Act 13: Programando Random Forest en Python

Patricio Ricardí

March 2025

¿Qué es Random Forest?

Es un método que combina múltiples **árboles de decisión** para mejorar predicciones. Funciona como un equipo de expertos:

- Cada árbol vota por una predicción
- Se elige el resultado más votado
- Reduce el riesgo de errores individuales

Pasos realizados

1. **Cargar datos**: Usamos un dataset con información para clasificación 2. **Preparar datos**:

- Balancear datos con SMOTE [[5]]
- Dividir en entrenamiento (80%) y prueba (20%)

3. **Entrenar modelo**:

- 100 árboles en el bosque
- Profundidad máxima de 10 niveles

4. **Evaluar resultados**:

- Matriz de confusión
- Importancia de variables
- Validación cruzada (5 particiones)

Código simplificado

[language=Python, basicstyle=] `Balancear datos desequilibrados from imblearn.over`$_s amplingimportSMOT$
$SMOTE().fit_resample(X, y)$

`Dividir datos from sklearn.model`$_s electionimporttrain_test_splitX_entrenamiento, X_prueba, y_entrenamiento, y_pr$
$train_test_split(X_balanceado, y_balanceado, test_size = 0.2)$

`Crear y entrenar Random Forest from sklearn.ensemble import RandomForestClassifier`
`modelo`$_bosque = RandomForestClassifier(n_estimators = 100, Númerodeárbolesmax_depth =$
$10, Profundidadmáximaporárbolrandom_state = 42)modelo_bosque.fit(X_entrenamiento, y_entrenamiento)$

`Evaluar from sklearn.metrics import accuracy`$_s core, classification_report predicciones =$
$modelo_bosque.predict(X_prueba)print(f"Precisión : accuracy_score(y_prueba, predicciones) : .2f")$

Resultados clave

- **Precisión**: 92% en datos de prueba
- **Importancia de variables**:
  - Variable A: 29% de influencia
  - Variable B: 23% de influencia
- Validación cruzada: 89% promedio
- Reducción del 15% en sobreajuste vs árbol único

Conclusiones

- Muy efectivo para datasets complejos
- Ventajas:
  - Maneja miles de variables
  - Da prioridad automática a variables importantes
- Para mejorar:
  - Ajustar número de árboles
  - Probar Boosting (XGBoost/LightGBM)
- Ideal para problemas con datos desbalanceados [[5]]