# Anteproyecto
# Predicción de Cuotas Deportivas mediante Aprendizaje Automático para Maximizar la Rentabilidad de la Casa de Apuestas

Mariana Valencia Cubillos
Leonard David Vivas Dallos
Tomás Escobar Rivera

June 5, 2025

## Contents

# 1 Pregunta de investigación y objetivos

## 1.1 Pregunta de investigación

¿Es posible entrenar un modelo de aprendizaje automático que prediga con precisión las cuotas de apuestas deportivas (victoria local, empate o visitante) a partir de resultados históricos y cuotas previas, ajustándolas con un margen que asegure la rentabilidad para la casa de apuestas?

Esta pregunta parte del hecho de que las cuotas reflejan estimaciones de probabilidad ajustadas para generar ganancia. El objetivo es replicar y automatizar este proceso de forma precisa y técnicamente robusta usando aprendizaje automático.

## 1.2 Objetivo general

Desarrollar un sistema con aprendizaje automático supervisado que prediga cuotas deportivas entre dos equipos, incorporando un margen de ganancia para asegurar la rentabilidad de la casa de apuestas, utilizando herramientas de Scikit-learn.

## 1.3 Objetivos específicos

1. Explorar, limpiar y transformar los datos históricos de resultados y cuotas de apuestas deportivas, aplicando técnicas de preprocesamiento compatibles con modelos de aprendizaje automático.
2. Diseñar y entrenar múltiples modelos supervisados (regresión lineal, árboles de decisión, métodos de ensamble) para predecir las cuotas de victoria local, empate y victoria visitante.
3. Evaluar y comparar el desempeño de los modelos utilizando métricas como MAE y RMSE, seleccionando el modelo con mejor capacidad predictiva.
4. Aplicar calibración y ajuste de márgenes (overround) sobre las probabilidades estimadas, generando cuotas competitivas y rentables desde la perspectiva de una casa de apuestas.
5. Simular y analizar el comportamiento del sistema en escenarios hipotéticos de apuestas para validar su viabilidad práctica.
6. Documentar el proceso usando herramientas como Scikit-learn y GitHub, resaltando las decisiones de diseño, los resultados obtenidos y las limitaciones del enfoque.

# 2 Metodología de investigación

Este proyecto aborda la predicción de cuotas de apuestas deportivas como un problema de regresión multisalida, en el que se busca estimar tres valores continuos correspondientes a las cuotas de victoria local, empate y victoria visitante. Para ello, se utilizarán algoritmos de aprendizaje automático supervisado implementados en la biblioteca Scikit-learn, aprovechando su robustez y facilidad de uso. El conjunto de entrenamiento estará compuesto por datos históricos que incluyen información del partido como equipos involucrados, goles anteriores, condición de localía y cuotas previas ofrecidas por casas de apuestas.

El proyecto se desarrollará en Python usando pandas y numpy para procesar datos, matplotlib y seaborn para visualización, y scikit-learn para entrenar y evaluar modelos. Se plantean tres fases: primero, entrenamiento de modelos base sin márgenes; luego, ajuste de overround para asegurar rentabilidad; y por último, simulación de escenarios financieros. Si es necesario, se usará XGBoost o LightGBM para mejorar el rendimiento.

## 2.1 Modelos propuestos

Por ahora los modelos candidatos a implementar y comparar en el proyecto incluyen:

- **Regresión lineal multisalida:** como modelo base para establecer un punto de referencia.

- **Árboles de decisión y Random Forest:** por su capacidad para capturar relaciones no lineales entre características.
- **Modelos de ensamble como Gradient Boosting y/o XGBoost:** debido a su alta capacidad predictiva.
- **Redes neuronales simples:** si el tiempo lo permite, se incluirá un modelo con MLP (Multi-Layer Perceptron) de scikit-learn.

Cada modelo será evaluado mediante validación cruzada (k-fold) y comparado según métricas de regresión para determinar el mejor candidato.

# 3 Datos y análisis previo

Este proyecto se basa en un conjunto de datos históricos de la Premier League que incluye resultados de partidos y cuotas ofrecidas por casas de apuestas. A partir de este, se realizará un análisis exploratorio y se extraerán variables relevantes para entrenar modelos de aprendizaje automático que permitan predecir las cuotas con base en las características del partido.

## 3.1 Fuente de datos

La base de datos utilizada en este proyecto proviene del portal Kaggle y contiene resultados históricos de la Premier League inglesa (EPL) entre 1993 y 2023. El archivo principal incluye información de los partidos como equipos involucrados, goles anotados, resultados finales y cuotas ofrecidas por distintas casas de apuestas (por ejemplo, Bet365, William Hill, Pinnacle, entre otras).

## 3.2 Variables

Las variables independientes consideradas incluyen información del partido como los equipos locales y visitantes (HomeTeam, AwayTeam), goles anotados (FTHG, FTAG), resultados previos, promedio de goles recientes, rachas de partidos y la condición de localía. También se podrán incluir cuotas históricas de encuentros similares. Estas variables serán transformadas según corresponda: las categóricas serán codificadas numéricamente y las numéricas podrán ser normalizadas o estandarizadas de acuerdo con el modelo utilizado.

La variable a predecir es un vector de tres valores continuos correspondientes a las cuotas ofrecidas por Bet365:

- B365H: cuota para la victoria del equipo local.
- B365D: cuota para el empate.
- B365A: cuota para la victoria del equipo visitante.

Dado que estas cuotas representan valores derivados de probabilidades implícitas, se analizarán también como probabilidades para realizar ajustes posteriores mediante overround.

# 4 Plan detallado

A continuación se describe el cronograma del proyecto, organizado por semanas y detallando las tareas principales en cada fase.

- **Semana 1–2: Exploración y preparación de datos.** Revisión del dataset, limpieza (valores nulos, formatos, duplicados), análisis exploratorio de cuotas y resultados, y construcción de variables como rachas y promedio de goles.
- **Semana 3: Entrenamiento de modelos base.** Implementación de regresión lineal multisalida y árboles de decisión. Evaluación preliminar con métricas como MAE y RMSE.

- **Semana 4: Modelos avanzados y validación.** Entrenamiento de modelos como Random Forest y XGBoost, ajuste de hiperparámetros (grid search) y validación cruzada (k-fold).
- **Semana 5: Cálculo de cuotas ajustadas (overround).** Conversión de predicciones a probabilidades implícitas, aplicación del margen de ganancia y validación de la coherencia de las cuotas.
- **Semana 6: Evaluación y simulación.** Simulación de escenarios de apuestas, cálculo de rentabilidad esperada y análisis de sensibilidad ante variaciones de entrada.
- **Semana 7: Documentación y entrega final.** Redacción del informe final, elaboración de visualizaciones, revisión y entrega del documento.

# 5 Implicaciones éticas

Este proyecto, aunque de carácter académico, aborda un tema con implicaciones éticas relevantes, dado que se enfoca en el sector de las apuestas deportivas. La automatización en la predicción de cuotas puede reforzar la asimetría de información entre casas de apuestas y usuarios, generando ventajas desproporcionadas para las primeras.

Además, se tendrá especial cuidado en evitar sesgos en los datos y en documentar claramente las limitaciones del sistema, enfatizando que los resultados solo deben interpretarse dentro del contexto de análisis técnico y no como una herramienta de predicción infalible o financiera.

# 6 Aspectos legales y comerciales

El proyecto presenta un alto potencial comercial si se adapta para ser utilizado en plataformas de casas de apuestas, ya que la automatización en la generación de cuotas, con un componente de rentabilidad incorporado, puede ofrecer ventajas significativas para los operadores del sector.

Sin embargo, su aplicación práctica en Colombia estaría sujeta a la regulación de entidades como Coljuegos, que supervisan el funcionamiento del sector de juegos de suerte y azar. Aunque este trabajo es netamente académico, si en el futuro se considera su implementación comercial, será necesario revisar el marco legal vigente, realizar un análisis de impacto y garantizar el uso responsable del modelo.

# References


[1] Galekwa, R. M., Tshimula, J. M., Tajeuna, E. G., & Kyamakya, K. (2024). *A Systematic Review of Machine Learning in Sports Betting: Techniques, Challenges, and Future Directions*. Disponible en: `https://arxiv.org/abs/2410.21484`

[2] Mandadapu, P. (2024). *The Evolution of Football Betting: A Machine Learning Approach to Match Outcome Forecasting and Bookmaker Odds Estimation*. Disponible en: `https://arxiv.org/abs/2403.16282`

[3] Hubáček, O., Šír, G., & Železný, F. (2019). *Exploiting Sports-Betting Market Using Machine Learning*. Disponible en: `https://www.researchgate.net/publication/331218530_Exploiting_sports-betting_market_using_machine_learning`

[4] Ren, Y., & Susnjak, T. (2022). *Predicting Football Match Outcomes with Explainable Machine Learning and the Kelly Index*. Disponible en: `https://arxiv.org/abs/2211.15734`

[5] Egidi, L., Pauli, F., & Torelli, N. (2018). *Combining Historical Data and Bookmakers' Odds in Modelling Football Scores*. Disponible en: `https://arxiv.org/abs/1802.08848`

[6] Chen, L. (2023). *Football Results and Betting Odds Data of EPL*. Dataset disponible en Kaggle: `https://www.kaggle.com/datasets/louischen7/football-results-and-betting-odds-data-of-epl`