COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

APUESTAS DEPORTIVAS COMO POTENCIAL ACTIVO FINANCIERO

Autor: Diego Urbano García
Director: Alejandro Rodríguez Gallego

MADRID | Junio 2023

RESUMEN

El rápido crecimiento de la industria de las apuestas deportivas ha motivado esta exhaustiva investigación sobre su potencial como activo financiero en términos de la relación riesgo-rentabilidad. Este trabajo de investigación explora esta idea en el contexto de las apuestas de La Liga. Para investigar esta hipótesis, se utilizaron diversas fuentes de datos públicos y se aplicaron técnicas de transformación de variables para extraer información relevante. Se desarrollaron modelos explicativos para comprender los factores subyacentes que influyen en los resultados de los partidos, seguidos de la construcción de modelos predictivos utilizando algoritmos diversos. Estos modelos predictivos sirvieron como base para la construcción de dos simulaciones financieras, una para victorias locales y otra para victorias visitantes. Las simulaciones tuvieron como objetivo evaluar la rentabilidad, para posteriormente determinar la dinámica de riesgo-rentabilidad de las apuestas deportivas, con un enfoque específico en su comparación con activos financieros convencionales como bonos y acciones.

Los resultados de esta investigación indican que las apuestas deportivas pueden generar beneficios. Sin embargo, al considerar la relación riesgo-rentabilidad, su desempeño es inferior en comparación con los activos financieros convencionales. El análisis de las simulaciones financieras revela que las ganancias generadas por las apuestas deportivas no compensan adecuadamente los riesgos asociados, especialmente en comparación con el desempeño de bonos y acciones. Esta investigación destaca la importancia de considerar las características de riesgo y rentabilidad al evaluar la idoneidad de las apuestas deportivas como opción de inversión.

PALABRAS CLAVE

Apuestas deportivas, activos financieros, modelos predictivos, aprendizaje automático, simulación de inversión, rentabilidad, riesgo.

ABSTRACT

The rapid growth of the sports betting industry has prompted this extensive examination of its potential as a financial asset in terms of risk-return tradeoff. This research paper explores this notion within the context of La Liga football bets. To investigate this hypothesis, various public data sources were utilized, and feature engineering techniques were applied to extract relevant information. Explicative models were developed to understand the underlying factors influencing match outcomes, followed by the construction of predictive models using diverse algorithms. These predictive models served as the foundation for building two financial simulations, one for home wins and another for away wins. The simulations aimed to assess the profitability, to later determine the risk-return dynamics of football bets, with a specific focus on their comparison to conventional assets like bonds and shares.

The findings of this research indicate that football bets can indeed yield profits. However, when considering the risk-return tradeoff, their performance falls short in comparison to traditional financial assets. The analysis of the financial simulations reveals that the returns generated by football bets may not adequately compensate for the associated risks, particularly when compared against the performance of bonds and shares. This research highlights the importance of considering risk and return characteristics when evaluating the suitability of football bets as an investment option.

KEYWORDS

Sports betting, financial asset, predictive model, machine learning, financial simulation, return, risk.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Introducción

Activos financieros y apuestas deportivas

En el ámbito financiero, se considera un activo a aquel instrumento o bien que posee un valor económico y puede generar rendimientos o flujos de efectivo futuros. Los activos financieros más comunes incluyen acciones, bonos, derivados o fondos de inversión. Estos activos son ampliamente estudiados y utilizados por inversores para diversificar sus carteras y lograr retornos en al inversión. Según Fama y French, los activos financieros se caracterizan por su capacidad para generar rendimientos esperados, reflejados en cambios en su valor o en pagos futuros, y por la existencia de un mercado en el que se pueden comprar y vender (Fama & French, 2004). Además, se espera que los activos financieros estén respaldados por un marco legal y regulador que garantice su funcionamiento adecuado y su liquidez.

Se ha planteado la posibilidad de que las apuestas deportivas puedan tener ciertas similitudes con los derivados financieros. En esta analogía, el evento deportivo sería el equivalente al subyacente en un derivado financiero, con los inversores especulando sobre el resultado de dicho evento. Sin embargo, es crucial señalar que las apuestas deportivas no cumplen con todos los criterios y características de los derivados financieros. Los derivados financieros se caracterizan por ser contratos estandarizados que se negocian en mercados regulados y organizados. En contraste, las apuestas deportivas se basan en eventos inciertos y no se negocian en un mercado de este tipo. Además, no están respaldadas por un marco legal y regulador comparable al de los activos financieros. Aunque las apuestas deportivas pueden generar rendimientos, su naturaleza y funcionamiento las diferencian notablemente de los activos financieros convencionales. Por lo tanto, a pesar de algunas similitudes superficiales con los derivados financieros, las apuestas deportivas no cumplen con los criterios convencionales para ser consideradas como un activo financiero.

Tomando como premisa que las apuestas deportivas no pueden ser consideradas como activos financieros por su inherente naturaleza, el propósito de este estudio consiste en investigar la posibilidad de considerar las apuestas deportivas como activos financieros, basándose en la evaluación de los rendimientos y riesgos que estas generan.

El sector de las apuestas deportivas y su funcionamiento

El sector de las apuestas deportivas en España ha presenciado un crecimiento significativo en años recientes, con un impulso marcado por el creciente interés en los deportes y la digitalización de los servicios de apuestas. Según un informe publicado por la Dirección General de Ordenación del Juego[1], el *Gross Gaming Revenue*, es decir, la cantidad total de dinero apostado por los jugadores menos los pagos de ganancias, experimentó un aumento de aproximadamente un 110% en el último trimestre de 2022 en comparación con el mismo período del año anterior. El fútbol se ha consolidado como el deporte que acapara la mayor cantidad de apuestas en España, tanto en términos de volumen de apuestas como de ingresos, representando el 77% de todas las apuestas deportivas en el país. Los actores más destacados en el sector de las apuestas deportivas en el mercado español incluyen grandes corporaciones de juego y apuestas internacionales como Bet365, William Hill y Bwin, así como operadores locales como Codere, Luckia y Sportium.

Para determinar las cuotas en las apuestas deportivas, las casas de apuestas utilizan modelos matemáticos y algoritmos que consideran una variedad de factores. Estos factores incluyen el rendimiento histórico de los equipos o jugadores, o las condiciones específicas del evento deportivo. Las casas de apuestas asignan un valor numérico a cada posible resultado del evento y expresan las cuotas en diferentes formatos, como decimal, fraccional o americano. Estas cuotas reflejan la probabilidad de que ocurra un resultado específico según el análisis de la casa de apuestas. Sin embargo, es importante destacar que las probabilidades implícitas calculadas a partir de las cuotas no representan una probabilidad exacta. Las casas de apuestas incorporan un margen de ganancia conocido como *vig*[2] o *juice*. Este margen se añade a las cuotas y garantiza beneficios para la casa de apuestas a largo plazo.

En España, las cuotas se expresan comúnmente en formato decimal, y se pueden transformar en probabilidades implícitas. Este proceso de conversión es relativamente sencillo y se rige por la fórmula, “Probabilidad Implícita = 1/Cuota Decimal”. Para ilustrar esta conversión, considérese el caso en el que la cuota decimal para la victoria de un equipo específico es de 2. Aplicando la fórmula previamente mencionada, se obtiene que la probabilidad implícita es de 1/2, lo que equivale a 0,5 o 50%. Este resultado indica

[1] https://www.ordenacionjuego.es/es/4_Informe_Trimestral_2022
[2] https://www.legalsportsreport.com/sports-betting/vigorish/

que, según la evaluación de la casa de apuestas, dicho equipo posee una probabilidad de victoria del 50%. Es crucial, sin embargo, reconocer que estas probabilidades implícitas también incorporan el *vig* o *juice* previamente mencionado, por lo que al sumarse las cuotas de los diferentes resultados de un partido se obtiene valores superiores al 100%.

Las cuotas en las apuestas deportivas representan la cantidad de euros que se recibirán por cada euro apostado en caso de una predicción acertada. Sin embargo, es fundamental comprender que si la apuesta resulta errónea, el apostante pierde la totalidad del monto apostado, subrayando así el carácter intrínsecamente arriesgado de este tipo de inversión.

Finalmente, es relevante subrayar que el objeto de estudio de esta investigación son las apuestas relacionadas con La Liga de Fútbol Profesional, la máxima competición futbolística a nivel nacional en España. La Liga comprende a 20 equipos, incluyendo entidades como el Real Madrid, Fútbol Club Barcelona y Atlético de Madrid, y se extiende a lo largo de 38 jornadas.

Revisión de la literatura

Dada la naturaleza de las apuestas deportivas, asociadas al juego y causantes de adicciones, no se trata de un tema que no ha sido objeto de numerosas investigaciones con este enfoque financiero, lo que supone que la literatura existente es algo limitada en comparación con temas más recurrentes en la literatura académica.

Aun así, se puede encontrar literatura que trata diferentes aspectos de las apuestas deportivas. Existen numerosos estudios de carácter social o psicológico, como el de Hing, Russell y Brown en el que se realizan encuestas a apostantes australianos para determinar factores de riesgo demográficos, de comportamiento y psicológicos, encontrando que los factores de riesgo más comunes en apuestas deportivas incluyen ser hombre, joven, tener bajos ingresos o haber nacido fuera de Australia, entre otros (Hinge, Russell & Browne, 2017). Otro estudio en este ámbito psicológico es el publicado por la *American Psychological Association* en el cual se estudia si los apostantes expertos en hockey pueden hacer predicciones superiores a las del azar tratando de determinar qué tipo de estrategias utilizan y si pueden obtener mayores ganancias monetarias que las del azar. Los resultados de este experimento sugieren que las llamadas "habilidades" de los apostadores deportivos son distorsiones cognitivas, lo cual opone un reto para la presente investigación (Cantinotti, Ladouceur, & Jacques, 2004).

También se encuentran artículos que analizan el fenómeno desde un punto de vista de análisis de mercado, como hacen López-González y Griffiths en su artículo que explora la integración de las apuestas deportivas en línea dentro de los sectores digital, deportivo y de juego, examinando cómo los mercados de datos, los *eSports*, los deportes virtuales, los juegos sociales, las herramientas de realidad inmersiva, los medios deportivos, y otros más están convergiendo todos alrededor de la actividad de apuestas (López-González & Griffiths, 2018). Griffiths cuenta con múltiples otros artículos en los que analiza contenido publicitario de casas de apuestas, en el que concluye que la amistad y el humor son las narrativas más recurrentes (Killick & Griffiths, 2022), o en los que trata las apuestas deportivas en vivo, afirmando que tienen el potencial de ser más perjudiciales que las apuestas tradicionales, debido a las características estructurales inherentes de las apuestas en vivo (Killick & Griffiths, 2019).

En cuanto a lo que se refiere estrictamente a la capacidad predictiva para eventos deportivos, el mayor éxito en este campo de investigación pertenece a las casas de

apuesta, al contar con una cantidad de información y un poder predictivo sobresalientes, que no están al alcance de investigadores. De todos modos, al existir ciertas bases de datos públicas, como las que van a ser a ser empleadas en esta investigación, podemos encontrar trabajos que tratan de predecir resultados de fútbol.

Igualmente se encuentran artículos académicos enfocados en la predicción de resultados de diversas competiciones de fútbol. En la primera fuente consultada, Ulmer y Fernández de la Universidad de Stanford utilizaron datos de 10 temporadas previas para predecir los resultados de las temporadas 2012/2013 y 2013/2014 de la English Premier League (Ulmer & Fernández, 2014). Sus variables de análisis incluyeron equipo local, equipo visitante, estado de forma de cada equipo y una variable de calificación cuyo método de obtención no fue revelado. A pesar de la utilización de 6 técnicas de clasificación, *Baseline*, *Naive Bayes*, *Hidden Markov Model*, *Support Vector Machine*, *Random Forest* y *One-vs-all Stochastic Gradient Descent*, se encontró que la predicción de empates era un desafío debido a su menor frecuencia. El modelo *One-vs-all* tuvo la mayor precisión, pero una baja eficacia en la predicción de empates. Los modelos *Random Forest* y *Support Vector Machine*, aunque con porcentajes de error ligeramente más altos, demostraron un rendimiento superior en la predicción de empates. A pesar de la limitación de datos, el estudio fue considerado exitoso. Sin embargo, la precisión de entre el 48% y 50% son inferiores a las obtenidos por otros investigadores.

Un tema recurrente en la literatura es la dificultad en la predicción de los empates. Una investigación aborda este desafío analizando la entropía de los partidos de la temporada 2012/2013 de la *English Premier League*, obteniendo un resultado de entropía de 0.9789, cercano a 1, lo que indica una alta aleatoriedad y dificultad en la predicción. Los datos utilizados provenían del portal Footballdata.co.uk y se limitaban a 3 variables: goles anotados, tiros a puerta y saques de esquina. A partir de estas variables, formaron 2 vectores, *el K-Past Performance* y el *Temporal Gradient K-Past Performance*, que sirvieron para la construcción de múltiples modelos con variaciones en los datos de entrada, el número de k y el algoritmo de predicción. Los mejores resultados se obtuvieron con el vector *Temporal Gradient K-Past Performance*, una *k* con valor de 7 y el algoritmo *Support Vector Machine* con el kernel *Radial Basis Function*, alcanzando una precisión general del 66%. Sin embargo, la precisión en la predicción de empates fue de solo 54%, y el *recall* de 23%. Aunque estos resultados no se consideraron exitosos, la muestra de evaluación puede haber sido insuficiente para extraer conclusiones definitivas.

Más en concreto, existen investigaciones en la literatura no sólo centradas en prestaciones predictivas, sino que también añaden el objetivo de conseguir ganancias mediante apuestas deportivas. En este campo específico se encuentra El estudio de Rodrigues y Pinto, que analizó 1.900 partidos de la *English Premier League* entre las temporadas 2013/2014 y 2018/2019 (Rodrigues & Pinto, 2022), incorporando datos del portal Footballdata.co.uk y del videojuego FIFA. Estos incluían estadísticas a nivel de equipo y de jugadores individuales, con un énfasis especial en los factores locales y visitantes.

Después de normalizar los datos, los autores seleccionan las variables más relevantes mediante un algoritmo denominado *Boruta*, que se basa en un *Random Forest*. Se encuentra que las cuotas de la casa de apuestas Bet365, la calificación general del equipo visitante y las victorias del equipo local en sus últimos partidos como las más significativas. Experimentaron con varios algoritmos, siendo el *Support Vector Machine*, una vez más, el que proporcionó la mejor precisión, con un 61%, aunque el desempeño en la predicción de empates fue bajo, con solo un 3% de acierto. Tras una nueva selección de variables mediante la técnica *Backwards Feature Selection*, y pruebas con combinaciones de estas variables selccionadas, la precisión aumentó hasta un 64% en los casos de *Support Vector Machine, Random Forest* y *Xgboost*. Asimismo, se observó un mejor rendimiento en la predicción de empates, alcanzando alrededor del 30%. El modelo con mayor rentabilidad fue el *Random Forest*, logrando un 26% de margen de beneficio.

La siguiente investigación consultada se centra en el desarrollo de un Sistema de Soporte de Decisión para minimizar el riesgo en apuestas deportivas (Gomes, Portela & Santos, 2021). Al igual que en otros estudios de la literatura, las variables son transformadas para su inclusión en los modelos, utilizando únicamente datos disponibles antes del inicio de los partidos. Las estadísticas transformadas incluyen el promedio de goles, tiros y victorias en partidos recientes, además de variables que determinan el número de victorias del equipo local y visitante en sus últimos 5 enfrentamientos. Para el análisis, se utilizaron 3 técnicas de *Machine Learning* y 2 técnicas de muestreo. Entre los 6 modelos generados, el que demostró un mejor rendimiento fue el *Support Vector Machine* con una partición porcentual, logrando una precisión del 51%. Posteriormente, se desarrolló el Sistema de Soporte de Decisión utilizando *Exsys Corvid*, una herramienta integrada en *WEKA*, y se crearon 4 bloques lógicos. El último bloque relaciona las puntuaciones obtenidas previamente para cada equipo y sugiere la mejor apuesta a realizar.

Finalmente, se validó el Sistema de Soporte de Decisión en 7 jornadas de la competición inglesa, es decir, 70 partidos. Siguiendo las indicaciones del sistema y apostando 100 euros en cada partido, se obtuvo un beneficio de 1.409 euros, un retorno del 20% de la inversión inicial de 7.000 euros. El rango de precisión en las apuestas varió entre el 30% y el 80% por jornada.

Habiendo revisado la literatura existente en el campo de la predicción de apuestas de fútbol, la contribución de este estudio reside en tratar las apuestas como si fueran un activo financiero convenvional. Si bien la construcción de un modelo que busque predecir correctamente resultados de partidos de fútbol no es una novedad, ni tampoco lo es el acto de hacer apuestas basadas en estos modelos, la innovación de este trabajo radica en desarrollar una simulación financiera rigurosa para luego tratar de medir con precisión y objetividad el riesgo inherente a estas apuestas. Posteriormente, se compara el rendimiento de estas apuestas con ciertos activos financieros. Este enfoque innovador ayudará a responder a la pregunta de si es posible realizar apuestas basándose en información abiertamente disponible y obtener un retorno razonable en relación con el riesgo asumido.

Además, este estudio propone un enfoque innovador para abordar el desafío de predecir empates en los partidos. Se construyen 2 modelos distintos, uno centrado en la predicción de victorias locales y otro en victorias de equipos visitantes. Este planteamiento reduce la exposición a las imprecisiones en la predicción de empates, que representan una amenaza potencial para la rentabilidad obtenida. Esta metodología diferenciada agrega una nueva dimensión a los métodos de predicción en el ámbito de las apuestas deportivas.

Metodología

El presente trabajo se centra en el análisis de las apuestas deportivas como posible activo financiero. La idea consiste en determinar si es posible obtener rentabilidades comparables a otros activos financieros, como las acciones o ciertos bonos, con un riesgo también comparable. Es importante tener en cuenta que, por definición, las apuestas no son un activo financiero real y que, por tanto, no están sujetas a las mismas reglas y controles que los productos financieros. No obstante, debido a la popularidad y accesibilidad de las apuestas deportivas, puede resultar interesante analizarlas como una posible opción de inversión.

Es importante señalar que el riesgo de las apuestas deportivas es muy elevado, ya que se trata de una actividad que implica la posibilidad de perder todo el dinero apostado en caso de que no se acierte el resultado. De hecho, el riesgo es tan notable que puede resultar difícil considerar las apuestas deportivas como una opción de inversión viable a largo plazo. No obstante, como compensación a esta pérdida total en caso de fallo, se encuentran apuestas con una probabilidad alta de que suceda el evento deseado y que pagan cuotas entre 1,10 y 1,20 por euro apostado. Consecuentemente, si se aplica una estrategia adecuada y se realiza un seguimiento riguroso de las estadísticas y resultados, es posible minimizar el riesgo y obtener rentabilidades aceptables. A esto hay que sumarle el hecho de que las apuestas deportivas no están expuestas a ciclos macroeconómicos, lo que podría suponer un factor interesante.

El objetivo final de este trabajo es analizar el potencial de las apuestas deportivas como opción de inversión y, en caso de que se obtengan resultados positivos, proponer una estrategia de inversión que permita maximizar las rentabilidades y minimizar los riesgos. Es importante tener en cuenta que se trata de un análisis experimental y que, por tanto, los resultados obtenidos no deben tomarse como una garantía de éxito. No obstante, se considera que este tipo de estudios pueden resultar interesantes para aquellos inversores que buscan diversificar sus carteras y explorar nuevas oportunidades de inversión.

En este trabajo se aborda la construcción de 2 modelos independientes, ambos de clasificación binaria, los cuáles se van a evaluar separadamente. A diferencia de un modelo de clasificación triple, donde se contemplan 3 opciones: victoria, derrota o empate, un modelo binario únicamente contempla 2 opciones, que variarán en función de cada uno de los dos mencionados modelos. Esta simplificación es debida a que la opción

de empate introduce una complejidad adicional en el modelo, ya que la probabilidad de un empate puede ser más difícil de predecir que la de una victoria o una derrota.

Durante las últimas 16 temporadas de La Liga, desde la 2006/2007 a la 2021/2022, el porcentaje de empates es menor del 25%, siendo las victorias visitantes el siguiente evento menos común superando el 28%. Como consecuencia, esto hace que se disponga de menos observaciones de empates para entrenar el modelo, lo que, a su vez dificulta la predicción precisa de su ocurrencia. Además, al revisar la literatura existente, se ha observado que diversos estudios han coincidido en que predecir los empates es uno de los mayores impedimentos a la hora de construir modelos predictivos precisos.

La complejidad adicional introducida por la opción del empate afecta también al proceso de *reporting* y evaluación del modelo, dado que muchas métricas de evaluación, como la matriz de confusión, el área debajo de la curva o el F1, pasan a ser más complejas y menos intuitivas. Es importante tener en cuenta esta complejidad adicional en la construcción del modelo, así como en la evaluación de su desempeño, para poder obtener resultados confiables y efectivos.

El primero de los modelos anteriormente mencionados es el modelo que predice victorias locales. Las 2 categorías existentes, por tanto, serán victoria local y no victoria local, que engloba el empate y la victoria visitante. Este modelo se caracteriza por su enfoque conservador, el cual se basa en la premisa de que la victoria local es el resultado más frecuente en encuentros la competición de La Liga, como se acaba de demostrar. Es importante destacar que, a pesar de que se espera una mayor precisión en este modelo, debido a la mayor ocurrencia de victorias locales, siempre deberá ser comparado con el modelo *naive*, de forma que se pueda determinar si ofrece mejores predicciones que la simple predicción de clase mayoritaria. Por otra parte, se esperan cuotas moderadas, ajustadas a esta mayor frecuencia de las victorias locales, pero una mayor tasa de acierto en comparación con el modelo de predicción de victorias visitantes.

El segundo modelo de predicción se enfoca en la victoria visitante, y considera 2 clases, victoria visitante y no victoria visitante, que incluye tanto el empate como la victoria local. Este modelo se diferencia del anterior por su enfoque oportunista, en el que se buscan explotar oportunidades en las que exista cierta certeza, atendiendo a la probabilidad dictada por el modelo, de que el equipo visitante saldrá victorioso en el encuentro deportivo. Es importante tener en cuenta que, a diferencia del modelo

conservador anterior, se espera una menor tasa de acierto en este modelo, lo que puede deberse a la menor ocurrencia de las victorias visitantes y a la dificultad de predecir estos resultados. No obstante, las cuotas más altas ofrecen una mayor recompensa económica en caso de que la predicción resulte acertada. Al igual que en el modelo anterior, se debe comparar la precisión de este modelo con el modelo *naive*, de forma que se pueda determinar si ofrece mejores predicciones que la simple predicción de la clase mayoritaria.

Se procederá a la construcción de modelos explicativos inicialmente, con la finalidad de obtener un mayor entendimiento sobre los patrones que rigen las victorias locales y visitantes. Posteriormente, se crearán modelos predictivos utilizando distintos algoritmos, cuyo principal propósito es alcanzar la mayor eficacia predictiva posible. Una vez desarrollados estos modelos, se llevará a cabo la evaluación de su rendimiento, permitiendo de esta manera identificar cuál de los modelos proporciona un mejor desempeño.

Con los modelos seleccionados, se desarrollará una simulación financiera en la que se realizarán apuestas basadas en las predicciones de dichos modelos. Esta simulación tomará la forma de un *backtesting*, es decir, un proceso que evalúa la efectividad de una estrategia o modelo predictivo basándose en datos históricos. Este término se utiliza comúnmente en los campos de las finanzas y la economía, especialmente en el comercio de activos financieros.

La estrategia definida se implementará durante las últimas temporadas disponibles, asegurándose de que estos datos no hayan sido utilizados para el entrenamiento de los modelos, ya que esto representaría un error significativo. Este *backtesting* permitirá determinar si la estrategia desarrollada resulta rentable, y posteriormente permitirá establecer una comparación en términos de retorno con una serie de activos financieros ordinarios.

Fuentes y extracción de datos

Fuentes de datos

Las fuentes empleadas para construir la base de datos de este trabajo son 2, Footballdata.co.uk [3] y SoFifa[4]. Antes de introducir ambos sitios web, destacar que, durante el proceso de investigación y recopilación de datos, se identificaron diversas fuentes de información que podrían haber sido utilizadas para mejorar la precisión de los modelos. Sin embargo, la mayoría de estas fuentes no fueron utilizadas por diversas razones. Por ejemplo, algunas fuentes requerían una subscripción de pago y, por lo tanto, no se consideró viable su uso en este proyecto. Este es el caso de StatsBomb[5] o Opta Sports[6]. Otras fuentes, como Transfermarkt[7], presentaban dificultades para obtener la información de manera eficiente y precisa. Aunque esta fuente de información podría haber añadido más datos al conjunto ya existente, se decidió que el esfuerzo necesario para recopilar y procesar estos datos no justificaba la posible mejora en la precisión de los modelos.

Primeramente, el sitio web Footballdata.co.uk proporciona una fuente rica de datos para investigadores y analistas interesados en estudiar partidos y ligas de fútbol en todo el mundo. El sitio web recopila y publica datos sobre diversos aspectos de los partidos de fútbol, incluyendo los marcadores, los goles, los tiros a puerta, la posesión y las cuotas de apuestas de diferentes casas de apuestas, las cuales son una fuente valiosa de información debido al conocimiento implícito que incluyen de las predicciones de las casas de apuestas. Los datos están disponibles en un formato estructurado y fácilmente accesible, lo que facilita a los investigadores su uso para sus estudios. Sin embargo, es necesario tener en cuenta que los datos pueden tener limitaciones, como pueden ser datos incompletos o valores faltantes para ciertas variables. Estas limitaciones y posibles sesgos pueden afectar la calidad y precisión de los resultados obtenidos a partir de los datos, y serán presentadas en profundidad posteriormente, al igual que la manera de lidiar con ellas.

[3] https://www.football-data.co.uk/spainm.php
[4] https://sofifa.com/teams?type=all&lg%5B%5D=53
[5] https://statsbomb.com/es/
[6] https://www.statsperform.com/opta/
[7] https://www.transfermarkt.es/

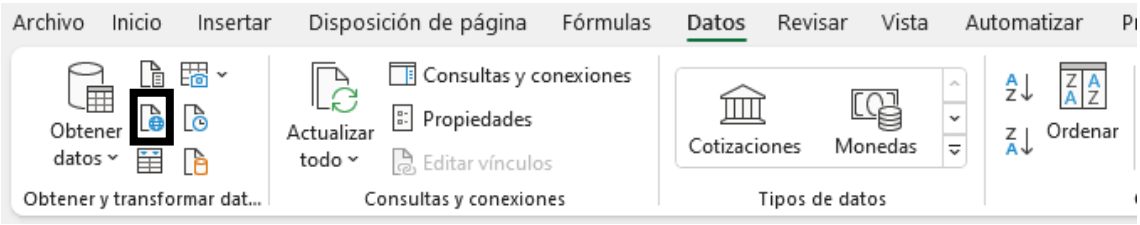
Por otro lado, SoFifa es una plataforma en línea que ofrece una amplia variedad de datos sobre el videojuego FIFA. FIFA es un popular videojuego de simulación de fútbol en el que los jugadores pueden crear y administrar equipos, jugar partidos y competir en ligas y torneos. Las valoraciones del FIFA pueden proporcionar una buena aproximación de la calidad y desempeño de un equipo. SoFifa utiliza datos del FIFA para proporcionar información detallada sobre equipos y jugadores de fútbol. Aunque SoFifa incluye datos de jugadores individuales, en este trabajo solo se han utilizado las valoraciones de los equipos, que son *overall*, defensa, medio del campo y ataque. La recopilación de datos detallados sobre cada jugador de cada equipo requeriría conocer los partidos jugados por cada uno, lo que resulta una tarea muy compleja. Además, las valoraciones de los equipos se actualizan y ajustan a lo largo de una misma temporada, en función del rendimiento del equipo. En este trabajo, se han descargado actualizaciones de forma mensual siempre que se encontrasen disponibles.

Extracción de datos

La recolección de datos proveniente de Footballdata.co.uk se ejecuta de forma bastante accesible. Se ingresa al sitio web, se selecciona la competición de interés, en este caso La Liga, y dentro de dicha sección, se pueden descargar los datos de cada temporada a través de hipervínculos en formato CSV. Dada la simplicidad y rapidez de esta operación, no se consideró necesario automatizar la extracción de datos. Sin embargo, al integrar los archivos CSV de todas las temporadas, se detectaron ciertas inconsistencias en la cantidad de variables dependiendo de la temporada, lo cual demandó un examen detallado y una correspondiente limpieza y preparación de datos previo a la integración completa.

Por otro lado, la obtención de datos de SoFIFA resultó más compleja, puesto que el sitio web no proporciona opciones de descarga directa. Frente a esta situación, se consideró apropiado desarrollar un código en Python para realizar *web scraping*, una técnica utilizada para extraer información de sitios web de manera automatizada. Después de diversos intentos, se concluyó que el *web scraping* no era viable, presumiblemente debido a las políticas de uso de la página. Por lo tanto, se adoptó un enfoque alternativo utilizando una función incorporada en Excel que permite extraer información de páginas web mediante su URL.

Figura 1- Función para obtener datos de web en Excel



Fuente: Excel

Aunque este proceso implicó un considerable esfuerzo, especialmente en términos de tiempo, emergió como la única opción viable para obtener los datos del videojuego FIFA. Se repitió esta operación alrededor de 10 veces por temporada, dado que se buscaban actualizaciones mensuales de las calificaciones de los equipos, lo cual supuso una fuerte inversión de tiempo al considerar que la obtención no es inmediata y requiere unos minutos para cada URL. Es relevante indicar que, durante la extracción de datos, se añadió una variable de fecha para facilitar la identificación y la posterior integración con la otra fuente de datos. Como la estructura y cantidad de variables se mantuvieron consistentes a lo largo de todas las temporadas, se pudo consolidar todos los datos en un solo archivo. Este proceso de integración, junto con algunas modificaciones como la eliminación de variables innecesarias o la reorganización del orden de las variables, se efectuó mediante *Power Query*[8], una herramienta de análisis de datos de Excel que facilita la importación, limpieza, transformación y preparación de datos para su análisis.

[8] https://support.microsoft.com/es-es/office/acerca-de-power-query-en-excel-7104fbee-9e62-4cb9-a02e-5bfb1a6c536a

Descripción de la muestra

La base de datos utilizada para el desarrollo de los modelos comprende partidos de la competición de la liga española desde la temporada 2006/2007 hasta la 2021/2022. Si bien Footballdata.co.uk proporciona información desde la temporada 1993/1994, la disponibilidad de datos de SoFifa comienza en la temporada 2006/2007. Considerando que cada temporada consiste en 38 jornadas, con 10 partidos por jornada, lo que equivale a 380 partidos por temporada, la muestra final suma un total de 6.080 partidos.

Es importante señalar de nuevo que el número de variables derivadas de las descargas de Footballdata.co.uk varía en función de la temporada, principalmente debido a las variables relacionadas con las cuotas de apuestas. Este aspecto será discutido con mayor detalle en la sección de tratamiento de datos. Sin embargo, las variables referentes a las estadísticas de los partidos permanecen constantes. Estas incluyen goles al descanso, goles al final del partido, tiros, tiros a puerta, faltas, saques de esquina, tarjetas amarillas y tarjetas rojas, cada una de ellas tanto para el equipo local como para el visitante. Adicionalmente, existe una variable categórica que refleja el resultado del partido, la cual es de suma importancia dado que representa la variable objetivo a predecir por los modelos.

En relación con los datos obtenidos de SoFifa, las variables disponibles incluyen el nombre del equipo, la valoración general, la valoración de ataque, la valoración del centro del campo, la valoración de defensa y una variable de fecha que indica la fecha de actualización de las valoraciones.

Procesamiento de los datos

Tratamiento de variables de cuotas de apuestas

A continuación, se va a detallar el tratamiento y procesamiento de los datos de cara a ser aportar la máxima utilidad y nivel de información a la hora de la construcción de los modelos.

En cuanto a los datos provenientes de Footballdata.co.uk, no se observa ningún valor faltante en las variables relacionadas con las estadísticas de los encuentros como goles, tiros, saques de esquina o tarjetas, entre otras. En cambio, los datos referidos las cuotas de las casas de apuestas sí se encuentras problemas. Primeramente, el número de variables ofrecidas en relación con cuotas de apuestas varían dependiendo de la temporada, siendo más limitadas en las temporadas más antiguas. Adicionalmente, existen valores *NA* y valores atípicos en estas variables de apuestas.

En los datos de las primeras temporadas de la muestra los datos relacionados con las casas de apuestas se limitan a cuotas de victoria local, empate y victoria visitante ofrecidas por diferentes casas de apuestas. En temporadas posteriores se añaden variables vinculadas a BetBrain[9], un sitio web que ofrece servicios de comparación de cuotas y pronósticos deportivos para apuestas deportivas. Estas nuevas variables incorporadas están relacionadas con el aspecto de agregador de cuotas de BetBrain, con variables como mayores, menores y medias de las cuotas ofrecidas por cualquier casa de apuestas, variables relacionadas con las cuotas ofrecidas de número de goles en los encuentros y variables relacionadas con diferentes modalidades de apuestas, como puede ser el hándicap asiático. Por último, en las 2 últimas temporadas, se ofrecen variables relacionadas también con el número de goles y con hándicap asiático, pero por casas de apuestas individualmente, en vez de valores máximos, mínimos y medios como se observa en las temporadas previas. Para lidiar con dicha incoherencia presente en los datos de las diferentes temporadas, se ha optado por un acercamiento prudente, manteniendo aquellas variables que están presentes a lo largo de todas las temporadas, y descartando aquellas que no lo están, dado que la imputación de valores no es viable.

Al examinar las variables vinculadas a cuotas de apuestas que se han mantenido, se observa que existen valores *NA* en las cuotas ofrecidas por las 9 casas de apuestas, lo cual

[9] https://www.betbrain.com/football-odds/today

supone un problema ya que no se pueden entrenar modelos con este tipo de valores. Más concretamente, entre las cuotas ofrecidas por William Hill se observa que hay más de 500 observaciones con NAs, por lo que se prescinde de las variables de cuotas de victoria local, empate y victoria visitante vinculadas a esta casa de apuestas. En cuanto a las casas restantes, ninguna supera los 15 *NAs*. Para evitar tener que eliminar las observaciones con *NA* que quedan, se opta por calcular la media y las desviación estándar para las cuotas disponibles en cada partido, obviando los valores *NA* en el cálculo. Posteriormente se eliminan todas las variables de las cuotas específicas, usadas para el cálculo, y se mantienen solo estas 6 nuevas variables, que son la media y la desviación estándar para victoria local, empate y victoria visitante. Adicionalmente, esta decisión de reducir el número de variables relacionadas con datos de casas de apuestas reduce el potencial riesgo de *overfitting* a la hora de entrenar el modelo.

Una vez calculadas la media y la desviación estándar para todos los partidos se advierte que existen valores desproporcionados e ilógicos en algunas observaciones. Al estudiar la causa de dichos valores se observa que existen valores atípicos entre las cuotas provenientes de la casa de apuestas Sportingbet. Este inconveniente está relacionado con el separador decimal, que no es leído correctamente. A modo de ejemplo, hay una observación en la que la cuota debería ser 1,111, pero aparece como 1.111, desvirtuando la media y la desviación estándar. Por este motivo, se decide suprimir las 3 variables relacionadas con Sportingbet como se ha hecho previamente con William Hill, ya que se sigue contando con 7 casas de apuestas para los cálculos de media y desviación estándar.

Una vez llevadas a cabo estas modificaciones se obtiene una base datos que integra toda la información de interés proveniente de Footballdata.co.uk, y que podrá ser unida a los datos obtenidos de SoFifa.

Identificación de equipos

Respecto a la transformación de variables, existe una específica que resulta indispensable: la correspondiente a las variables "HomeTeam" y "AwayTeam" de los datos procedentes de Footballdata.co.uk. Estas variables registran los nombres de los equipos que participan en un partido, ya sea como local o visitante. Debido a que son variables categóricas, es necesario transformarlas en variables numéricas antes de la construcción de cualquier modelo, dado que la mayoría de los algoritmos de aprendizaje automático requieren

exclusivamente variables numéricas para su correcto funcionamiento. La transformación se realiza mediante un proceso comúnmente denominado *one-hot encoding*[10].

El *one-hot encoding* consiste en la generación de variables *dummy*, es decir, variables dicotómicas que sólo pueden adoptar los valores 0 y 1. En el caso de estas variables, un 1 indica que el equipo correspondiente juega en condición de local o visitante, mientras que un 0 denota lo contrario. De esta forma, se crean 2 variables para cada equipo presente en los datos, una para identificar cuándo dicho equipo juega como local y otra cuando juega como visitante. Esta transformación genera 80 nuevas variables. Para evitar problemas derivados de la multicolinealidad perfecta, será necesario eliminar una de las variables *dummy* de uno de los equipos, tanto para los partidos en los que juega como local como en los que lo hace como visitante. La multicolinealidad es un fenómeno que se presenta cuando existen variables predictoras altamente correlacionadas. Al eliminar una de las variables binarias, no se pierde información, se reduce la correlación y la multicolinealidad. Por tanto, se designará al equipo denominado Alavés como categoría base, lo que implica que si todas las variables que identifican al equipo local o visitante en un partido son 0, significa que el equipo es el Alavés.

Una alternativa al *one-hot encoding* es el *ordinal encoding*, que consiste en reemplazar cada categoría por un número distinto. Esta técnica es relevante cuando existe un orden lógico en las categorías. Por ejemplo, en una variable que refleje el nivel de educación, el uso de esta técnica sería adecuado, dado que resultaría coherente asignar un valor mayor a las observaciones correspondientes a individuos con doctorado que a aquellos sin estudios universitarios. Sin embargo, en el contexto de este trabajo, establecer un orden coherente entre los equipos es una tarea compleja, ya que la posición relativa de los equipos varía a lo largo de las diferentes temporadas incluidas en el conjunto de datos.

Variables "*rolling-sum*"

Es pertinente resaltar que, en la condición actual de los datos, no sería factible implementar el modelo de clasificación. La razón es simple, si se construye el modelo con los datos tal como se encuentran, bastaría con observar el número de goles anotados por cada equipo al final del partido para obtener una predicción precisa. Sin embargo, esto no representaría un proceso de predicción legítimo, ya que ya se tendría conocimiento del resultado de cada partido.

[10] https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/

Por lo tanto, si se desea integrar información relacionada con las estadísticas de los partidos en el conjunto de datos que se utilizará para la predicción, será necesario realizar ciertas transformaciones en las variables.

En este trabajo en particular, se ha optado por generar sumas móviles que acumulan información de los últimos 5 partidos para todas las estadísticas disponibles. En concreto, se han calculado estas nuevas variables acumulativas basadas en 2 enfoques diferenciados.

Primero, se han computado las variables de suma móvil que acumulan las estadísticas de los últimos 5 partidos jugados por el equipo local y el visitante antes del partido en cuestión, sin importar si jugó como local o visitante en estos últimos 5 partidos. Como resultado, se observan variables como "home_rolling_shots", que registra el número de tiros realizados en los últimos 5 partidos que ha jugado el equipo local, o "away_rolling_corners", que indica el número de saques de esquina lanzados por el equipo que juega como visitante en el partido en cuestión durante los últimos 5 partidos que ha disputado.

Además, se han calculado las variables de suma móvil que acumulan las estadísticas de los últimos 5 partidos en los que el equipo local del partido en cuestión ha jugado de local, y de los últimos 5 partidos en los que el visitante ha jugado de visitante. Esto es potencialmente relevante porque puede ser que un equipo no haya obtenido buenos resultados en sus últimos 5 partidos, pero esto puede ser debido a que ha jugado como visitante en 4 de estos últimos 5 partidos. En contraste, si se examinan los resultados obtenidos en los últimos 5 partidos en los que ha jugado estrictamente como local, se pueden observar resultados más favorables. Por este motivo, para capturar las variaciones que puedan existir causadas por el factor local o visitante, se considera que estas variables pueden proporcionar una importante contribución de información, distinta a la aportada por las otras variables de suma móvil que no consideran este factor. Ejemplos de estas variables incluyen "home_rolling_fouls_home", que muestra las faltas cometidas por el equipo local en sus últimos 5 partidos como local, o "away_rolling_red_cards_away", que indica el número de tarjetas rojas vistas por el equipo visitante en sus últimos 5 partidos disputados como visitante.

Adicionalmente, se han creado variables relacionadas que indican el número de puntos obtenidos por cada equipo en sus últimos cinco partidos para ambos enfoques, es decir,

variables de puntos conseguidos en los últimos partidos teniendo en cuenta si se ha jugado de local o de visitante, y sin tenerlo en cuenta. Antes de calcular estas variables, se generan 2 variables que muestran el número de puntos obtenidos por el equipo local y el visitante en cada partido, otorgando 3 puntos en caso de victoria, 1 en caso de empate y 0 en caso de derrota. Como resultado, se generan 4 variables *rolling sum* nuevas, "home_rolling_points", "away_rolling_points", "home_rolling_points_home" y "away_rolling_points_away".

Finalmente, se calculan nuevas variables que hacen referencia a las estadísticas concedidas. Por ejemplo, se calcula el número de goles concedidos por cada equipo en los últimos 5 partidos, de nuevo para ambos enfoques. Se sigue el mismo procedimiento con otras variables de interés como los saques de esquina o los tiros, ya que si un equipo ha concedido muchos tiros o saques de esquina en sus últimos partidos puede ser un indicador de que su defensa está en un momento complicado y que sus oponentes son capaces de generarles ocasiones de gol con facilidad.

Finalmente, es importante destacar que el cálculo de las variables de suma móvil genera la aparición de valores faltantes. Esto es porque se ha decidido que, en las ocasiones en las que no existan suficientes partidos previos para su cálculo, se asignen valores NA. Concretamente, se encuentran 200 observaciones con valores NA en las variables que se refieren a las estadísticas acumulativas de los últimos 5 partidos en los que el equipo local ha jugado de local y el equipo visitante ha jugado fuera de casa.

Las primeras 100 observaciones con valores *NA* coinciden con los primeros 100 partidos del conjunto de datos, lo cual es coherente dado que un equipo necesita disputar aproximadamente 10 jornadas, 5 en casa y 5 fuera, para que existan datos suficientes para computar estas estadísticas. Las 100 observaciones restantes con valores *NA* están relacionadas con equipos recién ascendidos, que no cuentan con un historial previo en el conjunto de datos que permita calcular estas variables. Por ejemplo, el Córdoba Club de Fútbol ascendió a Primera División en la temporada 2014/2015, después de más de 40 años sin participar en la máxima categoría del fútbol español. Por esta razón, las 10 primeras observaciones relacionadas con el Córdoba Club de Fútbol incluirán algún valor *NA* en algunas de estas variables acumulativas. Esta misma situación se repite con otros equipos como Xerez, Hércules, Tenerife o Numancia, entre otros.

Si bien podría haberse recurrido a la imputación de valores para reemplazar estos valores faltantes, se ha considerado que no existe una metodología óptima para hacerlo. Por lo tanto, se ha optado por prescindir de estas observaciones con valores faltantes.

Es relevante mencionar que los inconvenientes explicados para las variables anteriores se repiten de manera idéntica en el caso de las variables relacionadas con las estadísticas acumulativas para los últimos 5 partidos, sin importar si se han jugado como local o visitante. Las únicas diferencias son que no son las primeras 100 observaciones las que tienen valores *NA*, sino las primeras 50, ya que solo se necesitan aproximadamente 5 partidos previos y no 10, y lo mismo sucede en el caso de los equipos recién ascendidos. Sin embargo, al eliminar las 200 observaciones anteriores con valores *NA*, desaparecen todos los valores *NA* en estas variables, ya que estas 100 observaciones están incluidas en aquellas 200.

Unión de los conjuntos de datos

La etapa final en la preparación del conjunto de datos para el entrenamiento y validación de modelos implica la consolidación de los datos procedentes de 2 fuentes distintas, Footballdata.co.uk y SoFifa. Un aspecto crítico en esta fase radica en la homogeneización de los nombres de los equipos en ambos conjuntos de datos, lo que es esencial para una unión efectiva. En esta operación, se identificaron 20 equipos cuyos nombres no concordaban en las 2 fuentes de datos, por lo que se procedió a una modificación precisa de los mismos para garantizar la consistencia. Dado que las diferencias de acentuación podrían resultar problemáticas, se realizó este proceso con un meticuloso cuidado. La elección se inclinó por ajustar los nombres en el conjunto de datos de SoFifa para que coincidieran exactamente con los proporcionados por Footballdata.co.uk.

Con los conjuntos de datos armonizados en términos de la nomenclatura de los equipos, se estaban preparados para el proceso de combinación, denominado *merge*. Inicialmente, los conjuntos de datos se ordenaron en función de las fechas correspondientes, llamadas “Date” en ambos conjuntos, en orden ascendente. Posteriormente, se creó un diccionario que mapeaba las fechas de los partidos con las fechas más cercanas pasadas disponibles en los datos de SoFifa. Este enfoque asegura la aplicación de las valoraciones más recientes disponibles para cada equipo antes de cada partido.

Después de esto, se añadió una nueva columna “Date2” al conjunto de datos de Footballdata.co.uk, que asignaba las fechas de los partidos a las fechas más cercanas y pasadas de SoFifa basándose en el diccionario previamente creado. Seguidamente, se efectuó la primera combinación de los conjuntos de datos, utilizando “HomeTeam” y “Date2” de Footballdata.co.uk y “TeamName” y “Date” de SoFifa como criterios de unión. Esto permitió la incorporación de las 4 variables de valoración del FIFA para el equipo local en el conjunto de datos que contenía estadísticas y cuotas de apuestas. Los nombres de las variables asociadas con las valoraciones del FIFA se modifican para que lleven el prefijo “Home” como por ejemplo “HomeAttack”.

Una segunda combinación se realizó entre el conjunto de datos resultante de la primera unión y el de SoFifa nuevamente, pero esta vez utilizando “AwayTeam”, en vez de “HomeTeam” como criterio de unión. Esta operación añadió las calificaciones del FIFA para el equipo visitante al conjunto de datos de los partidos. Asimismo, se renombraron las columnas asociadas a las valoraciones para que llevaran el prefijo “Away”.

Es esencial resaltar que el éxito de esta operación de unión requiere una gestión cuidadosa de los datos de fecha. En el proceso, surgieron errores relacionados con intentos de comparar objetos de fecha con objetos de tipo entero, lo cual se pudo prevenir asegurándose de que todos los objetos de fecha se convirtieran a *Unix timestamps* antes de cualquier operación de comparación o conversión a enteros. Esto subraya la importancia de mantener un formato de fecha consistente en ambos conjuntos de datos.

Modelos explicativos

Descripción de un modelo explicativo

Con el conjunto de datos adecuadamente procesado y transformado, el siguiente paso es la construcción de modelos explicativos. A diferencia de los modelos predictivos que se centran en la predicción de resultados futuros, los modelos explicativos se ocupan de comprender la naturaleza y las causas de un fenómeno. En este contexto, el objetivo principal es identificar las variables que explican el fenómeno en cuestión, determinar su significación estadística y analizar su impacto marginal en la probabilidad de ocurrencia de una victoria local o visitante, dependiendo del modelo específico

La regresión logística es un algoritmo de clasificación que posee un elevado carácter explicativo. Este algoritmo emplea una función logística o sigmoide para convertir una combinación lineal de variables predictoras en una probabilidad. Su naturaleza probabilística, que posibilita la obtención de las probabilidades de pertenencia a una clase específica, es uno de los aspectos que le confieren su carácter explicativo. Esto no sucede con otros algoritmos, como los árboles de decisión o los bosques aleatorios, que no proporcionan directamente probabilidades, aunque se pueden obtener estimaciones de estas. Adicionalmente, otra ventaja significativa de este algoritmo es que permite la interpretación de los coeficientes, ya que estos representan el cambio logarítmico en las probabilidades de pertenencia a una clase en relación con los predictores.

Antes de construir cualquier modelo, es fundamental dividir los datos en 2 conjuntos, el de entrenamiento y el de validación. El conjunto de entrenamiento se utiliza para entrenar el modelo, es decir, los coeficientes del modelo se determinan en función de estos datos. Por otro lado, el conjunto de validación se emplea para evaluar el rendimiento del modelo con datos que no se utilizaron en la fase de entrenamiento.

Es crucial tener en cuenta el carácter de serie temporal de los datos durante este proceso de división. Si se dividen los datos de forma aleatoria, se corre el riesgo de incluir en el conjunto de validación datos que no estarían disponibles en el momento de realizar una predicción. Por esta razón, se establece el 1 de agosto de 2018 como fecha para dicha división. De esta forma, las temporadas anteriores a esta fecha conforman el conjunto de entrenamiento, que abarca hasta 2006, y las temporadas posteriores conforman el conjunto de validación.

Modelo explicativo de victorias locales

Inicialmente, se desarrolla el modelo explicativo centrado en las victorias locales. Dada la naturaleza explicativa de este modelo, cuyo propósito es identificar qué variables contribuyen a explicar el fenómeno en cuestión, se opta por un enfoque simplificado, que prescinde de la validación cruzada y la regularización, metodologías que serán introducidas en etapas posteriores.

Tras la implementación del modelo, se obtienen algunas métricas de rendimiento significativas. El modelo reporta una precisión del 62%. Sin embargo, como se discutirá posteriormente, la precisión no necesariamente representa una medida de evaluación objetiva del rendimiento de un modelo. Al considerar los coeficientes y p-valores, elementos que son de gran relevancia en un modelo explicativo, se detecta un error y no es posible su obtención. Tras una investigación adicional, se identifica que el error es producto de la multicolinealidad, un fenómeno que surge debido a la alta correlación entre dos o más variables independientes, como se ha mencionado en el caso del *one-hot encoding*. Este alto grado de correlación dificulta la identificación del efecto individual de cada variable sobre la variable dependiente, impidiendo así el cálculo de los coeficientes (Daoud, 2017).

Ante este hallazgo, se procede a examinar los pares de variables que exhiben una alta correlación, específicamente, correlaciones superiores al 0,85 en valor absoluto.

Tabla 1- Pares de variables predictoras con correlación alta (>0,85)

Variable 1	Variable 2	Correlación
HomeOverall	HomeMidfield	**0.8765**
HomeOverall	HomeDefence	**0.8510**
HomeAttack	HomeMidfield	**0.8783**
HomeAttack	HomeDefence	**0.8713**
HomeMidfield	HomeDefence	**0.9205**
AwayOverall	AwayMidfield	**0.8773**
AwayAttack	AwayMidfield	**0.8796**
AwayAttack	AwayDefence	**0.8728**
AwayMidfield	AwayDefence	**0.9214**
home_rolling_points	home_rolling_points_conceded	**-0.9571**
away_rolling_points	away_rolling_points_conceded	**-0.9580**
home_rolling_points_home	home_rolling_points_conceded_home	**-0.9563**
home_rolling_points_conceded_home	away_rolling_points_conceded_away	**-0.9563**
Home_mean_odds	Home_std_odds	**0.8900**
Draw_mean_odds	Draw_std_odds	**0.8776**
Draw_mean_odds	Away_mean_odds	**0.8542**
Away_mean_odds	Away_std_odds	**0.9050**

Fuente: Elaboración propia

La tabla ilustra el origen de los problemas asociados con la alta correlación. En primera instancia, las correlaciones de mayor magnitud, y por ende más problemáticas, están asociadas con las variables de puntos concedidos. Estas correlaciones cercanas a la perfección, superiores a 0,95, se justifican dado que un partido tiene únicamente 3 posibles desenlaces, cada uno con un total de puntos asociado. En la mayoría de los casos, conocer simplemente el número de puntos obtenidos por un equipo permite deducir los puntos concedidos. Por este motivo, se toma la decisión de eliminar las 4 variables que reflejan los puntos concedidos, "home_rolling_points_conceded", "away_rolling_points_conceded", "home_rolling_points_conceded_home" y "away_rolling_points_conceded_away".

El segundo conjunto de variables con alta correlación está relacionado con las valoraciones del FIFA. En este caso, la solución es directa y sencilla. Se eligen "HomeMidfield" y "AwayMidfield" para su eliminación, ya que estas son las que se repiten con mayor frecuencia entre los pares de variables problemáticas mostrados en la tabla y, además, presentan correlaciones por encima de 0,92 en dos casos.

Finalmente, las variables asociadas con las cuotas de apuestas también requieren cierta evaluación. Las variables que reflejan los valores medios y las desviaciones de las cuotas presentan correlaciones entre sí que son alarmantes para el modelo. Las variables "Away_mean_odds" y "Away_std_odds" tienen una correlación superior a 0,9, y los pares asociados a la victoria local y el empate también muestran correlaciones excesivamente elevadas. Ante esta situación, la solución óptima implica la eliminación de las variables de los valores medios o de las desviaciones, al contener información redundante evidente a la luz de la correlación. Tras examinar la correlación de las 6 variables con la variable *target*, que indica la victoria o no del equipo local, se determina que es más conveniente eliminar las 3 variables vinculadas a las desviaciones de las cuotas, ya que sus correlaciones con la variable objetivo son inferiores a las de los valores medios de las cuotas. Por consiguiente, se eliminan las variables "Home_std_odds", "Draw_std_odds" y "Away_std_odds".

Tabla 2- Correlación de variables de cuotas de apuestas con el target del modelo de victorias locales

Variable	Correlación con Target (home)
Home_mean_odds	**-0.2822**
Home_std_odds	**-0.1924**
Draw_mean_odds	**0.1865**
Draw_std_odds	**0.1546**
Away_mean_odds	**0.3169**
Away_std_odds	**0.2483**

Fuente: Elaboración propia

Interpretación de coeficiente y p-valores para el modelo de victorias locales

Una vez abordado el problema de la multicolinealidad, es factible obtener tanto los coeficientes de regresión de cada variable como los respectivos p-valores.

Inicialmente, es pertinente destacar que el p-valor es un indicador de significatividad estadística. Esta significatividad hace referencia a la probabilidad de que una relación observada entre dos o más variables no sea producto del azar. Un p-valor es una medida de la evidencia en contra de una hipótesis nula, que es una afirmación que sostiene que no existe relación entre la variable predictora y la variable objetivo. Un p-valor inferior a 0,05 sugiere que hay menos de un 5% de probabilidad de que la relación entre la variable predictora y la variable objetivo sea casual, por lo que se considera estadísticamente significativa. Esto implica que la variable predictora es un factor relevante y determinante para la predicción del objetivo.

De los 147 p-valores proporcionados por el modelo de regresión, solamente 15 son estadísticamente significativos, es decir, tienen un p-valor inferior a 0,05. De hecho, todas las variables significativas son aquellas variables *dummies* resultantes de la *one-hot encoding* para la identificación de los equipos. El resto de las variables, tanto las *rolling sum*, como las valoraciones de FIFA y las cuotas de apuestas, presentan p-valores que oscilan entre 0,15 y 0,5. Aunque a priori estos valores indican que no existe evidencia suficiente para afirmar que hay una relación o efecto significativo entre la variable predictora y la variable objetivo, es importante tener en cuenta que esto no necesariamente implica que no exista ninguna asociación entre las variables. Además, las relaciones analizadas en una regresión son relaciones lineales, por lo que es plausible que algunas variables no tengan una relación lineal directa con la variable objetivo, pero sí sean

significativas si se consideran relaciones no lineales, como lo hace el modelo *Random Forest*, por ejemplo, que será empleado en la construcción del modelo predictivo.

Dada esta situación de escasez de variables significativas, la interpretación de los coeficientes no aporta excesivo valor, pero aun así es de interés la introducción del concepto. En una regresión logística, los coeficientes se representan como el logaritmo de las probabilidades de éxito, la probabilidad de pertenecer a la clase positiva, dividido por las probabilidades de fracaso, es decir, la probabilidad de pertenecer a la clase negativa. A estos coeficientes se les conoce como *log-odds*. Para interpretarlos en términos de efectos marginales, es necesario calcular la exponencial de cada uno de ellos, conocida como *odds-ratio*. Estas exponenciales se interpretan como cambios en la probabilidad de éxito. El odds-ratio indica la razón de probabilidades entre la clase positiva y la clase de referencia. Por ejemplo, un *odds-ratio* de 2 implica que las probabilidades de pertenecer a la clase positiva son el doble en comparación con la clase de referencia. De manera similar, un *odds-ratio* de 1,5 se asocia con un aumento del 50% en las probabilidades, mientras que un *odds-ratio* de 0,8 se relaciona con una disminución del 20% en las probabilidades. En el caso de las variables dicotómicas, la interpretación es similar, un *odds-ratio* mayor que 1 indica que la presencia de esa variable, en comparación con su ausencia, está asociada con un incremento en las probabilidades de pertenecer a la clase positiva. Si el *odds-ratio* es menor que 1, indica una disminución de las probabilidades, mientras que si es igual a 1, la variable dicotómica no tiene un efecto discernible sobre las probabilidades de éxito.

En el caso del modelo explicativo de victorias locales, los *odds-ratio* cercanos a 2 son aquellos asociados a las variables dicotómicas que identifican a equipos visitantes, que además poseen p-valores por debajo de 0,05. Ejemplos de estas son "AwayTeam_Hercules" o "AwayTeam_Murcia", *cuyos odds-ratio* indican que la presencia de estos equipos como visitante aumenta considerablemente las probabilidades de victoria del equipo local. En el otro extremo, formado por los *odds-ratio* de menor valor, se observa que también están asociados a variables estadísticamente significativas, pero en este caso la interpretación es contraria. Se encuentra la variables "HomeTeam_Cordoba" con un *odds-ratio* de 0,3, que implica que le presencia del Córdoba como equipo local disminuye la probabilidad de victoria local. "AwayTeam_Real Madrid" y "AwayTeam_Barcelona" también muestran *odds* menores

de 0, indicando una disminución en la probabilidad de victoria del oponente, lo cual tiene sentido dado su nivel de éxito en las últimas temporadas.

A la vista de los resultados obtenidos en cuanto a p-valores y coeficientes es evidente que el factor con mayor capacidad explicativa para el fenómeno de las victoria locales es el relacionado a la identificación de los equipos que se enfrentan en los partidos. En consecuencia, el mero conocimiento de los equipos que se enfrentan proporciona más información que las cuotas de apuesta promedio o el desempeño en los últimos partidos.

Comparación con modelo *naive* de victorias locales

Cuando se construye un modelo con el que se pretenda realizar predicciones es esencial conocer y comprender los datos, y consecuentemente, ser consciente de la precisión que se espera alcanzar. En cualquier tipo de predicción, el primer umbral que debe superarse es el conocido como modelo *naive*. Este término, comúnmente utilizado en aprendizaje automático, se refiere al modelo más simplificado posible, que no tiene en cuenta las posibles relaciones entre las variables predictoras y la variable dependiente. Puede ser descrito también como un modelo de "mínimo esfuerzo", dado que no incorpora ningún tipo de conocimiento ni algoritmo sofisticado.

Al tratarse de una clasificación, existen diferentes enfoques a este modelo simplista. Un posible acercamiento sería la predicción aleatoria, la cual asumiría que el resultado de los partidos de La Liga es completamente aleatorio y asignaría igual probabilidad a los posibles sucesos. De esta forma, al tener 2 posibles resultados, victoria local o no, asignaría un 50% de probabilidad respectivamente. Dicho esto, el modelo *naive* más apropiado para la predicción en cuestión es el de clase mayoritaria. Como su propio nombre indica, el único factor tomado en consideración es la clase mayoritaria de los datos de entrenamiento, asignándole esta a todos los partidos a predecir. Será la precisión brindada por la predicción de clase mayoritaria la que tendrá que ser superada por el modelo a construir, para poder así garantizar que este aporta valor. De no ser superada, implicaría que la complejidad añadida, tanto a nivel de tratamiento de datos como de algoritmo de predicción, sería inútil.

Si se considerasen los 3 resultados posibles de un partido, las victorias locales son la clase mayoritaria, superando ampliamente a los empates y a las victorias visitantes. La ventaja de jugar en casa ha sido un tema de gran interés en el mundo del deporte, especialmente en deportes de equipo como el fútbol, baloncesto y fútbol americano. Se encuentran

investigaciones que concluyen que existe una ventaja significativa al jugar en casa en estos deportes, aunque la magnitud de esta ventaja puede variar dependiendo de factores como el tamaño del estadio, la distancia recorrida por el visitante para llegar a la localidad donde se juega el partido o la altitud (Pollard & Gómez, 2014). Si se analizan los datos empleados en este modelo concreto, en los que se agrupan los empates y las victorias visitantes, la clase mayoritaria no son las victorias locales. La variable a predecir presenta un 1, es decir victorias locales, en el 49% de las observaciones, y un 0 en el 51% restante. De esta forma, la clase mayoritaria es la formada por victorias visitantes y empates. Como se ha introducido previamente, la precisión del modelo explicativo es del 62%, superando holgadamente el 51% que muestra la clase mayoritaria, ratificando así que la complejidad añadida por la regresión aporta cierto valor.

Modelo explicativo de victorias visitantes

A continuación, se procederá a replicar el procedimiento realizado para el modelo explicativo de victorias locales, pero esta vez para el modelo de victorias visitantes.

Al ejecutar el modelo se obtiene una precisión del 72%. Esta medida de desempeño resulta engañosa porque las victorias visitantes son la clase minoritaria, y un modelo que siempre prediga la no victoria local obtendría una precisión aparentemente favorable. De nuevo, aparece el contratiempo de la multicolinealidad, dado que los datos empleados son idénticos, a excepción de la variables objetivo, lo cual implica que las correlaciones son las mismas a las mostradas en la tabla previamente introducida. La solución será la empleada anteriormente, de forma que se prescinde de las variables asociadas a los puntos concedidos y las variables de valoraciones del FIFA *HomeMidfield* y *AwayMidfield*. En lo que respecta a las variables vinculadas a las cuotas de apuestas, se repite el escenario anteriormente descrito, en el cual los valores medios presentan correlaciones más altas con el objetivo. En consecuencia, se eliminarán las variables relacionadas con las desviaciones estándar.

Tabla 3. Correlación de variables de cuotas de apuestas con el target del modelo de victorias visitantes

Variable	Correlación con Target (away)
Home_mean_odds	0.3175
Home_std_odds	0.2431
Draw_mean_odds	-0.0888
Draw_std_odds	-0.0788

Away_mean_odds	-0.2445
Away_std_odds	-0.1761

Fuente: Elaboración propia

Los valores de p obtenidos en este nuevo modelo explicativo de victorias visitantes son parecidos a los observados en el modelo de victorias locales. De manera precisa, son 14 variables las que muestran valores de p inferiores a 0,05, y la mayoría de ellas están vinculadas a variables dicotómicas encargadas de identificar los equipos que participan en cada encuentro. Se destaca que una de estas variables significativas es "Draw_mean_odds", la cual contiene el promedio de las cuotas de empate ofrecidas por diversas casas de apuestas. Las variables relacionadas con las estadísticas acumulativas de diferentes parámetros tienen en su mayoría valores de p entre 0,4 y 0,5. Similar a lo realizado con el modelo previo, no se suprimen estas variables ya que no se descarta la existencia de relaciones no lineales que pueden ser capturadas por otros algoritmos.

Interpretación de coeficiente y p-valores para el modelo de victorias visitantes

En lo que respecta a los coeficientes, la situación es nuevamente similar a la observada en el modelo de victorias locales. Las variables con los *odds-ratios* más altos y bajos están asociadas a las variables *dummy* de identificación de equipos, y son significativas. Entre ellas se encuentran "HomeTeam_Cordoba" o "HomeTeam_Murcia", ambas con *odds-ratios* superiores a 2, y "AwayTeam_Leganes" o "AwayTeam_Tenerife", con *odds-ratios* inferiores a 0,5. Estas variables tienen un impacto notable en las probabilidades de victoria visitante. Una vez más, las variables de identificación de equipos son las que mejor explican las probabilidades de victoria visitante, superando de manera notable a las relacionadas con las cuotas de apuestas, el FIFA y las variables resultantes de los *rolling sum*.

Comparación con modelo *naive* de victorias visitantes

En este modelo enfocado en las victorias visitantes, la clase predominante es la representada por el valor 0, que incluye victorias locales y empates. El 71,6% de las observaciones en el conjunto de datos presentan un 0 en la variable objetivo, mientras que solo el 28% muestran un 1 en la misma. Al cotejar el porcentaje de la clase predominante con la precisión obtenida mediante la regresión logística, que es del 72,5%, se concluye que el modelo está proporcionando una utilidad, pero de manera limitada, ya que apenas supera el umbral presentado por la clase predominante en menos de un 1%. Sin embargo, es razonable considerar que el uso de algoritmos más sofisticados que la regresión,

capaces de identificar relaciones no lineales, y una posible selección de variables podrían resultar en un incremento en la precisión del modelo.

Modelos predictivos

Descripción de algoritmos utilizados

Ahora, contando con una mayor comprensión acerca de las variables que explican el fenómeno y sus efectos en la variable dependiente, es oportuno avanzar en la construcción de los modelos predictivos. El objetivo de estos modelos se centra en la precisión de las predicciones, no en la interpretación de las relaciones subyacentes. Este enfoque permite la utilización de algoritmos de mayor complejidad.

Se ajustarán un total de 16 modelos predictivos, de los cuales 8 estarán orientados a la predicción de victorias locales y los 8 restantes a la predicción de victorias visitantes. Se aplicarán 4 algoritmos de predicción diferentes: *Random Forest*, *Support Vector Machine* (*SVM*), *K-Nearest Neighbors* (*KNN*) y *Extreme Gradient Boosting* (*XGBoost*). La selección de estos algoritmos se basa en la revisión de la literatura, la cual evidencia que estos métodos presentan un alto rendimiento en este ámbito específico (Rodríguez & Pinto, 2022) (Gomes, Portela & Santos, 2021) (Pugsee & Pattawong, 2019).

El algoritmo *Random Forest* es un método de aprendizaje supervisado que integra múltiples árboles de decisión para generar predicciones. Cada árbol en el conjunto se entrena con una muestra aleatoria del conjunto de datos y proporciona una predicción. Posteriormente, las predicciones de todos los árboles se promedian para obtener la predicción final. Esto facilita el manejo de conjuntos de datos grandes y complejos, y también contribuye a reducir el sobreajuste. Los árboles en el bosque se construyen utilizando una combinación de subconjuntos de características y muestras aleatorias, lo cual potencia la diversidad y precisión del modelo.

El *SVM* es un método de aprendizaje supervisado que se utiliza para la clasificación y regresión. Su objetivo es encontrar un hiperplano en un espacio de alta dimensión que mejor separe las clases de datos. Utiliza vectores de soporte, que son muestras cercanas a la frontera de decisión, para determinar la ubicación óptima del hiperplano. *SVM* puede utilizar diferentes funciones de *kernel* para transformar los datos en espacios de mayor dimensionalidad, lo que permite manejar datos no linealmente separables. Es especialmente eficaz cuando el número de características es grande en comparación con el número de muestras.

El algoritmo *KNN* es un método de aprendizaje que basa en la idea de que las muestras que son similares entre sí en términos de características están más propensas a pertenecer a la misma clase. *KNN* clasifica un punto desconocido asignándole la clase más común entre sus k vecinos más cercanos en el espacio de características. El valor de k determina la influencia de los vecinos en la predicción. *KNN* es un método simple y fácil de entender, pero puede volverse computacionalmente caro en conjuntos de datos grandes.

Por último, el *XGBoost* es un algoritmo basado en árboles que utiliza un enfoque de impulso para mejorar la precisión de las predicciones. *XGBoost* construye una secuencia de árboles de decisión, cada uno de los cuales se enfoca en corregir los errores del modelo anterior. Esto permite capturar patrones más complejos en los datos y lograr una mayor capacidad de generalización. El algoritmo se destaca por su eficiencia computacional y su capacidad para manejar conjuntos de datos grandes y de alta dimensionalidad.

Tratamiento del factor temporal

Se emplearán 2 conjuntos de entrenamiento y 2 conjuntos de validación distintos para cada uno de los algoritmos implementados, diferenciándose estos en el número de temporadas incluidas. Cuando se utilizan datos históricos para entrenar un modelo predictivo, es crucial considerar que el desempeño y las condiciones de un fenómeno o evento pueden cambiar a lo largo del tiempo. Esto es especialmente relevante en los contextos deportivos, donde los equipos, jugadores, tácticas y reglas están en constante evolución. La inclusión de datos antiguos puede introducir información obsoleta o irrelevante para el contexto actual, lo que podría afectar la precisión del modelo y su capacidad para generalizar adecuadamente. Por esta razón, se construirán modelos que consideran 2 ventanas temporales diferentes.

En primer lugar, se emplearán todos los datos disponibles, utilizando las temporadas desde 2006/2007 hasta 2018/2019 como conjunto de entrenamiento, y las temporadas desde 2019/2020 hasta 2021/2022 como conjunto de validación. Por otra parte, se construirán modelos que únicamente toman en cuenta las 8 temporadas más recientes, desde 2013/2014 para el entrenamiento, y las temporadas 2020/2021 y 2021/2022 para la validación.

En todos y cada uno de los modelos ajustados en esta investigación se hace uso de la validación cruzada, también conocida como *cross-validation*. Esta consiste en dividir los datos de entrenamiento en diferentes subconjuntos permitiendo llevar a cabo múltiples

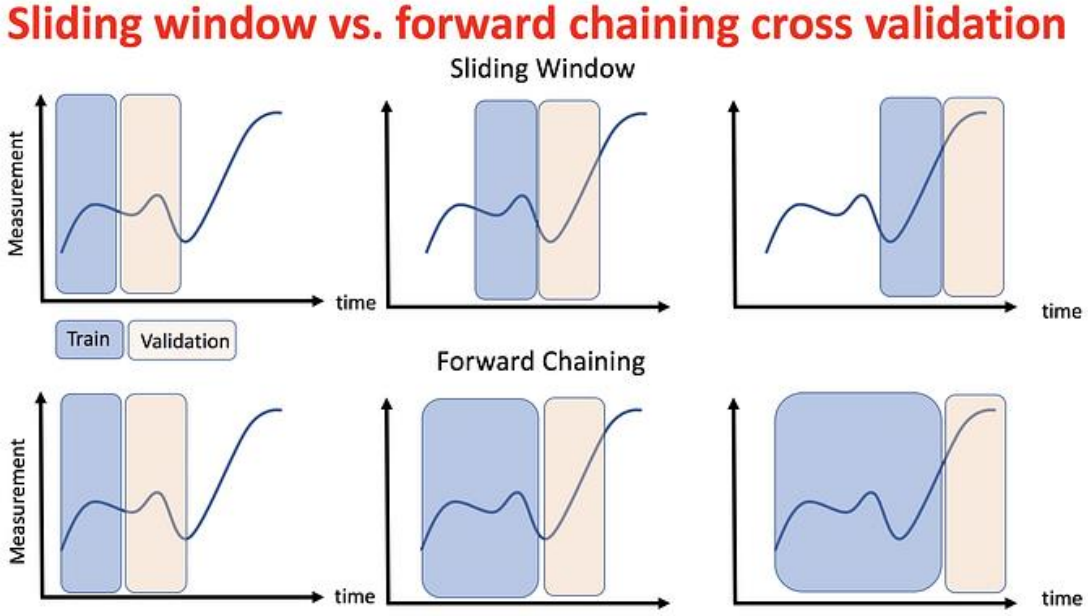
iteraciones de entrenamiento con los diferentes subconjuntos. Esta técnica, comúnmente usada en la construcción de modelos de aprendizaje automático, aborda el problema que aparece al partir los datos de entrenamiento y validación en un punto concreto, y aparte permite que el modelo sea evaluado con la totalidad de datos disponibles (Bergmeir & Benitez, 2012). Una forma habitual es la validación cruzada *k-folds*, en la que los datos se dividen en *k* subconjuntos de igual tamaño. Posteriormente, se realizan k iteraciones, donde en cada iteración, un subconjunto diferente se utiliza como conjunto de prueba, y los *k-1 folds* restantes se emplean como conjunto de entrenamiento. Una vez que todos los subconjuntos han sido considerados como conjunto de validación, se promedian los resultados de todas las iteraciones y se obtiene una estimación del rendimiento del modelo. Es importante puntualizar que en cada iteración el modelo se entrena desde cero, ya que de otra forma existiría un notable riesgo de *overfitting* al estar evaluando el modelo con datos que han sido utilizados para el entrenamiento en iteraciones previas.

Dada la naturaleza temporal de los datos de los modelo en cuestión, la partición de estos subconjuntos no puede ser aleatoria. Sería ilógico entrenar el modelo con información de partidos más recientes que los presentes en el conjunto de evaluación. A modo de ejemplo, a la hora de predecir la victoria local o visitante en el Real Betis contra el Valencia Club de Fútbol de la temporada 2019/2020, no sería lícito contar en el conjunto de entrenamiento con este mismo enfrentamiento, pero en las temporadas 2020/2021 y 2021/2022, dado que estaría empleando información futura y no disponible en el momento que tiene lugar el encuentro en la temporada 2019/2020. Para abordar este problema se emplea el *time series cross-validation*, en la que los datos se dividen en múltiples conjuntos de entrenamiento y prueba de manera secuencial, manteniendo el orden temporal de los datos. Cada conjunto de prueba se encuentra en un período posterior al conjunto de entrenamiento correspondiente, lo que garantiza que el modelo se evalúe en datos futuros no observados durante el entrenamiento. La técnica más comúnmente utilizada es la *TimeSeriesSplit*[11] proporcionada por la biblioteca *scikit-learn*. Esta estrategia de particionamiento, conocida como *forward chaining*, divide los datos en múltiples conjuntos de entrenamiento y prueba de manera progresiva. En cada iteración, el conjunto de entrenamiento se expande para incluir datos más recientes, mientras que el

[11] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

conjunto de prueba avanza en el tiempo, de tal manera que siempre se respete la cronología de los datos, es decir, no se utilicen datos futuros para predecir el pasado.

Figura 2- Comparación de validaciones cruzadas sliding window y forward chaining



Fuente: https://medium.com/@pradip.samuel/cross-validation-in-time-series-model-b07fbba65db7

Selección de variables

Dado que se dispone de 147 variables en los datos disponibles, y teniendo en cuenta que una proporción considerable de estas no son significativas, se considera apropiado realizar una selección de variables. Dependiendo de las características de cada modelo se emplean diferentes métodos para dicha selección.

En el caso del *Random Forest*, por ejemplo, se emplea el *Recursive Feature Elimination*[12], traducido como la eliminación recurisva de variables, aprovechando así la capacidad de *Random Forest* para calcular la importancia de las variables. También se emplea otro método de selección de variables, en el caso del *SVM*, conocido como *SelectKBest*[13], que ayuda a reducir la dimensionalidad de los datos mediante pruebas estadísticas como el análisis de varianza o la prueba de chi-cuadrado. Aparte de la selección de variables se utilizan otros procedimientos el *Randomized Search*[14], un algoritmo de búsqueda que selecciona de forma aleatoria un conjunto de combinaciones de hiperparámetros para evaluar, lo que lo hace eficiente y efectivo para encontrar una configuración óptima en problemas de optimización de modelos. Estos recursos son

[12] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
[13] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

empleados con el objetivo de construir modelos eficientes en cuanto a la utilización de recursos computacionales, ya que son 16 modelos los que se precisa construir, pero que sean suficientemente complejos para ofrecer captar las particularidades de los datos y ofrecer predicciones óptimas.

Elección de modelos predictivos

Tras la implementación de todos los modelos, se hace necesario seleccionar uno para la predicción de victorias locales y otro para las visitantes, con el fin de fundamentar la simulación financiera en estos modelos. Elegir el modelo basándose en la precisión sería un error considerable. La precisión representa simplemente la proporción de predicciones correctas, sin tomar en cuenta la clase de estas, lo que puede resultar engañoso cuando las clases están desequilibradas. Por ejemplo, una precisión del 70% en la predicción de victorias visitantes podría parecer exitosa a primera vista, pero si consideramos que el 71% de las observaciones pertenecen a la clase negativa, el modelo no estaría superando al modelo *naive*.

Una decisión más acertada sería utilizar la sensibilidad, que mide la proporción de predicciones positivas que son correctas. En otras palabras, indica la frecuencia con la que el modelo acierta cuando predice una victoria local, de manera que una mayor sensibilidad implicaría un mayor rendimiento esperado en las apuestas. Sin embargo, considerando el alto impacto económico de un falso positivo, es decir la predicción incorrecta de una victoria local o visitante, esta métrica tampoco resulta ideal. Además, fijarse solo en la sensibilidad podría generar un modelo demasiado conservador, que perdería oportunidades de incrementar la rentabilidad al no apostar en ciertos partidos. Por ello, resulta relevante considerar también la especificidad, que mide el acierto de un modelo al predecir la clase negativa.

La puntuación F1 se presenta como la métrica ideal para la selección del modelo óptimo. El *score F1* es la media armónica entre la sensibilidad y la especificidad, tratando los falsos positivos y falsos negativos con la misma importancia. Este se calcula como la media de la sensibilidad y la especificidad, variando entre 0 y 1. Dado el interés en una predicción precisa tanto de victorias locales o visitantes como de las derrotas y empates, la elección se basa en la puntuación F1 macro. A diferencia del *Weighted F1 score*, que pondera la puntuación en función de la proporción de cada clase, la métrica macro ofrece una evaluación libre de sesgos causados por el desequilibrio entre clases.

A continuación, se muestran 2 tablas, una para el modelo predictor de victorias locales y otra para el de victorias visitantes, con los *macro F1 scores* observados para los 4 algoritmos y 2 particiones temporales. "Duración 1" se refiere a los modelos construidos con las 16 temporadas disponibles y "Duración 2" a aquellos construidos con las últimas 8 temporadas.

Tabla 4- Puntuación F1 macro para los modelos predictivos de victorias locales

	Duración 1	Duración 2
Random Forest	0,65	0,66
SVM	0,66	0,66
KNN	0,59	0,60
XGBoost	0,66	0,67

Fuente: Elaboración propia

Tabla 5- Puntuación F1 macro para los modelos predictivos de victorias visitantes

	Duración 1	Duración 2
Random Forest	0,56	0,61
SVM	0,55	0,54
KNN	0,55	0,58
XGBoost	0,57	0,59

Fuente: Elaboración propia

En el caso de los modelos de predicción de victorias locales se observa que, menos el algoritmo *KNN*, todos ofrecen *F1 scores* parejos, sin grandes diferencias entre los modelos de diferente horizonte temporal, a pesar de que los basados en datos de menor duración tienden a ofrecer mejores prestaciones. El modelo elegido es el construido con el algoritmo *XGBoost* y con la "Duración 2", al superar en 0,01 a varios de los otros modelos. En cambio, en lo que respecta a los modelos de predicción de victorias visitantes se observan mayores diferencias tanto entre algoritmos como entre ventanas temporales, dado que las ventanas de menor duración ofrecen significativamente un mejor desempeño. El modelo óptimo es el ajustado mediante el *Random Forest* y con la "Duración 2", que supera en 0,02 al segundo mejor modelo.

Con los modelos seleccionados, antes de proceder con la simulación financiera, se persigue mejorar su rendimiento, sin considerar en gran medida el coste computacional, dado que el número de modelos a entrenar se ha reducido de 16 a 2. Durante el proceso

de optimización del modelo, se examinaron diversas técnicas y estrategias de mejora. Estas incluyeron la ampliación de las iteraciones y la expansión de la cuadrícula en la sintonización de hiperparámetros, el incremento del número de características en la selección de estas, la aplicación de la Eliminación Recursiva de Características con Validación Cruzada (*RFECV*), y el balanceo de los pesos de las clases. Adicionalmente, se probó la sustitución de RandomizedSearchCV, que realiza una búsqueda aleatoria en el espacio de hiperparámetros especificado, lo cual permite una exploración más eficiente y menos costosa, por GridSearchCV[15], que realiza una búsqueda exhaustiva a través de todas las posibles combinaciones de hiperparámetros especificados en la cuadrícula proporcionada.

No obstante, en el caso del modelo de victorias visitantes, ninguna de estas intervenciones resultó en mejoras en la precisión predictiva. Por otro lado, en el caso del modelo de victorias locales, el único cambio que tuvo un impacto considerable en el rendimiento del modelo fue la modificación del parámetro "*cv*" en el *RFECV*. Al aumentar "*cv*" de 3 a 10, se estableció un proceso de validación cruzada más riguroso dentro del *RFECV*. Este ajuste condujo a una selección de características más eficiente, y finalmente, a un ligero aumento en el rendimiento del modelo, mejorando en 0,02 puntos la puntuación F1 macro.

Esta experiencia resalta un principio fundamental en el aprendizaje automático, y es que la adición de complejidad no siempre conduce a un rendimiento mejorado del modelo. Esto puede parecer contraintuitivo, pero se debe a múltiples factores. Por ejemplo, la inclusión de un mayor número de características, o la búsqueda exhaustiva de hiperparámetros, puede conducir a un sobreajuste del modelo a los datos de entrenamiento, disminuyendo su capacidad para generalizar a nuevos datos (Ying, 2019). Además, cada conjunto de datos posee su propia estructura subyacente y ruido inherente. Más allá de cierto punto, ningún modelo puede mejorar su rendimiento simplemente aumentando su complejidad. Por lo tanto, la clave del éxito en el aprendizaje automático reside en encontrar un equilibrio entre la simplicidad y la complejidad del modelo, para representar de la mejor manera posible los datos proporcionados.

[15] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Resultados

Descripción de la simulación financiera

Una vez que se han seleccionado los modelos con el rendimiento predictivo más elevado para cada uno de los dos casos, se procede a implementar la simulación financiera. Esta simulación tiene como objetivo determinar si las habilidades predictivas de los modelos son suficientes para generar rentabilidad y, en caso afirmativo, calcular el valor de dicha rentabilidad.

La simulación financiera se basa en las predicciones proporcionadas por los modelos y en las cuotas ofrecidas por las casas de apuestas. Se realiza una simulación durante el periodo de validación, que comprende las 76 jornadas que suman las temporadas 2021/2021 y 2021/2022, para averiguar el rendimiento financiero derivado de la teórica estrategia de inversión basada en los modelos.

En primer lugar, es importante destacar que se mantendrá el umbral de clasificación predeterminado de 0,5 para ambos modelo. Esto significa que el modelo predice como victoria local, o visitante, cuando la probabilidad calculada es mayor de 0,5. Sin duda, modificar este umbral para optimizar la rentabilidad habría sido un ejercicio interesante y valioso. No obstante, es esencial subrayar que este umbral debe ser determinado antes de la validación del modelo. Sería un error grave probar el modelo con diferentes umbrales y seleccionar el que proporcione la mayor rentabilidad. El umbral debe ser tratado como otro hiperparámetro y debe ser determinado durante la fase de entrenamiento. Debido a la complejidad adicional que implicaría optimizar el umbral de clasificación, se ha optado por mantener el umbral estándar de 0,5.

A continuación, se describe en detalle el proceso seguido en esta simulación, que simula la estrategia de inversión. En primer lugar, las apuestas se estructuran por jornada, considerando cada una de ellas como un evento independiente, lo que facilita el cálculo posterior del retorno. Por lo tanto, en cada jornada se observan los partidos para los que los modelos predicen una victoria local o visitante y se reparte una cantidad de dinero entre ellos. Es crucial señalar que esta cantidad se introduce en cada jornada, es decir, que no se realiza ninguna reinversión, por lo que en cada jornada se introduce una cantidad de dinero independiente a la acumulada en las jornadas anteriores. A pesar de que esto podría requerir un mayor capital, esta decisión, respaldada por el hecho de tratar cada jornada como un evento independiente, está motivada por el riesgo que conllevaría

introducir solamente una cantidad al inicio. Si en una jornada se predice incorrectamente en todos los partidos en los que se apuesta, esto podría resultar en la pérdida de toda la inversión, lo que haría la estrategia significativamente más arriesgada. En esta simulación específica, se ha optado por una cantidad de 1.000€ por jornada, lo que no es determinante, ya que se mide la rentabilidad en porcentaje.

Para poder realizar estas apuestas en función de la jornada, ha sido necesario generar una variable que indique a qué jornada pertenece cada partido. Un desafío que surge en este contexto es el relacionado con los partidos aplazados. Varias circunstancias, como pueden ser las condiciones climáticas adversas o la participación de los equipos en otras competiciones, pueden llevar a la postergación de ciertos partidos a lo largo de la temporada. Estos partidos no se disputan junto con el resto de encuentros de la misma jornada, sino que se juegan significativamente antes o después. Esta situación tiene implicaciones en la realización de las apuestas, ya que no sería apropiado agrupar los partidos aplazados con el resto de su jornada debido a las diferencias temporales. Aunque se podrían haber agrupado con los partidos de la jornada más próxima en el tiempo, o incluso haberse considerado como jornadas independientes, lo que implicaría asumir un riesgo desmedido al apostar 1.000€ a un solo partido, estos partidos aplazados se han excluido a la hora de realizar las apuestas, ya que se considera este el enfoque más honesto y prudente. Siguiendo la metodología empleada por Caro en su estudio predictivo de La Liga (Caro, 2020), se recurrió a la página web Flashscore[16], un sitio web de seguimiento de resultados deportivos para identificar los encuentros aplazados. Gracias a la forma en que este sitio muestra los datos de los resultados de la Liga, fue posible detectar eficientemente estos partidos. Durante el periodo de los datos de *backtesting*, se detectaron 12 partidos aplazados en la temporada 2020/2021 y 8 en la 2021/2022.

Con los partidos de cada jornada en lo que se va a apostar ya determinados, el siguiente paso del proceso es determinar la cantidad que se invierte en cada uno de estos partidos. Si bien se podrían dividir los 1.000€ entre el número de partidos en dicha jornada, y distribuirlos equitativamente, se ha decidido que hacer uso de las probabilidades ofrecidas por los modelos para determinar estas cantidades. Se calcula la suma de las probabilidades de los partidos seleccionados en cada jornada, para luego dividir la probabilidad de un partido entre la suma de las probabilidades de todos los partidos seleccionados para esa

[16] https://www.flashscore.es/

jornada. Posteriormente, se multiplica este porcentaje por la cantidad fija de 1.000€ que se invierte en cada jornada, distribuyendo la inversión de forma proporcional a la probabilidad de éxito dictada por el modelo.

Una vez determinados los partidos y sus respectivas cantidades, se realizan las apuestas y se calculan las pérdidas o ganancias, comparando con los resultados reales observados en los partidos de La Liga. Si la predicción es correcta se gana la apuesta, de forma que se calculan las ganancias multiplicando la cantidad invertida en el partido por la cuota correspondiente, la de victoria local o visitante, en función del modelo que se esté empleando. Cabe destacar que las cuotas utilizadas en estos cálculos de ganancias son las cuotas medias utilizadas en los modelos. En el caso de que la predicción sea incorrecta se pierde la totalidad del dinero apostado ese partido. De esta forma, se obtiene un registro de perdidas o ganancias por jornada, que se acumulan dando lugar al beneficio o perdida final.

Finalmente, esta simulación podría beneficiarse de la diversificación debido a la aparente baja correlación entre los resultados de los partidos en una jornada y entre las jornadas, aunque no se disponga de cifras exactas para respaldar esta afirmación. En este sentido, la teoría de carteras de Markowitz puede ser relevante para el enfoque de diversificación en la simulación de apuestas deportivas. Esta teoría se centra en la combinación óptima de activos para maximizar el rendimiento esperado y minimizar el riesgo total de una cartera (Markowitz, 1952). Aplicada a esta simulación, sugiere que al seleccionar una combinación diversa de partidos en diferentes jornadas, se podría reducir la exposición al riesgo individual de cada partido y potencialmente mejorar los resultados generales de la simulación.

Resultados de la simulación financiera

Tras ejecutar la simulación se obtienen resultados similares en términos del beneficio bruto para la simulación que apuesta en victorias locales y para la que lo hace en victorias visitantes.

En la simulación enfocada en victorias locales, se apuesta en un total de 253 partidos, distribuidos en 75 jornadas, es decir, se apuesta en todas las jornadas de las 2 temporadas de la simulación de prueba excepto en la jornada 18 de la temporada 2020/2021. Se gana la apuesta en 171 partidos, lo que significa que se acierta el 68% de las apuestas realizadas. Los rendimientos por jornada son diversos, existen 8 jornadas en las que los

retornos son del -100%, mientras que también se observan 46 jornadas con retornos positivos, llegando a valores máximos del 83%. Al apostar en 75 jornadas, la cantidad total de dinero invertido es de 75.000€, y al sumar las cantidades obtenidas al final de cada una de las 75 jornadas, la cifra es de 79.192€. Esto supone un beneficio bruto de 4.192€, que, en comparación con el capital invertido, proporciona una rentabilidad del 5,6% a lo largo de las 2 temporadas. Sin embargo, como se detallará más adelante, esta tasa de retorno no es precisa, ya que no se están considerando ciertos factores en este momento.

La simulación centrada en victorias visitantes es más selectiva, siendo un modelo oportunista como se indicó inicialmente, de manera que solo realiza apuestas en 73 partidos pertenecientes a 50 jornadas diferentes. Se ganan las apuestas en 49 partidos, lo que implica que la proporción de aciertos en las apuestas realizadas es del 67%, ligeramente menor que en el caso de las victorias locales. Los retornos por jornada son más extremos que en la simulación anterior, lo cual es lógico considerando que las victorias visitantes ocurren con mucha menos frecuencia. Esto se traduce en mayores cuotas que contribuyen a mayores retornos. Dicho esto, se observan retornos del -100% en 12 de las 50 jornadas en las que se apuesta, por lo que en más del 20% de las jornadas en las que se apuesta se pierde la totalidad del dinero. Sin embargo, para las jornadas en las que se obtienen retornos positivos, estos son significativamente superiores a los observados en la simulación de victorias locales, siendo casi todos mayores al 50%. De hecho, en 10 jornadas específicas se obtienen rentabilidades por encima del 75%. La suma del capital invertido es de 50.000€, mientras que la suma de los resultados de todas las jornadas es de 54.199€, lo que se traduce en un beneficio bruto de 4.199€. En este caso, la rentabilidad obtenida a lo largo de las 2 temporadas, que de nuevo, no es precisa, es del 8,4%, dado que el capital invertido es considerablemente menor que en la otra simulación, pero el beneficio se mantiene casi idéntico.

Dado que ambas simulaciones arrojan beneficios, se considera apropiado combinar las apuestas proporcionadas por las 2 simulaciones con el objetivo de buscar una mayor rentabilidad. Este proceso es sencillo y consiste en unir las apuestas señaladas por los modelos y volver a determinar la distribución del capital de manera proporcional a las probabilidades de cada partido. Es importante tener en cuenta que no existe interferencia alguna entre los 2 modelos, ya que al revisarlos no se observa ningún partido en el que ambos modelos indiquen una apuesta, lo cual es coherente al predecir sucesos

prácticamente opuestos. Al ejecutar la simulación de la combinación de los modelos, se apuesta en las 76 jornadas y se obtiene una suma de 79.380€, resultando en un beneficio de 3.380€ o un retorno del 4,4%, que es menor al obtenido con las simulaciones de forma individual.

La razón de la menor rentabilidad se explica en gran parte por la dilución de la cantidad apostada en cada partido, especialmente en el caso de las apuestas a la victoria visitante. Al haber un mayor número de partidos para la distribución de la cantidad fija por jornada, el importe apostado es menor, por lo que las ganancias también disminuyen. Para entender claramente la penalización a la rentabilidad a la que se hace referencia, por ejemplo, a la jornada 13. En esta se apuesta a 5 partidos, 4 al local y 1 al visitante, con un beneficio en la jornada de 474€. En los modelos previos, se obtienen beneficios en la jornada de 326€ en la simulación de victorias locales y de 1.110€ en el caso de la victoria visitante. Si se suman estos beneficios y se dividen entre 2, dado que provienen de 2.000€ apostados, se observa un beneficio de 718€. La reasignación del capital en la combinación de modelos disminuye el beneficio en numerosas jornadas, especialmente al asignar menores cantidades a las victorias visitantes. Dado el nulo aporte a la rentabilidad total, se descarta el uso de la combinación de modelos para la comparación con activos financieros.

Cálculo de rentabilidad real

Para el objetivo final de comparar la rentabilidad obtenida en la simulación financiera con diferentes tipos de activos financieros, es necesario realizar ciertas transformaciones sobre la tasa de rentabilidad introducida anteriormente, que es el resultado de la simple división del capital final entre la inversión inicial. Primero, existen numerosos activos que asumen la reinversión automática del capital, como es el caso de la mayoría de los fondos de inversión o los *Exchange-Trade Funds*. Al intentar adaptar la simulación financiera en cuestión a este efecto de capitalización, se encuentran ciertas limitaciones. Se podría calcular una aproximación de una tasa que asuma reinversión multiplicando los retornos de cada jornada, al ser considerados como eventos independientes, pero al existir jornadas en las que el retorno es de -100%, no es posible obtener esta aproximación. Por lo tanto, no hay una manera viable de comparar esta simulación con activos que asumen reinversión de ganancias.

Otro factor que considerar es el periodo de la tasa de rentabilidad. Los porcentajes de retorno se expresan típicamente de forma anual, ya que proporcionan una medida estandarizada y comparable de rendimiento a lo largo de un año. La tasa de retorno anual

permite evaluar y comparar diferentes activos o estrategias en un marco temporal consistente. Por este motivo, se anualiza el retorno obtenido durante las dos temporadas con la siguiente formula.

Figura 3- Formula para la anualización de retornos

$$\text{Annual Return Formula} = \left(\frac{\text{Ending Value}}{\text{Initial Value}}\right)^{(1\ /\ \text{No. of Years})} - 1$$

Fuente: https://www.educba.com/annual-return-formula/

Se considera como capital inicial el total del capital invertido a lo largo de las 2 temporadas de simulación, es decir, el número de jornadas en las que se apuesta multiplicado por 1.000€. Aunque en realidad este capital no se requiere en su totalidad al inicio de la simulación, sino que se va aportando semanalmente, por motivos de simplificación del cálculo se asume dicha suposición. De igual manera, se considera el capital final como la suma de los valores obtenidos al final de cada jornada en la que se realizan apuestas.

En relación al número de años, se busca adaptar el cálculo a la naturaleza de la simulación de la siguiente manera, se parte del hecho de que un año consta de 52 semanas, y se asume que el retorno se ha obtenido en 76 semanas, las cuales componen las 2 temporadas de evaluación. Si se considera la duración real durante la cual se han obtenido los beneficios, esta sería de 1,67 años, entre septiembre de 2020 y mayo de 2022, equivalente a aproximadamente 88 semanas. Esta diferencia es consecuencia del período de descanso entre temporadas, durante el cual la simulación no puede funcionar. Los activos con los que se va a realizar la comparación se encuentran disponibles durante todos los días del año, sin un período de inactividad similar. Para que la comparación sea más consistente, se decide asumir que no existe descanso entre temporadas y usar el período de 76 semanas, que al dividirse entre el número de semanas en un año resulta en 1,46 años.

Al introducir los valores respectivos de la simulación de victorias locales y visitantes, se obtienen las tasas de rentabilidad anuales de 3,8% para la simulación local y de 5,7% para la visitante, las cuales serán utilizadas para la comparación con los activos financieros.

Comparación con activos convencionales

A continuación, se procede a contrastar estos retornos anuales con los de ciertos activos de distinta naturaleza, pero que no asumen la reinversión de las ganancias de capital, dado que la simulación implementada tampoco lo hace. Los activos seleccionados para la comparativa son bonos y acciones.

Antes de proceder con la comparación con los mencionados activos, es pertinente establecer una comparativa con la tasa libre de riesgo. Esta, también denominada como *risk-free rate*, es el rendimiento obtenido de una inversión sin riesgo de pérdida de capital. La existencia de esta tasa implica que, en la práctica, no se asume un riesgo a menos que se espere obtener un rendimiento mayor a la tasa libre de riesgo, por lo tanto, se considera como el rendimiento mínimo esperado de cualquier inversión. Se suele considerar la tasa libre de riesgo como el retorno ofrecido por los *Treasury Bills*, que son instrumentos de deuda a corto plazo emitidos por el gobierno de los Estados Unidos. Estos activos se caracterizan por su alta liquidez y se consideran de riesgo cercano a inexistente, ya que están respaldados por la solvencia del gobierno de los Estados Unidos. Para obtener la tasa libre de riesgo actual, se consulta la página de Fama French[17], donde se pueden encontrar la tasa de retorno de los *Treasury Bills* con vencimiento de 1 mes. El último dato disponible corresponde al mes de abril de 2023, y es de 0,35%, que al ser anualizado resulta en un 4,3%.

Esta cifra se sitúa entre las rentabilidades anuales de las simulaciones introducidas, lo cual implica que la relacionada a las victorias locales no supera el *risk-free rate*. A pesar de que no se han establecido cálculos concretos, el simple hecho de que las simulaciones se basan apuestas deportivas sirve para afirmar que el riesgo es significativamente mayor al de las *Treasury Bills*, lo que unido al menor retorno hacen de la simulación local una opción a desestimar. Por este motivo, no se establecerán posteriores comparaciones de la simulación de victorias locales con los activos propuestos. Por otro lado, la simulación de victorias visitantes brinda una tasa de retorno anual que se sitúa 1,4% por encima de la tasa libre de riesgo, superando el umbral fundamental que representa esta tasa sin riesgo.

[17] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Comparación con bonos

Seguidamente, se contrasta la simulación visitante con bonos. Estos son instrumentos financieros de deuda que los entes gubernamentales o corporativos emiten para adquirir capital, y que se agrupan dentro de los llamados activos de renta fija. Los inversores en bonos facilitan capital al emisor a cambio de pagos periódicos de intereses, a menos que se trate de un bono de cupón cero, y la restitución del principal al vencimiento del bono. En primer lugar, cabe puntualizar que se van a emplear bonos españoles, tanto soberanos como corporativos, dado que la simulación financiera se basa en la competición de fútbol del mismo país. Adicionalmente, es relevante indicar que se utiliza el *Yield to Maturity* (*YTM*) para establecer la comparación. El *YTM* es similar a una tasa interna de retorno anual para un bono, y se calcula considerando tanto los pagos de intereses periódicos como cualquier ganancia o pérdida de capital que resultaría de mantener el bono hasta su vencimiento. Se es consciente de que no es la métrica óptima para establecer esta comparación, ya que asume que los cupones obtenidos se reinvierten a la misma tasa que el *YTM* actual y también asume que el bono se conserva hasta el vencimiento, además de que la simulación de apuestas posee un perfil de riesgo y flujos de efectivo muy distintos a los de un bono, pero es la información más coherente para la comparación de la que se dispone. Finalmente, se emplean los últimos valores de los *YTM* disponibles, aunque lo idóneo sería obtener el *YTM* promedio para el periodo en el que se realiza la simulación, estableciendo así una comparación coherente en aspecto temporal.

En el caso de los bonos emitidos por el Estado de España, los rendimientos oscilan entre el 3,1%, en los de menor vencimiento, y el 3,9% en los de mayor vencimiento, según Factset. Asombrosamente, estas tasas resultan inferiores a la tasa libre de riesgo, pero los motivos detrás de esta situación se alejan del propósito de este estudio. Dicho esto, la simulación visitante supera el rendimiento ofrecido por los bonos del Estado.

En cuanto a los bonos corporativos, se ha tratado de encontrar un índice de bonos corporativos de empresas españolas. Los índices son comúnmente usados para la determinación de *benchmarking* de desempeño, dado que representan un amplio conjunto de activos, lo cual implica cierta diversificación. Sin embargo, el inconveniente es que el índice encontrado, *iBoxx EUR Spain*, no cuenta con información pública necesaria, como podría ser la falta de información sobre los componentes. Dada esta situación, se descarta el uso de dicho índice, y se selecciona un grupo de entidades españolas consolidadas, obteniendo información sobre los *YTM* de sus bonos.

Tabla 6- Información relevante de bonos emitidos por empresas españolas

Entidad	YTM	Fecha de emisión	Fecha de vencimiento	ISIN
Grifols	7,2%	oct-21	oct-28	XS2393001891
Telefónica	6,4%	mar-19	mar-49	US87938WAX11
Indra	6,1%	dic-16	dic-26	XS1542249559
Telefónica	5,3%	mar-17	mar-27	US87938WAT09
ACS	4,8%	jun-20	jun-25	XS2189592616
Merlin Properties	4,7%	jul-20	jul-27	XS2201946634
Repsol	4,4%	ene-16	ene-31	XS1352121724
Santander	3,8%	feb-06	feb-26	ES0413900129
Santander	3,7%	ene-23	ene-26	ES0413900905
Repsol	3,6%	dic-14	dic-26	XS1148073205
Iberdrola	3,5%	nov-22	nov-32	XS2558966953
Iberdrola	3,4%	nov-22	nov-28	XS2558916693

La comparación ideal entre estos bonos y la simulación visitante implicaría el uso de una *risk-adjusted metric*, es decir, una medida que no solo tenga en cuenta el retorno, sino que también considere el riesgo que conlleva la inversión. Sin embargo, este tipo de métricas se utilizan comúnmente en activos con cierta variabilidad en los retornos, como las acciones o los fondos de inversión. Los bonos se caracterizan por su baja variabilidad en retornos, lo que los convierte en activos de bajo riesgo en la mayoría de los casos, de forma que no es común calcular *risk-adjusted metrics* para estos.

Dado el bajo riesgo asociado a los bonos, lo que conduce a la no utilización de métricas ajustadas al riesgo, resultaría lógico postular que la simulación construida no se considera un activo atractivo en comparación con los bonos corporativos. A pesar de que dicha simulación ofrece un rendimiento superior a varios de estos bonos, dicho incremento no es suficientemente considerable como para compensar el riesgo adicional que se asume al invertir en la simulación.

No obstante, se procede a cuantificar el riesgo tanto de los bonos como de la simulación. En el caso de la simulación, al considerar cada jornada como un evento independiente y al observar el rendimiento obtenido de los 1.000€ invertidos en cada jornada, se evidencian situaciones extremas, que van desde un retorno del -100% en la jornada 24 de la temporada 2020/2021 a un retorno del 105% en la jornada posterior, poniendo de manifiesto una volatilidad de los retornos extremadamente elevada. Sin embargo, debido a que este enfoque podría parecer excesivo y poco comparable con un activo como un

bono, se decide calcular la variación de los retornos tomando el capital total, es decir, 50.000€ en lugar de los 1.000€ semanales.

Con este método, tras la primera jornada en la que se realizan apuestas, se sumará o restará a los 50.000€ en función de si se obtienen ganancias o pérdidas, y se calculará la variación, que proporcionará el retorno para la primera jornada. Al repetirse el mismo proceso todas las jornadas se obtiene un retorno por jornada. Este enfoque también presenta ciertas limitaciones, dado que el retorno por jornada está acotado por el hecho de que solo se pueden perder 1.000€, y también tiene un límite superior, dado que es difícil obtener más de 2.000€ en una jornada, lo que prácticamente impide que los retornos sean mayores del 4%. Sin embargo, este método de cálculo de retornos se considera más adecuado para la comparación con activos financieros.

Así, se calcula la desviación estándar de estos cambios en el valor total, que se pueden considerar una especie de retorno semanal, obteniéndose un valor de 1,4. Posteriormente, se seleccionan aleatoriamente dos de los bonos de la tabla presentada, el de ACS y el de Merlin Properties, y se calcula el equivalente de esta desviación estándar de los retornos semanales, utilizando el cambio de precios semanal para dichos bonos, obteniendo valores de 0,4 para ACS y 0,6 para Merlin Properties. Si bien estos cálculos pueden no ser perfectos debido a las características de los bonos, resultan suficientemente adecuados para demostrar que la variación en los retornos es significativamente más extrema en el caso de la simulación, lo que indica una mayor volatilidad, y por ende, un mayor riesgo, el cual no se ve compensado por un retorno significativamente mayor.

Comparación con acciones

Posteriormente, se compara la simulación con acciones, que son instrumentos financieros que representan una unidad de propiedad en una empresa, y suelen ser más volátiles que los bonos. Además, la comparación con este activo se va a poder realizar de forma más precisa y coherente. Primero, existe un índice de acciones en España, el IBEX35. Este índice representa las 35 empresas con mayor liquidez que cotizan en las 4 Bolsas Españolas, Madrid, Barcelona, Bilbao y Valencia. El IBEX 35 sirve como indicador clave de la economía española y del rendimiento de su mercado de valores, por lo que es utilizado como un *benchmark* o referencia para comparar el rendimiento de carteras o estrategias de inversión. En este caso, se dispone de información histórica del precio del índice, lo que permite calcular el retorno obtenido si se hubiese invertido durante el

mismo periodo de la simulación, y obtener así una comparación consistente temporalmente.

Dicho esto, se calcula el retorno obtenido entre la fecha de inicio y de final de la simulación. El primer partido en el que se apuesta es el 13 de septiembre de 2020, fecha en la que el precio del IBEX35 se sitúa en 6.951€. El último partido en el que se apuesta ocurre el 22 de mayo de 2022, ubicándose el IBEX35 en 8.626€. Los precios mencionados del índice resultan en un retorno anual del 10,4%, casi el doble del 5,7% de rentabilidad observado en la simulación visitante.

Es pertinente considerar el cálculo de las *risk-adjusted metrics*, ya que permiten una evaluación integral del retorno y el riesgo. La primera de estas métricas es el Ratio de Sharpe, una medida desarrollada para determinar el rendimiento excesivo que una inversión genera por unidad de riesgo asumido (Sharpe, 1994). Un valor más alto de este ratio indica un rendimiento ajustado al riesgo superior.

Figura 4- Ecuación del Ratio de Sharpe

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

where:

$R_p = \text{return of portfolio}$

$R_f = \text{risk-free rate}$

$\sigma_p = \text{standard deviation of the portfolio's excess return}$

Fuente: Investopedia

El Ratio de Sharpe implica restar la tasa libre de riesgo del rendimiento esperado de la inversión y dividir ese resultado por la desviación estándar de los rendimientos de la inversión. Se calcula un Ratio de Sharpe diario para el IBEX35 en el periodo indicado, y luego se anualiza, obteniendo un valor de 0,8. En el caso de la simulación, debido a su naturaleza específica, se opta por calcular un Ratio de Sharpe semanal basado en los retornos por jornada tal y como se establece en la comparación con los bonos, seguido de su anualización, lo que resulta en un valor de 0,6. En términos de eficiencia del riesgo-retorno, el IBEX35 se presenta como una inversión más favorable, pues su Ratio de Sharpe supera al de la simulación, lo que sugiere que brinda un retorno superior por cada unidad de riesgo asumido.

Finalmente, se procede al cálculo del Ratio de Sortino, que se enfoca específicamente en la volatilidad a la baja, es decir, en el riesgo asociado a los retornos negativos, conocido como *downside risk*. De este modo, proporciona una medición del rendimiento excesivo sobre la tasa libre de riesgo por unidad de volatilidad negativa (Sortino & Price, 1994).

Figura 5- Ecuación del Ratio de Sortino

$$\text{Sortino Ratio} = \frac{R_p - r_f}{\sigma_d}$$

where:

R_p = Actual or expected portfolio return

r_f = Risk-free rate

σ_d = Standard deviation of the downside

Fuente: Investopedia

Para su cálculo, se determina la desviación estándar de los retornos negativos tanto para la simulación como para el IBEX35. Los valores resultantes del Ratio de Sortino son 1,2 para el IBEX35 y 0,7 para la simulación. Estos valores corroboran la afirmación previamente postulada, en la que se argumenta que el riesgo adicional inherente a la simulación no se compensa con un incremento adecuado en el retorno.

A pesar de que los retornos por jornada están limitados por la inversión de 1.000€ en cada una de ellas, estos presentan una mayor volatilidad en comparación con el IBEX35. En consecuencia, el retorno por unidad de riesgo es inferior en la simulación, tanto en el Ratio de Sharpe como en el Ratio de Sortino.

Figura 6. Detalle del cálculo de los ratios de Sharpe y Sortino para la simulación de victorias visitantes

Average excess return (weekly)	0,12

Sharpe Ratio	
Return standard deviation (weekly)	1,37
Weekly Sharpe Ratio	0,09
Annualized Sharpe Ratio	0,6
*Weekly Sharpe * √52*	

Sortino Ratio	
Downside return standard deviation (weekly)	0,71
Weekly Sortino Ratio	0,10
Annualized Sortino Ratio	0,7
*Weekly Sortino * √52*	

Fuente: Elaboración propia

Figura 7. Detalle del cálculo de los ratios de Sharpe y Sortino para el IBEX35

Average excess return (daily)	0,05

Sharpe Ratio	
Return standard deviation (daily)	1,01
Daily Sharpe Ratio	0,05
Annualized Sharpe Ratio	0,8
*Daily Sharpe * √252*	

Sortino Ratio	
Downside return standard deviation (daily)	0,70
Daily Sortino Ratio	0,07
Annualized Sortino Ratio	1,2
*Daily Sortino * √252*	

Fuente: Elaboración propia

Conclusión

En respuesta a la pregunta central de investigación: ¿Pueden las apuestas deportivas ser consideradas como un activo financiero? La conclusión es negativa. El análisis evidencia que el retorno proporcionado por la simulación más rentable de las implementadas no logra ser suficientemente robusto como para ser considerado un activo que ofrezca rendimientos ajustados al riesgo comparables a activos convencionales tales como bonos o acciones. Este hecho revela la insuficiencia de los datos empleados para el entrenamiento de los modelos, tal como se evidencia por la escasez de variables significativas, aparte de las variables dicotómicas de identificación de los equipos.

Es relevante señalar que el presente estudio no ha tomado en consideración el impacto fiscal que podría incidir en la estrategia de inversión en las apuestas deportivas, factor que probablemente distanciaría aún más su rendimiento del ofrecido por los activos convencionales. No obstante, un desarrollo potencial de los modelos y la estrategia, que incorpore nuevos datos valiosos, quizá procedentes de otras ligas para ampliar las oportunidades de apuesta, junto a la optimización del *threshold* de clasificación para alcanzar un equilibrio que reduzca errores sin desaprovechar oportunidades rentables, podría eventualmente convertir la respuesta a la pregunta de investigación en un afirmativo.

Bibliografía

Bergmeir, C., & Benítez, J. M. (2012). *On the use of cross-validation for time series predictor evaluation*. Information Sciences, 191, 192-213.

Caro, A. (2020). *Modelos estadísticos para la predicción de La Liga 2019/2020*. Universidad de Sevilla, Grade en Estadística. España. Recuperado de: https://idus.us.es/handle/11441/114937

Cantinotti, M., Ladouceur, R., & Jacques, C. (2004*). Sports betting: Can gamblers beat randomness?* Psychology of Addictive Behaviors, 18(2), 143–147.

Daoud, J. I. (2017). *Multicollinearity and Regression Analysis*. Journal of Physics: Conference Series, 949.

Fama, E. F., & French, K. R. (2004). *The Capital Asset Pricing Model: Theory and Evidence*. Journal of Economic Perspectives, 18(3), 25-46.

Gomes, J., Portela, F., & Santos, M. F. (2021). *Decision Support System For Predicting Football Game Resul*t. Recuperado de: https://www.inase.org/library/2015/zakynthos/bypaper/COMPUTERS/COMPUTERS-57.pdf

Hing, N., Russell, A. M., & Browne, M. (2017). *Risk Factors for Gambling Problems on Online Electronic Gaming Machines, Race Betting and Sports Betting*. Frontiers in Psychology, 8.

Killick, E. A., & Griffiths, M. D. (2019). *In-Play Sports Betting: a Scoping Study*. International Journal of Mental Health and Addiction, 17, 1456–1495.

Killick, E. A., & Griffiths, M. D. (2022). *Sports Betting Advertising: A Systematic Review of Content Analysis Studies*. International Journal of Mental Health and Addiction, 20(1).

Lopez-Gonzalez, H., & Griffiths, M. D. (2018). Understanding the convergence of markets in online sports betting. The International Review for the Sociology of Sport, 53(7).

Markowitz, H. (1952). *Portfolio Selection*. The Journal of Finance, 7(1), 77-91

Pollard, R., & Gómez Ruano, M. Á. (2014). *Components of home advantage in 157 national soccer leagues worldwide.* International Journal of Sport and Exercise Psychology, 12(3), 218-233.

Pugsee, P., & Pattawong, P. (2019). *Football Match Result Prediction Using the Random Forest Classifier*. Proceedings of the 2nd International Conference on Big Data Technologies - ICBDT2019, 154-158. https://doi.org/10.1145/3358528.3358593

Rodrigues, F., & Pinto, Â. (2022). *Prediction of football match results with Machine Learning*. Procedia Computer Science, 204, 463-470. Recuperado de: https://doi.org/10.1016/j.procs.2022.08.057

Sharpe, W. F. (1994). *The Sharpe Ratio*. The Journal of Portfolio Management, 21(1), 49-58.

Sortino, F. A., & Price, L. N. (1994). *Performance measurement in a downside risk framework.* Journal of Investing, 3(3), 59-64.

Ulmer, B., & Fernandez, M. (2014*). Predicting Soccer Match Results in the English Premier League*. Recuperado de: https://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the

Ying, X. (2019). *An Overview of Overfitting and its Solutions*. Journal of Physics: Conference Series, 1168(2).

Anexos

Anexo 1- Código para la simulación de victorias visitantes

```
investment_per_jornada = 1000
capital = 0
investments = []
jornada_games = df_merged.groupby('Jornada')

for jornada, games in jornada_games:
    games_with_prediction = games[games['Prediction'] == 1]
    num_games = len(games_with_prediction)
    if num_games > 0:
        total_confidence = games_with_prediction['Probability'].sum()
        for _, game in games_with_prediction.iterrows():
            home_team = game['HomeTeam']
            away_team = game['AwayTeam']
            actual_outcome = game['Target']
            bet_amount = investment_per_jornada * (game['Probability'] / total_confidence)
            if game['Prediction'] == actual_outcome:
                money_made = bet_amount * game['Away_mean_odds']
                capital += money_made
            else:
                money_made = 0
            investment = {
                'Jornada': jornada,
                'Date': game['Date_x'],
                'HomeTeam': home_team,
                'AwayTeam': away_team,
                'BetAmount': bet_amount,
                'ActualOutcome': actual_outcome,
                'HomeTeamOdds': game['Away_mean_odds'],
                'MoneyMade': money_made
            }
```

```
        investments.append(investment)

temp_investments_df = pd.DataFrame(investments)
investment_count = temp_investments_df['Jornada'].nunique()
total_investment = investment_per_jornada * investment_count
total_profits = capital - total_investment
return_rate = total_profits / total_investment
years = investment_count / 38
annualized_return = ((1 + return_rate) ** (1 / years)) - 1
```

Anexo 2- Resultados de la simulación de partidos locales

Jornada	Fecha	Local	Visitante	Cantidad Apostada	Resultado	Cuota	Cantidad final
3	26/09/2020	Elche	Real Sociedad	1000	1	1,964	1964
4	30/09/2020	Huesca	Atletico Madrid	537	0	1,48	0
4	01/10/2020	Celta de Vigo	Barcelona	463	1	1,558	722
5	04/10/2020	Levante	Real Madrid	1000	1	1,51	1510
6	17/10/2020	Celta de Vigo	Atletico Madrid	1000	1	1,788	1788
8	31/10/2020	Alaves	Barcelona	1000	0	1,46	0
11	28/11/2020	Huesca	Sevilla	1000	1	2,01	2010
12	05/12/2020	Cadiz	Barcelona	1000	0	1,364	0
13	12/12/2020	Getafe	Sevilla	1000	1	2,11	2110
14	20/12/2020	Eibar	Real Madrid	1000	1	1,544	1544
15	22/12/2020	Valencia	Sevilla	528	1	2,06	1087
15	22/12/2020	Valladolid	Barcelona	472	1	1,394	658
16	30/12/2020	Elche	Real Madrid	1000	0	1,272	0
17	03/01/2021	Huesca	Barcelona	564	1	1,47	828
17	03/01/2021	Alaves	Atletico Madrid	436	1	1,57	685
18	09/01/2021	Granada	Barcelona	481	1	1,426	687
18	09/01/2021	Osasuna	Real Madrid	519	0	1,48	0
20	23/01/2021	Alaves	Real Madrid	447	1	1,574	704
20	24/01/2021	Elche	Barcelona	553	1	1,362	753
21	31/01/2021	Cadiz	Atletico Madrid	1000	1	1,472	1472
22	06/02/2021	Elche	Villarreal	304	0	1,57	0
22	06/02/2021	Huesca	Real Madrid	384	1	1,47	564
22	07/02/2021	Betis	Barcelona	313	1	1,534	480
24	20/02/2021	Valladolid	Real Madrid	1000	1	1,594	1594
25	28/02/2021	Cadiz	Betis	1000	1	2,056	2056
26	06/03/2021	Osasuna	Barcelona	507	1	1,522	772
26	06/03/2021	Elche	Sevilla	493	0	1,612	0
27	13/03/2021	Getafe	Atletico Madrid	1000	0	1,93	0
28	20/03/2021	Celta de Vigo	Real Madrid	538	1	1,744	937
28	21/03/2021	Real Sociedad	Barcelona	462	1	1,992	921
30	12/04/2021	Celta de Vigo	Sevilla	1000	1	2,208	2208
31	18/04/2021	Getafe	Real Madrid	1000	0	1,636	0
32	21/04/2021	Levante	Sevilla	447	1	1,66	742
32	21/04/2021	Cadiz	Real Madrid	553	1	1,382	764
34	01/05/2021	Elche	Atletico Madrid	465	1	1,414	657
34	02/05/2021	Valencia	Barcelona	535	1	1,374	736

36	11/05/2021	Levante	Barcelona	425	0	1,21	0
36	13/05/2021	Granada	Real Madrid	302	1	1,402	423
36	13/05/2021	Valladolid	Villarreal	274	1	1,968	539
37	16/05/2021	Athletic Bilbao	Real Madrid	1000	1	1,6	1600
38	22/05/2021	Eibar	Barcelona	344	1	1,666	573
38	22/05/2021	Valladolid	Atletico Madrid	362	1	1,338	484
38	22/05/2021	Osasuna	Real Sociedad	294	1	1,624	478
39	14/08/2021	Alaves	Real Madrid	1000	1	1,452	1452
40	21/08/2021	Athletic Bilbao	Barcelona	314	0	1,892	0
40	22/08/2021	Levante	Real Madrid	331	0	1,64	0
40	23/08/2021	Getafe	Sevilla	356	1	2,246	799
44	21/09/2021	Getafe	Atletico Madrid	455	1	1,686	767
44	23/09/2021	Cadiz	Barcelona	545	0	1,516	0
45	25/09/2021	Alaves	Atletico Madrid	1000	0	1,53	0
46	03/10/2021	Espanol	Real Madrid	1000	0	1,654	0
49	27/10/2021	Mallorca	Sevilla	457	0	1,78	0
49	27/10/2021	Rayo Vallecano	Barcelona	543	0	1,786	0
50	30/10/2021	Elche	Real Madrid	1000	1	1,442	1442
52	21/11/2021	Granada	Real Madrid	1000	1	1,468	1468
53	28/11/2021	Cadiz	Atletico Madrid	535	1	1,492	798
53	28/11/2021	Espanol	Real Sociedad	465	0	2,306	0
57	02/01/2022	Getafe	Real Madrid	1000	0	1,672	0
58	08/01/2022	Granada	Barcelona	1000	0	1,758	0
59	16/01/2022	Elche	Villarreal	1000	0	1,774	0
60	23/01/2022	Alaves	Barcelona	1000	1	1,566	1566
62	12/02/2022	Cadiz	Celta de Vigo	1000	0	2,056	0
63	19/02/2022	Osasuna	Atletico Madrid	501	1	2,092	1049
63	20/02/2022	Valencia	Barcelona	499	1	2,002	998
64	26/02/2022	Rayo Vallecano	Real Madrid	1000	1	1,66	1660
65	06/03/2022	Elche	Barcelona	1000	1	1,424	1424
66	14/03/2022	Mallorca	Real Madrid	1000	1	1,564	1564
69	09/04/2022	Mallorca	Atletico Madrid	478	0	1,94	0
69	10/04/2022	Levante	Barcelona	522	1	1,448	756
71	20/04/2022	Osasuna	Real Madrid	1000	1	1,674	1674
74	11/05/2022	Elche	Atletico Madrid	544	1	1,666	906
74	12/05/2022	Rayo Vallecano	Villarreal	456	1	1,81	826
75	15/05/2022	Getafe	Barcelona	1000	0	2,134	0