UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Aplicación de algoritmos de aprendizaje automática y ciencia de datos para la predicción de resultados de partidos de fútbol

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Martínez de la Rosa, Constantino

Tutor/a: Sánchez Anguix, Víctor

Cotutor/a: Alberola Oltra, Juan Miguel

CURSO ACADÉMICO: 2023/2024

Agradecimientos

A mis padres, mis hermanos y mis abuelos, por ser los pilares de mi vida. Este trabajo es tan vuestro como mío, porque sin vuestro amor, apoyo incondicional y confianza nada de esto sería posible. Gracias por estar siempre a mi lado.

A mi gente de mi pueblo, Montecorto, lugar de rincón de paz y fuente de inspiración inagotable. Este trabajo es un pequeño tributo a los momentos inolvidables y a las risas que me regaláis cada verano. Gracias por ser mi refugio y mi hogar.

Abstract

Sports events, especially those involving competition between teams, involve complex dynamics among the participants that ultimately affect the final outcome. Predicting the outcome of these sports events in advance is a complex task precisely because of these intricate dynamics that arise during the event, as well as the decisions made by all the participants. In many cases, even human experts can be wrong in their predictions, leading to surprising results against all expectations. This task presents challenges from the perspective of analytics and machine learning, as it is a problem that even human experts find difficult. In this project, we study the performance of different machine learning algorithms in predicting the outcome of professional football matches, with the goal of forecasting the final result (i.e., home win, draw, away win) based solely on information available before the match. By using statistics related to match information from the five major European leagues, two different approaches are developed to tackle the task using machine learning algorithms (neural networks and ensemble methods based on decision trees). As a result, an accuracy rate of 55% is achieved, and an analysis of some important features for predicting match outcomes is carried out.

Keywords: Machine learning, neural network, decision tree, classification problem, sports analytics, prediction.

Resumen

Los eventos deportivos, especialmente aquellos que involucran la competición entre equipos, involucran dinámicas complejas entre los participantes que finalmente afectan en el resultado final. Predecir de antemano el resultado de estos eventos deportivos resulta una tarea compleja precisamente por estas complejas dinámicas que aparecen en el evento, así por las decisiones que son tomadas por todos los participantes. En muchos casos, incluso expertos humanos erran en sus predicciones y se producen resultados sorprendentes ante todo pronóstico. La tarea presenta desafíos desde el punto de vista de la analítica y el aprendizaje automático, pues es un problema que incluso presenta dificultades a los expertos humanos. En este Trabajo de Fin de Grado estudiamos el rendimiento de diferentes algoritmos de aprendizaje automático en la tarea de predecir el resultado de partidos de fútbol profesional con el fin de intentar predecir el resultado final (i.e., victoria local, empate, victoria visitante) en base a información exclusivamente disponible antes del partido. Mediante el uso de estadísticas relacionadas con la información de los partidos de las 5 grandes ligas europeas, se desarrollan dos abordajes diferentes para tratar de resolver la tarea utilizando algoritmos de aprendizaje automático (redes neuronales y métodos de conjunto basados en árboles de decisión). Con todo esto, se logra obtener unos resultados del 55% de exactitud y se lleva a cabo el análisis de algunas características importantes para la predicción de los resultados de los partidos.

Palabras clave: Aprendizaje automático, red neuronal, árbol de decisión, problema de clasificación, analítica deportiva, predicción.

Índice

1. Introducción 9
   1.1 Motivación del trabajo 10
   1.2 Objetivos 10
2. Marco teórico 13
   2.1. Analítica deportiva 13
   2.2. Analítica deportiva en el fútbol 14
   2.3. Modelos de predicción de resultados 15
3. Marco legal y ético 19
4. Bases teóricas 21
   4.1 Métodos de conjunto 21
   4.2 Árboles de decisión 22
      4.2.1 Criterios de división univariante de un nodo 23
      4.2.2 LightGBM 24
   4.3 Redes neuronales Long-Short Term Memory (LSTM) 25
5. Desarrollo técnico del proyecto 27
   5.1 Fuentes de datos 27
   5.2 Análisis del dato y agrupación por equipos 29
   5.3 Cálculo de las medias ponderadas 30
   5.4 Análisis exploratorio de datos 31
   5.5 Reducción de la dimensionalidad 36
      5.5.1 Prueba de Kruskal-Wallis 38
      5.5.2 Estudio de las correlaciones y clustering de variables 41
      5.5.3 Selección de características 43
   5.6 Modelos empleados y preparación de los datos 44
      5.6.1 Abordaje 1: LightGBM 44
      5.6.2 Abordaje 2: Redes neuronales LSTM 45
6. Experimentos 47
   6.1 Puntos de referencia para modelos de predicción 47
   6.2 Experimentos con redes neuronales LSTM 48
      6.2.1 Red LSTM básica 49
      6.2.2 Búsqueda de hiperparámetros: Optimizador 51
      6.2.3 Búsqueda de hiper parámetros: Tasa de aprendizaje 53
      6.2.4 Búsqueda de hiperparámetros: Umbral y funciones de activación 56
      6.2.5 Arquitectura de la red neuronal 59

6.2.6 Resultados de la red neuronal LSTM óptima 61
6.3 Experimentos con LightGBM 64
6.3.1 Búsqueda en rejilla (Grid Search) 64
6.3.2 Resultados de LightGBM 65
7. Conclusiones 71
7.1 Limitaciones del trabajo 72
7.2 Trabajo futuro 72
Referencias 75
Anexo I: Objetivos de desarrollo sostenible 77
Anexo II 79

Figuras

Figura 1. Bagging vs boosting 22
Figura 2. Ejemplo simple de árbol de decisión basado en variables booleanas. 23
Figura 3. Evolución del Elo de diferentes clubes a lo largo de las temporadas. 28
Figura 4. Número de valores faltantes para variables que tienen más de 1000 valores faltantes. 29
Figura 5. Número de valores faltantes para variables que tienen menos de 1000 valores faltantes 29
Figura 6. Distribución de media de goles por resultado. 32
Figura 7. Distribución del número de goles marcados en un partido. 32
Figura 8. Gráfico de distribución de las medias ponderadas de elo y remates a portería. ..... **¡Error! Marcador no definido.**
Figura 9. Gráfico de caja y bigotes para goles contra media de tiros en partidos anteriores. ..... **¡Error! Marcador no definido.**
Figura 10. Gráfico de dispersión para la media ponderada de goles contra la cuota de victoria asignada por la casa Bet365. 34
Figura 11. Gráfico de barras para la categoría del elo frente a el número de victorias, empates y derrotas. 34
Figura 12. Gráficos de dispersión para la diferencia de medias ponderadas contra las cuotas de victoria. 35
Figura 13. Histogramas para algunas variables con posibilidad de ser estadísticamente significativas. 37
Figura 14. QQ-plots para las cuatro variables. 37
Figura 15. p-valores obtenidos en la prueba de Kruskal-Wallis. 39
Figura 16. Mapa de calor para las correlaciones de las variables del conjunto de datos. 41
Figura 17. Correlaciones de las variables del clúster 2. 42
Figura 18. Mapas de calor para los clústeres 6,7,8 y 9. 43
Figura 19. División de los datos para entrenamiento y prueba en el segundo abordaje. 46
Figura 20. Matriz de confusión para el clasificador aleatorio. 48
Figura 21. Evolución de la función de pérdida para la red neuronal simple. 50
Figura 22. Matriz de confusión para la red neuronal simple. 51
Figura 23. Evolución de la función de pérdida para entrenamiento y validación. 52
Figura 24. Resultados obtenidos para Adam, SGD y RMSProp. 52
Figura 25. Predicciones de Adam y RMSProp vs Valores reales. 53
Figura 26. Evolución de la función de pérdida para diferentes tasas de aprendizaje y optimizadores 54
Figura 27. Densidad del número de goles predichos por los modelos frente a la densidad real del número de goles. 55
Figura 28. Predicción del número de goles para Adam y RMSProp frente al número real de goles 56
Figura 29. Evolución de la función de pérdida en validación para los diferentes modelos. 57

1. Introducción

En los últimos años, la evolución de la ciencia de datos ha supuesto una revolución en la toma de decisiones en el mundo de la empresa. Dentro del campo del deporte, las diferentes técnicas desarrolladas para el análisis deportivo han tenido como consecuencia una inversión masiva para la obtención de grandes cantidades de datos (UNIR, 2024). Actualmente, podemos obtener bases de datos de gran volumen generadas por clubes, por deportistas e incluso por los aficionados. Estos datos se encuentran dentro de un abanico de información enorme: desde información sobre características fisiológicas sobre deportistas para analizar el rendimiento de un jugador, hasta el comportamiento de los seguidores a una determinada campaña publicitaria de un club (Navarro, 2024). Todo es analizable y todos los procesos dentro de una entidad deportiva pueden optimizarse.

En concreto, dentro del mundo del fútbol, la evolución del Big Data y del aprendizaje automático ha supuesto una importante mejora en la forma de competir de los clubes. Diferentes tipos de modelos de datos han sido creados para tratar de optimizar la toma de decisiones en los clubes (i.e. campañas de abonados, rentabilidad de fichajes, seguimiento de jugadores de categorías inferiores, estrategias dentro del juego, etc) (Herberger & Litke, 2021). De la misma manera, aficionados por este deporte han tratado de desarrollar sus propios modelos para entender y compartir información relevante sobre las diferentes competiciones futbolísticas (i.e. predicción de resultados, estimación del valor de jugadores, agrupación de jugadores, etc).

En este trabajo se estudia la aplicación de la ciencia de datos y los algoritmos de aprendizaje automático en la predicción de resultados de partidos de fútbol. En él, se han desarrollado modelos de clasificación multiclase (el equipo gana, empata o pierde). Estos modelos se basan principalmente en la utilización tanto de redes neuronales como de métodos de conjunto y técnicas de Gradient Boosting en árboles de decisión . Para ello, los modelos han sido entrenados empleando un conjunto de datos con información referente a la información de los diferentes equipos para todos los partidos de las últimas 5 temporadas (2017/18 – 2022/23) en las 5 grandes ligas europeas (Premier League, LaLiga, Ligue One, Bundesliga y Serie A). Para la evaluación de los resultados, se ha utilizado la información de la temporada 2023/24.

Para alcanzar este objetivo, se ha seguido una metodología de espiral iterativa. Comenzando por proponer soluciones simples y básicas, y después de un análisis riguroso de los resultados de los experimentos, se sugieren gradualmente soluciones más complejas. Luego, estos nuevos experimentos se evalúan y se proponen nuevamente nuevas soluciones, llevando el estudio hacia la solución óptima.

1.1 Motivación del trabajo

Las motivaciones detrás de este trabajo han sido diversas. En primer lugar, mi pasión por el fútbol fue un impulso clave para su realización. Desde pequeño, he sentido un gran interés por el fútbol, y aún más por las estadísticas y datos relacionados con los equipos y jugadores.

Anticipar los resultados de los partidos puede ser de gran ayuda para los equipos, ya que les permite prepararse mejor para los próximos enfrentamientos mediante el ajuste de estrategias de juego, entrenamientos y selección de jugadores. Por esta razón, contar con esta información es fundamental para los equipos de fútbol. Les posibilita modificar su estrategia y tácticas en función del resultado esperado, lo que podría aumentar sus probabilidades de éxito. Además, les ayuda a identificar a los jugadores más adecuados para el partido según sus habilidades y rendimiento previo, así como gestionar mejor los recursos, como el tiempo de entrenamiento, reuniones de equipo y periodos de descanso, en función del resultado anticipado. Esta información también es crucial para las casas de apuestas, las cuales la necesitan para establecer las cuotas, gestionar riesgos y mantener un balance rentable.

En segundo lugar, desde un punto de vista profesional, esta investigación se me presentó como una excelente oportunidad para iniciarme en el campo de la investigación y la Ciencia de Datos, que es el ámbito en el que quiero especializarme en el futuro. También me sentí motivado por el hecho de que la Analítica Deportiva es un área aún poco explorada. Dado que el mundo del fútbol está integrando cada vez más el Big Data y la Inteligencia Artificial, veo en ello una posible carrera profesional que combina mis dos pasiones: el fútbol y la Ciencia de Datos.

Como mencioné, fue una excelente oportunidad para adentrarme en una rama de la Ciencia de Datos que aún no ha sido ampliamente estudiada, lo cual me impulsó a aceptar el desafío. Sabía que el problema era complejo, ya que identificar una correlación entre los eventos de un partido de fútbol y su resultado no es una tarea sencilla debido a la alta dosis de azar y las limitaciones de los datos disponibles.

1.2 Objetivos

A menudo fijar un objetivo suele ser una tarea complicada en trabajos tan extensos y áreas tan amplias como el aprendizaje automático. El objetivo principal de este trabajo es desarrollar un modelo de aprendizaje automático capaz de predecir los resultados de los partidos de fútbol con la misma eficacia que los modelos mencionados en el marco teórico. Para lograr esto, se buscará diseñar un modelo predictivo que pueda determinar el resultado de un partido (victoria, empate o derrota) utilizando un conjunto de características seleccionadas previamente. La eficacia de este modelo se medirá a través de su precisión, es decir, el porcentaje de predicciones correctas en un conjunto de datos de prueba, además de otras métricas como la matriz de confusión, la precisión, la sensibilidad, la

especificidad y el F1-score, que ofrecerán una visión más detallada de su rendimiento.

Otro objetivo importante es identificar las características clave que influyen en la predicción de los resultados de fútbol. Esto implica analizar y seleccionar las variables que más afectan la precisión del modelo, lo que no solo mejorará su rendimiento, sino que también proporcionará una comprensión más profunda de los factores determinantes en los resultados de los partidos. Para medir este objetivo, se evaluará la importancia de cada característica utilizando técnicas como la importancia de características basada en árboles de decisión. Además, se valorará la mejora en la precisión del modelo después de aplicar técnicas de selección de características o reducción de dimensionalidad.

Finalmente, este trabajo también se propone comparar diferentes algoritmos de aprendizaje automático con el fin de identificar cuál es el más eficaz en la predicción de resultados de fútbol. Para ello, se evaluará el rendimiento de varios algoritmos. El éxito de este objetivo se medirá mediante la comparación de métricas de rendimiento como la precisión, el F1-score, la sensibilidad y la exactitud. Además, se tendrá en cuenta el tiempo de entrenamiento y la complejidad computacional de cada algoritmo, así como su capacidad de generalización, evaluada mediante técnicas de validación cruzada.

2. Marco teórico

En este capítulo se hace un análisis retrospectivo sobre trabajos con una temática relacionada o similar a la de este proyecto. El objetivo principal es encontrar ideas que nos puedan ser de utilidad y ver cuáles son las bases sobre las que podemos partir para la realización del Trabajo de Fin de Grado. En él, encontramos tres secciones diferentes: Analítica deportiva (donde se repasa qué es esta y la historia de la analítica deportiva), analítica deportiva en el fútbol (similar a la anterior, pero con un carácter específico para este deporte) y modelos de predicción de resultados (donde tratamos de introducir la problemática de la predicción a la vez que comparamos ideas, ventajas y desventajas de otros proyectos sobre esta temática).

Al final de este último apartado encontramos una tabla que resume los principales trabajos que se han documentado, mostrando las técnicas, los datos y los resultados obtenidos para cada uno de ellos.

2.1. Analítica deportiva

La analítica deportiva es el uso de métodos y técnicas de análisis de datos para mejorar la comprensión y el rendimiento en el ámbito deportivo. Esta disciplina abarca la recolección, procesamiento e interpretación de datos relacionados con el deporte, tales como estadísticas de rendimiento de jugadores, datos biométricos, información táctica y resultados de partidos (Navarro, 2024). A través de herramientas avanzadas como algoritmos de aprendizaje automático, análisis predictivo y visualización de datos, la analítica deportiva permite a entrenadores, equipos y organizaciones tomar decisiones informadas para optimizar estrategias de juego, mejorar la salud y el entrenamiento de los atletas, y aumentar el rendimiento general (Puig, 2024). Además, la analítica deportiva también se utiliza para mejorar la experiencia de los aficionados, optimizar operaciones comerciales y desarrollar estrategias de marketing efectivas en el mundo deportivo.

La historia de la analítica deportiva se remonta a la década de 1960, cuando los equipos de béisbol comenzaron a utilizar estadísticas básicas para evaluar el rendimiento de los jugadores. Sin embargo, la verdadera revolución comenzó a principios del siglo XXI con la publicación del libro (Lewis, 2003), que narra cómo el equipo de béisbol Oakland Athletics utilizó análisis estadísticos para competir con equipos con presupuestos mucho mayores. Este evento marcó un punto de inflexión y motivó a otros deportes a adoptar métodos similares.

La analítica deportiva se divide en varias áreas principales (UNIR, 2024), entre las que destacan:

- **Análisis de Rendimiento**: Implica el uso de datos para evaluar y mejorar el rendimiento de los atletas. Esto puede incluir el monitoreo de la condición física, el análisis de técnicas y la prevención de lesiones. Herramientas como GPS, sensores biométricos y cámaras de alta velocidad son comunes en este tipo de análisis. Existen varios estudios

llevados a cabo en esta área los cuales exploran la importancia de comprender las demandas fisiológicas específicas de los diferentes deportes y cómo estas pueden influir en el rendimiento de los jugadores (Drust, Atkinson, & Reilly, 2007; Bourdon, y otros, 2017)

- **Estrategia y Tácticas**: Los equipos utilizan análisis de datos para desarrollar y ajustar sus estrategias y tácticas. Esto puede incluir el análisis de los patrones de juego de los oponentes, la optimización de alineaciones y la planificación de jugadas (Puig, 2024). Se prevén años emocionantes para el análisis de estrategias y tácticas en el mundo del deporte ya que cada vez más datos estarán disponibles, permitiendo investigaciones más detalladas. La adopción de tecnologías de *big data* para la investigación en deporte podría proporcionar soluciones a algunos de los problemas clave (Rein & Memmert, 2016).

- **Marketing y conexión con el aficionado**: La analítica también se utiliza para mejorar la experiencia de los aficionados, aumentar la lealtad de estos y, en general, mejorar la gestión del club en la relación con sus aficionados. En esta área, los estudios más comunes vienen relacionados con el impacto que tienen tanto los precios de las entradas (Shapiro, Drayer, & Dwyer, 2016), como los éxitos conseguidos por el club (Tachis & Tzetzis, 2015) en la lealtad de los aficionados.

- **En el ámbito médico**: Esta tecnología no solo ayuda a predecir lesiones, sino que también aprovecha los datos de recuperación para comparar y mejorar los tratamientos. Además, permite estimar la trayectoria deportiva de un jugador basándose en sus estadísticas y antecedentes médicos, a menos que ocurra una lesión grave inesperada o surjan factores externos que el algoritmo no pueda considerar. En resumen, ofrece un control médico personalizado adaptado a las condiciones específicas de cada jugador, lo que traerá beneficios directos para su salud a mediano y largo plazo (UNIR, 2024).

- **Desde un punto de vista económico y financiero**: Correctamente implementado, favorece realizar una lectura correcta del mercado de fichajes, con un uso más eficiente de los recursos con el objetivo de lograr una plantilla más optimizada con la que afrontar, de forma competitiva, la temporada (Navarro, 2024).

2.2. Analítica deportiva en el fútbol

La analítica deportiva en el fútbol ha cobrado una relevancia crucial, distinguiéndose de su aplicación en otros deportes por la complejidad y la naturaleza dinámica del juego. Mientras que deportes como el béisbol y el baloncesto, con una estructura más segmentada y una mayor cantidad de datos discretos por evento, permiten un análisis más directo y estadísticamente significativo, el fútbol presenta desafíos únicos debido a su flujo continuo, la influencia de variables contextuales y la menor frecuencia de eventos decisivos

como goles (Acosta, 2022). En el fútbol, la analítica no solo abarca la evaluación del rendimiento individual y táctico en tiempo real, sino que también integra factores externos como las condiciones climáticas y el estado emocional de los jugadores, haciendo necesario el uso de modelos más sofisticados y una interpretación más cualitativa de los datos. Esta complejidad y diversidad en los tipos de datos requieren un enfoque multidisciplinario y técnicas avanzadas de análisis para obtener una ventaja competitiva y optimizar todas las facetas del juego, desde la estrategia en el campo hasta la gestión de los recursos humanos del equipo.

La analítica deportiva en el fútbol abarca todas las áreas mencionadas en el apartado 2.1. Sin embargo, en este deporte existe un área adicional muy extensa e importante: **la evaluación del valor y potencial de los jugadores**. Este campo permite a los clubes tomar decisiones informadas sobre fichajes, transferencias y renovaciones de contratos. Al utilizar datos objetivos, los clubes pueden minimizar riesgos y optimizar sus inversiones en jugadores. La identificación de jugadores sobreestimados, la identificación de jóvenes talentos es particularmente importante para clubes que buscan construir equipos competitivos a largo plazo y mantenerse a la vanguardia del talento emergente (Pappalardo, Cintia, Giannotti, & Pedreschi, 2019).

Algunos estudios se han centrado en explicar casos de éxito al aplicar la analítica deportiva en el mundo del deporte. En (Wilson, 2018), Jonathan Wilson explora la influencia duradera del FC Barcelona en el fútbol moderno, centrándose en dos de los entrenadores más destacados y controvertidos de la era contemporánea: Pep Guardiola y José Mourinho. A través de una narrativa detallada y perspicaz, Wilson analiza cómo estos dos entrenadores, con sus estilos contrastantes y filosofías de fútbol, han moldeado y redefinido el deporte a través de un análisis extendido y preciso.

2.3. Modelos de predicción de resultados

La predicción de resultados de partidos representa un desafío significativo debido a la complejidad y la naturaleza multifacética del deporte. En primer lugar, el fútbol es inherentemente impredecible debido a factores aleatorios y eventos fortuitos que pueden influir en el resultado, como errores arbitrales, lesiones inesperadas o condiciones climáticas adversas. Además, la dinámica del equipo, incluyendo tácticas de juego, estado físico y psicológico de los jugadores, así como la interacción entre ellos, añade otra capa de dificultad. También se deben considerar variables externas como el rendimiento histórico, las estadísticas individuales y colectivas, y los factores de localía. La combinación de estos elementos, junto con la necesidad de manejar grandes volúmenes de datos y aplicar modelos estadísticos o de aprendizaje automático, convierte la predicción en un reto complejo que requiere un enfoque multidisciplinario para lograr precisión y fiabilidad.

Los primeros estudios en el campo de la predicción de resultados de fútbol surgen durante la década de los noventa, entre los que destaca el realizado por (Dixon & Coles, 2002). En su trabajo, los autores desarrollan un modelo paramétrico basado en la distribución de Poisson para predecir los resultados de los partidos de

fútbol. Utilizando datos de la liga inglesa y de la FA Cup entre 1992 y 1995, demostraron que su modelo no solo describe los resultados históricos, sino que también puede identificar ineficiencias en el mercado de apuestas. Esto sugiere que las probabilidades estimadas mediante su modelo pueden superar a las cuotas ofrecidas por las casas de apuestas, presentando así una estrategia potencialmente rentable para apostar en partidos de fútbol. Sin embargo, una de las principales limitaciones del trabajo de Dixon y Coles es el hecho de trabajar únicamente con un histórico de resultados para cada equipo. Haciendo esto, se obvian otras características del fútbol que pueden llegar a ser importantes a la hora de predecir los resultados de un partido.

En (Heerde, Hardie, & Leeflang, 2015) los autores demuestran que los equipos en una mala racha son más propensos a perder sus próximos partidos. Utilizando un modelo estadístico que controla diversas variables, encuentran que los equipos con resultados negativos previos tienen una mayor probabilidad de continuar perdiendo, lo cual subraya la influencia psicológica y de confianza en el rendimiento deportivo. Sin embargo, el estudio se centra en los datos de seguimiento y algunos aspectos tácticos, pero podría no considerar plenamente factores externos, como la fatiga acumulada de los jugadores, las decisiones arbitrales, o los cambios estratégicos no reflejados en los datos de posición. El autor Mark Nesti destaca en su libro (Nesti, 2010) la importancia del estado anímico y la salud mental en el deporte, destacando cómo los fracasos en lo deportivo y en la vida personal del deportista pueden retroalimentarse generando un estado en el que es imposible trabajar.

En (Baboota & Kaur, 2018) se muestra el impacto positivo de utilizar información más sofisticada como la presencia de ciertos jugadores claves, las lesiones que se han producido en el equipo, la información relativa a los últimos partidos disputados por el equipo, etc. En él se emplean modelos más sencillos como *Gaussian Naive Bayes*, *Support Vector Machine* y *Random Forest*. Estos modelos ofrecen buenos resultados (>60% exactitud), aunque denotan la dificultad de predecir cuándo un partido va a acabar en empate. Estos resultados pueden ser consecuencia de la limitada cantidad de datos con la que se contaba, además de que el propio estudio reconoce que la inclusión de características más detalladas y sofisticadas podría mejorar los resultados.

Algunos estudios como (Stübinger, Benedikt, & Knoll, 2019) plantean la utilización de métodos de conjunto para la predicción de resultados deportivos. En él se plantea la técnica conocida como *Weighted Ensembled All Method*, que consiste en combinar múltiples modelos predictivos para mejorar la precisión de las predicciones. Este enfoque aprovecha la fortaleza de cada modelo individual, asignando pesos a sus predicciones de acuerdo con su rendimiento relativo. Otras técnicas empleadas para tratar de comprender la naturaleza de los datos y mejoras las predicciones ofrecidas por los modelos son las técnicas de reducción de la dimensionalidad. En (Tax & Joustra, 2015), los autores utilizaron la técnica de Análisis de Componentes Principales (PCA) para la reducción de dimensionalidad, seleccionando un umbral de varianza del 15% que demostró ser óptimo en sus

experimentos. Entre los modelos de clasificación probados, el clasificador *Naive Bayes* y el Perceptrón Multicapa (*Multilayer Perceptron, MLP*) mostraron una alta precisión cuando se combinaban con PCA.

El *Deep Learning* es un conjunto de algoritmos que permite aprender automáticamente múltiples niveles de representaciones de la distribución subyacente de los datos a modelar. En otras palabras, los algoritmos de aprendizaje profundo extraen automáticamente características de bajo y alto nivel necesarias para la clasificación, eliminando la necesidad de diseñar manualmente características, lo que puede ser un proceso laborioso y propenso a errores (Lauzon, 2012). Dentro de este campo, (Rahman, 2020) desarrolla un marco basado en redes neuronales profundas y artificiales para predecir resultados de partidos de fútbol, empleando técnicas de reducción de dimensionalidad como *embeddings* profundos y clasificadores *softmax*. Estos modelos, en particular las redes LSTM, han sido de gran utilidad en nuestro proyecto para manejar la secuencialidad y complejidad de los datos deportivos, mejorando significativamente la precisión de las predicciones. Sin embargo, el estudio de Rahman cuenta con una limitación muy importante, pues se trabaja con las estadísticas de selecciones nacionales, por lo que cuenta con una variabilidad muy limitada en los datos, además de un conjunto muy pequeño para la validación de los resultados.

Por último, cuando hablamos de aprendizaje automático una forma de tanto evaluar los resultados de nuestros modelos como de contribuir a la mejora de estos, las métricas de pérdida y evaluación son imprescindibles. En (Rodrigues & Pinto, 2022) los autores evaluaron el rendimiento de varios modelos de predicción de resultados de partidos de fútbol utilizando múltiples métricas de error. La exactitud la métrica principal, calculada como la proporción de predicciones correctas entre todas las predicciones realizadas. Además, utilizaron una matriz de confusión para proporcionar una visión detallada de los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, lo que permitió una comprensión más profunda de los errores específicos del modelo. Para evaluar la efectividad de las predicciones en el contexto de apuestas deportivas, se calcularon la tasa de éxito de apuestas y el margen de beneficio, ofreciendo una perspectiva sobre el rendimiento financiero potencial del modelo. Aunque se realizó un análisis de rendimiento a lo largo de la temporada, no se menciona un análisis exhaustivo de la estabilidad del modelo en diferentes contextos o condiciones, lo que podría ser crucial para su implementación en un sistema de soporte de decisiones.

En este trabajo, tratamos de combinar varias de las técnicas explicadas en este apartado para tratar de mejorar los resultados obtenidos hasta el momento.

Referencia	Datos	Modelo	Exactitud
(Dixon & Coles, 2002)	Premier League 1992-1995	Modelo de Distribución de Poisson	52%
(Heerde, Hardie, & Leeflang, 2015)	Premier League 1990-2002	Modelo basado en grafos y RNN	51’3%
(Baboota & Kaur, 2018)	Premier League 2005-2016	Gradient boosting, Random Forest, Gaussian Naïve Bayes y SVM	61’2%
(Stübinger, Benedikt, & Knoll, 2019)	5 grandes ligas 2006-2018	Gradient boosting, Random Forest, Gaussian Naïve Bayes, Regresión lineal y métodos de conjunto	>80% para clasificación binaria (no se tiene en cuenta empates)
(Tax & Joustra, 2015)	Eredivise, temporadas no especificadas	Naïve Bayes y Perceptrón multicapa + PCA	54%
(Rahman, 2020)	No especificado	Redes neuronales LSTM	63%
(Rodrigues & Pinto, 2022)	Premier League 2013-2019	Random Forest	58’7%

3. Marco legal y ético

En el desarrollo de este proyecto, hemos asegurado que la recolección, almacenamiento y uso de los datos cumplen con todas las normativas legales aplicables, tanto a nivel nacional como internacional. En particular, hemos prestado especial atención a las disposiciones de la Ley de Protección de Datos Personales y al Reglamento General de Protección de Datos (GDPR) de la Unión Europea, que establecen estándares estrictos para el manejo de información personal.

El GDPR, reconocido por su rigidez en la protección de los derechos de los individuos, establece en su Artículo 5 principios fundamentales para el tratamiento de datos personales, como la legalidad, lealtad y transparencia en su uso, y la minimización de datos, que requiere que solo se recolecten los datos estrictamente necesarios para los fines específicos del tratamiento. Además, el Artículo 6 del GDPR especifica que el tratamiento de datos personales es lícito solo si se basa en una de las condiciones legales enumeradas, como el consentimiento explícito del interesado o el cumplimiento de una obligación legal.

Es importante destacar que, aunque el GDPR no regula directamente la recolección y uso de estadísticas de eventos públicos, como los partidos de fútbol, siempre que estas estadísticas no incluyan información personal identificable, hemos tomado medidas para garantizar el cumplimiento de otras disposiciones relevantes. Por ejemplo, de acuerdo con el Considerando 26 del GDPR, los datos que han sido adecuadamente anonimizados, de manera que no puedan ser asociados con una persona física identificada o identificable, están fuera del ámbito de aplicación del reglamento. Esto nos permite utilizar datos anonimizados para análisis estadísticos sin incurrir en riesgos de incumplimiento.

Adicionalmente, el Artículo 89 del GDPR permite el uso de datos personales para fines estadísticos y de investigación siempre que se implementen salvaguardias adecuadas, como la pseudonimización o anonimización de los datos, para proteger los derechos y libertades de los individuos.

Hemos implementado políticas internas rigurosas para asegurar que todos los miembros del equipo estén plenamente conscientes de sus responsabilidades en cuanto a la protección de datos. Estas políticas incluyen procedimientos detallados para la recolección, tratamiento, almacenamiento y eliminación segura de datos, alineados con las mejores prácticas internacionales y las exigencias del GDPR.

Con este enfoque integral, no solo cumplimos con las obligaciones legales, sino que también reforzamos nuestro compromiso con la privacidad y seguridad de los datos, asegurando que el proyecto se lleve a cabo con los más altos estándares éticos y de cumplimiento normativo.

4. Bases teóricas

El aprendizaje automático, es una rama de la inteligencia artificial que permite a los sistemas informáticos aprender y mejorar automáticamente a partir de la experiencia, sin ser programados explícitamente para ello. Utiliza algoritmos y modelos estadísticos para analizar y detectar patrones en grandes volúmenes de datos, lo que permite a las máquinas tomar decisiones o hacer predicciones basadas en la información aprendida (Yousef & Allmer, 2014). Este enfoque es fundamental en una amplia gama de aplicaciones, desde la personalización de contenido en plataformas digitales hasta la conducción autónoma, y está transformando la manera en que interactuamos con la tecnología en nuestra vida cotidiana.

4.1 Métodos de conjunto

Los métodos de conjunto, también conocidos como *Ensemble Methods*, son técnicas que combinan múltiples modelos individuales para producir un modelo más robusto y preciso (Opitz & Maclin, 1999). La idea principal detrás de los métodos de ensemble es que, al combinar varios modelos, se pueden compensar las debilidades individuales de cada modelo y mejorar el rendimiento general.

En los últimos años, los métodos de conjunto han ganado una importancia notable en el campo del machine learning y la inteligencia artificial, convirtiéndose en una herramienta esencial para la resolución de problemas complejos y la mejora de la precisión de los modelos predictivos. Su capacidad para combinar múltiples modelos individuales, compensando así las debilidades y errores de cada uno, ha demostrado ser crucial en diversas aplicaciones, desde la predicción financiera y la detección de fraudes hasta el diagnóstico médico y el procesamiento del lenguaje natural. La eficacia estos modelos en competiciones de machine learning, como las organizadas por *Kaggle*[1], ha evidenciado su potencial para superar a modelos individuales, logrando un rendimiento superior y más robusto.

Existen múltiples tipos de métodos de conjunto (Zhang & Ma, 2012), aunque los 3 principales son los siguientes:

- **Bagging**. El *bagging* es una técnica que mejora la precisión del modelo al reducir la varianza (Opitz & Maclin, 1999). Este método implica la creación de múltiples versiones de un modelo base utilizando diferentes subconjuntos del conjunto de datos de entrenamiento, generados mediante muestreo con reemplazo. Cada modelo se entrena de forma independiente y las predicciones finales se obtienen mediante la combinación de las predicciones individuales, promediándolas en el caso de regresión o realizando una votación en el caso de clasificación.

- **Boosting**. El *boosting* es una técnica secuencial en la que cada nuevo modelo se construye para corregir los errores de los modelos

[1] https://www.kaggle.com/

anteriores (Zhang & Ma, 2012). En este enfoque, los modelos se entrenan uno tras otro, y cada modelo nuevo se centra en las instancias de datos que fueron mal clasificadas por los modelos anteriores.

- **Voting**. El *voting* es uno de los métodos de conjunto más simples, en el que las predicciones de varios modelos individuales se combinan para hacer la predicción final (Zhang & Ma, 2012). En el caso de clasificación, se puede utilizar la votación mayoritaria, donde la clase que recibe más votos es la predicción final, o la votación ponderada, donde las predicciones de cada modelo se ponderan según su precisión. A menudo se suele confundir este con *bagging*, sin embargo, estos modelos difieren en su enfoque y la manera de utilizar los datos. Mientras que el *voting* se centra en combinar las predicciones de varios modelos diferentes entrenados en el mismo conjunto de datos, el *bagging* se centra en reducir la varianza mediante la combinación de múltiples versiones de un mismo modelo entrenado con diferentes subconjuntos de datos.

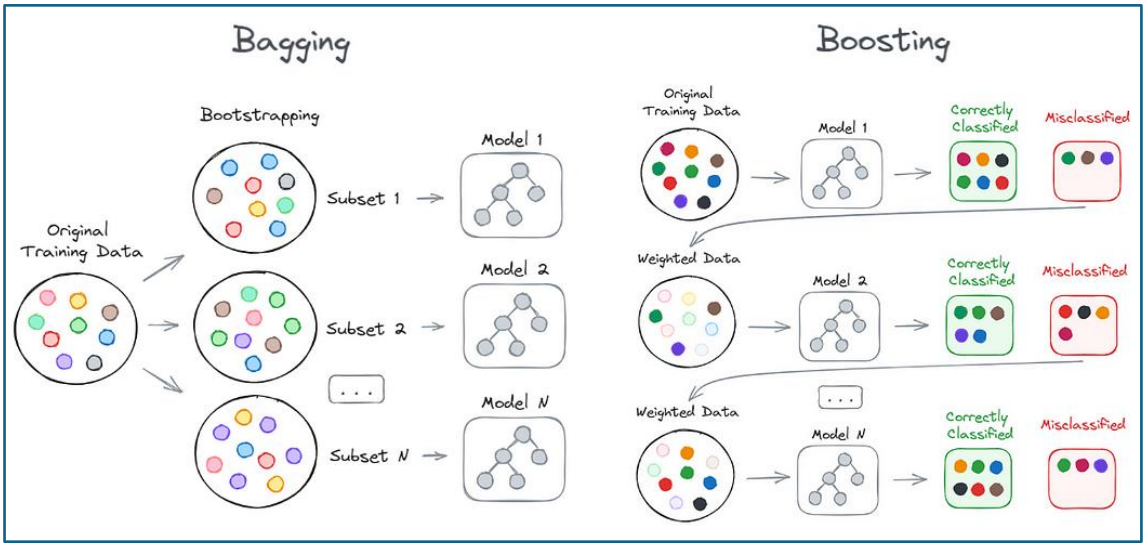

Figura 1. Bagging vs boosting (Dorfer, 2023)

4.2 Árboles de decisión

Un árbol de decisión es un clasificador expresado como una partición recursiva del espacio de instancias. Consiste en un conjunto de nodos unidos mediante aristas las cuales sirven para tomar las decisiones, es decir, el árbol de decisión es un árbol con direcciones (Chollet, 2021). Dentro de un árbol de decisión encontramos tres tipos de nodos: nodo raíz, nodos internos y hojas (también conocidos como nodos de decisión).

Los dos primeros tipos de nodos, raíces e internos, representan una operación lógica que permite dividir el espacio de instancias en dos o más subespacios. En los casos más simples, cada operación considera un solo atributo de la instancia, de manera que el espacio de instancias se particiona de acuerdo con el valor del atributo.

Las hojas, por su parte, representan el resultado final o la decisión tomada después de seguir el camino desde el nodo raíz a través de los nodos internos. En clasificación, una hoja puede representar una clase; en regresión, puede representar un valor numérico.

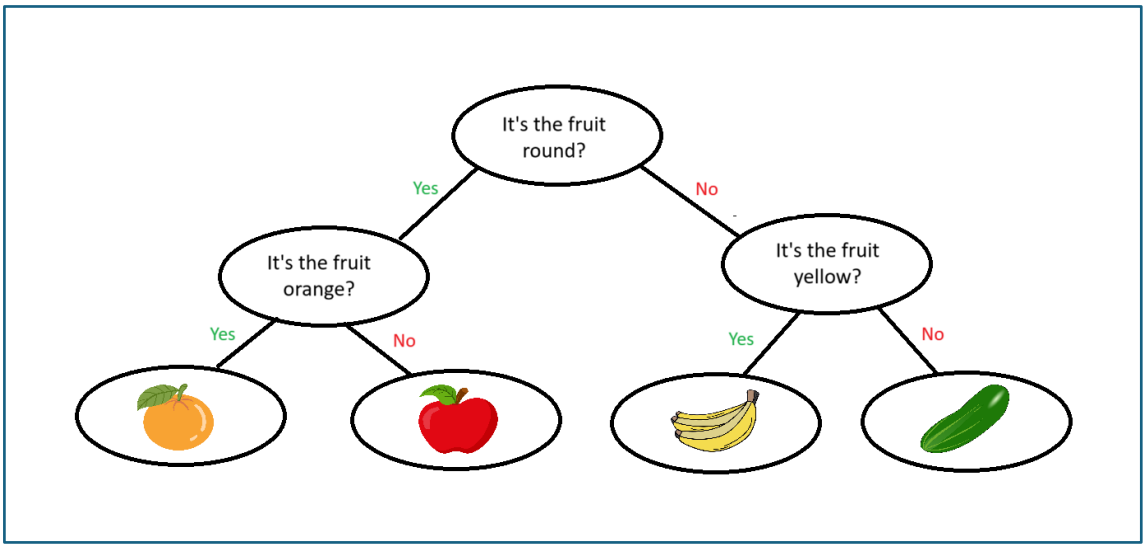

Figura 2. Ejemplo simple de árbol de decisión basado en variables booleanas.

4.2.1 Criterios de división univariante de un nodo

En la construcción de árboles de decisión en aprendizaje automático, la división de nodos es una etapa crucial que se realiza utilizando diversos criterios para garantizar la mejor segmentación posible de los datos. Existen múltiples criterios para llevar a cabo esta división, aunque entre estos destaca la división por la **entropía**.

La entropía H se define para un conjunto de datos con respecto a la distribución de las clases presentes en el conjunto. Para un conjunto de datos D con k clases posibles, la entropía se calcula como:

$$H(D) = -\sum_{i=1}^{k} p_i log_2(p_i)$$

donde p_i es la proporción de instancias en el conjunto D que pertenecen a la clase i.

Por lo tanto, una entropía baja (valores cercanos a 0) indica que el conjunto de datos es muy homogéneo (predomina una sola clase), mientras que una entropía alta indica que el conjunto de datos es muy heterogéneo (las clases están distribuidas de manera uniforme). Cuando se realiza una división en un nodo, el objetivo es reducir la entropía total del sistema. Para evaluar una posible división, se calcula la entropía ponderada de los nodos hijos resultantes de la división.

$$H_{ponderada} = \sum_{j=1}^{m} \frac{|D_j|}{|D|} H(D_j)$$

Donde:

- m es el número de nodos hijos.
- D_j es el conjunto de datos en el nodo hijo j.
- $|D_j|$ es el número de instancias en el nodo hijo j.
- $|D|$ es el número de instancias en el nodo padre.

La división que produce la mayor ganancia de información (o la mayor reducción de entropía) se selecciona como la mejor división para ese nodo, siendo la ganancia de información la diferencia entre la entropía del nodo padre y la entropía ponderada del nodo hijo.

4.2.2 LightGBM

LightGBM es una implementación eficiente del algoritmo de Gradient Boosting desarrollado por Microsoft. Está diseñado para ser altamente eficiente en términos de tiempo y uso de memoria, lo que lo hace adecuado para trabajar con grandes volúmenes de datos y características de alta dimensionalidad. (Ke, y otros, 2017). LightGBM utiliza un modelo basado en histogramas para construir los árboles de decisión. Esto significa que convierte los valores de las variables en intervalos discretos, lo que reduce drásticamente el tiempo de cálculo y el uso de memoria en comparación con los árboles de decisión tradicionales que evalúan cada posible punto de división.

Además, este algoritmo emplea dos técnicas fundamentales para lograr su eficiencia:

1. **Muestreo Unilateral Basado en Gradiente (GOSS):** Esta técnica mejora la eficiencia al disminuir el número de instancias de datos utilizadas para calcular la ganancia de información durante el entrenamiento. GOSS filtra una gran cantidad de instancias con gradientes pequeños, reteniendo principalmente aquellas con gradientes grandes, que son más relevantes para la estimación de la ganancia de información. Esto permite que se realice una estimación precisa utilizando un conjunto de datos más reducido, lo que reduce significativamente el tiempo de procesamiento (Ke, y otros, 2017).
2. **Agrupación Exclusiva de Características (EFB):** Esta técnica se utiliza para disminuir el número de características efectivas en el modelo. En entornos donde las características son esparsas, muchas de ellas son casi exclusivas, es decir, rara vez presentan valores diferentes de cero al mismo tiempo. EFB combina estas características mutuamente exclusivas en una

sola, lo que reduce la cantidad de características que el modelo necesita manejar sin comprometer la precisión. Esta reducción disminuye el tiempo de entrenamiento y el uso de memoria (Ke, y otros, 2017).

Estas técnicas permiten a LightGBM entrenar modelos de Gradient Boosting de manera mucho más rápida y eficiente en comparación con los métodos tradicionales.

4.3 Redes neuronales Long-Short Term Memory (LSTM)

Las redes neuronales artificiales (ANN) son un algoritmo de aprendizaje automático basado en unidades de computación a las cuales se les denomina “neuronas”. Estas neuronas son en realidad operaciones matemáticas que toman un número de entradas con valores reales y producen una única salida con valor real. Basadas en la conectividad entre las diferentes neuronas, estas redes pueden modelar un comportamiento global complejo (Staudemeyer & Morris, 2019).

Dentro del campo de las redes neuronales, se han desarrollado diferentes algoritmos para tratar de resolver problemas de aprendizaje automático. Un tipo de red neuronal son las redes neuronales recurrentes (RNN), las cuales tienen la capacidad de "recordar" la información de entradas anteriores a medida que procesan cada elemento de la secuencia. Estas redes neuronales son particularmente útiles para procesar secuencias de datos, como texto, series temporales o cualquier otro tipo de dato donde el orden y la dependencia temporal son importantes (Smagulova & James, 2019). Sin embargo, un problema común con las RNN tradicionales es que, a medida que la secuencia de datos se alarga, la capacidad de la red para "recordar" la información relevante de entradas anteriores se desvanece, lo que se conoce como el problema del "desvanecimiento del gradiente". Esto hace que las RNN tradicionales tengan dificultades para aprender dependencias a largo plazo en la secuencia.

Las redes LSTM fueron introducidas para resolver este problema. Una LSTM es una versión avanzada de las RNN que tiene una estructura de celda especial, diseñada para mantener y actualizar información durante largos periodos de tiempo (Staudemeyer & Morris, 2019). La celda LSTM tiene tres puertas principales:

- Puerta de entrada: Controla cuánta información nueva debe almacenarse en la celda.
- Puerta de olvido: Decide cuánta de la información almacenada debe olvidarse.

- Puerta de salida: Determina cuánta información de la celda debe usarse para generar la salida actual.

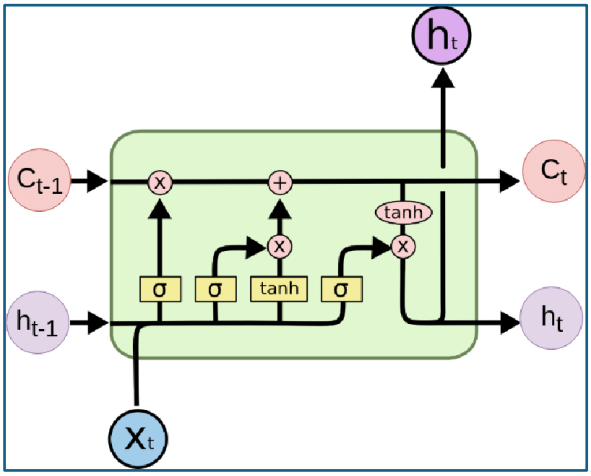

Figura 3. Arquitectura de una neurona en las redes neuronales LSTM (Varikuti, 2021)

5. Desarrollo técnico del proyecto

En este capítulo se detallará todo el proceso técnico llevado a cabo para la implementación de este trabajo. En primer lugar, se describirán las fuentes de datos utilizadas, proporcionando una visión clara de su origen y características. Posteriormente, se realizará un análisis exhaustivo de estos datos para identificar patrones y preparar la información para su uso posterior. A continuación, se abordará el proceso de transformación de los datos, asegurando que estén en el formato adecuado para su análisis. Finalmente, se presentarán los modelos que se emplearán en el proyecto, explicando su elección y cómo se aplicarán para alcanzar los objetivos planteados.

5.1 Fuentes de datos

Para el desarrollo de este proyecto se han utilizado tres conjuntos de datos, a partir de los cuales se ha llevado un detallado análisis de las variables incluidas en ellos. Este análisis nos ha permitido ser capaces de comprender la naturaleza de las variables, comprendiendo cuáles podrían ser útiles para la tarea que tratábamos de resolver. Posteriormente, se llevaron a cabo una serie de procesos de ingeniería de datos con el fin de crear nuevas características que pudieran aportar valor y, por lo tanto, tratar de mejorar la calidad de las predicciones.

El primero de los dos conjuntos de datos se obtuvo mediante fuentes de datos públicas, en concreto de la página de FBREF[2], la cual proporciona datos actualizados diariamente sobre estadísticas tanto a nivel de jugador como a nivel de equipo para los diferentes partidos que se juegan a lo largo de una temporada. Este contenía información relativa a las estadísticas por partido de los diferentes jugadores que juegan en las 5 grandes ligas europeas (Alemania, España, Francia, Inglaterra e Italia). Estas estadísticas muestran información sobre diferentes facetas del juego, encontrando estadísticas relacionadas con la faceta ofensiva (como podrían ser los remates a puerta), estadísticas relacionadas con la faceta defensiva (como podría ser el número de faltas), estadísticas relacionadas con la creación de juego (como podría ser el número de pases) e información contextual (como podría ser la liga y el equipo en el que juega el jugador). Todas estas estadísticas fueron recogidas para las últimas 5 temporadas, contando con datos desde 2018 hasta 2023. Ver Anexo II para conocer todas las variables del conjunto de datos.

season	game	B365H	B365D	B365A	BWH	BWD	BWA	IWH	IWD	IWA
18-19	Manchester United - Leicester City	1.57	3.9	7.5	1.53	4.0	7.5	1.55	3.8	7.0
18-19	Bournemouth – Cardiff City	1.9	3.6	4.5	1.9	3.4	4.4	1.9	3.5	4.1
18-19	Fullham – Crystal Palace	2.5	3.4	3	2.45	3.3	2.95	2.4	3.3	2.95
18-19	Huddersfield - Chelsea	6.5	4.0	1.61	6.25	3.9	1.57	6.2	4.0	1.55
18-19	Newcastle - Tottenham	3.9	3.5	2.04	3.8	3.5	2.0	3.7	3.35	2.05

Tabla 1. Estructura del conjunto de datos sobre cuotas de casas de apuestas.

[2] https://fbref.com/es/

El segundo de los conjuntos de datos empleados fue obtenido también mediante fuentes de acceso público, en concreto de Football-Data[3], una fuente de datos que ofrece información sobre los resultados de los partidos y las cuotas que ofrecían las casas de apuestas para estos partidos. La estructura de este conjunto de datos es muy sencilla, las dos primeras variables nos indican la temporada y el encuentro para los que se tienen las cuotas, mientras que el resto de las variables nos indica la cuota que le asignaba cada casa de apuesta a cada uno de los tres posibles resultados del partido (victoria del local, empate o victoria del visitante). Así, por ejemplo, la variable B365H es la cuota que ofrecía la casa de apuestas Bet365 a la victoria del equipo local. En la Tabla 1 se puede observar la forma de este conjunto de datos descrito.

El tercer y último de los conjuntos de datos se obtuvo de la página de Football Club Elo Rating[4]. El Elo Rating en fútbol es un sistema que mide la fortaleza de los equipos basándose en los resultados de sus partidos. Todos los equipos empiezan con una puntuación inicial, y esta se ajusta después de cada partido: sube si el equipo gana, baja si pierde, y cambia ligeramente si empata. Antes de cada partido, se calcula quién es el favorito según las puntuaciones actuales. Este sistema es útil porque permite comparar la fuerza de los equipos y se actualiza constantemente, reflejando su forma actual. En este conjunto de datos teníamos datos históricos sobre el Elo de los clubes durante los últimos años, actualizados semanalmente.

Figura 4. Evolución del Elo de diferentes clubes a lo largo de las temporadas.

[3] https://www.football-data.co.uk/data.php
[4] http://clubelo.com/

5.2 Análisis del dato y agrupación por equipos

Como ha quedado constatado en el apartado anterior, el conjunto de datos de las estadísticas de jugadores contaba con una gran cantidad de datos. El primer paso en todo problema de ciencia de datos es analizar la naturaleza de los datos y tratar de ver si existe algún tipo de anomalía. En este caso, la primera característica que tratamos de apreciar era la cantidad de datos faltantes, pues en el planteamiento que queríamos hacer del problema era fundamental no tener ausencia de información.

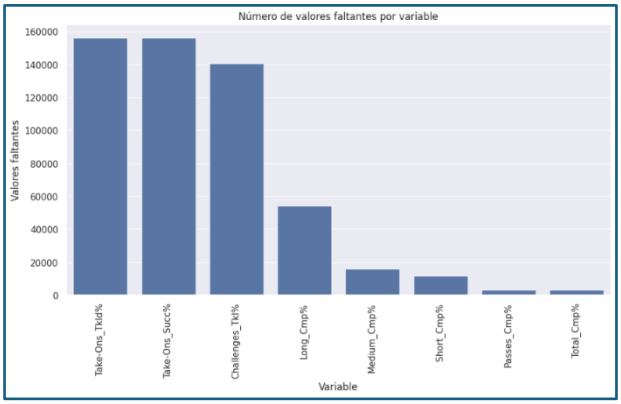

Figura 5. Número de valores faltantes para variables que tienen más de 1000 valores faltantes.

En la Figura 5 se aprecia que contábamos con una serie de variables las cuales contaban con una gran cantidad de valores faltantes. Al tratarse de información relacionada con porcentajes (% de entradas exitosas, % de pases acertados, etc.), que en caso de ser necesaria la mayoría seríamos capaces de calcular, decidimos eliminar estas variables de nuestro conjunto de datos.

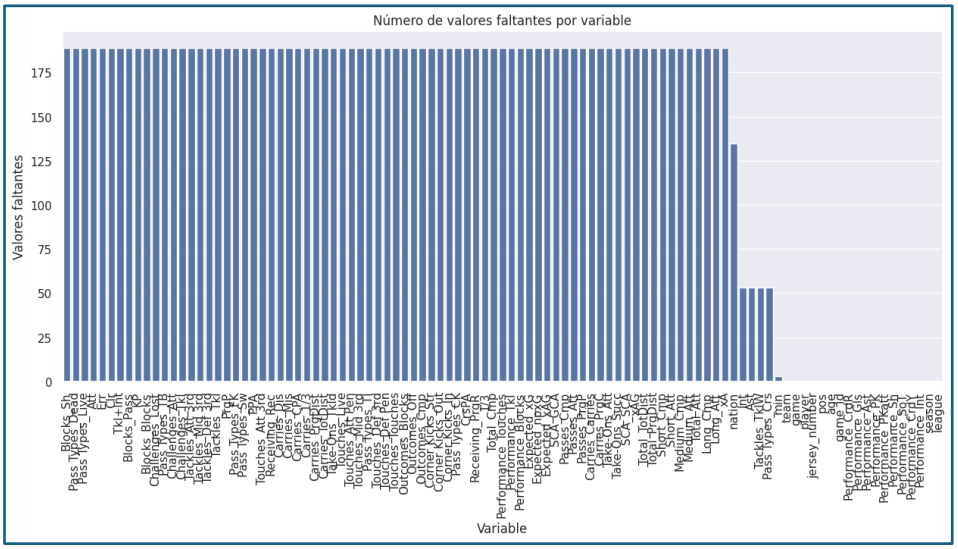

Figura 6. Número de valores faltantes para variables que tienen menos de 1000 valores faltantes

Por otro lado, en la Figura 6 se observa el número de valores faltantes del resto de variables (variables con menos de 1000 valores faltantes) se aprecia un patrón muy interesante, ya que muchas de estas cuentan exactamente con el mismo de número de datos faltantes: 189 datos faltantes. Además, existían otras variables las cuales contaban con un número muy pequeño de datos faltantes, como los goles, los tiros o las tarjetas amarillas que contaban con un único valor faltante. Analizando estas observaciones llegamos a descubrir que estas eran información sobre

jugadores que habían entrado al campo en el último minuto de juego, por lo que decidimos eliminar también estas observaciones de nuestro conjunto de datos, pues no tienen transcendencia en los partidos. Para eliminar estas observaciones, simplemente eliminamos las observaciones del conjunto de datos que tuvieran el valor de la variable "Minutos jugados" igual a 0.

Una vez teníamos nuestro conjunto de datos limpio de datos faltantes, y comprobado que no existe ningún dato que haya sido mal medido o haya habido algún tipo de error a la hora de guardarlo, la siguiente transformación necesaria era agrupar la información por equipos. Como sabemos, la información hasta este momento la teníamos mesurada individualmente por jugador, sin embargo, para la tarea de predicción de resultados no es suficiente. Por lo tanto, se llevó a cabo la agrupación de la información por equipo siguiendo la lógica de que si tengo una estadística (p.ej. número de pases) medida para varios jugadores de un equipo, la estadística a nivel de equipo para un partido será la suma de las estadísticas de los jugadores de ese equipo que disputaron el partido. Es decir, el número de goles que marca un equipo en un partido es la suma de los goles de cada uno de los jugadores de ese equipo en ese partido, el número de pases de un equipo es la suma de los pases de los jugadores, etc.

$$STAT_{equipo}(Partido) = \sum_{jugador \in equipo} STAT_{jugador}\,(Partido)$$

Además, agrupar la información por equipos nos permitió definir la variable objetivo que, hasta este momento, no teníamos. Sabiendo los goles que ha metido cada equipo por partido podemos calcular la diferencia de goles como la resta entre los goles del equipo local y los goles del equipo visitante. Esto a su vez nos permite definir si el equipo ha ganado el partido (diferencia de goles > 0), ha empatado (diferencia de goles = 0) o ha perdido (diferencia de goles < 0).

5.3 Cálculo de las medias ponderadas

En este trabajo se ha realizado una tarea de predicción a futuro, lo que comúnmente llamamos *forecasting*. Es por ello por lo que, la manera en la que encontramos el dato en nuestro conjunto, ya agrupado por equipos no nos sirve. Esto se debe a que, para una observación, nosotros tenemos el resultado del partido junto a las estadísticas del equipo en ese encuentro. Sin embargo, cuando tratamos de hacer predicciones de un partido de fútbol, no contamos con las estadísticas de ese partido, pero sí con las estadísticas de partidos anteriores. En (Heerde, Hardie, & Leeflang, 2015), se habla de cómo los estados de forma afectan a los resultados deportivos, de manera que se puede tratar los datos como una serie temporal y, por lo tanto, medir el estado de forma como una media ponderada de las estadísticas de los últimos partidos.

Por ello, basándonos en (Baboota & Kaur, 2018), transformamos nuestras variables de manera que ahora sean la media ponderada de los últimos encuentros. Es decir, en lugar de tener en una observación el número de goles de un partido, tendremos la media ponderada de goles de los anteriores partidos, asignándole una

ponderación mayor a los partidos más recientes. Matemáticamente podemos expresar este cálculo como:

$$X_t = \sum_{i=1}^{n} w_i * X_{t-i}$$

$$w_i = \frac{e^{-i}}{\sum_{j=1}^{n} e^{-j}}$$

Donde:

- X_t es el valor de la variable en un determinado período de tiempo t.
- n es el número de partidos que se emplean para calcular la media ponderada.
- w_i es la ponderación asignada al instante de tiempo.

Por lo tanto, debemos tener en cuenta dos consideraciones:

1. Se emplea la función exponencial en el cálculo de la ponderación porque es una manera efectiva de darle un mayor peso a los partidos más recientes. A pesar de esto, esta función no es estrictamente necesaria y se podría sustituir por otra forma de cálculo.
2. El cálculo se lleva a cabo utilizando los n partidos anteriores, por lo que los n primeros partidos de cada temporada no pueden ser calculados y simplemente se utilizan para realizar este cálculo de la media ponderada. Estas observaciones para las que no se puede realizar el cálculo no se emplean después en los modelos de aprendizaje automático.

Cabe recalcar que, tras haber hecho el cálculo de las medias ponderadas, se mantuvo en el conjunto de datos la variable correspondiente a los goles (Performance_Gls), pues se podía utilizar como variable objetivo en algunos algoritmos que tratan de resolver este problema.

5.4 Análisis exploratorio de datos

El análisis exploratorio de datos (EDA) es la fase inicial fundamental en cualquier proyecto de ciencia de datos. Su objetivo es obtener un conocimiento profundo del conjunto de datos, identificando patrones, relaciones y anomalías. A través de técnicas estadísticas descriptivas y visualizaciones, el EDA permite caracterizar las variables, detectar relaciones, validar supuestos y garantizar la calidad de los datos. En esta sección se presentan los resultados del EDA aplicado a nuestro conjunto de datos, proporcionando una base sólida para los análisis posteriores.

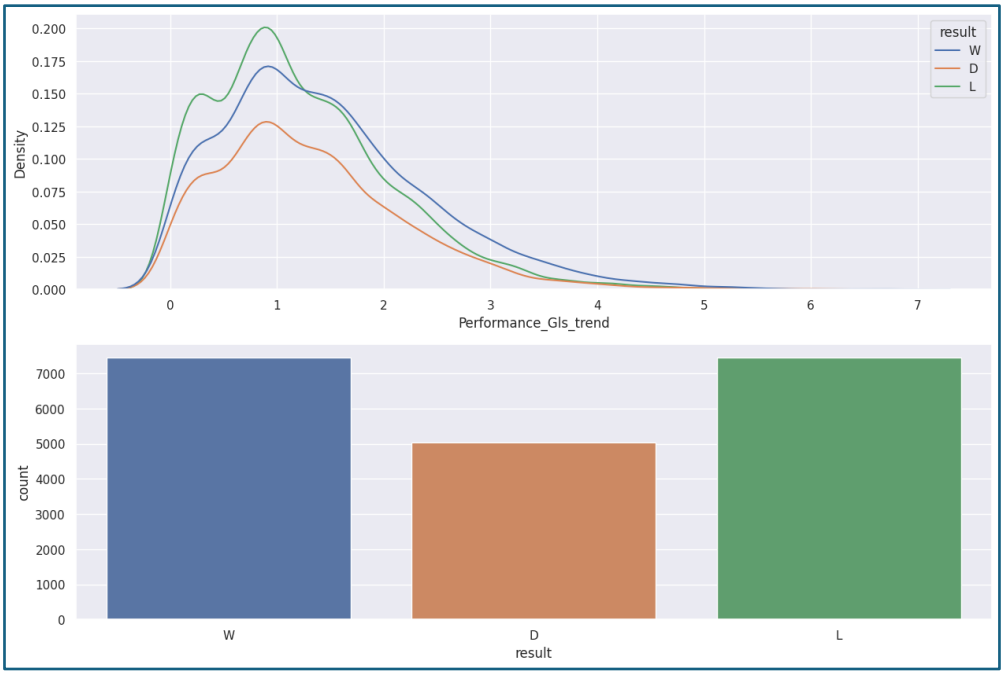

Figura 8. Distribución de media de goles por resultado.

En primer lugar, en la Figura 7 se puede apreciar la distribución de las medias ponderadas de los goles que marca un equipo para cada uno de los posibles resultados. En estas distribuciones se pueden observar varias ideas, la primera de ellas es que sí parece existir un factor que relacione la media ponderada de goles con el resultado del siguiente partido, pues cuando el valor de la media ponderada de goles es inferior a 1'2 se observa que el resultado con mayor densidad es el de derrota, mientras que cuando se supera este valor el resultado con mayor densidad es el de victoria. Esto puede ser un indicio de las conclusiones que saca Nesti en (Nesti, 2010). Por otro lado, también se aprecia que el resultado menos frecuente y que no destaca por encima de los demás es el empate, hecho que nos indica que muy probablemente sean los encuentros terminados en empate los más difíciles de clasificar o diferenciar.

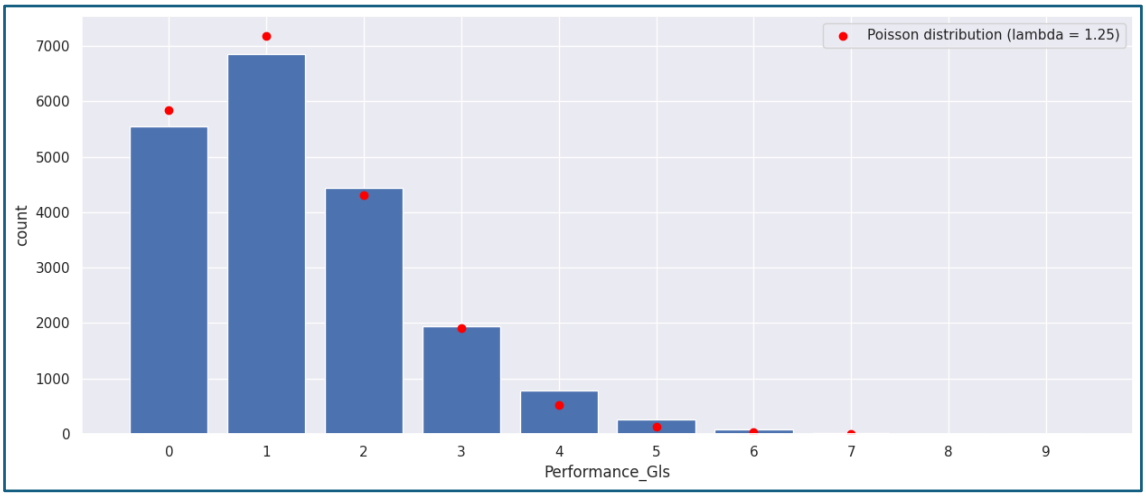

Figura 7. Distribución del número de goles marcados en un partido.

En la Figura 8 se observa la distribución del número de goles marcados en los diferentes partidos. La mayoría de los partidos registran entre 0 y 3 goles, siendo el

valor más frecuente 1 gol. Además, se evidencia que esta distribución se adapta a una distribución de Poisson con un valor λ = 1'25, por lo que a medida que aumenta el número de goles, la frecuencia disminuye considerablemente. Es raro ver partidos con 5 goles o más, lo cual nos hace suponer que llevar a cabo la predicción de grandes goleadas es una tarea muy compleja utilizando algoritmos de aprendizaje automático.

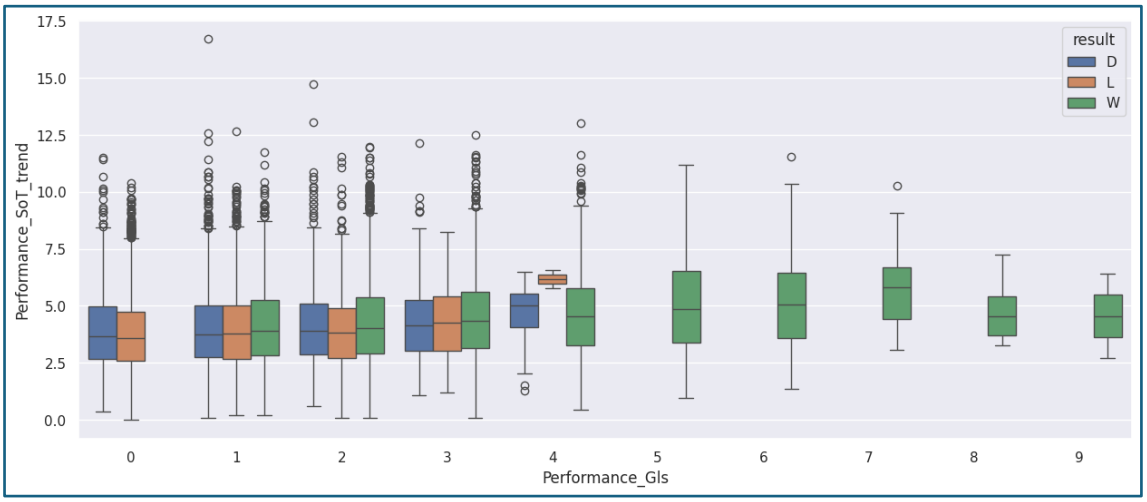

Figura 10. Gráfico de caja y bigotes para goles contra media de tiros en partidos anteriores.

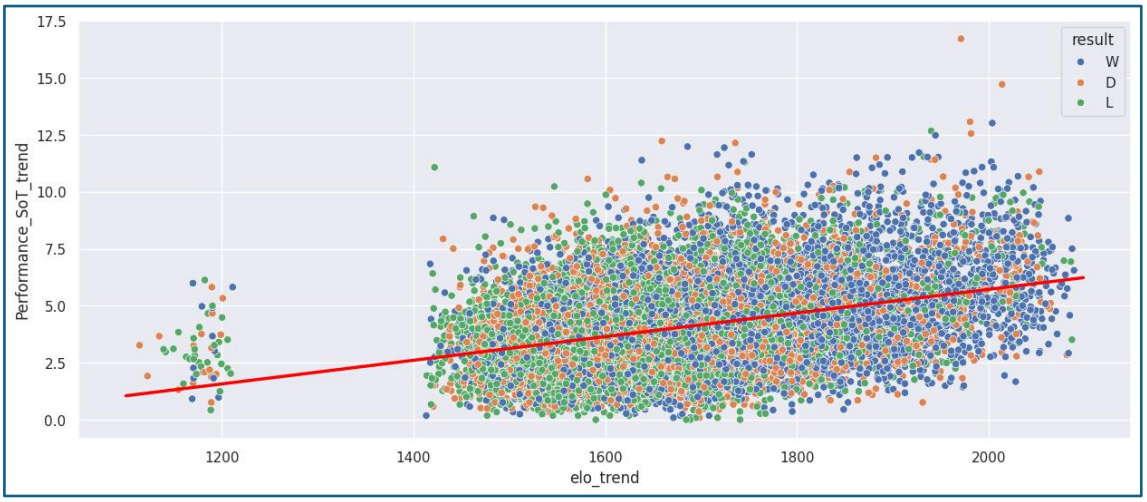

Figura 9. Gráfico de distribución de las medias ponderadas de ELO y remates a portería.

En la Figura 9 se han representado, para los diferentes resultados de un partido, el número de goles marcados frente a la media ponderada de disparos a portería en los anteriores partidos. Se pone en manifiesto que para un número de goles mayor o igual a 5, por lo general se tiene una media ponderada de disparos a puerta superior que para un número de goles menor que 5. Esto puede tener varias lecturas:

1. En primer lugar, podemos volver a citar a (Nesti, 2010) y reforzar sus conclusiones sobre los estados anímicos de los equipos.

2. Sin embargo, también podemos pensar que lo que se observa es fruto de las diferencias existentes entre los equipos que juegan en las grandes ligas europeas. Es decir, podría darse el caso de que los

equipos más fuertes tiendan a hacer más disparos a puerta, por lo que sus medias ponderadas son mayores y, además, marcan más goles.

En la Figura 10 encontramos un gráfico de puntos en el que se muestra la media ponderada del número de remates a puerta en los anteriores partidos frente a la media ponderada del ELO del equipo en esos partidos. Lo que se puede apreciar en este es que, efectivamente se observa una ligera relación lineal entre el nivel del equipo (representado por el ELO) y el promedio de disparos en los anteriores partidos, de manera que los mejores equipos tienden a realizar más disparos. Esto nos indica que trabajar con esta forma de medir el nivel del equipo puede ser de utilidad en nuestra tarea.

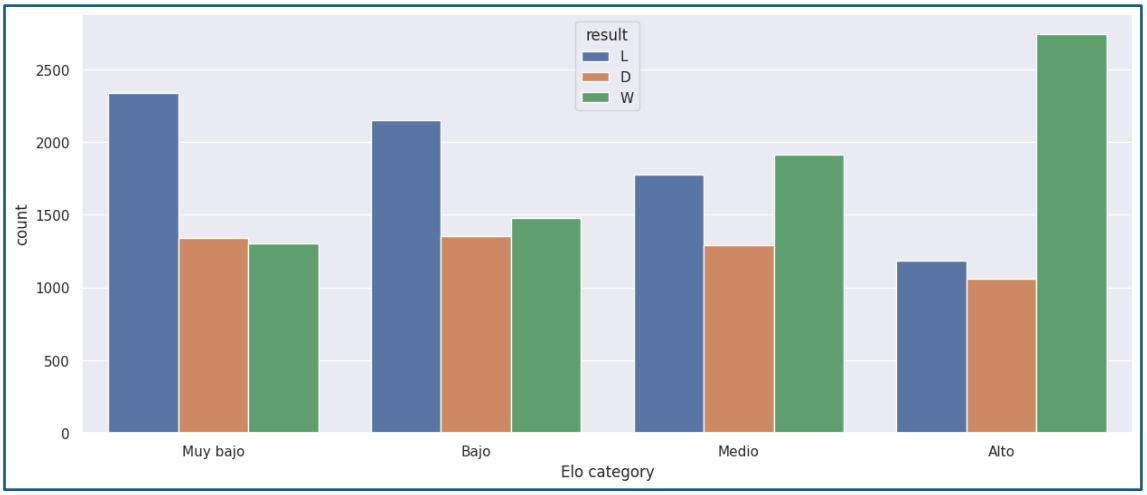

Figura 11. Gráfico de barras para la categoría del ELO frente a el número de victorias, empates y derrotas.

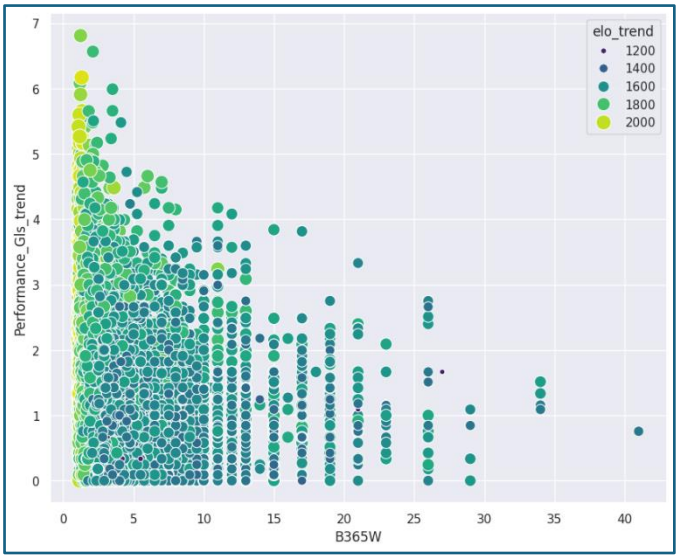

Figura 12. Gráfico de dispersión para la media ponderada de goles contra la cuota de victoria asignada por la casa Bet365.

En la Figura 11 se ha representado un gráfico de barras en el que se muestra el número de victorias, empates y derrotas dependiendo del nivel de elo de los equipos. Para ello, lo que se hizo fue dividir la variable de elo en cuatro categorías utilizando los cuartiles de la variable como puntos de límite. En el resultado obtenido se muestra cómo el número de victorias aumenta significativamente cuando la

categoría del elo es mayor, lo que nos indica que los equipos más fuertes tienden a ganar más partidos. A su misma vez, el número de derrotas disminuye cuando la categoría del elo es mayor. Por último, se observa que la cantidad de empates no muestra cambios significativos para los distintos niveles de elo, hecho que nos refuerza la hipótesis de que esta categoría es la más difícil de clasificar.

En la Figura 12 podemos observar la relación que existe entre la media ponderada de los goles en los anteriores partidos y la cuota asignada por la casa de apuestas Bet365. Se observa una tendencia general: a medida que aumenta la media ponderada de goles la cuota disminuye. Esto es lógico, ya que las casas de apuestas asignan cuotas más bajas a equipos que históricamente han demostrado un mejor rendimiento. Por otra parte, el color de los puntos, que representa el Elo, parece coincidir con la distribución de los puntos en el gráfico. Los equipos con un Elo más alto tienden a tener una media de goles más alta y cuotas más bajas. Esto sugiere que el Elo es un buen indicador del rendimiento general del equipo. Se evidencia también el sesgo que tienen las casas de apuestas para asignar cuotas más bajas a equipos con mayor elo, siendo muy complicado la predicción de una derrota de un equipo grande. En la Tabla 2 encontramos las cuotas promedio asignadas a cada uno de los posibles resultados de un partido para equipos de diferentes niveles de elo. Como se observa, existe una diferenciación clara en las cuotas asignadas para cada equipo.

Equipo	Cuota promedio de victoria	Cuota promedio de empate	Cuota promedio de derrota
Manchester City	1.357714	6.721943	12.360514
Bayern Munich	1.322792	6.869675	11.295195
Sevilla	2.431314	3.683314	4.259086
Dortmund	1.880968	4.737290	5.831097
Celta de Vigo	3.445371	3.646800	3.003771
West Ham	3.813486	3.950914	2.793143
Clermont Foot	4.039286	3.740714	2.334286
Arminia Bielefeld	5.974355	4.375161	1.907419

Tabla 2. Cuotas promedias asignadas por Bet365 para diferentes equipos.

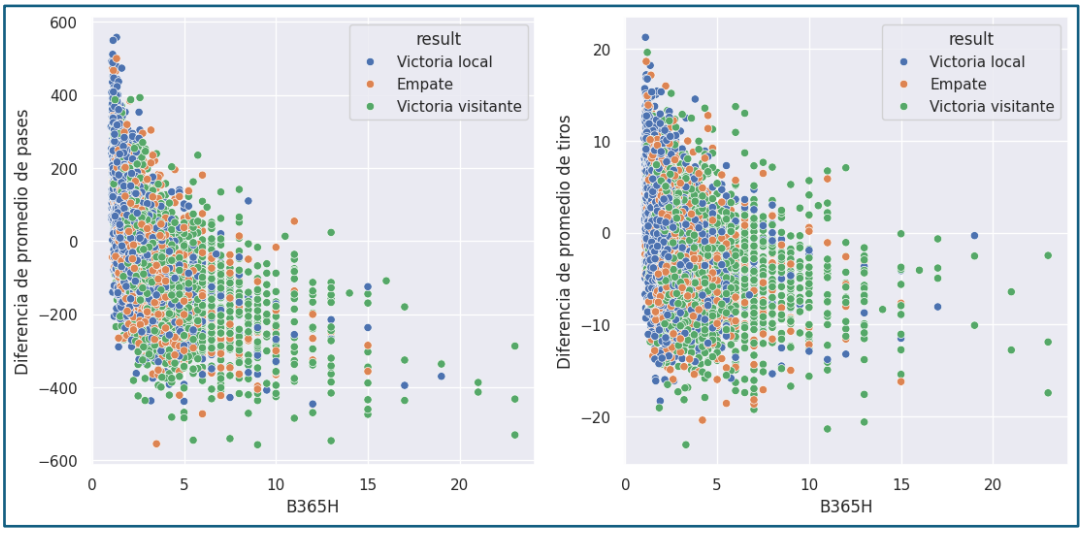

Figura 13. Gráficos de dispersión para la diferencia de medias ponderadas contra las cuotas de victoria.

Por último, en la Figura 13 se observa la representación de dos gráficos de dispersión. En estos se muestra representada la diferencia de las medias ponderadas del equipo local y del equipo visitante, contra las cuotas que asignaba la casa de apuestas Bet365 a la victoria del equipo local.

En el gráfico de la izquierda se observa que a medida que la cuota de Bet365 aumenta, la diferencia de pases tiende a ser más negativa, es decir, los equipos que tienen una mayor cuota suelen dar menos pases que sus rivales. Por otro lado, el gráfico de la derecha muestra que las diferencias en tiros tienden a ser más pequeñas, pero también se vuelve más negativa conforme la cuota aumenta. Sin embargo, la dispersión es menor en comparación con los pases. Estos gráficos indican que las estadísticas de juego como pases y tiros, junto con las cuotas de apuestas, pueden ser buenos indicadores del resultado del partido. Los equipos que mantienen la posesión y generan más pases o tiros tienen una mayor probabilidad de ganar, especialmente cuando las cuotas de las apuestas sugieren que son los favoritos.

En esta sección se han analizado algunas de las variables contenidas en nuestro conjunto de datos, como el número de goles, el elo de los equipos, las cuotas de casas de apuestas o las medias ponderadas de pases, remates a puerta y goles. Esto nos ha permitido tener una visión general del comportamiento de las variables que nos es de utilidad, junto con el siguiente apartado, para llevar a cabo la selección de características. Así, por ejemplo, algunas de las conclusiones que se obtienen de este análisis exploratorio del dato es la posibilidad de que nuestras variables tengan distribuciones no normales, que podamos encontrar problemas a la hora de predecir los empates de un partido de fútbol o que la aproximación que se basa en predecir el número de goles de cada equipo puede mostrar limitaciones a la hora de predecir resultados abultados.

5.5 Reducción de la dimensionalidad

La maldición de la dimensionalidad (*Curse of dimensionality*) es un término descrito por Richard Bellman (Bellman, 1957) que se refiere a la dificultad de encontrar una estructura latente, capaz de resolver el problema que nos aborde, debido a la gran cantidad de variables contenidas en un conjunto de datos. Las causas principales de que no nos sea fácil encontrar la solución con tantas variables es la dificultad de la interpretabilidad del conjunto de datos y la posibilidad de encontrar información redundante. Por ello, una de las preguntas que nos surgió una vez limpiados nuestros conjuntos de datos fue: ¿Son necesarias tantas variables para ser capaces de predecir el resultado de un partido de fútbol? Con el fin de resolver esta pregunta, se llevó a cabo un análisis detallado de los datos empleando diversas técnicas. Es de importancia recalcar que este análisis se llevó sobre las variables referidas a las estadísticas de los partidos.

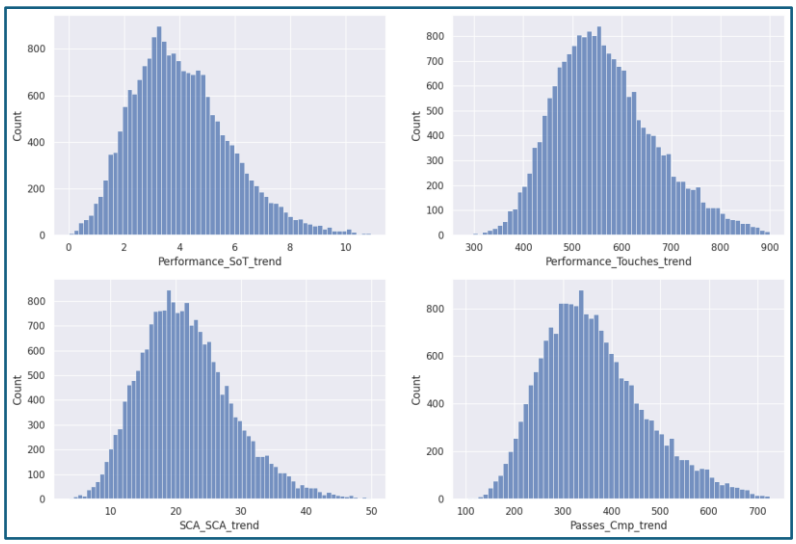

Figura 14. Histogramas para algunas variables con posibilidad de ser estadísticamente significativas.

En la Figura 14 observamos las distribuciones de otras cuatro variables, en concreto de las medias ponderadas de remates a portería, toques del balón, jugadas acabadas en gol y pases cortos. Se aprecia que, basándonos únicamente en los histogramas, podemos sospechar que las distribuciones podrían ser aproximadamente normales. Sin embargo, para una conclusión más sólida y confiable, es necesario realizar pruebas estadísticas formales de normalidad.

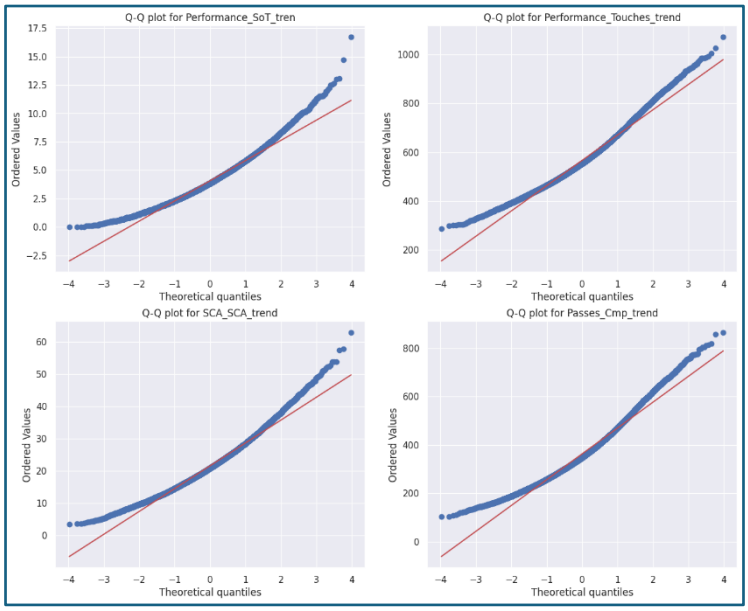

Figura 15. QQ-plots para las cuatro variables.

En la Figura 15 se muestran los QQ-plots para las 4 variables mencionadas anteriormente. Un QQ-plot es una herramienta gráfica que compara la distribución de un conjunto de datos con una distribución teórica, en este caso, una distribución normal. La idea es que, si los datos siguen aproximadamente la distribución teórica, los puntos del gráfico se alinearán a lo largo de una línea recta. Observando los gráficos podemos apreciar que:

1. En primer lugar, los puntos no se alinean perfectamente con la recta, lo que indica que la distribución de datos no es exactamente normal.

2. En segundo lugar, la curva es convexa, lo que nos sugiere que la distribución de los datos tiene colas más pesadas de lo que se esperaría en una distribución normal. Es decir, hay más valores extremos de lo que cabría esperar si los datos fueran normales.

En este caso, el conocer la no normalidad de nuestros datos es un paso clave a la hora de llevar a cabo nuestros análisis para la selección de variables, ya que algunas pruebas como el análisis de la varianza o el coeficiente de correlación de Pearson asumen normalidad en los datos. Por ello, nosotros aplicamos otra serie de pruebas que se adapten a nuestros datos.

Antes de mostrar las pruebas realizadas, es necesario matizar que en este proyecto se ha trabajado con el conjunto de datos transformado de dos maneras diferentes, aunque en ambos casos se haya llevado a cabo una clasificación:

- Por un lado, una de las formas con las que se ha trabajado ha sido utilizar una única observación por partido. En esta observación encontramos las medias ponderadas del equipo local, las medias ponderadas del equipo visitante, los elos de cada uno de los equipos y las cuotas asignadas a victoria del local, empate y victoria del visitante. En esta forma de trabajar con el conjunto de datos, nuestra variable objetivo es una variable categórica con 3 clases (victoria del equipo local, empate o victoria del equipo visitante), por lo que se lleva a cabo una clasificación simple.

- La otra forma de trabajar con los datos ha sido tener dos observaciones por partido: una para el equipo local y otra para el equipo visitante. En cada una de esas observaciones se tenía las medias ponderadas de las estadísticas de ese equipo, el elo del equipo junto con la diferencia de elo respecto al rival, y las cuotas que le asignaba cada casa de apuestas a la victoria, el empate y la derrota del equipo. En esta forma de trabajar, nuestra variable objetivo es el número de goles que va a marcar el equipo en el siguiente partido. La clasificación en victoria del local, empate o victoria del visitante se hace a posteriori utilizando estas predicciones del número de goles y un determinado valor para establecer los límites entre las categorías.

Ahora sí, podemos mostrar las pruebas realizadas para la selección de variables.

5.5.1 Prueba de Kruskal-Wallis

La prueba de Kruskal-Wallis es una prueba no paramétrica (esto es, cuando se quiere comparar poblaciones cuyas distribuciones no son normales), análoga a la prueba paramétrica ANOVA (Soto, 2013). Las hipótesis para la prueba de Kruskal-Wallis son:

- H0: Las medias o medianas de los n grupos son todas iguales.

- H1: Existe diferencia en al menos uno de los grupos.

Así, la prueba de Kruskal-Wallis proporciona información sobre la posible igualdad de medias o medianas entre grupos (3 o más) y permite rechazar esta hipótesis de igualdad cuando el valor de p sea menor de 0'05. En nuestro caso, realizamos esta prueba para cada una de las variables de nuestro conjunto de datos, dividiendo la variable en 3 grupos: Victoria del local, empate o victoria del visitante. De esta manera, si la prueba de Kruskal-Wallis devuelve un p-valor inferior a 0'05, podemos decir que sujetos a esta prueba esa variable es significativa para nuestro análisis. Como es evidente, en esta prueba estamos trabajando con el conjunto de datos que tiene una única observación por partido, por lo que para cada estadística tenemos dos variables: la del equipo local y la del equipo visitante. Es decir, para un partido contamos con media ponderada de los goles en los partidos anteriores del equipo local y con la media ponderada de los goles en los partidos anteriores del equipo visitante. En este caso, entenderemos que una variable no es estadísticamente significativa cuando su p-valor es superior a 0'05 tanto cuando se mide para el equipo local como cuando se mide para el equipo visitante.

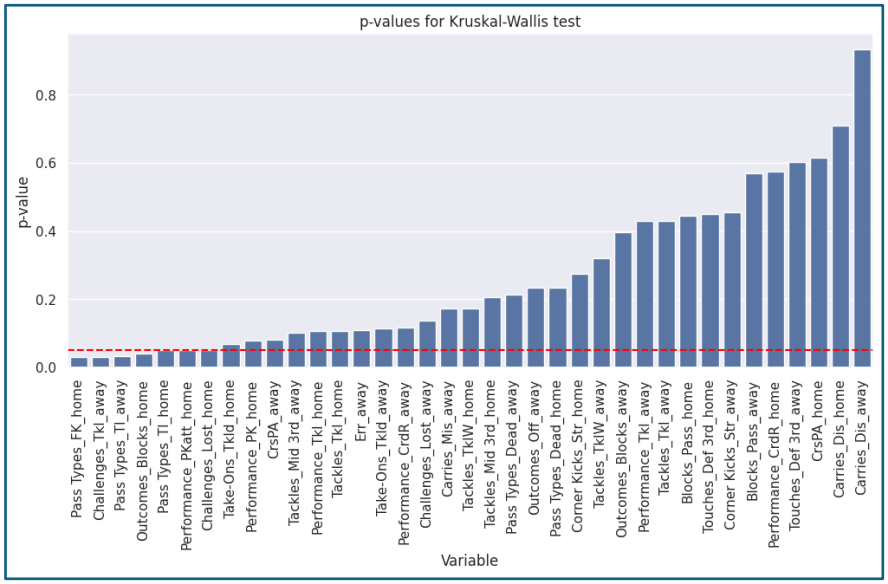

Figura 16. p-valores obtenidos en la prueba de Kruskal-Wallis.

En la Figura 16 se muestran las variables que han obtenido los mayores p-valores en la prueba de Kruskal-Wallis. Como se aprecia, existen bastantes variables que han obtenido un valor por encima del 0'05 nivel de significación. En la Tabla 3 se muestra la lista de variables que han obtenido un p-valor superior a 0'05 tanto cuando se miden para el equipo local como cuando se miden para el equipo visitante.

Variables no significativas tanto para el equipo local como para el visitante
Media de pases bloqueados
Media de distancia recorrida con balón
Media de duelos perdidos
Media de número de córneres sacados
Media de centros en el área rival
Media de pases con balón parado
Media de tarjetas rojas
Media de entradas realizadas
Media de entradas realizadas en el campo rival
Media de entradas realizadas con éxito
Media de entradas realizadas en campo propio

Tabla 3. Variables no significativas según la prueba de Kruskal-Wallis.

Si analizamos las variables, podemos llegar a encontrar sentido en que estas no tengan ningún tipo de relación con el resultado de los partidos, pues no son variables que puedan tener algún tipo de relación con la forma/estado anímico en el que se encuentra el equipo. También es posible que se encuentre que estas variables no son significativas debido a que algunas son variables con muy poca variación, por ejemplo, la media de tarjetas rojas en los anteriores partidos va a ser por lo general 0. En la siguiente Tabla 4 se muestra el valor medio de algunas de las variables.

Variable	Victoria local	Empate	Victoria visitante
Media de córneres sacados por el equipo local	0.074925	0.059863	0.064367
Media de tarjetas rojas del equipo local	0.099352	0.094959	0.095126
Media de distancia recorrida por el equipo visitante	9.125987	9.144316	9.138800
Media de entradas del equipo visitante	6.490818	6.473880	6.388767

Tabla 4. Medias obtenidas para dos variables del equipo local y dos variables del equipo visitante (ambas no significativas en la prueba Kruskal-Wallis)

Como se observa, los resultados obtenidos muestran la veracidad de nuestra hipótesis anterior, pues para las dos variables relacionadas con equipos locales se obtienen medias muy próximas a 0. Por otro lado, para las variables relacionadas con los equipos visitantes se observan valores muy similares, lo que indica poca variación entre grupos (ver Anexo para los resultados de la tabla completa).

5.5.2 Estudio de las correlaciones y clustering de variables

Tener muchas variables altamente correlacionadas en un modelo de aprendizaje automático puede generar problemas significativos y afectar la calidad de los resultados. Por ello, se realizó un estudio de las correlaciones de las variables para tratar de solucionar este problema. Para esta tarea, se utilizó el coeficiente de correlación de Spearman debido a la no normalidad de nuestros datos. En la Figura 17 se observa el mapa de calor para las correlaciones de las variables de nuestro conjunto de datos.

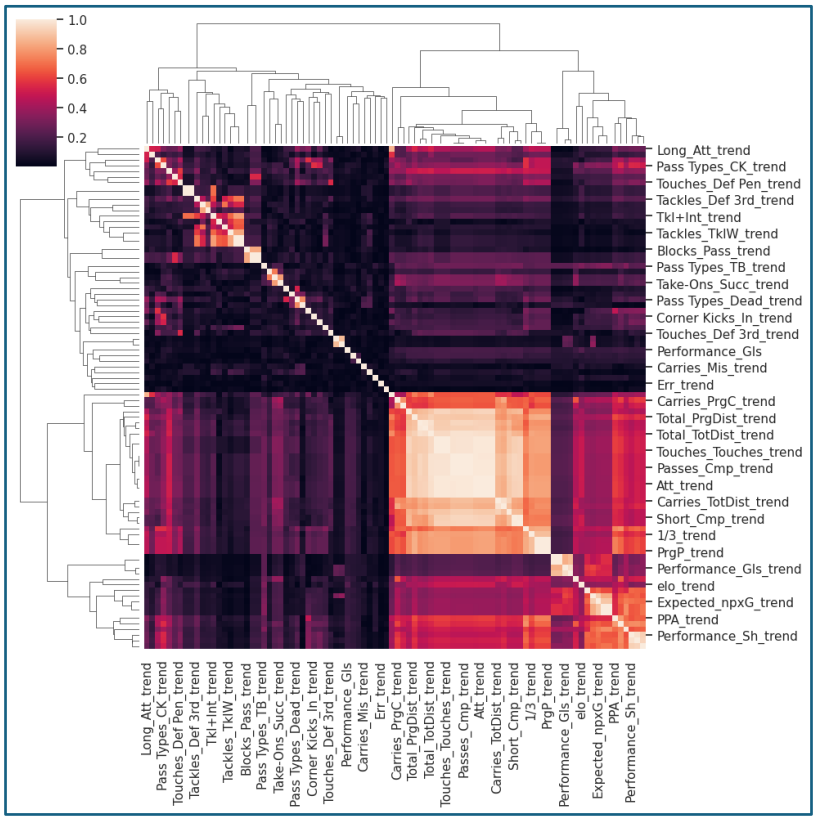

Figura 17. Mapa de calor para las correlaciones de las variables del conjunto de datos.

Como se puede apreciar, existe un gran grupo de variables altamente correlacionadas. Para tratar de solucionar este problema, se planteó hacer un clustering de variables, intentando comprender cuales eran las variables que formaban este grupo y porqué existe esta correlación. Para llevar a cabo este clustering se utilizó el algoritmo HDBSCAN, ya que este no asume ninguna distribución de los datos y puede identificar clústeres de forma arbitraria. En la Tabla 5 se muestran los clústeres que contenían más de una variable.

Clúster	N.º Variables
2	13
0	6
5	5
3	4
8	4
9	4
6	4
4	3
10	3
1	3
7	3

Tabla 5. Cantidad de variables por clúster.

Como se puede apreciar, encontramos un total de 11 clústeres en los que se agrupan las variables. El que destaca por encima del resto es el clúster número 2, que contiene un total de 13 variables. Para entender estas agrupaciones, se llevó un análisis de cada uno de los clústeres.

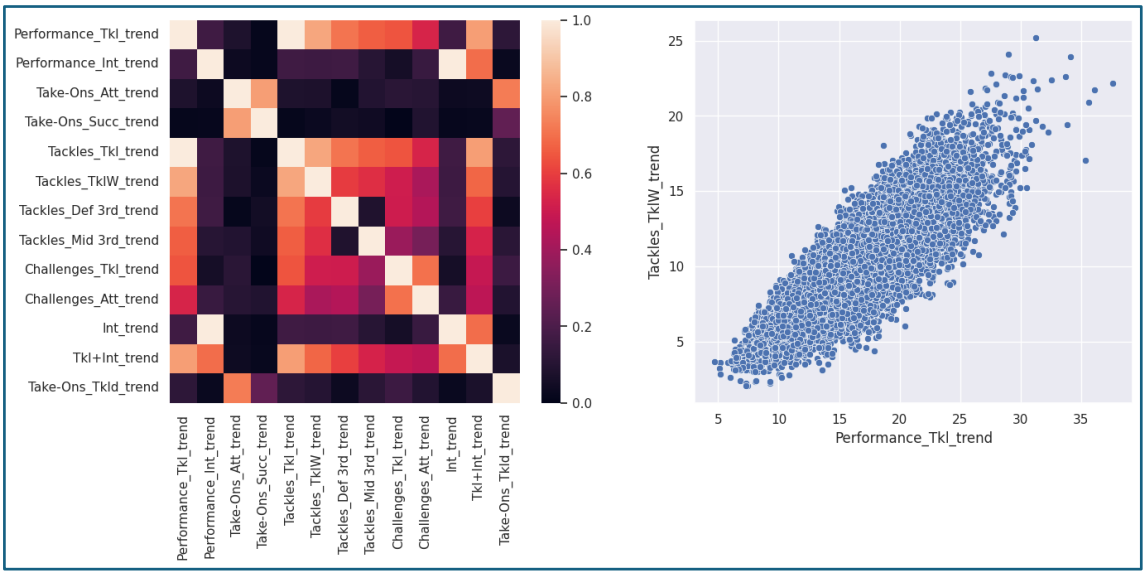

Figura 18. Correlaciones de las variables del clúster 2.

En la Figura 18 se pueden ver el mapa de calor de las correlaciones de las variables del segundo clúster. Se aprecia perfectamente que existen una serie de variables muy correlacionadas, las cuales la mayoría hacen referencia a las entradas que hace un equipo en un partido. Estas variables son la media de entradas realizadas, la media de entradas perdidas, la media de entradas ganadas, la media de entradas en campo propio y la media de entradas en campo rival. Como es lógico, estas variables están referidas a una faceta del juego y, por lo tanto, van a estar correlacionadas positivamente. Dicho de otra manera, cuantas más entradas haga un equipo, más entradas exitosas hará, más entradas en el centro del campo hará y más entradas en zona defensiva hará. Por lo tanto, podemos considerar que estas variables pueden representarse haciendo uso de una única variable: la media de entradas realizadas.

En la Figura 19 están representados los mapas de calor para las correlaciones de otros cuatro clústeres. Como se puede observar, las correlaciones entre las variables de un mismo clúster son prácticamente perfectas, lo que indica una redundancia significativa en los datos. Es necesario tomar medidas para reducir esta redundancia antes de proceder con cualquier modelo de aprendizaje automático. Para ello, se decidió seleccionar una serie de variables y eliminar el resto, ya que como muchas de estas representaban la misma información, con mantener una era suficiente. Las variables seleccionadas se explican en el siguiente apartado.

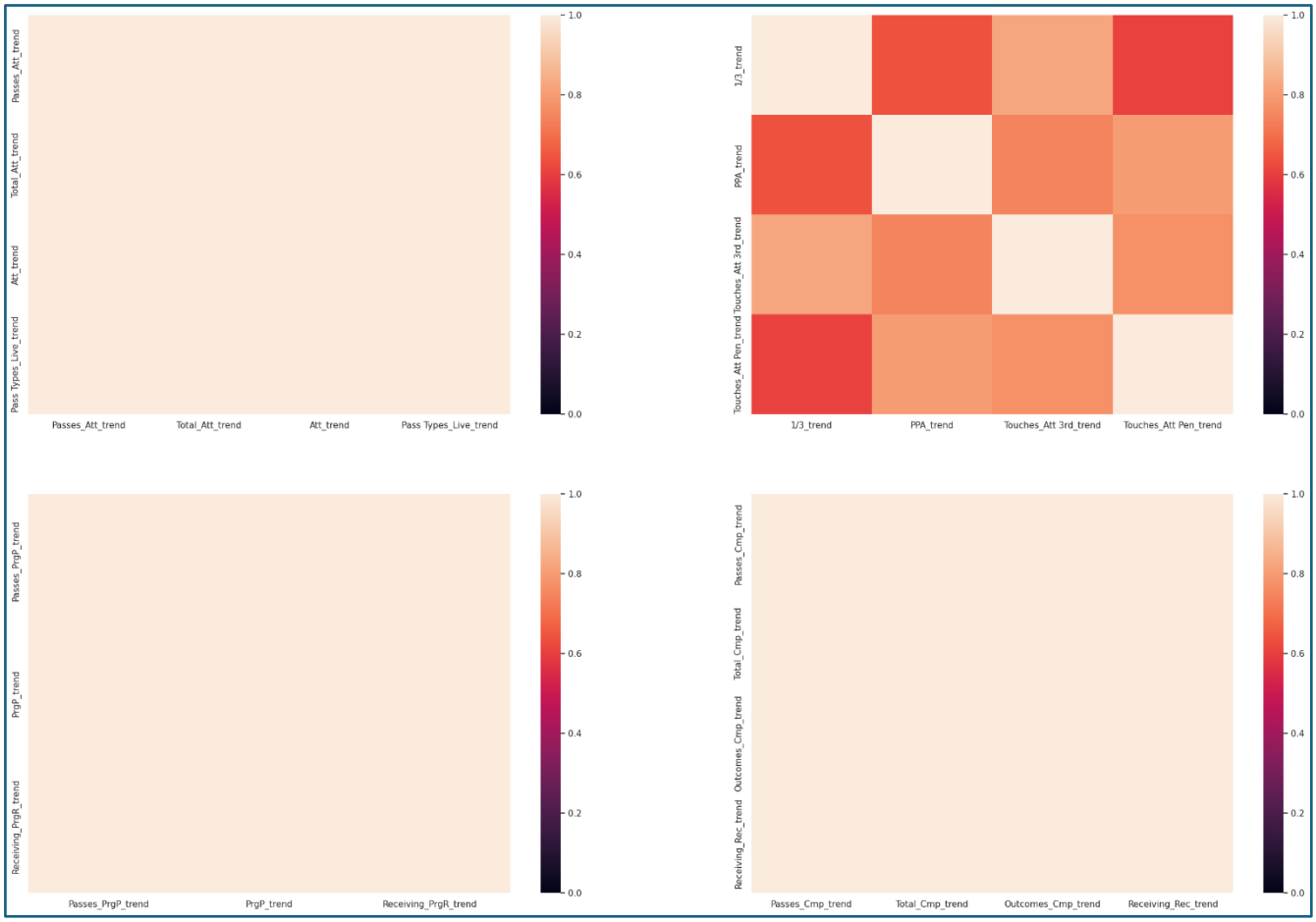

Figura 19. Mapas de calor para los clústeres 6,7,8 y 9.

5.5.3 Selección de características

Como resultado de los análisis y pruebas realizadas, se llevó a cabo la siguiente selección de variables:

- Las variables mencionadas en la Tabla 3 se eliminan del conjunto de datos debido a que no muestran ninguna significación en la prueba de Kruskal-Wallis, tanto para el equipo local como para el equipo visitante. Además, los coeficientes de correlación de estas variables con el número de goles que marca el equipo en el siguiente partido son muy próximos a 0, por lo que determinamos que estas no aportan información para ninguno de los dos abordajes con los que se trata de resolver el problema.

- La variable media de penaltis marcados se elimina del conjunto de datos ya que su información viene representada en la variable media de penaltis tirados.
- La variable media de pases bloqueados se elimina del conjunto de datos ya que su información viene representada en la variable media de bloqueos.
- Las variables de la media de goles esperados excluyendo penaltis y la media de goles-asistencias esperados son eliminadas del conjunto de datos ya que su información queda recogida en las variables de media de goles esperados y media de asistencias esperadas.
- La variable media de pases progresivos recibidos se elimina del conjunto de datos ya que su información viene contenida en la variable media de pases progresivos dados.
- Las variables de media de pases a media distancia con éxito, media de pases recibidos y media de pases a media distancia fallados se eliminan del conjunto de datos ya que su información queda recogida en la variable media de pases a media distancia intentados.
- Las variables de media de pases con el balón en movimiento y media de pases intentados son eliminadas del conjunto de datos ya que en la variable media de pases realizados se encuentra su información.
- Las variables de media de toques progresivos y media de toques con el balón en movimiento quedan fuera del conjunto de datos ya que su información es muy similar a la de la variable media de toques realizados.

En esta sección se han llevado a cabo una serie de pruebas y análisis con el objetivo de optimizar la información de nuestro conjunto de datos en un número de variables reducido. Todas las variables resultantes se pueden ver en Anexo II.

5.6 Modelos empleados y preparación de los datos

Para tratar de resolver nuestra tarea de predicción de resultados existen dos abordajes posibles. En esta sección se explica detalladamente ambos abordajes y la forma en la que se preparan los datos para cada uno de ellos.

5.6.1 Abordaje 1: LightGBM

El primer abordaje propuesto es emplear el algoritmo de LightGBM para llevar una clasificación multiclase en tres clases: victoria del equipo local, empate o victoria del equipo visitante. Para ello, se agrupan los datos del equipo local y visitante en una misma observación, quedando un conjunto de datos con la forma que se muestra en la Tabla 6.

Jornada	Partido	Estadísticas equipo local	Elo equipo local	Estadísticas equipo visitante	Elo equipo visitante	Cuotas	Resultado
4	Arsenal-Liverpool	Medias ponderadas del Arsenal de las jornadas 1, 2 y 3	Media del elo del Arsenal de las jornadas 1, 2 y 3	Medias ponderadas del Liverpool de las jornadas 1, 2 y 3	Media del elo del Liverpool de las jornadas 1, 2 y 3	Cuotas de las casas de apuestas	Resultado de ese partido
5	Everton-Arsenal	Medias ponderadas del Everton de las jornadas 2, 3 y 4	Media del elo del Everton de las jornadas 2, 3 y 4	Medias ponderadas del Arsenal de las jornadas 2, 3 y 4	Media del elo del Arsenal de las jornadas 2, 3 y 4	Cuotas de las casas de apuestas	Resultado de ese partido

Tabla 6. Forma del conjunto de datos para la predicción con LightGBM.

De este conjunto de datos resultante, se utilizan las estadísticas del equipo local, el elo del equipo local, las estadísticas del equipo visitante, el elo del equipo visitante y las cuotas de las casas de apuestas para predecir nuestra variable objetivo: el resultado del partido. Esta variable objetivo es una variable categórica con 3 clases: victoria del local, empate o victoria del visitante.

Para llevar a cabo la división de los datos en conjuntos de entrenamiento, validación y prueba, se aplica la función de `train_test_split` de la librería `scikit-learn` de Python, utilizando un 70% del conjunto de datos para entrenamiento, un 20% para validación y un 10% para prueba.

5.6.2 Abordaje 2: Redes neuronales LSTM

De otro lado, el segundo abordaje posible es llevar a cabo una clasificación de los partidos en tres clases (victoria local, empate, victoria visitante) pero desde un punto de vista diferente. En este caso no se trata de predecir qué va a ocurrir, sino de llevar a cabo la predicción de cuántos goles va a marcar cada equipo en el próximo partido.

A partir de esta predicción del número de goles se puede llevar a cabo la clasificación utilizando un determinado *threshold* δ para que sea posible considerar que existan empates. De esta manera, la predicción del resultado de un determinado partido seguiría la siguiente lógica:

$$Resultado(partido) = \begin{cases} Victoria\ local & si \quad \widehat{Goles\ local} - \widehat{Goles\ visitante} > \delta \\ Empate & si \quad -\delta \leq \widehat{Goles\ local} - \widehat{Goles\ visitante} \leq \delta \\ Victoria\ visitante & si \quad \widehat{Goles\ local} - \widehat{Goles\ visitante} < -\delta \end{cases}$$

Para llevar a cabo este planteamiento, tendremos dos observaciones por partido: una para el equipo local y otra para el equipo visitante. De esta manera, nuestro conjunto de datos tiene la forma que se muestra en la Tabla 7.

Jornada	Equipo	Partido	Estadísticas	Elo	Cuotas	Goles marcados
4	Arsenal	Arsenal-Liverpool	Medias ponderadas del Arsenal de las jornadas 1, 2 y 3	Media del elo del Arsenal de las jornadas 1, 2 y 3	Cuotas de las casas de apuestas	Goles marcados por el Arsenal en el partido
4	Liverpool	Arsenal-Liverpool	Medias ponderadas del Liverpool de las jornadas 1, 2 y 3	Media del elo del Liverpool de las jornadas 1, 2 y 3	Cuotas de las casas de apuestas	Goles marcados por el Liverpool en el partido

Tabla 7. Forma del conjunto de datos para la predicción del número de goles.

Este conjunto de datos se agrupa por equipos, ya que la predicción del número de goles se lleva a cabo como una serie temporal. Además, dentro de las estadísticas del elo, también se tiene en cuenta de diferencia entre el elo del equipo con el equipo rival, ya que es necesario tener en cuenta que dependiendo del rival el número de goles varia. Es decir, no es igual marcar dos goles a un equipo con un elo alto, como podrían ser el Manchester City o el Real Madrid, que marcar dos goles a otro equipo con un elo más bajo.

Además, la preparación del conjunto de datos tiene otra peculiaridad respecto a la forma de dividir los datos, pues al trabajar con series temporales es importante la dimensión temporal del dato. Así pues, en este caso se emplea 80% de las observaciones para entrenamiento y el 20% para prueba, seleccionadas en orden temporal, por equipo y temporada.

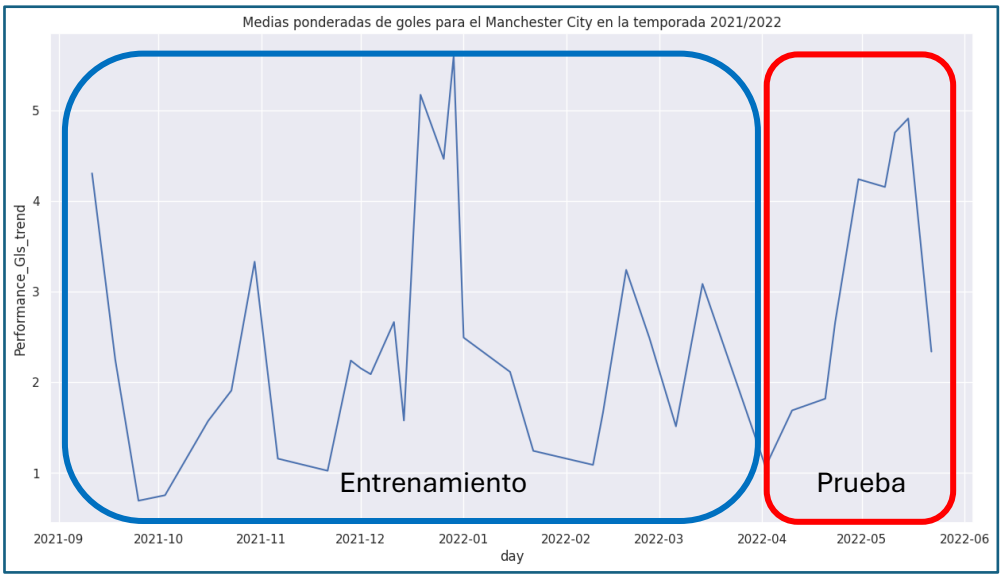

Figura 20. División de los datos para entrenamiento y prueba en el segundo abordaje.

6. Experimentos

Este capítulo describe los modelos implementados y los experimentos realizados para entrenarlos y evaluarlos. Primero, se explicarán los experimentos de referencia y de línea base. Luego, se detallará el proceso iterativo con el cual se han llevado a cabo diferentes experimentos para encontrar un modelo de predicción y un conjunto de datos óptimos.

6.1 Puntos de referencia para modelos de predicción

Dentro del aprendizaje automático, llamamos puntos de referencia a una serie de modelos muy sencillos de implementar, que en la mayoría de los casos no requieren de entrenamiento y que nos sirven para hacernos una idea de cómo de bien funcionan los modelos que hemos implementado. Es decir, si uno de estos modelos que no requieren entrenamiento es capaz de hacer predicciones iguales o mejores que las de nuestro modelo de aprendizaje automático, entonces es que algo no estamos haciendo bien.

Para comparar nuestros modelos, se han desarrollado dos modelos sencillos como puntos de referencia:

- **Modelo aleatorio**: Este modelo, como su propio nombre indica, hace predicciones aleatorias del resultado del partido. No tiene en cuenta ninguna de las variables de nuestro conjunto de datos, simplemente se limita a seleccionar para cada partido uno de los tres posibles resultados. Para que las predicciones sean más "realistas", se utiliza el número de victorias del equipo local, empates y victorias del equipo visitante para establecer una probabilidad de elección.
- **Modelo del resultado más frecuente**: Este modelo hace predicciones en función del resultado más frecuente que se dé. Así, por ejemplo, en nuestro caso el resultado más frecuente del conjunto de datos es la victoria del equipo local, por lo que nuestro modelo hará que todas las predicciones sean victoria del equipo local. Este modelo tiene algunos inconvenientes dependiendo el tipo de problema, ya que para conjuntos de datos desbalanceados puede ofrecer un porcentaje de acierto muy alto siendo un mal modelo. Sin embargo, para la tarea que nos ocupa es un buen punto de referencia.

En la Figura 21 se observa la matriz de confusión del modelo aleatorio, en la que 0 es la etiqueta asignada a las victorias del equipo local, 1 es la etiqueta asignada al empate y 2 es la etiqueta asignada a la victoria del equipo visitante. Como se puede apreciar, las predicciones de este modelo son malas. El modelo tiene dificultades para clasificar correctamente las instancias en sus respectivas categorías. Se observa una cantidad significativa de errores de clasificación, ya que ningún valor en la diagonal (donde los valores correctos deberían estar) es dominante en comparación con los valores fuera de la diagonal. Además, cabe recalcar que las clases en las que más confunde el modelo aleatorio son las dos clases dominantes (victoria del local y

victoria del visitante), cuando en la práctica es mejor confundir una victoria del equipo local con un empate que con una victoria del equipo visitante, ya que esto nos indica que nos hemos quedado "más cerca" de acertar.

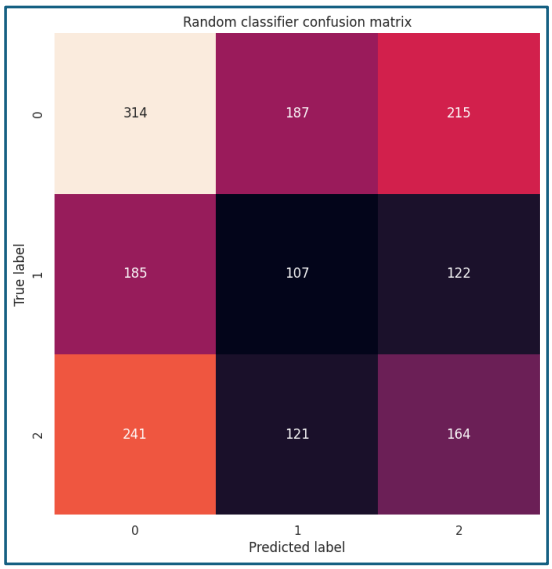

Figura 21. Matriz de confusión para el clasificador aleatorio.

Por otro lado, en la Tabla 8 se observa el informe de clasificación del modelo del resultado más frecuente. Como se ver, los resultados del modelo son malos, la exactitud del modelo es de 0.43, lo que significa que el 43% de las predicciones son correctas. Sin embargo, dado el desequilibrio en la clasificación (donde solo la victoria del equipo local es correctamente clasificada), esta métrica puede ser engañosa, ya que no refleja el pobre rendimiento en las otras clases.

		Precisión	Sensibilidad	F1-Score	Soporte
Clases	**Victoria local**	0.43	1.00	0.60	716
	Empate	0.00	0.00	0.00	414
	Victoria visitante	0.00	0.00	0.00	526
Métrica	**Exactitud**			0.43	1656
	Macro Average	0.14	0.33	0.20	1656
	Weighted Average	0.19	0.43	0.26	1656

Tabla 8. Informe de clasificación del modelo de resultado más frecuente.

6.2 Experimentos con redes neuronales LSTM

El primer algoritmo que utilizaremos para nuestras predicciones es la red neuronal Long-Short Term Memory. El objetivo de esta sección es encontrar el modelo de redes neuronales LSTM que mejor predice los resultados de los partidos, utilizando diferente combinación de hiperparámetros y arquitectura de red.

La red neuronal recibirá como entrada las medias ponderadas de los tres partidos anteriores de un equipo, las cuotas de las casas de apuestas, el promedio del elo en los partidos anteriores y la diferencia de elo estimada con el rival en los anteriores partidos. Con estas entradas, la red devolverá un valor para la predicción del número de goles en el siguiente partido de ese equipo. Con este número de goles

predicho, se hará la comparación con el número de goles predichos para el equipo rival y se hará una clasificación empleando un valor límite para considerar los empates.

Respecto a la diferencia de elo estimada, cabe mencionar la forma en la que se calcula esta. Como bien sabemos, a la hora de hacer predicciones no conocemos el valor de la diferencia de elo exacto. Sin embargo, conocemos los valores del elo en las jornadas anteriores y sabemos que este no es un valor que varíe mucho, por lo que utilizar la media ponderada del elo en las jornadas anteriores para calcular esta diferencia de elo es una práctica que puede ser eficiente (Hvattuma & Arntzen, 2010).

Uno de los hiperparámetros que tendremos como fijos y sobre el que no se harán experimentos es la función de activación de la capa de salida, para la cual se utilizará la Rectified Linear Unit (ReLu). Esta elección se debe a que, desde el punto de vista más realista, si lo que trata de predecir la red neuronal es el número de goles de un equipo, no tiene sentido que se predigan valores negativos, en todo caso el valor mínimo que se pueden predecir son 0 goles.

Por otro lado, para la evaluación de estos modelos se utilizarán diferentes métricas. Para la evaluación de la propia red neuronal se utilizará la función de pérdida del error medio cuadrático, ya que nos interesa obtener la red que mejor sea capaz de predecir el número de goles. Para la selección del mejor valor límite para la clasificación una vez predichos el número de goles utilizaremos otras métricas como la exactitud, el f1-score, la matriz de confusión o el informe de clasificación. Para que la evaluación sea justa, se define una semilla inicial a la hora de comenzar el entrenamiento de los diferentes modelos, de manera que los pesos de la red neuronal sean siempre los mismos al inicio del entrenamiento y los resultados de la red no puedan verse afectados por la aleatoriedad del proceso.

En los siguientes apartados se explican cada uno de los experimentos realizados y las conclusiones obtenidas.

6.2.1 Red LSTM básica

En primera instancia, se trato de trabajar con una red neuronal LSTM con todos los hiperparámetros por defecto para ver qué resultados se podían obtener y qué margen de mejora existía al respecto. Los hiperparámetros empleados fueron: una tasa de aprendizaje de 10^{-4}, un optimizador ADAM, una función de activación lineal y un umbral para la clasificación de 0’5 (suponemos que con una diferencia de 0’5 goles en la predicción del número de goles podemos afirmar que un equipo saldrá vencedor). En lo relacionado a la arquitectura de la red, en la siguiente tabla se hace una descripción detallada de esta:

Capa (Tipo)	Forma de Salida	Número de Parámetros
LSTM	(None, 3, 256)	345,088
LSTM	(None, 3, 128)	197,12
BATCH NORMALIZATION	(None, 3, 128)	512
LSTM	(None, 3, 64)	49,408
BATCH NORMALIZATION	(None, 3, 64)	256
LSTM	(None, 32)	12,416
BATCH NORMALIZATION	(None, 32)	128
DENSA	(None, 1)	33

Tabla 9. Estructura de la red neuronal simple.

En la Figura 22 se muestra la evolución de la función de pérdida para el conjunto de entrenamiento y de validación. La pérdida en el conjunto de entrenamiento disminuye constantemente a lo largo de las épocas, lo que indica que el modelo está aprendiendo cada vez mejor a ajustar los datos de entrenamiento. Sin embargo, la pérdida en el conjunto de validación comienza a aumentar significativamente a partir de aproximadamente la época 15. Esto es una clara señal de sobreajuste. El modelo está memorizando el conjunto de entrenamiento en lugar de generalizar a nuevos datos.

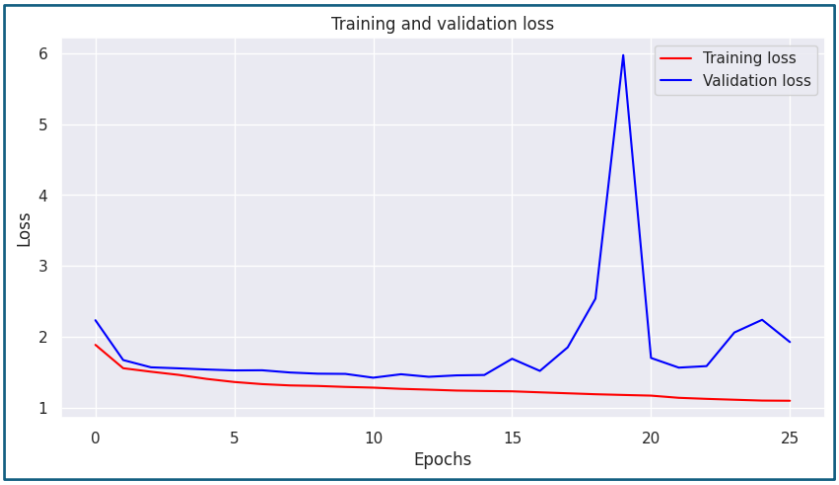

Figura 22. Evolución de la función de pérdida para la red neuronal simple.

En la tabla 10 podemos ver el informe de clasificación para este modelo con un umbral de 0’5. De esta tabla se pueden obtener varias conclusiones. En primer lugar, se observa que la precisión de las victorias (tanto locales como visitantes) es superior a 0’6 en ambos casos, pero la sensibilidad es muy baja (inferior 0’32 en ambos casos), lo que sugiere que el modelo es bueno para identificar correctamente los ejemplos de estas clases cuando los predice, pero omite muchos ejemplos positivos de esta clase. Es decir, cuando el modelo hace una predicción de una victoria, ya sea local o visitante, es probable que esta sea correcta.

Por otro lado, fijándonos en la sensibilidad de la clase de los empates podemos apreciar que esta tiene un valor de 0’8. Sin embargo, en la Figura 23 podemos observar la matriz de confusión para este modelo y ver cómo la gran mayoría de predicciones de este modelo corresponden a la clase de empate, por lo que, aunque se tenga un valor alto de sensibilidad, la precisión de esta clase es muy baja. Aproximadamente el 70% de los valores que clasifica este modelo como

empate son errores de clasificación. Para terminar, se ve que este modelo tiene un 0’41 de exactitud, por lo que vemos que se debe hacer una mejor selección de hiper parámetros ya que el modelo no mejora nuestro modelo de referencia de las clases mayoritarias.

		Precisión	Sensibilidad	F1-Score	Soporte
Clases	**Victoria local**	0.73	0.32	0.45	390
	Empate	0.29	0.80	0.42	220
	Victoria visitante	0.62	0.22	0.33	270
Métrica	**Exactitud**			0.41	880
	Macro Average	0.55	0.45	0.40	880
	Weighted Average	0.59	0.41	0.40	880

Tabla 10. Informe de clasificación para la red neuronal simple.

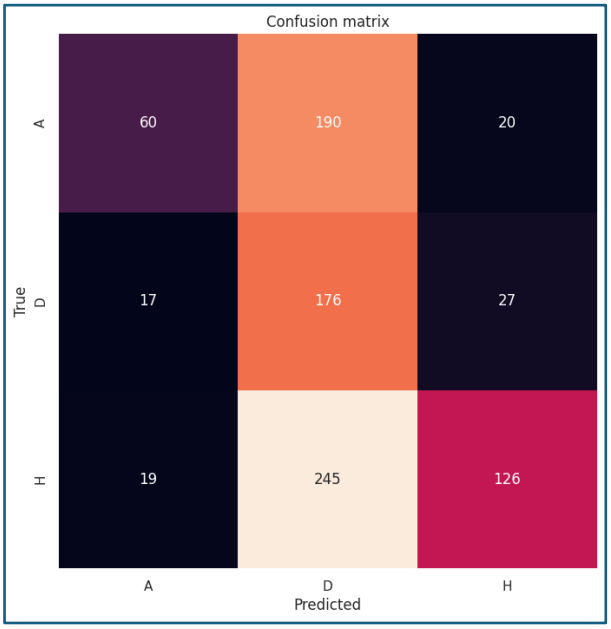

Figura 23. Matriz de confusión para la red neuronal simple.

6.2.2 Búsqueda de hiperparámetros: Optimizador

En este apartado tratamos de averiguar qué optimizador funciona mejor para el problema que tratamos de resolver. Para ello, además hemos disminuido el valor del umbral para llevar a cabo la clasificación ya que como hemos visto en el apartado anterior se tendía a clasificar en empate demasiados partidos. Para este apartado el umbral tiene un valor de 0’25.

En la Figura 24 se observa la evolución de la función de pérdida en entrenamiento y validación para los tres optimizadores. Se observa que para el entrenamiento SGD la disminución de la función de pérdida es más lenta que la de los otros dos optimizadores. Por otro lado, con el gráfico del conjunto de validación se aprecia un sobreajuste claro para los tres optimizadores. RMSPROP parece el más adecuado, pues muestra un sobreajuste similar a los otros optimizadores, pero con una tendencia ligeramente más estable. Con una elección diferente del resto de hiper parámetros puede obtenerse una mejor evolución del modelo.

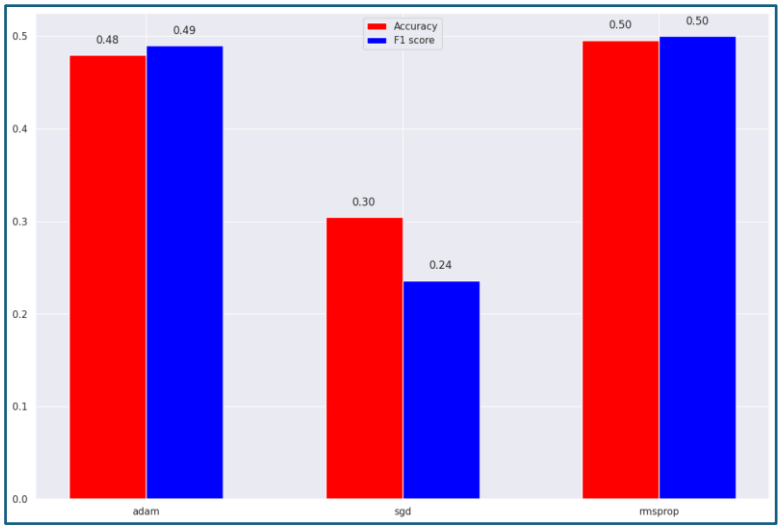

Figura 25. Resultados obtenidos para Adam, SGD y RMSProp.

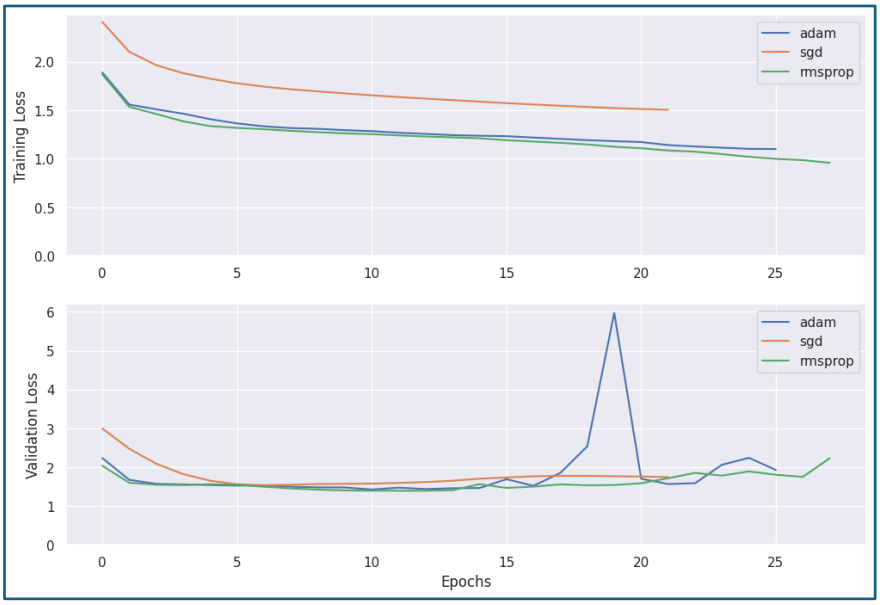

Figura 24. Evolución de la función de pérdida para entrenamiento y validación.

En la Figura 23 se observan los valores de exactitud y F1-Score para obtenidos para los tres optimizadores. Los resultados concuerdan con lo que se aprecia en la Figura 23, los optimizadores de Adam y RMSProp obtienen mejores resultados frente a SGD el cual obtiene un 30% de exactitud y un 24% de F1-Score, lo que significa que por la evolución del modelo con este optimizador no se mejoran ni siquiera los resultados del modelo aleatorio. Por otro lado, Adam y RMSProp obtienen un nivel de exactitud del 50% aproximadamente (48% para Adam y 50% para RMSProp) y un 50% de F1-Score (49% para Adam y 50% para RMSProp). Se observa que RMSProp obtiene resultados ligeramente mejores que Adam.

En la Figura 25 encontramos las predicciones para el número de goles de los optimizadores Adam y RMSProp frente al número de goles real que se marcaron en algunos de los partidos del conjunto de prueba. Basándonos en la información del gráfico, podemos concluir que tanto Adam como RMSProp son capaces de capturar cierta tendencia en los datos, es decir, los picos y valles en las predicciones tienden

a coincidir con los picos y valles en los valores reales. Sin embargo, ambos algoritmos presentan limitaciones en términos de precisión y generalización.

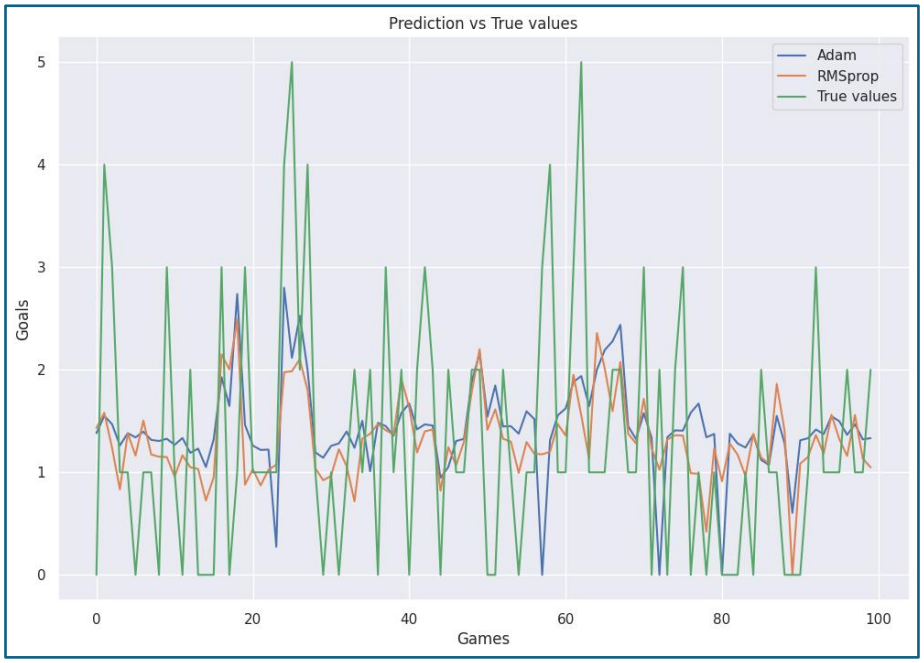

Figura 26. Predicciones de Adam y RMSProp vs Valores reales.

A consecuencia de las conclusiones extraídas, descartamos SGD como un posible optimizador óptimo. En los siguientes apartados probaremos más hiper parámetros, los cuales nos ayudarán a determinar cuál de los dos optimizadores es el óptimo. Por el momento, RMSProp parece el más adecuado.

6.2.3 Búsqueda de hiper parámetros: Tasa de aprendizaje

En este apartado tratamos de encontrar la tasa de aprendizaje de nuestra red neuronal que optimiza la clasificación de esta. Para ello, se han probado las diferentes combinaciones de modelos con los optimizadores de Adam y RMSProp; y con tasas de aprendizaje que van desde 0'1 hasta 10^{-4}.

En la Figura 26 se ha representado la evolución de estos modelos a lo largo del entrenamiento, tanto para el conjunto de entrenamiento como para el de validación. Parece apreciarse que las tasas de aprendizaje de 0'1 no ofrecen buenos resultados, pues la evolución de estas es irregular y en validación se aprecia que tienden al sobreajuste, encontrando picos a lo largo de las diferentes épocas. La tasa de aprendizaje que parece óptima es 0'001, ya que se observa en entrenamiento que es para ambos modelos el que mantiene una tendencia buena en entrenamiento. Además, en la validación los modelos que alcanzan el mínimo son los que tienen esta tasa de aprendizaje (como se puede observar en las diez primeras épocas de entrenamiento).

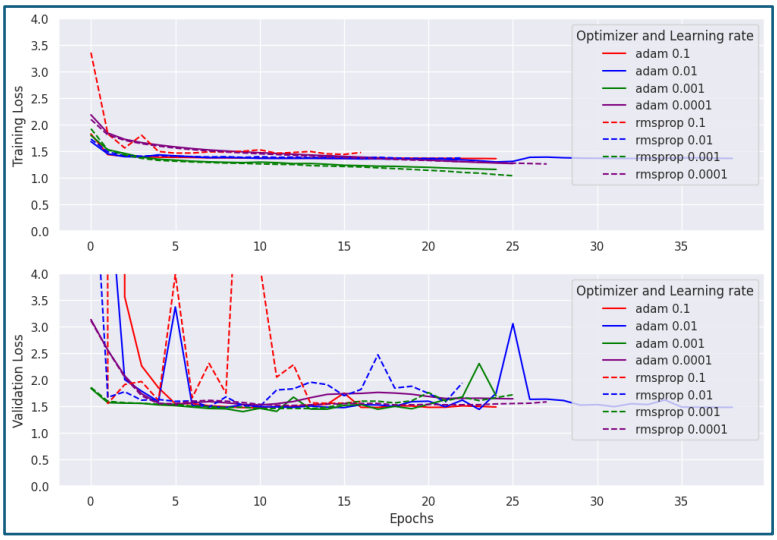

Figura 27. Evolución de la función de pérdida para diferentes tasas de aprendizaje y optimizadores.

En la Tabla 11 encontramos las métricas obtenidas para los diferentes modelos que se han creado. Para ambos optimizadores, la tasa de aprendizaje de 0'001 parece ofrecer un rendimiento óptimo. Esto se observa en las métricas de exactitud, F1 Score y en las precisiones y sensibilidades de las clases. Además, el optimizador Adam generalmente ofrece una mejor precisión y F1 Score que RMSProp para la mayoría de las tasas de aprendizaje, sugiriendo que podría ser más adecuado para este problema. En general, la precisión y sensibilidad para las victorias del equipo local son más altas en comparación con las victorias del equipo visitante y los empates, lo que nos muestra que el modelo está más seguro al predecir este resultado.

Optimizador	Tasa de aprendizaje	Exactitud	F1 Score	Precisión local	Precisión empate	Precisión visitante	Sensibilidad local	Sensibilidad empate	Sensibilidad visitante
Adam	**0'1**	0'44	0'44	0'66	0'31	0'58	0'33	0'74	0'35
Adam	**0'01**	0'45	0'45	0'66	0'30	0'62	0'38	0'67	0'46
Adam	**0'001**	0'49	0'49	0'64	0'32	0'57	0'53	0'55	0'40
Adam	**0'0001**	0'35	0'34	0'61	0'26	0'46	0'23	0'72	0'24
RMSProp	**0'1**	0'40	0'40	0'68	0'28	0'54	0'27	0'74	0'30
RMSProp	**0'01**	0'39	0'39	0'63	0'27	0'52	0'27	0'67	0'32
RMSProp	**0'001**	0'45	0'45	0'64	0'30	0'67	0'41	0'69	0'31
RMSProp	**0'0001**	0'39	0'39	0'58	0'26	0'44	0'35	0'48	0'38

Tabla 11. Métricas obtenidas para las diferentes tasas de aprendizaje y optimizadores.

En la Figura 27 se ha representado la densidad del número de goles predichos por algunos de los modelos frente al la densidad real del número de goles marcados en los partidos. Se observa en este gráfico dos comportamientos muy diferentes:

- Por un lado, RMSProp con tasa de aprendizaje de 0'001 y Adam con tasa de aprendizaje 0'1 tienden a acumular todas sus predicciones cerca del valor 1, lo que nos indica que estos modelos se están ajustando para predecir el valor mayoritario. Probablemente RMSProp con tasa de aprendizaje 0'001 reduce el error medio cuadrático en la predicción ya que tiende a, cuando existen dudas en la predicción o no se encuentra un patrón claro para predecir, hacer esta predicción por el valor media del número de goles. Esto no nos

interesa en absoluto ya que, para conseguir un modelo robusto, lo ideal es capaz de detectar situaciones en las que el número de goles varíe, y en este caso estamos observando que la varianza de nuestras predicciones es muy pequeña.

- Por otro lado, Adam con 0'001 de tasa de aprendizaje y RMSProp con 0'0001 de tasa de aprendizaje muestran un comportamiento diferente a la hora de hacer las predicciones. El número de goles predicho no se agrupa tanto en torno al valor 1, sino que tiene su media un poco más alta (en torno a 1'5 goles) y más variabilidad en el número de goles. Con los resultados obtenidos, parece que Adam con 0'001 es el modelo que mejor es capaz de predecir el número de goles del siguiente encuentro de un equipo. Para RMSProp parece que la tasa de aprendizaje de 0'0001 no es suficiente para que se realice un buen entrenamiento, pues existen muchos resultados a los que le asigna un número de goles igual a 0, lo que puede ser consecuencia de falta de entrenamiento del modelo.

Para todos los modelos se observa que es muy difícil predecir cuando un equipo va a golear a otro, pues el número más alto de goles predichos lo ofrece Adam con 0'001 de tasa de aprendizaje el cual predice hasta 3 goles por partido, cuando la realidad es que el máximo número de goles en un partido para el conjunto de prueba es de 7 goles.

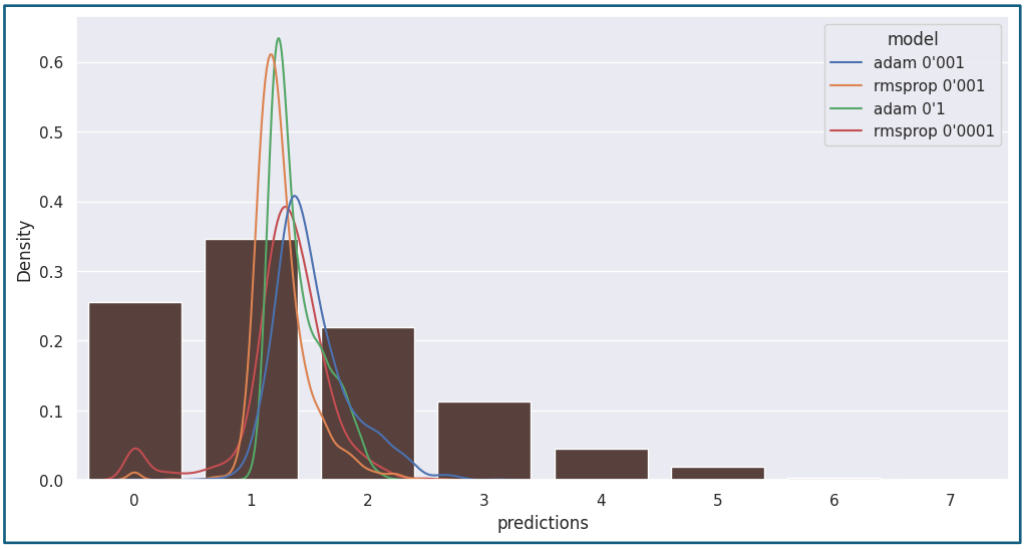

Figura 28. Densidad del número de goles predichos por los modelos frente a la densidad real del número de goles.

En la Figura 28 se representa una comparación entre las predicciones de goles realizadas por dos algoritmos de optimización, Adam y RMSProp, y los valores reales de goles en cuatro categorías diferentes (1,2,3 y 4 goles). Se observa que tanto Adam como RMSProp muestran una considerable variabilidad en sus predicciones a lo largo de las diferentes categorías, lo que indica que ninguno de los dos algoritmos es capaz de predecir con una precisión perfecta el número de goles en todos los casos. Ambos algoritmos tienden a subestimar el número de goles, especialmente en las categorías con valores más altos. Adam parece ser el algoritmo que mejor se ajusta en la mayoría de los casos al número de goles marcados en realidad, sin

embargo, puede ocurrir debido a la aleatoriedad. Una conclusión clara que sí que se manifiesta a lo largo de este apartado es que la tasa de aprendizaje de 0’001 parece ser la óptima independientemente del optimizador, por lo que en los siguientes experimentos será la utilizada.

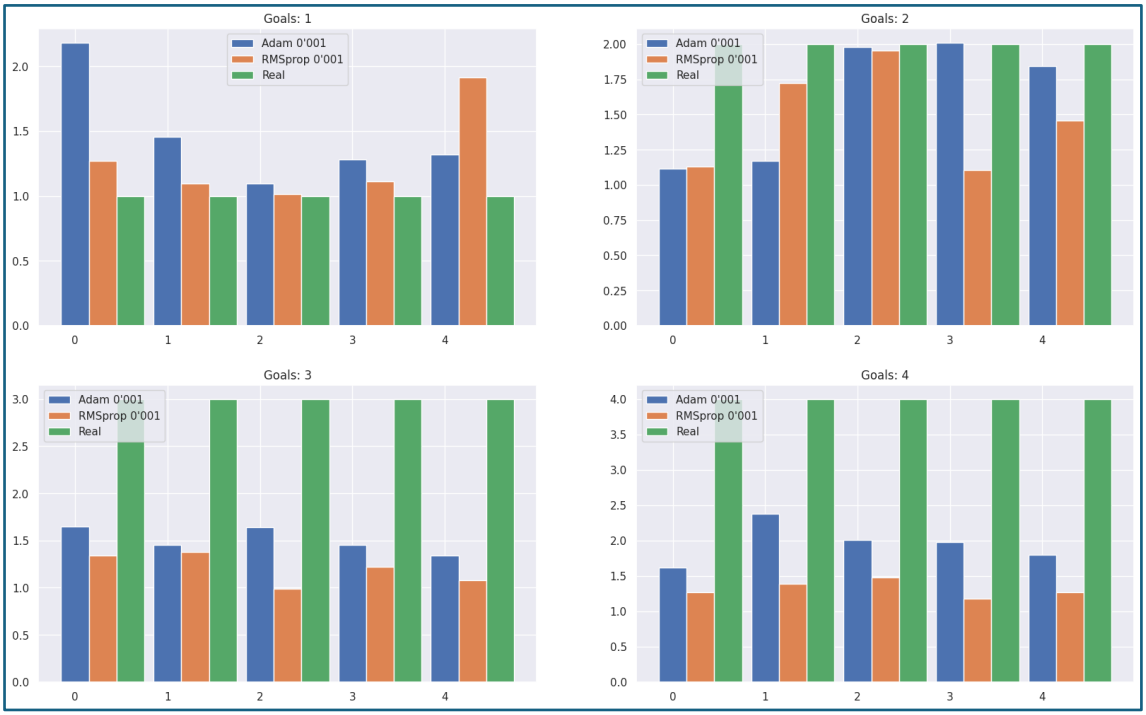

Figura 29. Predicción del número de goles para Adam y RMSProp frente al número real de goles.

6.2.4 Búsqueda de hiperparámetros: Umbral y funciones de activación

En este apartado tratamos de encontrar la función de activación de nuestra red neuronal y el umbral de decisión que optimiza la clasificación. Para ello, se han probado las diferentes combinaciones de modelos con cuatro funciones de activación distintas: lineal, tangente hiperbólica, sigmoide y ReLu. Además, después de entrenar las redes neuronales con cada una de estas funciones de activación se ha llevado a cabo la clasificación con distintos umbrales que abarcan desde 0’01 hasta 0’5.

En la Figura 29 encontramos representadas las evoluciones de la función de pérdida de los diferentes modelos para la validación. Algunos modelos muestran picos de pérdida pronunciados, lo cual sugiere que el modelo podría estar sobre ajustándose temporalmente o encontrando puntos de inestabilidad durante el entrenamiento. Las combinaciones más estables que encontramos son Adam con la función sigmoide y RMSProp con las funciones lineal y ReLu. Generalmente, RMSProp parece proporcionar una pérdida de validación más estable y en algunos casos más baja.

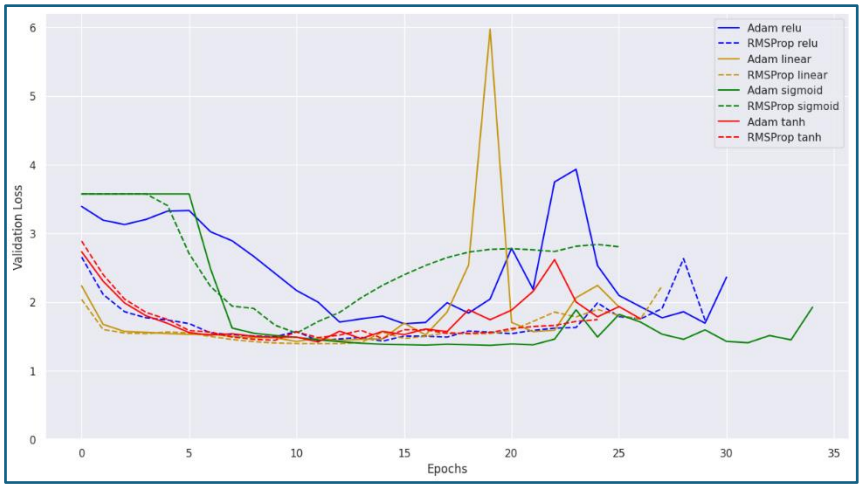

Figura 30. Evolución de la función de pérdida en validación para los diferentes modelos.

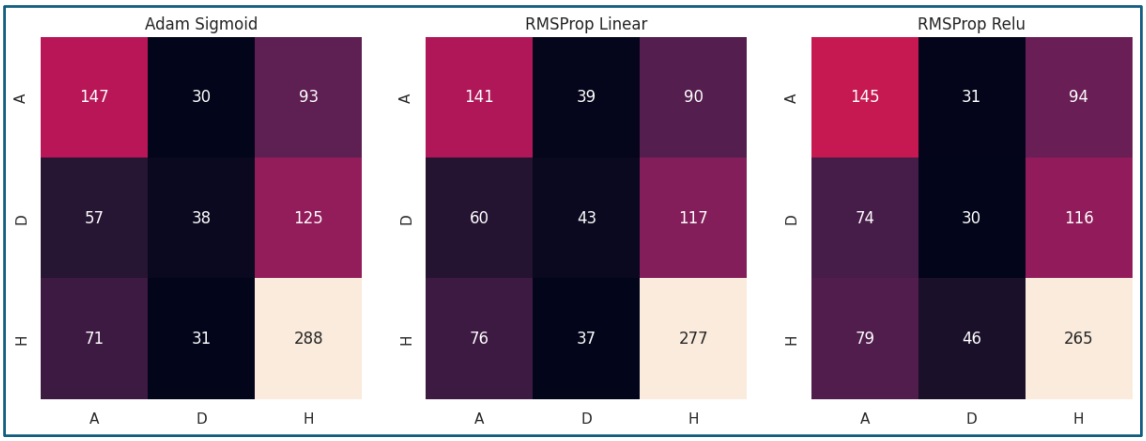

Figura 30. Matrices de confusión para tres modelos diferentes.

Para comparar los resultados que proporcionaban estos modelos, se seleccionó los tres mencionados anteriormente. En la Figura 30 se muestran las matrices de confusión para los modelos, utilizando un umbral de 0'1 para la clasificación. Los tres modelos parecen ofrecer un rendimiento muy similar, siendo modelos relativamente buenos a la hora de clasificar las victorias de los equipos visitantes y las victorias de los equipos locales, y mostrando dificultades a la hora de detectar los empates. El modelo que utiliza el optimizador Adam con la función de activación sigmoide parece se ligeramente superior a los otros dos, teniendo una mayor cantidad de aciertos. En la siguiente tabla se pueden apreciar las métricas obtenidas con la clasificación.

Optimizador	**Función de activación**	**Exactitud**	**F1 Score**	**Precisión**		
				Local	**Empate**	**Visitante**
Adam	Sigmoide	0'54	0'47	0'57	0'38	0'54
RMSProp	Lineal	0'52	0'47	0'57	0'36	0'51
RMSProp	ReLu	0'50	0'44	0'56	0'28	0'49

Tabla 12. Resultados para los modelos de Adam y RMSProp con diferentes funciones de activación y umbral de decisión de 0'1.

Se entiende que el optimizador Adam con la función Sigmoide es el modelo que mejores resultados ofrece, teniendo 0'54 de exactitud (superando a RMSProp con las funciones lineal y ReLu), 0'47 de macro F1 Score y una precisión para las

victorias del equipo local y visitante superiores al 0’5. Dados los resultados, consideramos este modelo que utiliza el optimizador Adam y la función de activación Sigmoide como el más adecuado para la resolución de nuestro problema. Para tratar de averiguar qué umbral de decisión es el que mejores resultados ofrece, se ha representado en la Figura 31 las métricas de Exactitud y F1 Score para diferentes umbrales de decisión. Como se puede observar, para valores extremos del umbral (tanto bajos como altos) es para los que peores resultados se obtienen. Por un lado, si se establece un umbral muy bajo (como es el caso del 0’01) se obtiene un valor de exactitud bueno (>0’5) ya que estamos haciendo predicciones para las clases mayoritarias (victoria del local o victoria del visitante). Sin embargo, se aprecia que para este umbral bajo el valor del F1 Score disminuye notablemente debido a que se clasifican muy pocos valores como empate. Es decir, si establecemos un umbral de clasificación muy bajo, se clasifican muy pocos valores como empate ya que se requiere una precisión muy alta del modelo para que se acierten los empates. Este umbral no nos interesa, pues nuestro objetivo es intentar predecir cualquiera de los tres posibles resultados de un partido, y utilizando este umbral estamos prácticamente ignorando uno de ellos.

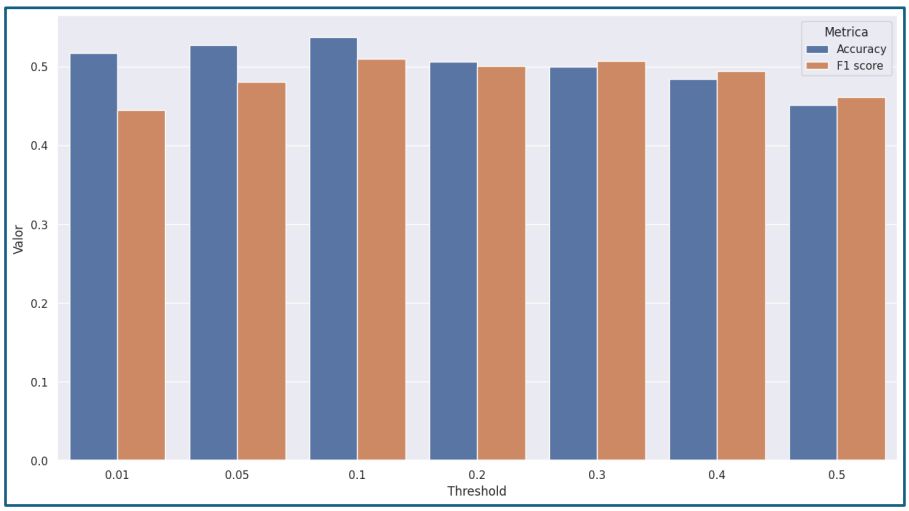

Figura 31. Exactitud y F1 Score para diferentes umbrales de clasificación (Optimizador Adam y función Sigmoide).

Por otro lado, cuando se establece un umbral de decisión muy alto, como en este caso el umbral de 0’5, vemos como disminuye considerablemente la exactitud del modelo. Esto es debido a que muchas observaciones pasan ahora a estar clasificadas en la categoría empate. A pesar de esto, que la exactitud y el F1 Score estén ambos alrededor del 0’45 es buen indicador de que nuestra red neuronal es capaz de identificar los partidos con una victoria del local o visitante muy claras. El umbral de decisión óptimo parece encontrarse alrededor del 0’1, donde se maximiza tanto la exactitud como el F1 Score.

En la Tabla 13 se encuentran las métricas obtenidas con el modelo que utiliza optimizador Adam y función de activación Sigmoide. Se observa que las métricas son muy similares para los diferentes umbrales, sin embargo, determinamos que el mejor umbral es el de 0’12, ya que es el que maximiza el macro F1 Score y obtiene el segundo mejor resultado en la métrica de exactitud.

Umbral	Exactitud	macro F1 Score
0.09	0.535227	0.466678
0.10	0.537500	0.473517
0.11	0.532955	0.473232
0.12	0.534091	0.480313
0.13	0.530682	0.479929
0.14	0.521591	0.473197

Tabla 13. Métricas obtenidas para los diferentes umbrales de clasificación.

Hasta este apartado hemos determinado el mejor modelo es el que utiliza optimizador Adam, con función de activación Sigmoide, tasa de aprendizaje 0'001 y umbral de clasificación de 0'12. En el siguiente apartado se comparan diferentes arquitecturas de red y se determina el modelo de red neuronal final.

6.2.5 Arquitectura de la red neuronal

En este apartado se comparan diferentes arquitecturas de redes neuronales LSTM con el objetivo de ver cuál se ajusta mejor a nuestro problema de clasificación. En la Tabla 14 encontramos las 5 arquitecturas distintas de redes neuronales con las que se trabaja en este apartado.

Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
LSTM-128 Batch Norm LSTM-64 Batch Norm LSTM-32 Batch Norm Capa densa-1	LSTM-128 Batch Norm Dropout-0'5 LSTM-64 Batch Norm Dropout-0'5 LSTM-32 Batch Norm Dropout-0'5 Capa densa-1	LSTM-128 Dropout-0'5 LSTM-64 Dropout-0'5 LSTM-32 Dropout-0'5 Capa densa-1	LSTM-256 LSTM-128 Dropout-0'5 LSTM-64 Dropout-0'5 LSTM-32 Dropout-0'5 Capa densa-1	LSTM-256 LSTM-128 Batch Norm Dropout-0'5 LSTM-64 Dropout-0'5 LSTM-32 Dropout-0'5 Capa densa-1

Tabla 14. Arquitecturas de los 5 modelos probados.

En la Figura 32 se muestra la evolución de la función de pérdida en entrenamiento y validación para los 5 modelos propuestos. A simple vista nos llama la atención un modelo por encima del resto: el modelo 1, el cual se observa que en entrenamiento es el que mejor evoluciona, mientras que en validación presenta serios problemas, lo que es un claro indicador de sobre ajuste. Si nos fijamos, el único modelo que no utiliza la técnica de Dropout es el único que muestra síntomas claros de sobre ajuste, por lo que podemos deducir que el uso de esta técnica es de utilidad en el problema que nos ocupa. El resto de los modelos parecen presentar una evolución más estable, destacando el modelo 3 que llega casi a las 100 épocas de entrenamiento.

El uso de normalización del batch parece tener un efecto significativo en el tiempo de entrenamiento, pues se observa en la Tabla 15 que los modelos que emplean esta técnica tardan menos en finalizar el entrenamiento. El modelo 5, contando con una estructura muy similar a la del modelo 4 tarda aproximadamente

30 segundos menos en realizar el entrenamiento. También hay que remarcar que tiene un error medio cuadrático ligeramente superior.

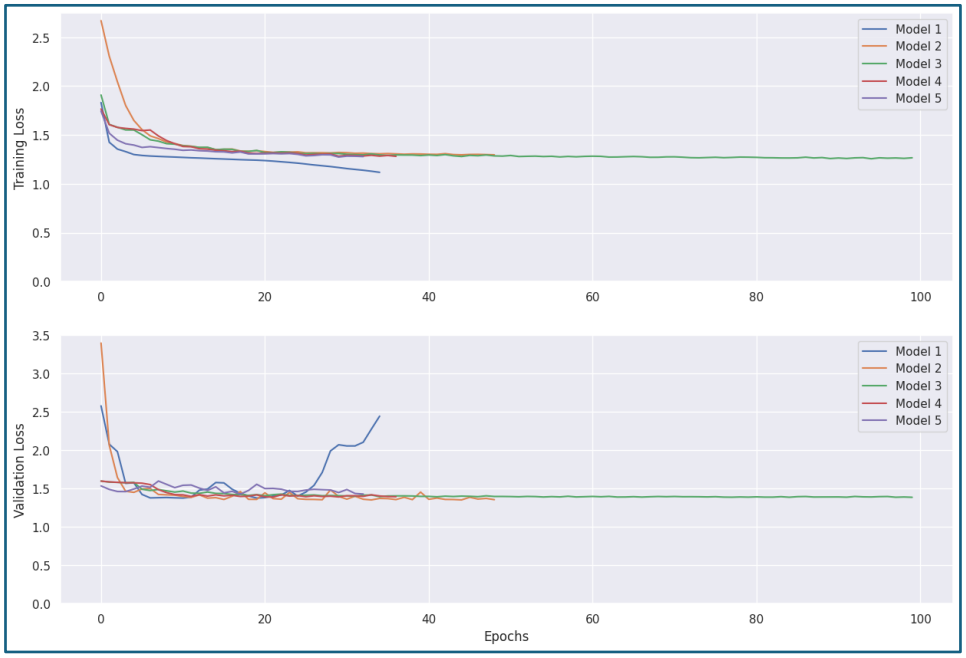

Figura 32. Evolución de la función de pérdida en entramiento y validación para los modelos propuestos.

Modelo	Tiempo de entramiento	MSE	Exactitud	F1 Score
Modelo 1	141’23 segundos	1’3774	0’509	0’4389
Modelo 2	212’71 segundos	1’353	0’527	0’4567
Modelo 3	325’28 segundos	1’3874	0’531	0’4823
Modelo 4	319’48 segundos	1’3889	0’519	0’4631
Modelo 5	295’67 segundos	1’4282	0’534	0’4772

Tabla 15. Tiempo de entrenamiento y métricas de cada modelo.

Como se observa en la tabla, a pesar de que el modelo 5 es el que mayor error medio cuadrático tiene, es el que mejores valores de exactitud, macro F1 Score y tiempo de entrenamiento presenta en conjunto. Esto es consecuencia de que el modelo es capaz de predecir mejor cuando un equipo va a marcar más goles que otro, pero es peor prediciendo el número de goles. Es decir, el modelo es el peor de los 5 prediciendo el valor exacto de número de goles ya que es el que menos variabilidad tiene en las predicciones, pero si que detecta mejor que el resto cuando un equipo va a marcar más goles que otro (aunque no prediga correctamente cuánto es esta cantidad de goles). Si observamos la Figura 33, podemos ver la variabilidad de las predicciones del número de goles de cada modelo. En ella se aprecia claramente como existe una relación directa entre la complejidad del modelo y la variabilidad de las predicciones. Mientras que el modelo más simple (modelo 1) cuenta con un abanico de predicciones desde los 0 goles hasta los 3’5 goles, el modelo más complejo ve el suyo reducido de los 0’8 goles aproximadamente hasta

casi los 2’5. A priori podríamos pensar que tener una variabilidad de goles mayor beneficia a nuestro modelo, sin embargo, debemos tener en cuenta varios factores:

- Por un lado, el tener una variabilidad mayor no implica que estas predicciones sean más correctas que las de un modelo con menor variabilidad, simplemente hace más probable que se acierte un resultado con mucha cantidad de goles.
- Por otra parte, nuestra tarea no es ser capaces de predecir cuántos goles va a meter cada equipo, sino tratar de adivinar si ganará el equipo local, habrá un empate o ganará el equipo visitante.

Por lo tanto, para nuestra tarea de clasificación escogemos el modelo 5 como la arquitectura más apropiada para ella.

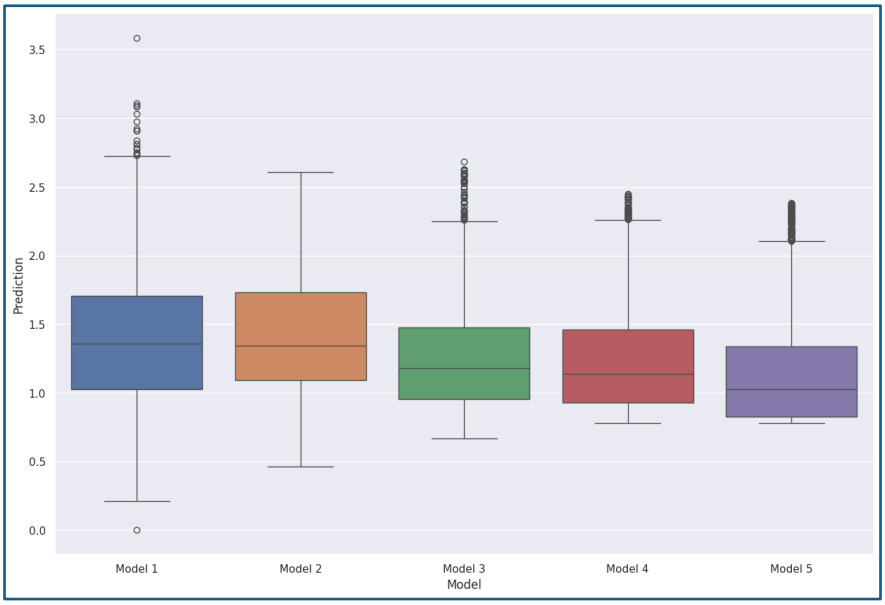

Figura 33. Gráfico de caja y bigotes para las predicciones del número de goles de cada modelo.

6.2.6 Resultados de la red neuronal LSTM óptima

La arquitectura seleccionada como mejor modelo es la del modelo 5, que cuenta con las capas que se pueden ver en la tabla 15. Como resultado de todos los experimentos realizados, se ha seleccionado como función de activación la Sigmoide, el optimizador Adam y una tasa de aprendizaje de 0’001. Para el umbral de clasificación se ha escogido el valor 0’1. En este apartado se analizan los resultados que ofrece este modelo, tratando de comprenderlos y obtener conclusiones de ellos.

En la Figura 34 se han representado dos mapas de calor, uno con la exactitud para cada uno de los resultados de un partido y otro con el número de casos que hay en el conjunto de prueba. Es decir, el 0’62 de la primera columna y segunda fila del mapa de calor de la izquierda significa que de los partidos que terminan 1-0, el modelo tiene un acierto del 62%. Mientras que el valor que hay en la misma casilla en el mapa de calor de la derecha (82) representa el número de encuentros que terminan con este resultado. Se puede ver cómo el modelo ofrece muy buenos resultados a la hora de pronosticar los resultados favorables al equipo local, no

solamente porque estos sean la clase mayoritaria, sino porque dentro de cada resultado tiene un porcentaje de acierto alto, superior al 60% en prácticamente todos y llegando a valores superiores al 80% para los resultados más abultados.

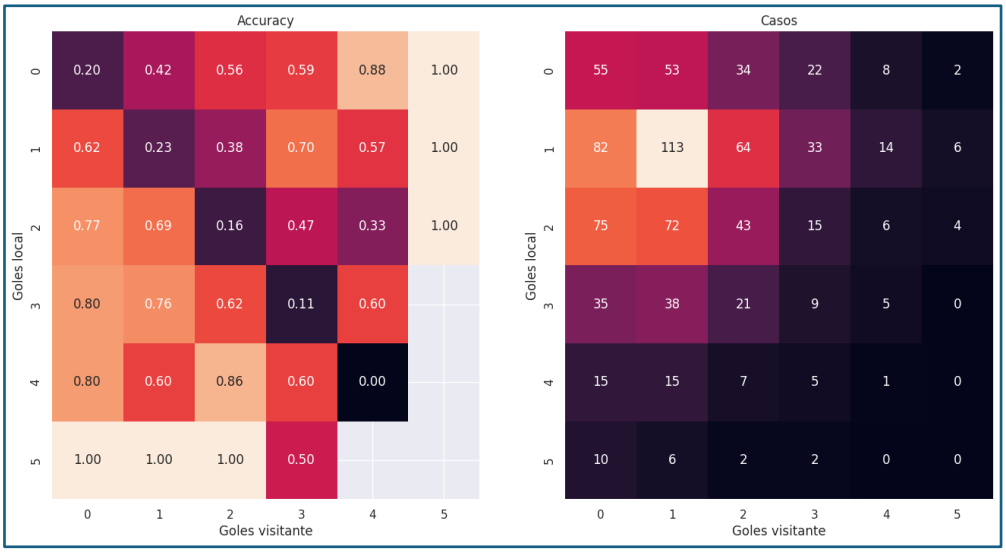

Figura 34. Mapas de calor para la exactitud y el número de casos de cada resultado.

Además, el modelo también es capaz de identificar con facilidad en la mayoría de los casos cuando la victoria se da para el equipo visitante, presentando dificultados solamente cuando en los encuentros hay pocos goles y el visitante gana por la mínima (0-1 o 1-2). Se evidencia también las limitaciones que presenta el modelo a la hora de predecir los empates, teniendo menos de un 23% de acierto para todos los posibles empates.

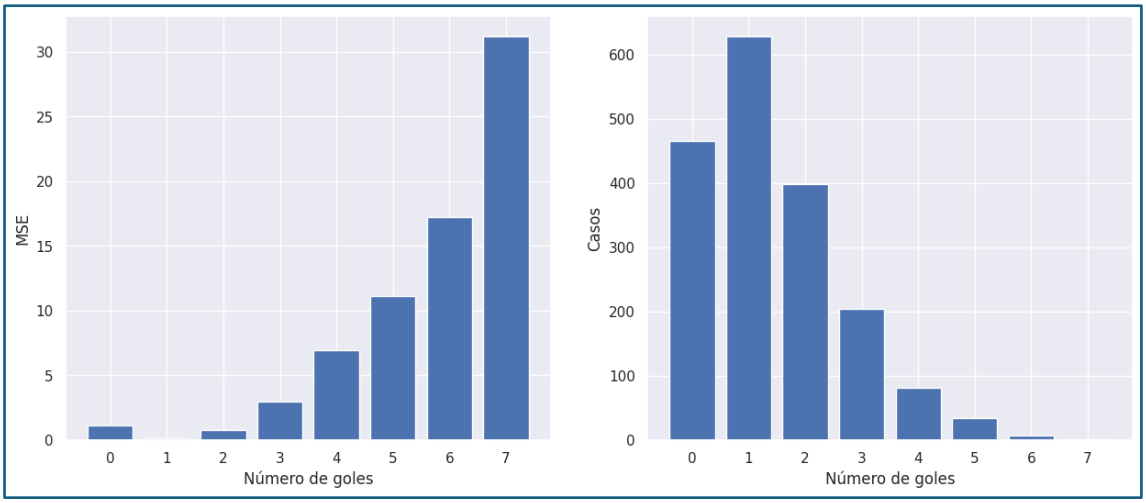

Figura 35. MSE para diferente número de goles en un partido.

En la Figura 35 se ha representado dos gráficos de barras, uno (el de la izquierda) para el error media cuadrático para el diferente número de goles que se ha metido en cada partido, es decir, cada barra representa el error medio cuadrático de los partidos en los que se ha metido esa cantidad de goles. El otro gráfico de barras representa el número de casos en los que se ha metido esa cantidad de goles. Así, podemos observar cómo ambos gráficos son completamente opuestos. El número de goles más común que mete un equipo en un partido es 1, y a su vez en

esta cifra para la que se tiene el menor error medio cuadrático (0'16 de MSE), mientras que para el caso menos común (partidos en los que se meten 7 goles) el error medio cuadrático supera las 30 unidades. Esto quiere decir que nuestro modelo es incapaz de encontrar la forma de anticipar que hay una goleada, lo que puede ser debido al desbalanceo existente en la cantidad de goles que se mete en cada partido (en el 82% de los partidos se marcan dos goles o menos).

En la Tabla 16 podemos ver el informe de clasificación de nuestro modelo. Vemos como es el mejor que se ha podido obtener, llegando a un 55% de exactitud. Los resultados que se aprecian nos dejan conclusiones muy similares a las que hemos obtenido hasta ahora: la clase que mejor se predice es la victoria del equipo local, el modelo cumple con el objetivo propuesto para este trabajo decentemente y la clase más difícil de predecir es el empate, para la cual se tiene un valor de sensibilidad de solamente 0'2.

		Precisión	**Sensibilidad**	**F1 Score**	**Soporte**
Clase	**A**	0'53	0'53	0'53	270
	D	0'37	0'2	0'26	220
	H	0'58	0'73	0'64	390
Métrica	**Exactitud**			0'55	880
	Macro Average	0'49	0'49	0'48	880
	Weighted Average	0'51	0'53	0'51	880

Tabla 16. Informe de clasificación del modelo LSTM final.

Por último, en la Tabla 17 podemos ver los equipos para los que más se aciertan las victorias y las derrotas como local y visitante. Se aprecia que el modelo sabe diferenciar perfectamente algunos equipos a la hora de hacer predicciones de estos. Por ejemplo, cuando se trata de predecir que van a ganar, tanto jugando de locales como de visitante, el modelo adivina más del 90% de los partidos de equipos como el Liverpool, Real Madrid, Manchester City… Es decir, el modelo está siendo capaz de detectar las características de equipos que por lo general son de un nivel más alto y hacer predicciones sobre estos.

	Victoria		Derrota	
	Equipo	Porcentaje	Equipo	Porcentaje
Jugando como local	Atleti	100%	Aston Villa	100%
	Liverpool	100%	Norwich City	100%
	Real Madrid	100%	Newcastle	100%
	AC Milan	100%	Torino	100%
	Barcelona	100%	Genoa	100%
	Man City	100%	Sassuolo	75%
	Inter Milan	100%	Watford	75%
Jugando como visitante	Napoles	100%	Mallorca	100%
	Inter Milan	100%	Leeds Utd	100%
	Man City	100%	Sampdoria	100%
	Real Madrid	100%	Osasuna	100%
	PSG	100%	Southampton	100%
	Liverpool	90%	Angers	85%
	Barça	83%	Bordeaux	83%

Tabla 17. Equipos para los que más se acierta cada tipo de resultado.

De la misma manera, para los equipos que más pierden tanto de local como de visitante ocurre el proceso contrario. Los equipos para los que más se acierta son equipos que por lo general están asociados a un nivel más bajo que los anteriores (Norwich City, Watford, Bordeaux, Mallorca…). Esto quiere decir que nuestro modelo es capaz de diferenciar las características de estos equipos de un nivel inferior y pronosticar las derrotas de estos.

6.3 Experimentos con LightGBM

El segundo de los algoritmos con el que tratamos de llevar a cabo nuestra tarea de clasificación es LightGBM. El objetivo de esta sección es encontrar el modelo de este algoritmo que mejor es capaz de resolver nuestro problema de predicción de resultados.

Para la tarea de clasificación, LightGBM recibirá las medias ponderadas de los partidos anteriores, tanto del equipo local como del visitante. Recibirá también las estadísticas relacionadas con los elos de cada equipo antes del partido y las cuotas que asocian las casas de apuestas a cada uno de los 3 resultados posibles. Con esta información, el modelo devolverá un valor para la clasificación: victoria del local, empate o victoria del visitante.

Por otro lado, para la evaluación de estos modelos se utilizarán diferentes métricas. Para el desempeño del modelo en la clasificación utilizaremos métricas como la exactitud, el F1 Score, la matriz de confusión o el informe de clasificación. Además, para la evaluación de las probabilidades asignadas por el modelo se emplearán funciones como la pérdida logarítmica (log loss)

En los siguientes apartados se explican los diferentes experimentos llevados a cabo con este algoritmo.

6.3.1 Búsqueda en rejilla (Grid Search)

La búsqueda en rejilla es una técnica utilizada en el aprendizaje automático para encontrar el mejor conjunto de hiperparámetros para un modelo. El proceso que aplica esta técnica es sencillo:

Primero, se define un espacio de búsqueda, que es una cuadrícula de todos los posibles valores de hiperparámetros que se desee probar. Cada punto en la cuadrícula corresponde a una combinación específica de valores de hiperparámetros.

Después, la búsqueda en rejilla entrena y evalúa el modelo usando cada combinación posible de hiperparámetros. Esto se hace normalmente utilizando validación cruzada para asegurar que el modelo se evalúe de manera justa y que los resultados no dependan de una única partición del conjunto de datos.

Por último, después de evaluar todas las combinaciones posibles, se selecciona el conjunto de hiperparámetros que produce el mejor rendimiento según una métrica de evaluación específica. En nuestro caso la métrica empleada para la

selección del conjunto de hiperparámetros es la exactitud del modelo, ya que nuestro objetivo es predecir el mayor número de resultados posible.

En la Tabla 18 se encuentran los hiperparámetros seleccionados por la búsqueda en red.

Modelo	Tasa de aprendizaje	Profundidad máxima	N.º Estimadores	N.º hojas
LightGBM	0'01	No límite	150	40

Tabla 18. Hiperparámetros seleccionados por la búsqueda en rejilla.

En la Figura 36 se ha representado la matriz de confusión obtenida con estos hiperparámetros del modelo. Como se puede apreciar, el problema de la identificación de los empates se acentúa en esta aproximación utilizando LightGBM. El modelo sólo predice 27 muestras como empate, y de esas realmente son empates reales 9, por lo que tiene un precisión y sensibilidad muy bajas.

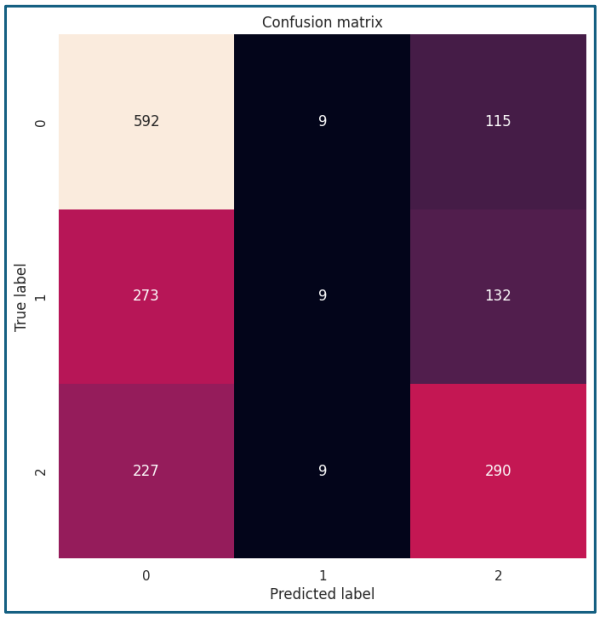

Figura 36. Matriz de confusión para el modelo de LightGBM.

En el siguiente apartado se evalúa y se analizan los resultados proporcionados por este modelo.

6.3.2 Resultados de LightGBM

La búsqueda en rejilla nos ha proporcionado una combinación óptima de hiperparámetros para nuestra tarea, sin embargo, los resultados de nuestro modelo no distan mucho de lo conseguido con redes neuronales LSTM.

En la Figura 37 se ha representado la matriz de confusión del modelo normalizada por filas y por columnas. Como se puede ver, el modelo cuenta con unas precisiones considerablemente buenas para las victorias tanto de los equipos visitantes como de los equipos locales. El punto débil del modelo es la predicción de los empates, para la cual se cuenta con sólo un 0'02 de sensibilidad, lo que nos indica que este modelo es incapaz de predecir prácticamente ningún empate.

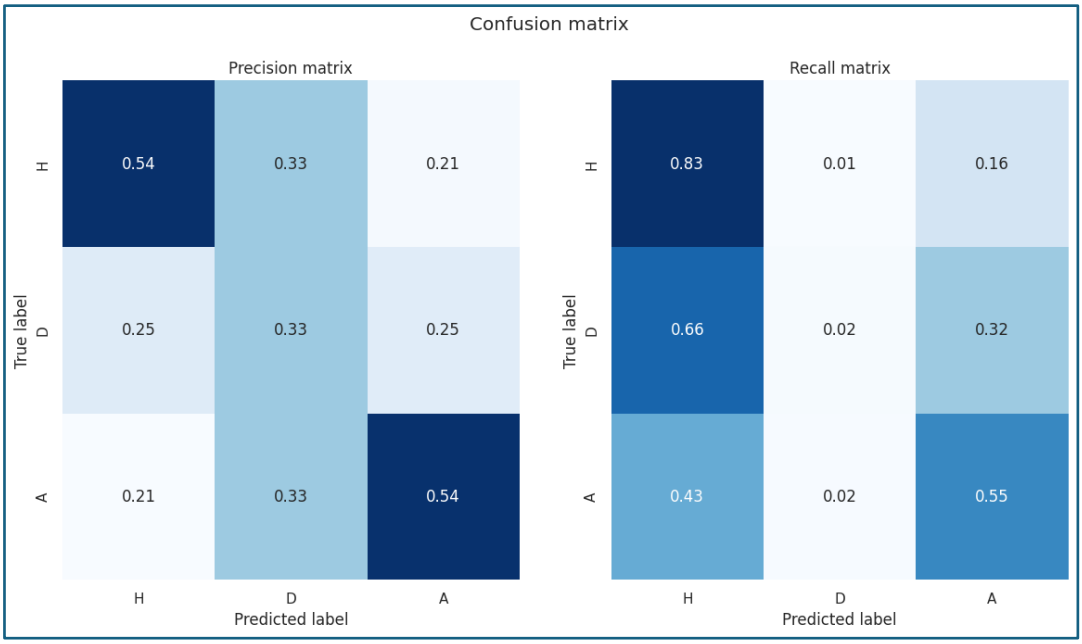

Figura 37. Matriz de confusión del modelo LightGBM, normalizada por filas y por columnas.

En la Figura 38 se ha representado la importancia de las variables para hacer las predicciones. Se observan 4 variables que destacan por encima del resto: las relacionadas con las cuotas de las casas de apuestas para cada resultado y la variable de la tendencia de elo del equipo local. Estas variables son las que se utilizan principalmente para llevar a cabo las particiones dentro de los árboles de decisión de nuestro algoritmo. Nos llama la atención el resto de las variables que se aprecian, pues aunque cabría esperar que las variables que más se utilizaran para clasificar fueran las relacionadas con la fase ofensiva del juego, encontramos otras variables relacionadas con la creación de jugadas y con la fase defensiva.

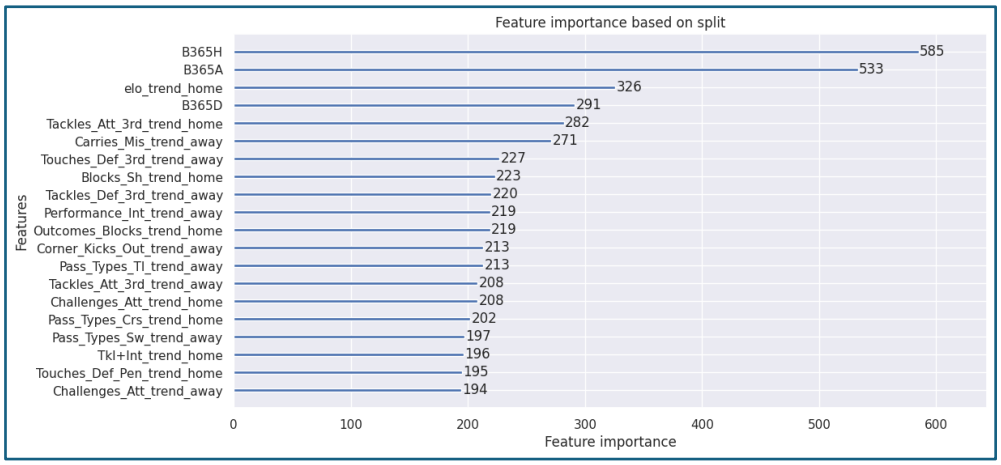

Figura 38. Importancia de las variables basada en el número de divisiones.

En la Figura 39 se ha representado el promedio de probabilidad que le asigna a cada resultado el modelo para los diferentes partidos que tenemos. El valor de Goal_Diff del eje X representa la diferencia de goles entre el local y el visitante en los partidos, es decir, las barras de Goal_Diff = 2 representan las probabilidades que se le asigna en promedio a cada uno de los resultados en los partidos en los que el equipo local gana con dos goles de ventaja al equipo visitante. Así, observamos que

el modelo en líneas generales parece ser bastante coherente asignando las probabilidades, pues para diferencias de goles positivas asigna probabilidades más altas a la victoria del equipo local y para diferencias de goles negativas las asigna a la victoria del equipo visitante. Donde muestra principalmente incoherencias es a la hora de asignar probabilidades en los empates, donde vemos que el modelo opta por asignar mayoritariamente la probabilidad a la victoria del equipo local. Es decir, en casos de empate, donde el modelo no consigue identificar patrones para identificar este resultado, el modelo asigna probabilidades similares al número de observaciones de cada clase.

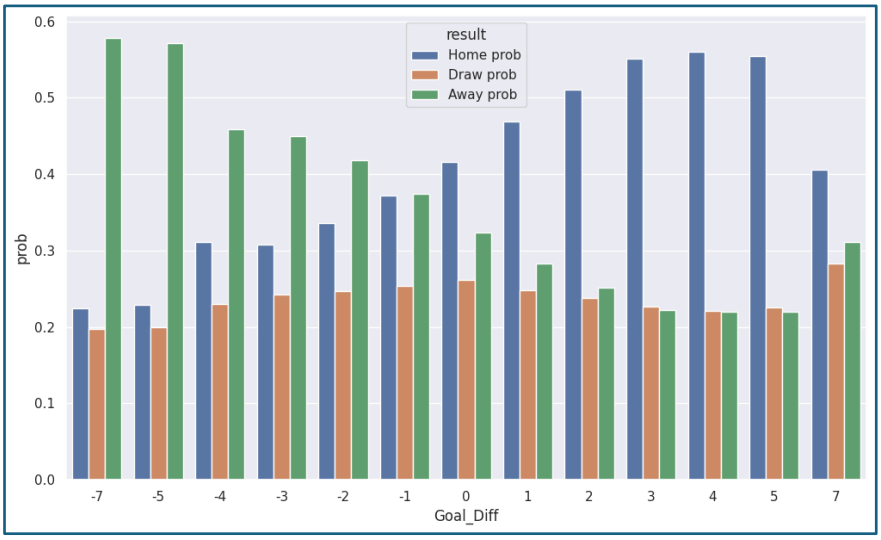

Figura 39. Media de probabilidades asignada a cada resultado para cada tipo de partido.

En la Tabla se ha representado a los equipos a los que se les asigna una mayor probabilidad de victoria y de derrotas, jugando tanto de local como de visitante. Al igual que ocurría con el modelo de red neuronal LSTM, parece que el modelo sabe diferenciar perfectamente algunos equipos a la hora de hacer predicciones de estos, asignando probabilidades de victoria más altas a equipos de un nivel superior, como podrían ser el Barcelona, Manchester City o Bayern de Múnich, y probabilidades de derrotas más altas a equipos como el SPAL, el Bournemouth o el Elche.

	Victoria local		Victoria visitante	
	Equipo	Probabilidad	Equipo	Probabilidad
Jugando como local	Bayern	0’715	West Brom	0’633
	Man City	0’695	Cardiff City	0’625
	Barcelona	0’685	Frosinone	0’569
	PSG	0’679	SPAL	0’511
	Real Madrid	0’662	Bochum	0’509
	Liverpool	0’639	Sheffield Utd	0’492
	Juventus	0’613	Chievo	0’488
Jugando como visitante	Brescia	0’668	PSG	0’598
	Bournemouth	0’647	Man City	0’579
	Paderborn	0’638	Bayern	0’579
	Crotone	0’621	Liverpool	0’556
	Elche	0’614	Real Madrid	0’551
	SPAL	0’589	Barcelona	0’525
	Norwich City	0’586	Inter Milan	0’521

Tabla 19. Probabilidades de victoria y derrotas para los equipos.

6.3.3 Comparación con LSTM y puntos de referencia

En este último apartado de la experimentación tratamos de comparar los mejores modelos obtenidos con ambos abordajes llevado a cabo en el proyecto, además de los resultados que se obtienen con los puntos de referencia explicados anteriormente.

En lo referido al tiempo de entrenamiento, LightGBM se muestra mucho más eficiente. En la Figura 40 se ha representado las densidades de los tiempos de entrenamiento obtenidos tras haber entrenado cada modelo 50 veces. Se observa que LightGBM cuenta tanto con una media (15 segundos de entrenamiento aproximadamente) como con una varianza inferior. Por su lado, las redes neuronales tardan, en promedio, 10 veces más en completar el entrenamiento. Desde este criterio, LightGBM es mucho mejor opción que LSTM.

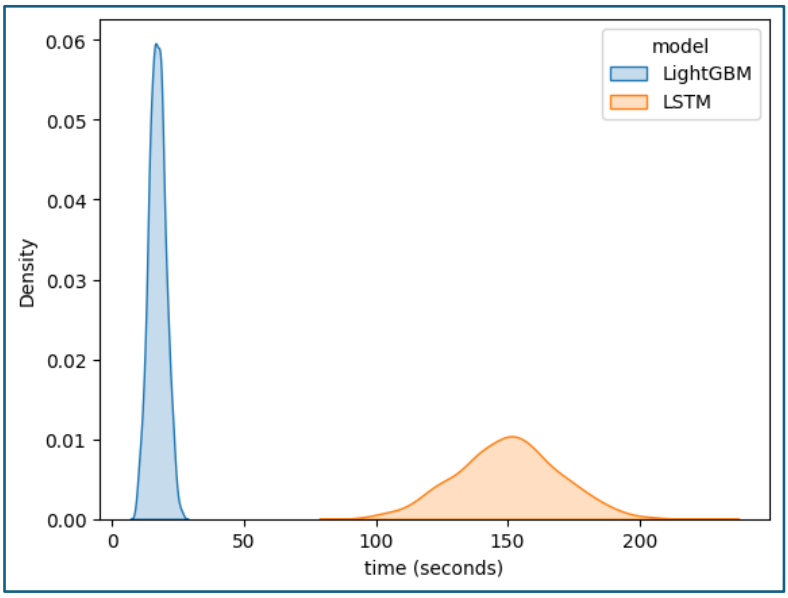

Figura 40. Gráfico de densidad para los tiempos de entrenamiento de la red neuronal LSTM y LightGBM.

Por otro lado, podemos comparar la eficiencia de los modelos a la hora de llevar a cabo las predicciones. En la Tabla 20 se muestra una comparación de las diferentes métricas que se obtiene para cada modelo, mientras que en la Figura 41 se ha representado las matrices de confusión para los modelos de LSTM y de LightGBM. Como se puede observar, tanto el modelo de LSTM como el de LightGBM obtienen mejores resultados para todas las métricas que los modelos que tomamos como puntos de referencia, lo que nos indica que estos modelos siguen una lógica y no son modelos arbitrarios.

Se aprecia que LightGBM obtiene un valor más alto para la exactitud del modelo. Sin embargo, el valor del F1 Score es mejor para el modelo de LSTM, lo que nos indica que este modelo generaliza mejor para las tres clases. Si observamos las precisiones para cada una de las tres clases, se puede deducir que se obtienen unos resultados similares, siendo ligeramente superior para la predicción de los empates la red LSTM.

Además, observando la Figura 41 se observa que el modelo de redes neuronales hace muchas más predicciones para la clase minoritaria, por lo que obtiene un valor de sensibilidad mucho mayor.

Modelo	Exactitud	F1 Score	Tiempo	Precisión local	Precisión empate	Precisión visitante
Aleatorio	0'35	0'34	-	0'42	0'26	0'33
Resultado más frecuente	0'43	0'2	-	1	0	0
LSTM	0'53	0'48	152 s	0'58	0'37	0'53
LightGBM	0'55	0'41	17 s	0'54	0'33	0'54

Tabla 20. Métricas obtenidas para los modelos finales.

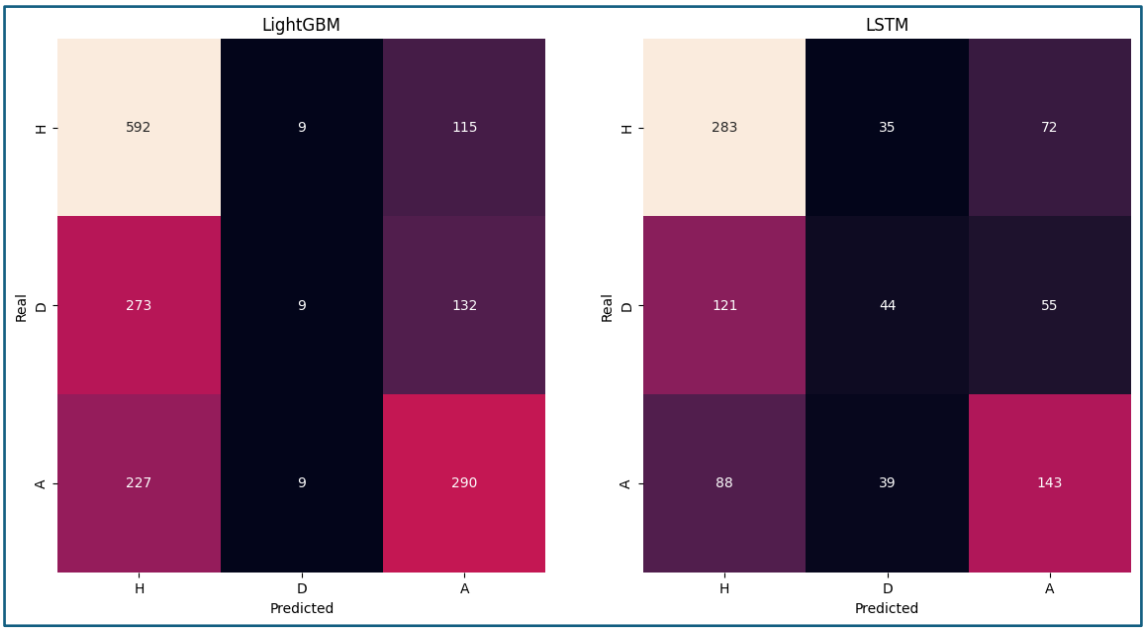

Figura 41. Matrices de confusión para los modelos de LSTM y de LightGBM.

En general, ambos modelos consiguen obtener unos buenos resultados a la hora de predecir las victorias de alguno de los dos equipos, y muestran limitaciones a la hora de predecir los empates. Sin embargo, el modelo que emplea redes neuronales muestra más flexibilidad a la hora de llevar a cabo la predicción de esta clase, por lo que para trabajos futuros puede tener más recorrido a la hora de mejorar los resultados.

7. Conclusiones

En este capítulo se discuten las conclusiones del trabajo realizado. A diferencia de las conclusiones obtenidas con los resultados de los modelos, estas conclusiones se centrarán en una perspectiva más general.

Durante este trabajo se han desarrollado dos tipos modelos de aprendizaje automático para la clasificación de resultados de fútbol. Los modelos de redes neuronales LSTM han ofrecido un rendimiento sorprendente a la hora de clasificar, pues la tarea de predecir el número de goles no parecía ser la mejor estrategia. Por su lado, los modelos que utilizan el algoritmo LightGBM los cuales ofrecen resultados ligeramente superiores a los de redes neuronales. Es también destacable que el tiempo invertido en la optimización de los modelos de redes neuronales es inferior al de los modelos de LightGBM.

Comparando con los trabajo relacionados que se han tenido como referencia en el marco teórico, los modelos desarrollados durante este trabajo consiguen combinar varias de las técnicas de estos y obtener unos buenos resultados. Todos los modelos, tanto los mencionados en el marco teórico como los desarrollados en este trabajo obtienen unos valores muy similares en métricas como exactitud o el F1 Score, sim embargo, se debe tener en cuenta la diferencia en el tipo de datos utilizados en cada trabajo. El hecho de conseguir unos resultados igual de buenos con un conjunto de datos tan limitado como el que se ha empleado en este trabajo es uno de los puntos fuertes del trabajo. También existen otros trabajos con los que es muy difícil comparar los modelos pues el enfoque y los resultados que emplean son muy diferentes a los desarrollados en este proyecto.

Relacionado con los objetivos propuestos al comienzo del trabajo, se ha conseguido cumplir gran parte de estos:

- En primer lugar, el objetivo de crear un modelo de aprendizaje automático que prediga con la misma eficiencia que los del marco teórico se ha cumplido, pues nuestro modelo llega a tener una exactitud de hasta el 55% de los partidos.
- En segundo lugar, también se ha logrado implementar diferentes modelos con los dos algoritmos y se ha logrado llegar al modelo que optimiza cada algoritmo, comparando con métricas y resultados los modelos entre ellos.
- Por último, en cuanto a la identificación de las variables más importantes, se ha llevado un análisis con el algoritmo de LightGBM en el que se ha visto qué variables influían en la clasificación con este algoritmo. Por parte de los modelos de redes neuronales, no se ha sido tan preciso en este análisis.

Relacionado con la identificación de variables significativas, en este trabajo se ha llevado a cabo un análisis exploratorio en el que cabe recalcar el uso de datos que no contaban con una distribución normal. Como resultado de este análisis, se

han podido extraer conclusiones sobre ciertas relaciones ocultas o no intuitivas que existen entre ciertas estadísticas de partidos anteriores y el resultado de un partido de fútbol. Estas relaciones nos han ayudado a proponer soluciones para mejorar el rendimiento predictivo de nuestros modelos.

7.1 Limitaciones del trabajo

A lo largo del desarrollo de este proyecto, nos hemos encontrado diferentes dificultades.

La principal dificultad ha sido la dificultad de encontrar un conjunto de datos que pudiera adaptarse a nuestro problema. La mayoría de las fuentes de datos sobre analítica deportiva, hoy en día, son privada o de pago con limitaciones para desarrollar proyectos alrededor de ellas. Una de las tareas que más carga de trabajo ha supuesto ha sido encontrar unas fuentes de datos públicas y que tuvieran información relevante. Además, las fuentes de datos encontradas contaban solamente con datos para las últimas 5 temporadas, por lo que la cantidad de datos era muy limitada.

Otras de las limitaciones del trabajo ha sido la inexperiencia personal en tareas de predicción a futuro (*forecasting*). Esta inexperiencia ha influido en la precisión de las predicciones y en la capacidad para identificar y mitigar posibles errores en los modelos. Sin embargo, este proyecto ha sido una valiosa oportunidad para aprender y mejorar mis habilidades en este campo.

Por último, el hecho de que el fútbol sea un deporte con un alto grado de aleatoriedad e incertidumbre, donde factores imprevistos pueden influir en el resultado de un partido ha sido otra de las limitaciones del proyecto. Variables como el estado físico de los jugadores, las decisiones tácticas del entrenador durante el partido o los eventos fortuitos (como un gol tempranero) pueden alterar significativamente el resultado, lo que es muy difícil mesurar y tener en cuenta a la hora de llevar a cabo las predicciones.

7.2 Trabajo futuro

A lo largo del desarrollo de este proyecto, han ido surgiendo nuevas ideas par formas de plantear este problema, las cuales por diversos motivos no ha sido posible desarrollar. En este apartado se ofrecen algunas ideas de posibles mejoras para futuros trabajos.

En primer lugar, la idea más obvia es tratar de conseguir un conjunto de datos más consistente, con mayor cantidad de datos y otras variables de las que no disponía el utilizado en este proyecto. Algunas de estas variables podrían ser, por ejemplo, las lesiones de cada equipo. Además, con una mayor cantidad de datos podría tratar de medirse el resultado más frecuente en los enfrentamientos entre equipos y utilizarlo para tratar hacer predicciones.

Otra idea que podría realizarse es tratar de analizar y predecir los resultados de los equipos en conjuntos separados. Si bien en este proyecto se ha tratado de

hacer predicciones con un carácter general, podría intentarse dividir los equipos en clústeres (ya sea por nivel de juego, por estilo o incluso por ambas) y ver si se consigue hacer predicciones más precisas para cada grupo de equipos.

Por último, en este proyecto se ha tratado de representar el estado de forma de los equipos haciendo uso de las medias ponderadas. Otro buen indicador del estado de forma o la racha con las que los equipos llegan a los partidos es el estado de ánimo de la afición de cara al partido. Podría intentarse obtener los mensajes que dejan los aficionados de cada equipo los días anteriores a un partido en redes sociales y, utilizando análisis de sentimiento, utilizar estos sentimientos como un medidor del estado de forma de los equipos.

Referencias

Acosta, N. G. (2022). *Analítica de datos y su influencia sobre la gerencia deportiva.*

Baboota, R., & Kaur, H. (2018). *Predictive analysis and modelling football results using machine learning approach for English Premier League.*

Bellman, R. (1957). *Dynamic Programming.* Princeton: Princeton University Press.

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., . . . Cable, N. T. (2017). Monitoring Athlete Training Loads: Consensus Statement.

Chollet, F. (2021). *Machine Learning with Python.* Manning.

Dixon, M. J., & Coles, S. G. (2002). *Modelling association football scores and inefficiencies in the football betting market.*

Dorfer, T. A. (2023). *Medium*. Obtenido de https://pub.towardsai.net/bagging-vs-boosting-the-power-of-ensemble-methods-in-machine-learning-6404e33524e6

Drust, B., Atkinson, G., & Reilly, T. (2007). Future Perspectives in the Evaluation of the Physiological Demands of Soccer.

Heerde, H. J., Hardie, B. G., & Leeflang, P. S. (2015). *Momentum in Soccer: Controlling for Hands of Time.*

Herberger, T. A., & Litke, C. (2021). *The Impact of Big Data and Sports Analytics on Professional Football: A Systematic Literature Review.* Springer.

Hvattuma, L. M., & Arntzen, H. (2010). *Using ELO ratings for match result prediction in association football.*

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.*

Lauzon, F. Q. (2012). *An introduction to Deep Learning.*

Lewis, M. (2003). *Moneyball: The art of winning an unfair game.* WW Norton & Company.

Navarro, S. (2024). *Big Data en el fútbol: Herramientas y aplicaciones.*

Nesti, M. (2010). *Psychology in football: Working with elite and professional players.* Routledge.

Opitz, D., & Maclin, R. (1999). *Popular Ensemble Methods: An Empirical Study.*

Pappalardo, L., Cintia, P., Giannotti, F., & Pedreschi, D. (2019). *A data-driven approach to scout football players based on their career paths.*

Puig, D. (2024). El 'nuevo' Castellón que llega al fútbol profesional... de la mano del Big data.

Rahman, M. A. (2020). *A deep learning framework for football match prediction.*

Rein, R., & Memmert, D. (2016). *Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science.*

Rodrigues, F., & Pinto, Â. (2022). *Prediction of football match results with Machine Learning.*

Shapiro, S. L., Drayer, J., & Dwyer, B. (2016). Examining Consumer Perceptions of Demand-Based Ticket Pricing in Sport. *Sport Marketing Quarterly*, 24-46.

Smagulova, K., & James, A. P. (2019). *A survey on LSTM memristive neural network architectures and applications.*

Soto, P. J. (2013). *Contraste de hipótesis. Comparación de más de dos medias independiente mediante pruebas no paramétricas: Prueba de Kruskal-Wallis.*

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM: a tutorial into Long Short-Term Memory Recurrent Neural Networks.*

Stübinger, J., Benedikt, & Knoll, J. (2019). *Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics.*

Tachis, S., & Tzetzis, G. (2015). The Relationship Among Fans' Involvement, Psychological Commitment, and Loyalty in Professional Team Sports.

Tax, N., & Joustra, Y. (2015). *Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach.*

UNIR. (2024). Obtenido de UNIR: https://www.unir.net/ingenieria/revista/big-data-futbol/

Varikuti, M. (2021). *Medium*. Obtenido de https://mohitv.medium.com/lstm-networks-75d44ac8280f

Wilson, J. (2018). *The Barcelona Legacy: Guardiola, Mourinho and the Fight For Football's Soul.*

Yousef, M., & Allmer, J. (2014). *miRNomics: MicroRNA Biology and Computational Analysis.*

Zhang, C., & Ma, Y. (2012). *Ensemble Machine Learning: Methods and applications.*

Anexo I: Objetivos de desarrollo sostenible

En la siguiente tabla se muestran los diferentes objetivos del desarrollo sostenible junto con la importancia que tiene este trabajo para cada uno de ellos.

	Relación			
Objetivos del desarrollo sostenible	**Alta**	**Media**	**Baja**	**No relacionado**
Fin de la pobreza			X	
Hambre cero				X
Salud y bienestar			X	
Educación de calidad		X		
Igualdad de género			X	
Agua limpia y saneamiento				X
Energía asequible y no contaminante				X
Trabajo decente y crecimiento		X		
Industria, innovación e infraestructura	X			
Reducción de las desigualdades		X		
Ciudades y comunidades sostenibles			X	
Producción y consumo responsable				X
Acción por el clima				X
Vida submarina				X
Vida de ecosistemas				X
Paz, justicia e instituciones sólidas				X
Alianzas para lograr los objetivos			X	

Justificación de la relación de los ODS con el trabajo desarrollado en el TFG:

El objetivo de Industria, innovación e infraestructura es el que más relacionado está con el trabajo desarrollado. La predicción de resultados de fútbol requiere el desarrollo y aplicación de algoritmos de aprendizaje automático. Esto no solo impulsa el desarrollo tecnológico en el ámbito deportivo, sino que también pueden tener aplicaciones en otros sectores industriales.

En relación con el objetivo de educación de calidad, el desarrollo de este tipo de proyectos puede contribuir a la educación de futuros analistas deportivos y a la mejora de la calidad educativa en disciplinas relacionadas con la tecnología y el deporte. Además, la accesibilidad a tecnologías de predicción puede democratizar el acceso al conocimiento y herramientas avanzadas en comunidades menos favorecidas o en equipos con menor presupuestos, lo que está relacionado con el objetivo de reducción de las desigualdades.

Por otra parte, el sector deportivo es un motor importante de crecimiento económico, y las innovaciones tecnológicas en la predicción de resultados pueden crear nuevas oportunidades de empleo en áreas como el análisis de datos, la gestión deportiva, y la tecnología. Proyectos como este podrían influir positivamente en la creación de trabajos decentes y en el fomento de un crecimiento económico

sostenido, especialmente en mercados emergentes donde el fútbol es una industria en crecimiento, lo que está relacionado directamente con el objetivo de trabajo decente y crecimiento económico.

Anexo II

Nombre de la variable	Descripción	Tipo de dato
league	Liga a la que pertenece el partido	Texto
season	Temporada del partido	Texto
game	Fecha y equipos del partido	Texto
team	Nombre del equipo	Texto
player	Nombre del jugador	Texto
jersey_number	Número de camiseta del jugador	Numérico
nation	Nacionalidad del jugador	Texto
pos	Posición del jugador	Texto
age	Edad del jugador	Numérico
game_id	Identificador único del partido	Texto
min	Minutos jugados	Numérico
Performance_Gls	Goles anotados	Numérico
Performance_Ast	Asistencias	Numérico
Performance_PK	Penales marcados	Numérico
Performance_PKatt	Penales intentados	Numérico
Performance_Sh	Disparos	Numérico
Performance_SoT	Disparos a puerta	Numérico
Performance_CrdY	Tarjetas amarillas	Numérico
Performance_CrdR	Tarjetas rojas	Numérico
Performance_Touches	Toques del balón	Numérico
Performance_Tkl	Entradas	Numérico
Performance_Int	Intercepciones	Numérico
Performance_Blocks	Bloqueos	Numérico
Expected_xG	Goles esperados	Numérico
Expected_npxG	Goles esperados sin contar penaltis	Numérico
Expected_xAG	Asistencias esperadas	Numérico
SCA_SCA	Acciones que llevan a disparos	Numérico
SCA_GCA	Acciones que llevan a goles	Numérico
Passes_Cmp	Pases completados	Numérico
Passes_Att	Pases intentados	Numérico
Passes_PrgP	Pases progresivos	Numérico
Carries_Carries	Conducciones	Numérico
Carries_PrgC	Conducciones progresivas	Numérico
Take-Ons_Att	Regates intentados	Numérico
Take-Ons_Succ	Regates exitosos	Numérico
Total_Cmp	Total de pases completados	Numérico
Total_Cmp%	Porcentaje de pases completados	Numérico
Total_Att	Total de pases intentados	Numérico
Total_TotDist	Distancia total abarcada por pases completados	Numérico
Total_PrgDist	Distancia total abarcada por pases progresivos completados	Numérico
Short_Cmp	Pases cortos completados	Numérico
Short_Att	Pases cortos intentados	Numérico
Short_Cmp%	Porcentaje de pases cortos completados	Numérico
Medium_Cmp	Pases a media distancia completados	Numérico
Medium_Att	Pases a media distancia intentados	Numérico
Medium_Cmp%	Porcentaje de pases a media distancia completados	Numérico
Long_Cmp	Pases largos completados	Numérico
Long_Att	Pases largos intentados	Numérico
Long_Cmp%	Porcentaje de pases largos completados	Numérico
KP	Pases clave	Numérico
1/3	Pases en el último tercio de campo	Numérico
PPA	Pases dentro del área rival	Numérico

CrsPA	Centro dentro del área rival	Numérico
Tackles_TklW	Entradas exitosas	Numérico
Tackles_Def 3rd	Entradas en el primer tercio	Numérico
Tackles_Mid 3rd	Entradas en el medio del campo	Numérico
Tackles_Att 3rd	Entradas en el último tercio	Numérico
Challenges_Tkl	Entradas realizadas a delanteros	Numérico
Challenges_Att	Entradas intentadas a delanteros	Numérico
Challenges_Lost	Entradas sin éxito a delanteros	Numérico
Blocks_Sh	Disparos bloqueados	Numérico
Blocks_Pass	Pases bloqueados	Numérico
Err	Errores que llevan a disparo del rival	Numérico
Pass Types_Att	Pases intentados	Numérico
Pass_Types_Live	Pases con el balón en movimiento	Numérico
Pass_Types_Dead	Pases con el balón parado	Numérico
Pass_Types_FK	Pases desde tiro libre	Numérico
Pass_Types_Sw	Pases cruzados por alto	Numérico
Pass_Types_CK	Centros de corner	Numérico
Touches_Def Pen	Toques en el área propia	Numérico
Touches_Def 3rd	Toques en el primer tercio	Numérico
Touches_Mid 3rd	Toques en el centro del campo	Numérico
Touches_Att 3rd	Toques en el último tercio	Numérico
Touches_Att Pen	Toques en el área rival	Numérico
Take-Ons_Succ	Regates exitosos	Numérico
Carries_TotDist	Distancia total recorrida con el balón	Numérico
Carries_1/3	Número de conducciones en el último tercio	Numérico

Tabla 21. Variables del conjunto de datos de FBREF.

Variable	Victoria local	Empate	Victoria visitante
Blocks_Pass_trend_home	7.549524	7.611048	7.544125
Carries_Dis_trend_home	9.079822	9.162299	9.091951
Challenges_Lost_trend_home	9.219529	9.260503	9.070542
Corner Kicks_Str_trend_home	0.074925	0.059863	0.064367
CrsPA_trend_home	1.971871	1.954247	1.992197
Pass Types_Dead_trend_home	38.369822	38.304234	38.548112
Performance_CrdR_trend_home	0.099352	0.094959	0.095126
Performance_Tkl_trend_home	16.225437	16.301305	16.065807
Tackles_Mid 3rd_trend_home	6.304781	6.294067	6.358108
Tackles_Tkl_trend_home	16.225437	16.301305	16.065807
Tackles_TklW_trend_home	9.593898	9.638919	9.492898
Take-Ons_Tkld_trend_home	6.879123	6.918108	7.038680
Touches_Def 3rd_trend_home	156.447663	156.004122	156.960073
Blocks_Pass_trend_away	7.617517	7.592710	7.648831
Carries_Dis_trend_away	9.125987	9.144316	9.138800
Challenges_Lost_trend_away	8.874345	8.971225	9.024376
Corner Kicks_Str_trend_away	0.075696	0.063913	0.075969
CrsPA_trend_away	2.202585	2.131512	2.128399
Pass Types_Dead_trend_away	39.307384	39.154511	39.003231
Performance_CrdR_trend_away	0.086300	0.089673	0.094216
Performance_Tkl_trend_away	15.935174	16.105973	16.110208
Tackles_Mid 3rd_trend_away	6.490818	6.473880	6.388767
Tackles_Tkl_trend_away	15.935174	16.105973	16.110208
Tackles_TklW_trend_away	9.458979	9.581510	9.539620
Take-Ons_Tkld_trend_away	7.161902	7.038946	7.036262
Touches_Def 3rd_trend_away	152.799917	152.174517	152.506030
Blocks_Pass_trend_home	7.549524	7.611048	7.544125

Carries_Dis_trend_home	9.079822	9.162299	9.091951

Tabla 22. Medias de las variables no significativas según el test de Kruskal-Wallis para los distintos grupos.