Decision Tree

Decision tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables.

In this technique, we split the population or sample into two or more homogeneous sets based on the 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.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas leaf nodes are the output of those decisions and do not contain any further branches. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems, the most significant split in input variable.

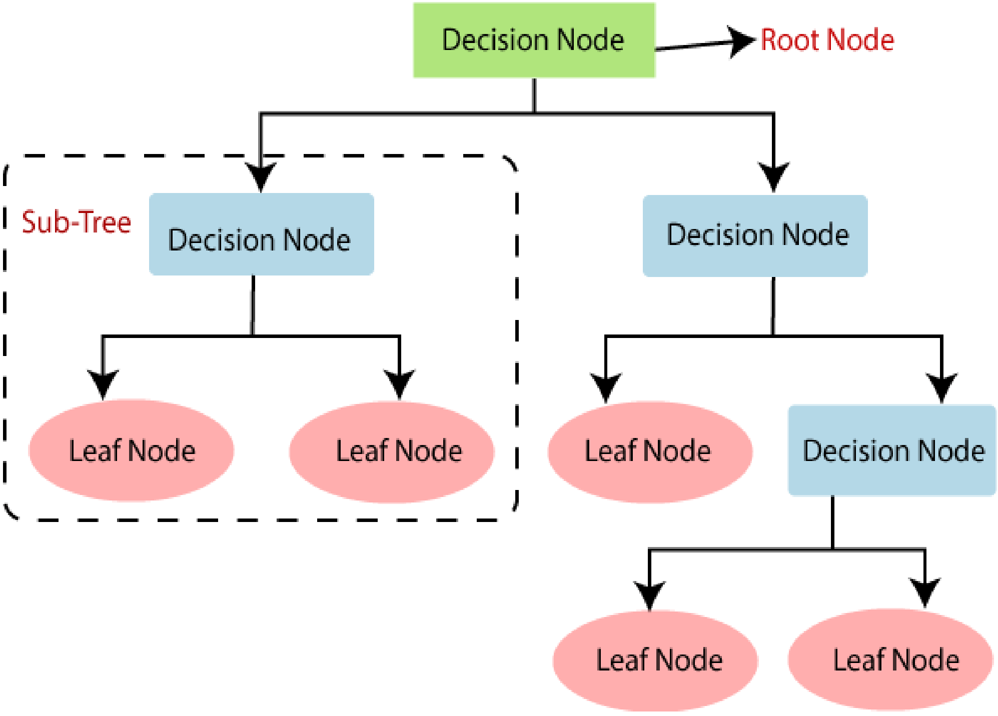
There are two types of Decision Tree and it is based on the type of target variable we have.They are

1. Categorical variable Decision Tree
2. Continous variable Decision Tree

Categorical variable Decision Tree: Decision Tree which has a categorical target variable then it is called a Categorical variable decision tree

Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

How does it works?



In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. The complete process can be better understood using the below steps:

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 contain 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 call the final node as a leaf node.

The primary challenge in the Decision Tree implementation is to identify the attributes which we consider as the root node and each level. This process is known as attributes selection measures.

There are 2 popular attribute selection measures. They are as follows:-
1. Information gain
2. Gini index

Advantages of Decision Tree

1.Easy to Understand
2.Useful in Data exploration
3.Less data cleaning required:
4.Data type is not a constraint
5.Non Parametric Method

Disadvantages

1. Over-fitting.
2. Not fit for continuous variables.
3.The decision tree contains lots of layers, which makes it complex.
4.For more class labels, the computational complexity of the decision tree may increase.

Random Forest

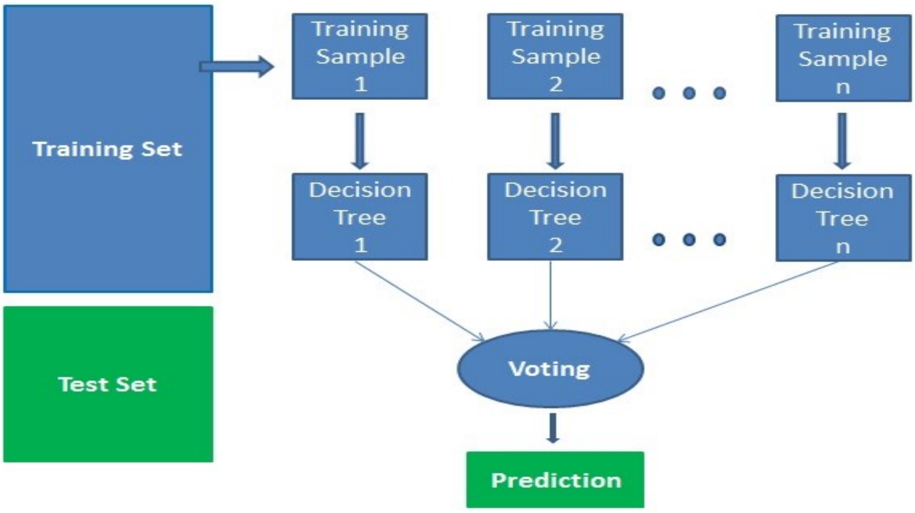
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest" is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes on predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over-fitting.

How does the algorithm work?

It works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.



Advantages

1. Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
2. It does not suffer from the over fitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
3. The algorithm can be used in both classification and regression problems.
4. Random forests can also handle missing values. There are two ways to handle these, using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
5. You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

Disadvantages

1. Random forests are slow in generating predictions because they have multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform vote on it. This whole process is time-consuming.
2. The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

Random Forests vs Decision Trees

1. Random forests is a set of multiple decision trees.
2. Deep decision trees may suffer from over-fitting, but random forests prevent over-fitting by creating trees on random subsets.
3. Decision trees are computationally faster.
4. Random forests are difficult to interpret, while a decision tree is easily

interpretable and can be converted to rules.