UNIVERSITAT JAUME I

Grado en Matemática Computacional

Trabajo Final de Grado

Análisis estadístico de las cinco grandes ligas de fútbol europeas

Autor:
Nicolás Camañes Antolín

Tutor académico:
Marina Martínez García

Fecha de lectura: 23 de Junio de 2023
Curso académico 2022/2023

Resumen

Este documento se corresponde con el trabajo realizado para la asignatura MT1054 - Treball Final de Grau del grado en Matemáticas Computacionales de la Univestat Jaume I (UJI).

El proyecto se basa en el empleo de diferentes técnicas de minería de datos. El análisis de las componentes principales, para reducir la dimensión de un conjunto de datos. Los contrastes de hipótesis para evaluar los datos. La regresión logística, para predecir la ocurrencia de un evento binario. Y, la regresión ordinal, para predecir variables categóricas. El empleo de estas técnicas tiene como propósito realizar el estudio de datos sobre las diferentes ligas europeas de fútbol. A parte de la implementación y descripción de las técnicas empleadas, en el documento se encuentran explicados los fundamentos teóricos en los que se sustentan dichos procedimientos. Además, se exponen los resultados obtenidos en la implementación de dichos procedimientos.

Palabras clave

Regresión logística, Regresión ordinal, Análisis de las Componentes Principales y Fútbol.

Keywords

Logistic regression, Ordinal regression, Principal Component Analysis and Football.

Índice general

1. Introducción **7**

1.1. Contexto y motivación del proyecto 7

1.2. Objetivos del proyecto 9

2. Desarrollo del TFG **11**

2.1. Análisis descriptivo 11

2.1.1. Introducción 11

2.1.2. Fundamentos teóricos 12

2.1.3. Resultados obtenidos 18

2.2. Regresión logística 26

2.2.1. Introducción 26

2.2.2. Fundamentos teóricos 27

2.2.3. Resultados obtenidos 34

2.3. Regresión ordinal 42

2.3.1. Introducción 42

2.3.2. Fundamentos teóricos 42

2.3.3. Resultados obtenidos 43

3. Conclusiones **49**

A. Descripción del dataset **53**

B. Repositorio de GitHub con los programas implementados **57**

Capítulo 1

Introducción

En este primer capítulo, se lleva a cabo la descripción del contexto en el que el proyecto ha sido realizado. Además también se explica la motivación. Y se concluye con una breve explicación de los objetivos a satisfacer.

En el segundo capítulo, se encuentra el desarrollo del proyecto realizado. El capítulo está dividido en tres secciones diferentes, donde cada una de estas secciones sigue la misma estructura. En primer lugar se introducen el motivo de la realización y sus objetivos. A continuación se detallan todos los fundamentos teóricos matemáticos en los que se basan los algoritmos empleados. Y finalmente, se exponen los resultados obtenidos en cada una de las pruebas realizando una valoración de estos.

En el tercer capítulo se presentan las conclusiones resultantes de la realización del proyecto tanto generales como personales.

Por último, el documento cuenta con un anexo con dos secciones. En la primera se encuentra una explicación detallada de los datos mediante los cuales se ha construido el dataset. En la segunda, se encuentra el enlace para poder acceder tanto a los ficheros de la base de datos usada como a los programas creados mediante Python a través de la aplicación web Jupyter Notebook.

1.1. Contexto y motivación del proyecto

El fútbol es uno de los deportes más seguidos en toda europa. Son millones las personas que, no sólo los siguen sino que además, viven y basan sus negocios entorno a este deporte. La Liga española, la Premier League inglesa, la Serie A italiana, la Ligue 1 francesa y la Bun-

desliga alemana, son reconocidas como las competiciones con más rivalidad y emoción a nivel internacional. Es por ello, que en ellas se encuentran los mejores jugadores del mundo. Debido a su popularidad y como consecuencia del avance diario de las tecnologías, la cantidad de información y datos estadísticos que se generan para poder ser analizados por los expertos está sufriendo un crecimiento considerable en los últimos tiempos.

El análisis estadístico deportivo tuvo sus orígenes en el béisbol, durante la década de 1860, cuando Henry Chadwick comenzó a realizar anotaciones acerca de los bateos, lanzamientos y carreras de los partidos para su posterior estudio. No obstante, no fue hasta la década de 1970 cuando este ámbito de las matemáticas empezó a popularizarse. En la actualidad, sigue habiendo un creciente interés en el análisis estadístico en este ámbito. Como consecuencia de la gran cantidad de datos que se recogen, para muchas empresas tanto dentro como fuera del del deporte se han abierto nuevas oportunidades de mercado en cuanto a la aplicación de métodos estadísticos para el análisis y comprensión de diferentes aspectos del juego. Gracias a estas herramientas, se hace posible identificar variables relevantes, evaluar la influencia de distintos factores en los resultados o pronosticar el rendimiento futuro de los equipos y jugadores entre otras muchas cosas.

A través de este proyecto, se pretende llevar a cabo un análisis estadístico de las cinco grandes ligas europeas de fútbol en los últimos cinco años. Obteniendo información que contribuya a una mejor comprensión del rendimiento de los equipos y las diferentes ligas.

Por último, considero importante mencionar las dos inspiraciones que dieron lugar a un interés personal en el análisis estadístico aplicado al ámbito deportivo. En primer lugar, el actual propietario del equipo de fútbol CD Castellón, Haralabos Voulgaris quien adquirió el club en 2022. El greco canadiense, es conocido por haberse convertido en un apostador profesional y experto en el análisis estadístico en el ámbito del baloncesto. Su gran renombre en la NBA se debe a que es considerado como uno de los pioneros en la realización de análisis estadísticos y algoritmos de aprendizaje automático para la evaluación de equipos y jugadores. Consiguiendo, como fruto resultante de sus actividades, una gran fortuna. En segundo lugar, la película *Moneyball*, dirigida por Bennett Miller y estrenada en el año 2011. La trama, protagonizada por Brad Pitt, consiste en la historia real en la que un entrenador de béisbol y un economista se juntan y deciden emplear el análisis estadístico para identificar jugadores infravalorados que pueden tener un rendimiento significativo en los resultados del equipo. La película muestra como la aplicación de la estadística deportiva desafía las convenciones establecidas en el béisbol y genera resistencia por parte de los miembros más tradicionales de la industria. A medida que el equipo implementa estrategias basadas en el análisis estadístico, los resultados comienzan a ser positivos y los Oakland Athletics logran una racha exitosa en la temporada.

1.2. Objetivos del proyecto

Mediante la realización de este proyecto se pretende el aprendizaje de los conceptos matemáticos de los algoritmos implementados, la programación necesaria para poder implementar los diferentes algoritmos y el análisis de los resultados pertinente para cada prueba.

Capítulo 2

Desarrollo del TFG

2.1. Análisis descriptivo

2.1.1. Introducción

En esta sección se pretende una primera toma de contacto con los datos. La base de datos se encuentra explicada detalladamente en el anexo de este documento. Resumidamente, se trata de los datos recopilados acerca de 62 variables parametrizables al concluir una temporada de fútbol regular, sobre cada uno de los equipos que ha participado durante las cinco últimas temporadas finalizadas en las cinco grandes ligas de fútbol europeas: Premier League, La Liga, Serie A, Ligue 1 y Bundesliga. Para una mayor claridad se dividirá la sección en dos fases.

En la primera fase, lo que se pretende hacer es comprobar si existen diferencias entre los equipos en función de su posición en la clasificación. Para ello, en primer lugar se evaluará, mediante el Test de Kruskal-Wallis, si es posible encontrar diferencias tanto entre los equipos de una misma liga como entre las diferentes ligas. Y, en caso de encontrar diferencias, mediante el Test Post Hoc de Tukey se tratará de reagrupar estos grupos. Este análisis se realizará únicamente teniendo en cuenta los datos registrados en la temporada 2021/2022 sobre: goles a favor (GoalsFavor), tiros a puerta (ShotsOnTarget), pases completados con éxito (PassesCompleted) y faltas realizadas (FoulsCommited). La elección de estas variables se debe a que son cuatro de las variables más generales y populares acercca de los equipos del dataset.

Para realizar esto en cada liga se dividirán los equipos en cuatro grupos, en función de su posición final (Rk) en la temporada de acuerdo con el siguiente criterio:

- Para **La Liga**, **Premier League**, **Serie A** y **Ligue 1**
$$\begin{cases} \text{Grupo 1: Del 1 al 5} \\ \text{Grupo 2: Del 6 al 10} \\ \text{Grupo 3: Del 11 al 15} \\ \text{Grupo 4: Del 16 al 20} \end{cases}$$

- Para **Bundesliga**
$$\begin{cases} \text{Grupo 1: Del 1 al 4} \\ \text{Grupo 2: Del 5 al 9} \\ \text{Grupo 3: Del 10 al 14} \\ \text{Grupo 4: Del 15 al 18} \end{cases}$$

Además, también se graficarán los boxplots correspondientes a cada comparación para ayudar a interpretar los resultados obtenidos en los contrastes de hipótesis.

En la segunda fase, los datos con los que se ha trabajado se han obtenido juntando los datos de todas las ligas y de las cinco temporadas registradas en el dataset. Para cada uno de los equipos, se va a trabajar con las siguientes variables: Puntos, Victorias, Empates, Derrotas, Goles a favor, Goles en contra, Asistencias, Tiros, Pases completados, Regates, Toques totales, Faltas, Interceptaciones, Posesiones, Edad de los jugadores, Corners y Duelos aéreos ganados. Cada una de las variables se ha obtenido realizando la media de todas las temporadas registradas para cada uno de los equipos. Por ejemplo, si un equipo sólo ha participado en tres temporadas, la variable Derrotas contiene la media de derrotas registradas por ese equipo en las tres temporadas en las que ha participado. La elección de estas variables en este caso se basa en que se trata del conjunto de variables más genérico de los equipos, es decir, no sirven para reflejar aspectos muy específicos del juego. Para continuar con el análisis, se ha generado el mapa de correlaciones y se ha usado la técnica del PCA para poder identificar los patrones ocultos en los datos y, además, poder graficar los equipos gracias a la reducción de la dimensionalidad.

2.1.2. Fundamentos teóricos

Test de Shapiro-Wilk

El objetivo de los análisis de la normalidad, o test de normalidad, es comprobar si los datos provienen de una distribución normal. Para llevarlo a cabo, se realizará un contraste de hipótesis mediante el test de Shapiro-Wilk [15]. Se considerarán las siguientes hipótesis:

$$\begin{cases} H_0 : \text{La distribución de la variable es Normal} \\ H_A : \text{La distribución de la variable NO es Normal} \end{cases}$$

Para este contraste de hipótesis, el estadístico que se empleará será el siguiente:

$$W = \frac{D^2}{nS^2}$$

Donde D es la suma de las diferencias corregidas, que se calculan realizando las diferencias entre: el primero y el último, el segundo y el penúltimo, el tercero y el antepenúltimo y así sucesivamente. Y, una vez calculadas estas diferencias, se corrigen con los valores de las tablas de Shapiro-Wilk y se suman para obtener el valor de D. S^2 hace referencia a la varianza de la muestra y por último, n hace referencia al tamaño de la muestra.

Una vez calculado el estadístico, se calcula el valor del p-valor correspondiente, que tendrá un valor entre 0 y 1, se podrá tomar la decisión de cual de las dos hipótesis aceptar. Se podrá rechazar la hipótesis nula (H_0), si el p-valor obtenido es menor que 0.05. De lo contrario, se aceptará la hipótesis nula.

Test Kruskal-Wallis

Si tras realizar el test anterior, se ha obtenido que los datos no provienen de una distribución normal. No se podrá aplicar una ANOVA para evaluar la existencia de diferencias en la medina, para los diferentes grupos.

En su lugar, es conveniente realizar el test de Kruskal-Wallis [10] y [14]. Se trata de una alternativa a la ANOVA, pero en este no es necesario que los datos sigan una distribución normal. Este test consiste en un contraste de hipótesis. Lo que se realiza es comparar las diferentes muestras con tal de comprobar si están equilibradas. Es decir, se comprueba si los datos pertenecen a una misma distribución. Para este contraste, se consideran las dos siguientes hipótesis:

$$\begin{cases} H_0 : \text{Misma mediana para todos los grupos.} \\ H_A : \text{Al menos uno de los grupos tiene una mediana diferente al resto.} \end{cases}$$

La fórmula para calcular el estadístico de este test se corresponde con la siguiente:

$$H = \left(\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(N+1)$$

Donde N es el tamaño total de toda la muestra, k es la cantidad de grupos en la muestra, R_i es la suma de rangos de cada uno de los grupos y n_i es el tamaño de cada uno de los grupos.

Si existe homogeneidad entre las distribuciones de los k grupos, el estadístico se distribuirá según una ji-cuadrado con k-1 grados de libertad. Una vez calculado el estadístico, se obtendrá el p-valor correspondiente a dicho valor obtenido, que tendrá un valor entre 0 y 1. El requisito para poder rechazar la hipótesis nula (H_0) y, por consecuencia, aceptar la hipótesis alternativa (H_A), será que el p-valor sea menor que 0.05.

Test Post Hoc de Tukey

El objetivo de esta prueba es, entre los diferentes grupos que forman el dataset, identificar entre cuales de ellos hay diferencias significativas y entre cuales no hay apenas diferencias. Es más, se podrá llegar a encontrar posibles agrupaciones entre diferentes grupos similares de tal forma que se dé con combinaciones de grupos con características semejantes.

Para esto, se empleará el test post hoc de Tukey [13] y [9]. Se trata de una prueba post hoc mediante la cual se puede identificar qué media es significativamente diferente respecto a las demás. Esto se realiza comparando las medias de los grupos par a par. Por tanto, las hipótesis del contraste serán las siguientes:

$$\begin{cases} H_0 : \mu_i = \mu_j \text{ con i=1, ...,k y j=1, ...,k} \neq \text{i} \\ H_1 : \text{Al menos alguna de las medias es diferente al resto} \end{cases}$$

El estadístico del contraste se calcula de la siguiente forma:

$$W = q_{(\alpha,glee,t)} \sqrt{\frac{CMee}{r}}$$

Donde $q_{(\alpha,glee,t)}$ es un valor de las tablas de Tukey que depende del nivel de significación (α), los grados de libertad del error experimental (glee) y el número de tratamientos (t). CMee es el cuadrado medio del error experimental y r es el número de repeticiones de las medias de los tratamientos a ser comparados.

Una vez obtenido el estadístico y calculado el p-valor correspondiente, se rechazará la hipostesis nula, y por tanto se aceptará que las medias son diferentes, cuando se obtenga un p-valor inferior a 0,05.

Cabe tener en cuenta que, si realizando el test de Kruskal-Wallis no se rechaza la hipótesis nula, en el test de Tukey no se encontrarán diferencias significativas entre los grupos. La finalidad

de realizar este test, a continuación del test de Kruskal-Wallis, es obtener más información de utilidad y conocer mejor el conjunto de datos.

Análisis de las componentes principales

El análisis de las componentes principales es una técnica estadística que tiene como finalidad encontrar un nuevo conjunto de variables, a las que se llamará componentes principales (CP), que sean una combinación lineal de las variables originales y que expliquen la mayor parte de la varianza en los datos. Esta técnica resultará de gran utilidad para identificar patrones en los datos, describir un conjunto de datos en términos de nuevas variables no correlacionadas y, por ende, reducir la dimensionalidad de estos. Por tanto, el objetivo principal es transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas. Estas, representan la mayor variabilidad de los datos originales y sirven para representar la misma cantidad de información con un número menor de variables.[6]

En esta técnica, se parte de un conjunto de variables originales x_i y se obtienen las nuevas componentes y_i que resultarán de la combinación lineal de las x_i. Estas nuevas variables representarán los patrones más importantes de los datos y permitirán reducir la cantidad de variables necesarias para describirlos.

$$(x_1, x_2, ..., x_m) \Rightarrow (y_1, y_2, ..., y_m)$$

Las características que deben cumplir las nuevas componentes principales es que deberán estar incorrelacionadas y sus varianzas deberán ir decreciendo progresivamente. Es decir, y_1 será la componente principal con mayor varianza mientras y_m será la que menos varianza tenga.

$$y_i = \sum_{j=1}^{m} a_{ij}x_j = a_i x = (a_{i1}, a_{i2}, ..., a_{im}) \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \quad \text{con} \quad ||a_i|| = 1 \quad \forall i = 1, ..., m$$

Se deberá tener en cuenta la técnica de los Multiplicadores de Lagrange para resolver el problema de maximización de varianzas respectivas de cada nueva variable creada y_i.

En primer lugar se deberá buscar a_1 tal que Var(y_1) sea máxima.

$$Var(y_1) = Var(a_1^t x) = a_1^t \Sigma a_1$$

Donde Σ hace referencia a la matriz de covarianzas:

$$\Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \cdots & Cov(x_1, x_m) \\ Cov(x_2, x_1) & Var(x_2) & \cdots & Cov(x_2, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_m, x_1) & Cov(x_m, x_2) & \cdots & Var(x_m) \end{pmatrix}$$

Luego, se deberá encontrar max(a_1^t Σ a_1) donde $a_1^t a_1 = 1$.

Se construye la función L(a_1) = a_1^t Σ a_1 - λ ($a_1^t a_1$ - 1). Para maximizarla se deriva e iguala a cero para encontrar el punto crítico:

$$\frac{\partial L}{\partial a_1} = \Sigma a_1 + a_1^t \Sigma - \lambda a_1 - \lambda a_1^t = \Sigma a_1 + \Sigma a_1 - \lambda I a_1 - \lambda I a_1 = 0$$

$$\frac{\partial L}{\partial a_1} = 2\Sigma a_1 - 2\lambda I a_1 = 2(\Sigma - \lambda I)a_1 = 0 \Rightarrow (\Sigma - \lambda I)a_1 = 0$$

Por el Teorema Rouché-Fröbenius $\Rightarrow$ $/\Sigma/ - \lambda = 0 \rightarrow \lambda$ es un valor propio de $\sum$. Como la matriz de covarianzas, Σ es de orden m y definida positiva, tendrá m valores propios $\lambda_1 > \lambda_2 > ... > \lambda_m$.

Desarrollando $(\Sigma - \lambda I)a_1 = 0$ se obtiene que $\sum a_1 = \lambda I a_1$

Luego, teniendo en cuenta lo anterior y que $a_1^t a_1 = 1$, se calcula:

$$Var(y_1) = Var(a_1^t x) = a_1^t \Sigma a_1 = a_1^t \lambda I a_1 = \lambda a_1^t a_1 = \lambda$$

Para hacer que Var(y_1) sea máxima, $Var(y_1) = \lambda \rightarrow$ se toma λ_1. Además, a_1 será el vector propio asociado a λ_1

Se calcula,

$$\begin{aligned} y_1 &= a_1^t x \\ y_2 &= a_2^t x \\ &\vdots \\ y_m &= a_m^t x \end{aligned}$$

Por último, como se quieren las nuevas componentes incorrelacionadas, se necesitará que

$$cov(y_1, y_2) = 0$$

Por definición, $cov(y_2, y_1) = cov(a_2^t x, a_1^t x) = a_2^t cov(x, x) a_1 = a_2^t E((x - \mu)(x - \mu))a_1 = a_2^t \Sigma a_1$

Se quiere que $a_2^t \Sigma a_1 = 0$. Para ver cuando se da esta condición se parte de (2.1) y se desarrolla de la siguiente forma:

$$\Sigma a_1 = \lambda a_1 \tag{2.1}$$

$$a_2^t(\lambda a_1) = \lambda a_2^t a_1 = 0$$

$$a_2^t a_1 = 0 \rightarrow a_1 \perp a_2$$

Luego para encontrar Var(y_2) se deberá cumplir la condición $\begin{cases} ||a_2|| = 1 \\ a_2 \perp a_1 \rightarrow a_2^t a_1 = 0 \end{cases}$

Llegados a este punto, para calcular el vector propio a_2 se tomará:

$$L(a_2) = a_2^t \Sigma a_2 - \lambda(a_2^t a_2 - 1) - \delta(a_2^t a_1) \tag{2.2}$$

Derivando el lagrangiano, como previamente (2.3). A continuación, se multiplicará por a_1^t (2.4). Y, se simplificará (2.5) y (2.6):

$$\frac{\partial L}{\partial a_2} = 2\Sigma a_2 - 2\lambda a_2 - \delta a_1 = 0 \tag{2.3}$$

$$2a_1^t \Sigma a_2 - 2a_1^t \lambda a_2 - \delta a_1^t a_1 = 0 \tag{2.4}$$

$$2a_1^t \Sigma a_2 \delta = 0 \tag{2.5}$$

$$\delta = 2a_1^t \Sigma a_2 \tag{2.6}$$

Sustituyendo δ en (2.3) y teniendo en cuenta que a_1 y a_2 son ortogonales, se obtiene que:

$$\frac{\partial L}{\partial a_2} = 2\Sigma a_2 - 2\lambda a_2 - 2a_1^t \Sigma a_2 a_1 = 0$$

$$\frac{\partial L}{\partial a_2} = 2(\Sigma - \lambda)a_2 = 0$$

Se cumple el mismo requisito que se había establecido para a_2 que para a_1. Esto es debido a la construcción que se ha establecido, por tanto irán saliendo vectores ortogonales.

En resumen,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}; \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix};$$

$$y = Ax \rightarrow Var(y_i) = \lambda_i \rightarrow \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r \end{bmatrix}$$

$$\Lambda = Var(y) = A^t Var(x) A = A^t \Sigma A \rightarrow \Sigma = A \Lambda A^{-1}$$

Donde A es una matriz ortogonal que satisface $a_i^t a_i = 1 \ \forall i = 1, ..., m$. Y además, también se cumplirá que $A^{-1} = a_t$

2.1.3. Resultados obtenidos

Para la primera fase de la sección, lo primero en realizarse ha sido la generación de seis boxplots por cada una de las variables, con la finalidad de ayudar a la interpretación y comprobación de los resultados que se obtendrán en los contrastes de hipótesis. A continuación, se ha realizado el Test de Shapito-Wilk [2], donde se ha obtenido que los datos de todas las variables estudiadas no representan una distribución normal. Seguidamente a esto, se ha realizado el Test de Kruskal-Wallis [3]. Y finalmente, el Test Post Hoc de Tukey [16], en los casos correspondientes. Pasemos a analizar los resultados obtenidos en estos dos últimos tests para cada una de las variables junto con sus boxplots correspondientes.

En primer lugar, los goles a favor (GoalsFavor). El valor mínimo de esta variable en la temporada 2021/2022 se corresponde con 23, conseguido por el Norwich City, de la Premier League. Mientras que el valor máximo fue de 99, conseguido por el Manchester City, también de la Premier League. En los diagramas de cajas generados, se aprecia que los equipos que ocupan un puesto en la clasificación más alto, son los más goleadores. Algo que resulta congruente, ya que es lo que determina el resultado de un partido.

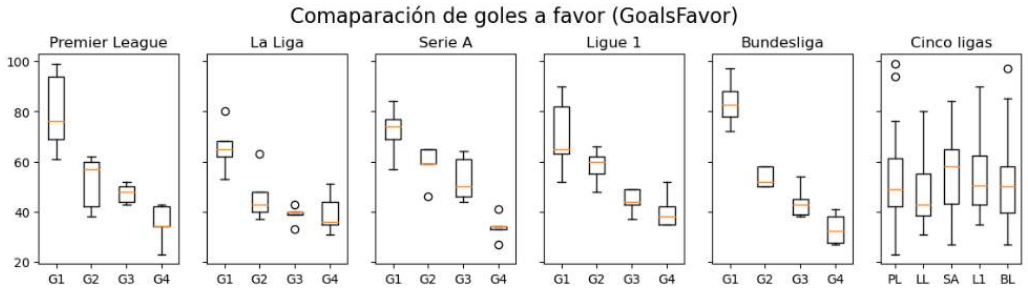

Figura 2.1: Boxplots de la variable GoalsFavor para comparar entre los cuatro grupos definidos dentro de cada liga (cinco primeros boxplots) y entre los equipos de las cinco ligas (último boxplot).

Tras realizar los seis Test de Kruskal-Wallis se ha obtenido que dentro de las cinco ligas existen diferencias significativas entre los distintos grupos para la variable GoalsFavor. En cambio, en los datos de la comparación de las cinco ligas no se han encontrado evidencias de la existencia de diferentes grupos. A continuación, se muestra una tabla con los p-valores obtenidos.

Premier League	La Liga	Serie A	Ligue 1	Bundesliga	Cinco ligas
0.00276	0.01143	0.00257	0.00376	0.00241	0.46291

Cuadro 2.1: P-valores resultantes en el Test de Kruskal-Wallis para GoalsFavor entre los cuatro grupos definidos en función de la clasificación de los equipos al final de la temporada dentro de cada liga y entre las cinco ligas.

Seguidamente, para comparar las medias de los equipos de cada grupo dentro de cada liga, se ha realizado la Prueba de Tukey. En la tabla del cuadro 2.2 se pueden ver los resultados para los goles a favor.

	Agrupaciones resultantes				
Grupos	Premier League	La Liga	Serie A	Ligue 1	Bundesliga
G1	a	a	a	a	a
G2	b	b	a, b	a, b	b
G3	b	b	b	b, c	b, c
G4	b	b	c	c	c

Cuadro 2.2: Grupos homogéneos resultantes sobre la variable GoalsFavor para los grupos definidos dentro de cada liga.

Tal y como ha quedado representado, tanto en la Premier como en la Liga, se aprecia que el Grupo 1 marca muchos goles por temporada en comparación con el resto de grupos, entre los que no se han encontrado diferencias significativas. En el resto de ligas, los grupos intermedios provocan que no haya dos agrupaciones claramente diferenciadas, ya que cada grupo no presenta diferencias significativas con su anterior o posterior, pero sí con el resto.

En segundo lugar, los tiros a puerta (ShotsOnTarget). En este caso, los valores se encuentran en un rango de 96 a 255. De igual forma que en el caso anterior, los boxplots reflejan que los equipos que más tiros a puerta realizan son los que mejor posición en su ranking logran al concluir la temporada.

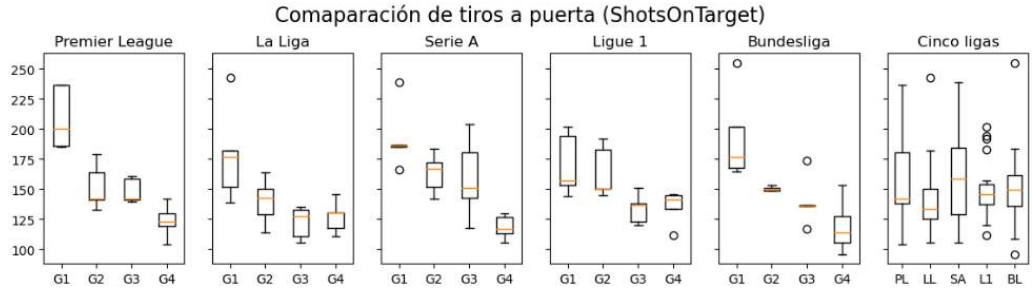

Figura 2.2: Boxplots de la variable ShotsOnTarget para comparar entre los cuatro grupos definidos dentro de cada liga (cinco primeros boxplots) y entre los equipos de las cinco ligas (último boxplot).

Después de llevar a cabo los seis Test de Kruskal-Wallis se ha obtenido que en las cinco ligas existen diferencias significativas entre los grupos definidos dentro de cada liga para la variable ShotsOnTarget. Mientras que, en los datos de la comparación de las cinco ligas no se han encontrado evidencias de la existencia de diferentes grupos. En el cuadro 2.3, se muestra

una tabla con los p-valores obtenidos.

Premier League	La Liga	Serie A	Ligue 1	Bundesliga	Cinco ligas
0.00260	0.01851	0.00641	0.01435	0.02035	0.22563

Cuadro 2.3: P-valores resultantes en el Test de Kruskal-Wallis para ShotsOnTarget entre los cuatro grupos definidos en función de la clasificación de los equipos al final de la temporada dentro de cada liga y entre las cinco ligas.

Posteriormente, se ha procedido con la realización de la Prueba de Tukey. En la tabla adjunta se pueden ver los resultados obtenidos para los tiros a puerta.

	Agrupaciones resultantes				
Grupos	Premier League	La Liga	Serie A	Ligue 1	Bundesliga
G1	a	a	a	a	a
G2	b	a, b	a, b	a, b	a, b
G3	b	b	a, b, c	b	b
G4	b	b	c	a, b	b

Cuadro 2.4: Grupos homogéneos resultantes sobre la variable ShotsOnTarget para los grupos definidos dentro de cada liga.

Analizando el cuadro 2.4, en la Premier League el Grupo 1 se encuentra muy diferenciado de el resto de grupos, entre los que apenas hay diferencias. En La Liga y en la Bundesliga, hay claramente dos agrupaciones resultantes. Una formada por el Grupo 1 y la otra por el Grupo 3 y 4, donde el Grupo 2 se situaría en la zona intermedia entre las dos agrupaciones. En el caso de la Serie A, hay muy pocas diferencias entre los grupos, ya que cada grupo se podría agrupar con su siguiente o anterior. Por último, en la Ligue 1, tampoco hay diferencias muy notables entre los grupos, pero es destacable el hecho de que los equipos del Grupo 4 hayan realizado más tiros a puerta que los equipos del Grupo 3.

En tercer lugar, los pases completado con éxito (PassesCompleted). Con un rango de valores que va de 9181 a 24406. En este caso, analizando la figura 2.3 se puede deducir lo mismo que en los boxplots anteriores. Es decir, los equipos que mayor cantidad de pases realizan son aquellos que mejor se clasifican. Pero, para esta variable, se puede observar la peculiaridad de que los equipos que se forman el Grupo 1 de cada liga tienen un rango de valores más amplio que el resto de grupos de la liga.

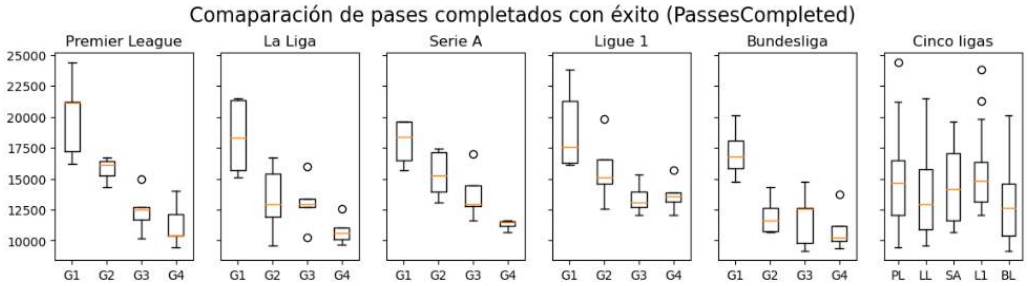

Figura 2.3: Boxplots de la variable PassesCompleted para comparar entre los cuatro grupos definidos dentro de cada liga (cinco primeros boxplots) y entre los equipos de las cinco ligas (último boxplot).

Una vez completados los seis Test de Kruskal-Wallis se ha obtenido que en las cinco ligas existen diferencias significativas entre los grupos internos para la variable PassesCompleted. En cambio, en los datos de la comparación de las cinco ligas no se han encontrado evidencias de la existencia de diferentes grupos. A continuación, se muestra una tabla con los p-valores obtenidos.

Premier League	La Liga	Serie A	Ligue 1	Bundesliga	Cinco ligas
0.00151	0.01317	0.00234	0.01574	0.02160	0.43680

Cuadro 2.5: P-valores resultantes en el Test de Kruskal-Wallis para PassesCompleted entre los cuatro grupos definidos en función de la clasificación de los equipos al final de la temporada dentro de cada liga y entre las cinco ligas.

A continuación, de igual forma que en los casos anteriores, se ha realizado la Prueba de Tukey. En la tabla adjunta se pueden ver los resultados para los pases completados con éxito.

	Agrupaciones resultantes				
Grupos	Premier League	La Liga	Serie A	Ligue 1	Bundesliga
G1	a	a	a	a	a
G2	b	b	a, b	a, b	b
G3	b, c	b	b, c	b	b
G4	c	b	c	b	b

Cuadro 2.6: Grupos homogéneos resultantes sobre la variable PassesCompleted para los grupos definidos dentro de cada liga.

Se puede deducir, interpretando los resultados obtenidos, que en todas las ligas el Grupo 1 se encuentra diferenciado del resto. En el caso de la Premier League y la Ligue 1, entre los dos últimos grupos no se han encontrado diferencias significativas. Y, tanto en la Serie A cómo en

la Ligue 1, en el Grupo 2 no se han encontrado diferencias significativas con respecto al Grupo 1 y 3.

Por último, en cuarto lugar, las faltas realizadas (FoultsCommited). Con un valor mínimo de 305, conseguido por el Bayern de Munich de la Bundesliga, y un valor máximo de 641, conseguido por el Valencia de La Liga.

Observando los boxplots generados para esta variable, se aprecia que, por norma general los equipos que menos faltas realizan son los que más alto ranking ocupan en la clasificación final de la temporada. Además, en el último boxplot, se han obtenido unas medias aparentemente desiguales, comprobemos mediante los test si para esta variable es posible encontrar diferencias significativas entre las medias de las cinco ligas.

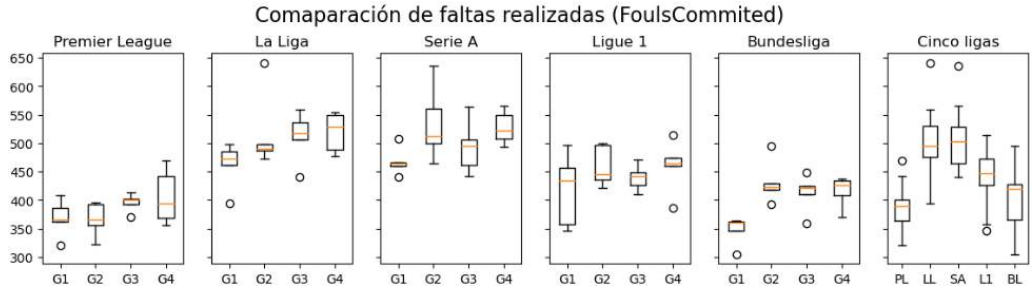

Figura 2.4: Boxplots de la variable FoulsCommited para comparar entre los cuatro grupos definidos dentro de cada liga (cinco primeros boxplots) y entre los equipos de las cinco ligas (último boxplot).

Tras ejecutar los seis Test de Kruskal-Wallis, a diferencia de los casos anteriores, en este se ha obtenido que dentro de las cinco ligas no existen diferencias significativas entre los cuatro grupos de cada una. En cambio, en los datos de la comparación de las cinco ligas sí se han encontrado evidencias de la existencia de diferencias entre las medias. A continuación se muestra una tabla con los p-valores obtenidos en el test.

Premier League	La Liga	Serie A	Ligue 1	Bundesliga	Cinco ligas
0.22028	0.16603	0.09647	0.43587	0.06085	7.4948e-10

Cuadro 2.7: P-valores resultantes en el Test de Kruskal-Wallis para FoulsCommited entre los cuatro grupos definidos en función de la clasificación de los equipos al final de la temporada dentro de cada liga y entre las cinco ligas.

Seguidamente, se ha realizado la Prueba de Tukey. En la tabla del cuadro 2.8 se pueden ver los resultados para las faltas realizadas, donde se han evaluado las posibles agrupaciones entre las cinco ligas, y no de forma interna a cada liga como en los casos anteriores.

	Premier League	La Liga	Serie A	Ligue 1	Bundesliga
Agrupaciones	a	b	b	c	a, c

Cuadro 2.8: Grupos homogéneos resultantes sobre la variable FoultsCommited para encontrar posibles agrupaciones entre las cinco ligas

En el recuadro 2.8, se refleja que, las ligas entre las que no se han encontrado diferencias significativas en cuanto a las faltas cometidas se corresponden con las siguientes agrupaciones: Premier League - Bundesliga, Bundesliga - Ligue 1 y La Liga - Serie A. Observando el boxplot de la figura 2.4, se aprecia que la Premier League es la liga en la que menos faltas se comenten. De forma contraria, la Serie A y La Liga son las ligas en las que más faltas se realizan. Además, se puede ver en el boxplot que en todas las ligas, los equipos que menos faltas realizan son los pertenecientes al Grupo 1 en cada liga.

Para la segunda fase de la sección, en primer lugar se ha generado el mapa de correlaciones para poder identificar aquellas variables que estén más relacionadas entre sí. Gracias al mapa de correlaciones, también es posible cuantificar estas relaciones ya que la intensidad del color de los recuadros refleja la dependencia existente entre las variables.

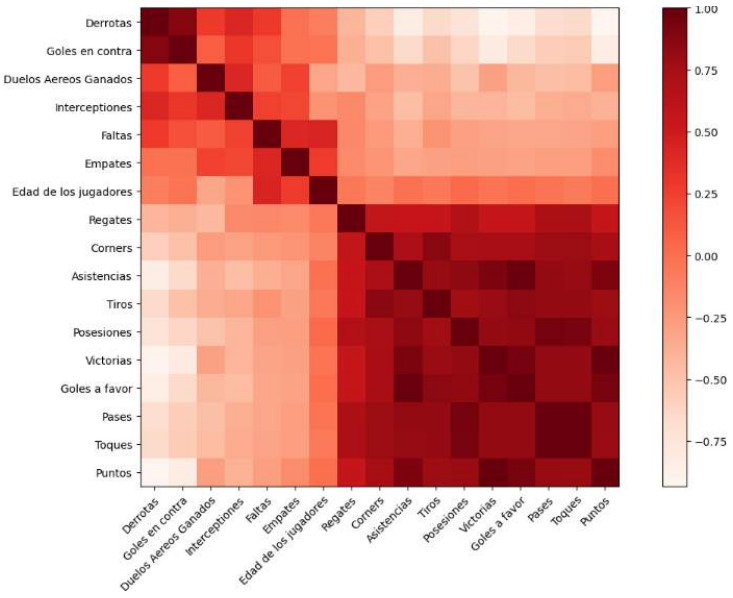

Figura 2.5: Mapa de correlaciones

Como se puede ver, hay tres grupos de dependencia directa bastante claros. El primero, asociado a las derrotas, que refleja que las derrotas en una temporada, están muy relacionadas con el número de goles en contra, duelos aéreos ganados, interceptaciones y faltas realizadas.

En segundo lugar, los empates, están relacionados con las faltas realizadas y la edad de los jugadores. Por último, el tercer grupo, es el que tiene las relaciones de dependencia más altas y el que está asociado a las victorias, donde se comprueba que están muy relacionadas con la cantidad de toques, pases, goles a favor, posesiones, tiros, asistencias, córners y regates. En cuanto a la dependencia inversa entre variables, se ha obtenido que las variables derrotas y goles en contra tienen una clara dependencia inversa con los puntos, goles a favor, victorias, toques, pases asistencias y tiros.

Seguidamente se ha implementado el PCA correspondiente. A continuación se puede ver el gráfico de sedimentación generado. Teniendo en cuenta que el objetivo es reducir la muestra a dos variables para poder graficar todos los equipos en un plano. Resultará determinante de este gráfico poder saber cuanta información es posible explicar únicamente empleando las dos primeras componentes principales. Como se puede observar en la figura 2.6, con las dos primeras componentes se estaría explicando el 68.77 % de la información total.

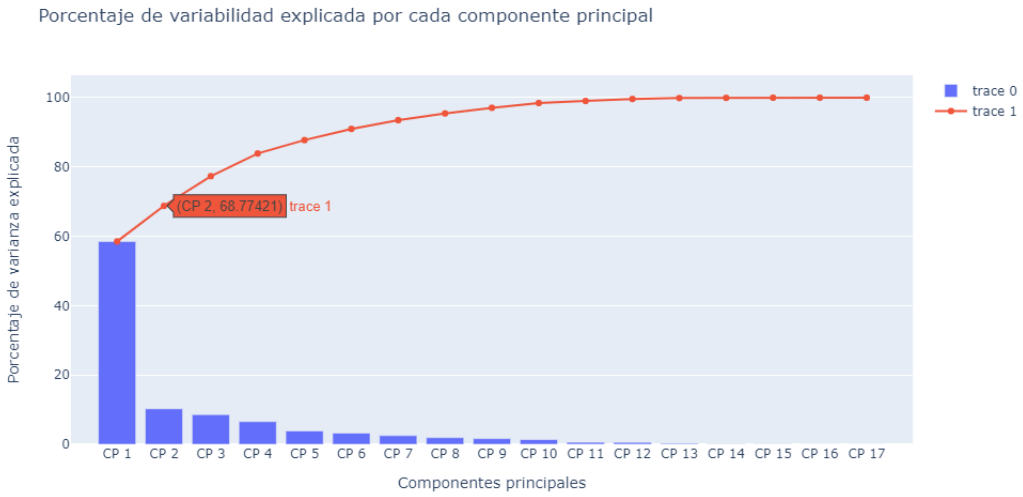

Figura 2.6: Gráfico de sedimentación de las componentes principales.

Para entender mejor cómo se han formado cada una de las dos componentes principales, en la siguiente tabla se puede ver los pesos que, para cada componente, se ha dado a cada una de las variables. En función de los pesos obtenidos, la componente CP1 ayudará a diferenciar los equipos en función de las variables: Puntos, Victorias, Derrotas, Goles a favor, Goles en contra, Asistencias, Tiros, Pases, Regates, Toques, Posesiones, Córners y Duelos aéreos ganados. Mientras que la componente CP2 ayudará a diferenciar en función de las variables: Empates, Faltas y Edad de los jugadores.

Pesos de las variables					
Variables	CP1	CP2	Variables	CP1	CP2
Puntos	0.2992	-0.0838	Regates	0.2116	0.0242
Victorias	0.3023	-0.0243	Toques	0.2919	0.0585
Empates	-0.0971	-0.481	Faltas	-0.1154	-0.4946
Derrotas	-0.2714	0.1989	Intercepciones	-0.1511	0.1164
Goles a favor	0.3017	-0.0169	Posesiones	0.293	-0.0143
Goles en contra	-0.231	0.1977	Edad de los jugadores	-0.0065	-0.6225
Asistencias	0.2984	0.0093	Córners	0.2589	0.0869
Tiros	0.2748	0.0462	Duelos Aéreos Ganados	-0.1454	0.1437
Pases	0.2947	0.0403			

Cuadro 2.9: Pesos de las dos primeras componentes principales resultantes del PCA

A continuación, se ha representado gráficamente todos los equipos de las cinco ligas en función de las dos primeras componentes principales. Donde, cada liga es representada por un color diferente. El símbolo que representa cada equipo se ha utilizado para diferenciar: los equipos que durante las cinco temporadas han jugado al menos una vez las competiciones europeas Champions League o Europa League (triángulo hacia arriba) , los que durante las cinco temporadas se han mantenido siempre a mitad tabla (círculo) y los que en alguna de las cinco temporadas han descendido o jugado el playoff de descenso (triángulo hacía abajo). Los equipos Schalke 04, Wolfsburg, Saint-Étienne y Granada tienen la particularidad de que durante las cinco temporadas han estado una vez en competiciones europeas y otra en descenso. Por lo tanto, se ha decidido representarlos mediante el círculo.

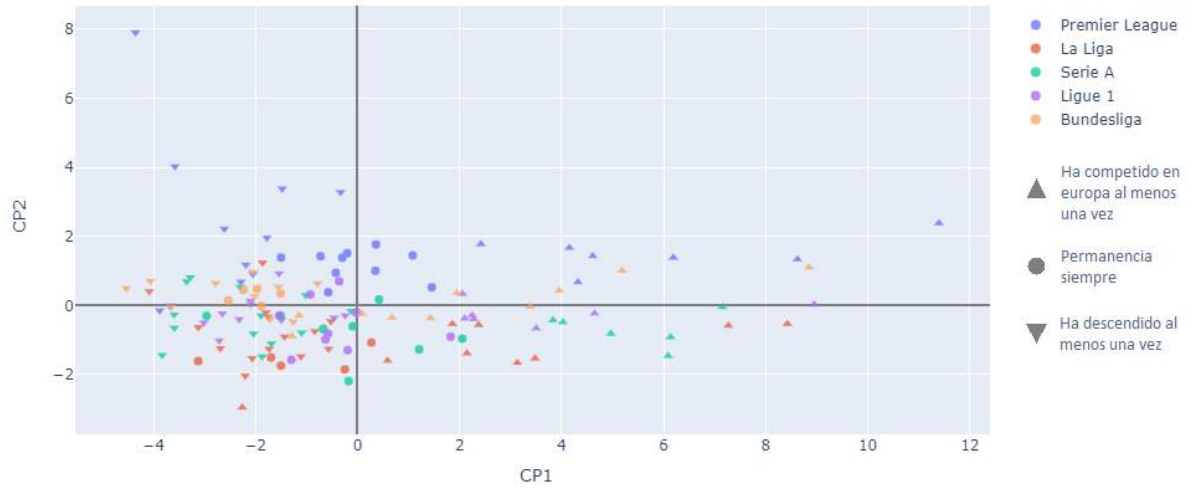

Figura 2.7: Gráfico con los equipos en función del nuevo espacio bidimensional dado por el PCA, donde se diferencia a los equipos por liga (color) y por participaciones en competiciones europeas o promociones de descenso (forma).

Por un lado, como se puede observar en el gráfico, la componente CP1 permite diferenciar a los equipos en función de su clasificación dentro de cada liga. Ya que, en la parte positiva del eje de abscisas se han representado la mayor parte de los equipos que han jugado competiciones europeas y en la parte negativa se han quedado los equipos que han descendido en alguna temporada. Interpretando este eje, se puede ver que las ligas Premier League, La Liga, Serie A y Bundesliga tienen una aparente cantidad de equipos muy parecida en cada parte del ranking, pese a ser la Premier la que más destaca con sus equipos. En cambio, la Ligue 1 tiene un equipo a la altura de la élite del resto de ligas, que se trata del Paris S-G, pero el resto de equipos son muy inferiores en comparación con el resto de ligas. Por otro lado, la componente CP2 ayuda a diferenciar entre las diferentes ligas, ya que la Premier League y la Bundesliga han sido representadas en la parte positiva del eje de abscisas mientras que el resto de ligas tienen la mayoría de sus equipos en la parte negativa de este eje.

Teniendo en cuenta los resultados obtenidos en las dos fases de esta sección. En la primera fase dimos con que las faltas cometidas (FoultsCommited) no permite diferencia a los equipos en función de su ranking dentro de una misma liga, pero sí que dimos con que esta variable era útil para diferenciar entre diferentes ligas. De hecho, vimos que en la Premier y en la Bundesliga se realizaban una cantidad de faltas similar y, a la vez, inferior al del resto de ligas. De acuerdo con esto, en el PCA hemos vuelto a obtener que esa variable es útil para separar entre diferentes ligas. Donde hemos visto que las la Premier y Bundesliga tienen valores similares para la CP2 y diferentes al resto de ligas. En este caso valores superiores, ya que el peso que en la CP2 se le ha dado a dicha variable es negativo.

2.2. Regresión logística

2.2.1. Introducción

En esta sección lo que se pretende es averiguar aquellos indicadores cuantificables del juego en la temporada de un equipo de cualquiera de las cinco ligas que están relacionados con el estilo de juego y resultan determinantes en que dicho equipo acabe la temporada clasificándose para jugar competiciones europeas.

Los datos empleados para esta sección se corresponden con los 490 registros correspondientes con las cinco últimas temporadas finalizadas en las cinco grandes ligas de fútbol europeas. Para poder disponer de estos datos, ha sido necesario realizar un proceso de limpieza para seleccionar aquellas variables que permiten reflejar el estilo de juego de un equipo, y no están tan relacionadas aparentemente con el resultado final como lo es el caso de los puntos, victorias, empates, goles a favor o goles en contra entre otras. Además, se ha añadido una variable binaria que indica si al final de la temporada dicho equipo clasificó para competiciones europeas (Campions

League o Europa League) en su correspondiente liga. Las variables empleadas para esto han sido: posesión, penaltis realizados, tarjetas amarillas, tarjetas rojas, paradas, tiros a puerta, distancia media de los tiros, pases cortos, pases medios, pases largos, pases en movimiento, pases a balón parado, pases de tiro libre, pases entre defensas, pases a lo ancho del campo, pases desde centros, pases de saque de banda, pases de córner, regates intentados, toques en la zona 1, toques en la zona 2, toques en la zona 3, toques en la zona 4, toques en la zona 5, duelos aéreos ganados, faltas realizadas, fueras de juego y interceptaciones.

Como parte del proceso del análisis, se empezará generando un mapa de correlaciones y el gráfico bidimensional resultante de aplicar el algoritmo de PCA con las variables de esta sección. Para completar el análisis se realizarán diferentes implementaciones de un algoritmo de regresión logística para así poder analizar los gráficos de odds ratios resultantes. La primera implementación se realizará empleando todas las variables. La segunda, con las variables seleccionadas mediante un método de selección de variables. Y, la tercera, con las componentes principales resultantes del PCA implementado.

2.2.2. Fundamentos teóricos

La regresión logística se trata de una técnica de gran utilidad para la predicción de eventos binarios en función de un conjunto de variables predictoras. En esta sección se explicarán los conceptos y métodos clave para su implementación como son: la probabilidad condicional, el ratio de probabilidades, la función logarítmica, el método de la máxima verosimilitud y el método de Newton-Raphson. Además, también se detallarán los fundamentos teóricos del método Stepwise y el concepto de pseudo-R cuadrado. La teoría de los apartados "Función de probabilidad" y "Método de máxima verosimilitud" de esta sección han sido realizados consultando la referencia [12].

Probabilidad condicional

Es importante recordar que la probabilidad de que se de un suceso X, teniendo en cuenta que se ha dado un suceso Y, se representa de la siguiente forma:

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

Ratio de probabilidades

Primero definimos los dos siguiente términos:

- $P_X \Rightarrow$ Probabilidad de éxito cumpliéndose la condición X.
- $P_Y \Rightarrow$ Probabilidad de éxito cumpliéndose la condición Y.

Teniendo en cuenta lo anterior, el ratio de probabilidades (odds)[1] se trata del cociente entre los casos de éxito sobre los casos de fracaso en el suceso estudiado y para cada grupo. Dicho indicador puede estar en un rango de valores entre 0 y ∞. Y, se representa de la siguiente forma:

$$\left.\begin{array}{l} odds_{éxito,X} = \dfrac{P_X}{1-P_X} \\ odds_{éxito,Y} = \dfrac{P_Y}{1-P_Y} \end{array}\right\} \Rightarrow \; odds_{ratio} = \frac{odds_{éxito,X}}{odds_{éxito,Y}} \text{ con } odds \in [0, +\infty)$$

Otra medida que será necesaria tener en cuenta es el intervalo de confianza (IC). Este intervalo refleja la precisión del odd ratio asociado. Cuanto más grande sea el intervalo de confianza, menos precisa será la estimación del odd ratio obtenido. Más detalladamente, el IC refleja la variación de resultado posible en la estimación del odd ratio. Por ejemplo, si el intervalo de confianza es [1.5, 1.7] para un odd ratio de 1.6, significará que de 100 hipotéticos estudios realizados, en 95 de ellos se obtendrán valores dentro del intervalo. Mientras que, en los 5 estudios restantes se obtendrán valores fuera del intervalo.

Para poder interpretar los resultados obtenidos en las dos medidas anteriores, será necesario tener en cuenta que se trata de un cociente de sucesos. Por tanto, los intervalos de confianza que contengan el valor 1 reflejarán que dicho parámetro o variable no tiene significancia estadística. Es más, los intervalos superiores a 1 indicarán una asociación positiva entre las variables (y serán consideradas como factores de riesgo). Y, los intervalos de confianza inferiores a 1 reflejarán una asociación negativa entre las variables (y serán consideradas como factores de contención).[5]

Función de probabilidad

Para poder implementar este modelo, será necesario conocer la función de probabilidad. Se partirá de una serie de observaciones $\{(x_i, y_i)\}_{i=1}^{n}$. Donde $x_i = (x_{i1},\ x_{i2},\ ...,\ x_{ik})$, siendo k el número de variables predictoras. Cada x_{ij} representa el dato numérico en la posición j-ésima asociado a la variable i-ésima, que será usado para predecir la variable binaria del resultado y_i. Al ser binaria podrá tomar los valores $y_i \in \{0, 1\}$. Y, n hace referencia a la cantidad de datos de los que se dispondrá.

Teniendo en cuenta la notación empleada, se tratará de definir una función que permita obtener el valor de $P(y_i = 1)$ a partir de las variables predictoras $x_{i1}, ..., x_{ik}$. Por tanto, la función de probabilidad se podrá definir del siguiente modo:

$$P(y_i = 1) = \frac{1}{1 + e^{-(\alpha+\beta_1 x_{i1}+...+\beta_k x_{ik})}}$$

De forma simplificada, también se podrá expresar por:

$$P(y_i = 1) = \frac{1}{1 + e^{-(\alpha+\beta' x_i)}} \text{ considerando } \beta' = (\beta_1, ..., \beta_k)$$

Gracias a esta función, será posible analizar que variables influyen en el resultado y la calidad del modelo implementado. Además, a partir de la fórmula recién definida, se puede calcular:

$$1 - P(y_i = 1) = \frac{e^{-(\alpha+\beta' x_i)}}{1 + e^{-(\alpha+\beta' x_i)}} = \frac{1}{1 + e^{\alpha+\beta' x_i}}$$

Por último, al tratarse de una probabilidad, es decir, $P(y_i = 1) \in [0, 1]$. Usando la notación $P_i = P(y_i = 1)$, aplicando el logaritmo también se podrá definir la fórmula:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta' \cdot x_i \quad \text{donde} \quad \begin{cases} \frac{P_i}{1 - P_i} \in [0, 1] \Rightarrow ln\left(\frac{P_i}{1 - P_i}\right) \in (-\infty, 0] \\ \frac{P_i}{1 - P_i} \in [1, \infty) \Rightarrow ln\left(\frac{P_i}{1 - P_i}\right) \in [0, \infty) \end{cases}$$

Gráficamente, la función logística definida en este apartado tendrá el siguiente aspecto.

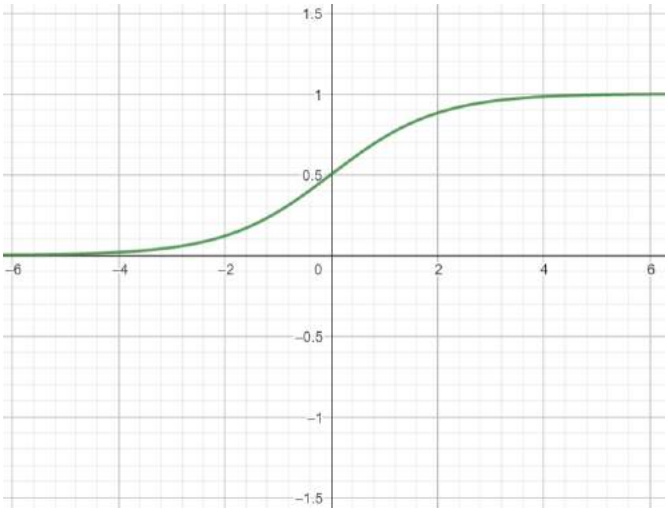

Figura 2.8: Gráfica con la función logística para el modelo de regresión creada mediante Geogebra.

Método de máxima verosimilitud

Este método se usa para estimar los parámetros de la regresión logística. Teniendo en cuenta todas estas variables nombradas y siguiendo con la notación empleada anteriormente, se define el entorno como la probabilidad conjunta:

$$L = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i} \tag{2.7}$$

Para desarrollar este entorno, primero se definen los siguientes resultados que serán necesarios en el transcurso:

I) $ln(\frac{P_i}{1-P_i}) = \alpha + \sum_{j=1}^{k} \beta_j x_{ij} = \alpha + \beta' x_i$

II) $P_i = \frac{e^{\alpha+\beta x_i}}{1+e^{-(\alpha+\beta' x_i)}} = \frac{1}{1+e^{-(\alpha+\beta' x_i)}}$

III) $1 - P_i = 1 - \frac{1}{1+e^{-(\alpha+\beta' x_i)}} = \frac{1+e^{-(\alpha+\beta' x_i)}-1}{1+e^{-(\alpha+\beta' x_i)}} = \frac{1}{1+e^{\alpha+\beta' x_i}}$

IV) $ln(1 - P_i) = ln(\frac{1}{1+e^{\alpha+\beta' x_i}}) = ln(1) - ln(1 + e^{\alpha+\beta' x_i}) = -ln(1 + e^{\alpha+\beta' x_i})$

V) $\frac{\partial P_i}{\partial \beta_j} = \frac{-e^{-(\alpha+\beta' x_i)}}{(1+e^{-(\alpha+\beta' x_i)})^2} = -\frac{1}{1+e^{-(\alpha+\beta' x_i)}} \frac{e^{-(\alpha+\beta' x_i)}}{1+e^{-(\alpha+\beta' x_i)}} x_{ij} = -P_i(1-P_i)x_{ij}$

Como el nombre del método indica, lo que se busca es la máxima verosimilitud. Luego, se tratará de encontrar $\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial \beta_j} = 0$. Para ello se deberán calcular las dos respectivas derivadas:

$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} -\frac{e^{\alpha+\beta' x_i}}{1+e^{\alpha+\beta' x_i}} + y_i = \sum_{i=1}^{n}(y_i - P_i)$

$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} -\frac{e^{\alpha+\beta' x_i}}{1+e^{\alpha+\beta' x_i}} x_{ij} + y_i x_{ij} = \sum_{i=1}^{n} x_{ij}(y_i - P_i)$

Definiendo, $\beta_0 = \alpha$ y $x_{i0} = 1 \ \forall i = 1, ..., n$ se puede definir y simplificar (primero usando el producto escalar y luego la notación vectorial) una fórmula general para todas las derivadas de la siguiente forma:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \cdot (y_i - P_i) = x_j * (y_i - P_i) \Rightarrow \begin{pmatrix} \frac{\partial l}{\partial \beta_0} \\ \vdots \\ \frac{\partial l}{\partial \beta_k} \end{pmatrix} = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} x_i \cdot (y_i - P_i)$$

A continuación se aplicará al entorno definido (3.7) el logaritmo neperiano y los resultados enumerados previamente tal y como sigue:

$$\begin{aligned} l = ln(L(\alpha, \beta')) = ln(\prod_{i=1}^{n} P_i^{y_i}(1-P_i)^{1-y_i}) = \sum_{i=1}^{n} y_i ln(P_i) + (1-y_i) ln(1-P_i) = \\ = \sum_{i=1}^{n} ln(1-P_i) + y_i(ln(P_i) - ln(1-P_i)) \overset{(I)}{=} \sum_{i=1}^{n} ln(1-P_i) + y_i(\alpha + \sum_{j=1}^{k} \beta_j x_{ij}) = \\ \overset{(IV)}{=} \sum_{i=1}^{n} -ln(1+e^{\alpha+\beta' x_i}) + y_i(\alpha + \sum_{j=1}^{k} \beta_j x_{ij}) \end{aligned}$$

Se calcula la segunda derivada:

$$\frac{\partial^2 l}{\partial \beta_j^2} = \sum_{i=1}^{n} -x_{ij} \frac{\partial P_i}{\partial \beta_k} = \sum_{i=1}^{n} x_{ij} P_i(1-P_i) x_{ij}$$

Simplificando en notación vectorial se llega a:

$$\frac{\partial^2 l}{\partial \beta'^2} = \sum_{i=1}^{n} x_i P_i(1-P_i)x_i^t = x \cdot W(\beta') \cdot x^t \text{ donde, } W(\beta') = diag(P_i(1-P_i))_{i=1}^n$$

El motivo de haber calculado la segunda derivada es porque la primera derivada no necesariamente será analítica. Y, para resolver la ecuación se empleará el método de Newton-Raphson.

Método de Newton-Raphson

El método de Newton-Raphson será de gran utilidad para encontrar los valores óptimos de los parámetros que maximizan la función de verosimilitud. El método consiste en un algoritmo para hallar las raíces de una función no lineal. Su metodología se basa en aproximar la función por una recta tangente a la curva en cada iteración e ir dando con la intersección de esta recta con el eje de abscisas. Para poder aplicar este método es necesario partir de una función continua y diferenciable con derivada también continua y diferenciable en el intervalo donde se busca la raíz. Se parte de un punto inicial β_0 y se repite el proceso hasta lograr la precisión deseada.

$$\beta_{n+1} = \beta_n - \frac{f(\beta_n)}{f'(\beta_n)} \rightarrow \triangle\ \beta = -\frac{f(\beta_n)}{f'(\beta_n)} \text{ donde, en nuestro caso tendremos:}$$

$$\left.\begin{aligned} f(\beta) &= \frac{\partial l}{\partial \beta} = x \cdot (y - P(\beta)) \\ f'(\beta) &= \frac{\partial^2 l}{\partial \beta^2} = x \cdot W(\beta) \cdot x^t \end{aligned}\right\} \Rightarrow\ \triangle\ \beta = (xW(\beta)x^t)^{-1} \cdot (x(y-P))$$

En nuestro caso el punto inicial será $\beta_0 = \vec{0}$ y se deberá iterar hasta dar con una aproximación de los parámetros de la regresión logística.

Más detalladamente, los pasos que se deberán seguir son los siguientes. En primer lugar, se deberá definir la función de verosimilitud L. En segundo lugar, se deberá definir la probabilidad de cada observación de los P_i. En tercer lugar, se deberá calcular la matriz diagonal $W(\beta)$. Seguidamente, se definirá la función logística. Y, finalmente, se usará el método de Newton-Raphson para resolver el problema.

Algoritmos Stepwise para la selección de variables

Los algoritmos de selección de variables resultan de gran utilidad para mejorar el rendimiento y la eficiencia de los modelos. Las razones principales por las que se usan estos algoritmos son: la eliminación de las variables menos relevantes, la mejora en el rendimiento del modelo, el ahorro de recursos computacionales y la información aportada a la hora de interpretar los resultados de un modelo. A pesar de las razones anteriores, es importante tener en cuenta que no todos los algoritmos de selección de variables son adecuados para todos los modelos. Y, los resultados obtenidos es conveniente que sean evaluados y validados con un cierto criterio. En el caso de la regresión logística, el método Stepwise es apropiado debido a su secuencialidad. [8] [7]

- **Método Forward**

En este método se parte de un modelo sin variables y en cada iteración se introduce la variable más significativa que aún no esté introducida en el modelo. Este procedimiento iterativo se realiza hasta satisfacer alguna regla de parada.

- **Método Backward**

En este método se parte de un modelo con todas las variables incluidas y en cada iteración se elimina la variable menos significativa hasta que no proceda eliminar ninguna variable más. Este método se podría decir que es el contrario al método forward.

- **Método Stepwise**

Este método se trata de una combinación de los dos métodos anteriores. Se comienza con un modelo sin variables (como en el método forward) y en cada iteración se van introduciendo nuevas variables en función de su nivel de significación. Además en cada iteración, una variable que ha sido introducida en una iteración anterior puede ser eliminada en función de su influencia para el modelo (como en el método backward).

En el caso de R existen funciones ya creadas que hacen mucho más fácil emplear el algoritmo y iterar hasta encontrar la cantidad de variables óptima, gracias a la información que aportan indicadores como AIC o BIC. En el caso de Python no existen aún estas funciones. Para poder emplear este algoritmo se ha programado un método que seleccionará las variables más influyentes en función de los p-valores de cada una.

Concepto de Pseudo-R Squared

En la regresión lineal, para evaluar la bondad de ajuste, es decir lo bien que se ajusta un modelo a los datos, se emplea el coeficiente de determinación R^2. Ya que, representa la proporción de varianza de los resultados que puede ser explicada por el modelo. En cambio, cuando la variable resultado es categórica no puede usarse este coeficiente como medida, y para evaluar la bondad de ajuste se emplea el pseudo-R cuadrado.

El concepto de pseudo-R [11] cuadrado se utiliza en modelos de regresión que tienen variables categóricas como variables objetivo. Este indicador proporciona una medida de la bondad de ajuste del modelo en relación con las variables categóricas. En otras palabras, el pseudo-R cuadrado proporciona una medida de la mejora en el ajuste del modelo en comparación con el modelo de referencia. El rango de posibles valores va de 0 a 1. Cuando más cercano a 1 sea quedará decir que mejor se ajusta el modelo en comparación con el modelo nulo.

El método que se usará para calcular este indicador se conoce como pseudo-R cuadrado de McFadden, que la fórmula para calcularlos se corresponde con la siguiente:

$$R^2_{McF} = 1 - \frac{\log(L_M)}{\log(L_0)} \text{ donde } \begin{cases} L_M \text{ representa la función de verosimilitud del modelo ajustada} \\ L_0 \text{ representa la función de verosimilitud del modelo nulo} \end{cases}$$

2.2.3. Resultados obtenidos

En primer lugar, lo que se ha realizado en esta sección es generar un mapa de correlaciones para poder analizar las relaciones entre las variables usadas. Ya que no son las mismas que en la sección anterior. Además se ha añadido una variable binaria llamada "Europe"que indica con 0 o 1 si cada equipo acabó la temporada clasificando para competiciones europeas o no (Campions League o Europa League), de este modo resulta más fácil ver las relaciones lineales de ésta variable con el resto. Esto se corresponde con la figura 2.9

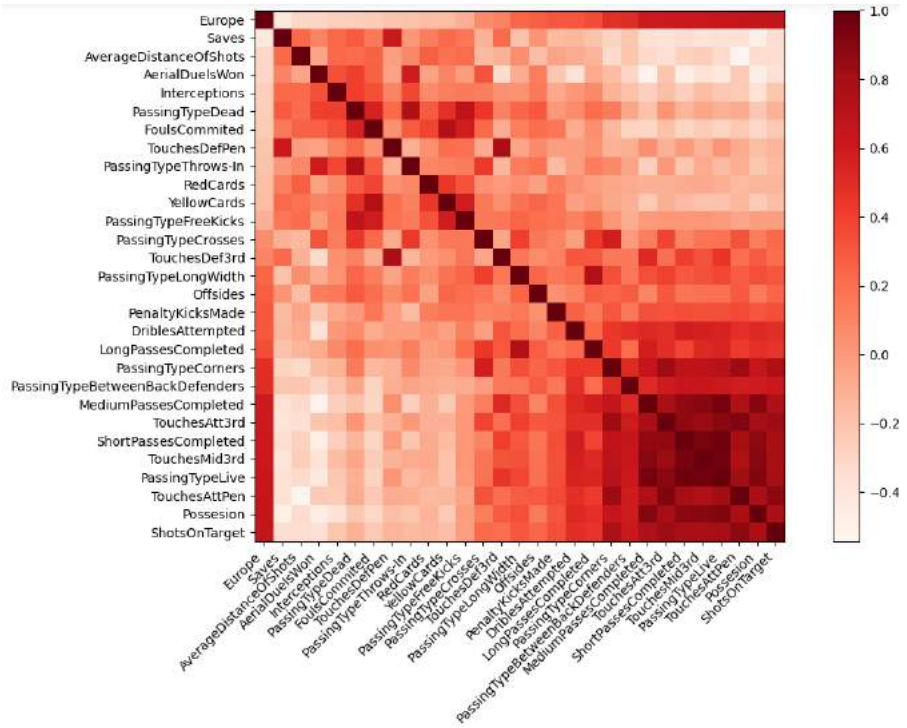

Figura 2.9: Mapa de correlaciones

Analizando los resultados, vemos que la variable que indica si un equipo va a europa tiene una relación directa con variables que reflejan un estilo de juego ofensivo y dominante como son: tiros a puerta (ShotsOnTarget), posesion (Possesion), toques en la zona 5 (TouchesAttPen), pases en movimiento (PassingTypeLive), toques en la zona 3 (TouchesMid3rd), pases cortos completados (ShortPassesCompleted), toques en la zona 4 (TouchesAtt3rd) y pases medios completados (MediumPassesCompleted). También es importante tener en cuenta que la variable "Europe"tiene una fuerte relación inversa con las variables: paradas (Saves), distancia media de los tiros (AverageDistanceOfShots) y los duelos aéreos ganados (AverageDistanceOfShots).

Gracias a los resultados obtenidos, es interesante el hecho de que las variables que tienen una alta relación directa con la variable que estamos estudiado, también tienen una alta relación directa en entre ellas, esto queda reflejado por el recuadro más oscuro que se forma en la parte inferior derecha del mapa de correlaciones. Y, es más, las variables que tienen una alta relación directa con "Europe"tienen una alta relación inversa con las mismas variables con las que dicha variable tiene una relación inversa.

Para continuar con el análisis, se ha usado el algoritmo PCA visto anteriormente. En este caso, al haber más variables, la cantidad de información que se puede llegar a explicar única-

mente con las dos componentes principales se corresponde con el 50.25 %, tal y como se aprecia en el gráfico de la figura 2.10. Cabe destacar que este porcentaje es muy inferior a lo que podría considerarse apropiado, que sería de entorno al 70 % o superior. Esto se debe a la cantidad de variables que se está intentando reducir en este caso.

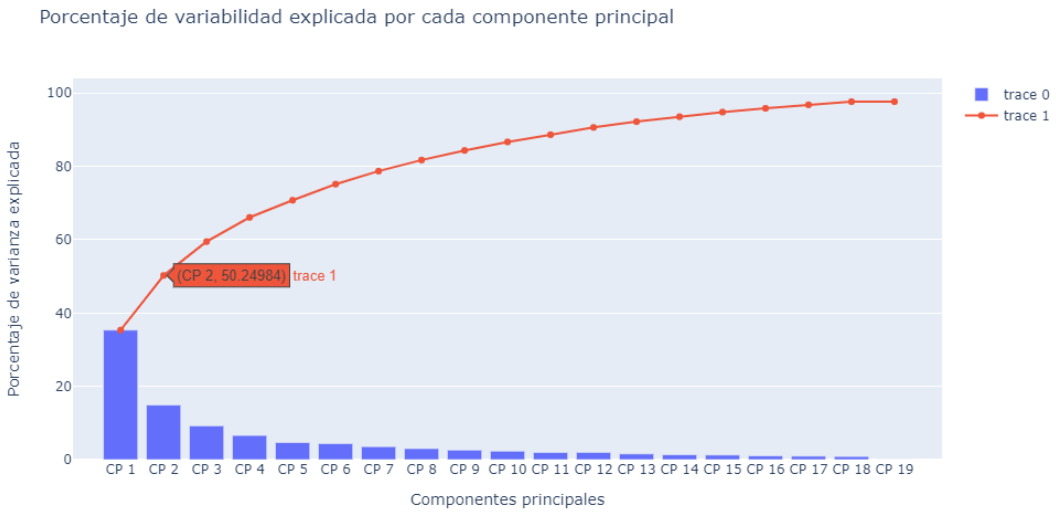

Figura 2.10: Gráfico de sedimentación de las componentes principales.

Tal y como se puede apreciar en la tabla del cuadro 2.10, en la componente CP1 resultante se le da un peso mayor a las variables: pases cortos completados (ShortPassesCompleted), pases medios completados (MediumPassesCompleted), pases en movimiento (PassingTypeLive), córners (PassingTypeCorner), toques en la zona 3 (TouchesMid3rd), toques en la zona 4 (TouchesAtt3rd) y toques en la zona 5 (TouchesAttPen). En cambio, en la componente CP2 se le da un mayor peso a las variables: tarjetas amarillas (YellowCards), pases largos completados (LongPassesCompleted), pases desde tiros libres (PassingTypeFreeKicks), saques de banda (PassingTypeThrows-In), faltas (FoulsCommited), fueras de juego (Offsides) y interceptaciones (Interceptions).

Pesos de las variables					
Variables	CP1	CP2	Variables	CP1	CP2
Posesión	-0.2949	-0.0446	Pases a lo ancho del campo	-0.1270	0.2290
Penaltis realizados	-0.1172	0.0842	Pases cruzados	-0.0732	0.2770
Tarjetas amarillas	0.0692	0.3215	Saques de banda	0.0563	0.2865
Tarjetas rojas	0.0474	0.1834	Córners	-0.2510	0.1197
Paradas	0.139	0.0924	Regates intentados	-0.1876	0.0541
Tiros a puerta	-0.2800	0.0115	Toques en la zona 1	0.0286	0.0950
Distancia media de los tiros	0.1261	0.083	Toques en la zona 2	-0.1108	0.0726
Pases cortos completados	-0.2963	-0.0108	Toques en la zona 3	-0.2949	0.0238
Pases medios completados	-0.2933	-0.0242	Toques en la zona 4	-0.2898	0.0799
Pases largos completados	-0.1749	0.2077	Toques en la zona 5	-0.2843	0.0273
Pases en movimiento	-0.3078	0.0136	Duelos aéreos ganados	0.1419	0.1637
Pases desde balón parado	0.0485	0.435	Faltas	0.1013	0.3498
Pases desde tiros libres	0.0121	0.3416	Fueras de juego	-0.0717	0.2078
Pases entre defensas	-0.2131	-0.0205	Interceptaciones	0.0905	0.2123

Cuadro 2.10: Pesos de las dos primeras componentes principales resultantes del PCA

La representación bidimensional de todas las temporadas de todos los equipos daría lugar a la gráfica de la figura 2.11. Se han representado con un triángulo hacia arriba todas las temporadas de los equipos que acabaron clasificándose para competiciones europeas y con un círculo el resto.

Figura 2.11: Gráfico bidimensional de los 490 datos correspondientes a las cinco temporadas registradas, donde se ha diferenciado a cada registro por la liga a la que pertenece (color) y por haber o no acabado quedado en puestos clasificatorios para competiciones europeas (forma).

Analizando el gráfico, se puede ver que las variables a las cuales la componente CP1 les da más valor tienen una gran influencia en el rendimiento del equipo, y por ende, su clasificación para competiciones europeas. En cambio, mediante la variables de la componente CP2, no es posible sacar conclusiones a partir de este gráfico acerca de su influencia en la clasificación para competiciones europeas. Pero, del mismo modo que resultó en el PCA correspondiente con el

gráfico de la figura 2.7, las variables de la componente CP2 son de gran ayuda para diferenciar entre ligas.

Seguidamente, tal y como se corresponde en esta sección, se ha implementado el modelo de regresión logística para las 28 variables de los 490 registros de temporadas de las cinco ligas durante los cinco años registrados. Donde se ha obtenido un Pseuedo-R cuadrado de 0,635. El motivo de esta implementación es poder analizar la influencia de las variables en que un equipo se clasifique para competiciones europeas en su correspondiente liga, que se corresponde con el gráfico de odds ratios con sus respectivos intervalos de confianza de la figura 2.12

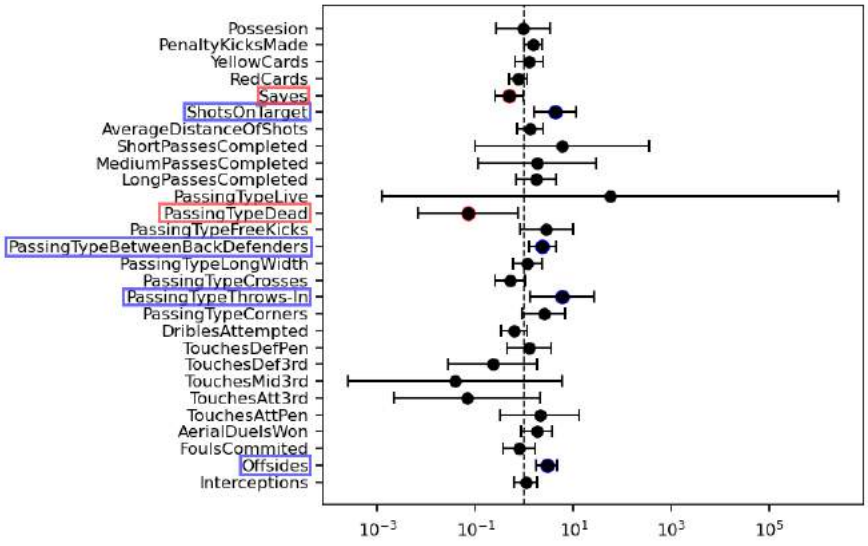

Figura 2.12: Gráfico de odds ratios con escala logarítmica y sus correspondientes intervalos de confianza. Donde se han destacado las variables de contención (rojo) y las de riesgo (azul).

Los resultados obtenidos son, por un lado, que las variables tiros a puerta (ShotsOnTarget), pases entre los defensas (PassingTypeBetweenBackDefenders), pases de saque de banda (PassingTypeThrows-In) y los fueras de juego (Offsides) influyen positivamente en que un equipo se clasifique para competiciones europeas. Es decir, a mayores los registros de estas variables, más probable será que el equipo se clasifique para alguna competición europea. Estas variables son llamadas variables de riesgo. Del lado contrario, las variables paradas (Saves) y pases desde balón parado (PassingTypeDead) influyen negativamente en que un equipo se clasifique para competiciones europeas. Es decir, a mayor cantidad de esta variable, más difícil será que el equipo se clasifique esa temporada. A estas variables se les denotará por variables de contención.

Para mejorar los resultados obtenidos, se ha iterado mediante el método Stepwise para obtener aquellas variables más relevantes del dataset y que más eficiente hacen el modelo. En

la tabla del cuadro 2.11 se pueden ver las variables seleccionadas junto con los coeficientes y odd ratios obtenidos en una nueva implementación de un modelo logístico usando únicamente las variables seleccionadas.

Coeficientes y odd ratios de las variables seleccionadas		
Variables	Coeficientes	Odd ratios
Término independiente	-1.7260	0.1780
Possesion	0.7450	2.1063
ShotsOnTarget	1.3999	4.0547
Saves	-1.0837	0.3383
Offsides	1.0316	2.8057
PassingTypeDead	-0.8993	0.4068
PassingTypeBetweenBackDefenders	0.6779	3.2316

Cuadro 2.11: Coeficientes y odd ratios correspondientes a las variables seleccionadas en el método Stepwise

Como refleja la tabla, tras realizar las iteraciones correspondientes con el método Stepwise, las viables seleccionadas han sido: tiros a puerta (ShotsOnTarget), paradas (Saves), posesión (Possesion), fueras de juego (Offsides), pases a balón parado (PassingTypeDead) y pases entre los defensas (PassingTypeBetweenBackDefenders). Y, el gráfico de odd ratios resultante se corresponde con el de la figura 2.13

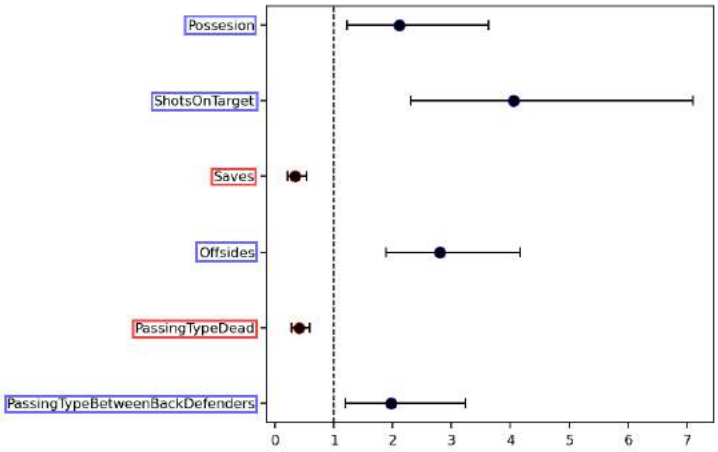

Figura 2.13: Gráfico de odds ratios y sus correspondientes intervalos de confianza, correspondiente con el modelo realizando una selección de variables previamente. Donde se han destacado las variables de contención (rojo) y las de riesgo (azul).

En comparación con el gráfico obtenido sin realizar la selección se han obtenido algunos cambios. En esta implementación se ha obtenido que las variables Possesion, ShotsOnTarget, Offsides y PassingTypeBetweenBackDefenders son de riesgo y las variables Saves y PassingTy-

peDead son de contención. Resultado bastante congruente ya que, teniendo en cuenta los resultados obtenidos en el mapa de correlaciones correspondiente con la figura 2.9 realizado al principio de esta sección, las variables ShotsOnTarget y Possesion son las dos variables que más relación directa presentan con la variable binaria Europe. No sólo eso, sino que la variable Saves ha sido identificada como variable de contención y, en dicho mapa de correlaciones, se obtuvo que es la variable que más relación inversa presenta con la variable binaria Europe.

Además, en la nueva implementación del modelo se ha obtenido un Pseudo-R cuadrado de 0.595. Con lo cual, queda reflejado que mediante el método Stepwise hemos conseguido seleccionar las variables más influyentes en el modelo sólo perdiendo un 4 % de la calidad del resultado. Al tener ahora muchas menos variables, resulta más sencillo representar la función de probabilidad del modelo, que quedará de la siguiente forma:

Considerando el cambio de nomenclatura: P como Possesion, SOT como ShotsOnTarget, S como Saves, O como Offsides, PTD como PassingTypeDead y PTBBD como PassingTypeBetweenBackDefenders

$$P(Y=1) = \frac{1}{1+e^{1{,}7260-0{,}7450\cdot P-1{,}3999\cdot SOT+1{,}0837\cdot S-1{,}0316\cdot O+0{,}8993\cdot PTD-0{,}6779\cdot PTBBD}}$$

Por último, se ha decidido implementar un modelo usando como variables predictoras las seis primeras componentes principales que se obtuvieron en la realización del PCA al principio de esta sección. Con estas seis componentes es posible representar, reflejado en el gráfico de sedimentación de la figura 2.10, el 75.163 % de la información total. En esta implementación, los coeficientes y odd ratios obtenidos se corresponden con los del cuadro 2.12.

Coeficientes y odd ratios de las componenetes principales		
Variables	Coeficientes	Odd ratios
Término independiente	-1.4360	0.237884
CP1	-0.8781	0.4156
CP2	-0.2627	0.7690
CP3	-0.3988	0.6711
CP4	-0.3066	0.7359
CP5	-0.2873	0.7502
CP6	0.2209	1.247198

Cuadro 2.12: Coeficientes y odd ratios correspondientes a las componentes principales obtenidas en el PCA con las cuales se ha implementado un modelo de regresión logística.

Siguiendo con el procedimiento realizado en el resto de implementaciones anteriores, se ha generado el correspondiente gráfico de odd ratios con sus intervalos de confianza respectivos. Esto se corresponde con el gráfico de la figura 2.14

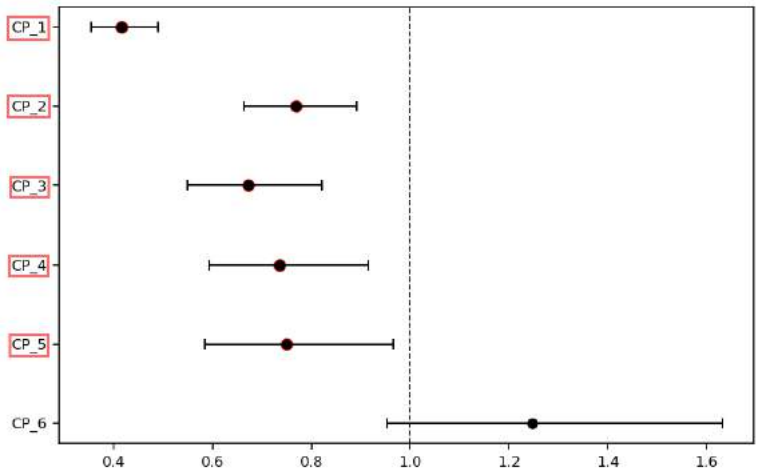

Figura 2.14: Gráfico de odds ratios y sus correspondientes intervalos de confianza, correspondiente con el modelo realizando una selección de variables previamente. Donde se han destacado las variables de contención (rojo) y las de riesgo (azul).

En este caso, las cinco primeras componentes principales han sido consideradas como factores de contención. Tal y como se aprecia, la primera componente será la más determinante a la hora de clasificar o no para competiciones europeas, puesto que es la que tiene el intervalo de confianza más alejado del punto x=1. Hilando con resultados obtenidos con anterioridad, en el gráfico de la figura 2.11 correspondiente con la representación gráfica de los registros en función de las dos primeras componentes principales, vimos que los equipos que menos valor tenían en la CP1 eran los que se clasificaban para competiciones europeas. Esto coincide con los resultados obtenidos en este gráfico, ya que al ser considerada como un factor de contención, se estará reflejando que los equipos que más altos tengan los valores de esta componente, más probabilidad tendrán de finalizar la temporada fuera de puestos clasificatorios para competiciones europeas.

El Pseudo-R cuadrado obtenido en esta última implementación ha sido de 0.5177. Es decir, se ha perdido un 11 % de calidad respecto al modelo implementado con todas las variables. Y, para finalizar, la función de probabilidad correspondiente con esta implementación es la siguiente:

$$P(Y=1) = \frac{1}{1+e^{1{,}436+0{,}8781\cdot CP1+0{,}2627\cdot CP2+0{,}3988\cdot CP3+0{,}3066\cdot CP4+0{,}2873\cdot CP5-0{,}2209\cdot CP6}}$$

2.3. Regresión ordinal

2.3.1. Introducción

En esta sección se pretende realizar la implementación de un modelo de regresión ordinal. Para este caso, la variable objetivo será una variable categórica con tres posibles valores: Europa, Permanencia y Descenso. En función de que al final de cada temporada registrada el equipo acabó clasificándose para una de las dos competiciones europeas (Champions League y Europa League), acabó en la zona de media tabla, o acabó en la zona de descenso de su correspondiente liga. En este caso, las variables usadas como variables predictoras han sido las mismas que en la sección anterior, que se corresponden con los siguientes datos de las últimas cinco temporadas concluidas de las cinco grandes ligas europeas: posesión, penaltis realizados, tarjetas amarillas, tarjetas rojas, paradas, tiros a puerta, distancia media de los tiros, pases cortos, pases medios, pases largos, pases en movimiento, pases a balón parado, pases de tiro libre, pases entre defensas, pases a lo ancho del campo, pases desde centros, pases de saque de banda, pases de córner, regates intentados, toques en la zona 1, toques en la zona 2, toques en la zona 3, toques en la zona 4, toques en la zona 5, duelos aéreos ganados, faltas realizadas, fueras de juego y interceptaciones.

En el siguiente apartado de la sección se detallarán los fundamentos teóricos de la regresión ordinal. Y, para concluir, se presentarán los resultados obtenidos en la implementación del modelo antes y después de aplicar un algoritmo de selección de variables.

2.3.2. Fundamentos teóricos

La regresión ordinal [4] es un técnica muy similar a la regresión logística. Estas técnicas se diferencian en que en el caso de la regresión ordinal, se tendrán más de dos respuestas categóricas cómo variable objetivo. De forma similar al logístico, se tendrá una serie de observaciones $\{(x_i, y_i)\}_{i=1}^{n}$ donde $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$, siendo p el número de variables estimadoras y n el tamaño del dataset. Pero en éste caso, cada una de las variables objetivo y_i, podrá tomar K valores categóricos diferentes.

Función de probabilidad

Para representar los K posibles valores categóricos de la variable objetivo se empleará la siguiente notación $R = 0, 1, .., K - 1$.

En primer lugar, el modelo se representará de forma:

$$P(R \geq k|x_i) = \frac{1}{1 + e^{-(\alpha_k + \sum_{i=1}^{p} \beta_i x_i)}} \text{ para } k = 0, 1, ..., K-1$$

De este modo, el odds ratio quedará de la siguiente manera:

$$odds(R \geq k|x_i) = \frac{P(R \geq k|x_i)}{P(R \leq k|x_i)} = e^{\alpha_k + \sum_{i=1}^{p} \beta_i x_j}$$

Por último, se podrá definir la función de probabilidad de acuerdo con la siguiente fórmula:

$$L = \prod_{j=1}^{n} \prod_{k=0}^{K-1} P(R = k|X)^{y_{jk}}, \text{ donde } y_{jk} = \begin{cases} 1, & \text{si } y_j = k \\ 0, & \text{resto} \end{cases}$$

Método de máxima verosimilitud

Para la regresión ordinal, la función de verosimilitud se definirá del siguiente modo:

$$L(\alpha, \beta|Y, X) = \ldots = \prod_{i=1}^{n} \prod_{j=2}^{K-1} \left(\frac{1}{1 + e^{-(\alpha_1 + \beta' x_j)}} \right)^{\delta_{i1}} \left(\frac{1}{1 + e^{-(\alpha_j + \beta' x_j)}} - \frac{1}{1 + e^{-(\alpha_{j-1} + \beta' x_j)}} \right)^{\delta_{ij}}$$

$$\text{donde } \delta_{ij} = \begin{cases} 1, & \text{si el i-ésimo individuo muestra } Y = y_j \\ 0, & \text{en otro caso} \end{cases}$$

Cabe destacar que α_j y $\beta' = (\beta_1, ..., \beta p)$ son los parámetros que se desea estimar. Del mismo modo que en la regresión logística, para realizar la estimación de dichos parámetros, se puede realizar iterando mediante el método de Newton-Raphson explicado en los fundamentos teóricos de la sección anterior.

2.3.3. Resultados obtenidos

Una vez implementado el modelo de regresión ordinal, donde se ha obtenido un Pseudo-R cuadrado de 0.4247, se han calculado los odds ratios y los intervalos de confianza para cada variable para cada una de las dos comparaciones realizadas.

La primera comparación se corresponde a la comparación de la categoría 0 (Descenso) frente a la categoría 1 (Permanencia). Por tanto, los odds ratios representan cómo cambia la probabilidad de descender frente a la probabilidad de permanecer en la primera división de cada liga para cada variable independiente usada en el modelo. El gráfico generado para representar lo anterior se corresponde con el de la figura 2.15.

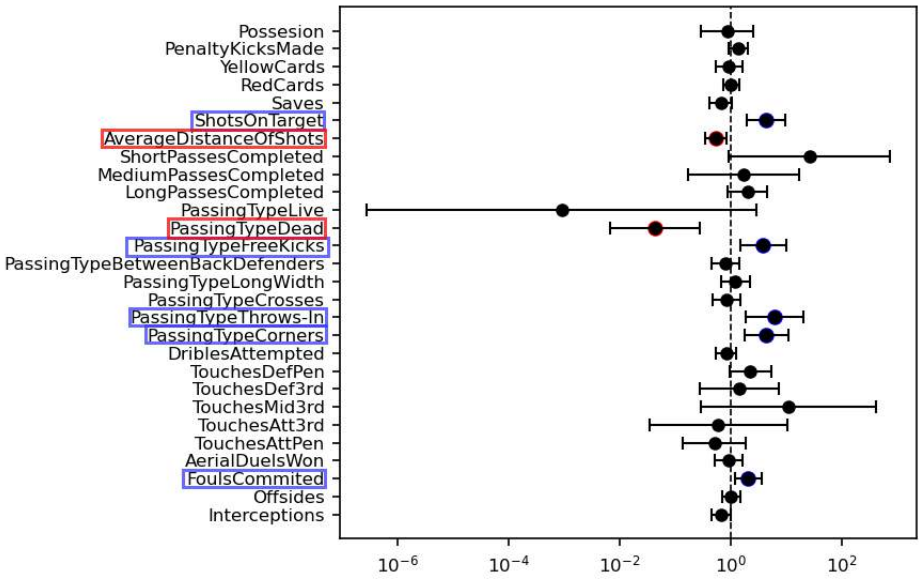

Figura 2.15: Gráfico de los odds ratios resultantes de la comparación Descender frente a Permanencia, con sus correspondientes intervalos de confianza. Donde se han destacado las variables de contención (rojo) y las de riesgo (azul). Se encuentra escalado por el logaritmo.

Como se observa en el gráfico, las variables que ayudan a diferenciar los equipos pertenecientes a cada categoría son: ShotsOnTarget, AverageDistanceOfShots, PassingTypeDead, PassingTypeFreeKicks, PassingTypeThrows-In, PassingTypeCorners y FoulsCommited. En el gráfico se han representado en rojo, si se tratan de variables de contención, o en azul, si se tratan de variables de riesgo.

La segunda comparación realizada se corresponde con con la comparación de la categoría 0 (Descenso) frente a la categoría 2 (Europa). Por tanto, los odds ratios representan cómo cambia la probabilidad de descender frente a la probabilidad de clasificarse para competiciones europeas en la primera división de cada liga para cada variable independiente usada en el modelo. El gráfico generado para representar lo anterior se corresponde con el de la figura 2.16.

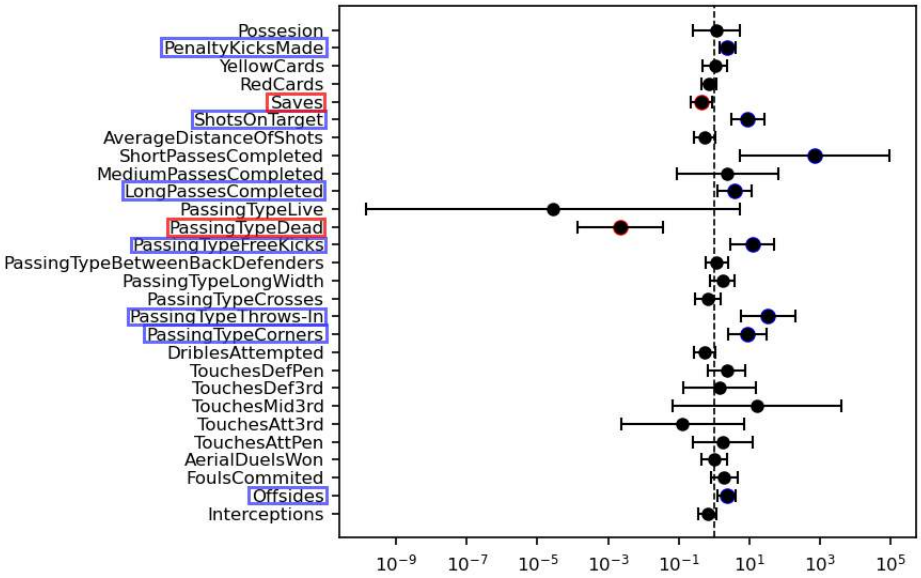

Figura 2.16: Gráfico de los odds ratios resultantes de la comparación Descender frente a Europa, con sus correspondientes intervalos de confianza. Donde se han destacado las variables de contención (rojo) y las de riesgo (azul). Se encuentra escalado por el logaritmo.

En este caso, las variables que permiten diferenciar los equipos de cada categoría son: PenaltyKicksMade, Saves, ShotsOnTarget, ShortPassesCompleted, LongPassesCompleted, PassingTypeDead, PassingTypeFreeKicks, PassingTypeThrows-In, PassingTypeCorners y Offsides. De el mismo modo que en el gráfico de la otra comparación realizada, las variables de contención se han representado de color rojo y las de riesgo de color azul. Como se puede observar, son muchas más las variables que ayudan a diferenciar una categoría de la otra en comparación con el caso anterior. Esto refleja que será mucho más fácil diferenciar entre los equipos que descienden y los que se clasifican para competiciones europeas, en lugar de entre los que descienden y los que acaban la temporada a mitad tabla.

Tras esto, se ha implementado un método Stepwise (explicado en la sección 2.2) para poder seleccionar las variables más influyentes en el modelo. y, con las variables seleccionadas se ha implementado un nuevo modelo de regresión ordinal usando esta selección como variables predictores. En la tabla del cuadro 2.13 se muestran las variables seleccionadas en el algoritmo de selección junto con los coeficientes y odd ratios obtenidos para cada una de las dos comparaciones que se realizan en el modelo tras realizar su implementación. Mediante los coeficientes resultantes, se pueden construir las dos funciones de probabilidad, tal y como se ha realizado en la regresión logística, correspondientes a las dos comparaciones: $P(Permanencia) = P(Y = 1)$ y $P(Europa) = P(Y = 2)$. En este caso, no se ha representado puesto que al haber más variables resultaría más largo para poderse mostrar sobre el papel.

Coeficientes y odd ratios para la comparación Descender frente a Pemanencia (y=1)				
Variables	Coeficientes (y=1)	Odd ratios (y=1)	Coeficientes (y=2)	Odd ratios (y=2)
Término independiente	1.8863	6.5949	0.4721	1.6034
ShotsOnTarget	1.4588	4.3007	2.7310	15.3480
Saves	-0.4324	0.6490	-1.1314	0.3226
Interceptions	-0.2775	0.7576	-0.4761	0.6212
PenaltyKicksMade	0.2527	1.2875	0.6860	1.9858
TouchesAtt3rd	-1.7775	0.1690	-2.5549	0.0777
AverageDistOfShots	-0.5846	0.5573	-1.0889	0.3366
PassingTypeLongWidh	0.2971	1.3460	0.6324	1.8821
ShortPassesCompleted	0.8686	2.3836	1.9833	7.2668
FoultsCommited	0.4070	1.5023	0.3047	1.3562
TouchesDef3rd	-0.2274	0.7966	-0.6162	0.5400

Cuadro 2.13: Coeficientes y odd ratios resultantes de las dos comparaciones realizadas empleando las variables seleccionadas mediante el método de selección de variables.

A continuación se han generados las dos gráficas de los odd ratios junto con sus correspondientes intervalos de confianza. Para así poder analizar las variables de contención y riesgo más influyentes en el modelo. El gráfico representativo de la comparación Descenso frente a Permanencia (y=1) se corresponde con la figura 2.17. Mientras que el gráfico representativo de la comparación Descenso frente a Europa (y=2) se corresponde con la figura 2.18

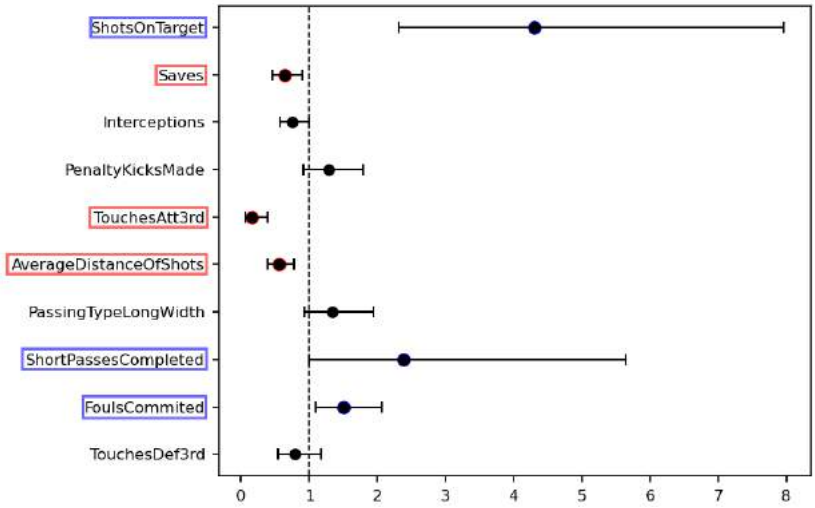

Figura 2.17: Gráfico de los odds ratios resultantes de la comparación Descender frente a Permanencia, con sus correspondientes intervalos de confianza. Donde se han destacado las variables de contención (rojo) y las de riesgo (azul). Implementación realizada después de iterar mediante un algoritmo de selección de variables.

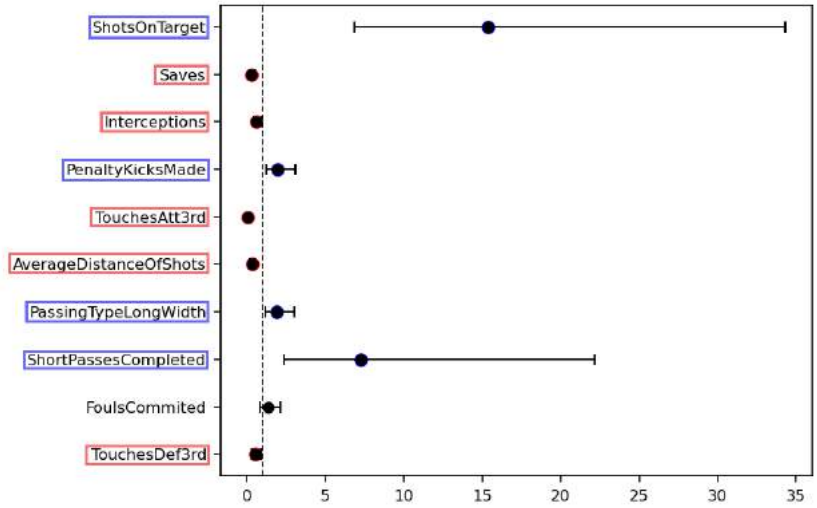

Figura 2.18: Gráfico de los odds ratios resultantes de la comparación Descender frente a Europa, con sus correspondientes intervalos de confianza. Donde se han destacado las variables de contención (rojo) y las de riesgo (azul). Implementación realizada después de iterar mediante un algoritmo de selección de variables.

En estas dos gráficas generadas, se observa con más claridad que son más las variables que permiten diferencia entre un equipo que acaba la temporada en puestos de descenso y uno que acaba en puesto de competiciones europeas, que entre un equipo que acaba en puesto de descenso y uno que acaba en puestos de permanencia. Además, al haber seleccionado las variables más influyentes en el modelo, las clasificaciones obtenidas sobre si una variable es de contención o de riesgo (colores en las gráficas) será más fiable y útil que en los resultados de la implementación realizada anteriormente.

Es importante recalcar que las dos variables de riesgo, tiros a puerta (ShotsOnTarget) y pases cortos completados (ShortPassesCompleted), serán las más determinantes a la hora de que un equipo pertenzca a cada uno de lo grupos de la variable objetivo estudiados en esta sección. Puesto que son las que más alejadas se encuentran del punto x=1 en el gráfico.

Por último, es importante tener en cuenta que en esta última implementación del modelo se ha obtenido un Pseudo-R cuadrado de 0.356, que comparado con el de la implementación realizada con todas las variables refleja que se ha perdido un 6 % de la calidad del modelo mediante la eliminación de las 18 variables menos influyentes.

Capítulo 3

Conclusiones

Resumiendo, el proyecto de análisis estadístico de las cinco grandes ligas de fútbol europeas ha supuesto una experiencia enriquecedora. A lo largo de su realización ha sido posible adquirir gran cantidad de conocimientos sobre la minería de datos, los cuales eran desconocidos por mi hasta el momento.

Durante este proceso, se ha realizado el aprendizaje de las fases para la realizar una limpieza de datos precisa y efectiva, así como la creación de modelos, algoritmos y métodos que resultan de gran utilidad en el análisis estadístico. Además, mediante la realización de estas técnicas se me ha brindado la oportunidad de comprender los fundamentos teóricos matemáticos en los que se sustentan.

Gracias a la puesta en práctica de estas técnicas, se ha podido determinar los factores que permiten diferenciar entre las diferentes ligas estudiadas. También se han podido identificar los equipos en la clasificación de su correspondiente liga, así como los factores más determinantes para el descenso, permanencia en la mitad de la tabla o la clasificación para competiciones europeas en la siguiente temporada.

Los resultados obtenidos me han permitido obtener una visión más clara y profunda de las características distintivas de cada liga y los factores clave que influyen en el rendimiento de los equipos.

En definitiva, el proyecto de análisis estadístico realizado en el marco de la asignatura MT1054 - Treball Final de Grau me ha permitido adquirir conocimientos y habilidades fundamentales en la minería de datos aplicada al fútbol. Los resultados obtenidos han contribuido a una mayor comprensión de las ligas y equipos estudiados.

Bibliografía

[1] Juan Pablo Aguilar Ticona, María Belen Arriaga Gutiérrez, Ninfa Marlen Chaves Torres, and Diana Reyna Zeballos Rivas. Entendiendo la odds ratio. *Scientifica*, 2018.

[2] Joaquín Amat Rodrigo. Análisis de normalidad con python. `https://www.cienciadedatos.net/documentos/pystats06-analisis-normalidad-python.html`. Consulta: 10 de Abril de 2023.

[3] Joaquín Amat Rodrigo. Test kruskal-wallis. `https://www.cienciadedatos.net/documentos/20_Kruskal-Wallis_test.html`.

[4] M. Arias Benítez. Regresión ordinal y sus aplicaciones. Trabajo Fin de Grado Inédito, 2018.

[5] Sergio Alexis Dominguez-Lara. El odds ratio y su interpretación como magnitud del efecto en investigación. *Educación Médica*, 19(1):65–66, enero 2018. Consultado el 16 de abril de 2021.

[6] Juan Gabriel Gomila and Frogames Support Team. Curso completo de machine learning: Data science en python. `https://www.udemy.com/share/106hnS3@c6gwnN_R54EELRT6FsUL3WiJPrQNk5qToYG99_b2j-mrgzsZ2CDl_OGcu0fZJr5Y8A==/`.

[7] Ana González Vidal. Selección de variables: Una revisión de métodos existentes. `http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1263.pdf`.

[8] Oscar Eduardo Gualdrón Guerrero. Desarrollo de diferentes métodos de selección de variables para sistemas multisensoriales. `https://www.tdx.cat/bitstream/handle/10803/8473/Tesis_Oscar_Gualdron.pdf?...1`, 2007.

[9] J. I. Guerrero. Anova prueba de tukey. `https://es.slideshare.net/JaimeIncaGuerrero/anova-prueba-de-tukey`. Recuperado el 10 de abril de 2023.

[10] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, December 1952.

[11] Rafael Niño Alfaro. Predicción de las razones de un estudiante en su selección de una escuela mediante regresión logística. https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_NINO%20ALFARO.pdf, 2021.

[12] Daniel Peña. *ANALISIS DE DATOS MULTIVARIANTES*. MCGRAW-HILL, España, 2002.

[13] A. Pérez López, D. Hierro, and M. Pereda. Estudio de la influencia de los tiempos de decisión, el género y la edad sobre la cooperación humana en los juegos de bienes públicos noviembre 2019. https://oa.upm.es/57351/1/TFG_ANTONIO_PEREZ_LOPEZ_DEL_HIERRO.pdf.

[14] Sidney Siegel and N. John Castellan Jr. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 2nd edition, 1988.

[15] Universitat de Barcelona. Contrastes de normalidad. http://www.ub.edu/aplica_infor/spss/cap5-6.htm.

[16] ZACH. How to perform tukey's test in python. https://www.statology.org/tukey-test-python/, December 18 2020.

Anexo A

Descripción del dataset

Para poder realizar un análisis estadístico de diferentes ligas de fútbol es necesario disponer de un registro de datos, para así, poder realizar los estudios y las comparaciones pertinentes. Para este proyecto, se han tomado los datos de la página web: https://fbref.com/es/. Se trata de un sitio web dedicado al seguimiento de estadísticas de equipos y jugadores de fútbol de todo el mundo. Es uno de las webs gratuitas más completas existentes actualmente.

Teniendo en cuenta los objetivos, los datos que se descargaron fueron referentes a las estadísticas correspondientes a las últimas cinco temporadas finalizadas: 17/18, 18/19, 19/20, 20/21 y 21/22. Y, correspondientes a las cinco grandes ligas europeas de fútbol: Premier League (Inglaterra), La Liga (España), Serie A (Italia), Ligue 1 (Francia) y Bundesliga (Alemania).

Por todas las temporadas de todos los equipos, se realizó una selección de una parte de los datos registrados, ya que eran demasiados los datos disponibles y no iban a ser necesarios todos. Estos datos seleccionados se pueden agrupar en los siguientes bloques: Posición, Equipo, Estadísticas Generales, Tiempo de juego, Tarjetas, Porteros, Tiros, Pases, Regates, Toques, Duelos aéreos y Otros. Y, dentro de cada bloque se pueden encontrar la siguiente información:

- **Posición** (Rk) (Entero)
- **Equipo** (Squad) (Cadena de caracteres)
- **Estadísticas Generales** (Enteros): Partidos jugados (MachesPlayed), victorias (Wins), empates (Draws), derrotas (Loses), goles a favor (GoalsFavor), goles en contra (GoalsAgainst), diferencia de goles (GoalsDifference), asistencias (Assists), puntos (Pts), público (Attendance), jugadores usados (Players Used), edad media (AverageAge) y posesión (Possesion).

- **Tiempo de juego** (Enteros): Comienzos (Starts), Minutos (Min).
- **Penaltis** (Enteros): Realizados (PenaltyKicksMade) y recibidos (PenaltyKicksAttended).
- **Tarjetas** (Enteros): Amarillas (YellowCards), segundas tarjetas amarillas (2nd YellowCards) y rojas (RedCards).
- **Porteros** (Enteros): Porteros usados (GoalkepersUsed), tiros a puerta en contra (ShotsOnTargetAgainst), paradas (Saves) y cantidad de porterías a cero (CleanSheets).
- **Tiros**(Enteros): Totales (TotalShots), a puerta (ShotsOnTarget), distancia media de los tiros (AverageDistanceOfShots) y tiros libres(FreeKicks).
- **Pases** (Enteros):
  - **General**: Completados (TotalPassesCompleted), intentados (TotalPassesAttended), distancia total de los pases completados (TotalDistOfPassesCompleted), fallados (PassesOff) y bloqueados (PassesBlocks).
  - **Cortos**: Completados (ShortPassesCompleted) y intentados (ShortPassesAttempted).
  - **Medios**: Completados (MediumPassesCompleted) y intentados (MediumPassesAttempted).
  - **Largos**: Completados (LongPassesCompleted) y intentados (LongPassesAttempted).
  - **Tipos**: En movimiento (PassengTypeLive), a balón parado (PassingTypeDead), de tiro libre (PassengTypeFreeKick), entre defensas (PassingTypeBetweenBackDefenders), a lo ancho del campo (PassingTypeLongWidth), desde centros (PassingTypeCrosses), de saque de banda (PassingTypeThrows-In) y de córner (PassingTypeCorners).
- **Regates** (Enteros): Completados (DriblesSuccessfully) y intentados(DriblesAttempted).
- **Toques** (Enteros): Totales (TotalTouches), en la zona 1 (TouchesDefPen), en la zona 2 (TouchesDef3rd), en la zona 3 (TouchesMid3rd), en la zona 4 (TouchesAtt3rd), en la zona 5 (TouchesAttPen) y en movimiento (TouchesLive).

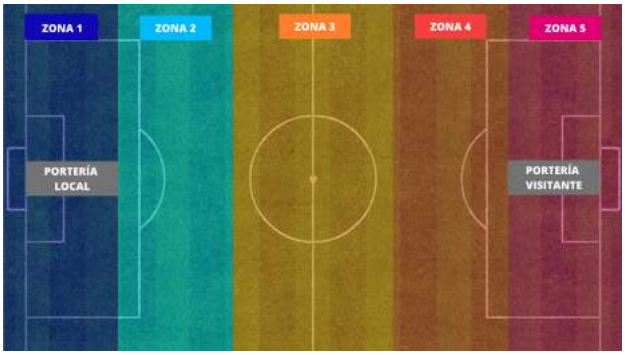

Figura A.1: Dibujo de las zonas de campo

- **Duelos aéreos** (Enteros): Ganados (AerialDuelsWon) y perdidos (AerialDuelsLost).
- **Otros** (Enteros): Faltas realizadas (FoulsCommited), goles en propia (OwnGoals), fueras de juego (Offsides) y interceptaciones (Interceptions).

Anexo B

Repositorio de GitHub con los programas implementados

Los ficheros con los datos empleados y los programas creados en Python mediante Jupyter Notebook han sido subidos a un repositorio de GitHub al cual se puede acceder mediante el siguiente enlace:

https://github.com/nicolascamanes/TFG_Nicolas_Camanes_Antolin.git

Dicho repositorio consta de dos carpetas. La primera, "datasets" contiene cinco ficheros .xlsx (hoja de cálculo de Microsoft Excel) que se corresponden con los datos usados con las cinco ligas, un fichero por liga. La segunda carpeta, "programas en jupyter notebook" que contiene los ficheros .ipynb (Notebook de Jupyter) en los que se encuentra el código programado en Python para las diferentes implementaciones. Por último, en el repositorio se encuentran cuatro ficheros .html que se corresponden con los ficheros de Jupyter Notebook descargados en formato HTML para tener la posibilidad de descargar los programas ejecutados para poder visualizar tanto el código como los resultados de las ejecuciones en el navegador. Es importante mencionar que muchos de los conocimientos para realizar la programación fueron aprendidos mediante la realización de un curso de la plataforma Udemy [6]. Muchos otros fueron aprendidos de diferentes fuentes referenciadas en el documento.