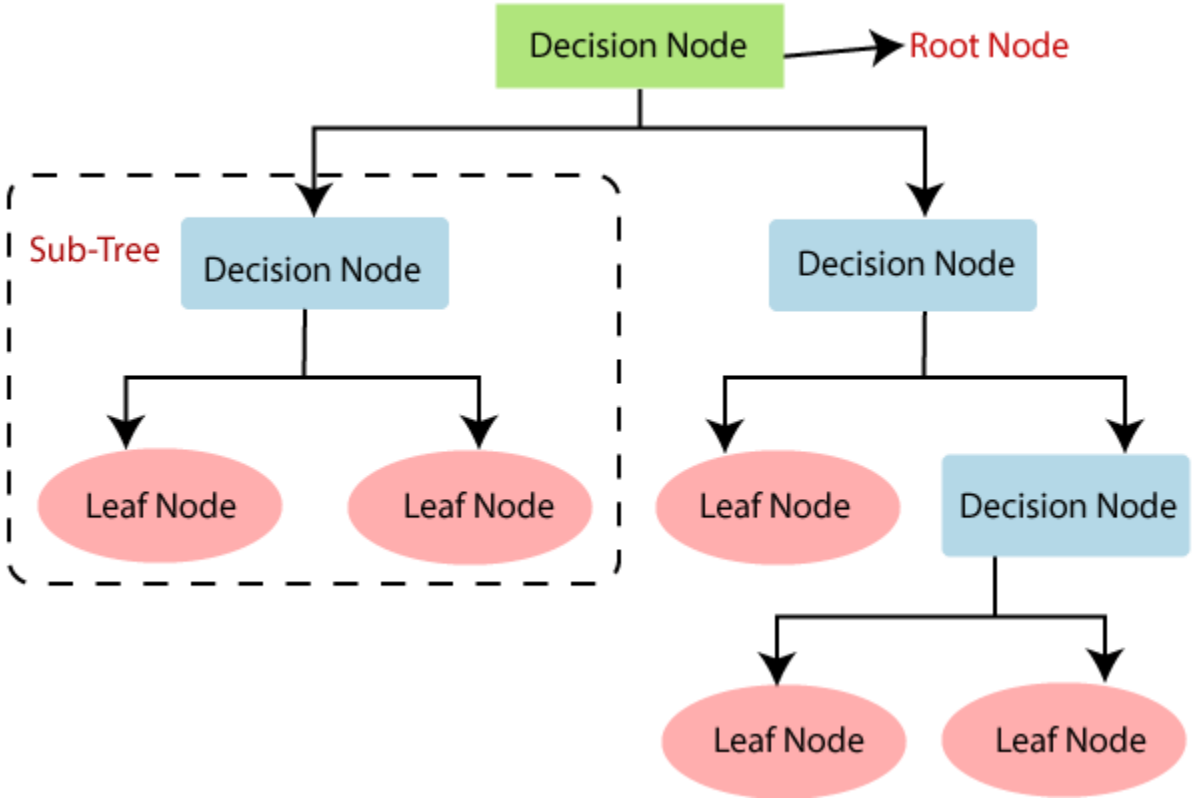
Decision Tree Algorithm

Decision tree Algorithm

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- It is a tree-structured classifier, where,
  - Internal nodes represent the features of a dataset,
  - Branches represent the decision rules and
  - Each leaf node represents the outcome.

- **It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.**
- It is called a decision tree because, a question and a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- To build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
- A decision tree simply asks a question and based on the answer **(Yes/No),** it further splits the tree into subtrees.

Decision tree



Why use Decision Trees?

•Below are the two reasons for using the Decision tree:

1.Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.

2.The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

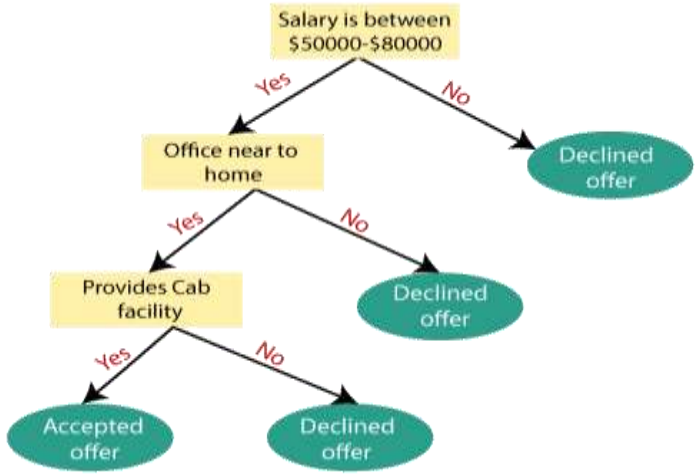
- **Root Node:** The root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning**: Pruning is the process of removing unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

How does the Decision Tree algorithm Work?

- In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree.
- This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
- For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further.
- It continues the process until it reaches the leaf node of the tree.
- The complete process can be better understood using the given algorithm:

Example - 1

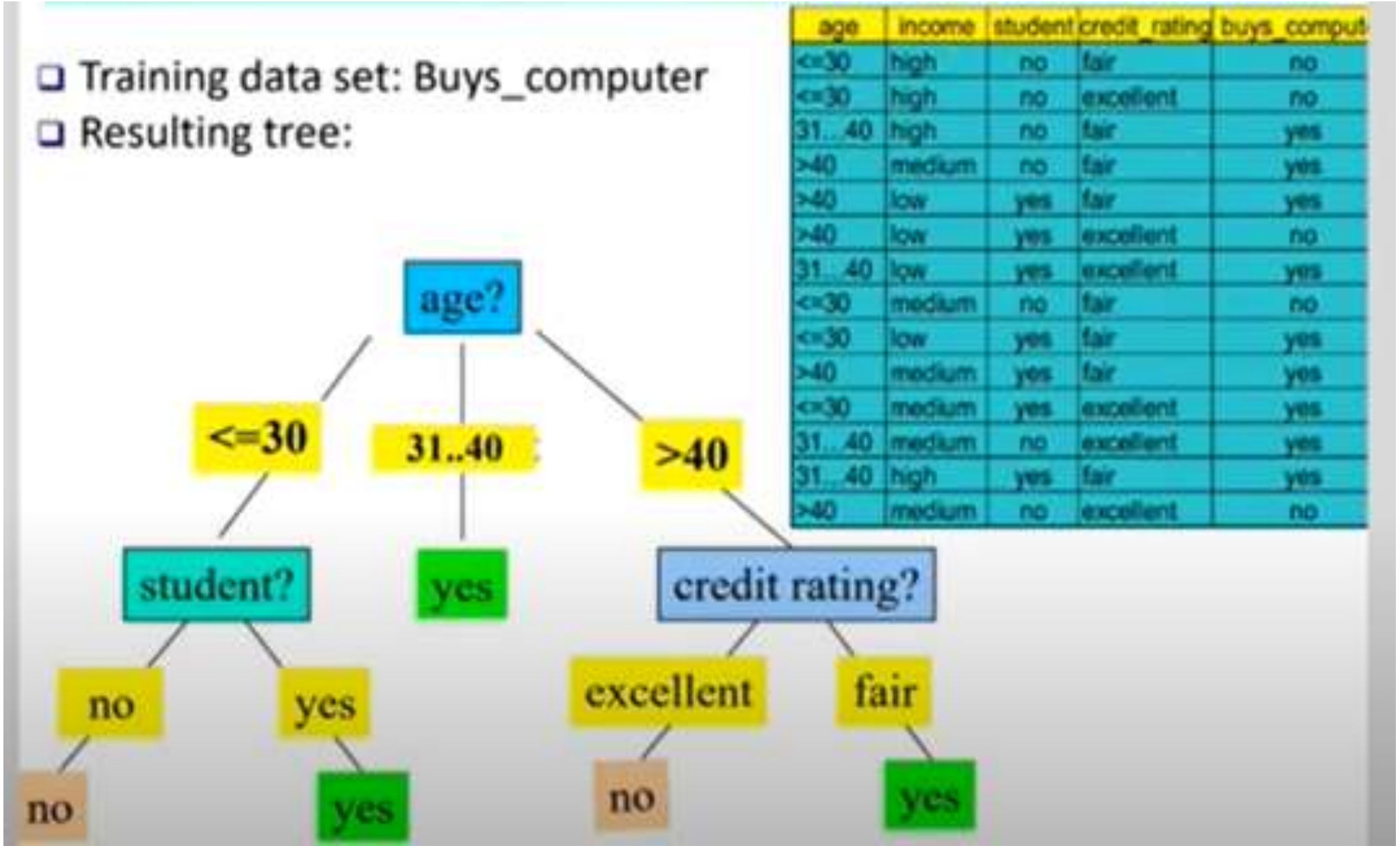
- Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not.
- So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM).
- The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels.
- The next decision node further gets split into one decision node (Cab facility) and one leaf node.
- Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offers). Consider the below diagram:



Example -2

- Training data set: Buys_computer
- Resulting tree:

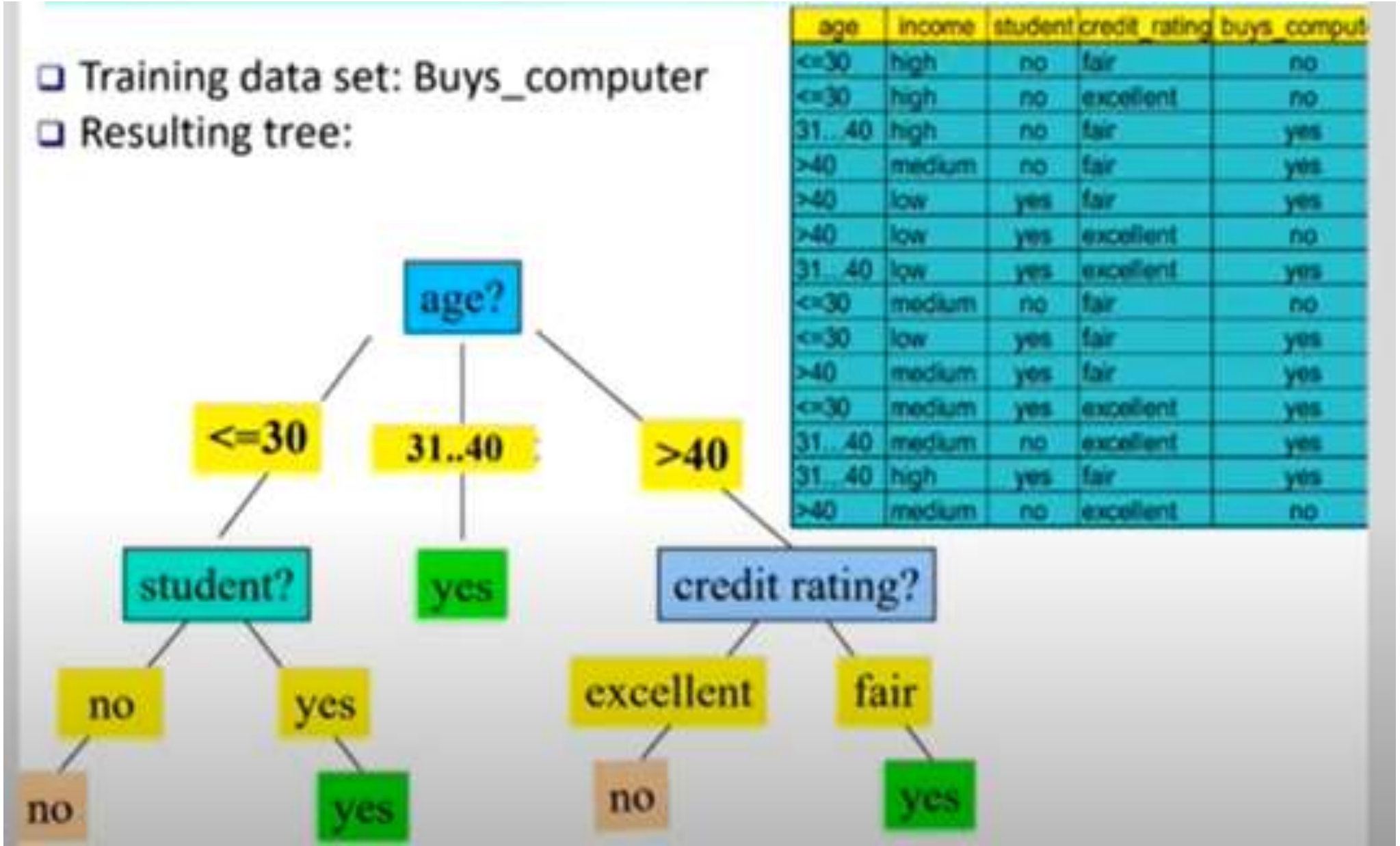
age	income	student	credit_rating	buys_comput
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no



Example -2

- Training data set: Buys_computer
- Resulting tree:

age	income	student	credit_rating	buys_comput
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no



Attribute Selection Measures

- While implementing a Decision tree, the main issue arises in how to select the best attribute for the root node and sub-nodes.
- So, to solve such problems there is a technique which is called as **Attribute selection measure** or **ASM.** By this measurement, we can easily select the best attribute for the nodes of the tree.
- There are two popular techniques for ASM, which are:

  - **Information Gain**
  - **Gini Index**

1. Information Gain

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first.
- It can be calculated using the below formula

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

- **Entropy:** Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data.
- Entropy can be calculated as:

$$Entropy(s) = -P(yes)\log_2 P(yes) - P(no)\log_2 P(no)$$

Where,

- S= Total number of samples
- P(yes)= probability of yes
- P(no)= probability of no

Example

Day	outlook	temperature	humidity	wind	playtennis
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rain	mild	high	weak	yes
D5	rain	cool	normal	weak	yes
D6	rain	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rain	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rain	mild	high	strong	no

Solution

$$\begin{aligned}\text{Entropy}\,(S) &= -p_{\text{yes}}\log_2(p_{\text{yes}}) - p_{\text{no}}\log_2(p_{\text{no}}) \\ &= -(9/14)\times\log_2(9/14) - (5/14)\times\log_2(5/14) \\ &= 0.9405\end{aligned}$$

- Entropy is measured in bits.
- If there are only two possible classes, entropy values can range from 0 to 1.
- For n classes, entropy ranges from 0 to log2(n).

Some Examples:

$$Entropy([14+, 0-]) = -14/14\ log_2(14/14) - 0\ log_2(0) = 0$$

$$Entropy([9+, 5-]) = -9/14\ log_2(9/14) - 5/14\ log_2(5/14) = 0.94$$

$$Entropy([7+, 7-]) = -7/14\ log_2(7/14) - 7/14\ log_2(7/14) = 1/2 + 1/2 = 1$$

Entropy & Information Gain

INFORMATION GAIN MEASURES THE EXPECTED REDUCTION IN ENTROPY

- Given entropy as a measure of the impurity in a collection of training examples, the ***information gain,*** is simply the expected reduction in entropy caused by partitioning the examples according to an attribute.
- More precisely, the information gain, ***Gain(S, A)*** of ***an*** attribute **A,** relative to a collection of examples ***S,*** is defined as,

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Example - 1

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Values(Wind) = Weak, Strong$$

$$S = [9+, 5-]$$

$$S_{Weak} \leftarrow [6+, 2-]$$

$$S_{Strong} \leftarrow [3+, 3-]$$

$$Gain(S, Wind) = Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - (8/14)Entropy(S_{Weak}) - (6/14)Entropy(S_{Strong})$$

$$= 0.940 - (8/14)0.811 - (6/14)1.00$$

$$= 0.048$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attribute: Outlook

$$Values(Outlook) = Sunny, Overcast, Rain$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$S_{Sunny} \leftarrow [2+, 3-] \qquad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$S_{Overcast} \leftarrow [4+, 0-] \qquad Entropy(S_{Overcast}) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$$

$$S_{Rain} \leftarrow [3+, 2-] \qquad Entropy(S_{Rain}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$$

$$Gain(S, Outlook) = Entropy(S) - \sum_{v \in \{Sunny, Overcast, Rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Outlook)$$

$$= Entropy(S) - \frac{5}{14}Entropy(S_{Sunny}) - \frac{4}{14}Entropy(S_{Overcast})$$

$$- \frac{5}{14}Entropy(S_{Rain})$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14}0.971 - \frac{4}{14}0 - \frac{5}{14}0.971 = 0.2464$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attribute: Temp

$Values\ (Temp) = Hot, Mild, Cool$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$

$S_{Hot} \leftarrow [2+, 2-]$ $\quad Entropy(S_{Hot}) = -\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4} = 1.0$

$S_{Mild} \leftarrow [4+, 2-]$ $\quad Entropy(S_{Mild}) = -\frac{4}{6} log_2 \frac{4}{6} - \frac{2}{6} log_2 \frac{2}{6} = 0.9183$

$S_{Cool} \leftarrow [3+, 1-]$ $\quad Entropy(S_{Cool}) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} = 0.8113$

$$Gain\ (S, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Temp) = Entropy(S) - \frac{4}{14} Entropy(S_{Hot}) - \frac{6}{14} Entropy(S_{Mild}) - \frac{4}{14} Entropy(S_{Cool})$$

$$Gain(S, Temp) = 0.94 - \frac{4}{14} 1.0 - \frac{6}{14} 0.9183 - \frac{4}{14} 0.8113 = 0.028$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attribute: Humidity

$$Values\,(Humidity) = High, Normal$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$S_{High} \leftarrow [3+, 4-] \qquad Entropy(S_{High}) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.9852$$

$$S_{Normal} \leftarrow [6+, 1-] \qquad Entropy(S_{Normal}) = -\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7} = 0.5916$$

$$Gain\,(S, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Humidity)$$

$$= Entropy(S) - \frac{7}{14}Entropy(S_{High}) - \frac{7}{14}Entropy(S_{Normal})$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14}0.9852 - \frac{7}{14}0.5916 = 0.1516$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attribute: Wind

$Values\ (Wind) = Strong, Weak$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$

$S_{Strong} \leftarrow [3+, 3-]$ $\quad Entropy(S_{Strong}) = 1.0$

$S_{Weak} \leftarrow [6+, 2-]$ $\quad Entropy(S_{Weak}) = -\frac{6}{8} log_2 \frac{6}{8} - \frac{2}{8} log_2 \frac{2}{8} = 0.8113$

$$Gain\ (S, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Wind) = Entropy(S) - \frac{6}{14} Entropy(S_{Strong}) - \frac{8}{14} Entropy(S_{Weak})$$

$$Gain(S, Wind) = 0.94 - \frac{6}{14}\ 1.0 - \frac{8}{14}\ 0.8113 = 0.0478$$

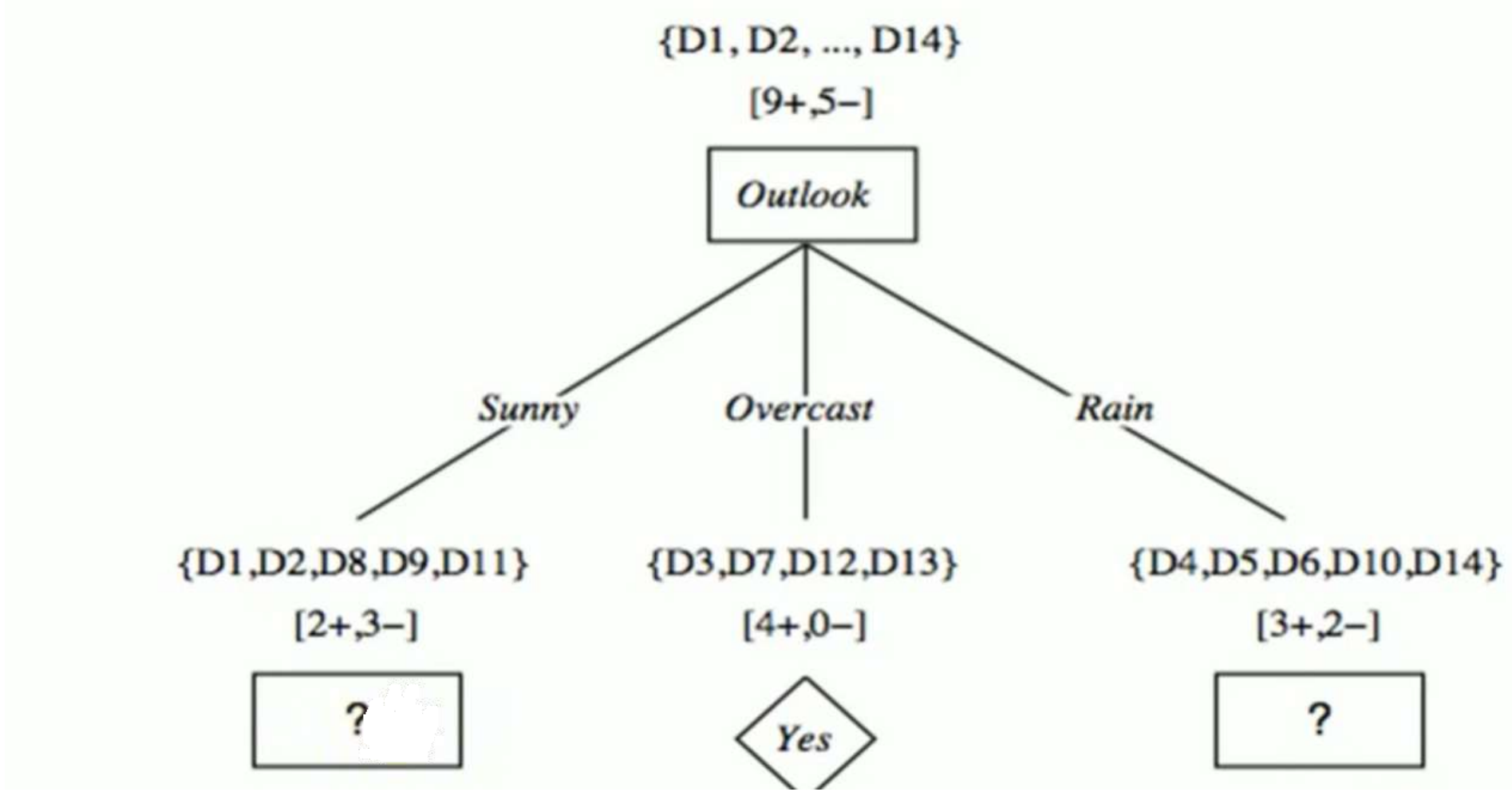
Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Gain(S, Outlook) = 0.2464$$

$$Gain(S, Temp) = 0.0289$$

$$Gain(S, Humidity) = 0.1516$$

$$Gain(S, Wind) = 0.0478$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
{D1,D2,D8,D9,D11}
[2+,3−]
{D3,D7,D12,D13}
[4+,0−]
{D4,D5,D6,D10,D14}
[3+,2−]
?
Yes
?

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

Attribute: Temp

$Values\ (Temp) = Hot, Mild, Cool$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{Hot} \leftarrow [0+, 2-]$ $\quad Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Mild}) = 1.0$

$S_{Cool} \leftarrow [1+, 0-]$ $\quad Entropy(S_{Cool}) = 0.0$

$$Gain\ (S_{Sunny}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Temp)$$

$$= Entropy(S) - \frac{2}{5}Entropy(S_{Hot}) - \frac{2}{5}Entropy(S_{Mild})$$

$$- \frac{1}{5}Entropy(S_{Cool})$$

$$Gain(S_{sunny}, Temp) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1 - \frac{1}{5}0.0 = 0.570$$

Day	Temp	Humidity	Wind	Play Tennis
Dl	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
Dl1	Mild	Normal	Strong	Yes

Attribute: Humidity

$Values\,(Humidity) = High, Normal$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{high} \leftarrow [0+, 3-]$ $\quad Entropy(S_{High}) = 0.0$

$S_{Normal} \leftarrow [2+, 0-]$ $\quad Entropy(S_{Normal}) = 0.0$

$$Gain\,(S_{Sunny}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Humidity) = Entropy(S) - \frac{3}{5}Entropy(S_{High}) - \frac{2}{5}Entropy(S_{Normal})$$

$$Gain(S_{sunny}, Humidity) = 0.97 - \frac{3}{5}\,0.0 - \frac{2}{5}0.0 = 0.97$$

Day	Temp	Humidity	Wind	Play Tennis
Dl	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
Dl1	Mild	Normal	Strong	Yes

Attribute: Wind

$Values\,(Wind) = Strong, Weak$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{Strong} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Strong}) = 1.0$

$S_{Weak} \leftarrow [1+, 2-]$ $\quad Entropy(S_{Weak}) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9183$

$$Gain\,(S_{Sunny}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

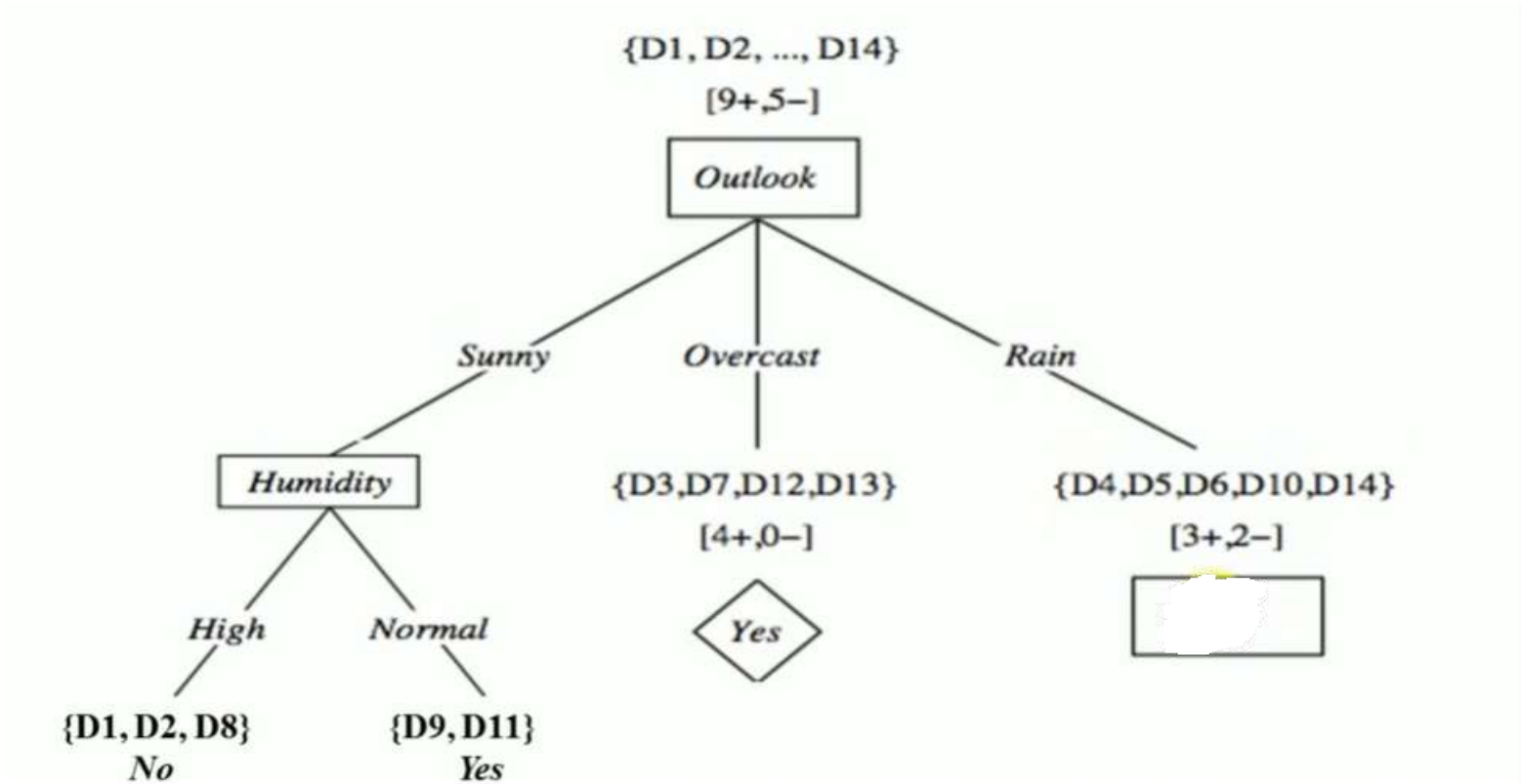
$$Gain(S_{sunny}, Wind) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}\,0.918 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

$$Gain(S_{sunny}, Temp) = 0.570$$

$$Gain(S_{sunny}, Humidity) = 0.97$$

$$Gain(S_{sunny}, Wind) = 0.0192$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
Humidity
{D3,D7,D12,D13}
[4+,0−]
{D4,D5,D6,D10,D14}
[3+,2−]
High
Normal
Yes
{D1, D2, D8}
No
{D9, D11}
Yes

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

Attribute: Temp

$Values\ (Temp) = Hot, Mild, Cool$

$S_{Rain} = [3+, 2-]$ $\quad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$

$S_{Hot} \leftarrow [0+, 0-]$ $\quad Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [2+, 1-]$ $\quad Entropy(S_{Mild}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$

$S_{Cool} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Cool}) = 1.0$

$$Gain\ (S_{Rain}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Temp)$$

$$= Entropy(S) - \frac{0}{5}Entropy(S_{Hot}) - \frac{3}{5}Entropy(S_{Mild})$$

$$- \frac{2}{5}Entropy(S_{Cool})$$

$$Gain(S_{Rain}, Temp) = 0.97 - \frac{0}{5}0.0 - \frac{3}{5}\,0.918 - \frac{2}{5}\,1.0 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

Attribute: Humidity

$Values\ (Humidity) = High, Normal$

$S_{Rain} = [3+, 2-]$ $\quad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$

$S_{High} \leftarrow [1+, 1-]$ $\quad Entropy(S_{High}) = 1.0$

$S_{Normal} \leftarrow [2+, 1-]$ $\quad Entropy(S_{Normal}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$

$$Gain\ (S_{Rain}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Humidity) = Entropy(S) - \frac{2}{5}Entropy(S_{High}) - \frac{3}{5}Entropy(S_{Normal})$$

$$Gain(S_{Rain}, Humidity) = 0.97 - \frac{2}{5}\ 1.0 - \frac{3}{5}0.918 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

Attribute: Wind

$Values\ (wind) = Strong, Weak$

$S_{Rain} = [3+, 2-]$

$S_{Strong} \leftarrow [0+, 2-]$

$S_{Weak} \leftarrow [3+, 0-]$

$Entropy(S_{Sunny}) = -\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5} = 0.97$

$Entropy(S_{Strong}) = 0.0$

$Entropy(S_{weak}) = 0.0$

$$Gain\ (S_{Rain}, Wind) = Entropy(S) - \sum_{v \in (Strong, Weak)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Wind) = Entropy(S) - \frac{2}{5} Entropy(S_{Strong}) - \frac{3}{5} Entropy(S_{Weak})$$

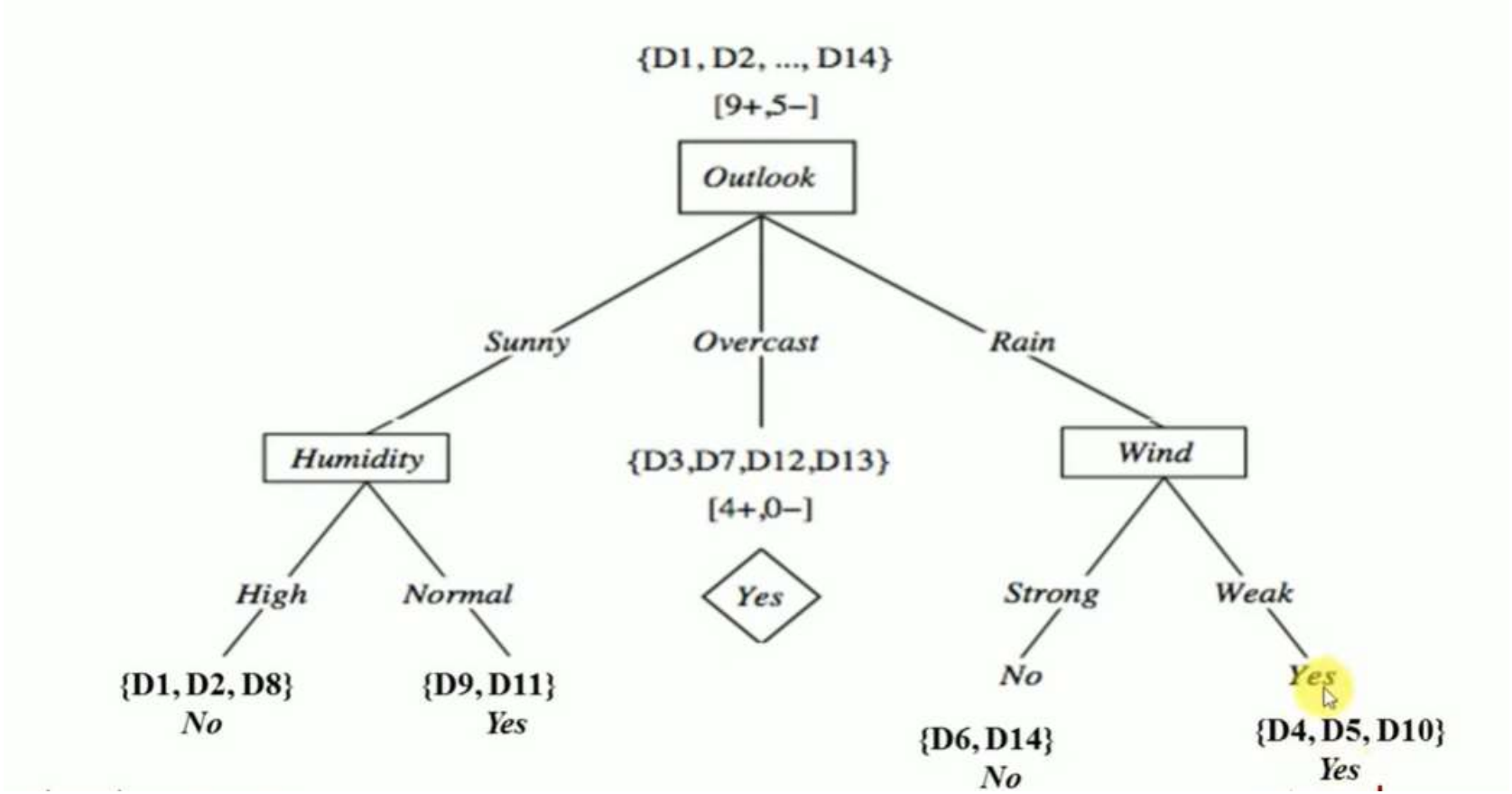
$$Gain(S_{Rain}, Wind) = 0.97 - \frac{2}{5}\ 0.0 - \frac{3}{5} 0.0 = 0.97$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

$$Gain(S_{Rain}, Wind) = 0.97$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
Humidity
{D3,D7,D12,D13}
[4+,0−]
Wind
High
Normal
Yes
Strong
Weak
{D1, D2, D8}
No
{D9, D11}
Yes
No
{D6, D14}
No
Yes
{D4, D5, D10}
Yes

Example 2

Find Entropy – Given Probabilities

- Given

$$p_1 = 0.1,\ p_2 = 0.2,\ p_3 = 0.3\ and\ p_4 = 0.4$$

- Find the Entropy ?

Solution

$p_1 = 0.1,\ p_2 = 0.2,\ p_3 = 0.3\ and\ p_4 = 0.4$

- $Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i)$
- $Entropy = -p_1 * \log_2(p_1) - p_2 * \log_2(p_2) - p_3 * \log_2(p_3) - p_4 * \log_2(p_4)$
- $Entropy = -0.1 * \log_2(0.1) - 0.2 * \log_2(0.2) - 0.3 * \log_2(0.3) - 0.4 * \log_2(0.4)$
- $Entropy = -0.1 * (-3.322) - 0.2 * (-2.322) - 0.3 * (-1.736) - 0.4 * (-1.322)$
- $Entropy = 0.3322 + 0.4644 + 0.5208 + 0.5288$
- $Entropy = 1.8462$

$$\log_2(0.1) = \frac{\log(0.1)}{\log(2)} = \frac{-1}{0.3010} = -3.322$$

Example - 3

- Consider the following data, where the Y label is whether or not the child goes out to play

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Solution

- **Step 1:** Calculate the IG (information gain) for each attribute (feature)

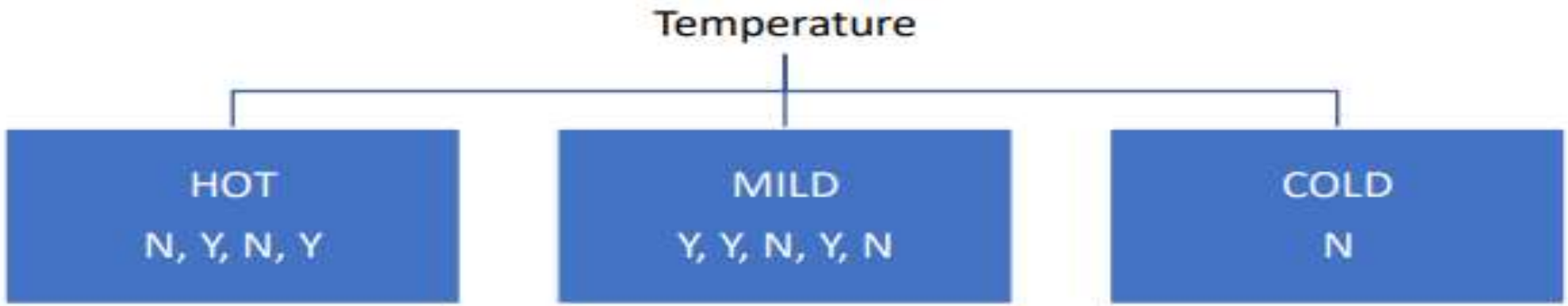
$$\text{Initial entropy} = H(Y) = -\sum_y P(Y = y) \log_2 P(Y = y)$$

$$= -P(Y = yes) \log_2 P(Y = yes) - P(Y = no) \log_2 P(Y = no)$$

$$= -(0.5) \log_2 (0.5) - (0.5) \log_2 (0.5)$$

$$= 1$$

Temperature:



Total entropy of this division is:

$$H(Y \mid temp) = -\sum_{x} P(temp = x) \sum_{y} P(Y = y \mid temp = x) \log_2 P(Y = y \mid temp = x)$$

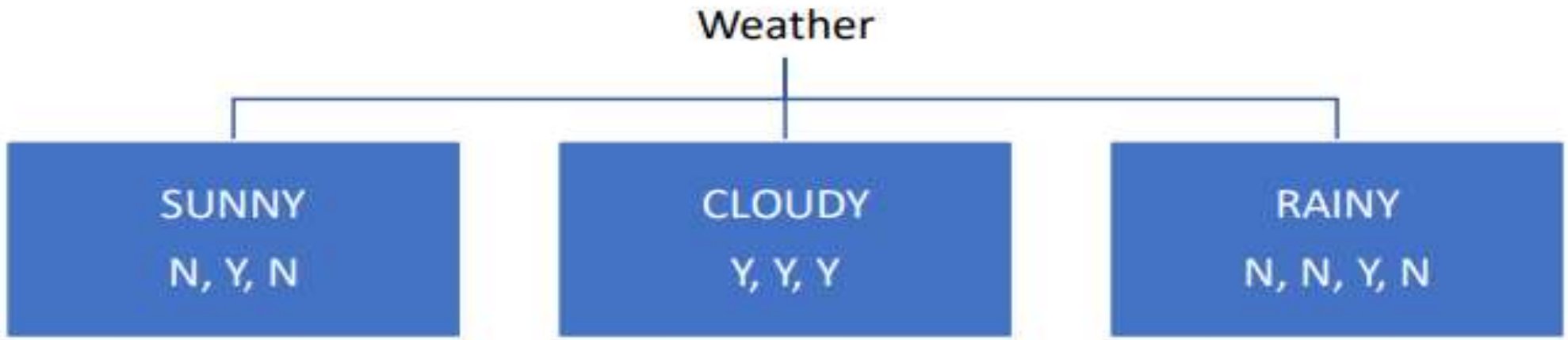
$$= -(P(temp = H) \sum_y P(Y = y \mid temp = H) \log_2 P(Y = y \mid temp = H) + P(temp = M) \sum_y P(Y = y \mid temp = M) \log_2 P(Y = y \mid temp = M) + P(temp = C) \sum_y P(Y = y \mid temp = C) \log_2 P(Y = y \mid temp = C))$$

$$= -((0.4)((\tfrac{1}{2}) \log_2 (\tfrac{1}{2}) + (\tfrac{1}{2}) \log_2 (\tfrac{1}{2})) + (0.5)((\tfrac{3}{5}) \log_2 (\tfrac{3}{5}) + (\tfrac{2}{5}) \log_2 (\tfrac{2}{5})) + (0.1)((1) \log_2 (1) + (0) \log_2 (0)))$$

$$= 0.7884$$

IG(Y, temp) = 1 – 0.7884 = 0.2116

Weather:



Total entropy of this division is:

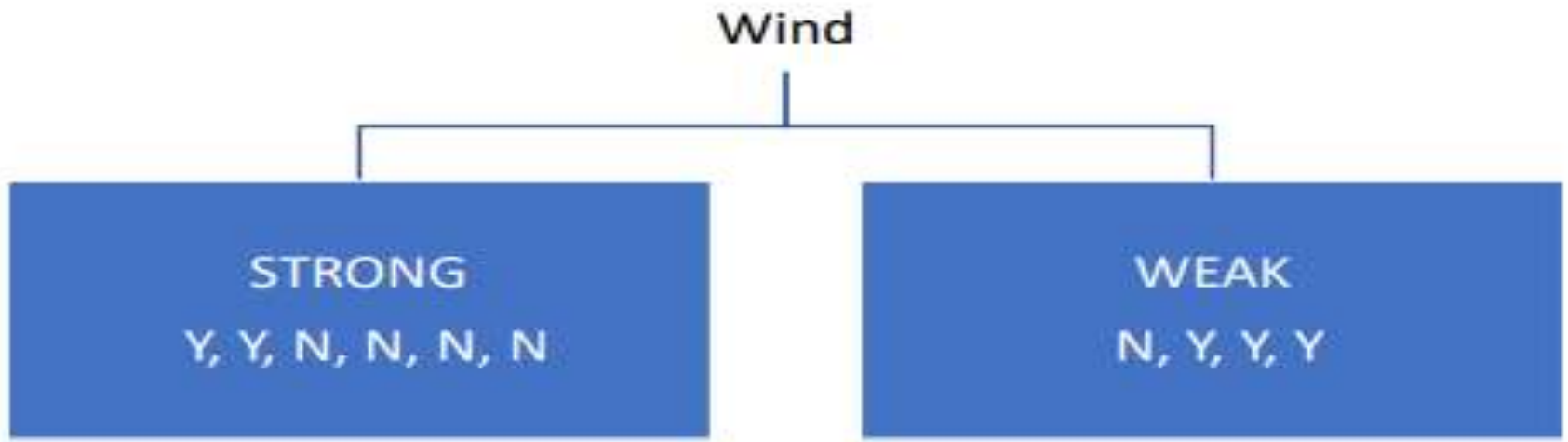
$$H(Y \mid weather) = -\sum_{x} P(weather = x) \sum_{y} P(Y = y \mid weather = x) \log_2 P(Y = y \mid weather = x)$$

$$= -(P(weather = S) \textstyle\sum_y P(Y = y \mid weather = S) \log_2 P(Y = y \mid weather = S) +$$
$$P(weather = C) \textstyle\sum_y P(Y = y \mid weather = C) \log_2 P(Y = y \mid weather = C) +$$
$$P(weather = R) \textstyle\sum_y P(Y = y \mid weather = R) \log_2 P(Y = y \mid weather = R))$$

$$= -((0.3)((\tfrac{1}{3}) \log_2 (\tfrac{1}{3}) + (\tfrac{2}{3}) \log_2 (\tfrac{2}{3})) + (0.3)((1) \log_2 (1) + (0) \log_2 (0)) +$$
$$(0.4)((\tfrac{1}{4}) \log_2 (\tfrac{1}{4}) + (\tfrac{3}{4}) \log_2 (\tfrac{3}{4})))$$

$$= 0.6$$

Wind:



Total entropy of this division is:

$$H(Y \mid wind) = -\sum_{x} P(wind = x) \sum_{y} P(Y = y \mid wind = x) \log_2 P(Y = y \mid wind = x)$$

$$= -(P(wind = S) \sum_y P(Y = y \mid wind = S) \log_2 P(Y = y \mid wind = S) + P(wind = W) \sum_y P(Y = y \mid wind = W) \log_2 P(Y = y \mid wind = W)$$

$$= -((0.6)(\tfrac{2}{6}) \log_2 \left(\tfrac{2}{6}\right) + (\tfrac{4}{6}) \log_2 (\tfrac{4}{6})) + (0.4)((\tfrac{1}{4}) \log_2 \left(\tfrac{1}{4}\right) + \left(\tfrac{3}{4}\right) \log_2 (\tfrac{3}{4}))$$

$$= 0.8755$$

IG(Y, wind) = 1 − 0.8755 = 0.1245

Step 2: Choose which feature to split with.

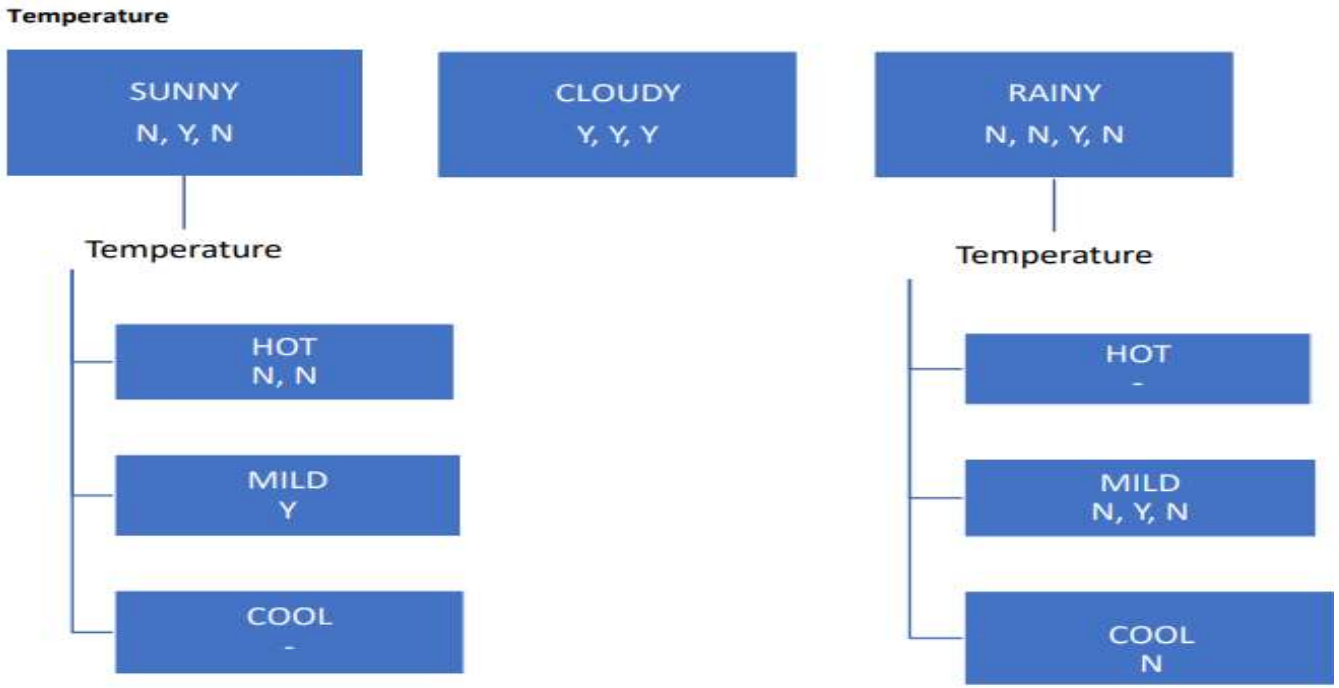
IG(Y, wind) = 0.1245

IG(Y, hum) = 0.1349

IG(Y, weather) = 0.4

IG(Y, temp) = 0.2116

Step 3: Repeat for each level



Entropy of "Sunny" node = $-((\frac{1}{3})\log_2(\frac{1}{3}) + (\frac{2}{3})\log_2(\frac{2}{3})) = 0.9183$

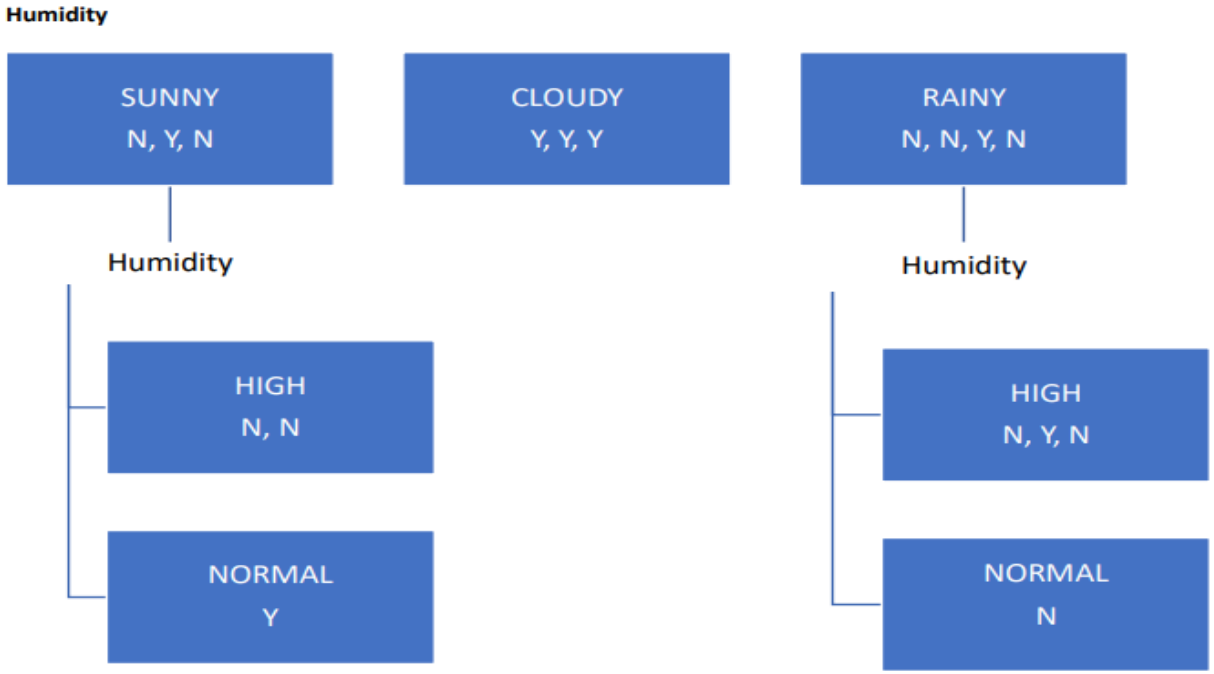
Entropy of its children = 0

IG = 0.9183

Entropy of "Rainy" node = $-((\frac{1}{4})\log_2(\frac{1}{4}) + (\frac{3}{4})\log_2(\frac{3}{4})) = 0.8113$

Entropy of children = $-(\frac{3}{4})((\frac{1}{3})\log_2(\frac{1}{3}) + (\frac{2}{3})\log_2(\frac{2}{3})) + 0 = 0.6887$

IG = 0.1226



Entropy of "Sunny" node = $-((\frac{1}{3})\log_2(\frac{1}{3}) + (\frac{2}{3})\log_2(\frac{2}{3})) = 0.9183$

Entropy of its children = 0

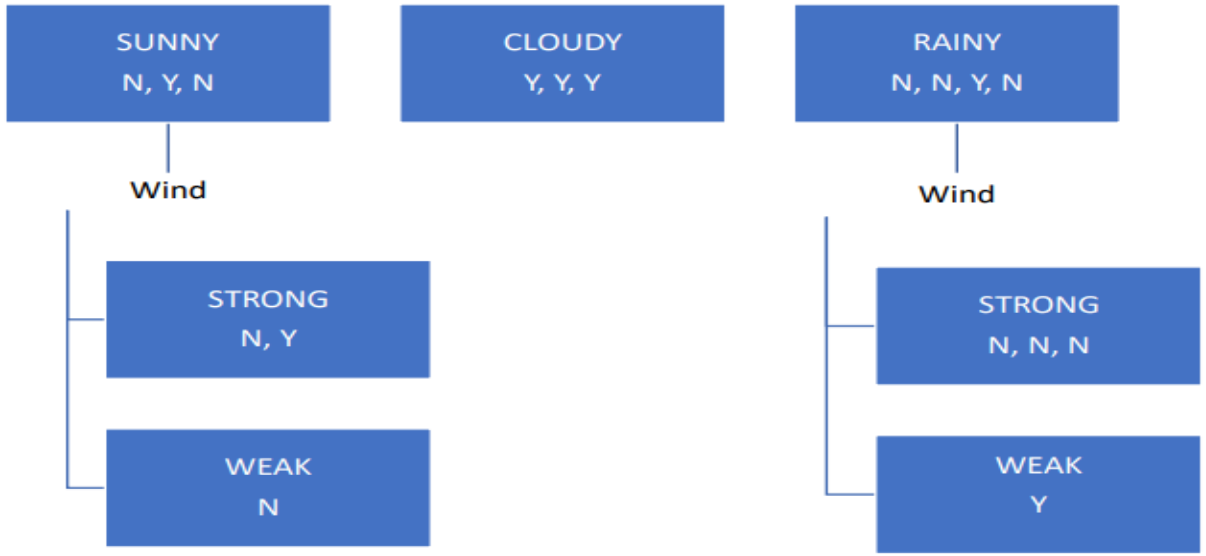
IG = 0.9183

Entropy of "Rainy" node = $-((\frac{1}{4})\log_2(\frac{1}{4}) + (\frac{3}{4})\log_2(\frac{3}{4})) = 0.8113$

Entropy of children = $-(\frac{3}{4})((\frac{1}{3})\log_2(\frac{1}{3}) + (\frac{2}{3})\log_2(\frac{2}{3})) + 0 = 0.6887$

IG = 0.1226

Wind



Entropy of "Sunny" node = $-((\frac{1}{3})\log_2(\frac{1}{3}) + (\frac{2}{3})\log_2(\frac{2}{3})) = 0.9183$

Entropy of its children = $-(\frac{2}{3})((\frac{1}{2})\log_2(\frac{1}{2}) + (\frac{1}{2})\log_2(\frac{1}{2})) + 0 = 0.6667$

IG = 0.2516

Entropy of "Rainy" node = $-((\frac{1}{4})\log_2(\frac{1}{4}) + (\frac{3}{4})\log_2(\frac{3}{4})) = 0.8113$

Entropy of children = 0

IG = 0.8113

Step 4: Choose a Feature for each node.

"Sunny node":

IG(Y, weather) = IG(humidity) = 0.9183

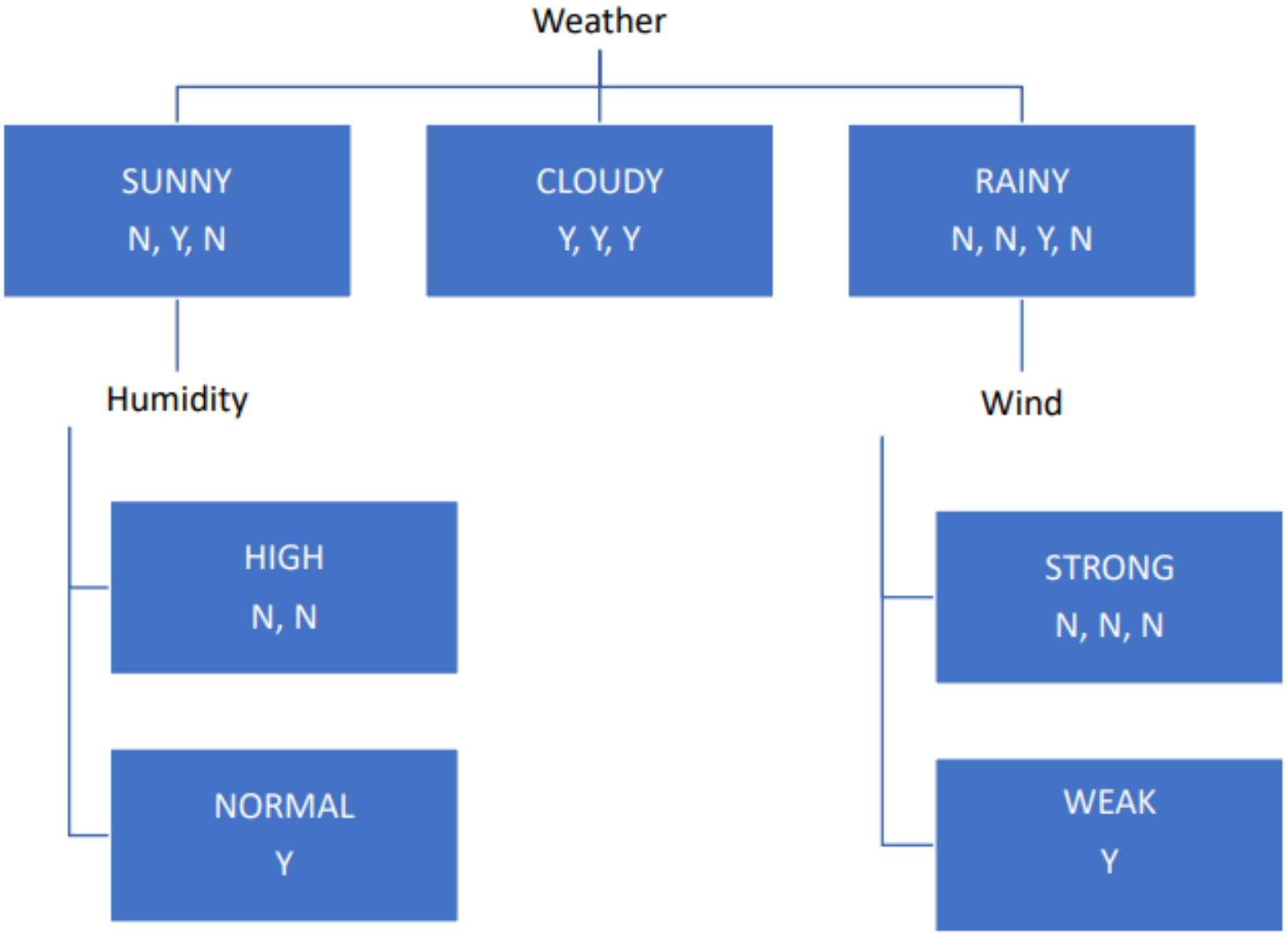
IG(Y, wind) = 0.2516

"Rainy node":

IG(Y, weather) = IG(Y, humidity) = 0.1226

IG(Y, wind) = 0.8113

Step – 5 - Final Tree



Example - 6

- Consider the table given. It represent factors that affect whether John would go out to play golf or not. Using the data in the table, build a decision tree to model that can be used to predict if John would play golf or not.

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

Step 1: Determine the Decision Column

- Since decision trees are used for classification, you need to determine the classes that are the basis for the decision.

•

In this case, it is the last column, that is ***Play Golf*** column with classes ***Yes*** and ***No***.

- To determine the rootNode we need to compute the entropy. To do this, we create a frequency table for the classes (the **Yes/No** column).

Play Golf(14)	
Yes	No
9	5

Step 2: Calculating Entropy for the classes (Play Golf)

- In this step, you need to calculate the entropy for the Play Golf column and the calculation step is given below.

$$Entropy(PlayGolf) = \boldsymbol{E}(5,9)$$

$$\boldsymbol{E(PlayGolf)} = E(5,9)$$

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

$$= -(0.357\ \log_2 0.357) - (0.643\ \log_2 0.643)$$

$$= \mathbf{0.94}$$

Step 3: Calculate Entropy for Other Attributes After Split

For the other four attributes, we need to calculate the entropy after each of the split.

- E(PlayGolf, Outloook)
- E(PlayGolf, Temperature)
- E(PlayGolf, Humidity)
- E(PlayGolf,Windy)

The entropy for two variables is calculated using the formula.

$$Entropy(S,T) = \sum_{c \in T} P(c)E(c)$$

- The easiest way to approach this calculation is to create a frequency table for the two variables, that is PlayGolf and Outlook.

This frequency table is given below:

		PlayGolf(14)		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5

Table 3: Frequency Table for Outlook

Using this table, we can then calculate E(PlayGolf, Outlook), which would then be given by the formula below

$$E(PlayGolf, Outlook) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

Let's go ahead to calculate E(3,2)

We would not need to calculate the second and the third terms! This is because

E(4, 0) = 0

E(2,3) = E(3,2)

$$E(Sunny) = E(3,2)$$

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

$$= -(0.60\ \log_2 0.60) - (0.40\log_2 0.40)$$

$$= -(0.60 * 0.737) - (0.40 * 0.529)$$

$$= \mathbf{0.971}$$

Just for clarification, let's show the the calculation steps

The calculation steps for E(4,0):

$$E(Overcast) = E(4,0)$$

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

$$= -(0) - (0)$$

$$= \mathbf{0}$$

The calculation step for E(2,3) is given below

$$E(Rainy) = E(2,3)$$

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

$$= -(0.40\ \log_2 0.40) - (0.6\log_2 0.60)$$

$$= \mathbf{0.971}$$

Time to put it all together.

We go ahead to calculate the **E**(PlayGolf, Outlook) by substituting the values we calculated from **E**(Sunny), **E**(Overcast) and **E**(Rainy) in the equation:

$$E(PlayGolf, Outlook) = P(Sunny)\,E(3,2) + P(Overcast)\,E(4,0) + P(rainy)\,E(2,3)$$

$$E(PlayGolf, Outlook) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

$$= \frac{5}{14}0.971 + \frac{4}{14}0.0 + \frac{5}{14}0.971$$

$$= 0.357 * 0.971 + 0.0 + 0.357 * 0.971$$

$$= 0.693$$

E(PlayGolf, Temperature) Calculation

Just like in the previous calculation, the calculation of E(PlayGolf, Temperature) is given below. It

It is easier to do if you form the frequency table for the split for Temperature as shown.

		PlayGolf(14)		
		Yes	No	
Temperature	Hot	2	2	4
	Cold	3	1	4
	Mild	4	2	6

Table 4: Frequency Table for Temperature

E(PlayGolf, Temperature) = P(Hot) E(2,2) + P(Cold) E(3,1) + P(Mild) E(4,2)

$$E(\text{PlayGolf, Temperature}) = \mathbf{4/14} * \mathbf{E}(\text{Hot}) + \mathbf{4/14} * \mathbf{E}(\text{Cold}) + \mathbf{6/14} * \mathbf{E}(\text{Mild})$$

$$E(\text{PlayGolf, Temperature}) = \mathbf{4/14} * \mathbf{E}(2, 2) + \mathbf{4/14} * \mathbf{E}(3, 1) + \mathbf{6/14} * \mathbf{E}(4, 2)$$

$$E(\text{PlayGolf, Temperature}) = 4/14 * -(2/4 \log 2/4) - (2/4 \log 2/4)$$
$$+ 4/14 * -(3/4 \log 3/4) - (1/4 \log 1/4)$$
$$+ 6/14 * -(4/6 \log 4/6) - (2/6 \log 2/6)$$

$$E(\text{PlayGolf, Temperature}) = 5/14 * 1.0$$
$$+ 4/14 * 1.811$$
$$+ 5/14 * 0.918$$
$$= \mathbf{0.911}$$

E(PlayGolf, Humidity) Calculation

Just like in the previous calculation, the calculation of E(PlayGolf, Humidity) is given below. It
It is easier to do if you form the frequency table for the split for Humidity as shown.

		PlayGolf(14)		
		Yes	No	
Humidity	High	3	4	7
	Normal	6	1	7

Table 5: Frequency Table for Humidity

$$E(\text{PlayGolf, Humidity}) = \mathbf{7/14} * E(\text{High}) + \mathbf{7/14} * E(\text{Normal})$$

$$E(\text{PlayGolf, Humidity}) = \mathbf{7/14} * \mathbf{E}(3, 2) + \mathbf{7/14} * \mathbf{E}(4, 0)$$

$$E(\text{PlayGolf, Humidity}) = 7/14 * -(3/7 \log 3/7) - (4/7 \log 4/7)$$
$$+ 7/14 * -(6/7 \log 6/7) - (1/7 \log 1/7)$$

$$E(\text{PlayGolf, Humidity}) = 7/14 * 0.985$$
$$+ 7/14 * 0.592$$
$$= \mathbf{0.788}$$

E(PlayGolf, Windy) Calculation

Just like in the previous calculation, the calculation of E(PlayGolf, Windy) is given below. It

It is easier to do if you form the frequency table for the split for Windy as shown.

		PlayGolf(14)		
		Yes	No	
Windy	TRUE	3	3	6
	FALSE	6	2	8

Table 6: Frequency Table for Windy

$$E(\text{PlayGolf, Windy}) = \mathbf{6/14} * \mathbf{E}(\text{True}) + \mathbf{8/14} * \mathbf{E}(\text{False})$$

$$E(\text{PlayGolf, Windy}) = \mathbf{6/14} * \mathbf{E}(3, 3) + \mathbf{8/14} * \mathbf{E}(6, 2)$$

$$E(\text{PlayGolf, Windy}) = 6/14 * -(3/6 \log 3/6) - (3/6 \log 3/6)$$
$$+ 8/14 * -(6/8 \log 6/8) - (2/8 \log 2/8)$$

$$E(\text{PlayGolf, Windy}) = 6/14 * 1.0$$
$$+ 8/14 * 0.811$$
$$= \mathbf{0.892}$$

- So now that we have all the entropies for all four attributes, let's go ahead to summarize them as shown in below:

1. E(PlayGolf, Outloook) = **0.693**
2. E(PlayGolf, Temperature) = **0.911**
3. E(PlayGolf, Humidity) = **0.788**
4. E(PlayGolf,Windy) = **0.892**

Step 4: Calculating Information Gain for Each Split

- The next step is to calculate the information gain for each of the attributes.

- The information gain is calculated from the split using each of the attributes. Then the attribute with the largest information gain is used for the split.

The information gain is calculated using the formula:

Gain(S,T) = Entropy(S) – Entropy(S,T)

For example, the information gain after spliting using the Outlook attibute is given by:

Gain(PlayGolf, Outlook) = Entropy(PlayGolf) – Entropy(PlayGolf, Outlook)

So let's go ahead to do the calculation

Gain(PlayGolf, Outlook) = Entropy(PlayGolf) – Entropy(PlayGolf, Outlook)

= 0.94 – 0.693 **= 0.247**

Gain(PlayGolf, Temperature) = Entropy(PlayGolf) – Entropy(PlayGolf, Temparature)

= 0.94 – 0.911 **= 0.029**

Gain(PlayGolf, Humidity) = Entropy(PlayGolf) – Entropy(PlayGolf, Humidity)

= 0.94 – 0.788 **= 0.152**

Gain(PlayGolf, Windy) = Entropy(PlayGolf) – Entropy(PlayGolf, Windy)

= 0.94 – 0.892 **= 0.048**

Having calculated all the information gain, we now choose the attribute that gives the highest information gain after the split.

Step 5: Perform the First Split

- **Draw the First Split of the Decision Tree**

- Now that we have all the information gain, we then split the tree based on the attribute with the highest information gain.

- From our calculation, the highest information gain comes from Outlook. Therefore, the split will look like this:

Figure 2: Decision Tree after first split

Now that we have the first stage of the decison tree, we see that we have one leaf node. But we still need to split the tree further.

To do that, we need to also split the original table to create sub tables.

This sub tables are given in below.

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

Table 7: Initial Split using Outlook

From Table 3, we could see that the Overcast outlook requires no further split because it is just one homogeneous group. So, we have a leaf node.

Step 6: Perform Further Splits

- The Sunny and the Rainy attributes needs to be split
- The Rainy outlook can be split using either Temperature, Humidity or Windy.
- What attribute would best be used for this split? Why?
- **Humidity**. Because it produces homogenous groups.

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Table 8: Split using Humidity

- The Rainy attribute could be split using **High** and **Normal** attributes and that would give us the tree as shown.

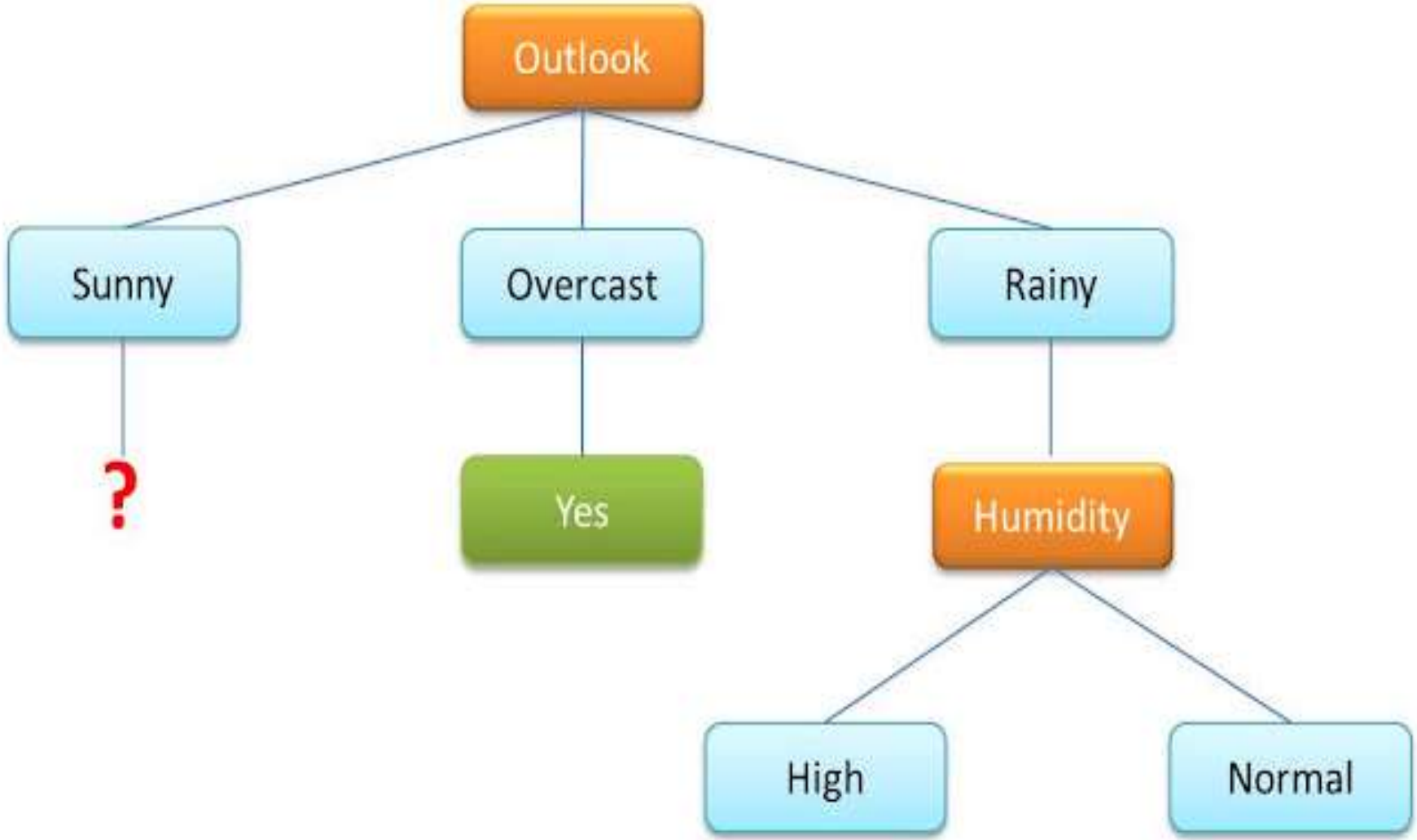


Figure 3: Split using the Humidity Attribute

- Let’s now go ahead to do the same thing for the Sunny outlook.

- The Rainy outlook can be split using either Temperature, Humidity or Windy.

- What attribute would best be used for this split? Why?

- **Windy** . Because it produces homogeneous groups.

Outlook	Temperature	Humidity	Windy	Play Golf
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes

Sunny	Cool	Normal	TRUE	No
Sunny	Mild	High	TRUE	No

Table 9: Split using Windy Attribute

- If we do the split using the Windy attribute, we would have the final tree that would require no further splitting! This is shown in the next Figure.

Step 7: Complete the Decision Tree

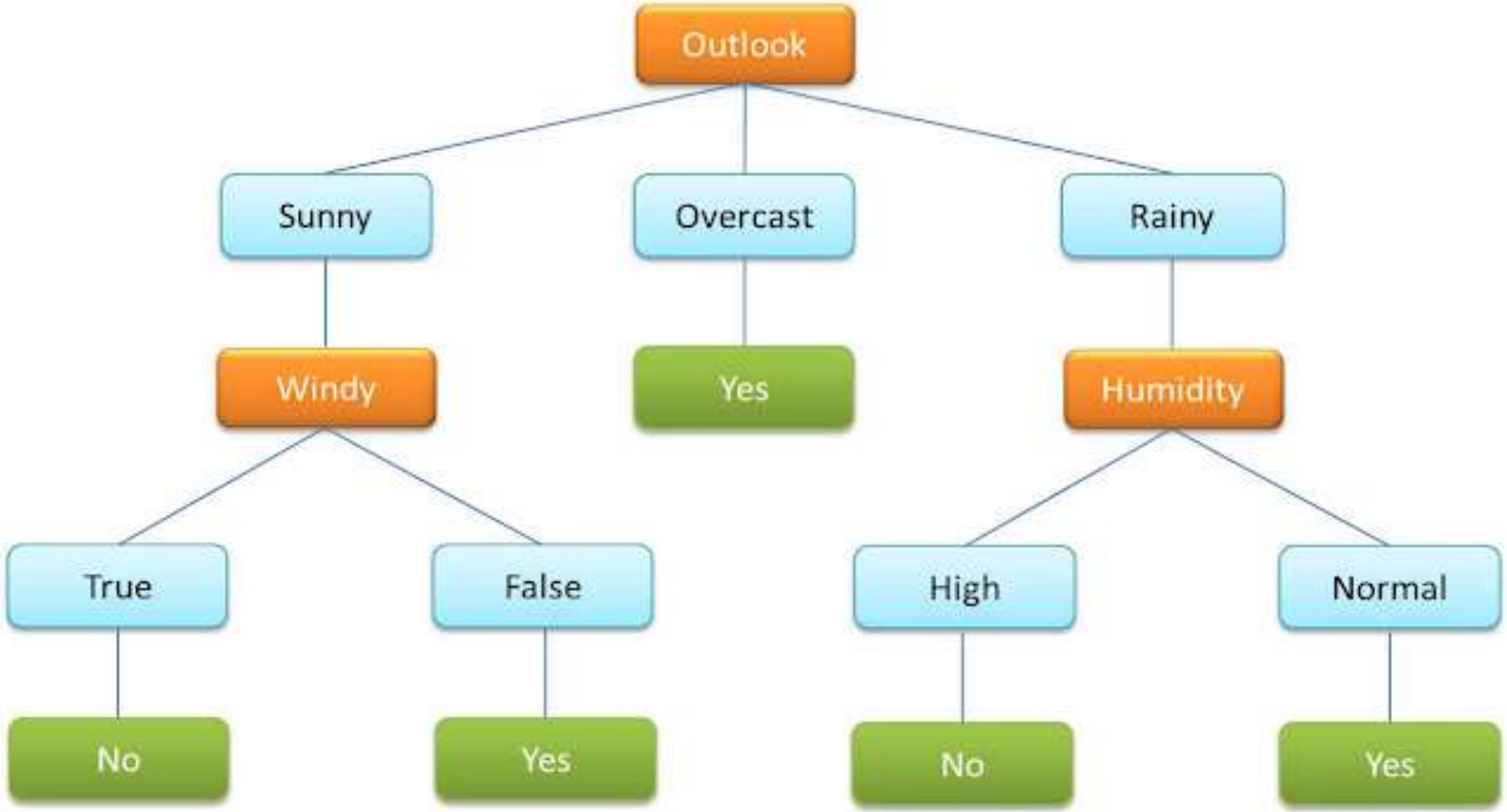


Figure 4: Final Decision Tree

2. Gini Index

- **Gini index** is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.

- An attribute with the low Gini index should be preferred

- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.

- Gini index can be calculated using the below formula:

$$Gini\ Index = 1 - \sum_j P_j^2$$

Pruning: Getting an Optimal Decision Tree

- ***Pruning*** *is a process of deleting unnecessary nodes from a tree to get the optimal decision tree.*

- A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset.

- Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as **Pruning**

- . There are mainly two types of tree **pruning** technology used:

1. **Cost Complexity Pruning**
2. **Reduced Error Pruning.**

Advantages of the Decision Tree

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

Disadvantages of the Decision Tree

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the **Random Forest algorithm**.
- For more class labels, the computational complexity of the decision tree may increase.

Decision Tree: - In Short

Basic algorithm (a greedy algorithm)

- Tree is constructed in a top-down (from general to specific) recursive divide-and-conquer manner
- At start, all the training examples are at the root
- Attributes are categorical (if continuous-valued, discretization in advance)
- Examples are partitioned recursively based on selected attributes
- Attributes are selected based on heuristic or statistical measure (e.g., information gain)

When to stop

- All example for a given node belong to the same class (pure), or
- No remaining attributes to select from, or
- majority voting to determine class label for the node
- No examples left