****

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

ÍNDICE GENERAL

ÍNDICE DE FIGURAS 4

ÍNDICE DE TABLAS 5

RESUMEN 6

ABSTRACT 6

1. INTRODUCCIÓN 7

1.1 Motivación y contexto 7

1.2 Planteamiento del problema 9

1.2.1 Incertidumbre en los Resultados Deportivos 9

1.2.2 Limitaciones de los Modelos Matemáticos y Estadísticos 10

1.2.3 Valoración del Éxito de los Equipos 10

1.3 Objetivos del Trabajo 11

2. ESTADO DE LA CUESTIÓN 13

2.1 El Mercado de las Apuestas Deportivas 13

2.1.1 Funcionamiento de las Cuotas y el Mercado de Apuestas 13

2.1.2 Impacto Económico y Social 15

2.2 El Futbol y las Apuestas 15

2.2.1 Tipos de Apuestas en el Futbol 16

2.2.2 Impacto Económico de las Apuestas en el Fútbol 16

2.3 Marco teórico del trabajo 17

2.3.1 Modelos Probabilísticos en el Deporte 17

2.3.2 Modelos de Conteo 18

2.3.3 Distribución de Poisson: Definición y Propiedades 19

2.3.4 Modelo de Poisson Doble en el Fútbol 20

2.3.5 Estimación del Número Medio de Goles 23

2.3.6 Modelo de Poisson Bivariante 23

2.3.7 Modelos Lineales Generalizados 25

2.3.8 Regresión de Poisson 26

2.3.9 Supuestos y Limitaciones 28

2.4 Trabajos Relacionados 28

3. ASPECTOS METODOLÓGICOS 31

3.1 Metodología 31

3.1.1 Planificación 31

3.2 Tecnologías Empleadas 32

3.2.1 Seguimiento del Trabajo 32

3.2.2 Lenguajes y Entornos de Desarrollo 33

3.2.3 Librerías y Frameworks 33

4. DESARROLLO DEL TRABAJO 34

4.1 Adquisición, Análisis y Procesamiento de Datos 34

4.1.1 Descripción y Preproceso de los Datos 35

4.1.2 Análisis Exploratorio de los Datos (EDA) 38

4.2 Desarrollo de los Modelos 42

5. CONCLUSIONES 43

6. REFERENCIAS 44

6.1 Bibliografía 44

ANEXOS 46

Anexo A. Fundamentos Estadísticos 46

A.1 Variables Aleatorias 46

A.2 Distribuciones de Probabilidad 46

A.3 Momentos Estadísticos 48

Anexo B. Descripción de Variables del Dataset 49

B.1 Dataset Original 49

B.1 Dataset Ampliado 51

# ÍNDICE DE FIGURAS

[Figura 1 - Distribuciones de Poisson para distintos valores de λ 19](#_Toc200394237)

[Figura 2. Proporción de victorias locales, empates y victorias visitantes en el dataset 39](#_Toc200394238)

[Figura 2. Proporción de victorias locales, empates y victorias visitantes en el dataset 39](#_Toc200394239)

[Figura 3. Frecuencia conjunta de marcadores (Goles Local vs Visitante) 40](#_Toc200394240)

[Figura 4. Frecuencia conjunta de marcadores con restricción de 0 a 6 goles por equipo 40](#_Toc200394241)

[Figura 5. Distribución de cuotas para los tres posibles resultados del mercado 1X2 41](#_Toc200394242)

# ÍNDICE DE TABLAS

[Tabla 1 – Dataset básico de partidos desde la temporada 03/04 hasta 23/24 36](#_Toc200394243)

[Tabla 2 - Dataset aumentado de partidos (columnas 1-7) 37](#_Toc200394244)

[Tabla 3 - Dataset aumentado de partidos (columnas 8-13) 37](#_Toc200394245)

[Tabla 4 - Dataset aumentado de partidos (columnas 14-17) 37](#_Toc200394246)

[Tabla 5 - Dataset aumentado de partidos (columnas 18-25) 37](#_Toc200394247)

# RESUMEN

Breve resumen del TFG en español. Se recomienda describir en pocas palabras, no más de dos párrafos, la temática, el trabajo desarrollado y la conclusión.

# ABSTRACT

Brief summary of the TFG in English. It is recommended to describe in a few words, no more than two paragraphs, the topic, the work developed and the conclusion.

# 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]](#footnote-1). En España, según el informe de KPMG (2023) [[1](#_6.1_Bibliografía)], 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]](#footnote-2). Además, se espera que el mercado mundial de apuestas deportivas siga creciendo a un ritmo cercano al 9% para los próximos años[[3]](#footnote-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 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](#_6.1_Bibliografía)] y de Dixon y Coles (1997) [[3](#_6.1_Bibliografía)], hasta investigaciones más recientes como el de Nguyen (2021) [[4](#_6.1_Bibliografía)], 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 significativo 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](#_6.1_Bibliografía)], 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 ligas como LaLiga 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](#_6.1_Bibliografía)], 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 está 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](#_6.1_Bibliografía)], este problema puede reducir drásticamente la capacidad predictiva si no se aplican técnicas de validación adecuadas. El sobreajuste es 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]](#footnote-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](#_6.1_Bibliografía)] 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 un modelo estadístico basado en la distribución de Poisson para realizar predicciones de resultados de fútbol, y analizar tanto sus capacidades como sus limitaciones. Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

* Investigar distintos enfoques analíticos aplicados al modelado de goles en fútbol.
* Calcular, mediante el modelo, las probabilidades de los tres resultados posibles en un partido: victoria local, empate y victoria visitante.
* Evaluar variantes del modelo clásico de Poisson para determinar si alguna mejora la precisión de las predicciones.
* Analizar críticamente los supuestos y la metodología del modelo utilizando datos de la liga española.
* Incorporar variables explicativas adicionales que puedan enriquecer el modelo basado en la distribución de Poisson.
* Estudiar la robustez del modelo bivariante ante cambios en las condiciones o características de los partidos.
* Examinar las limitaciones del modelo y su efectividad en contextos de alta incertidumbre, como el mercado de apuestas deportivas.

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 El Mercado de las Apuestas Deportivas

A lo largo de los años, as 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.[[5]](#footnote-5) 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.

2.1.1 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](#_6.1_Bibliografía)], este margen promedio suele estar entre el 3 % y el 5 %, y constituye una ventaja estructural para los operadores.

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](#_6.1_Bibliografía)] 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.

2.1.2 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.[[6]](#footnote-6)

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.[[7]](#footnote-7) 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.[[8]](#footnote-8) 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.

## 2.2 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.[[9]](#footnote-9) 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

2.2.1 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 apuestaconsiste 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.

2.2.2 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 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](#_6.1_Bibliografía)]. 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.[[10]](#footnote-10)

## 2.3 Marco teórico del trabajo

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.

2.3.1 Modelos Probabilísticos en el Deporte

El uso de distribuciones de probabilidad para modelar resultados deportivos tiene una larga trayectoria. En fútbol, Moroney (1956) [[10](#_6.1_Bibliografía)] 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 es 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](#_6.1_Bibliografía)] introdujo un modelo doble de Poisson que considera las características ofensivas y defensivas de los equipos. En 1997, Dixon y Coles [[3](#_6.1_Bibliografía)] refinaron este modelo para mejorar la predicción de empates. Más adelante, Karlis y Ntzoufras (2003) [[12](#_6.1_Bibliografía)] propusieron un modelo bivariante que tiene en cuenta la dependencia entre los goles de ambos equipos. Recientemente, Loukas et al. (2024) [[11](#_6.1_Bibliografía)] incorporaron 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](#_2.2_Trabajos_Relacionados), donde se analizará su impacto metodológico y sus aplicaciones en el modelado estadístico de resultados de fútbol.

2.3.2 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.

2.3.3 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 sigue una distribución de Poisson con parámetro si su función de masa de probabilidad viene dada por:

El parámetro representa la tasa media de ocurrencia de eventos en el intervalo considerado, es decir, 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](#_6.2_Índice_de) se muestran representaciones típicas de la distribución de Poisson para diferentes valores del parámetro λ.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

[Figura 1](#_6.2_Índice_de) - Distribuciones de Poisson para distintos valores de λ

El eje horizontal (x) representa los posibles valores que puede tomar la variable aleatoria . 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 ), la distribución de Poisson puede aproximarse mediante una distribución normal con media y varianza .

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.

2.3.4 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](#_6.1_Bibliografía)], 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:

donde e son el número de goles del equipo *i*  (local) y del *j*  (visitante), respectivamente, y y 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:

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

Los parámetros se estiman por el **método de máxima verosimilitud**, a partir de los datos históricos de partidos. Este método, conocido como EMV por sus siglas en español (*Maximum Likelihood Estimation*, MLE, en inglés), es una técnica fundamental en estadística para estimar los parámetros de un modelo probabilístico. Su objetivo es encontrar los valores de los parámetros que hacen más verosímil (probable) la observación del conjunto de datos disponibles bajo el modelo considerado.

Si es la función de probabilidad del modelo para una observación con parámetro , el estimador se obtiene como:

En el contexto de la distribución de Poisson, el EMV se utiliza para estimar el parámetro λ, que representa la media (y varianza) del número de goles. En modelos más avanzados como el Poisson doble o el bivariante, el método se emplea para ajustar múltiples parámetros como las fuerzas ofensivas y defensivas de los equipos, optimizando su valor para que el modelo se ajuste lo mejor posible a los datos históricos.

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.

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

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 . 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 estar 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](#_6.1_Bibliografía)] introdujeron una serie de ajustes. En primer lugar, propusieron un ajuste específico para resultados bajos, ajustando la probabilidad conjunta 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 y el visitante 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:

Con esta información, es posible construir una **matriz de probabilidades** conjuntas 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
* Empate: todos los casos en los que
* Victoria visitante: todos los casos en los que

Formalmente, se puede expresar como:

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%.

2.3.5 Estimación del Número Medio de Goles

2.3.6 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 , que representan recuentos, estén correlacionadas. Para ello, se define a partir de tres variables aleatorias independientes:

De forma que:

La variable  actúa como un componente común que introduce dependencia entre , de modo que cuando se obtiene una correlación positiva entre ambas. En cambio, si , se recupera el caso independiente, equivalente al modelo de Poisson doble. Una ventaja fundamental de esta construcción es que las distribuciones marginales de siguen siendo de Poisson, con medias , respectivamente, mientras que la covarianza entre ambas variables es precisamente .

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 que refleja la influencia del término común.

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 , que refleja el efecto de un componente común que afecta simultáneamente a ambos equipos.

Este modelo fue aplicado al fútbol por Karlis y Ntzoufras (2003) [[12](#_6.1_Bibliografía)], 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 ​ 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:

Esta es la formulación introducida por Kocherlakota en 1992 [[15](#_6.1_Bibliografía)] y utilizada en modelos como el de Karlis y Ntzoufras (2003) [[12](#_6.1_Bibliografía)]. El sumatorio recorre todos los posibles valores del componente común , introduciendo una dependencia explícita entre los goles de ambos equipos. El parámetro ​ actúa como fuente de correlación. Si , se recupera el caso de independencia.

2.3.7 Modelos Lineales Generalizados

Los Modelos Lineales Generalizados (GLM, por sus siglas en inglés *Generalized Linear Models*) [[15](#_6.1_Bibliografía)] 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 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:

Donde es el parámetro canónico, es un parámetro de dispersión, y son las funciones específicas de cada distribución.

**2. Función de enlace**

Se utiliza una función de enlace que conecta la media de la variable dependiente con una combinación lineal de los predictores:

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 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.

2.3.8 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 se relaciona con las variables explicativas mediante una función de enlace logarítmica:

donde es el número de goles esperados por un equipo en el partido, representa la media esperada de goles, son las variables explicativas, y son los coeficientes del modelo.

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](#_6.1_Bibliografía)], donde los goles esperados se estimaban solo con parámetros fijos, estudios más recientes como el de Loukas et al. (2024) [[11](#_6.1_Bibliografía)] 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 precisión 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.

2.3.9 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](#_6.1_Bibliografía)] 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](#_6.1_Bibliografía)], 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.

## 2.4 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.

Maher (1982) [[2](#_6.1_Bibliografía)] 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 χ², 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).

En el trabajo de Dixon & Coles (1997) [[3](#_6.1_Bibliografía)], 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 %.

Karlis y Ntzoufras (2003) [[12](#_6.1_Bibliografía)] 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.

Penn y Donnelly (2022) [[14](#_6.1_Bibliografía)] 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.

Finalmente, Loukas et al. (2024) [[11](#_6.1_Bibliografía)] 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]](#footnote-11), que ofrece información histórica 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](#_4._DESARROLLO_DEL), mientras que la fase 5 será abordada en el apartado [5. Conclusiones](#_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.

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](https://www.football-data.co.uk/spainm.php), 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](https://github.com/anaigs/tfg_maco_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](#_Anexo_B._Descripción). 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](https://www.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.

Tabla

El contenido generado por IA puede ser incorrecto.

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.](#_Anexo_B._Descripción) A continuación, se presenta una vista preliminar del dataset ampliado.

Tabla

El contenido generado por IA puede ser incorrecto.

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

Tabla

El contenido generado por IA puede ser incorrecto.

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

Tabla

El contenido generado por IA puede ser incorrecto.

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

Tabla

El contenido generado por IA puede ser incorrecto.

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.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Figura 2. Proporción de victorias locales, empates y victorias visitantes en el dataset

Figura 2. 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.

Tabla

El contenido generado por IA puede ser incorrecto.

Figura 3. Frecuencia conjunta de marcadores (Goles Local vs Visitante)

Calendario

El contenido generado por IA puede ser incorrecto.

Figura 4. 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.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Figura 5. 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

# 5. CONCLUSIONES

# 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.

# 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 que puede tomar una variable aleatoria discreta la probabilidad . Debe cumplir que:

**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:

**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 sigue una distribución normal con media μ y desviación estándar σ si su función de densidad de probabilidad es:

Sus propiedades son:

**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 .

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

Sus propiedades son:

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 con valores y probabilidades asociadas se define como:

Para una variable aleatoria continua con función de densidad , se define como:

**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:

La varianza para una variable aleatoria continua se define como:

## 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](https://www.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

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/ [↑](#footnote-ref-1)
2. El País. (2023, 28 septiembre). El fútbol profesional duplica su peso en el PIB español en una década. https://elpais.com/ [↑](#footnote-ref-2)
3. Statista. (2024, 12 marzo). Valor de mercado de las apuestas deportivas en el mundo. https://es.statista.com/ [↑](#footnote-ref-3)
4. ESPN staff (2014, 11 Junio). Soccer Power Index explained. https://www.espn.com/ [↑](#footnote-ref-4)
5. La evolución histórica de las apuestas deportivas: de los juegos antiguos a la era digital. (2023, 21 junio). Fox Sports. https://www.foxsports.com. [↑](#footnote-ref-5)
6. Fernández, R. (2025, 13 febrero). Las apuestas y los juegos de azar en el mundo: Datos estadísticos. https://es.statista.com/ [↑](#footnote-ref-6)
7. 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 [↑](#footnote-ref-7)
8. Zuppello, M. (2024, 30 septiembre). Adicción a las apuestas online en Brasil. Infobae. https://www.infobae.com [↑](#footnote-ref-8)
9. (2024, abril) El Pilón. Los favoritos de los aficionados: Los deportes más populares para apostar. https://elpilon.com.co [↑](#footnote-ref-9)
10. 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 [↑](#footnote-ref-10)
11. Football-Data.co.uk. (2024). Spain football data files. https://www.footballdata.co.uk/spainm.php [↑](#footnote-ref-11)