Machine Learning

Attribute Selection Measures

- **Information Gain -** Changes in entropy after the segmentation of a dataset based on an attribute

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

Entropy: It measure the impurity in a given attribute

$$Entropy(s) = -P(yes)\log 2\ P(yes) - P(no)\ \log 2\ P(no)$$

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

- **Gini Index -** Measure of impurity or purity used while creating a decision tree

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

Pruning -

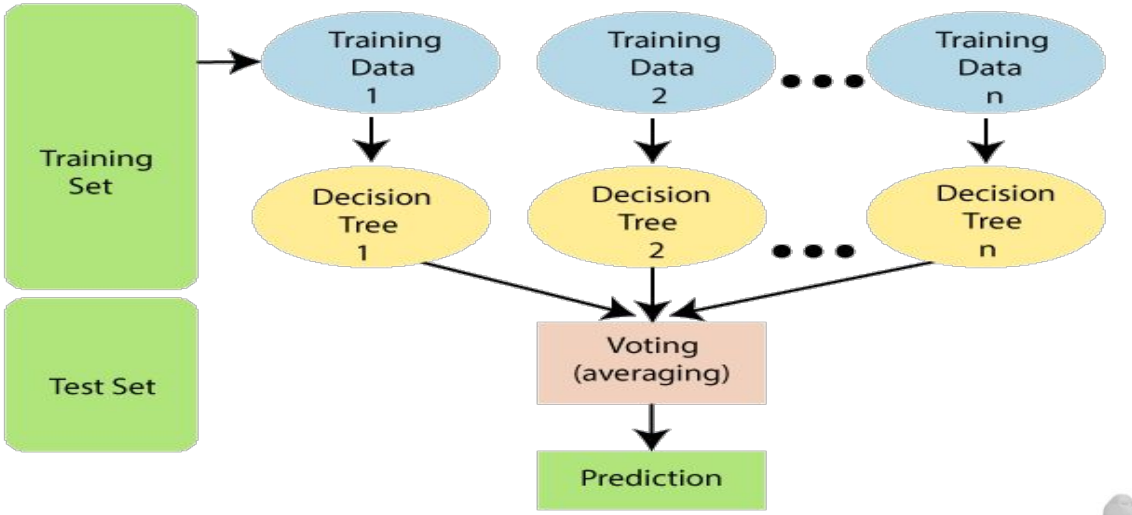
A process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

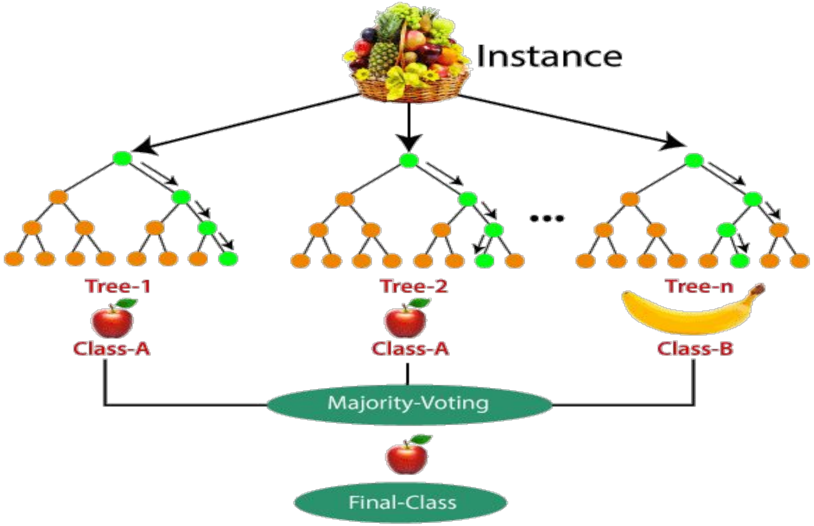
- **Cost Complexity Pruning**

- **Reduced Error Pruning**.

Random Forest -

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.

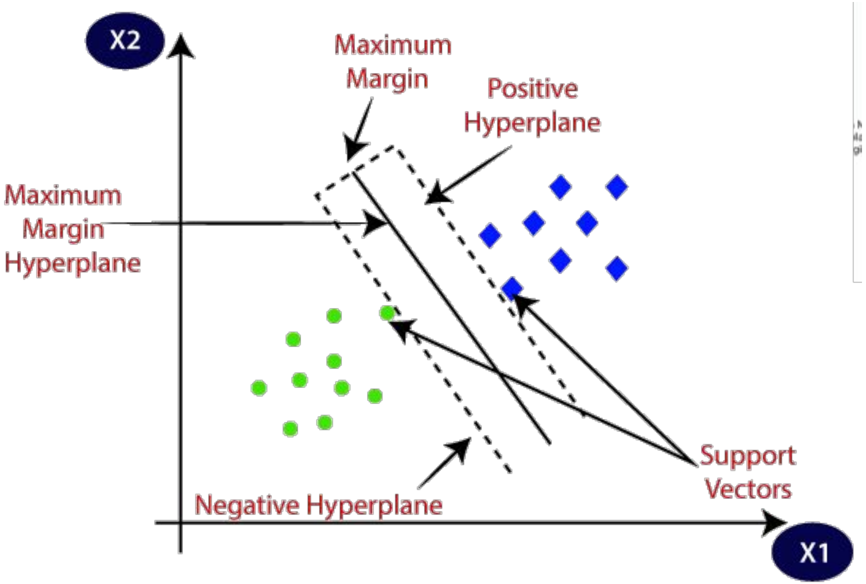




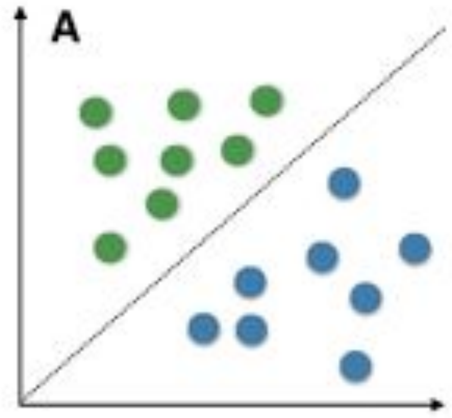
$$g(x)= f_0(x)+ f_1(x)+ f_2(x)+....$$

Random forest uses **Bagging or Bootstrap Aggregation** technique of ensemble learning in which aggregated decision tree runs in parallel and do not interact with each other.

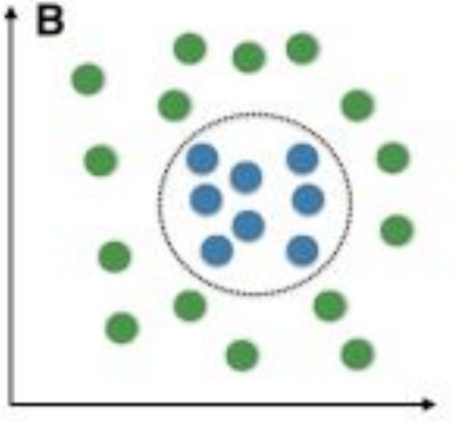
Support Vector Machine -



- Kernel
- Hyperplane
- Boundary line
- Support vectors



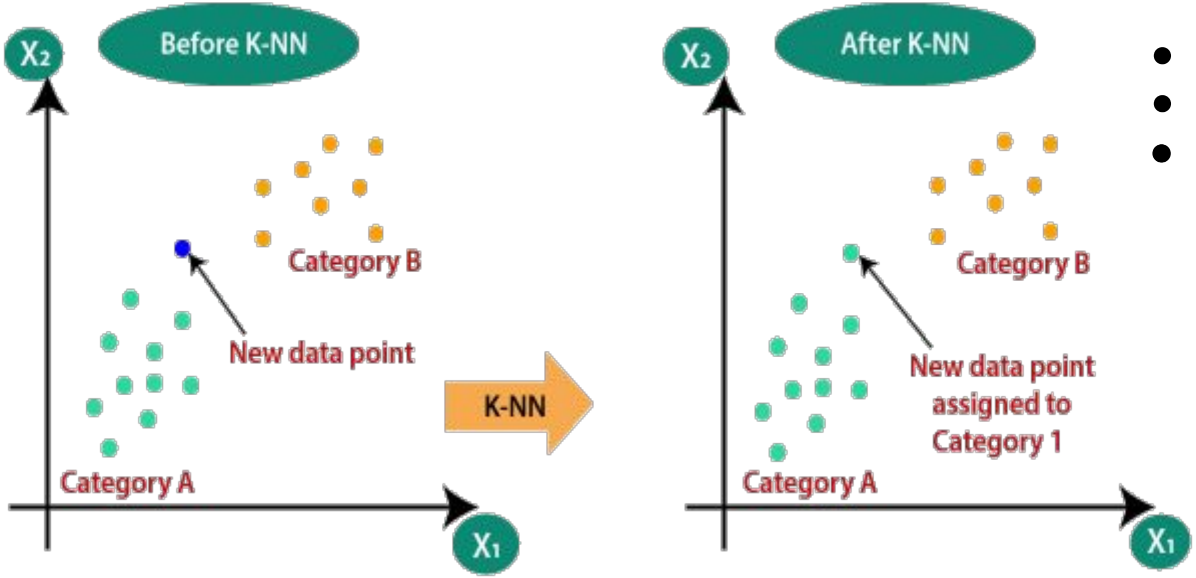
Linear SVM



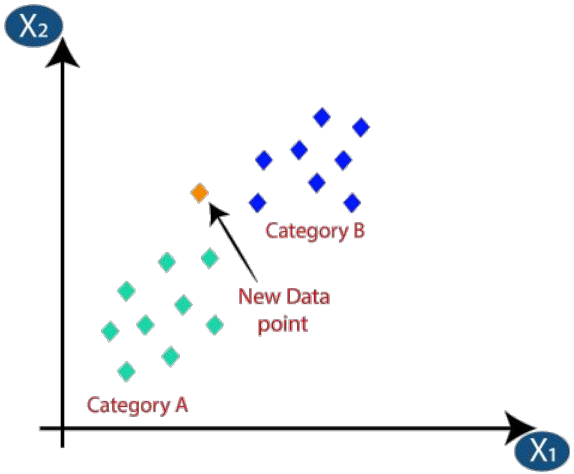
Non-Linear SVM

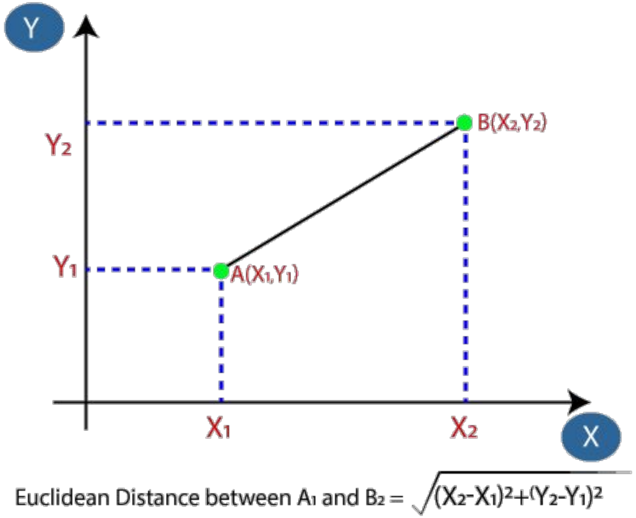
The main goal of SVR is to consider the maximum data points within the boundary lines and the hyperplane (best-fit line) must contain a maximum number of datapoints

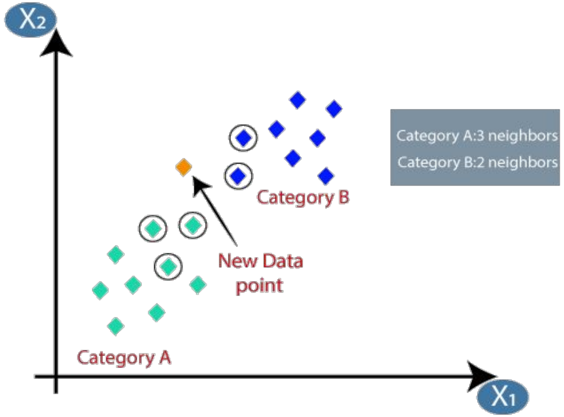
K-Nearest Neighbor(KNN) -



- lazy learner algorithm
- non-parametric algorithm
- Based on the similarity







K is the number of nearest neighbors

Here k=5, as new datapoint is having 3 nearest neighbors from category A so it belongs to Category A

Naïve Bayes -

- Based on **Bayes theorem**
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where, P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B, **P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true, **P(A) is Prior Probability**: Probability of hypothesis before observing the evidence, **P(B) is Marginal Probability**: Probability of Evidence.

Steps -

1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posterior probability.

Types -

- **Gaussian**
- **Multinomial**
- **Bernoulli**