Random Forest

(Bosques Aleatorios)

Analítica de Datos, Universidad de San Andrés

Si encuentran algún error en el documento o hay alguna duda, mandenmé un mail a rodriguezf@udesa.edu.ar y lo revisamos.

1. Introducción a Random Forest

Random Forest es una técnica de aprendizaje automático que combina múltiples árboles de decisión para crear un modelo más robusto y preciso. A diferencia de CART, que utiliza un solo árbol de decisión basado en el índice de Gini y tiende a ajustarse demasiado a los datos de entrenamiento (overfitting), Random Forest construye varios árboles y combina sus predicciones para obtener un resultado final con mejor capacidad de generalización.

El algoritmo funciona de la siguiente manera:

1. Selecciona aleatoriamente un subconjunto de los datos de entrenamiento (con reemplazo) para cada árbol.
2. Para cada árbol, selecciona aleatoriamente un subconjunto de las variables predictoras.
3. Construye un árbol de decisión completo usando CART con los datos y variables seleccionadas.
4. Repite los pasos anteriores para crear múltiples árboles.
5. Para hacer una predicción, cada árbol vota y se toma la predicción más común (en clasificación) o el promedio (en regresión).

Las ventajas principales de Random Forest son:

- Reduce el overfitting al promediar múltiples árboles.
- Maneja bien tanto variables numéricas como categóricas.
- Es robusto a outliers y ruido en los datos.

- Proporciona una medida de importancia de variables.
- No requiere normalización de datos.

2. Descripción del dataset

El análisis se realizó sobre el dataset **Wine** de scikit-learn, que contiene información química de 178 vinos de tres variedades diferentes. Cada observación corresponde a una muestra de vino, con 13 variables numéricas (por ejemplo: alcohol, magnesio, fenoles, color, etc.) y una variable objetivo categórica (`target`) que indica la clase de vino (0, 1 o 2).

Para detalles completos de las variables y primeras filas del dataset, consultar el anexo del notebook.

3. Análisis exploratorio de datos (EDA)

- **Datos nulos:** No se encontraron datos nulos en el dataset.
- **Estadísticas descriptivas:** Las variables presentan diferentes escalas y rangos. Algunas variables muestran mayor dispersión (por ejemplo, `proline` y `color intensity`).
- **Distribución de clases:** Las tres clases de vino están representadas con 59 muestras de la clase 0, 71 muestras de la clase 1 y 48 muestras de la clase 2, mostrando un desbalance moderado entre las categorías.
- **Visualizaciones:** Se realizaron histogramas, boxplots y gráficos de correlación para explorar la distribución y relación entre variables (consultar anexo del notebook para gráficos y comentarios).

4. Modelado y resultados

Se entrenaron y compararon dos modelos principales:

- **Árbol de Decisión (CART):** Se implementó un árbol de decisión utilizando el criterio de Gini para la división de nodos. Recordemos que esto puede overfittear dado que CART es un algoritmo greedy.

- **Random Forest:** Se entrenó un modelo con 100 árboles, lo que permitió mejorar la capacidad de generalización y reducir el overfitting. Random Forest alcanzó una precisión perfecta.
- **Métricas de desempeño:** Con un test size del 20 % (36 muestras), el árbol de decisión se equivocó en 2 casos mientras que Random Forest solo se equivocó en 0 casos. Esto puede ser debido a que el Random Forest es un modelo más robusto y generaliza mejor los datos.

¿Qué pasa si hacemos que el test size es más grande? ¿Qué modelo será capaz de generalizar mejor los datos?

A continuación está el código del notebook. Es mejor si pueden bajarlo y correrlo en su máquina para poder modificarlo y explorarlo mejor.