Random Forest

Minería de Datos (CC3074) - 2026

Random Forest

Semestre 01, 2026

Problema

Inestabilidad

- Los árboles de decisión son muy sensibles a los datos.
- Pequeños cambios en el dataset pueden producir árboles muy distintos.
- Esto genera modelos inestables.

Alta Varianza

- Un árbol puede ajustarse demasiado al conjunto de entrenamiento.
- Aprende patrones específicos que no se repiten en datos nuevos.
- Esto produce overfitting.

Los árboles individuales tienen alta varianza.

Aprendizaje en Conjunto

En lugar de usar un solo modelo, usamos muchos modelos.

La idea es simple:

- Entrenar múltiples árboles
- Combinar sus predicciones
- Obtener un modelo más estable

Principio

Muchos modelos débiles → un modelo fuerte.

Random Forest

Definición

Random Forest es un método ensemble basado en árboles de decisión.

Construye múltiples árboles y combina sus predicciones.

Cada árbol se entrena con una versión ligeramente distinta de los datos.

Intuición

Imagina que 100 expertos responden una pregunta.

Cada uno comete errores.

Pero si tomamos el promedio de sus respuestas, el resultado suele ser más confiable.

Random Forest hace exactamente eso con árboles.

¿Cómo funciona?

Random Forest introduce dos fuentes de aleatoriedad:

1. Muestreo aleatorio de datos
2. Muestreo aleatorio de variables

Esto hace que los árboles sean diferentes entre sí.

Paso 1: Bootstrap

Muestreo con reemplazo

Cada árbol se entrena con una muestra distinta del dataset.

Este muestreo se llama bootstrap.

Características:

- Se seleccionan observaciones con reemplazo
- Algunas observaciones se repiten
- Otras quedan fuera

Consecuencia

Cada árbol ve un dataset distinto.

Esto reduce la correlación entre árboles.

Menor correlación → mejor ensemble.

Paso 2: Selección Aleatoria de Variables

Subconjunto de atributos

En cada división del árbol:

- No se consideran todas las variables
- Solo un subconjunto aleatorio

Esto introduce diversidad entre árboles.

Ejemplo

Si tenemos 10 variables:

En cada nodo el algoritmo podría evaluar solo 3.

Esto produce árboles diferentes.

Construcción del Bosque

Entrenamiento

El proceso se repite muchas veces:

1. Crear muestra bootstrap
2. Entrenar un árbol
3. Repetir cientos de veces

El resultado es un bosque de árboles.

Predicción

Clasificación

Cada árbol vota una clase.

La predicción final es la clase más votada.

Regresión

Cada árbol produce un valor numérico.

La predicción final es el promedio de las predicciones.

Ventajas de Random Forest

Mayor estabilidad

- Reduce la varianza del modelo
- Menos sensible a cambios en los datos
- Mejora la generalización

Mejor desempeño

- Suele superar a árboles individuales
- Maneja relaciones complejas
- Funciona bien sin mucho ajuste

Importancia de variables

Random Forest permite estimar:

Qué variables son más importantes para la predicción.

Limitaciones

Menor interpretabilidad

Un árbol es fácil de explicar.

Un bosque con 500 árboles no lo es.

Mayor costo computacional

Entrenar muchos árboles requiere:

- más memoria
- más tiempo de cálculo

Hiperparámetros

Número de árboles

`n_estimators`

Cantidad de árboles en el bosque.

Más árboles → mayor estabilidad.

Variables por división

`max_features`

Número de variables evaluadas en cada nodo.

Profundidad del árbol

`max_depth`

Limita el tamaño de cada árbol.

Relación con Árboles de Decisión

Un Random Forest es:

Muchos árboles de decisión combinados.

Los árboles individuales son modelos base débiles.

El bosque es un modelo fuerte.

Comparación

| Modelo | Estabilidad | Interpretabilidad | | ---------- | ---------- | ---------- | | Árbol | Baja | Alta | | Random Forest | Alta | Media |