Paso 1: Conjunto de Datos

Supongamos que tenemos el siguiente conjunto de datos que indica si una persona comprará un producto basado en dos características: **Edad** (Joven, Adulto, Senior) y **Ingreso** (Alto, Bajo).

Edad	Ingreso	Compra
Joven	Bajo	No
Joven	Alto	Sí
Adulto	Bajo	No
Adulto	Alto	Sí
Senior	Bajo	No
Senior	Alto	No

En total, hay 6 instancias: 2 "Sí" y 4 "No".

Paso 2: Calcular la Entropía del Nodo Raíz

Primero, calculamos la **entropía** del nodo raíz (antes de dividir los datos).

- Hay 2 compras "Sí" y 4 "No".
- Las probabilidades son: $p_{Si} = 2/6 = 0.33$ y $p_{No} = 4/6 = 0.67$.

La fórmula de la entropía es:

$$H(S) = -(p_{Si} \log_2(p_{Si}) + p_{No} \log_2(p_{No}))$$

Sustituyendo los valores:

$$H(S) = -(0.33 \log_2(0.33) + 0.67 \log_2(0.67))$$

Calculando:

$$H(S) = -(0.33 \cdot -1.585 + 0.67 \cdot -0.585)$$

$$H(S) = 0.528 + 0.392 = 0.920$$

Paso 3: Calcular la Ganancia de Información para Cada Atributo

Ahora calculamos la **ganancia de información** para los atributos **Edad** e **Ingreso** para determinar cuál divide mejor los datos.

1. Dividir por el atributo "Edad"

- **Edad = Joven**: 2 instancias (1 "Sí", 1 "No")
- **Edad = Adulto**: 2 instancias (1 "Sí", 1 "No")
- **Edad = Senior**: 2 instancias (0 "Sí", 2 "No")

Entropía para cada subconjunto:

1. Para **Edad = Joven**: $p_{Sí} = 0.5, p_{No} = 0.5$:

$$H(Joven) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = 1$$

2. Para **Edad = Adulto**: $p_{Sí} = 0.5, p_{No} = 0.5$:

$$H(Adulto) = 1$$

3. Para **Edad = Senior**: $p_{Sí} = 0, p_{No} = 1$:

$$H(Senior) = 0$$

Entropía ponderada para "Edad":

$$H(Edad) = \frac{2}{6} \cdot 1 + \frac{2}{6} \cdot 1 + \frac{2}{6} \cdot 0 = \frac{4}{6} = 0.667$$

Ganancia de Información para "Edad":

$$IG(Edad) = H(S) - H(Edad) = 0.920 - 0.667 = 0.253$$

2. Dividir por el atributo "Ingreso"

- **Ingreso = Bajo**: 3 instancias (0 "Sí", 3 "No")
- **Ingreso = Alto**: 3 instancias (2 "Sí", 1 "No")

Entropía para cada subconjunto:

1. Para **Ingreso = Bajo**: $p_{Sí} = 0, p_{No} = 1$:

$$H(Bajo) = 0$$

2. Para **Ingreso = Alto**: $p_{Sí} = \frac{2}{3}, p_{No} = \frac{1}{3}$:

$$H(Alto) = -(\frac{2}{3} \log_2(\frac{2}{3}) + \frac{1}{3} \log_2(\frac{1}{3})) = 0.918$$

Entropía ponderada para "Ingreso":

$$H(Ingreso) = \frac{3}{6} \cdot 0 + \frac{3}{6} \cdot 0.918 = 0.459$$

Ganancia de Información para "Ingreso":

$$IG(Ingreso) = H(S) - H(Ingreso) = 0.920 - 0.459 = 0.461$$

Paso 4: Elegir el Atributo con Mayor Ganancia de Información

- Ganancia de información para **Edad**: 0.253

- Ganancia de información para **Ingreso**: 0.461

Dado que la ganancia de información es mayor para **Ingreso**, este es el atributo que se usará para la primera división del árbol de decisión.

Paso 5: Continuar la Construcción del Árbol

Una vez que hemos dividido por **Ingreso**, seguiríamos aplicando el mismo proceso en cada subconjunto (Bajo y Alto), dividiendo los datos utilizando los atributos restantes hasta que todos los nodos sean puros o se cumpla un criterio de parada.

Este proceso muestra cómo el árbol de decisión utiliza la entropía y la ganancia de información para seleccionar los mejores atributos en cada paso, con el objetivo de dividir los datos de manera que minimice la impureza en los nodos hijos.