Regression:

Linear- (one variable input, one variable output)
Multiple-(multiple variable input , one variable output)

Evaluation:

- Mean_absolute_error
- Mean_squared_error
- Root_mean_squared_error

Aplications:

- Weather prediction
- House price prediction

Classification:

Evaluation parameters:

- Recall/TPR/Sensitivity=TP/(TP+FN)
- Specificity=TN/(TN+FP)
- FPR=FP/(TN+FP)=1-Specificity
- Confusion matrix
- Precision=TP/(TP+FP)
- Accuracy- no of samples correctly classified=(TP+TN)/(TP+TN+FP+FN)
- Classification report

Applications:

- Medical disease diagnosis
- Covid positive or negative based on symptoms

AUC ROC curve:

(Area under the curve) and (Receiver operating characteristics)
If 1-> clearly able to distinguish
If 0.7 -> (existence of FP,FN)
If 0.5-> Not able to distinguish
If 0-> It is predicting opposite labels

SVM:

- Support vectors- These are the points of one class that are nearer to the points of other class and thus help in deciding the margin between classes
- Can be used for both classification and regression.
- Mostly used for binary classification

Naive bayes:

Advantages:

- Easy to determine the class using the mathematical formula given the frequency distribution of various classes as per attributes

Disadvantages:

- Class conditional Independence

Applications:

- Spam/ham same
- Text classification
- Sentimental Analysis (Ex: positive/negative feedback)
- Recommender systems

KNN:

- Lazy learner
- Even if k is small the distance between all the points if the dataset and the given point should be calculated the sorted to find the k nearest neighbours
- Consumes lot of time during testing phase, training phase just stores data

Apriori:

Quantification is done by

a) Support
b) Confidence
c) Lift

How to improve apriori

FP growth:(method)

K Means:

- Mutually exclusive clusters
- Measure used is mean(sy)or median for cluster centroids
- Voronoi cells
- Iterative relocation technique : The data points are assigned to clusters iteratively to improve the quality of clusters and minimize squared error within the cluster and maximizes the separation between clusters
- Expectation-Maximization
- E step is to assign data point to cluster
- M step is to calculate the centroid

- Normalise it to stop one attribute from overweighting other(units are different)

Disadvantages:

- No of clusters should be known in advance(else trial and error)
- Sensitive to noise(different clusters for pure dataset and the one with noise)
- Different initialization values ->different clusters(because it settles in local optimum rather than global optimum)
- Spherical shape
- Gives focus to bigger clusters to minimize the variation

Applications:

- Market / customer segmentation
- Handwriting recognition
- Document clustering
- Image recognition

Evaluation Method:

a)Elbow method

- Choosing right k using SSE(sum of squared errors) where it forms an elbow and flattens out

b)Silhouette Analysis

- Degree of separation between clusters

  a=intra cluster distance
  b=inter cluster distance

  Coefficient =(b-a)/max(a,b)
  If 0, diff clusters are near to each other
  If 1,well apart and clearly distinguishable
  If -1,overlapped