: número de victorias, número de derrotas, puntos obtenidos, etc.
4. *Teams Pt.2* - Información avanzada sobre cada equipo.
5. *Players* - Contiene una gran cantidad de variables útiles para el trabajo, como el número de goles, asistencias, pases, regates, acciones defensivas, etc.

El archivo *Players.csv* parecía una buena elección para la base de datos de este TFG ya que ofrecía un gran número de variables sobre cada jugador. Sin embargo, solo se disponía de los datos de la *Premier League* en la temporada 2018/2019 para descargar de manera gratuita y para acceder al resto de información había que suscribirse. Finalmente, también se descartó esta página como fuente de datos.

- **kaggle.com** [16]

  Kaggle es una plataforma en línea ampliamente utilizada por científicos de datos de todo el mundo. Ofrece una amplia gama de recursos y herramientas relacionados con ciencia de datos. Se buscó en varias bases de datos relacionadas con el fútbol y las cinco grandes ligas europeas, encontrando sobre todo información a nivel de equipo, pero lo interesante eran estadísticas a nivel de jugador. El conjunto de datos usado en el TFG de Mario Garrido [2] es un conjunto de *Kaggle* llamado *Soccer players values and their statistics* [17], el cual contenía más de 200 variables interesantes sobre cada jugador y fue el que condujo a la base de datos finalmente utilizada.

3.2. Obtención del conjunto de datos

El conjunto de datos de *Kaggle Soccer players values and their statistics* parecía una buena opción para este trabajo y se realizó una pequeña investigación sobre cómo se obtuvo. Eran datos combinados entre las páginas *transfermarkt.es* y *fbref.com* obtenidos mediante *scraping* y recopilando la información para 3 temporadas (*2017-2018*, *2018-2019* y *2019-2020*). Se descargó el código *python* con el cual se realizó el *scraping* y se hicieron modificaciones. Mediante el *script data_scrapping.py* se recopilan las estadísticas de todos los jugadores de las cinco grandes ligas europeas desde el año 2010 hasta el 2022, recogiendo la información únicamente de la web *fbref.com*. El resultado es un archivo *all_data.csv* donde cada fila corresponde a un jugador. Contiene más de 200 variables con diferentes estadísticas, incluyendo la temporada, el equipo, la posición y la liga en la que jugó. Cuenta con un total de 30693 jugadores. Cabe destacar que hay nombres de jugadores repetidos, que corresponden a los que jugaron en más de una temporada o más de un equipo.

Capítulo 4

Exploración de datos

4.1. Conjunto de datos inicial

Originalmente se disponía de 30693 observaciones con datos desde la temporada 2010-2011 hasta la temporada 2021-2022. A partir de la temporada del 2017 se contaba con un mayor número de variables registradas para cada jugador y no era necesario tener tantas observaciones. Por tanto, se decidió recortar el conjunto de datos, permaneciendo la información desde la temporada 2017-2018. En definitiva, se contaba con un total de 12584 jugadores en 5 temporadas, siendo una cantidad de datos razonable para realizar el análisis. Se disponía de 205 variables, incluyendo el nombre de jugador, la posición, la liga en la que jugó, goles marcados, asistencias, minutos jugados, etc.

4.2. Descripción del conjunto de datos

Se realizó un análisis de los datos con tablas y gráficos básicos para describir el conjunto de observaciones resultante.

4.2.1. Número de partidos

Se comenzó analizando el número de partidos en los que participó cada jugador, dividiendo por temporada y por liga. El objetivo era establecer un número mínimo de partidos razonable que un jugador tuvo que jugar para entrar al estudio, ya que si ha participado poco en los partidos no se pueden sacar conclusiones acerca de su rendimiento y sus estadísticas. El cuantil 0.25 es el valor que deja por debajo al 25 % de las observaciones. Se calculó ese valor para cada liga y cada temporada, recogiéndose en la Tabla 4.1.

	Liga					Temporada				
	Bundesliga	LaLiga	Ligue 1	Premier League	Serie A	2017-2018	2018-2019	2019-2020	2020-2021	2021-2022
Cuantil	10	10	9	11	10	10	10	9	10	10

Tabla 4.1: *Cuantil 0.25 de partidos jugados por liga y temporada*

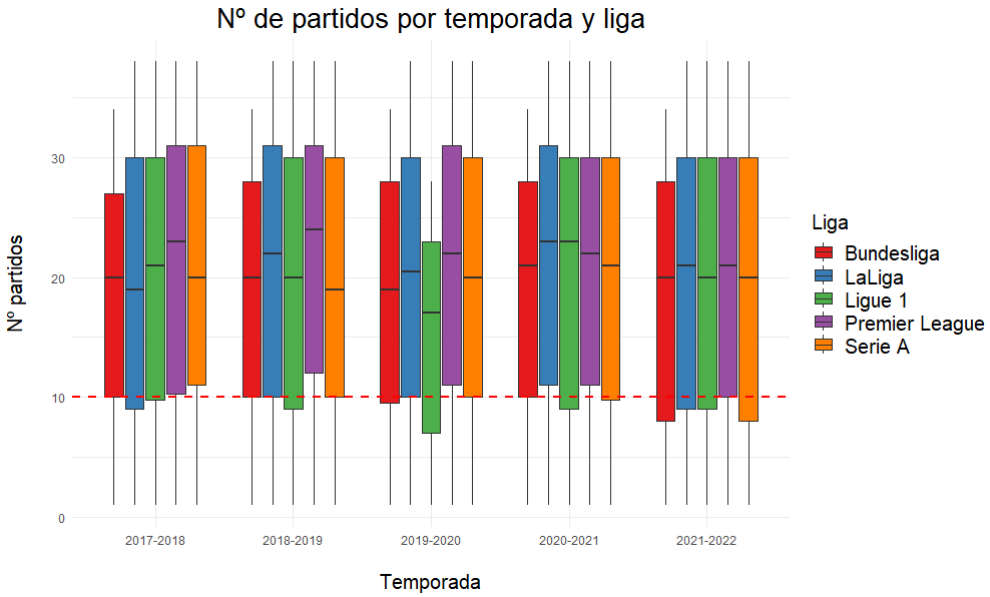

Figura 4.1: *Distribución del número de partidos por temporada y liga*

A la vista de la tabla y el cuantil 0.25 utilizando todos los datos, parece razonable utilizar el valor 10 como umbral para el número de partidos jugados. Por tanto, se eliminaron del conjunto de datos los jugadores con menos de 10 partidos en una temporada, quedando así el 75 % de las observaciones. El conjunto resultante cuenta con 9472 jugadores.

4.2.2. Número de jugadores

Se ha obtenido la distribución de los 9472 jugadores según la liga y la temporada por separado, obteniendo la Tabla 4.2 y las Figuras 4.2 y 4.3. En la Tabla 4.3 se muestra la distribución conjunta de los jugadores según la liga y la temporada.

	Liga					Temporada				
	Bundesliga	LaLiga	Ligue 1	Premier League	Serie A	2017-2018	2018-2019	2019-2020	2020-2021	2021-2022
Nº	1715	2009	1886	1842	2020	1840	1820	1834	1961	2017

Tabla 4.2: *Número de jugadores por liga y por temporada*

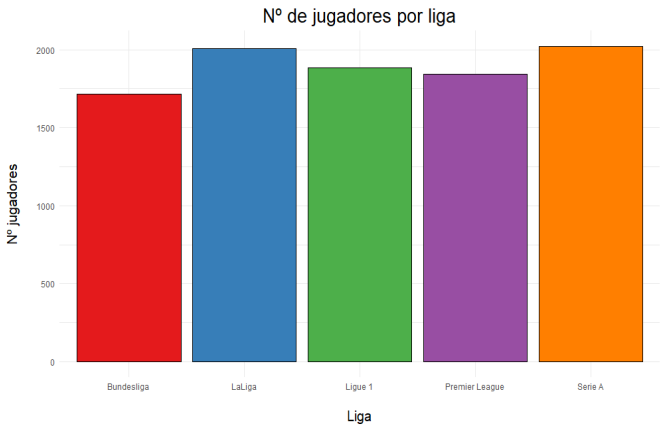

Figura 4.2: *Distribución de los jugadores por liga*

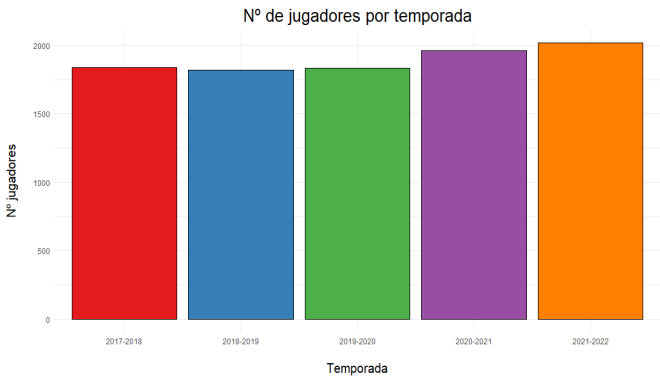

Figura 4.3: *Distribución de los jugadores por temporada*

	Bundesliga	LaLiga	Ligue 1	Premier League	Serie A	Total
2017-2018	328	396	372	361	383	1840
2018-2019	320	379	372	363	386	1820
2019-2020	347	392	331	362	402	1834
2020-2021	357	415	392	374	423	1961
2021-2022	363	427	419	382	426	2017
Total	1715	2009	1886	1842	2020	9472

Tabla 4.3: *Distribución de los jugadores por liga y temporada*

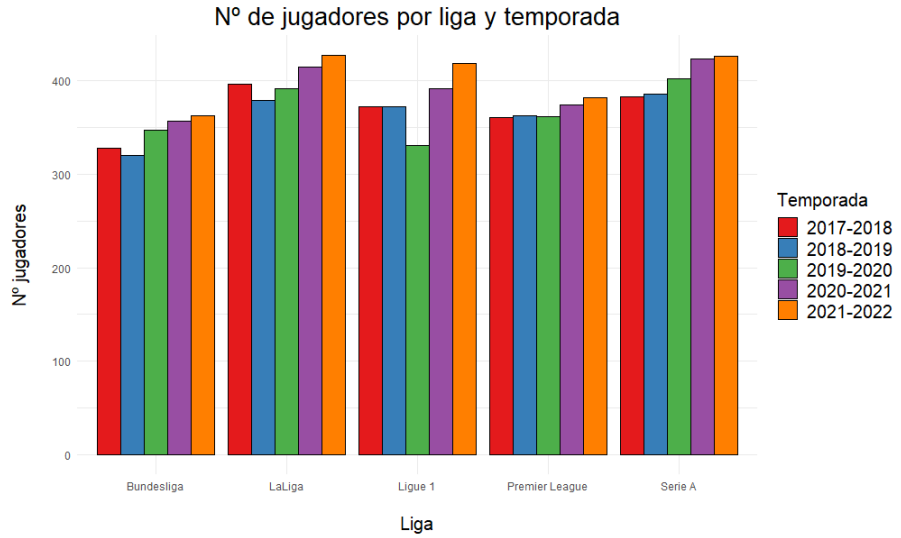

Figura 4.4: *Distribución de los jugadores por liga y temporada*

El p-valor del *test de independencia* χ^2 de la Tabla 4.3 es 0.9738, no se rechaza que haya homogeneidad en el número de jugadores teniendo en cuenta la liga y la temporada.

4.2.3. Posición de los jugadores

En la Tabla 4.4 se muestra el número de jugadores por posición en cada una de las cinco ligas y las proporciones correspondientes. Las abreviaturas utilizadas representan las siguientes posiciones: DF (defensa), MF (centrocampista), FW (delantero) y GK (portero). Se utilizará el análisis clúster para formar grupos de jugadores y compararlos con las posiciones reales.

	DF	MF	FW	GK
Bundesliga	605	592	427	91
LaLiga	710	715	486	98
Ligue 1	655	627	498	106
Premier League	650	628	471	93
Serie A	758	685	467	110

	DF	MF	FW	GK
Bundesliga	0.35	0.35	0.25	0.05
LaLiga	0.35	0.36	0.24	0.05
Ligue 1	0.35	0.33	0.26	0.06
Premier League	0.35	0.34	0.26	0.05
Serie A	0.38	0.34	0.23	0.05

Tabla 4.4: *Número y porcentaje de jugadores por posición y liga*

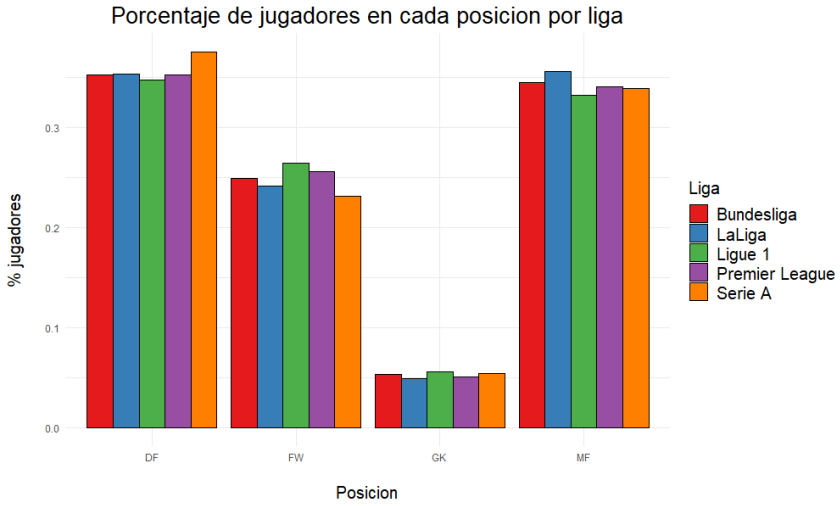

Figura 4.5: *Distribución de la posición de los jugadores por liga*

Las posiciones en las que más jugadores hay son defensa y centrocampista, mientras que en la que menos hay es en la posición de portero, como era de esperar. El p-valor del *test* χ^2 para la tabla del número de jugadores es 0.5568, no hay asociación entre la liga y la posición. La composición de cada liga en cuanto a las posiciones del campo es homogénea y aproximadamente: 35 % defensas, 35 % centrocampistas, 25 % delanteros y 5 % porteros.

4.2.4. Edad de los jugadores

La edad de los jugadores es un factor importante en el desempeño del fútbol de élite. A continuación se muestra un resumen de la variable *age* (edad del jugador) y la media de edad por posición.

Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo
15	23	26	26.01	29	42

Posición	DF	FW	GK	MF
Media	26.3	25.6	28.3	25.7

Tabla 4.5: *Resumen de la edad*

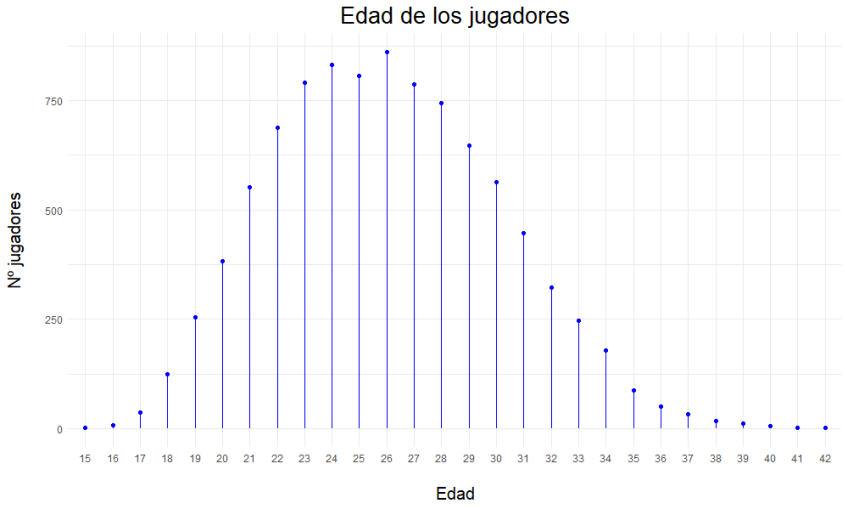

Figura 4.6: *Distribución de la edad*

La mayoría de los jugadores están entre 23 y 29 años, con una media de 26. Los porteros son los que tienen una media de edad más alta, ya que en esta posición es muy importante la experiencia de juego. En el resto de posiciones se tiene una edad promedio más baja porque se requiere un mayor despliegue físico.

4.2.5. Número de jugadores extranjeros

Resulta interesante observar la proporción de jugadores extranjeros que hay en cada liga como se muestra a continuación:

	2017-2018	2018-2019	2019-2020	2020-2021	2021-2022	Media
Bundesliga	0.582	0.581	0.605	0.622	0.590	0.596
LaLiga	0.439	0.451	0.423	0.402	0.459	0.435
Ligue 1	0.581	0.562	0.559	0.546	0.594	0.568
Premier League	0.695	0.719	0.682	0.671	0.694	0.692
Serie A	0.564	0.593	0.622	0.638	0.671	0.618
Media	0.572	0.581	0.578	0.576	0.601	

Tabla 4.6: *Proporción de jugadores extranjeros por liga y temporada*

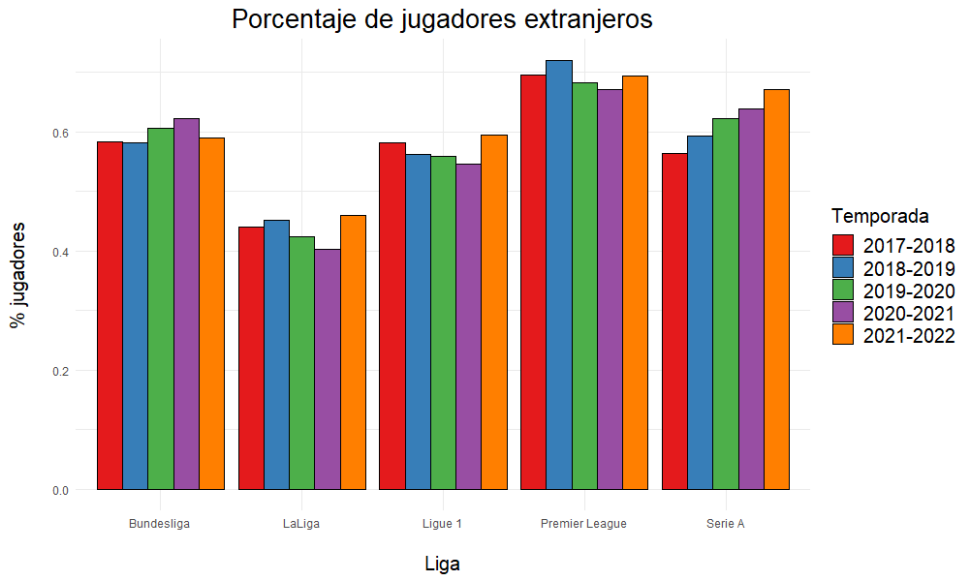

Figura 4.7: *Distribución del número de jugadores extranjeros*

LaLiga es la competición con mayor porcentaje de jugadores de la misma nacionalidad, mientras que la *Premier League* es la competición con mayor porcentaje de jugadores extranjeros. Cabe destacar que en este análisis se ha realizado sobre los jugadores que tienen como mínimo 10 partidos jugados, por lo que los porcentajes pueden diferir de un análisis realizado con todos los jugadores.

ANOVA para la proporción de jugadores extranjeros

Se realizó un *análisis de la varianza* (ANOVA) utilizando el software *R* para evaluar la influencia de la temporada y la liga en el porcentaje de jugadores extranjeros. Primero se realizó un *ANOVA de dos factores* cuyos resultados se encuentran en la Tabla 4.7. Como la variable *season* no es significativa (p-valor alto) se realizó un *ANOVA de un factor* solo con la liga y los resultados se muestran en la Tabla 4.8. El p-valor de la liga es muy bajo, por lo que se rechaza que haya la misma proporción de extranjeros en cada liga.

	gl	sum err	media err	F	p-valor
temporada	4	0.00264	0.00066	1.04	0.417
liga	4	0.17704	0.04426	69.77	$6.4 \cdot 10^{-10}$
Residuos	16	0.01015	0.00063		

Tabla 4.7: *Tabla ANOVA de dos factores*

	gl	sum err	media err	F	p-valor
liga	4	0.0.17704	0.04426	69.22	$1.99 \cdot 10^{-11}$
Residuos	20	0.01279	0.00064		

Tabla 4.8: *Tabla ANOVA de un factor*

Para verificar que el ANOVA realizado es válido necesitamos comprobar que se cumplen las condiciones de normalidad, homogeneidad de varianzas e independencia de las observaciones.

Normalidad

Para comprobar la normalidad podemos utilizar ver cómo se distribuyen los residuos realizando un histograma y un *Q-Q plot*. También se puede utilizar el *test de normalidad Shapiro-Wilk*.

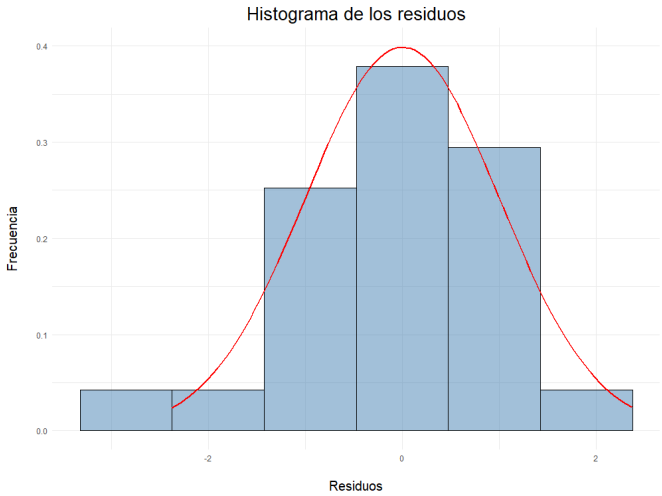

Figura 4.8: *Histograma de los residuos*

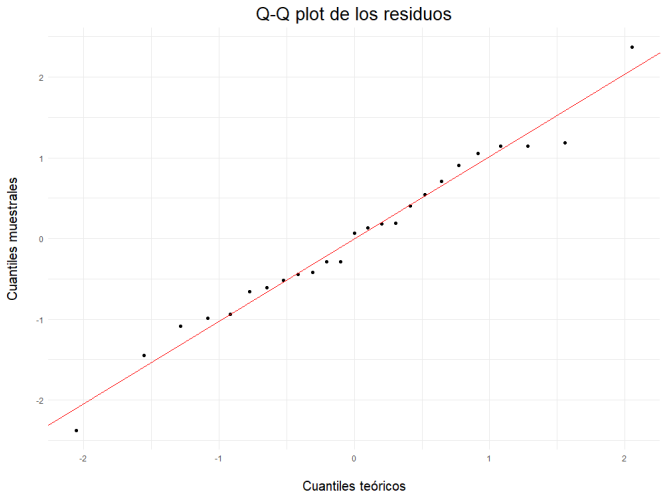

Figura 4.9: *Q-Q plot de los residuos*

Parece que los residuos se distribuyen con forma de una *Normal* y en el *Q-Q plot* los puntos se ajustan a la recta. Realizando el *test de Shapiro-Wilk* se obtiene un p-valor de 0.9504, que es muy alto. Por tanto, no se rechaza que haya normalidad.

Homogeneidad de varianzas

Se realiza el gráfico de *residuos vs predichos*, observando si la varianza se mantiene constante. También se realiza el *text de Levene*.

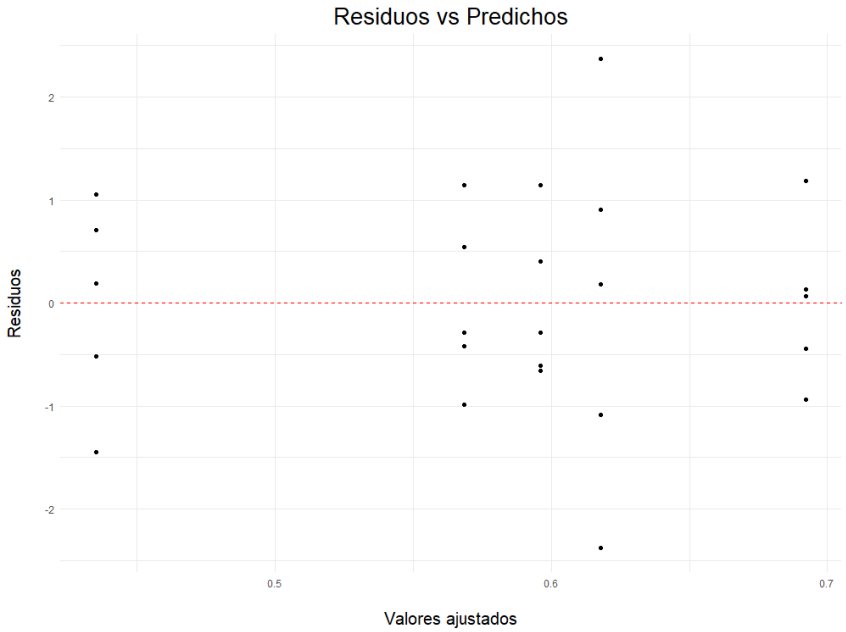

Figura 4.10: *Gráfico residuos vs predichos*

En el gráfico no se observa ningún patrón y parece que la varianza se mantiene constante. El p-valor del *test de Levene* es 0.3376, por lo que no se rechaza que haya homogeneidad de varianzas.

Independencia

Se supone la independencia por la propia naturaleza del estudio. Las ligas europeas son conocidas por atraer a futbolistas de diferentes países y culturas, se puede suponer que se distribuyen de manera independiente. Se concluye que el ANOVA realizado es válido.

Análisis post-hoc

Como el factor *liga* ha resultado significativo y el ANOVA es válido se puede realizar el *test de Tukey* para observar entre qué grupos hay diferencias significativas. En la Tabla 4.9 se observan los resultados de dicho test.

Comparación	diff	lwr	upr	p adj
LaLiga-Bundesliga	-0.161	-0.209	-0.113	0.000
Ligue 1-Bundesliga	-0.028	-0.076	0.020	0.438
Premier League-Bundesliga	0.096	0.048	0.144	0.000
Serie A-Bundesliga	0.022	-0.026	0.070	0.659
Ligue 1-LaLiga	0.133	0.085	0.181	0.001
Premier League-LaLiga	0.257	0.209	0.305	0.000
Serie A-L Liga	0.183	0.135	0.231	0.000
Premier League-Ligue 1	0.124	0.076	0.172	0.000
Serie A-Ligue 1	0.049	0.002	0.097	0.041
Serie A-Premier League	-0.075	-0.122	-0.027	0.001

Tabla 4.9: *Comparaciones post-hoc mediante el test de Tukey*

Se observa que hay diferencias significativas prácticamente entre todas las ligas. Las únicas que no difieren significativamente entre sí son *Ligue 1-Bundesliga* y *Serie A-Bundesliga*. Se podría pensar entonces que también es igual la proporción de extranjeros entre las ligas *Ligue 1* y *Serie A*, pero el p-valor del test que compara esos dos grupos es 0.041, por lo que son significativamente diferentes, aunque está al borde de la significación 0.05.

La *Premier League* es significativamente diferente al resto de ligas en cuanto a proporción de jugadores extranjeros. Es la liga que cuenta con más dinero de las cinco [18]. Dada su popularidad y su poder adquisitivo es capaz de atraer a jugadores de todos los países, aumentando así la proporción de jugadores extranjeros. La liga española (*LaLiga*) es la que menos extranjeros tiene.

Capítulo 5

Análisis clúster

En este capítulo se utilizan técnicas de análisis clúster para ver posibles agrupaciones de jugadores. Se usarán los métodos de *Ward* y *K-Medias* por estar entre los más extendidos. Los porteros no se han tenido en cuenta en el análisis. Se clasificará a los individuos en grupos utilizando diferentes conjuntos de variables, comparando si esa clasificación se corresponde con las posiciones reales de los jugadores o incluso con la liga o la temporada. Se explorará la clasificación de los jugadores utilizando un número k de clústeres mayor que 3 (que corresponde a las etiquetas de las posiciones clásicas de jugadores), permitiendo identificar diferentes perfiles (ofensivos y defensivos) dentro de un mismo grupo. Se utilizará el *análisis de correspondencias* para representar conjuntamente los clústeres con las posiciones reales, la temporada y la liga para obtener conclusiones.

5.1. Variables

En el Capítulo 4 se eliminaron del conjunto de datos a los jugadores que no habían participado en 10 partidos o más durante una misma temporada, obteniendo un conjunto final que cuenta con un total de 9472 jugadores y 205 variables. Sin tener en cuenta a los porteros el conjunto de datos se reduce a 8974 observaciones y 164 variables (ya que 41 variables son solo de los porteros). De esas variables, hay 141 que son numéricas y pueden resultar interesantes para el análisis, mientras que el resto, o bien no son de interés, o bien son categóricas y no se incluyen en el *clustering* (como el nombre del jugador, la posición, la liga, ...).

5.1.1. Selección de variables inicial

Antes de comenzar el análisis clúster es necesario realizar una selección de variables coherente para explicar la variabilidad de los datos. Las 141 variables de las que se dispone se pueden agrupar de la siguiente forma:

- **Estadísticas estándar**: variables básicas como el número de goles y asistencias o el número de tarjetas amarillas y rojas. También contienen variables categóricas como el nombre del jugador y su equipo.
- **Estadísticas de tiro**: evalúan la capacidad ofensiva que tiene un jugador. En teoría deberían servir para distinguir a los delanteros del resto de posiciones e incluso identificar centrocampistas con actitud ofensiva.
- **Estadísticas de pase**: analizan la habilidad de un jugador para mover el balón por el campo y distribuir el juego. Los centrocampistas deberían tener unos valores distintos a los del resto de posiciones en pases en profundidad o pases de gol.
- **Estadísticas de creación de tiros y de goles**: miden el talento de un jugador para crear acciones ofensivas. A priori deberían permitir distinguir delanteros y centrocampistas ofensivos de otras posiciones del campo.
- **Estadísticas de defensa**: valoran el nivel de destreza de un jugador en acciones defensivas. Se supone que permiten distinguir a los defensas y centrocampistas defensivos del resto, ya que son los que más acciones defensivas realizan.
- **Estadísticas de posesión**: relacionadas con el talento de un jugador para retener el balón. Los centrocampistas e incluso los defensas centrales deberían destacar y tener valores altos en estas estadísticas.
- **Estadísticas de tiempo de juego**: variables como el número de minutos por partido, partidos jugados, suplencias, etc.
- **Otras estadísticas**: estadísticas variadas como el número de fueras de juego o los goles en propia puerta.

A priori las variables más importantes son las de tiro, pase, creación de tiros y goles, defensa y posesión. Tras analizar las variables disponibles, se decide crear un subconjunto de 79 de ellas (eligiendo algunas de cada grupo) para el análisis clúster por ser las que parecen más importantes para separar a los jugadores en clústeres. La lista completa de las variables seleccionadas con sus explicaciones se pueden encontrar en el Apéndice A. En definitiva, se dispone de dos conjuntos de variables:

- ***Conjunto 1***: todas las variables numéricas del conjunto de datos. 141 variables.
- ***Conjunto 2***: 79 variables elegidas a criterio del estudiante del *Conjunto 1*.

5.1.2. Reducción del número de variables con PCA

Se continúa con el *Conjunto 2* de variables del Apartado 5.1.1 (79 variables) y se realiza un PCA con el objetivo de reducir el número de variables. El análisis que se hace es *normado*, ya que las variables de las que se dispone están medidas en unidades diferentes.

En la Tabla 5.1 se muestran los 20 primeros autovalores con el porcentaje de inercia que explican junto con el *scree plot* correspondiente en la Figura 5.1.

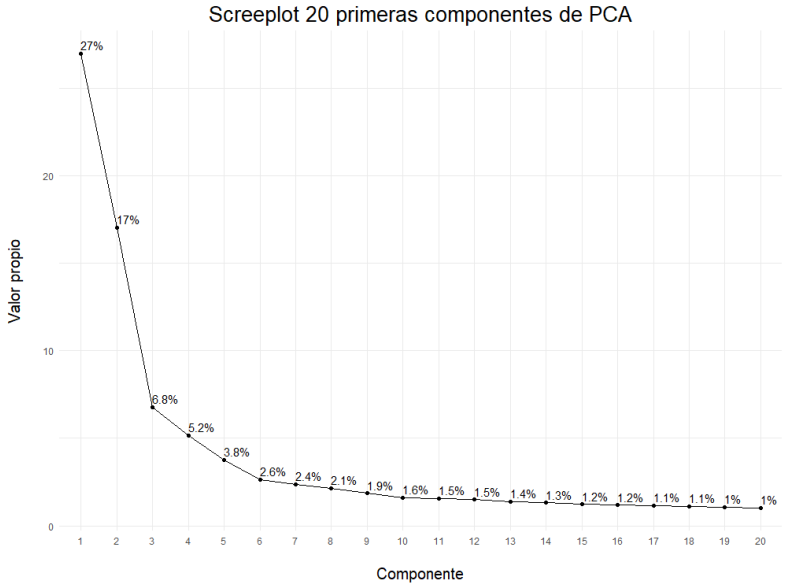

Figura 5.1: *Scree plot de PCA*

n	λ	% acumulado
1	21.32	27.0
2	13.46	44.0
3	5.36	50.8
4	4.08	56.0
5	2.97	59.7
6	2.08	62.4
7	1.88	64.8
8	1.69	66.9
9	1.46	68.8
10	1.25	70.3
11	1.22	71.9
12	1.18	73.4
13	1.10	74.8
14	1.04	76.1
15	0.97	77.3
16	0.93	78.5
17	0.90	79.6
18	0.86	80.7
19	0.83	81.7
20	0.79	82.8

Tabla 5.1: *PCA: 20 primeros autovalores*

Para seleccionar el número de componentes que retener podemos utilizar los criterios vistos en el *Marco teórico*. Observando el *scree plot* habría que retener 3 componentes principales, pero solo se alcanza a explicar un 50 % de la variabilidad total, que es un porcentaje bastante bajo. Si se aplica la *Regla de Kaiser* habría que retener 14 componentes, que es el número de autovalores mayores que 1, y si se pretenende fijar un porcentaje de varianza suficientemente alto (85 % de la variabilidad) hacen falta más de 20 componentes para alcanzarlo. En definitiva, se puede conseguir una reducción de variables pero no a un número pequeño, además de la pérdida de información e interpretabilidad que conlleva. Se descarta la idea de reducir el número de variables utilizando PCA.

5.1.3. Reducción del número de variables con las correlaciones

También se utiliza el *Conjunto 2* de variables del Apartado 5.1.1. Se puede calcular la matriz de correlaciones de las 79 variables con el fin de identificar las que estén altamente correladas y simplificar la selección. En la Tabla 5.2 se muestran las parejas de variables con correlación mayor que 0.9. En el Apéndice B se puede encontrar la representación gráfica de la matriz de correlaciones.

Las parejas de variables que se muestran en la tabla están altamente correlacionadas y es necesario valorar si se pueden eliminar.

Variable 1	Variable 2	Corr
passes_completed	passes	0.982
passes_completed	passes_live	0.985
passes	passes_live	0.977
corner_kicks	sca_passes_dead	0.911
tackles	tackles_won	0.930
passes_completed	touches	0.968
passes	touches	0.991
passes_live	touches	0.967
touches_def_pen_area	touches_def_3rd	0.923
passes_completed	touches_mid_3rd	0.912
passes	touches_mid_3rd	0.905
passes_live	touches_mid_3rd	0.921
passes_completed	touches_live_ball	0.968
passes	touches_live_ball	0.991
passes_live	touches_live_ball	0.967
touches	touches_live_ball	1.00
passes_live	carries	0.907
carries_distance	carries_progressive_distance	0.955
passes_completed	passes_received	0.902
passes_live	passes_received	0.921
carries	passes_received	0.956
shots_per90	shots_on_target_per90	0.907
passes_pct	passes_pct_medium	0.912
assisted_shots	sca_per90	0.923
sca_passes_live	sca_per90	0.937

Tabla 5.2: *Variables altamente correladas*

- **Estadísticas de tiro** (*shots_per90*, *shots_on_target_per90*): Se mantiene la variable *shots_on_target_per90* por representar mejor la capacidad ofensiva.
- **Estadísticas de pase** (*passes_completed*, *passes_live*, *passes*, *passes_pct*, *passes_pct_medium*, *corner_kicks*, *assisted_shots*): Parece razonable mantener la variable *passes_completed*, ya que parece el indicador más directo del rendimiento del jugador en términos de precisión en los pases.
- **Estadísticas de creación de tiros** (*sca_passes_dead*, *sca_passes_live*, *sca_per90*): Están relacionadas con las variables *corner_kicks* y *assisted_shots*. Se decide mantener la variable *sca_per90* por ser la que mejor recoge todas las acciones de creación de tiros. El resto de variables se eliminan.
- **Estadísticas de defensa** (*tackles*, *tackles_won*): Se elimina la variable *tackles* y se mantiene *tackles_won* por representar mejor la capacidad defensiva de un jugador.
- **Estadísticas de posesión** (*touches, touches_def_pen_area, touches_def_3rd, touches_mid_3rd, touches_live_ball, carries, carries_distance, carries_progressive_distance, passes_received*): Están también muy relacionadas con las variables de pases. Se opta por conservar las variables *touches*, *touches_def_3rd*, *touches_mid_3rd* y *carries* porque parecen las más indicadas para separar los perfiles de los jugadores.

En definitiva, se eliminan las siguientes variables: *passes_live, passes_pct_medium, corner_kicks, assisted_shots, touches_def_pen_area, touches_live_ball, carries_distance, carries_progressive_distance, passes_received, sca_passes_live, sca_passes_dead, tackles, shots_per90*. De esta manera se ha realizado una reducción de variables sin una gran pérdida de información e interpretabilidad. Quedan un total de 64 variables para realizar el *clustering*.

Al realizar el análisis clúster con las variables restantes no se obtienen buenos resultados. El *Método de Ward* no es capaz de crear 3 clústeres y que cada uno esté asociado a una posición del campo (defensa, centrocampista y delantero). Además, con el *Método de las K-Medias* se obtiene un error de clasificación alto. Por tanto, también se rechaza la idea de reducir las variables utilizando las correlaciones.

5.1.4. Selección de variables mediante un algoritmo

A la vista de los resultados anteriores, se plantea la idea de utilizar un algoritmo para seleccionar el conjunto de variables que mejor clasifique a los jugadores en 3 grupos: defensas, centrocampistas y delanteros. Se ha creado un método de selección de variables paso a paso. Este concepto indica que se realiza un proceso de selección de variables que cuenta con agregación de variables (paso hacia delante) y eliminación de variables (paso hacia atrás). Se parte de un conjunto de variables candidatas y se pretende obtener el subconjunto que minimice el error de clasificación cometido. Se puede utilizar el *Método de las K-Medias* o el *Método de Ward* como método de *clustering*, aunque hay que destacar que el *clustering jerárquico* es mucho más costoso computacionalmente por el cálculo de la matriz de distancias. Para el cálculo del error de clasificación se crean 3 clústeres y se asigna cada uno a una posición del campo según la posición mayoritaria de los jugadores que se han incluido en ese grupo. El error se calcula como la proporción de jugadores asignados a un clúster cuya posición no es la que se asigna al clúster. A continuación se presenta el pseudocódigo del algoritmo.

Algoritmo: SELECCIÓN DE VARIABLES CON MÉTODO PASO A PASO

Funcion `seleccion_variables_paso_a_paso(`*posiciones_reales, datos, n_inicio, variables_inicio*`):`

Entrada:

- **posiciones_reales**: array de posiciones reales de los jugadores
- **datos**: matriz de datos con las variables elegidas para el *clustering*
- **n_inicio**: número de variables iniciales en caso de elegirlas con PCA
- **variables_inicio**: array con las variables iniciales (si no se obtienen con PCA)

Salida:

- **mejores_variables**: conjunto de variables seleccionado y error asociado

---

```
// Obtener las variables candidatas
variables_candidatas ← datos.nombres_columnas

// Seleccionar las variables iniciales
if variables_inicio == null then
    variables_inicio ← seleccionar_variables_PCA(datos, n_inicio)
end
variables_usadas ← variables_inicio
datos_actuales ← datos[:, variables_inicio]

// Calcular el error inicial
posiciones_cluster ← calcular_cluster(datos_actuales)
error ← calcular_error(posiciones_reales, posiciones_cluster)

// Inicializar la lista de las mejores variables y el error cometido
mejores_variables ← null
mejores_variables.error ← error
mejores_variables.variables ← variables_inicio

// Iterar hacia adelante y hacia atrás hasta que no haya cambios
anteriores_variables ← null
while mejores_variables.variables ≠ anteriores_variables do
    anteriores_variables ← mejores_variables.variables
    mejores_variables ← seleccion_hacia_delante(posiciones_reales, datos,
      variables_candidatas, variables_usadas, mejores_variables)
    mejores_variables ← eliminacion_hacia_atras(posiciones_reales, datos,
      variables_candidatas, mejores_variables)
end

return mejores_variables
```

Algoritmo: SELECCIÓN DE VARIABLES CON MÉTODO PASO A PASO

Funcion `seleccion_hacia_delante(`*posiciones_reales, datos, variables_candidatas, variables_usadas, mejores_variables*`):`

Entrada:

- **posiciones_reales**: array de posiciones reales de los jugadores
- **datos**: matriz de datos con las variables elegidas para el *clustering*
- **variables_candidatas**: conjunto completo de variables

Modifica:

- **variables_usadas**: conjunto de variables usadas
- **mejores_variables**: conjunto de variables seleccionado y error asociado

---

```
// Separar la dupla en dos variables
variables_seleccionadas ← mejores_variables.variables
error_inicial ← mejores_variables.error

// Obtener las variables no seleccionadas previamente
no_seleccionadas ← variables_candidatas.eliminar(variables_usadas)

// Inicializar el conjunto de la mejor variable y el error cometido
mejor_variable ← null
mejor_variable.error ← 1
mejor_variable.variable ← null

// Buscar la variable con la que se obtenga mejor error al ser
   incluida
error ← error_inicial
variable ←  no_seleccionadas.head()
// Se podría cambiar por cualquier estrategia de selección aleatoria o
   heurística

while variable ≠ null and error > 0.0 do
    no_seleccionadas.eliminar(variable)
    variables_actuales ← variables_seleccionadas.agregar(variable)
    datos_actuales ← datos[:, variables_actuales]
    posiciones_cluster ← calcular_cluster(datos_actuales)
    error ← calcular_error(posiciones_reales, posiciones_cluster)
    if error < mejor_variable.error then
        mejor_variable.error ← error
        mejor_variable.variable ← variable
    end
    variable ← no_seleccionadas.head()
end

if mejor_variable.error < error_inicial then
    mejores_variables.error ← mejor_variable.error
    mejores_variables.variables.insertar(mejor_variable.variable)
    variables_usadas.insertar(mejor_variable.variable)
end

return
```

Algoritmo: SELECCIÓN DE VARIABLES CON MÉTODO PASO A PASO

Funcion `eliminacion_hacia_atras(`*posiciones_reales, datos, variables_candidatas, mejores_variables*`):`

Entrada:

- **posiciones_reales**: array de posiciones reales de los jugadores
- **datos**: matriz de datos con las variables elegidas para el *clustering*
- **variables_candidatas**: conjunto completo de variables

Modifica:

- **mejores_variables**: conjunto de variables seleccionado y error asociado

---

```
// Separar la dupla en dos variables
variables_seleccionadas ← mejores_variables.variables
error_inicial ← mejores_variables.error

// Inicializar el conjunto de la mejor variable y el error cometido
mejor_variable ← null
mejor_variable.error ← 1
mejor_variable.variable ← null

// Buscar la variable con la que se obtenga mejor error al ser
   incluida
error ← error_inicial
variable ← variables_seleccionadas.head()
// Se podría cambiar por cualquier estrategia de selección aleatoria o
   heurística

while variable ≠ null and error > 0.0 do
    variables_seleccionadas.eliminar(variable)
    variables_actuales ← variables_seleccionadas.eliminar(variable)
    datos_actuales ← datos[:, variables_actuales]
    posiciones_cluster ← calcular_cluster(datos_actuales)
    error ← calcular_error(posiciones_reales, posiciones_cluster)
    if error < mejor_variable.error then
        mejor_variable.error ← error
        mejor_variable.variable ← variable
    end
    variable ← variables_seleccionadas.head()
end

if mejores_variables.error < error_inicial then
    mejores_variables.error ← mejor_variable.error
    mejores_variables.variables.eliminar(mejor_variable.variable)
end

return
```

Conjuntos de datos candidatos

Se ha aplicado el algoritmo bajo diferentes condiciones y algunos resultados se recogen en la Tabla 5.3. Los conjuntos de variables de partida usados son el *Conjunto 1* (141 variables) o el *Conjunto 2* (79 variables) del Apartado 5.1.1 y se usan los métodos de *Ward* y *K-Medias*. La selección inicial de variables varía de un caso a otro. Cuando se indican "k más contributivas al PCA" se refiere a las que más contribuyen al PCA con el conjunto inicial dado. En todos los casos se utilizan las 8974 observaciones de jugadores (defensas, centrocampistas y delanteros, sin porteros).

CASO	Método	Conjunto Inicial	Variables Iniciales	Conjunto Final	Error
1	K-Medias	Conjunto 2	10 más contributivas al PCA	13 variables	0.097
2		Conjunto 2	5 más contributivas al PCA	13 variables	0.088
3		Conjunto 1	10 más contributivas al PCA	18 variables	0.102
4		Conjunto 1	10 más contributivas al PCA	18 variables	0.102
5	Ward	Conjunto 2	10 más contributivas al PCA	12 variables	0.350
6		Conjunto 2	5 más contributivas al PCA	6 variables	0.349
7		Conjunto 2	13 variables finales CASO 2	14 variables	0.100

Tabla 5.3: *Aplicación del algoritmo de selección de variables*

Observando las trazas de los casos (por problemas de espacio solo se mencionarán más adelante las que se consideran más relevantes y se incluyen en un apéndice) y la información de la tabla se pueden extraer algunas conclusiones:

- Se obtienen mejores resultados con el conjunto reducido de variables que con el que contiene todas las variables, aunque a priori parezca contradictorio. Con ambos conjuntos el algoritmo comienza con el mismo camino y en algún momento se desvía incluyendo una variable del *Conjunto 1* que no está en el *Conjunto 2*, consiguiendo una mejora de forma parcial. Pero al final de los caminos, el conjunto seleccionado con el *Conjunto 2* obtiene menor error de clasificación.
- El algoritmo no funciona bien con el *Método de Ward*, obteniéndose errores de clasificación bastante altos. Por ello, se decidió aplicar el algoritmo con *Ward* pero comenzando con un conjunto de variables que obtiene una buena clasificación con las *K-Medias*, como se ve en el *CASO 7* (comienza con las finales del *CASO 2*).
- Los conjuntos finales que obtienen una buena clasificación tienen al menos una variable de cada tipo de las que parecían más importantes a priori (tiro, pase, creación de tiros y goles, defensa y posesión).
- Los resultados son sensibles a la selección de las variables con las que se comienza, aunque la mayoría de variables iniciales se van eliminando en los pasos de selección hacia atrás.

5.1.5. Conjuntos de variables finales

Se decide elegir como conjuntos finales de variables los que se obtienen utilizando el algoritmo, concretamente los de los casos 2 y 7 de la Tabla 5.3 por ser los que cometen un menor error de clasificación con cada método. Los caminos seguidos para llegar a los conjuntos de variables finales se pueden encontrar en el Apéndice C.

- ***Conjunto K-Medias***: comete un error de clasificación con la posición principal de los jugadores menor del 10 %. Cuenta con las siguientes 13 variables: *shots_per90, tackles_mid_3rd, offsides, challenge_tackles_pct, throw_ins, touches_att_pen_area, touches_def_3rd, ball_recoveries, progressive_passes, passes_dead, touches_def_pen_area, sca_passes_dead, plus_minus_wowy*. En total hay 1 variable de tiro, 2 de pase, 1 de creación de tiros y goles, 3 de defensa, 4 de posesión, 1 de tiempo de juego y 1 de otras estadísticas.
- ***Conjunto Ward***: comete un error (cortando en 3 clústeres) cercano al 10 %. Contiene 14 variables, siendo 13 de ellas las mismas del conjunto de *K-Medias*, pero además incluye la variable *carries_into_penalty_area* (otra variable de posesión).

5.2. Clustering

En el Apartado 5.1.5 se obtienen los conjuntos de variables *Conjunto K-Medias* y *Conjunto Ward* que solo difieren en una variable. Se utilizarán para aplicar el *Método de las K-Medias* y el *Método de Ward* respectivamente.

Se crearán tablas de contingencia en las que se comparan los clústeres que se crean con las posiciones principales y las dobles posiciones de los jugadores. Se reordenarán las filas de las tablas de manera que los elementos de las tablas de las posiciones principales aparezcan en la diagonal principal, pudiendo calcular el error como la proporción de jugadores fuera de esa diagonal. Si se tiene en cuenta la segunda posición del jugador, se cuenta como acierto si un jugador se ha clasificado en alguna de las dos posiciones (la principal o la secundaria). Cabe destacar que había jugadores en el conjunto de datos con doble posición *DF-FW*, pero en estos casos se ha tenido en cuenta únicamente la posición principal, ya sea *DF* (defensa) o *FW* (delantero). También se comparan los clústeres con la temporada y con la liga para comprobar si existe alguna relación.

5.2.1. Método de Ward

Utilizando la selección de variables del *Conjunto Ward* se obtiene el dendrograma de la Figura 5.2. Se puede cortar el dendrograma por diferentes alturas, obteniendo diferentes clasificaciones de los individuos.

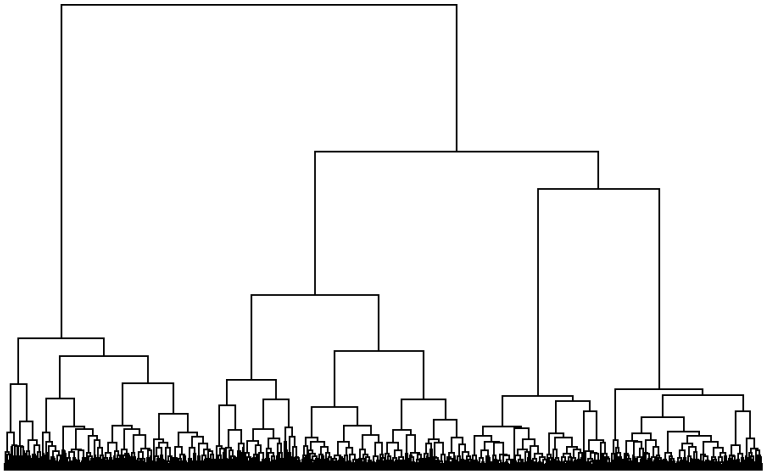

Figura 5.2: *Dendrograma con Método de Ward*

Cortando en k=3 grupos

Se corta el dendrograma de manera que se obtienen 3 clústeres. En la Figura 5.3 se representan con colores los grupos que se forman. El grupo rojo corresponde al grupo mayormente formado por delanteros, el verde por centrocampistas y el azul por defensas. Cabe destacar que los nodos finales (las hojas) son jugadores, pero no se han incluido las etiquetas por ser demasiadas y no se pueden leer bien.

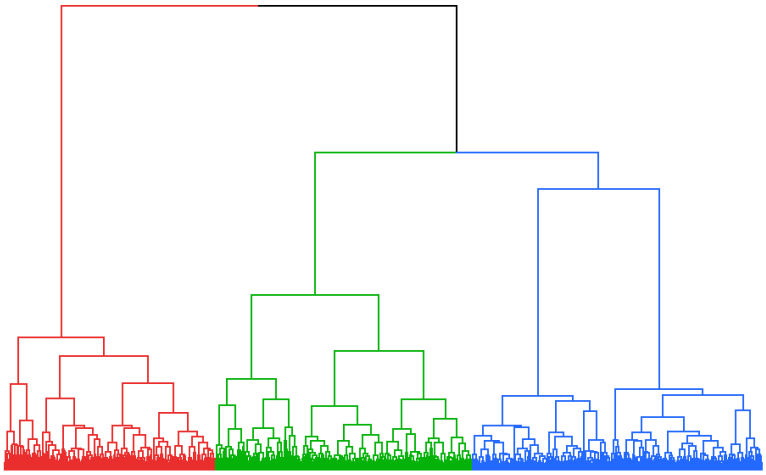

Figura 5.3: *Dendrograma con Método de Ward cortando en k=3 grupos*

En las tablas de contingencia de la Tabla 5.4 se cruzan los clústeres creados y las posiciones reales de los jugadores. Debajo de las tablas aparece el error de clasificación y el p-valor del *test de independencia* χ^2.

Clúster	Posición		
	DF	MF	FW
2	3302	95	27
3	64	2716	264
1	12	436	2058
error: 0.1001		**p-valor** ≈ 0	

Clúster	Posición				
	DF	DF,MF	MF	MF,FW	FW
2	3015	350	28	11	20
3	35	175	1983	806	45
1	9	17	63	1157	1260
error: 0.0254			**p-valor** ≈ 0		

Tabla 5.4: *Tablas de contingencia comparando 3 clústeres con la posición (Ward)*

El p-valor del *test de independencia* χ^2 es prácticamente 0 para ambas tablas, se puede realizar un *análisis de correspondencias*. En la Figura 5.4 se muestran los SCA de las tablas. Se puede observar la asociación entre los grupos formados y las posiciones del campo: el *clúster 2* está relacionado con los defensas, el *clúster 3* con los centrocampistas y el *clúster 1* con los delanteros. Las dobles posiciones se sitúan entre medias de las posiciones principales. Se prueba a cortar en $k = 5$ grupos para comprobar si existe alguna correspondencia con las 5 posiciones de las que se dispone.

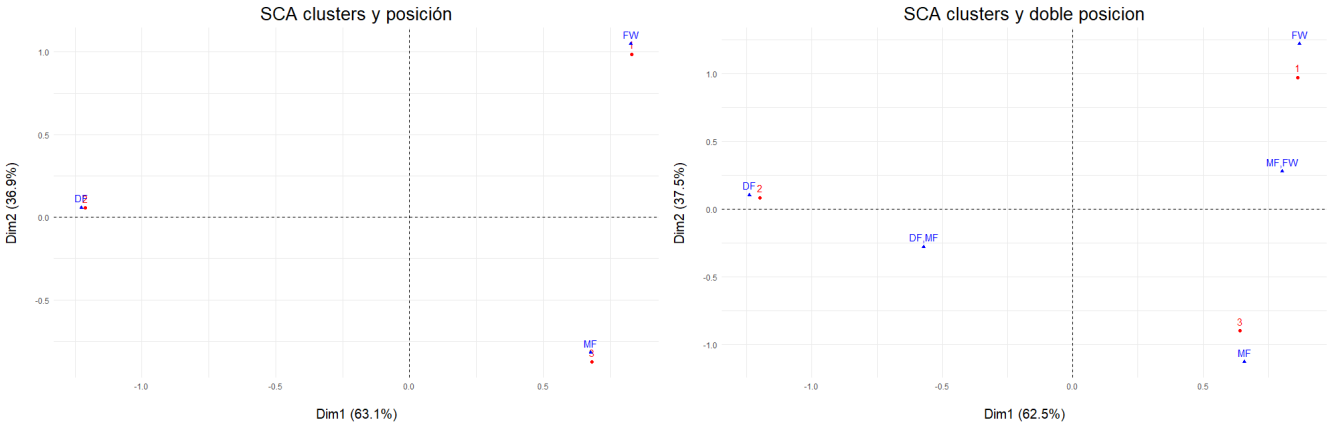

Figura 5.4: *SCA del método de Ward con 3 clústeres (Tabla 5.4)*

Cortando en k=5 grupos

Con 5 grupos se obtiene la Tabla 5.5 y el SCA de la Figura 5.5, que corresponde a la tabla de la derecha. El clúster de defensas se ha separado en dos grupos con diferentes perfiles. El *clúster 2* está formado por defensas laterales, mientras que el *clúster 5* está formado por defensas centrales. En el SCA aparecen ambos grupos muy relacionados con la defensa. También se ha dividido el clúster de centrocampistas en los *clústeres 3* y *4*. Se puede apreciar en el SCA que el *clúster 3* está desplazado hacia la doble posición *MF-FW*, pero no hay una correspondencia completa de 5 posiciones con los 5 grupos creados.

Se puede calcular un error de clasificación de la segunda tabla como la proporción de elementos fuera de la diagonal principal, obteniéndose un valor muy alto (0.474). Esto se debe a que se está utilizando un conjunto de variables que obtiene una buena clasificación

para las 3 posiciones principales, pero no se puede extender a 5 clústeres. Gran parte del error se debe a que los dos grupos de defensas están formados por defensas puros y por la dificultad para distinguir los centrocampistas ofensivos de los delanteros. No hay clústeres en los que predominen jugadores con doble posición. Si se quiere obtener un buen error de clasificación para las 5 posiciones se podría utilizar el algoritmo añadiendo las dobles posiciones para el cálculo del error.

Clúster	**Posición**		
	DF	**MF**	**FW**
2	1724	37	27
5	1578	58	0
4	35	1856	139
3	29	860	125
1	12	436	2058
error: 0.1001		**p-valor** ≈ 0	

Clúster	**Posición**				
	DF	**DF,MF**	**MF**	**MF,FW**	**FW**
2	1513	243	1	11	20
5	1502	107	27	0	0
4	19	126	1464	397	24
3	16	49	519	409	21
1	9	17	63	1157	1260
error: 0.4704			**p-valor** ≈ 0		

Tabla 5.5: *Tablas de contingencia comparando 5 clústeres con la posición (Ward)*

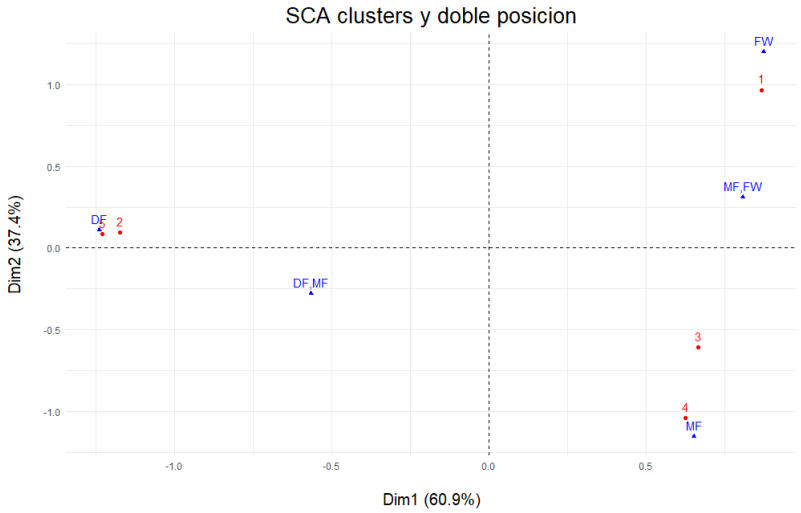

Figura 5.5: *SCA del método de Ward con 5 clústeres (Tabla 5.5)*

5.2.2. Método de las K-Medias

En esta sección se usan las variables del *Conjunto K-Medias* para obtener los clústeres. Se podrá observar que la agrupación de los datos es muy similar a la que se obtiene con el *Método de Ward* y, por tanto, no se incluyen los *análisis de correspondencias simples* ya que son muy similares a los anteriores.

Elección del número de clústeres

Antes de aplicar el *clustering* se puede observar el número de grupos k que crear. Para ello se puede aplicar el *método del codo* y el *método de la silueta*, como se observa en la Figura 5.6.

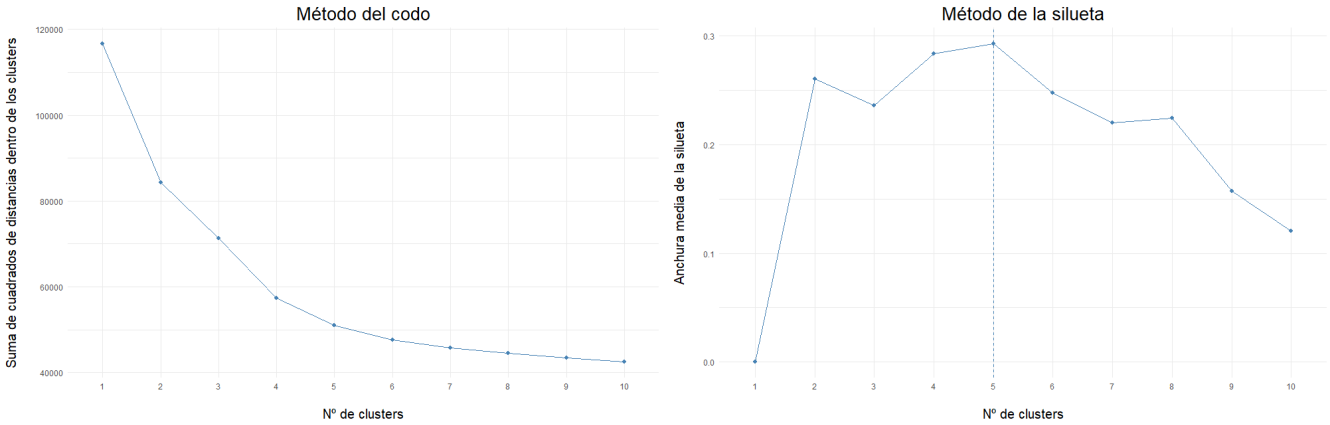

Figura 5.6: *Método del codo y silueta Conjunto K-Medias*

El codo del primer gráfico se sitúa en $k = 4$. Los coeficientes promedio de silueta son bastante parecidos desde $k = 2$ hasta $k = 6$ y valor óptimo se encuentra con 5 grupos. Se probará a aplicar el *clustering* con $k = 3$ y 5 grupos como con el *Método de Ward*.

Creando k=3 grupos

En la Tabla 5.6 se puede observar la clasificación realizada con el *Método de las K-Medias* con 3 clústeres. Es similar a la clasificación que se realizó con el *clustering jerárquico*. El error cometido con la primera posición es 0.0883 y si se tiene en cuenta la segunda se reduce a 0.0195.

Clúster	**Posición**		
	DF	**MF**	**FW**
3	3295	59	12
2	72	2825	275
1	11	363	2062
error: 0.0883		**p-valor** ≈ 0	

Clúster	**Posición**				
	DF	**DF,MF**	**MF**	**MF,FW**	**FW**
3	3010	328	14	4	10
2	38	206	2013	872	43
1	11	8	47	1098	1272
error: 0.0195			**p-valor** ≈ 0		

Tabla 5.6: *Tablas de contingencia comparando 3 clústeres con la posición (K-Medias)*

Creando k=5 grupos

Ocurre lo mismo que con el anterior método. Se forman dos grupos de defensas y los centrocampistas se dividen en un grupo de centrocampistas más ofensivos y otro más defensivo. También se comete un error alto de clasificación al comparar con las 5 posiciones.

Clúster	Posición		
	DF	MF	FW
2	1681	28	0
4	1602	22	13
1	61	2213	163
5	25	685	218
3	9	299	1955
error: 0.0934		**p-valor** ≈ 0	

Clúster	Posición				
	DF	DF,MF	MF	MF,FW	FW
2	1603	97	9	0	0
4	1396	226	1	4	10
1	34	177	1679	520	27
5	17	35	351	478	47
3	9	7	34	972	1241
error: 0.418			**p-valor** ≈ 0		

Tabla 5.7: *Tablas de contingencia comparando 5 clústeres con la posición (K-Medias)*

Se usa el *análisis en componentes principales* para representar las observaciones en dos dimensiones utilizando el *Conjunto K-Medias*. En la Figura 5.7 se colorea a los individuos según su posición principal, mientras que en la Figura 5.8 se colorea según el clúster.

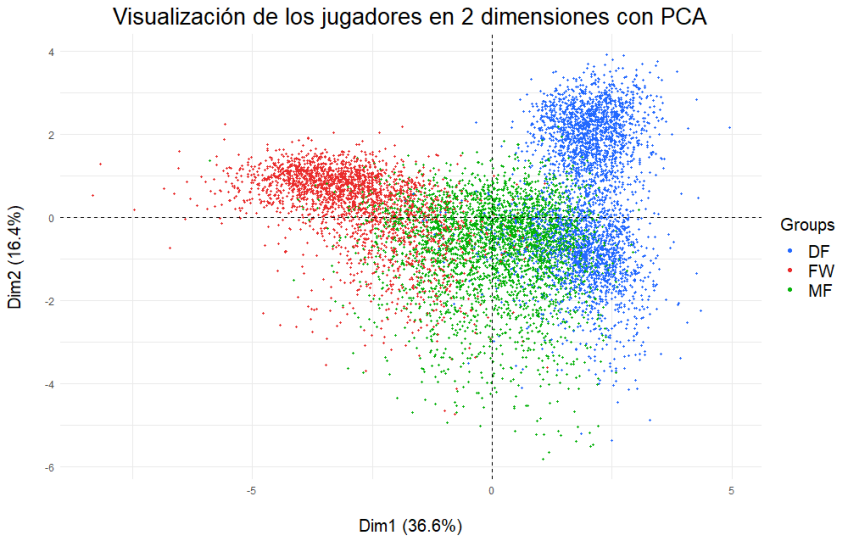

Figura 5.7: *Jugadores en 2 dimensiones con PCA (coloreando posición)*

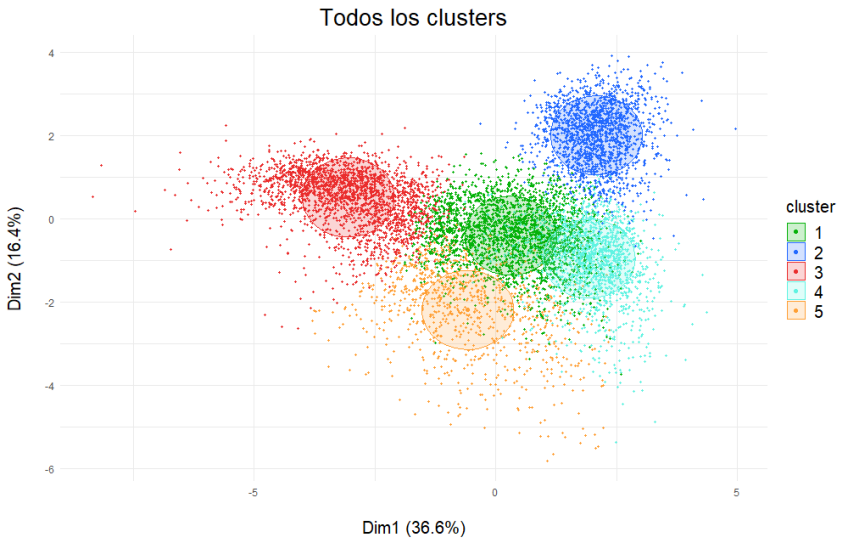

Figura 5.8: *Jugadores en 2 dimensiones con PCA (coloreando clúster)*

Utilizando PCA también se pueden visualizar las proyecciones de las variables junto con los centroides, de manera que se observan las relaciones entre las variables y los clústeres creados, como se muestra en la Figura 5.9. Se pueden observar las siguientes relaciones:

- ***Clúster 2***: muy asociado las entradas exitosas y toques en el área defensiva y el tercio del campo propio. Está mayormente formado por defensas centrales.
- ***Clúster 4***: relacionado sobre todo con recuperaciones de balón y saques de banda. Este grupo está formado sobre todo por defensas laterales.
- ***Clúster 1***: vinculado a las entradas en el centro del campo, dar pases progresivos y a realizar saques a balón parado. Sobre todo lo componen centrocampistas defensivos, aunque también cuenta con un gran número de centrocampistas más ofensivos.
- ***Clúster 5***: conectado con la creación de tiros a balón parado. Mayormente formado por centrocampistas ofensivos y delanteros.
- ***Clúster 3***: muy relacionado con los fueras de juego, los tiros a puerta y los toques en el área rival, que están muy vinculados a los delanteros.

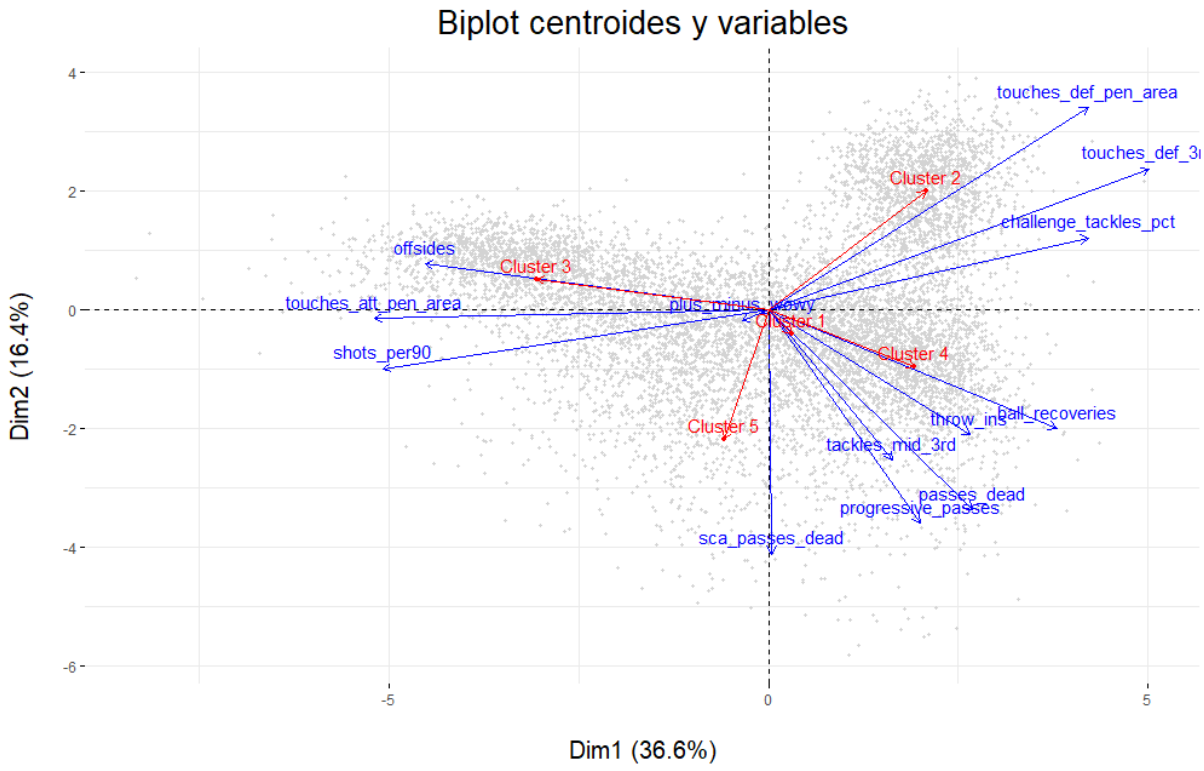

Figura 5.9: *Biplot de las variables y centroides K-Medias*

A continuación se presentan los clústeres obtenidos por separado, con algunos nombres de los jugadores que los conforman. Hay que destacar que algunos nombres de los jugadores aparecen más de una vez en un mismo grupo o en grupos distintos debido a que hay

jugadores que estuvieron presentes en más de una de las 5 temporadas que conforman el estudio. También se presentan tablas con una muestra aleatoria de 15 jugadores incluidos en cada clúster junto con las posiciones obtenidas de la página *transfermarkt.es*. La codificación para la posición utilizada en las tablas es la siguiente, obtenida del diario Marca [19]:

- LI - Lateral izquierdo
- LD - Lateral derecho
- DFC - Defensa central
- MCD - Mediocentro defensivo
- MCO - Mediocentro ofensivo
- EI - Extremo izquierdo
- ED - Extremo derecho
- DC - Delantero centro

Cabe destacar que las posiciones mencionadas no son todas las que existen. Se podría ser más técnico y afinar más las posiciones, sobre todo en el centro del campo, pero no es necesario para este trabajo.

Clúster 2 - Defensas centrales

En la Figura 5.10 se puede encontrar el clúster formado sobre todo por defensas centrales, como el español Iñigo Martínez o los franceses Lucas Hernández y Dayot Upamecano. En la Tabla 5.8 se puede observar que en la muestra aleatoria de jugadores todos son defensas centrales.

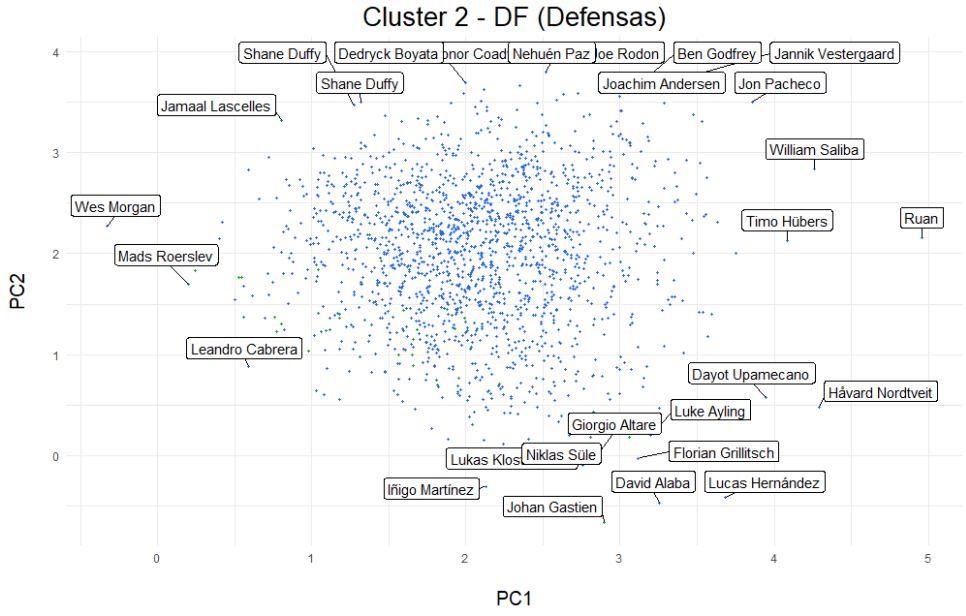

Figura 5.10: *Clúster 2 con K-Medias y k=5*

Jugador	Posición
Fabian Schär	DFC
Marvin Friedrich	DFC
Patric	DFC
Rúben Semedo	DFC
Mexer	DFC
Matthias Ginter	DFC
Dayot Upamecano	DFC
Fernando Calero	DFC
Waldemar Anton	DFC
Mats Hummels	DFC
Dylan Bronn	DFC
Liam Cooper	DFC
Milan Bišević	DFC
Craig Cathcart	DFC
Gianluca Mancini	DFC

Tabla 5.8: *Muestra clúster 2 con K-Medias y k=5*

Clúster 4 - Defensas laterales

El grupo de defensas laterales se encuentra en la Figura 5.11, con nombres destacados como Reece James, Alexander-Arnold o Marcelo. La mayoría de los jugadores de la muestra son laterales, aunque también se puede encontrar algún defensa central.

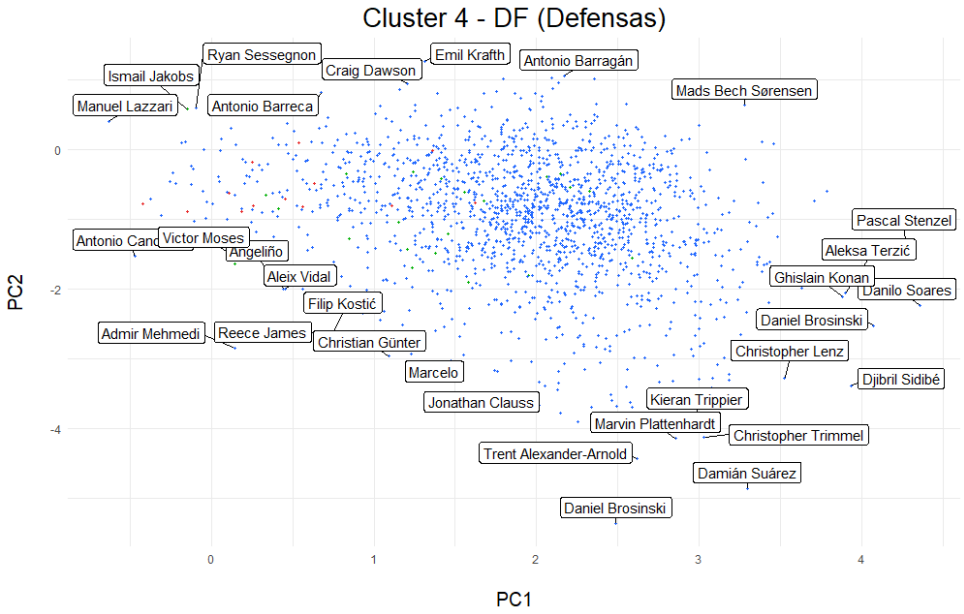

Figura 5.11: *Clúster 4 con K-Medias y k=5*

Jugador	Posición
Deiver Machado	LI
Mitchell Weiser	LD
Christophe Jallet	LD
Japhet Tanganga	DFC
Federico Barba	DFC
Iago	LI
Alex Sandro	LI
Patrick van Aanholt	LI
Sébastien Corchia	LD
Peter Pekarík	LD
Antonio Barragán	LD
Bastian Oczipka	LI
Miguel Trauco	LI
Arthur Masuaku	LI
Marc Muniesa	DFC

Tabla 5.9: *Muestra clúster 4 con K-Medias y k=5*

Clúster 1 - Centrocampistas defensivos

Se pueden encontrar en la Figura 5.12. Aparecen jugadores como el mediocentro defensivo Marco Verratti o Thiago Alcántara. Además se puede ver que Thiago aparece más de una vez debido a que estuvo presente en más de una temporada.

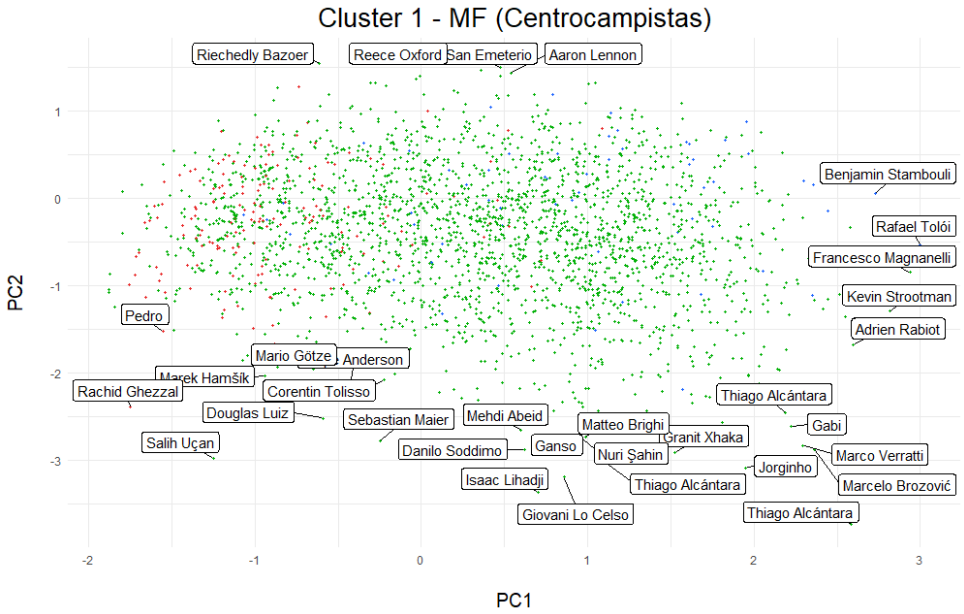

Figura 5.12: *Clúster 1 con K-Medias y k=5*

Jugador	Posición
Mijat Gaćinović	MCO
Romain Del Castillo	MCO
Valon Behrami	MCD
Dominick Drexler	MCO
Yacine Adli	MCO
Julian Draxler	MCO
Étienne Didot	MCD
Nicolò Brighenti	DFC
Ladislav Krejčí	MCD
Marco D'Alessandro	ED
Lucas Tousart	MCD
Renaud Ripart	ED
Sebastian Vasiliadis	MCD
Mateusz Klich	MCD
Granit Xhaka	MCD

Tabla 5.10: *Muestra clúster 1 con K-Medias y k=5*

Clúster 5 - Centrocampistas ofensivos

En este grupo se ven jugadores como Toni Kroos, jugador del Real Madrid C.F., o Cesc Fàbregas. También se pueden ver algunos extremos con grandes nombres como Lionel Messi

o Neymar, que por sus estilos de juego se les puede colocar de mediapunta (mediocentro ofensivo). Se pueden ver en la Figura 5.13.

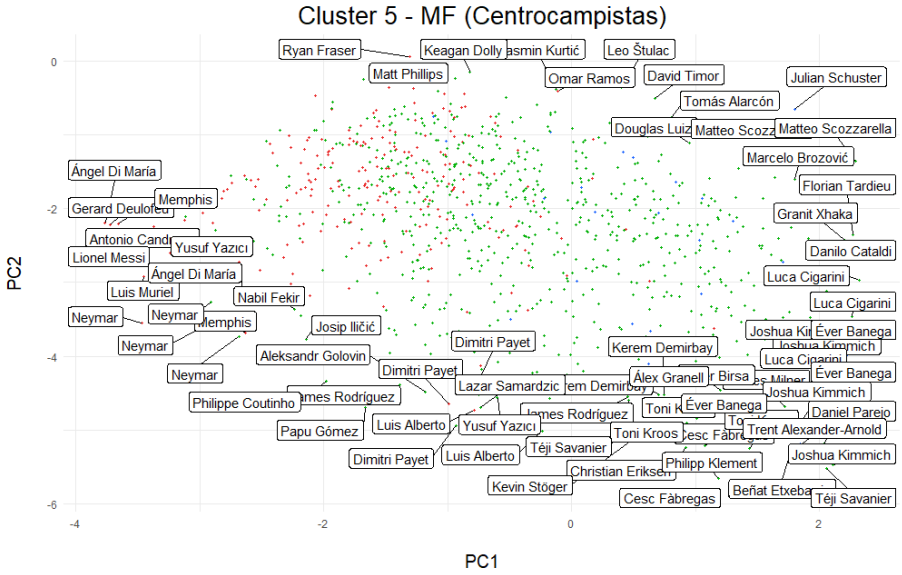

Figura 5.13: *Clúster 5 con K-Medias y k=5*

Jugador	Posición
Ryad Boudebouz	MCO
Jason Berthomier	MCO
Rodrigo De Paul	MCD
Frédéric Sammaritano	MCO
Alex Oxlade-Chamberlain	MCD
Andrea Cossu	MCO
Robert Snodgrass	ED
Téji Savanier	MCO
Luka Modrić	MCO
Grigoris Kastanos	MCO
Pasquale Schiattarella	MCD
Daniel Parejo	MCO
Ángel Di María	ED
Vincenzo Grifo	EI
Nicola Sansone	EI

Tabla 5.11: *Muestra clúster 5 con K-Medias y k=5*

Clúster 3 - Delanteros

Este último clúster se encuentra en la Figura 5.14 y cuenta con jugadores con un gran potencial ofensivo. Hay grandes figuras como Lionel Messi (en distintas temporadas), Cristiano Ronaldo o Kylian Mbappé.

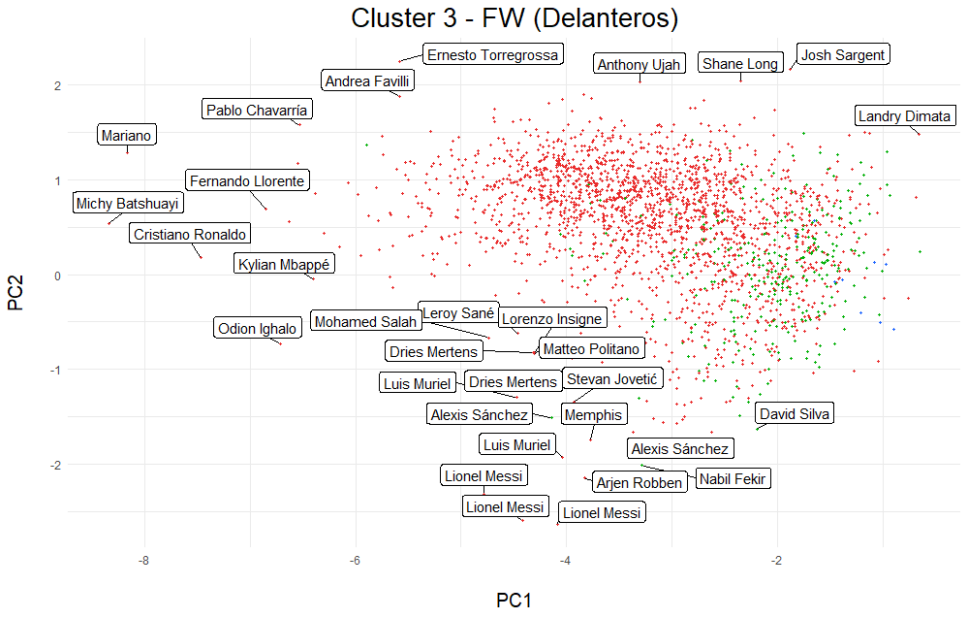

Figura 5.14: *Clúster 3 con K-Medias y k=5*

Jugador	Posición
Kevin-Prince Boateng	MCO
Nabil Fekir	MCO
Luis Suárez	DC
Mario Gómez	DC
Roberto Inglese	DC
Antonio Sanabria	DC
Alexis Claude-Maurice	MCO
Davie Selke	DC
Moses Simon	EI
Orji Okwonkwo	ED
Nuno da Costa	DC
Adama Traoré	ED
Lautaro Martínez	DC
Samuel Di Carmine	DC
Dominic Calvert-Lewin	DC

Tabla 5.12: *Muestra clúster 3 con K-Medias y k=5*

5.2.3. Comparación con la temporada

A partir de los resultados obtenidos con los dos métodos anteriores se puede observar que se agrupan los jugadores prácticamente de la misma manera. Se plantea la pregunta de si existe alguna relación entre las 5 temporadas que componen el estudio y los clústeres obtenidos.

Con k=3 clústeres

Se obtiene la siguiente tabla de contingencia correspondiente de comparar los 3 grupos creados con la temporada.

Clúster	**Temporada**					**Total**
	2017-2018	**2018-2019**	**2019-2020**	**2020-2021**	**2021-2022**	
1	478	499	494	505	530	2506
2	652	648	652	730	742	3424
3	617	573	593	626	635	3044
Total	1747	1720	1739	1861	1907	8974

Tabla 5.13: *Tabla de contingencia comparando 3 clústeres con la temporada (Ward)*

El p-valor del *test de independencia* χ^2 es 0.8160, se puede asumir que la temporada es independiente del grupo. No obstante, se realiza un MCA con los 3 clústeres creados con el *Método de Ward*, las posiciones reales y la temporada. También se ha probado con las *K-Medias* y se obtienen resultados muy parecidos, solo se incluyen los de *Ward*.

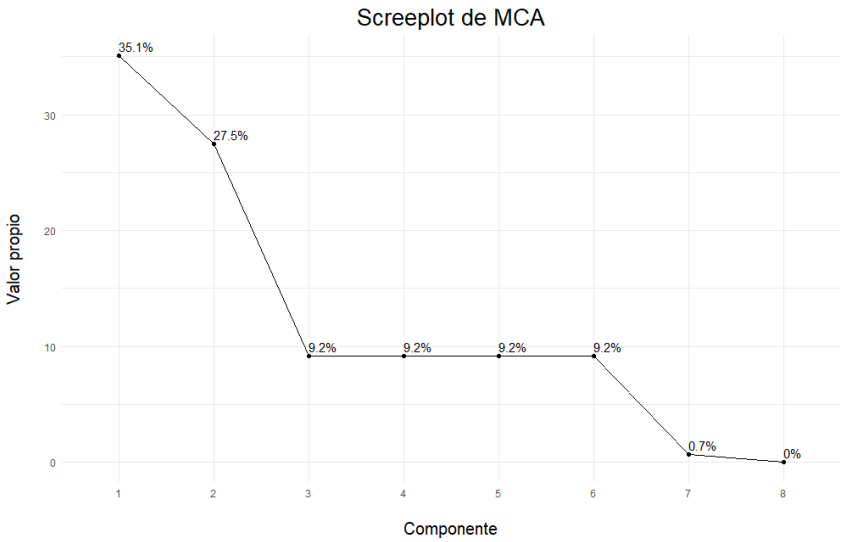

Figura 5.15: *Scree plot MCA con la temporada*

Categoría	Dim1	Dim2	Dim3
DF	0.99	0.00	0.00
FW	0.31	0.67	0.00
MF	0.33	0.65	0.00
2017-2018	0.00	0.00	0.22
2018-2019	0.00	0.00	0.10
2019-2020	0.00	0.00	0.89
2020-2021	0.00	0.00	0.04
2021-2022	0.00	0.00	0.00
1	0.34	0.64	0.00
2	0.99	0.00	0.00
3	0.31	0.68	0.00

Tabla 5.14: *Calidad de representación MCA con la temporada*

Observando el *scree plot* de la Figura 5.15 parece razonable extraer 3 componentes, que es donde se encuentra el codo del gráfico. Por tanto, se representan los *biplot* correspondientes a las dimensiones *1-2* y a las dimensiones *2-3*. Observando los gráficos y las calidades de representación se concluye que la temporada no tiene asociación con los clústeres creados ni las posiciones del campo.

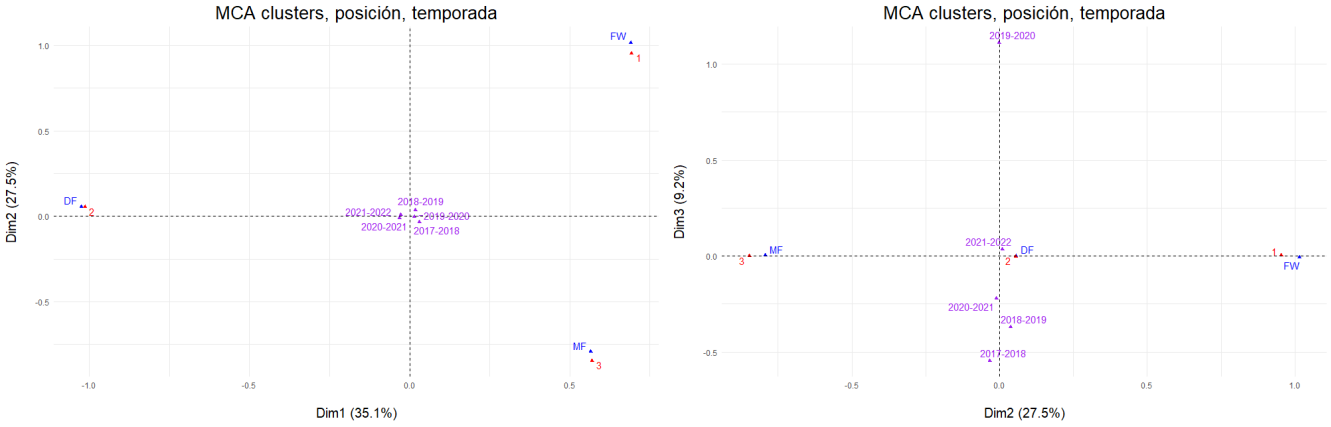

Figura 5.16: *Biplot MCA con la temporada con k=3 (Dimensiones 1-2 y 2-3)*

Con k=5 clústeres

También se prueba a buscar una correspondencia de las temporadas con los 5 clústeres creados previamente, aunque a priori no parece que se vaya a encontrar alguna relación. La tabla de contingencia correspondiente de comparar 5 clústeres con la temporada se encuentra en la Tabla 5.15. El p-valor del *test de independencia* χ^2 es 0.9669. También se han extraído 3 dimensiones para el MCA y los *biplot* correspondientes se encuentra en la Figura 5.17. Tampoco se observa ninguna relación de las posiciones con la temporada, ya que las temporadas están todas muy cercas del centro y las posiciones y clústeres se encuentran en otros lugares del *biplot*.

Clúster	**Temporada**					**Total**
	2017-2018	**2018-2019**	**2019-2020**	**2020-2021**	**2021-2022**	
1	478	499	494	505	530	2506
2	347	340	342	368	391	1788
3	197	191	193	219	214	1014
4	420	382	400	407	421	2030
5	305	308	310	362	351	1636
Total	1747	1720	1739	1861	1907	8974

Tabla 5.15: *Tabla de contingencia comparando 5 clústeres con la temporada (Ward)*

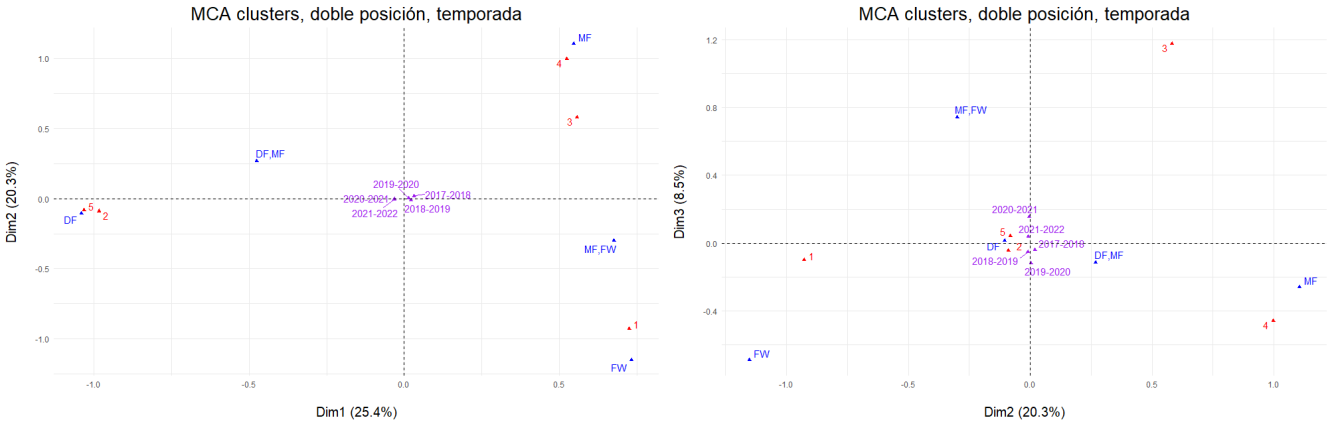

Figura 5.17: *Biplot MCA con la temporada con k=5 (Dimensiones 1-2 y 2-3)*

5.2.4. Comparación con las ligas

Se realiza el MCA utilizando como variables los clústeres creados con el *Método de Ward* (con $k = 3$ y $k = 5$), las posiciones y las diferentes ligas para comprobar si existe alguna correspondencia entre ellas.

Con k=3 clústeres

Se comparan los clústeres creados con la liga en la Tabla 5.16, con un p-valor del *test* χ^2 de 0.3350, que es alto. Por tanto, a priori no existe relación entre los clústeres y las ligas. Se realiza el MCA y se muestra el *scree plot* (Figura 5.18) junto con la Tabla 5.17, que contiene las calidades de representación de las categorías. En este caso también parece razonable extraer las 3 primeras componentes, recogiendo el 71.8 % de la inercia total.

Clúster	**Liga**					**Total**
	Bundesliga	**LaLiga**	**Ligue 1**	**Premier League**	**Serie A**	
1	480	539	486	499	502	2506
2	625	719	661	662	757	3424
3	519	653	633	588	651	3044
Total	1624	1911	1780	1749	1910	8974

Tabla 5.16: *Tabla de contingencia comparando 5 clústeres con la liga (Ward)*

En la Figura 5.19 se muestran los *biplot* correspondientes a las dimensiones *1-2* y *2-3*. Se puede observar una fuerte relación entre las posiciones y los clústeres. En cuanto a las ligas, se sitúan todas en el centro del *biplot*. Utilizando la tercera dimensión tampoco parece que haya alguna relación, por lo que no se puede concluir que haya alguna asociación entre los clústeres formados y las diferentes ligas.

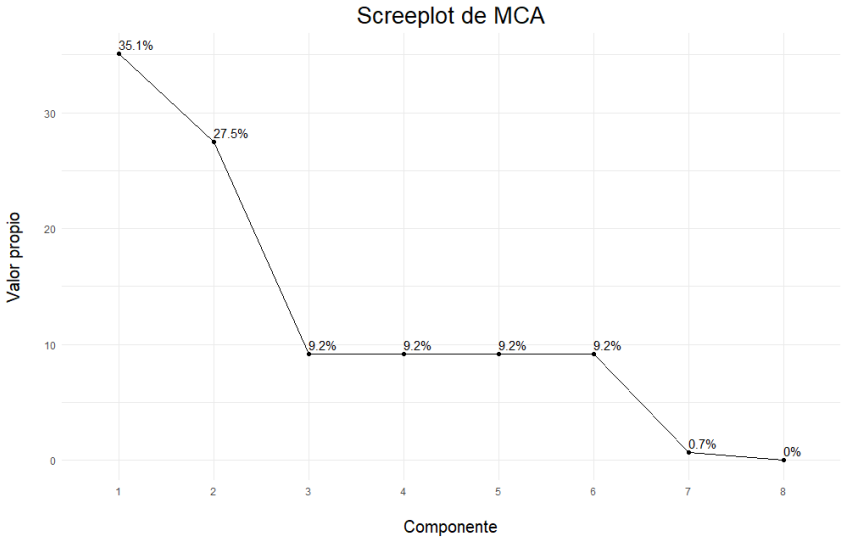

Figura 5.18: *Scree plot MCA con la liga*

Categoría	Dim1	Dim2	Dim3
DF	1.00	0.00	0.00
FW	0.31	0.67	0.00
MF	0.33	0.65	0.00
Bundesliga	0.00	0.00	0.32
LaLiga	0.00	0.00	0.19
Ligue 1	0.00	0.00	0.72
Premier League	0.00	0.00	0.00
Serie A	0.00	0.00	0.01
1	0.34	0.64	0.00
2	1.00	0.00	0.00
3	0.30	0.68	0.00

Tabla 5.17: *Calidad de representación MCA con la liga*

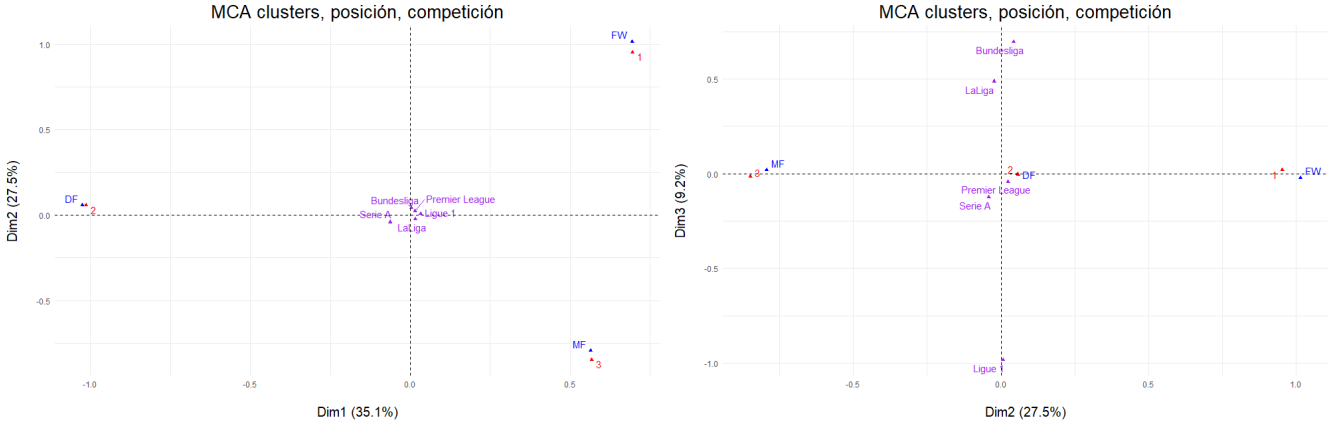

Figura 5.19: *Biplot MCA con la liga con k=3 (Dimensiones 1-2 y 2-3)*

Con k=5 clústeres

Se comparan los 5 clústeres con las ligas en la Tabla 5.18, con un p-valor de 0.5039, que es bastante alto y por tanto no se rechaza la independencia.

Clúster	**Liga**					**Total**
	Bundesliga	**LaLiga**	**Ligue 1**	**Premier League**	**Serie A**	
1	480	539	486	499	502	2506
2	304	383	363	344	394	1788
3	172	225	205	193	219	1014
4	347	428	428	395	432	2030
5	321	336	298	318	363	1636
Total	1624	1911	1780	1749	1910	8974

Tabla 5.18: *Tabla de contingencia comparando 5 clústeres con la liga (Ward)*

Realizando un MCA y extrayendo 3 componentes se obtienen los *biplot* en las Figuras 5.20 y 5.21. Fijándonos en la primera dimensión (*eje x* de la Figura 5.20) se puede observar una fuerte asociación entre los *clústeres 2* y *5* con el grupo de los defensas (lado izquierdo del gráfico). A la derecha se encuentran los centrocampistas y los delanteros junto con los *clústeres 1*, *3* y *4*. La segunda dimensión (*eje y*) es la que permite distinguir los delanteros de los centrocampistas, situando a los delanteros en la parte inferior del gráfico y a los jugadores menos ofensivos en la parte superior. Sin embargo, no se puede extraer ninguna conclusión de que una liga esté asociada más a un clúster en concreto, ni siquiera utilizando la tercera dimensión.

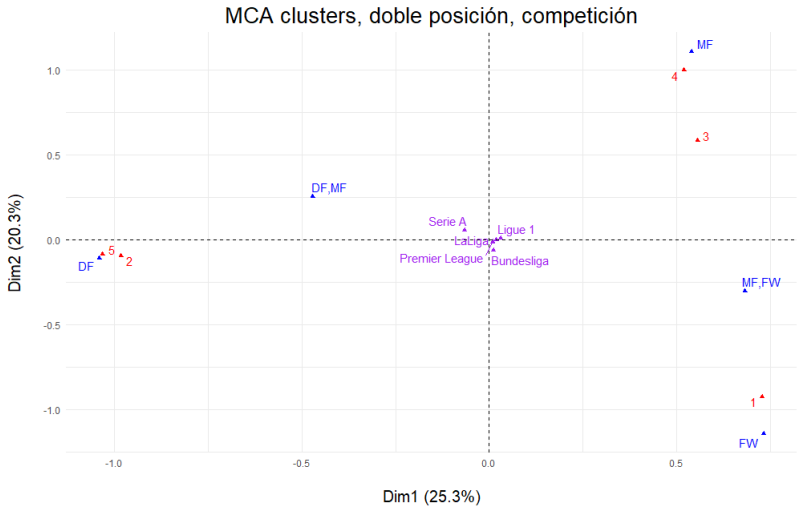

Figura 5.20: *Biplot MCA con la liga con k=5 (Dimensiones 1-2)*

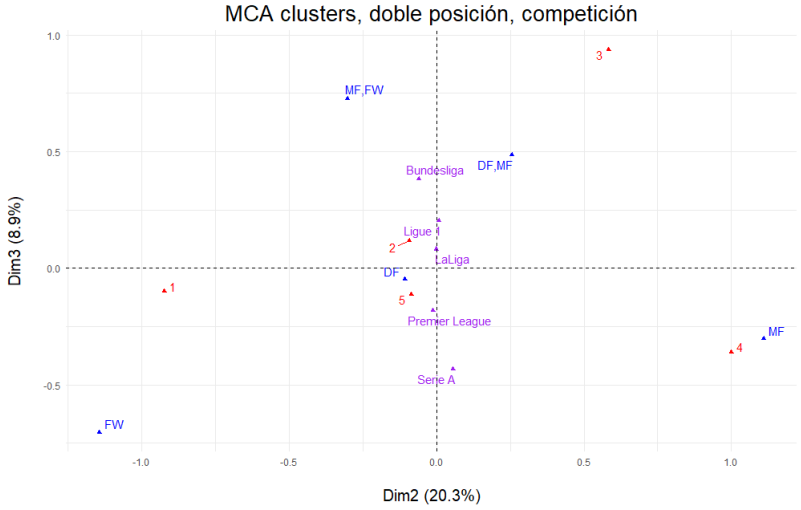

Figura 5.21: *Biplot MCA con la liga con k=5 (Dimensiones 2-3)*

Cabe destacar que si no se aplica el algoritmo de selección de variables y se utiliza la lista completa de variables del Apéndice A tampoco se obtiene una asociación entre clústeres y liga.

Capítulo 6

Conclusiones y trabajo futuro

En este TFG se han utilizado estadísticas de jugadores de fútbol de las cinco grandes ligas europeas desde la temporada *2017-2018* hasta la *2021-2022*. Los datos se han obtenido de la web *fbref.com* mediante *scraping* con el lenguaje *python*.

Se ha probado si un método de clasificación no supervisada como el análisis clúster puede generar resultados buenos al clasificar a los jugadores en grupos determinados por una variable que no se ha considerado para el *clustering*, como es la posición. Utilizando un método de selección de variables se obtiene una buena clasificación para 3 clústeres, haciendo que cada uno corresponda a una posición del campo (defensa, centrocampista y delantero). Sin utilizar métodos tan dirigidos a la buena clasificación se han obtenido resultados aceptables en algunos casos.

Como métodos de crear los clústeres se ha utilizado el *Método de Ward* del *clustering jerárquico*, en el que los grupos se van dividiendo o formando sucesivamente, y el *Método de las K-Medias* del *clustering no jerárquico*, donde los individuos se clasifican en un número k de grupos previamente fijado. El algoritmo de selección de variables empleado funciona bien con las *K-Medias* pero tiene muchas dificultades con *Ward* debido a que no se fija previamente un número k de grupos que crear que se adapten a los datos. Por eso se decidió comenzar el algoritmo de *Ward* con los resultados finales del algoritmo con *K-Medias*, obteniendo una mejor clasificación.

Observando el SCA de la Figura 5.4 parecía que se podía obtener una correspondencia de 5 clústeres con 5 posiciones. Sin embargo, al realizar la clasificación no se obtienen buenos resultados. Para mejorar la clasificación habría que utilizar el método de selección de variables utilizando la doble posición en lugar de las posiciones principales individuales. Es decir, el método de selección de variables depende de la "variable respuesta".

No se obtiene ninguna relación entre los clústeres creados con la liga ni con la temporada. Si no se utiliza el método de selección de variables tampoco se puede ver ninguna asociación. Por tanto, se concluye que no hay diferencias significativas en el juego de las cinco grandes ligas europeas.

A partir de este TFG surgieron nuevas propuestas para trabajos futuros. Antes de plasmar la versión final en el trabajo se habían implementado diferentes versiones del algoritmo para seleccionar las variables. En un principio el algoritmo solo contaba con pasos hacia delante. Luego se mejoró haciendo que al alcanzar el límite hacia delante se fueran eliminando algunas variables hasta que no se mejorara la solución, para volver a hacer los pasos hacia delante. Se descubrió que el algoritmo funciona mejor si se da un paso hacia delante y uno hacia atrás.

Una propuesta para un trabajo futuro sería analizar el funcionamiento del algoritmo y ver cómo se podría mejorar de cara a obtener las mejores variables que clasifiquen a los individuos en grupos previamente etiquetados utilizando clasificación no supervisada. En caso de estar relacionado con el fútbol, analizar la importancia de que se incluyan en el modelo estadísticas de los diferentes tipos importantes que se han mencionado (tiro, pase, creación de tiros y goles, defensa y posesión) para poder separar los perfiles de los jugadores.

Con el *Método de Ward* no funciona bien el algoritmo de selección de variables paso a paso. Sería necesario analizar por qué no se obtienen buenos resultados de clasificación y crear otro tipo de algoritmo que permita obtener un buen error de clasificación con *Ward*.

Bibliografía

[1] Statista Research Department (2023). *Big Five - statistics & facts*. Accesible: https://www.statista.com/topics/5909/-big-five/#topicOverview. Accedido el 19/06/2023.

[2] Mario Garrido Tapias (2022). *Uso de técnicas de clustering para encontrar perfiles de jugadores en una competición de fútbol profesional*. Accesible: https://uvadoc.uva.es/handle/10324/57954. Accedido el 19/06/2023.

[3] ElDesmarque (2022). *¿Cuáles son las ligas de fútbol más importantes de Europa?* Accesible: https://www.eldesmarque.com/futbol/20230112/ranking-mejores-ligas-futbol-g00g_21597651.html. Accedido el 19/06/2023.

[4] Iván Fuente (2017). *Formas de entender el fútbol. Capítulo 1: El Catenaccio*. Accesible: https://www.marca.com/blogs/desde-el-aula/2017/01/31/formas-de-entender-el-futbol-capitulo-1.html. Accedido el 19/06/2023.

[5] Miguel Alejandro Fernández Temprano (2021). *Tema 1. Análisis en componentes principales*. Apuntes de la asignatura de Análisis de Datos del Grado en Estadística. Universidad de Valladolid.

[6] Mohsen Hesami and A. Maxwell P. Jones (2020). *Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. Applied Microbiology and Biotechnology*, 104, 9449-9485. Accesible: https://www.researchgate.net/figure/An-example-of-principal-component-analysis-PCA-for-a-two-dimensional-data-set_fig2_344399773. Accedido el 19/06/2023.

[7] Paloma Recuero de los Santos (2018). *Python para todos: Tutorial de PCA en 5 sencillos pasos*. Accesible: https://empresas.blogthinkbig.com/python-para-todos-tutorial-de-pca-en-5. Accedido el 19/06/2023.

[8] Rukshan Pramoditha (2022). *How to Select the Best Number of Principal Components for the Dataset*. Accesible: https://towardsdatascience.com/how-to-select-the-best-number-of-principal-components-for-the-dataset-287e64b14c6d. Accedido el 19/06/2023.

[9] Miguel Alejandro Fernández Temprano (2021). *Tema 2. Análisis de correspondencias*. Apuntes de la asignatura de Análisis de Datos del Grado en Estadística. Universidad de Valladolid.

[10] Michael Greenacre (2008). *La práctica del análisis de correspondencias. Fundacion BBVA*. ISBN: 978-84-96515-71-0. Accesible: https://www.fbbva.es/wp-content/uploads/2017/05/dat/DE_2008_practica_analisis_correspondencias.pdf. Accedido el 19/06/2023.

[11] Alboukadel Kassambara (2017). *CA - Correspondence Analysis in R: Essentials*. Accesible: http://www.sthda.com/english/articles/31-principal-

component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials. Accedido el 19/06/2023.

[12] Luis Ángel García Escudero (2021). *Tema 4. Clasificación no supervisada*. Apuntes de la asignatura de Análisis de Datos del Grado en Estadística. Universidad de Valladolid.

[13] Matthew J. Oldach (2019). *10 Tips for Choosing the Optimal Number of Clusters*. Accesible: https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92. Accedido el 19/06/2023.

[14] *Football Data*. Accesible: http://www.football-data.co.uk. Accedido el 19/06/2023.

[15] *FootyStats*. Accesible: https://footystats.org. Accedido el 19/06/2023.

[16] *Kaggle*. Accesible: https://www.kaggle.com. Accedido el 19/06/2023.

[17] Rafal Stępień (2021). *Soccer players values and their statistics*. Accesible: https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics. Accedido el 19/06/2023.

[18] Statista Research Department (2023). *Revenue of the Big Five soccer leagues in Europe from 2011/12 to 2020/21, with a forecast to 2022/23, by country*. Accesible: https://www.statista.com/statistics/261218/big-five-european-soccer-leagues-revenue. Accedido el 19/06/2023.

[19] Víctor Ayora (2019). *Conoce todas las posiciones de los jugadores de FIFA*. Accesible: https://www.marca.com/esports/fifa/2019/03/22/5c951fa4268e3eaf348b45c1.html. Accedido el 26/06/2023.

Apéndice A

Lista de variables

A continuación se muestra el conjunto inicial de variables seleccionadas para el *clustering*. Algunas variables están marcadas con un asterisco (*). Son las variables elegidas que se dividen entre la variable *minutes_90s* ($\frac{minutos\ jugados}{90}$) de manera que la estadística tiene un valor relativo en relación al tiempo de juego. El resto de variables se mantienen sin modificar.

Estadísticas estándar

1. **goals_per90** - número de goles cada 90 minutos
2. **assists_per90** - número de asistencias cada 90 minutos

Estadísticas de tiro

3. **shots_per90** - número de tiros cada 90 minutos
4. **shots_on_target_per90** - número de tiros a puerta cada 90 minutos
5. **goals_per_shot** - número de goles por tiro
6. **average_shot_distance** - distancia media entre el tirador y la portería de todos los tiros
7. **shots_free_kicks** - número de tiros desde tiro libre *

Estadísticas de pase

8. **passes_completed** - número de pases completados
9. **passes** - número de pases intentados *
10. **passes_pct** - % de pases completados
11. **passes_pct_short** - % de pases cortos completados
12. **passes_pct_medium** - % de pases a media distancia completados
13. **passes_pct_long** - % de pases largos completados

14. **assisted_shots** - número de pases que asisten un tiro *
15. **passes_into_final_third** - número de pases completados que entran en el último tercio del campo rival *
16. **pass_xa** - número de pases que se convierten en asistencia de gol *
17. **passes_live** - número de pases durante el juego *
18. **passes_dead** - número de pases a balón parado *
19. **passes_free_kicks** - número de pases desde tiro libre *
20. **through_balls** - número de pases que van entre los defensores del equipo rival y crean oportunidad de gol
21. **passes_switches** - número de pases de mas de 40 yardas a lo ancho (cambios de orientacion del juego) *
22. **crosses** - número de centros intentados *
23. **throw_ins** - número de saques de banda *
24. **corner_kicks** - número de corners *

Estadísticas creación de tiros y goles

25. **sca_per90** - número de acciones ofensivas que llevan a un tiro cada 90 minutos
26. **sca_passes_live** - número de pases durante el juego que llevan a un tiro *
27. **sca_passes_dead** - número de pases a balón parado que llevan a un tiro *
28. **sca_take_ons** - número de regates que llevan a un tiro *
29. **sca_shots** - número de tiros que llevan a otro tiro *
30. **sca_fouled** - número de faltas recibidas que llevan a un tiro *
31. **sca_defense** - número de acciones defensivas que llevan a un tiro *
32. **gca_per90** - número de acciones ofensivas que llevan a un gol cada 90 minutos
33. **gca_passes_live** - número de pases durante el juego que llevan a un gol *
34. **gca_passes_dead** - número de pases a balón parado que llevan a un gol *
35. **gca_take_ons** - número de regates que llevan a un gol *
36. **gca_shots** - número de tiros que llevan a otro tiro que se convierte en gol *
37. **gca_fouled** - número de faltas recibidas que llevan a un gol *
38. **gca_defense** - número de acciones defensivas que llevan a un gol *

Estadísticas de defensa

39. **tackles** - número de entradas intentadas *
40. **tackles_won** - número de entradas exitosas *
41. **tackles_def_3rd** - número de entradas en el tercio defensivo *

42. **tackles_mid_3rd** - número de entradas en el tercio medio *
43. **tackles_att_3rd** - número de entradas en el tercio ofensivo *
44. **challenge_tackles_pct** - % de entradas exitosas a un jugador que intenta regatear
45. **blocks** - número de balones bloqueados estando en la trayectoria del balón *
46. **blocked_shots** - número de tiros bloqueados estando en la trayectoria del balón *
47. **blocked_passes** - número de pases bloqueados estando en la trayectoria del balón *
48. **interceptions** - número de balones interceptados *
49. **tackles_interceptions** - número de entradas + intercepciones *
50. **clearances** - número de despejes *
51. **errors** - número de errores que llevan a un disparo del oponente *
52. **ball_recoveries** - número de balones recuperados *

Estadísticas de posesión

53. **touches** - número de veces que un jugador toca el balón *
54. **touches_def_pen_area** - número de toques en el area defensiva *
55. **touches_def_3rd** - número de toques en el tercio defensivo *
56. **touches_mid_3rd** - número de toques en el tercio medio *
57. **touches_att_3rd** - número de toques en el tercio ofensivo *
58. **touches_att_pen_area** - número de toques en el área rival *
59. **touches_live_ball** - número de toques durante el juego *
60. **take_ons_won_pct** - % de regates exitosos
61. **carries** - número de de conducciones *
62. **carries_distance** - distancia recorrida conduciendo el balón *
63. **carries_progressive_distance** - distancia recorrida conduciendo el balón hacia la portería rival *
64. **carries_into_final_third** - número de conducciones que entra en el tercio ofensivo *
65. **carries_into_penalty_area** - número de conducciones que entran al área rival *
66. **miscontrols** - número de veces que el jugador falla intentando controlar el balón *
67. **dispossessed** - número de veces que el jugador pierde el balón tras una entrada rival *
68. **passes_received** - número de pases recibidos *
69. **progressive_carries** - número de conducciones hacia la portería rival *
70. **progressive_passes** - número de pases hacia la portería rival *
71. **progressive_passes_received** - número de pases progresivos recibidos *

Estadísticas de tiempo de juego

72. **plus_minus_wowy** - número de goles cada 90 minutos a favor - número de goles cada 90 minutos en contra cuando el jugador está jugando

Otras estadísticas

73. **fouls** - número de faltas cometidas *
74. **fouled** - número de faltas recibidas *
75. **offsides** - número de fuera de juego *
76. **pens_won** - número de penaltis que recibe el jugador *
77. **pens_conceded** - número de penaltis que hace el jugador *
78. **own_goals** - número de goles en propia puerta *
79. **aerials_won_pct** - % de duelos aéreos ganados

Apéndice B

Matriz de correlaciones

La Figura B.1 corresponde a la matriz de correlaciones de las 79 variables seleccionadas.

Figura B.1: *Matriz de correlaciones*

Apéndice C

Trazas del algoritmo de selección de variables

Traza del *Conjunto K-Medias*

Algoritmo de selección de variables con *Método de las K-Medias* como método de crear los clústeres, partiendo del *Conjunto 2* (79 variables) y comenzando con las 5 que más contribuyen al PCA.

```
VARIABLES INICIALES (5 VARIABLES)
    touches_live_ball, touches, passes, passes_completed, sca_per90

CAMINO SEGUIDO POR EL ALGORITMO
    INICIO
    SELECCION HACIA DELANTE: Entra sca_passes_live - error 0.373
    SELECCION HACIA ATRAS: Sale touches - error 0.351
    SELECCION HACIA DELANTE: Entra shots_per90 - error 0.321
    SELECCION HACIA ATRAS: Sale passes_completed - error 0.316
    SELECCION HACIA DELANTE: Entra average_shot_distance - error 0.299
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE: Entra tackles_mid_3rd - error 0.278
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE: Entra offsides - error 0.256
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE: Entra challenge_tackles_pct - error 0.241
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE: Entra throw_ins - error 0.176
    SELECCION HACIA ATRAS: Sale sca_per90 - error 0.122
    SELECCION HACIA DELANTE: Entra touches_att_pen_area - error 0.116
    SELECCION HACIA ATRAS: Sale average_shot_distance - error 0.113
```

```
    SELECCION HACIA DELANTE: Entra touches_def_3rd - error 0.108
    SELECCION HACIA ATRAS: Sale touches_live_ball - error 0.108
    SELECCION HACIA DELANTE: Entra ball_recoveries - error 0.105
    SELECCION HACIA ATRAS: Sale sca_passes_live - error 0.099
    SELECCION HACIA DELANTE: Entra progressive_passes - error 0.098
    SELECCION HACIA ATRAS: Sale passes - error 0.096
    SELECCION HACIA DELANTE: Entra passes_dead - error 0.094
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE: Entra touches_def_pen_area - error 0.091
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE: Entra sca_passes_dead - error 0.089
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE: Entra plus_minus_wowy - error 0.088
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE:
    SELECCION HACIA ATRAS:
    FIN

VARIABLES SELECCIONADAS (13 VARIABLES)
    shots_per90, tackles_mid_3rd, offsides, challenge_tackles_pct,
    throw_ins, touches_att_pen_area, touches_def_3rd, progressive_passes,
    passes_dead, touches_def_pen_area, sca_passes_dead, plus_minus_wowy

ERROR DE CLASIFICACIÓN: 0.0883
```

Traza del *Conjunto Ward*

Algoritmo de selección de variables con *Método de Ward* como método de crear los clústeres, partiendo del *Conjunto 2* (79 variables) y comenzando con las 13 variables finales del *Conjunto K-Medias*.

```
VARIABLES INICIALES (13 VARIABLES)
    shots_per90, tackles_mid_3rd, offsides, challenge_tackles_pct,
    throw_ins, touches_att_pen_area, touches_def_3rd, progressive_passes,
    passes_dead, touches_def_pen_area, sca_passes_dead, plus_minus_wowy

CAMINO SEGUIDO POR EL ALGORITMO
    INICIO
    SELECCION HACIA DELANTE: Entra carries_into_penalty_area - error 0.1
    SELECCION HACIA ATRAS:
    SELECCION HACIA DELANTE:
    SELECCION HACIA ATRAS:
    FIN

VARIABLES SELECCIONADAS (14 VARIABLES)
    shots_per90, tackles_mid_3rd, offsides, challenge_tackles_pct,
    throw_ins, touches_att_pen_area, touches_def_3rd, progressive_passes,
    passes_dead, touches_def_pen_area, sca_passes_dead, plus_minus_wowy,
    carries_into_penalty_area

ERROR DE CLASIFICACIÓN: 0.1001
```

Apéndice D

Código fuente y datos utilizados

El código fuente utilizado, hecho en *R* y *python*, y los datos se pueden encontrar en el siguiente repositorio de *OneDrive*: `https://uvaes-my.sharepoint.com/:f:/g/personal/victor_mulero_estudiantes_uva_es/Ep2AVYxQx25PstY_7BZwgPQBG9DDdWtubOxr9il7tuWxVA?e=bnrzZJ`.