Decision Tree Classifier

Category - Supervised Learning

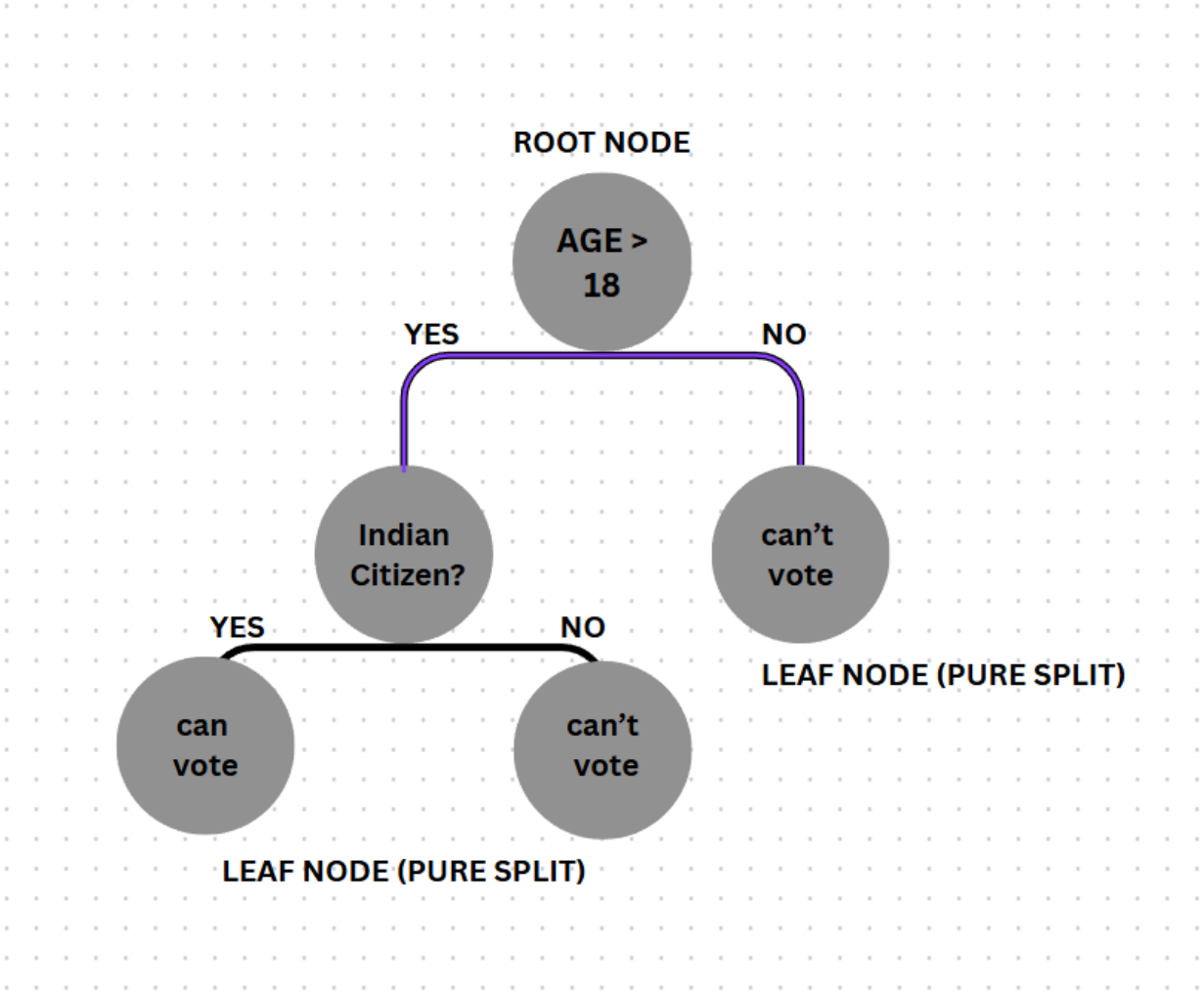
Class - Classification and Regression

Author - Siddhant Nautiyal

Credits - ChatGPT-4-turbo for mathematics, Canva for illustrations, Kaggle for the dataset, and Notion for notetaking.

Introduction

Decision Trees work on identifying the root node and creating a tree to traverse from top to bottom to determine the right class.



The idea behind the algorithm is to determine the root node and branch out until leaf nodes with pure splits are derived.

Criterion

Decision trees have various criteria to determine the best ways to split. The most important criteria are **Information Gain**, **Gini Index**, and **Entropy**.

Gini Index and Entropy

The Gini Index and Entropy are **two measures used in decision tree algorithms to evaluate the impurity or randomness in a dataset**. The Gini Index is a linear measure that calculates the probability of misclassifying a randomly chosen element in a set, while Entropy is a logarithmic measure that quantifies the amount of uncertainty or randomness in a set.

Entropy Formula:

$$H(S) = -(p_+ \log_2 p_+ + p_- \log_2 p_-)$$

where:

- p_+ is the probability of the positive class.
- p_- is the probability of the negative class.

Gini Impurity Formula:

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

For binary classification ($n = 2$), it simplifies to:

$$GI = 1 - (p_+^2 + p_-^2)$$

where:

- p_+ is the probability of the positive class.
- p_- is the probability of the negative class.

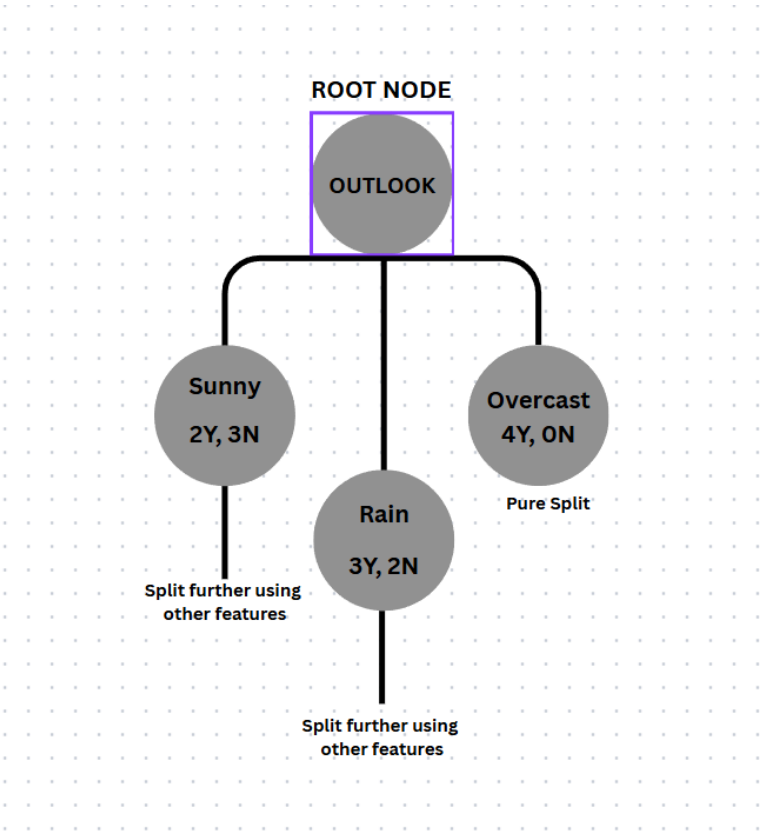
Information Gain

Information Gains help in determining the best root node.

	Outlook	Temperature	Humidity	Wind	Play Tennis
1					
2	Sunny	Hot	High	Weak	No
3	Sunny	Hot	High	Strong	No
4	Overcast	Hot	High	Weak	Yes
5	Rain	Mild	High	Weak	Yes
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Cool	Normal	Strong	No
8	Overcast	Cool	Normal	Strong	Yes
9	Sunny	Mild	High	Weak	No
10	Sunny	Cool	Normal	Weak	Yes
11	Rain	Mild	Normal	Weak	Yes
12	Sunny	Mild	Normal	Strong	Yes
13	Overcast	Mild	High	Strong	Yes
14	Overcast	Hot	Normal	Weak	Yes
15	Rain	Mild	High	Strong	No

DATASET: test-play-tennis SOURCE: kaggle

Consider “**outlook**” as the root node.



Step 1: Count Total Samples

The dataset consists of **14 instances**, where:

- **"Yes" (Play Tennis) = 9**
- **"No" (Play Tennis) = 5**

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

$$= -\left(0.643 \times -0.637 + 0.357 \times -1.485\right)$$

$$= -(-0.41 - 0.53) = 0.94$$

$$Gini(S) = 1 - \left(\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right)$$

$$= 1 - (0.413 + 0.127) = 1 - 0.54 = 0.46$$

Step 2: Calculate Entropy and Gini for Each Outlook Category

We split the dataset based on "Outlook" into **Sunny, Overcast, and Rain**.

Sunny Subset:

Outlook	Play Tennis
Sunny	No
Sunny	No
Sunny	No
Sunny	Yes
Sunny	Yes

Total: **5 instances**

- "Yes" = **2**, "No" = **3**

$$Entropy(Sunny) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right)$$

$$= -(0.4 \times -1.322 + 0.6 \times -0.737) = 0.971$$

$$Gini(Sunny) = 1 - ((2/5)^2 + (3/5)^2) = 1 - (0.16 + 0.36) = 0.48$$

Overcast Subset:

Outlook	Play Tennis
Overcast	Yes
Overcast	Yes
Overcast	Yes
Overcast	Yes

Total: **4 instances**

- "Yes" = **4**, "No" = **0**

$$Entropy(Overcast) = -\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right)$$

$$= -(1 \times 0 + 0) = 0$$

$$Gini(Overcast) = 1 - ((4/4)^2 + (0/4)^2) = 1 - (1 + 0) = 0$$

Rain Subset:

Outlook	Play Tennis
Rain	Yes
Rain	Yes
Rain	No
Rain	Yes
Rain	No

Total: **5 instances**

- "Yes" = **3**, "No" = **2**

$$Entropy(Rain) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)$$

$$= -(0.6 \times -0.737 + 0.4 \times -1.322) = 0.971$$

$$Gini(Rain) = 1 - ((3/5)^2 + (2/5)^2) = 1 - (0.36 + 0.16) = 0.48$$

Step 3: Calculate Weighted Entropy and Gini Index

Weighted Entropy

$$Entropy_{Outlook} = \left(\frac{5}{14} \times 0.971\right) + \left(\frac{4}{14} \times 0\right) + \left(\frac{5}{14} \times 0.971\right)$$
$$= (0.346) + (0) + (0.346) = 0.69$$

Weighted Gini Index

$$Gini_{Outlook} = \left(\frac{5}{14} \times 0.48\right) + \left(\frac{4}{14} \times 0\right) + \left(\frac{5}{14} \times 0.48\right)$$
$$= (0.171) + (0) + (0.171) = 0.34$$

CONCLUSION:

- **Entropy(Outlook) = 0.69**
- **Gini Index(Outlook) = 0.34**

This means
"Outlook" is a good choice for the root node as it provides a significant information gain

Mathematical Observations

- Entropy ranges between $(0, 1)$ for a binary classification problem, 0 being a perfect split and 1 being the most impure split.
- For a non-binary classification, it will range between $(0, log_2 C)$ where **C** is the number of classes.
- Gini Impurity ranges between $(0, 0.5)$ for a binary classification problem.
- For a non-binary classification, it will range between $(0, 1 - 1/C)$ where **C** is the number of classes.

- It is better to use GI since it does not require the computation of logarithmic functions.