UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Identificación de estrategias de juego en ligas europeas de fútbol

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Lopez Sanchez, Julio Moises

Tutor/a: Conchado Peiró, Andrea

Cotutor/a: Jabaloyes Vivas, José Manuel

CURSO ACADÉMICO: 2021/2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Identificación de estrategias de juego en ligas europeas de fútbol

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Lopez Sanchez, Julio Moises

Tutor/a: Conchado Peiró, Andrea

Cotutor/a: Jabaloyes Vivas, José Manuel

CURSO ACADÉMICO: 2021/2022

Resumen

Hoy en día, el análisis de datos está presente en todos los sectores, y el deporte no es indiferente a esta revolución. Desde la predicción de un aspecto concreto del juego, como los lanzamientos de penaltis en fútbol, hasta la elección de los fichajes o la optimización de los entrenamientos para adaptarlos a un contrincante concreto, como aplica la jugadora de bádminton Carolina Marín, son muchos y muy diversos los ejemplos de la aplicación de diferentes herramientas de análisis de datos aplicados al mundo del deporte.

En el presente trabajo, se desarrolla un análisis de las estrategias de juego en las principales ligas europeas de fútbol. A partir de las estadísticas de cada equipo en cada partido jugado (ataques, centros, posesión, pases…), y a partir del análisis de componentes principales (PCA), se obtendrán relaciones entre ellas para detectar estas diferentes estrategias.

Una vez se identifiquen estas estrategias, se analizará su capacidad de éxito en términos generales durante el periodo temporal de los datos a partir de una clasificación de los registros por medio de un análisis Clúster.

En la segunda parte del trabajo, se intentará predecir el éxito en los partidos de cada equipo teniendo en cuenta las estadísticas de cada equipo en cada partido para comprobar si realmente a partir de estas estadísticas es posible determinarlo. El éxito en el partido se definirá en base a si el partido ha sido ganado o no ha sido ganado (empatado o perdido).

Palabras clave: estrategias, tácticas, enfrentamiento, predicciones, éxito en el partido.

Resum

Hui en dia, l'anàlisi de dades està present en tots els sectors, i l'esport no és indiferent a esta revolució. Des de la predicció d'un aspecte concret del joc, com els llançaments de penals en futbol, fins a l'elecció dels fitxatges o l'optimització dels entrenaments per a adaptar-los a un contrincant concret, com aplica la jugadora de bàdminton Carolina Marín, són molts i molt diversos els exemples de l'aplicació de diferents ferramentes d'anàlisi de dades aplicats al món de l'esport.

En el present treball, es desenrotlla una anàlisi de les estratègies de joc en les principals lligues europees de futbol. A partir de les estadístiques de cada equip en cada partit jugat (atacs, centres, possessió, pases...) , i a partir de l'anàlisi de components principals (PCA) , s'obtindran relacions entre elles per a detectar estes diferents estratègies.

Una vegada s'identifiquen estes estratègies, s'analitzarà la seua capacitat d'èxit en termes generals durant el període temporal de les dades a partir d'una classificació dels registres per mitjà d'una anàlisi Cluster.

En la segona part del treball, s'intentarà predir l'èxit en els partits de cada equip tenint en compte les estadístiques de cada equip en cada partit per a comprovar si realment a partir d'estes estadístiques és possible determinar-ho. L'èxit en el partit es definirà basant-se en si el partit ha sigut guanyat o no ha sigut guanyat (empatat o perdut) . Paraules clau: estratègies, tàctiques, enfrontament, prediccions, èxit en el partit.

Paraules clau: estratègies, tàctiques, enfrontament, èxit en el partit.

Abstract

Nowadays, data analysis is present in all sectors, and sports is not indifferent to this revolution. From the prediction of a specific aspect of the game, such as penalty kicks in soccer, to the choice of signings or the optimization of training sessions to adapt them to a specific opponent, as applied by the badminton player Carolina Marín, there are many and very diverse examples of the application of different data analysis tools applied to the world of sports.

In this paper, an analysis of the game strategies in the main European soccer leagues is developed. From the statistics of each team in each match played (attacks, crosses, possession, passes...), and from the principal component analysis (PCA), relationships between them will be obtained to detect these different strategies.

Once these strategies have been identified, their ability to succeed in general terms during the period of the data will be analyzed from a classification of the records by means of a cluster analysis.

In the second part of the work, we will try to predict the success in the matches of each team taking into account the statistics of each team in each match to check if it is really possible to determine it from these statistics. The success in the match will be defined based on whether the match has been won or not won (drawn or lost).

Keywords: strategies, tactics, confrontation, predictions, match success.

Índice de Memoria

Índice de los Anexos

Índice de Figuras

Índice de Tablas

MEMORIA

1. Motivación

A lo largo de los últimos años, uno de los puntos clave para el desarrollo de cualquier industria ha sido la aplicación de técnicas avanzadas de análisis de datos masivos con objetivos tan distintos como optimizar procesos, detectar fallos de forma temprana o tomar decisiones de negocio basadas en la evidencia que muestran estos datos. Estas técnicas han permitido dar importantes pasos adelante en cuanto al conocimiento sobre los procesos internos y externos o las operativas de cualquier industria y al acierto en la toma decisiones; a esta revolución se la conoce comúnmente como la revolución del *Big Data*.

El término *Big Data* hace referencia a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

Esta revolución permite a empresas y organizaciones aprovechar mucho mejor sus datos y utilizarlos para identificar nuevas y mejores oportunidades, ya que permiten responder preguntas que hasta este momento ni siquiera se planteaban.

Aplicado al deporte, el análisis de datos se centra en áreas tanto dentro del terreno de juego (conocerse a uno mismo y al rival) como fuera de la pista (evitar lesiones o captar talento).

El equipo pionero en la aplicación de estos análisis, tal y como narra la película *Moneyball* (2011), son los Oakland Athletics; este modesto equipo aplicó el análisis de las estadísticas para decidir cuáles son los mejores fichajes posibles para formar una plantilla equilibrada con poco presupuesto. A partir de ese momento, son muchos los equipos que han adoptado técnicas de análisis de datos; tanto es así que, en 2017, Deloitte calculó que un 97% de equipos de la MLB (*Major League Baseball*) y un 80% de equipos de la NBA (*National Basketball League*) aplicaban estas técnicas [1].

En el año 2018, el mercado global de análisis de datos en deportes se estimó en más de 750 millones de dólares, cifrando las expectativas de crecimiento anual en más de un 30% anual entre 2019 y 2025. Tal es el papel del análisis de datos en el deporte que David R. Sáez, CEO de Sport Data Campus, afirma que “El *Big Data* y la analítica avanzada están revolucionando el mundo del deporte y el perfil del analista de datos con *Big Data* se está convirtiendo en esencial en muchas entidades deportivas, siempre en aras de encontrar mejoras y ventajas competitivas”.

Entre los ejemplos más conocidos de análisis de datos aplicados al deporte se encuentran, además de su aplicación para la confección de plantillas en numerosos deportes, la predicción de los lanzamientos de penaltis en el fútbol o el estudio de las estrategias y el juego de los rivales en bádminton que aplica Carolina Marín.

En particular en el fútbol, el deporte más extendido del mundo, si bien no ha sido un deporte pionero en el uso de estas técnicas, en los últimos años muchos clubes han llegado a crear departamentos propios de analítica de datos conformados por científicos e ingenieros para optimizar su rendimiento tanto dentro como fuera de la pista. Diferentes empresas especializadas en la recogida y procesamiento de datos deportivos han facilitado el trabajo de estos clubes y federaciones para el posterior análisis de esos datos.

También investigadores han desarrollado muy diversos trabajos aplicando técnicas de análisis de datos al fútbol, como la predicción de acciones de gol en la liga de fútbol estadounidense [2], o el análisis de jugadores y posiciones para optimizar la decisión a l ahora de fichar a un jugador [3].

En el ámbito puramente futbolístico, y contextualizando el presente trabajo, existen grandes diferencias en las estrategias de los diferentes equipos; mientras algunos equipos se deciden por jugar en su propio campo y ser equipos muy defensivos que intentan aprovechar los contraataques, otros equipos buscan controlar mucho más el balón y generar oportunidades a través de oportunidades más largas.

Sin embargo, equipos históricos han conseguido ser exitosos utilizando estrategias muy diferentes, por lo que a priori no se puede determinar una estrategia única que todos los equipos deban perseguir si quieren ser equipos triunfadores en cualquier época.

Hasta el momento, pocos trabajos científicos han abordado la problemática de estimar modelos predictivos del éxito de los partidos de fútbol en función de las estrategias. El presente trabajo pretende analizar esta cuestión y aportar luz en esta línea de investigación.

Los resultados obtenidos ayudarán tanto a equipos y profesionales del mundo del deporte a desarrollar la planificación deportiva, como a analistas de fútbol y aficionados a conocer más en profundidad lo que dicen los datos sobre el deporte que les apasiona cuando se analizan con técnicas profesionales de análisis de datos.

Por otro lado, hoy en día se puede ver en muchos partidos que la realización proporciona en diferentes momentos del encuentro la probabilidad de que gane un equipo u otro gracias a los análisis proporcionados por empresas como Driblab. Son especialmente conocidos 2 casos que han ocurrido en el año 2022 en los que el resultado del partido contradijo la clara predicción que se obtuvo con el análisis de datos:

- Final Open de Australia 2022: en el partido entre Rafael Nadal y Daniil Medvedev, durante el tercer set y siendo el resultado muy positivo para el tenista ruso, se estimaron las posibilidades de victoria del tenista español en un 4%.

- Semifinal Liga de Campeones 2022: en el partido de vuelta en la que se enfrentaban el Real Madrid CF y el Manchester City, en el minuto 89 del partido, la probabilidad de clasificación del Real Madrid CF para la final del torneo se estimó en un 1% [4].

Aunque es cierto que en estos casos el jugador o equipo que menos posibilidades tenía terminó “venciendo a la estadística”, este tipo de predicciones no dejan de ser una contextualización y un apoyo gráfico sobre la realidad que es probable que se dé y aportan un valor añadido tanto a los comentaristas como a los aficionados a la hora de disfrutar del partido.

Desde la Organización de las Naciones Unidas, se valoran las contribuciones que el deporte, y especialmente el fútbol por su gran presencia e influencia internacional, hace y puede hacer a los Objetivos de Desarrollo Sostenible (ODS) [5]. Este trabajo avanza en la línea de desarrollo sostenible relacionado con la innovación, en este caso aplicada en el mundo del fútbol, por la aplicación de técnicas de análisis de datos a nuevas áreas del deporte. Además, este avance puede generar curiosidad en los aficionados por conocer las técnicas utilizadas para obtener los análisis que ven en los diferentes medios e informarse o formarse en estas técnicas.

2. Objetivos

El objetivo global del presente trabajo es analizar las diferentes estrategias de juego de las grandes ligas europeas, y para alcanzar este objetivo se han definido los siguientes objetivos específicos:

Objetivo 1. Identificar estrategias de juego predominantes en base a la información disponible sobre el juego del equipo en el partido.

A partir de los datos de lo que se tiene, se hará una primera selección de variables no relacionadas directamente entre sí y, posteriormente, se realizará un Análisis de Componentes Principales para determinar los grupos de variables que representan cada una de estas estrategias.

Posteriormente se llevará a cabo un análisis clúster de los registros para determinar qué partidos se caracterizan por qué estrategias.

Objetivo 2. Analizar la capacidad predictiva de la información disponible del juego del equipo en el éxito del partido.

Definiendo el éxito en el partido en base al resultado (partido ganado o partido no ganado), se utilizarán diferentes técnicas de predicción para predecir, a partir de los datos del partido, el éxito de cada equipo en cada partido. Estas predicciones se realizarán para todo el conjunto de los datos, así como para cada liga de forma individual.

3. Metodología

En este apartado se describe el programa informático utilizado, así como las metodologías empleadas para la consecución de los objetivos del presente trabajo.

3.1. Software utilizado

El software R es un ambiente de programación formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones. Una de las principales características de este software, y que lo hace muy potente y extendido, es el hecho de que sea gratuito y de código abierto, por lo que son los propios usuarios los que pueden crear librerías y compartirlas con el resto de los usuarios de la herramienta.

RStudio es un entorno de desarrollo integrado (IDE) para R. Incluye una consola, un editor que resalta la sintaxis y admite la ejecución directa del código, así como herramientas para el trazado, el historial, la depuración y la gestión del espacio de trabajo [6].

Las librerías que se utilizarán a lo largo del presente trabajo son las siguientes:

- Readxl: librería utilizada para leer archivos Excel con R [7].
    - Función: read_excel.
- Ggplot2: librería utilizada para realizar gráficas de los resultados obtenidos [8].
    - Funciones: ggplot, aes.
- MVA: libraría utilizada para realizar análisis multivariantes [9].
    - Funciones:
- Biotools: librería utilizada para realizar y evaluar el análisis clúster, el análisis discriminante, entre otras técnicas de análisis de datos [10].
    - Funciones: boxM,
- MASS: libraría utilizada para realizar el análisis discriminante lineal y predicciones [11].
    - Funciones: lda, predict.
- ROCR: librería utilizada para crear curvas ROC [12].
    - Funciones: performance, prediction, plot.
- MDATOOLS: librería utilizada para llevar a cabo el análisis discriminante de mínimos cuadrados parciales [13].
    - Funciones: plsda.
- randomForest: librería utilizada para llevar a cabo análisis *random forest* tanto de clasificación como de regresión [14].
    - Funciones: randomForest, importance, varImpPlot.
- rpart: librería utilizada para crear árboles de clasificación [15].
    - Funciones: rpart, plotcp, prune.rpart, printcp.
- rpart.plot: librería diseñada para crear gráficos para modelos obtenidos con la librería *rpart* [16].
    - Funciones: prp.

- e1071: librería utilizada para análisis como las máquinas de soporte vectorial y el clasificador de Naive-Bayes [17].
    - Funciones: kurtosis, skewness, naiveBayes, svm.

3.2. Aprendizaje no supervisado

El aprendizaje no supervisado es una serie de técnicas de Aprendizaje Automático en los que los algoritmos trabajan sobre datos en los que la variable respuesta es desconocida.

En el presente trabajo se trabaja por diferentes técnicas de aprendizaje no supervisado para un análisis exploratorio de los datos y el análisis del éxito de las diferentes estrategias. Las técnicas de aprendizaje no supervisado utilizadas serán el análisis de componentes principales y el análisis clúster.

3.2.1. Análisis de Componentes Principales

El Análisis de Componentes Principales (por sus siglas en inglés, PCA), es una técnica clásica de aprendizaje no supervisado. Puede utilizarse para una serie de fines, como son:

- Análisis exploratorio de los datos:

  Tanto para descubrir las relaciones existentes entre las variables y observaciones analizadas como para descubrir la existencia de datos atípicos [18].

- Reducción de la dimensionalidad:

  El análisis de componentes principales utiliza la correlación entre las variables para crear nuevas variables, denominadas componentes principales, que explican la mayor parte de la variabilidad; la variabilidad explicada por cada componente principal será mayor que la posteriores, así como la varianza explicada por cada componente principal explicará variabilidad no explicada por las componentes anteriores.

  De esta forma, estas componentes principales pueden utilizarse como nuevas variables, con la ventaja de que no están correlacionadas entre sí. Teniendo en cuenta que las componentes principales de menor orden son aquellas de contienen el aspecto más importante de la información, será importante determinar cuántas componentes principales se desean utilizar a lo largo del análisis.

Entrando en detalle en la explicación técnica de este método, en el PCA se parte, como en todos los análisis de datos multivariados, es una matriz de datos **X**, Esta matriz **X** tiene dimensiones **N**x**K**, siendo **N** el número de individuos o registros, y **K** el número de variables explicativas. En la Figura 1 se puede ver una descripción gráfica de la descomposición de la matriz de datos inicial **X** en la matriz de componentes principales (por medio de los *scores* **T** y los *loadings* **P**) y en la matriz de residuos E que contiene toda la información que no se encuentra contenida en las componentes principales.

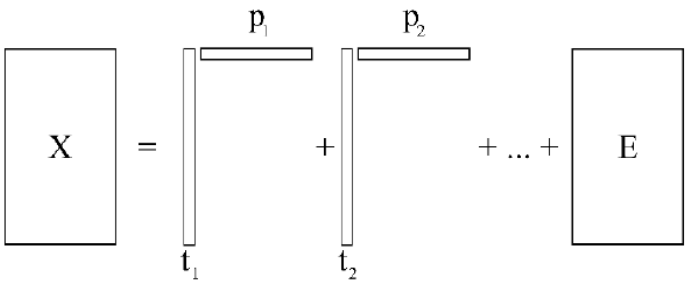

Figura 1. *Descomposición gráfica de la matriz de datos X en scores (T), loadings (P) y residuos (E). Figura tomada de [19].*

De esta forma, la base de datos original **X** se expresa mediante nuevas variables ortogonales (independientes entre sí) con una transformación a un nuevo sistema de coordenadas formado por los **T** *scores* y los **P** *loadings*. Con esta transformación, la proyección de **X** en un subespacio de dimensión A por medio de la matriz de proyecciones **P** genera las coordenadas del individuo en el subespacio **T**.

A las columnas en **T** se las conoce como *scores* o vectores de puntuación ($\mathbf{t_a}$) y las columnas en **P** se las conoce como *loadings* o vectores de carga ($\mathbf{p_a}$). Por último, las desviaciones de las proyecciones con respecto a las coordenadas originales se encuentran incluidas en la matriz de residuos **E** [19].

Con el resultado de este análisis pueden analizarse diferentes aspectos, entre ellos puede llevarse a cabo el estudio de las relaciones entre las variables explicativas y las componentes principales, y entre las componentes principales y los individuos. En concreto, el *Score Plot* permite visualizar los *scores* de las observaciones en las componentes principales seleccionadas; y el *Loading Bi-plot* permite analizar la relación entre las componentes principales seleccionadas y las variables en el espacio de las X.

3.2.2. Análisis Clúster

El Análisis Clúster es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los diferentes grupos.

Las soluciones de los análisis clúster no son únicas, ya que la conformación de los grupos depende del procedimiento escogido y sus elementos. Por otro lado, la solución clúster depende completamente de las variables escogidas, de modo que la adición o eliminación de variables relevantes puede tener un gran impacto en la solución final.

Dado un conjunto de puntos (elementos u objetos) $x = \{x_1, x_2, \dots, x_n\}$ de tamaño n, un clúster c_j es un conjunto de puntos que, basados en una medida de proximidad, son similares entre sí. El *clustering* es un proceso que permite dividir un conjunto en k grupos $c = \{c_1, c_2, \dots, c_k\}$ de datos distintos, por medio de algún criterio de agrupamiento, como una función de coste o algún otro tipo de regla de asociación [20].

De esta forma, el principio fundamental del clustering es garantizar que los grupos sean lo más heterogéneos entre sí, pero que los elementos del grupo sean lo más homogéneos posibles, basados en un criterio de optimización. En otras palabras, lo que se busca es minimizar la distancia o similaridad intra-cluster (cohesión) $mín\ d(x_1, x_2)$, y a la vez maximizar la distancia o similaridad inter-cluster (separación) $máx\ d(c_1, c_3)$ [21]. En la Figura 2 se puede ver esta relación de forma gráfica.

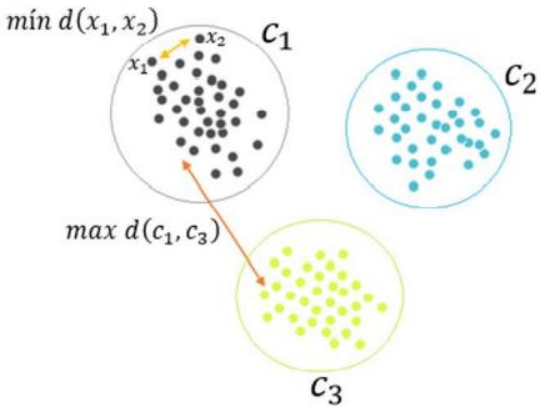

Figura 2. *Representación gráfica del objetivo del análisis clúster.*

En concreto, en este trabajo se utiliza un análisis clúster jerárquico. Este método entrega una jerarquía de divisiones del conjunto de elementos en grupos, que puede representarse en forma de árbol de soluciones o dendograma (Figura 3), en el cuál en el nivel más alto se encuentra un solo grupo que se subdivide a medida que se desciende en este dendograma. En este punto, es el analista el que debe decidir el número de clústeres que desea obtener.

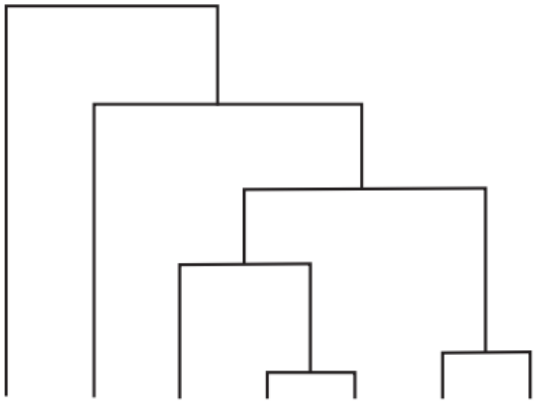

Figura 3. *Ejemplo de estructura de dendograma.*

Para este trabajo se ha aplicado para calcular los clústeres el método de Ward debido a que tiende a formar clústeres más compactos y de igual tamaño y forma en comparación con otros métodos y es poco sensible a *outliers* o valores atípicos. Este método une los individuos buscando minimizar la varianza dentro de cada grupo. Para ello, la función objetivo de la que parte el método de Ward es la suma de cuadrados de las desviaciones con respecto a la media del grupo:

$$ESS = \sum_{i=1}^{n} x_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad [1]$$

Siendo x_i el valor de la variable x para el individuo i, y n el número de individuos dentro del grupo o clúster [22].

Como se puede apreciar en la Ecuación 1, la medida de distancia utilizada para el cálculo del método de Ward es la distancia euclídea, definida por la ecuación 2:

$$d_E(P,Q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_i - q_i)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad [2]$$

Siendo P y Q dos puntos, que pueden corresponder a dos registros o a dos clústeres, e i el número de dimensiones, que correspondería al número de variables.

Para determinar el número de clústeres se aplica el algoritmo *k-means* para identificar el valor a partir del cual la reducción de la suma total de cuadrados intra-cluster (Ecuación 1 aplicada a los registros dentro de un mismo clúster) deja de ser sustancial; a esta técnica se la suele conocer como método del codo o *elbow method.*

3.3. Aprendizaje supervisado

El aprendizaje supervisado es una serie de técnicas de Aprendizaje Automático en los que los algoritmos trabajan sobre datos en los que la variable respuesta es conocida, lo que quiere decir que están previamente clasificados.

Al utilizar estas técnicas, por lo general se cuenta con un conjunto de datos de entrenamiento, a partir del cual el algoritmo de aprendizaje supervisado crea un modelo predictivo de las variables respuesta a partir de los datos de entrada o variables explicativas, y un conjunto de datos de validación, utilizado para evaluar la eficacia del modelo creado.

En el presente trabajo se utilizarán diferentes técnicas de análisis supervisado, como el Análisis Discriminante o *Random Forest* como con el objetivo de predecir el éxito (Ganado o No ganado) de un equipo en un partido a partir de los datos de entrada explicados en el apartado 4 (Análisis de la base de datos).

3.3.1. Análisis Discriminante

El análisis discriminante es una técnica estadística que sirve para clasificar individuos u objetos según el grupo al que es más probable que pertenezcan. Esta probabilidad se establece a partir de la observación de diferentes variables, las cuales, a diferencia del grupo de pertenencia, deben ser directamente observables.

Los objetivos del análisis discriminante se pueden resumir en:

- Describir las características que distinguen a los individuos de un grupo.
- Clasificar a nuevos individuos en los grupos que ya están diferenciados [23].

El resultado del análisis discriminante lineal, el que se aplicará en este trabajo, es una función discriminante que será una combinación lineal de las variables discriminantes que minimice los errores de clasificación.

No obstante, para efectuar de forma correcta el análisis, debe partirse de una serie de supuestos:

1. Disponemos de una matriz que contiene una variable categórica, donde se recoge el grupo de pertenencia, y el resto de las variables son de intervalo o de razón.

2. Debe haber al menos dos grupos y cada uno de los grupos debe contener al menos dos individuos.

3. El número de variables discriminantes debe ser menor que el número de individuos menos 2. Esto es, si $(X_1, \dots, X_p)'$ es el vector de variables, tiene que verificarse que $p < (N-2)$, siendo N el número de objetos.

4. Ausencia de multicolinealidad. Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.

5. Igualdad de matrices de varianzas-covarianzas. Las matrices de varianzas-covarianzas dentro de cada grupo deben ser aproximadamente iguales.

6. Normalidad multivariante. Las variables deben seguir una distribución normal multivariante.

Como se ha dicho previamente, se quiere encontrar una función discriminante capaz de minimizar la variabilidad dentro de los grupos y maximizar la variabilidad entre los grupos que sea combinación lineal de las p variables de las que se dispone:

$$D = \omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_p X_p \qquad [3]$$

siendo D la función discriminante, ω los coeficientes correspondientes a cada variable, X el valor de cada variable y p el número de variables explicativas [24].

Dicho esto, el objetivo es hallar los coeficientes ω para cada una de las variables.

Con esta función discriminante puede llevarse a cabo la clasificación de un nuevo registro. Suponiendo que se quiere clasificar un individuo p entre dos grupos cuya observación viene dada por $x_0 = (x_0, \dots, x_0)'$. Entonces se calcula la función discriminante d_0, sustituyendo los valores correspondientes de las p variables en ella.

Se puede calcular la frontera discriminante a partir de la siguiente ecuación:

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2} \quad [4]$$

siendo:

$$\bar{D}_I = \omega_1 \bar{X}_1 + \cdots + \omega_p \bar{X}_p \quad [5]$$

$$\bar{D}_{II} = \omega_1 \bar{X}_1 + \cdots + \omega_p \bar{X}_p \quad [6]$$

De esta forma, se clasificará la observación en el grupo 1 si

$$d_0 < C \quad [7]$$

Y se clasificará la observación en el grupo 2 si

$$d_0 > C \quad [8]$$

3.3.2. Análisis discriminante de mínimos cuadrados parciales

El Análisis Discriminante de Mínimos Cuadrados Parciales (PLS-DA, por sus siglas en inglés, *Partial least Squares – Discriminant Analysis)* es un método de regresión lineal (PLS) que se combina con el análisis discriminante (DA).

Esta variante del método PLS, cuyo modelo puede verse en la Ecuación 9, utiliza el algoritmo de la regresión PLS para explicar y predecir la pertenencia de observaciones a varias clases, mediante la creación de una matriz Y que tiene tantas columnas como clases la base de datos. En cada columna, se asigna un valor de 1 a los individuos asociados a la clase ligada a dicha columna, y un valor de 0 al resto de individuos. Después, se construye el modelo como un PLS "normal".

$$\begin{aligned} X &= TP^T + E \\ Y &= TC^T + F \end{aligned} \quad [9]$$

Siendo X una matriz de dimensiones NxM de variables explicativas (M) y los registros (N), Y es una matriz MxP de variables respuesta (P) e individuos (M). T es la matriz de proyecciones o *scores*. P y C son las matrices de puntuaciones factoriales o *loadings* y E y F son las matrices de errores.

De esta forma, y tal y como se representa en la Figura 4, el modelo consta de una matriz X con los datos de las variables explicativa y una matriz Y conformada por las variables explicadas, y en la que habrá una variable explicada por cada clase posible. En este caso, al tener la variable explicada 2 posibles valores en función de si se ha ganado o no el partido, en la matriz Y habrá 2 variables respuesta distintas.

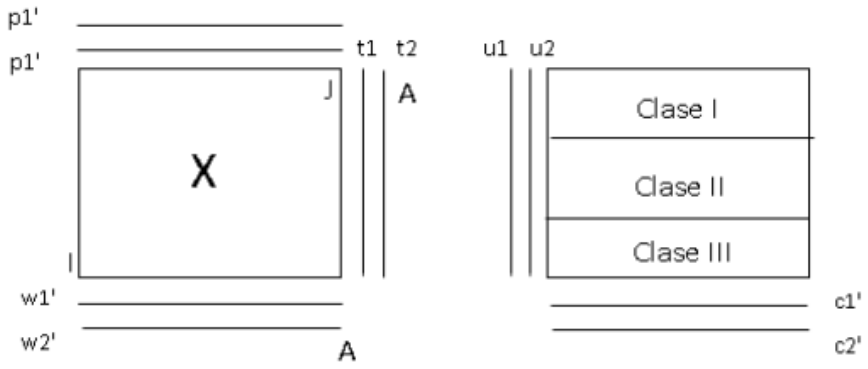

Figura 4. *Esquema gráfico de los elementos del modelo PLS-DA. Tomada de [25].*

3.3.3. Árboles de clasificación

El objetivo de este método de clasificación en crear un modelo que predice el valor de la variable explicada o clase en función de las variables de entrada. En la Figura 5 se puede ver un ejemplo de árbol de decisión de clasificación que sirve para decidir si se concede o no un préstamo en base a una serie de valores de entrada (años en el empleo actual, antecedentes penales…).

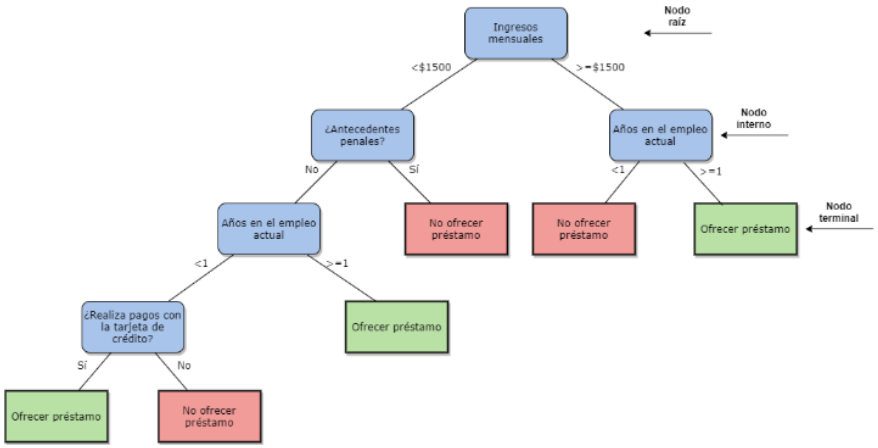

Figura 5. *Ejemplo de árbol de clasificación aplicado a un caso en el que se decide si se ofrece o no un préstamo. Tomada de [26].*

El aprendizaje basado en árboles de decisión es una de las técnicas más eficaces para la clasificación supervisada y se basa en la construcción de un árbol de decisión a partir de un conjunto de datos de entrenamiento, del que se conoce la clase a la que pertenecen los registros. La estructura de un árbol de clasificación es similar a la de un diagrama de flujo, donde cada nodo interno representa una decisión binaria a partir de la que obtener un resultado o avanzar hacia un nuevo nodo interno hasta alcanzar un nodo terminal que prediga la clase a la que pertenece el registro. El nodo superior en un árbol es el nodo raíz [27].

Las principales ventajas que ofrecen los árboles de clasificación se encuentran las siguientes:

- Permiten una fácil interpretación de los resultados.
- No necesitan mucho tratamiento de los datos, como la normalización o la eliminación de valores en blanco.
- Es una técnica capaz de manejar datos numéricos y categorizados.
- Buen funcionamiento con grandes conjuntos de datos, ya que el coste computacional es bajo en comparación con otras técnicas.

El principal problema de los árboles de clasificación es el sobreajuste o sobreentrenamiento, es decir, crear un árbol demasiado complejo que no generalice bien a partir de los datos de entrenamiento. Para evitar esta circunstancia se realiza una poda del árbol total obtenido en primera instancia en el punto donde minimice menos el error en las observaciones a partir de la validación cruzada o *xerror*.

Para realizar esta poda del árbol, el parámetro fundamental en el estadístico Cp, o *parámetro de complejidad*, el cual indica que durante la generación del árbol no se intente ninguna división que no reduzca la falta de ajuste global en, al menos, el valor del parámetro. Por ejemplo, si este parámetro de complejidad tiene un valor de 0,01, no se intentará ninguna división que reduzca la falta de ajuste global del modelo en, al menos, 0,01.

3.3.4. Random forest

El método Random Forest, desarrollado por Breiman y Adele Cutler [28], es un algoritmo predictivo de aprendizaje no supervisado que, por medio de la técnica de *Bagging* [29], combina diferentes árboles de clasificación. La diferencia con los árboles de clasificación es que, en este caso, cada árbol se genera a partir de un subconjunto de registros de la base de datos seleccionado al azar [30]. El proceso seguido por este algoritmo es el siguiente:

1. A partir de la base de datos completa, se seleccionan al azar registros para crear diferentes subconjuntos de datos.

2. Para cada subconjunto de datos, se genera un árbol de clasificación. Estos árboles, al generarse a partir de individuos diferentes, son distintos.

3. Para cada nodo del árbol, se eligen aleatoriamente un número de variables en las que basar la decisión. Estos árboles de dejan sin podar.

4. Una vez tenemos todos los árboles, a partir de un voto mayoritario se clasifican los registros de forma que, si la mayoría de los árboles lo clasifican como positivo o negativo, el algoritmo lo clasificará como positivo o negativo, respectivamente.

Este método, al igual que sucede con el análisis discriminante y los árboles de clasificación, permite hacer análisis más allá de la tasa de aciertos obtenida ya que, para cada una de las variables, se obtiene una medida de la importancia de esta variable en la predicción de los resultados y una medida de la estructura interna de los datos.

La primera de estas medidas es especialmente importante porque permite conocer las variables más importantes y, por lo tanto, reducir la dimensionalidad de los datos al permitirnos centrar nuestra atención en aquellas variables predictoras que más influyen y, por lo tanto, mejor clasifican.

3.3.5. Naive-Bayes

El método o clasificador Naive-Bayes o “Bayesiano ingenuo” es un clasificador probabilístico basado en el teorema de Bayes que se resume en la independencia entre las variables explicativas, es decir, que cada una de las características contribuye de manera independiente a la probabilidad de que un patrón de datos corresponda a una clase u otra, independientemente de las otras variables.

Se pueden encontrar 2 características clave que poseen los métodos bayesianos:

- Cada ejemplo observado modifica la probabilidad de que la hipótesis formulada sea correcta; es decir, la probabilidad estimada para la hipótesis disminuirá o aumentará en función de la mayor o menor coincidencia con el conjunto de datos utilizado.

- Se trata de un método robusto al ruido que pueda tener el conjunto de datos de entrenamiento y a la posibilidad de tener registros erróneos o incompletos.

La idea de usar el teorema de Bayes en un problema de aprendizaje automático es que se puede estimar las probabilidades a posteriori de cualquier hipótesis consistente con el conjunto de datos de entrenamiento para poder escoger la hipótesis más probable [31].

Desde un punto de vista matemático, si la descripción de un individuo viene dada por loa valores $(X_1, X_2, \dots, X_n)$, la hipótesis más probable será aquella que cumpla:

$$v_{MAP} = argmax_{v_i \in V} P(v_j | a_1, \dots, a_n) \quad [10]$$

Esta fórmula quiere decir que la probabilidad de que, conocidos los valores que describen a ese ejemplo, este pertenezca a la clase v_j (donde v_j es el valor de la función de clasificación en el conjunto finito V). A partir del teorema de Bayes:

$$v_{MAP} = argmax_{v_i \in V} \frac{P(a_1,\dots,a_n|v_j)P(v_j)}{P(a_1,\dots,a_n)} = argmax_{v_i \in V} P(a_1, \dots, a_n|v_j)P(v_j) \quad [11]$$

Para simplificar el cálculo de este término se recurre a la hipótesis de independencia condicional para factorizar la probabilidad: *Los valores a_j que describen un atributo de un ejemplo cualquiera son independientes entre sí conocido el valor de la categoría a la que pertenecen*. Así la probabilidad a la que pertenecen es el producto de las probabilidades de cada valor por separado [31]:

$$P(a_1, \dots, a_n|v_j) = \prod_i P(a_i|v_j) \quad [12]$$

Un problema que puede darse al aplicar la técnica de Naive-Bayes es el problema de la probabilidad cero, que sucede cuando el conjunto de datos de entrenamiento no contiene o contiene pocas muestras con una clase determinada de la variable respuesta. Esto provoca que, al aplicar el modelo obtenido a otro conjunto de datos, la probabilidad de que el modelo prediga esa clase sea nula o muy baja.

Para solucionar este problema se puede utilizar el suavizado de Laplace, el cual aumenta las probabilidades de las clases no presentes o poco presentes artificialmente para evitar este posible problema de probabilidad cero.

En el conjunto de datos del presente trabajo hay aproximadamente la mitad de los registros de partidos ganados que de partidos no ganados. Si bien no es un desbalance excesivo, se ha aplicado este suavizado para comprobar si los resultados mejoran o si, por el contrario, no tienen ningún efecto notable.

3.3.6. Máquinas de Soporte Vectorial (SVM)

Las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés de *Support Vector Machines*) son un conjunto de algoritmos de aprendizaje supervisado que, a partir de un conjunto de datos de entrenamiento, puede construir un modelo de prediga la clase a la que pertenece un nuevo individuo.

Su funcionamiento se basa en encontrar el mejor hiperplano que separa todos los puntos de dos clases diferentes, maximizando el margen entre las dos clases. Los vectores soporte son los puntos que marcan la separación en el hiperplano, como puede verse en la Figura 6.

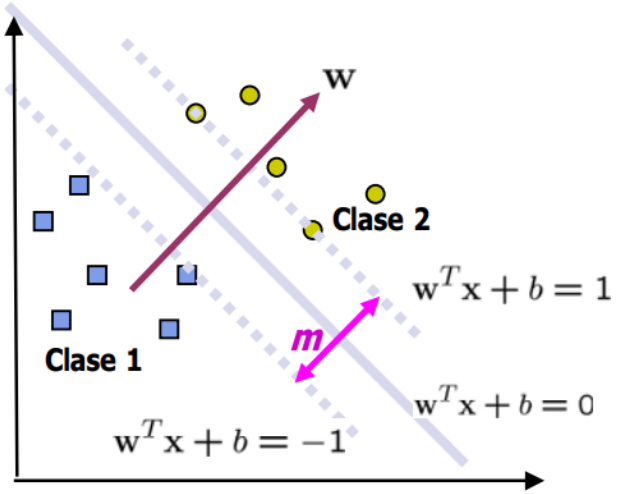

Figura 6. *Representación gráfica de un problema de clasificación 2D resuelto mediante SVM. Tomada de [32].*

Las máquinas de soporte vectorial pertenecen a la categoría de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos, ya sea en el espacio original del conjunto de entrenamiento, si estos son separables, o en el espacio transformado, si los ejemplos no son linealmente separables en el espacio original, como se puede ver en la Figura 6.

En concreto, la búsqueda de este hiperplano de separación en estos espacios transformados se hará utilizando las funciones *kernel* de forma implícita.

En lo que respecta a las fortalezas, las principales son que el entrenamiento es relativamente fácil, no hay óptimo local como sucede con las redes neuronales y se pueden usar datos como cadenas de caracteres como entrada. No obstante, también tienen una debilidad, y es que se necesita una buena función *kernel*, es decir, se necesitan metodologías eficientes para sintonizar los parámetros de inicialización de la SVM [32].

Los principales parámetros configurables de las máquinas de soporte vectorial son los siguientes:

- **Kernel:** se trata de un conjunto de funciones matemáticas utilizada por el algoritmo de las máquinas de soporte vectorial. Hay diferentes posibilidades (radial, lineal, sigmoide…) con las que se puede configurar este parámetro.

- **Coste:** este parámetro indica el coste de realizar una clasificación errónea de cada ejemplo del subconjunto de datos del entrenamiento.

- **Gamma:** este parámetro define hasta dónde llega la influencia de un solo subconjunto de entrenamiento, de forma que cuanto más alto sea el valor, menor influencia tendrá. Puede entenderse como la inversa del radio de influencia de las muestras seleccionadas por el modelo como vectores de soporte.

3.3.7. Entrenamiento y validación

En todos los análisis predictivos se dividirá el conjunto de datos original en dos subconjuntos de datos diferentes: el subconjunto de entrenamiento y el subconjunto de validación.

El primero de ellos, formado por el 70% de los registros seleccionados de forma aleatoria, se utiliza para entrenar un modelo predictivo; y el segundo de ellos, conformado por el restante 30% de los datos, se utiliza para evaluar el modelo generado con registros no utilizados para la construcción de este.

4. Análisis de la base de datos

Este apartado contiene la descripción de la base de datos con la que se cuenta, así como la explicación de la selección de variables y el tratamiento de los datos que se realiza.

4.1. Descripción de la base datos

La base de datos inicial cuenta con datos detallados de todos los partidos de la temporada 2018-2019 de las consideradas 5 grandes ligas europeas: Serie A italiana, La Liga española, Premier League inglesa, Bundesliga alemana y Ligue 1 francesa.

De cada uno de estos partidos tenemos, en dos registros distintos, las estadísticas de cada equipo competidor. En concreto, en cada registro encontramos la variable cualitativa GANADO, que tiene valor 1 si el equipo ha ganado el partido. Además, encontramos una gran cantidad de variables cuantitativas que reflejan aspectos del juego (minutos de posesión, número de ataques posicionales, número de contraataques…).

En total tenemos 3650 registros y 68 variables (1 cualitativa y 67 cuantitativas), cuya descripción puede verse en el Anexo 1. Estos registros son considerados independientes, a pesar de que los partidos se jueguen por los mismos partidos y tengan un orden específico.

4.2. Selección de variables

Teniendo el objetivo inicial de detectar estrategias a partir de los datos, se seleccionaron las variables siguiendo una serie de criterios:

- No tener en cuenta goles marcados o recibidos, ya que el objetivo a analizar es el éxito del partido y estas dos variables estarían altamente correlacionadas con ello.
- Descartar variables de éxito. Existen grupos de 3 variables relacionadas entre ellas (p. ej. Ataques, ataques efectivos y porcentaje de ataques efectivos), de entre las cuales 2 de ellas dependen de la efectividad del equipo más que de la estrategia elegida. Por lo tanto, las variables que dependan del éxito a la hora de llevar a cabo la estrategia se descartan.
- Variables altamente correlacionadas entre sí o que sean combinación lineal entre ellas (p. ej. La posesión es la suma de la posesión en campo propio y la posesión en campo contrario).

Siguiendo este criterio, las variables seleccionadas para hacer el análisis son las siguientes:

Faltas
Posesión campo propio
Posesión campo contrario
Tiros
Ataques posicionales
Contraataques
Pases
Pases en profundidad y de finalización
Centros
Disputas por arriba
Regates
Entradas
Intercepciones

Intercepciones en campo rival
Rechaces
Rechaces en campo contrario
Pérdidas
Pérdidas en campo propio
Recuperaciones en campo propio
Recuperaciones en campo contrario

Tabla 1. *Variables seleccionadas para los análisis posteriores.*

4.2.1. Análisis descriptivo y estudio de la normalidad de las variables

A continuación, se estudia la normalidad de las variables a través del coeficiente de asimetría y el coeficiente de curtosis. Para apoyar este análisis se muestran algunos de los estadísticos clave de cada una de las variables: la media, la desviación típica, el valor mínimo y el valor máximo.

El coeficiente de asimetría analiza la proximidad de los datos a la media de estos. Un valor de 0 en este coeficiente indica que la distribución en perfectamente simétrica, mientras que un valor superior indica que los datos tienen una asimetría positiva (desplazado hacia valores mayores que la media) y un valor inferior indica una asimetría negativa (desplazado hacia valores menores que la media).

Con respecto al coeficiente de curtosis, este es una medida de forma que mide cuán escarpada o achatada es una distribución. Un valor de 0 indica una distribución normal, un valor superior a 0 indica una distribución más apuntada y un valor inferior a 0 indica una distribución menos apuntada.

Variable	Media	Desviación típica	Mínimo	Máximo	Coeficiente de asimetría	Coeficiente de curtosis
Faltas	12,35	4,024	0	28	0,353	0,111
Posesión campo propio	651,79	236,896	91,31	1749,48	0,279	0,427
Posesión campo contrario	974,9	329,766	124,3	2461,0	0,594	0,400
Tiros	12,06	4,952	0	43	0,653	0,755
Ataques posicionales	62,3	10,977	28	105	0,130	-0,062
Contraataques	13,66	4,287	2	32	0,365	0,124
Pases	482,5	124,788	199	1075	0,663	0,583
Pases en profundidad y de finalización	12,35	6,660	0	49	1,010	1,492
Centros	13,47	6,687	0	45	0,915	1,090
Disputas por arriba	44,67	15,854	8	122	0,603	0,486
Regates	26,17	8,066	5	69	0,533	0,529
Entradas	33,56	8,928	9	77	0,463	0,370
Intercepciones	48,08	10,702	17	104	0,417	0,410
Intercepciones en campo rival	9,705	4,113	0	27	0,550	0,397
Rechaces	63,91	12,102	29	120	0,346	0,218
Rechaces en campo contrario	23,19	8,164	3	59	0,441	0,089

Pérdidas	71,16	8,958	37	103	0,075	0,060
Pérdidas en campo propio	14,31	5,499	1	37	0,474	0,168
Recuperaciones en campo propio	44,31	7,185	23	72	0,219	0,151
Recuperaciones en campo contrario	9,332	4,193	0	29	0,611	0,322

Tabla 2. *Tabla resumen de los estadísticos de las variables seleccionadas (media, desviación típica, mínimo, máximo, coeficiente de asimetría y el coeficiente de curtosis).*

Como se puede ver en la Tabla 2, tanto el coeficiente de asimetría como el coeficiente de curtosis tienen, por lo general, valores ligeramente positivos. Con estos valores, si bien indican que la distribución de la mayoría de las variables es ligeramente apuntada y con asimetría positiva, puede decirse que las variables siguen una distribución cercana a una distribución normal.

Esta ligera asimetría positiva queda explicada por el hecho de que, por naturaleza, los datos tienen un límite inferior de 0 (no puede haber un número de tiros o un tiempo negativo de posesión), mientras que puede haber datos especialmente altos de estas estadísticas.

Para finalizar, cabe mencionar que se ha analizado la normalidad de estas variables mediante el test de Kolmogorov-Smirnov; no obstante, debido al alto número de registros el test sale significativo a pesar de que, como se puede ver en los coeficientes de asimetría y curtosis, no existe tanta diferencia con una distribución normal.

5. Análisis de resultados

A lo largo de este apartado se presentan los resultados obtenidos al intentar dar respuesta a los objetivos planteados en el apartado correspondiente, desde la obtención de las estrategias hasta el enfrentamiento entre las diferentes estrategias.

5.1. Objetivo 1: Identificar perfiles de estrategia a partir de los datos de juego

El objetivo de esta sección es analizar las estrategias de juego de las grandes ligas europeas de fútbol. Para identificar estas estrategias, se utilizará un análisis de componentes principales con las variables seleccionadas en el apartado anterior.

En primer lugar, y teniendo en cuenta que se tienen variables con medias muy distintas que podrían añadir errores en los resultados, normalizamos las variables de forma que tengan desviación típica de 1 y media 0 (estandarización a *z-score*).

Posteriormente, se analiza la correlación entre las variables seleccionadas para analizar relaciones entre ellas. A pesar de que se detectan varias correlaciones fuertes (de hasta 0,6) entre algunos pares de variables, como puede verse en la Figura 6, se decide no descartar ninguna otra variable partiendo de la hipótesis es que pueden formar parte de estrategias de juego distintas.

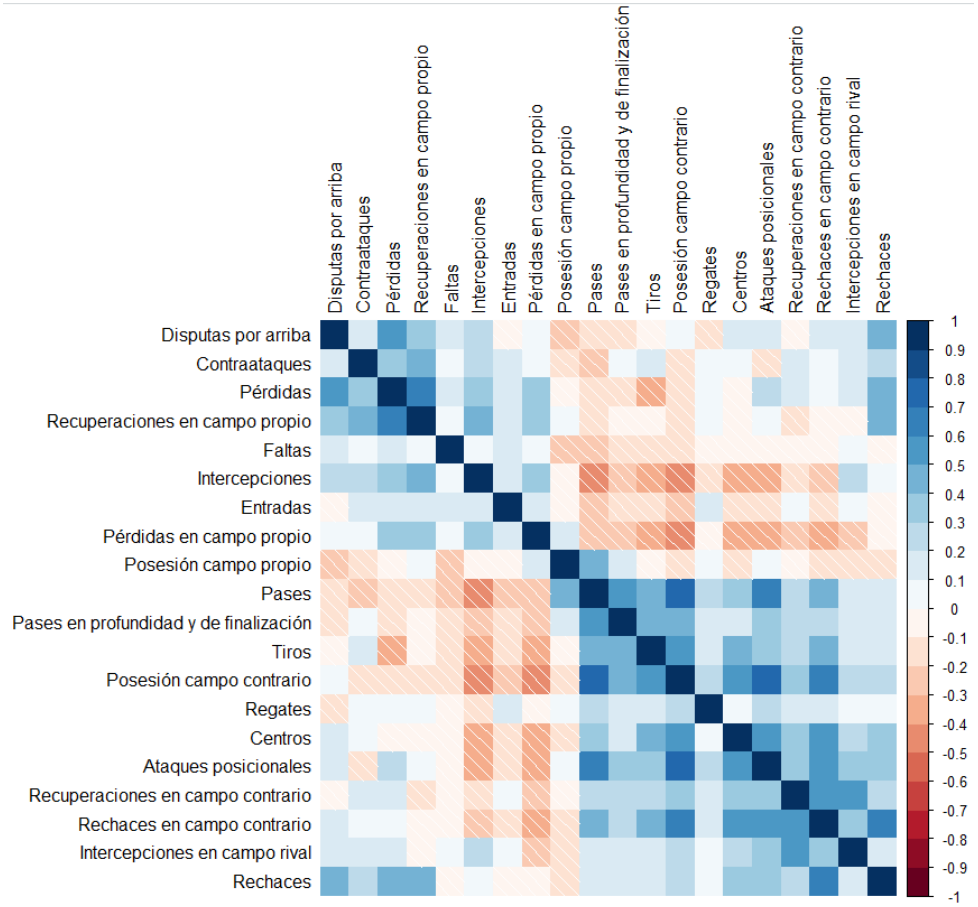

Figura 7. *Matriz de correlaciones de las variables seleccionadas, utilizando un código de colores.*

Este cierto grado de correlación existente, que puede verse en la Figura 6, demuestra la adecuación del análisis de componentes principales en el siguiente análisis.

5.1.1. Componentes principales

Una vez se han estandarizado las variables y analizado las correspondencias entre ellas, se realiza el análisis de componentes principales haciendo uso de la función *princomp()*.

Atendiendo a los valores propios de cada componente principal, y siguiendo al criterio de Kaiser, sería conveniente seleccionar 6 componentes principales.

Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7	Comp. 8
2,300	1,806	1,353	1,201	1,099	1,041	0,905	0,901

Tabla 3. *Valores propios correspondientes a las 6 primeras componentes principales.*

No obstante, al analizar las puntuaciones factoriales (*loadings*) de cada variable en cada componente principal para asignar una estrategia de juego a cada una de las componentes principales, reflejadas en la Tabla 3, se puede ver cómo las componentes 4 y 5 corresponderían a la misma estrategia de juego. Al mismo tiempo, la componente 6, definida por 3 variables no refleja una estrategia, sino el hecho de tener una alta intensidad y muchas individualidades dentro del terreno de juego.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Faltas			**0.343**			**0.408**
Posesión campo propio		-0.179	**-0.539**		-0.213	
Posesión campo contrario	**-0.386**					
Tiros	-0.297			-0.125	**0.371**	-0.186
Ataques posicionales	**-0.345**	-0.110	-0.110	-0.155	-0.190	0.222
Contraataques		**0.314**		**0.348**	**0.425**	-0.211
Pases	**-0.330**	-0.134	**-0.332**		-0.125	
Pases en profundidad y de finalización	-0.232		-0.237	-0.267	0.209	-0.221
Centros	-0.297		0.155	0.152		
Disputas por arriba		**0.384**		**0.364**		
Regates	-0.128		-0.225	-0.289	0.106	**0.499**
Entradas	0.120			**-0.416**	0.125	**0.476**
Intercepciones	0.219	0.263		-0.233	-0.221	-0.293
Intercepciones en campo rival	-0.167	0.195	0.159	**-0.352**	**-0.478**	-0.196
Rechaces	-0.168	**0.377**	-0.101	-0.185		
Rechaces en campo contrario	**-0.337**	0.189				
Pérdidas		**0.453**	-0.174		-0.188	0.167
Pérdidas en campo propio	0.256	0.114	**-0.320**		-0.104	0.115
Recuperaciones en campo propio		**0.381**	**-0.344**		0.260	
Recuperaciones en campo contrario	-0.245	0.120	0.163	**0.353**	**-0.316**	

Tabla 4. *Scores de todas las variables incluidas en el estudio en cada una de las 6 componentes principales que cumplen con el criterio de Kaiser.*

En la Tabla 4 puede verse el gráfico de las puntuaciones factoriales de las componentes 1 y 2. Como puede observarse, apoyándose en la Tabla 5 como resumen de las variables representativas de cada componente, con valores altos en la **componente 1** (en **azul** en la Figura 8) y cercanas entre sí, pueden verse las variables "Posesión en campo contrario", "Ataques posicionales" y "Pases". Estas reflejan una estrategia basada en mantener el balón más allá de la línea del medio campo (por parte de los mediocampistas especialmente) y crear ataques basados en un gran número de pases y poco verticales para crear huecos en la defensa rival y aprovechar la oportunidad o llegar poco a poco a la portería rival.

Con respecto a las variables con valores altos en la **componente 2** (en **verde** en la Figura 8), se puede ver un grupo de variables formado por "Contraataques", "Pérdidas" y "Recuperaciones en campo propio"; estas variables reflejan una estrategia basada en la defensa cerca de la portería propia y, al recuperar el balón arriesgar con los pases (de ahí el valor alto en las pérdidas) para conseguir forma contraataques rápidas hacia la portería rival.

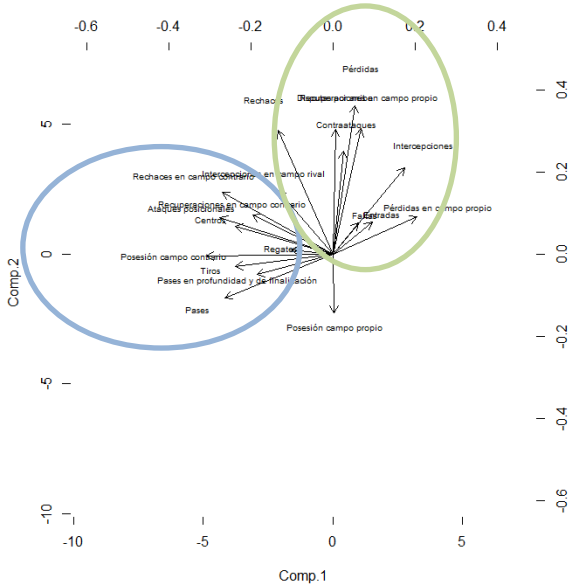

Figura 8. *Gráfico de las puntuaciones factoriales de las componentes principales 1 y 2.*

Más allá de estos dos grupos de variables, y a partir de la Figura 9, se puede ver la relación de otros dos grupos de variables. Por un lado, y atendiendo a la **componente 3** (en **naranja** en la Figura 8), puede verse el grupo formado por las variables "Posesión en campo propio", "Pérdidas en campo propio", "Recuperaciones en campo propio" y "Pases" en el espacio de los valores negativos de la componente 3 y cercanos al 0 en la componente 4. Este grupo de variables reflejan una estrategia en la que el equipo mantiene la posesión del balón cerca de su portería (por parte de los defensas y mediocampistas) y creen ataques desde muy lejos de la portería rival (de ahí que las recuperaciones y pérdidas se den en campo propio).

Por otro lado, atendiendo a la **componente 4** (en **rojo** en la Figura 9), puede verse en el espacio de los valores negativos de la componente 4 cómo las variables "Contraataques", "intercepciones en campo rival" y "recuperaciones en campo contrario" se encuentran cercanas y en direcciones similares. El conjunto de estas variables refleja una estrategia basada en presionar al equipo rival lejos de su portería por medio de los delanteros y mediocampistas, intentar recuperar el balón (lo que explica las variables de recuperaciones e intercepciones) y generar un ataque rápido.

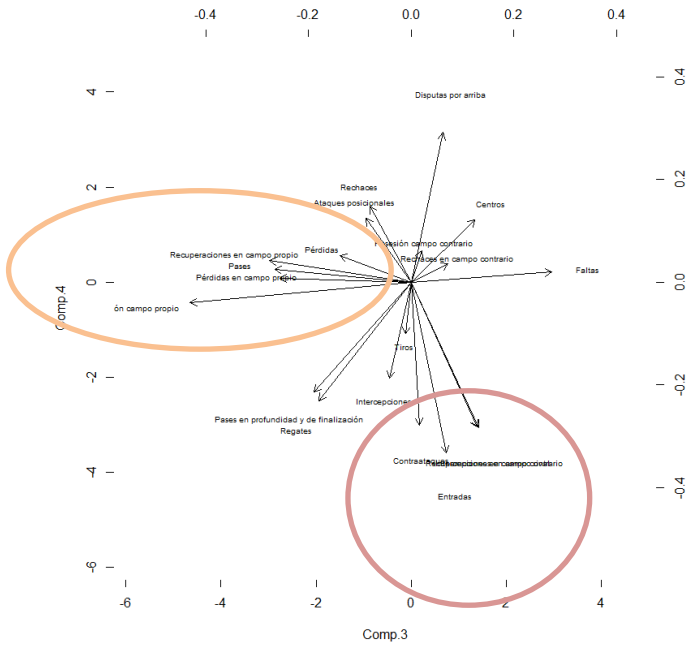

Figura 9. *Gráfico de las puntuaciones factoriales de las componentes principales 3 y 4.*

Teniendo esto en cuenta, se decide continuar el trabajo con las 4 primeras componentes principales. Para poder explicar las estrategias que representan cada una de componente principal ellas, se destacan las variables que más peso tienen dentro de cada una de ellas.

	Variables	**Estrategia**
Comp. 1	Posesión campo contrario, Ataques posicionales, pases, rechaces en campo contrario	Posesión en campo contrario
Comp. 2	Contraataques, Recuperaciones en campo propio, disputas por arriba, rechaces, pérdidas	Defensa en campo propio y contraataques
Comp. 3	Posesión en campo propio, recuperaciones en campo propio, pases, faltas, pérdidas en campo propio	Posesión en campo propio
Comp. 4	Entradas, Disputas por arriba, intercepciones en campo rival, recuperaciones en campo contrario, contraataques	Presión en campo contrario y contraataques

Tabla 5. *Variables principales de las componentes principales estudiadas y estrategia representada por cada una.*

En vista a estas variables, y a modo de resumen, las estrategias predominantes que representaría cada una de ellas sería:

- Componente 1: posesión en el campo contrario.
  - Esta estrategia define a un equipo que ha jugado controlando el balón y buscando un ataque desde el campo rival, cerca de la portería contraria.
- Componente 2: defensa en campo propio y contraataques rápidos.
  - Esta estrategia define a un equipo que ha jugado atrás, defendiendo en campo propio y saliendo al contraataque para encontrar al rival mal organizado.
- Componente 3: posesión en campo propio.
  - Esta estrategia define a un equipo que ha jugado controlando el balón cerca de su propia portería y atacando a partir de una posesión prolongada.
- Componente 4: presión al equipo rival en campo contrario y contrataques rápidos.
  - Esta estrategia definiría a un equipo cuya estrategia es presionar al rival para impedirle que se acerque a su portería y salir al contraataque al conseguir un robo (intercepción o recuperación) en campo rival. Los altos valores en entradas y disputas por arriba reflejan la intensidad a la que juega este tipo de equipo.

5.1.2. Análisis Clúster

Una vez identificadas las principales estrategias de juego presentes en los datos por medio del PCA, el siguiente paso lógico es clasificar los partidos en clústeres para poder analizar el éxito de las estrategias definidas previamente por la relación entre estas y los clústeres.

Para llevar a cabo este análisis clúster no se utilizan las variables originales, si no que se utilizan las componentes principales analizadas apartado anterior. En concreto, se utilizarán las 4 primeras componentes principales ya que, como se ha explicado en el anterior apartado, son aquellas que representan claramente una estrategia definida.

Al final de este análisis se espera tener la relación de partidos ganados y no ganados por cada uno de los clústeres, así como las puntuaciones de los clústeres para cada una de las componentes; de esta forma, se podrá analizar el éxito general de las estrategias.

5.1.2.1. Análisis clúster jerárquico

Para llevar a cabo el análisis clúster jerárquico se utiliza el método Ward puesto que tiende a formar clústeres más compactos y de igual tamaño y forma que otros métodos de agrupación como el método de la media, o del centroide.

Este método, que utiliza las distancias euclídeas, agrupa los registros minimizando la varianza, o la suma de los cuadrados de error, dentro de cada uno de los clústeres. De esta forma obtenemos el dendograma de la Figura 10, que determina la división en los diferentes clústeres.

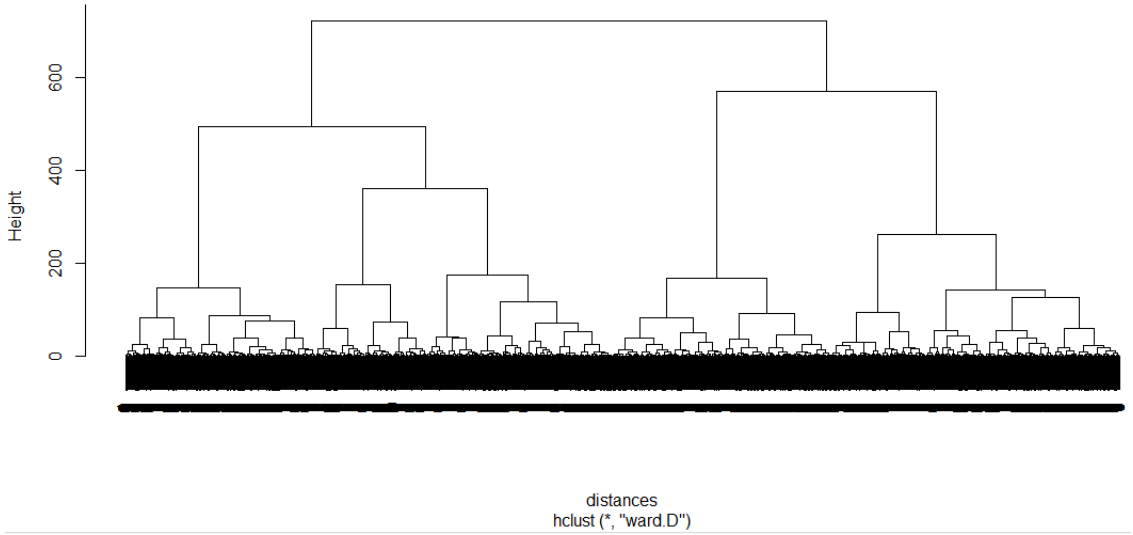

Figura 10. *Dendograma obtenido a partir de la función "hclust", utilizando el método Ward y distancias euclidianas.*

No obstante, y puesto que el análisis clúster jerárquico no determina el número óptimo de clústeres, en este punto es necesario decidir el número de clústeres a analizar.

5.1.2.2. Número óptimo de clústeres

Para obtener este número óptimo de clústeres a analizar, utilizamos el algoritmo *k-means* para identificar el valor a partir del cual la reducción de la suma total de cuadrados intra-cluster deja de ser sustancial.

Este algoritmo está basado en el método del codo, o *elbow method*, en el cual se minimiza la suma total de cuadrados intra-cluster. El resultado de este método es una gráfica en la que se muestra cómo la suma total de cuadrados disminuye considerablemente menos a partir de un

número determinado de clústeres, por lo que ese número será el valor de referencia a la hora de seleccionar el número óptimo de clústeres.

El código del algoritmo puede verse en el Anexo 2, y en la Figura 11 puede verse la gráfica que resulta de aplicar este método al caso que nos ocupa.

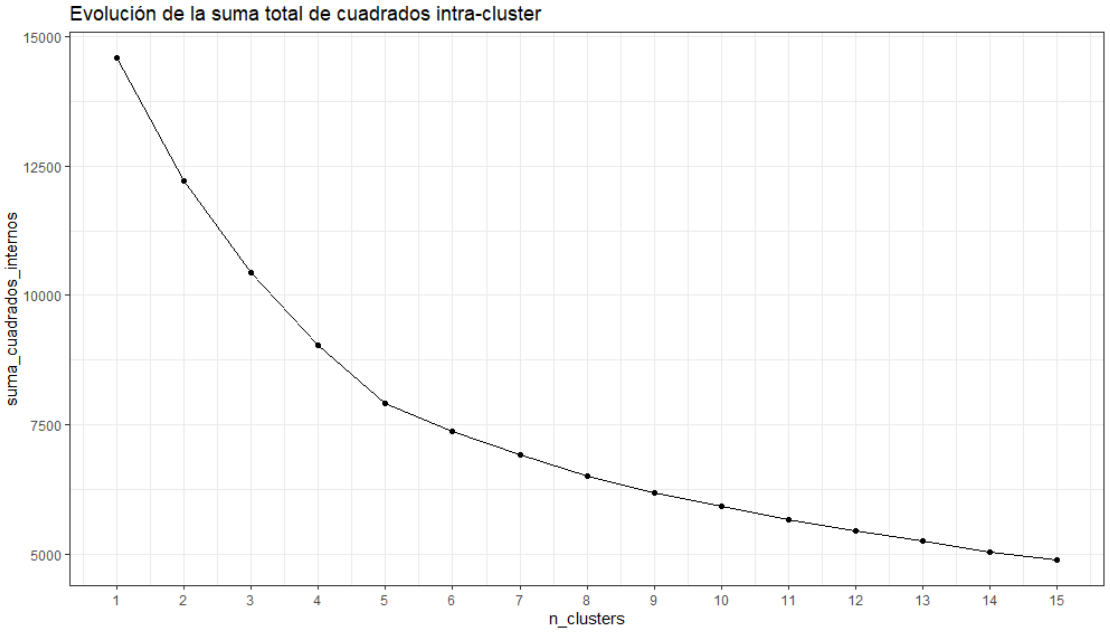

Figura 11. *Gráfica de la suma total de cuadrados intra-cluster frente al número de clústeres.*

En vista a la Figura 11, y atendiendo al codo que se forma en esta progresión, se puede deducir que el número óptimo de clústeres serían 4 o 5. Para determinar cuál de los 2 es más adecuado en este caso, se observa el número de registros que se obtendrían en cada clúster en cada uno de los casos.

El resultado que se obtiene es el reflejado en la Tabla 6.

	Solución con 4 clústeres	**Solución con 5 clústeres**
Clúster 1	801	801
Clúster 2	1062	1062
Clúster 3	699	699
Clúster 4	1088	678
Clúster 5		410

Tabla 6. *Reparto de registros utilizando 4 y 5 clústeres.*

Como puede verse, al utilizarse 5 clústeres los grupos se encuentran muy desbalanceados, incluso habiendo un clúster con más del doble de registros que otro clúster. Por el contrario, al utilizar 4 clústeres el balance entre ellos es mucho mejor a pesar de las lógicas diferencias que se encuentran.

5.1.2.3. Estrategias representativas de los clústeres

Una vez decidido el número de clústeres y hecha la asignación de los registros a éstos, es el momento de analizar la estrategia predominante representada por cada uno de los clústeres.

Para ello, se analizan los *scores* de cada uno de los clústeres sobre cada una de las componentes principales que se pueden ver en la Tabla 7.

Clúster	Componente 1	Componente 2	Componente 3	Componente 4
1	1,639	-1,461	0,540	0,180
2	0,946	0,770	-1,020	0,629
3	-0,509	1,143	1,374	-0,004
4	-1,803	-0,411	-0,284	-0,745

Tabla 7. *Scores de cada uno de los clústeres en cada una de las componentes principales.*

Con estos datos, podemos describir las estrategias de los diferentes clústeres de la siguiente forma:

El clúster 1 se caracteriza por tener un valor alto (el mayor de todos los clústeres) en la componente 1, por lo que se caracteriza especialmente por ser registros de partidos en los que el equipo se ha centrado en tener mucha posesión en el cerca de la portería rival y crear ataques lentos con el balón buscando huecos en la defensa rival.

También cuadra el valor negativo en la componente 2, ya que esta componente representa una estrategia basada en defender en campo propio y salir al contraataque, estrategia opuesta a la estrategia representada por la primera componente.

El clúster 2 tiene el valor máximo de los clústeres en la componente 4; por lo tanto, podría decirse que se trata de registros caracterizados por ejercer presión al equipo rival en el campo contrario e intentar recuperar el balón para hacer un ataque rápido.

También obtiene un valor alto, pero no máximos, en las componentes 1 y 2. Esto puede ser explicado y coherente con los resultados porque en la componente 1 tiene mucho peso la posesión en el campo contrario y en la componente 2 tienen mucho peso los contrataques o ataques rápidos, siendo dos características compartidas o cercanas a aquella estrategia representada por la componente 4.

El clúster 3 tiene un valor alto, y máximo, tanto en la componente 2 como en la componente 3. Esto puede indicar que estos registros se caracterizan con partidos en los que el equipo ha defendido en su propio campo (cerca de su propia portería) y, cuando ha conseguido recuperar el balón, o bien ha salido al contraataque o bien ha preferido mantener la calma, controlar el balón y construir un ataque posicional.

El clúster 4 no tiene valores positivos en ninguna de las componentes principales. Esto indica que son registros sin una estrategia clara o consolidada. Podría indicar partidos donde el equipo no se limita a centrarse en su estrategia, si no que adapta su juego a las diferentes situaciones que se dan a lo largo del encuentro.

5.1.2.4. Éxito general de las estrategias

Por último, y como paso previo al análisis en detalle del éxito de las estrategias, se analiza el éxito general de las estrategias. En la Tabla 8 se puede ver la proporción de partidos ganados y no ganados por cada uno de los clústeres.

	GANADO	
Clúster	**0**	**1**
1	0,727	0,273
2	0,638	0,362
3	0,711	0,289
4	0,496	0,504

Tabla 8. *Proporción de los partidos ganados (GANADO=1) y no ganados (GANADO=0) por clúster.*

Se evalúan los resultados reflejados en la tabla, aplicamos una prueba de chi-cuadrado de Pearson a las proporciones de la Tabla 8. El p-valor obtenido es de 0,986, por lo que no puede rechazarse la hipótesis nula (que asume la igualdad entre las proporciones) y decirse que las proporciones obtenidas sean significativamente diferentes entre sí.

A partir de estos datos se puede ver, a nivel descriptivo, cómo el clúster 4 es el clúster más exitoso en términos generales, y cabe recordar que este clúster no se caracteriza por ninguna de las estrategias representadas por las componentes principales. Por lo tanto, podría deducirse de estos resultados que no existe una estrategia cerrada más exitosa que el resto, si no que el éxito proviene de que el equipo sepa adaptarse a las diferentes situaciones que se dan a lo largo del partido.

5.2. Objetivo 2: Analizar la capacidad predictiva de las estrategias de juego en el éxito del partido

En el apartado anterior se identificaron los clústeres en base a las estrategias predominantes, pero la variable protagonista del objetivo 2, GANADO, no se incluyó ya que el análisis se centró la relación de esta variable con los clústeres de forma descriptiva. Debido a que en este apartado quedó de manifiesto que no existe un clúster claramente superior, este apartado pondrá el foco el analizar la capacidad predictiva de las estadísticas de juego originales en el partido y no de los clústeres o las componentes principales del primer apartado.

En las siguientes páginas se utilizarán diferentes técnicas de análisis de datos para predecir, en base a las estadísticas originales de un equipo en un partido, el éxito del equipo en éste.

El enfoque estudiado pone el foco en el partido, ya que se utilizarán los datos de este para intentar predecir el éxito del equipo en cada partido, siendo la variable a predecir GANADO. GANADO es una variable categórica que indica si el equipo ha ganado (GANADO = 1) o no ha ganado (GANADO = 0) el partido. Con respecto a las variables explicativas, en este análisis se utilizará la misma selección de variables que se han utilizado para el objetivo 1 por las razones expuestas en el Apartado 4.2 (Selección de variables).

Para esta predicción se utilizarán diferentes técnicas de predicción para obtener la mejor predicción posible y poder analizar cuál es la más adecuada para realizar estas predicciones. Además de la predicción general, incluyendo los partidos de las 5 ligas, también se llevarán a cabo las predicciones por liga para saber si alguna de las ligas resulta ser más “predecibles” que otras o si no existe una diferencia clara entre ellas.

5.2.1. Análisis Discriminante

En primer lugar, se lleva a cabo un análisis discriminante, en el que podremos analizar tanto el éxito en la predicción como los coeficientes lineales de las diferentes variables.

A partir de este análisis se analizan dos aspectos de los resultados: el éxito en la predicción y los pesos de las variables en ésta.

Éxito general en la predicción

Como se puede ver en la Figura 12 se analiza la curva ROC obtenida a partir de los resultados. En esta el área bajo la curva que se obtiene es de 0.79, lo que indica que, teniendo un registro de partido ganado y un registro de partido no ganado, la probabilidad de que los clasifique correctamente es del 79%.

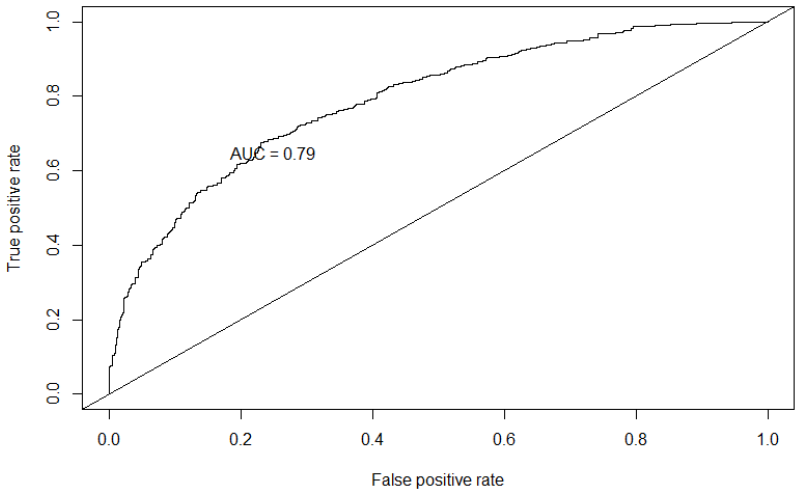

Figura 12. *Curva ROC obtenida a partir del Análisis Discriminante.*

Además de la curva ROC, en la figura 13 se puede ver la gráfica que relaciona el resultado del modelo del análisis discriminante con el éxito del partido. En este se puede ver cómo, a pesar de haber solapamiento entre los dos posibles valores, se ve claramente cómo el valor medio del modelo para un registro de partido ganado es superior al valor para un registro de partido no ganado.

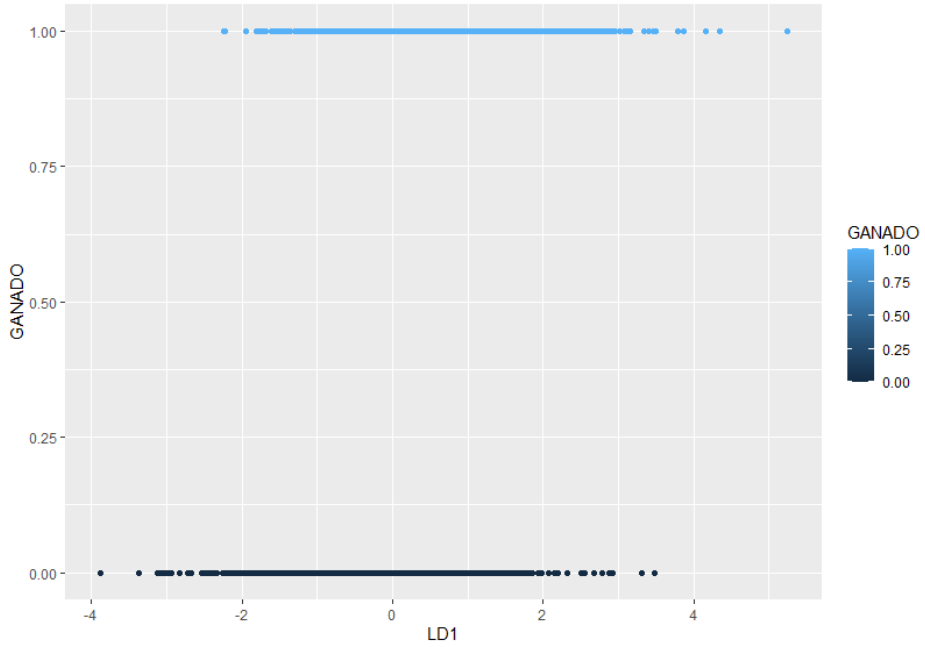

Figura 13. *Gráfico que relaciona el valor de la variable GANADO (0 para partidos no ganados o 1 para partidos ganados) con el valor de la función discriminante.*

En la predicción llevada a cabo con el set de datos de prueba, la matriz de confusión obtenida puede consultarse en la siguiente tabla:

		Valores reales	
		No ganado	Ganado
Valores predichos	No ganado	626	182
	Ganado	101	220

Tabla 9. *Matriz de confusión obtenida a partir del Análisis Discriminante.*

Con estos resultados se obtiene un 74.93% de aciertos, con una tasa de aciertos de partidos ganados del 54.72% y una tasa de aciertos de partidos no ganados del 86.1%. Teniendo en cuenta que en los valores reales del conjunto de prueba el porcentaje de partidos es de un 36% y el de no ganados un 64%, los resultados tienen un acierto 18 puntos porcentuales mayor que el azar.

Por lo tanto, en base a estos resultados no puede decirse que sean unos resultados especialmente buenos, aunque sí es cierto que el modelo tiene cierta capacidad predictiva. Teniendo en cuenta los resultados del Objetivo 1 del presente trabajo, el ajuste no es excelente para la predicción de partidos ganados debido a la elevada incertidumbre asociada a esta variable objetivo.

Peso de las variables en las predicciones

Atendiendo a los coeficientes discriminantes lineales de cada una de las variables, puede verse las que mayor peso tienen en las predicciones, que serán aquellas que mayor valor absoluto tengan. Teniendo en cuenta que aquellas con valor positivo influirán más a la hora de predecir los partidos ganados (GANADO = 1) y aquellos con valor negativo serán aquellos que más influirán en predecir los partidos no ganados (GANADO = 0), las variables de mayor peso son las siguientes:

Partidos ganados	Partidos no ganados
Posesión en campo propio **(LD: 0,392)**	Centros **(LD: -0,562)**
Tiros **(LD: 0,420)**	Pérdidas **(LD: -0,525)**
Pases en profundidad y de finalización **(LD: 0,541)**	
Recuperaciones en campo propio **(LD: 0,336)**	

Tabla 10. *Variables con mayor peso a la hora de predecir los partidos ganados (izquierda) y los partidos no ganados (derecha) a partir del análisis discriminante y sus coeficientes.*

Atendiendo a las variables de peso para los partidos ganados, éstas son consistentes con los resultados del objetivo 1 ya que no corresponden a ninguna estrategia clara.

Con respecto a las variables que más influyen en los partidos no ganados son el número de pérdidas, lo cual indudablemente tiene sentido, y el número de centros (se consideran como centros los pases aéreos desde una banda del campo de juego hacia el área); esta última, aunque a priori podría no tener sentido lógico directo, es cierto que, de entre todos los tipos de ataques, es el ataque que menos eficacia tiene a la hora de meter un gol.

Resultados de las predicciones

Por último, los resultados generales y para las diferentes ligas son los siguientes:

	Predicción general	% acierto en partidos ganados	% acierto en partidos no ganados
General	74.93%	54.72%	86.11%
España	67.76%	45.88%	80.71%
Reino Unido	73.33%	59.52%	81.56%
Alemania	73.37%	61.53%	79.83%
Italia	70.67%	51.11%	82.7%
Francia	72.89%	43.84%	86.84%

Tabla 11. *Resultados generales y por ligas obtenidos a partir del Análisis Discriminante.*

Si bien en la mayoría de las ligas las variables con más peso son las mismas que las obtenidas a partir del conjunto general de datos, profundizaremos en los resultados de la liga española.

En la liga española destacan, para la predicción de los partidos ganados, el número de intercepciones totales, y para la predicción de los partidos no ganados, el número de intercepciones en campo del rival. Este último valor sorprende a priori pero puede estar explicado por la posición de los equipos en liga, ya que el FC Barcelona y el Atlético de Madrid, equipos que suelen defender en su propio campo, fueron los dos partidos con más victorias.

A partir de estos resultados se puede ver que los mejores resultados de predicción general se tienen con el conjunto de datos de todas las ligas, aunque sean valores bastante cercanos excepto para la liga española, la cual tiene las peores predicciones con diferencia. Sin embargo, el mejor resultado de predicción para los partidos ganados la tiene la liga alemana (7 puntos porcentuales mayor que la predicción general) y la liga francesa la mejor predicción para los partidos no ganados, aunque similar al resultado general. al.

5.2.2. Análisis discriminante de mínimos cuadrados parciales

Utilizando la función *plsda* [13], configurando 2 componentes principales debido a los 2 niveles que la variable respuesta puede tener y aplicando validación cruzada, obtenemos los gráficos necesarios para analizar la importancia de las variables, así como las predicciones generales para cada una de las ligas.

Importancia de las variables

A continuación, en la Figura 14 se representan las puntuaciones factoriales de las variables para cada componente.

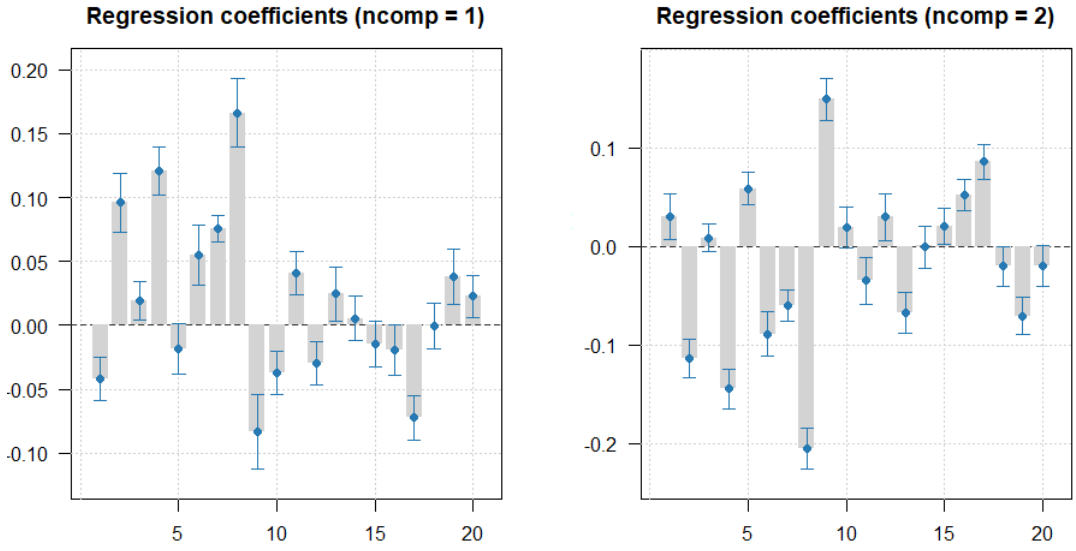

***Figura 14**. Gráfica que muestra las puntuaciones factoriales, o loadings, de cada variable en la componente 1 (izquierda) y la componente 2 (derecha).*

Viendo las variables a las que corresponden cada uno de los índices, las variables que más peso tienen en la componente 1 son el tiempo de posesión en campo propio, el número de tiros y el número de pases; para la componente 2 las variables que más peso tienen son el número de centros y las pérdidas de balón. En vista a los resultados del análisis discriminante, puede decirse que la componente 1 es representativa de los registros de partidos ganados, mientras que la componente 2 son representativos de los registros de los partidos no ganados.

Comparando los resultados con aquellos obtenidos en el análisis discriminante, puede verse que las variables que mayor peso tienen en el análisis PLS-DA son similares a aquellas que mayor importancia tienen en el análisis discriminante.

Resultados de la predicción

En Tabla 12 se muestra un resumen de los resultados obtenidos al aplicar el modelo PLS-DA al conjunto general de resultados.

	Predicción general (Tasa de acierto)	% acierto en partidos ganados (Especificidad)	% acierto en partidos no ganados (Sensibilidad)
General	72,40%	47,20%	87,20%

***Tabla 12**. Tabla resumen de los resultados obtenidos con el análisis PLS-DA.*

En estos resultados puede verse la especificidad, sensibilidad y la precisión de las predicciones. En concreto la precisión obtenida es de un 72,4%, la especificidad (tasa de acierto para los partidos ganados) de un 47,2% y la sensibilidad (tasa de acierto para partidos no ganados) de un 87,2%.

Estos resultados, en comparación con los resultados del análisis discriminante, reflejan un peor resultado para la predicción de partidos ganados y un mejor resultado para la predicción de partidos no ganados. Esto parece indicar que este método tiende a predecir un mayor número de partidos no ganados, generando un mayor número de falsos negativos.

5.2.3. Random Forest

Por medio de esta técnica, al igual que sucedía con el análisis discriminante, es posible analizar también, además de los resultados finales de las predicciones, la importancia de las variables a la hora de realizar estas predicciones.

Éxito general en la predicción

Utilizando la función *randomforest* [14], se configura el parámetro *mtry* (número de variables muestreadas aleatoriamente en cada división que se realiza) con un valor de 4. Este valor es debido a que 4 es aproximadamente la raíz cuadrada del número de variables explicativas, 25. Con esta configuración, los resultados generales obtenidos son los siguientes:

		Valores reales	
		No ganado	Ganado
Valores predichos	No ganado	622	254
	Ganado	54	176

***Tabla 13.** Matriz de confusión obtenida a partir del método de random forest.*

Con estos resultados se obtiene un 72.15% de aciertos, con una tasa de aciertos de partidos ganados del 40.93% y una tasa de aciertos de partidos no ganados del 92.01%. Teniendo en cuenta que en los valores reales del conjunto de prueba el porcentaje de partidos es de un 39% y el de no ganados un 61%, los resultados tienen un acierto similar al que se obtendría por azar con las probabilidades del conjunto de pruebas. Dicho esto, no puede decirse que sean unos resultados buenos para esta predicción. Por el contrario, el porcentaje de acierto de los partidos no ganados sube hasta el 92%, 31 puntos porcentuales por encima del azar.

Atendiendo a estos resultados, se puede decir que el método de *random forest*, al igual que el análisis PLS-DA, tiende a catalogar como partidos no ganados la mayoría de los registros, por lo que se obtiene una cantidad muy alta de falsos negativos (en este caso casi un 23%). Esto hace, a priori, de este método un método inadecuado para la predicción del éxito en el partido.

A continuación, se analizan las importancias de las variables en estas predicciones, para posteriormente analizar los resultados por ligas.

Importancia de las variables en las predicciones

En la figura 15 se puede ver, a la izquierda, la gráfica relativa a la importancia de las variables en el modelo generado. Por un lado, en la gráfica de la izquierda se puede ver que las variables que ayudan a mejorar la precisión de las predicciones son: el número de pases en profundidad y de finalización, el número de centros y de tiros. Por otro lado, a la derecha puede verse que las variables que más ayudan a reducir el índice de Gini son: el número de pases en profundidad y de finalización, el tiempo de posesión en campo propio y el número de centros, tiros y pases.

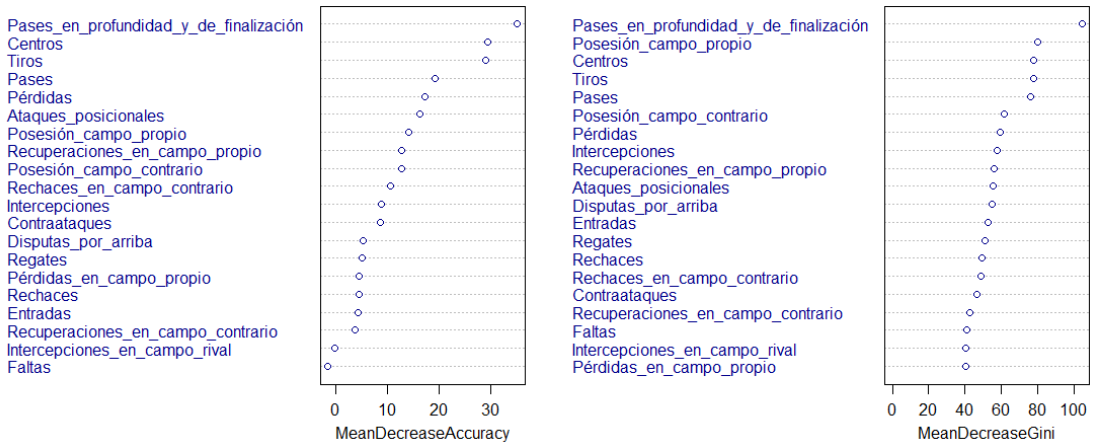

Figura 15. *Gráfica de importancia de las variables a la hora de mejorar la precisión (izquierda) y mejorar el índice de Gini (derecha).*

Comparando los resultados con los obtenidos a partir del análisis discriminante, las variables con peso o importancia en las predicciones coinciden en su mayoría, lo que hace que ambos análisis sean consistentes entre ellos.

Resultados de las predicciones

Como se ha comentado previamente, como último paso se analizan los resultados para las diferentes ligas y se comparan por los resultados generales.

	Predicción general	% acierto en partidos ganados	% acierto en partidos no ganados
General	72.11%	40.93%	92.01%
España	63.11%	28.23%	84.29%
Reino Unido	72.89%	52.38%	85.11%
Alemania	73.37%	56.92%	82.35%
Italia	68.89%	36.05%	89.21%
Francia	72.44%	36.99%	89.47%

Tabla 14. *Resultados generales y por ligas obtenidos a partir del método Random Forest.*

Al igual que sucedía con el análisis discriminante, en todas las ligas las variables más importantes coincidan con las del análisis general excepto en la liga española. En este caso, lo que cambia en estas ligas es el orden de estas variables, siendo la más importante en la liga española el número de centros y en la liga italiana el número de tiros.

A partir de estos resultados se puede ver que los resultados entre las diferentes ligas son muy dispares, con diferencias de hasta 28 puntos porcentuales en la predicción de partidos ganados entre la liga española y la liga alemana. Dicho esto, la mejor predicción general y entre los partidos ganados el mayor porcentaje de acierto se consigue para la liga alemana; de hecho, esta predicción puede considerarse incluso mejor que la del análisis discriminante.

A pesar de haber comentado previamente que este método tiende a generar muchos falsos negativos (falsos partidos no ganados), en el caso de la liga alemana sí parece un método válido para llevar a cabo estas predicciones.

5.2.4. Árboles de clasificación

La siguiente técnica de predicción que se utilizará es la de los árboles de clasificación. En este método se genera un árbol de decisión a partir de los registros de entrenamiento (70% de los registros totales), y este se corta para optimizar el error de predicción.

Generación del árbol

En este apartado se explicará paso por paso el proceso de generación del árbol de clasificación utilizado para llevar a cabo la predicción general, siendo el proceso para cada una de las ligas similar.

1. Creación del árbol máximo.

A partir de la función *rpart* [15], el árbol máximo se obtiene configurando un valor muy bajo del estadístico Cp, en este caso 0,01. Con respecto al resto de parámetros configurables, como número mínimo de observaciones en un nodo interno o en un nodo terminal, no se configuran para no limitar la construcción del árbol máximo. El árbol máximo generado puede verse en la Figura 16.

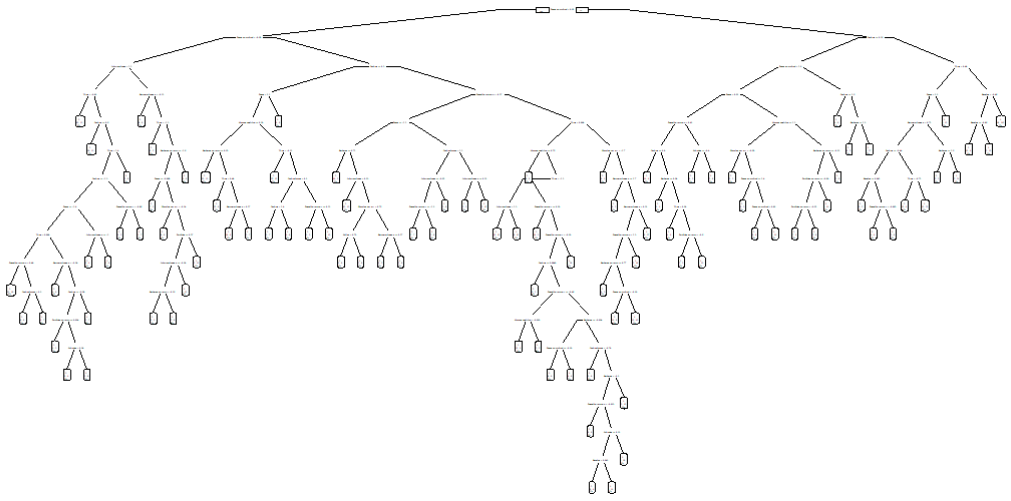

Figura 16. *Árbol de clasificación máximo.*

2. Poda del árbol

Para hacer la poda del árbol, a partir de la función *rpart* se obtiene la tabla de la Figura 17, que relaciona el valor del estadístico Cp con el número de divisiones presentes en el árbol en cada corte, el error relativo o error cuadrado medio y el *xerror* (error en las observaciones obtenido a partir de la validación cruzada). También puede verse en la Figura 18 la evolución del *xerror* en función del estadístico Cp y las divisiones del árbol.

Para podar el árbol de una forma óptima, se escoge el valor del estadístico Cp que optimice el valor de *xerror*. En este caso, ese valor de Cp que optimiza el valor de *xerror* es 0.0078370.

	CP	nsplit	rel error	xerror
1	0.1212121	0	1.00000	1.00000
2	0.0245559	1	0.87879	0.90073
3	0.0195925	3	0.82968	0.85580
4	0.0078370	7	0.75131	0.80669
5	0.0076628	9	0.73563	0.81609
6	0.0073145	12	0.71264	0.82027
7	0.0062696	15	0.69070	0.82550
8	0.0052247	17	0.67816	0.84326
9	0.0047022	20	0.66249	0.84744
10	0.0041797	22	0.65308	0.84848
11	0.0036573	33	0.60502	0.84744
12	0.0031348	35	0.59770	0.85475
13	0.0027865	41	0.57889	0.84953
14	0.0020899	48	0.55904	0.86834
15	0.0018286	63	0.51829	0.86625
16	0.0017416	68	0.50888	0.86625
17	0.0015674	75	0.49530	0.86625
18	0.0013932	79	0.48798	0.87461
19	0.0010449	93	0.45768	0.88715
20	0.0010000	96	0.45455	0.91327

Figura 17. *Tabla que relaciona el valor del estadístico Cp con el número de divisiones, el error relativo y el xerror. Recuadro rojo: valores óptimos escogidos.*

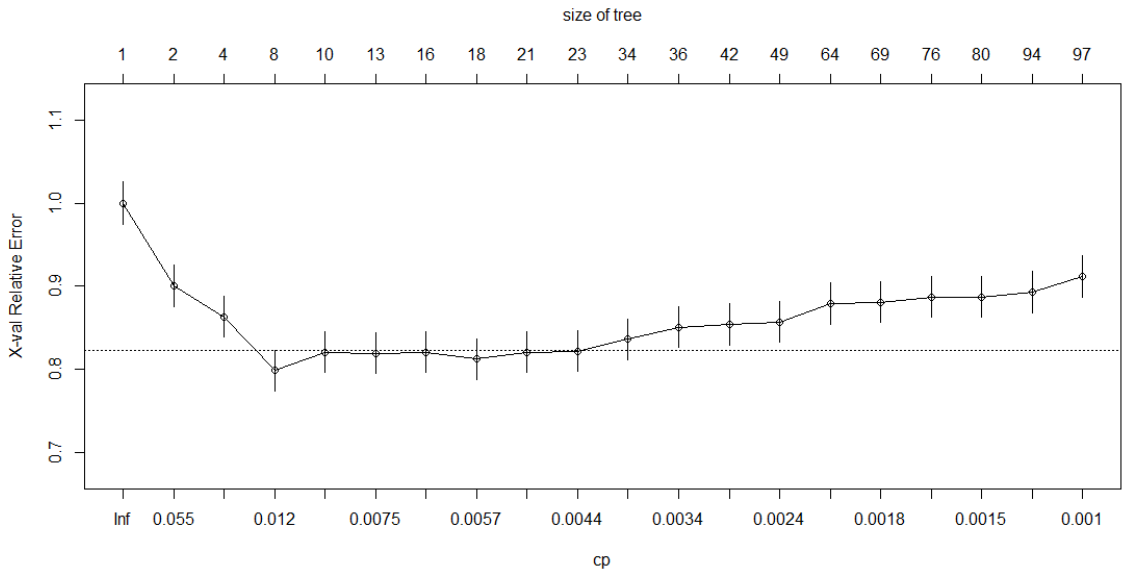

Figura 18. *Gráfica que relaciona el valor del estadístico Cp con el número de divisiones, el error relativo y el xerror.*

3. Obtención y análisis del árbol

Al utilizar el valor de 0.0078370 al estadístico Cp a la función *prune.rpart* se obtiene el árbol podado y óptimo de la Figura 19.

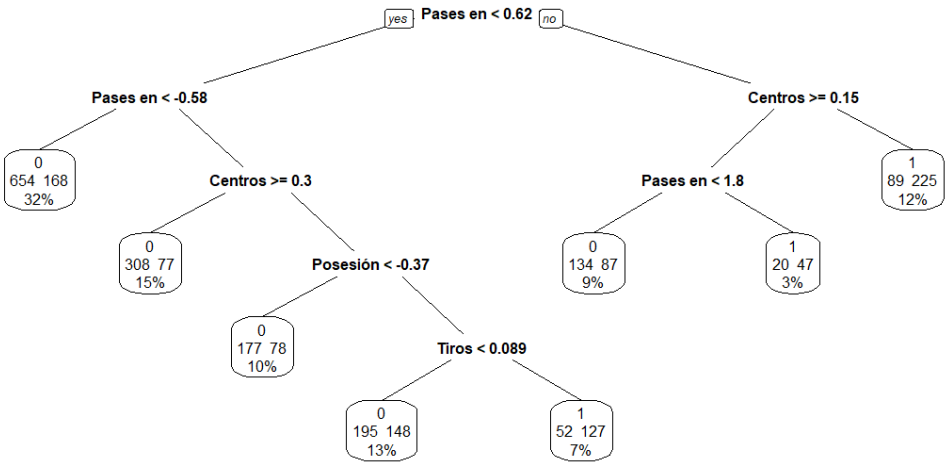

Figura 19. *Árbol de clasificación óptimo.*

A partir de este árbol de clasificación se puede ver cómo las variables clave a la hora de clasificar un registro son el número de pases, el número de centros, el tiempo de posesión en campo propio y el número de tiros. Atendiendo a las variables y a los valores que aparecen en los diferentes nodos, una vez invertida la transformación z-score, se puede describir el razonamiento lógico del árbol a la hora de predecir la clase con las siguientes reglas (correspondientes con los diferentes "caminos" del árbol de clasificación):

- Si el número de pases es mayor de 560 y el número de centros es menor o igual a 14, o el número de centros es mayor de 14 y el de pases es mayor que 707, el partido será clasificado como GANADO.
- Si el número de pases se encuentra entre 560 y 410, el de centros es mayor de 15. la posesión en campo propio es superior a 9 minutos y 30 segundos y el número de tiros es mayor de 12, el partido será clasificado como GANADO.
- De lo contrario, el partido será clasificado como NO GANADO.

<u>Resultados de las predicciones</u>

A partir de este árbol óptimo, los resultados generales que se obtienen son los expuestos en la Tabla 15.

		Valores reales	
		No ganado	Ganado
Valores predichos	No ganado	585	244
	Ganado	83	152

Tabla 15. *Matriz de confusión obtenida a partir del árbol de clasificación.*

Con estos resultados se obtiene un 69.27% de aciertos, con una tasa de aciertos de partidos ganados del 38.38% y una tasa de aciertos de partidos no ganados del 87.57%. Teniendo en cuenta que en los valores reales del conjunto de prueba el porcentaje de partidos es de un 37%

y el de no ganados un 63%, los resultados tienen un acierto similar al que se obtendría por azar con las probabilidades del conjunto de pruebas. Por lo tanto, al igual que con el método de *Random Forest*, no puede decirse que sean unos resultados buenos para esta predicción.

Los resultados parecen indicar que este método, como *Random Forest*, tiende a catalogar como partidos no ganados la mayoría de los registros, por lo que se obtiene una cantidad demasiado alta de falsos negativos; lo que convierte a este método en un método inadecuado a priori para la predicción del éxito en el partido.

Por último, y obteniendo un árbol óptimo particular para cada una de las ligas, se analiza cómo cambia el rendimiento de este método para predecir el éxito en los partidos en las diferentes ligas.

	Predicción general	% acierto en partidos ganados	% acierto en partidos no ganados
General	69.27%	38.38%	87.57%
España	66.81%	22.35%	93.62%
Reino Unido	71.12%	57.78%	79.58%
Alemania	67.65%	50.00%	75.86%
Italia	63.32%	41.11%	77.70%
Francia	69.40%	36.49%	85.90%

Tabla 16. *Resultados generales y por ligas obtenidos a partir del árbol de clasificación.*

Con respecto a los árboles de clasificación óptimos obtenidos, hay bastante diferencia entre las variables clave para las diferentes ligas. Mientras que en la liga española y la alemana las variables clave son las mismas que en el árbol obtenido para la predicción general, en la liga inglesa el número de contraataques y en la liga italiana el número de recuperaciones son clave a la hora de clasificar los registros.

En el caso de la liga francesa, las variables son completamente distintas al árbol de clasificación general, siendo las variables clave en este caso el número de pases en profundidad y de finalización, el número de intercepciones, el tiempo de posesión en campo propio y el número de centros, contraataques y regates. El árbol de clasificación correspondiente a esta liga se puede ver en la Figura 20.

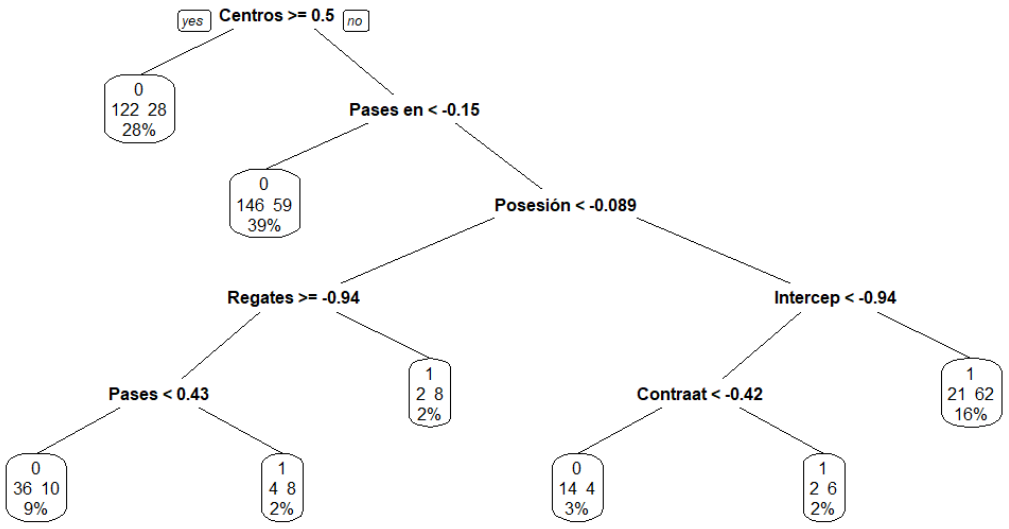

Figura 20. *Árbol de clasificación óptimo para el subconjunto de datos de Francia.*

En el caso de algunas ligas, los resultados generan una cantidad demasiado alta de falsos partidos no ganados, como es el caso de las ligas española y francesa; mientras que en otros casos se obtienen resultados similares a otros métodos como el análisis discriminante, pero con una tasa de aciertos tanto generales como para partidos ganados y no ganados menor.

Dicho esto, a pesar de que podría ser un método aceptable para predecir el éxito de los partidos en la liga inglesa, no sería el mejor método puesto que se obtienen mejores resultados con el análisis discriminante.

5.2.5. Naive-Bayes

Las siguientes técnicas de predicción tienen el inconveniente de no ser interpretables como pueden ser las técnicas aplicadas anteriores, por lo que los únicos datos que pueden obtenerse son las tasas de acierto en las predicciones. Estas técnicas son las de Naive-Bayes y las máquinas de soporte vectorial (SVM, por las siglas en inglés de *Support Vector Machines*).

En el caso de Naive-Bayes, clasificador probabilístico basado en el teorema de Bayes explicado en el apartado de Metodología de este mismo trabajo, se utiliza la función *naiveBayes* de la librería *e1071* los resultados generales que se obtienen son los siguientes:

		Valores reales	
		No ganado	Ganado
Valores predichos	No ganado	568	213
	Ganado	113	194

Tabla 17. *Matriz de confusión obtenida a partir del método de Naive-Bayes.*

Con estos resultados se obtiene un 70.04% de aciertos, con una tasa de aciertos de partidos ganados del 47.67% y una tasa de aciertos de partidos no ganados del 83.41%. Teniendo en cuenta que en los valores reales del conjunto de prueba el porcentaje de partidos es de un 37% y el de no ganados un 63%, los resultados tienen un acierto mayor al que se obtendría por azar. En comparación con otros métodos, este método sería, en términos generales, mejor que *random forest* y los árboles de clasificación, pero peor que el análisis discriminante.

Aplicando este método a cada una de las ligas por separado, los resultados obtenidos son los siguientes:

	Predicción general	% acierto en partidos ganados	% acierto en partidos no ganados
General	70.04%	47.67%	83.41%
España	70.61%	47.56%	83.56%
Reino Unido	60.85%	45.10%	72.93%
Alemania	68.42%	51.25%	80.91%
Italia	68.95%	46.05%	80.56%
Francia	66.40%	39.58%	83.12%

Tabla 18. *Resultados generales y por ligas obtenidos a partir del método de Naive-Bayes.*

Posteriormente, se han repetido las predicciones aplicando un suavizado de Laplace. No obstante, los resultados obtenidos han sido idénticos a los obtenidos sin él debido a las características de los datos.

Los resultados obtenidos a partir de este método son similares a los obtenidos con el análisis discriminante, pero con un rendimiento menor de forma que no parece el mejor método para hacer la predicción para ninguna de las ligas.

Por otro, y como hemos comentado previamente, el punto más negativo con respecto a este método y a las máquinas de soporte vectorial es la falta de interpretabilidad del resultado; por lo que, independientemente de los resultados, sería mucho más complicado aplicar estos resultados a la disciplina de un equipo.

5.2.6. Máquinas de Soporte Vectorial (*Support Vector Machines, SVM*)

Como última técnica de clasificación se utilizan las máquinas de soporte vectorial, o SVM por sus siglas en inglés (*Support Vector Machines*). Para realizar este análisis se utiliza la función *svm* de la librería *e1071* con 3 parámetros configurados:

- **Kernel:** para este trabajo se han probado aquellas opciones para las que no se necesita conocimiento previo, obteniendo los mejores resultados para el kernel radial.
- **Coste:** después de probar con varios valores de coste, finalmente los mejores resultados se han obtenido con un valor de 10 en este parámetro.
- **Gamma:** Se utiliza un valor de 0,05 por ser la inversa del número de variables explicativas. También se han probado valores distintos cercanos a este, pero los resultados obtenidos son peores que con el valor inicial.

Con esta configuración final, obteniendo los siguientes resultados:

		Valores reales	
		No ganado	Ganado
Valores predichos	No ganado	558	188
	Ganado	118	192

Tabla 19. *Matriz de confusión obtenida a partir de las máquinas de soporte vectorial.*

Con estos resultados se obtiene un 71.02% de aciertos, con una tasa de aciertos de partidos ganados del 50.05% y una tasa de aciertos de partidos no ganados del 82.54%. Teniendo en cuenta que en los valores reales del conjunto de prueba el porcentaje de partidos es de un 36% y el de no ganados un 64%, los resultados tienen un acierto mayor al que se obtendría por azar con las probabilidades del conjunto de pruebas. En comparación con otros métodos, este método sería similar en cuanto a resultados con el método de Naive-Bayes, con una tasa de acierto para partidos ganados 3 puntos porcentuales mayor y una tasa de acierto para partidos no ganados 2 puntos menor que el anterior método.

Aplicando este método a cada una de las ligas, los resultados obtenidos son los siguientes:

	Predicción general	% acierto en partidos ganados	% acierto en partidos no ganados
General	71.02%	47.67%	83.41%
España	74.44%	44.58%	92.14%
Reino Unido	69.57%	55.32%	79.41%
Alemania	74.40%	55.42%	87.10%
Italia	70.32%	41.18%	88.81%
Francia	72.25%	45.35%	88.65%

Tabla 20. *Resultados generales y por ligas obtenidos a partir de las máquinas de soporte vectorial.*

Como se puede ver a partir de estos resultados finales, con este método se consiguen unos resultados variables entre las ligas, pero en general pueden considerarse buenos y comparables a los mejores resultados obtenidos con otros métodos. El principal problema que tendría este método es, al igual que ocurre con el método de Naive-Bayes, la falta de interpretabilidad del proceso de predicción del éxito del partido.

5.2.7. Comparación de los resultados obtenidos en los modelos de predicción

Una vez han sido obtenido y analizados todos los resultados, en esta sección se compara la bondad de ajuste y predictores más representativos para una valoración crítica.

Como se ha podido ver, si bien existen métodos que han resultado negativos para las predicciones para el conjunto completo de los datos, como el método de *random forest* o de Naive-Bayes, estos métodos pueden resultar útiles a la hora de realizar predicciones a un nivel de granularidad de los datos diferentes. Por lo tanto, parece claro afirmar que no existe un método mejor en términos generales, y es necesario seleccionar el método que mejor se ajuste al enfoque considerado y al conjunto de datos concreto.

Dicho esto, como resumen final de los resultados es conveniente seleccionar aquel método que resulte más adecuado para cada una de las casuísticas, que en este caso corresponden a cada una de las ligas y al conjunto de ligas.

Como se puede observar, en algunos casos los resultados no difieren en gran medida entre los 2 o 3 mejores métodos para predecir un resultado, por lo que se tendrán en consideración 3 aspectos a la hora de seleccionar el mejor método en cada caso:

- **La predicción general**: puesto que, al fin y al cabo, es el resultado más directo del éxito de un modelo predictivo.

- **La interpretabilidad de los resultados:** a igualdad de resultados, siempre será conveniente utilizar un método a partir del cual pueda obtenerse información valiosa más allá de la predicción numérica.

- **La predicción de los partidos ganados:** debido a la menor cantidad de los partidos clasificados como ganados (aproximadamente un tercio de los registros) y al valor extra que supone la predicción correcta de un partido ganado frente a la predicción correcta de un partido no ganado, una mejor predicción de este conjunto de partidos se considerará un aspecto positivo del modelo.

Dicho esto, la tabla final con los métodos escogidos para cada caso se puede ver en la Tabla 21.

	Predicción general	% acierto en partidos ganados	% acierto en partidos no ganados	Método
General	74.93%	54.72%	86.11%	Análisis Discriminante
España	74.44%	44.58%	92.14%	Máquinas de soporte vectorial
Reino Unido	73.33%	59.52%	81.56%	Análisis Discriminante
Alemania	73.37%	61.53%	79.83%	Análisis Discriminante
Italia	70.67%	51.11%	82.7%	Análisis Discriminante
Francia	72.89%	43.84%	86.84%	Análisis Discriminante

Tabla 21. *Resultados obtenidos con el método elegido para cada una de las ligas.*

En base a esta tabla final, podría decirse que el método más exitoso a la hora de predecir el éxito en el partido ha sido el análisis discriminante puesto que, excepto para la liga española, es el mejor predictor tanto en términos generales como para el resto de las ligas de forma individual.

Por otro lado, aunque la tasa de acierto general no es especialmente distinta entre los casos (apenas 4 puntos porcentuales de diferencia entre el mejor y el peor), sí lo hacen las tasas de acierto para partidos ganados (hasta casi 18 puntos) y para partidos no ganados (casi 13 puntos).

Más allá de las predicciones numéricas y los métodos, otro punto interesante a analizar son las variables significativas para cada uno de estos métodos escogidos para cada liga. En la Tabla 22 se puede ver la importancia de las variables en cada uno de los métodos que mejor funcionan para cada una de las ligas. Debido a que el mejor modelo predictivo para la liga española es una máquina de soporte vectorial, se exponen las variables más importantes reflejadas al aplicar el análisis discriminante.

	General	España	Reino Unido	Alemania	Francia
Variables	**Análisis discriminante**	**Análisis discriminante**	**Análisis discriminante**	**Análisis discriminante**	**Análisis discriminante**
Posesión en campo propio	Alta (G)	Media (G)	Media (G)	Alta (G)	Alta (G)
Tiros	Alta (G)	Media (G)	Alta (G)	Media (G)	Media (G)
Pases en profundidad y de finalización	Alta (G)	Alta (G)	Alta (G)	Alta (G)	Alta (G)
Recuperaciones en campo propio	Alta (G)	Baja	Alta (G)	Baja	Media (G)
Centros	Alta (NG)	Alta (NG)	Alta (NG)	Alta (NG)	Alta (NG)
Pérdidas	Alta (NG)	Alta (NG)	Alta (NG)	Alta (NG)	Alta (NG)
Disputas por arriba	Media (G)	Media (G)	Alta (G)	Baja	Media (G)
Faltas	Baja	Baja	Baja	Alta (G)	Baja
Regates	Baja	Baja	Baja	Alta (G)	Baja
Intercepciones	Media (G)	Alta (G)	Baja	Alta (G)	Baja
Rechaces en campo contrario	Media (NG)	Media (NG)	Baja	Alta (NG)	Media (NG)

Tabla 22. *Nivel de importancia de las variables clave para cada método escogido.*

Nota: entre paréntesis aparece "G" si es clave para la predicción de partidos ganado, y "NG" si lo es para partidos no ganados.

Como puede verse, como regla general las variables que tienen una importancia alta la tienen en todos o en la mayoría de los casos; estas variables son el tiempo de posesión en campo propio, el número de tiros, el número de pases en profundidad y de finalización, el número de centros y el número de pérdidas. Lo mismo sucede con las variables con una importancia baja, como el número de faltas y el número de regates.

El único caso que se diferencia sustancialmente del resto es la liga alemana. En esta liga, algunas variables como el número de faltas, regates, intercepciones y rechaces en campo contrario tienen una importancia alta mientras que para el resto de las ligas su importancia es baja. Sucede también al contrario, puesto que algunas variables con una importancia baja para la liga alemana, como las recuperaciones en campo propio, tienen una importancia alta o media para el resto de ligas.

Como se conoce, algunas variables afectan más a la predicción de los partidos ganados y otras a la predicción de los partidos no ganados. En la tabla, por medio de las letras G (Ganados) y NG (No Ganados) que se encuentran junto con el nivel de importancia, se puede ver cómo el sentido de esta importancia es compartido para todas las variables en todas las ligas y en el caso general.

5.2.8. Dificultad en la predicción de la posición final en liga

En un principio, el trabajo también contenía la predicción de la posición final en liga a partir de los datos de los partidos. No obstante, los resultados obtenidos fueron bastante pobres, teniendo una tasa de éxito de menos de un 15%.

Este resultado puede deberse a que, el fin y al cabo, la posición en liga es el resultado acumulado de todos los partidos que juega un equipo, por lo que hacer la predicción de la posición final en base a los datos de un solo partido puede ser complicado. Además, considerando esta razón, cobraría más protagonismo las dinámicas de los equipos a lo largo de la temporada.

Debido a este mal rendimiento, se consideró finalmente eliminar este enfoque puesto que no aportaba apenas valor al conjunto del trabajo más allá de la conclusión de no ser predecible la posición en liga a partir de los datos de los partidos.

El mal rendimiento de estas predicciones puede observarse en la Tabla 23, donde se puede ver el resultado general con el método de *random forest* para la predicción de la posición. En esta tabla aparecen las posiciones reales y predichas de los equipos, teniendo en cuenta que todas las ligas tienen un total de 20 equipos excepto la liga alemana, que tiene 18 equipos.

		Posición real																			
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Posición predicha	**1**	**34**	19	3	9	7	1	1	1	1	4	1	0	0	0	0	1	0	1	0	0
	2	3	**6**	4	6	3	2	1	3	2	1	6	2	0	2	0	0	2	0	1	0
	3	11	6	**19**	5	6	3	5	3	5	6	2	6	1	3	0	0	3	2	0	1
	4	2	0	2	**6**	4	4	4	5	1	2	3	1	0	2	0	3	3	2	1	1
	5	0	4	1	5	**6**	2	2	5	1	2	1	1	0	2	2	3	1	0	1	1
	6	1	2	1	5	6	**12**	0	3	4	4	5	6	4	0	5	5	5	3	3	3
	7	0	4	1	1	1	0	**3**	4	1	3	2	1	1	1	0	4	1	1	2	2
	8	0	0	3	2	2	7	6	**3**	0	1	4	1	1	2	2	5	3	0	1	1
	9	2	4	0	5	4	6	4	4	**5**	1	4	1	1	2	2	5	3	0	1	1
	10	0	2	0	1	3	1	5	3	5	**3**	2	2	2	1	2	0	1	8	4	0
	11	1	0	0	0	0	2	0	4	3	1	**2**	2	4	4	0	0	1	3	0	0

12	0	1	3	0	1	1	6	3	2	2	4	**13**	3	6	4	8	7	6	3	6
13	1	2	1	4	2	2	3	1	6	3	2	4	**9**	7	4	11	4	6	5	5
14	0	0	1	2	0	1	6	4	3	3	5	6	4	**7**	2	5	5	5	2	2
15	0	1	1	2	3	5	3	3	1	4	10	5	11	1	**16**	9	5	6	8	3
16	0	1	0	0	2	2	2	2	2	4	10	5	11	1	16	**9**	5	6	8	3
17	0	1	0	0	1	4	0	3	3	3	5	1	3	5	2	1	**1**	3	6	2
18	0	0	0	3	4	1	2	2	0	2	4	2	5	2	3	5	5	**10**	3	4
19	0	0	1	2	1	1	4	1	0	0	0	0	0	0	1	1	1	1	**2**	2
20	0	0	0	1	0	3	2	3	0	1	2	0	1	6	2	5	7	1	4	**5**

Tabla 23. *Matriz de confusión obtenida a partir del método Random Forest para la predicción de la posición final de liga.*

Con los resultados mostrados en esta tabla, se puede ver claramente cómo no es una buena predicción de la posición en liga, siendo la tasa de acierto general de un 14,65%. Este mal resultado puede ser explicado a partir de los resultados de los anteriores partidos apartados:

- **Capacidad predictiva de las variables:** igual que la capacidad predictiva de las variables no es excelente para predecir el éxito en un partido, es lógico que tampoco lo sea para la predicción de la posición de liga.

- **Relación de las variables con la posición de liga:** si bien las estadísticas del partido tienen una relación de primer orden con el éxito del partido, su relación con la posición de liga podemos definirla como de segundo nivel ya que esta variable a predecir depende de todos los partidos, y no de uno sólo.

6. Limitaciones del trabajo

La limitación principal se basa en la propia naturaleza del deporte. En el fútbol, como sucede con cualquier otro deporte, los datos son capaces de medir los aspectos del juego puramente cuantificables; no obstante, existen otros aspectos que no son fácilmente cuantificables, como las condiciones climatológicas, o son directamente imposibles de cuantificar, como el estado de forma o de ánimo de los jugadores, el acierto de los entrenadores y la actuación arbitral, entre otros.

En este punto es donde juegan un papel fundamental los expertos en el deporte, desde los entrenadores, preparadores deportivos propios de cada equipo, hasta los analistas o periodistas. De esta forma, los análisis podrían nutrirse de los comentarios de especialistas en la materia, haciendo de los resultados una fuente mayor de información más certera; y, por el contrario, estos resultados podrían ser una ayuda para los miembros del club y una información muy interesante para los analistas y periodistas especializados.

Centrándonos en el contenido propio del trabajo, se ha de tener en cuenta que se han supuesto los registros independientes entre sí. Al tener en cuenta esta suposición se están descartando otras variables que pueden afectar al rendimiento de un equipo a lo largo de la liga, como puede ser las diferentes dinámicas del equipo a lo largo de la temporada (conocidos popularmente como rachas), las lesiones o las expulsiones o sanciones. De ser posible añadir algunas de estas variables, es probable que la predicción del éxito en sus dos enfoques (partido y temporada) mejore.

Otra limitación es relativa al acceso a la información. Si bien es posible conseguir estadísticas básicas de los partidos en diferentes fuentes, tener acceso a una base de datos con datos tan completos de todos los partidos de las 5 grandes ligas europeas es algo complicado de conseguir. En el caso de este trabajo contábamos con los datos de una temporada completa, pero en el caso de querer hacer análisis en un intervalo temporal mayor (abarcando varias temporadas) o haber podido incluir otros datos más complejos, por ejemplo, información relativa al día a día de cada equipo (lesiones, sanciones…) habría sido muy complicada de conseguir toda la información para alguien que no se dedica al mundo del deporte.

7. Conclusiones

Una vez presentados los resultados del trabajo, a continuación se exponen las conclusiones tanto en la parte más exploratoria del trabajo, como son el PCA y el análisis clúster, como en la parte predictora del trabajo con las diferentes técnicas utilizadas.

En referencia a los primeros pasos del trabajo, en los que se realiza una selección de variables a partir de un conjunto de variables mayor, cabe mencionar que esta selección se ha realizado varias veces añadiendo y eliminando algunas variables en base al análisis exploratorio y a los primeros resultados obtenidos hasta llegar al conjunto de variables final. Con esto se pone en evidencia la enorme importancia que tiene la preparación previa de los datos, ya que los resultados pueden cambiar sensiblemente si del conjunto de variables se añade o se elimina alguna de las variables.

En la primera mitad del trabajo, a partir de los datos de los partidos, se han podido deducir las diferentes estrategias que han utilizado los equipos en cada partido. En concreto se han identificado 4 componentes principales correspondientes a diferentes estrategias: tener la posesión del balón en el campo del rival buscando huecos para atacar, defender en el campo propio y salir al contraataque, tener la posesión en el propio campo y crear ataques a partir de tener mucha posesión y, por último, presionar al rival cerca de su portería para evitar que puedan avanzar y crear ataques rápidos al robar el balón.

A continuación, se identificaron 4 clústeres distintos de partidos representados por las estrategias explicadas. El primero de ellos está caracterizado por mantener posesiones en el campo del rival y buscar huecos para acercarse a la portería rival. El segundo de ellos se caracteriza por ejercer presión al rival cerca de su portería para intentar robar el balón y hacer ataques rápidos. El tercero de los clústeres está representado por la combinación de dos de las estrategias identificadas, siendo partidos en los que el equipo ha defendido en su propio campo y, al recuperar el control del balón, o bien ha salido al contraataque o bien ha creado ataques a partir de mantener una posesión larga. En lo que respecta al cuarto y último clúster, este no está representado por ninguna de las estrategias, por lo que son partidos donde el equipo no ha tenido una estrategia definida y constante a lo largo del partido.

Siendo este último clúster el más exitoso a nivel descriptivo, que no tenga una estrategia clara puede interpretarse como que son partidos donde los equipos han sabido ajustar su estrategia al momento o la dinámica del partido (jugar con un jugador más o menos, lesiones, cambios en la estrategia del otro equipo), mientras que aquellos que son fieles a una estrategia independientemente del momento del partido tienden a tener un menor éxito.

Con respecto al objetivo de predicción, los resultados pueden ser considerados relativamente buenos, pero no excelentes. Una razón para el rendimiento de estas predicciones radica en la elevada incertidumbre asociada a esta variable objetivo.

Si bien el ajuste global de los modelos al conjunto de datos no es excelente, sí que se ha observado cómo las variables clave a la hora de determinar el éxito en las estrategias son similares entre las técnicas con resultados interpretables (análisis discriminante, *random forest* y árboles de clasificación) y son, salvo excepciones, iguales para cada una de las ligas y para la predicción general. Estas variables clave, como puede verse en la Tabla 22, son el tiempo de posesión en campo propio, el número de pases en profundidad y de finalización, el número de centros y el número de pérdidas. Esto indica que las predicciones, aunque podrían ser mejores, son robustas y consistentes entre ellas y todas apuntan en una dirección parecida.

A nivel global, es interesante atender a las diferencias entre los modelos de predicción para las diferentes ligas. Aunque es cierto que finalmente el análisis discriminante ha sido la técnica más exitosa para casi todos los casos, queda patente la importancia de probar y ajustar diferentes modelos para cada casuística que se desee estudiar, ya que los modelos que funcionan bien para unos casos pueden no hacerlo para otros, incluso si la naturaleza de los datos es la misma o los datos de una casuística no son más que un subconjunto de ese conjunto de datos.

Por último, y a raíz de las predicciones intentadas de la posición final en liga a partir de los datos de los partidos, este trabajo se ha encontrado con la otra cara de la moneda en cuanto a predicciones se refiere. Este lado es el de intentar predecir aspectos que, a bien son especialmente difíciles de predecir, o directamente no se puedan llegar a predecir con una tasa de acierto suficientemente alta como para ser de utilidad.

8. Líneas futuras y posibilidades

En el punto actual del trabajo, pueden plantearse diferentes posibilidades o líneas futuras en las que profundizar para obtener tanto mejores resultados con los mismos objetivos como cambiar el enfoque de algunos de los objetivos o plantear nuevos.

De manera general, una forma de buscar nuevos resultados que pueden parecer útiles es probar con otro conjunto de variables al seleccionado y a los probados previamente. Una nueva criba que podría hacerse es eliminar del conjunto variables que tengan una relación con la calidad del equipo, como pueden ser los tiros o los centros, ya que dos equipos que jueguen con una estrategia similar conseguirán llegar a realizar más tiros o centros en función de la calidad de sus plantillas.

Por otro lado, y con apoyo en lo escrito en las limitaciones, con el objetivo de llevar a cabo un análisis más complejo sería un siguiente paso abandonar la suposición de independencia en los registros para poder tener en cuenta las dinámicas de los equipos a lo largo de la temporada.

También relacionado con las limitaciones expuestas previamente, sería conveniente intentar ampliar la información con la que se cuenta con datos que reflejen las situaciones descritas en las limitaciones, como pueden ser las lesiones o las condiciones climatológicas.

Poniendo el foco en el primer objetivo, y con la vista puesta en la aplicación concreta para un analista de deportes o un club de fútbol, podría ser interesante analizar los patrones de relaciones entre variables en un subconjunto de los datos relativo a una de las ligas (incluso a un solo equipo). Esto ayudaría al equipo a obtener más información a la hora de planificar los entrenamientos y pivotar su estrategia en base a las conclusiones que se obtengan.

Centrándonos en el segundo objetivo, para obtener mejores resultados en el objetivo de las predicciones, sería conveniente llevar a cabo estas predicciones con otras técnicas más avanzadas de predicción, como las redes neuronales o versiones más avanzadas de técnicas propias del análisis multivariante o la minería de datos.

En vista general, sería conveniente poder contar con datos de más de una temporada para poder conocer si los resultados para una temporada son consistentes con otras temporadas o únicamente son válidos a lo largo de una temporada tanto a nivel general como para cada una de las ligas.

9. Referencias

[1] Deloitte. *Sports Tech Innovation in the Start-up Nation*. 2017. Enlace: *https://www2.deloitte.com/content/dam/Deloitte/il/Documents/finance/sport_tech_report_short.pdf* (Acceso: 28 de Agosto de 2022).

[2] Altarriba-Bartés A. et al. *Analysis of the winning probability and the scoring actions in the American professional soccer championship*. RICYDE. Revista Internacional de Ciencias del Deporte 16, 67-84. 2020

[3] Malagón M. *Machine Learning en el mundo del fútbol.* Trabajo de Fin de Máster. Universidad Politécnica de Valencia, Valencia, España. 2019

[4] Sánchez B. (6 de mayo, 2022). Artículo "Por qué el Real Madrid y Rafael Nadal son los únicos capaces de pulverizar los pronósticos del Big Data". Enlace: *https://www.elespanol.com/deportes/20220506/real-madrid-rafael-nadal-pronosticos-big-data/670183389_0.html* . (Acceso: 29 de agosto de 2022)

[5] RStudio. RStudio. 2018. Enlace: *https://www.rstudio.com/products/rstudio/.* (último acceso: 26 de agosto de 2022).

[6] Ahmed A., Picco I. (6 de abril, 2020). Artículo "Las contribuciones tangibles del deporte a los ODS". Enlace: *https://www.un.org/es/cr%C3%B3nica-onu/las-contribuciones-tangibles-del-deporte-los-ods* (Acceso: 5 de septiembre de 2022)

[7] Wickham H., Bryan J. readxl: Read Excel. Files. R package version 1.4.0. 2022. Enlace: *https://CRAN.R-project.org/package=readxl*

[8] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[9] Everitt B., Hothorn T. MVA: An Introduction to Applied Multivariate Analysis with R. R package version 1.0-8. 2022. Enlace: *https://CRAN.R-project.org/package=MVA*

[10] da Silva, A. R. biotools: Tools for Biometry and Applied Statistics in Agricultural Science. R package version 4.2. 2021. Enlace: *https://cran.r-project.org/package=biotools*

[11] Venables, W. N., Ripley, B. D. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0. 2022

[12] Sing T., Sander O., Beerenwinkel N., Lengauer T. "ROCR: visualizing classifier performance in R." _Bioinformatics_, *21*(20), 7881. 2005. Enlace: *http://rocr.bioinf.mpi-sb.mpg.de*.

[13] Kucheryavskiy S. "mdatools - R package for chemometrics." _Chemometrics and Intelligent Laboratory Systems_, *198*. 2020. Enlace: *https://doi.org/10.1016/j.chemolab.2020.103937*.

[14] Liaw A., Wiener M. Classification and Regression by randomForest. R News 2(3), 18--22. 2002.

[15] Therneau T., Atkinson B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. 2019. Enlace: *https://CRAN.R-project.org/package=rpart*

[16] Milborrow S. rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.9. 2020. Enlace: *https://CRAN.R-project.org/package=rpart.plot*

[17] Meyer D., Dimitriadou E., Hornik K., Weingessel A. and Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-6. 2021. Enlace: *https://CRAN.R-project.org/package=e1071*

[18] Wold S., Esbensen K., Geladi, P. *Principal component analysis. Chemometrics and intelligent laboratory systems.* 1987.

[19] Dunn, K. *Process Improvement Using Data*. *https://learnche.org/pid/latentvariable-modelling/principal-component-analysis/some-properties-of-pcamodels.* 2019 (Acceso: 29 de agosto de 2022)

[20] Duda R. O., Hart P. E., and Stork D. G., *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000

[21] CHEN, M.; HAN, J.; YU, Philip S. *Data mining: an overview from a database perspective. Knowledge and data Engineering*, IEEE Transactions, 8(6), 866-883, 1996. Enlace: *http://www.nyu.edu/classes/jcf/g22.3033-002/handouts/chen96data.pdf*

[22] Ward, Joe H. *Hierarchical Grouping to Optimize an Objective Function*. Journal of the American Statistical Association, 58(301), 236-244, 1963

[23] Cea, M. A. *Análisis Discriminante*. Centro de Investigaciones Sociológicas. CIS. 2019.

[24] Fisher, R. *The use of multiple measurements in taxonomic problems*. Annals of Eugenics. 1936.

[25] Trecet R. *Aplicación de métodos estadísticos multivariantes para la evaluación de la calidad en jamones, por medio de estructuras de datos N-dimensionales.* Trabajo de Fin de Máster. Universidad Politécnica de Valencia, Valencia, España. 2019.

[26] Sanabria A. (19 de mayo, 2020). Artículo "Una introducción a los árboles de decisión". Enlace: *https://www.grupodabia.com/post/2020-05-19-arbol-de-decision/s* (Acceso: 30 de agosto de 2022)

[27] Berk, R.A.: *Classification and regression trees (CART).* In: Berk, R.A. (ed.) Statistical Learning from a Regression Perspective, 129–186. Springer, Cham. 2016.

[28] Breiman, L. *Random forests.* Machine learning, 45, 5-32. 2001.

[29] Rodrigo-Amat, R. *Clustering y heatmaps: aprendizaje no supervisado. Medidas de distancia. Escala de las variables.* 2017. *https://rpubs.com/Joaquin_AR/310338*. (Acceso: 1 de septiembre de 2022).

[30] Santana, E. *Machine Learning con R*. 2014. Enlace: *http://apuntesr.blogspot.com/2014/11/ejemplo-de-random-forest.html* (último acceso: 2 de septiembre de 2022).

[31] Luque C.M. *Clasificadores bayesianos. El algoritmo Naïve Bayes Resumen*. Apuntes de clase, Universidad de Nebrija. 2003.

[32] Betancour G. *Las máquinas de soporte vectorial (SVMs).* Scientia Et Technica, 9(27), 67–72. 2005. *doi:10.22517/23447214.6895*.

ANEXOS

Anexo 1. Descripción de variables

A continuación, en la Tabla 24 puede verse una breve descripción de cada una de las variables que conforman la base de datos original, incluidas aquellas que finalmente se han descartado de los análisis posteriores.

Variable	Descripción
LIGA	Liga a la que pertenece el registro: Serie A (IT), LaLiga (ES), Bundesliga (AL), Inglaterra (IN), Ligue 1 (FR).
POSICIÓN	Posición final de liga del equipo al que corresponde el registro.
COD. EQUIPO	Código alfanumérico que identifica al equipo al que corresponde el equipo. Código conformado con las primeras letras de la liga y la posición final del equipo (Ej. El equipo italiano Chievo Verona, al terminar en el puesto 20, su código es el "IT20")
CHAMPION	Variable categórica que indica si, a final de la temporada, el equipo al que corresponde el registro se ha clasificado para la UEFA Champions League (1 si se clasificado, 0 si no).
DESCENSO	Variable categórica que indica si, a final de la temporada, el equipo al que corresponde el registro ha descendido de categoría (1 si ha descendido, 0 si no).
NOMBRE	Nombre del equipo
PARTIDO	Resultado del partido con nombre de los equipos y goles marcados.
CASA	Variable categórica que tiene valor C si el equipo al que corresponde el registro juega en casa y F si juega fuera.
GANADO	Variable categórica que tiene valor 1 si el equipo ha ganado el partido y 0 si no lo ha ganado.
GOLES MARCADOS - RESULTADO	Goles válidos marcados.
GOLES ENCAJADOS	Goles válidos encajados.
TARJETA AMARILLA	Tarjetas amarillas recibidas
TARJETA ROJA	Tarjetas rojas recibidas
SAQUES DE ESQUINA	Saques de esquina sacados.

FUERAS DE JUEGO	Fuera de juego cometidos.
FALTAS	Faltas cometidas.
FALTAS - RIVAL	Faltas recibidas.
POSESIÓN DEL BALÓN, MIN	Tiempo de juego con la posición del balón (minutos).
POSESIÓN DEL BALÓN EN CAMPO PROPIO	Tiempo de juego con la posesión del balón en campo propio (en segundos).
POSESIÓN DEL BALÓN EN CAMPO CONTRARIO	Tiempo de juego con la posesión del balón en campo contrario (en segundos).
POSESIÓN DEL BALÓN EN EL ÚLTIMO TERCIO DEL CAMPO	Tiempo de juego con la posesión del balón en el último tercio del campo (en segundos).
TIROS	Tiros realizados.
TIROS A PORTERÍA	Tiros realizados
ATAQUES	Ataques totales
ATAQUES POSICIONALES / ELABORADOS	Ataques posicionales.
ATAQUES CON TIROS - POSICIONALES	Ataques posicionales finalizados con tiro.
ATAQUES CON TIROS - ATAQUES ELABORADOS, %	Porcentaje de ataques posicionales, o elaborados, finalizados con tiro.
CONTRAATAQUES	Contrataques totales.
CONTRAATAQUES FINALIZADOS CON TIRO	Contraataques finalizados con tiro.
ATAQUES CON TIROS - CONTRAATAQUES, %	Porcentaje de contraataque, o elaborados, finalizados con tiro.
ACCIONES A BALÓN PARADO A FAVOR	Acciones a balón parado iniciadas.
ATAQUES CON TIROS - ACCIONES A BALÓN PARADO	Ataques en acciones a balón parado finalizados con tiro.
ATAQUES CON TIROS - ACCIONES A BALÓN PARADO, %	Porcentaje de ataques en acciones a balón parado finalizados con tiro.
PASES	Pases realizados.

PASES PRECISOS	Pases efectivos.
PASES PRECISOS, %	Porcentaje de pases efectivos.
PASES EN PROFUNDIDAD Y DE FINALIZACIÓN	Pases en profundidad y de finalización.
PASES EN PROFUNDIDAD Y DE FINALIZACIÓN - EFECTIVO	Pases en profundidad y de finalización efectivos.
CENTROS	Centros realizados.
CENTROS EFECTIVO	Centros efectivos.
CENTROS - EFECTIVO, %	Porcentaje de centros efectivos.
DISPUTAS	Disputas del partido.
BALÓNES DIVIDIDOS GANADOS	Balones divididos ganados.
DISPUTAS GANADAS, %	Porcentaje de disputas ganadas.
DISPUTAS DEFENSIVAS	Disputas en el propio campo (zona de defensa).
DISPUTAS EN DEFENSA / GANADAS	Disputas ganadas en el propio campo (zona de defensa).
DISPUTAS EN DEFENSA / GANADAS, %	Porcentaje de disputas en campo propio ganadas
DISPUTAS EN ATAQUE	Disputas en campo contrario (zona de ataque).
DISPUTAS EN ATAQUE / GANADAS	Disputas ganadas en campo contrario (zona de ataque).
DISPUTAS EN ATAQUE / GANADAS, %	Porcentaje de disputas en campo contrario.
DISPUTAS POR ARRIBA	Disputas por arriba (en altura, jugadas con la cabeza).
DISPUTA POR ARRIBA GANADA	Disputas por arriba ganadas.
DISPUTAS POR ARRIBA %	Porcentaje de disputas por arriba ganadas.
REGATES	Regates intentados.
REGATES EFECTIVOS	Regates completados.
REGATES %	Porcentaje de regates completados.
ENTRADAS	Entradas realizadas.

ENTRADAS EFECTIVAS	Entradas realizadas con éxito.
BALÓNES ROBADOS %	Porcentaje de balones robados, o porcentaje de entradas efectivas.
INTERCEPTACIONES	Interceptaciones de pases rivales.
INTERCEPTACIONES / EN CAMPO RIVAL	Interceptaciones de pases rivales en campo rival.
RECHACES	Rechaces totales.
RECHACES - EN CAMPO CONTRARIO	Rechaces en campo contrario.
PÉRDIDAS	Pérdidas totales.
PÉRDIDAS / EN CAMPO PROPIO	Pérdidas en campo propio.
BALONES RECUPERADOS	Balones recuperados totales.
RECUPERACIÓN / EN CAMPO PROPIO	Recuperaciones en campo propio.
RECUPERACIÓN / EN CAMPO CONTRARIO	Recuperaciones en campo contrario.
PASES ADELANTE (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases realizados hacia adelante (considerando aquellos pases realizados en dirección hacia la portería rival).
PASES VOLVER (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases realizados hacia atrás (considerando aquellos pases realizados en dirección hacia la propia portería).
PASES A LA IZQUIERDA (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases realizados hacia la izquierda (considerando aquellos pases realizados en dirección a la línea de banda de la izquierda mirando hacia la portería rival).
PASES A LA DERECHA (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases realizados hacia la derecha (considerando aquellos pases realizados en dirección a la línea de banda de la derecha mirando hacia la portería rival).
PASES ADELANTE EFECTIVO (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases efectivos hacia adelante (considerando aquellos pases realizados en dirección hacia la portería rival).
PASES VOLVER EFECTIVO (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases efectivos hacia atrás (considerando aquellos pases realizados en dirección hacia la propia portería).

PASES A LA IZQUIERDA EFECTIVO (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases efectivos hacia la izquierda (considerando aquellos pases realizados en dirección a la línea de banda de la izquierda mirando hacia la portería rival).
PASES A LA DERECHA EFECTIVO (ÁNGULO DE CAPTURA - 180 GRADOS)	Pases efectivos hacia la derecha (considerando aquellos pases realizados en dirección a la línea de banda de la derecha mirando hacia la portería rival).

Tabla 24. Descripciones de las variables de la base de datos original.

Anexo 2. Código del método del codo (*elbow method*)

El código utilizado para la determinación del número óptimo de clústeres por medio del método del codo, o *elbow method*, es el siguiente:

```
calcular_totwithinss <- function(n_clusters, Data_Cluster_M_Scale, iter.max=1000, nstart=50){
  # Esta función aplica el algoritmo kmeans y devuelve la suma total de
  # cuadrados internos.
  cluster_kmeans <- kmeans(centers = n_clusters, x = Data_Cluster_M_Scale, iter.max = iter.max,
                           nstart = nstart)
  return(cluster_kmeans$tot.withinss)
}

# Se aplica esta función con para diferentes valores de k
total_withinss <- map_dbl(.x = 1:15,
                          .f = calcular_totwithinss,
                          Data_Cluster_M_Scale)
total_withinss


data.frame(n_clusters = 1:15, suma_cuadrados_internos = total_withinss) %>%
  ggplot(aes(x = n_clusters, y = suma_cuadrados_internos)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = 1:15) +
  labs(title = "Evolución de la suma total de cuadrados intra-cluster") +
  theme_bw()
```