Actividad: Random Forest en Python

Gloria

Marzo 2025

Introducción

Random Forest es un algoritmo de aprendizaje supervisado basado en árboles de decisión. Combina múltiples árboles de decisión y vota por la clase más frecuente en clasificación. Su capacidad para reducir el sobreajuste y mejorar la precisión lo hace ideal para conjuntos de datos complejos y desbalanceados.

Metodología

Se utilizó el dataset de transacciones de tarjetas de crédito disponible en Kaggle. Este dataset contiene una alta desproporción entre clases (fraude y no fraude). Se realizó el siguiente procedimiento:

- Se cargó y analizó el dataset para identificar el desbalance entre clases.
- Se aplicó escalamiento a los datos utilizando la técnica StandardScaler.
- Se dividió el conjunto de datos en entrenamiento y prueba (70% - 30%).
- Se entrenó un modelo de Random Forest con 100 árboles, bootstrap habilitado y selección de características aleatorias con `max_features='sqrt'`.
- Se evaluó el modelo usando métricas de clasificación como matriz de confusión, precisión, recall y f1-score.
- Se comparó el rendimiento con un modelo base de regresión logística.
- Se generaron curvas ROC y se calcularon los valores AUC para ambos modelos.

Resultados

El modelo Random Forest obtuvo un **recall de 0.80** y un **f1-score de 0.87** para la clase positiva (fraude), superando al modelo base de regresión logística que obtuvo un recall de 0.62 y f1-score de 0.73. Esto muestra una mejora importante en la detección de fraudes reales.

Conclusión

Random Forest resultó ser un modelo eficaz para detectar fraudes en datos desbalanceados. Superó significativamente al modelo de regresión logística, especialmente en la clase minoritaria, lo cual es fundamental en problemas reales como la detección de fraude financiero.