

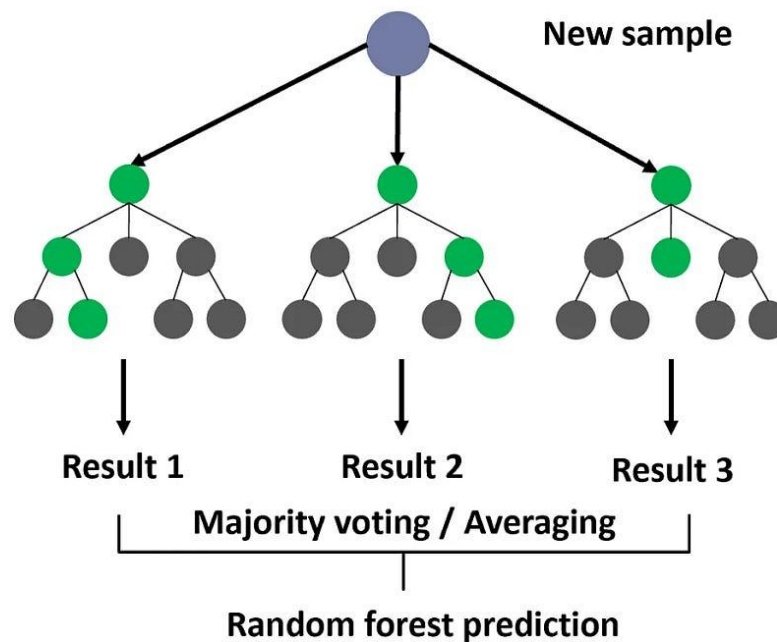
¿Qué es Random Forest?

Random Forest es un algoritmo de aprendizaje automático de uso común, registrado por Leo Breiman y Adele Cutler, que combina el resultado de múltiples árboles de decisión para llegar a un resultado único. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión.

¿Para qué sirve?

Random Forest es útil en una variedad de aplicaciones, incluyendo:

- **Clasificación:** Identificar categorías dentro de un conjunto de datos (ejemplo: detección de spam en correos electrónicos).
 - **Regresión:** Predecir valores numéricos (ejemplo: estimación de precios de viviendas).
 - **Selección de características:** Identificar las variables más importantes en un conjunto de datos.
 - **Detección de anomalías:** Identificar valores atípicos en grandes volúmenes de datos.
- Modelado de series temporales:** Predecir tendencias con datos históricos.



¿Cómo se aplica? (Métodos de implementación)

1. Recolección de datos

Se obtiene un conjunto de datos con variables independientes (**features**) y una variable objetivo (**target**).

2. Preprocesamiento

- Limpieza de datos (manejo de valores nulos, eliminación de outliers).
- Codificación de variables categóricas si es necesario.
- Normalización o estandarización de datos en caso de regresión.

3. Entrenamiento del modelo

- Se crean múltiples árboles de decisión entrenados con diferentes subconjuntos de datos.
- Cada árbol realiza una predicción y luego se combinan los resultados (promedio para regresión y votación para clasificación).

4. Evaluación del modelo

- Se mide la precisión con métricas como **accuracy**, **F1-score**, **RMSE**, **R²** dependiendo del tipo de problema.
- Se usa validación cruzada para verificar la generalización del modelo.

5. Ajuste de hiperparámetros

Parámetros clave de Random Forest:

- **n_estimators** (cantidad de árboles).
- **max_depth** (profundidad máxima de los árboles).
- **min_samples_split** (mínimo de muestras necesarias para dividir un nodo).
- **max_features** (número de características consideradas en cada división).