SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

Redesigning Common Mind & Business Towards Excellence

3Purpose Process People Culture

www.snsgroups.com

1st GenAI Powered Design Thinking FrameWork

Build an Entrepreneurial Mindset Through Our Design Thinking FrameWork

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

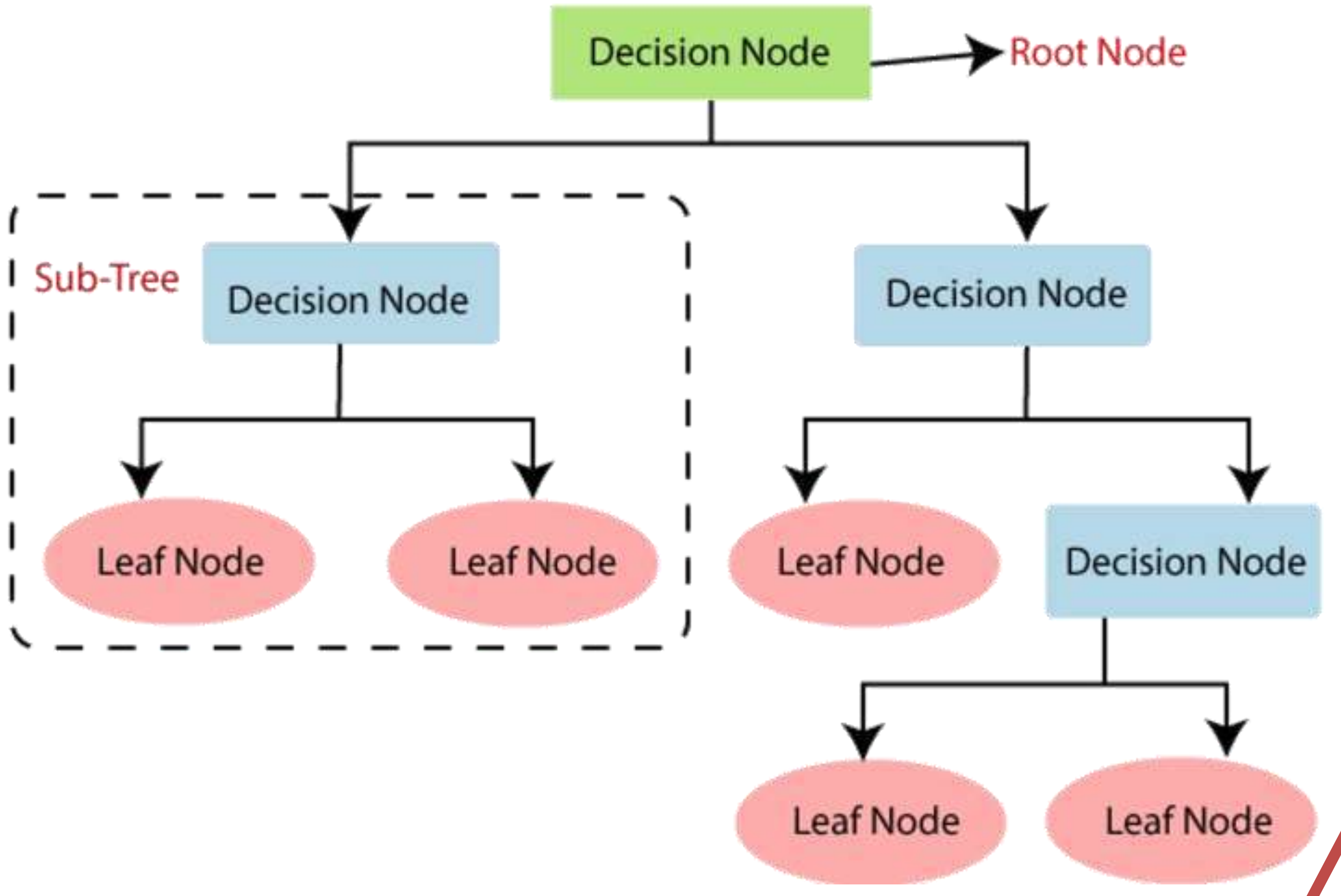
23AMB201 - MACHINE LEARNING

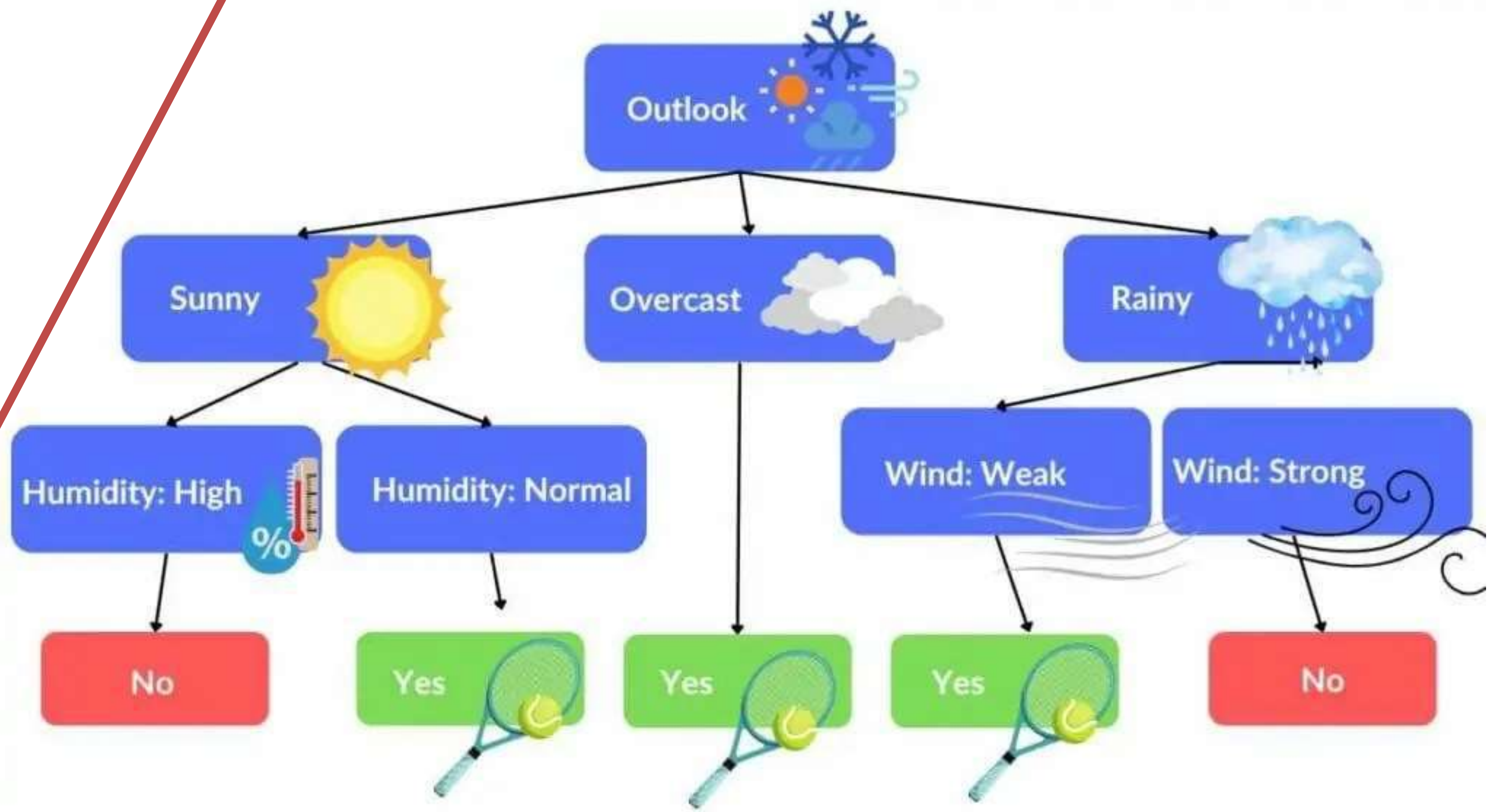
II YEAR IV SEM

UNIT III – GENERATIVE MODELS AND BOOSTING

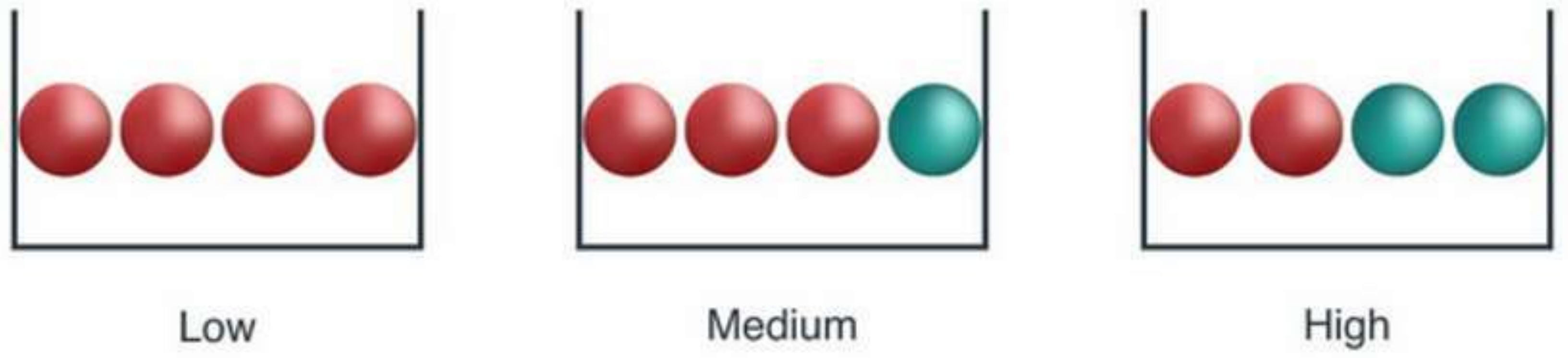
TOPIC 17 – Decision Tree – CART Algorithm

Decision tree





Why Decision Tree?



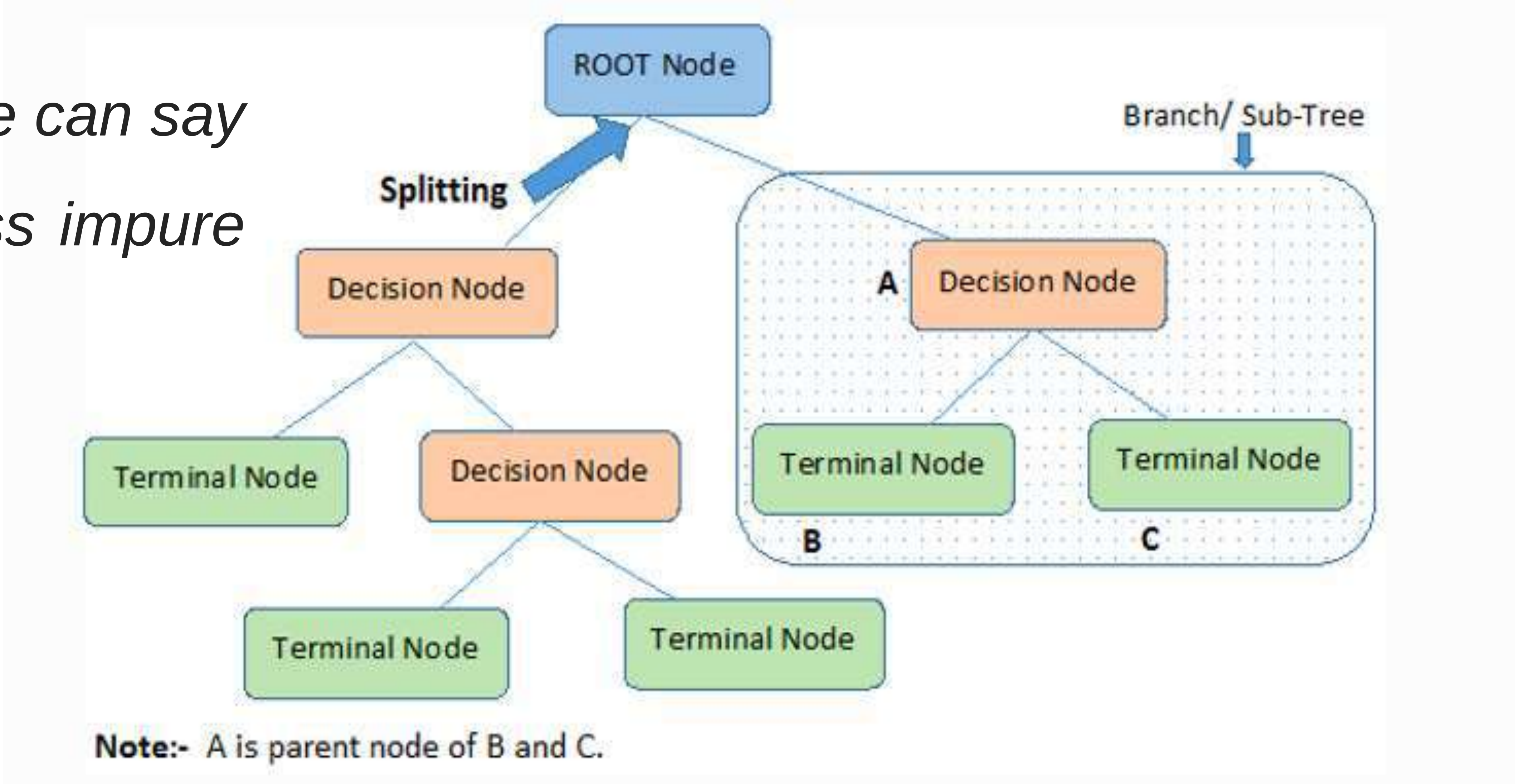
Decision tree algorithm is a tree where nodes represents features(attributes), branch represents decision(rule) and leaf nodes represents outcomes(discrete and continuous).

As information is measure of purity, so we can say that left bowl is pure node, middle is less impure and right is more impure.

Left Bowl: Purity

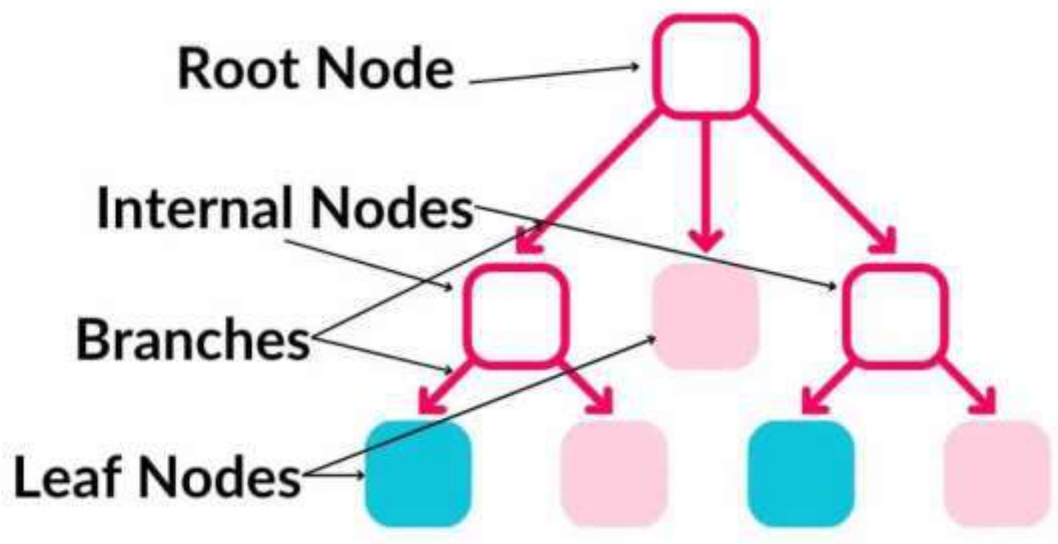
Middle Bowl: Less impurity

Right Bowl: High impurity



Components of Decision Tree

1. **Root Node:** The topmost node in a decision tree. It represents the entire dataset and is split into two or more homogeneous sets.
2. **Internal Nodes:** Nodes within the tree that represent decision points. Each internal node corresponds to an attribute test.
3. **Leaf Nodes (Terminal Nodes):** Nodes at the end of the branches, which provide the outcome of the decision path. For classification tasks, they represent class labels. For regression, tasks represent continuous values.
4. **Branches:** Arrows connecting nodes, representing the outcome of a test and leading to the next node or leaf.



Common Algorithms

1.Iterative Dichotomiser - ID 3

2.Classification and regression tree - CART

1.Iterative Dichotomiser - ID 3

1. Entropy – To find Uncertainty or impurity
2. Information Gain –Maximum Information

2. Dataset: Imbalanced Class

3. Want more information

4. Understanding information gain is crucial

5. Use case:

1. Text Classification
2. Medical Diagnosis

6. Algorithm: ID3, C4.5

1.Classification and regression tree – CART

1. Gini Index – To find Impurity

2. Faster and simpler splits

3. For Large Dataset

4. Simpler calculation

5. Classes are well balanced

6. Use Case:

1. Spam Detection
2. Credit Scoring
3. Fraud Detection

Difference between ID3 and CART

Criterion	Formula	Used In	Best When	Computational Complexity
Gini Index	$1 - \sum p_i^2$	CART	Faster, large datasets	Lower (No log calculations)
Entropy	$-\sum p_i \log_2 p_i$	ID3, C4.5	Imbalanced datasets, information gain	Higher (Log calculations involved)

👉 If speed matters → Use Gini

👉 If class imbalance is a concern → Use Entropy

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Gini = $1 - \Sigma (Pi)^2$ for i=1 to number of classes

Outlook

Outlook	Yes	No	# Instances
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

Gini index (outlook=sunny)= $1-(2/5)^2-(3/5)^2 = 1- 0.16–0.36 = 0.48$

Gini index(outlook=overcast)= $1- (4/4)^2-(0/4)^2 = 1- 1- 0 = 0$

Gini index(outlook=rainfall)= $1- (3/5)^2 -(2/5)^2 = 1- 0.36- 0.16 = 0.48$

Now , we will calculate the weighted sum of Gini index for outlook features,

Gini(outlook) = (5/14)*0.48 + (4/14) *0 + (5/14)*0.48 = 0.342

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Gini = $1 - \Sigma (Pi)^2$ for i=1 to number of classes

Temperature

Temperature	Yes	No	# Instances
hot	2	2	4
cool	3	1	4
mild	4	2	6

Gini(temperature=hot) = $1-(2/4)^2-(2/4)^2 = 0.5$

Gini(temperature=cool) = $1-(3/4)^2-(1/4)^2 = 0.375$

Gini(temperature=mild) = $1-(4/6)^2-(2/6)^2 = 0.445$

Now, the weighted sum of Gini index for temperature features can be calculated as,

Gini(temperature)= (4/14) *0.5 + (4/14) *0.375 + (6/14) *0.445 =0.439

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Gini = $1 - \Sigma (Pi)^2$ for i=1 to number of classes

Humidity

Humidity	Yes	No	# Instances
high	3	4	7
Normal	6	1	7

$\text{Gini(humidity=high)} = 1-(3/7)^2-(4/7)^2 = 0.489$

$\text{Gini(humidity=normal)} = 1-(6/7)^2-(1/7)^2 = 0.244$

Now, the weighted sum of Gini index for humidity features can be calculated as, $\text{Gini(humidity)} = (7/14) *0.489 + (7/14) *0.244=0.367$

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Gini = 1 – Σ $(Pi)^2$ for i=1 to number of classes

Wind

wind	Yes	No	# Instances
weak	6	2	8
strong	3	3	6

Gini(wind=weak)= $1-(6/8)^2-(2/8)^2 = 0.375$

Gini(wind=strong)= $1-(3/6)^2-(3/6)^2= 0.5$

Now, the weighted sum of Gini index for wind features can be calculated as,

Gini(wind) = (8/14) *0.375 + (6/14) *0.5=0.428

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

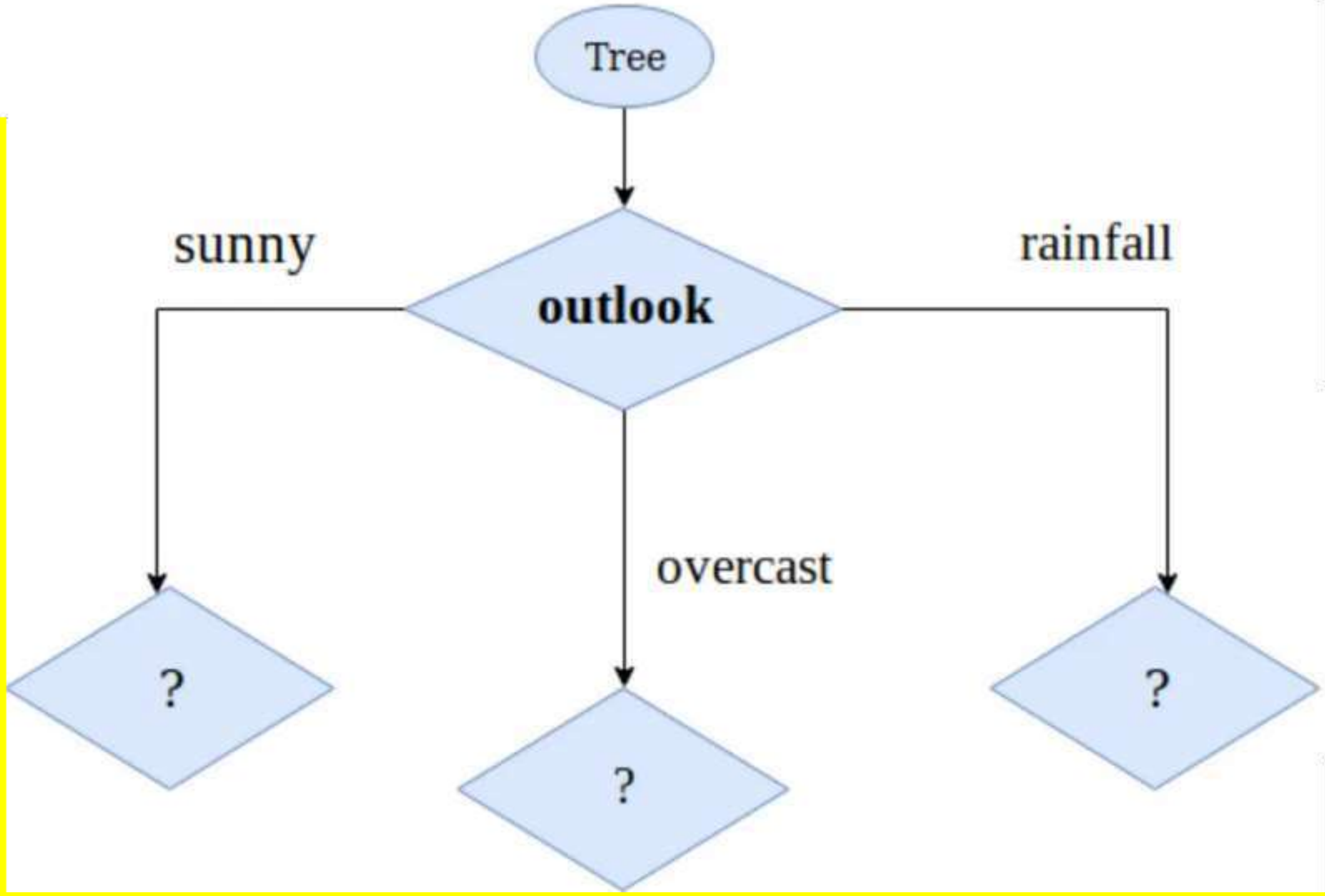
Decision for root node

Features	Gini Index
outlook	0.342
temperature	0.439
humidity	0.367
wind	0.428

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

From table, you can seen that Gini index for outlook feature is lowest. So w get our root node.



Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Gini of temperature for sunny outlook

Temperature	Yes	No	Number of instances
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

Gini(Outlook=Sunny and Temp.=Hot) = $1 - (0/2)^2 - (2/2)^2 = 0$

Gini(Outlook=Sunny and Temp.=Cool) = $1 - (1/1)^2 - (0/1)^2 = 0$

Gini(Outlook=Sunny and Temp.=Mild) = $1 - (1/2)^2 - (1/2)^2 = 1 - 0.25 - 0.25 = 0.5$

Gini(Outlook=Sunny and Temp.) = $(2/5)\times0 + (1/5)\times0 + (2/5)\times0.5 = 0.2$

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Gini Index for humidity on sunny outlook

Humidity	Yes	No	# Instances
high	0	3	3
Normal	2	0	2

Gini(outlook=sunny & humidity=high) = $1-(0/3)^2-(3/3)^2 = 0$

Gini(outlook=sunny & humidity=normal) = $1-(2/2)^2-(0/2)^2 = 0$

Now, the weighted sum of Gini index for humidity on sunny outlook features can be calculated as,

Gini(outlook = sunny & humidity) = (3/5) *0 + (2/5) *0=0

Use Case: Weather Dataset – Predict play or not

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Gini Index for wind on sunny outlook

wind	Yes	No	# Instances
weak	1	2	3
strong	1	1	2

Gini(outlook=sunny & wind=weak) = 1-(1/3)^2-(2/3)^2 = 0.44

Gini(outlook=sunny & wind=strong) = 1-(1/2)^2-(1/2)^2 = 0.5

Now, the weighted sum of Gini index for wind on sunny outlook features can be calculated as,

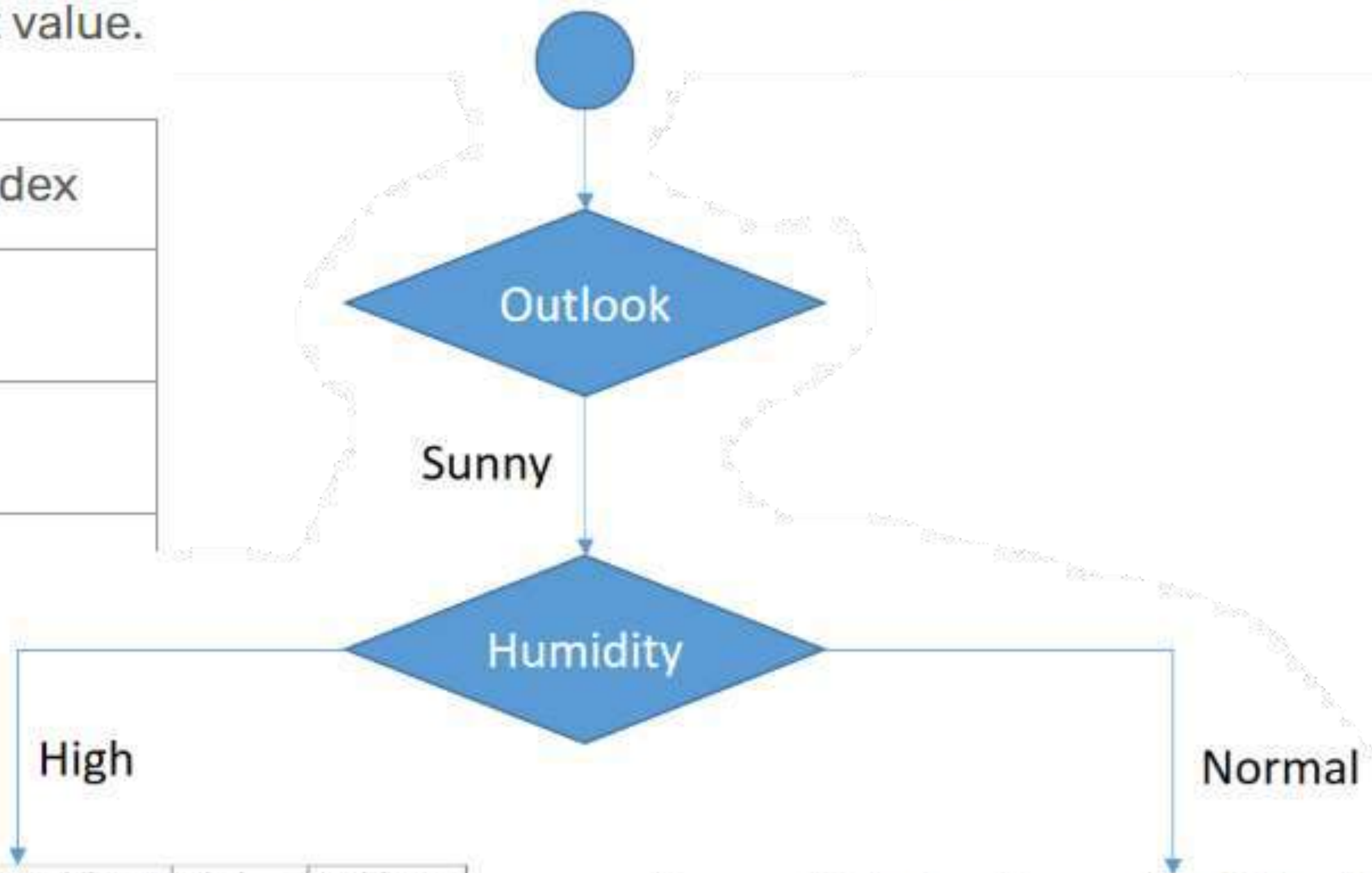
Gini(outlook = sunny and wind) = (3/5) *0.44 + (2/5) *0.5=0.266+0.2= 0.466

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Decision for sunny outlook

We've calculated gini index scores for feature when outlook is sunny. The winner is humidity because it has the lowest value.

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

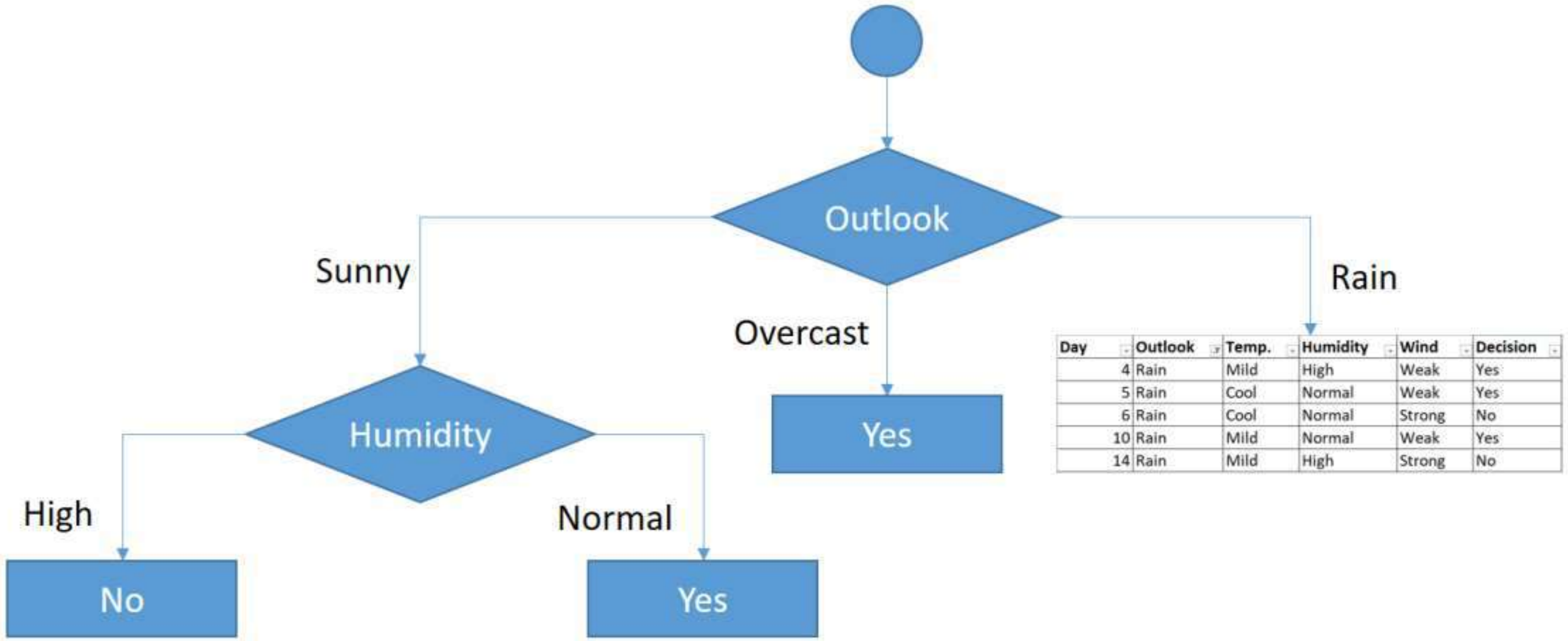


Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

Day	Outlook	Temp.	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Now,Lets focus on sub data for overcast outlook feature.

Day	outlook	temperature	humidity	wind	decision
3	overcast	hot	high	weak	Yes
7	overcast	cool	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes



As, you can see from the above table all the decision for overcast outlook feature is always ‘Yes’. Then Gini index for each feature is 0, means it is a leaf nodes.

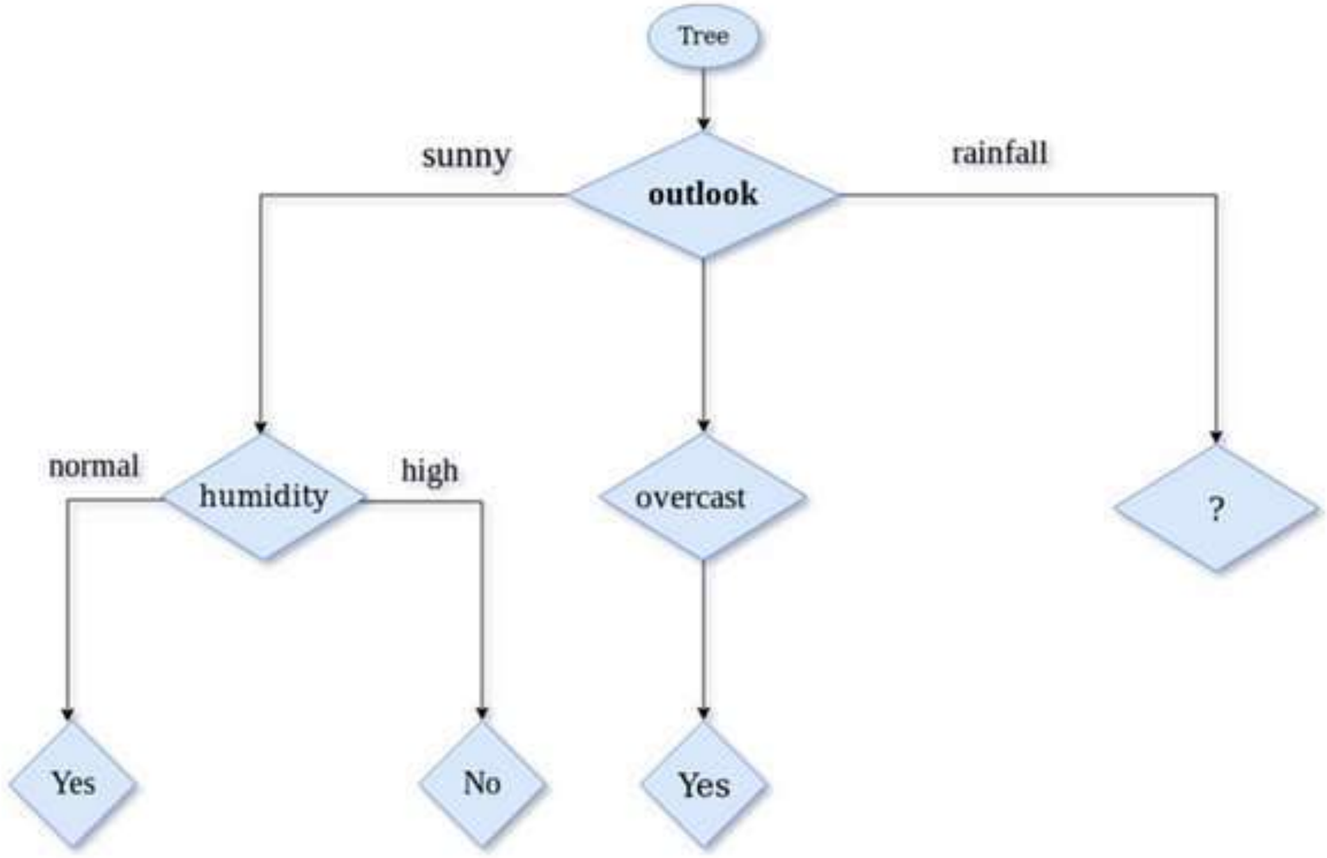
Use Case: Weather Dataset – Predict play or not

Now,Lets focus on sub data for high and normal humidity feature.

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No

Day	outlook	temperature	humidity	wind	decision
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

From the given two table, the decision is always 'No' when humidity is 'high' and decision is always 'Yes' when humidity is 'normal'. So we got leaf node. now decision tree can be viewed as,



Use Case: Weather Dataset – Predict play or not

Now,Lets focus on sub data for rainfall outlook feature.

we need to find the Gini index for temperature,humidity and wind feature respectively.

Day	outlook	temperature	humidity	wind	Decision
4	rain	mild	high	weak	Yes
5	rain	cool	normal	weak	Yes
6	rain	cool	normal	strong	No
10	rain	mild	normal	weak	Yes
14	rain	mild	high	strong	No

Gini index for temperature for rainfall outlook

temperature	Yes	No	# Instances
cool	1	1	2
mild	2	1	3

Gini(outlook=rainfall and temp.=Cool) = 1 — (1/2)2 — (1/2)2 = 0.5

Gini(outlook=rainfall and temp.=Mild) = 1 — (2/3)2 — (1/3)2 = 0.444

Gini(outlook=rainfall and temp.) = (2/5)*0.5 + (3/5)*0.444 = 0.466

Use Case: Weather Dataset – Predict play or not

Now,Lets focus on sub data for rainfall outlook feature.

we need to find the Gini index for temperature,humidity and wind feature respectively.

Day	outlook	temperature	humidity	wind	Decision
4	rain	mild	high	weak	Yes
5	rain	cool	normal	weak	Yes
6	rain	cool	normal	strong	No
10	rain	mild	normal	weak	Yes
14	rain	mild	high	strong	No

Gini index for humidity for rainfall outlook

humidity	Yes	No	# Instances
high	1	1	2
normal	2	1	3

Gini(outlook=rainfall and humidity=high) = 1 — (1/2)2 — (1/2)2 = 0.5

Gini(outlook=rainfall and humidity=normal) = 1 — (2/3)2 — (1/3)2 = 0.444

Gini(Outlook=rainfall and humidity) = (2/5)*(0.5 + (3/5)*0.444 = 0.466

Use Case: Weather Dataset – Predict play or not

Now,Lets focus on sub data for rainfall outlook feature.

we need to find the Gini index for temperature,humidity and wind feature respectively.

Day	outlook	temperature	humidity	wind	Decision
4	rain	mild	high	weak	Yes
5	rain	cool	normal	weak	Yes
6	rain	cool	normal	strong	No
10	rain	mild	normal	weak	Yes
14	rain	mild	high	strong	No

Gini index for wind for rainfall outlook feature

wind	Yes	No	# Instances
weak	3	0	3
strong	0	2	2

Gini(outlook=rainfall and wind=weak) = 1 — (3/3)2 — (0/3)2 = 0

Gini(outlook=rainfall and wind=strong) = 1 — (0/2)2 — (2/2)2 = 0

Gini(outlook=rainfall and wind) = (3/5)*0 + (2/5)*0 = 0

Use Case: Weather Dataset – Predict play or not

Now,Lets focus on sub data for rainfall outlook feature.

we need to find the Gini index for temperature,humidity and wind feature respectively.

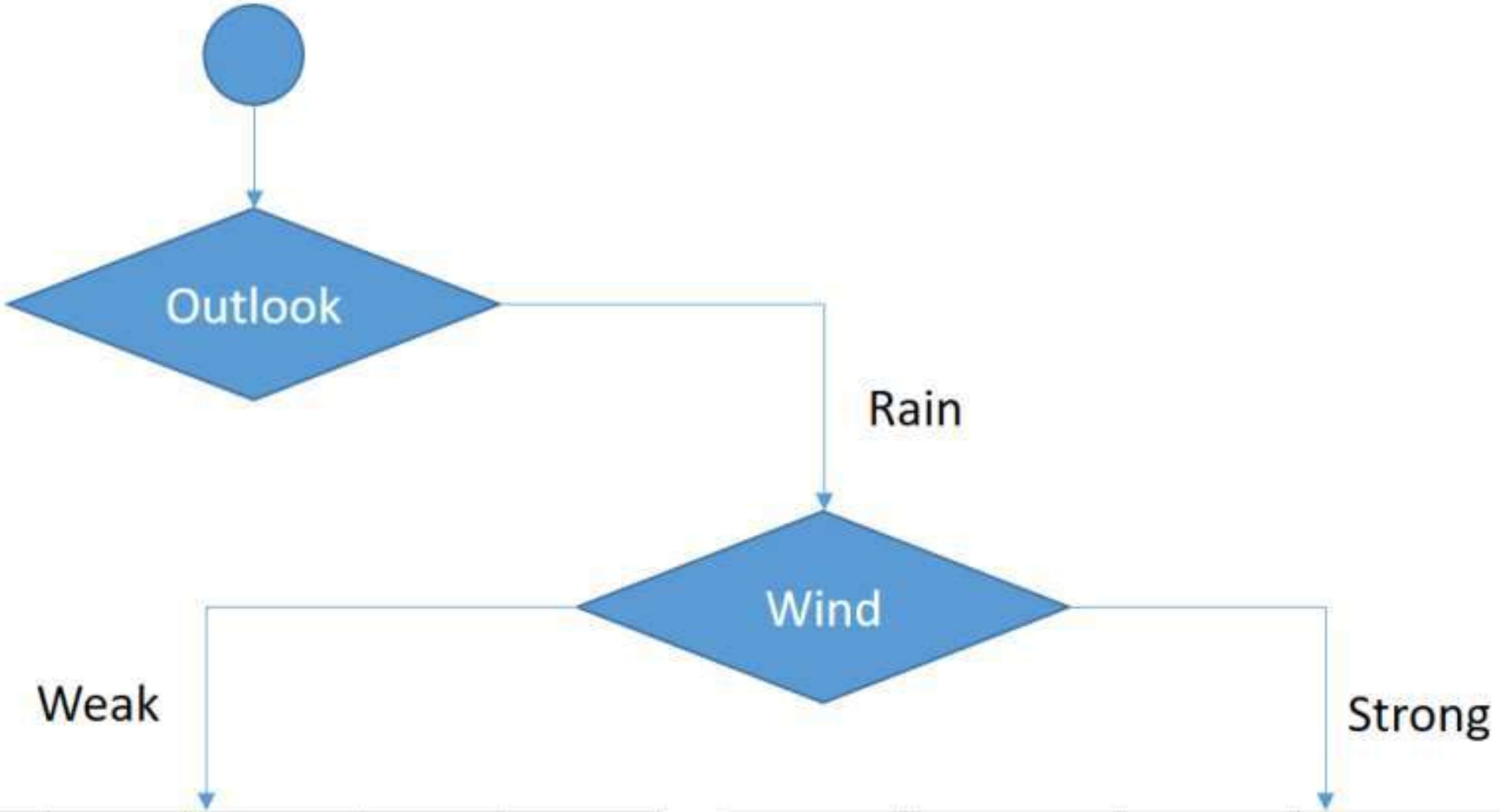
Day	outlook	temperature	humidity	wind	Decision
4	rain	mild	high	weak	Yes
5	rain	cool	normal	weak	Yes
6	rain	cool	normal	strong	No
10	rain	mild	normal	weak	Yes
14	rain	mild	high	strong	No

Decision on rainfall outlook factor

Features	Gini Index
temperature	0.466
humidity	0.466
wind	0

we have calculated the Gini index of all the features when the outlook is rainfall. You can infer that wind has lowest value. so next node will be wind.

Outcome

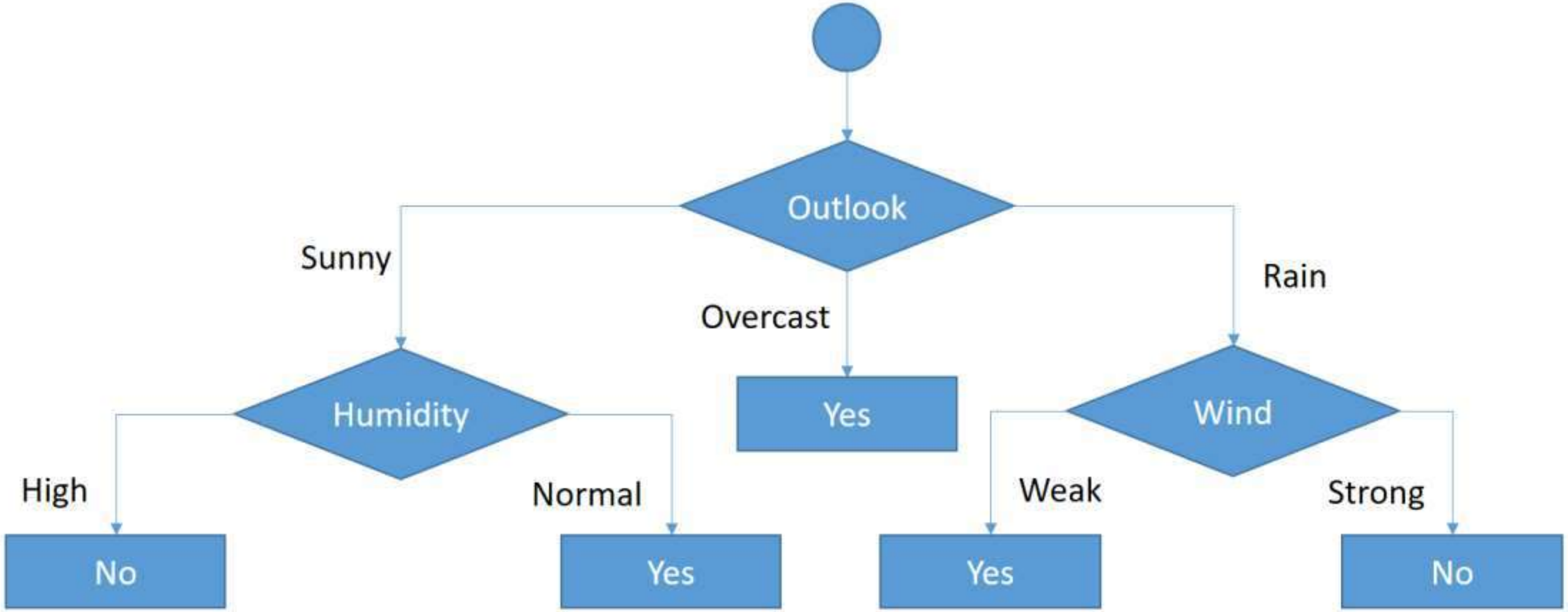


Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
6	Rain	Cool	Normal	Strong	No
14	Rain	Mild	High	Strong	No

Use Case: Weather Dataset – Predict play or not

As seen, decision is always yes when wind is weak. On the other hand, decision is always no if wind is strong. This means that this branch is over.



References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, ―Learning from Data, AML Book Publishers, 2012.

P. Flach, ―Machine Learning: The art and science of algorithms that make sense of data‖, Cambridge University Press, 2012.

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-example/#google_vignette

https://medium.com/@singhakshay.etw69/decision-tree-algorithm-cart-e61032794927