Malagón-Selma, P., Debón, A., & Ferrer, A. (2022). Modelos de Machine Learning y estadística multivariante para predecir la posición de los equipos de primera división. *Journal of Sports Economics & Management*, 12(1), 3-22.

MODELOS DE MACHINE LEARNING Y ESTADÍSTICA MULTIVARIANTE PARA PREDECIR LA POSICIÓN DE LOS EQUIPOS DE PRIMERA DIVISIÓN

Machine Learning and Multivariate Statistics models to predict the position of the first division teams

Pilar Malagón-Selma[1], Ana Debón[1], Alberto Ferrer[2]

1 Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Spain
2 Grupo de Ingeniería Estadística Multivariante. Departamento de Estadística e Investigación Operativa Aplicadas y Calidad. Universitat Politècnica de València, Spain

RESUMEN: Esta investigación tiene como objetivo encontrar qué modelos de Machine Learning y Estadística Multivariante tienen una mayor capacidad de predicción de la posición de los equipos al final de la temporada en la tabla clasificatoria. Se han utilizado las acciones de juego de los equipos que compitieron en la primera división de la Bundesliga, Premier League, LaLiga, Ligue 1 y Serie A, a lo largo de la temporada 2018-2019. Los equipos mal clasificados por el mejor de los modelos, el *Random Forest* con datos equilibrados, fueron analizados en profundidad para determinar las acciones del juego que provocaban el error de clasificación. Los resultados indican que, generalmente, la efectividad de cara a portería y la posesión del balón son las variables en las que más difieren los equipos mal clasificados respecto a los valores medios de las variables en su grupo real. En conclusión, esta investigación muestra cómo las técnicas de Machine Learning y Estadística Multivariante se pueden utilizar con éxito para predecir la clasificación final de los equipos que compiten en las mejores ligas del mundo.

PALABRAS CLAVE: Machine Learning, Doble Validación Cruzada, equilibrado de datos, fútbol, estadísticas de juego.

ABSTRACT: *This research aims to find which models of Machine Learning and Multivariate Statistics have a greater predictive capacity when deciding what the team's classification will be at the end of the season. The teams that competed in the first division of the Bundesliga, Premier League, LaLiga, Ligue 1 and Serie A throughout the 2018-2019 season have been studied. The badly classified teams by the best of the models, the Random Forest with balanced data, were analyzed in-depth to determine the game's actions that caused the classification error. The results indicate that, generally, the effectiveness in front of goal and the possession of the ball are the statistics in which badly classified teams differ the most with the average of their real position. In conclusion, this research shows how Machine Learning and Multivariate Statistical techniques can be used successfully to discriminate between Top and Bottom teams competing in the best leagues in the world.*

KEY WORDS: Machine Learning, Doble Cross Validation, Balance data, football, game statistics.

Recibido/received: 30-10-2021 Aceptado/accepted: 18-12-2021

Contact details:

Corresponding author

Pilar Malagón-Selma	Ana Debón	Alberto Ferrer
pimasel@doctor.upv.es	andeau@eio.upv.es	aferrer@eio.upv.es
Universitat Politècnica de València	Universitat Politècnica de València	Universitat Politècnica de València
Edificio 7A	Edificio 7A	Edificio 7A
Camí de Vera, s/n	Camí de Vera, s/n	Camí de Vera, s/n
46022 Valencia	46022 Valencia	46022 Valencia

1. Introducción

Con al menos 158 años de historia, el fútbol sigue siendo el deporte hegemónico de la sociedad actual. Así lo demuestran los datos de audiencia televisiva de los acontecimientos futbolísticos más importantes del mundo: la final de UEFA Champions League de 2018 en la que se enfrentaban Real Madrid y Liverpool alcanzó un récord de audiencia con 350 millones de telespectadores. Este partido fue retransmitido en 226 países y obtuvo más del doble de audiencia que la Super Bowl, vista por 172 millones de telespectadores.

La popularidad de este deporte ha provocado que los equipos de fútbol generen enormes ingresos. Según la consultora Deloitte, solo los 20 clubes con mayor facturación del mundo, todos ellos pertenecientes a los Cinco Grandes (Premier League, Bundelisga, LaLiga, Serie A y Ligue 1), superaron los 9.283 millones de ingresos en la temporada 2018/2019 (Ajadi, Burton, Dwyer, Hammond, y Ross, 2020).

Las cantidades ingentes de millones que se mueven en esta industria y que se centran especialmente en la élite de este deporte ha provocado que exista una extensa literatura sobre la eficiencia en el fútbol (Boscá, Liern, Martínez, y Sala, 2009; Espita-Escuer y Garcia-Cebrian, 2008; Zambom-Ferraresi, García-Cebrián, Lera-López, y Iráizoz, 2017), ya que existe una notable diferencia entre los ingresos que perciben los equipos que alcanzan puestos de Champions League y aquellos que descienden a segunda división de sus ligas.

En los últimos años se han estudiado las variables relacionadas con acciones ofensivas, defensivas, de gol y de pase que resultan estadísticamente significativas a la hora de determinar los equipos que ganan, pierden o empatan un partido de fútbol (Castellano, Casamichana, y Lago, 2012; Lago-Peñas, Lago-Ballesteros, Dellal, y Gómez, 2010; Liu, Gómez, Gonçalves, y Sampaio, 2016). Estos estudios incorporan en algunos casos variables categóricas como la calidad de los equipos, la estrategia de juego o si los equipos compiten como visitantes o locales (Lago, 2009; Lucey, Oliver, Carr, Roth, y Matthews, 2013; Taylor, Mellalieu, James, y Shearer, 2008). La búsqueda de las variables que más contribuyen al ranking final de los equipos al final de la temporada también ha sido objeto de estudio (Brito de Souza, López-Del Campo, Blanco-Pita, Resta, y Del Coso, 2019; Lago-Ballesteros y Lago-Peñas, 2010; Oberstone, 2009). Para estos análisis, los equipos se han clasificado en tres posiciones: equipos que ocupan las cuatro primeras posiciones de sus ligas, las tres últimas posiciones (y, por lo tanto, descienden a segunda división) y los que quedan en mitad de la tabla.

Los análisis citados anteriormente fueron realizados a partir de la aplicación de técnicas univariantes (Brito de Souza et al., 2019; Lago-Ballesteros y Lago-Peñas, 2010), mediante técnicas de regresión (Oberstone, 2009) o análisis discriminante (Castellano et al., 2012; Lago-Peñas, et al., 2010; Liu et al., 2016). Sin embargo, estas investigaciones se centraban en encontrar las variables más discriminantes entre las diferentes posiciones analizadas sin profundizar en la predicción, exceptuando a Oberstone (2009), quien utilizó un modelo de regresión múltiple para predecir los puntos de los equipos de la Premier League

en la temporada 2007-2008. De acuerdo con sus resultados, el poder explicativo del modelo fue de un $R^2=0.990$ y un p-valor<0.001. Para ello, únicamente utilizó 6 variables: efectividad, porcentaje de goles marcados desde fuera del área, ratio de pases largos y cortos, centros exitosos, goles concedidos por partido y tarjetas amarillas. Además, en su artículo también realizó el análisis de la varianza (ANOVA) de un factor para seleccionar las variables más influyentes, pero no se utilizó ningún modelo de clasificación para predecir las posiciones de los equipos.

El objetivo de este trabajo es el de utilizar técnicas de Machine Learning y Estadística Multivariante para predecir la posición que ocuparán los equipos al final de la temporada. Para ello se han utilizado las acciones de juego como variables predictoras. A partir de este estudio, se propondrá la técnica que mejor diferencie entre las posiciones y se estudiará el peso que tienen las variables identificadas por autores previos en la clasificación de los equipos (Brito de Souza et al., 2019; Lago-Ballesteros y Lago-Peñas, 2010; Oberstone, 2009). Este trabajo cobra relevancia al discriminar entre razones objetivas que son susceptibles de ser analizadas, como es el caso de las variables de rendimiento, y que influyen en la clasificación final de los equipos de fútbol, y razones puramente debidas al azar, pues a veces la pelota entra o no entra. Se considera que los resultados y conclusiones de este artículo pueden aportar información de gran utilidad a gestores deportivos, ya que permite utilizar indicadores objetivos para valorar el rendimiento de entrenadores y jugadores, mejorando así el proceso de toma de decisiones, como podría ser la de rescindir el contrato de un entrenador o su renovación.

Este artículo está estructurado en cinco secciones. Tras la introducción, la segunda sección está dedicada a describir la base de datos y resumir cómo se han aplicado las técnicas de Machine Learning y Estadística Multivariante. La tercera sección muestra los resultados de la predicción y el análisis de los equipos clasificados incorrectamente. La cuarta y la quinta sección comprenden, respectivamente, la discusión de los resultados y las conclusiones alcanzadas.

2. Material y Metodología

Esta sección presenta la base de datos analizada y los métodos de análisis utilizados. En primer lugar, se realizó el análisis exploratorio de los datos mediante el análisis de componentes principales (PCA) (Wold, Esbensen, y Geladi, 1987). El PCA es una técnica de análisis no supervisado comúnmente utilizada para la exploración de los datos. Entre sus múltiples ventajas destacan: la posibilidad de detectar valores atípicos, así como la de visualizar la estructura de correlación de las variables.

A continuación, se seleccionaron cinco técnicas de análisis supervisado con el objetivo de determinar el mejor modelo predictivo. Los modelos utilizados fueron: Árboles de clasificación y regresión (CART) (Breiman, Friedman, Olshen, y Stone, 1984), Random Forest (RF) (Breiman, 2001), Naïve Bayes (Maron, 1961), K-vecinos más cercanos (K-NN) (Altman, 1992) y Análisis Discriminante de Mínimos Cuadrados Parciales (PLS-DA) (Wold, Johansson, y Cocchi, 1993).

Una vez realizada la predicción y escogido el mejor modelo, se utilizó el gráfico de radar (Kolence y Kiviat, 1973) para comparar las estadísticas de los equipos mal clasificados con el valor de la posición promedio a la que pertenecían. De esta forma es posible comprobar visualmente el comportamiento de los equipos. Se utilizó el programa informático R para analizar la base de datos (Team, R Core, 2019). R es un software libre a partir del cual pueden realizarse una gran variedad de métodos estadísticos y gráficos.

2.1. Base de datos

Se recopilaron datos sobre el rendimiento de los equipos de fútbol que compitieron en las ligas europeas (LaLiga, Premier League, Bundesliga, Serie A y Ligue 1) a lo largo de la temporada 2018-2019. Las fuentes de Internet consultadas fueron: WhoScored, FBref y Fichajes.net. Por tanto, la base de datos utilizada para el análisis estaba formada por 98 observaciones (los equipos de fútbol), 48 variables explicativas (variables de rendimiento) y por la variable respuesta "ranking". Con vistas al objetivo propuesto, los equipos de fútbol fueron etiquetados según su posición final en la liga nacional. Se definieron tres posiciones: “Alto” para aquellos clubs cuya clasificación les permitía participar en la Champions League (20 equipos), “Bajo” para los clubes que descendieron a segunda división (15 equipos), y “Centro” para el resto (63 equipos). La Tabla 1 muestra las variables utilizadas para llevar a cabo el análisis predictivo.

Tabla 1. Variables clasificadas por el tipo de acción de juego y sus abreviaturas**.**

Tipo de variables	**Acciones de juego y abreviaturas**
Variables relacionadas con acciones defensivas	Remates concedidos bloqueados (SCB), Recuperaciones (R), Portería a cero (CS), Intercepciones (I), Remates concedidos desde dentro del área (SCTI), Remates concedidos desde fuera del área (SCTO), Entradas ganadas (TW), Entradas perdidas (TL), Despejes (Cl) y Precisión de entradas (TA), Tarjetas amarillas (YC), Faltas concedidas (FC) y Penaltis condedidos (PC)
Variables relacionadas con acciones ofensivas	Córneres ganados (CW), Centros fallados (CU), Centros exitosos (SC), Regates exitosos (DS), Faltas recibidas (FW), Penaltis recibidos (PT), Regates fallados (DU), Precisión de córner (CrA) y Precisión de regates (DrA)
Variables relacionadas acciones de gol o creación	Precisión de goles (GA), Goles desde dentro del área (GIB), Pases clave (KP), Penaltis recibidos (PT), Goles de tiro libre directo (DFKG), Remates desde fuera de portería (SOT), Remates bloqueados (SB), Precisión de remate (SA), Asistencias (A) y Remates con dirección a portería (ST).
Variables relacionadas con acciones de pase o posesión	Precisión de pase (PA), Porcentaje de posesión (AP), Duelos perdidos (DlL), Duelos ganados (DlW), Pases fallados en el campo del rival (PUOpp), Pases exitosos en el campo del rival (PSOpp), Pases largos exitosos (SLP), Pases cortos fallados (PUS), Pases cortos exitosos (PSS), Porcentaje de pases largos exitosos (LPS), Pases largos fallados (ULP), Duelos aéreos ganados (ADL), Porcentaje de duelos aéreos ganados (ADA), Porcentaje de duelos ganados (Dl_A), Pases por 90 mins (P_90), Porcentaje de pases exitosos en el campo del rival (PAOppH) y Porcentaje de pases exitosos en el propio campo (PAOwnH)

2.2. Métodos de análisis

Métodos de aprendizaje no supervisado

El análisis exploratorio de los datos se llevó a cabo a partir del análisis de componentes principales (PCA). Esta técnica de aprendizaje no supervisado utiliza la correlación entre las variables para crear otras nuevas (componentes principales). Como resultado se obtiene un número menor de variables que no se encuentran correlacionadas entre sí y que explican la mayor parte de la variabilidad de los datos (Wold et al., 1987). Utilizando las componentes principales (CP) resultantes en los ejes podemos obtener representaciones de los datos que en ocasiones permiten explorar la separabilidad de las clases.

Métodos de aprendizaje supervisado

Con el objetivo de seleccionar el modelo con mayor capacidad predictiva se han utilizado diferentes técnicas de aprendizaje supervisado, procedentes tanto del Machine Learning (Random Forest, Naive Bayes, Árbol de clasificación, Vecino más cercano), como de la Estadística Multivariante (PLS-DA). Estas técnicas emplean un conjunto de datos (conjunto de entrenamiento) en el que se incluyen datos de entrada y valores de respuesta. A partir de ellas, se busca crear un modelo que pueda realizar predicciones de los valores de respuesta para un nuevo conjunto de datos (conjunto de prueba) independiente del conjunto de entrenamiento, y que se utilizará para la validación del modelo (Hormozi, Hormozi, y Nohooji, 2012). Son técnicas que, en su mayoría, pueden aplicarse tanto a regresión como a clasificación, aunque en este artículo se aplicaran para clasificación.

Árboles de clasificación y regresión (CART) es el nombre que recibe el algoritmo del árbol de decisión (Breiman et al., 1984). El modelo está estructurado sobre una secuencia de preguntas, a partir de las cuales se crea el árbol. El árbol se construye a través de "nodos" que dividen los datos en función de sus características hasta el nodo "hoja" donde se clasifican los datos por clase y probabilidad según el camino tomado (Nisbet, Elder, y Miner, 2009).

El algoritmo Random Forest (Breiman, 2001) es un método de conjunto[1] (Opitz y Maclin, 1999) que utiliza la técnica Bagging[2] para combinar árboles aleatoriamente y mejorar la capacidad de predicción (Breiman, 1996). Además, cada split se realiza a partir de un subconjunto de variables aleatoriamente seleccionado de entre todas. De manera sucinta, el proceso comienza con la selección aleatoria de los individuos para crear un conjunto de datos diferente. Cada conjunto de datos crea un árbol de decisiones y se utilizan como conjuntos de entrenamiento para predecir un subconjunto que no se ha utilizado para el entrenamiento. Finalmente, los individuos de prueba se clasifican de acuerdo con cómo se han predicho los individuos en la mayoría de los árboles.

Naïve Bayes es una técnica basada en el teorema de Bayes que explica cómo a partir de un evento conocido que toma condiciones particulares es posible conocer la probabilidad de

[1] Los métodos de conjunto utilizan múltiples algoritmos para lograr una mejor predicción.
[2] Realiza un entrenamiento repetido del conjunto de datos a través de un subconjunto aleatorio.

que también ocurra otro evento con características similares. Esta técnica supone que las variables son independientes entre ellas (Maron, 1961).

K-vecinos más cercanos (K-NN) es un algoritmo de los llamados *lazy learners*. A partir de la información obtenida en la etapa de entrenamiento, los individuos se clasifican según la clase de *K* vecinos más cercanos. El valor de *K* lo determina el analista e indica el número de observaciones que debe utilizar el algoritmo para clasificar a un individuo. La métrica de distancia a utilizar suele ser la distancia euclídea (Altman, 1992).

El análisis discriminante de mínimos cuadrados parciales (PLS-DA) es una variante del PLS. Los modelos PLS encuentran variables latentes en el espacio X (predictores) y el espacio Y (respuestas) con máxima covarianza. El resultado es un número de variables latentes que explican tanto la variabilidad en el espacio X como la relación entre los espacios X e Y. La primera variable latente tiene más información que la segunda, la segunda variable latente más que la tercera, y así sucesivamente (Höskuldsson, 1988). Una de las principales ventajas sobre otros modelos predictivos es que admiten regresores correlacionados, dando lugar a modelos fácilmente interpretables.

2.3. Técnica de sobremuestreo de minorías ponderadas por mayoría (MWMOTE)

Como se indicó anteriormente los equipos fueron etiquetados en base a su posición en el ranking al final de la temporada (Bajo, Centro y Alto). Sin embargo, de acuerdo con los criterios utilizados el conjunto de datos resultó desequilibrado, es decir, las clases no se encontraban representadas por igual.

El principal inconveniente de trabajar con una base de datos desequilibrada es que los algoritmos tienden a clasificar los equipos en la clase mayoritaria pues conduce a minimizar la tasa de error. Por lo tanto, para tratar de discriminar las clases minoritarias, se consideró el uso de una técnica de equilibrado de las clases. Tras comprobar el rendimiento de diferentes métodos de balanceo equilibrado de datos (Barua, Islam, Yao, y Kazuyuki, 2014; Cordón, García, Fernández, y Herrera, 2018), se seleccionó la técnica de sobremuestreo de minorías ponderadas por mayoría (MWMOTE) por ofrecer los mejores resultados.

La técnica MWMOTE se divide en tres etapas. En la primera, identifica la clase minoritaria en el conjunto de datos y recopila información relevante sobre los individuos minoritarios. En la segunda etapa, calcula los pesos de las observaciones de acuerdo con tres ítems: los individuos más cercanos al conglomerado mayoritario, los subgrupos de clase minoritaria que se encuentran en grupos dispersos, y las observaciones de clase minoritaria cerca de un grupo de mayoría compacta, que tendrán mayor peso que el resto. En la tercera etapa, la técnica crea el nuevo conjunto de datos sintéticos a partir de las ponderaciones. Para la aplicación de esta técnica de equilibrado de las clases, se ha utilizado el paquete de R imbalance (Cordón et al., 2018). A través de la función con el mismo nombre, es posible construir el modelo siguiendo las especificaciones explicadas anteriormente. En primer lugar, se utiliza el argumento *KNoisy* para filtrar aquellos individuos del conjunto minoritario que se encuentran rodeados únicamente por individuos de clase mayoritaria,

así se elimina el posible ruido de los datos y también se evita que el nuevo conjunto de datos lo contenga. En segundo lugar, el parámetro *KMajority* detecta la posición de los individuos limítrofes de la clase mayoritaria. A partir de esta información, se asignan los pesos, ya que estas observaciones se consideran más difíciles de aprender que los individuos rodeados por las mismas observaciones de clase. En tercer lugar, a través del parámetro *KMinority*, se indica el número de muestras que se van a utilizar para crear los individuos sintéticos. En cuarto lugar, se ha utilizado el parámetro *cclustering* para designar el espacio que las nuevas muestras van a ocupar. Los valores utilizados en la investigación cambiaron para las clases Alto y Bajo pues, según lo descrito por la bibliografía, no existen cifras genéricas, sino que deben investigarse aquellos valores que ofrezcan un mejor resultado en cada caso (Barua et al., 2014). Se generaron 30 observaciones sintéticas para los equipos Alto y 24 para los Bajo, ya que de acuerdo con (Japkowicz, 2000) el número de muestras sintéticas creadas debe ser el 200 por cien del original.

2.3. Validación de los métodos de aprendizaje supervisado

La técnica de doble validación cruzada (2CV) se utilizó para el ajuste y la evaluación de los modelos (ver Figura 1). Esta técnica se seleccionó porque permite optimizar los parámetros y evaluar el modelo con conjuntos de datos sin solapamientos evitando así problemas de sobreajuste (Stone, 1974). En la 2VC la base de datos se divide aleatoriamente en F subgrupos. A partir de esta división, se crea el conjunto de entrenamiento con F-1 subgrupos (80% de los datos) que se utiliza para la obtención del modelo óptimo, y el conjunto de validación (subgrupo restante) que se reserva para la validación (20% de los datos). Una vez realizada la primera división (VC2), el conjunto de entrenamiento (VC1) vuelve a separarse de forma aleatoria en dos nuevos conjuntos: el de calibración, formado por F-1 subgrupos (80% de los datos)); y el conjunto de prueba, formado por el subgrupo restante (20% de los datos)). El conjunto de entrenamiento (VC1) se utiliza para la optimización de los hiperparámetros: *mtry*, número de variables en cada árbol en el algoritmo Random Forest; *K*, número de vecinos más cercanos que utiliza cada observación para su clasificación en el algoritmo K-NN; en el clasificador Naïve Bayes, el parámetro *usekernel* elige entre usar una estimación de densidad de kernel o una estimación de densidad gaussiana; el parámetro *Cp* indica cuándo el árbol de clasificación debe dejar de crecer y, por lo tanto, deja de agregar variables al árbol de decisión; *VL*, número óptimo de variables latentes para construir el modelo PLS-DA. En la VC1 se calcula el modelo óptimo que se utilizará para realizar la predicción en el conjunto de validación (CV2). Este proceso se repite F veces, utilizando conjuntos de datos no solapados para entrenar y validar el modelo, concluyendo cuando todos los individuos han estado una vez (solamente una vez) en ambos grupos (Szymanska, Saccenti, Smilde, y Westerhuis, 2012; Westerhuis et al., 2008).

La técnica de sobremuestreo se aplicó al conjunto de entrenamiento en cada iteración. Por tanto, mientras que los parámetros óptimos fueron calculados a partir de una base de datos formada por observaciones reales y creadas mediante la técnica MWMOTE, la validación del modelo se realizó únicamente sobre datos reales. Investigaciones anteriores

defienden que esta es la forma correcta de validar los resultados en un contexto de clases desequilibradas (Santos, Soares, Abreu, Araujo, y Santos, 2018).

Una vez finalizadas las *F* repeticiones en VC1 y VC2, se promedia el valor del Coeficiente de Correlación de Matthews (CCM) obtenido en cada bucle de cada modelo (ver Figura 1). El CCM servirá para evaluar el modelo con mayor capacidad predictiva. Este coeficiente utiliza el resultado de la matriz de confusión para calcular la calidad de la predicción:

$$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FP)}}$$

Donde TP y TN son, respectivamente, el número de verdaderos positivos y verdaderos negativos, y FP y FN el número de falsos positivos y falsos negativos, respectivamente. El valor del CCM puede oscilar entre -1 y 1, donde 1 representa una predicción perfecta y -1 es indicativo de que no existe relación alguna entre las observaciones y la predicción (Matthews, 1975).

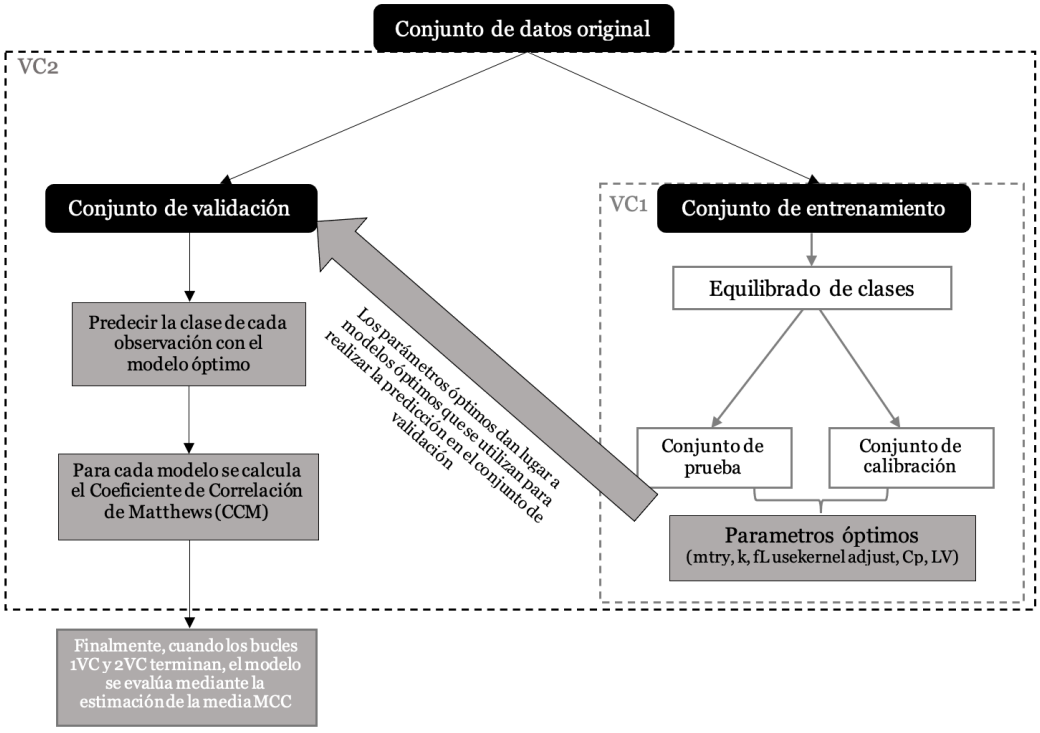

Figura 1. Diagrama de la doble validación cruzada utilizada para evaluar los modelos de clasificación.

2.4. Gráfico de radar

Una vez realizada la predicción se utilizará el gráfico de radar para analizar los equipos mal clasificados. El gráfico de radar es una herramienta de gran utilidad que permite representar datos multivariados en dos dimensiones (Saary, 2008). Estos gráficos se caracterizan por su forma circular y por los radios que se proyectan desde el punto central. Los valores de las variables se escalan según la longitud de los radios y los valores se trazan sobre un plano bidimensional (Budsaba, Smith y Riviere, 2000). Esto permite comparar de forma rápida y sencilla varias observaciones a la vez en función de algunas variables o

criterios. En los últimos años esta herramienta descriptiva ha tomado notoriedad en el análisis de datos deportivos. Empresas como Opta (Mark y Sormaz, 2019), Statbomb (Knutson, 2020) o Driblab (Driblab, 2020) utilizan a menudo estas figuras para evaluar equipos y jugadores (Pérez, 2019). También varios investigadores han comenzado a ilustrar las variables de rendimiento mediante gráficos de radar (Liu et al., 2016; Liu, Yi, Giménez, Gómez y Lago-Peñas, 2015; Oberstone, 2009).

En nuestro caso, las acciones de juego sobre las que se realizará el análisis son aquellas que en artículos previos (Brito de Souza et al., 2019; Lago-Ballesteros y Lago-Peñas, 2010; Oberstone, 2009) fueron destacadas como variables estadísticamente significativas para diferenciar entre posiciones (ver Tabla 2).

Tabla 2. Variables estadísticamente significativas para diferenciar entre los equipos etiquetados como Alto, Centro y Bajo.

Variables	**Lago-Ballesteros y Lago-Peñas (2010)**	**Oberstone (2009)**	**Brito de Souza et al. (2019)**
Variables relacionadas con acciones defensivas	-	YC y FC	SCTI, SCTO, R, YC, FC y PC
Variables relacionadas con acciones ofensivas	-	CW	CW, PT y FW
Variables relacionadas con acciones de pase o posesión	AP	PA, PSS y LPS	PA
Variables relacionadas acciones de gol o creación	ST, GA y A	ST y GA	SA, DFKG y GA

Tarjetas amarillas (YC), Faltas concedidas (FC), Remates concedidos desde dentro del área (SCTI), Remates concedidos desde fuera del área (SCTO), Recuperaciones (R), Penaltis concedidos (PC), Córneres ganados (CW), Penaltis recibidos (PT), Faltas recibidas (FW), Porcentaje de posesión (AP), Precisión de pase (PA), Pases cortos exitosos (PSS), Porcentaje de pases largos exitosos (LPS), Remates con dirección a portería (ST), Precisión de goles (GA), Asistencias (A), Precisión de remate (SA), Goles de tiro libre directo (DFKG).

3.Resultados

El estudio exploratorio se llevó a cabo mediante el análisis de componentes principales (PCA). A continuación, se realizó la predicción de los equipos mediante cinco modelos de aprendizaje supervisado (árbol de clasificación, algoritmo *random forest*, Naïve Bayes, k-vecinos más cercanos y PLS-DA). Los equipos mal clasificados fueron estudiados con la finalidad de conocer las acciones de juego que conducían a error a los algoritmos propuestos.

3.1. Aplicación de métodos de aprendizaje no supervisados

El PCA fue obtenido a partir de la biblioteca FactoMineR (Le, Josse, y Husson, 2008). La Figura 2 muestra la proyección de los equipos en el plano de las dos primeras componentes principales (CP), con indicación de su categoría en el ranking, de forma que es posible visualizar la relación de los equipos entre grupos y dentro de los grupos.

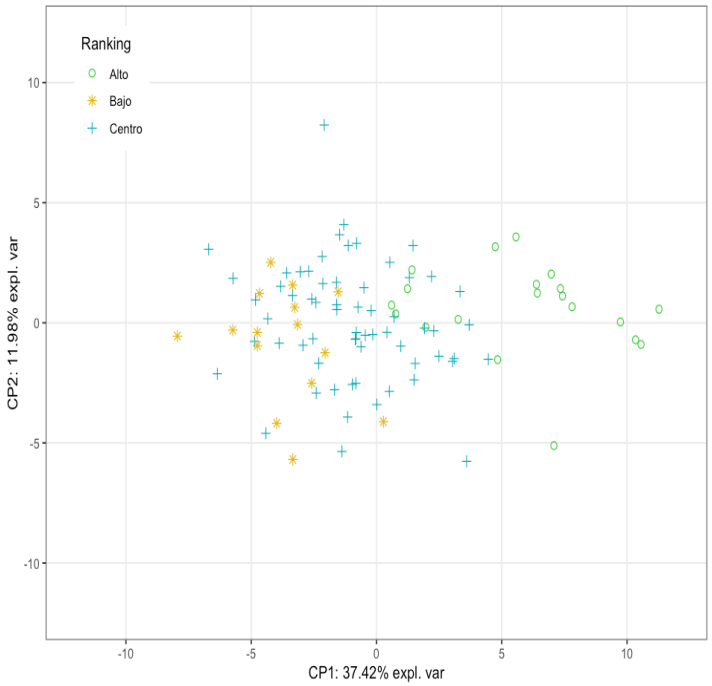

Figura 2. Diagrama de dispersión de las puntuaciones (scores) de los equipos en las dos primeras componentes principales (a distribución de los equipos en función del ranking; se proyectan en el CP1 / CP2) con indicación de su categoría en el ranking.

Es posible observar cómo a lo largo de la CP_1 se separan los equipos Alto y Bajo, mientras que los equipos Centro aparecen entre los dos anteriores, solapándose parcialmente con ambos grupos. Además, se muestra gráficamente que la base de datos tiene las clases desequilibradas, siendo la clase Centro la más numerosa.

3.2. Aplicación de métodos de aprendizaje supervisados

Tras haber aplicado la metodología descrita en la sección 2.3, la Tabla 2 muestra la media de los Coeficiente de Correlación de Mathews (CCM) de las F=5 repeticiones para cada modelo obtenido mediante la librería caret (Kuhn, 2020) de R. Este proceso fue repetido sobre los datos desequilibrados con el objetivo de comprobar qué metodología resultaba más eficiente.

Tabla 3. Valores del Coeficiente de Correlación de Matthews de los modelos de aprendizaje supervisado para datos desequilibrados y equilibrados

Modelos	Árbol de Clasificación	Random Forest	K-vecinos cercanos	Naïve Bayes	PLS-DA
Desequilibrado	0,558	0,663	0,633	0,615	0,631
Equilibrado	0,615	**0,722**	0,508	0,643	0,536

Según los resultados de la Tabla 2, ninguno de los modelos proporciona una clasificación destacada. Sin embargo, es posible concluir que trabajar con datos equilibrados proporciona, en algunos casos como en Árboles de Clasificación, Naïve Bayes y Random Forest, mejores resultados para el conjunto de validación. El algoritmo *Random Forest* es el modelo que mejor clasifica (destacar que este análisis fue repetido en varias ocasiones

y en todas ellas el algoritmo *Random Forest* con el conjunto de datos equilibrado fue el modelo que aportaba mejores resultados). La Tabla 3 muestra la matriz de confusión del algoritmo *Random Forest* para las F=5 repeticiones.

Tabla 4. Matriz de confusión general del algoritmo Random Forest.

Predichos/Observados	Bajo	Centro	Alto
Bajo	**7**	2	0
Centro	8	**55**	5
Alto	0	3	**15**

Los resultados de la Tabla 3 completan la información aportada por los valores del CCM (Tabla 2). En la diagonal principal de la Tabla 3 se indica el número de equipos que han sido clasificados adecuadamente. Los equipos fuera de la diagonal son aquellos cuya predicción ha sido errónea. De los 60 equipos Centro, 2 han sido confundidos con equipos Bajo y 3 con Alto; de los 20 equipos Alto, 5 han sido clasificados como Centro; por último, de los 15 equipos Bajos, 8 se han clasificado como Centro, siendo en esta clase donde mayor es el porcentaje de confusión. Estas confusiones entre equipos de clases contiguas ya se intuían en el análisis PCA exploratorio de la Figura 2, sobre todo, entre las clases Bajo y Centro. Sin embargo, es importante destacar que no ha habido confusión entre clases extremas, es decir, ningún equipo Bajo ha sido clasificado como Alto, ni viceversa.

3.3. Gráfico de radar

La Figura 3 muestra un ejemplo de las variables que serán utilizadas para llevar a cabo el análisis comparativo, así como la apariencia del gráfico de radar creado mediante el paquete plotly (Sievert, 2020) de R. En los ejes del gráfico de radar se han indicado las variables que en artículos previos han resultado estadísticamente significativas a la hora de diferenciar entre posiciones.

La Figura 3 destaca cómo los equipos Alto son aquellos que realizan un mayor número de acciones de gol/creación (GA, ST, SA, A y DFKG), ofensivas (CW, PT) y de posesión/pase (PA, AP, PSS y LPS) y un menor número de acciones defensivas (YC, FC, SCTI, SCTO, PC) a excepción de la variable número de recuperaciones (R) cuya media es mayor para los equipos Alto. Por el contrario, los equipos Bajo realizan un menor número de acciones de gol/creación (GA, ST, SA, A y DFKG), ofensivas (CW, PT y FW) y de posesión/pase (PA, AP, PSS y LPS) y un mayor número de acciones defensivas (YC, FC, SCTI, SCTO, PC). Las medias de los equipos Centro en estas variables se encuentran, en general, entre las de los equipos Alto y Bajo. Como excepciones se observa que el número de faltas recibidas (FW) es la misma para los tres grupos, y que la media de los equipos Centro y Bajo es la misma en el caso de faltas realizadas (FC) y tarjetas amarillas recibidas (YC).

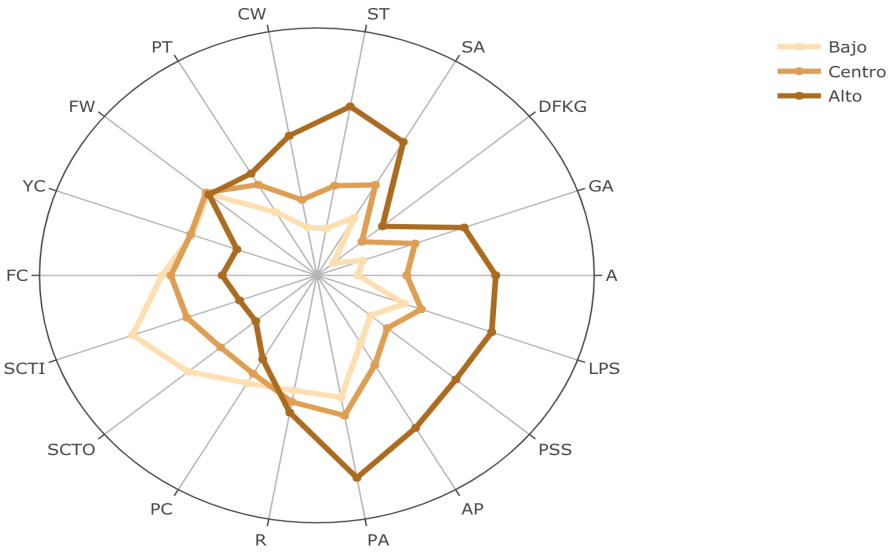

Figura 3. Gráfico de radar utilizado para la comparación de los valores medios de las acciones del juego estadísticamente significativas para diferenciar entre posiciones de los equipos Bajo, Centro y Alto.

Con el objetivo de profundizar en la predicción realizada por los modelos y comprender la razón de los errores cometidos, se han estudiado aquellos equipos mal clasificados por el algoritmo Random Forest. Se han analizado conjuntamente los 2 equipos Centro clasificados como Bajo (Figura 4), los 8 equipos Bajo clasificados como Centro (Figura 5), los 3 equipos Centro clasificados como Alto (Figura 6) y los 5 equipos Alto clasificados como Centro (Figura 7). Sobre el gráfico de radar, se han proyectado las estadísticas de cada equipo y el número promedio de acciones que realizaron los equipos Bajo, Centro y Alto.

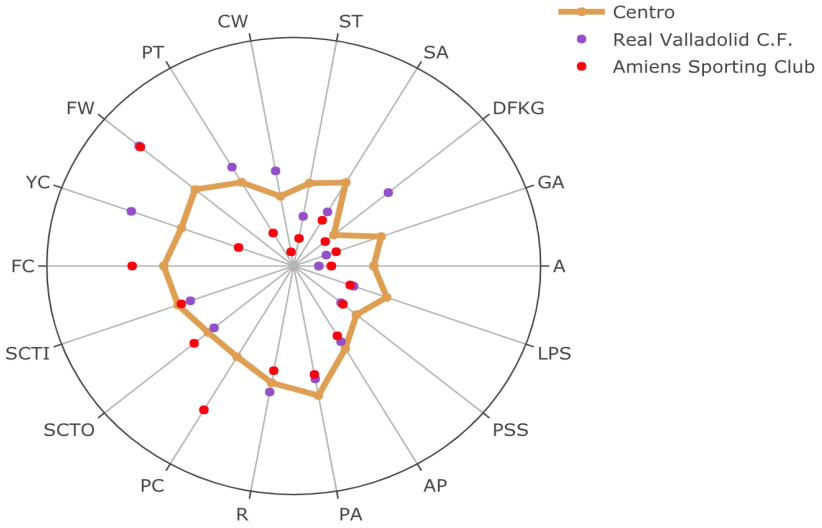

Figura 4. Gráfico de radar para la comparación de los equipos mal clasificados como Bajo con los valores medios de las acciones del juego (estadísticamente significativas para diferenciar entre posiciones) de los equipos Centro.

La Figura 4 muestra las variables de juego de los dos equipos Centro (posición real) que fueron clasificados como Bajo (posición predicha): Real Valladolid C.F. (punto violeta) y Amiens Sporting Club (punto rojo).

A partir del gráfico de radar de la Figura 4 se observa que ambos equipos ejecutaron un número de acciones de gol/creación (GA, ST, SA y A) menor que la media de los equipos Centro. Además, el Amiens Sporting Club (punto rojo) llevó a cabo menos acciones de posesión/pase (PA, AP, PSS y LPS) y más acciones defensivas (PC, SCTO y FC) que el promedio de los equipos Centro.

La Figura 5 muestra los valores de las variables de rendimiento de los 8 equipos Bajo (posición real) que fueron clasificados como Centro (posición predicha): Rayo Vallecano de Madrid (punto granate), Sociedad Deportiva Huesca (punto azul oscuro), Girona F.C. (punto amarillo), Dijon Football Côte d'Or (punto verde), 1.F.C. Núremberg (punto negro), Fulham (punto naranja), En Avant de Guingamp (punto rosa) y Cardiff City F.C. (punto azul celeste).

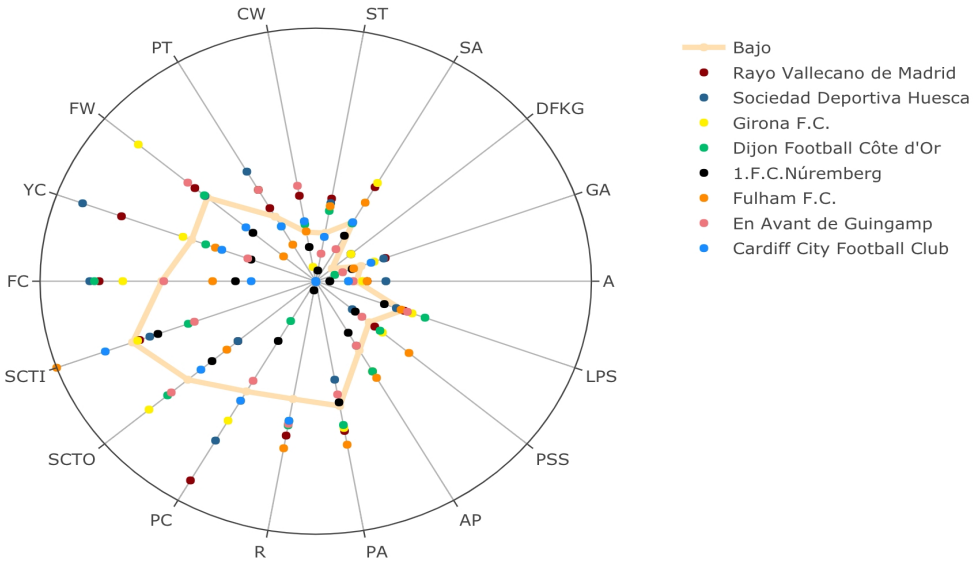

Figura 5. Gráfico de radar para la comparación de los equipos mal clasificados como Centro con los valores medios de las acciones del juego (estadísticamente significativas para diferenciar entre posiciones) de los equipos Bajo.

A la luz de la Figura 5 se comprueba que los equipos que compitieron en LaLiga (Rayo Vallecano de Madrid (punto granate), Sociedad Deportiva Huesca (punto azul oscuro) y Girona F.C. (punto amarillo)) realizaron un número más elevado de acciones de gol/creación (GA, ST, SA y A) y de pase/posesión (AP) que el promedio de los equipos Bajo. Además, el Rayo Vallecano de Madrid y de la Sociedad Deportiva Huesca recibieron un menor número de remates desde fuera del área (SCTO) que el promedio de los equipos Bajo. El Dijon F.C.O (punto verde) llevó a cabo un mayor número de acciones de pase/posesión (PA, AP y LPS) y un menor número de acciones defensivas (YC y SCTI) que el promedio de los equipos Bajo. El 1.F.C. Núremberg (punto negro) y el Cardiff City F.C. también realizaron pocas acciones defensivas (PC, SCTO, SCTI, FC y YC) respecto al promedio de los equipos Bajo. El resto de los equipos, Fulham F.C. (punto naranja) y En

Avant de Guingamp (punto rosa), presentaron diferencias con la posición Bajo (posición real) en las acciones de juego de gol/creación (GA, ST, SA y A), pase/posesión (AP) y defensivas (SCTI, SCTO).

La Figura 6 proporciona información de las variables de rendimiento de los 3 equipos Centro (posición real) clasificados como Alto (posición predicha): Real Betis Balompié (punto verde), Olympique de Marsella (punto azul celeste) y Associazione Calcio Milan (punto granate).

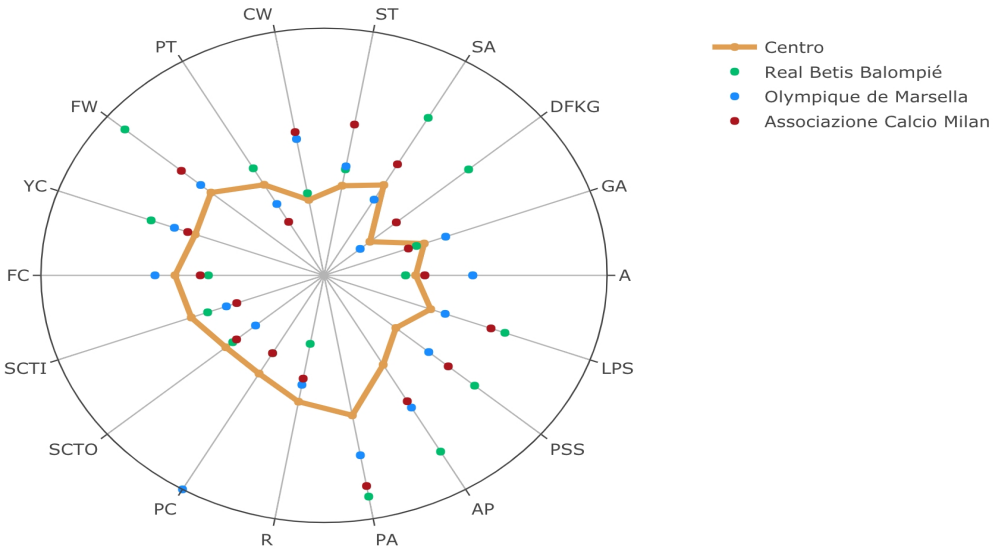

Figura 6. Gráfico de radar para la comparación de los equipos mal clasificados como Alto con los valores medios de las acciones del juego (estadísticamente significativas para diferenciar entre posiciones) de los equipos Centro.

La Figura 6 destaca que los tres equipos Centro clasificados erróneamente como Alto realizaron más acciones de pase/posesión (PA, AP, PSS y LPS) y menos defensivas (R, SCTI y SCTO) que la media de los equipos Centro (posición real). Además, el Olympique de Marsella (punto azul celeste) destaca por su elevado número de acciones de gol/creación (GA y A) en comparación con el resto. Ocurre lo mismo con el Real Betis Balompié (punto verde) y la Associazione Calcio Milan (punto granate), respecto a las acciones de gol/creación (ST, SA y DFKG).

La Figura 7 muestra los valores de las variables de rendimiento de los 5 equipos Alto (posición real) que fueron clasificados como Centro (posición predicha): Association Sportive de Saint-Étienne (punto verde oscuro), Lille Olympique Sporting Club (punto azul oscuro), RasenBallsport Leipzig (punto rojo), Club Atlético de Madrid (punto azul claro) y Valencia Club de Fútbol (punto naranja).

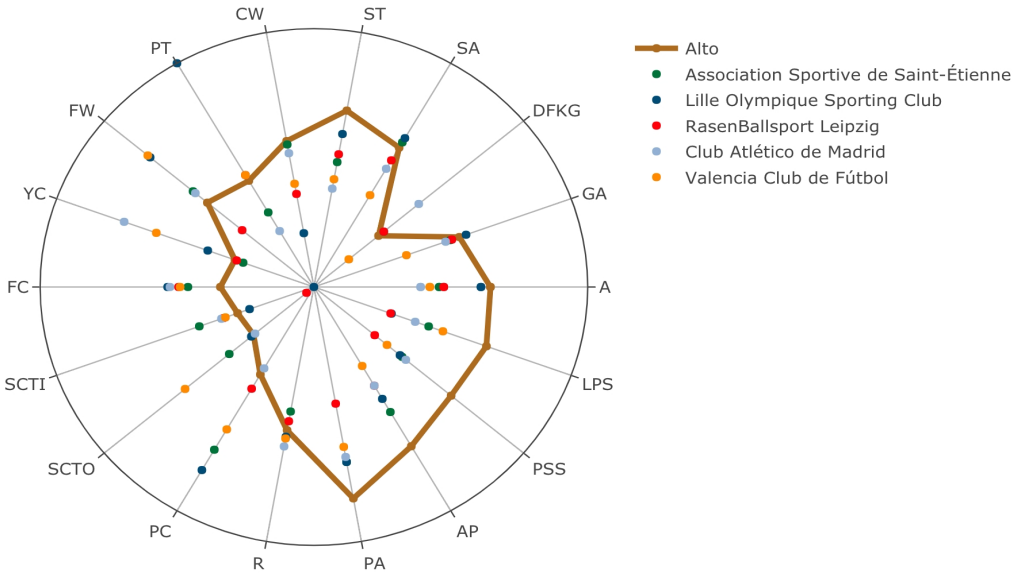

Figura 7. Gráfico de radar para la comparación de los equipos mal clasificados como Centro con los valores medios de las acciones del juego (estadísticamente significativas para diferenciar entre posiciones) de los equipos Alto.

La Figura 7 pone de manifiesto que los cinco equipos clasificados erróneamente realizaron menos acciones de pase/posesión (PA, AP, PSS y LPS) y ofensivas (ST y CW), y más defensivas (FC) que la media de los equipos Alto (posición real). En general, los equipos tuvieron un comportamiento defensivo similar, puesto que recibieron un elevado número de remates a puerta desde dentro del área (SCTI) (Association Sportive de Saint-Étienne (punto verde oscuro), Valencia Club de Fútbol (punto naranja) y el Club Atlético de Madrid (punto azul claro)), y desde fuera del área (SCTO) (Association Sportive de Saint-Étienne (punto verde oscuro) y Valencia Club de Fútbol (punto naranja)). Finalmente, en el caso del Association Sportive de Saint-Étienne (punto verde oscuro), Lille Olympique Sporting Club (punto azul oscuro), RasenBallsport Leipzig (punto rojo) y Valencia Club de Fútbol (punto naranja), también realizaron un elevado número de penaltis (PC).

4. Discusión

El objetivo de este artículo es utilizar técnicas de Machine Learning y Estadística Multivariante para predecir la posición en la clasificación final de los equipos de fútbol profesional que compitieron en las cinco grandes ligas (Premier League, Bundelisga, LaLiga, Serie A y Ligue 1) a lo largo de la temporada 2018/2019. Además de proponer el mejor modelo predictivo, este artículo analiza los equipos de fútbol mal clasificados en base a las acciones de juego identificadas en artículos previos (Brito de Souza et al., 2019; Lago-Ballesteros y Lago-Peñas, 2010; Oberstone, 2009). Este análisis estadístico cobra relevancia por tratarse de un estudio llevado a cabo con datos de los equipos que compitieron en la primera división de las cinco ligas más importantes del mundo. Además, se han utilizado técnicas de Machine Learning y Estadística Multivariante, y el lenguaje de programación de R para realizar los cálculos, aportando un método de análisis diferente al utilizado en la mayoría de los estudios precedentes. Hasta donde sabemos,

este análisis es el primero en utilizar métodos de Machine Learning y Estadística Multivariante para predecir la posición final de los equipos que compitieron en las cinco grandes ligas, concluyendo que el algoritmo Random Forest, usando conjuntos de datos equilibrados, es el mejor modelo para llevar a cabo la predicción de las posiciones. Se recomienda el uso de la doble validación cruzada para evitar el sobreajuste de los datos.

Aunque la predicción de las posiciones no ha sido correcta en el 100%, se destaca que no se han obtenido resultados extraños (Tabla 3), como, por ejemplo: que un equipo Alto sea clasificado como Bajo (y viceversa), que un equipo que ha ocupado los primeros puestos del centro de la tabla haya sido predicho como un equipo que descendió al final de la temporada, o que un equipo que ocupo los últimos puestos de la tabla fuera predicho como un equipo que optó a competir la Champions League. Al contrario, equipos como el Real Valladolid C.F. y el Amiens Sporting Club, que finalizaron la temporada en decimosexto y decimoquinto lugar (empatado a puntos con el decimosexto), respectivamente, fueron predichos como equipos Bajo, siendo los equipos que descendieron a segunda división los que ocupaban las posiciones decimoctava, decimonovena y vigésima. El Dijon F.C.O. terminó la temporada como decimoctavo. Sin embargo, en la Ligue 1 ocupar esta posición no supone el descenso inmediato, sino que el equipo debe participar en una eliminatoria para optar a mantenerse en primera división. En esta temporada, el Dijon F.C.O. no descendió. Del mismo modo, el Olympique de Marsella y la Associazione Calcio Milan quedaron en quinto puesto y, por tanto, fueron considerados como Centro, mientras que el Valencia Club de Fútbol y la Association Sportive de Saint-Étienne quedaron en la cuarta posición y fueron etiquetados como Alto.

Para profundizar en los errores de predicción se han utilizado las variables identificadas en artículos anteriores como estadísticamente significativas a la hora de diferenciar entre la posición de los equipos (Figuras 4, 5, 6 y 7). Los equipos mal clasificados mostraron que tenían valores de efectividad (GA) distintos a la media de su posición en todos los casos, exceptuando los equipos Alto clasificados como Centro. En el caso de las variables relacionadas con acciones de pase/posesión, la mayoría de los equipos mal clasificados diferían con la media de su posición, destacando la posesión (AP) y la efectividad de pase (PA). De entre las variables defensivas analizadas, el número de remates recibidos desde fuera del área (SCTO) y desde dentro (SCTI) son las variables con las que los equipos mal clasificados más difieren respecto a los valores medios de su clase.

El fútbol es un deporte donde el azar juega un papel importante y donde interactúan múltiples factores, lo que hace difícil juzgar el rendimiento de un equipo en base a un solo indicador, como podría ser su posición en la tabla al final de la temporada. Este trabajo cobra relevancia al ofrecer una metodología para modelar la parte objetiva del fútbol (rendimiento del equipo). La metodología de análisis descrita es una herramienta útil para directores deportivos ya que permite cuantificar el rendimiento de entrenadores y jugadores más allá del resultado al final de la temporada.

El análisis de los equipos mal predichos (ver apartado 3.3) muestra que, en ocasiones, el rendimiento esperado (posición predicha por los modelos predictivos) puede diferir del rendimiento real (posición final en la tabla). El sentido de estas discrepancias entre la

posición predicha y la real puede ayudar en el proceso de toma de decisiones de los gestores deportivos. Por ejemplo, imaginemos un equipo que generó un número razonable de ocasiones de gol, que recibió pocos disparos y que tuvo bastante posesión durante la temporada. Según el modelo predictivo se esperaría que acabase en posiciones medias de la tabla. Sin embargo, si con estos números el equipo desciende, podría argumentarse que este descenso podría ser el resultado de la mala suerte (factor azar) y no de que el equipo lo haya hecho mal. Si el director deportivo de este equipo (que lo ha hecho bien, pero ha descendido) tiene que decidir si rescinde al entrenador, podría encontrar argumentos para mantenerle en el puesto ya que, de acuerdo con el análisis realizado, en condiciones normales se esperaría que sus buenos indicadores en las variables estudiadas se tradujesen en buenos resultados deportivos.

Este análisis demuestra que en el proceso de toma de decisiones la observación de un solo indicador como es la posición del equipo a lo largo de la temporada puede no ser el mejor indicador para juzgar el rendimiento del equipo o entrenador.

5. Conclusión

Tras la discusión de los resultados se concluye, en primer lugar, que el uso de técnicas de análisis de datos es de gran utilidad a la hora de elaborar la estrategia de un equipo, ya que estos análisis sirven de guía para entrenadores y analistas que podrían comprobar si el juego de sus equipos hace peligrar su permanencia en primera división o el mantenerse en puestos de Champions League. Para ello, el gráfico de radar ha resultado de gran utilidad, puesto que el análisis de los equipos mal clasificados ha mostrado que su actuación a lo largo de la temporada fue más propia de equipos que terminaron en otras posiciones.

En segundo lugar, este artículo se presenta como complementario a otros estudios previos cuyo objetivo se centraba en establecer las acciones de juego que más contribuyen al éxito o fracaso de los equipos de fútbol.

Este artículo podría ser de gran utilidad para gestores deportivos, analistas y expertos en materia futbolística, ya que a lo largo del artículo se ha demostrado que la actuación de los equipos no siempre ofrece los resultados esperados. Por tanto, este análisis permite a entrenadores y directivos valorar el desempeño de los equipos de fútbol al final de la temporada, más allá de la clasificación, así como encontrar de forma rápida y visual las debilidades de los equipos gracias al uso de los gráficos de radar.

6. Agradecimientos

Los autores quieren agradecer a la Universitat Politècnica de València el apoyo económico a través de la beca FPI-UPV (PAID-01-19). Además, agradecen el trabajo de los dos revisores anónimos y el editor, cuyas sugerencias mejoraron el manuscrito original.

7. Referencias

Ajadi, T., Burton, Z., Dwyer, M., Hammond, T., & Ross, C. (2020). *Deloitte Football Money League 2020 - Eye on the prize*. Recuperado de: https://www2.deloitte.com/bg/en/pages/finance/articles/football-money-league-2020.html

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175-185.

Barua, S., Islam, M. M., Yao, X., & Kazuyuki, M. (2014). MWMOTE—majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering, 26*(2), 405-425.

Boscá, J. E., Liern, V., Martínez, A., & Sala, R. (2009). Increasing offensive or defensive efficiency? An analysis of Italian and Spanish football. *Omega, 37*(1), 63-78.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterrey, CA: Routledge.

Brito de Souza, D., López-Del Campo, R., Blanco-Pita, H., Resta, R., & Del Coso, J. (2019). An extensive comparative analysis of successful and unsuccessful football teams in LaLiga. *Frontiers in Psychology, 10*, 2566.

Budsaba, K., Smith, C. E., & Riviere, J. E. (2000). Compass plots: a combination of star plot and analysis of means to visualize significant interactions in complex toxicology studies. *Toxicol Methods, 10*(4), 313-332.

Castellano, J., Casamichana, D., & Lago, C. (2012). The use of match statistics that discriminate between successful and unsuccessful soccer teams. *Journal of Human Kinetics, 31*, 139.

Cordón, I., García, S., Fernández, A., & Herrera, F. (2018). Imbalance: Oversampling algorithms for imbalanced classification in R. *Knowledge-Based Systems, 161*, 329-341.

Driblab. (2020). *Player Analysis*. Recuperado de: https://www.driblab.com/servicios-driblab/player-analysis/

Espitia-Escuer, M., & Garcia-Cebrian, L. I. (2008). Measuring the efficiency of spanish first-division soccer teams. *European Sport Management Quarterly, 8*(3), 229-246.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics, 2*(3), 211-228.

Hormozi, H., Hormozi, E., & Nohooji, H. R. (2012). The classification of the applicable machine learning methods in robot manipulators. *International Journal of Machine Learning and Computing, 2*(5), 560-563.

Japkowicz, N. (2000, julio). Learning from imbalanced data sets: a comparison of various strategies. *En AAAI workshop on learning from imbalanced data sets*. (Vol. *68*, págs. 10-15). Menlo Park: AAAI Press.

Knutson, T. (2020). *StatsBomb presenta sus nuevas visualizaciones*. Recuperado de: https://statsbomb.com/es/2020/01/statsbomb-presenta-sus-nuevas-visualizaciones/

Malagón-Selma, P., Debón, A., & Ferrer, A. (2022). Modelos de Machine Learning y estadística multivariante para predecir la posición de los equipos de primera división. *Journal of Sports Economics & Management*, 12(1), 3-22.

Kolence, K. W., & Kiviat, P. J. (1973). Software unit profiles & Kiviat figures. *ACM SIGMETRICS Performance Evaluation Review, 2*(3), 2-12.

Kuhn, M. (2020). *Caret: classification and regression training. R package version 6.0-86.*

Lago, C. (2009). The influence of match location, quality of opposition, and match status on possession strategies in professional association football. *Journal of Sports Sciences, 27*(13), 1463-1469.

Lago-Ballesteros, J., & Lago-Peñas, C. (2010). Performance in team sports: identifying the keys to success in soccer. *Journal of Human Kinetics, 25*(1), 85-91.

Lago-Peñas, C., Lago-Ballesteros, J., Dellal, A., & Gómez, M. (2010). Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. *Journal of Sports Science & Medicine, 9*(2), 288-293.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: a package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1-18.

Liu, H., Gómez, M.-A., Gonçalves, B., & Sampaio, J. (2016). Technical performance and match-to-match variation in elite football teams. *Journal of Sports Sciences, 34*(6), 509-518.

Liu, H., Yi, Q., Giménez, J.-V., Gómez, M.-A., & Lago-Peñas, C. (2015). Performance profiles of football teams in the UEFA Champions League considering situational efficiency. *International Journal of Performance Analysis in Sport, 15*(1), 371-390.

Lucey, P., Oliver, D., Carr, P., Roth, J., & Matthews, I. (2013, agosto). Assessing team strategy using spatiotemporal data. *En Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (págs. 1366-1374). Chicago: Association for Computing Machinery.

Mark, C., & Sormaz, M. (2019). *Clustering playing styles in the modern day full-back*. Recuperado de: https://www.statsperform.com/resource/clustering-playing-styles-in-the-modern-day-full-back/

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM), 8*(3), 404-417.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure, 405*(2), 442-451.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

Oberstone, J. (2009). Differentiating the top english premier league football clubs from the rest of the pack: identifying the keys to success. *Journal of Quantitative Analysis in Sports, 5*(3).

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169-198.

Pérez, D. (2021). *Radares en fútbol: para qué sirven y por qué están de moda*. Recuperado de: https://objetivoanalista.com/radares-futbol/

Saary, M. J. (2008). Radar plots: a useful way for presenting multivariate health care data. *Journal of Clinical Epidemiology, 61*(4), 311-317.

Malagón-Selma, P., Debón, A., & Ferrer, A. (2022). Modelos de Machine Learning y estadística multivariante para predecir la posición de los equipos de primera división. *Journal of Sports Economics & Management*, 12(1), 3-22.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Computational Intelligence Magazine, 13*(4), 59-76.

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Florida: Chapman and Hall/CRC.

Stone, M. (1974). Cross validatory choice and assessement of statistical predictions. *Journal of Royal Statistical Society B (Methodological)*, 36(2), 111-147.

Szymanska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics, 8*(1), 3-16.

Taylor, J. B., Mellalieu, S. D., James, N., & Shearer, D. A. (2008). The influence of match location, quality of opposition, and match status on technical performance in professional association football. *Journal of Sports Sciences, 26*(9), 885-895.

Team, R. C. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Westerhuis, J. A., Hoefsloot, H. C., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J., . . . van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*(1), 81-89.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1-3), 37-52.

Wold, S., Johansson, E., & Cocchi, M. (1993). PLS: partial least squares projections to latent structures. En H. Hubinyi (Ed.), *3D QSAR in Drug Design: Theory, Methods and Applications*. (págs. 523-550). Leiden, Paises Bajos: ESCOM Science Publishers.

Zambom-Ferraresi, F., García-Cebrián, L. I., Lera-López, F., & Iráizoz, B. (2017). Performance evaluation in the UEFA Champions League. *Journal of Sports Economics, 18*(5), 448-470.