Decision Tree

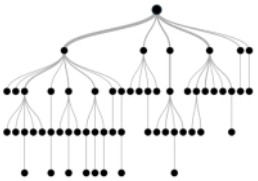
13 October 2025 22:32

1 Introduction

- A **Decision Tree** is a supervised learning algorithm.

Decision Tree → classification
→ Regression

- It splits data into branches based on **feature values** to make predictions, resembling a flowchart or tree structure.



Types of Decision Tree:

1. ID3
2. CART

Key idea:

- At each step, it chooses the **feature and threshold** that best separates the data into pure (homogeneous) subsets.

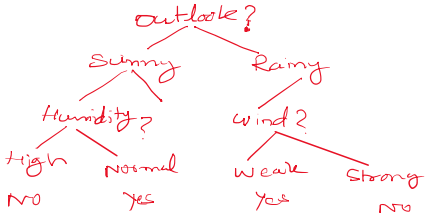
2 Why Decision Trees?

- ☑ Easy to understand and interpret
- ☑ No need for feature scaling
- ☑ Works with numerical and categorical data
- ☑ Can handle non-linear relationships
- ⚠ But:
- ✗ Can overfit easily
- ✗ Sensitive to small changes in data
- ✗ Greedy nature — may not find the global optimum

3 Anatomy of a Decision Tree

- **Root Node:** The first node where splitting starts (entire dataset).
- **Internal Nodes:** Nodes where the data is split further.
- **Leaf Nodes (Terminal Nodes):** Represent class labels (for classification) or predicted values (for regression).

Example (Classification):



4 How Does a Decision Tree Work?

It recursively partitions data using the **best feature** based on a chosen **splitting criterion**.

For Classification:`

Common metrics:

1. Entropy and Gini Impurity → pure split
2. Information Gain → Feature Section

For Regression:

Metric: Variance Reduction (mean Squared Error Reduction)

1. Entropy

Entropy measures the impurity or randomness:

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

Example : Play Tennis Examples Dataset

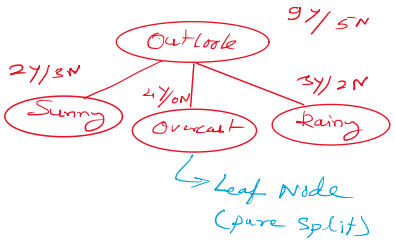
	Outlook	Temperature	PlayTennis
1	Sunny	Hot	No
2	Sunny	Hot	No
3	Overcast	Hot	Yes
4	Rainy	Mild	Yes
5	Rainy	Cool	Yes
6	Rainy	Cool	No
7	Overcast	Cool	Yes
8	Sunny	Mild	No
9	Sunny	Cool	Yes
10	Rainy	Mild	Yes
11	Sunny	Mild	Yes
12	Overcast	Mild	Yes
13	Overcast	Hot	Yes
14	Rainy	Mild	No

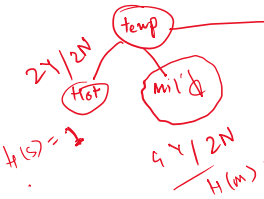
Frequency tables

⇒ Outlook

	Yes	No
Sunny	2	3
Overcast	4	0
Rainy	3	2
	9	5

Temperature	Y	N
Hot	2	2
mild	4	2
cool	3	1
	9	5 = 14





$$H(Hot) = -\left[\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right]$$
$$= -\left[\frac{1}{2}\log_2 1/2 + \frac{1}{2}\log_2 1/2\right]$$
$$= -[0.5 \times -1 + 0.5 \times -1]$$
$$= -[-0.5 - 0.5)$$
$$= 1$$

At Node Sunny:

$$\text{Entropy } H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$
$$= -(p_y \log p_y + p_N \log_2 p_N)$$
$$= -\left[\frac{2}{5}\log_2 2/5 + \frac{3}{5}\log_2 3/5\right]$$
$$= 0.159 + 0.132$$
$$= 0.291$$

Entropy of Node Overcast (4Y/0N)

$$H(ovc) = -\frac{4}{4}\log 4/4 - \frac{0}{4}\log 0/4$$
$$= 0$$

Entropy of Node Rainy (3Y/2N)

$$H(Rainy) = -\frac{3}{5}\log_2 3/5 - \frac{2}{5}\log_2 2/5$$
$$= 0.291$$

(2) Gini Impurity

$$G.I. = 1 - \sum_{i=1}^{c} (p_i)^2$$

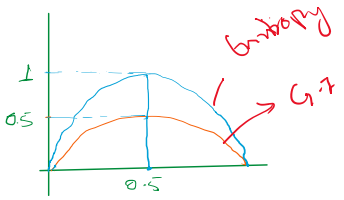
Rainy

3Y/2N

$$G.I. = 1 - [(p_y)^2 + (p_N)^2]$$
$$= 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]$$

$= 1 - [0.36 + 0.16]$

$= 1 - 0.44 = 0.56$



34/BD

$H(S) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$

$= 1$

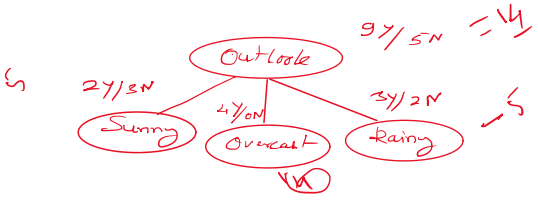
$G.I. = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2$

$= 1 - \frac{1}{4} - \frac{1}{4} = 0.5$

(b) Information Gain

$$\text{Information Gain } (IG) = H(S)_{\text{parent}} - \sum_{k \in \text{children}} \frac{n_k}{n} \times H(S)_k$$

Goal: Maximize Information Gain.



$H(\text{Parent}) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

$= -(0.643)(-0.637) - (0.357)(-1.485)$

$= 0.4095 + 0.5301$

$= 0.94$

$H(\text{Sunny}) = 0.291$

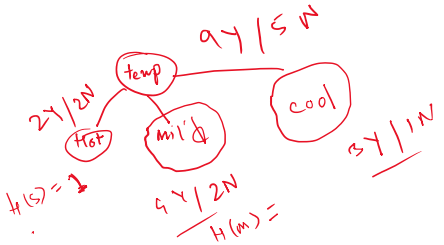
$H(\text{Outlook}) = 0$

$H(\text{Rainy}) = 0.291$

$\text{Information Gain} = 0.94 - \left[\frac{5}{14} \times 0.291 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.291\right]$

$= 0.94 - 0.2078$

$= 0.752$



H(Hot) = 1

H(mild) = 0.918

H(cool) = 0.8

H(Temp) = 0.94

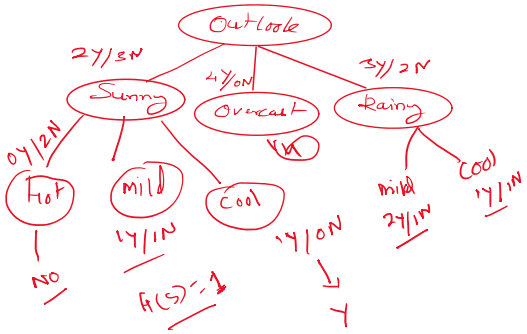
$$\text{Info Gain} = 0.94 - \left[\frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.8\right]$$

$$= 0.94 - [0.28 + 0.42 * 0.918 + 0.28 \times 0.8]$$

$$= 0.932$$

I.G.(Outlook) = 0.732 ✓

I.G.(Temp) = 0.932



	Outlook	Temperature	PlayTennis
1	Sunny	Hot	No
2	Sunny	Hot	No
3	Overcast	Hot	Yes
4	Rainy	Mild	Yes
5	Rainy	Cool	Yes
6	Rainy	Cool	No
7	Overcast	Cool	Yes
8	Sunny	Mild	No
9	Sunny	Cool	Yes
10	Rainy	Mild	Yes
11	Sunny	Mild	Yes
12	Overcast	Mild	Yes
13	Overcast	Hot	Yes
14	Rainy	Mild	No