Classification- Decision trees and Naive Bayes

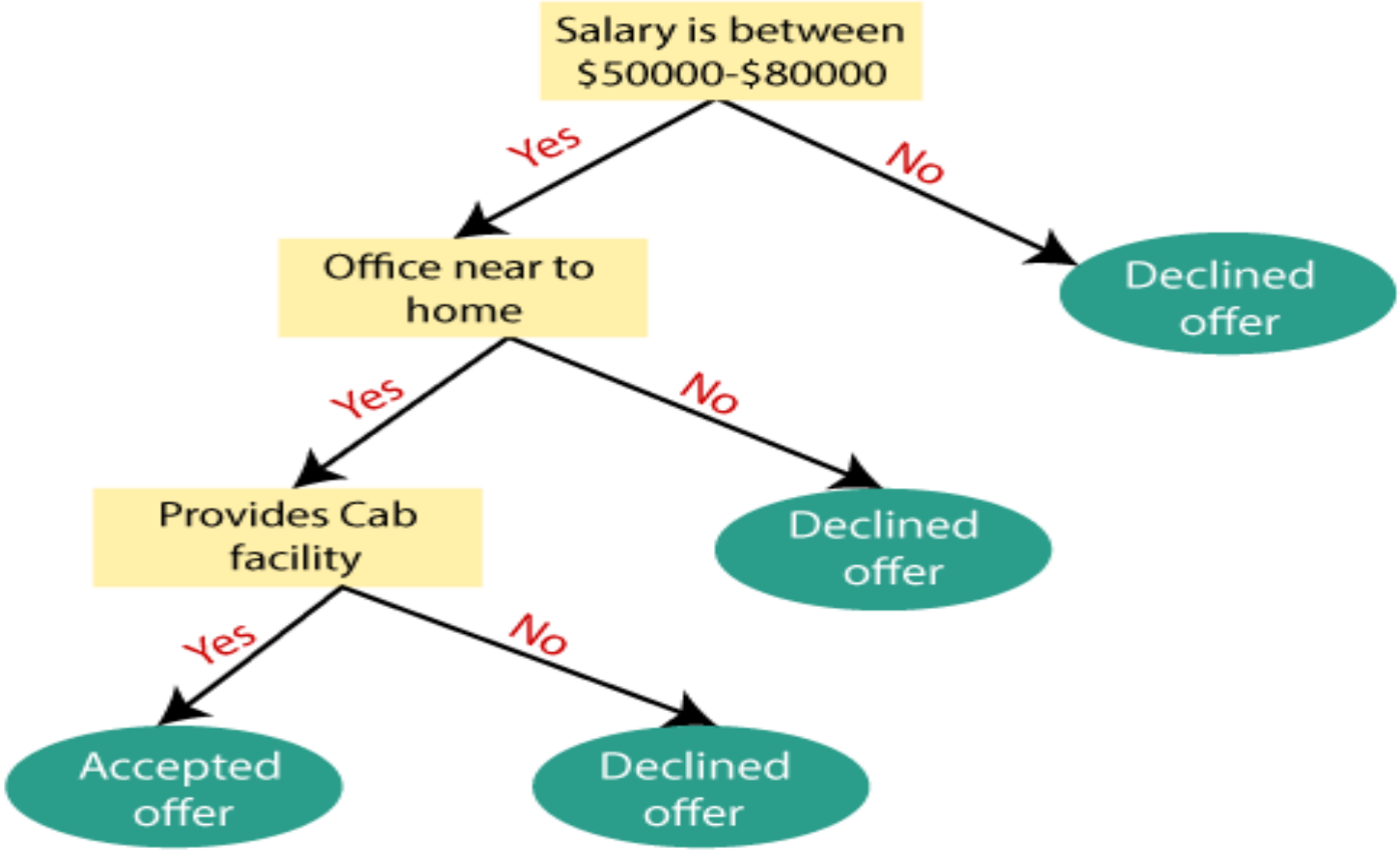
By

Suchita Patil

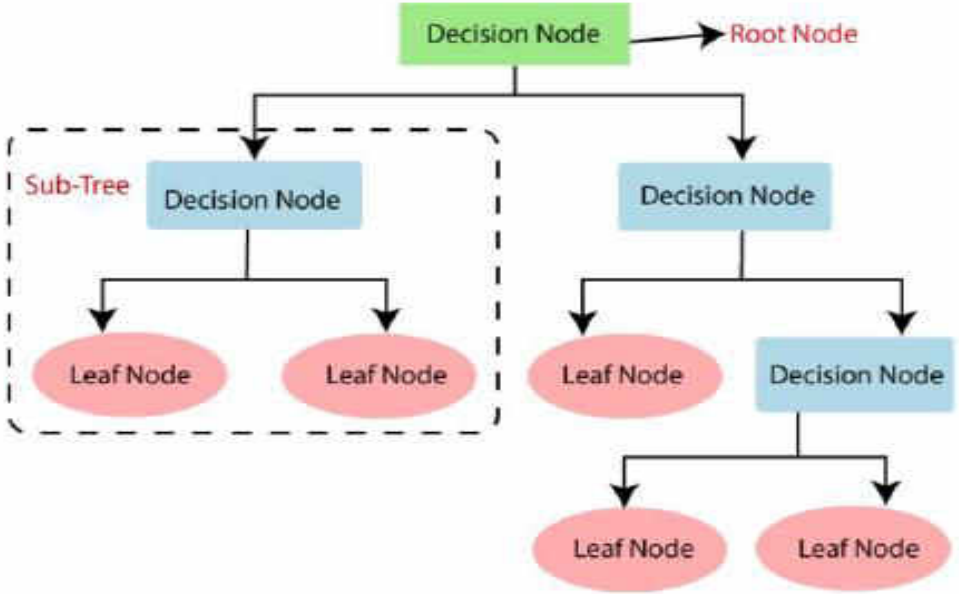
Contents:

- Decision trees – definition, terminology, the need, advantages, and limitations.
- Constructing and understanding decision trees.
- Common problems with decision trees.
- Decision tree algorithms – ID3, CART, random forest, examples.
- Naïve Bayes classifier.
- Instance-based classifier – K–Nearest Neighbour classifier.

Example


Salary is between $50000-$80000
Yes
No
Office near to home
Declined offer
Yes
No
Provides Cab facility
Declined offer
Yes
No
Accepted offer
Declined offer

Decision trees naturally induce rules that can be used in data classification and prediction.



- **Definition:**

  Decision Tree is a Supervised learning technique that can be used for both classification and Regression 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.

Terminology

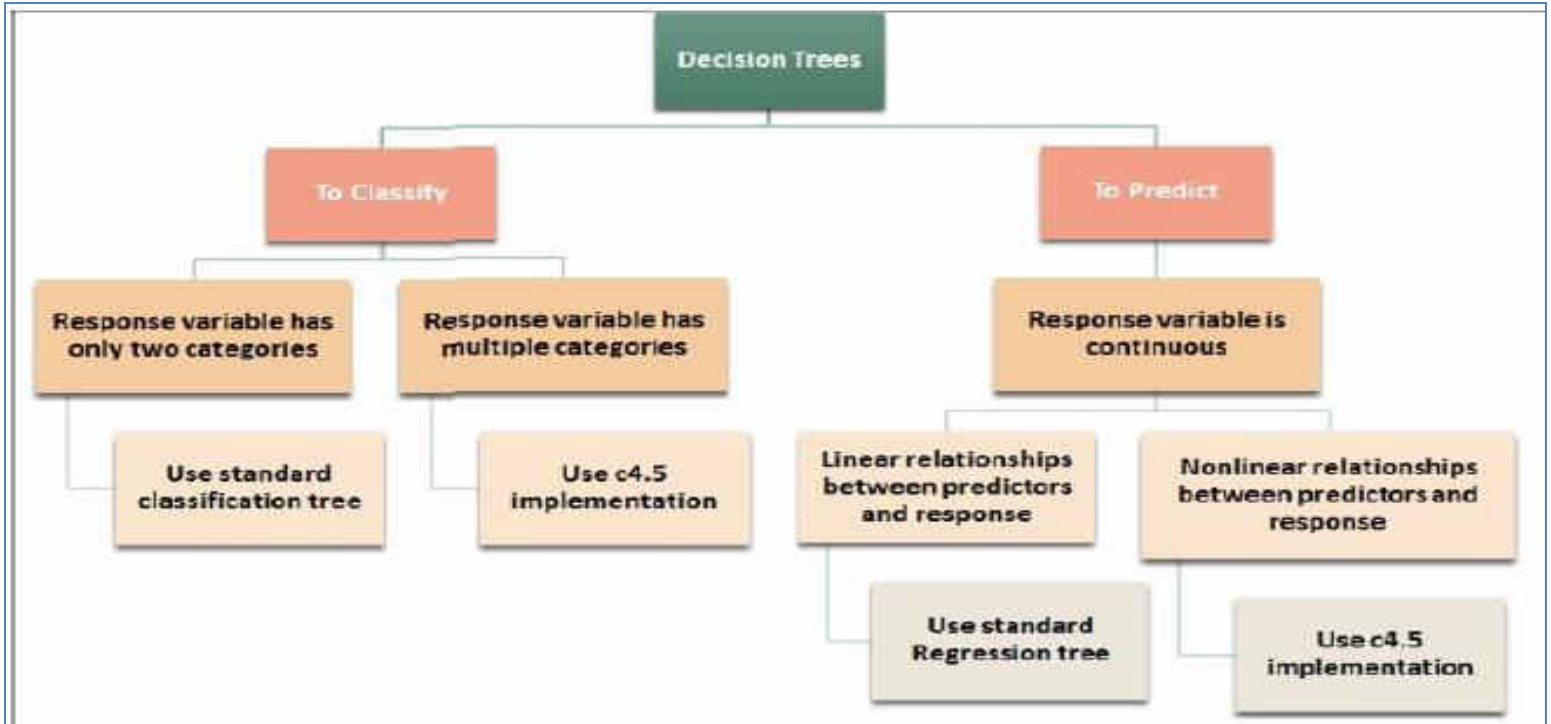
1. **Root Node**: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
2. **Leaf Node**: Leaf nodes are the final output node that represents the value of the target attribute, and the tree cannot be segregated further after getting a leaf node.
3. **Decision node**: Every non-leaf node denotes a representation of the attribute value
4. **Splitting**: Splitting is the process of dividing the decision node/root node into subnodes according to the given conditions.
5. **Branch/Subtree**: A tree formed by splitting the tree.

Terminology

6. **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
7. **Parent/Child node**: One node is parent node of other if it is above of the node (closer to the root node).
8. **Siblings:** Two nodes are sibling if they share same parent node.
9. **Edge:** Connection between two nodes.
10. **Degree:** Count of sub tree of any node.
11. **Height of a node:** Number of edges from node to leaf node having longest edges in the middle.
12. **Height of tree:** Height of root node.

The Need

- Decision trees are used for classification and regression. Two types of trees are used in this context:
- Classification trees
- Regression trees



Advantages of Decision Tree

1. Easy to explain to others. It does not need any complex mathematical knowledge to understand the result.
2. Can handle irrelevant attributes.
3. Can capture nonlinear relationships in the data.
4. Decision tree learning, or construction of tree is fast process. It uses greedy algorithms to create trees. Also, prediction by using tree is fast.
5. Normalization and other cleaning on data is not needed.
6. It is not based on assumptions. In linear methods, we have to test some assumptions before building the model. No assumptions are needed in decision tree learning.
7. Can handle both numerical and categorical attributes.
8. They are easy to understand and interpret
9. Decision trees do not require complex data preparation
10. They can handle highly dimensional data and also operate large datasets

Limitations/ Disadvantage of Decision Tree

1. Decision tree needs to stop somewhere. Otherwise they will give no error on training data and high error on test data.
2. Prone to overfitting.
3. Require some kind of measurement as to how well they are doing.
4. Can create biased learned trees if some classes dominate.
5. The decision tree contains lots of layers, which makes it complex.
6. For more class labels, the computational complexity of the decision tree may increase.

Constructing a Decision tree:

- 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 move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

Algorithm

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

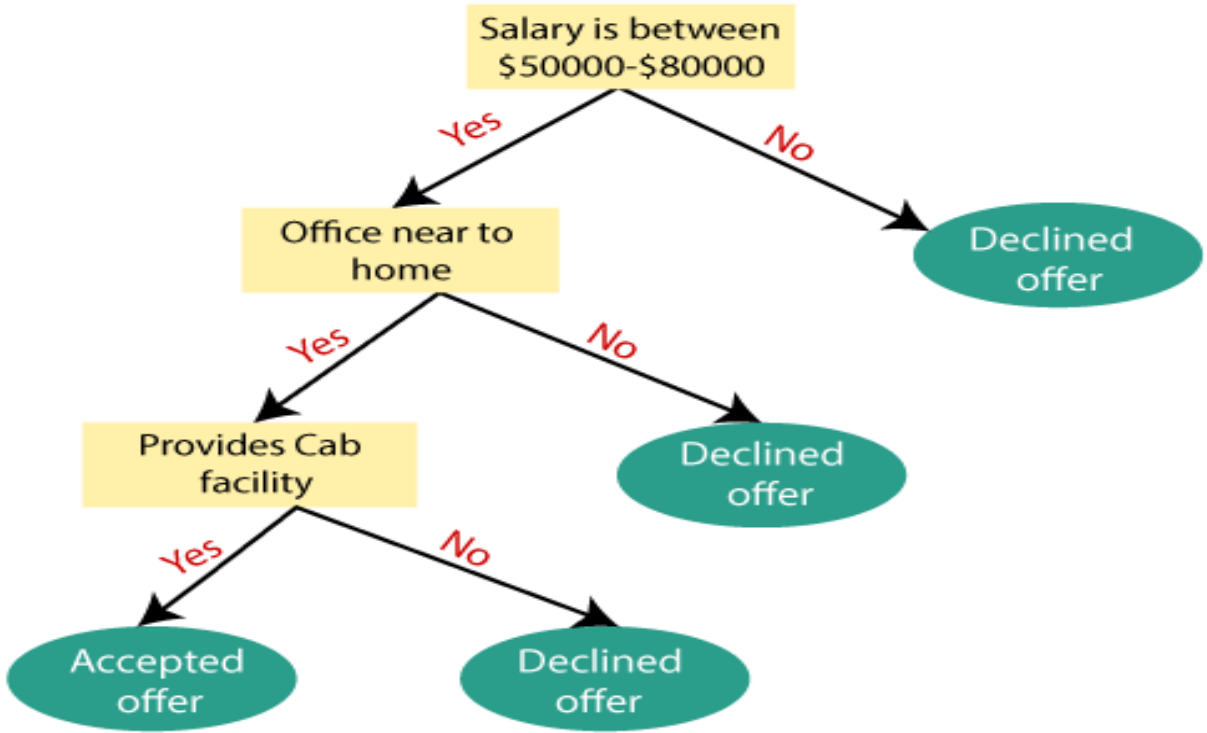
Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3.

Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example: Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. The decision tree starts with the root node. 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.



Considerations for constructing Decision trees

The key to constructing Decision trees is knowing where to split them. To do this, we need to be clear on the following:

- Which attribute to start and which attribute to apply subsequently?
- When do we stop building the Decision tree (that is to avoid over-fitting)?

Choosing the appropriate attribute(s): (ASM-Attribute Selection Measure)

There are three different ways to identify the best-suited attributes:

- Information Gain uses Entropy
- Gini index
- Gain ratio

Decision tree Algorithms
1. ID3 (Iterative Dichotomiser 3)

• The core algorithm for building decision trees is called **ID3 by J. R. Quinlan which** employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses *Entropy and Information Gain to construct a decision* tree.

• **Steps in ID3 algorithm:**

1. It begins with the original set S as the root node.
2. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates **Entropy(H) and Information gain(IG) of this attribute.**
3. It then selects the attribute which has the smallest Entropy or Largest Information gain.
4. The set S is then split by the selected attribute to produce a subset of the data.
5. The algorithm continues to recur on each subset, considering only attributes never selected before.

Information Gain

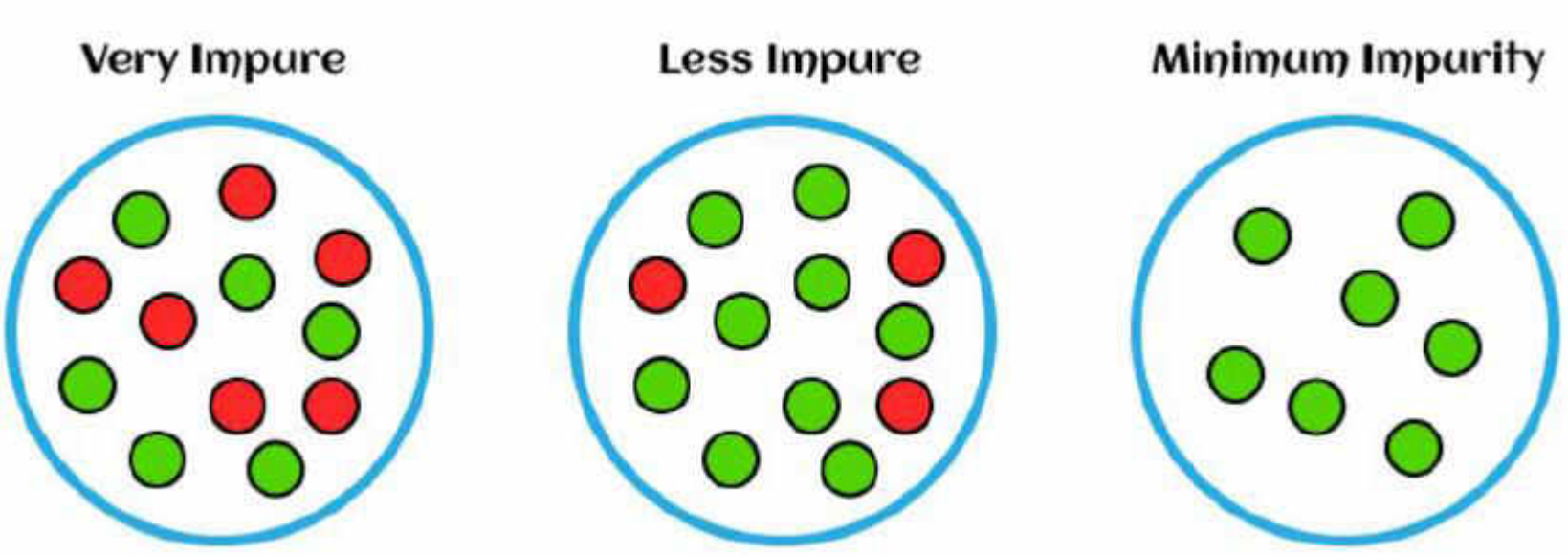
- When we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information

gain is a measure of this change in entropy.

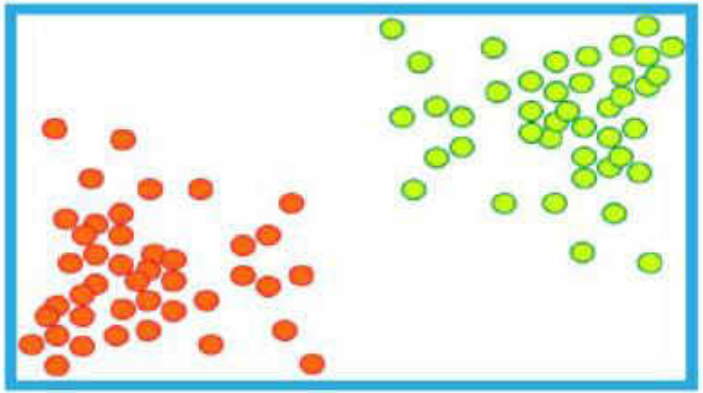
- ***Definition:*** Suppose S is a set of instances, A is an attribute**,** Sv is the subset of S with A = v, and Values (A) is the set of all possible values of A, then

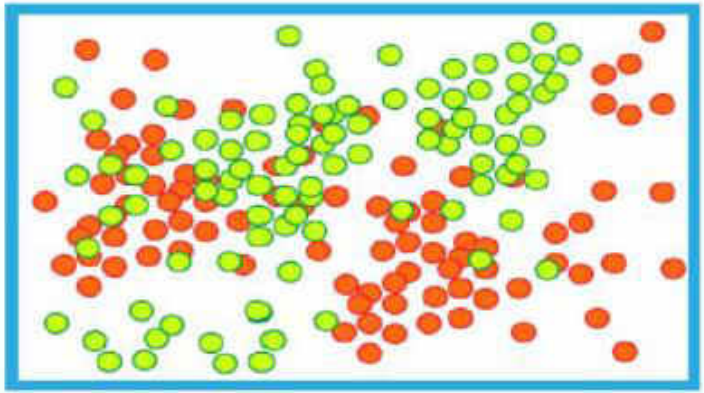
$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|}.Entropy(S_v)$$

- **Entropy is the measurement of disorder or impurities in the** information processed in machine learning. It determines how a decision tree chooses to split data.



- When information is processed in the system, then every piece of information has a specific value to make and can be used to draw conclusions from it. So if it is easier to draw a valuable conclusion from a piece of information, then entropy will be lower in Machine Learning, or if entropy is higher, then it will be difficult to draw any conclusion from that piece of information.
- **example: flipping a coin. When we flip a coin, then there can be two outcomes.** However, it is difficult to conclude what would be the exact outcome while flipping a coin because there is no direct relation between flipping a coin and its outcomes. There is a 50% probability of both outcomes; then, in such scenarios, entropy would be high. This is the essence of entropy in machine learning.





Mathematical Formula for Entropy

• Consider a data set having a total number of N classes,

then the entropy (E) can be determined with the formula below:

• Where;

Pi = Probability of randomly selecting an example in class I;

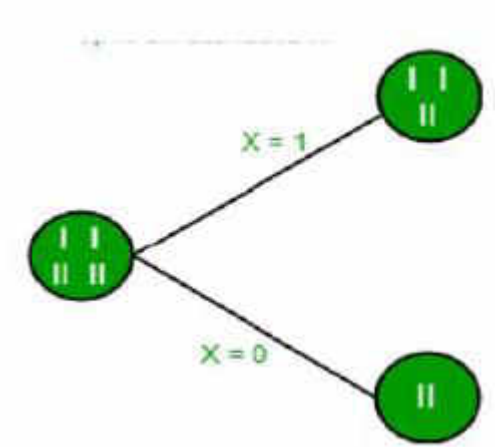
$$E = -\sum_{i=1}^{N} P_i \log_2 P_i$$

E.g.Draw a Decision Tree for the following data using Information gain.

- Training set: 3 features and 2 classes
- To build a decision tree using Information gain. We will take each of the feature and calculate the information for each feature.

Split on feature X

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II



$E_{parent} = 1$

$E_{child1} = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3)$

$= 0.5284 + 0.39$

$= 0.9184$

$E_{child2} = 0$

$GAIN = 1 - (3/4)(0.9184) - (1/4)(0) = 0.3112$

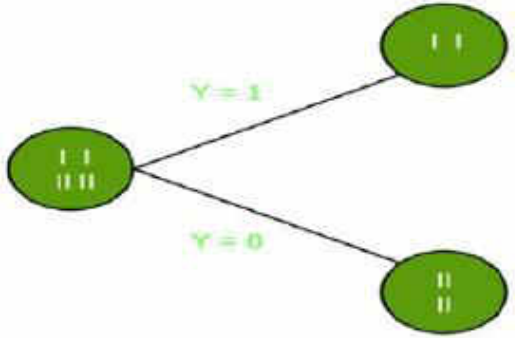
- **Split on feature Y**

$E_{parent} = 1$

$E_{child1} = 0$

$E_{child2} = 0$

GAIN = 1 - (1/2)(0) - (1/2)(0) = 1



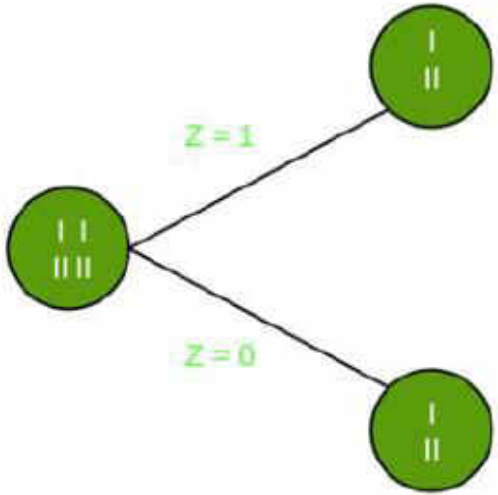
- **Split on feature Z**

$E_{parent} = 1$

$E_{child1} = 1$

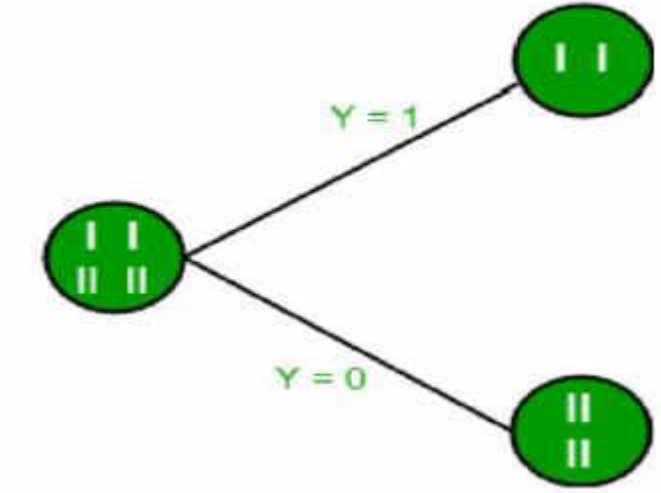
$E_{child2} = 1$

GAIN = 1 - (1/2)(1) - (1/2)(1) = 1



From the above images we can see that the information gain is maximum when we make a split on feature Y. So, for the root node best suited feature is feature Y. Now we can see that while splitting the dataset by feature Y, the child contains pure subset of the target variable. So we don't need to further split the dataset.

• The final tree for the above dataset would be look like this:



Data set

Day	Weather	Temp.	Humidity	Wind	Decision
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

Calculate IG of Weather

- **Step1: Entropy of entire dataset**

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

- **Step2: Entropy of all attributes:**
- **Entropy of Sunny** $\{+2,-3\} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$
- **Entropy of Cloudy** $\{+4,-0\} = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$
- **Entropy of Rain** $\{+3,-2\} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$
- **Information Gain= Entropy(whole data)** $-\frac{5}{14}Ent(S) - \frac{4}{14}Ent(C) - \frac{5}{14}Ent(R)$

$= 0.246$

Calculate IG of Temperature

- **Step1: Entropy of entire dataset**

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

- **Step2: Entropy of all attributes:**
- **Entropy of Hot** $\{+2,-2\} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$
- **Entropy of Mild** $\{+4,-2\} = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.91$
- **Entropy of Cold** $\{+3,-1\} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{3}\log_2\frac{1}{3} = 0.81$
- **Information Gain= Entropy(whole data)** $-\frac{4}{14}\text{Ent(H)} - \frac{6}{14}\text{Ent(M)} - \frac{4}{14}\text{Ent(C)}$

$= 0.029$