# ÁRBOLES DE DECISIÓN

## INTRODUCCIÓN

En Machine Learning, uno de los objetivos principales es tomar decisiones automáticas a partir de datos. Los árboles de decisión de clasificación son uno de los algoritmos más intuitivos para lograrlo, ya que imitan la forma en que los humanos toman decisiones: haciendo preguntas secuenciales.

Un árbol de decisión permite:

- Analizar un conjunto de datos
- Formular preguntas basadas en sus características
- Dividir los datos paso a paso
- Llegar a una decisión final (clase)

Su principal fortaleza es que no funciona como una “caja negra”, sino que puede explicarse y visualizarse fácilmente. En problemas de clasificación, el objetivo del árbol es asignar una observación a una categoría discreta, por ejemplo:

- Spam / No spam
- Aprobado / Reprobado
- Enfermo / Sano
- Presente / Ausente

Para entender cómo funciona un árbol de decisión, es esencial conocer su vocabulario básico.

**Nodo raíz (Root Node)**

- Es el primer nodo del árbol
- Contiene todos los datos
- Realiza la primera pregunta
- Produce la división más importante del conjunto

**Nodo de decisión (Decision Node)**

- Nodo intermedio del árbol
- Contiene una condición lógica
- Tiene dos o más ramas de salida

**Nodo hoja (Leaf Node)**

- Nodo final del árbol
- No contiene preguntas
- Representa una clase final

**Rama (Branch)**

- Conecta nodos
- Representa una respuesta posible a una pregunta
- Puede ser binaria (Sí / No) o multiclase

# ALGORITMO

El entrenamiento de un árbol de decisión es un proceso jerárquico y recursivo. Su objetivo principal es dividir el conjunto de datos de forma inteligente, de manera que se logre la mejor separación posible entre las clases. La idea central del algoritmo puede resumirse así: en cada paso, el árbol selecciona la pregunta o condición que mejor separa los datos y repite este procedimiento hasta que ya no es necesario seguir dividiendo.

El proceso inicia considerando todo el conjunto de entrenamiento como el nodo raíz del árbol. A partir de ahí, en cada nodo se evalúan las características disponibles y se determina cuál de ellas, junto con un umbral de separación, produce la mayor ganancia de información. Esta ganancia se calcula utilizando un criterio de partición definido por el usuario (por ejemplo, entropía o índice Gini).

Aunque no es un requisito teórico de los árboles de decisión, muchas implementaciones prácticas —como la de scikit-learn— se limitan a divisiones binarias, ya que considerar más de dos particiones incrementa considerablemente el costo computacional.

Idealmente, las características deberían ser categóricas; sin embargo, cuando los valores son continuos, estos se discretizan para poder realizar las divisiones. Con base en los valores del atributo

seleccionado, los registros se distribuyen de manera recursiva hacia los nodos hijos. Para decidir el orden y la importancia de los atributos como nodos raíz o internos, el algoritmo utiliza métodos estadísticos que permiten evaluar qué divisiones aportan mayor información al modelo.

La ecuación que mide la cantidad de información es la siguiente:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} \times Entropy(S_v) \quad (1)$$

Donde

$$Entropy(x) = -p(x) \log_2(p(x)) \quad (2)$$

Para ejemplificar el algoritmo, utilizaremos un ejemplo muy sencillo. Suponga que tenemos un conjunto de datos sobre personas a las que les autorizaron un crédito y a la que no (Figura 1).

| ID | Estabilidad Laboral | Casado | Sueldo Arriba de 20k | Crédito |
|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 |
| 2 | 0 | 0 | 1 | 0 |
| 3 | 1 | 0 | 0 | 0 |

Fig. 1. Conjunto de datos

Seleccionaremos las variables y mediremos la cantidad de información; comenzaremos con la variable “Estabilidad Laboral”.

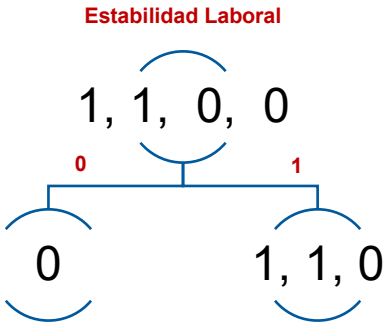


Fig. 2. División del conjunto de datos con respecto a “Estabilidad Laboral”

Calcularemos la ganancia de información de la siguiente manera:

$$Gain(S, Estabilidad\ Laboral) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} \times Entropy(S_v)$$

En donde $Entropy(S)$ es la entropía de todo el conjunto de datos (nodo raíz), la cual se calcula de la siguiente manera:

$$Entropy(S) = -[p(1)\log_2 p(1) + p(0)\log_2 p(0)]$$

Donde

- $p(1)$ es la probabilidad de ocurrencia del valor “1” en el nodo raíz
- $p(0)$ es la probabilidad de ocurrencia del valor “0” en el nodo raíz

Entonces:

$$Entropy(S) = -\left[\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right]$$

$$Entropy(S) = -\left[\frac{2}{4}(-1) + \frac{2}{4}(-1)\right]$$

$$Entropy(S) = 1$$

En este caso la entropía es máxima porque los valores “0” y “1” tienen la misma probabilidad de ocurrencia, indicando que la incertidumbre es máxima.

Ahora, calcularemos la ganancia de información en cada nodo:

$$\sum_{v \in Values(A)} \frac{|S_v|}{S} \times Entropy(S_v)$$

$$= \frac{|S_{Estabilidad\ Laboral=0}|}{S} \times Entropy(S_{Estabilidad\ Laboral=0}) + \frac{|S_{Estabilidad\ Laboral=1}|}{S} \times Entropy(S_{Estabilidad\ Laboral=1})$$

En donde $|S_{Estabilidad\ Laboral=0}|$ es el tamaño de muestras que contiene el nodo “Estabilidad Laboral =0” y $|S_{Estabilidad\ Laboral=1}|$ es el tamaño de muestras que contiene el nodo “Estabilidad Laboral =1”; y $S$ es el tamaño total de muestras en el nodo raíz. De acuerdo con la Figura 2, tenemos que $|S_{Estabilidad\ Laboral=0}| = 1$, $|S_{Estabilidad\ Laboral=1}| = 3$ y $S = 4$, por lo tanto:

$$\sum_{v \in Values(A)} \frac{|S_v|}{S} \times Entropy(S_v)$$

$$= \frac{1}{4} \times Entropy(S_{Estabilidad\ Laboral=0}) + \frac{3}{4} \times Entropy(S_{Estabilidad\ Laboral=1})$$

Ahora calcularemos $Entropy(S_{Estabilidad\ Laboral=0})$ y $Entropy(S_{Estabilidad\ Laboral=1})$ las cuales son las entropías de cada subconjunto de la variable “Estabilidad Laboral”.

$$Entropy(S_{Estabilidad\ Laboral=0}) = -[p(0)\log_2 p(0) + p(1)\log_2 p(1)]$$

De acuerdo con la Figura 2, para “Estabilidad Laboral” es igual a cero, tenemos solo una muestra de “0” y ninguna de “1”, por lo tanto:

$$Entropy(S_{Estabilidad\ Laboral=0}) = -\left[\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$Entropy(S_{Estabilidad\ Laboral=0}) = 0$$

Realizaremos lo mismo para $Entropy(S_{Estabilidad\ Laboral=1})$:

$$Entropy(S_{Estabilidad\ Laboral=1}) = -[p(0)\log_2 p(0) + p(1)\log_2 p(1)]$$

$$Entropy(S_{Estabilidad\ Laboral=1}) = -\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right]$$

$$Entropy(S_{Estabilidad\ Laboral=1}) = -\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right]$$

$$Entropy(S_{Estabilidad\ Laboral=1}) = 0.9181$$

Luego entonces:

$$\sum_{v \in Values(A)} \frac{|S_v|}{S} \times Entropy(S_v) = \frac{1}{4} \times (0) + \frac{3}{4} \times (0.9181)$$

$$\sum_{v \in Values(A)} \frac{|S_v|}{S} \times Entropy(S_v) = 0.6887$$

La ganancia de información entonces es:

$$Gain(S, Estabilidad\ Laboral) = 1 - 0.6887$$

$$Gain(S, Estabilidad\ Laboral) = 0.3112$$

La división del nodo raíz al utilizar “Estabilidad Laboral” queda como:

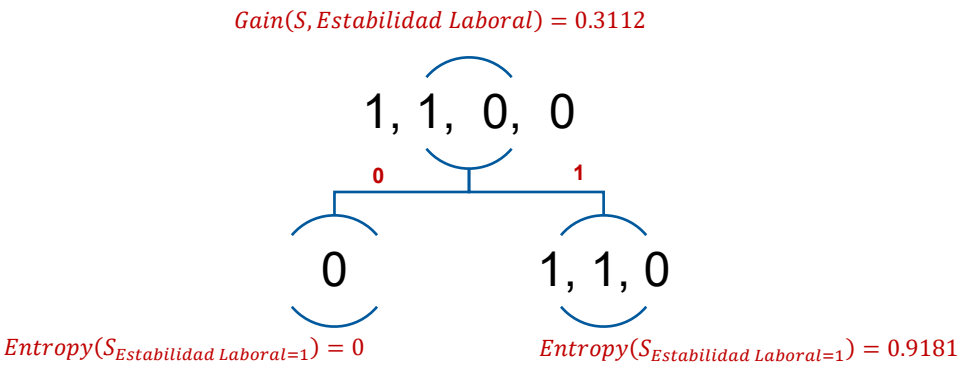


Fig. 3. División con $Gain(S, Estabilidad\ Laboral) = 0.3112$

En donde la ganancia de información es de 0.3112 y la entropía de cada nodo hijo es de 0 y 0.9181 para cuando Estabilidad Laboral es 0 y 1, respectivamente.

Al hacer lo mismo con las variables “Casado” y “Sueldo arriba de 20k” tenemos que:

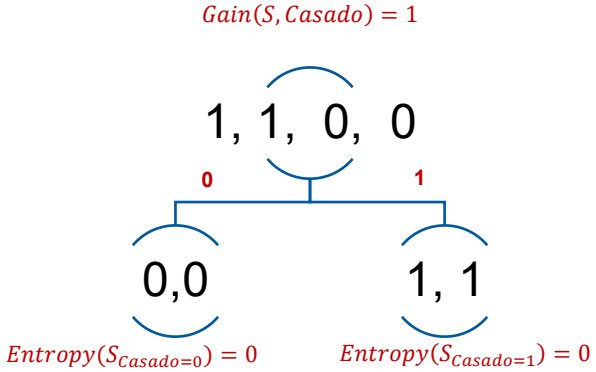


Fig. 4. División con $Gain(S, Casado) = 1$

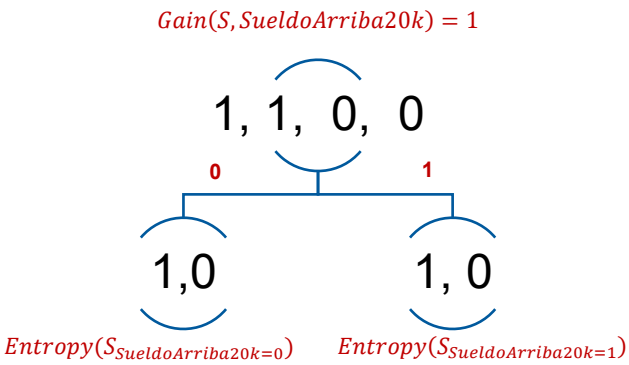


Fig. 5. División con $Gain(S, SueldoArriba20k) = 1$

Las Figuras 3, 4 y 5 muestran las ganancias de información al utilizar las tres variables para dividir el nodo raíz. Podemos observar que las variables “Casado” y “Sueldo Ariba de 20k” tienen una ganancia de información máxima, sin embargo, la división al utilizar la variable “Sueldo Arriba de 20k” los nodos hijos presentan un valor muy alto de entropía, por lo que se dice que no están puros. Por lo tanto, la variable a utilizar en la primera división será “Casado” y nuestro árbol de decisión quedaría de la siguiente manera:

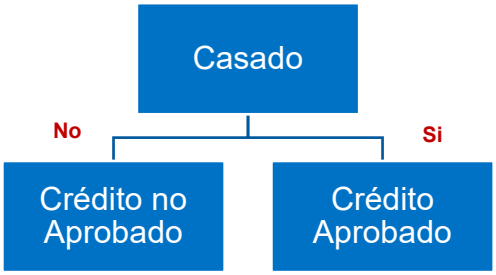


Fig. 6. Árbol de decisión para aprobación de crédito