Clase N°5. Algoritmo de bosque aleatorio

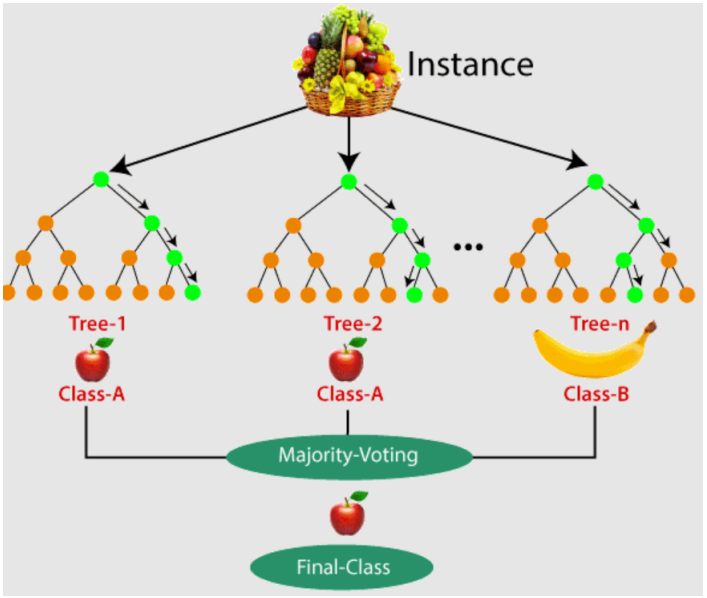
Un **bosque aleatorio** es una combinación de árboles de decisión seleccionados en forma aleatoria, los cuales permiten obtener un resultado global de la predicción.

Para generar el algoritmo, se utiliza el ***método de bagging*** o algoritmo de segmentación recursiva, donde se divide el conjunto de datos en diferentes muestras formadas por subconjuntos elegidos de manera aleatoria.

Luego, cada árbol es entrenado en forma independiente para obtener resultados preliminares, los cuales son evaluados mediante un sistema de votos en los **bosques aleatorios de clasificación** y una estimación porcentual en los **bosques aleatorios de regresión.**

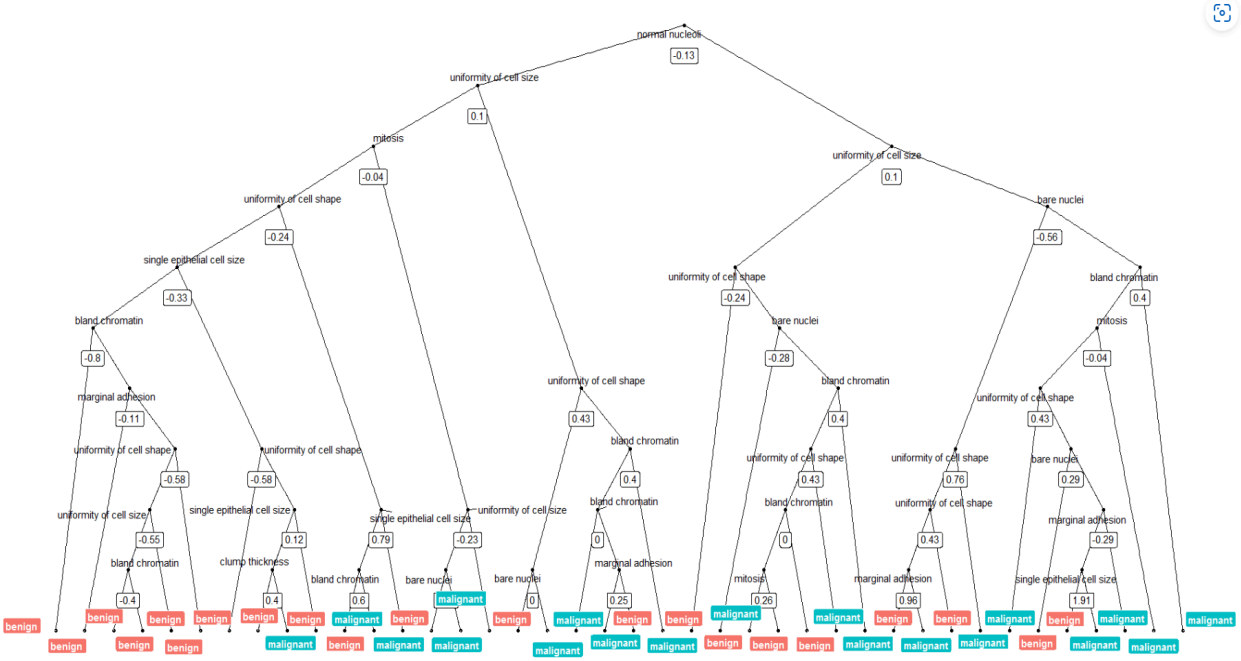
De la misma forma que en los árboles de clasificación y de regresión se cumplen las mismas etapas presentes en un modelo de machine learning: limpieza de los datos, partición, entrenamiento, evaluación y validación del modelo empleado.

EJEMPLO 1. Algoritmo RandomForestClassifier para clasificar frutas.

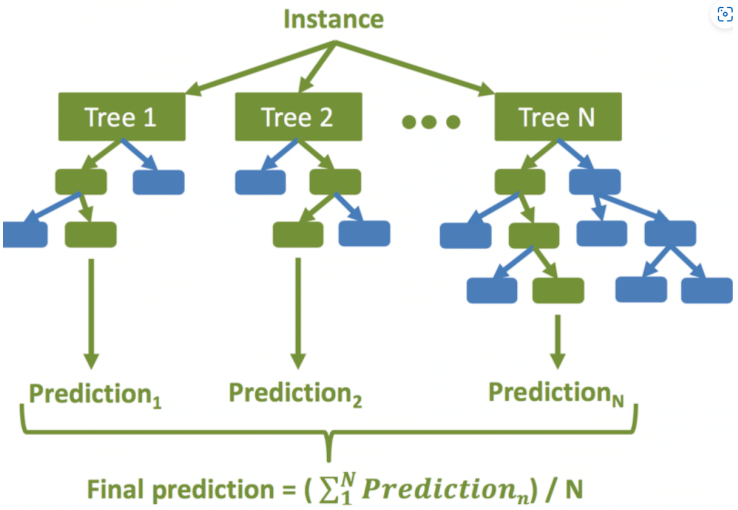


Partiendo de un conjunto de frutas que serán clasificadas por el algoritmo de bosque aleatorio, se puede observar que en el primer árbol se clasifica la categoría A:" manzana roja" al igual que en el segundo árbol. Por el contrario, en el último árbol se clasifica la categoría B: "banana". Aplicando el **método de votación** se tiene que el elemento de la categoría A: manzana roja es el que más se repite, por lo tanto es la respuesta de la predicción del conjunto de árboles de decisión.

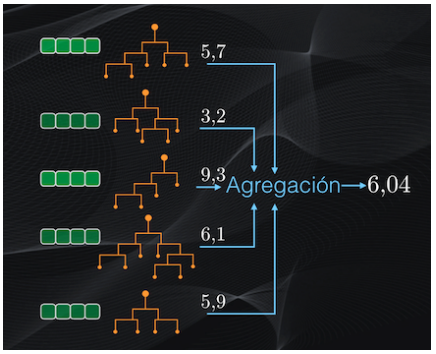
EJEMPLO 2. Algoritmo RandomForestClassifier para clasificar tipo de tumor.



EJEMPLO 3. Algoritmo RandomForestRegressor para predecir variables numéricas continuas.



Para poder hallar el valor de la predicción del bosque aleatorio de regresión, se calcula **el promedio** de los valores parciales de las predicciones de cada árbol de regresión.



Ventajas y desventajas del modelo

Beneficios	Desafíos
• Disminuye el riesgo de overfitting o sobreajuste.	• Requiere mucho tiempo de procesamiento.
• Es flexible ante la implementación de métricas de validación.	• Necesita más recursos de almacenamiento.
	• Mayor complejidad de la estructura del algoritmo.

Hiperparámetros de RandomForestClassifier

RandomForestClassifier (*n_estimators=100,* determina la cantidad de árboles.

criterion='gini',

max_depth=None,

min_samples_split=2,

min_samples_leaf=1,

min_weight_fraction_leaf=0.0,

max_features='sqrt',

max_leaf_nodes=None,

min_impurity_decrease=0.0,

bootstrap=True)

RandomForestRegressor (*n_estimators=100, determina la cantidad de árboles.*

criterion='squared_error',

max_depth=None,

min_samples_split=2,

min_samples_leaf=1,

min_weight_fraction_leaf=0.0,

max_features='sqrt',

max_leaf_nodes=None,

min_impurity_decrease=0.0,

bootstrap=True)

ACTIVIDAD

Considerando la base de datos analizada en la clase sobre indicadores socioeconómicos de la República Argentina, siendo la variable respuesta “poverty” y los factores: “school dropout” y “birth mortal”, ejecutar un bosque aleatorio considerando:

- Caso 1: 50 árboles de decisión.
- Caso 2: 70 árboles de decisión.
- Caso 3: 95 árboles de decisión.

Evaluar en cuál de todos los supuestos se logra reducir el (MSE) error cuadrático medio.