Machine learning y deep learning: bosques aleatorios

Rolando Gonzales Martinez, PhD

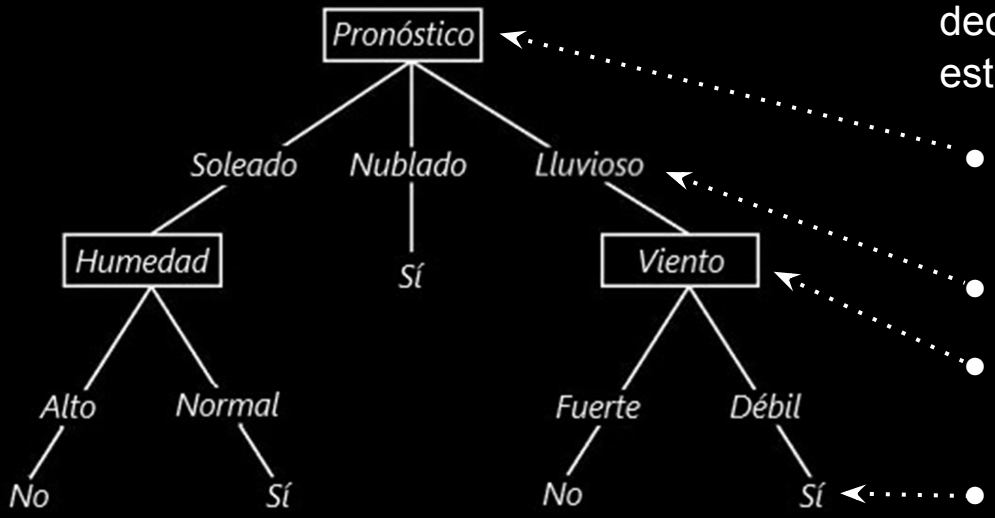
Fellow postdoctoral Marie Skłodowska-Curie

Universidad de Groningen (Países Bajos)

Investigador (researcher)

Iniciativa de Pobreza y Desarrollo Humano de la Universidad de Oxford (UK)

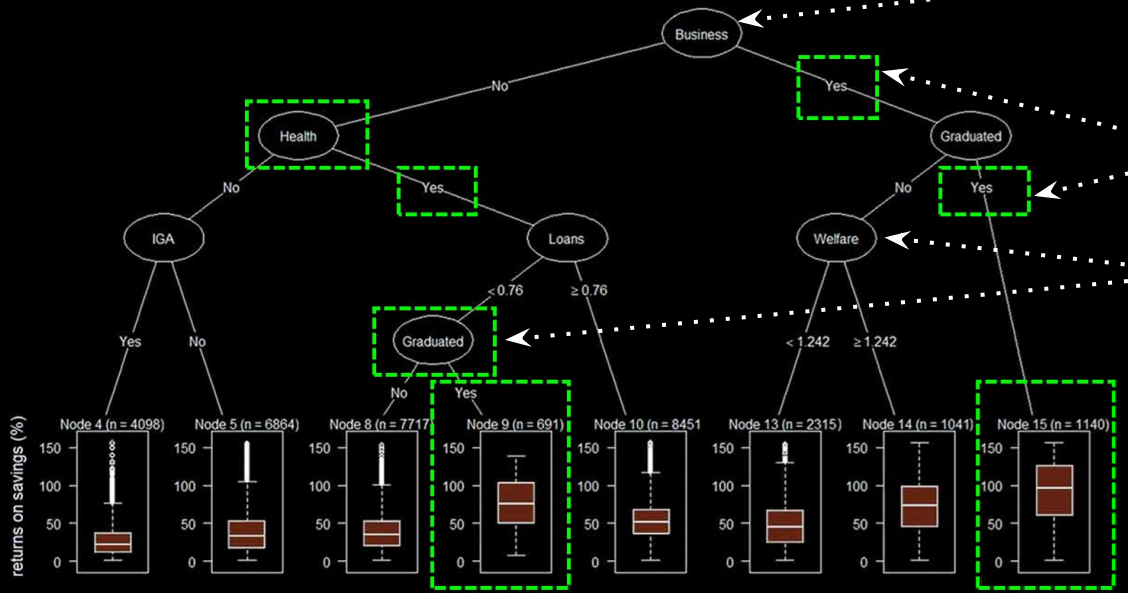
Árbol de decisión



Un árbol de decisión es una estructura jerárquica en la que se realiza una serie de decisiones secuenciales. TIene la siguiente estructura:

- **Raíz:** El nodo superior del árbol donde comienza la división de los datos.
- **Ramas:** que representan las divisiones
- **Nodos:** representan opciones de división para realizar divisiones adicionales
- **Hojas:** Representan las decisiones finales o las clasificaciones.

Árbol de decisión



Raíz: El nodo superior del árbol donde comienza la división de los datos.

Ramas: que representan las divisiones

Nodos: representan opciones de división para realizar divisiones adicionales

Hojas: Representan las decisiones finales o las clasificaciones.

Árbol de decisión: estadigrafo de Gini

- El Gini es una medida de pureza utilizada en la construcción de árboles de decisión.
- Se usa para evaluar la calidad de una división o partición en los nodos del árbol. El objetivo del índice de Gini es determinar cuán mezclados están los datos
- El Gini mide la proporción de clases en un nodo particular.

$$Gini = 1 - (p_1^2 + p_2^2)$$

El índice de Gini oscila entre 0 y 0.5.

- **Gini = 0**: Indica que todas las instancias en el nodo pertenecen a una sola clase, es decir, el nodo es *puro*.
- **Gini cercano a 0.5**: Indica que las instancias en el nodo están distribuidas equitativamente entre las clases, lo que significa que el nodo es *impuro*.

Árbol de decisión: Nodo Raiz

El nodo inicial del árbol de decisión se escoge mediante una evaluación de todas las características (features) y posibles umbrales, seleccionando la combinación que minimice la impureza de los nodos hijos resultantes:

$$D_{\text{izq}} \text{ con las instancias donde } x_i \leq umbral.$$

$$D_{\text{der}} \text{ con las instancias donde } x_i > umbral.$$

$$I_{\text{hijos}}(umbral) = \frac{|D_{\text{izq}}|}{|D|} \cdot I(D_{\text{izq}}) + \frac{|D_{\text{der}}|}{|D|} \cdot I(D_{\text{der}})$$

$$x^*, umbral^* = \arg \min_{x_i, umbral} (I_{\text{hijos}}(umbral))$$

Árbol de decisión: Algoritmo

1. Selección de la Característica y el Umbral de partición:

- **Para cada característica**: El algoritmo evalúa todos los posibles puntos de división (umbrales).
- **Para cada umbral posible**: Se separan las muestras en dos grupos: uno donde la característica es menor o igual al umbral, y otro donde es mayor.
- **Cálculo del Gini**: Se calcula el Gini para los ramas resultantes de la división.

2. Elección del Mejor Umbral de partición:

- **Comparación de la pureza**: Se comparan los Ginis de las posibles divisiones. El objetivo es minimizar la impureza (Gini < 0.5) en los nodos hijos.
- **Selección del umbral**: Se selecciona el umbral que produce la mayor reducción del Gini (reducción impureza). Esto significa que se selecciona el umbral que resulta en los nodos hijos más puros posibles.

Árbol de decisión: Algoritmo

3. División del Nodo:

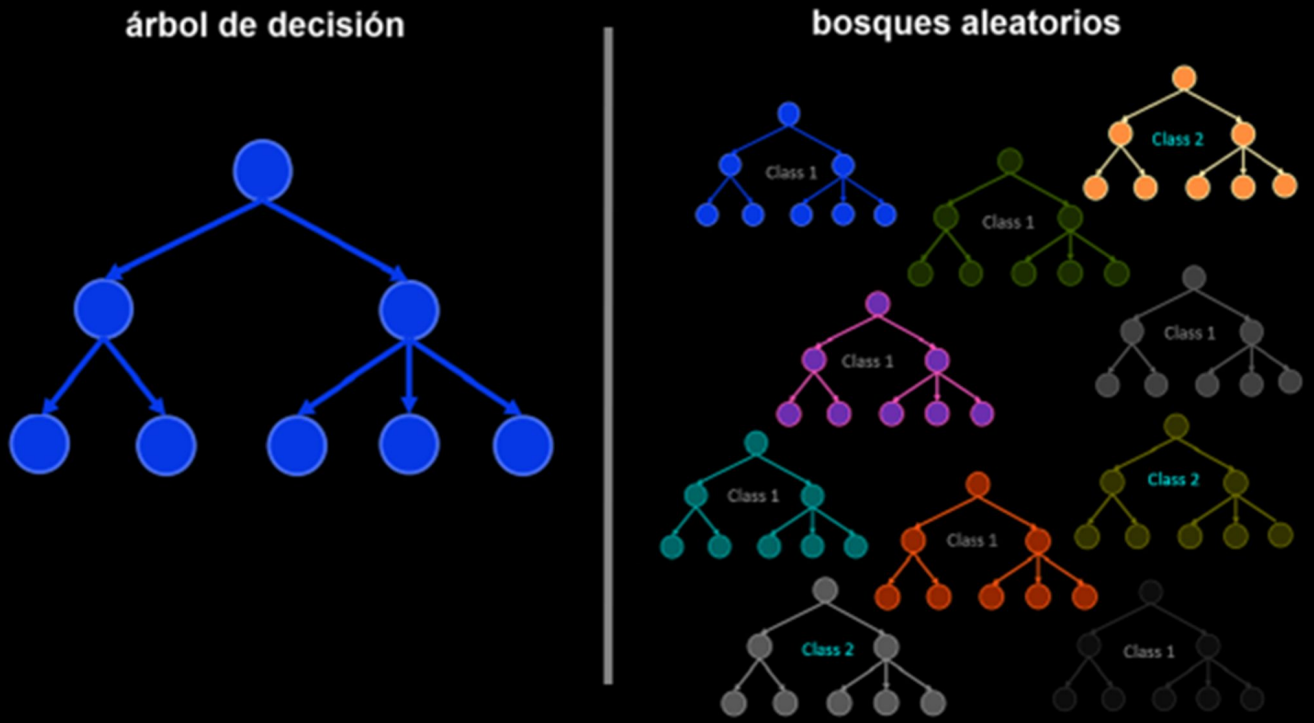
- Una vez seleccionado el umbral óptimo, el nodo se divide en dos ramas: una para las instancias que cumplen con la condición (es decir, donde el valor de la característica es menor o igual al umbral) y otra para las que no (donde el valor es mayor al umbral).

4. Repetición del Proceso:

- Este proceso se repite recursivamente para cada nodo hijo hasta que se cumpla un criterio de detención (por ejemplo, cuando todos los nodos son puros, o cuando no se puede reducir significativamente la impureza).

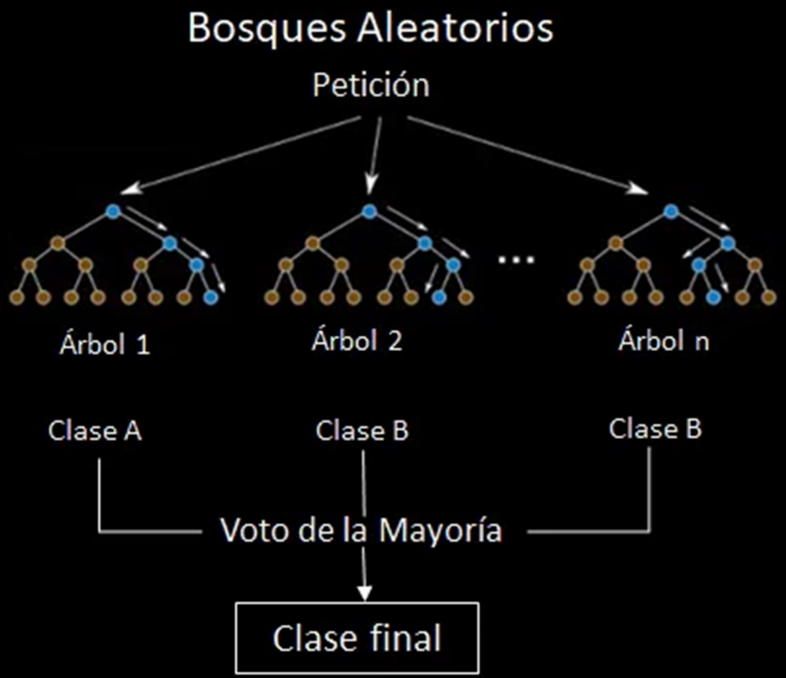
Bosques aleatorios

Los bosques aleatorios son algoritmos de aprendizaje supervisado que combina la salida de múltiples árboles de decisión.



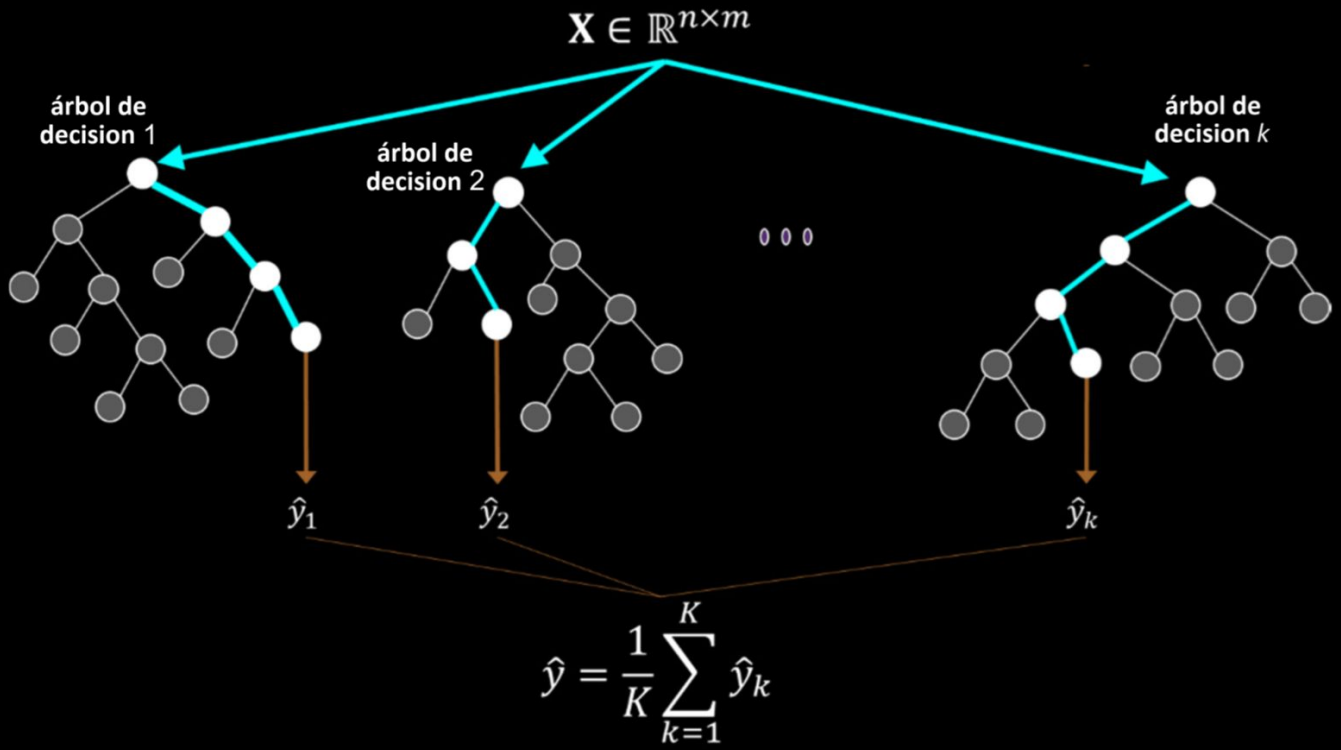
Bosques aleatorios

La clasificación final es el resultado de la agregación de las clasificaciones individuales de los árboles de decisión individuales:



Bosques aleatorios

Por ejemplo: en el caso de datos continuos, el resultado final del bosque aleatorio es el promedio de las predicciones individuales de cada árbol:



Bosques aleatorios

- Cada árbol en el bosque se entrena usando una **muestra diferente** de los datos original. Estas muestras se obtienen mediante muestreo con reemplazo.
- En cada árbol, se considera solamente un subconjunto aleatorio de características (features) para decidir la mejor división.

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

$$\{T_1, T_2, \ldots, T_B\} \quad \begin{matrix} D_b \subseteq D \\ \mathcal{X}_b \subseteq \mathcal{X} \end{matrix}$$

$$\hat{y}_b(x') = T_b(x') \quad T_b : \mathcal{X} \to \mathcal{Y}$$

$$\hat{y}(x') = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b(x')$$

$$\hat{y}(x') = \arg\max_{c \in \mathcal{Y}} \sum_{b=1}^{B} \mathbb{I}\left[\hat{y}_b(x') = c\right]$$

$$\mathcal{Y} = \{c_1, c_2, \ldots, c_k\}$$

$$\hat{y}(x') = \text{moda}\left(\hat{y}_1(x'), \hat{y}_2(x'), \ldots, \hat{y}_B(x')\right)$$

Bosques aleatorios

Ventajas

- Se aplica con inputs y outputs categóricos y continuos
- Es paralelizable
- Es robusto ante valores atípicos
- Funciona bien con datos desbalanceados
- Cuanto mayor es la cantidad de árboles, más preciso es el resultado, pero hay que tener cuidado con el sobreajuste
- Puede generalizar mejor debido al uso de múltiples árboles de decisión

Desventajas:

- Interpretabilidad: los modelos de bosque aleatorios no son fácilmente interpretables
- Para conjuntos de datos muy grandes, el tamaño de los árboles puede consumir mucha memoria.
- Puede tender al sobreajuste, por lo que es necesario ajustar los hiperparámetros.

Bosques aleatorios: medidas de evaluación

- El soporte para una clase C es simplemente el número de ejemplos en el conjunto de prueba que pertenecen a esa clase. Un soporte alto en una clase permite calcular métricas más robustas y confiables para esa clase.
- El Promedio Macro y el Promedio Ponderado son dos formas de promediar las métricas de rendimiento (como precisión, recall, F1-score) a través de todas las clases en un problema de clasificación.

$$\text{Soporte}(C_i) = N_i$$

$$\text{Promedio Ponderado} = \frac{\sum_{i=1}^{N} (\text{Soporte}_i \times \text{Métrica}_i)}{\sum_{i=1}^{N} \text{Soporte}_i}$$

$$\text{Promedio Macro} = \frac{1}{N} \sum_{i=1}^{N} \text{Métrica}_i$$

Bosques aleatorios: medidas de evaluación

- El Promedio Macro trata a todas las clases por igual, independientemente de cuántas instancias haya en cada clase. Es útil cuando cuando se busca que el modelo tenga un rendimiento equilibrado en todas las clases.
- El Promedio Ponderado da más importancia a las clases con más instancias. Es útil en clases desbalanceadas porque refleje el rendimiento en las clases más representadas cuando estas son de interés.

$$\text{Soporte}(C_i) = N_i$$

$$\text{Promedio Ponderado} = \frac{\sum_{i=1}^{N} (\text{Soporte}_i \times \text{Métrica}_i)}{\sum_{i=1}^{N} \text{Soporte}_i}$$

$$\text{Promedio Macro} = \frac{1}{N} \sum_{i=1}^{N} \text{Métrica}_i$$

Bosques aleatorios: cálculo de importancia de features

El cálculo de la importancia de las variables en un modelo de **Random Forests** se basa en la **reducción de la impureza** (Gini < 0.5) en los t-nodos de los árboles de decisión.

$$\text{Gini}(t) = 1 - \sum_{i=1}^{C} p_i^2$$

$$\Delta\text{Impureza} = \text{Impureza}(t) - \left(\frac{N_L}{N} \times \text{Impureza}(t_L) + \frac{N_R}{N} \times \text{Impureza}(t_R) \right)$$

$$\text{Importancia}(x_j) = \sum_{t \in \text{nodos donde } x_j \text{ es utilizado}} \Delta\text{Impureza}(t) \qquad \text{árbol } k$$

$$\text{Importancia}(x_j) = \frac{1}{T} \sum_{k=1}^{T} \text{Importancia}(x_j, \text{árbol } k)$$