Vamos a realizar un ejemplo completo de la construcción de un árbol de decisión utilizando el **índice de Gini** como medida de impureza. El índice de Gini se usa para medir la probabilidad de clasificar incorrectamente una instancia seleccionada al azar si fuera etiquetada según la distribución de clases del nodo.

Paso 1: Conjunto de Datos

Utilizaremos el mismo conjunto de datos anterior, 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

Paso 2: Calcular el Índice de Gini del Nodo Raíz

El índice de Gini se calcula como:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

Donde p_i es la proporción de ejemplos de la clase i.

En el nodo raíz, tenemos 6 instancias en total: 2 "Sí" y 4 "No". Las probabilidades son:

$$p_{Si} = \frac{2}{6} = 0.33, \quad p_{No} = \frac{4}{6} = 0.67$$

Aplicamos la fórmula del índice de Gini:

$$Gini(S) = 1 - (0.33^2 + 0.67^2) = 1 - (0.1089 + 0.4489) = 1 - 0.5578 = 0.4422$$

Paso 3: Calcular el Índice de Gini para Cada Atributo

Ahora, calculamos el índice de Gini después de dividir los datos por los atributos **Edad** e **Ingreso** para determinar cuál es el mejor para dividir.

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")

Índice de Gini para cada subconjunto:

1. Para **Edad = Joven**:

$$Gini(Joven) = 1 - (0.5^2 + 0.5^2) = 1 - (0.25 + 0.25) = 0.5$$

2. Para **Edad = Adulto**:

$$Gini(Adulto) = 1 - (0.5^2 + 0.5^2) = 0.5$$

3. Para **Edad = Senior**:

$$Gini(Senior) = 1 - (0^2 + 1^2) = 0$$

Índice de Gini ponderado para "Edad":

$$Gini(Edad) = \frac{2}{6} \cdot 0.5 + \frac{2}{6} \cdot 0.5 + \frac{2}{6} \cdot 0 = 0.333$$

2. Dividir por el Atributo "Ingreso"

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

Índice de Gini para cada subconjunto:

1. Para **Ingreso = Bajo**:

$$Gini(Bajo) = 1 - (0^2 + 1^2) = 0$$

2. Para **Ingreso = Alto**:

$$Gini(Alto) = 1 - ((\frac{2}{3})^2 + (\frac{1}{3})^2) = 1 - (0.444 + 0.111) = 0.444$$

Índice de Gini ponderado para "Ingreso":

$$Gini(Ingreso) = \frac{3}{6} \cdot 0 + \frac{3}{6} \cdot 0.444 = 0.222$$

Paso 4: Comparar los Índices de Gini y Elegir el Mejor Atributo

- Índice de Gini para **Edad**: 0.333
- Índice de Gini para **Ingreso**: 0.222

Dado que el índice de Gini es menor para **Ingreso**, este atributo ofrece la mejor división inicial, ya que minimiza la impureza en los nodos hijos.

Paso 5: Dividir los Datos por "Ingreso"

Ahora, dividimos los datos según el atributo **Ingreso**:

1. **Ingreso = Bajo**: Todas las instancias son "No", por lo que este nodo es puro y no requiere más divisiones.
2. **Ingreso = Alto**: Este nodo tiene 2 instancias de "Sí" y 1 de "No". Aquí podemos continuar dividiendo los datos basándonos en otros atributos, como **Edad**.

Este ejemplo muestra cómo el algoritmo de árboles de decisión utiliza el **índice de Gini** para medir la impureza de los nodos y cómo se busca minimizar esta impureza en cada paso para construir el árbol de decisión. La división se realiza eligiendo el atributo que produce la menor impureza en los nodos hijos.