Proyecto Final – Computación Científica

Técnicas de Ensamble con Datos Reales de Crímenes en los Ángeles

Oscar Daniel Arcos Realpe - 614222710

2025

Resumen

Este proyecto final de la asignatura Computación Científica tiene como objetivo aplicar y comparar distintas técnicas de aprendizaje automático basadas en árboles para resolver tareas de clasificación y regresión utilizando un conjunto de datos reales sobre crímenes en Los Ángeles. Se evaluaron modelos de Bagging, Random Forest y Boosting, y se exploraron diferentes variables objetivo para ampliar el análisis. El trabajo se desarrolló en Google Colab utilizando bibliotecas como `Pandas`, `scikit-learn` y `Matplotlib`, y está estructurado en tres secciones: predicción del estado del caso (`Status`), predicción del tipo de crimen (`Part 1-2`) y predicción de la edad de la víctima (tanto en forma exacta como por rangos).

Introducción

Los métodos de ensamble como Bagging y Boosting se han consolidado como herramientas fundamentales en el aprendizaje automático moderno debido a su capacidad para mejorar la precisión y robustez de los modelos base. El presente trabajo tiene como motivación aplicar estas técnicas sobre un conjunto de datos públicos de crímenes cometidos en Los Ángeles, con el fin de explorar su capacidad predictiva en diferentes escenarios. A lo largo del semestre se estudiaron múltiples herramientas computacionales y bibliotecas de Python, que fueron empleadas para abordar este proyecto de forma integral. El objetivo principal es comparar el rendimiento de estos modelos bajo diferentes tipos de problemas: clasificación multiclase, clasificación binaria y regresión.

Sección 1: Predicción del estado del caso (`Status`)

La primera tarea planteada consistió en predecir la variable `Status`, que representa el estado final del caso (por ejemplo, arresto realizado, pendiente, cerrado, etc.). Se trató como un problema de clasificación multiclase con clases desbalanceadas, donde la clase 3 representaba la mayoría de los casos. Se entrenaron y evaluaron los modelos de Árbol de Decisión, Bagging, Random Forest y Boosting.

A pesar de que todos los modelos alcanzaron una precisión global cercana al 80 %, el análisis detallado reveló un fuerte sesgo hacia la clase dominante. El modelo Bagging logró un

mejor equilibrio entre clases, especialmente al elevar los valores de f1-score en las clases menos representadas. Random Forest, aunque teóricamente más robusto, no superó al Bagging en este caso debido al desbalance del conjunto de datos. Boosting, por su parte, mostró una alta precisión global, pero bajo desempeño en clases minoritarias, lo que sugiere un sobreajuste a los patrones dominantes.

Sección 2: Predicción del tipo de crimen (`Part 1-2`)

En esta segunda parte del proyecto se replanteó el problema predictivo utilizando como variable objetivo `Part 1-2`, que clasifica los crímenes en dos categorías: Parte 1 (crímenes graves) y Parte 2 (crímenes menores). Esta reformulación permitió trabajar con una variable binaria más balanceada y significativa desde el punto de vista operativo.

Los modelos entrenados fueron los mismos que en la sección anterior: Árbol de Decisión, Bagging, Random Forest y Boosting. Todos los modelos alcanzaron un rendimiento prácticamente perfecto, con valores de precisión y f1-score cercanos a 1.0. Esto sugiere que las variables disponibles en el conjunto de datos permiten predecir con gran certeza la gravedad del crimen.

Sección 3.1: Predicción de la edad exacta de la víctima (Regresión)

La tercera sección del proyecto cambió el enfoque de clasificación a regresión, planteando como objetivo predecir la edad exacta de la víctima (`Vict Age`). Para ello, se utilizaron versiones de regresión de los modelos ya trabajados: Árbol de Regresión, Bagging Regressor, Random Forest Regressor y Gradient Boosting Regressor.

Las métricas evaluadas fueron el error absoluto medio (MAE), el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Todos los modelos obtuvieron valores de MAE cercanos a 0.06 y MSE alrededor de 0.03, mientras que el mejor R^2 lo obtuvo Random Forest con un 36.8 % de varianza explicada.

Sección 3.2: Clasificación por rangos de edad

Como complemento al enfoque de regresión, se diseñó una segunda estrategia para abordar la variable `Vict Age`, esta vez mediante clasificación. Para ello, se agruparon las edades en cuatro rangos: 0–17, 18–35, 36–60 y 61+. Esta reformulación convirtió el problema en una clasificación multiclase, mejorando la interpretabilidad de los resultados.

Se utilizaron los modelos Árbol de Decisión, Bagging y Random Forest. El modelo Boosting fue descartado debido a problemas técnicos relacionados con el bajo número de registros en una de las clases, lo que generaba errores al ajustar pesos internos.

El rendimiento de los tres modelos fue perfecto en términos de precisión y f1-score, lo que indica que los rangos de edad fueron bien capturados a partir de las variables predictoras disponibles.

Conclusiones generales

Este proyecto permitió aplicar, comparar y analizar diferentes modelos de ensamble (Bagging, Random Forest y Boosting) en contextos de clasificación multiclase, clasificación binaria y regresión. Se trabajó con un conjunto de datos reales sobre crímenes en Los Ángeles, y se exploraron múltiples variables objetivo para obtener una visión más completa de las capacidades de los modelos.

Entre los hallazgos principales se destacan:

- La importancia de revisar el balance de clases al interpretar métricas como la precisión global.
- La superioridad del Bagging frente a otros modelos en tareas desbalanceadas.
- El excelente desempeño de todos los modelos en problemas bien estructurados, como la predicción de `Part 1-2`.
- El uso de técnicas de regresión mostró limitaciones ante una variable tan dispersa como la edad, pero los resultados fueron razonables y se reforzaron al reformular el problema como clasificación por rangos.

En conjunto, el trabajo evidenció la aplicabilidad de métodos computacionales avanzados en problemas del mundo real, y fortaleció la comprensión práctica de técnicas de ensamble vistas durante el curso.

Reflexión final

Este proyecto no solo permitió aplicar modelos de aprendizaje automático, sino también reflexionar sobre su potencial uso en entornos reales como la prevención del crimen. Aunque los modelos matemáticos no pueden reemplazar el juicio humano ni el contexto legal, sí pueden apoyar la toma de decisiones, el análisis histórico de patrones delictivos y la asignación estratégica de recursos policiales. Esta experiencia demuestra el valor de la computación científica como puente entre los datos y la acción.