Universidad de Valladolid

Facultad de Ciencias

TRABAJO DE FIN DE GRADO

GRADO EN ESTADISTICA

APLICACIÓN DE TÉCNICAS DE CLASIFICACIÓN A DATOS DE LIGAS EUROPEAS DE FÚTBOL

Autor

Juan González Magdalena

Tutores

Miguel Alejandro Fernández Temprano
José Belarmino Pulido Junquera

Resumen

El fútbol es uno de los deportes más practicados y populares a nivel mundial. Este deporte tiene una complejidad táctica muy elevada por lo que se requiere un análisis detallado y precisos para un mejor entendimiento de las características que componen un jugador. Este trabajo de fin de grado se centra en el refinamiento de técnicas de clústering con el objetivo principal de mejorar la precisión en la clasificación de jugadores en cinco perfiles específicos.

Para ello se han obtenido diferentes estadísticas *in-game* de las 5 grandes ligas europeas entre las temporadas 2017-2018 y 2022-2023. Con ellas realiza una comparación entre técnicas de clasificación supervisadas y no supervisadas para evaluar su eficacia a la hora de identificar de perfiles de jugadores, buscando minimizar el error de clasificación. Además, se lleva a cabo un análisis comparativo entre los clústers obtenidos y las cinco grandes ligas europeas.

La optimización de estas técnicas en el contexto futbolístico tiene como finalidad extraer patrones interesantes sobre cada perfil de jugador, dando así una visión global sobre las dinámicas de los diferentes perfiles a lo largo de las temporadas en las grandes ligas europeas.

Abstract

Football is one of the most widely practiced and popular sports worldwide. This sport involves a high level of tactical complexity, demanding detailed and precise analysis to better understand the characteristics that make a player. This undergraduate thesis focuses on refining clustering techniques with the primary objective of improving the accuracy in classifying players into five specific profiles.

For this purpose, different in-game statistics from the five major European leagues between the 2017-2018 and 2022-2023 seasons were collected. A comparison is made between supervised and unsupervised classification techniques to evaluate their effectiveness in identifying player profiles, aiming to minimize classification error. Additionally, a comparative analysis between the obtained clusters and the five major European leagues is conducted.

The optimization of these techniques in the football context aims to extract interesting patterns about each player profile, thus providing a global overview of the dynamics of the different profiles throughout the seasons in the major.European leagues.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción y Objetivos

1.1. Introducción

Tradicionalmente el fútbol se inició en Europa más específicamente en Inglaterra a finales del siglo XIX. Durante los primeros años del desarrollo de este deporte los sistemas tácticos eran escasos, incluso inexistentes. A medida que fueron pasando los años el fútbol se fue profesionalizando poco a poco y con ello la importancia táctica dentro de este deporte, hasta día de hoy donde se caracteriza por su complejidad táctica y la importancia de habilidades individuales.

Durante las últimas décadas se han ido recopilando multitud de diferentes datos sobre este deporte. En el contexto actual del fútbol moderno, la evaluación de jugadores y equipos ha evolucionado significativamente en parte gracias al uso de tecnologías avanzadas para mejorar el rendimiento y la toma de decisiones. Algunas de estas aplicaciones podrían ser el análisis de rendimiento, scouting y reclutamiento [6], estrategias de juego, prevención de lesiones [11], entre muchas otras.

En este trabajo nos enfocaremos en la utilización de estadísticas in-game como son tiros completados, pases completados, goles a favor, entre muchas otras para ver que perfiles de jugadores podemos obtener. Además aplicaremos diferentes técnicas de clústering y de clasificación supervisada con el objetivo de mejorar la identificación de variables relevantes para la clasificación de jugadores en distintos perfiles mencionados anteriormente. La exploración de estas técnicas busca no solo obtener una clasificación más precisa, sino también comprender mejor las características distintivas de los jugadores en las cinco grandes ligas europeas.

Este estudio es una continuación de las investigaciones previas realizadas por Víctor Mulero [14] y Mario Garrido [20], quienes exploraron la distinción entre las cinco principales ligas europeas en términos de estilos de juego predominantes. El trabajo de Mulero, en particular, se centró en responder a la pregunta sobre la diferencia de estilo de juego de las 5 grandes ligas europeas.

1.2. Objetivos

Este trabajo tiene como objetivo principal el refinamiento de técnicas de clústering y clasificación supervisada, buscando mejorar las técnicas de agrupamiento y clasificación utilizadas anteriormente, con el fin de obtener una clasificación más precisa de los jugadores en 5 perfiles diferentes: DEFENSA, MEDIO-CENTRO DEFENSIVO, MEDIO CENTRO, MEDIO-CENTRO OFENSIVO y DELANTERO. Para conseguir este objetivo se proponen varias tareas:

Trataremos de analizar las diferentes variables que influyen en la clasificación de los jugadores en los diferentes perfiles mencionados en las cinco grandes ligas europeas, considerando aspectos como tiros a puerta, goles marcados, entradas realizadas con éxito, entre otros.

Finalmente, se comparan técnicas de clasificación tanto supervisadas como no supervisadas para evaluar su eficacia en la identificación de perfiles de jugadores, buscando minimizar el error de clasificación. Se llevará a cabo un análisis comparativo entre los clústers obtenidos y las cinco grandes ligas, con el objetivo de extraer información valiosa sobre los diferentes patrones y características distintivas de cada liga.

Este trabajo de fin de grado trata de optimizar las diferentes técnicas aplicadas en el contexto futbolístico, con el objetivo de extraer patrones interesantes sobre cada uno de los perfiles analizados reduciendo al máximo su error de clasificación. Con esto se pretende brindar un visión sobre las dinámicas de los diferentes perfiles en las 5 grandes ligas europeas.

1.3. Asignaturas relacionadas

El desarrollo de este Trabajo de Fin de Grado ha requerido la aplicación de conceptos aprendidos en diversas asignaturas a lo largo del grado. Entre las más relevantes se encuentran Análisis Multivariante, Análisis de Datos y Regresión y ANOVA. A continuación detallaremos más en profundidad los conceptos de las asignaturas a este trabajo.

- **Análisis de Datos**: Asignatura impartida en 3º de grado que asienta las bases de técnicas como el ACP (Análisis de Componentes Principales) y el análisis de correspondencias a la par que da una introducción a técnicas de clústering.
- **Análisis Multivariante**: Asignatura impartida en 3º de grado, donde se profundiza los conceptos aprendidos en la asignatura de Análisis de Datos como técnicas de clústering. Además de añadir conceptos como el análisis de correspondencias y análisis factorial.
- **Regresión y ANOVA** Asignatura Impartida en 3º de grado que enseña conceptos sobre el análisis de la varianza (ANOVA) de las diferentes formas que existen.

Cabe destacar que no solo se han utilizados conceptos aprendidos durante la carrera. Durante el desarrollo de este trabajo se han aprendido conceptos relacionados con estas asignaturas en los cuales no se ha profundizado como es en Método de selección de variables Stepwise.

Capítulo 2

Marco Teórico

Durante este capítulo trataremos diferentes conceptos utilizados en la resolución del problema planteado.

2.1. PCA

El análisis en componentes principales trata de un procedimiento multivariante de interdependencia. Es decir que no hay variable respuesta [22]. Se trata de estudiar relaciones entre las variables y los individuos presentes en el conjunto de datos, no predecir una variable a partir de otras.

El objetivo fundamental es reducir la dimensión del conjunto de datos minimizando la pérdida de información. Es es posible construyendo un nuevo conjunto de variables a partir de combinaciones lineales de las variables originales que recojan la mayor cantidad de información posible. Muchas veces se utiliza como paso previo en otras técnicas multivariantes como Regresión o Análisis Cluster.

2.1.1. Procedimiento del PCA

El PCA se basa en la construcción de nuevas variables como combinaciones lineales de las variables originales. Estas nuevas variables son las denominadas componentes principales. Estas componentes principales están ordenadas de forma que la primera componente principal captura la mayor variabilidad de los datos, la segunda componente captura la mayor variabilidad restante y así hasta completar toda la variabilidad permitida.

Partiendo de un conjunto de datos X de dimensión $n \times k$, donde n es el número de observaciones y k es el número de variables, se busca una matriz U de autovectores y una matriz Δ de autovalores tales que:

$$X'X = U\Delta U'$$

Denominaremos $X'X$ como la matriz de Covarianzas de los datos centrados X. Una vez definidos los conceptos especificaremos por pasos el procedimiento para realizar el PCA.

Paso 1: Centrado de Datos

Restamos la media de cada variable para centrar los datos alrededor del origen de la siguiente forma.

$$X_{\text{centrado}} = X - \overline{X}$$

Paso 2: Calculo de la Matriz de Covarianzas

$$\text{Cov}(X) = \frac{1}{n-1} X'_{\text{centrado}} X_{\text{centrado}}$$

Paso 3: Descomposicion en Valores Singulares (SVD)

Calcular los autovalores y autovectores de la matriz de covarianzas.

Paso 4: Construcción de Componentes Principales

Los autovectores obtenidos son los coeficientes de las variables iniciales en las componentes. Las proyecciones de los datos sobre estos vectores proporcionan las nuevas variables reducidas.

2.1.2. Formas de representación

En esta sección se describen los posibles gráficos que se pueden realizar para un PCA, utilizaremos R con el conjunto de datos USArrest para realizar las diferentes representaciones. Este conjunto de datos contiene estadísticas sobre el número de arrestos por asalto, asesinato y violación, así como el porcentaje de urbanización, en cada uno de los 50 estados de EE. UU. en 1973.

Scree Plot

El Scree Plot 2.1 contiene en el eje x las diferentes dimensiones ordenadas de menor a mayor. En el eje y se muestra la proporción de variabilidad explicada por cada componente principal.

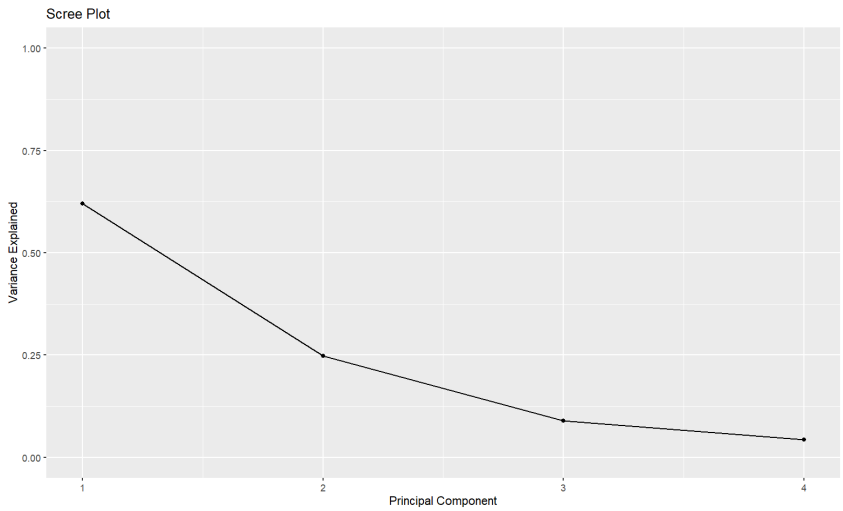


Figura 2.1: Ejemplo scree plot USArrest

Gráfico de Cargas

El gráfico de cargas (loadings plot) 2.2 nos proporciona información sobre cómo las variables originales contribuyen a los componentes principales. Las variables con cargas similares estarán cerca unas de otras en el gráfico, lo que sugiere que están correlacionadas. Si las variables se encuentran próximas al circulo, es decir, alejadas del origen, contribuyen significativamente a la variabilidad explicada por las componentes principales.

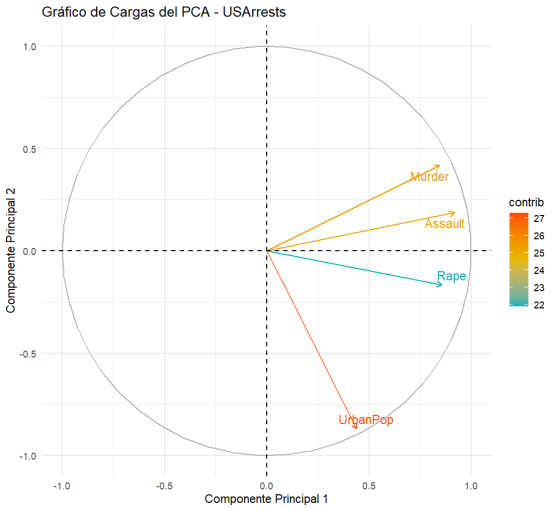


Figura 2.2: Ejemplo gráfico de cargas (Loadings) USArrests

Biplot

El biplot representa simultáneamente las proyecciones de los individuos y las variables en el espacio de las dos primeras componentes principales.

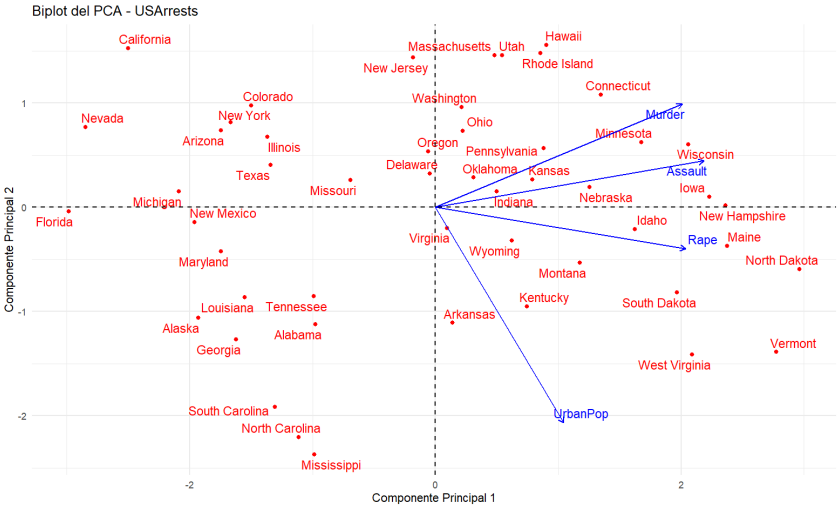


Figura 2.3: Ejemplo biplot USArrests

2.2. Análisis de Correspondencias

El análisis de correspondencias es un procedimiento multivariante de interdependencia, al igual que el ACP como ya comentamos en la sección anterior. Esto quiero decir que no hay variable respuesta [21], pero en este caso a diferencia del ACP, se consideran variables categóricas. Se busca estudiar las relaciones entre las variables y sus categorías.

El objetivo fundamental del análisis de correspondencias es representar en un número reducido de dimensiones, las similitudes y diferencias existentes entre las categorías de las variables para poder estudiar posibles asociaciones entre las mismas.

Dependiendo del número de variables categóricas podemos diferenciar dos Análisis de correspondencias diferentes:

- **Análisis de Correspondencias Simple**: Disponemos únicamente de dos variables categóricas.
- **Análisis de Correspondencias Múltiple**: Disponemos de más de dos variables categóricas.

2.2.1. Análisis de Correspondencias simple

Test de independencia X2

La siguiente tabla 2.1 muestra una tabla de contingencia teórica, donde r es el número de filas (rows), c el de columnas columnas (columns) y n_{++} el número total de individuos.

X \Y					
		Cl. 1	...	Cl. c	n_{1+}
	Cl. 1	n_{11}	...	n_{1c}	n_{1+}
⋮	⋮	⋮		⋮	⋮
	Cl. r	n_{r1}	...	n_{rc}	n_{r+}
	n_{+1}	n_{+1}	...	n_{+c}	n_{++}

Cuadro 2.1: Tabla de Contingencia Teórica

En la tabla anterior 2.1 n_{ij} es el número de individuos de la muestra que están en la clase i de la variable X y en la clase j de la variable Y . El símbolo + en un subíndice quiere decir que se ha sumado en los valores de ese índice. De ese modo n1+ es el número de individuos que están en la clase 1 de la variable X.

Tenemos que la hipótesis de independencia es

$$H_0 : p_{ij} = p_{i+} \times p_{+j}, \quad \forall i = 1, \ldots, r,\ j = 1, \ldots, c$$

El estadístico X2 compara las diferencias entre ambos valores en la tabla del modo siguiente:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n_{++}}\right)^2}{\frac{n_{i+}n_{+j}}{n_{++}}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\text{Obs}_{ij} - \text{Esp}_{ij})^2}{\text{Esp}_{ij}} \text{ asint.} \sim \chi^2_{(r-1)(c-1)}$$

El cuadrado del numerador evita que las diferencias positivas se compensen con las negativas mientras que el denominador da más relevancia a las diferencias que se producen en celdas con valores esperados bajos.

Calculando el p_valor nos daría si rechazamos la hipótesis de independencia o no.

Perfiles fila y Columna

Vemos ahora que hemos rechazado la Hipótesis de Independencia la relación que existe entre ellas. Para ello vamos a analizar los denominados perfiles fila y columna.

Perfil de fila i

$$\left(\frac{n_{i1}}{n_{i+}}, \frac{n_{i2}}{n_{i+}}, \ldots, \frac{n_{ic}}{n_{i+}}\right)$$

Perfil de columna j

$$\left(\frac{n_{1j}}{n_{+j}}, \frac{n_{2j}}{n_{+j}}, \ldots, \frac{n_{rj}}{n_{+j}}\right)$$

Cabe destacar que los perfiles tanto fila como columna no tienen todos el mismo peso.

Calculo de la distancia Chi-Cuadrado

La distancia que se utiliza para medir las diferencias entre los perfiles es la denominada distancia χ^2, que es una distancia euclídea ponderada entre los perfiles.

Distancia Chi-Cuadrado entre perfiles de fila

$$d^2_{\chi^2}(i, i') = \sum_{j=1}^{c} \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}}\right)^2$$

Distancia Chi-Cuadrado entre perfiles de columna

$$d^2_{\chi^2}(j, j') = \sum_{i=1}^{r} \frac{1}{f_{i+}} \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{ij'}}{f_{+j'}}\right)^2$$

Inercia Total

La inercia total es una medida de la variabilidad total en la tabla de contingencia. La calcularemos como la suma de los cuadrados de las distancias de los perfiles al centro de gravedad teniendo en cuenta además los pesos de cada perfil.

$$\text{Inercia total} = \sum_{i=1}^{r} \text{peso}(f_{\text{fila}i}) d^2_{\chi^2}(f_{\text{fila}i}, G_{\text{filas}}) = \sum_{j=1}^{c} \text{peso}(\text{col}_j) d^2_{\chi^2}(\text{col}_j, G_{\text{col}})$$

Equivalentemente podemos escribir la inercia como:

$$\text{Inercia total} = \frac{\chi^2}{n_{++}}$$

Podemos descomponer la inercia total calculada en componentes principales. De esta forma obtenemos los autovalores y autovectores que representan la dirección de máxima variabilidad. Utilizando multiplicadores de Lagrange para maximizar una función bajo restricciones se obtienen las siguientes ecuaciones características.

Proyecciones en Ejes Principales

Finalmente se proyectan los perfiles de filas y columnas sobre los ejes principales obtenidos en el paso anterior. Las coordenadas de las proyecciones se calculan utilizando los autovectores correspondientes. A continuación podemos ver las formulas del calculo de proyecciones para ambos ejes:

$$\hat{v}_\alpha = (D_r^{-1}F)D_c^{-1}u_\alpha$$

$$\hat{u}_\alpha = (D_c^{-1}F')D_r^{-1}v_\alpha$$

2.2.2. Representación de las proyecciones

Para ejemplificar la representación de este apartado se ha usado el conjunto de datos *HairEyeColor* de R.

Screeplot

El Screeplot en la figura 2.4 representa los valores propios/varianzas ordenados de mayor a menor.

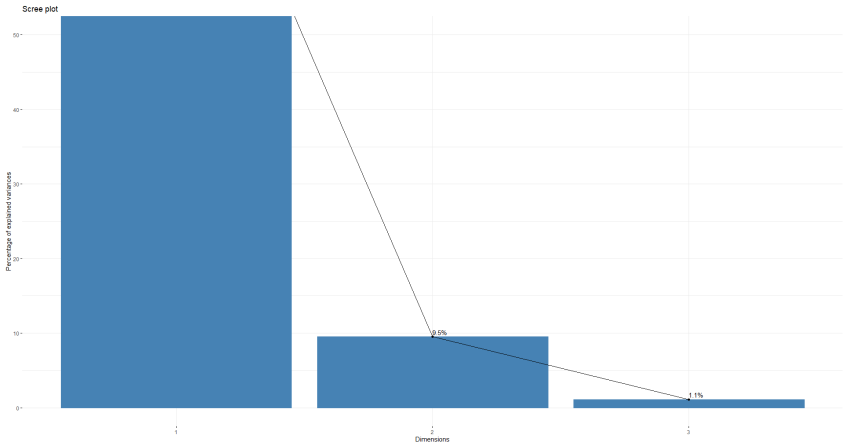


Figura 2.4: Screeplot *HairEyeColor* ejemplo

Biplot

Biplot es una visualización gráfica de filas y columnas en 2 dimensiones. Las filas están representadas por puntos azules y las columnas por triángulos rojos. La distancia entre cualquier punto de fila o columna da una medida de su similitud (o disimilitud). Los puntos similares están más próximos en el gráfico. La relación entre la proximidad entre los perfiles fila y columna similares en el gráfico suele indicar un relación. En la figura 2.5 se muestra un ejemplo.

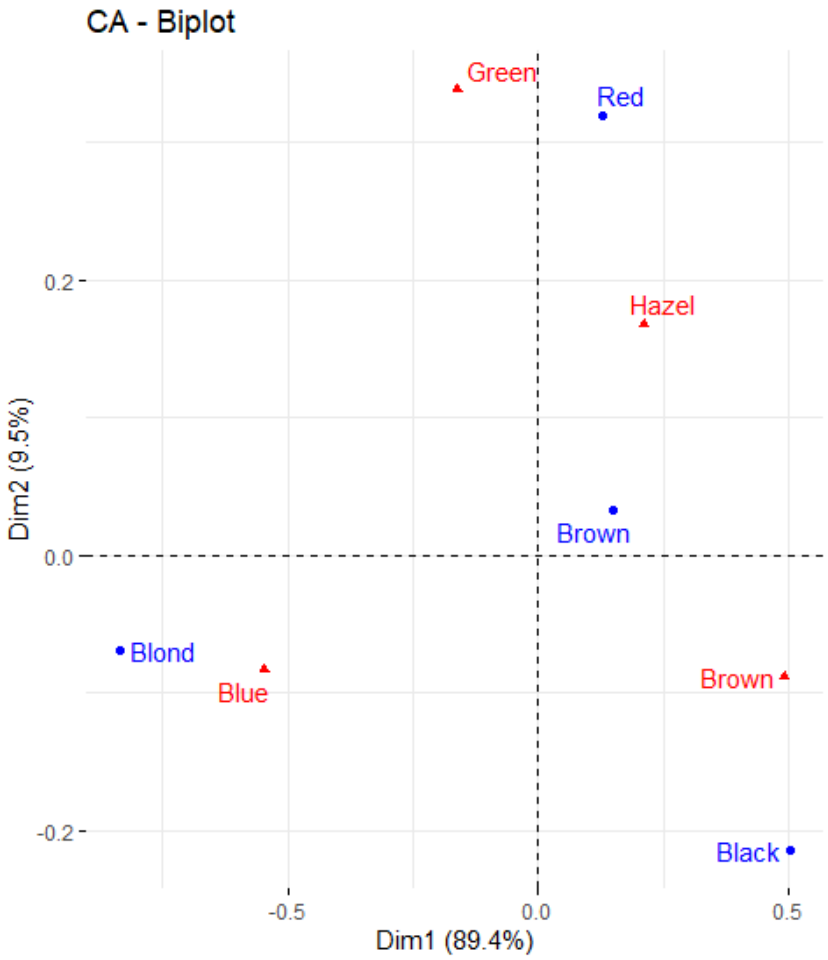


Figura 2.5: Biplot *HairEyeColor* perfiles fila columna

2.3. Métodos de selección de variables

A lo largo de esta sección vamos a ver métodos parcialmente heurísticos [10]. Estos algoritmos se basan en la comparación de modelos con diferentes variables donde se eligen las mejores.

2.3.1. Selección hacia adelante (Forward Selection)

En el método de selección hacia delante parte de un modelo vacio y en cada iteración se añade la variable más significativa de acuerdo a un criterio. Lo podemos esquematizar en los siguientes pasos:

Paso 1: Se genera el modelo nulo $y = \beta_0 + \epsilon$.
Paso 2: Se generan un número n de modelos, donde cada uno introduce una de las variables no presentes en el modelo. Se comparan estos n modelos y el modelo del paso anterior entre si, seleccionando el mejor de ellos según el criterio de comparación establecido (por ejemplo, el que minimiza la suma de cuadrados residual, RSS).
Paso 3: El modelo seleccionado en el paso 2 se utiliza como base para generar nuevos modelos, añadiendo una nueva variable no seleccionada al modelo. Posteriormente se comparan estos modelos y el modelo base seleccionado en el paso anterior, eligiendo el mejor según el mismo criterio de comparación.
Paso 4: Se repiten los pasos 2 y 3 hasta que no haya una mejora significativa al incluir nuevas variables. El algoritmo se detiene por un criterio de parada, es decir, si el mejor modelo en un paso es el mismo que el del paso anterior, indicando que no se mejora al añadir más variables.

2.3.2. Algoritmo stepwise hacia atrás (Backward elimination)

En el método de selección de variables hacia atrás se basa en la introducción de todas las variables en un modelo y después se van excluyendo una tras otra. Se puede esquematizar en los siguientes pasos:

Paso 1: Se genera el modelo completo $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
Paso 2: Se calculan los modelos con $p-1$ variables, eliminando una variable a la vez del modelo actual. Se comparan estos p modelos y el modelo del paso anterior, seleccionando el que tiene el menor impacto en el ajuste, es decir, la variable con el menor Z-score.
Paso 3: El modelo seleccionado en el paso 2 se utiliza como base para generar nuevos modelos, eliminando una variable a la vez de entre las restantes. Se comparan estos modelos y el modelo base, eligiendo el mejor según el mismo criterio de comparación.
Paso 4: El proceso se repite iterativamente, eliminando variables del modelo, hasta que la eliminación de cualquier variable adicional no mejore significativamente el modelo. El algoritmo se detiene cuando el mejor modelo en un paso es el mismo que el del paso anterior, indicando que no se mejora al eliminar más variables.

2.4. Métodos de clasificación

Podemos diferenciar dos tipos de clasificación a la hora de clasificar individuos. Pueden ser métodos supervisados o no supervisados.

- Clasificación Supervisada: Este tipo de clasificación cuenta con un conocimiento a priori, es decir para la tarea de clasificar un objeto dentro de una categoría o clase contamos con individuos ya clasificados (objetos agrupados que tienen características comunes). Podemos encontrar diferentes técnicas como Random Forest, SVM (Support Vector Machine), LDA (Análisis de Discriminante Lineal), etc.

- Clasificación No Supervisada: A diferencia de la supervisada no contamos con conocimiento a priori, por lo que tendremos un área de entrenamiento disponible para la tarea de clasificación. A la clasificación no supervisada se la suele llamar también clústering. En este tipo de clasificación contamos con "objetos" o muestras que tiene un conjunto de características, de las que no sabemos a que clase o categoría pertenece, entonces la finalidad es el descubrimiento de grupos de "objetos" cuyas características afines nos permitan separar las diferentes clases. Podemos distinguir entre dos tipos de Clústering, jerárquico y no jerárquico.

2.4.1. Clústering jerárquico

Los métodos de clústering jerárquico se basan en la construcción de una estructura en la que los elementos se van agrupando en subconjuntos cada vez mayores hasta que todos pertenezcan al mismo conjunto . De esta forma no se muestra un agrupamiento, sino las relaciones de proximidad entre los elementos [1]. Existen dos tipos diferentes:

- Disociativo: Se parte de un único clúster al que pertenecen todos los elementos. En cada iteración se escoge un clúster y se divide.

- Aglomerativo: Inicialmente se forman clústers individuales, cada uno de los cuales contiene a un único elemento. En cada iteración se unen los dos clúster más próximos. El procedimiento finaliza cuando solo haya un clúster.

A continuación, vamos a profundizar en el clústering jerárquico aglomerativo, para ello vamos a definir los conceptos de índice de disimilaridad e índice de agrupación.

- Índice de disimilaridad: Mide que tan diferentes son dos clústers entre sí. Se expresa como d(x,y) y cumple las siguientes propiedades: $d(x, y) > 0$, $d(x, x) = 0$ y $d(x, y) = d(y, x)$.

- Índice de agregación: Mide que tan apropiado es juntar dos clústers en un paso determinado del proceso de agrupamiento. Se define como $\delta(A, B)$ donde A y B son clústers. Uno de los métodos más conocido es el método de Ward.

Para la visualización de los clústers creados a partir del del clústering jerárquico se usa el dendograma. Muestra cómo se van agrupando los objetos o clusters a medida que se aumenta el nivel de disimilaridad. A continuación en la figura 2.6 se muestra un ejemplo.

2.4.2. Método de las K-Medias

K-medias es un método de agrupamiento [24], que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

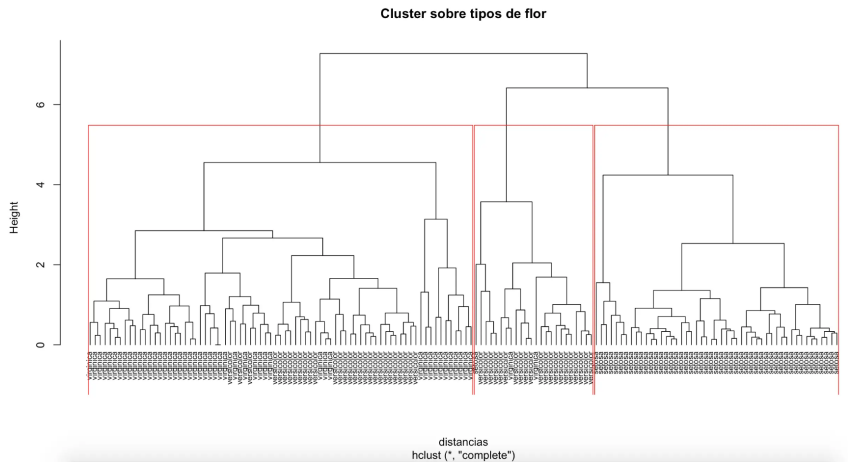


Figura 2.6: Ejemplo dendograma sobre los tipos de flor [2]

Algoritmo

Podemos dividir el algoritmo en tres pasos diferenciados:

Paso 1: Elegir k centros al azar (por ejemplo, k observaciones de nuestros datos) $C_1^0, \ldots, C_k^0$ y particionamos las observaciones en k grupos $I_1^0, \ldots, I_k^0$ usando las observaciones más próximas a cada uno de los centros.

Paso 2: Los nuevos centros $C_1^1, \ldots, C_k^1$ se obtienen calculando las medias de las observaciones en cada uno de los grupos $I_1^0, \ldots, I_k^0$. Se obtiene una nueva partición $I_1^1, \ldots, I_k^1$ usando criterios de proximidad.

Paso 3: Repetir los pasos 2 y 3 hasta que todas las particiones se estabilicen.

Elección del Número de Clústers

Es un problema complicado porque depende de la aplicación que el usuario tenga en mente. Pero existen diferentes métodos que nos ayudan a tomar decisiones "razonables" dependiendo del conjunto de datos. Realizaremos los ejemplo con el conjunto de datos de R USArrests.

Método del codo

El método del codo es uno de los métodos más populares para determinar este valor óptimo de k. Para determinar el número óptimo de clústers, tenemos que seleccionar el valor de k en el *codo*, es decir, el punto después del cual la distorsión/inercia comienza a disminuir de forma lineal.
En la figura 2.8 vemos como se representan la suma de los cuadrados en cada clúster. A la vista de esta figura podemos deducir que el número óptimo es de tres clústers.

Método de la silueta

El Método de la silueta también es un método para encontrar el número óptimo de clústers. El valor de silueta es una medida de cuán similar es un objeto a su propio clúster (cohesión) en comparación con otros clústers (separación). Este valor oscila entre [1, -1], donde un valor alto indica que el objeto se corresponde bien con su propio clúster y no a los clústers vecinos. Si la mayoría de los objetos tienen un valor alto, entonces la configuración de clústers es apropiada. En caso contrario indicaría un exceso en el número de clústers, por lo que se debería reducir.

Para obtener la puntuación de silueta hay que calcular los coeficientes de silueta para cada uno de los puntos y realizar el promedio de todas las muestras. Los pasos para encontrar el coeficiente de silueta de un i-ésimo punto:

- a (i): la distancia promedio de ese punto con todos los demás puntos en los mismos grupos.
- b (i): la distancia promedio de ese punto con todos los puntos en el grupo más cercano a su grupo.
- s (i) - coeficiente de silueta o i-ésimo punto utilizando la fórmula mencionada a continuación. $s(x) = \frac{b(x)-a(x)}{\text{máx}\{a(x),b(x)\}}$

Después de calcular el coeficiente de silueta de cada punto en el conjunto de datos, vamos a representarlo con un ejemplo 2.7. Podemos ver que en este caso el numero óptimo de clústers es k = 2. Con un valor de silueta de 0.4.

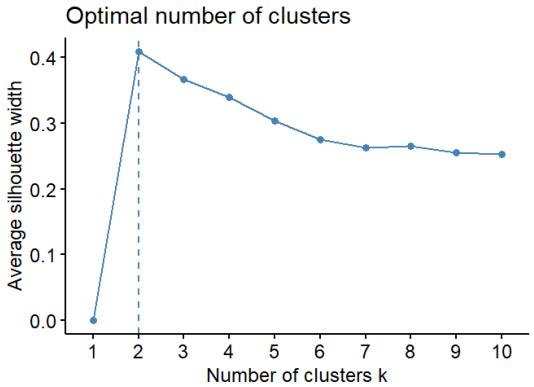


Figura 2.7: Método de la silueta

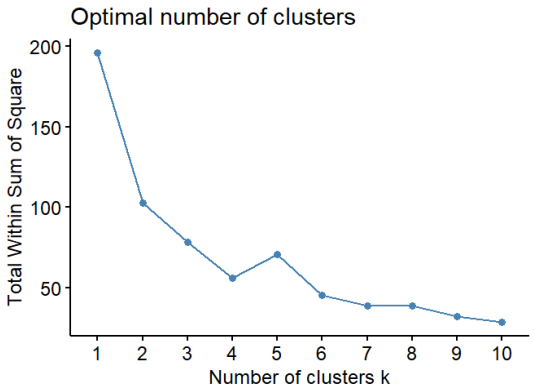


Figura 2.8: Método del codo selección de clústers óptimo

Capítulo 3

Datos

En esta sección vamos a tratar la procedencia de los datos que utilizaremos a lo largo de todo el trabajo asi como su procedencia, las modificaciones realizadas al dataset original e información general sobre el conjunto de datos que utilizaremos

3.1. Bases de datos disponibles

Disponemos del conjunto de datos utilizado en otros proyectos [20] y trabajos [14] con las características necesarias para abordar el problema planteado. Aún así se ha deseado explorar otras alternativas que pudiesen mejorar al conjunto de datos ya conocido. Se barajaron dos opciones que se describen a continuación

StatsBomb

StatsBomb [17] es una plataforma líder en el análisis de datos deportivos, enfocada en proporcionar datos precisos y detallados para facilitar la toma de decisiones en el fútbol. Dentro de esta plataforma encontramos un apartado de datos gratuitos donde se ofrecen multitud de ligas entre las cuales encontramos *"The Big Five"*.

StatsBomb ofrece una colección extensa de datos de fútbol a través de su repositorio de datos abiertos en GitHub [18]. Estos datos, disponibles gratuitamente, incluyen información detallada sobre eventos en partidos de fútbol. El acceso a estos datos es proporcionado mediante su propia librería de R llamada StatsBombR, diseñada para facilitar la descarga y el análisis de la información disponible.

La librería StatsBombR permite a los usuarios acceder a una variedad de datos históricos de partidos, así como métricas avanzadas y análisis tácticos. StatsBomb ha desarrollado varios tutoriales detallados sobre el uso de estos datos y su libreria. Estos tutoriales cubren desde los conceptos básicos para los principiantes hasta técnicas más avanzadas para los usuarios experimentados.

Sin embargo, a pesar de que los datos disponibles nos eran suficientes para llevar a cabo nuestro estudio, se descartó debido a la necesidad de comprender los formatos y estructuras de los datos que utiliza StatsBombR.

footystats.org

Footystats.org es una plataforma en línea que ofrece estadísticas detalladas tanto de partidos en tiempo real como partidos pasados, abarcando más de 1500 ligas alrededor del

mundo [5]. Esta plataforma ofrece algunos partidos de diferentes ligas de forma gratuita para descargar en formato CSV. También ofrece una API en las que podemos obtener un amplia gama de estadísticas para la mayoría de ligas a nivel mundial en formato JSON.

Se descartó la idea de usar esta base de datos debido a la complejidad en la utilidad de la API para obtener los datos necesarios para abordar este trabajo y el costo de 30 euros mensuales para usar este servicio.

Finalmente debido a que estás opciones no suponían mejoras al conjunto de datos de partida, se ha optado por utilizar el conjunto de datos inicial.

3.2. Procedencia del conjunto de datos

Los datos obtenidos para el desarrollo de este proyecto han sido obtenidos de la página web de kaggle.com.

3.2.1. Kaggle.com

Kaggle, es una comunidad en línea de científicos de datos y profesionales del aprendizaje automático. Kaggle permite publicar conjuntos de datos, explorar y crear modelos en un entorno de ciencia de datos basado en la web.

Tras la exploración de diversos conjuntos de datos de diversas fuentes mencionadas en el apartado anterior, obtamos por el conjunto de datos *"Soccer players values and their statistics"*. Utilizado en el TFG de Victor Mulero [14]. Debido a que recopilaba toda la información requeridas para nuestro analisis. Número de goles, número de asistencias, pases realizados, etc. Para las 5 grandes ligas europeas. Además de sus fiabilidad ya que había sido utilizada en otros proyectos.

Este conjunto de datos combina información detallada de dos fuentes principales, transfermarkt.de y fbref.com, abarcando tres temporadas de fútbol(2017/18, 2018/19 y 2019/20). El objetivo del conjunto se datos es analizar y modelar los valores de los jugadores de fútbol y los factores determinantes en el mercado de transferencias.

El conjunto de datos fusionado contiene las siguientes características:

- Información de jugadores: Incluye nombre, edad, nacionalidad, posición y equipo.
- Estadísticas de rendimiento: Datos detallados sobre rendimiento en el campo como goles, asistencias, minutos jugados, pases, interceptaciones, entre otras métricas relevantes.
- Valores de mercado: Valores estimados de los jugadores en el mercado de transferencias, proporcionados por transfermarkt.de.
- Detalles de traspasos: Información sobre transferencias de jugadores, incluyendo costos de transferencias y movimientos entre clubes.

3.3. Modificaciones realizadas al conjunto de datos

Como hemos mencionado anteriormente el conjunto de datos *"Soccer players values and their statistics"*. Recopilaba información de varios ámbitos. El ámbito que nos interesa son las estadísticas procedentes de fbref.com. Únicamente para 3 temporadas (2017-2018, 2018-2019 y 2019-2020).

Mediante la utilización del código proporcionado por RSKRIEGS [9] recopilamos las estadísticas de las 5 ligas para las temporadas 2017-2018 hasta la temporada 2022-2023. Recopilamos únicamente información desde la temporada 2017-2018 ya que a partir de esta cantidad de variables recogidas por la página fbref.com cambia y pasan a recopilarse 205 variables [14]. Únicamente obteniendo datos de fbref.com. El resultado es un único dataset con más de 15000 observaciones y las más de 200 variables mencionadas anteriormente.

Capítulo 4

Exploración de los datos

Durante este capítulo realizaremos una exploración sobre el conjunto de datos

4.1. Conjunto de datos inicial

Como hemos mencionado anteriormente disponíamos de un total de 15372 jugadores durante 5 temporadas, lo cual representa una cantidad de datos razonable para llevar a cabo el análisis. Las 205 variables incluían información como el nombre del jugador, su posición, la liga en la que jugó, los goles marcados, las asistencias, los minutos jugados, entre otros aspectos relevantes.

4.2. Análisis

Durante esta sección realizaremos un análisis descriptivo sobre el dataset con el objetivo de poder entender más en profundidad el conjunto de datos.

4.2.1. Número de partidos

Hay que tener en cuenta que estos jugadores puede que no hayan jugado algún partido o no lleguen a un número de partidos suficientemente elevado para contemplarlos en el análisis posterior. Vamos a 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. Utilizaremos como medida de corte el cuantil 0.25 es el valor que deja por debajo al 25 % de las observaciones. Calculando para las diferentes temporadas obtenemos los resultados en la siguiente tabla.

Temporada	2017-2018	2018-2019	2019-2020	2020-2021	2021-2022	2022-2023
Partidos	10	10	9	10	9	9

Cuadro 4.1: Cuantil 0.25 de partidos por temporada

Podemos observar en la tabla 4.2 como el número de partidos mínimo para el cuantil 25 % en las 5 ligas es de unos 10 partidos de media. Por lo que la tomaremos como medida de corte, eliminando asi a los jugadores que no cumplan con los 10 partidos.

Liga	Bundesliga	Premier League	La Liga	Ligue 1	Serie A
Partidos	10	11	10	8	10

Cuadro 4.2: Cuantil 0.25 de partidos por liga

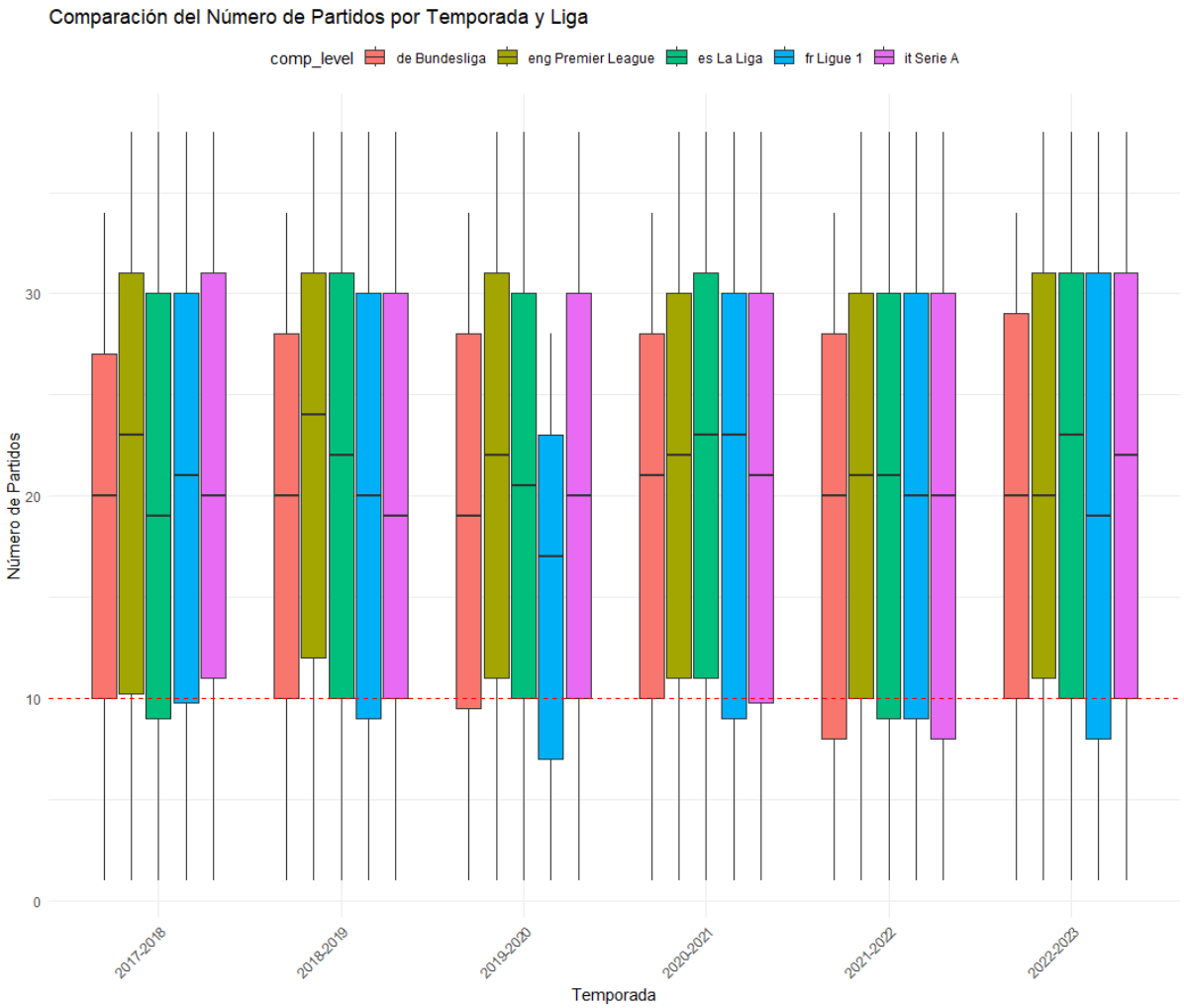


Figura 4.1: BoxPlot número de partidos por temporada y liga

Si observamos esta figura 4.1 observamos que La ligue 1 es la más perjudicada dejando fuera a más observaciones de esa liga que del resto. La temporada que más valores deja fuera es la temporada 2019-2020 y 2021-2022. Si eliminamos las observaciones que no llegan a 10 partidos nos quedaría un total de 11513 observaciones. Estas serán las observaciones con las que trabajaremos en los siguientes apartados.

4.2.2. Jugadores por temporada

Para poder realizar un análisis coherente necesitamos saber el número de jugadores que tenemos por cada temporada y por cada liga. Es importante para saber si el análisis es equilibrado, es decir, se tiene un numero similar de jugadores por cada temporada y en cada liga. Se muestran los resultados en la tabla 4.3.

Podemos observar una tendencia ascendente en las últimas 4 temporadas. Esto se

	2017-2018	2018-2019	2019-2020	2020-2021	2021-2022	2022-2023
Bundesliga	328	320	347	357	363	367
Premier League	361	363	362	374	382	416
La Liga	396	379	392	415	427	424
Ligue 1	372	372	331	392	419	411
Serie A	383	386	402	423	426	423

Cuadro 4.3: Jugadores por temporada y liga

puede deber al incremento a la carga de partidos en las ultimas temporadas [4]. Lo que hace que los equipos tenga que tener más jugadores en sus plantillas a medida que avanzan las temporadas.

Además podemos observar una clara diferencia entre el número de jugadores de la Bundesliga y el resto de ligas y de La Premier y la Ligue 1 respecto a la Serie A y La Liga. Esto se puede deber a diversas razones, diversidad de número de equipos en cada liga, numero de jugadores máximos en cada plantilla. Es importante saber si los jugadores por liga y temporada tienen la propiedad de homogeneidad. Para ello realizaremos el test de independencia. Obtenemos un p-valor = 0.9805. Esto implica que no hay diferencias significativas en el número de jugadores dependiendo de la temporada y liga.

4.2.3. Posiciones de los jugadores

Vamos a observar como se distribuyen los jugadores a lo largo del terreno de juego de dos formas diferentes. 1º Tomando como referencia únicamente su posición principal. Por lo que analizaremos: Porteros (GK), Defensas (DF), Centrocampistas (MF) y Delanteros (FW). 2º Tomando como referencia las posiciones dobles, es decir, las posiciones principales más (DF-MF) y (MF-FW).

Centrándonos primero en la distribución de en 3 perfiles (excluyendo al portero) de todas las ligas a lo largo de las 6 temporadas se puede ver en la tabla 4.4.

Posición	GK	DF	MF	FW
Nº Jugadores	616	4120	3902	2875
Proporción	5.3 %	35.78 %	33.9 %	24.8 %

Cuadro 4.4: Jugadores por posición principal

Si observamos los 5 perfiles excluyendo al portero en la tabla 4.5 vemos que sigue siendo bastante homogéneo. Pero destaca la escasez de Jugadores de perfil DF-MF.

Posición	GK	DF	MF	FW	DF-MF	MF-FW
Nº Jugadores	616	3720	2506	1611	669	2391
Proporción	5.3 %	32.3 %	21.76 %	14 %	5.8 %	20.76 %

Cuadro 4.5: Jugadores por posición (Doble Posición)

En la figura 4.2 se presenta un gráfico de barras que muestra la frecuencia de jugadores según su posición principal y si disponen o no de una posición secundaria. Las barras en color no incluidas en la leyenda representan la cantidad de jugadores que tienen solo una posición principal sin una posición secundaria. Las barras apiladas en diferentes colores sobre cada posición principal representan las frecuencias de los jugadores que tienen una posición secundaria específica. Esto permite observar tanto las posiciones únicas como la distribución de posiciones secundarias para cada posición principal.

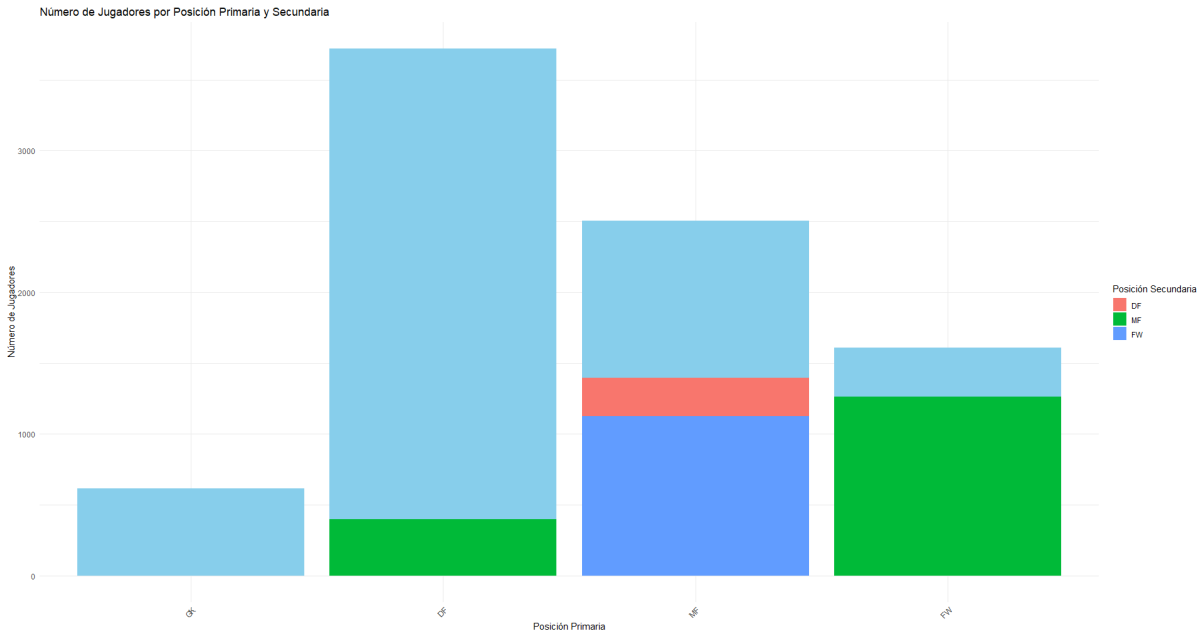
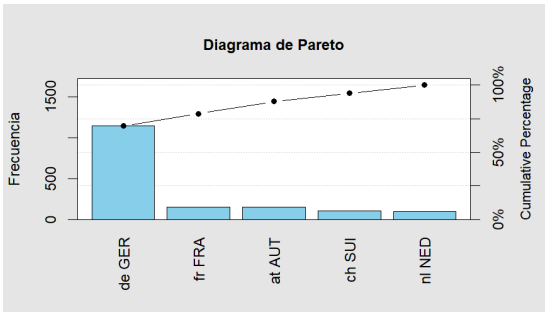


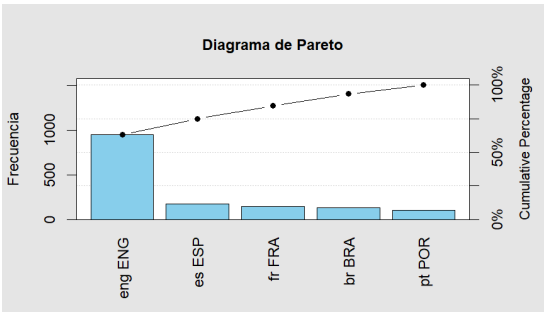
Figura 4.2: Gráfico de barras apiladas posiciones principales y secundarias

4.2.4. Nacionalidad de los jugadores en las ligas

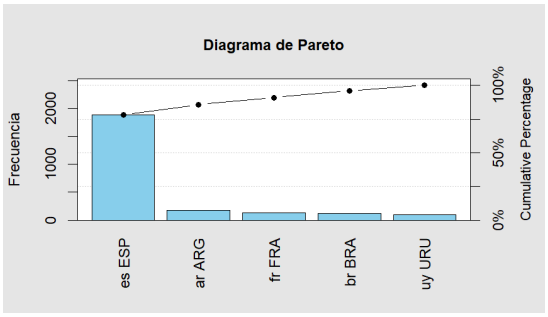
Al analizar las nacionalidades representadas en las diferentes ligas de fútbol, podemos obtener una visión más completa de la dinámica global del deporte más popular del mundo. En este sentido, la Figura 4.3 presenta un análisis visual de la distribución de nacionalidades dentro de varias ligas de fútbol desde la temporada 2017-2018 hasta la 2022-2023.



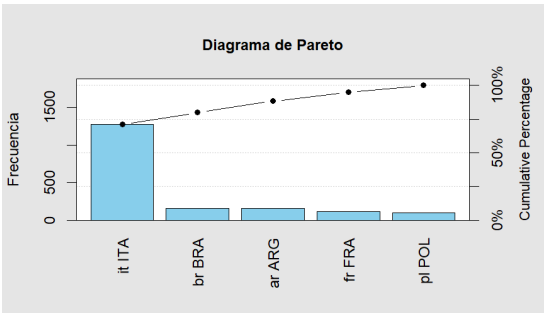
(a) Liga Alemania



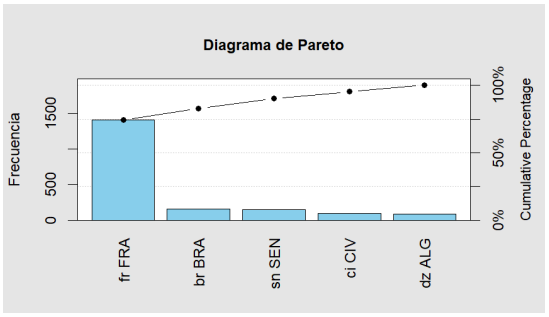
(b) Liga Inglesa



(c) Liga Española



(d) Liga Italiana



(e) Liga Francesa

Figura 4.3: Nacionalidades 5 grandes ligas europeas

Observamos que la nacionalidad más común en cada liga suele ser la del país de origen de dicha liga. Este patrón refleja la tendencia de las ligas a atraer principalmente a jugadores de su propia nacionalidad. Algo totalmente esperado. Pasemos a analizar a la proporción de jugadores extranjeros por temporada y liga.

Temporada	Bundesliga	Premier League	La Liga	Ligue 1	Serie A
2017-2018	0.5823171	0.6952909	0.4393939	0.5806452	0.5639687
2018-2019	0.5812500	0.7190083	0.4511873	0.5618280	0.5932642
2019-2020	0.6051873	0.6823204	0.4234694	0.5589124	0.6218905
2020-2021	0.6218487	0.6711230	0.4024096	0.5459184	0.6382979
2021-2022	0.5895317	0.6937173	0.4590164	0.5942721	0.6713615
2022-2023	0.5912807	0.7043269	0.4410377	0.6058394	0.6808511

Cuadro 4.6: Proporción de jugadores extranjeros por temporada y liga

Atendiendo a la tabla 4.6 vemos que la liga con más proporción de extranjeros es la Premier League y la que menos es La Liga. Esto se puede deber en gran parte a que la Premier League es la liga con mayor gasto en el mercado de fichajes con un gasto de 2.737 millones en comparación a los 432 millones gastados por La Liga en el verano de 2023 [19]. Esto hace que La Liga haga uso de las canteras y de jugadores nacionales. A continuación se presenta la figura 4.4, donde se muestra cómo ha cambiado la proporción de jugadores extranjeros en las principales ligas de fútbol europeas a los largo de las temporadas analizadas.

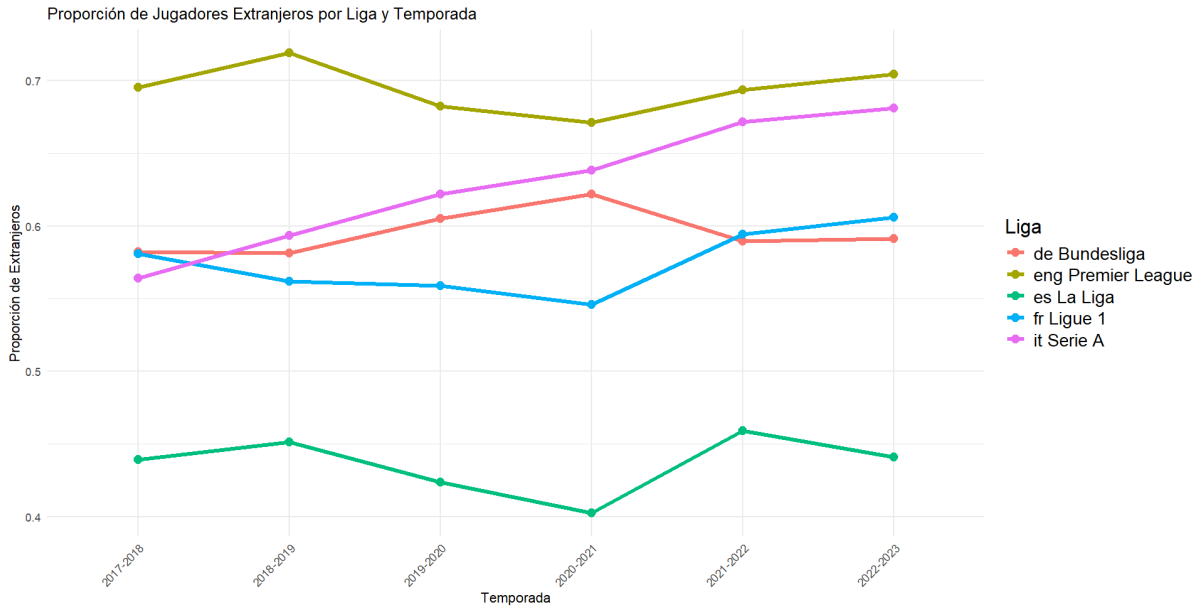


Figura 4.4: Gráfico líneas múltiple proporción de jugadores extranjeros por liga y temporada

4.2.5. Análisis de la edad

En esta sección vamos a analizar la edad de los diferentes jugadores. Ya que es un factor crucial en el deporte de elite.

Podemos observar en la 4.7b que la edad de los jugadores varía considerablemente entre las diferentes posiciones, con los porteros siendo, en promedio, los más veteranos, con una media de 28,5 años mientras que los delanteros son ligeramente más jóvenes, con una media de 25,6 años. La distribución de las edades en general es simétrica y está centrada

Estadística	Valor
Mínimo	15.00
Primer cuartil	23.00
Mediana	26.00
Media	25.99
Tercer cuartil	29.00
Máximo	42.00

(a) Resumen estadístico

Posición	Edad
GK	28.51438
DF	26.19335
MF	25.65809
FW	25.61200

(b) Edad media por posición

Cuadro 4.7: Estadísticas sobre la edad

alrededor de los 26 años, lo que sugiere que los jugadores alcanzan su plenitud física y táctica alrededor de esta edad. Esto también indica una mezcla saludable de juventud y experiencia en el conjunto de jugadores analizados.

En la tabla 4.7 se observa que las medias obtenidas para cada posición son similares pero comprobemos si podemos decir que la media de los jugadores en cada posición es la misma. Para ello se realiza un ANOVA.

Para poder realizar el ANOVA se deben cumplir 4 características:

- Continuidad: La variable ha de ser continua. En nuestro caso la edad, esta es una variable numérica continua
- Independencia: Los valores esperados son independientes entre sí.
- Normalidad: Los datos de los grupos deben tener una distribución normal.
- Homogeneidad: Las varianzas de cada grupo deben ser aproximadamente iguales

Realizamos el ANOVA para la edad en cada posición. A continuación se muestran los resultados obtenidos en la tabla 4.8 tras realizar el ANOVA de un factor. Observamos que la variable position es significativa obtenemos un pvalor muy bajo. Esto nos lleva a rechazar la hipótesis de igualdad de edad en cada posición.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Posición	5	6778	1355.6	81.98	<2e-16 ***
Residuos	11507	190271	16.5		

Cuadro 4.8: ANOVA de un factor edad y posición

Para verificar el ANOVA que hemos realizado vamos a comprobar si se cumplen las condiciones mencionadas anteriormente.

Normalidad

Para verificar la normalidad se realiza una representación de los residuos. En la figura 4.5 observamos que la normalidad es aceptable. Debido a la cantidad de los datos el test de normalidad se va a rechazar en prácticamente cualquier muestra. Se acepta la hipótesis de normalidad

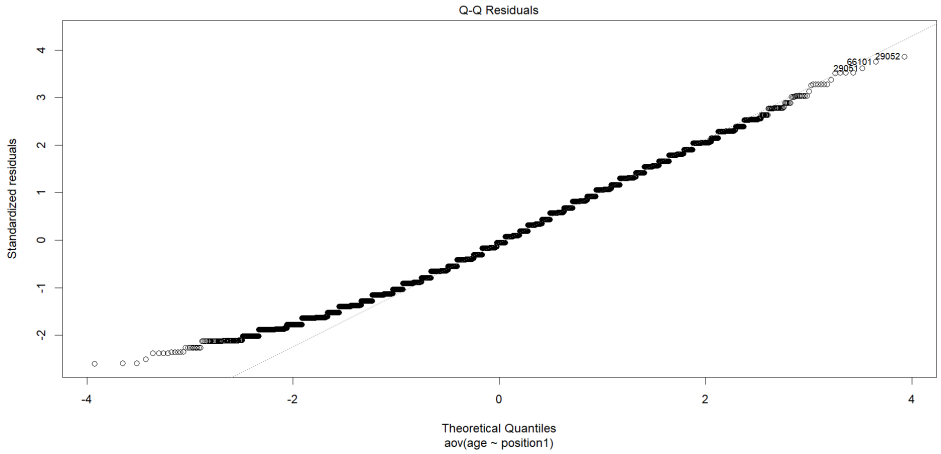


Figura 4.5: Normalidad de los residuos

Homogeneidad

Realizamos el gráfico de residuos vs valores ajustados para comprobar que se mantiene constante la igualdad de Varianzas. Observamos que la figura 4.6 no tiene una forma de embudo [12], esto nos hace aceptar la hipótesis de homogeneidad.

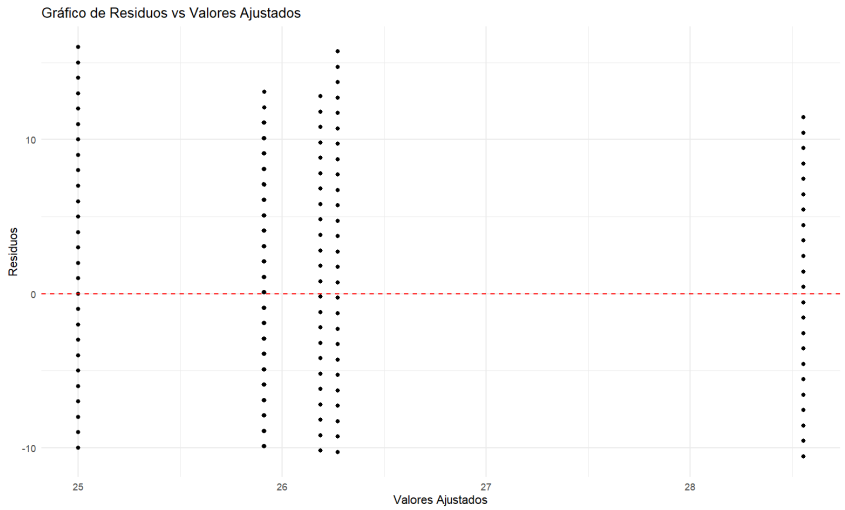


Figura 4.6: Gráfico de residuos vs valores Ajustados

Análisis post-hoc

Como el factor posición ha resultado significativo y hemos comprobado que el ANOVA es válido se procede a realizar el Test de Duncan, los resultados se encuentran en la tabla 4.9.

Posición	Edad promedio	Grupo a	Grupo b	Grupo c
GK	28.55682	a		
DF	26.23083		b	
MF	25.68580			c
FW	25.62539			c

Cuadro 4.9: Resultados del análisis post-hoc de Duncan para las edades promedio según la posición

Los resultados indican que los porteros denotados por la letra 'a' tienen una edad promedio significativamente mayor comparada con los otros grupos. Los defensas denotadas con la letra 'b', indica que son significativamente más jóvenes que los porteros, pero mayores que los medio-centros y delanteros. Los medio-centros y delanteros tienen edades promedio muy similares compartiendo la misma letra de grupo 'c'. Esto indica no hay diferencias significativas en sus edades.

Capítulo 5

Analisis Clúster

Durante esta capítulo se utilizarán técnicas para la selección de variables como PCA y algoritmos Stepwise. Además se usarán métodos clúster para ver las agrupaciones de los jugadores.

5.1. Variables

Tras el análisis descriptivo del conjunto de datos vimos que los porteros no iban a ser relevantes para nuestro objetivo, por lo cual eliminamos estas observaciones y las variables relacionadas a estos. Quedando un total de 10912 observaciones y 164 variables.

En el análisis clúster es primordial hacer una buena selección de variables iniciales, por lo cual antes de empezar nos tomaremos un tiempo para ver qué métodos podemos utilizar para hacerlo lo más eficiente posible.

Podemos observar que las variables numéricas restantes se pueden dividir en diferentes categorías tal y como se muestra a continuación. Las variables de cada categoría se detallan en el Anexo A.

5.1.1. Clasificación de variables

- **Estadísticas Estándar:** Incluyen variables básicas como el número de goles y asistencias, así como el número de tarjetas amarillas y rojas. También contemplan variables categóricas como el nombre del jugador y su equipo.
- **Estadísticas de Tiro:** Evalúan la capacidad ofensiva de un jugador, lo que debería ayudar a distinguir a los delanteros del resto de las posiciones e incluso identificar a centrocampistas con inclinación ofensiva.
- **Estadísticas de Pase:** Analizan la habilidad de un jugador para mover el balón por el campo y distribuir el juego. Los centrocampistas, en particular, deberían mostrar valores distintos en aspectos como pases en profundidad o pases que resulten en gol.
- **Estadísticas de Creación de Tiros y de Goles:** Miden el talento de un jugador para crear acciones ofensivas, lo que idealmente permite distinguir a delanteros y centrocampistas ofensivos de otras posiciones en el campo.

- **Estadísticas de Defensa:** Valoran el nivel de destreza de un jugador en acciones defensivas, lo que supuestamente permite diferenciar a los defensores y centrocampistas defensivos del resto, ya que son los que más intervenciones defensivas realizan.
- **Estadísticas de Posesión:** Relacionadas con el talento de un jugador para retener el balón. Los centrocampistas y defensas centrales deberían destacar en estas estadísticas, mostrando valores altos.
- **Estadísticas de Tiempo de Juego:** Incluyen variables como el número de minutos por partido, partidos jugados, suplencias, entre otros.
- **Otras Estadísticas:** Estadísticas variadas como el número de fueras de juego o los goles en propia puerta.

Vamos a utilizar diferentes conjuntos de datos para la aplicación de diferentes técnicas de reducción de variables dentro del dataset. La primera forma que vamos a seleccionar las variables es usando todas las variables disponibles. No se realiza ninguna excepción se utilizan las 151 variables (excluyendo las variables categóricas y redundantes del dataset con 164 variables).

La segunda forma es la selección manual. Donde tras analizar todas las variables se ha utilizado un subconjunto de variables reuniendo alguna de cada grupo. La lista es la misma que en el TFG de Victor Mulero [14].

Obtenemos 2 conjuntos de datos:

- Conjunto 1: Todas las variables numéricas.
- Conjunto 2: Variables elegidas en el conocimiento del problema.

5.1.2. Reducción de dimensionalidad

En esta sección usaremos el conjunto de datos 2 para realizar un PCA con el objetivo de reducir el número de dimensiones y obtener las variables que más contribuyen al total de dimensiones de diferentes formas.

Podemos observar en la tabla 5.2 la representación de los 20 primeros autovectores con el porcentaje de inercia acumulada y la figura 5.1 que muestra el screeplot correspondiente al PCA.

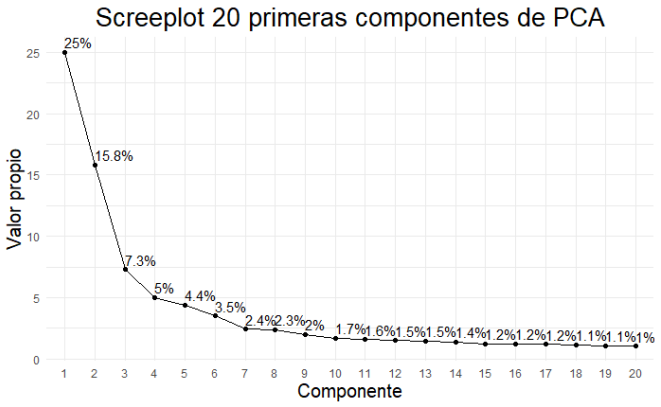


Figura 5.1: Screeplot del conjunto de datos 2

n	valor propio	% explicado
1	19.698	24.93479
2	12.325	40.53655
3	5.779	47.85185
4	3.958	52.86226
5	3.432	57.20667
6	2.741	60.67628
7	1.942	63.13410
8	1.862	65.49108
9	1.565	67.47212
10	1.347	69.17773
11	1.232	70.73766
12	1.196	72.25137
13	1.179	73.74413
14	1.086	75.11911
15	0.962	76.33698
16	0.949	77.53885
17	0.936	78.72309
18	0.910	79.87554
19	0.861	80.96488
20	0.809	81.98896

Figura 5.2: Valores propios y porcentaje explicado

Atendiendo a los diferentes criterios vistos el el marco teórico. Para explicar un variabilidad superior al 85 % se necesitan más de 20 componentes. Según la regla del codo observando el screeplot 5.1 deberíamos utilizar únicamente 3 componentes. Finalmente atendiendo a la regla de los autovectores, si observamos la 5.2 se seleccionan únicamente 14 componentes, lo que explicaría una variabilidad del 75 %. Tras este análisis se decide descartar el PCA como método reducir el número de variables.

En la figura 5.3 podemos observar las variables más contributivas a las dos primeras dimensiones. Como hemos mencionado anteriormente la falta de variabilidad que explican estas dos dimensiones no resulta interesante. Por ello se calcula la contribución al total de dimensiones de las variables. Estas nos servirán como variables iniciales para el algoritmo de selección de variables que utilizaremos en próximas secciones.

Utilizaremos como medida de importancia la contribución total basada en cargas cuadradas. Este enfoque nos es útil para medir cuánto contribuye cada variable a la variabilidad total explicada por todos los componentes principales. La lista es la siguiente:

1. Las 5 más Contributivas: `progressive_passes_received`, `sca_passes_live`, `passes_into_final_third`, `progressive_carries`, `passes_free_kicks`
2. Las 10 más Contributivas: `progressive_passes_received`, `sca_passes_live`, `carries_into_final_third`, `passes_into_final_third`, `progressive_carries`, `passes_free_kicks`, `shots_free_kicks`, `passes_pct_medium`,`passes_pct_short`, `carries_into_final_third`, `carries_into_penalty_area`

Además también calculamos las variables más contributivas de acuerdo a la suma de las magnitudes absolutas de las cargas para cada variable. A continuación se muestra la lista:

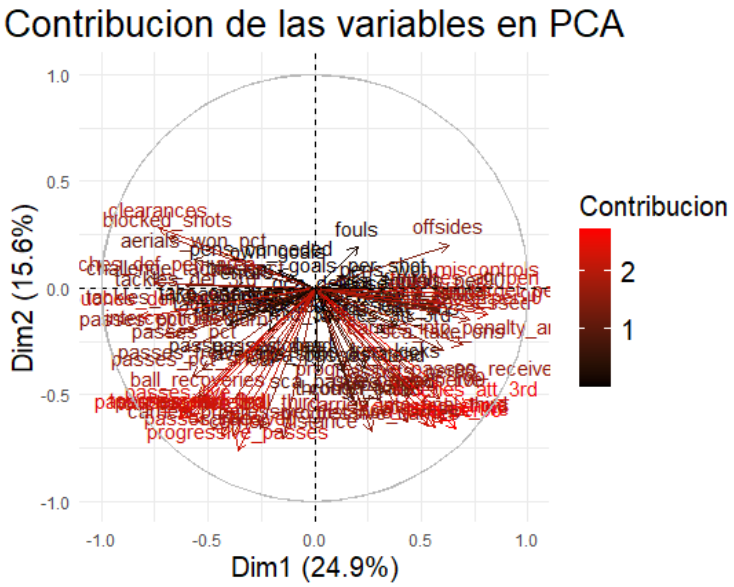


Figura 5.3: Gráfico de Cargas 2 primeras dimensiones

- Las 5 más Contributivas: `tackles_att_3rd`, `sca_defense`, `gca_defense`, `challenge_tackles_pct`
- Las 10 más Contributivas: `tackles_att_3rd`, `sca_defense`, `gca_defense`, `challenge_tackles_pct`, `corner_kicks`, `plus_minus_wowy`, `carries`, `own_goals` , `gca_per90`, `passes_free_kicks`

5.1.3. Selección de variables iniciales mediante Random Forest

Nuestro objetivo es identificar, a través del algoritmo de selección de variables Stepwise, aquellas variables que minimicen significativamente los errores de clasificación utilizando los algoritmos K-means y Ward. Aplicaremos la técnica de clasificación supervisada random forest a los Conjuntos de Datos 1 y 2 previamente mencionados, con el fin de determinar las 5 y 10 variables más importantes para el modelo. Estas variables seleccionadas se utilizarán como variables iniciales en el algoritmo Stepwise.

Hemos seleccionado random forest sobre otras técnicas de aprendizaje supervisado por 2 razones principales:

- Robustez y Precisión: Al ser es un algoritmo de aprendizaje supervisado robusto, combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste.
- Importancia de Variables: Una de las características más útiles de Random Forest es su capacidad para evaluar la importancia de las variables. Lo logra evaluando como la permutación de cada variable afecta la precisión del modelo.

Al utilizar Random Forest, se pueden identificar las variables más relevantes según su importancia al modelo. Estas variables seleccionadas pretenden reducir el error de clasificación y proporcionar un conjunto inicial más eficiente para el algoritmo Stepwise.

Se ha utilizado un Random Forest con 50 árboles, *maxbins* por defecto, donde los árboles crecen hasta el tamaño máximo y una profundidad máxima por defecto.

En la tabla 5.1 podemos ver las 10 variables más importantes para el modelo para predecir las posiciones del conjunto con todas las variables numéricas (Conjunto 1). Observamos que de entre las 10 variables más importantes para el random forest no se encuentran en el conjunto de datos con variables seleccionadas atendiendo al conocimiento del problema (Conjunto 2).

Variable	**Importancia**
touches_def_3rd	525.260661
throw_ins	423.026445
touches_def_pen_area	325.867982
clearances	309.690708
progressive_passes_received	267.607411
passes_dead	189.810802
shots	184.900320
miscontrols	182.112800
ball_recoveries	169.366520
passes_completed_medium	169.040291

Cuadro 5.1: Importancia de las variables en el modelo (Conjunto 1)

Repetimos el proceso para el conjunto de datos 2. En la tabla 5.2 Obtenemos las 10 variables con mayor importancia para el modelo de este conjunto de datos.

Variable	**Importancia**
clearances	691.394304
throw_ins	469.357224
touches_def_3rd	347.943399
touches_att_pen_area	323.308345
passes_dead	293.752655
shots_per90	289.666024
miscontrols	288.131663
touches_def_pen_area	269.998722
touches_att_3rd	232.223369
dispossessed	217.238101

Cuadro 5.2: Importancia de las variables en el modelo (Conjunto 2)

5.1.4. Algoritmo Stepwise

Observando los resultados obtenidos en el PCA, vamos a realizar un algoritmo para la selección de variables más influyentes a la hora de clasificar los jugadores en los 5 perfiles especificados anteriormente.

Se ha utilizado un método de selección de variables paso a paso como se ha explicado en el marco teórico. Este método consta de dos fases muy definidas. La primera sería la selección de variables que cuenta con agregación de variables (paso hacia delante) y la segunda sería la eliminación de variables (paso hacia atrás). Inicialmente se parte de un conjunto de variables candidatas. Estas variables van a ser las que hemos ido obteniendo en los apartados anteriores por diferentes métodos.

El objetivo principal del método es obtener obtener el subconjunto de variables que minimice el error de clasificación cometido. En un primer lugar vamos a utilizar el método de las K-medias para hacer todos las cálculos de error, ya que en el trabajo de Victor Mulero [14] vimos que los errores de clasificación obtenidos usando el método de Ward eran muy superiores. Si añadimos esto al mayor coste computacional del método al ser un clustering jerárquico, no merece la pena aplicarlos a todos los casos que realizaremos. Se utilizara el método de WARD con el caso que menor error de clasificación obtengamos con el método de las K-medias.

Para calcular el error de clasificación, se crean cinco clústers y cada uno se asigna a una posición del campo en función de la posición mayoritaria de los jugadores en ese grupo. El error se calcula dividiendo la cantidad de jugadores asignados a un clúster cuya posición no es la que se le asignó.

Podemos ver el psudocódigo del algoritmo Stepwise con el cálculo del error con K-medias a continuación.

Algorithm 1 ALGORITMO STEPWISE

Función: stepwise_variable_selection

Require: real_positions, cluster_data, n_start=10,initial_variables=NULL
Ensure: best_variables

print('Algoritmo para seleccionar las mejores variables')
variables ← **colnames**(cluster_data)
if is.null(initial_variables) **then**
  PCA ← **prcomp**(cluster_data)
  initial_variables ← variables[**head**(**order**(**rowSums**(**get_pca**(PCA)$cos2[, 1 : 2]$), **decreasing=TRUE**), n_start)]
end if
print('Variables iniciales:')
print(initial_variables)
actual_data ← cluster_data[, initial_variables]
initial_error ← **calculate_error**(real_positions, actual_data)
best_variables ← {error=initial_error, variables=initial_variables}
aux ← c()
while !identical(best_variables$variables$, aux) **do**
  aux ← best_variables$variables$
  best_variables ← **forward_selection**(real_positions, cluster_data, variables, best_variables)
  best_variables ← **backward_selection**(real_positions, cluster_data, variables, best_variables)
end while
return best_variables

Algorithm 2 SELECCIÓN HACIA DELANTE

Función: forward_selection

Require: real_positions, cluster_data, variables, best_variables
Ensure: updated_best_variables

```
print('SELECCIÓN HACIA DELANTE')
selected_variables ← best_variables$variables
initial_error ← best_variables$error
not_selected ← setdiff(variables, selected_variables)
best_variables ← {error=1, variable=NULL}
for variable in not_selected do
  actual_data ← cluster_data[, c(selected_variables, variable)]
  error ← calculate_error(real_positions, actual_data)
  if error <best_variables$error then
    best_variables ← {error=error, variable=variable}
  end if
end for
if best_variables['error'] <initial_error then
  print(paste('Entra ', best_variables['variable'], ' - error ',
  round(best_variables['error'], 3), sep=''))
  initial_error ← best_variables['error']
  selected_variables ← c(selected_variables, best_variables['variable'])
end if
return {error=initial_error, variables=selected_variables}
```

Algorithm 3 SELECCIÓN HACIA ATRÁS

Función: backward_selection

Require: real_positions, cluster_data, variables, best_variables
Ensure: updated_best_variables

```
print('SELECCIÓN HACIA ATRÁS')
selected_variables ← best_variables$variables
initial_error ← best_variables$error
best_variables ← {error=1, variable=NULL}
for variable in selected_variables do
  actual_data ← cluster_data[, setdiff(selected_variables, variable)]
  error ← calculate_error(real_positions, actual_data)
  if error <best_variables$error then
    best_variables ← {error=error, variable=variable}
  end if
end for
if best_variables$error <initial_error then
  print(paste('Sale ', best_variables$variable, ' - error ', round(best_variables$error,
  3), sep=''))
  initial_error ← best_variables$error
  selected_variables ← setdiff(selected_variables, best_variables$variable)
end if
return {error=initial_error, variables=selected_variables}
```

Conjuntos de variables candidatos

Se ha aplicado este algoritmo sobre diferentes condiciones para los dos conjuntos de datos creados. Se han utilizado como variables iniciales aquellas que hemos ido especificando en los apartados anteriores. Cabe recalcar que lo que buscamos es la clasificación en 5 clústers diferentes, donde se pretende que cada clúster se asocie a cada una de las 5 posiciones. Las tres posiciones principales y las posiciones intermedias DF-MF,MF-FW.

A continuación se muestran las tablas 5.3 que muestran los resultados obtenidos para el conjunto de datos 1 y 2 con variables iniciales como la Suma de Magnitudes Absolutas de las Cargas. La tabla 5.4 que muestra la aplicación del algoritmo con las variables iniciales como Contribución Total Basada en Cargas Cuadradas y finalmente la tabla 5.5 que muestra la aplicación del algoritmo StepWise con las variables iniciales como la importancia al modelo a través de un Random Forest. Todas estas tablas tratan de minimizar el error mediante el método de las K-medias. Atendiendo a los siguientes resultados decidiremos qué caso usar para la aplicación del método de WARD.

CASO	**Método**	**Conjunto Inicial**	**Variables Iniciales**	**Conjunto Final**	**Error**
1	K-Medias	Conjunto 1	10 más contributivas al PCA	25 variables	0.3103922
2	K-Medias	Conjunto 1	5 más contributivas al PCA	22 variables	0.2530
3	K-Medias	Conjunto 2	10 más contributivas al PCA	20 variables	0.3241386
4	K-Medias	Conjunto 2	5 más contributivas al PCA	12 variables	0.2919721

Cuadro 5.3: Resultado algoritmo stepwise de selección de variables con Kmedias y variables iniciales como contribución total basada en cargas cuadradas

A priori se observa que tanto el conjunto de datos utilizado, el método utilizado como selección de variables iniciales y el número de variables iniciales afectan significativamente en el número de variables en el conjunto final y el error de clasificación resultante.

- Podemos observar que la media de error entre ambos conjuntos es menor en el conjunto de datos 1. Esto se puede deber a que algunas variables importantes para la minimización del error no se han tenido en cuenta en el conjunto de datos seleccionadas atendiendo el conocimiento del problema.
- Observamos que en general la utilización de menos variables iniciales nos proporciona un menor error de clasificación con respecto a la utilización de 10 variables iniciales.

CASO	Método	Conjunto Inicial	Variables Iniciales	Conjunto Final	Error
5	K-Medias	Conjunto 1	10 más contributivas al PCA	21 variables	0.256415
6	K-Medias	Conjunto 1	5 más contributivas al PCA	22 variables	0.2565066
7	K-Medias	Conjunto 2	10 más contributivas al PCA	18 variables	0.2741019
8	K-Medias	Conjunto 2	5 más contributivas al PCA	10 variables	0.4419905

Cuadro 5.4: Resultado algoritmo stepwise de selección de variables con Kmedias y variables iniciales como suma de magnitudes absolutas de las cargas

CASO	Método	Conjunto Inicial	Variables Iniciales	Conjunto Final	Error
9	K-Medias	Conjunto 1	10 con Mayor Importancia	18 variables	0.3154326
10	K-Medias	Conjunto 1	5 con Mayor Importancia	10 variables	0.327621
11	K-Medias	Conjunto 2	10 con Mayor Importancia	16 variables	0.3279875
12	K-Medias	Conjunto 2	5 con Mayor Importancia	12 variables	0.333761

Cuadro 5.5: Resultado algoritmo stepwise de selección de Variables con Kmedias y variables iniciales como importancia en un random forest

- Para el Conjunto 1, empezar con más variables iniciales (10 en lugar de 5) obtenemos mejores resultados en términos de error.
- Vemos que mantener un conjunto final de variables relativamente amplio tiende a producir mejores resultados, aunque el número exacto óptimo puede variar según el conjunto de datos.
- Si comparamos las trazas para todas las ejecuciones observamos que las iteraciones que realizan los casos con variables iniciales como las más importantes a un *random forest* son mucho menores que el resto. Esto se debe principalmente que para estos casos el algoritmo no realiza ninguna expulsión de variables, indicando así la relevancia de estas variables a la minimización del error de clasificación.

A continuación, se muestra una tabla con los resultados obtenidos de aplicar el método de Ward con las variables finales para el Caso 2.

CASO	Método	Conjunto Inicial	Variables Iniciales	Conjunto Final	Error
13	Ward	Conjunto 1	22 Variables Caso 2	23 variables	0.3583211

Cuadro 5.6: Resultado algoritmo stepwise de selección de variables con el método de Ward

Vemos que el error producido por el método de Ward es un error bastante alto en comparación a los casos donde se ha aplicado el método de las K-medias.

Observamos que tanto el conjunto inicial utilizado como la selección de variables iniciales son muy importantes a la hora de la reducción del error.

Comparativa de variables:

Vamos a seleccionar el caso 2 que es el que tiene un menor error de clasificación para las K-medias 0,2530242, el caso número 4 que tiene un mayor error de clasificación pero cuenta únicamente 12 variables seleccionadas para el conjunto final. Se selecciona también el Caso 13, obtenido con las variables finales del Caso 2 obtenido con el método de las k-medias.

Este análisis permite comparar no solo la cantidad y tipo de variables seleccionadas, sino también cómo estas variables contribuyen al rendimiento del modelo, medido por el error obtenido.

A continuación, se presenta la tabla 5.7 que detalla las variables únicas seleccionadas por el algoritmo en ambos casos, marcando con una "X"las variables presentes en cada caso.

En el análisis de las variables seleccionadas por el algoritmo en los casos 2, 4 y 13, se observa una distribución que cubre la mayoría categorías estadísticas. En el Caso 2 y 13, las variables seleccionadas cubren un amplio espectro de estadísticas estándar, estadísticas de tiro, pase, creación de tiros y goles, defensa, posesión y tiempo de juego. Esto indica que el algoritmo está capturando las diferentes habilidades de los jugadores, lo que da pie una mejor división en las cinco posiciones de campo. Variables como *own_goals*, *minutes_per_game*, y *unused_subs* destacan la inclusión de estadísticas estándar y tiempo de juego, mientras que variables como *shots_per90*, *npxg*, y *passes_into_final_third* reflejan capacidades ofensivas y de pase.

Por otro lado, el Caso 4 presenta una selección más concentrada de variables, principalmente en las categorías de defensa y posesión. Variables como *tackles_mid_3rd*, *clearances*, y *blocks* indican un enfoque claro en el desempeño defensivo, mientras que *dispossessed*, *carries_into_penalty_area*, y *touches* reflejan aspectos de posesión. Esta concentración sugiere que el algoritmo en el Caso 4 prioriza características específicas que pueden ser imprescindibles para ciertos perfiles como defensores y centrocampistas defensivos.

Los Casos 2 y 13 ofrecen una visión más global del desempeño del jugador, abarcando múltiples aspectos del juego, mientras que el Caso 4 se enfoca en áreas específicas, pro-

Cuadro 5.7: Comparación de variables seleccionadas entre Caso 2, Caso 4 y Caso 13

Variable	**Caso 2**	**Caso 4**	**Caso 13**
passes_into_final_third	X	X	X
own_goals	X	X	X
challenges_lost	X		X
touches_def_3rd	X		X
shots_per90	X	X	X
aerials_lost	X		X
passes_pct_long	X		X
dispossessed	X	X	X
tackles_mid_3rd	X	X	X
goals_pens	X		X
carries_progressive_distance	X		X
minutes_per_game	X		X
unused_subs	X		X
challenge_tackles_pct	X	X	X
offsides	X	X	X
clearances	X	X	X
passes_dead	X		X
gca_defense	X		X
fouled	X		X
npxg	X		X
corner_kicks_straight	X		X
npxg_per_shot	X		X
carries_into_penalty_area		X	
blocks		X	
touches		X	
plus_minus_wowy		X	
goals_per_shot			X

porcionando una evaluación más especializada.

5.2. Clústering

En el apartado anterior hemos conseguido 3 conjuntos de datos con el menor error más reducido de variables candidatas. Vamos a crear tablas de contingencia para las dobles posiciones para calcular el error de jugadores mal asignados y poder proporcionar una buena medida de error.

5.2.1. Ward

En esta sección utilizáramos los conjuntos de variables obtenidos en el apartado anterior, el caso 13. Utilizando el conjunto de datos asociado al Caso 13 obtenemos el

dendograma cortado de forma que se obtienen 5 clústers 5.4.

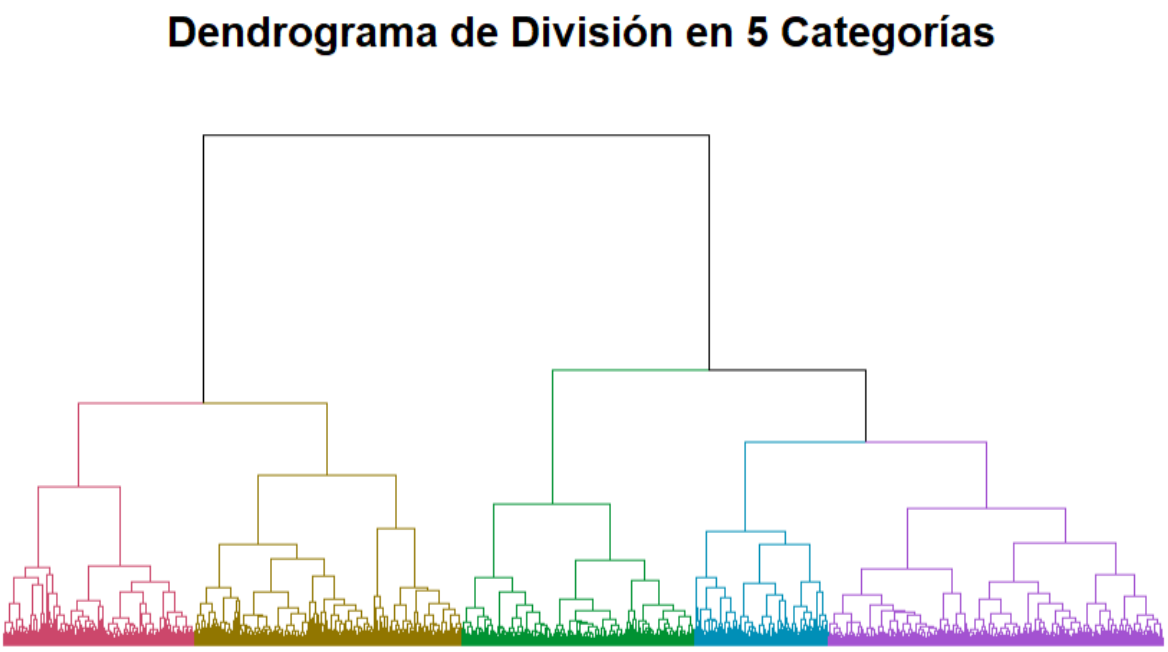


Figura 5.4: Dendograma 5 clústers

En las tabla de contingencia 5.8 se cruzan los cinco clústeres creados y las posiciones reales de los jugadores. Además en la tabla 5.9 se muestra los errores de clasificación por clúster. Se observa que los clústers suelen tener un error de clasificación bastante alto salvo para el clúster 5 que asocia con los DF con un error al inferior 10 %. El clúster que mayor error comete es el número 2, asociado a centrocampistas defensivos con un error superior al 50 %.

Clústers	**DF**	**DF,MF**	**MF**	**MF,FW**	**FW**
5	1995	129	70	1	0
2	1424	359	1237	104	19
4	223	79	761	175	18
3	93	106	427	1575	317
1	2	6	16	529	1247
Error de clasificación: 0.3583					

Cuadro 5.8: Valores reales vs Clústers

Clústers	**Error**
5	0.0911
2	0.5469
4	0.3941
3	0.3745
1	0.3072

Cuadro 5.9: Errores de clasificación por clúster

Si realizamos el test de independencia para la tabla 5.8 el pvalor asociado es prácticamente 0. Se decide realizar un análisis de correspondencias para analizar las asociaciones ente los clústers creados y las posiciones reales. En la figura 5.5 se muestra el SCA de la tabla. Se puede observar una asociación entre los clústers y las posiciones reales. El *clúster 1* se asocia de forma muy directa con los delanteros (FW), el *clúster 3* con los centrocampistas ofensivos (MF,FW). Por otro lado el *clúster 4* presenta una asociación con los mediocentros (MF), el *clúster 2* con los centrocampistas defensivos (DF,MF)y el *clúster 5* una asociación algo más débil con los defensas centrales (DF).

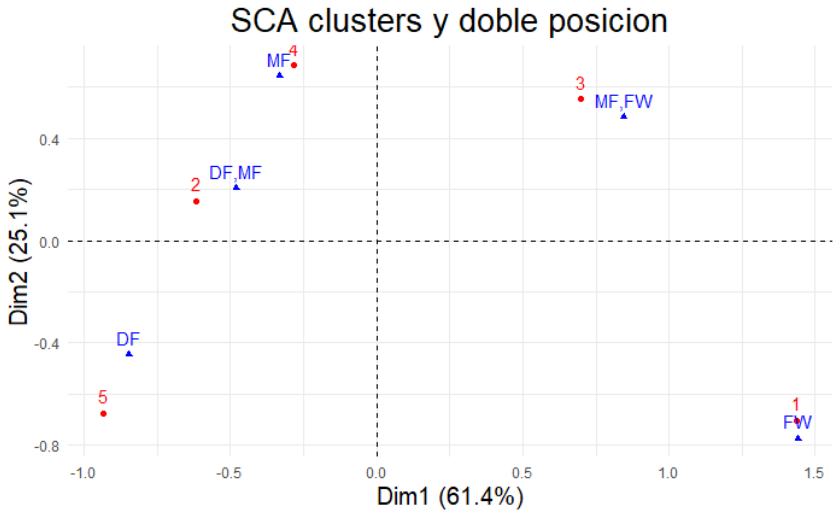


Figura 5.5: Biplot 5 Clústers Método de Ward

5.2.2. K-medias

En esta sección utilizáramos los conjuntos de variables obtenidos en el apartado anterior, el Caso 2 y el Caso 4. Se han realizado todos los cálculos para ambos casos pero por tema de espacio mostraremos únicamente la ejecución para el caso 2.

Debido a la intencionalidad de seleccionar el mismo número de clústers como posiciones, con el objetivo de asociar cada clúster una posición concreta se va a omitir la selección del número de clústers.

Creación de 5 clústers

En la tabla 5.10 se puede apreciar la clasificación realizada por el método de las K-medias con 5 clústers. Obtenemos un error de clasificación de 0.2122. Este error de clasificación se obtiene mediante la suma de los máximos valores de cada fila divididos entre el total. De esta forma no se realiza una asociación 1 a 1 clúster perfil. Si decidimos asociar un clúster a cada posición el error global cambia. Se calcula como la suma de elementos en la diagonal de la tabla 5.10 divido entre el total. Dando como resultado un valor de 0,3082.

clústers	DF	DF,MF	MF	MF,FW	FW
1	3111	330	41	4	7
2	491	45	33	2	3
3	81	209	2012	208	7
5	54	94	420	1707	308
4	0	1	5	463	1276

Error de clasificación: 0.2122
P-valor del test chi-cuadrado: 0

Cuadro 5.10: Caso 2: Tabla de contingencia 5 clústers y 5 posiciones

Observando la tabla vemos que no tenemos ningún clúster que se asocia al perfil DF-MF. Si calculamos los errores individuales de cada clúster (clúster 2 asociado a DF) obtenemos la tabla 5.11.

clúster	Error de Clasificación
1	0.1094
2	0.1446
3	0.2006
4	0.2688
5	0.3391

Cuadro 5.11: Errores de clasificación por clúster

De los datos presentados en la tabla 5.11, se concluye que los clústers que más errores producen son los clústers 4 y 5. Específicamente, el clúster 5 presenta el mayor error de clasificación con un valor de 0,3391, seguido por el clúster 4 con un error de 0,2688.

El clúster 5 se caracteriza por clasificar predominantemente a medio centros ofensivos. Sin embargo, clasifica un número considerable de delanteros y medio centros. Lo que da respuesta a ese error tan elevado.

El clúster 4, se caracteriza por la clasificación de delanteros. Aunque dentro de este clúster se encuentran un 30 % de medio centros ofensivos. Lo que indicaría ese error tan elevado.

A continuación realizamos un análisis de correspondencias para comprender mejor la agrupación y las posibles relaciones entre los clústers y las categorías de jugadores.

Analizando el biplot 5.6 para las dobles posiciones vemos que existen correspondencias entre los clústers y los perfiles. Los clústers 1 y 2 se asocian a defensas, el clúster 3 a medio centros, el clúster 5 a medio centros ofensivos y por último el clúster 4 a delanteros.

Es importante destacar que no se encontró ningún clúster que mantenga una relación de correspondencia clara con los centrocampistas defensivos (DF-MF). Esta ausencia sugiere que las variables usadas en el caso 2 no están capturando adecuadamente las características clave de este perfil.

A la vista de estos resultados vamos a experimentar con k=4 y k=6 para ver si podemos encontrar un clúster que se asocie mejor al perfil DF,MF.

Creación de 4 clústers

Se repite el mismo proceso que para la creacción de 5 clústers. En la tabla 5.12 podemos observar el clústering con k-medias con 4 clústers. Con un error de clasificación de 0,3093.

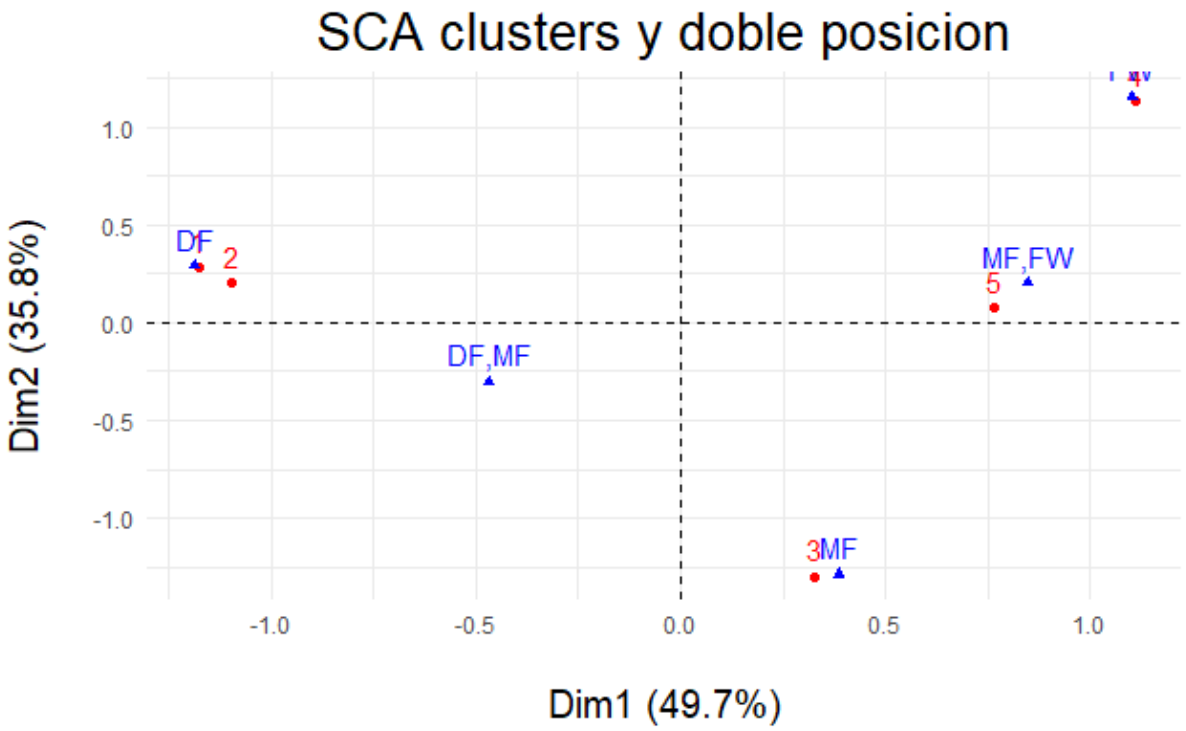


Figura 5.6: Caso 2: Biplot dos dimensiones para 5 clústers y 5 perfiles

clústers	DF	DF,MF	MF	MF,FW	FW
4	3525	337	47	4	6
1	146	241	2017	198	6
3	66	100	442	1718	312
2	0	1	5	464	1277

Error de clasificación: 0.3093
P-valor del test chi-cuadrado: 0

Cuadro 5.12: Caso 2: Tabla de contingencia 4 clústers y 5 posiciones

Realizando un análisis de correspondencias los 4 clústers creados obtenemos la representación 5.7. Podemos observar que se no se producen desplazamiento de clústers con respecto a la figura 5.6. Lo que se produce es la eliminación del clúster asociado a la posición de centrocampistas atacantes (MF-FW). Cabe destacar que no se ha eliminado ninguno de los clústers asociado al perfil de DF, esto es lo que buscábamos al realizar la agrupación con 4 clústers.

Creación de 6 clústers

Podemos observar en la figura 5.8 el biplot con 6 clústers y 5 perfiles. Vemos que se crea un clúster más en la posición defensiva manteniendo aquellos asociados a las otras posiciones. Este nuevo clúster se sitúa entre medias de los perfiles DF y DF-MF. Lo que nos lleva a pensar que este clúster tendrá estadísticas asociadas tanto a defensores como a mediocentros defensivos.

Realizando un PCA podemos visualizar también los proyecciones de las variables junto

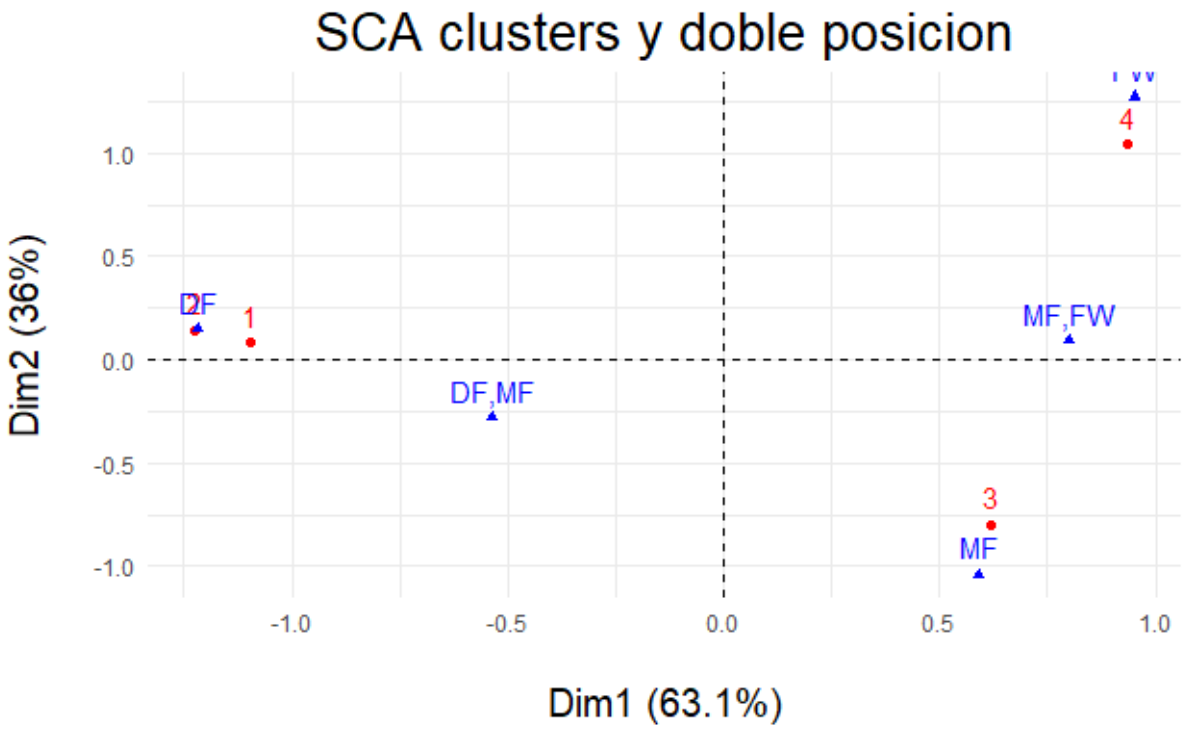


Figura 5.7: Caso 2: Biplot dos dimensiones para 4 clústers y 5 perfiles

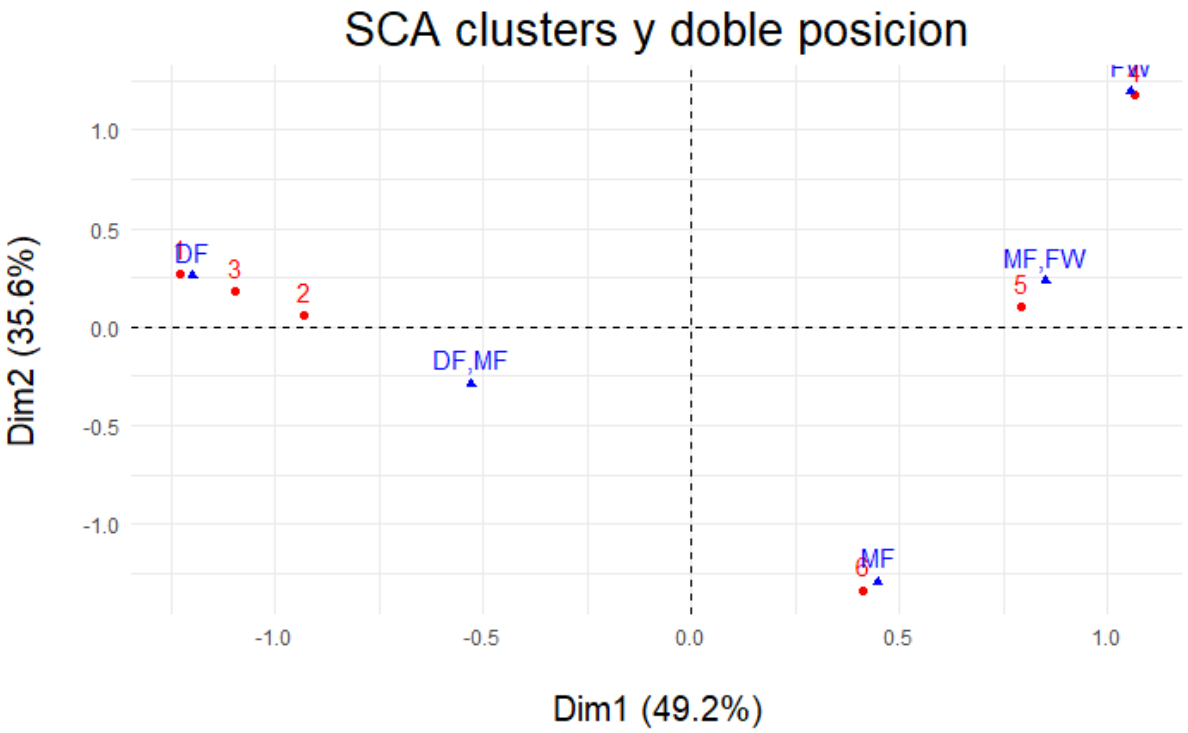


Figura 5.8: Caso 2: Biplot dos dimensiones para 6 clústers y 5 perfiles

con los centroides de los 5 clústers. De forma que podemos observar las relaciones creadas entre los diferentes clústers y variables. Vamos a describir las variables relacionadas con cada clúster de acuerdo a la figura.

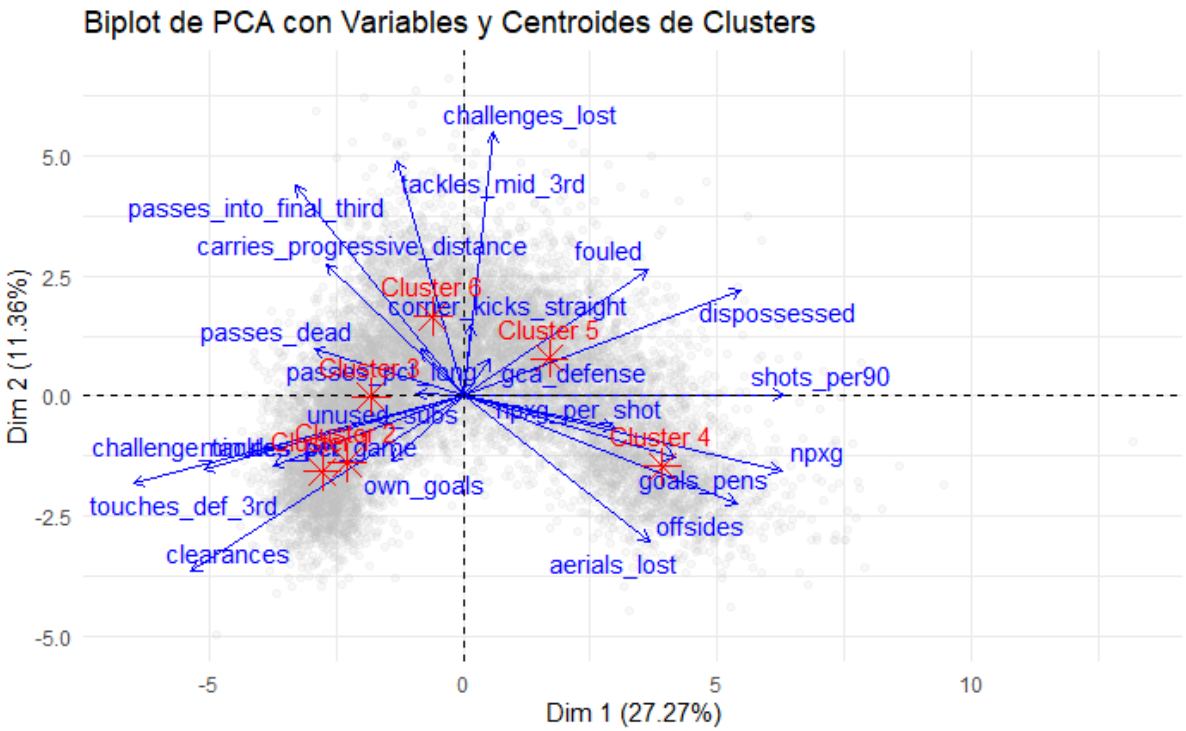


Figura 5.9: Biplot Centroides y Variables

- **Clúster 1**: Para este clúster encontramos variables como toques en el tercio del propio campo (touches_def_3rd), despejes (clearances), Porcentaje de entradas realizadas (challenge_tackles_pct).Estas variables muy asociadas a perfiles defensivos especialmente en el área más cercana a su propia portería (DFC).

- **Clúster 2**: El Clúster 2 también muestra una asociación significativa con variables defensivas. Las variables encontradas en este clúster complementan las observadas en el Clúster 1, reforzando la idea de que ambos clústers representan perfiles defensivos.

- **Clúster 3**: En este clúster encontramos variables como pases a el ultimo tercio del campo (passes_into_final_third), pases muertos, entradas en el último tercio del campo. Variables muy relacionada con medio centros puros- (MC), reflejando así su rol en la distribución del balón y transición del juego desde la defensa hacia el ataque.

- **Clúster 4**: Este clúster está asociado con variables como fueras de juego (offsides), número de goles esperados (npxG), goles de penaltis (goals pens) y tiros cada 90 minutos (shots per90). Estas métricas son propias de delanteros centro y extremos, indicando que los jugadores en el Clúster 4 son predominantemente ofensivos, centrados en la finalización de jugadas y en la generación constante de oportunidades de gol.

- **Clúster 5** En este clúster, encontramos variables como balones perdidos (dispossessed) y faltas recibidas (fouled). Estas variables son características de un mediocentro ofensivo, quienes frecuentemente se encuentran en situaciones de uno contra uno, perdiendo y recuperando el balón, y sufriendo faltas debido a su papel creativo

en el campo. Los jugadores en el Clúster 5 tienen una función más avanzada en el mediocampo, contribuyendo tanto en la creación de juego como en situaciones ofensivas.

Clústers individuales

A raíz de la interpretación de los clústers, procederemos a representarlos individualmente y a identificar los jugadores que pertenecen a cada uno de ellos. Para lograr esto, seleccionaremos aleatoriamente entre 10 y 15 jugadores de cada clúster y mencionaremos sus nombres en el diagrama correspondiente a cada clúster individual. La información sobre estos jugadores será obtenida de la base de datos de Transfermarkt [23].

Clúster 1: Defensas

En la figura 5.10 podemos encontrar jugadores en su mayoría de primera linea de defensa, como son (DFC) y (LD, LI). Podemos destacar nombres como Ashley Young, José Holebas.

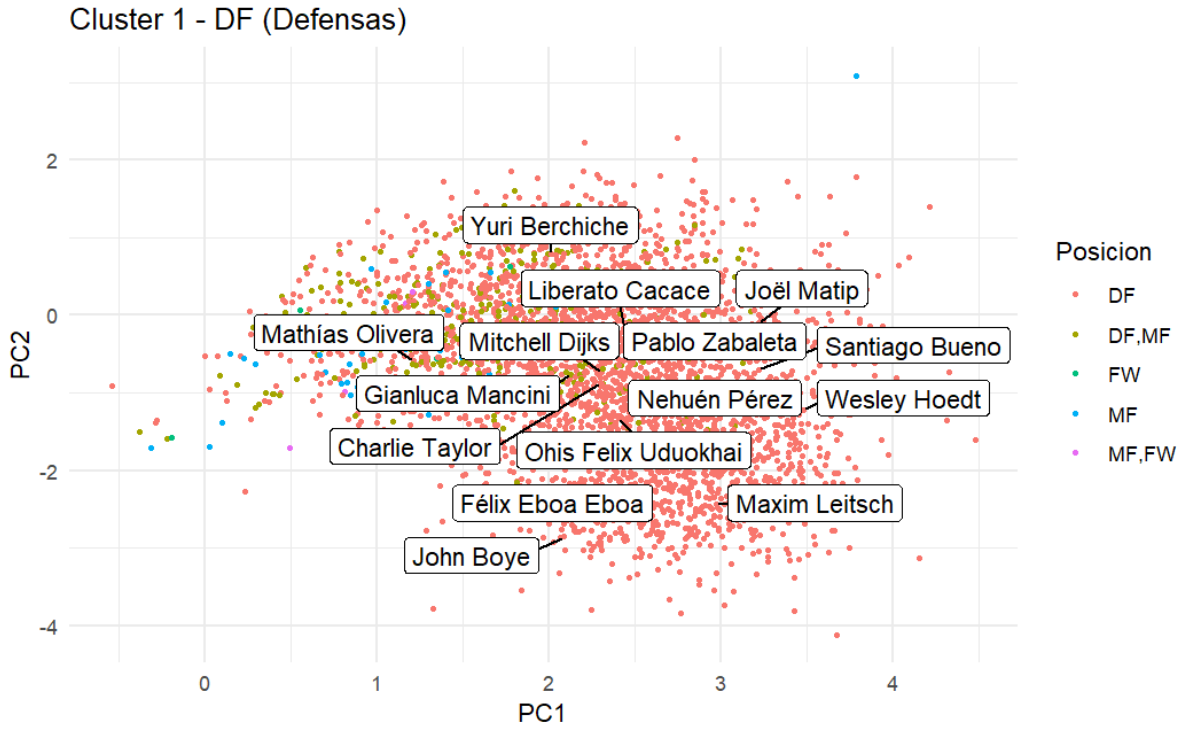


Figura 5.10: Clúster 1: Defensas Centrales

Clúster 2: Defensas

En la figura 5.11 podemos encontrar jugadores similares a los de la figura 5.10. En este cluúter podemos encontrar nombres como Jorge Pulido (DFC) o Marc Guei (DFC). Aunque también podemos encontrar algún MF, como Mikel San José.

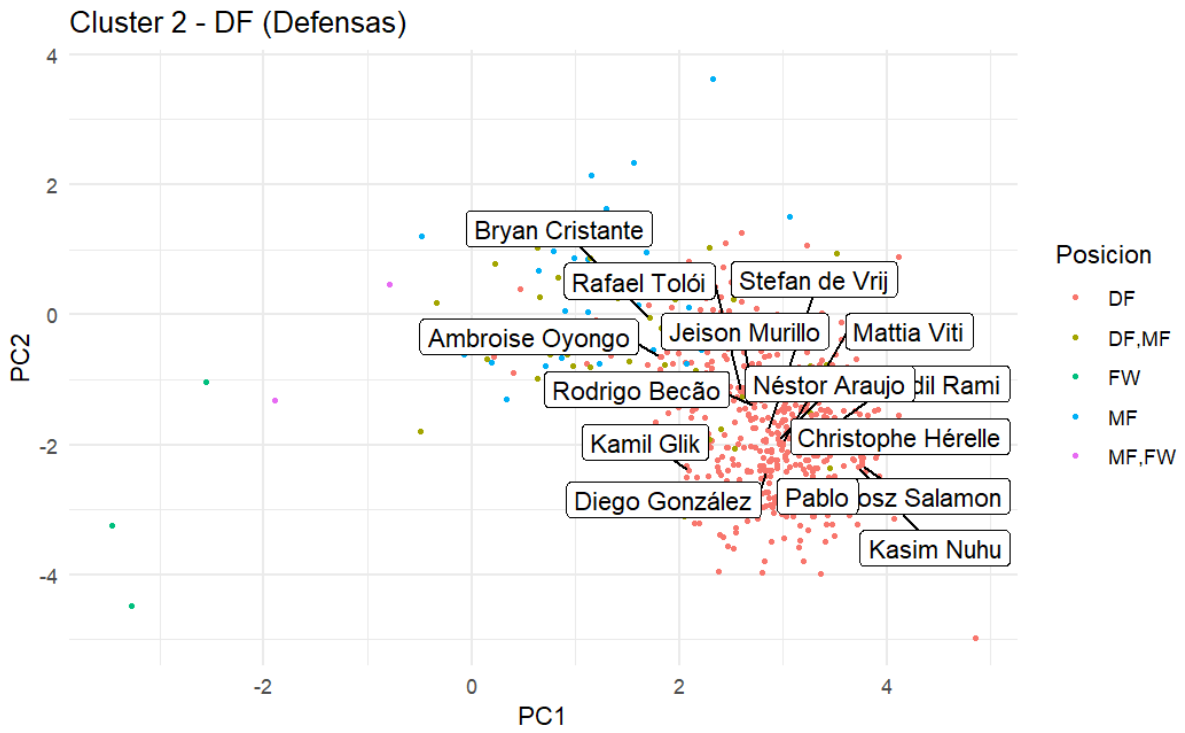


Figura 5.11: Clúster 2: Defensas Centrales

Clúster 3: Medio Centros

En la figura 5.12 podemos encontrar lo que se conoce centrocampistas puros. Aquí encontramos nombres como Rodri, Sergio Busquets, David Silva,quienes, de acuerdo con los datos obtenidos de Transfermarkt [23], están clasificados como medio centros defensivos y medio centros ofensivos, respectivamente. Sin embargo, en nuestro conjunto de datos, todos estos jugadores están etiquetados de manera uniforme como MF (centrocampistas).

Clúster 4: Delanteros Centro

En la figura 5.13 podemos encontrar el perfil de DC puro con algunos extremos. Destacan los nombres de Andrea Belotti, Diego Costa y Antoine Griezman. Estos son las clara representación de delanteros centro puros.

Clúster 5: Medio Centros Ofensivos

En la figura 5.14 se representan los centrocampistas ofensivos, categorizados en el *clúster 5*. Este clúster incluye jugadores que desempeñan roles tanto en el medio-campo como en el ataque, combinando habilidades creativas y ofensivas. Entre los nombres destacados se encuentran Kingsley Coman, Bukayo Saka y Gelson Martins.

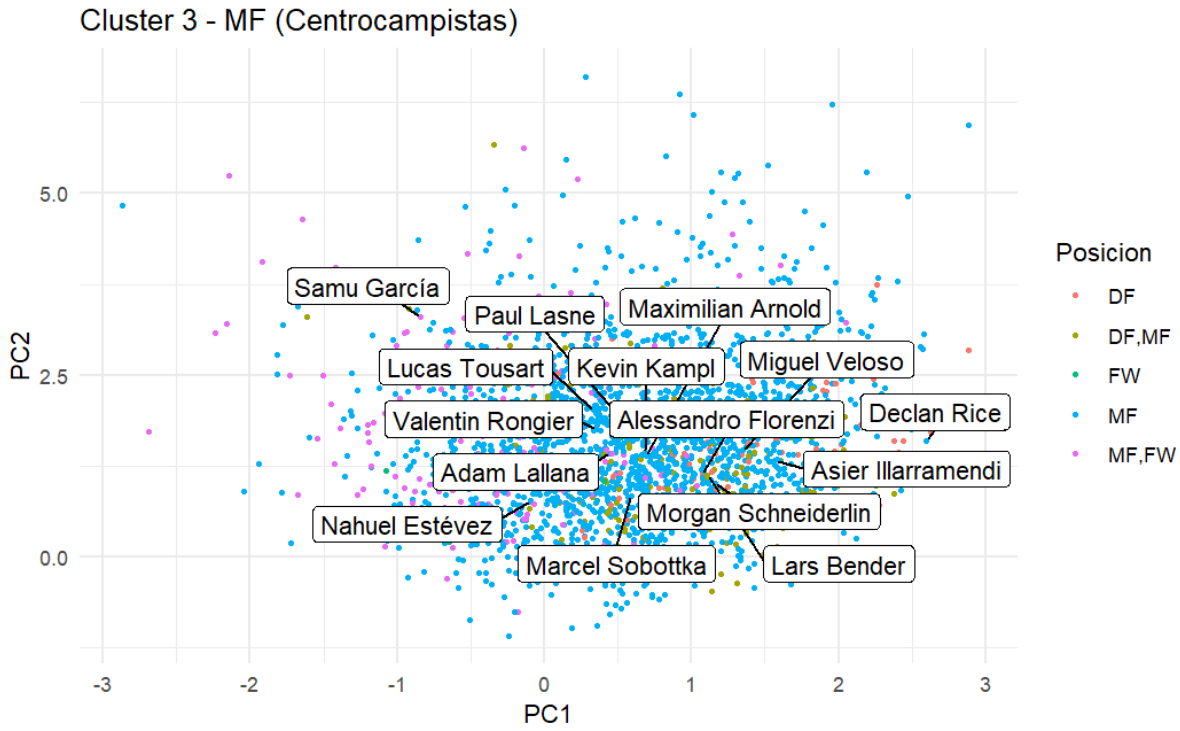


Figura 5.12: Clúster 3: Medio centros

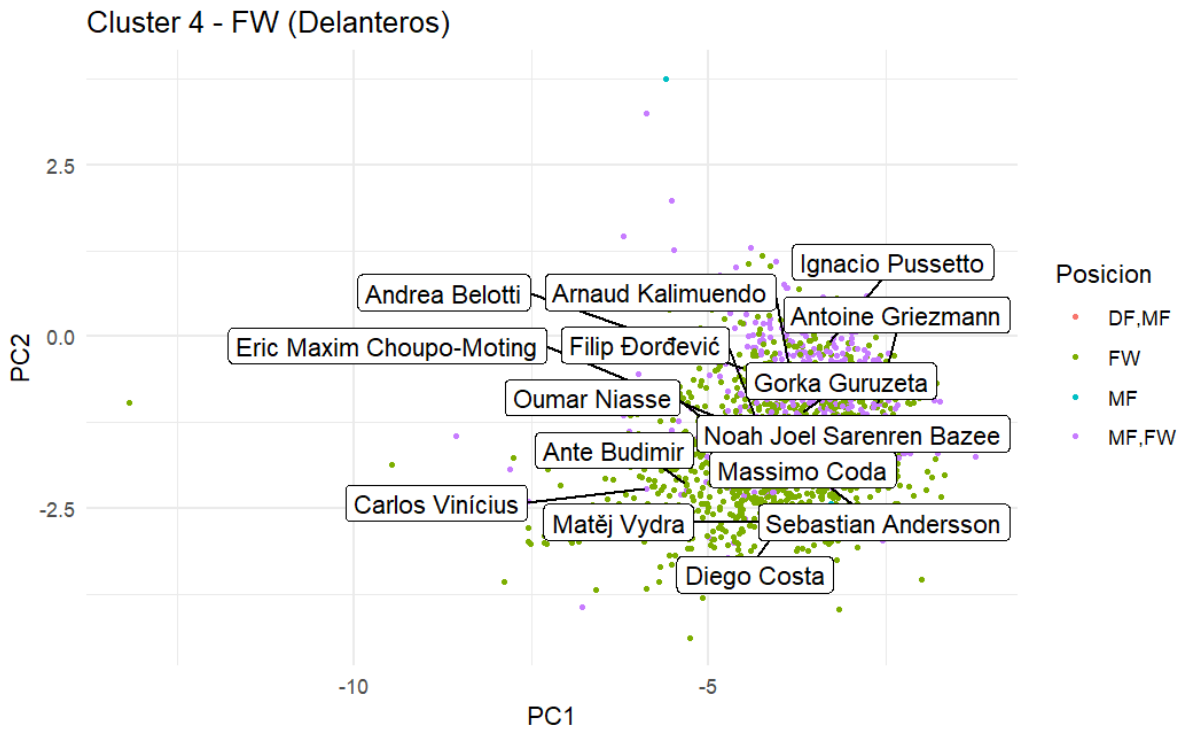


Figura 5.13: Clúster 4: Delanteros Centro

5.2.3. Comparación con las ligas

A partir de de los resultados obtenidos con ambos métodos vemos que la agrupación es muy similar. Veamos ahora si existe alguna relación entre los clústers y las ligas que

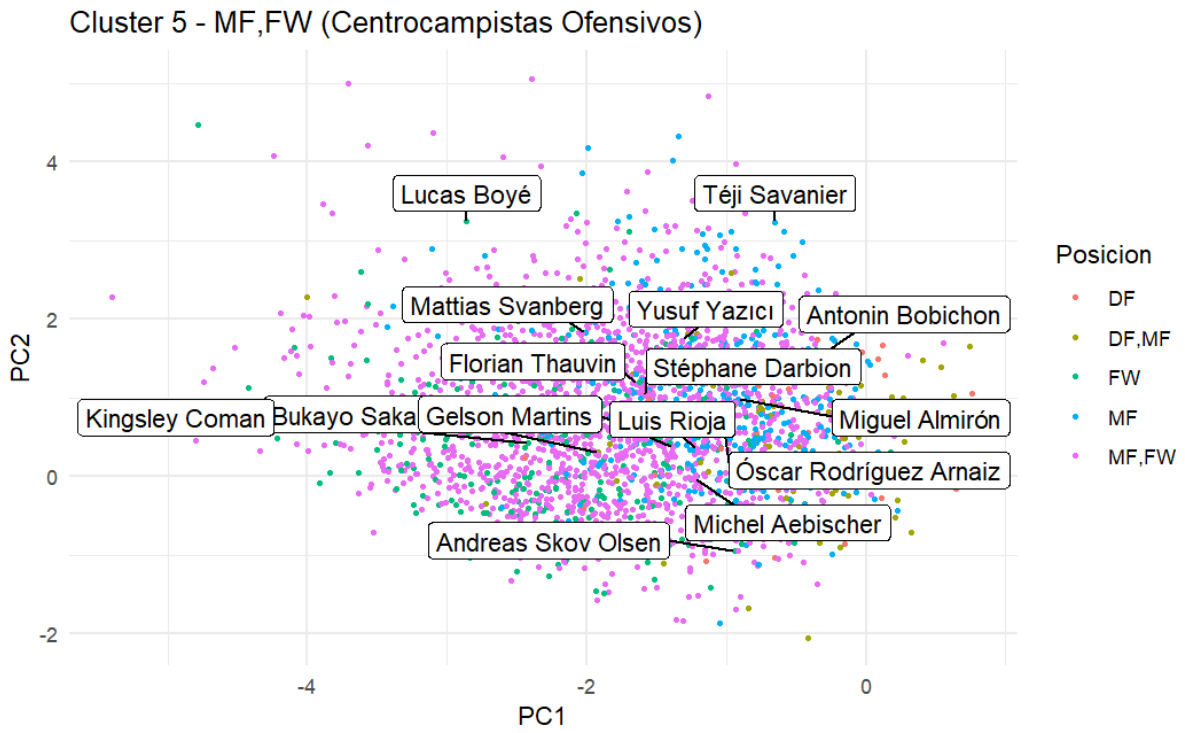


Figura 5.14: Clúster 5: Medio Centros Ofensivos

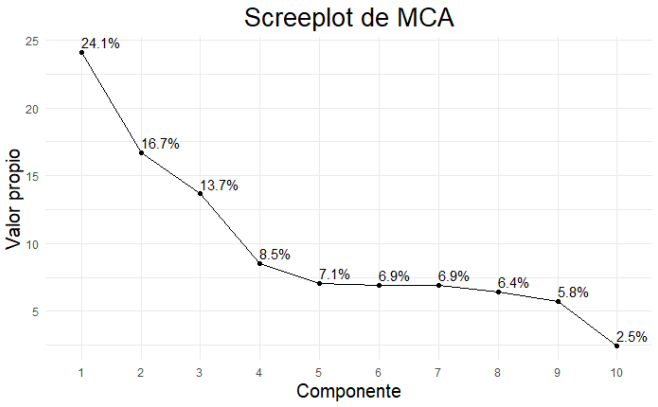
componen el estudio. Realizaremos este análisis con 5 clústers tanto para el método de Ward como el K-medias. Por similitud de resultados solo se incluye el análisis de correspondencias múltiple para el método de Ward

A continuación se muestra la tabla de contingencia asociada los 5 clústers obtenidos y las 5 ligas 5.13.

	Bundesliga	**LaLiga**	**Ligue 1**	**Premier League**	**Serie A**
1	386	387	335	349	343
2	549	693	658	579	664
3	477	519	524	481	517
4	163	265	254	290	284
5	401	449	411	431	503

Cuadro 5.13: Tabla contingencia comparando 5 clústers y las ligas (Ward)

El *pvalor* asociado al test de independencia χ^2 es de $1{,}134e-06$, un valor que rechaza la hipótesis nula de independencia entre grupos. A raíz de esto se plantea un análisis de correspondencias múltiple con los clústers creados, las posiciones reales y las ligas.



(a) Scree plot MCA ligas

	Dim 1	Dim 2	Dim 3	Dim 4
DF	0.582	0.267	0.082	0.015
DF,MF	0.032	0.012	0.004	0.483
FW	0.548	0.282	0.122	0.000
MF	0.065	0.383	0.381	0.024
MF,FW	0.353	0.198	0.360	0.003
Bundesliga	0.006	0.000	0.019	0.171
LaLiga	0.000	0.000	0.000	0.003
Ligue 1	0.000	0.007	0.007	0.008
Premier League	0.000	0.000	0.006	0.028
Serie A	0.008	0.004	0.015	0.138
1	0.617	0.258	0.082	0.001
2	0.298	0.032	0.142	0.243
3	0.262	0.281	0.342	0.001
4	0.023	0.211	0.209	0.193
5	0.373	0.323	0.163	0.038

(b) Calidad de representación en 4 dimensiones

Figura 5.15: Comparación entre el scree plot y la calidad de representación

Atendiendo al *Scree plot* obtenido 5.15a vemos que el codo se encuentra en la 4 componente, por lo que parece razonable extraer 4 componentes. Si observamos la tabla 5.15b que representa la calidad de representación para las 4 componentes extraídas vemos que las posiciones reales están bien representadas en las primeras dos dimensiones, así como los diferentes clústers salvo los clústers 2,3,4 que tienen una distribución más equilibrada a lo largo de todas la dimensiones. En contraste, las ligas presentan valores muy bajos en todas las dimensiones, indicando que no contribuyen significativamente a la varianza explicada. A continuación se representan 3 *biplots* para las dimensiones 1-2, 2-3 y 3-4. Observando la figura 5.16. No podemos sacar ninguna conclusión sobre asociación entre los clústers creados y las posiciones reales.

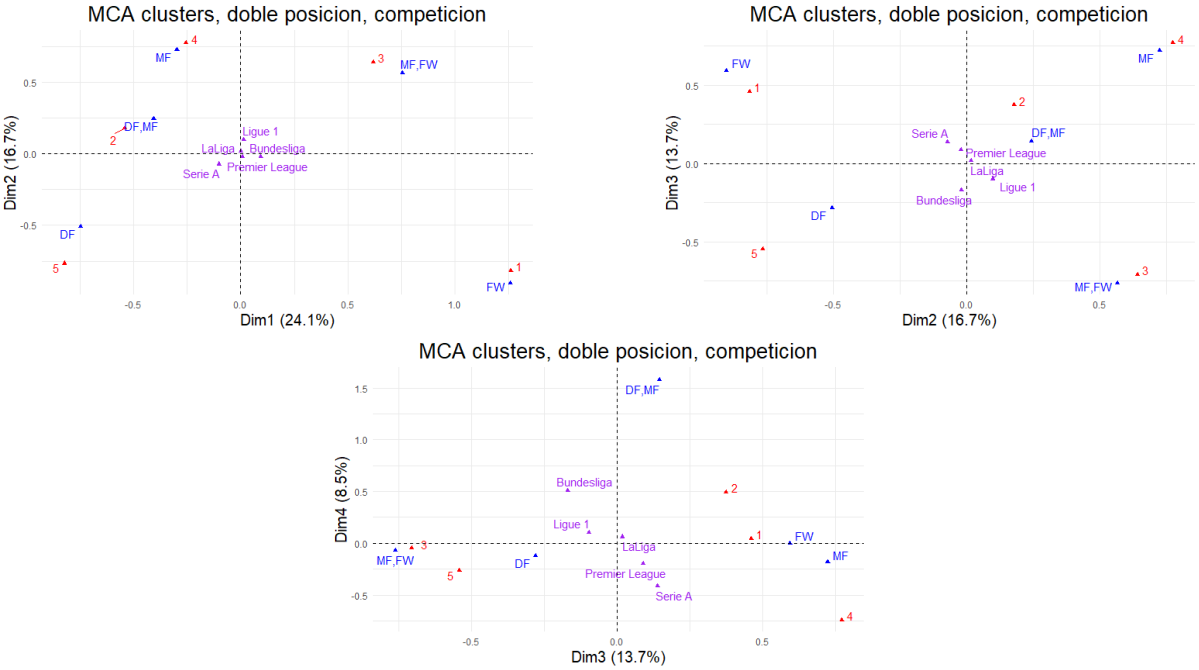


Figura 5.16: Biplot MCA con la liga (Dimensiones 1-2, 2-3, 3-4)

Capítulo 6

Conclusiones

En este trabajo de fin de grado se han utilizado diferentes estadísticas *in-game* de jugadores de fútbol de las cinco grandes ligas europeas desde la temporada 2017-28 hasta la 2022-2023. Estos datos se han obtenido haciendo scrapping con python de la pagina web fbref.com [3].

Durante el desarrollo de este trabajo fin de grado se ha tratado de aplicar una técnica de clasificación no supervisada como es el clústering para generar buenos resultados a la hora de clasificar a los diferentes jugadores en 5 perfiles diferentes asociados a la posición, sin tener en cuenta la misma. Para ello se ha tratado de optimizar un algoritmo de selección de variables mediante la utilización de dos conjuntos de datos y 6 conjuntos variables iniciales diferentes. Cuatro de estos seis se han obtenido mediante la utilización métodos no supervisado atendiendo a dos criterios diferentes. Los dos restantes se han obtenido mediante la utilización de un método de clasificación supervisada como es el *Random Forest*. Esta optimización ha dado algunos resultados aceptables para 5 clústers haciendo que cada clúster corresponda a una posición diferente.

Se han utilizado los métodos de *Ward* y K-medias para las creación de los clústers. El algoritmo de selección de variables funciona bien para el método de las k-medias pero es muy lento para el método de *Ward*. Por eso se decidió hacer la optimización únicamente con el método de las k-medias de esta forma solo aplicaríamos el método de Ward para el mejor caso obtenido.

Si observamos tanto la figura 5.6 como la figura 5.5 se observa que para el método de las *K-medias* no se llega a realizar una correspondencia completa entre cada clúster y cada posición en el campo. Pero si evaluamos su error observamos que es el más bajo obtenido. Por el contrario, con el método de Ward conseguíamos tener una correspondencia de las 5 clústers con las 5 posiciones. Sin embargo al realizar la clasificación se obtienen resultados bastante peores que con el método de las K-medias.

Finalmente se ha observado que para tener una buena correspondencia entre posiciones y perfiles tanto para el método de Ward como para el de K-medias las variables seleccionadas tienen que compartir un rasgo, tienen que abarcar la mayoría del espectro de las 7 categorías en las que se dividen las variables. (Estándar, Tiro, Pase, Creación de Tiros y Goles, Defensa Posesión y Tiempo de juego).

Referencias

[1] Christopher Expósito Izquierdo y Airam Expósito Márquez y Israel López Plata y Belén Melián Batista y J. Marcos Moreno Vega. *Clustering jerárquico*. Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna. 2023.

[2] Diego Calvo. *Análisis cluster jerárquico en R*. https://www.diegocalvo.es/analisis-cluster-jerarquico-en-r/. 2023.

[3] FBref. *Football Statistics and History — FBref.com*. FBref.com, https://fbref.com/. n.d.

[4] FIFPRO. *Investigación FIFPRO: Crece la cantidad de partidos consecutivos en el fútbol masculino*. URL: %5Curl%7Bhttps://fifpro.org/es/apoyar-a-los-y-las-futbolistas/salud-y-rendimiento/carga-de-trabajo-del-futbolista/investigacion-fifpro-crece-la-cantidad-de-partidos-consecutivos-en-el-futbol-masculino/%7D (visitado 15-04-2024).

[5] FootyStats. *FootyStats - Detailed Football Statistics*. 2024. URL: https://footystats.org/.

[6] Asociación de Futbolistas Españoles. *Análisis de Datos aplicados al Scouting*. 2019. (Visitado 21-06-2024).

[7] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. New York: Springer, 2009, pág. 745. ISBN: 9780387848570, 9780387848587. DOI: 10.1007/978-0-387-84858-7. URL: https://doi.org/10.1007/978-0-387-84858-7.

[8] ICHI.PRO. *Método de Silueta mejor que el método del codo para encontrar clústeres óptimos*. 2024. URL: https://ichi.pro/es/metodo-de-silueta-mejor-que-el-metodo-del-codo-para-encontrar-clusteres-optimos-61080390822033.

[9] RS Kriegs. *Modelling Football Players Values on a Transfer Market*. 2024. URL: https://github.com/RSKriegs/Modelling-Football-Players-Values-on-a-Transfer-Market.

[10] Jennifer Roque López. «TÉCNICAS DE SELECCIÓN DE VARIABLES EN REGRESIÓN LINEAL MÚLTIPLE». A thesis submitted in conformity with the requirements for the MSc in Economics, Finance and Computer Science. Tesis de mtría. University of Huelva & International University of Andalusia, 2021.

[11] Sara Martí. «El big data como diagnóstico y prevención de las lesiones deportivas». En: *Economía3* (2022). (Visitado 21-06-2024).

[12] Itziar Fernández Martinez. *ANOVA un Factor*. Diapositivas. Grado en Estadística y Grado en Matemáticas, Universidad de Valladolid. 2021.

[13] Alan Miller. *Subset Selection in Regression*. 2nd Edition. New York: Chapman y Hall/CRC, 2002, pág. 256. ISBN: 9780429119187. DOI: 10.1201/9781420035933. URL: https://doi.org/10.1201/9781420035933.

[14] Víctor Mulero Merino. *Estudio de técnicas de clustering aplicadas a una competición profesional de fútbol*. Trabajo de Fin de Grado, Curso 2022-2023. 2023.

[15] Statologos. *Análisis de componentes principales en R: ejemplo paso a paso*. 2021. URL: https://statologos.com/analisis-de-componentes-principales-en-r/.

[16] Statologos. *Una guía para el uso de pruebas post hoc con ANOVA*. 2021. URL: https://statologos.com/pruebas-anova-post-hoc/#google_vignette.

[17] StatsBomb. *Free Data*. 2024. URL: https://statsbomb.com/what-we-do/hub/free-data/.

[18] StatsBomb. *StatsBomb Open Data*. 2024. URL: https://github.com/statsbomb/open-data.

[19] Rubén P. Suárez. *Así está el ranking de gasto en fichajes por competición: LaLiga, fuera del top 5*. [Accessed 1-Jun-2024]. 2023. URL: https://masfichajes.com/premier-league/2023-09-02/ranking-gasto-fichajes-ligas/.

[20] Mario Garrido Tapias. *Uso de técnicas de clustering para encontrar perfiles de jugadores en una competición de fútbol profesional*. Trabajo de Fin de Grado. 2022.

[21] Miguel Alejandro Fernández Temprano. *Análisis de correspondencias*. Slides. Diapositivas. Grado en Estadística y Grado en Matemáticas, Universidad de Valladolid. 2021.

[22] Miguel Alejandro Fernández Temprano. *Análisis en Componentes Principales*. Slides. Diapositivas. Grado en Estadística y Grado en Matemáticas, Universidad de Valladolid. 2021.

[23] Transfermarkt. *Detalles de Jugador - Búsqueda de Jugador*. 2024. URL: https://www.transfermarkt.es/detailsuche/spielerdetail/suche.

[24] Wikipedia. *K-medias*. 2023. URL: https://es.wikipedia.org/wiki/K-medias (visitado 21-06-2024).

ANEXOS

ANEXO A: Variables

A continuación se muestran las variables numéricas de los conjunto de datos. El conjunto 1 se compone de todas las variables numéricas que se van a definir a continuación. El conjunto 2 esta compuesto por aquellas resaltadas en negrita. Además, algunas variables están marcadas con un asterisco (*). Son las variables elegidas que se dividen entre la variable minutes_90s (minutos jugados 90) de manera que la estadística tiene un valor relativo en relación al tiempo de juego.

Estadísticas estándar:

age – edad
birth_year – año de nacimiento
games – número de partidos
games_starts – número de partidos comenzados
minutes – minutos jugados
minutes_90s – minutos jugados / 90
goals – goles marcados
assists – asistencias
goals_assists – goles + asistencias
goals_pens – goles marcados que no son de penalti
pens_made – penaltis marcados
pens_att – penaltis lanzados
cards_yellow – número de tarjetas amarillas
cards_red – número de tarjetas rojas
goals_per90 – goles cada 90 minutos
assists_per90 – asistencias cada 90 minutos
xg – goles esperados *
npxg – goles esperados que no son de penalti *
xg_assist – goles esperados a partir de un pase que lleva a un tiro *
npxg_xg_assist – goles esperados que no son de penalti + asistencias *
xg_per90 – goles esperados cada 90 minutos
xg_assist_per90 – asistencias esperadas cada 90 minutos
xg_xg_assist_per90 – goles esperados + asistencias esperadas cada 90 minutos
npxg_per90 – goles esperados que no son de penalti cada 90 minutos
npxg_xg_assist_per90 – goles esperados que no son de penalti + asistencias cada 90 minutos
npxg_per_shot – goles esperados que no son de penalti por disparo *

xg_net – goles - goles esperados *
npxg_net – goles que no son de penalti - goles esperados que no son de penalti *

Estadísticas de tiro:

shots – tiros totales *
shots_on_target – tiros a puerta *
shots_on_target_pct – porcentaje de tiros a puerta
shots_per90 – tiros cada 90 minutos
shots_on_target_per90 – tiros a puerta cada 90 minutos
goals_per_shot – goles por tiro
goals_per_shot_on_target – goles por tiro a puerta
average_shot_distance – distancia media entre el tirador y la portería de todos los tiros
shots_free_kicks – tiros libres *

Estadísticas de pase:

passes_completed – pases completados *
passes – pases intentados *
passes_pct – porcentaje de pases completados
passes_total_distance – distancia total de los pases completados
passes_progressive_distance – distancia total de los pases completados hacia la portería rival
passes_completed_short – pases completados entre 5 y 15 yardas (cortos) *
passes_short – pases cortos intentados *
passes_pct_short – porcentaje de pases cortos completados
passes_completed_medium – pases completados entre 15 y 30 yardas (a media distancia) *
passes_medium – pases a media distancia intentados *
passes_pct_medium – porcentaje de pases a media distancia completados
passes_completed_long – pases completados de mas de 30 yardas (largos) *
passes_long – pases largos intentados *
passes_pct_long – porcentaje de pases largos completados
xg_assist_net – asistencias - goles esperados asistidos
assisted_shots – pases que asisten un tiro *
passes_into_final_third – pases completados que entran en 1/3 del campo (hacia la portería rival) *
passes_into_penalty_area – pases completados que entran en el área rival *
crosses_into_penalty_area – numero de centros completados que entran en el área rival *
pass_xa – pases que se convierten en asistencia de gol *
passes_live – numero de pases durante el juego *
passes_dead – numero de pases a balón parado (tiros libres, corners, bandas, ...) *
passes_free_kicks – pases desde tiros libres
through_balls – pases que van entre los defensores del equipo rival y crean oportunidad de gol
passes_switches – pases de mas de 40 yardas a lo ancho (cambios de orientacion del juego) *
crosses – numero total de centros intentados *

throw_ins – numero de saques de banda *
corner_kicks – numero de corners *
corner_kicks_in – numero de corners con efecto hacia dentro *
corner_kicks_out – numero de corners con efecto hacia fuera *
corner_kicks_straight – numero de corners sin efecto *
passes_offsides – pases a un jugador en fuera de juego *
passes_blocked – pases bloqueados por un oponente que estaba en la trayectoria *

Estadísticas creación de tiros y goles:
sca – acciones ofensivas que llevan a un tiro (pases, regates, faltas recibidas...) *
sca_per90 – acciones ofensivas que llevan a un tiro cada 90 minutos
sca_passes_live – pases durante el juego que llevan a un tiro *
sca_passes_dead – pases a balón parado que llevan a un tiro (tiros libres, corners, bandas, ...) *
sca_take_ons – regates que llevan a un tiro *
sca_shots – tiros que llevan a otro tiro *
sca_fouled – faltas recibidas que llevan a un tiro *
sca_defense – acciones defensivas que llevan a un tiro *
gca – acciones ofensivas que llevan a un gol (pases, regates, faltas recibidas...) *
gca_per90 – acciones ofensivas que llevan a un gol cada 90 minutos
gca_passes_live – pases durante el juego que llevan a un gol *
gca_passes_dead – pases a balón parado que llevan a un gol (tiros libres, corners, bandas, ...) *
gca_take_ons – regates que llevan a un gol *
gca_shots – tiros que llevan a otro tiro que se convierte en gol *
gca_fouled – faltas recibidas que llevan a un gol *
gca_defense – acciones defensivas que llevan a un gol *

Estadísticas de defensa:
tackles – numero de entradas intentadas *
tackles_won – numero de entradas exitosas (recupera la posesión de manera limpia)*
tackles_def_3rd – entradas en el 1/3 defensivo *
tackles_mid_3rd – entradas en el 1/3 medio *
tackles_att_3rd – entradas en el 1/3 ofensivo *
challenge_tackles – entradas exitosas a un jugador que intenta regatear *
challenges – entradas totales a un jugador que intenta regatear *
challenge_tackles_pct – porcentaje de entradas exitosas a un jugador que intenta regatear
challenges_lost – entradas no exitosas a un jugador que intenta regatear *
blocks – balones bloqueados estando en la trayectoria del balón *
blocked_shots – tiros bloqueados estando en la trayectoria del balón *
blocked_passes – pases bloqueados estando en la trayectoria del balón *
interceptions – balones interceptados *
tackles_interceptions – entradas + intercepciones *
clearances - despejes *
errors – errores que llevan a un disparo del oponente *
ball_recoveries – balones recuperados *

Estadísticas de posesión:

touches – numero de veces que un jugador toca el balón (recibir, moverse y pasar cuenta como 1. Balones parados también cuentan como 1 toque) *
touches_def_pen_area – toques en el área defensiva *
touches_def_3rd – toques en el 1/3 defensivo *
touches_mid_3rd – toques en el 1/3 medio *
touches_att_3rd – toques en el 1/3 ofensivo *
touches_att_pen_area – toques en el área rival *
touches_live_ball – toques durante el juego (no cuentan balones parados) *
take_ons – regates intentados *
take_ons_won – regates exitosos *
take_ons_won_pct – porcentaje de regates exitosos
take_ons_tackled – regates no exitosos (se lleva el balón el oponente con una entrada) *
take_ons_tackled_pct – porcentaje de regates no exitosos
carries – numero de conducciones *
carries_distance – distancia recorrida conduciendo el balón *
carries_progressive_distance – distancia recorrida conduciendo el balón hacia la portería rival *
carries_into_final_third – numero de conducciones que entra en el 1/3 ofensivo *
carries_into_penalty_area – numero de conducciones que entran al área rival *
miscontrols – numero de veces que el jugador falla intentando controlar el balón *
dispossessed – numero de veces que el jugador pierde el balón tras una entrada rival *
passes_received – numero de pases recibidos *
progressive_carries – conducciones hacia la portería rival *
progressive_passes – pases hacia la portería rival (no cuentan en el 40 % del campo del equipo propio) *
progressive_passes_received – pases progresivos recibidos *

Estadísticas de tiempo de juego:
minutes_per_game – minutos por partido
minutes_pct – porcentaje de minutos del jugador sobre el total del equipo
minutes_per_start – minutos por partido que comienza jugando
games_complete – partidos completos jugados
games_subs – partidos que el jugador no comienza (suplente)
minutes_per_sub – minutos jugados por partido que es sustituto
unused_subs – partidos que no ha salido a jugar estando de sustituto
points_per_game – puntos obtenidos por el equipo en partidos en los que el jugador ha jugado
on_goals_for – goles marcados por el equipo cuando el jugador esta jugando
on_goals_against – goles marcados por el rival cuando el jugador esta jugando
plus_minus – goles - goles recibidos cuando el jugador esta jugando *
plus_minus_per90 – goles - goles recibidos cuando el jugador esta jugando cada 90 minutos
plus_minus_wowy – goles cada 90 minutos a favor - goles cada 90 minutos en contra cuando el jugador esta jugando *
on_xg_for – goles esperados por el equipo cuando el jugador esta jugando *
on_xg_against - goles esperados recibidos cuando el jugador esta jugando *
xg_plus_minus – goles esperados marcados - goles esperados recibidos cuando el jugador

esta jugando *
xg_plus_minus_per90 - goles esperados marcados - goles esperados recibidos cuando el jugador esta jugando cada 90 minutos
xg_plus_minus_wowy – goles esperados cada 90 minutos a favor - goles esperados cada 90 minutos en contra cuando el jugador esta jugando *

Otras estadísticas:
cards_yellow_red – roja por doble tarjeta amarilla
fouls – numero de faltas cometidas *
fouled – numero de faltas recibidas *
offsides – numero de fueras de juego *
pens_won – numero de penaltis que recibe el jugador *
pens_conceded – numero de penaltis que hace el jugador *
own_goals – goles en propia puerta *
aerials_won – duelos aéreos ganados *
aerials_lost – duelos aéreos perdidos *
aerials_won_pct – porcentaje de duelos aéreos ganados

ANEXO B: Trazas de Algoritmo StepWise

Varibles Iniciales con Cargas Cuadrado

Algoritmo para seleccionar las mejores variables Variables iniciales: progressive_passes_received, sca_passes_live, passes_into_final_third, passes_free_kicks, progressive_carries, shots_free_kicks, passes_pct_medium, passes_pct_short, carries_into_final_third, carries_into_penalty_area

SELECCION HACIA DELANTE
Entra throw_ins - error ±0,482
SELECCION HACIA ATRAS
Sale passes_free_kicks - error 0.471
SELECCION HACIA DELANTE
Entra touches_def_pen_area - error 0.411
SELECCION HACIA ATRAS
Sale sca_passes_live - error 0.405
SELECCION HACIA DELANTE
Entra challenges_lost - error 0.387
SELECCION HACIA ATRAS
Sale carries_into_penalty_area - error 0.38
SELECCION HACIA DELANTE
Entra offsides - error 0.365
SELECCION HACIA ATRAS
Sale carries_into_final_third - error 0.357
SELECCION HACIA DELANTE
Entra xg - error 0.341
SELECCION HACIA ATRAS
Sale shots_free_kicks - error 0.337
SELECCION HACIA DELANTE
Entra touches_def_3rd - error 0.33
SELECCION HACIA ATRAS
Sale passes_pct_medium - error 0.328
SELECCION HACIA DELANTE
Entra tackles_def_3rd - error 0.324
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra passes_medium - error 0.322
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra goals_pens - error 0.322
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra dispossessed - error 0.321
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra aerials_won - error 0.32
SELECCION HACIA ATRAS

SELECCION HACIA DELANTE
Entra passes_pct_long - error 0.319
SELECCION HACIA ATRAS
Sale passes_pct_short - error 0.319
SELECCION HACIA DELANTE
Entra blocks - error 0.317
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra passes_pct_medium - error 0.316
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra crosses_into_penalty_area - error 0.315
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra on_goals_for - error 0.314
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra cards_red - error 0.313
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra fouled - error 0.313
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra cards_yellow - error 0.312
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra touches_att_pen_area - error 0.311
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra carries - error 0.311
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra plus_minus - error 0.31
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Warning: Quick-TRANSfer stage steps exceeded maximum (= 545600)
Warning: Quick-TRANSfer stage steps exceeded maximum (= 545600)
SELECCION HACIA ATRAS

Varibles Iniciales Importancia de Random Forest

Algoritmo para seleccionar las mejores variables Variables iniciales: touches_def_3rd, throw_ins, touches_def_pen_area, clearances, progressive_passes_received, passes_dead, shots, miscontrols, ball_recoveries, passes_completed_medium

SELECCION HACIA DELANTE
Entra passes_pct_short - error 0.338
SELECCION HACIA ATRAS

Sale passes_dead - error 0.338
SELECCION HACIA DELANTE
Entra offsides - error 0.332
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra crosses_into_penalty_area - error 0.326
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra dispossessed - error 0.325
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra tackles - error 0.324
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra passes_pct_long - error 0.324
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra passes_blocked - error 0.321
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra gca_take_ons - error 0.321
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra touches_att_pen_area - error 0.32
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra aerials_won - error 0.319
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra goals - error 0.318
SELECCION HACIA ATRAS
Sale shots - error 0.318
SELECCION HACIA DELANTE
Entra games_starts - error 0.316
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
Entra challenge_tackles_pct - error 0.315
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
SELECCION HACIA ATRAS

Método de Ward

Algoritmo para seleccionar las mejores variables

Variables iniciales:
passes_into_final_third, own_goals, challenges_lost, touches_def_3rd, shots_per90, aerials_lost,

passes_pct_long, dispossessed, tackles_mid_3rd, goals_pens, carries_progressive_distance, minutes_per_game, unused_subs , challenge_tackles_pct , offsides, clearances , passes_dead , gca_defense, fouled, npxg , corner_kicks_straight, npxg_per_shot

SELECCION HACIA DELANTE
Entra goals_per_shot - error 0.358
SELECCION HACIA ATRAS
SELECCION HACIA DELANTE
SELECCION HACIA ATRAS

ANEXO C

A continuación se presenta un enlace con el código utilizado para los ejemplos del capítulo 2, el análisis descriptivo realizado en el capítulo 4 y el desarrollo del capítulo 5.

`https://uvaes.sharepoint.com/sites/CdigoTFGESTJuanGonzlezMagdalena/Documentos%20compartidos/Forms/AllItems.aspx`