Aplicación de métodos de clasificación en la predicción de la jornada futbolística de la Liga Inglesa en la temporada 2022/2023

Ariana Chacón Navarro[1], Marco Espinoza Marín[1], Alexis Monjarrez Calderon
arianachacon.navarro@ucr.ac.cr, marco.espinozamarin@ucr.ac.cr,
alexis.monjarrez@ucr.ac.cr

RESUMEN

En la actualidad, la Premier League inglesa es una de las ligas más competitivas e importantes del mundo, lo que hace cada vez más relevante la predicción de los resultados de sus partidos. Es por eso por lo que el objetivo de este estudio es predecir el resultado de una jornada posterior de la liga inglesa para la temporada 2022/23 mediante métodos de clasificación. Para emplear este análisis se utilizó un conjunto de datos que comprende 1900 partidos de la Premier League de las últimas cinco temporadas. Para la clasificación se emplearon las técnicas de regresión logística multinomial, árboles de decisión, k-vecinos más cercanos (KNN) y también modelos de ensamble, específicamente bagging y bosques aleatorios. Para cada una de estas técnicas de clasificación se emplearon métodos de calibración con el propósito de lograr la máxima precisión posible. El método KNN, presentó el mejor nivel de clasificación. Este método mostró el menor porcentaje de falsos positivos para las categorías "ganó", "perdió" y "empató" y el que mejor predijo, en menor cantidad de jornadas. Se concluyó que los modelos propuestos no lograron predecir correctamente los resultados de empate, lo que indica un rendimiento deficiente. Por lo tanto, es importante para futuros estudios la inclusión de variables que permitan mejorar la precisión de la predicción de resultados de fútbol.

Palabras clave: fútbol, métodos de clasificación, árboles de decisión, validación cruzada.

INTRODUCCIÓN

La Liga Premier Inglesa, es considerada como uno de los campeonatos de fútbol más apasionantes y seguidos a nivel mundial, el cual atrajo a millones de espectadores en cada uno de sus partidos. Según datos de Jakegellerman (2023), sólo en Estados Unidos, la audiencia total por partido ascendió a 527,000 espectadores durante la temporada 2022-2023, lo que implicó más de 200 millones de espectadores durante los 380 partidos de la temporada, sin contar los partidos de copa nacional o campeonato europeo. Este inmenso interés generó el auge de industrias centradas en la predicción de resultados de partidos, tales como las casas de apuestas y los medios de comunicación deportivos. En este escenario, los métodos de clasificación se posicionaron como una herramienta relevante para predecir los desenlaces de los encuentros futbolísticos y explorar cómo estas técnicas se aplicaron para pronosticar los resultados de la jornada futbolística durante la temporada 2022/2023, constituyendo el principal eje de este artículo.

[1] Estudiantes de la Escuela de Estadística de la Universidad de Costa Rica

Por esta razón, y la disponibilidad de datos hoy en día sobre el estado del equipo, dinero que cuenta para contratar jugadores y datos sobre la plantilla, ha popularizado el uso de modelos de clasificación. Además, actualmente los equipos de alto calibre invierten en infraestructura como complejos deportivos, médicos de alto calibre y planes estructurados y estrictos de nutrición que permiten el más alto rendimiento de los jugadores, todo esto con el único fin de tener resultados positivos en los partidos de fútbol durante la temporada regular. Es por esto que, los modelos de clasificación también permiten identificar fortalezas y debilidades en un equipo para tomar decisiones antes o durante un partido y que pueden ser decisivos en el resultado final de un partido (Bunker & Thabtah, 2019).

También los directores de los equipos y los expertos en análisis deportivo emiten pronósticos sobre el resultado de los partidos mediante técnicas de clasificación para intentar predecir el desenlace de los encuentros de fútbol. Estas predicciones se realizan con base en datos históricos y dependen de diferentes factores como las estadísticas del equipo, las estadísticas individuales de los jugadores, entre otros. Estos datos son empleados por los directivos y gerentes de los clubes para determinar qué se requiere para ganar el partido y quién tiene mayores posibilidades de hacerlo. En años recientes, la predicción de los resultados de los partidos de fútbol se ha vuelto muy popular (Zaveri, Shah, Tiwari, Shinde & Kumar Teli, 2018).

Un estudio que pretendía predecir cuál equipo iba a ganar el mundial 2018, utilizó varios métodos de clasificación para predecir el ganador, un ejemplo de estos es el de bosques aleatorios, para el cual, según los investigadores, tuvo un mejor rendimiento de clasificación en comparación con los modelos de regresión. Para efectos de la investigación, los autores utilizaron variables tanto de las selecciones clasificadas al mundial, como variables de los jugadores convocados a la cita mundialista (Groll, Ley, Schauberger, & Van Eetvelde, 2018).

Por otro lado, también se han realizado modelos de clasificación en sistemas competitivos parecidos al fútbol, es así el caso del fútbol australiano ya que según (Robertson, Back & Bartlett, 2015) los métodos de clasificación permiten encontrar áreas deportivas que deberían ser priorizadas, es por eso que la exploración de estos estudios han logrado determinar qué factores como tiros a marco de la oposición, tiempo en posesión del balón y los tiros totales a marco, sean importantes en que un equipo gane o pierda un partido.

Por otra parte, la aplicación de diversas técnicas de análisis multinivel es fundamental en este tipo de estudios. Un ejemplo de esto es el estudio realizado por Lopes (2019), en este estudio, se aplica una variedad de técnicas para predecir si ambos equipos marcarán goles en un partido de fútbol. Utilizando datos del campeonato portugués (LIGA NOS) desde la temporada 1994/1995 hasta la 2018/2019, Lopes (2019) presenta varios algoritmos de clasificación, como k-Vecinos más cercanos (K-NN), máquinas vectoriales de soporte (SVM), árboles de decisión y regresión lineal, y evalúa cuál se ajusta mejor en este caso de estudio. De estos, el (K-NN) demostró tener la mejor precisión de predicción de los resultados, con alrededor del 54%. Este enfoque de predicción basado en análisis multinivel no solo ha demostrado ser efectivo, sino que también proporciona una valiosa comprensión de los factores que pueden influir en los resultados de los partidos de fútbol.

Sin embargo, muchos estudios tienen limitaciones en sus pronósticos, como el uso de la regresión logística, que sólo proporciona dos resultados posibles: “victoria en casa” o “victoria fuera”, mientras que un partido de fútbol puede tener tres posibles resultados: “victoria en casa”, “victoria fuera” o “empate”, en los cuales se usa la regresión multinomial. Algunas investigaciones han empleado el algoritmo de bosques aleatorios pero su precisión de predicción es sólo del 63,4%. (Yoel & Sani , 2019)

Existen estudios como el de Prasetio et al., (2016), quienes utilizaron la regresión logística y características como la ofensiva y defensa de los equipos locales y visitantes para hacer sus predicciones, logrando una precisión del 69,5%. Sin embargo, a pesar de que la regresión logística obtuvo una alta precisión del 93%, se vio limitada por la obtención de sólo dos valores en los resultados. Otros investigadores, como Yezus (2014), han utilizado técnicas como KNN y bosques aleatorios, utilizando elementos como la consistencia en goles, concentración, motivación, diferencia de goles, diferencia de puntuación e historia para desarrollar clasificadores, obteniendo precisión de 55,8% y 63,4% respectivamente.

Es por eso que el objetivo general de esta investigación es utilizar modelos de clasificación para predecir los resultados de las jornadas de la liga inglesa para la temporada 2022/23. Mientras que los objetivos específicos son calibrar los modelos de clasificación mediante la validación cruzada y ajustar un modelo de clasificación para predecir el resultado final de la jornada posterior.

METODOLOGÍA

El presente estudio se llevó a cabo a partir de una base de datos que abarca las cinco temporadas más recientes de la Premier League, específicamente desde la temporada 2018/2019 hasta la 2022/2023 con 1900 observaciones. Los datos se extrajeron de “Footystats”, una plataforma reconocida por su amplio repertorio de estadísticas de ligas de fútbol a nivel global. Para cada encuentro entre dos equipos, se tenían detalles exhaustivos del partido disputado, así como variables relevantes de equipos y jugadores, y los resultados de los partidos. Cabe destacar que la base poseía únicamente variables numéricas de las cuales se seleccionaron 27 variables, las cuales correspondieron a edad media de los jugadores en casa y visitantes (Average age home team & away), la calificación media de los jugadores de los equipos en casa y visitante (Average rating overall home & away team), rendimiento promedio de un equipo local y visitante en términos de puntos ganados por partido(Pre_PPG home & Away), puntos en promedio el equipo local y visitante ha obtenido en sus partidos jugados en su propio campo (home_ppg & away_ppg), número total de tiros de esquina en el equipo local y visitante (Home &Away corner count), número total de tarjetas amarillas que los jugadores del equipo local y visitante han recibido durante el partido(home & away team yellow cards), tiros directo a marco en el equipo local y visitante (home & away shots on target), tiros que no fueron dirigidos a la portería del equipo contrario durante el partido para el equipo local y visitante (Home & Away team shots off target), número total de faltas cometidas por el equipo visitante y el equipo local (home & away team fouls), porcentaje de tiempo durante el cual el equipo local y visitante tuvo control del balón en el partido (home & away team possesion), cantidad estimada de goles que el equipo local y visitante podrían haber marcado, considerando la calidad de las oportunidades de gol que generalmente crean antes del partido(Home & Away Pre-Match) y cantidad estimada

de goles que el equipo visitante y equipo local podrían haber marcado, considerando la calidad de las oportunidades de gol que han creado(team a & team b xg). Es también relevante mencionar que la variable respuesta correspondió a si el equipo local ganó, perdió o empató un partido.

Es importante resaltar que, dada la naturaleza de las variables, se seleccionaron los métodos de clasificación que mejor se ajustaban a estas características. El indicador de desempeño que se utilizó fue el porcentaje de falsos ganes para cada clase de la variable respuesta. En este contexto, los falsos ganes corresponden al porcentaje de casos negativos incorrectamente clasificados como positivos (Corso, 2009).

Inicialmente, para llevar a cabo la investigación, se utilizó la última temporada disponible, la que corresponde a la temporada 2022/2023 que inició en agosto 2022 y finalizó en mayo del 2023. La temporada abarca 38 jornadas, de las cuales en cada una se juegan 10 partidos diferentes. Se realizó la validación cruzada con el objetivo de calibrar los modelos que se aplicaron a lo largo de la investigación, para esto, se fijó la base de entrenamiento en 20 jornadas y la jornada 21, como validación, se utilizó de esta manera debido a que el objetivo de la investigación era predecir la siguiente jornada tomando las jornadas previas, por lo tanto, era necesario mantener el orden de las jornadas. Una vez con la base de entrenamiento, se calibraron los parámetros para cada modelo utilizado.

Seguidamente, se utilizó la regresión logística multinomial, una técnica estadística que es empleada para analizar datos donde la variable dependiente es categórica y tiene más de una categoría. Esta técnica de modelado proporciona la probabilidad de que una observación se clasifique en una de las categorías (Harrell, 2015). Para la validación cruzada, se trabajó con un modelo inicial que incluía 26 variables predictoras y se procedió a eliminar una variable a la vez, esto en diez iteraciones. En cada paso, se evaluaron los indicadores de desempeño de falsos ganes, falsas pérdidas y falsos empates. Al final, se conservaron únicamente las variables que mostraron el mejor indicador de desempeño.

Posteriormente, se utilizó el método de árboles de decisión, que permitió representar y categorizar una serie de condiciones que ocurren de forma sucesiva, es decir que dichos árboles fueron categorizados como aprendizaje basado en similitudes (Orea,S et. al., 2005). Para la validación cruzada de este modelo, se fijó el mínimo de observaciones necesarios para una partición en 2 y también se fijó la cantidad mínima de observaciones en los nodos terminales el cual es 1 (Alvarado, 2023). Luego de esto, se calibró la profundidad del árbol y el parámetro de complejidad.

Por otra parte, se utilizó el ensamble de modelos, la cual es una técnica de minería de datos que tiene como objetivo aumentar la eficiencia en las predicciones, tomando como base modelos ajustados previamente. Se seleccionaron dos métodos: bagging y bosques aleatorios. Para la calibración del método de bagging se fijaron las reglas duras, las cuales fueron: 100 como el mínimo de observaciones para una partición, 50 observaciones mínimas para ser considerado un nodo terminal, el parámetro de complejidad en 0.003 y la máxima profundidad en 12 divisiones (Alvarado, 2023). Consiguientemente, se calibró el número de árboles para determinar la cantidad adecuada a usar. Por otro lado, el método de bosques aleatorios se utilizó

la agregación de bootstrap para mezclar diferentes árboles, y se fijaron el tamaño del nodo terminal en 50 y máximo de números de nodos, donde cada árbol fue construido con observaciones y variables aleatorios en cada nodo.

Por último, se utilizó KNN. Este método de clasificación busca encontrar a los vecinos más cercanos del vector de características que posea la observación (Lall & Sharma., 1996). Primeramente, se procedió a estandarizar las variables, una acción necesaria dada la distintas escalas en la que cada variable se encontraba. Al poner todas las variables en la misma escala, se evitó que las variables con rangos más amplios dominaran el proceso de clasificación. (Rojas-Jimenez, n.d.). Para calibrar el algoritmo de vecinos más cercanos, se experimentó variando el número de vecinos considerados, probando con 3, 5, 7 y 9 vecinos. El rendimiento de cada configuración se evaluó utilizando los indicadores de desempeño: falsos positivos de las clases ganó, empató y perdió. Seguidamente, se realizó la predicción de los resultados de las jornadas posteriores a partir de la jornada anterior.

Para el análisis y procesamiento de los datos de las cinco técnicas se utilizó el software estadístico R (R Core Team, 2023) en su versión 4.3.1 y las siguientes librerías: *readxl* (Wickham y Bryan, 2022), *class* (Venables et al., 2002), *adabag* (Alfaro et al., 2013), *randomforest* (Liaw y Wiener, 2002), *modeest* (Poncet, 2019), *ISLR* (James et al., 2021), *rattle* (Williams, 2011), *dplyr* (Wickham et al., 2020) y nnet (Ripley, 2023).

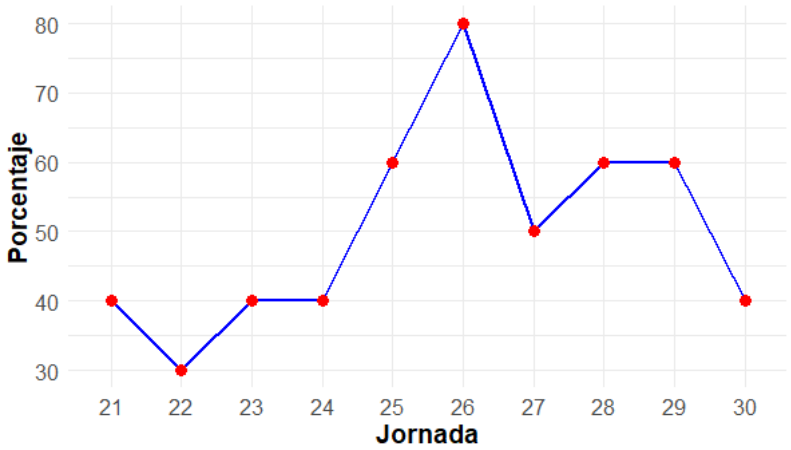
RESULTADOS

De manera inicial, se realizó la validación cruzada del modelo logístico multinomial, en el que se encontró que el modelo con 15 y 7 variables generaba indicadores idénticos en comparación con el modelo con sólo 3 variables. Específicamente, la categoría "ganó" no registró falsos ganes, mientras que la categoría "perdió" presentó un 33% de ellos. Sorprendentemente, la categoría "empató" registró un 100% de falsos empates. Esto sugirió dado que la mayor complejidad del modelo de 15 variables y 7 variables no aportaba mejoras significativas en las predicciones como se observa en la Figura 1 (Ver Figura 1 en Anexos).

De manera similar, el modelo con 7 variables mostró un rendimiento similar al del modelo más simple, que usaba 3 variables. Dado este escenario, se optó por elegir el modelo con 3 variables predictoras, ya que esto favoreció la clasificación de los resultados de las jornadas, y a su vez, redujo la posibilidad de sobreajuste. El modelo más simple se basa en tres variables: 1) cantidad estimada de goles que el equipo local podría haber marcado, considerando la calidad de las oportunidades de gol que generalmente crean antes del partido; 2) la cantidad estimada de goles que el equipo visitante podría haber marcado, tomando en cuenta la calidad de las oportunidades de gol que han creado y 3) la cantidad estimada de goles que el equipo visitante podría haber marcado, tomando la calidad de las oportunidades de gol que generalmente crean antes del partido.

Al realizar la predicción de las jornadas 21 a 30, se observó que en general la precisión de clasificación fue deficiente. Sin embargo, en la jornada 26, el modelo clasificó correctamente el 80% de los datos. (Ver Gráfico 1).

Gráfico 1. *Porcentaje de Precisión del Modelo Multinomial: Jornadas 21-30 de la Premier League utilizando un modelo con 3 variables*



Por otro lado, para el método de árboles de decisión, se procedió a calibrar el modelo mediante validación cruzada. Para el parámetro de complejidad (cp) se estableció un punto inicial de 0.0005 y por cada iteración de 1 a 10, se le sumó 0.01 a este parámetro, para el cual se encontró que los indicadores de desempeño se estabilizaron al determinar un parámetro de complejidad de 0.0025 dado que la cantidad de falsos positivos se redujo en comparación con otros valores. Consiguientemente, se calibró la profundidad máxima del árbol (maxdepth), para el cual se determinó una profundidad de 5 divisiones, de la cual obtuvo diferencia entre las diferentes profundidades que en un inicio iban en un rango de a 1 a 7, ya que, a partir de una profundidad de 5 divisiones, se logró que los indicadores tuvieran un mejor desempeño en el rendimiento de la clasificación. Las variables donde el árbol ganó mayor información a la hora de dividir los nodos para construir dichos árboles con los parámetros designados son calificación promedio de los jugadores del equipo visitante, número total de disparos realizados por el equipo local, cantidad estimada de goles que el equipo visitante podría haber marcado, punto en promedio que el equipo local ha acumulado, punto en promedio que el equipo visitante ha acumulado creando así un árbol de profundidad 5 (Ver Figura 2 en Anexos).

Teniendo en cuenta los aspectos mencionados anteriormente, se implementó el método de árboles de decisión cuidadosamente ajustados. Las jornadas de la 1 a la 20 fueron seleccionadas como conjunto de entrenamiento para predecir los resultados de la jornada siguiente, en este caso la jornada 21 y así sucesivamente. Como se puede ver en la Tabla 2, se presentaron los resultados obtenidos, en la jornada 26 se observó una reducción en los falsos positivos, aunque no se logró una eliminación completa de estos. Sin embargo, al analizar la predicción para la jornada 38 se encontró que sus indicadores mejoraron, pero, no necesariamente más jornadas en la base de entrenamiento hicieron que el modelo clasificó mejor.

Tabla 2. *Indicadores de desempeño para árboles de decisión*

Jornada de validación	FP-Ganó	FP-Perdió	FP-Empató
26	22.2%	0%	100%
38	0%	0%	100%

Como parte de las otras técnicas, la primera en ser utilizada correspondió a bagging y en la cual se incluyeron los árboles de decisión con los parámetros fijados anteriormente. Al realizar la validación cruzada para este algoritmo, se encontró que la cantidad de árboles adecuados es de 4 árboles finales donde los indicadores de desempeño mejoraban, y el cual se utilizó para posteriormente ajustar los 4 árboles de decisión.

Por último, se calibraron los bosques aleatorios, para el cual se determinaron los parámetros de número de árboles en 3, y el número de variables aleatorias para dividir en cuatro variables, posteriormente como se observó en el cuadro 4 (Ver anexos) que los árboles aleatorios en todos los casos predijeron que empataron cuando en realidad no, mientras que 50% de ellos fueron clasificados como que perdieron cuando en realidad no perdieron, lo cual indica que el desempeño de este es bajo.

Por otro lado, para la validación cruzada del método bagging se determinó un parámetro de calibración de 6 árboles, el cual en comparación con el método de árboles aleatorios tuvo un peor rendimiento, debido a que en las tres categorías clasificó mal con un porcentaje de falsos negativos muy alto.

Por otra parte, para el método de KNN, se calibró la cantidad de vecinos donde los indicadores de desempeño para las 3 categorías en la variable respuesta era el más bajo. Como se observó en la figura 3 (Ver Anexos) los Falsos pérdidas para la clase de perdió, se mantuvo en 100% en todos los casos lo que significó que en todos los vecinos el modelo clasificó qué perdió el siguiente partido pero en realidad no perdieron, mientras que para los falsos ganes para la clase de ganó con el mínimo de vecinos se mantuvo alto, no obstante cuando se utilizaron 9 vecinos más cercanos, el desempeño de este mejoró, por lo tanto el modelo de KNN se calibró en 9 vecinos más cercanos. Consiguientemente, se tomó a partir de la jornada 20 para predecir el resultado de la jornada 21 y jornadas posteriores. Se encontró que no necesariamente más jornadas en la base de entrenamiento el modelo va a clasificar mejor, ya que cuando se utilizó hasta la jornada 22 en la base de entrenamiento para ajustar el modelo y al validarlo con la jornada 23, los indicadores de desempeño mostraron que clasificó mejor que cuando se predijo la última jornada de la temporada.

Por último, en la Tabla 3 presenta lo obtenido en los indicadores de desempeño a través de la validación cruzada, se muestra que indistintamente de cuál método se utilice los falsos positivos en la categoría empate, siempre va a estar mal clasificado. Cabe resaltar que el modelo de k-vecinos fue el que tuvo un mejor rendimiento en la predicción con menos jornadas, dado que sus falsas pérdidas son nulos.

Tabla 3. *Indicadores de desempeño para los modelos ajustados*

Técnicas	Jornada	Falsos ganes	Falsos pérdidas	Falsos empates
Multinomial	26	11%	0%	100%
Árboles de decisión	26	0%	0%	100%
K-vecinos más cercanos	22	0%	0%	100%

CONCLUSIONES

Primeramente, según Groll et. al (2018) en su investigación, los árboles de decisión tuvieron un mejor rendimiento en la predicción del ganador del mundial 2018; sin embargo, este modelo usado para predecir el resultado de una jornada de la temporada regular de la liga inglesa no tuvo el mejor rendimiento en comparación con el resto y además, los indicadores de desempeño demostraron que el árbol de decisión ajustado en todos los casos clasifica los resultados como empates cuando en realidad no fueron empates. Por otro lado, según lo planteado por Lopes (2019) el método de KNN tuvo un mejor rendimiento de clasificación, es el caso de esta investigación, dado que el modelo de KNN fue el que mejor logró clasificar los resultados con una cantidad menor de jornadas.

De igual forma Hastie et,.al, (2009) establecen que un un modelo de regresión multinomial con una gran cantidad de variables puede proporcionar una alta precisión en los datos de entrenamiento, pero corre el riesgo de que el modelo presente sobreajuste. No obstante, los indicadores de desempeño muestran resultados deficientes al usar un modelo con una menor cantidad de variables.

A pesar de eso, lo que se encontró con respecto a los indicadores de desempeño, es que los modelos planteados no predijeron bien los resultados de la clase de empate. Por lo tanto, el rendimiento general de los modelos es bastante malo. Una de estas limitantes puede ser la cantidad de datos disponibles por temporada o las variables que se utilizaron para la clasificación.

Para futuros estudios, se recomienda llevar a cabo una investigación más exhaustiva que permita la inclusión de variables adicionales. Estas podrían abarcar el valor de la plantilla de cada equipo, su poder de ataque y defensa, así como también considerar datos relacionados con el uso del Asistente de Video para Árbitros (VAR). Esta última es una herramienta frecuentemente empleada durante los últimos cinco años en el ámbito futbolístico para analizar y resolver situaciones de juego consideradas polémicas. También sería importante considerar en investigaciones posteriores, el uso de las últimas jornadas para entrenar los modelos, con el fin de tener un mejor rendimiento de los indicadores de desempeño.

Por otra parte, se sugiere considerar el uso del porcentaje de victorias del equipo local frente al visitante como uno de los atributos. Un ejemplo exitoso de la implementación de esta

variable es el caso en la investigación empleada por Zaveri et al. (2018). Esta variable resultó ser muy valiosa en, incrementando la precisión del modelo casi en un 8%. Esta mejora puede atribuirse al hecho de que cada equipo tiene un estilo de juego único y a menudo tiene rivales específicos. También es posible explorar otros enfoques. Por ejemplo, sería interesante analizar la cantidad total de goles que se anticipa sean anotados por ambos equipos durante un partido. Este último aspecto es muy popular en las casas de apuestas. (Lopes, 2019).

BIBLIOGRAFÍA

Alfaro, E., Gamez, M. y Garcia, N.(2013). adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software, 54(2), 1-35. URL http://www.jstatsoft.org/v54/i02/

Alvarado, R (2023). Introducción al Análisis Multivariado: Capítulo V: Validación [Presentación completa]. Universidad de Costa Rica. Presentación del curso.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. Córdoba: Universidad Tecnológica Nacional, Facultad Regional Córdoba.

D. Prasetio, Harlili,H. (2016) "Predicting football match results with logistic regression," in 4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2016.

Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2018). Prediction of the FIFA World Cup 2018 – A random forest approach with an emphasis on estimated team ability parameters. *ResearchGate*.

Harrell , F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. Springer International Publishing. https://link.springer.com/book/10.1007/978-3-319-19425-7

Hastie, Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. https://hastie.su.domains/Papers/ESLII.pdf

Jakegellerman. (2023, June 1). *NBC SPORTS COMPLETES FIRST DECADE AS EXCLUSIVE U.S. HOME OF PREMIER LEAGUE, HIGHLIGHTED BY AUTHENTIC, INNOVATIVE COVERAGE & RECORD AUDIENCES - NBC Sports Pressbox*. NBC Sports Pressbox. https://nbcsportsgrouppressbox.com/2023/06/01/nbc-sports-completes-first-decade-as-exclusive-u-s-home-of-premier-league-highlighted-by-authentic-innovative-coverage-record-audiences/#:~:text=In%202022%2D23%2C%20NBC%20Sports,from%20last%20season%20(510%2C000).

Lall, U., y Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water resources research*, *32*(3), 679-693. https://www.researchgate.net/publication/222797622_A_Nearest_Neighbor_Bootstrap_For_Resampling_Hydrologic_Time_Series

Liaw, A. y Wiener, M. (2002). "Classification and Regression by randomForest." R News, 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/

Lopes, A. (2019). *Application of Machine Learning Algorithms for Automatic Classification of Problems Football*. Lopes, Adão. https://www.academia.edu/es/38937812/Application_of_Machine_Learning_Algorithms_for_Automatic_Classification_of_Problems_Football

A. Yezus, and A. Igoshkin, (2014). "Predicting outcome of soccer matches using machine learning," Saint-Petersburg. Univ.

Orea, S, Vargas, A., y Alonso, M. (2005). Minería de datos: El algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*, *779*(73), 33. https://d1wqtxts1xzle7.cloudfront.net/34203825/e1-with-cover-page-v2.pdf?Expires=1658626420&Signature=BvHWLhnhPVKbGkqMmeG3~GG2Qzae4sbJSetwd4dj36m-ddslDVTupmFZUIh1skWzFUQMq1qoI~Wzy-sHElBslnP9F7mm1AvMJFxv-uPvk8vdOrE5krK1GHKP3dZCcnBLUVDqFXSqJ0K0LU5HTvp5c~6YFFStLocg9-YtQxHFwWBX33SQ0u8QGlw~rhqQ1DGXczTC5XDJ3cVtH9lrHzIW~ALdFo1ShEJm3rcVsDRlJfMO34mYCOJ113tXknweksAABs1~PCOWF5Nut1LfH4-vmjWiwET5vFD-GTgZoHLtWUeawK8pnKNG1N03ucP0Pryi5uPbm-DbGciSfMZHp7VNbg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Poncet, P. (2019). _modeest: Mode Estimation_. R package version 2.4.0, https://CRAN.R-project.org/package=modeest

Robertson, S., Back, & Bartlett, J. W. (2015). Explaining match outcome in elite Australian Rules football using team performance indicators. *Journal of Sports Sciences*, *34*(7), 637–644. https://doi.org/10.1080/02640414.2015.1066026

Ripley, B. (2023). Feed-Forward Neural Networks and Multinomial Log-Linear Models [R package nnet version 7.3-19]. R-Project.org. https://cran.r-project.org/package=nnet

Rojas-Jimenez, K. (n.d.). *Capítulo 5 Transformación, Estandarización e Imputación de Datos | Ciencia de Datos para Ciencias Naturales*. https://bookdown.org/keilor_rojas/CienciaDatos/transformaci%C3%B3n-estandarizaci%C3%B3n-e-imputaci%C3%B3n-de-datos.html#:~:text=5.2%20Estandarizaci%C3%B3n%20de%20datos,respecto%20a%20una%20escala%20com%C3%BAn.

Venables, W. N. y Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wickham,H., Hester,J. y Francois,R. (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

Wickham, H., François, R., Henry, L. y Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.0. https://CRAN.R-project.org/package=dplyr

Williams, G. J. (2011). Data Mining with Rattle and R: The art of excavating data for knowledge discovery, series Use R! Springer. https://rd.springer.com/book/10.1007/978-1-4419-9890-3.

Yoel F. Alfredo, Sani M. Isa. (2019) "Football Match Prediction with Tree Based Model Classification", International Journal of Intelligent Systems and Applications(IJISA), Vol.11, No.7, pp.20-28, 2019. DOI: 10.5815/ijisa.2019.07.03

Zaveri, N., Shah, U., Tiwari, S., Shinde, P., & Kumar Teli, L. (2018). Prediction of Football Match Score and Decision Making Process. *International Journal on Recent and Innovation Trends in Computing and Communication*.

ANEXOS

Figura 1. *Indicadores de desempeño con modelo logístico multinomial ajustando la cantidad de variables*

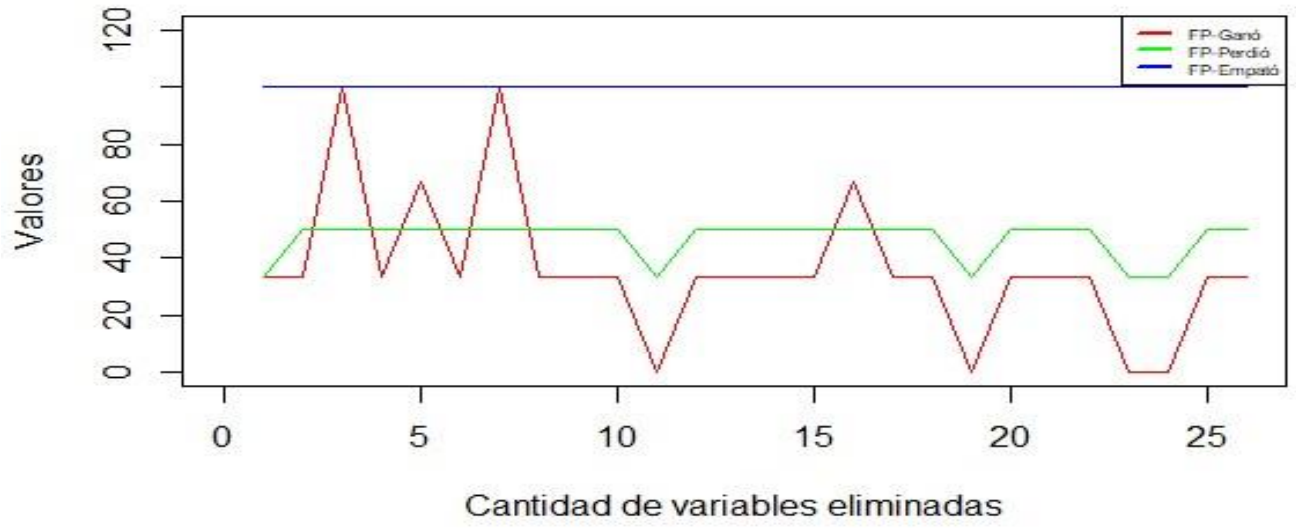


Figura 2

Ilustración de un árbol de decisión con los parámetros ajustados

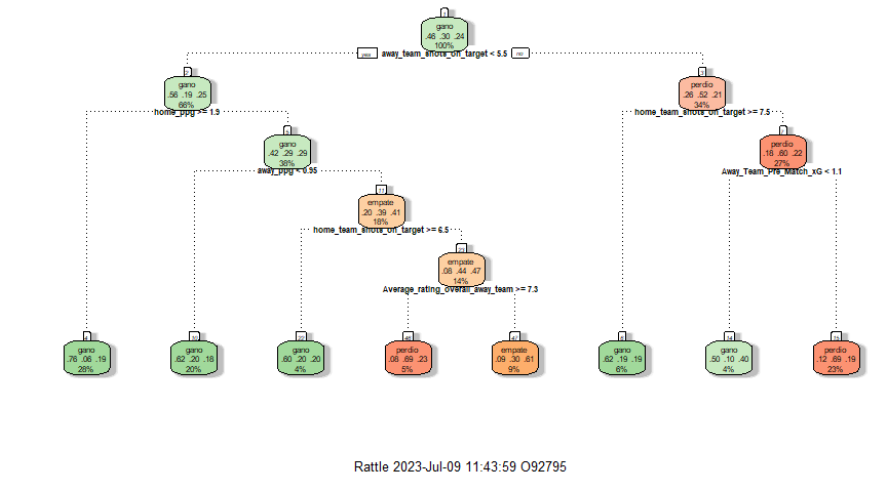
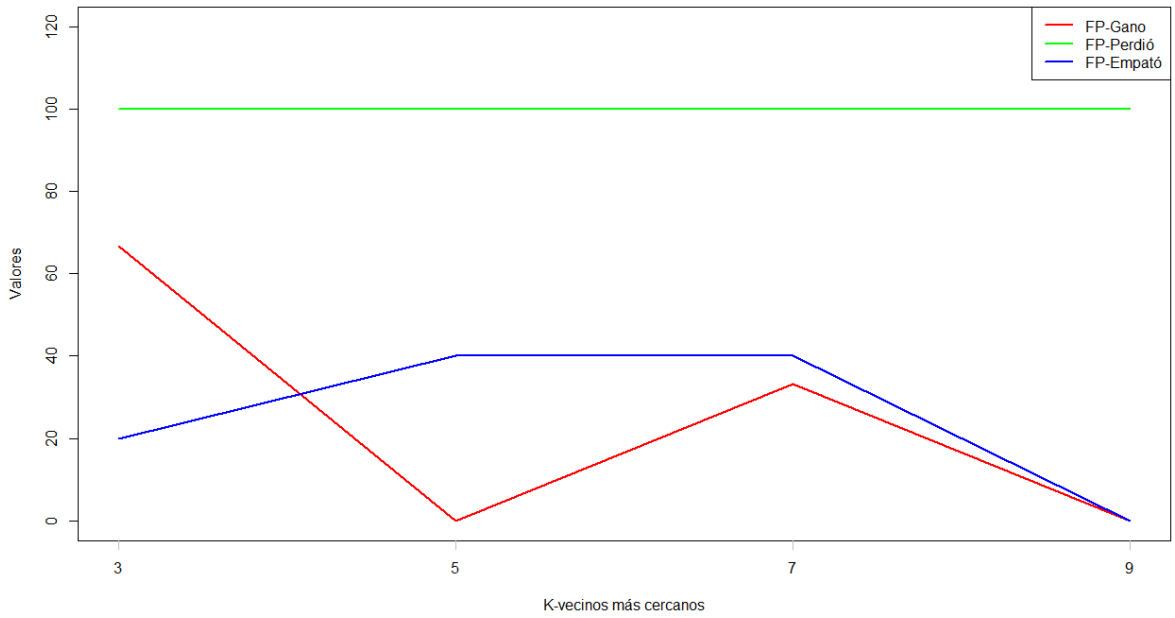


Figura 3.

Indicadores de desempeño para el método KNN



Cuadro 4.

Indicadores de desempeño para las diferentes técnicas de ensamble

	Falsos Positivos (ganó) %	Falsos Positivos (perdió) %	Falsos positivos (empató) %
Bosques aleatorios	*0*	*50*	*100*
Bagging para árboles de decisión	*33.33*	*66.66*	*100*