Informe final

Proyecto de Ciencia de Datos

Presentado por:
Esteban Camargo Cañas
Ana Sofia Villada Ospina

Programa:
Ingeniería mecatrónica, Universidad EIA, Envigado

Fecha:
Junio,2025

1. Introducción

El presente informe documenta el desarrollo de un proyecto de ciencia de datos enfocado en la predicción de resultados de apuestas deportivas. Este trabajo se enmarca dentro del curso de Programación 2025-1, cumpliendo con todos los requerimientos especificados: formulación de un problema de clasificación, tratamiento de datos reales, desarrollo y comparación de modelos predictivos.

El objetivo principal del proyecto es construir un modelo de clasificación capaz de predecir si una apuesta será ganadora (is_win) a partir de información tomada en Kaggle del evento deportivo, el tipo de apuesta, las cuotas, el monto apostado, el deporte, entre otros atributos.

2. Definición del problema

El problema abordado corresponde a una tarea de clasificación binaria, donde se busca predecir si una apuesta tendrá un resultado ganador (is_win = 1) o no (is_win = 0). Este problema es altamente relevante en la industria de las apuestas deportivas, ya que permite modelar y analizar el riesgo basado en datos históricos.

3. Descripción del Dataset

- Fuente: Kaggle (https://www.kaggle.com/)
- Nombre del archivo: bets.csv
- Instancias: 100,000
- Atributos: 15
- Variable objetivo: is_win (booleano)
- Datos faltantes: La columna sport_category presentaba un 4.8% de valores faltantes, los cuales fueron imputados.

Los atributos incluyen características del evento (tipo de deporte, tipo de apuesta), del usuario (monto apostado, ganancia) y del contexto (cuota, rentabilidad estimada).

4. Análisis Exploratorio

Se desarrollaron los siguientes pasos para comprender la estructura del dataset:

- Estadísticas descriptivas para variables numéricas (stake, odds, gain, etc.).
- Histogramas para estudiar distribuciones individuales.
- Mapa de calor para estudiar correlaciones entre variables numéricas.
- Diagramas de dispersión para explorar relaciones entre odds, stake, gain y profitability.
- Detección de valores extremos, los cuales se mantuvieron dado el dominio de aplicación (apuestas pueden tener valores altos).

Este análisis permitió identificar patrones de comportamiento y variables potencialmente predictivas.

5. Preprocesamiento

Se llevaron a cabo las siguientes etapas:

- Eliminación de duplicados.
- Imputación de valores faltantes en columnas categóricas con la moda.
- Codificación de variables categóricas usando LabelEncoder.
- Estandarización de variables numéricas con StandardScaler.

- División del dataset en entrenamiento (80%) y prueba (20%).

6. Modelos Utilizados

Se seleccionaron dos algoritmos de clasificación altamente efectivos:

Modelo 1: Random Forest

- Se usó GridSearchCV para buscar hiperparámetros óptimos.
- Se evaluó con matriz de confusión, precisión, recall, F1-score.
- Se analizó su curva de aprendizaje para detectar overfitting.

Modelo 2: XGBoost

- Optimizaciones con GridSearchCV.
- Evaluación con las mismas métricas.
- Curva de aprendizaje incluida para comparación.

7. Evaluación y Comparación

Modelo	Precisión	Recall	F1-score
Random Forest	1.0000	1.0000	1.0000
XGBoost	0.999938	0.9999863	0.999914

Ambos modelos mostraron excelente desempeño. Aunque las métricas fueron cercanas, **XGBoost** obtuvo un F1-score ligeramente superior, lo que sugiere un mejor equilibrio entre precisión y recall.

Se utilizó además un análisis de concordancia para verificar la consistencia en las predicciones entre ambos modelos.

8. Diagnóstico y Recomendaciones

Ambos modelos presentan indicios leves de overfitting, especialmente Random Forest. Como pasos futuros se recomienda:

- Aumentar la complejidad del dataset con atributos derivados (por ejemplo, interacciones entre stake y odds).
- Aplicar técnicas de validación cruzada más robustas (como K-fold con estratificación).
- Usar SHAP o LIME para interpretar la importancia de variables en XGBoost.
- Evaluar el desempeño en un entorno de producción con datos en tiempo real.

9. Conclusiones

- El dataset es altamente adecuado para problemas de clasificación binaria.
- El proceso de limpieza y preprocesamiento fue esencial para obtener buenos resultados.
- Ambos modelos fueron efectivos, pero se recomienda **XGBoost** como modelo final por su desempeño y capacidad de generalización.
- Este trabajo demuestra la utilidad de la ciencia de datos en contextos aplicados como las apuestas deportivas.