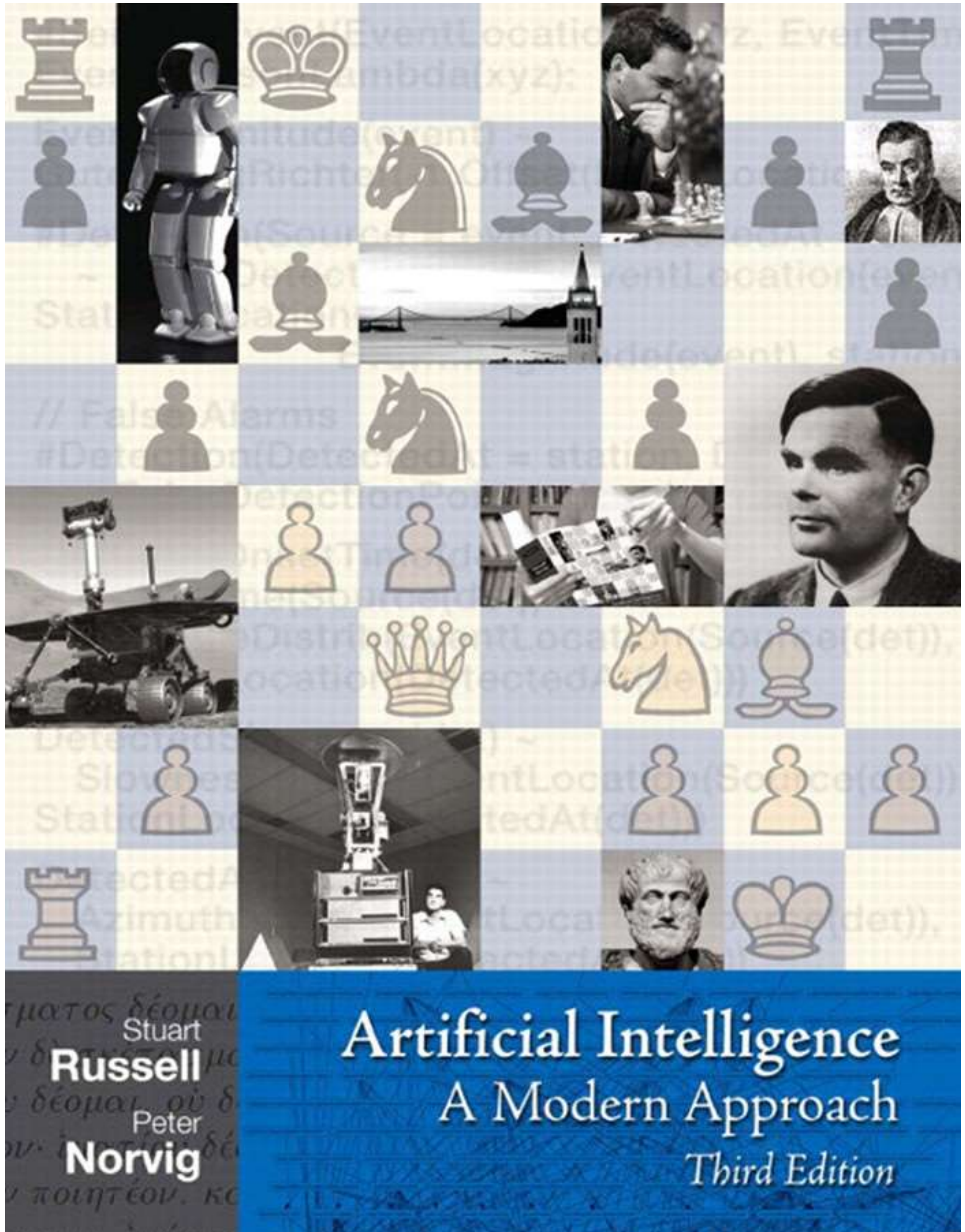
NEIL GOGTE INSTITUTE OF TECHNOLOGY
&
KESHAV MEMORIAL ENGINEERING COLLEGE

ARTIFICIAL INTELLIGENCE
(PC 502CSM)

Deepika.M

Assistant Professor, CSE(AIML)

NGIT


Stuart
Russell
Peter
Norvig
Artificial Intelligence
A Modern Approach
Third Edition

UNIT-IV

Learning

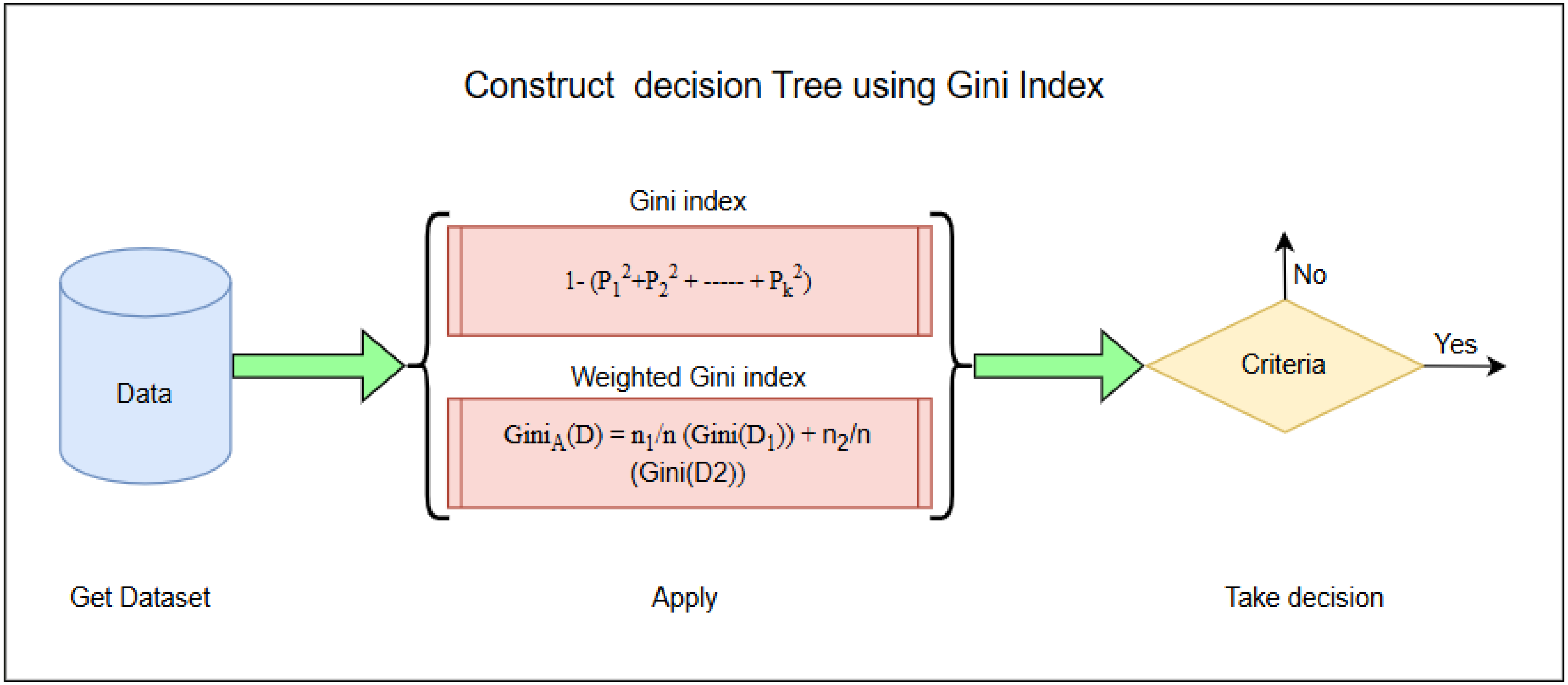
DECISION TREES

Gini index

- The **Gini index** measures impurity or inequality frequently used in decision tree algorithms.
- It quantifies the probability of misclassifying a randomly chosen element if it were randomly labeled according to the distribution of labels in a particular node.
- The equation for the Gini index is as follows:

$$GiniIndex = 1 - (p_1^2 + p_2^2 + ... + p_k^2)$$

where *p*1, *p*2, ..., *pk* are the probabilities of each class in the node.



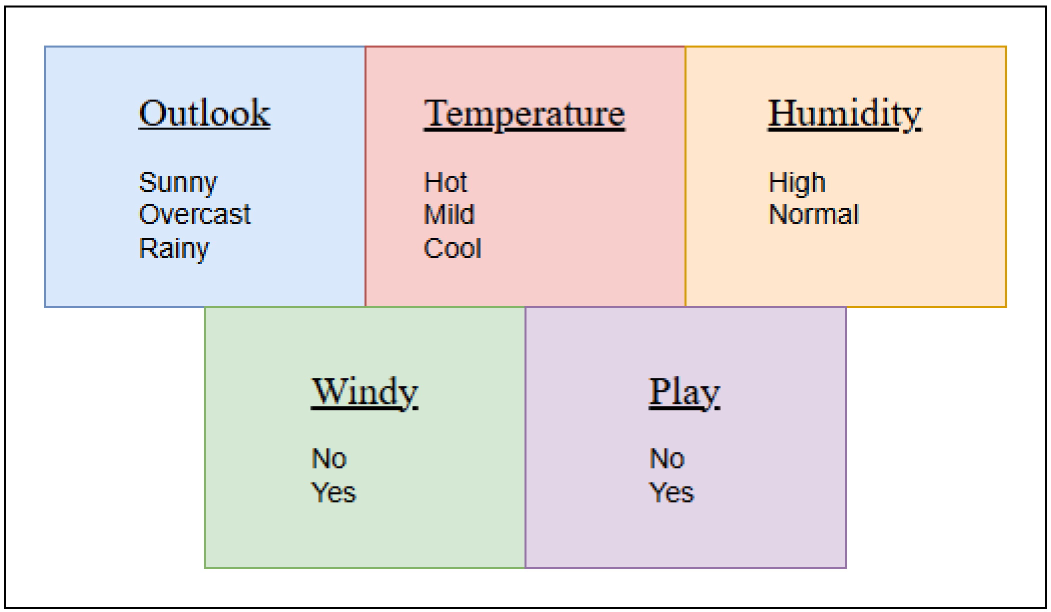
Gini Index visualization

- The ***GiniA*****(*D*)** represents the weighted Gini index for the entire dataset ***D***. It's a measure of impurity or inequality in the dataset, considering the weighted average of the impurities of two subsets, *D*1 and *D*2.
- *n*1: This is the number of instances (data points) in subset *D*1.
- *n*2: This is the number of instances (data points) in subset *D*2.
- *n*: The total number of instances in the entire dataset ***D*****(*n*=*n*1+*n*2**).
- *Gini*(*D*1): This is the Gini index of subset *D*1, which quantifies the impurity or uncertainty of class labels in D1. A lower Gini index indicates higher purity.
- *Gini*(*D*2): This is the Gini index of subset *D*2, similar to Gini(D1)*Gini*(*D*1), but for the other subset.

Dataset

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Overcast	Hot	High	No	Yes
Rainy	Mild	High	No	Yes
Rainy	Cool	Normal	No	Yes
Rainy	Cool	Normal	Yes	No
Overcast	Cool	Normal	Yes	Yes
Sunny	Mild	High	No	No
Sunny	Cool	Normal	No	Yes
Rainy	Mild	Normal	No	Yes
Sunny	Mild	Normal	Yes	Yes
Overcast	Mild	High	Yes	Yes
Overcast	Hot	Normal	No	Yes
Rainy	Mild	High	Yes	No

- To calculate the Gini index for each attribute and construct a decision tree, we'll start by analyzing the given data and calculating the Gini index for each attribute at the first step. **We have four attributes in the above dataset:**



Attributes to calculate gini index

Steps to construct a decision tree

Step 1

Calculate Gini index for each attribute

Calculate weighted Gini index for each attribute

Calculate Gini index for **Outlook**

For Sunny:

- Play=No count: 3
- Play=Yes count: 2
- Gini index for Sunny:
  - $= 1 - (2/5)^2 - (3/5)^2$
  - $= 1 - 4/25 - 9/25$
  - $= 1 - 13/25$
  - $= 12/25$

For Overcast:

- Play=No count: 0
- Play=Yes count: 4
- Gini index for Overcast:
  - $= 1 - (0/4)^2 - (4/4)^2$
  - $= 1 - 0/16 - 16/16$
  - $= 0$

For Rainy:

- Play=No count: 3
- Play=Yes count: 2
- Gini index for Rainy:
  - $= 1 - (3/5)^2 - (2/5)^2$
  - $= 1 - 9/25 - 4/25$
  - $= 12/25$

Calculate weighted Gini index for **Outlook**

$$(5/14) * (12/25) + (4/14) * 0 + (5/14) * (12/25) = 0.342$$

Calculate Gini index for **Windy**

For No:

- Play=No count: 2
- Play=Yes count: 6
  - $= 1 - (6/8)^2 - (2/8)^2 = 0.375$

For Yes:

- Play=No count: 3
- Play=Yes count: 3
  - $= 1 - (3/6)^2 - (3/6)^2 = 0.5$

Calculate weighted Gini index for **Windy**

$$(8/14) * (3/8) + (6/14) * (1/2) = 0.428$$

Calculate Gini index for **Temperature**

For Hot:

- Play=No count: 2
- Play=Yes count: 2
- Gini index for Hot:
  - $= 1 - (2/4)^2 - (2/4)^2 = 0.5$

For Mild:

- Play=No count: 2
- Play=Yes count: 4
- Gini index for Mild:
  - $= 1 - (2/6)^2 - (4/6)^2 = 4/9$

For Cool:

- Play=No count: 1
- Play=Yes count: 3
- Gini index for Cool:
  - $= 1 - (1/4)^2 - (3/4)^2 = 0.375$

Calculate weighted Gini index for **Temperature**

$$(4/14) * 0.5 + (6/14) * (4/9) + (4/14) * (0.375) = 0.4404$$

Calculate Gini index for **Humidity**

For High:

- Play=No count: 4
- Play=Yes count: 3
  - $= 1 - (3/7)^2 - (4/7)^2 = 0.4898$

For Normal:

- Play=No count: 1
- Play=Yes count: 6
  - $= 1 - (6/7)^2 - (1/7)^2 = 0.2449$

Calculate weighted Gini index for **Humidity**

$$(7/14) * 0.4898 + (7/14) * 0.2449 = 0.2449 + 0.2449 = 0.4898$$

Step 2

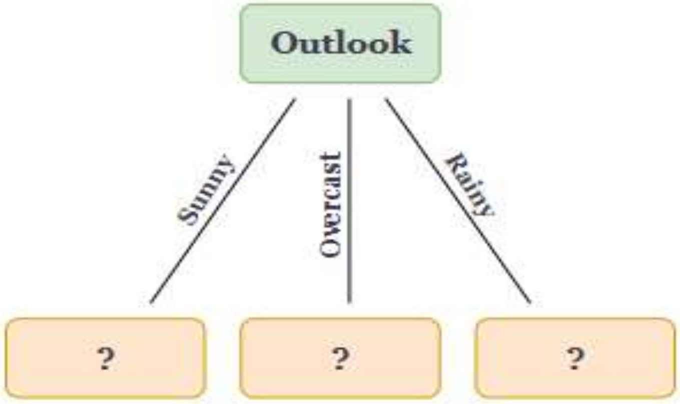
Take decision base on calculated result for root node.

Take a decision base on the calculated result for the root node

Now we have the Gini index calculations for each attribute at the first step:

- Outlook: 0.3429
- Temperature: 0.4404
- Humidity: 0.4898
- Windy: 0.4286

The attribute with the lowest Gini index is Outlook, so it would be selected as the root of the decision tree in the next step.



Step 3

Extract the dataset under the selected root node for each subtree.

Extract the dataset under the selected root node for each subtree.

- Outlook -> Sunny

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Sunny	Mild	High	No	No
Sunny	Cool	Normal	No	Yes
Sunny	Mild	Normal	Yes	Yes

- Outlook -> Overcast

Outlook	Temperature	Humidity	Windy	Play
Overcast	Hot	High	No	Yes
Overcast	Cool	Normal	Yes	Yes
Overcast	Mild	High	Yes	Yes
Overcast	Hot	Normal	No	Yes

- Outlook -> Rainy

Outlook	Temperature	Humidity	Windy	Play
Rainy	Mild	High	No	Yes
Rainy	Cool	Normal	No	Yes
Rainy	Cool	Normal	Yes	No
Rainy	Mild	Normal	No	Yes
Rainy	Mild	High	Yes	No

Step 4

Repeat **Step1**, **Step2** and **Step3** for each subtree until we reach the leaf node

Here we have three sub branches:

- Sunny
- Overcast
- Rainy

After repeating step1, step2 and step3, we will find these calculated results for leaf node

Outlook -> Sunny

- Temperature: 0.44
- Humidity: 0
- Windy: 0.44

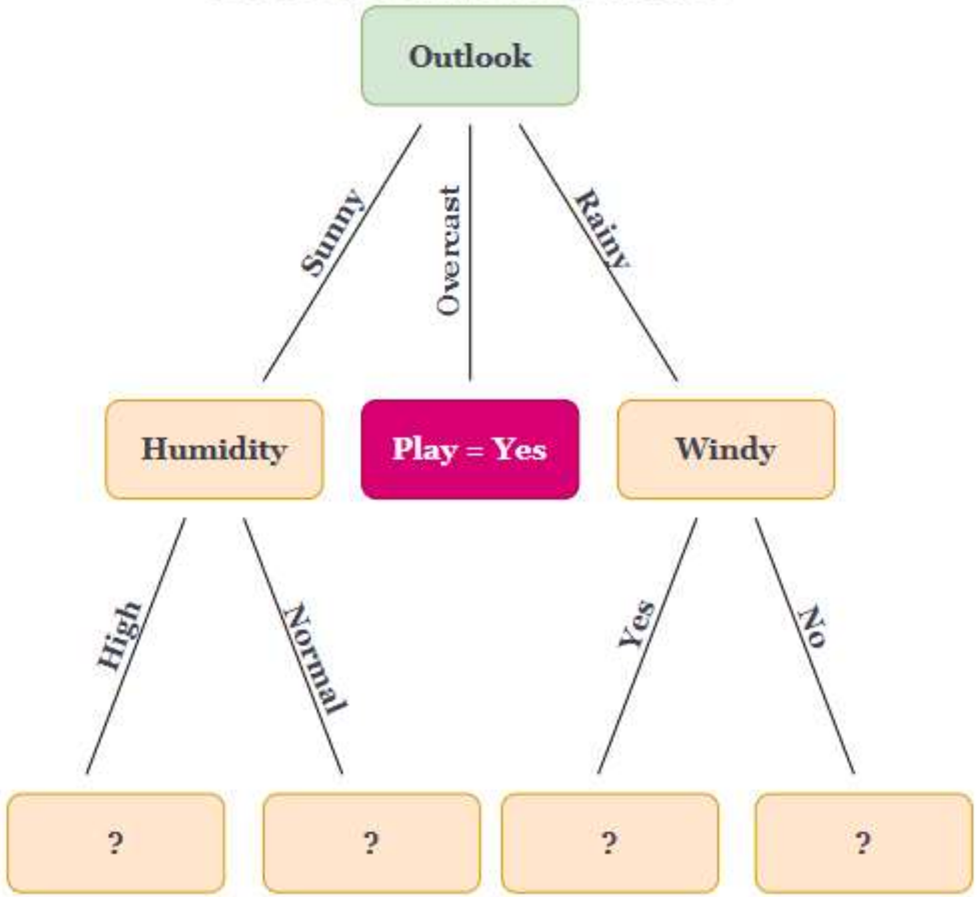
Outlook -> Overcast

- Temperature: 0
- Humidity: 0
- Windy: 0

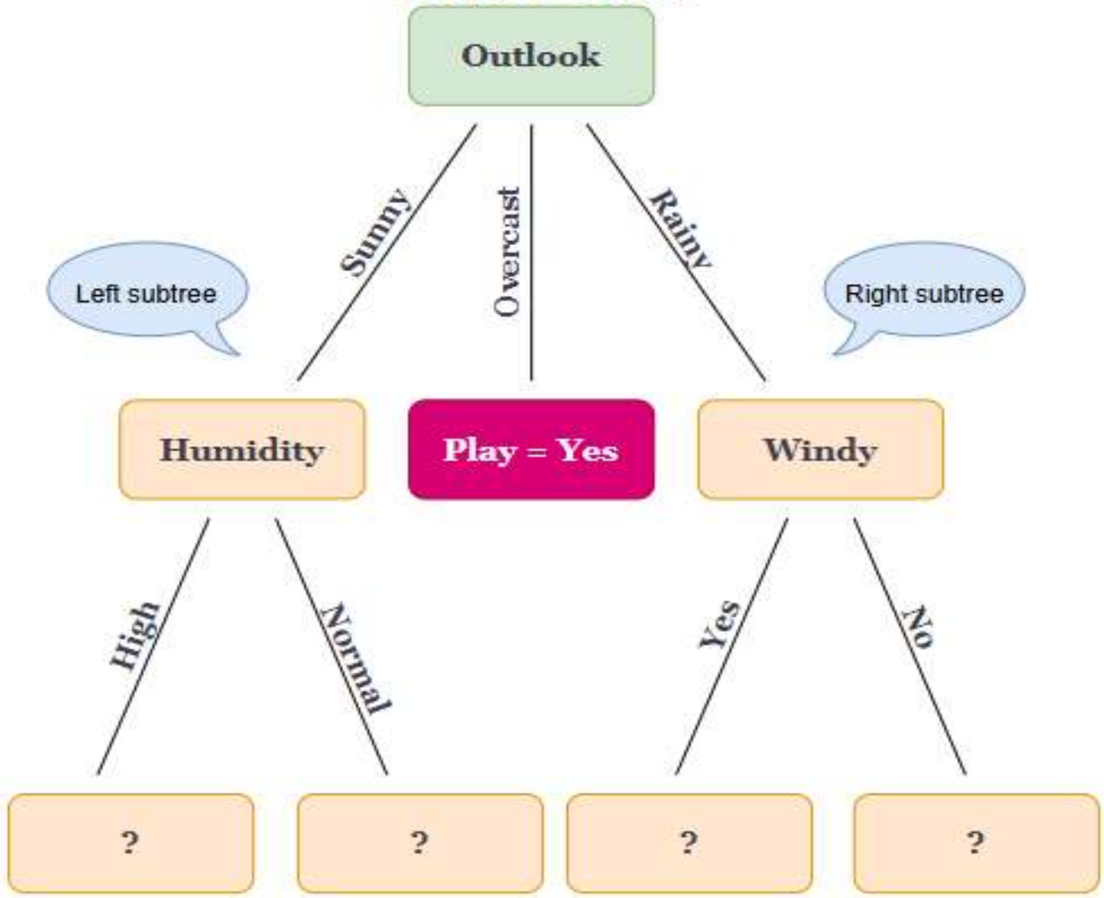
Outlook -> Rainy

- Temperature: 0.464
- Humidity: 0.464
- Windy: 0

Tree at this moment



Repeat the same steps for the subtrees



Extract the dataset under the selected root node for each attribute.

- Humidity -> High

Humidity	Temperature	Windy	Play
High	Hot	No	No
High	Hot	Yes	No
High	Mild	No	No

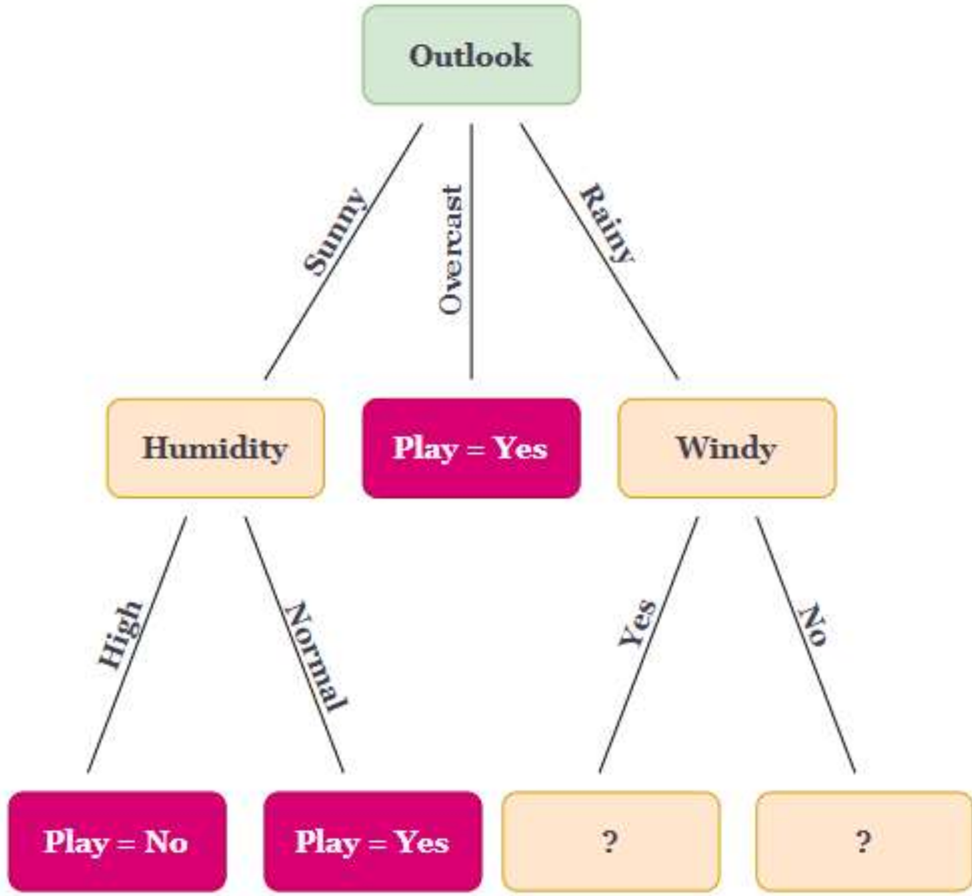
- Humidity -> Normal

Humidity	Temperature	Windy	Play
Normal	Cool	No	Yes
Normal	Mild	Yes	Yes

We can repeat step1 , step2 and step3 for above dataset of we can observe that for every case under

- Humidity -> High
    - Play= No
- Humidity -> Normal
    - Play = Yes

Tree at this moment



Extract the dataset under the selected root node for each attribute.

- Windy -> Yes

Windy	Temperature	Humidity	Play
Yes	Mild	High	No
Yes	Cool	Normal	No
Yes	Mild	Normal	No

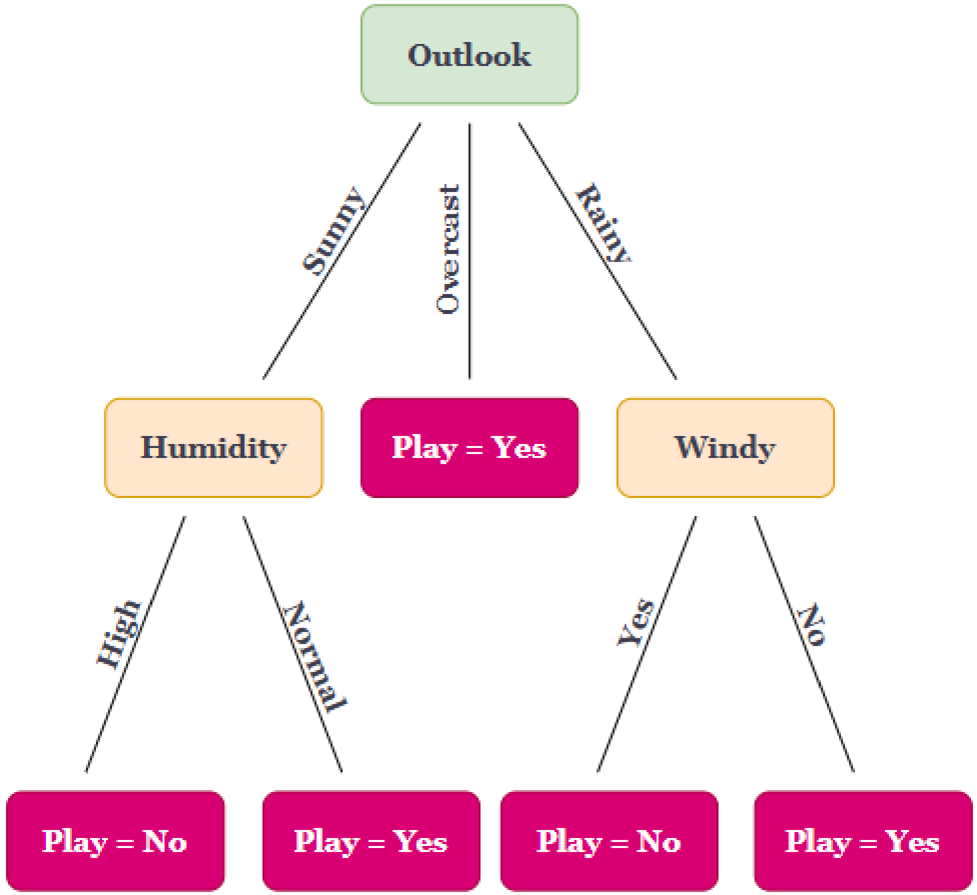
- Windy -> No

Windy	Temperature	Humidity	Play
No	Mild	High	Yes
No	Cool	Normal	Yes
No	Mild	Normal	Yes

We can repeat step1 , step2 and step3 for above dataset of we can observe that for every case under

- Windy -> Yes
  - Play= No
- Windy -> No
  - Play = Yes

Final Tree


Outlook
Sunny
Overcast
Rainy
Humidity
Play = Yes
Windy
High
Normal
Yes
No
Play = No
Play = Yes
Play = No
Play = Yes

- By leveraging the Gini index, which measures the impurity of a node, we were able to determine the best splitting criteria for creating an effective decision tree model.
- This approach allowed us to make informed decisions based on the purity and predictive power of each node in the tree.
- The Gini index offers a valuable tool for decision tree construction, enabling us to efficiently handle categorical and numerical features.