MODELADO ESTADÍSTICO DE RESULTADOS DE FÚTBOL MEDIANTE LA DISTRIBUCIÓN DE POISSON: APLICACIÓN PRÁCTICA AL MERCADO DE APUESTAS DEPORTIVAS

MODALIDAD DEL TFG: CONVENCIONAL

CONVOCATORIA: EXTRAORDINARIA

ALUMNA: ANA ISABEL GONZÁLEZ SAHAGÚN

TUTORA: MARÍA DEL MAR ANGULO MARTÍNEZ

GRADO: MATEMÁTICAS COMPUTACIONALES

AGRADECIMIENTOS

Cuando decidí cambiarme al doble grado, pasé un par de años complicados. Llegué a cursar hasta cuatro asignaturas de Matemáticas por cuatrimestre, además de las cinco propias de Ingeniería, lo que supuso un gran reto para mí. Por eso, quiero agradecer especialmente a mi profesor particular de Matemáticas, Jota.

Jota dedicó incontables horas a estudiar mis apuntes y a enseñarme de la forma más clara y eficiente posible, siempre con el objetivo de que me diese tiempo a sacar todo adelante. Ha sido un profesor excepcional, y sin él no podría haber llegado hasta aquí.

También quiero dar las gracias a mis abuelos, a los dos que me han acompañado durante toda la carrera, y a los dos que ya no están. Aunque no hayan podido verme graduar, estoy segura de que estarían muy orgullosos de mí.

Y, por supuesto, quiero agradecer profundamente a Mar, por su dedicación y su pasión por la docencia. Gracias a ella, alumnas como yo desarrollamos no solo conocimientos, sino también amor por lo que hacemos.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Este Trabajo de Fin de Grado desarrolla y analiza distintos modelos estadísticos basados en la distribución de Poisson para predecir resultados de partidos de fútbol y su aplicación práctica en el mercado de apuestas deportivas. Se implementan cuatro modelos principales: Poisson simple, Poisson doble, Poisson bivariante y regresión de Poisson, utilizando datos históricos de LaLiga. A través de un análisis exploratorio de datos, construcción de modelos y simulaciones de apuestas, se evalúa la capacidad predictiva de cada enfoque.

Los resultados muestran que, aunque el modelo Poisson doble presenta el mejor ajuste estadístico, la regresión de Poisson ofrece un mejor equilibrio en la predicción de empates y mayor rentabilidad práctica en simulaciones de apuestas. El trabajo concluye que, en contextos como el mercado de apuestas, modelos más flexibles y contextuales pueden superar a los enfoques teóricamente más ajustados, evidenciando la utilidad de la estadística aplicada en entornos de alta incertidumbre.

ABSTRACT

This study Project develops and analyzes various statistical models based on the Poisson distribution to predict football match outcomes and evaluate their practical application in the sports betting market. Four main models are implemented: simple Poisson, double Poisson, bivariate Poisson, and Poisson regression, using historical data from the Spanish competition LaLiga. Through exploratory data analysis, model building, and betting simulations, the predictive capacity of each approach is assessed.

Results indicate that while the double Poisson model provides the best statistical fit, the Poisson regression model achieves a better balance in draw predictions and yields the highest simulated profitability. The study concludes that in practical contexts like betting markets, flexible and contextual models can outperform more theoretically accurate ones, demonstrating the value of applied statistics in highly uncertain environments.

1. INTRODUCCIÓN

1.1 Motivación y contexto

El deporte juega un papel importante en el crecimiento económico y el desarrollo social de muchos países. No solo aporta ingresos a través de eventos, derechos de televisión o patrocinios, sino que también crea empleo y estimula la actividad comercial. Al movilizar a millones de personas en todo el mundo, se convierte en un fenómeno de gran relevancia social. Entre todas las disciplinas deportivas, el fútbol cuenta con el mayor número de seguidores a nivel mundial, con una audiencia estimada de más de 4.000 millones de personas.[1]. En España, según el informe de KPMG (2023) [1], el fútbol profesional generó más de 18.300 millones de euros durante la temporada 2021/22, lo que equivale al 1,44 % del PIB nacional.

El fútbol ocupa un lugar privilegiado como uno de los deportes que más apuestas generan, gracias su enorme base de seguidores y la cantidad de competiciones a lo largo del año. Solo en la temporada 2021/22, los aficionados españoles destinaron cerca de 2.954 millones de euros a quinielas y apuestas vinculadas al fútbol[2]. Además, se espera que el mercado mundial de apuestas deportivas siga creciendo a un ritmo cercano al 9% para los próximos años[3], gracias a la digitalización y a la expansión de plataformas legales. Este sector se basa, en gran medida, en la capacidad de anticipar los resultados de los partidos. Las casas de apuestas utilizan métodos estadísticos para calcular cuotas que reflejan probabilidades estimadas, mientras que muchos usuarios buscan identificar oportunidades para apostar con ventaja. Por eso, contar con modelos que permitan estimar con precisión las probabilidades de cada resultado tiene un valor práctico tanto para operadores como para analistas.

Sin embargo, predecir los resultados de un partido de fútbol es una tarea compleja debido a la alta incertidumbre inherente al juego. A diferencia de otros deportes con marcadores más amplios (por ejemplo, el baloncesto), el fútbol se caracteriza por ser un deporte de anotaciones escasas, donde un solo gol puede decidir un encuentro. Este alto grado de aleatoriedad, junto

---

[1] ESPN. (2021, 18 abril). El fútbol es el deporte más popular del mundo: más de 4 mil millones de aficionados lo siguen. ESPN Deportes. https://espndeportes.espn.com/

[2] El País. (2023, 28 septiembre). El fútbol profesional duplica su peso en el PIB español en una década. https://elpais.com/

[3] Statista. (2024, 12 marzo). Valor de mercado de las apuestas deportivas en el mundo. https://es.statista.com/

con factores externos como las decisiones arbitrales o el estado físico de los jugadores, hace que sea muy difícil prever con exactitud lo que ocurrirá en un partido. En términos técnicos, el fútbol es un suceso *altamente variable*. Ante este nivel de incertidumbre, surge la necesidad de herramientas estadísticas que permitan estimar con mayor precisión las probabilidades de los distintos resultados posibles.

Uno de los enfoques más utilizados para modelar resultados de fútbol desde una perspectiva estadística es el uso de la distribución de Poisson. Esta herramienta permite estimar el número de goles que puede marcar un equipo en un partido, tratándolos como sucesos que ocurren de forma independiente en un intervalo fijo de tiempo. Dado que el fútbol es un deporte con pocas anotaciones por encuentro, esta distribución resulta especialmente adecuada. Se asume que cada equipo tiene una tasa de gol media y que los goles se producen de forma aleatoria, pero con cierta regularidad. Aunque esta suposición simplifica la realidad, ya que en un partido los equipos interactúan y ejercen influencia entre sí, numerosos estudios han respaldado el uso del modelo de Poisson. Desde los primeros trabajos de Maher (1982) [2] y de Dixon y Coles (1997) [3], hasta investigaciones más recientes como el de Nguyen (2021) [4], se ha demostrado que este modelo ofrece una representación razonablemente precisa de la distribución de los marcadores en distintas competiciones. A lo largo del tiempo, se han desarrollado variantes que incorporan factores como la fortaleza ofensiva y defensiva de los equipos o el efecto de jugar en casa. Estos modelos han sido utilizados tanto por analistas deportivos como por operadores de apuestas para calcular probabilidades y estimar cuotas. Su ventaja principal es que, con pocos parámetros, permiten generar estimaciones sobre resultados posibles y evaluar la probabilidad de que se produzcan.

Teniendo en cuenta lo anterior, este Trabajo de Fin de Grado tiene como objetivo desarrollar un modelo estadístico basado en la distribución de Poisson para predecir los goles y resultados en partidos de fútbol, y analizar su utilidad práctica en el contexto del mercado de apuestas deportivas. Este modelo será implementado para estimar la probabilidad de los distintos resultados posibles en un partido, considerando factores como el rendimiento ofensivo y defensivo de los equipos o la influencia de jugar como local. Posteriormente, se compararán las predicciones obtenidas con las cuotas reales ofrecidas por las casas de apuestas, con el fin de evaluar si existen diferencias significativas que puedan representar oportunidades de valor. Este enfoque permite aplicar herramientas estadísticas de forma práctica en un contexto de alta incertidumbre, y aporta una base cuantitativa para analizar el funcionamiento del mercado de apuestas deportivas.

1.2 Planteamiento del problema

La predicción de resultados en el fútbol representa un reto estadístico debido a la complejidad y la variabilidad que caracterizan a este deporte. Aunque existen herramientas matemáticas que permiten estimar probabilidades, la naturaleza aleatoria del juego y las limitaciones inherentes a los modelos empleados generan incertidumbre sobre la fiabilidad de estas predicciones. En este apartado se analiza la problemática principal a la que se enfrenta cualquier intento de modelado estadístico del fútbol: desde la incertidumbre propia de los resultados deportivos, pasando por las limitaciones de los modelos, hasta la necesidad de definir adecuadamente qué significa el éxito de un equipo para poder construir estimaciones coherentes y útiles.

1.2.1 Incertidumbre en los Resultados Deportivos

El fútbol, al ser un deporte de baja puntuación, presenta una elevada incertidumbre en sus resultados. La diferencia entre ganar o perder puede depender de una sola jugada, lo que otorga al azar un papel importante, incluso cuando se enfrentan equipos de distintos niveles. Factores externos como el clima, el arbitraje o el estado físico de los jugadores aumentan esta variabilidad, dificultando cualquier predicción precisa.

Esta aleatoriedad ha sido analizada en profundidad por Aoki et al. (2017) [5], quienes introducen un coeficiente de habilidad (φ) para medir el peso relativo de la suerte y la habilidad en competiciones deportivas. Su estudio, basado en más de 1.500 temporadas de ligas de fútbol, baloncesto, voleibol y balonmano, demuestra que el fútbol es uno de los deportes donde el azar tiene mayor protagonismo. En torno al 7 % de las temporadas de fútbol analizadas se comportan como si fueran completamente aleatorias. Además, muestran que, en competiciones como La Liga o la Premier League, basta con eliminar tres o cuatro equipos dominantes para que el resto de la competición se asemeje a una lotería.

Este nivel de imprevisibilidad hace que los modelos deterministas sean poco útiles, y justifica el uso de herramientas probabilísticas. Modelos como el de Poisson permiten incorporar la aleatoriedad inherente al juego, estimando distribuciones de resultados posibles en lugar de un único marcador, lo cual es más realista y útil para la toma de decisiones en contextos como las apuestas deportivas.

1.2.2 Limitaciones de los Modelos Matemáticos y Estadísticos

Los modelos matemáticos y estadísticos aplicados al fútbol, como los basados en la distribución de Poisson, suelen requerir suposiciones simplificadas para poder aplicarse. Uno de los principales supuestos es que se debe asumir que los goles se generan de forma independiente y con una tasa constante a lo largo del tiempo. Sin embargo, esta hipótesis no siempre refleja la realidad del juego, donde factores como el estado físico de los jugadores, las decisiones tácticas o las condiciones climáticas pueden influir significativamente en el desarrollo del partido. Además, muchos modelos no consideran adecuadamente la posibilidad de empates, lo que limita su precisión en competiciones donde este resultado es común. Tal como señalan Dixon y Coles (1997) [3], introducir una ligera dependencia entre las anotaciones puede mejorar significativamente la capacidad predictiva del modelo.

Otra limitación importante es la dependencia de datos incompletos o imperfectos. La calidad de las predicciones generadas por estos modelos se encuentra estrechamente ligada a la disponibilidad y precisión de los datos utilizados. En el contexto del fútbol, factores como lesiones, sanciones, cambios en la alineación o incluso el estado emocional de los jugadores son difíciles de cuantificar y, a menudo, no se incluyen en los conjuntos de datos. Esta omisión puede llevar a predicciones que no capturan adecuadamente la complejidad del juego real.

Además, existe el riesgo de sobreajuste (*overfitting*) en modelos complejos, donde el modelo se ajusta demasiado a los datos históricos, capturando ruido en lugar de patrones reales. Esto reduce su capacidad para generalizar y predecir resultados futuros con precisión. Según Groll et al. (2018) [6], este problema puede reducir drásticamente la capacidad predictiva si no se aplican técnicas de validación adecuadas. El sobreajuste resulta especialmente problemático cuando se dispone de conjuntos de datos limitados o cuando los modelos incorporan un gran número de parámetros sin una validación adecuada.

1.2.3 Valoración del Éxito de los Equipos

Es importante establecer una definición adecuada del éxito para construir modelos estadísticos que sean útiles y coherentes. La forma más directa para medir el desempeño de un equipo es el resultado directo de un partido (victoria, empate, o derrota). Sin embargo, este enfoque puede resultar ineficiente dependiendo del enfoque del estudio. Existen diversas métricas que se utilizan para medir el éxito en un equipo, como por ejemplo la diferencia de goles, el porcentaje de victorias, o la clasificación final en una competición. Por tanto, es necesario analizar previamente qué indicadores permiten representar de forma más precisa el comportamiento de un equipo y su desempeño sostenido en el tiempo.

Cada estudio utiliza unas métricas diferentes para medir el rendimiento de un equipo en el fútbol. Una de ellas es el ***Soccer Power Index (SPI)***, creado por ***FiveThirtyEight***. Este índice combina información histórica con resultados recientes y proporciona una evaluación cuantitativa de la calidad de los equipos. El modelo tiene en cuenta tanto la capacidad ofensiva como defensiva, estimando cuántos goles se espera que marque o reciba un equipo frente a un rival promedio. Los partidos más recientes tienen mayor peso en el cálculo, aunque los datos anteriores también se consideran para mantener una base sólida de referencia. De esta forma, el SPI busca ofrecer una valoración global que capture la fuerza relativa de los equipos en distintos contextos.[4]

Otra métrica ampliamente utilizada en el análisis moderno es el modelo de goles esperados, conocido como **xG (*expected goals*)**. A diferencia de contar los goles realmente marcados, esta métrica evalúa la calidad de cada oportunidad de gol, considerando aspectos como la posición desde la que se dispara, el tipo de pase que precede al remate, la forma del disparo o la situación del juego. El valor de xG refleja la probabilidad de que una acción concreta termine en gol. De este modo, un equipo con un xG alto, aunque no haya anotado muchos goles, puede considerarse ofensivamente productivo por haber generado ocasiones de alta calidad. Estudios como ***Beyond Expected Goals*** del MIT [9] han demostrado que el xG es un predictor más fiable del rendimiento ofensivo futuro que los goles marcados. No obstante, su cálculo requiere una recopilación detallada y precisa de datos, lo que representa una dificultad adicional a la hora de implementarlo en modelos estadísticos a gran escala.

1.3 Objetivos del Trabajo

El objetivo principal de este Trabajo de Fin de Grado es desarrollar y evaluar distintos modelos estadísticos basados en la distribución de Poisson para predecir resultados de fútbol, analizando sus capacidades predictivas, limitaciones estructurales y utilidad práctica en el contexto del mercado de apuestas deportivas. Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

- Realizar un análisis exploratorio de los datos históricos con el fin de identificar patrones y tendencias que puedan orientar y mejorar el desarrollo de los modelos estadísticos.
- Modelar los goles históricos por equipo utilizando el modelo de Poisson simple.

[4] ESPN staff (2014, 11 Junio). Soccer Power Index explained. https://www.espn.com/

- Estimar las probabilidades de victoria local, empate y victoria visitante con el modelo de Poisson doble.
- Mejorar las predicciones del modelo clásico incorporando dependencia entre goles con el modelo de Poisson bivariante.
- Incluir variables explicativas (rendimiento reciente, cuotas, etc.) para enriquecer el modelo mediante regresión de Poisson.
- Evaluar cuantitativamente cada modelo mediante métricas de ajuste y precisión.
- Analizar la aplicabilidad práctica de los modelos simulando estrategias de apuestas.
- Comparar el rendimiento predictivo y económico de los distintos enfoques.
- Identificar las principales limitaciones y supuestos de cada modelo.

Se trata, en definitiva, no solo de construir una base teórica sólida para el modelado de resultados de fútbol, sino también de explorar su utilidad práctica en un entorno real como el de las apuestas deportivas, donde la precisión estadística puede traducirse en una ventaja económica.

2. ESTADO DE LA CUESTIÓN

2.1 Marco Contextual

2.1.1 El Futbol y las Apuestas

El fútbol es el deporte más apostado a nivel mundial, representando hasta el 86% de las apuestas deportivas en algunos países.[5] Su popularidad y la gran cantidad de partidos disputados a lo largo del año lo convierten en el mercado más activo dentro de la industria

Tipos de Apuestas en el Futbol

En las últimas décadas, las apuestas deportivas en este deporte han experimentado una transformación significativa, impulsada por el desarrollo de plataformas digitales y su legalización progresiva en distintos países. Aunque existen muchos tipos de apuestas dentro del mundo del fútbol, este trabajo se centrará en una de las formas más comunes y utilizadas, la **apuesta 1X2**.

Esta apuesta consiste en predecir qué equipo ganará el partido o si el resultado será un empate. A cada una de estas tres opciones se le asigna una cuota, que indica cuánto puede ganar un usuario por cada unidad apostada si su predicción resulta correcta. Estas cuotas no son fijas, sino que pueden cambiar entre diferentes casas de apuestas e incluso variar con el tiempo, en función de factores como el historial reciente de los equipos, las alineaciones previstas, posibles lesiones y el volumen de apuestas recibido por cada resultado.

Incluir las cuotas ofrecidas por las casas de apuestas como variable explicativa en el modelo puede aportar información adicional de valor. Estas cuotas sintetizan de forma implícita el conocimiento agregado del mercado y las expectativas sobre el resultado del partido, por lo que su incorporación podría mejorar la capacidad predictiva del modelo basado en la distribución de Poisson.

Impacto Económico de las Apuestas en el Fútbol

Torneos como la Copa Mundial y la UEFA Champions League están entre los más populares del fútbol y tienen un gran impacto económico en el mercado de las apuestas. Durante estas

[5] (2024, abril) El Pilón. Los favoritos de los aficionados: Los deportes más populares para apostar. https://elpilon.com.co

competiciones, el volumen de apuestas aumenta considerablemente, lo que provoca cambios importantes en la actividad económica asociada. En España, se estima que las apuestas deportivas representan cerca del 1 % del PIB [1]. Además, la influencia del fútbol no se limita al sector del juego, sino que también tiene un efecto relevante en la economía nacional, especialmente en la generación de empleo. Solo en España, el fútbol contribuye con más de 194.000 empleos a jornada completa, incluyendo tanto puestos directos como indirectos relacionados con este deporte.[6]

2.1.2 El Mercado de las Apuestas Deportivas

A lo largo de los años, las apuestas deportivas han experimentado una gran evolución. Lo que en sus orígenes eran prácticas informales, hoy se ha convertido en una industria con un fuerte impacto económico. Uno de los hitos más importantes en esta evolución fue la legalización y regulación de las apuestas en Inglaterra durante el siglo XVIII.[7] En las últimas décadas, el desarrollo tecnológico y la expansión de Internet han impulsado el crecimiento de plataformas digitales, haciendo que apostar en línea sea cada vez más fácil y accesible. Este apartado busca ofrecer una visión general del estado actual del mercado de apuestas deportivas.

Funcionamiento de las Cuotas y el Mercado de Apuestas

En el mercado de apuestas deportivas, las cuotas representan las probabilidades implícitas de los posibles resultados de un evento. Estas cuotas indican el retorno económico que recibiría el usuario si su apuesta fuese ganadora. Además. las casas de apuestas aplican un margen de beneficio, conocido como "*book margin"*, que les permite garantizar una rentabilidad independientemente del resultado del evento. Este margen surge porque la suma de las probabilidades implícitas suele ser superior al 100 %. Según Robbins (2022) [8], este margen promedio suele estar entre el 3 % y el 5 %, y constituye una ventaja estructural para los operadores.

---

[6] Villar, G. (2023, 13 octubre). La cara B de la riqueza que genera el fútbol: el 43% del gasto de los aficionados va a las apuestas online. Relevo. https://www.relevo.com

[7] La evolución histórica de las apuestas deportivas: de los juegos antiguos a la era digital. (2023, 21 junio). Fox Sports. https://www.foxsports.com.

Si bien las cuotas se calculan inicialmente a partir de modelos estadísticos, también se ajustan dinámicamente en función del comportamiento de los usuarios y del volumen de apuestas en cada resultado. Además, las casas consideran sesgos del mercado y factores de liquidez para equilibrar su exposición al riesgo y reducir posibles pérdidas. En consecuencia, las cuotas no reflejan únicamente estimaciones objetivas de probabilidad, sino también decisiones estratégicas de carácter comercial. El resultado es un sistema donde las cuotas actúan como precios de mercado, pero con una distorsión incorporada para asegurar la rentabilidad del operador.

Tipos de casas de apuestas

Principalmente, existen dos modelos de casas de apuestas en el mercado actual. El más común es el de las casas de apuestas tradicionales o de **contrapartida**, donde la propia casa establece las cuotas y acepta apuestas directamente de los usuarios. En este sistema, el apostante juega contra la casa, que asume el riesgo y obtiene beneficios mediante el margen aplicado a las cuotas. Algunos ejemplos representativos de este modelo son Bet365, William Hill o Bwin.

Por otro lado, las casas de apuestas de **intercambio** permiten que los propios usuarios apuesten entre sí. En este caso, los jugadores pueden proponer sus propias cuotas y aceptar las apuestas de otros, mediante que la plataforma actúa únicamente como intermediaria. Su beneficio proviene de aplicar una comisión sobre las ganancias de los apostantes. Betfair es el ejemplo más conocido de este sistema.

Ajustes de Cuotas y Margen de Beneficio: Sesgos e Ineficiencias

Uno de los principales debates en el análisis del mercado de apuestas deportivas gira en torno a si las cuotas ofrecidas por las casas de apuestas reflejan con precisión las probabilidades reales de los resultados. El concepto de eficiencia del mercado, aplicado al ámbito de las apuestas deportivas, se refiere a la idea de que las cuotas ofrecidas por las casas de apuestas reflejan toda la información disponible sobre un evento, de forma que no es posible obtener beneficios sistemáticos apostando. En un mercado eficiente, las cuotas serían una representación ajustada del conocimiento disponible, y no existiría la posibilidad de obtener beneficios sistemáticos. Sin embargo, existen diversos estudios que han cuestionado esta suposición.

El trabajo de Vlastakis, Dotsis y Markellos (2009) [7] representa uno de los estudios más completos sobre la eficiencia del mercado europeo de apuestas de fútbol. Analizando más de 12.000 partidos y cuotas de seis grandes casas de apuestas (tanto online como tradicionales),

los autores identifican la existencia de oportunidades de arbitraje limitadas pero altamente rentables, especialmente cuando se combinan cuotas de diferentes operadores. Entre los sesgos detectados, destacan el conocido "favo*urite–longshot bias*", donde las apuestas a favoritos tienden a estar mejor valoradas que las apuestas a no favoritos, y la sobreestimación sistemática de la ventaja de jugar en casa. Esta última genera lo que los autores denominan el *"away-favourite bias*", es decir, una infravaloración de los equipos favoritos que juegan como visitantes. Estos desajustes en las cuotas no solo reflejan errores en la estimación de probabilidades, sino también posibles influencias del comportamiento de los apostadores, que pueden sesgar las cuotas por motivos emocionales o preferencias locales.

Impacto Económico y Social

Actualmente, se estima que el mercado global de apuestas deportivas genera más de 90 millones de dólares al año, una cifra que continúa creciendo debido a la legalización en diversos países y la expansión de las plataformas digitales. En regiones como Europa y América del Norte, este sector representa un gran porcentaje la industria del entretenimiento y juegos de azar. De hecho, se estima que para el año 2030, el mercado de las apuestas deportivas alcance un valor de 608.410 millones de dólares.[8]

Sin embargo, la accesibilidad de las apuestas en línea también ha traído consecuencias negativas, como el aumento de casos de adicción al juego, especialmente entre los jóvenes. Varios estudios calculan que aproximadamente 80 millones de adultos en todo el mundo padecen problemas relacionados con el juego.[9] En países como Brasil, se ha encendido la alarma por el alto nivel de gasto en apuestas en línea, que supera los 3.200 millones de euros al mes, lo que equivale al 20 % del total de salarios del país.[10] Por este motivo, aunque los modelos predictivos pueden ser útiles para potenciar el desarrollo del sector, su aplicación debe ir acompañada de una visión responsable que evite reforzar conductas adictivas.

---

[8] Fernández, R. (2025, 13 febrero). Las apuestas y los juegos de azar en el mundo: Datos estadísticos. https://es.statista.com/

[9] Mouzo, J. (2024, 24 octubre). La amenaza de llevar un casino en el bolsillo: 80 millones de adultos sufren adicción al juego. El País. https://elpais.com

[10] Zuppello, M. (2024, 30 septiembre). Adicción a las apuestas online en Brasil. Infobae. https://www.infobae.com

2.2 Marco Teórico

En este apartado se presentan todos los conceptos teóricos necesarios para entender el desarrollo de este trabajo, así como las herramientas estadísticas que sustentan el modelo propuesto.

2.2.1 Modelos Probabilísticos en el Deporte

El uso de distribuciones de probabilidad para modelar resultados deportivos tiene una larga trayectoria. En fútbol, Moroney (1956) [10] fue uno de los primeros en señalar que el número de goles podía modelarse mediante una distribución de Poisson, debido a su naturaleza discreta y a la baja frecuencia de goles.

Desde entonces, la estadística se ha aplicado a diferentes aspectos del deporte, no solo para analizar goles, sino también tarjetas, saques de esquina, lesiones, posesiones o sustituciones. El enfoque probabilístico permite construir modelos más realistas y útiles para la predicción y la toma de decisiones. A diferencia de una predicción determinista, que intenta anticipar un único resultado concreto, una predicción probabilística estima la distribución de probabilidad de todos los posibles desenlaces. Esta diferencia resulta especialmente importante en deportes como el fútbol, donde los resultados son altamente inciertos y están condicionados por múltiples factores aleatorios como decisiones arbitrales o situaciones imprevistas del juego.

Este enfoque no solo facilita la comprensión de la variabilidad del fútbol, sino que también permite generar estimaciones objetivas sobre las probabilidades de victoria, empate o derrota. Estas estimaciones son fundamentales en ámbitos como las apuestas deportivas, donde la precisión de los modelos puede traducirse en una ventaja económica.

A partir del trabajo inicial de Moroney, distintos autores han propuesto mejoras al modelo de Poisson clásico. Maher (1982) [2] introdujo un modelo doble de Poisson que considera las características ofensivas y defensivas de los equipos. En 1997, Dixon y Coles [3] refinaron este modelo para mejorar la predicción de empates. Más adelante, Karlis y Ntzoufras (2003) [12] propusieron un modelo bivariante que tiene en cuenta la dependencia entre los goles de ambos equipos. Recientemente, Loukas et al. (2024) [11] han incorporado variables explicativas mediante regresión de Poisson, lo que permite capturar factores adicionales que afectan al rendimiento de los equipos. Esta evolución refleja cómo el modelado estadístico en el fútbol ha ido ganando complejidad y precisión con el tiempo.

En conjunto, estos trabajos representan distintos intentos de superar las limitaciones del modelo clásico y serán tratados con mayor detalle en el apartado 2.2 Trabajos Relacionados, donde se analizará su impacto metodológico y sus aplicaciones en el modelado estadístico de resultados de fútbol.

2.2.2 Modelos de Conteo

Los modelos de conteo se utilizan para describir sucesos en los que se necesita analizar cuántas veces ocurre un determinado evento dentro de un intervalo fijo de tiempo. En estos casos, la variable de interés es discreta y no negativa, como por ejemplo el número de accidentes en una carretera por semana, el número de llamadas a un centro de atención por hora o el número de fallos técnicos en una máquina durante un mes.

Algunas de las distribuciones más utilizadas para modelar este tipo de situaciones son la distribución de Poisson y la distribución binomial negativa. La distribución de Poisson es adecuada cuando los eventos ocurren de forma independiente, con una tasa constante y baja frecuencia. Es un modelo simple, con un único parámetro que representa la tasa media de ocurrencia. Sin embargo, su uso está limitado a contextos en los que la varianza del número de eventos es igual a la media.

Cuando los datos presentan una dispersión mayor que la esperada bajo un modelo de Poisson, es decir, cuando la varianza excede a la media (*overdispersion*), resulta más apropiado utilizar la distribución binomial negativa. Este modelo introduce un segundo parámetro que permite ajustar la varianza por separado, ofreciendo una mayor flexibilidad ante datos que presentan mayor variabilidad. Esta propiedad resulta especialmente útil en situaciones donde la ocurrencia de un evento tiende a aumentar la probabilidad de que se repita, lo que se conoce como efecto contagio o dependencia positiva entre eventos.

Los modelos de conteo, por tanto, permiten representar de forma realista procesos en los que los eventos son discretos y no continuos, y su elección debe considerar el comportamiento estadístico de los datos. Si los eventos son raros, independientes y con frecuencia constante, un modelo de Poisson puede ser suficiente, pero si hay mayor variabilidad o dependencia entre eventos, se debe optar por alternativas más flexibles como la binomial negativa.

2.2.3 Distribución de Poisson: Definición y Propiedades

La distribución de Poisson es una distribución de probabilidad que pueden seguir ciertas variables aleatorias discretas. Es uno de los modelos más utilizados para describir el número de veces que ocurre un evento en un intervalo fijo de tiempo o espacio, siempre que dichos eventos sean independientes entre sí y ocurran con una tasa media constante.

Formalmente, una variable aleatoria X sigue una distribución de Poisson con parámetro $\lambda > 0$ si su función de masa de probabilidad viene dada por:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{-x}}{x!}, \quad x = 0,1,2,\ldots$$

El parámetro λ representa la tasa media de ocurrencia de eventos en el intervalo considerado, es decir, λ es tanto la media como la varianza de la distribución. Las principales propiedades de la distribución de Poisson son:

1. Número de eventos discreto: solo toma valores enteros no negativos.
2. Rareza: adecuada para modelar eventos poco frecuentes.
3. Independencia: los eventos deben ser independientes unos de otros.
4. Homogeneidad temporal o espacial: la tasa λ debe mantenerse constante en el intervalo analizado.
5. Propiedad aditiva: la suma de variables independientes con distribución de Poisson también sigue una distribución de Poisson, con parámetro igual a la suma de los parámetros individuales.

En la Figura 1 se muestran representaciones típicas de la distribución de Poisson para diferentes valores del parámetro λ.

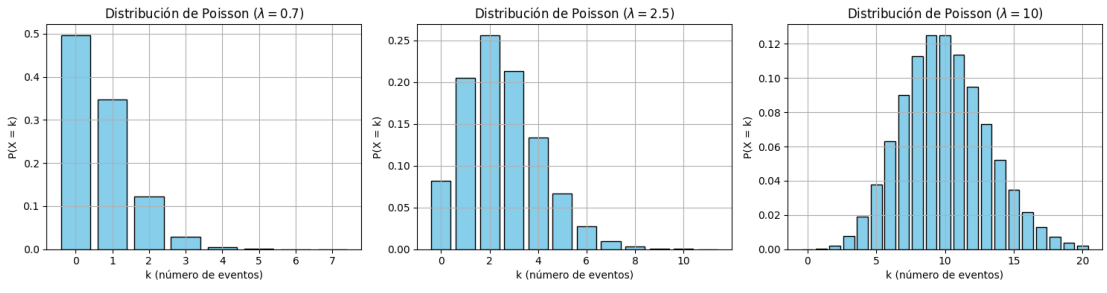

Figura 1. Distribuciones de Poisson para distintos valores de λ

El eje horizontal (x) representa los posibles valores que puede tomar la variable aleatoria X. El eje vertical (y) muestra la probabilidad de que ocurra cada uno de esos valores, según la función de masa de probabilidad. Cuando λ es suficientemente grande (habitualmente a partir de $\lambda \geq$

9), la distribución de Poisson puede aproximarse mediante una distribución normal con media $\mu = \lambda$ y varianza $\sigma^2 = \lambda$.

La distribución de Poisson se aplica habitualmente en contextos como el número de llamadas en una centralita, accidentes de tráfico en una carretera, errores tipográficos en una página, o cualquier fenómeno donde los eventos son discretos, esporádicos y distribuidos aleatoriamente en el tiempo o el espacio.

Por sus propiedades, la distribución de Poisson resulta especialmente adecuada para modelar fenómenos como los goles en fútbol, ya que estos suelen producirse en baja cantidad, de forma discreta y con una frecuencia media relativamente constante a lo largo de los partidos.

Cuando se pretende predecir el resultado de un partido, es común utilizar un **modelo de Poisson doble**, que modela de forma independiente el número de goles anotados por cada equipo mediante dos distribuciones de Poisson separadas. Este enfoque será desarrollado en detalle en el siguiente apartado.

2.2.4 Estimación del Número Medio de Goles

La estimación del número medio de goles es fundamental en la construcción de modelos probabilísticos para el fútbol. En el contexto de los modelos de Poisson, esta media representa la intensidad del proceso que genera los goles, y es clave para calcular cualquier probabilidad derivada del modelo. En función de la complejidad del modelo utilizado, se emplean diferentes técnicas de estimación. Entre ellas destacan el método de máxima verosimilitud, adecuado para modelos sencillos y moderadamente parametrizados, y el algoritmo Expectation-Maximization (EM), empleado en contextos más complejos con variables latentes o correlaciones.

Además, se introduce la **idea de credibilidad**, un concepto procedente de la estadística actuarial, que permite ponderar la información específica de un equipo con respecto a la media global de la competición. De esta forma, se evita que los parámetros se vean excesivamente influenciados por resultados atípicos o escasa información, mejorando así la estabilidad del modelo predictivo.

Método de Máxima Verosimilitud

El método de máxima verosimilitud, conocido como EMV por sus siglas en español (en inglés Maximum Likelihood Estimation, MLE), es una de las herramientas fundamentales en estadística para la estimación de parámetros. Su objetivo es encontrar aquellos valores que

maximizan la probabilidad de observar el conjunto de datos disponibles, bajo el supuesto de que estos datos se han generado según un modelo probabilístico concreto.

En el contexto del fútbol, se aplica principalmente para estimar el número medio de goles esperados por cada equipo en un partido, suponiendo que el número de goles sigue una distribución de Poisson. Este enfoque asume que los goles son eventos raros, discretos e independientes, que ocurren con una determinada tasa constante durante el transcurso del partido.

En el caso del modelo de Poisson, donde el número de goles sigue una distribución $Poisson(\lambda)$, este estimador busca el valor de λ que mejor explica los goles observados históricamente. Es decir, se trata de encontrar la media de la distribución que maximiza la probabilidad de haber observado los resultados reales de los partidos.

Formalmente, si se tiene una muestra de partidos con datos observados $x_1, x_2, \dots, x_n$, y se desea estimar el parámetro θ (por ejemplo, λ para la media de goles esperada), el estimador de máxima verosimilitud $\hat{\theta}$ se obtiene como:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i; \theta)$$

Donde $(f(x_i; \theta))$ es la función de masa de probabilidad correspondiente (en este caso, de Poisson).

En modelos como el Poisson doble, este método permite ajustar simultáneamente las fuerzas ofensivas y defensivas de cada equipo, junto con el efecto del factor campo. El EMV permite encontrar los valores de estos parámetros que maximizan la verosimilitud conjunta del conjunto de resultados históricos, proporcionando así una base sólida para realizar predicciones y análisis en competiciones deportivas.

Algoritmo Expectation-Maximization

El algoritmo Expectation-Maximization (EM) es una técnica iterativa ampliamente utilizada para obtener estimaciones de máxima verosimilitud en situaciones en las que los datos disponibles son incompletos o contienen variables latentes (variables no observables directamente). Fue propuesto por Dempster, Laird y Rubin (1977) [16], y desde entonces ha demostrado ser una herramienta especialmente eficaz en numerosos contextos estadísticos.

En el caso del modelado estadístico de resultados de fútbol, el algoritmo EM resulta particularmente útil cuando el modelo incorpora dependencias no observables directamente.

Esto sucede, por ejemplo, en el modelo de Poisson bivariante, donde se introduce una correlación explícita entre los goles del equipo local y visitante. En estos escenarios, la función de verosimilitud no puede optimizarse de forma directa, lo que justifica la necesidad de utilizar un procedimiento como el EM.

El algoritmo EM alterna entre dos pasos iterativos:

- **Paso E (Expectation):** Se calcula la esperanza del logaritmo de la verosimilitud completa, condicionada a los datos observados y al valor actual de los parámetros. Este paso estima los valores esperados de las variables latentes.
- **Paso M (Maximization):** Se maximizan esas esperanzas con respecto a los parámetros del modelo, actualizando así sus estimaciones.

Formalmente, sea $\theta^{(t)}$ el valor actual de los parámetros, *Y,* el conjunto de datos observados y *Z* el conjunto de variables latentes. El algoritmo consiste en la iteración de los siguientes pasos:

$$Paso\ E\text{: } Q\left(\theta \mid \theta^{(t)}\right) = E_{Z|Y,\theta^{(t)}}[\log p(Y, Z|\theta)]$$

$$Paso\ M\text{: } \theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta \mid \theta^{(t)}\right)$$

Esta dinámica permite obtener estimaciones estables y consistentes incluso cuando el modelo presenta cierta complejidad estructural.

En modelos aplicados al fútbol, el EM se ha empleado con éxito en trabajos como el de Karlis y Ntzoufras (2003) [12], donde se introduce una dependencia entre los goles del equipo local y visitante, mejorando así la representación de la realidad competitiva. En este caso, se utiliza un modelo de Poisson bivariante, en el que se descompone el número de goles en tres componentes:

$$X = U + W, \quad Y = V + W$$

Donde $U \sim \text{Poisson}(\lambda_1)$, representa los goles locales independientes, $V \sim \text{Poisson}(\lambda_2)$ representa los goles visitantes independientes, y $W \sim \text{Poisson}(\lambda_3)$ representa el componente común que introduce correlación.

Aquí, *X* e *Y* son las variables observadas, mientras que *W* es una variable latente. La verosimilitud conjunta del modelo se expresa como:

$$P(X = x, Y = y) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \sum_{k=0}^{\min(x,y)} \frac{\lambda_1^{x-k}}{(x-k)!} \cdot \frac{\lambda_2^{y-k}}{(y-k)!} \cdot \frac{\lambda_3^{k}}{k!}$$

Dado que no se observan directamente los valores de *W*, se aplica el algoritmo EM para estimar los parámetros $\lambda_1, \lambda_2\ y\ \lambda_3$.

En el paso E, se calcula la esperanza del número de goles compartidos $E[\,W \mid X = x, Y = y\,]$,

utilizando la distribución condicional de *W* dado *x* e *y*. En el paso M, estas esperanzas se utilizan para actualizar las estimaciones de los parámetros del modelo.

Además, el EM es compatible con mejoras como la ponderación temporal exponencial, que da mayor peso a los partidos recientes.

2.2.5 Modelo de Poisson Doble en el Fútbol

La distribución de Poisson es una de las herramientas estadísticas más utilizadas para modelar el número de goles que marca cada equipo en un partido de fútbol. La idea consiste en tratar los goles como eventos discretos, independientes y poco frecuentes que ocurren dentro de un intervalo fijo de tiempo. Estos supuestos, aunque simplifican la realidad del deporte, han demostrado ser razonablemente válidos desde un punto de vista empírico y han sido respaldados por múltiples estudios, como el de Maher (1982) [2], uno de los trabajos pioneros en este campo.

Este enfoque asume que cada equipo genera goles de forma independiente respecto a su oponente, y que el número de goles marcados por el equipo local y el visitante sigue una distribución de Poisson univariante. Este planteamiento se denomina modelo de **Poisson doble**, ya que utiliza dos variables aleatorias independientes: una para los goles del equipo local y otra para los del visitante. Formalmente, se expresa como:

$$X \sim Poisson(\lambda_{ij}), \quad Y \sim Poisson(\mu_{ij})$$

donde X e Y son el número de goles del equipo i (local) y del j (visitante), respectivamente, y λ_{ij} y μ_{ij} son las medias esperadas de goles en ese partido.

Cada una de estas medias se modela como una combinación de tres factores: la capacidad ofensiva del equipo que ataca, la capacidad defensiva del equipo contrario y, en el caso del equipo local, la ventaja de jugar en casa. Las medias esperadas pueden escribirse como:

$$\lambda_{ij} = \alpha_i \cdot \beta_j \cdot \gamma$$

$$\mu_{ij} = \alpha_j \cdot \beta_i$$

donde α_i, α_j, β_i y β_j son respectivamente las fuerzas ofensivas de cada equipo, y γ representa el efecto del factor campo (*home advantage*).

Los parámetros del modelo $\alpha, \beta\ y\ \gamma$ se estiman a partir de datos históricos mediante el **método de máxima verosimilitud**. Este procedimiento busca los valores que hacen más probable la observación de los resultados reales bajo el modelo considerado, ajustando las tasas de goles esperadas de manera que reflejen lo mejor posible el comportamiento observado de los equipos.

Una vez ajustado el modelo, es posible ya calcular la probabilidad de que un equipo marque exactamente k goles en un partido concreto utilizando la función de masa de probabilidad de la distribución de Poisson:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{-x}}{x!}$$

Por ejemplo, si tras el ajuste se estima que el Real Madrid jugando en casa contra el Sevilla tiene una media esperada de 2.3 goles (λ=2.3), la probabilidad de que marque exactamente 0, 1, 2, 3... goles se puede obtener aplicando directamente esta fórmula para cada valor de x.

A pesar de su sencillez, el modelo de Maher ha demostrado ser eficaz para describir la distribución de resultados en competiciones de fútbol. Sin embargo, presenta algunas limitaciones, especialmente al modelar empates con pocos goles como el 0–0 o el 1–1. Estos resultados tienden a quedar subestimados por el modelo, lo que puede afectar negativamente la calidad de las predicciones, especialmente en contextos donde los empates son frecuentes o relevantes, como en mercados de apuestas. Además, el modelo clásico trata todos los partidos con el mismo peso, sin considerar que el rendimiento de los equipos puede cambiar a lo largo del tiempo.

Para superar estas limitaciones, en 1997 Dixon y Coles [3] introdujeron una serie de ajustes. En primer lugar, propusieron un ajuste específico para resultados bajos, ajustando la probabilidad conjunta $P(X = i, Y = j)$ cuando i y j toman valores pequeños. Esta modificación mejora la capacidad del modelo para reflejar la frecuencia real de empates cerrados.

Además, incorporaron una ponderación temporal exponencial que asigna mayor peso a los partidos más recientes. Esta técnica permite que el modelo se adapte mejor al rendimiento actual de los equipos. También propusieron métodos para reducir el sobreajuste cuando se trata de modelar conjuntos de datos pequeños o equipos con poca información, mediante la agrupación o restricción de parámetros.

Gracias a estas modificaciones, el modelo de Dixon y Coles no solo mejora la capacidad predictiva, sino que también permite estimar con mayor realismo las probabilidades de victoria,

empate o derrota. Su efectividad fue comprobada en análisis retrospectivos y aplicaciones en mercados de apuestas, mostrando un mejor rendimiento que el modelo de Poisson clásico.

No obstante, este modelo mejorado sigue sin resolver uno de los supuestos estructurales del modelo de Maher, la independencia entre los goles anotados por ambos equipos en un partido. Esta limitación fue tratada posteriormente en 2003 mediante el **modelo de Poisson bivariante**, que introduce una correlación explícita entre las dos variables y que será analizado en detalle en el siguiente apartado.

Cálculo de Probabilidades de Victoria, Empate y Derrota

Una vez estimadas las medias esperadas de goles para el equipo local (λ_{ij}) y el visitante (μ_{ij}) mediante el modelo de Poisson doble, es posible calcular la distribución conjunta de marcadores asumida la independencia entre los goles de ambos equipos. En este caso, la probabilidad de que el partido termine con *i* goles locales y *j* goles visitantes se obtiene como el producto de dos distribuciones de Poisson independientes:

$$P(X = i, Y = j) = \frac{e^{-\lambda}\lambda^{i}}{i!} \cdot \frac{e^{-\mu}\mu^{j}}{j!}$$

Con esta información, es posible construir una **matriz de probabilidades** conjuntas $P(i,j)$ donde las filas representan los goles del equipo local y las columnas los del visitante. Esta matriz permite visualizar la distribución completa de marcadores posibles en un partido concreto. Para calcular la probabilidad de cada desenlace (victoria local, empate, o victoria visitante), se suman los elementos correspondientes de la matriz:

- Victoria local: todos los casos en los que $i > j$
- Empate: todos los casos en los que $i = j$
- Victoria visitante: todos los casos en los que $i < j$

Formalmente, se puede expresar como:

$$P(\textit{victoria local}) = \sum_{i=0}^{k} \sum_{j=0}^{i-1} P(X = i, Y = j)$$

$$P(\textit{empate}) = \sum_{i=0}^{k} P(X = i, Y = i)$$

$$P(\textit{victoria visitante}) = \sum_{j=0}^{k} \sum_{i=0}^{j-1} P(X = i, Y = j)$$

donde k es el número máximo de goles considerado (por ejemplo, 6 o 7), elegido de forma que la masa de probabilidad acumulada sea superior al 99%.

2.2.6 Modelo de Poisson Bivariante

El modelo de Poisson bivariante introduce una mejora respecto al modelo de Poisson doble al incorporar correlación entre los goles del equipo local y del visitante. Esta dependencia permite modelar de forma más realista los empates, que tienden a estar subestimados por los modelos independientes, como el de Maher (1982). Así, el enfoque bivariante corrige una de las limitaciones más importantes del modelo clásico: la suposición de independencia entre los goles marcados por ambos equipos.

Los modelos bivariantes permiten analizar conjuntamente dos variables aleatorias relacionadas. A diferencia de los modelos univariantes, que estudian cada variable por separado, el enfoque bivariante tiene en cuenta posibles dependencias entre ambas, siendo útil cuando los fenómenos estudiados están correlacionados entre sí.

En términos generales, el modelo de Poisson bivariante extiende la distribución de Poisson a dos dimensiones discretas, permitiendo que dos variables aleatorias (X, Y), que representan recuentos, estén correlacionadas. Para ello, se define a partir de tres variables aleatorias independientes:

$$X_1 \sim \text{Poisson}(\lambda_1) \quad X_2 \sim \text{Poisson}(\lambda_2) \quad X_3 \sim \text{Poisson}(\lambda_3)$$

De forma que:

$$X = X_1 + X_3 \qquad Y = X_2 + X_3$$

La variable X_3 actúa como un componente común que introduce dependencia entre X e Y, de modo que cuando $\lambda_3 > 0$ se obtiene una correlación positiva entre ambas. En cambio, si $\lambda_3 = 0$, se recupera el caso independiente, equivalente al modelo de Poisson doble. Una ventaja fundamental de esta construcción es que las distribuciones marginales de X e Y siguen siendo de Poisson, con medias $\lambda_1 + \lambda_3$ y $\lambda_2 + \lambda_3$, respectivamente, mientras que la covarianza entre ambas variables es precisamente λ_3.

La función de masa de probabilidad conjunta del modelo presenta una forma más compleja que en el caso independiente, ya que incluye una suma hasta $min(x, y)$ que refleja la influencia del término común.

$$P(X = x, Y = y) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \cdot \frac{\lambda_1^x \lambda_2^y}{x!\ y!} \cdot \sum_{k=0}^{\min(x,y)} \binom{x}{k}\binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k$$

En el contexto del fútbol, este tipo de modelos permite representar de forma más realista el resultado de un partido, ya que considera de manera conjunta los goles marcados por el equipo local y por el visitante. Esto corrige una de las principales limitaciones del modelo Poisson clásico, que asume independencia entre los goles de ambos equipos. La distribución Poisson bivariante introduce esta dependencia a través del parámetro λ_3, que refleja el efecto de un componente común que afecta simultáneamente a ambos equipos.

Para ajustar el modelo a los datos reales, se estiman los parámetros $\lambda_1, \lambda_2, y\ \lambda_3$ mediante el algoritmo Expectation-Maximization (EM). Este permite tratar la complejidad de la función de verosimilitud al incorporar un componente común no observable, actualizando iterativamente los parámetros hasta la convergencia.

Este modelo fue aplicado al fútbol por Karlis y Ntzoufras (2003) [12], quienes demostraron que la introducción de una correlación, aunque moderada, mejora notablemente la capacidad del modelo para estimar la frecuencia de empates, especialmente en ligas donde los resultados cerrados como 0–0 o 1–1 son comunes. Además, encontraron que la probabilidad de empate es sistemáticamente subestimada por los modelos independientes, y que el modelo bivariante corrige esta desviación con un ajuste relativamente sencillo.

Cálculo de Probabilidades de Victoria, Empate y Derrota

Una vez estimados los parámetros $\lambda_1, \lambda_2, y\ \lambda_3$ la probabilidad conjunta de que un partido termine con *i* goles del equipo local y *j* del visitante se calcula utilizando la función de masa de probabilidad del modelo bivariante de Poisson:

$$P(X = i, Y = j) = \sum_{k=0}^{\min(i,j)} \frac{e^{-(\lambda_1+\lambda_2+\lambda_3)} \lambda_1^{i-k} \lambda_2^{j-k} \lambda_3^k}{(i-k)!\,(j-k)!\,k!}$$

Esta es la formulación introducida por Kocherlakota en 1992 [17] y utilizada en modelos como el de Karlis y Ntzoufras (2003) [12]. El sumatorio recorre todos los posibles valores del componente común $X_3 = k$, introduciendo una dependencia explícita entre los goles de ambos equipos. El parámetro λ_3 actúa como fuente de correlación. Si $\lambda_3 = 0$, se recupera el caso de independencia.

2.2.7 Modelos Lineales Generalizados

Los Modelos Lineales Generalizados (GLM, por sus siglas en inglés *Generalized Linear Models*) [15] son una extensión del modelo lineal clásico que permite trabajar con variables dependientes que no necesariamente siguen una distribución normal. Esta familia de modelos resulta útil para analizar datos de conteo, proporciones o tiempos de espera. Formalmente, un GLM está compuesto por tres elementos fundamentales:

1. Distribución de la Familia Exponencial

La variable dependiente Y se asume que sigue una distribución perteneciente a la familia exponencial, como la normal, binomial, Poisson, gamma, entre otras. En general, esta familia tiene la siguiente forma canónica:

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

Donde θ es el parámetro canónico, ϕ es un parámetro de dispersión, y $a(\cdot), b(\cdot)\ y\ c(\cdot)$ son las funciones específicas de cada distribución.

2. Función de enlace

Se utiliza una función de enlace $g(\cdot)$ que conecta la media de la variable dependiente $\mu = \mathbb{E}[Y]$ con una combinación lineal de los predictores:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

La elección de la función de enlace depende del tipo de distribución: por ejemplo, la función logarítmica se usa para la Poisson, la función logit para la binomial, y la identidad para la normal.

3. Estructura Lineal en los Predictores

El predictor lineal η_i define como una combinación lineal de las variables independientes del modelo, permitiendo interpretar los efectos individuales de cada variable sobre la respuesta.

La formulación de los modelos lineales generalizados ofrece una gran flexibilidad para modelar relaciones no lineales entre variables, manteniendo a su vez una interpretación estadística clara y una base matemática sólida.

Selección de Variables Explicativas

En la construcción de modelos GLM es crucial seleccionar adecuadamente las variables explicativas. Se pueden emplear métodos automáticos como selección hacia adelante, hacia atrás o por pasos, basados generalmente en criterios como AIC o BIC. Además, es importante evaluar la significancia estadística de los coeficientes y la relevancia teórica de las variables para evitar el sobreajuste o la inclusión de predictores irrelevantes.

Multicolinealidad

Aunque los GLM permiten gran flexibilidad, siguen siendo sensibles al problema de multicolinealidad. Cuando las variables explicativas están altamente correlacionadas entre sí, la estimación de los coeficientes puede volverse inestable y difícil de interpretar. Esto puede detectarse mediante indicadores como el factor de inflación de la varianza (VIF), que mide cuánto se incrementa la varianza de un coeficiente debido a la colinealidad con otras variables. Si se detecta multicolinealidad, puede abordarse mediante regularización o eliminación de variables redundantes.

Evaluación del Modelo

La evaluación de los GLM se puede realizar mediante métricas específicas según el tipo de variable dependiente. En el caso de variables de conteo (como los goles), se utiliza la log-verosimilitud y sus derivados como el AIC (Akaike Information Criterion) o BIC (Bayesian Information Criterion). Además, en modelos donde se realiza clasificación (por ejemplo, predicción del resultado del partido), se pueden usar métricas como accuracy, precisión, recall o F1-score, evaluando el ajuste del modelo tanto a nivel probabilístico como en términos de predicciones categóricas. Todas estas métricas serán explicadas en el apartado **2.2.10 Métricas de Evaluación**.

2.2.8 Regresión de Poisson

La regresión de Poisson es un caso particular de modelo lineal generalizado en el que la variable dependiente representa un conteo de eventos y se modela mediante una distribución de Poisson. En el contexto del fútbol, se utiliza para predecir el número de goles que marcará cada equipo en un partido, teniendo en cuenta características adicionales del encuentro.

En este modelo, la media esperada de goles $\lambda_i = \mathbb{E}[Y_i]$ se relaciona con las variables explicativas mediante una función de enlace logarítmica:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

donde $Y_i \sim Poisson(\lambda_i)$ es el número de goles esperados por un equipo en el partido, λ_i representa la media esperada de goles, x_{ij} son las variables explicativas, y β_j son los coeficientes del modelo.

Para ajustar estos coeficientes a partir de los datos observados, se utiliza el método de máxima verosimilitud (MLE). Este procedimiento permite estimar los parámetros del modelo buscando aquellos valores que maximizan la probabilidad de observar los resultados reales del conjunto de partidos, dada la estructura de Poisson asumida para las variables dependientes.

Este tipo de regresión resulta especialmente útil porque permite modelar situaciones más realistas y dinámicas, en las que la tasa de goles esperados no es constante, sino que depende del contexto del partido. A diferencia de trabajos más clásicos como el de Maher (1982) [2], donde los goles esperados se estimaban solo con parámetros fijos, estudios más recientes como el de Loukas et al. (2024) [11] amplían esta idea mediante un modelo doble de regresión de Poisson, que estima los goles esperados del equipo local y del visitante por separado.

En dicho enfoque, el parámetro λ se vincula con múltiples variables a través de una función logarítmica, lo que permite incorporar información contextual como el rendimiento reciente, la ventaja de jugar en casa o la calidad del rival. Una de las principales ventajas de este enfoque es que puede adaptarse fácilmente a nuevos datos y actualizarse a lo largo de la temporada para reflejar mejor la evolución de los equipos. Además, permite incluir variables adicionales como las cuotas de apuestas, la posición en la liga o métricas avanzadas como los goles esperados (xG), mejorando la capacidad predictiva del modelo. Sin embargo, este enfoque también asume que los goles de un equipo son independientes de los del rival, lo cual puede no reflejar completamente la realidad de un partido.

Cálculo de Probabilidades de Victoria, Empate y Derrota

En este contexto, el cálculo de las probabilidades de victoria, empate y derrota sigue el mismo esquema que en el modelo de Poisson doble, ya que se asume también la independencia entre los goles de ambos equipos. La diferencia clave radica en que los valores de λ y μ no son parámetros fijos, sino que se obtienen como funciones de variables explicativas a través del modelo de regresión. Así, las medias esperadas de goles se ajustan dinámicamente en función de factores como la localía, el rendimiento reciente o la calidad del rival.

Una vez calculadas estas medias para un partido concreto, se puede aplicar directamente la distribución conjunta de Poisson para estimar la matriz de resultados y derivar las probabilidades asociadas a cada desenlace. Esto permite que el modelo sea más flexible y capaz de adaptarse a diferentes contextos competitivos o temporales.

2.2.9 Supuestos y Limitaciones

El modelo clásico de Poisson aplicado al fútbol parte de dos supuestos fundamentales. Por un lado, que los goles de cada equipo son generados de forma independiente, y por otro, que la tasa de anotación permanece constante durante los 90 minutos de juego. Ambos supuestos permiten simplificar el desarrollo del modelo, pero también introducen limitaciones.

El supuesto de independencia implica que el número de goles marcados por un equipo no depende del comportamiento del rival. Sin embargo, estudios posteriores han cuestionado esta idea al encontrar evidencias de correlación entre los goles de ambos equipos, especialmente en partidos con marcadores bajos o desiguales. Esta posible interdependencia puede generar sesgos en la estimación de ciertas probabilidades, como los empates, y ha sido abordada en trabajos como los de Dixon y Coles (1997) [3] mediante correcciones específicas.

Otro aspecto discutido es la suposición de que la varianza es igual a la media, que es una propiedad inherente a la distribución de Poisson. En datos reales, sin embargo, es común observar una dispersión mayor, fenómeno conocido como *overdispersion*. Tal como señala Pollard (1985) [13], este exceso de variabilidad puede deberse tanto a factores tácticos o contextuales no modelados como a errores de especificación.

Por último, el modelo clásico asume que los parámetros de ataque y defensa de los equipos son constantes a lo largo del tiempo, lo que no siempre refleja la realidad dinámica del fútbol. El rendimiento de los equipos puede cambiar a lo largo de la temporada, afectando a su capacidad ofensiva y defensiva. Para abordar este problema, algunas extensiones introducen ponderaciones temporales o incluso estructuras dinámicas en los parámetros.

2.2.10 Métricas de Evaluación

Para evaluar la calidad de los modelos predictivos utilizados, se han empleado diferentes métricas que capturan tanto el rendimiento estadístico como el impacto práctico en un entorno simulado de apuestas deportivas. A continuación, se describen las métricas principales utilizadas.

Evaluación de modelos de clasificación

La calidad de un modelo de clasificación se mide mediante métricas derivadas de la matriz de confusión, que compara las predicciones del modelo con los resultados reales. La matriz de confusión se compone de cuatro elementos clave:

- Verdaderos Positivos (TP): Casos correctamente clasificados como positivos.
- Falsos Positivos (FP): Casos incorrectamente clasificados como positivos.
- Verdaderos Negativos (TN): Casos correctamente clasificados como negativos.
- Falsos Negativos (FN): Casos incorrectamente clasificados como negativos.

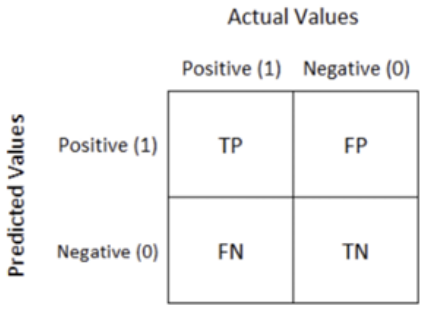

Figura 2. Matriz de confusión de dos clases

Accuracy

La accuracy (o exactitud) mide la proporción de predicciones correctas sobre el total de casos evaluados.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precisión

La precisión indica qué proporción de las predicciones positivas realizadas por el modelo fueron realmente correctas.

$$Precisión = \frac{TP}{TP + FP}$$

Recall

El recall (también conocido como sensibilidad o tasa de verdaderos positivos) representa la capacidad del modelo para detectar todos los casos positivos reales.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score

El F1-score es una medida de equilibrio entre la precisión y el recall, especialmente útil en situaciones de desbalance de clases.

$$Recall = 2 \times \frac{Precisión \times Recall}{Precisión + Recall}$$

En este trabajo, se calcula el F1 macro como promedio del F1-score para cada clase (victoria local, empate, victoria visitante), proporcionando una visión más justa del rendimiento general. Esta métrica penaliza los modelos que solo aciertan una clase mayoritaria (por ejemplo, solo victorias locales), aunque tengan buena accuracy. Diversos estudios han señalado que, en escenarios con clases desbalanceadas, métricas como el F1-score o las curvas de precisión-recall proporcionan una evaluación más representativa que otras métricas. En particular, Saito y Rehmsmeier (2015) [18] argumentan que la curva de precisión-recall resulta más informativa en estos casos, ya que refleja mejor la proporción de aciertos entre las predicciones de distintas clases.

Log-verosimilitud

La log-verosimilitud evalúa cuán bien las probabilidades asignadas por el modelo se ajustan a los resultados reales observados. Se calcula sumando el logaritmo de la probabilidad que el modelo asignó al resultado correcto en cada partido.

$$\log L = \sum_{i=1}^{n} \log P\,(\,y_i \mid \text{modelo}\,)$$

También se calcula su versión media:

$$\log L_{\text{media}} = \frac{1}{n}\sum_{i=1}^{n} \log P\,(\,y_i \mid \text{modelo}\,)$$

Un valor de log-verosimilitud más cercano a cero indica un mejor ajuste probabilístico. Esta métrica es fundamental cuando los modelos no solo hacen predicciones “planas”, sino que asignan distribuciones de probabilidad.

AIC y BIC

Para comparar modelos de distinta complejidad, se utilizan criterios de información como el AIC (Akaike Information Criterion) y el BIC (Bayesian Information Criterion), que penalizan la complejidad del modelo para evitar el sobreajuste.

$$\mathbf{AIC} = 2k - 2\log L$$

$$\mathbf{BIC} = k \cdot \log(n) - 2\log L$$

Donde k es el número de parámetros del modelo y n es el número de observaciones. Ambos criterios combinan el ajuste del modelo (medido por la log-verosimilitud) con una penalización por complejidad. Cuanto mayor sea $\log L$, mejor el ajuste, pero si el modelo es demasiado complejo (mayor k), AIC y BIC lo penalizarán.

Estrategia de Apuestas Simulada

Con el fin de evaluar la utilidad práctica de los modelos en el contexto de las apuestas deportivas, se aplicó una estrategia de simulación basada en sus predicciones. Para cada partido, se realizó una apuesta al resultado que el modelo consideraba más probable. Cuando la probabilidad predicha era alta (igual o superior al 60 %), se apostaban 2 €; en caso contrario, solo 1 €.

La simulación permite calcular indicadores clave como el total apostado, el total ganado, el beneficio neto y la rentabilidad obtenida. Esta última refleja si, en términos económicos, el modelo habría resultado rentable aplicando dicha estrategia, proporcionando así una medida complementaria a las métricas estadísticas tradicionales.

2.3 Trabajos Relacionados

El uso de modelos estadísticos basados en la distribución de Poisson para analizar resultados de fútbol ha sido ampliamente estudiado desde la década de los ochenta. A continuación, se presenta una revisión cronológica de los trabajos más relevantes en esta línea, destacando sus aportaciones metodológicas y principales resultados.

Título: *Modelling Association Football Scores*

Autor: M. J. Maher (1982)

Maher (1982) [2] introdujo el modelo Poisson doble independiente, donde los goles de cada equipo se modelan como variables Poisson independientes, con medias determinadas por parámetros de ataque, defensa y ventaja local. Aplicado a la liga inglesa (1971–1974), el modelo ofreció un buen ajuste: en 19 de 24 pruebas χ^2, los resultados no fueron significativos al 5 %, indicando concordancia entre observados y esperados. No obstante, tendía a subestimar empates bajos (0–0 y 1–1).

Titulo: *Modelling Association Football Scores and Inefficiencies in the Football Betting Market*

Autores: M. J. Dixon y S. G. Coles (1997)

En el trabajo de Dixon & Coles (1997) [3], los autores parten del modelo de Maher y lo refinan en dos direcciones clave. Para empezar, añaden un parámetro de corrección para inflar la probabilidad de marcadores bajos, solventando el sesgo detectado por Maher. Además, introducen un factor de ponderación temporal que asigna mayor peso a los partidos recientes, permitiendo que la fuerza de los equipos evolucione. El ajuste por máxima verosimilitud muestra mejoras sustanciales en la predicción de empates y, aplicado al mercado de apuestas de la Primera División inglesa, obtiene un beneficio teórico cercano al 7 %.

Título: Analysis of Sports Data by Using Bivariate Poisson Models

Autores: Dimitris Karlis e Ioannis Ntzoufras (2003)

Karlis y Ntzoufras (2003) [12] propusieron un modelo Poisson bivariante que permite capturar la dependencia entre los goles de ambos equipos mediante un parámetro de covarianza. Utilizaron el algoritmo EM para estimar los parámetros y corrigieron el exceso de empates mediante una inflación en la diagonal de la distribución conjunta. Aplicado a datos de la Serie A italiana (temporada 1991–1992), su modelo mejoró significativamente el ajuste respecto al Poisson doble, especialmente en la predicción de empates: por ejemplo, el número de empates 1–1 fue estimado con precisión exacta (58 observados, 58 estimados), frente a solo 33 predichos por el modelo independiente.

Título: *Analysis of a Double Poisson Model for Predicting Football Results in Euro 2020*

Autores: Matthew J. Penn y Christl A. Donnelly (2022)

Penn y Donnelly (2022) [14] en su artículo *"Analysis of a double Poisson model for Predicting Football Results in Euro 2020"*, analizaron desde un enfoque teórico y aplicado la validez actual del modelo Poisson doble. El estudio no solo confirmó la plena vigencia del modelo, sino que también aportó resultados novedosos, como la derivación de condiciones matemáticas precisas que garantizan la existencia y unicidad de los estimadores de máxima verosimilitud. Además, los autores aplicaron el modelo a datos de selecciones europeas entre 2018 y 2021, excluyendo equipos con muy bajo rendimiento, como San Marino, para evitar distorsiones en los parámetros. El modelo obtuvo el mejor rendimiento en el concurso de predicción organizado por la Royal Statistical Society para la Euro 2020, alcanzando una log-verosimilitud de −39.33, la más alta entre los participantes. La predicción del número de goles marcados y encajados por cada selección fue especialmente precisa: 47 de las 48 predicciones se situaron dentro del intervalo de confianza del 95 %. Estos resultados refuerzan la utilidad práctica del modelo Poisson doble como herramienta predictiva fiable, incluso frente a enfoques más complejos. No obstante, el artículo también señala limitaciones importantes, como la sobrevaloración de los resultados frente a rivales muy débiles, lo cual puede sesgar las estimaciones si no se filtran adecuadamente los datos.

Título: *Predicting Football Match Results Using a Poisson Regression Model*

Autores: Konstantinos Loukas, Dimitrios Karapiperis, Georgios Feretzakis y Vassilios S. Verykios (2024)

Finalmente, Loukas et al. (2024) [11] aplicaron una regresión Poisson clásica a la Premier League 2022–2023, incluyendo parámetros de ataque, defensa y ventaja local. Evaluado en diferentes escenarios, el modelo alcanzó un 75 % de precisión al predecir los goles dentro de ±1 respecto al marcador real, aunque persistía cierta infraestimación de los empates a 0.

La literatura sobre modelado de resultados de fútbol ha evolucionado desde enfoques independientes hasta modelos que integran dependencia, dinámica temporal y correcciones específicas. Estas mejoras han permitido aumentar la precisión predictiva y detectar ineficiencias en el mercado de apuestas, consolidando la utilidad práctica de estos modelos basados en la distribución de Poisson.

3. ASPECTOS METODOLÓGICOS

En este apartado se explican la metodología seguida para llevar a cabo la investigación, la redacción del trabajo y el desarrollo de la parte práctica. Se describen las técnicas utilizadas, y además se justifica la elección de las herramientas empleadas.

3.1 Metodología

El Trabajo de Fin de Grado se dividió en varias fases, cada una con un objetivo claro. Esta organización facilitó un desarrollo estructurado del proyecto, permitió un seguimiento preciso del progreso y una documentación ordenada.

3.1.1 Planificación

El desarrollo del trabajo se estructuró en cinco fases principales, cada una con funciones específicas orientadas a facilitar una evolución progresiva y coherente del proyecto:

1. Planificación: Se definieron los objetivos generales, el alcance del trabajo y un cronograma orientativo. También se identificaron los recursos necesarios y se organizaron las tareas a realizar para facilitar el desarrollo progresivo del proyecto.

2. Investigación: Se realizó una revisión bibliográfica centrada en modelos estadísticos de predicción de resultados en fútbol basados en la distribución de Poisson. Esta fase permitió seleccionar los métodos más adecuados y justificar su uso en el marco del trabajo.

3. Adquisición, análisis y procesamiento de datos: Se recopilaron datos relevantes de competiciones oficiales a través de Football-Data.co.uk[11], que ofrece información histórica

[11] Football-Data.co.uk. (2024). Spain football data files.
https://www.footballdata.co.uk/spainm.php

detallada sobre ligas de fútbol europeas. Los datos fueron posteriormente limpiados, transformados y explorados para detectar patrones, inconsistencias y variables relevantes.

4. Desarrollo del modelo: Se implementaron tres enfoques estadísticos diferentes basados en la distribución de Poisson: el modelo Poisson doble básico, el modelo Poisson bivariante y un modelo de regresión Poisson.

5. Evaluación y análisis de resultados: Se analizaron los resultados obtenidos a partir de los modelos, comparando su capacidad predictiva. Además, se valoró su aplicabilidad al ámbito de las apuestas deportivas y se discutieron sus principales limitaciones.

Las fases 3 y 4 se han desarrollado con mayor detalle en el apartado 4. Desarrollo del Trabajo, mientras que la fase 5 será abordada en el apartado 5. Conclusiones.

3.2 Tecnologías Empleadas

Durante el desarrollo del trabajo se utilizaron diversas herramientas tecnológicas que facilitaron tanto la organización y seguimiento del proyecto como la implementación práctica de los modelos estadísticos. A continuación, se detallan las principales tecnologías empleadas.

3.2.1 Seguimiento del Trabajo

Para el control de versiones y la organización del proyecto se utilizó **GitHub**, una plataforma ampliamente empleada en entornos académicos y profesionales. El repositorio permitió registrar los cambios realizados en el código, mantener un historial de desarrollo y facilitar el trabajo estructurado y progresivo. Gracias a esta herramienta fue posible documentar adecuadamente cada avance, comparar versiones anteriores y garantizar la trazabilidad de los resultados.

3.2.2 Lenguajes y Entornos de Desarrollo

El lenguaje de programación utilizado fue **Python (**versión 3.10.7**)**, debido a su flexibilidad, sintaxis clara y amplio soporte en el ámbito de la estadística y el análisis de datos. El desarrollo del código se llevó a cabo múltiples **Jupyter Notebooks**, un entorno interactivo que permitió combinar celdas de código, texto explicativo y visualizaciones, facilitando así la comprensión del flujo de trabajo y la reproducibilidad del análisis.

3.2.3 Librerías y Frameworks

Para implementar los modelos estadísticos, manipular los datos y visualizar los resultados se emplearon diversas librerías especializadas del ecosistema de Python. Entre las más destacadas se encuentran:

- **Pandas (2.2.3):** para la lectura, limpieza y manipulación de datos en formato tabular.
- **NumPy (2.2.5):** para operaciones numéricas y trabajo con arrays multidimensionales.
- **Matplotlib (3.10.0)** y **Seaborn (0.13.2**): para la creación de gráficos y análisis visual exploratorio.
- **SciPy (1.15.2):** para la estimación de modelos estadísticos
- **Scikit-learn (1.5.0):** utilizada para el cálculo de métricas de evaluación y validación de modelos.
- **Statsmodels (0.14.4)**: empleada para el ajuste de modelos de regresión de poisson.

En conjunto, el uso de estas tecnologías permitió abordar de forma estructurada y eficiente todas las fases del trabajo, desde la organización y el control del proyecto hasta la implementación, validación y análisis de los modelos estadísticos.

4. DESARROLLO DEL TRABAJO

Este capítulo describe con detalle el proceso práctico llevado a cabo para implementar y evaluar distintos modelos estadísticos basados en la distribución de Poisson, con el objetivo de modelar y predecir resultados de partidos de fútbol. Se presentan tanto la adquisición y preparación de los datos como la construcción de los modelos y el análisis de sus resultados.

El código fuente desarrollado para este proyecto se encuentra publicado en el repositorio de GitHub accesible en el siguiente enlace: https://github.com/anaigs/tfg_maco_github.

4.1 Adquisición, Análisis y Procesamiento de Datos

En este apartado se detallan los pasos seguidos para la adquisición, limpieza y procesamiento de los datos necesarios para la construcción y evaluación de los modelos estadísticos. Los datos originales se obtuvieron a través del portal Football-Data.co.uk, una fuente reconocida por ofrecer información detallada y estructurada sobre resultados y estadísticas de múltiples ligas de fútbol europeas, en formato CSV. El proceso se dividió en tres etapas principales:

1. Preprocesamiento de datos: Unificación de los datos, estandarización de formatos y limpieza de valores nulos o inconsistentes.
2. Generación de variables adicionales: Se calcularon nuevas variables que permiten capturar características ofensivas, defensivas y contextuales de los equipos, necesarias para el modelo de Regresión de Poisson.
3. Análisis exploratorio: Se realizó un primer estudio visual y estadístico de los datos para detectar patrones, distribuciones y posibles anomalías.

Para el desarrollo de estas etapas se utilizaron tres notebooks principales, todos disponibles en el repositorio público del proyecto en GitHub. Los notebooks son:

1. clean_dataset.ipynb: Encargado de la limpieza y estructuración inicial de los datos.
2. add_values.ipynb: Responsable de la creación de variables derivadas y enriquecimiento del dataset.
3. analisis.ipynb: Dedicado al análisis exploratorio y visualización de los datos procesados.

4.1.1 Descripción y Preproceso de los Datos

Los datos utilizados en este trabajo corresponden exclusivamente a partidos de la Primera División de la liga española. La primera temporada considerada en el estudio es la temporada 2003-2004, correspondiente al primer año en el que se dispone información de cuotas de apuestas históricas, una variable clave para el enfoque adoptado. Estas cuotas no solo permiten evaluar la precisión del modelo en comparación con las estimaciones del mercado, sino que también se incorporan como variables explicativas dentro del modelo de regresión de Poisson.

En total, el dataset contiene información de 7980 partidos, abarcando 21 temporadas, con 20 equipos y 380 partidos por temporada.

A continuación, se describen las principales variables utilizadas en el dataset final, tras el proceso de limpieza y transformación. La lista completa de las variables originales puede consultarse en el Anexo B. Cabe señalar que algunas variables solo están disponibles en determinadas temporadas, por lo que su presencia en el conjunto de datos puede variar. Las variables que se describen a continuación están presentes de forma consistente en todas las temporadas incluidas en el análisis.

Estructura y Procesamiento del Dataset Básico

Para más información sobre las variables de los datasets, se puede consultar el archivo de referencia de la plataforma: Football-Data.co.uk./notes.txt

Date: Fecha del partido

Season: Temporada

HomeTeam: Equipo local

AwayTeam: Equipo visitante

FTHG: Goles del equipo local (*Full Time Home Goals*)

FTAG: Goles del equipo visitante (*Full Time Away Goals*)

FTR: Resultado del partido (H = Victoria local, D = Empate, A = Victoria visitante)

AvgH: Cuotas promedio del mercado (victoria local)

AvgD: Cuotas promedio del mercado (empate)

AvgA: Cuotas promedio del mercado (victoria visitante)

En esta fase se llevó a cabo una selección de variables fundamentales para el modelado de los goles mediante la distribución de Poisson. Se conservaron aquellas variables que recogen la información esencial del encuentro: los equipos participantes, el resultado final del partido y la condición de localía, necesaria para estimar la ventaja de jugar en casa.

Debido a que no todos los partidos cuentan con datos de todas las casas disponibles, se optó por utilizar los promedios de las cuotas disponibles para cada resultado, (victoria local, empate y victoria visitante), representados por las variables AvgH, AvgD y AvgA. Esta aproximación ofrece una estimación más estable y comparable, lo que facilita evaluar la precisión del modelo y estimar el posible beneficio teórico en un entorno real de apuestas.

	Date	Season	HomeTeam	AwayTeam	FTHG	FTAG	FTR	AvgH	AvgD	AvgA
0	2003-08-30	2003-04	Albacete	Osasuna	0	2	A	2.21	3.06	2.99
1	2003-08-30	2003-04	Ath Bilbao	Barcelona	0	1	A	2.64	3.13	2.42
2	2003-08-30	2003-04	Espanol	Sociedad	1	1	D	2.58	3.10	2.48
3	2003-08-30	2003-04	Malaga	Villarreal	0	0	D	2.27	3.08	2.88
4	2003-08-30	2003-04	Real Madrid	Betis	2	1	H	1.38	4.00	7.18
...	...	...	...	...	...	...	...	...	...	...
7975	2024-05-25	2023-24	Real Madrid	Betis	0	0	D	5.90	2.30	6.03
7976	2024-05-26	2023-24	Getafe	Mallorca	1	2	A	3.01	2.51	3.25
7977	2024-05-26	2023-24	Celta	Valencia	2	2	D	3.36	2.27	3.59
7978	2024-05-26	2023-24	Las Palmas	Alaves	1	1	D	2.75	2.63	3.28
7979	2024-05-26	2023-24	Sevilla	Barcelona	1	2	A	2.34	3.94	3.67

7980 rows × 10 columns

Tabla 1 – Dataset básico de partidos desde la temporada 03/04 hasta 23/24

Estructura y Procesamiento del Dataset Ampliado

El conjunto de datos ampliado se construyó a partir del dataset básico mediante la generación de variables adicionales diseñadas para enriquecer el modelado estadístico, especialmente en el contexto de la regresión de Poisson.

En este conjunto se incluyen métricas que recogen el rendimiento reciente del equipo analizado y de su rival (como promedios de goles, victorias o diferencias de goles), así como el historial de enfrentamientos directos entre ambos. También se incorporan transformaciones de las cuotas de apuestas adaptadas a la perspectiva de cada equipo. Estas variables permiten capturar mejor el contexto competitivo de cada partido y mejorar la capacidad predictiva del modelo. La lista completa de variables derivadas se encuentra disponible en el Anexo B.2. A continuación, se presenta una vista preliminar del dataset ampliado.

	season	date	team	rival_team	home_adv	last_season_team	last_season_rival
0	2004-05	2004-08-28	Ath Madrid	Malaga	1	7	10
1	2004-05	2004-08-28	Espanol	La Coruna	1	17	3
2	2004-05	2004-08-28	Numancia	Betis	1	21	9
3	2004-05	2004-08-29	Mallorca	Real Madrid	1	11	4
4	2004-05	2004-08-29	Osasuna	Ath Bilbao	1	12	5

Tabla 2 - Dataset aumentado de partidos (columnas 1-7)

	pct_wins	avg_goals_scored	avg_goals_received	goal_difference	pct_wins_rival	avg_goals_scored_rival
0	0.4	1.7	1.9	-2	0.3	1.2
1	0.5	1.7	1.4	3	0.6	1.5
2	0.0	0.0	0.0	0	0.4	1.1
3	0.6	1.9	1.7	2	0.3	1.6
4	0.2	1.0	1.3	-3	0.4	1.8

Tabla 3 - Dataset aumentado de partidos (columnas 8-13)

	avg_goals_received_rival	goal_difference_rival	pct_wins_vs_rival	avg_goals_scored_vs_rival
0	1.3	-1	0.5	1.5
1	0.7	8	0.5	1.5
2	1.3	-2	0.0	0.0
3	2.3	-7	0.5	2.0
4	1.4	4	0.0	1.0

Tabla 4 - Dataset aumentado de partidos (columnas 14-17)

	avg_goals_received_vs_rival	goal_difference_vs_rival	AvgH	AvgD	AvgA	goals_team	goals_rival	result
0	1.5	0	1.75	3.20	4.30	2	0	1
1	1.0	1	2.80	3.16	2.32	1	1	0
2	0.0	0	2.50	3.12	2.56	1	1	0
3	2.5	-1	4.42	3.25	1.75	0	1	-1
4	1.5	-1	2.31	3.13	2.77	1	1	0

Tabla 5 - Dataset aumentado de partidos (columnas 18-25)

Importancia de estas Métricas

Estas nuevas métricas son fundamentales porque permiten incorporar información contextual y dinámica al modelo, reflejando no solo el rendimiento global de los equipos, sino también su forma reciente y su historial frente a rivales concretos. Esto aporta una visión más realista del comportamiento esperado en cada partido, lo que mejora significativamente la precisión de las predicciones.

4.1.2 Análisis Exploratorio de los Datos (EDA)

Antes de proceder al modelado estadístico, se llevó a cabo un análisis exploratorio de los datos (EDA, por sus siglas en inglés) con el objetivo de comprender mejor la estructura del dataset, detectar posibles anomalías y observar patrones relevantes. Esta fase permite identificar distribuciones, relaciones entre variables y comportamientos atípicos, lo cual resulta clave para validar la calidad del conjunto de datos y orientar adecuadamente el diseño del modelo.

Información General del Dataset

El dataset utilizado contiene un total de 7.980 partidos disputados en 21 temporadas de LaLiga, desde la 2003/04 hasta la 2023/24. Cada temporada cuenta con 20 equipos que juegan 38 partidos cada uno (19 como local y 19 como visitante), lo que equivale a 380 encuentros por temporada. En este periodo han participado 41 equipos distintos, aunque con diferentes niveles de continuidad.

Para el modelado estadístico de los goles, se han utilizado únicamente las cinco temporadas más recientes, dejando la última temporada para la evaluación del modelo. Esta decisión se debe a consideraciones de eficiencia computacional y a que el rendimiento de los equipos varía significativamente entre temporadas, por lo que los datos antiguos pierden relevancia predictiva.

Ventaja de Jugar en Casa

Uno de los primeros aspectos analizados fue la diferencia de rendimiento entre los equipos que juegan como local y visitante. Los equipos locales marcan una media de 1,531 goles por partido, mientras que los visitantes anotan 1,130. Esta diferencia refleja el conocido efecto campo, que otorga una ligera ventaja al conjunto que actúa en casa. Esta tendencia está presente de forma consistente a lo largo de las temporadas analizadas y ha sido ampliamente reconocida tanto en estudios estadísticos como en análisis deportivos. El hecho de jugar en casa influye no solo en la capacidad ofensiva, sino también en el resultado final del encuentro.

También se examinó la relación entre los goles anotados y el resultado del partido desde la perspectiva del equipo local. La correlación entre los goles del equipo local y su victoria fue de 0,635, lo que indica una asociación directa significativa. En cambio, los goles del visitante se relacionan negativamente con la probabilidad de victoria local, con una correlación de –0,484. Este comportamiento se ve reflejado visualmente en la Figura 2, donde las victorias en casa (H) son el desenlace más frecuente, seguidas por las victorias visitantes (A) y los empates (D). Este

patrón justifica la inclusión del efecto de localía como parámetro específico dentro del modelo estadístico.

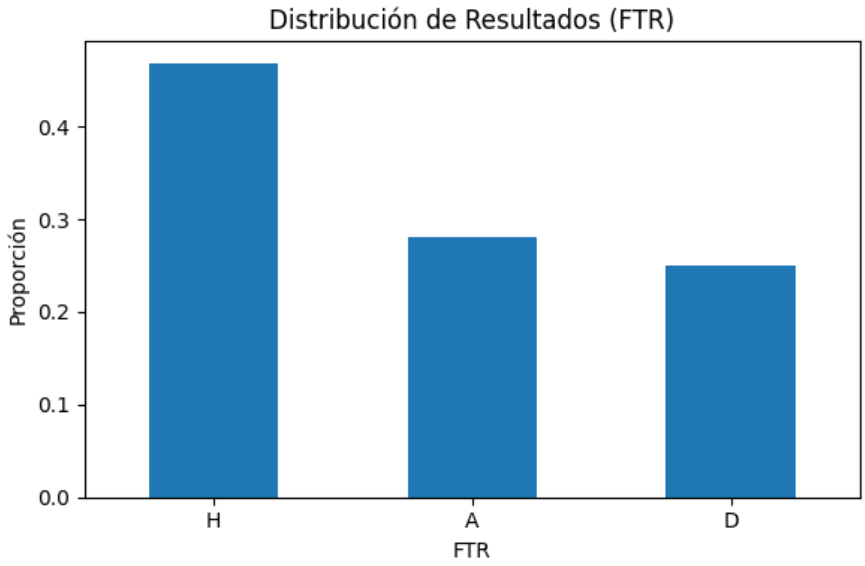

Figura 3. Proporción de victorias locales, empates y victorias visitantes en el dataset

Variación entre Temporadas Consecutivas

Un segundo bloque de análisis se centró en evaluar la estabilidad del rendimiento de los equipos entre temporadas consecutivas. El promedio de variación de goles anotados por equipo entre temporadas consecutivas es de 8,07 goles, mientras que la variación media en la posición final en la tabla es de 3,61 puestos. Estos datos evidencian una importante variabilidad interanual, tanto a nivel ofensivo como competitivo, lo que justifica la decisión de limitar el entrenamiento de los modelos a las temporadas más recientes. Incluir temporadas demasiado antiguas podría introducir ruido y dificultar la adaptación del modelo a las dinámicas actuales del campeonato.

Frecuencia de Marcadores

Con el objetivo de comprender mejor la distribución de los resultados posibles, se elaboraron matrices de calor que representan la frecuencia conjunta de marcadores en función del número de goles anotados por el equipo local y el visitante. Estas visualizaciones muestran que los resultados más comunes se encuentran en la franja baja de anotación, especialmente en torno a marcadores como 1–0, 1–1 y 2–1.

En la Figura 3 se muestra la matriz completa, sin restricciones, incluyendo todos los partidos con resultados de hasta diez goles por equipo. Para facilitar la lectura y enfocarse en los

marcadores más frecuentes, en la Figura 4 se presenta una versión restringida a partidos donde ambos equipos marcaron entre 0 y 6 goles. En esta representación se aprecia con mayor claridad que la mayor densidad se encuentra en los valores cercanos a la diagonal, lo que refleja la alta incidencia de empates y victorias por la mínima diferencia.

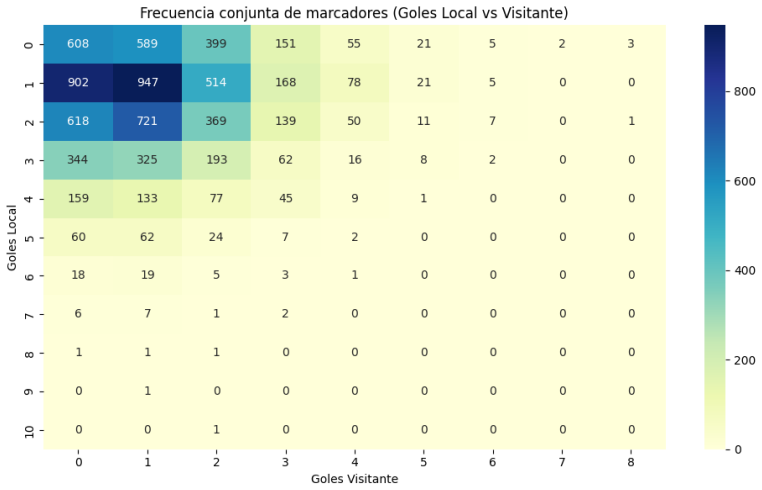

Figura 4. Frecuencia conjunta de marcadores (Goles Local vs Visitante)

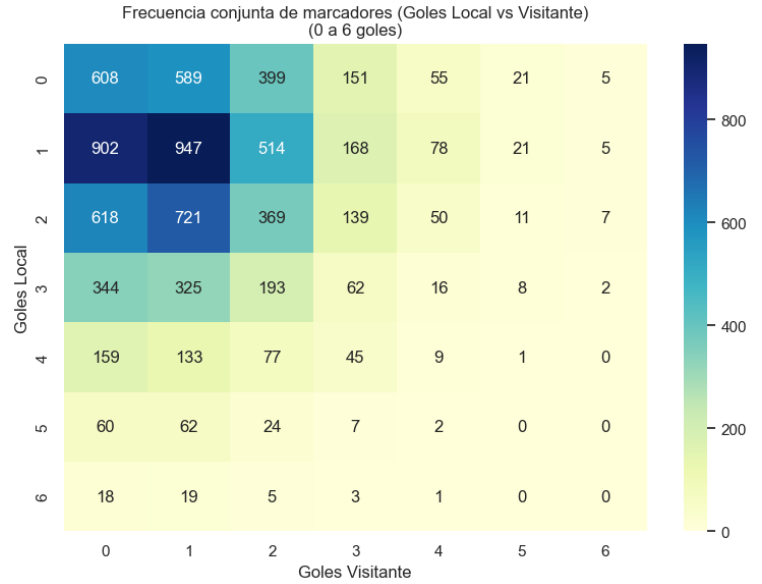

Figura 5. Frecuencia conjunta de marcadores con restricción de 0 a 6 goles por equipo

Este tipo de análisis resulta fundamental para el diseño y ajuste de modelos estadísticos, ya que permite identificar qué combinaciones de goles son más probables en la práctica. Así, se pueden construir modelos más realistas, que reproduzcan con mayor precisión la distribución empírica de los marcadores y mejoren la calidad de las predicciones.

Distribución de las Cuotas de Apuestas

Finalmente, se estudió la distribución de las cuotas de apuestas asociadas a los distintos resultados del mercado 1X2. Las cuotas promedio de victoria local (AvgH) se concentran en valores relativamente bajos, lo cual refleja la mayor probabilidad asignada por el mercado a este resultado. En contraste, las cuotas de victoria visitante (AvgA) presentan una mayor dispersión y tienden a situarse en niveles intermedios. Por su parte, las cuotas de empate (AvgD) son, en general, más elevadas, lo que indica que el mercado percibe este desenlace como el menos probable.

En la Figura 5 se muestra la distribución de densidad de estas tres variables, permitiendo visualizar de forma clara la diferencia en la forma y concentración de cada una. Las curvas confirman que el mercado atribuye una mayor probabilidad al triunfo del equipo local, con menor incertidumbre respecto a este resultado en comparación con los otros dos.

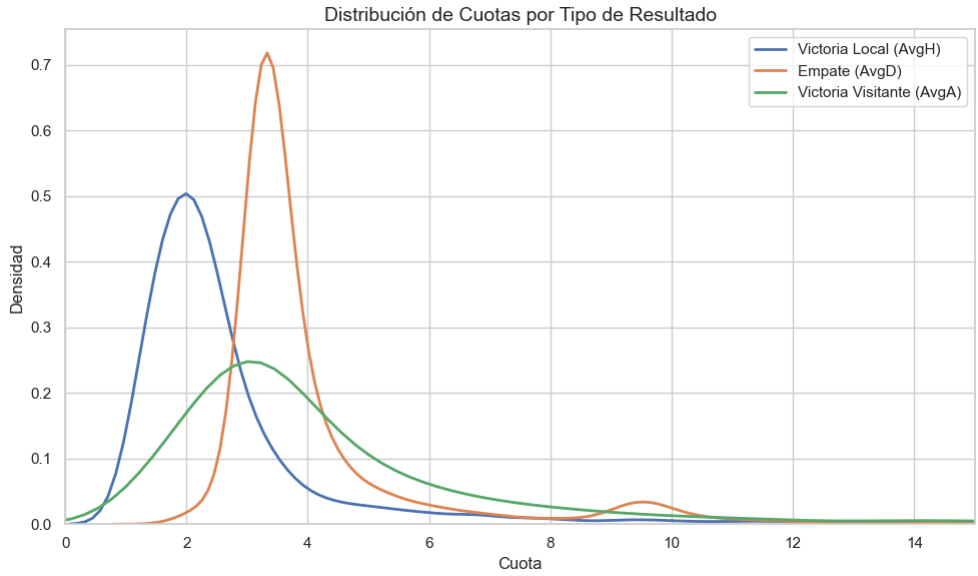

Figura 6. Distribución de cuotas para los tres posibles resultados del mercado 1X2

Esta información no solo es útil como variable explicativa dentro del modelo de regresión de Poisson, sino que también puede emplearse para diseñar estrategias de apuestas. Comparar las probabilidades implícitas en las cuotas con las probabilidades estimadas por el modelo permite detectar posibles ineficiencias en el mercado, lo que abre la puerta a la identificación de apuestas con valor esperado positivo.

4.2 Desarrollo de los Modelos

En este apartado se presentan los diferentes modelos estadísticos implementados para estimar resultados de fútbol mediante la distribución de Poisson. Cada modelo ha sido desarrollado en un notebook independiente, utilizando Python y datos históricos de LaLiga. Los modelos aplicados para el análisis de goles y la predicción de resultados se agrupan en las siguientes categorías:

- Modelo Poisson Simple: una aproximación general que estima el número de goles promedio por equipo local y visitante.
- Modelo Poisson Doble: una extensión que tiene en cuenta las características individuales de los equipos, estimando parámetros ofensivos, defensivos y de localía.
- Modelo Poisson Bivariante: un enfoque que incorpora la dependencia estadística entre los goles de ambos equipos, ajustando mejor los empates.
- Modelo de Regresión de Poisson: una formulación más flexible que permite incorporar múltiples variables explicativas, incluyendo información histórica y de mercado.

A continuación, se detalla el proceso seguido para el desarrollo de los modelos, los resultados obtenidos y las principales conclusiones prácticas derivadas de su uso.

4.2.1 Modelo Poisson Simple

El modelo Poisson simple parte de la hipótesis de que el número de goles anotados por un equipo como local o visitante puede modelarse mediante una distribución de Poisson con un único parámetro λ, calculado a partir de la media histórica de goles. Este enfoque no tiene en cuenta el rival ni otros factores del partido.

Para su implementación, se calcularon las medias de goles por equipo usando datos de la temporada 2022/23, y se evaluó el ajuste del modelo comparando las distribuciones teóricas con los goles reales de la temporada 2023/24. Además, se calcularon intervalos de confianza al 95 % para cada valor de λ, con el fin de valorar la estabilidad de las estimaciones y la variabilidad en el comportamiento goleador de cada equipo.

Resultados

En el caso del Real Madrid, se estimó un λ de 2.32 para los partidos como local y un λ de 1.63 para los partidos como visitante. Las distribuciones de Poisson correspondientes reflejan

adecuadamente el patrón general de goles marcados en ambas situaciones. En la Figura 6 se aprecia una mayor probabilidad de anotar 2 o 3 goles en casa y 1 o 2 goles fuera.

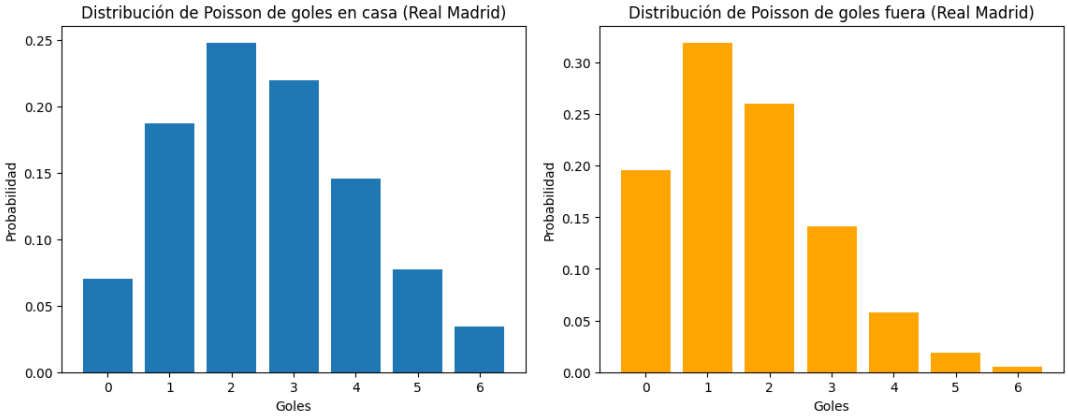

Figura 7. Distribución de Poisson de goles para el Real Madrid (2022-2023)

La **Figura 7** compara las distribuciones teóricas de goles entre varios equipos actuando como locales. Se observa que equipos caracterizados por su alto rendimiento como Real Madrid y Barcelona presentan mayor probabilidad de marcar más de 2 goles, mientras que Sevilla o Valencia tienden a concentrarse en valores más bajos.

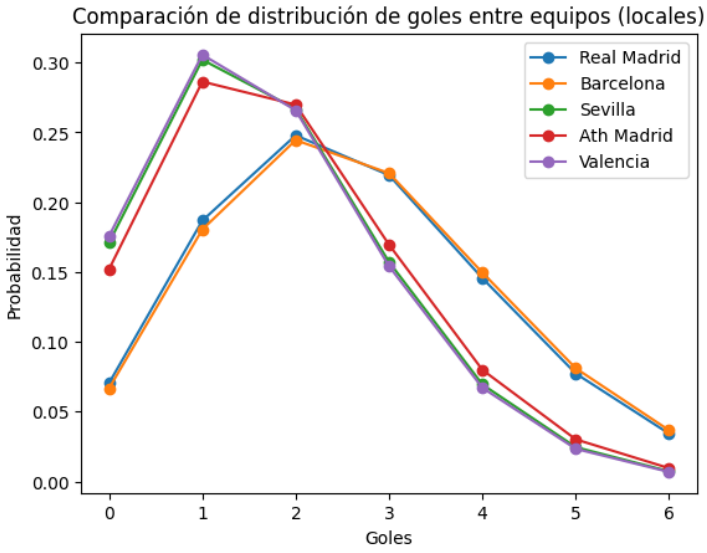

Figura 8. Comparación entre distribuciones de goles entre equipos como local

En la temporada 2023/24, el modelo ajustado con datos de 2022/23 resultó con una log-verosimilitud media por partido de –1.83 para el Real Madrid. La **Figura 8** muestra esta métrica para seis equipos, tanto como locales como visitantes.

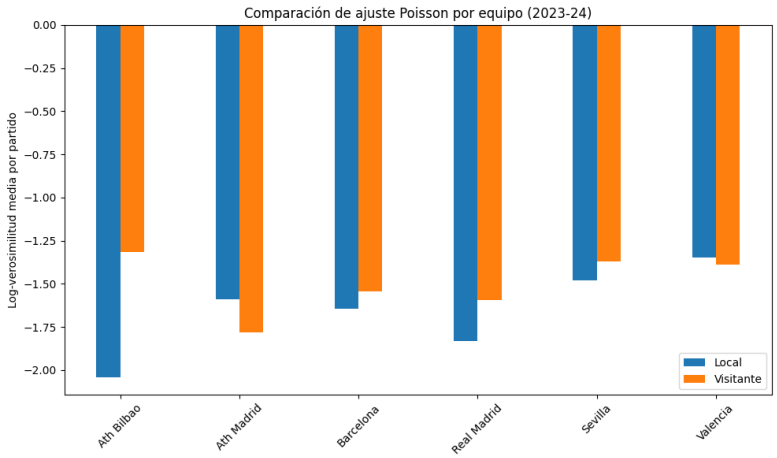

Figura 9. Comparación del ajuste de Poisson por equipo (2023-2024)

Se observa que el modelo tiende a ajustarse ligeramente mejor para los partidos jugados fuera de casa, lo que puede deberse a una menor dispersión en los resultados de visitante o a un sesgo del modelo en la estimación de λ como local.

Finalmente, la **Figura 9** muestra, para estos mismos equipos, la comparación entre la distribución teórica de Poisson y la distribución empírica de goles observados en la temporada 2023/24.

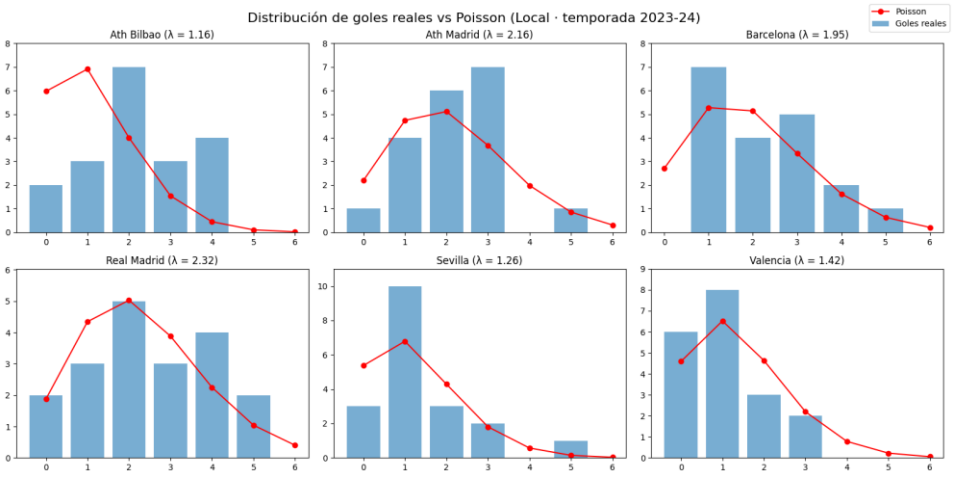

Figura 10. Distribución de goles reales vs Poisson (2023-2024)

Se observa que, en general, el modelo reproduce bien la forma de la distribución, aunque en algunos equipos tiende a sobrestimar o subestimar ciertos valores concretos de goles, como el 0 o el 4.

Además de las medias de goles (λ), se calcularon los intervalos de confianza al 95 % para cada equipo, tanto en condición de local como de visitante, utilizando los datos de la temporada 2022/23. Un intervalo de confianza del 95 % significa que, si se repitiera el procedimiento de

estimación muchas veces con diferentes muestras, aproximadamente el 95 % de los intervalos construidos contendrían al verdadero valor del parámetro λ. Un intervalo más amplio sugiere una mayor variabilidad en el comportamiento goleador del equipo y, por tanto, una estimación menos precisa. Por el contrario, un intervalo más estrecho refleja una mayor estabilidad en los resultados. La **Tabla 6** recoge las estimaciones obtenidas junto con sus respectivos intervalos.

Equipo	λ Local	IC95% Local	λ Visitante	IC95% Visitante
Ath Bilbao	1.16	[0.67, 1.64]	1.32	[0.80, 1.83]
Ath Madrid	2.16	[1.50, 2.82]	1.53	[0.97, 2.08]
Barcelona	1.95	[1.32, 2.57]	1.74	[1.14, 2.33]
Real Madrid	2.32	[1.63, 3.00]	1.63	[1.06, 2.21]
Sevilla	1.26	[0.76, 1.77]	1.21	[0.72, 1.71]
Valencia	1.42	[0.89, 1.96]	0.79	[0.39, 1.19]

Tabla 6. Estimación de λ e intervalos de confianza al 95 % para goles como local y visitante (temporada 2022/23)

Estos resultados muestran que algunos equipos tienen medias estimadas más estables (como Barcelona o Sevilla), mientras que en otros casos (como Real Madrid o Ath. Madrid) los intervalos son más amplios, lo que sugiere una mayor variabilidad ofensiva en los partidos analizados

Conclusión

El modelo Poisson simple permite estimar la distribución de goles a partir de medias históricas, diferenciando entre encuentros disputados en casa y fuera. En general, el ajuste obtenido es razonable, especialmente en aquellos equipos cuyo rendimiento ha sido más regular a lo largo del tiempo.

Sin embargo, la sencillez del modelo también limita su capacidad para adaptarse a situaciones más complejas. En algunos casos, se observan desviaciones respecto a los datos reales, y los intervalos de confianza muestran que la precisión de las estimaciones disminuye cuando existe mayor variabilidad. Por ello, aunque resulta útil como punto de partida, se hace necesario avanzar hacia modelos más completos que tengan en cuenta otros factores relevantes.

4.2.2 Modelo Poisson Doble

El modelo Poisson doble introduce una mejora respecto al modelo simple al estimar el número de goles en función de las características ofensivas y defensivas de cada equipo, además del efecto de jugar en casa. Para ello, se definen dos distribuciones de Poisson independientes: una para los goles del equipo local y otra para los del visitante, cada una con su media específica calculada a partir de estos factores.

El modelo se entrenó con datos de las temporadas 2019/20 a 2022/23, y se ajustaron los parámetros mediante la maximización de la log-verosimilitud. Se estimaron un parámetro de ataque y otro de defensa para cada equipo, además de un parámetro adicional para la ventaja de jugar como local (γ). Una vez ajustado el modelo, se utilizaron los parámetros obtenidos para predecir los resultados de la temporada 2023/24.

Resultados

El modelo se entrenó con 1.520 partidos y se validó sobre 342 de los 380 encuentros de la temporada 2023/24, excluyendo los que involucraban a un equipo recién ascendido sin historial disponible. El parámetro estimado para la ventaja de jugar en casa fue $\gamma = 0.2987$, lo que confirma la existencia de un efecto local positivo, consistente con lo observado en los datos exploratorios.

A partir de los goles esperados para cada partido, se construyó la matriz de probabilidad conjunta de goles y se calcularon las probabilidades asociadas a los tres posibles resultados: victoria local, empate o victoria visitante. El modelo clasificó correctamente el resultado más probable en el 53,2 % de los partidos. La **Figura 10** muestra la matriz de confusión correspondiente a los resultados del modelo

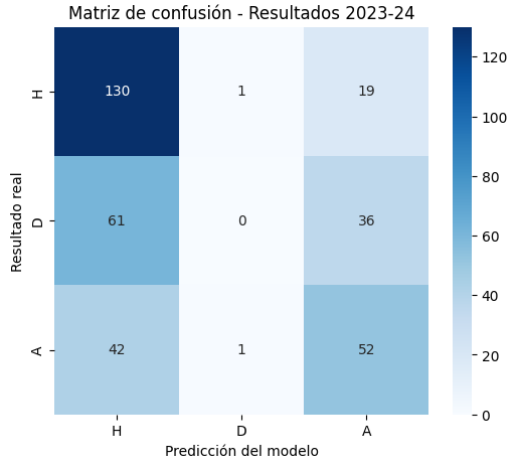

Figura 11. Matriz de confusión para el modelo Poisson doble (2023/24)

Se observa un buen nivel de acierto en victorias locales, aunque el modelo tiene dificultades para predecir empates, algo ya señalado por Maher (1982) [2] como una limitación inherente a los modelos basados en Poisson independientes.

El rendimiento del modelo se evaluó también mediante la log-verosimilitud, que alcanzó un valor total de –337.9462 y una media por partido de –0.9881. Estas cifras indican una mejora clara respecto al modelo simple. Además, se calculó el F1-score macro, que toma en cuenta el equilibrio entre precisión y cobertura en cada clase y es especialmente útil cuando las categorías están desbalanceadas. En este caso, el valor fue de 0.398, penalizado por el hecho de que el modelo no logró predecir correctamente ningún empate.

Por último, se simuló una estrategia de apuestas basada en las predicciones del modelo. Se apostó en todos los partidos de la temporada 2023/24, asignando 1 € por apuesta de baja confianza y 2 € cuando la probabilidad estimada del resultado era igual o superior al 60 %. Esta estrategia generó una rentabilidad del 39,25 %, con un total invertido de 412 € y un beneficio neto de 161,69 €, como se resume en la **Tabla 7.**

Aciertos	Total apostado	Total ganado	Beneficio neto	Rentabilidad
182 de 342	412,00 €	573,69 €	+161,69 €	39,25 %

Tabla 7. Estrategia de apuestas basada en el modelo Poisson doble (2023/24)

Conclusión

El modelo Poisson doble supone un avance relevante respecto al enfoque simple, ya que permite diferenciar el comportamiento ofensivo y defensivo de los equipos e incorporar el efecto de jugar en casa. Los resultados muestran una mejora en el ajuste y en la capacidad predictiva, especialmente en las categorías de victoria local y visitante.

No obstante, la suposición de independencia entre los goles de ambos equipos limita su precisión en la predicción de empates. A pesar de ello, el modelo ofrece un rendimiento sólido tanto desde el punto de vista estadístico como aplicado, demostrando su utilidad en el análisis de resultados y su posible aplicación práctica en contextos como el mercado de apuestas deportivas.

4.2.3 Modelo Poisson Bivariante

El modelo Poisson bivariante supone una mejora respecto al modelo doble, al incorporar una componente común que permite capturar la dependencia entre los goles anotados por ambos equipos. Esta característica resulta especialmente útil para representar con mayor precisión la frecuencia de empates.

A diferencia del modelo doble, donde los goles del equipo local y visitante se modelan mediante dos distribuciones de Poisson independientes, aquí se emplea una formulación conjunta que introduce correlación entre ambas variables. Para ello, se utiliza una convolución de tres distribuciones de Poisson: una para cada equipo (λ_1 y λ_2) y una tercera (λ_3) que actúa como factor de dependencia entre ambas. Esta estructura permite capturar la relación estadística entre los goles de los dos equipos, lo que mejora el ajuste en situaciones donde los resultados están más equilibrados, como los empates.

El modelo se entrenó utilizando datos de las temporadas 2019/20 a 2022/23. Para cada equipo se estimaron parámetros de ataque, defensa y un par adicional de parámetros (φ) que controlan la dependencia entre los goles en función de si el equipo juega como local o visitante, siguiendo la propuesta de Karlis y Ntzoufras (2003) [12]. Además, se incluye el parámetro γ para recoger la ventaja de jugar en casa. La estimación de todos los parámetros se realizó mediante la maximización de la log-verosimilitud. Para ello, se utilizó el algoritmo L-BFGS-B, un método de optimización numérica eficiente en problemas con muchos parámetros y restricciones.

Resultados

Las predicciones se realizaron sobre 342 de los 380 partidos de la temporada 2023/24. Se excluyeron aquellos partidos en los que participaba un equipo recién ascendido para el que no se disponía de datos históricos suficientes. A partir de las medias de goles estimadas y del término de dependencia, se construyó una matriz de probabilidades conjuntas para cada partido. De esta matriz se extrajeron las probabilidades asociadas a victoria local, empate y victoria visitante.

El modelo acertó el resultado del partido en el 53,2 % de los casos. La **Figura 11** muestra la matriz de confusión resultante. Se observa una mejora en la predicción de empates respecto al modelo Poisson doble, aunque el rendimiento en esta clase sigue siendo bajo. Esta mejora es coherente con lo que se espera al introducir dependencia entre los goles de ambos equipos.

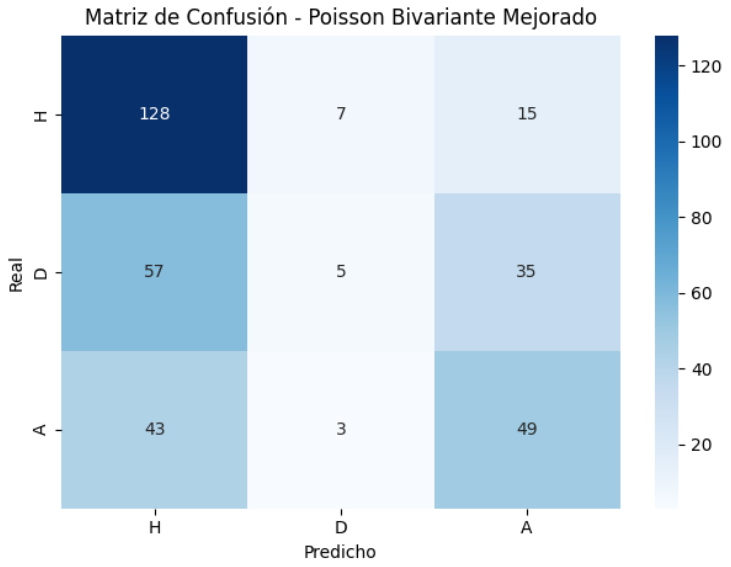

Figura 12. Matriz de confusión del modelo Poisson bivariante (2023/24)

La evaluación mediante el F1-score macro, que promedia el rendimiento en cada clase sin ponderar por frecuencia, resultó en un valor de 0.424. A pesar de la mejora respecto al modelo anterior, el modelo sigue penalizado por una baja capacidad para acertar los empates (recall = 5 % para esa clase).

En cuanto al ajuste del modelo, la log-verosimilitud total fue de –998,68, lo que se traduce en una log-verosimilitud media por partido de –2,6281, inferior a la obtenida con el modelo Poisson doble. Este resultado refleja un peor ajuste a nivel de distribución exacta de goles, lo cual es esperable dada la mayor complejidad del modelo y el riesgo de sobreajuste con más parámetros. Aun así, el valor medio estimado del componente de dependencia λ_3 fue de 0.1935, lo que confirma la existencia de correlación entre los goles de ambos equipos.

Por último, se simuló una estrategia de apuestas utilizando el mismo criterio aplicado en el modelo Poisson doble. Los resultados obtenidos muestran una rentabilidad del **47,37 %**, con 182 aciertos de 342 apuestas, un gasto total de 421 € y un beneficio neto de 199,45 €, tal y como se resume en la Tabla 8.

Aciertos	**Total apostado**	**Total ganado**	**Beneficio neto**	**Rentabilidad**
182 de 342	421,00 €	620,45 €	+199,45 €	47,37 %

Tabla 8. Estrategia de apuestas con el modelo Poisson bivariante (2023/24)

Conclusión

El modelo Poisson bivariante representa un avance importante al incorporar correlación entre los goles locales y visitantes, lo que permite una mejor representación de los empates, una de las principales limitaciones del modelo doble. Aunque el rendimiento general en términos de precisión es similar, se observa una mejora en el equilibrio de clasificación entre clases, reflejada en un F1-score macro más alto.

A pesar de que el ajuste global medido por log-verosimilitud es inferior al del modelo anterior, el Poisson bivariante demuestra un mejor comportamiento en escenarios prácticos como la simulación de apuestas, donde alcanza una rentabilidad superior. Esto sugiere que, aunque más complejo, el modelo aporta valor en aplicaciones donde importa más la calidad de las decisiones que el ajuste exacto a los datos.

4.2.4 Modelo de Regresión de Poisson

El modelo de regresión de Poisson permite estimar el número de goles esperados en función de distintas variables explicativas, lo que lo hace más flexible y adaptado al contexto que los modelos anteriores. Para ello, se ajustan dos regresiones separadas: una para los goles del equipo local y otra para los del visitante.

El entrenamiento del modelo se realizó con datos de las temporadas 2019/20 a 2022/23. Entre las variables utilizadas se incluyeron estadísticas tanto del equipo como de su rival, como el porcentaje de victorias recientes, la media de goles marcados y la diferencia de goles, entre otras. Además, se representó la identidad de cada equipo mediante variables binarias mediante codificación one-hot. El ajuste se realizó mediante máxima verosimilitud usando un modelo lineal generalizado con distribución de Poisson, implementado con la clase *GLM* de la librería *statsmodels*. Tras el entrenamiento, el modelo asignó coeficientes nulos a 7 variables, eliminándolas efectivamente del modelo.

Resultados

Las predicciones se realizaron sobre los 380 partidos de la temporada 2023/24. A partir de los goles esperados (λ) estimados para el equipo local y el visitante, se construyó una matriz de probabilidades conjunta para cada partido. De dicha matriz se extrajeron las probabilidades asociadas a victoria local, empate y victoria visitante.

El modelo acertó el resultado del partido en el 49,2 % de los casos. La **Figura 12** muestra la matriz de confusión correspondiente, en la que se aprecia una mayor capacidad para predecir empates frente a los modelos anteriores.

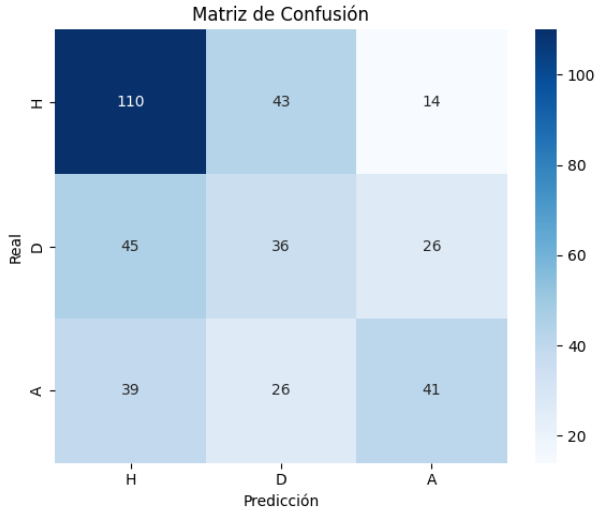

Figura 13. Matriz de confusión del modelo de regresión de Poisson (2023/24)

La evaluación mediante el F1-score terminó con un valor de 0.46, superior al alcanzado por los modelos Poisson doble y bivariante. El rendimiento es más equilibrado entre las tres clases, especialmente en la predicción de empates, lo que sugiere que la inclusión de variables explicativas proporciona al modelo una mayor capacidad de adaptación a diferentes contextos.

En términos de ajuste, el modelo presentó una log-verosimilitud total de −1118,49 y una media por partido de −2,9434. Estos valores, algo inferiores a los de modelos estructurales más simples, reflejan que, aunque el modelo es más flexible y eficaz en clasificación, pierde capacidad para ajustarse a la distribución exacta de goles, probablemente debido a la mayor complejidad y número de parámetros.

Por último, se aplicó la misma estrategia de apuestas utilizada en los modelos anteriores. Los resultados muestran una rentabilidad del 69,56 %, con 187 aciertos de 380 partidos, un total invertido de 475 € y un beneficio neto de 339,84 €, tal y como se resume en la **Tabla 9**.

Aciertos	Total apostado	Total ganado	Beneficio neto	Rentabilidad
187 de 380	475,00 €	814,84 €	+339,84 €	71,55 %

Tabla 9. Estrategia de apuestas con el modelo de regresión de Poisson (2023/24)

Conclusión

El modelo de regresión de Poisson destaca por su capacidad para incorporar información contextual relevante y adaptarse a la dinámica entre equipos. Aunque su precisión global es ligeramente inferior a la de los modelos estructurales, presenta un mejor equilibrio entre clases y una mayor capacidad para predecir empates.

A nivel de log-verosimilitud, el ajuste es inferior, lo que indica que el modelo no reproduce con tanta precisión la distribución exacta de goles. No obstante, su rendimiento práctico, especialmente en la simulación de apuestas, es notablemente superior, lo que sugiere que su mayor flexibilidad compensa la pérdida de ajuste teórico en aplicaciones reales.

4.2.5 Comparación entre Modelos

En este apartado se comparan los resultados obtenidos por los diferentes modelos desarrollados, considerando tres dimensiones principales: la capacidad de acierto por tipo de resultado, el rendimiento en una estrategia simulada de apuestas y el valor de la log-verosimilitud, como medida del ajuste probabilístico. Las siguientes tablas resumen estos aspectos clave.

En la **Tabla 10** se puede ver que el modelo de regresión de Poisson es el único que predice un número significativo de empates (37), mientras que el modelo bivariante y el modelo de Poisson doble alcanzan 5 y 0 ciertos respectivamente en esta categoría. A pesar de ello, estos modelos tienen un mayor número de aciertos en victorias y derrotas.

Modelo	Local	Empate	Visitante	Accuracy	F1-score
Poisson Doble	130	0	52	53,2 %	0,398
Poisson Bivariante	128	5	49	53,2 %	0,424
Regresión de Poisson	110	36	41	49,2 %	0,462

Tabla 10. Número de aciertos por tipo de resultado (2023/24)

En la **Tabla 11** se presentan los resultados de la simulación de apuestas. Entre ellas destaca el modelo de regresión de Poisson, que obtiene la mayor rentabilidad (71,55%) y el mayor beneficio neto.

Modelo	Aciertos	Apostado	Ganado	Beneficio neto	Rentabilidad
Poisson Doble	182	412,00 €	573,69 €	+161,69 €	39,25 %
Poisson Bivariante	182	421,00 €	620,45 €	+199,45 €	47,37 %
Regresión de Poisson	188	475,00 €	814,84 €	+349,84 €	71,55 %

Tabla 11. Resultados de la simulación de apuestas por modelo

En la **Tabla 12**, que muestra los valores de log-verosimilitud, el modelo Poisson doble alcanza el mejor ajuste probabilístico a los datos, con la log-verosimilitud media más alta (–0.9881). Le sigue el modelo simple, mientras que el bivariante y el de regresión obtienen peores resultados en esta métrica. Esto sugiere que, aunque estos modelos puedan mejorar en términos de predicción o rentabilidad, no necesariamente ofrecen una mejor estimación de las distribuciones de probabilidad subyacentes.

Modelo	Log-verosimilitud total	Log-verosimilitud media
Poisson Simple	–897,13 (aprox.)	–1,8326
Poisson Doble	–337,95	–0,9881
Poisson Bivariante	–998,68	–2,6281
Regresión de Poisson	–1118,49	–2,9434

Tabla 12. Log-verosimilitud de los modelos (2023/24)

5. CONCLUSIONES

Este capítulo resume los principales hallazgos del trabajo, así como una reflexión crítica sobre los resultados obtenidos, las limitaciones encontradas durante el proceso y propuestas de investigación futura.

5.1 Objetivos Planteados y Grado de Cumplimento

El objetivo general del trabajo ha sido desarrollar y evaluar modelos estadísticos basados en la distribución de Poisson para predecir resultados de fútbol y analizar su utilidad en el mercado de apuestas deportivas. A lo largo del proyecto se han cumplido, en distinto grado, los objetivos específicos planteados al inicio. Estos se detallan a continuación junto con su nivel de cumplimento.

Realizar un Análisis Exploratorio de los Datos Históricos

Este objetivo se cumplió mediante un estudio detallado de los datos disponibles, que permitió identificar patrones relevantes, como la ventaja de jugar en casa, la frecuencia de resultados más comunes, la variabilidad interanual de los equipos y la distribución de cuotas de apuestas. Este análisis fue fundamental para orientar el desarrollo posterior de los modelos y seleccionar las variables más adecuadas.

Modelar los Goles Históricos por Equipo con el Modelo de Poisson Simple

Se cumplió este objetivo mediante el cálculo de medias de goles como local y visitante a partir de la temporada 2022/23, y su evaluación sobre los datos de 2023/24. El modelo permitió estimar la distribución de goles de forma razonable, especialmente en equipos con comportamiento estable. También se calcularon intervalos de confianza al 95 % para valorar la variabilidad de los goles de cada equipo. Sin embargo, al no considerar el rival ni otros factores contextuales, mostró limitaciones en precisión y capacidad predictiva, lo que confirma su utilidad como punto de partida, pero no como modelo final.

Desarrollo de los Modelos

Se implementaron tres modelos avanzados con el objetivo de mejorar progresivamente la capacidad predictiva y la aplicabilidad práctica del enfoque basado en la distribución de Poisson. El modelo Poisson doble incorporó parámetros ofensivos, defensivos y el efecto de

localía, logrando una mejora notable en la log-verosimilitud respecto al modelo simple, aunque mostró limitaciones en la predicción de empates. El modelo Poisson bivariante, al introducir correlación entre los goles de ambos equipos, mejoró específicamente la estimación de empates y aumentó la rentabilidad en la simulación de apuestas, a pesar de presentar un ajuste estadístico algo inferior. Por último, la regresión de Poisson permitió incorporar variables explicativas como cuotas y rendimiento reciente, ofreciendo un enfoque más flexible y contextualizado. Aunque su log-verosimilitud fue más baja, fue el modelo más rentable y con un mejor equilibrio entre clases, lo que lo convierte en el más eficaz desde el punto de vista aplicado.

Evaluación de los Modelos

Se evaluó el rendimiento de los modelos mediante distintas métricas y enfoques, con el objetivo de analizarlos desde múltiples perspectivas. Entre las métricas utilizadas se encuentran la log-verosimilitud, la accuracy, el F1-score macro y una simulación de estrategia de apuestas. Esta combinación permitió no solo valorar el ajuste estadístico de los modelos, sino también su utilidad práctica en contextos reales. Gracias a este análisis comparativo, se pudo observar cómo, a pesar de que algunas métricas como la log-verosimilitud empeoraban en modelos más complejos, su rentabilidad práctica aumentaba significativamente. Esta diferencia de comportamiento según la métrica permitió extraer conclusiones más completas y relevantes sobre el valor real de cada modelo en función del objetivo final.

Analizar la Aplicabilidad Práctica en el Mercado de Apuestas Deportivas

Los resultados muestran que los modelos desarrollados pueden detectar oportunidades de valor en el mercado de apuestas. Aunque la estrategia de apuestas utilizada fue sencilla, basada en apostar al resultado más probable con mayor confianza cuando la estimación era alta, se obtuvieron resultados positivos en términos de rentabilidad. De cara al futuro, sería interesante explorar estrategias más completas, con algoritmos que ajusten de forma más precisa la cantidad a apostar según las probabilidades estimadas y otros factores, con el objetivo de mejorar la gestión y los beneficios obtenidos.

Comparación entre Modelos y Limitaciones

La comparación entre los distintos modelos permitió observar que un mejor ajuste estadístico, medido por la log-verosimilitud, no siempre se traduce en una mayor rentabilidad práctica. Esto resalta la importancia de considerar el contexto de aplicación al evaluar un modelo. Además,

durante el desarrollo se identificaron las principales limitaciones y supuestos de cada enfoque, lo que ayudó a interpretar los resultados de forma más precisa y realista. Estas limitaciones se detallarán en el siguiente apartado, ya que su análisis no solo permite entender el alcance de los modelos, sino que también señala caminos claros para futuras mejoras y desarrollos más robustos.

5.2 Limitaciones

Limitaciones técnicas

El desarrollo y evaluación de los modelos estadísticos presentó varias limitaciones técnicas que condicionaron el alcance del trabajo. En primer lugar, el coste computacional del entrenamiento fue considerable, especialmente en el caso del modelo Poisson bivariante, cuya ejecución superó las tres horas. Esto se debe a que, para cada partido, había que calcular varios parámetros relacionados con el ataque, la defensa y la relación entre los goles de ambos equipos, lo que hacía que el proceso fuera mucho más lento. Por esta razón, no fue posible utilizar las 20 temporadas completas disponibles en el dataset, limitando el entrenamiento a las cinco más recientes. Además, no se pudo implementar una validación secuencial partido a partido, que habría permitido simular un entorno de predicción más realista en el que los modelos se actualizasen de forma continua con nuevos datos.

Además, una de las dificultades más relevantes fue el desbalance de clases en los resultados de los partidos. Las victorias locales eran considerablemente más frecuentes que empates o victorias visitantes, lo que provocó que los modelos tendieran a aprender principalmente esta clase dominante y redujeran su capacidad para predecir empates con precisión.

Limitaciones contextuales

Más allá de las cuestiones técnicas, también hay limitaciones propias del contexto que afectan a cualquier modelo de predicción en el ámbito del fútbol. Se trata de un deporte con mucha variabilidad y un alto componente aleatorio, donde cualquier acción inesperada puede cambiar el curso de un partido. Además, no se incluyeron factores externos relevantes como lesiones, sanciones, decisiones arbitrales o condiciones meteorológicas, que pueden tener un impacto significativo en el desenlace de los encuentros. Por otro lado, las cuotas de apuestas utilizadas como referencia en algunos modelos están influenciadas por el comportamiento del mercado y no reflejan únicamente probabilidades objetivas, lo que introduce ciertos ruidos y sesgos en las predicciones.

5.3 Discusión y Análisis Crítico

Comparación con la Bibliografía

Uno de los aspectos señalados en la bibliografía, como en el trabajo de Maher (1982) [2], es la dificultad de predecir empates mediante modelos Poisson que asumen independencia entre los goles de los equipos. Este mismo problema se ha reflejado en los resultados obtenidos con el modelo Poisson doble, que no fue capaz de predecir correctamente ningún empate en la temporada analizada, lo que confirma una de sus limitaciones más conocidas.

Además, muchos trabajos previos evalúan el rendimiento de los modelos desde un enfoque probabilístico, midiendo la capacidad para predecir el número exacto de goles mediante intervalos de confianza o pruebas de bondad de ajuste, como la chi cuadrado. Por ejemplo, en el estudio de Loukas et al. (2024) [11], el modelo propuesto alcanzó un 75 % de precisión al predecir los goles dentro de un margen de ±1 respecto al marcador real. Sin embargo, este tipo de margen, aunque aceptable estadísticamente, puede ser demasiado amplio en el contexto del fútbol, donde un solo gol de diferencia puede cambiar completamente el resultado del partido. De hecho, como se observó en la **Figura 5**, los tres marcadores más frecuentes en el dataset fueron 1–1, 1–0 y 2–1, lo que muestra la alta sensibilidad del resultado al número exacto de goles.

Por ello, este trabajo propone un enfoque de validación complementario al habitual en la literatura, centrado no solo en la precisión de los goles predichos, sino también en la clasificación del resultado final (victoria local, empate o victoria visitante) y, especialmente, en su utilidad práctica mediante una simulación de estrategia de apuestas. Este enfoque diferencial aporta un valor añadido al análisis, ya que permite evaluar el modelo desde una perspectiva más aplicada, directamente vinculada con decisiones reales en el mercado de apuestas.

Comparación entre Modelos

Una de las principales conclusiones del trabajo es que el modelo con mejor ajuste estadístico (Poisson doble) no fue el que obtuvo mejores resultados económicos. En cambio, la regresión de Poisson, a pesar de tener una log-verosimilitud inferior, fue el modelo más rentable en la simulación de apuestas. Este resultado refuerza la idea de que, en entornos prácticos como el mercado de apuestas deportivas, es más útil incorporar información contextual relevante que simplemente maximizar el ajuste teórico del modelo. Una posible explicación de la menor log-verosimilitud en modelos más complejos como el de regresión es que esta métrica puede penalizar la presencia de muchos parámetros, ya que mide cómo de bien el modelo reproduce

la distribución exacta de los datos, evitando el sobreajuste. Por tanto, aunque el modelo sea más flexible y ofrezca mejores resultados en la práctica, su complejidad puede afectar negativamente en este tipo de medidas estadísticas.

Un aspecto importante fue la dificultad general para predecir empates, algo ya señalado en la literatura y confirmado en los resultados obtenidos. Sin embargo, fue precisamente en esta categoría donde se identificó un mayor margen de beneficio, debido a que las cuotas asociadas a los empates tienden a ser más elevadas y, en muchos casos, están subestimadas por las casas de apuestas. Desde una perspectiva práctica, esto implica que acertar un menor número de empates bien valorados puede ser más rentable que predecir con acierto resultados más frecuentes pero con menor retorno, como las victorias locales.

La inclusión de variables explicativas en el modelo de regresión permitió mejorar el equilibrio entre clases (victoria, empate y derrota), lo que se reflejó en un F1-score más alto. Esta métrica, que evalúa de forma equitativa el rendimiento en cada clase, resultó ser más informativa que la accuracy en un escenario claramente desbalanceado. Aunque tanto el modelo de regresión como el bivariante mostraron una ligera reducción en accuracy respecto al modelo Poisson doble, esta pérdida fue compensada por un mejor rendimiento global y una mayor capacidad para predecir empates, lo que tuvo un impacto directo en la rentabilidad.

Tal y como se muestra en la Figura 6, los empates presentan una media de cuotas significativamente superior, lo que podría explicarse por la tendencia del mercado a favorecer apuestas hacia un ganador, ajustando las cuotas en función del volumen apostado. Esto genera una posible ineficiencia que puede ser aprovechada por modelos bien calibrados. En este contexto, utilizar métricas como el F1-score permite captar mejor este tipo de oportunidades, mientras que medidas más generales como la accuracy pueden ocultar comportamientos relevantes. En conjunto, los resultados del trabajo ponen de manifiesto que, para construir modelos útiles en el ámbito de las apuestas, es fundamental ir más allá del ajuste estadístico tradicional y evaluar su comportamiento en términos de impacto económico real.

5.4 Propuesta de Trabajos Futuros

Existen varias líneas de trabajo que podrían desarrollarse a partir de este proyecto. Una de ellas podría ser el diseño de un modelo mixto que combine la estructura del modelo Poisson bivariante y su capacidad de detectar dependencia entre los goles de ambos equipos, con la flexibilidad de la regresión de Poisson, que permite incorporar variables explicativas de contexto. Además, sería interesante explorar datasets que incluyan métricas como los *expected*

goals (xG), ya que estudios como el de Spearman (2018) del MIT [9] han mostrado que esta variable es un mejor indicador del rendimiento ofensivo que los goles reales. Otra mejora relevante sería implementar un sistema de validación secuencial, en el que los modelos se reentrenen partido a partido, lo que permitiría adaptarse mejor a las dinámicas recientes de los equipos y simular un entorno de predicción más realista.

5.5 Conclusión Personal

Cuando decidí estudiar el doble grado con Matemáticas, sabía que no quería dedicarme a las matemáticas en el sentido más académico o tradicional. Sin embargo, esta carrera me ha dado algo mucho más valioso: la capacidad de enfrentarme a problemas complejos, analizar situaciones desde una perspectiva lógica y encontrar soluciones incluso cuando no hay un camino evidente. Son precisamente estas habilidades, las que marcan la diferencia entre un ingeniero y un buen ingeniero.

Este TFG me ha permitido poner todo eso en práctica. Elegí un tema que me apasiona y lo abordé desde una perspectiva matemática, investigando conceptos nuevos, planteando preguntas y buscando soluciones. Para mí, ha sido la manera perfecta de cerrar esta etapa universitaria, aplicando lo aprendido a un problema real que conecta teoría y práctica.

Estoy convencida de que esta capacidad de análisis y resolución me va a acompañar a lo largo de mi vida profesional. Y, sobre todo, estoy muy agradecida de haber tenido la oportunidad de desarrollar este proyecto y de haber recorrido este camino académico que, sin duda, volvería a elegir.

6. REFERENCIAS

Todas las referencias bibliográficas de este Trabajo de Fin de Grado quedan registradas en este apartado siguiendo la normativa APA7

6.1 Bibliografía

[1] KPMG Asesores S.L. (2023). Impacto socioeconómico del fútbol profesional en España. KPMG.

[2] Maher, M. J. (1982). Modelling association football scores. Statistica Neerlandica

[3] Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Applied Statistics.

[4] Nguyen, Q. (2021). Poisson Modeling and Predicting English Premier League Goal Scoring. Loyola University Chicago

[5] Aoki, R., Assunção, R., & Vaz de Melo, P. O. S. (2017). Luck is Hard to Beat: The Difficulty of Sports Prediction. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[6] Groll, A., Schauberger, G., & Tutz, G. (2018). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression.

[7] Vlastakis, N., Dotsis, G., & Markellos, R. N. (2009). How efficient is the European football betting market? Evidence from arbitrage and trading strategies. Journal of Forecasting.

[8] Robbins, T. R. (2022). On the Efficiency of Sports Betting Markets. Decision Sciences Institute.

[9] Spearman, W. (2018). Beyond Expected Goals. Sports Analytics Conference. MIT Sloan, Boston MA.

[10] Moroney, M. J. (1956). Facts from figures. Penguin Books.

[11] Loukas, K., Karapiperis, D., Feretzakis, G., & Verykios, V. S. (2024). Predicting football match results using a Poisson regression model. Applied Sciences.

[12] Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. The Statistician

[13] Pollard, R. (1985). *Home advantage in soccer: A retrospective analysis*. Journal of Sports Sciences

[14] Penn, M. J., & Donnelly, C. A. (2022). Analysis of a double Poisson model for Predicting football results in Euro 2020. PLOS ONE.

[15] Müller, M. (2004). Generalized linear models. En J. Gentle, W. Härdle, & Y. Mori (Eds.), Handbook of Computational Statistics (Vol. 1). Springer.

[16] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B.

[17] Kocherlakota, S., & Kocherlakota, K. (1992). Bivariate discrete distributions. Marcel Dekker.

[18] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE.

ANEXOS

En este apartado se adjunta todo el material citado que ocupe demasiado espacio para ser incluido en el cuerpo de este Trabajo de Fin de Grado o que pueda desviar la atención del lector.

Anexo A. Fundamentos Estadísticos

A.1 Variables Aleatorias

Variable aleatoria

Una variable aleatoria es una función que asigna un valor numérico a cada resultado posible de un experimento aleatorio. Puede ser de dos tipos principales: discreta o continua, según el conjunto de valores que pueda tomar.

Variable aleatoria discreta

Una variable aleatoria discreta toma un número finito o numerable de valores. Ejemplos típicos son el número de goles en un partido o el número de tarjetas en un encuentro. Estas variables suelen representarse mediante una función de masa de probabilidad.

Variable aleatoria continua

Una variable aleatoria continua puede tomar infinitos valores dentro de un intervalo. Por ejemplo, la distancia recorrida por un jugador en kilómetros o la posesión del balón en porcentaje. Este tipo de variable se describe mediante una función de densidad de probabilidad.

A.2 Distribuciones de Probabilidad

Distribución de probabilidad

La distribución de probabilidad describe cómo se asignan las probabilidades a los distintos valores posibles que puede tomar una variable aleatoria. En el caso de variables discretas, se representa mediante una función de masa de probabilidad; para variables continuas, se utiliza una función de densidad.

Función de masa de probabilidad (pmf)

Es una función que asocia a cada valor x_i que puede tomar una variable aleatoria discreta la probabilidad $P(X = x_i)$. Debe cumplir que:

$$P(X = x_i) \geq 0 \quad \text{y} \quad \sum_i P(X = x_i) = 1$$

Función de densidad de probabilidad (pdf)

Es la función que describe la distribución de una variable aleatoria continua. La probabilidad de que la variable tome un valor exacto es cero; en cambio, se calcula la probabilidad de que esté dentro de un intervalo. Debe cumplir que:

$$f(x) \geq 0 \quad \text{y} \quad \int_{-\infty}^{\infty} f(x)\ dx = 1$$

Distribución Normal

Es una distribución de probabilidad continua con forma de campana, simétrica respecto a su media. Es especialmente útil cuando se estudia la distribución de fenómenos aleatorios que resultan de la suma de múltiples factores independientes. Un ejemplo es la distribución de las alturas de los estudiantes de una clase, o el tiempo que tarda una persona en llegar al trabajo.

Una variable aleatoria continua X sigue una distribución normal con media μ y desviación estándar σ si su función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in R$$

Sus propiedades son:

$$E[X] = \mu, \quad \text{Var}(X) = \sigma^2$$

Distribución Binomial Negativa

Es una distribución de probabilidad discreta que modela el número de fracasos que se producen antes de alcanzar un número fijo de éxitos en una secuencia de ensayos independientes, cada uno con la misma probabilidad de éxito. A diferencia de la binomial clásica, que fija el número de ensayos, en la binomial negativa se fija el número de éxitos deseados. Si se lanza una moneda

hasta obtener 3 caras, la binomial negativa modela cuántas cruces aparecen antes de lograr esos 3 éxitos, con probabilidad de éxito $p = 0.5$.

Una variable aleatoria X sigue una distribución binomial negativa con parámetros r y p si su función de masa de probabilidad es:

$$P(X = k) = \binom{k + r - 1}{k} (1 - p)^k p^r$$

Sus propiedades son:

$$E[X] = \frac{r(1 - p)}{p}, \quad \text{Var}(X) = \frac{r(1 - p)}{p^2}$$

A.3 Momentos Estadísticos

Esperanza matemática (media esperada)

Es el valor promedio que se espera obtener al repetir un experimento aleatorio un número indeterminado de veces. Representa el centro de gravedad de la distribución de probabilidad.

Esperanza matemática para una variable aleatoria discreta X con valores x_i y probabilidades asociadas $P(X = x_i)$ se define como:

$$E[X] = \sum_i x_i \cdot P(X = x_i)$$

Para una variable aleatoria continua con función de densidad $f(x)$, se define como:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Varianza

La varianza mide la dispersión de los valores de una variable aleatoria respecto a su media esperada. Es útil para evaluar la incertidumbre o variabilidad de un resultado.

La varianza para una variable aleatoria discreta se mide como:

$$\text{Var}(X) = E[(X - E[X])^2] = \sum_i (x_i - E[X])^2 \cdot P(X = x_i)$$

La varianza para una variable aleatoria continua se define como:

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - E[X])^2 \cdot f(x)\, dx$$

Anexo B. Descripción de Variables del Dataset

Este anexo presenta en detalle todas las variables utilizadas a lo largo del trabajo, tanto las provenientes del conjunto de datos original como las generadas posteriormente para el análisis estadístico.

B.1 Dataset Original

Este anexo recoge todas las variables disponibles en el conjunto de datos original obtenido de Football-Data.co.uk. Se agrupan en diferentes categorías según su función y disponibilidad.

1. Información básica del partido: Describen el contexto general del encuentro.

Div: División (por ejemplo, SP1 = Primera División española)

Date: Fecha en la que se disputó el partido

Time: Hora de inicio del encuentro

HomeTeam: Nombre del equipo local

AwayTeam: Nombre del equipo visitante

2. Resultados del partido: Incluye el marcador al descanso y el resultado del partido.

FTHG: Goles del equipo local (Full Time Home Goals)

FTAG: Goles del equipo visitante (Full Time Away Goals)

FTR: Resultado final (H = victoria local, D = empate, A = victoria visitante)

HTHG: Goles del equipo local al descanso (Half Time Home Goals)

HTAG: Goles del equipo visitante al descanso (Half Time Away Goals)

HTR: Resultado al descanso (H, D, A)

3. Estadísticas avanzadas del partido: Contiene métricas adicionales sobre el rendimiento de los equipos. Estas variables no están disponibles en todas las temporadas.

HS, AS: Número de disparos realizados por el equipo local y visitante

HST, AST: Disparos a puerta del equipo local y visitante

HC, AC: Saques de esquina ejecutados por el equipo local y visitante

HY, AY: Tarjetas amarillas recibidas por el equipo local y visitante

HR, AR: Tarjetas rojas recibidas por el equipo local y visitante

4. Cuotas de apuestas – Mercado 1X2: Recogen la valoración de distintos operadores de apuestas sobre los posibles resultados del partido.

B365H, B365D, B365A: Cuotas ofrecidas por Bet365 para victoria local, empate y victoria visitante

PSH, PSD, PSA: Cuotas ofrecidas por Pinnacle

WHH, WHD, WHA: Cuotas ofrecidas por William Hill

AvgH, AvgD, AvgA: Cuotas promedio del mercado para cada resultado

5. Cuotas de apuestas – Mercado Over/Under: Indican la valoración del mercado sobre el número total de goles anotados en el partido.

B365>2.5, B365<2.5: Cuotas de Bet365 para más/menos de 2.5 goles

Max>2.5, Max<2.5: Cuotas máximas del mercado

Avg>2.5, Avg<2.5: Cuotas promedio del mercado

6. Cuotas de apuestas – Hándicap asiático: Cuotas relativas a apuestas con hándicap, una modalidad que ajusta el equilibrio entre los equipos.

B365AHH, B365AHA: Cuotas de Bet365 para hándicap asiático (local y visitante)

BbMxAHH, BbMxAHA: Máximas cuotas del mercado

AvgAHH, AvgAHA: Cuotas promedio del mercado

7. Otras casas de apuestas

Además de los operadores mencionados, el dataset incluye cuotas provenientes de otras casas como Sporting Odds, Ladbrokes, Gamebookers, Stanleybet, entre otras.

B.1 Dataset Ampliado

Este anexo incluye las variables derivadas a partir del conjunto de datos original. Han sido generadas específicamente para su uso en los modelos estadísticos, en particular en el modelo de regresión de Poisson. Todas las variables están definidas desde la perspectiva de un equipo en cada partido.

1. Información general del partido: Describe el contexto del encuentro y las características previas de los equipos implicados.

season: Temporada en la que se disputó el partido.

date: Fecha del partido.

team: Nombre del equipo analizado.

rival_team: Nombre del equipo rival.

home_adv: Indicador binario que señala si el equipo analizado jugó como local (1) o como visitante (0).

last_season_team: Posición final del equipo analizado en la temporada anterior.

last_season_rival: Posición final del equipo rival en la temporada anterior.

2. Rendimiento reciente del equipo analizado: Resumen estadístico del desempeño del equipo en sus últimos 10 partidos de la temporada.

pct_wins: Porcentaje de victorias en los últimos 10 partidos.

avg_goals_scored: Promedio de goles anotados en los últimos 10 partidos.

avg_goals_received: Promedio de goles encajados en los últimos 10 partidos.

goal_difference: Diferencia total de goles en los últimos 10 partidos.

3. Rendimiento reciente del equipo rival: Métricas equivalentes al equipo rival en sus últimos 10 partidos.

pct_wins_rival: Porcentaje de victorias del rival.

avg_goals_scored_rival: Promedio de goles anotados por el rival.

avg_goals_received_rival: Promedio de goles encajados por el rival.

goal_difference_rival: Diferencia de goles del rival.

4. Historial entre ambos equipos: Desempeño del equipo analizado frente al mismo rival en los últimos 5 enfrentamientos directos.

pct_wins_vs_rival: Porcentaje de victorias frente al rival.

avg_goals_scored_vs_rival: Goles promedio anotados al rival.

avg_goals_received_vs_rival: Goles promedio recibidos del rival.

goal_difference_vs_rival: Diferencia total de goles frente al rival.

5. Cuotas de apuestas: Valoración del mercado sobre el partido, transformada en variables desde la perspectiva del equipo analizado.

AvgWin: Cuota promedio para la victoria del equipo.

AvgLoss: Cuota promedio para la derrota del equipo.

AvgDraw: Cuota promedio para el empate.

AvgAHWin: Cuota promedio para la victoria del equipo con hándicap asiático.

AvgAHLoss: Cuota promedio para la derrota del equipo con hándicap asiático.

6. Información del resultado del partido: Resultado real del encuentro codificado para su uso como variable dependiente en los modelos.

goals_team: Número de goles anotados por el equipo analizado.

goals_rival: Número de goles anotados por el rival.

result: Resultado final desde la perspectiva del equipo analizado:

1: Victoria

0: Empate

–1: Derrota

Anexo C. Código del Desarrollo

C.1 Modelo de Poisson Simple

Comparación de distribuciones teóricas entre equipos

Genera y representa las distribuciones de Poisson para distintos equipos locales, comparando visualmente sus medias goleadoras.

```python
teams_to_compare = ["Real Madrid", "Barcelona", "Sevilla", "Ath Madrid", "Valencia"]
colors = ["blue", "green", "red", "orange", "purple"]

for team, color in zip(teams_to_compare, colors):
    lmbda = mean_goals_home[team]
    plt.plot(x, poisson.pmf(x, mu=lmbda), label=team, marker='o')

plt.title("Comparación de distribución de goles entre equipos (locales)")
plt.xlabel("Goles")
plt.ylabel("Probabilidad")
plt.legend()
plt.show()best_params = grid_search.best_params_
```

Cálculo de la log-verosimilitud media por equipo

Calcula el grado de ajuste del modelo estimando la log-verosimilitud media de los goles observados frente a los valores teóricos por equipo y condición (local/visitante).

```python
equipos_frecuentes = ['Ath Bilbao', 'Ath Madrid', 'Barcelona', 'Real Madrid', 'Sevilla', 'Valencia']
# Diccionarios para almacenar resultados
loglik_home = {}
loglik_away = {}
```

```python
for equipo in equipos_frecuentes:
    home_22_23 = df[(df['Season'] == '2022-23') & (df['HomeTeam'] == equipo)]
    home_23_24 = df[(df['Season'] == '2023-24') & (df['HomeTeam'] == equipo)]
    away_22_23 = df[(df['Season'] == '2022-23') & (df['AwayTeam'] == equipo)]
    away_23_24 = df[(df['Season'] == '2023-24') & (df['AwayTeam'] == equipo)]

    if len(home_22_23) == 0 or len(home_23_24) == 0 or len(away_22_23) == 0 or len(away_23_24) == 0:
        continue

    lambda_home = home_22_23['FTHG'].mean()
    lambda_away = away_22_23['FTAG'].mean()

    home_23_24 = home_23_24.copy()
    home_23_24['PoissonProb'] = poisson.pmf(home_23_24['FTHG'], mu=lambda_home)
    loglik_home[equipo] = home_23_24['PoissonProb'].apply(lambda p: log(p) if p > 0 else 0).mean()

    away_23_24 = away_23_24.copy()
    away_23_24['PoissonProb'] = poisson.pmf(away_23_24['FTAG'], mu=lambda_away)
    loglik_away[equipo] = away_23_24['PoissonProb'].apply(lambda p: log(p) if p > 0 else 0).mean()
```

Cálculo de intervalos de confianza al 95 % para λ

Estima los intervalos de confianza del 95 % para los valores medios de goles por equipo, proporcionando una medida de precisión para cada λ.

```python
# Calcular el intervalo de confianza al 95% para la media de goles en casa y fuera de cada equipo
confianza = 0.95
z = norm.ppf(1 - (1 - confianza) / 2)
```

```python
for equipo in equipos_frecuentes:
    # Goles como local
    goles_local = df[(df['HomeTeam'] == equipo) & (df['Season'] == '2022-23')]['FTHG']
    n_local = len(goles_local)
    lambda_local = goles_local.mean()
    se_local = (lambda_local / n_local) ** 0.5
    ic_inf_local = lambda_local - z * se_local
    ic_sup_local = lambda_local + z * se_local
    # Goles como visitante
    goles_visit = df[(df['AwayTeam'] == equipo) & (df['Season'] == '2022-23')]['FTAG']
    n_visit = len(goles_visit)
    lambda_visit = goles_visit.mean()
    se_visit = (lambda_visit / n_visit) ** 0.5
    ic_inf_visit = lambda_visit - z * se_visit
    ic_sup_visit = lambda_visit + z * se_visit
```

C.2 Modelo de Poisson Doble

Función de log-verosimilitud del modelo Poisson doble

Log-verosimilitud total del modelo en función de los parámetros de ataque, defensa y ventaja.

```python
def logverosimilitud(params):
    ataque = params[:n]
    defensa = params[n:2*n]
    gamma = params[-1]
    log_lik = 0
    for _, row in df_train.iterrows():
        i = idx[row['HomeTeam']]
        j = idx[row['AwayTeam']]
        g_local = row['FTHG']
        g_visitante = row['FTAG']
        lambda_ij = np.exp(np.clip(ataque[i] - defensa[j] + gamma, -10, 10))
        mu_ji = np.exp(np.clip(ataque[j] - defensa[i], -10, 10))
        log_lik += poisson.logpmf(g_local, lambda_ij)
        log_lik += poisson.logpmf(g_visitante, mu_ji)
    return -log_lik
```

Optimización de parámetros

Optimiza los parámetros del modelo mediante la minimización de la log-verosimilitud negativa.

```python
res = minimize(
    logverosimilitud,
    x0,
    method='L-BFGS-B',
    bounds=bounds,
    options={'maxiter': 50, 'disp': True}
)
# Extraer resultados
ataque = dict(zip(equipos, res.x[:n]))
defensa = dict(zip(equipos, res.x[n:2*n]))
gamma = res.x[-1]
```

Predicción de resultados y clasificación

Calcula las probabilidades de victoria local, empate y victoria visitante para cada partido a partir de los parámetros estimados.

```python
for _, row in df_test.iterrows():
    ...
    lambda_home = np.exp(ataque[equipo_local] - defensa[equipo_visitante] + gamma)
    mu_away = np.exp(ataque[equipo_visitante] - defensa[equipo_local])

    prob_home = poisson.pmf(range(max_goals), lambda_home)
    prob_away = poisson.pmf(range(max_goals), mu_away)
    matriz = np.outer(prob_home, prob_away)

    p_home_win = np.tril(matriz, -1).sum()
    p_draw = np.trace(matriz)
    p_away_win = np.triu(matriz, 1).sum()
```

Cálculo de log-verosimilitud media

Evalúa el ajuste del modelo sobre los datos reales mediante la log-verosimilitud media por partido.

```python
df_preds['log_prob_real'] = df_preds.apply(
    lambda row: np.log(
        row['P_H'] if row['FTR_real'] == 'H' else
        row['P_D'] if row['FTR_real'] == 'D' else
        row['P_A']
    ), axis=1
)

log_likelihood_total = df_preds['log_prob_real'].sum()
log_likelihood_media = log_likelihood_total / len(df_preds)
```

Cálculo de log-verosimilitud media

Simula una estrategia de apuestas basada en las predicciones del modelo y calcula su rentabilidad final.

```python
df_apuestas['stake'] = df_apuestas['p_pred'].apply(lambda p: 2 if p >= 0.6 else 1)
df_apuestas['ganancia'] = df_apuestas.apply(
    lambda row: row['cuota_usada'] * row['stake'] if row['acierto'] == 1 else 0,
    axis=1
)

total_apostado = df_apuestas['stake'].sum()
total_ganado = df_apuestas['ganancia'].sum()
beneficio_neto = total_ganado - total_apostado
rentabilidad = (beneficio_neto / total_apostado) * 100
```

C.3 Modelo de Poisson Bivariante

Función de verosimilitud bivariante

Define la función de masa de probabilidad bivariante, que combina tres distribuciones de Poisson. Calcula la probabilidad conjunta de que un partido termine con un número concreto de goles del equipo local *x* y visitante *y*, teniendo en cuenta un componente de dependencia λ_3.

```python
def poisson_bivariante(x, y, lambda1, lambda2, lambda3):
    prob = 0.0
    for z in range(0, min(x, y)+1):
        prob += poisson.pmf(z, lambda3) * poisson.pmf(x-z, lambda1) * poisson.pmf(y-z, lambda2)
    return max(prob, 1e-10)
```

Aquí se define la función de log-verosimilitud a maximizar, que combina los parámetros de ataque, defensa, ventaja local (γ) y dependencia (φ), para calcular la probabilidad logarítmica del conjunto de partidos de entrenamiento.

```python
def logverosimilitud_ext(params):
    ataque = params[:n]
    defensa = params[n:2*n]
    gamma = params[2*n]
    phi_home = params[2*n+1:3*n+1]
    phi_away = params[3*n+1:]
    log_lik = 0
    for _, row in df_train.iterrows():
        i = idx[row['HomeTeam']]
        j = idx[row['AwayTeam']]
        g_home = row['FTHG']
        g_away = row['FTAG']
        lambda_home = np.exp(ataque[i] - defensa[j] + gamma)
        lambda_away = np.exp(ataque[j] - defensa[i])
        lambda3 = np.exp(phi_home[i] + phi_away[j])
        lambda1 = max(lambda_home - lambda3, 1e-5)
        lambda2 = max(lambda_away - lambda3, 1e-5)
        p = poisson_bivariante(g_home, g_away, lambda1, lambda2, lambda3)
        log_lik += np.log(p)
    return -log_lik
```

Ajuste del modelo

Este bloque inicializa los parámetros de forma aleatoria y los ajusta utilizando el algoritmo L-BFGS-B, con restricciones en los valores posibles de cada parámetro. Una vez entrenado, se extraen los vectores estimados de ataque, defensa, dependencia y γ.

```python
np.random.seed(42)
x0 = np.concatenate([
    np.random.normal(0, 0.1, n),  # ataque
    np.random.normal(0, 0.1, n),  # defensa
    [0.1],                        # gamma
    np.random.normal(0, 0.1, n),  # phi_home
    np.random.normal(0, 0.1, n)   # phi_away
])
bounds = [(-5, 5)] * (2*n) + [(-1, 1)] + [(-2, 2)] * (2*n)
res = minimize(logverosimilitud_ext, x0, method='L-BFGS-B', bounds=bounds)
ataque = dict(zip(equipos, res.x[:n]))
defensa = dict(zip(equipos, res.x[n:2*n]))
gamma = res.x[2*n]
phi_home = dict(zip(equipos, res.x[2*n+1:3*n+1]))
phi_away = dict(zip(equipos, res.x[3*n+1:]))
```

Generación de predicciones

El siguiente bloque aplica los parámetros del modelo para estimar, para cada partido, la probabilidad de victoria local, empate o visitante. La predicción corresponde al resultado con mayor probabilidad.

```python
max_goals = 6
preds = []
```

```python
for _, row in df_test.iterrows():
    home, away = row['HomeTeam'], row['AwayTeam']
    if home not in ataque or away not in ataque:
        continue
    lambda_home = np.exp(ataque[home] - defensa[away] + gamma)
    lambda_away = np.exp(ataque[away] - defensa[home])
    lambda3 = np.exp(phi_home[home] + phi_away[away])
    lambda1 = max(lambda_home - lambda3, 1e-5)
    lambda2 = max(lambda_away - lambda3, 1e-5)

    matriz = np.zeros((max_goals, max_goals))
    for i in range(max_goals):
        for j in range(max_goals):
            matriz[i, j] = poisson_bivariante(i, j, lambda1, lambda2, lambda3)

    p_home = np.tril(matriz, -1).sum()
    p_draw = np.trace(matriz)
    p_away = np.triu(matriz, 1).sum()
    pred_label = ['H', 'D', 'A'][np.argmax([p_home, p_draw, p_away])]

    preds.append({
        'HomeTeam': home,
        'AwayTeam': away,
        'FTR_real': row['FTR'],
        'Pred': pred_label
    })

df_preds = pd.DataFrame(preds)
```

C.4 Regresión de Poisson

Preparación de datos y codificación de variables

Se seleccionan las temporadas de entrenamiento, se codifican los equipos con variables binarias y se extraen variables estadísticas del equipo y su rival.

```python
# Cargar dataset y preparar datos
df = pd.read_csv("../datasets/dataset_transformado.csv")
df['date'] = pd.to_datetime(df['date'])
df['season'] = df['season'].astype(str)

# Dividir entre entrenamiento y test
temporadas_entrenamiento = ['2019-20', '2020-21', '2021-22', '2022-23']
df_train = df[df['season'].isin(temporadas_entrenamiento)]
df_test = df[df['season'] == '2023-24']

# Codificación one-hot para equipos
equipos = sorted(pd.unique(df[['team', 'rival_team']].values.ravel()))
encoder = OneHotEncoder(categories=[equipos], drop=None, sparse_output=False, handle_unknown='ignore')
X_team = encoder.fit_transform(df_train[['team']])
X_rival = encoder.transform(df_train[['rival_team']].rename(columns={'rival_team': 'team'}))

# Selección de variables numéricas
features_numericas = [
    'home_adv',
    'pct_wins', 'avg_goals_scored', 'avg_goals_received', 'goal_difference',
    'pct_wins_rival', 'avg_goals_scored_rival', 'avg_goals_received_rival', 'goal_difference_rival',
    'pct_wins_vs_rival',           'avg_goals_scored_vs_rival',           'avg_goals_received_vs_rival',
'goal_difference_vs_rival',
    'AvgH', 'AvgD', 'AvgA'
]
X_features = df_train[features_numericas].values

# Matriz final de entrenamiento
X_train = np.hstack([X_team, X_rival, X_features])
y_train_home = df_train['goals_team']
y_train_away = df_train['goals_rival']
```

Entrenamiento del modelo

Se ajustan dos regresiones de Poisson, una para los goles del equipo local y otra para el visitante, usando PoissonRegressor.

```python
from sklearn.linear_model import PoissonRegressor

# Ajuste del modelo para goles locales y visitantes

model_home = PoissonRegressor(alpha=1e-12, max_iter=1000)

model_away = PoissonRegressor(alpha=1e-12, max_iter=1000)

model_home.fit(X_train, y_train_home)

model_away.fit(X_train, y_train_away)
```

Predicción y construcción de probabilidades

Se comparan las predicciones con los resultados reales mediante métricas de clasificación.

```python
# Preparación de los datos de test

X_team_test = encoder.transform(df_test[['team']])

X_rival_test = encoder.transform(df_test[['rival_team']].rename(columns={'rival_team': 'team'}))

X_features_test = df_test[features_numericas].values

X_test = np.hstack([X_team_test, X_rival_test, X_features_test])

# Predicción de goles esperados

lambda_home = model_home.predict(X_test)

mu_away = model_away.predict(X_test)
```

Generación de resultados y evaluación

Se simula una estrategia de apuestas basada en las predicciones y se calcula la rentabilidad obtenida.

```python
from scipy.stats import poisson

max_goals = 10

factor_calibracion = 0.97

umbral_empate = 0.29

# Predicciones

predicciones = []
```

```python
for idx, row in enumerate(df_test.itertuples(index=False)):
    lambda_cal = lambda_home[idx] * factor_calibracion
    mu_cal = mu_away[idx] * factor_calibracion

    matriz = np.outer(
        poisson.pmf(range(max_goals), lambda_cal),
        poisson.pmf(range(max_goals), mu_cal)
    )

    p_home_win = np.tril(matriz, -1).sum()
    p_draw = np.trace(matriz)
    p_away_win = np.triu(matriz, 1).sum()

    if p_draw >= umbral_empate:
        pred_label = 'D'
    else:
        pred_label = 'H' if p_home_win > p_away_win else 'A'

    predicciones.append({
        'team': row.team,
        'rival_team': row.rival_team,
        'result_real': row.result,
        'P_H': p_home_win,
        'P_D': p_draw,
        'P_A': p_away_win,
        'Pred': pred_label
    })

df_preds_poisson_reg = pd.DataFrame(predicciones)
```