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

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

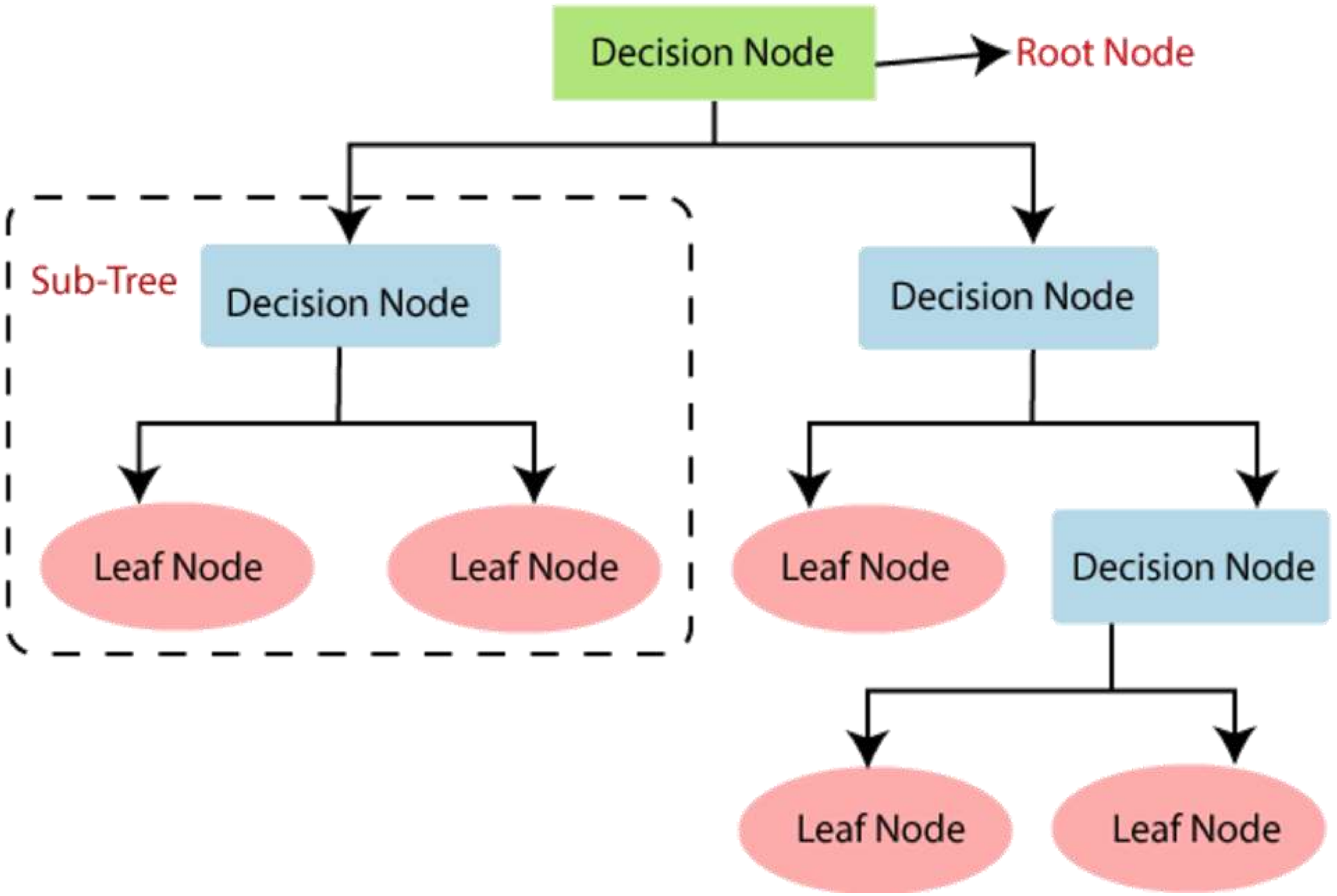
II YEAR IV SEM

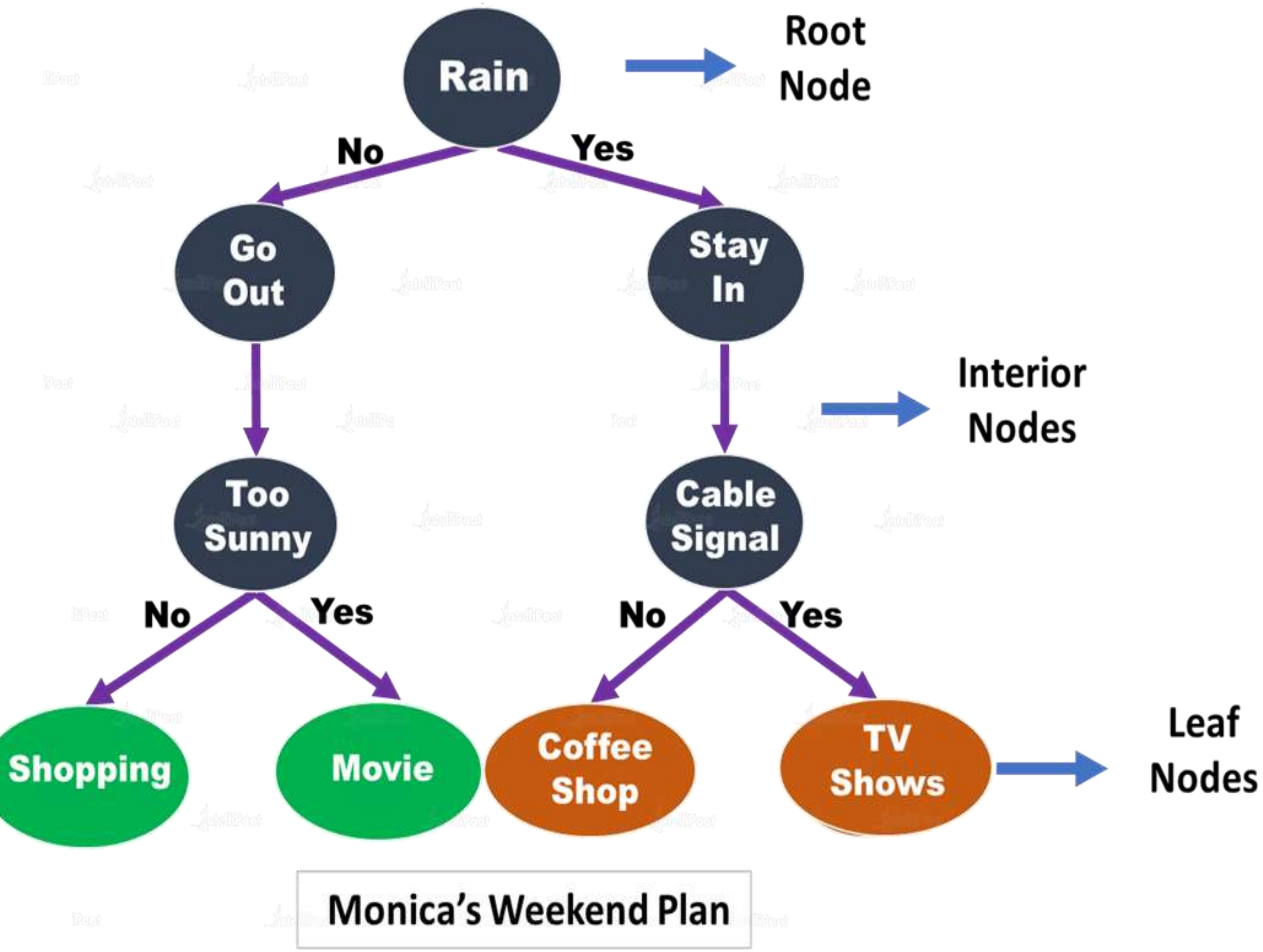
UNIT III – GENERATIVE MODELS AND BOOSTING

TOPIC 17 – Decision Tree - Training and Visualizing a Decision Tree-Entropy

Decision tree

1. Decision Trees are supervised learning algorithms used for classification and regression problems. They work by creating a model that predicts the value of a target variable based on several input variables.
2. The model is a tree-like structure, with each internal node representing a "test" on an attribute, each branch representing the outcome of the test, and each leaf node representing a class label.





Step 1: Begin the tree with the root node S, which includes the whole dataset.

Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset.

Step 3: Subdivide the S into subsets containing potential values for the best qualities.

Step 4: Create the decision tree node with the best attribute.

Step 5: Create new decision trees recursively using the subsets of the dataset obtained in step3. Continue this procedure until you reach a point where you can no longer categorize the nodes and refer to the last node as a leaf node.

The biggest challenge that emerges while developing a Decision tree is how to choose the optimal attribute for the root node and sub-nodes. To tackle such challenges, a technique known as Attribute selection measure, or ASM, is used.

Two Techniques:

1. Information Gain
2. Gini Index

1. **Information Gain:** Information Gain IG(A) tells us how much uncertainty in S was reduced after splitting set S on attribute A

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where,

- H(S) - Entropy of set S.
- T - The subsets created from splitting set S by attribute A such that

$S = \bigcup_{t \in T} t$

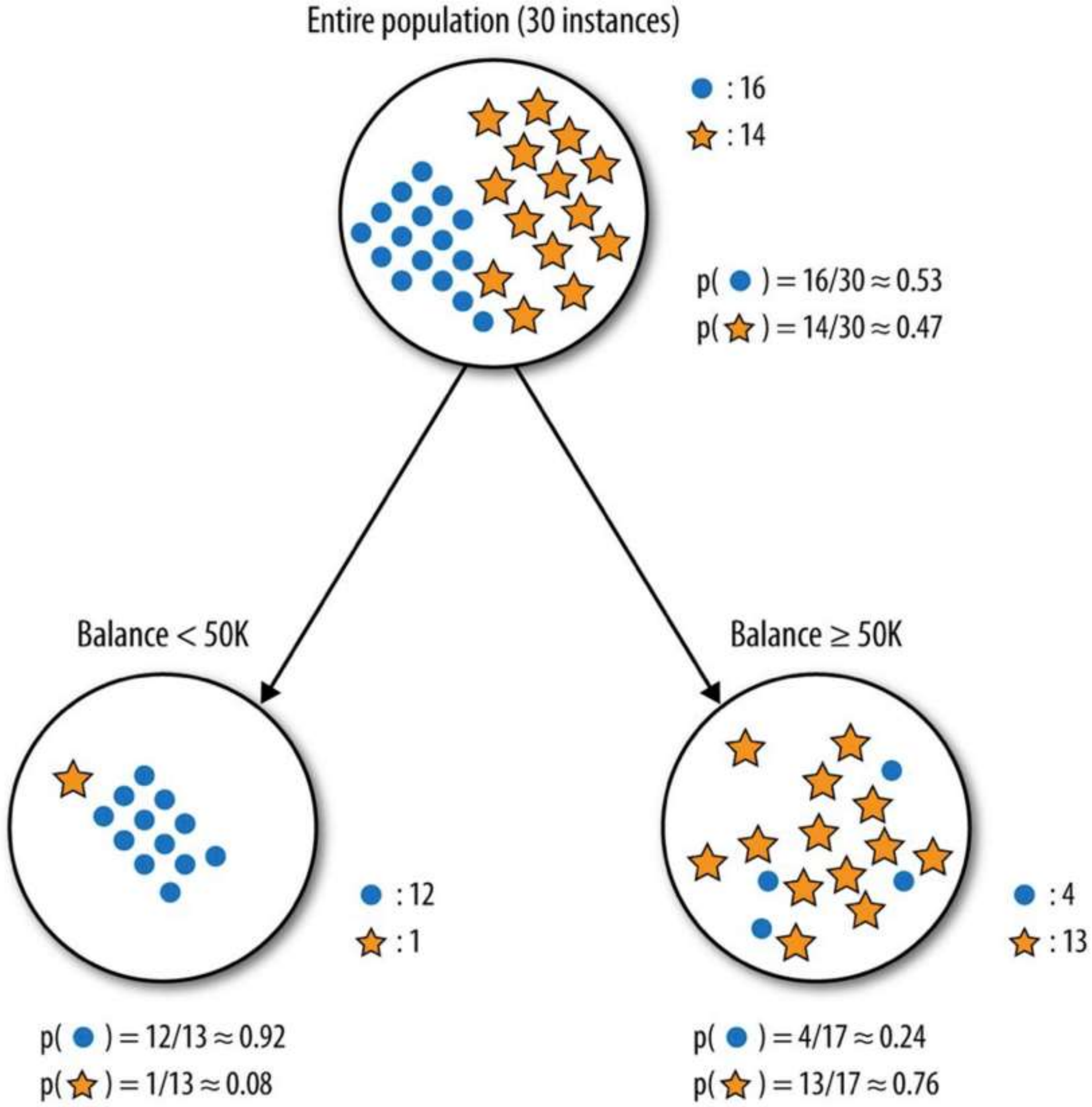
- p(t) - The proportion of the number of elements in t to the number of elements in set S.
- H(t) - Entropy of subset t.

2. **Entropy:** Entropy is a measure of the amount of uncertainty in the dataset S. Entropy can be calculated as:

$$H(S) = \sum_{c \in C} -p(c)\log_2 p(c)$$

The attribute with the smallest entropy is used to split the set S on that particular iteration. Entropy = 0 implies it is of pure class, that means all are of same category.

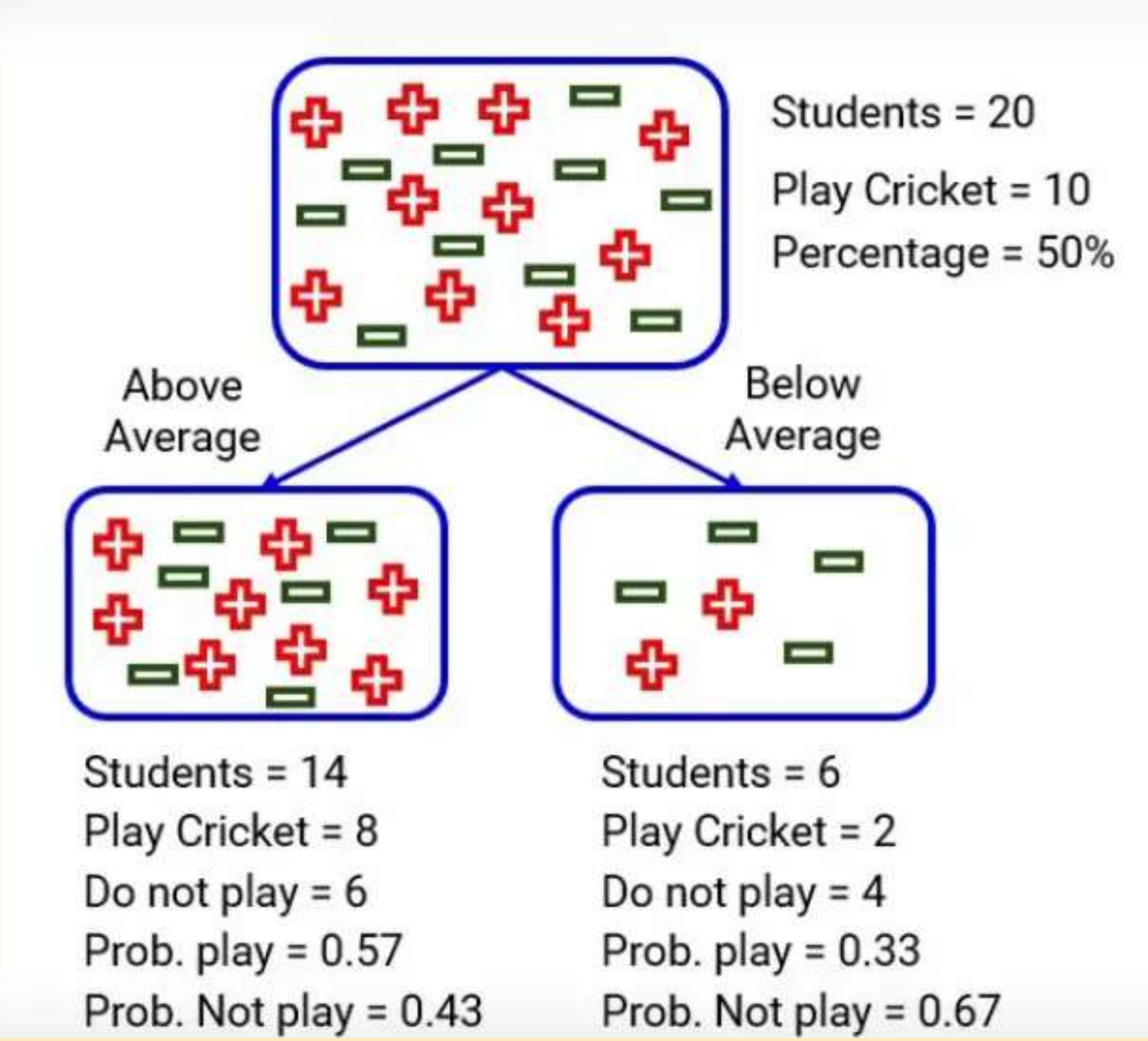
Samples


Entire population (30 instances)
: 16
: 14
p( ) = 16/30 ≈ 0.53
p( ) = 14/30 ≈ 0.47
Balance < 50K
Balance ≥ 50K
: 12
: 1
p( ) = 12/13 ≈ 0.92
p( ) = 1/13 ≈ 0.08
: 4
: 13
p( ) = 4/17 ≈ 0.24
p( ) = 13/17 ≈ 0.76

CART: Gini Index

1. It only generates binary splits, whereas the CART method generates binary splits using the Gini index.

Gini Index= 1- ∑jPj2



Terminologies Related to the Regression Analysis:

1. Root Node: This node gets divided into different homogeneous nodes. It represents the entire sample.
2. Splitting: It is the process of splitting or dividing a node into two or more sub-nodes.
3. Interior Nodes: They represent different tests on an attribute.
4. Branches: They hold the outcomes of those tests.
5. Leaf Nodes: When the nodes can’t be split further, they are called leaf nodes.
6. Parent and Child Nodes: The node from which sub-nodes are created is called a parent node. And, the sub-nodes are called the child nodes.

Using ID3 Algorithm to build a Decision Tree to predict the weather

ID3 algorithm, stands for **Iterative Dichotomiser 3**, is a **classification algorithm** that follows a **greedy approach** of **building a decision tree** by selecting a best attribute that yields **maximum Information Gain (IG) or minimum Entropy (H)**.

What are the steps in ID3 algorithm?

1. Calculate entropy for dataset.
2. For each attribute/feature
    1. Calculate entropy for all its categorical values.
    2. Calculate information gain for the feature.
3. Find the feature with maximum information gain.
4. Repeat it until we get the desired tree.

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

ID3: Finding best feature for root

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

Here, dataset is of binary classes(yes and no), where 9 out of 14 are "yes" and 5 out of 14 are "no".

Complete entropy of dataset is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (9/14) * log2(9/14) - (5/14) * log2(5/14)

= - (-0.41) - (-0.53)

= 0.94

First Attribute - Outlook

Categorical values - sunny, overcast and rain

H(Outlook=sunny) = -(2/5)*log(2/5)-(3/5)*log(3/5) =0.971

H(Outlook=rain) = -(3/5)*log(3/5)-(2/5)*log(2/5) =0.971

H(Outlook=overcast) = -(4/4)*log(4/4)-0 = 0

Average Entropy Information for Outlook -

I(Outlook) = p(sunny) * H(Outlook=sunny) + p(rain) * H(Outlook=rain) + p(overcast) * H(Outlook=overcast)

= (5/14)*0.971 + (5/14)*0.971 + (4/14)*0

= 0.693

Information Gain = H(S) - I(Outlook)

= 0.94 - 0.693 = 0.247

ID3: Finding best feature for root

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

Here, dataset is of binary classes(yes and no), where 9 out of 14 are "yes" and 5 out of 14 are "no".

Complete entropy of dataset is -

H(S) = 0.94

Second Attribute - Temperature

Categorical values - hot, mild, cool

H(Temperature=hot) = -(2/4)*log(2/4)-(2/4)*log(2/4) = 1

H(Temperature=cool) = -(3/4)*log(3/4)-(1/4)*log(1/4) = 0.811

H(Temperature=mild) = -(4/6)*log(4/6)-(2/6)*log(2/6) = 0.9179

Average Entropy Information for Temperature -

I(Temperature) = p(hot)*H(Temperature=hot) + p(mild)*H(Temperature=mild) + p(cool)*H(Temperature=cool)

= (4/14)*1 + (6/14)*0.9179 + (4/14)*0.811

= 0.9108

Information Gain = H(S) - I(Temperature)

= 0.94 - 0.9108

= 0.0292

ID3: Finding best feature for root

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

Here, dataset is of binary classes(yes and no), where 9 out of 14 are "yes" and 5 out of 14 are "no".

Complete entropy of dataset is -

H(S) = 0.94

Third Attribute - Humidity

Categorical values - high, normal

H(Humidity=high) = -(3/7)*log(3/7)-(4/7)*log(4/7) = 0.983

H(Humidity=normal) = -(6/7)*log(6/7)-(1/7)*log(1/7) = 0.591

Average Entropy Information for Humidity -

I(Humidity) = p(high)*H(Humidity=high) + p(normal)*H(Humidity=normal)

= (7/14)*0.983 + (7/14)*0.591

= 0.787

Information Gain = H(S) - I(Humidity)

= 0.94 - 0.787

= 0.153

ID3: Finding best feature for root

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

Here, dataset is of binary classes(yes and no), where 9 out of 14 are "yes" and 5 out of 14 are "no".

Complete entropy of dataset is -

H(S) = 0.94

Fourth Attribute - Wind

Categorical values - weak, strong

H(Wind=weak) = -(6/8)*log(6/8)-(2/8)*log(2/8) = 0.811

H(Wind=strong) = -(3/6)*log(3/6)-(3/6)*log(3/6) = 1

Average Entropy Information for Wind -

I(Wind) = p(weak)*H(Wind=weak) + p(strong)*H(Wind=strong)

= (8/14)*0.811 + (6/14)*1

= 0.892

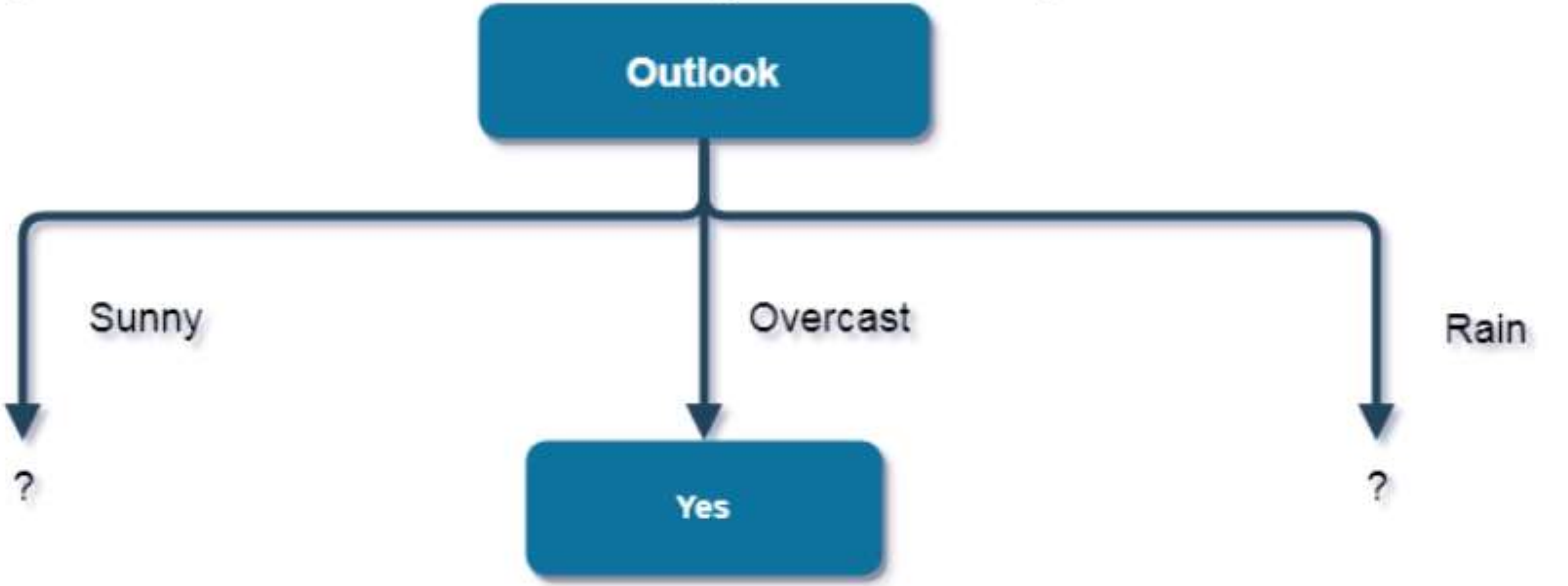
Information Gain = H(S) - I(Wind)

= 0.94 - 0.892

= 0.048

ID3: Finding best feature for root

Here, the attribute with maximum information gain is Outlook. So, the decision tree built so far -



Here, when Outlook == overcast, it is of pure class(Yes). Now, we have to repeat same procedure for the data with rows consist of Outlook value as Sunny and then for Outlook value as Rain.

Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [1, 2, 8, 9, 11]}.

Complete entropy of Sunny is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (2/5) * log2(2/5) - (3/5) * log2(3/5)

= 0.971

First Attribute - Temperature

Categorical values - hot, mild, cool

H(Sunny, Temperature=hot) = -0-(2/2)*log(2/2) = 0

H(Sunny, Temperature=cool) = -(1)*log(1)- 0 = 0

H(Sunny, Temperature=mild) = -(1/2)*log(1/2)-(1/2)*log(1/2) = 1

Average Entropy Information for Temperature -

I(Sunny, Temperature) = p(Sunny, hot)*H(Sunny, Temperature=hot) + p(Sunny, mild)*H(Sunny, Temperature=mild) + p(Sunny, cool)*H(Sunny, Temperature=cool)

= (2/5)*0 + (1/5)*0 + (2/5)*1

= 0.4

Information Gain = H(Sunny) - I(Sunny, Temperature)

= 0.971 - 0.4

= 0.571

Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [1, 2, 8, 9, 11]}.

Complete entropy of Sunny is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (2/5) * log2(2/5) - (3/5) * log2(3/5)

= 0.971

Second Attribute - Humidity

Categorical values - high, normal

H(Sunny, Humidity=high) = - 0 - (3/3)*log(3/3) = 0

H(Sunny, Humidity=normal) = -(2/2)*log(2/2)-0 = 0

Average Entropy Information for Humidity -

I(Sunny, Humidity) = p(Sunny, high)*H(Sunny, Humidity=high) + p(Sunny, normal)*H(Sunny, Humidity=normal)

= (3/5)*0 + (2/5)*0

= 0

Information Gain = H(Sunny) - I(Sunny, Humidity)

= 0.971 - 0

= 0.971

Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [1, 2, 8, 9, 11]}.

Complete entropy of Sunny is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (2/5) * log2(2/5) - (3/5) * log2(3/5)

= 0.971

Third Attribute - Wind

Categorical values - weak, strong

H(Sunny, Wind=weak) = -(1/3)*log(1/3)-(2/3)*log(2/3) = 0.918

H(Sunny, Wind=strong) = -(1/2)*log(1/2)-(1/2)*log(1/2) = 1

Average Entropy Information for Wind -

I(Sunny, Wind) = p(Sunny, weak)*H(Sunny, Wind=weak) + p(Sunny, strong)*H(Sunny, Wind=strong)

= (3/5)*0.918 + (2/5)*1

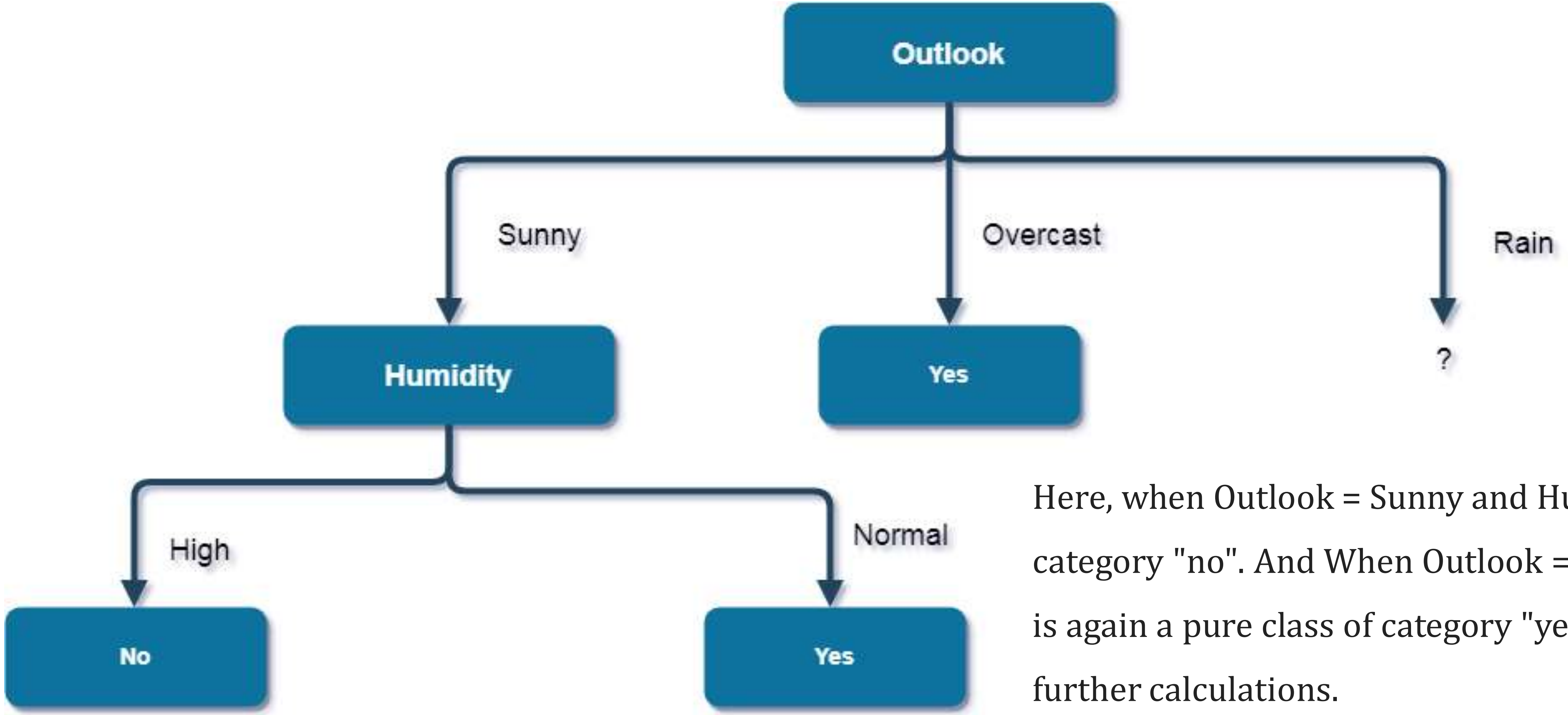
= 0.9508

Information Gain = H(Sunny) - I(Sunny, Wind)

= 0.971 - 0.9508

= 0.0202

Here, the attribute with maximum information gain is Humidity. So, the decision tree built so far



Here, when Outlook = Sunny and Humidity = High, it is a pure class of category "no". And When Outlook = Sunny and Humidity = Normal, it is again a pure class of category "yes". Therefore, we don't need to do further calculations.

Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [4, 5, 6, 10, 14]}.

Complete entropy of Rain is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (3/5) * log(3/5) - (2/5) * log(2/5)

= 0.971

First Attribute - Temperature

Categorical values - mild, cool

H(Rain, Temperature=cool) = -(1/2)*log(1/2)- (1/2)*log(1/2) = 1

H(Rain, Temperature=mild) = -(2/3)*log(2/3)-(1/3)*log(1/3) = 0.918

Average Entropy Information for Temperature -

I(Rain, Temperature) = p(Rain, mild)*H(Rain, Temperature=mild) + p(Rain, cool)*H(Rain, Temperature=cool)

= (2/5)*1 + (3/5)*0.918

= 0.9508

Information Gain = H(Rain) - I(Rain, Temperature)

= 0.971 - 0.9508

= 0.0202

Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [4, 5, 6, 10, 14]}.

Complete entropy of Rain is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (3/5) * log(3/5) - (2/5) * log(2/5)

= 0.971

Second Attribute - Wind

Categorical values - weak, strong

H(Wind=weak) = -(3/3)*log(3/3)-0 = 0

H(Wind=strong) = 0-(2/2)*log(2/2) = 0

Average Entropy Information for Wind -

I(Wind) = p(Rain, weak)*H(Rain, Wind=weak) + p(Rain, strong)*H(Rain, Wind=strong)

= (3/5)*0 + (2/5)*0

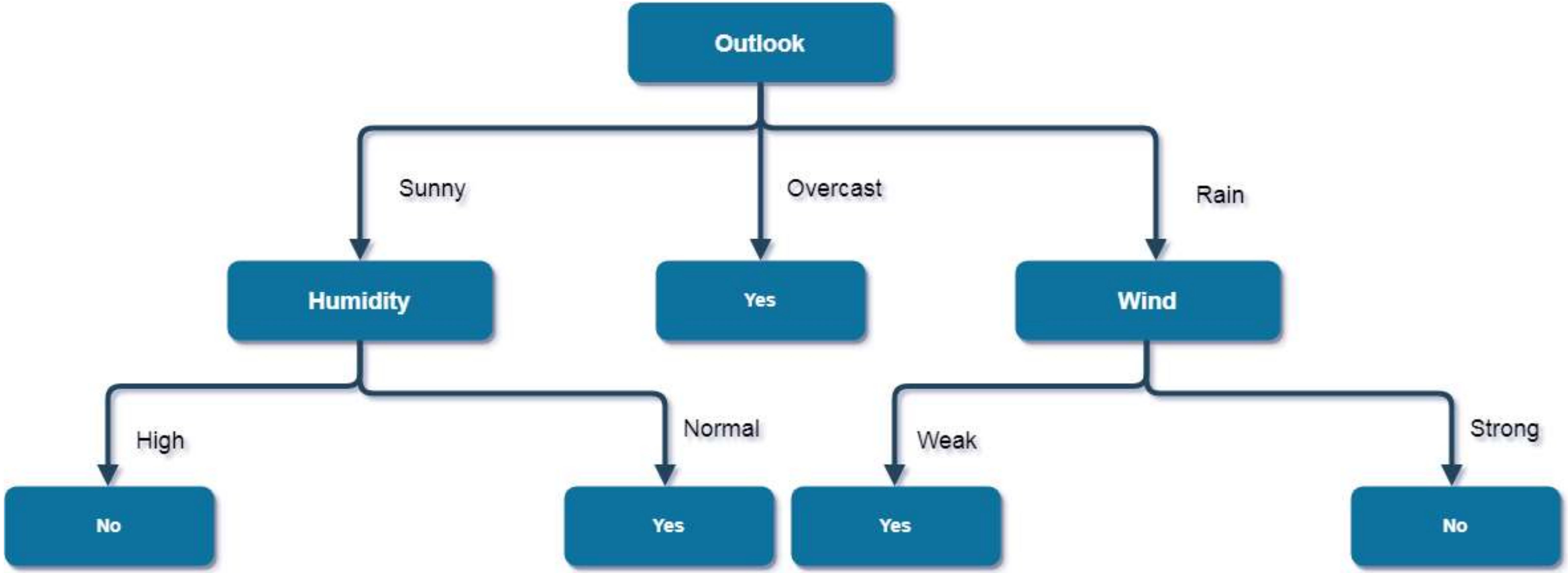
= 0

Information Gain = H(Rain) - I(Rain, Wind)

= 0.971 - 0

= 0.971

Here, the attribute with maximum information gain is Wind. So, the decision tree built so far



Here, when Outlook = Rain and Wind = Strong, it is a pure class of category "no". And When Outlook = Rain and Wind = Weak, it is again a pure class of category "yes". And this is our final desired tree for the given dataset.

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.

W3school.com

https://discourse.opengenus.org/t/using-id3-algorithm-to-build-a-decision-tree-to-predict-the-weather/3343