**Important Questions**

**UNIT –1**

1. **Explain different types of machine learning systems**

**Machine learning**

**Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions.**

Types of Machine Learning

1. **Supervised Machine Learning**

In the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output.

**The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).** Some real-world applications of supervised learning are **Risk Assessment, Fraud Detection, Spam filtering,** etc.

Supervised machine learning can be classified into two types of problems, which are given below:

* **Classification**
* **Regression**

2.**Unsupervised Machine Learning**

 In unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

**The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.**Machines are instructed to find the hidden patterns from the input dataset.

Unsupervised Learning can be further classified into two types, which are given below:

* **Clustering**
* **Association**

## **3. Semi-Supervised Learning**

**Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning**. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

## **4. Reinforcement Learning**

**Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.** Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

**2)Describe in detail various features and Challenges in Machine Learning**

# Challenges

### 1**. Poor Quality of Data**

Data plays a significant role in the machine learning process. One of the significant issues that machine learning professionals face is the absence of good quality data. Unclean and noisy data can make the whole process extremely exhausting.

Therefore, we need to ensure that the process of data preprocessing which includes removing outliers, filtering missing values, and removing unwanted features, is done with the utmost level of perfection.

### 2. **Underfitting of Training Data**

This process occurs when data is unable to establish an accurate relationship between input and output variables. It simply means trying to fit in undersized jeans. It signifies the data is too simple to establish a precise relationship. To overcome this issue:

* *Maximize the training time*
* *Enhance the complexity of the model*

### 3. **Lack of Training Data**

The most important task you need to do in the machine learning process is to train the data to achieve an accurate output. . Less amount training data will produce inaccurate or too biased predictions.we need to ensure that Machine learning algorithms are trained with sufficient amounts of data.

### **4. Slow Implementation**

This is one of the common issues faced by machine learning professionals. The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time. Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results.

### 5**. Lack of skilled resources**

Although Machine Learning and Artificial Intelligence are continuously growing in the market, still these industries are fresher in comparison to others. The absence of skilled resources in the form of manpower is also an issue. Hence, we need manpower having in-depth knowledge of mathematics, science, and technologies for developing and managing scientific substances for machine learning.

### 1- THE ABILITY TO PERFORM AUTOMATED DATA VISUALIZATION

### Machine learning offers a number of tools that provide rich snippets of data which can be applied to both unstructured and structured data. With the help of user-friendly automated data visualization platforms in machine learning, businesses can obtain a wealth of new insights in an effort to increase productivity in their processes.

### 2- AUTOMATION AT ITS BEST

One of the biggest characteristics of machine learning is its ability to automate repetitive tasks and thus, increasing productivity. A huge number of organizations are already using machine learning-powered paperwork and email automation.

### 3- ACCURATE DATA ANALYSIS

Traditionally, data analysis has always been encompassing trial and error method, an approach which becomes impossible when we are working with large and heterogeneous datasets. Machine learning comes as the best solution to all these issues by offering effective alternatives to analyzing massive volumes of data. By developing efficient and fast algorithms, as well as, data-driven models for processing of data in real-time, machine learning is able to generate accurate analysis and results.

### 4- BUSINESS INTELLIGENCE AT ITS BEST

Machine learning characteristics, when merged with big data analytical work, can generate extreme levels of business intelligence with the help of which several different industries are making strategic initiatives. From retail to financial services to healthcare, and many more – machine learning has already become one of the most effective technologies to boost business operations.

### 5- CUSTOMER ENGAGEMENT LIKE NEVER BEFORE

For any business, one of the most crucial ways to drive engagement, promote brand loyalty and establish long-lasting customer relationships is by triggering meaningful conversations with its target customer base. Machine learning plays a critical role in enabling businesses and brands to spark more valuable conversations in terms of customer engagement

**3)Discuss any four examples of machine learning applications**

### **1. Image Recognition:**

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's **face detection** and **recognition algorithm**.

### It is based on the Facebook project named "**Deep Face**," which is responsible for face recognition and person identification in the picture.

### 2. Speech Recognition

While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, machine learning algorithms are widely used by various applications of speech recognition. **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

* **Real Time location** of the vehicle form Google Map app and sensors
* **Average time has taken** on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

4. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

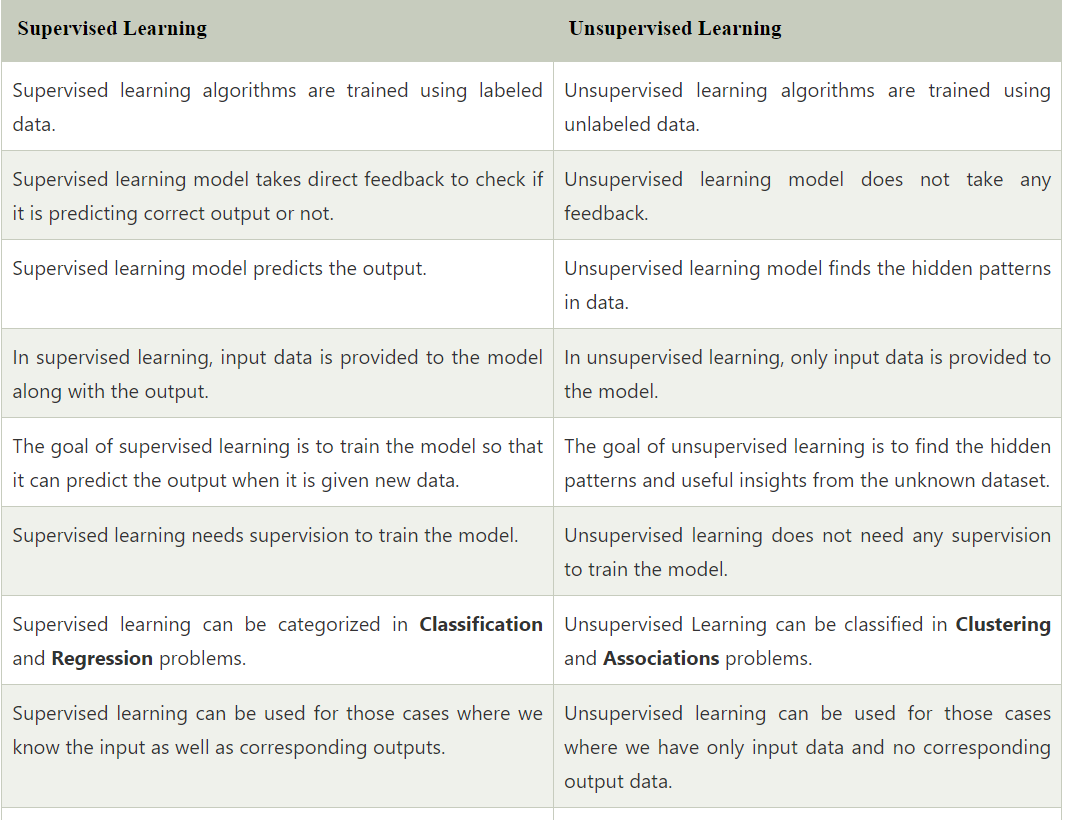
Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

### **5.Self-driving cars:**

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

**4.Compare Supervised and unsupervised learning**

****

**5. Explain Training and Test Loss Trade-offs in Statistical Learning**

In statistical learning, training and test loss trade-offs refer to the relationship between the performance of a machine learning model on the data it was trained on (training loss) and its performance on new, unseen data (test loss). The goal of a machine learning model is to generalize well to new data, which means that its performance on the test data should be similar to its performance on the training data.

**Training loss** is the error that the model makes on the training data during the training process. The training process involves adjusting the model's parameters to minimize the training loss. The model is evaluated on the training data after each iteration or epoch to monitor its performance.

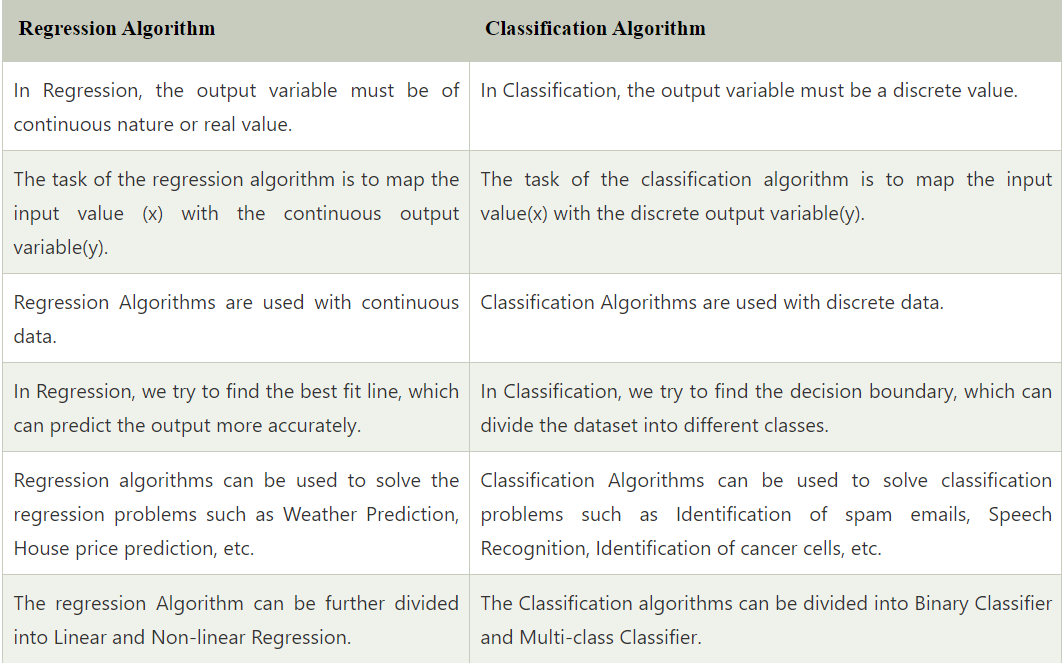
**Test loss** is the error that the model makes on new, unseen data that was not used during the training process. The test data is used to evaluate the model's generalization performance, which is its ability to perform well on new data.

**The trade-off** between training and test loss arises because optimizing the model for training loss alone can lead to overfitting. Overfitting occurs when the model becomes too complex and fits the noise in the training data, resulting in poor generalization performance on the test data. To avoid overfitting, the model must balance the training and test loss.

One way to achieve this balance is through regularization techniques, such as L1 and L2 regularization, which penalize large weights and reduce the complexity of the model. Another way is to use early stopping, which stops the training process when the performance on the validation data starts to degrade.

**UNIT –2**

1. **Compare Classification with regression with an example**



**Classification:**

**Example:** The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

**Regression:**

**Example:** Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

**2.Illustrate K-Nearest Neighbour classification Algorithm with Example**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

* K-NN is a **non-parametric algorithm**,
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

**ALGORITHM:**

**Step 1** − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

**Step 2** − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

**Step 3** − For each point in the test data do the following −

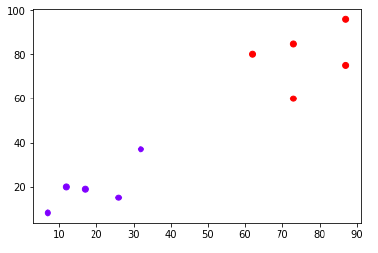
* **3.1** − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
* **3.2** − Now, based on the distance value, sort them in ascending order.
* **3.3** − Next, it will choose the top K rows from the sorted array.
* **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows.

**Step 4** − End

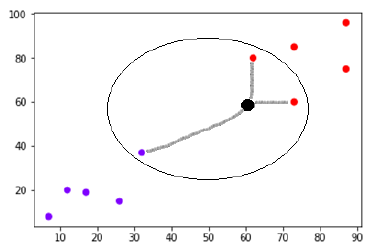
### **Example**

The following is an example to understand the concept of K and working of KNN algorithm −

Suppose we have a dataset which can be plotted as follows −



Now, we need to classify new data point with black dot (at point 60,60) into blue or red class. We are assuming K = 3 i.e. it would find three nearest data points. It is shown in the next diagram −



We can see in the above diagram the three nearest neighbors of the data point with black dot. Among those three, two of them lies in Red class hence the black dot will also be assigned in red class.

**3.Apply Decision Tree induction to classify the data**

To apply decision tree induction to classify data, we need a dataset with labeled examples and a set of features to use for classification. Decision tree induction is a supervised learning algorithm, which means it requires a labeled dataset for training.

Here's an example of how to apply decision tree induction to classify data:

Suppose we have a dataset of patients with a certain disease, and we want to build a decision tree to predict whether a new patient will have the disease based on their symptoms.

Our dataset has the following features:

• Age: the age of the patient (numeric)

• Gender: the gender of the patient (binary: male/female)

• Fever: whether the patient has a fever or not (binary: yes/no)

• Cough: whether the patient has a cough or not (binary: yes/no)

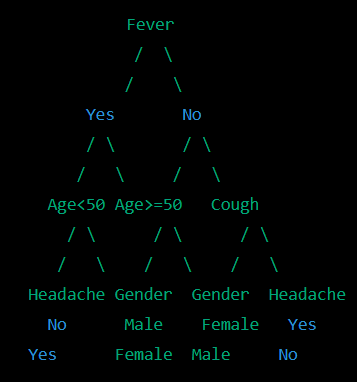
• Headache: whether the patient has a headache or not (binary: yes/no)

The dataset also includes a label indicating whether each patient has the disease or not (binary: yes/no).

To apply decision tree induction to this dataset, we first split the data into training and test sets. We randomly select 70% of the data for training and 30% for testing.

Next, we build a decision tree using the training data. The decision tree induction algorithm works by recursively partitioning the data based on the most informative feature at each step. There are several algorithms for decision tree induction, but a common one is the ID3 algorithm.

After building the decision tree, we evaluate its performance on the test data. We can compute various performance metrics, such as accuracy, precision, recall, and F1 score, to assess how well the decision tree performs on the test data.

For example, let's say our decision tree has the following structure: 

This decision tree partitions the data based on fever, age, cough, gender, and headache. We can use this decision tree to predict whether a new patient will have the disease based on their symptoms.

For example, suppose we have a new patient who is a 40-year-old male with fever, cough, and headache. We can follow the decision tree to predict whether he has the disease:

• Fever: Yes -> Age < 50 -> Headache: Yes -> No disease

Therefore, we predict that this patient does not have the disease.

That's an example of how to apply decision tree induction to classify data. Decision tree induction is a powerful and interpretable algorithm for classification and can be used in many machine learning applications. However, it can be sensitive to noisy and irrelevant features, and can suffer from overfitting if the tree is too complex.

**4.Explain Support Vector Machine Algorithm with an example.**

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

**Example:** SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat. Consider the below diagram:

69.2M

1.4K

Features of Java - Javatpoint

**Next**

**Stay**



SVM algorithm can be used for **Face detection, image classification, text categorization,** etc.

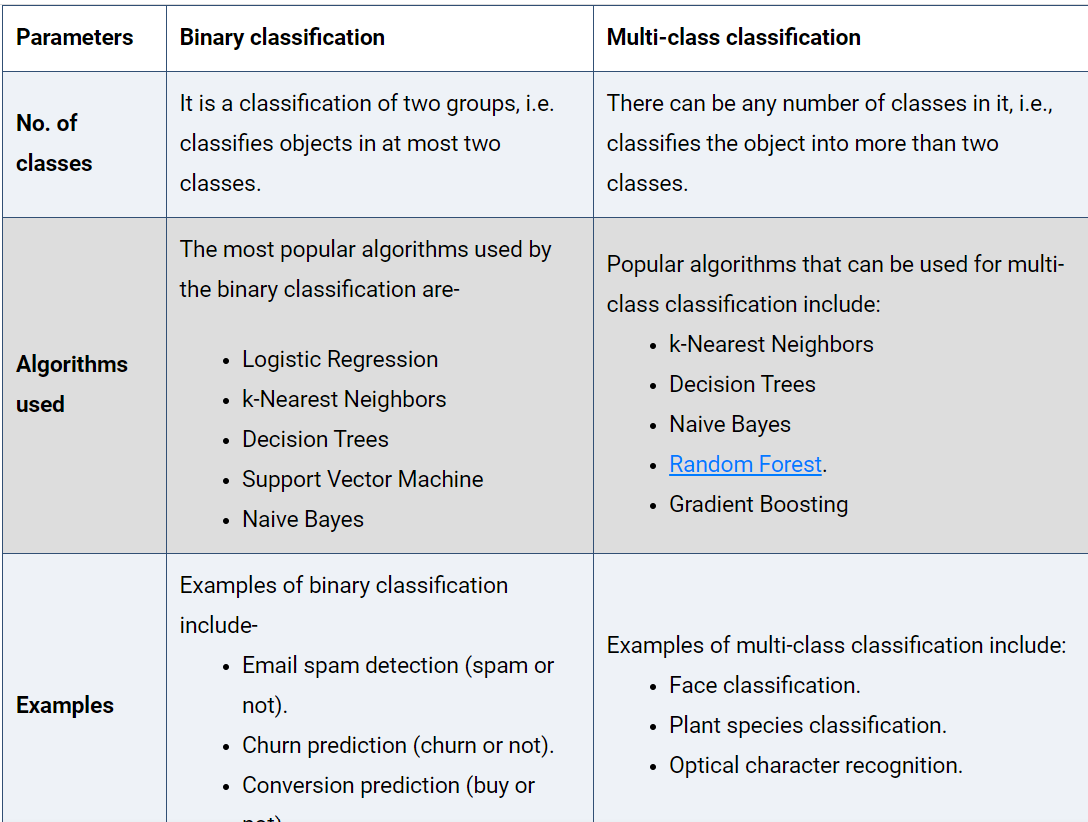
**5)Examine the Binary Classification & Multi Class Classification**

## **Binary Classification**

It is a process or task of classification, in which a given data is being classified into two classes.  It’s basically a kind of prediction about which of two groups the thing belongs to.

## **Multiclass Classification**

Multi-class classification is the task of classifying elements into different classes. Unlike binary, it doesn’t restrict itself to any number of classes.

****

**UNIT –3**

1. **What is Ensemble modeling? Discuss about Bagging and Boosting.**

Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different [modeling algorithms](https://www.sciencedirect.com/topics/computer-science/modeling-algorithm" \o "Learn more about modeling algorithms from ScienceDirect's AI-generated Topic Pages) or using different training data sets.

| **S.NO** | **Bagging** | **Boosting** |
| --- | --- | --- |
| **1.** | **The simplest way of combining predictions that  belong to the same type.** | **A way of combining predictions that  belong to the different types.** |
| **2.** | **Aim to decrease variance, not bias.** | **Aim to decrease bias, not variance.** |
| **3.** | **Each model receives equal weight.** | **Models are weighted according to their performance.** |
| **4.** | **Each model is built independently.** | **New models are influenced  by the performance of previously built models.** |
| **5.** | **Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.** | **Every new subset contains the elements that were misclassified by previous models.** |
| **6.** | **Bagging tries to solve the over-fitting problem.** | **Boosting tries to reduce bias.** |
| **7.** | **If the classifier is unstable (high variance), then apply bagging.** | **If the classifier is stable and simple (high bias) the apply boosting.** |
| **8.** | **In this base classifiers are trained parallelly.** | **In this base classifiers are trained sequentially.** |
| **9** | **Example: The Random forest model uses Bagging.** | **Example: The AdaBoost uses Boosting techniques** |

1. **Why do we need the random forest algorithm? Explain its advantages and disadvantages**.

# Random Forest Algorithm

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.

Below are some points that explain why we should use the Random Forest algorithm:

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.
* The Working process can be explained in the below steps and diagram:

## **How does Random Forest algorithm work?**

* **Step-1:** Select random K data points from the training set.
* **Step-2:** Build the decision trees associated with the selected data points (Subsets).
* **Step-3:** Choose the number N for decision trees that you want to build.
* **Step-4:** Repeat Step 1 & 2.

**Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

## **Advantages of Random Forest**

* Random Forest is capable of performing both Classification and Regression tasks.
* It is capable of handling large datasets with high dimensionality.
* It enhances the accuracy of the model and prevents the overfitting issue.

## **Disadvantages of Random Forest**

* Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

**3)Describe about Voting Classifiers**

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

Voting Classifier supports two types of votings.

1. **Hard Voting:**In hard voting, the predicted output class is a class with the highest majority of votes i.e the class which had the highest probability of being predicted by each of the classifiers. Suppose three classifiers predicted the *output class(A, A, B)*, so here the majority predicted *A* as output. Hence *A* will be the final prediction.
2. **Soft Voting:**In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class *A = (0.30, 0.47, 0.53)* and *B = (0.20, 0.32, 0.40)*. So the average for class *A is 0.4333* and *B is 0.3067*, the winner is clearly class *A* because it had the highest probability averaged by each classifier.

The advantage of voting classifiers is that they can improve the accuracy and robustness of the model by combining the predictions of multiple classifiers. This can help to reduce overfitting and improve the generalization of the model.