DECISION TREE : CART and C4.5

When to use GINI INDEX and when to use GINI RATIO ?

Prepared by Sanjiban S Roy, 5th Aug,2025

🧠 Decision Tree Attribute Selection Measures – Full Tutorial

✳️ Introduction

When building a decision tree, the most critical step is selecting the attribute that best splits the data at each node. The quality of a split is measured using various mathematical criteria:

- **ID3** uses **Information Gain**
- **C4.5** uses **Gain Ratio**
- **CART** uses **Gini Index**

Let's go through each method in depth with the **formulas and a worked example** using the AllElectronics dataset.

📐 FORMULAS FIRST

🔷 1. Information Gain (used in ID3)

Entropy (S):

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Where:

- p_i: proportion of class i in set S
- c: number of classes

Information Gain (IG):

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

Where:

- A: attribute
- S_v: subset of S for which attribute $A = v$

🔷 2. Gain Ratio (used in C4.5)

Split Info:

$$SplitInfo(S, A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot \log_2\left(\frac{|S_v|}{|S|}\right)$$

Gain Ratio:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

🔷 3. Gini Index (used in CART)

Gini Impurity:

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

Gini Index for a Split:

$$Gini_{split}(A) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Gini(S_v)$$

📋 Dataset: AllElectronics

We'll use this subset for simplicity:

ID	Age	Buys_computer
1	youth	no
2	youth	no
3	middle-aged	yes
4	senior	yes
5	senior	yes
6	senior	no
7	middle-aged	yes
8	youth	no
9	youth	yes
10	senior	yes
11	youth	yes
12	middle-aged	yes
13	middle-aged	yes
14	senior	no

- Total records: 14
- Yes = 9
- No = 5

🔍 Example: Attribute = Age

Values = {**youth**, **middle-aged**, **senior**}

🔷 1. Information Gain (ID3)

Step 1: Calculate Entropy(S)

$$Entropy(S) = -(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14})$$

$$= -(0.6439 \cdot \log_2(0.6439) + 0.3571 \cdot \log_2(0.3571)) = 0.940$$

Step 2: Compute Entropy for Age values

- **Youth** (5 samples): 2 Yes, 3 No

$$Entropy = -(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}) = 0.971$$

- **Middle-aged** (4 samples): 4 Yes, 0 No

$$Entropy = 0$$

- **Senior** (5 samples): 3 Yes, 2 No

$$Entropy = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}) = 0.971$$

Step 3: Information Gain

$$Gain(S, Age) = 0.940 - (\frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0.971)$$

$$= 0.940 - (0.3475 + 0 + 0.3475) = 0.940 - 0.695 = 0.245$$

✅ **Information Gain (Age) = 0.245**

🔷 2. Gain Ratio (C4.5)

$$SplitInfo(Age) = -(\frac{5}{14}\log_2\frac{5}{14} + \frac{4}{14}\log_2\frac{4}{14} + \frac{5}{14}\log_2\frac{5}{14})$$

$$= -(0.357 \cdot (-1.485) + 0.286 \cdot (-1.807) + 0.357 \cdot (-1.485)) = 1.577$$

$$GainRatio(Age) = \frac{0.245}{1.577} = 0.155$$

✅ **Gain Ratio (Age) = 0.155**

🔷 3. Gini Index (CART)

Now evaluate the **best binary split**, e.g.,

Split: `{youth, senior}` **and** `{middle-aged}`

- Group 1 (youth + senior): 10 samples → 5 Yes, 5 No

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

- Group 2 (middle-aged): 4 samples → 4 Yes, 0 No

$$Gini = 0$$

Weighted Gini:

$$Gini_{split} = \frac{10}{14} \cdot 0.5 + \frac{4}{14} \cdot 0 = 0.357$$

✅ **Gini Index for best split on Age = 0.357**

✅ Summary Table

Attribute	Info Gain (ID3)	Gain Ratio (C4.5)	Gini Index (CART)
Age	0.245	0.155	0.357

- **Use ID3** when you want pure entropy-based selection.
- **Use C4.5** when you want to avoid bias toward attributes with many values (uses Gain Ratio).
- **Use CART** when building binary trees (uses Gini Index and binary splits).

Prepared by Sanjiban S Roy, 5th Aug, 2025

Prepared by Sanjiban S Roy, 5th Aug, 2025