Proyecto Análisis de Ingresos en Retail

Introducción:

En este bootcamp de Data Science, una de las formas de evaluar el aprendizaje fue mediante el desarrollo de actividades denominadas CORE. Estas actividades incrementaban su dificultad a medida que avanzaba el curso. En esta ocasión, el CORE consiste en aplicar los conocimientos adquiridos de manera progresiva, culminando en el desarrollo del archivo presente en este repositorio: Core_Clasificacion_Basada_en_arboles.ipynb.

Objetivo General:

Predecir ingresos usando Python y ML en una tienda de retail.

Objetivo específicos

- Aplicar análisis exploratorio de datos (EDA) para comprender la estructura y calidad del dataset.
- Interpretar las variables disponibles para evaluar su relevancia en la predicción del ingreso.
- Detectar y tratar valores atípicos y datos inconsistentes que puedan afectar el rendimiento del modelo.
- Desarrollar y evaluar modelos de Machine Learning que permitan predecir con precisión los ingresos generados, aportando así herramientas para la toma de decisiones estratégicas.

Descripción del Conjunto de Datos:

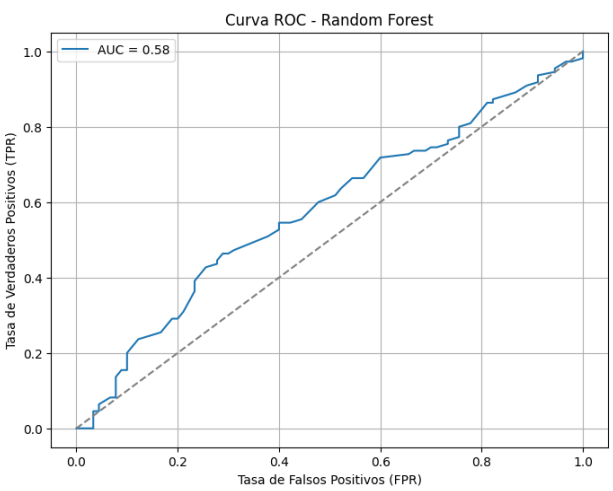
- Total de registros: 1.000 transacciones.
- Variables principales: Edad, Género, Categoría de producto, Cantidad comprada, Precio unitario, Fecha.
- Variable objetivo: Ingreso Total (posteriormente transformado a binario: alto o bajo según mediana).

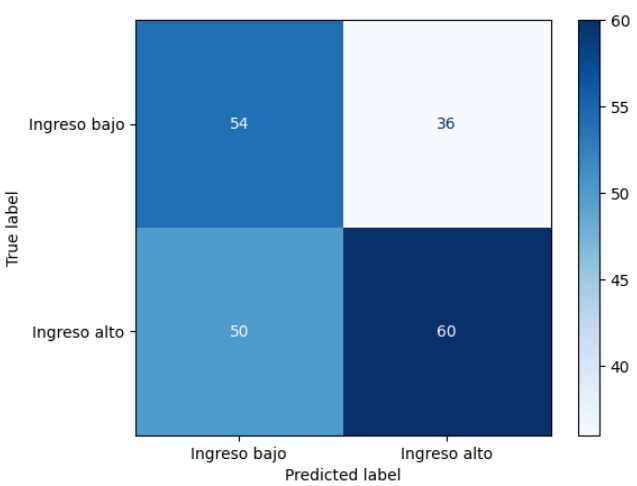
Principales Análisis y Hallazgos:

- Colinealidad: El ingreso total estaba altamente correlacionado con Price per Unit, generando data leakage.
- Overfitting inicial: Modelos como Random Forest predecían con RMSE = 0 debido a la fuga de datos.
- Ajuste: Se eliminó Price per Unit y se reformuló el problema como clasificación binaria.

Visualizaciones Clave

- Matriz de Confusión: El modelo logró 60 verdaderos positivos y 54 verdaderos negativos, pero presentó 86 errores en total.
- Curva ROC: AUC = 0.58, lo que indica una capacidad baja para distinguir entre ingresos altos y bajos.





Conclusiones y Recomendaciones

- El modelo actual tiene un rendimiento limitado (AUC bajo).
- Se recomienda:
  - Recolectar más datos o variables adicionales (promociones, fidelización, historial).Probar otros modelos (e.g. Gradient Boosting, redes neuronales).
  - Realizar ajuste de hiperparámetros y validación más robusta.
  - Explorar el uso de clasificación multiclase (bajo / medio / alto ingreso).

Creado por Hector Acevedo Astudillo - 2025