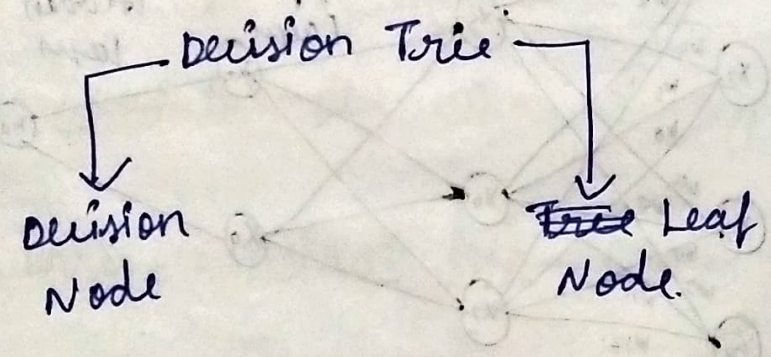
Decision Tree:

> It is a supervised learning technique that can be used for both classification and Regression problems.

> It is a tree-structure classifier where internal nodes represent the features of dataset, branch represent the decision rules and each leaf node represent the outcomes.



Decision Node } → It is used to make decision and have multiple branches.

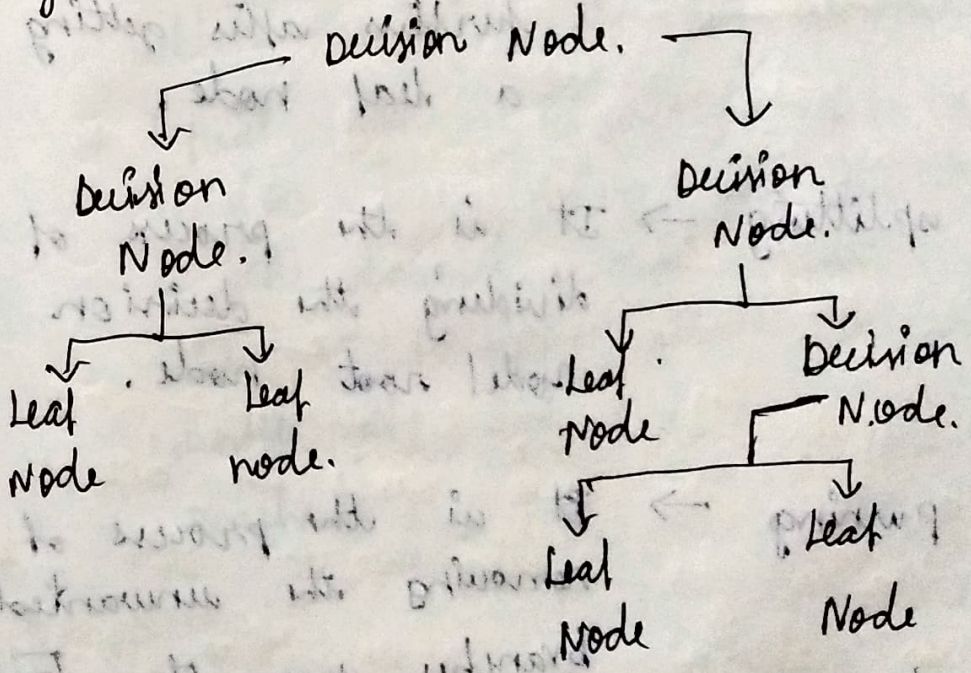
Leaf Node } → It gives the o/p of those decisions.

> It is a graphical representation for getting all the possible solutions to a problem / decision based on given condition.

> In order to build a tree, we use the 'CART Algorithem'.

> A decision tree simply asks a question, and based on the answer (yes/NO), it further split the tree into subtrees.

Diagram :



Terminologies:

Root Node → It is from where the decision tree starts.

→ It represents the entire data set.

→ Divided into Multiple branches.

Leaf Node → They are the final o/p node, and the tree cannot be segregated further after getting a leaf node.

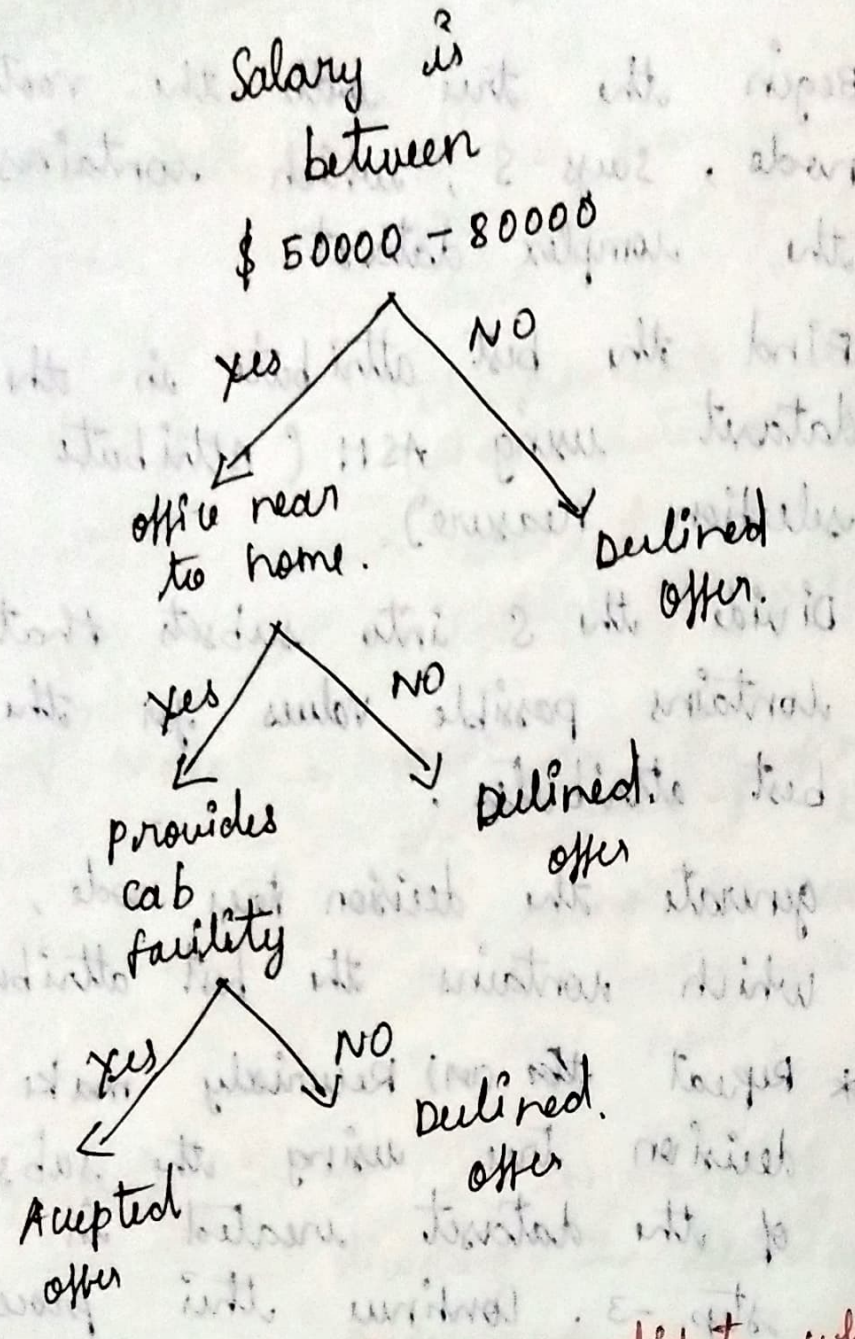
splitting → It is the process of dividing the decision node/ root node.

Pruning → It is the process of removing the unwanted branches from the tree.

Algorithm :

* Begin the tree with the root node, says S, which contains the complex dataset.
* Find the best attribute in the dataset using ASM (Attribute selection Measure).
* Divide the S into subsets that contains possible values for the best attributes.
* generate the decision tree node, which contains the best attribute
* Repeat (or) Recursively make new decision tree using the subset of the dataset created in step -3. Continue this process untill the leaf node reached.

Eg :



Q) 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 many nodes.

Attribute selection : Measures :

* Information Gain
* Gini Index.

Advantages :

* It is simple to understand.
* It can be very useful for solving decision-related problems.
* It helps to think about all the possible outcomes for a problem.
* It is less requirement of data cleaning.

Disadvantages :

* It may have overfitting issue.
* It contains lot of layers, which makes it complex.

4. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.

2.6 Random Forests

- Random forest is a famous system learning set of rules that belongs to the supervised getting to know method. It may be used for both classification and regression issues in ML. It is based totally on the concept of ensemble studying, that's a process of combining multiple classifiers to solve a complex problem and to enhance the overall performance of the model.
- As the call indicates, "Random forest is a classifier that incorporates some of choice timber on diverse 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 primarily based on most of the people's votes of predictions, and it predicts the very last output.
- The more wider variety of trees within the forest results in better accuracy and prevents the hassle of overfitting.

2.6.1 How Does Random Forest Algorithm Work ?

- Random forest works in two-section first is to create the random woodland by combining N selection trees and second is to make predictions for each tree created inside the first segment.
- The working technique may be explained within the below steps and diagram :

Step - 1 : Select random K statistics points from the schooling set.

Step - 2 : Build the selection trees associated with the selected information points (Subsets).

Step - 3 : Choose the wide variety N for selection trees which we want to build.

Step - 4 : Repeat step 1 and 2.

Step - 5 : For new factors, locate the predictions of each choice tree and assign the new records factors to the category that wins most people's votes.

- The working of the set of rules may be higher understood by the underneath example :
- Example : Suppose there may be a dataset that includes more than one fruit photo. So, this dataset is given to the random wooded area classifier. The dataset is divided into subsets and given to every decision tree. During the training section, each decision tree produces a prediction end result and while a brand new

statistics point occurs, then primarily based on the majority of consequences, the random forest classifier predicts the final decision. Consider the underneath picture :

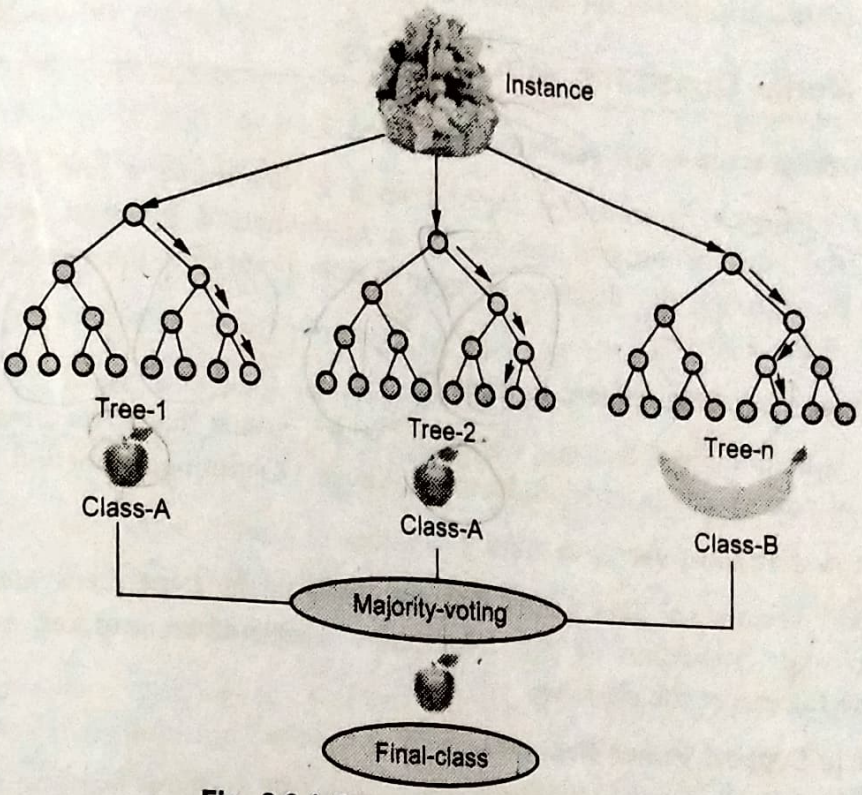


Fig. 2.6.1 Example of random forest

2.6.2 Applications of Random Forest

There are specifically 4 sectors where random forest normally used :

1. **Banking :** Banking zone in general uses this algorithm for the identification of loan danger.
2. **Medicine :** With the assistance of this set of rules, disorder traits and risks of the disorder may be recognized.
3. **Land use :** We can perceive the areas of comparable land use with the aid of this algorithm.
4. **Marketing :** Marketing tendencies can be recognized by the usage of this algorithm.

2.6.3 Advantages of Random Forest

Random forest is able to appearing both classification and regression responsibilities.

- It is capable of managing large datasets with high dimensionality.
- It enhances the accuracy of the version and forestalls the overfitting trouble.