Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

Group-B: Assignment No. 1

Title: Predict the price of the Uber ride from a given pickup point to the agreed drop-off location.

Problem Statement: Predict the price of the Uber ride from a given pickup point to the agreed drop-off location.

Perform the following tasks:

1. Pre-process the dataset.
2. Identify outliers.
3. Check the correlation.
4. Implement linear regression and random forest regression models.
5. Evaluate the models and compare their respective scores like R2, RMSE, etc.

Objective:

1. Pre-process the dataset
2. Identify outliers
3. Check the correlation
4. Implement linear regression and random forest regression models
5. Predict the price of the Uber ride from a given pickup point to the agreed drop-off location

Theory:

- **What Is Data Preprocessing and Why Do We Need It?**

  For machine learning algorithms to work, it's necessary to convert raw data into a clean data set, which means we must convert the data set to numeric data.

- **How to preprocess data in Python step-by-step**

Step 1: Load data in Pandas.
Step 2: Drop columns that aren't useful.
Step 3: Drop rows with missing values.
Step 4: Create dummy variables.
Step 5: Take care of missing data.
Step 6: Convert the data frame to NumPy.
Step 7: Divide the data set into training data and test data.

- **Identify outliers**

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

An Outlier is a data item/object that deviates significantly from the rest. They can be caused by measurement or execution errors.

- **Detecting the outliers**

Outliers can be detected using visualization, implementing mathematical formulas on the dataset, or using the statistical approach.

- **What is Correlation?**

- **Positive Correlation**: Both variables change in the same direction.
- **Neutral Correlation**: No relationship in the change of the variables.
- **Negative Correlation**: Variables change in opposite directions.

- **Covariance**

Variables can be related by a linear relationship. This relationship can be summarized between two variables, called the covariance.

- **What is Regression?**

Regression analysis is a statistical method that helps us to understand the relationship between dependent and one or more independent variables**.**

- **Dependent Variable:** This is the Main Factor that we are trying to predict.
- **Independent Variable:** These are the variables that have a relationship with the dependent variable.

- **Types of Regression Analysis**

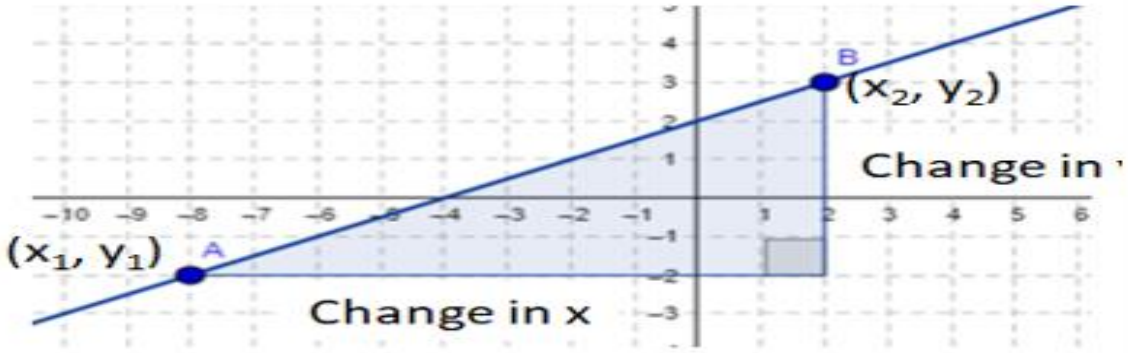
1. **Simple Linear Regression**

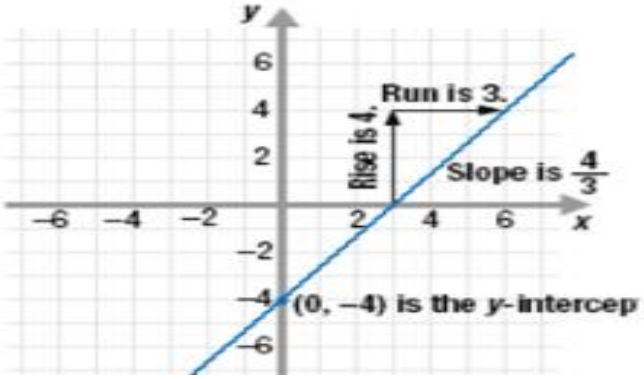
Simple Linear Regression uses the slope-intercept form, where our model needs to find the optimal value for both slope and intercept.

There are many equations to represent a straight line, we will stick with the common equation.

$$y = b_o + b_1 x$$

$$\text{slope} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\text{Change in y}}{\text{Change in x}}$$





In machine learning, every algorithm has a cost function, and in simple linear regression, the goal of our algorithm is to find a minimal value for the cost function. In linear regression (LR), we have many cost functions, but the most used cost function is MSE(Mean Squared Error). It is also known as a Least Squared Method.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

2. Multiple Linear Regression

In multiple linear regression instead of having a single independent variable, the model has multiple independent variables to predict the dependent variable.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ldots + b_nx_n$$

here bo is the y-intercept

b1,b2,b3,b4…,bn are slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - y_j|$$

The final step is to evaluate the performance of the algorithm. For regression algorithms, three evaluation metrics are commonly used:

1. **Mean Absolute Error** (MAE) is the mean of the absolute value of the errors. It is calculated as:
2. **Mean Squared Error** (MSE) is the mean of the squared errors and is calculated as:

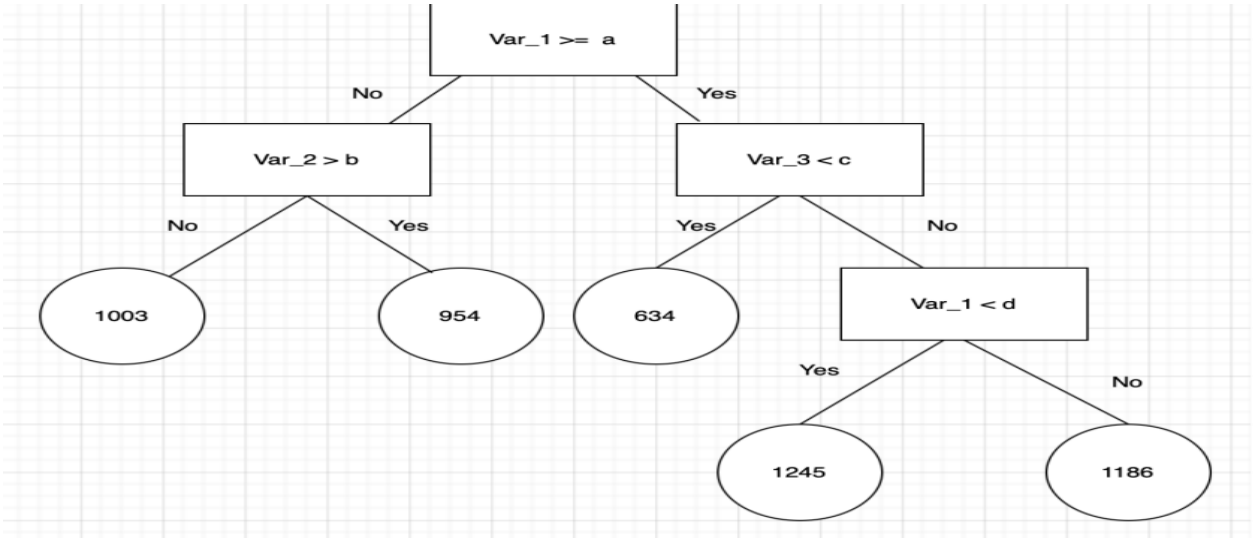
$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

3. **Root Mean Squared Error** (RMSE) is the square root of the mean of the squared errors:

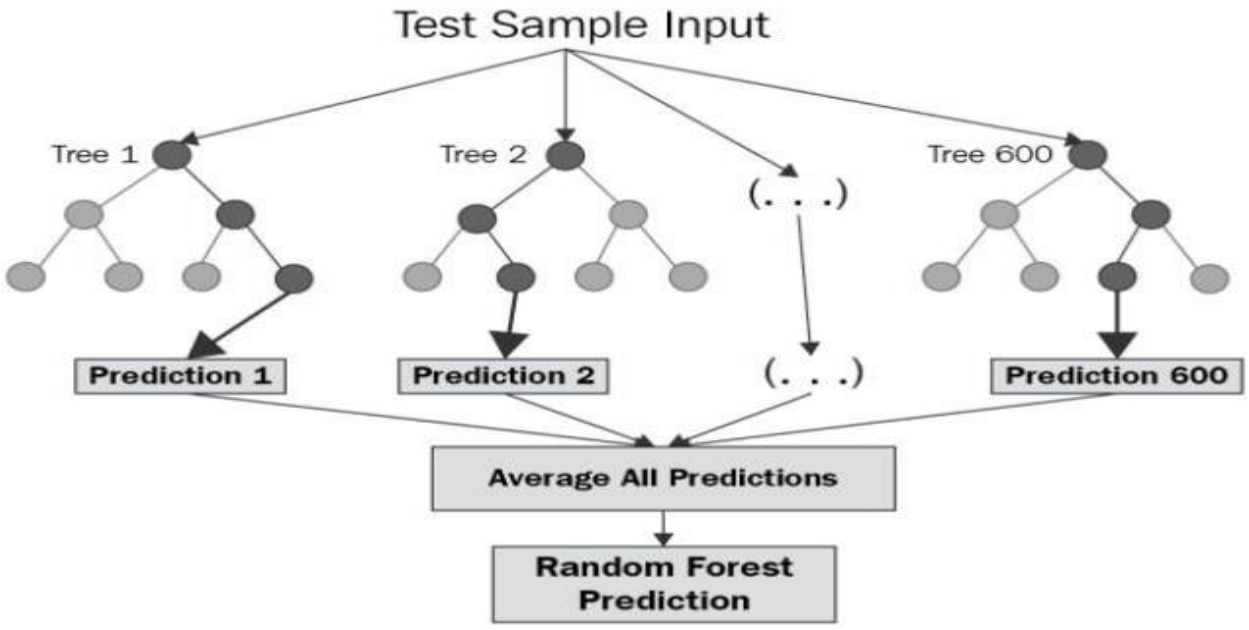
- **Random Forest Regression Models**

  Decision Trees are used for both regression and classification problems. They start with the root of the tree and follow splits based on variable outcomes until a leaf node is reached and the result is given. An example of a decision tree is below.

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)



- **Ensemble learning** is the process of using multiple models, trained over the same data, averaging the results of each model ultimately finding a more powerful predictive/classification result.
- **Bootstrapping** is the process of randomly sampling subsets of a dataset over a given number of iterations and a given number of variables.
- **Random Forest Regression** is a supervised learning algorithm that uses an ensemble learning method for regression.



To get a better understanding of the Random Forest algorithm, let's walk through

the steps:
Step 1: Pick at random k data points from the training set.
Step 2: Build a decision tree associated with these k data points.
Step 3: Choose the number N of trees you want to build and repeat steps 1 and 2.
Step 4: For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

Conclusion:

We studied how to Predict the price of the Uber ride from a given pickup point to the agreed drop-off location using linear regression and random forest regression models. Also Evaluated the models and compared their respective scores like R2, and RMSE.

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

Group-B: Assignment No. 2

Title: Implementation of K-Means Clustering/Hierarchical Clustering on Sales Data

Problem Statement: Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method.

Objective:

To understand and implement:

- K-Means clustering
- Determine the number of clusters using the elbow method

Theory:

- **What is the K-Means Algorithm?**

  K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

  It is an iterative algorithm that divides the unlabeled dataset into k different clusters

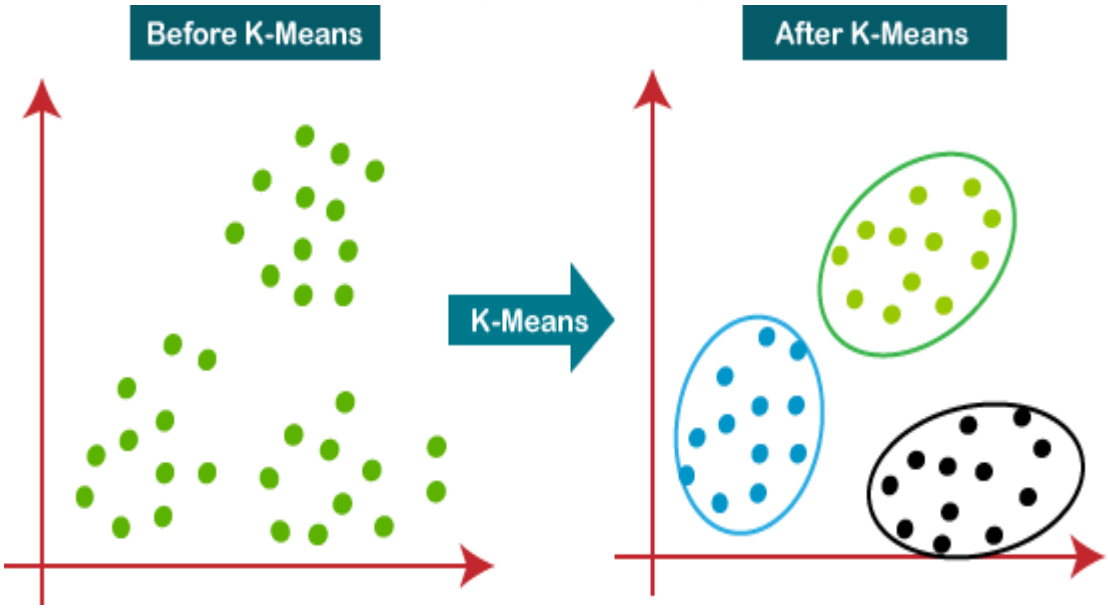
  The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K centre points by an iterative process.
- Assign each data point to its closest K-center.

  The below diagram explains the working of the K-means Clustering Algorithm:

- **Algorithms:**

  The working of the K-Means algorithm is explained in the below steps:



Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

Step 1: Select the number K to decide the number of clusters.
Step 2: Select random K points or centroids.
Step 3: Assign each data point to its closest centroid, which will form the predefined K clusters.
Step 4: Calculate the variance and place a new centroid of each cluster.
Step-5: Repeat the third step, which means reassigning each datapoint to the new closest centroid of each cluster.
Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
Step-7: The model is ready.

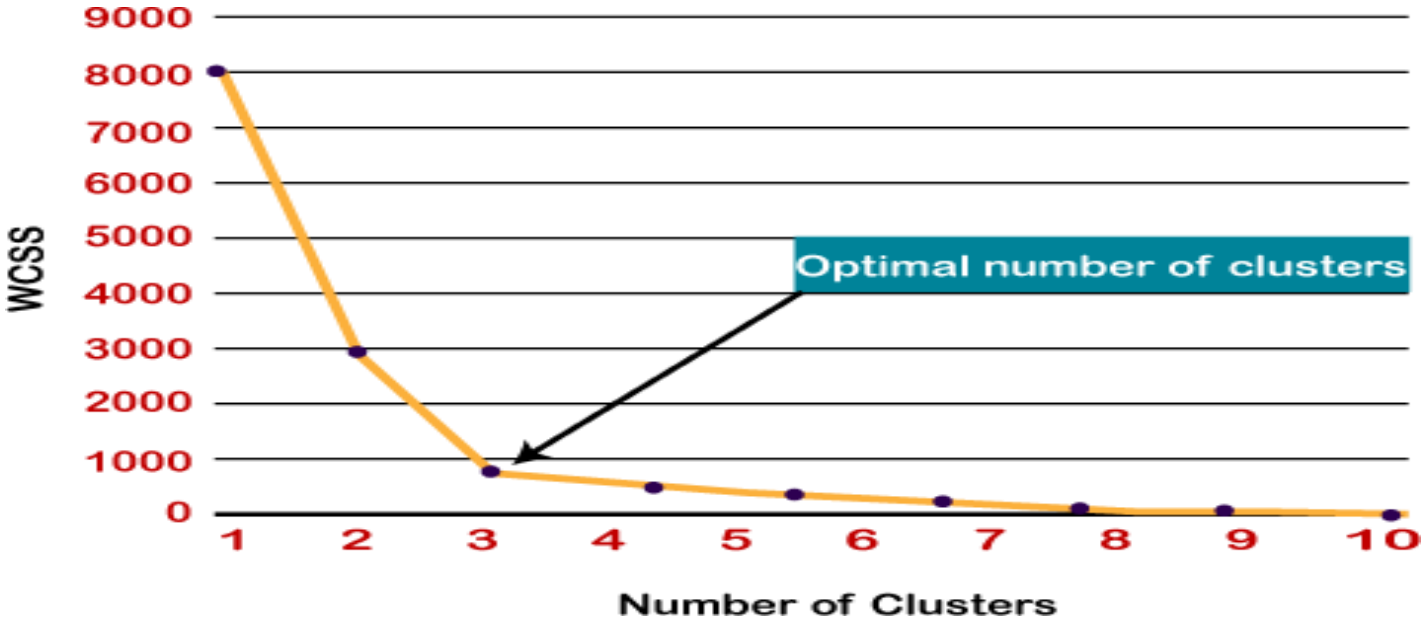
- **How to choose the value of "K number of clusters" in K-means Clustering?**
  The performance of the K-means clustering algorithm depends upon the highly efficient clusters that it forms. However, choosing the optimal number of clusters is a big task.
  The method is given below:
  - **Elbow Method:**

  The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value.
  WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster.
  The formula is given below:
  WCSS= ∑Pi in Cluster1 distance (Pi C1)^2 +∑Pi in Cluster2distance (Pi C2)^2+∑Pi in CLuster3 distance (Pi C3)^2

  To find the optimal value of clusters, the elbow method follows the below steps:
  Step 1: It executes the K-means clustering on a given dataset for different K values.
  Step 2: For each value of K, calculate the WCSS value.
  Step 3: Plot a curve between calculated WCSS values and the number of clusters K.
  Step 4: The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

The graph for the elbow method looks like the below image:



- **Applications of K-means Clustering**

1. K-means clustering is applied in the Call Detail Record (CDR) Analysis.
2. It is used in the clustering of documents to identify compatible documents in the same place.
3. It is deployed to classify the sounds based on their identical patterns and segregate malformation in them.

- **K-means vs Hierarchical Clustering**

1. K-means clustering produces a specific number of clusters for the disarranged and flat dataset, whereas Hierarchical clustering builds a hierarchy of clusters.
2. K-means are highly sensitive to noise in the dataset and perform better than Hierarchical clustering where it is less sensitive to noise in a dataset.
3. K-means are good for a large dataset and Hierarchical clustering is good for small datasets.

Conclusion:

Thus, we have understood what Clustering is, K-mean clustering. We also Determined the number of clusters using the elbow method and Implemented K-mean clustering.

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

Group-B: Assignment No. 3

Title: Implementation of K-Nearest Neighbors Algorithm

Problem Statement:

Implement the K-Nearest Neighbors algorithm on the diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.

Objective:

To understand and implement:

- K-Nearest Neighbors.
- Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.

Theory:

- **K-Nearest Neighbour (KNN) Algorithm for Machine Learning**

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning techniques.
- The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most like the available categories.
- The K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems.

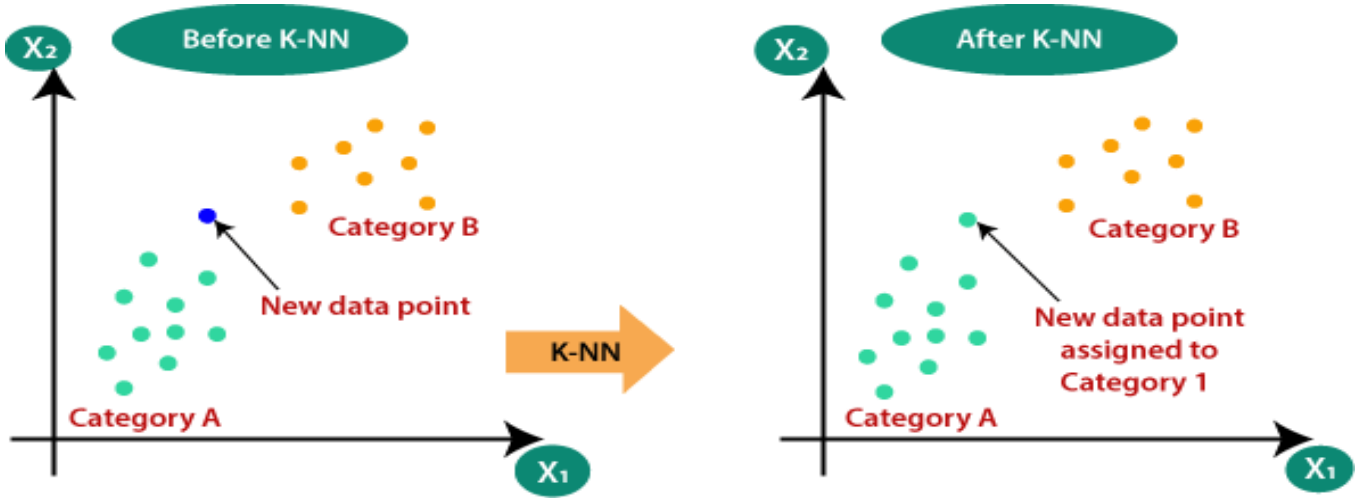
- **Why do we need a K-NN Algorithm**?

  Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories?

  To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

  Consider the below diagram:

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)



- **How does K-NN work?**

  Step 1: Select the number K of the neighbours
  Step 2: Calculate the Euclidean distance of K number of neighbours
  Step 3: Take the K nearest neighbours as per the calculated Euclidean distance.
  Step 4: Among these k neighbours, count the number of the data points in each category.
  Step 5: Assign the new data points to that category for which the number of neighbours is maximum.
  Step 6: Our model is ready.

- **Advantages of KNN Algorithm:**

- It is simple to implement.
- It is robust to noisy training data

- **Disadvantages of KNN Algorithm:**

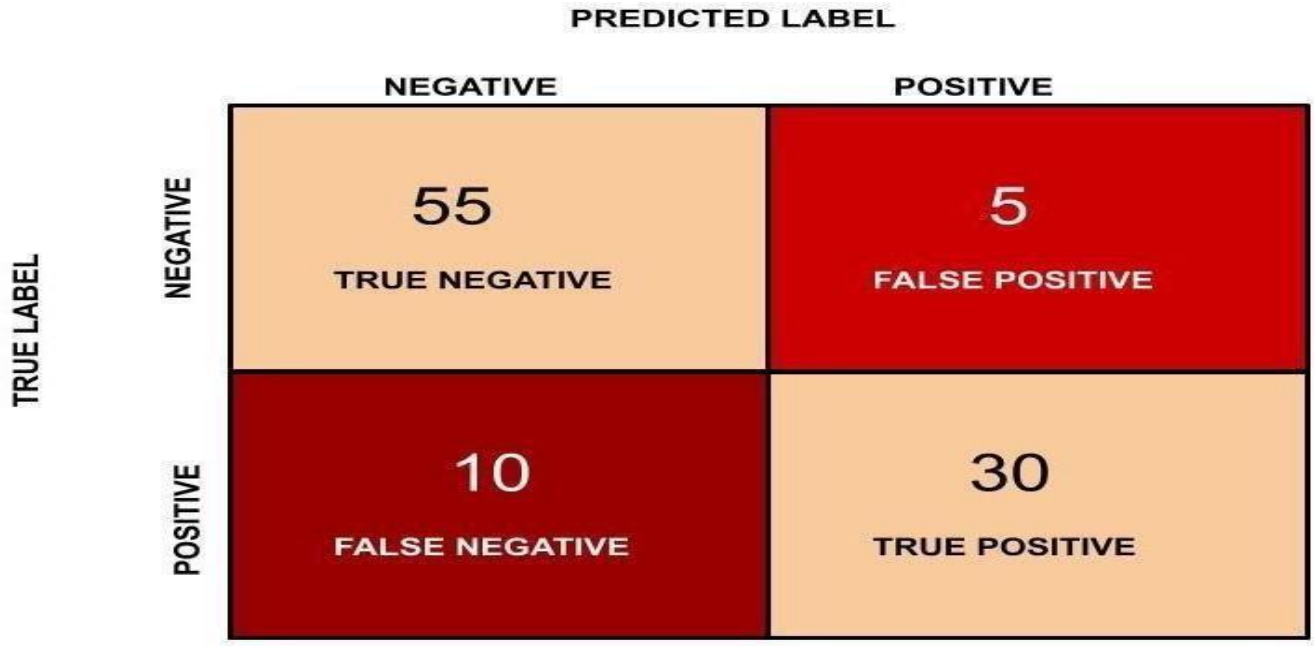
- Always needs to determine the value of K which may be complex sometimes.
- The computation cost is high.

- **Steps to implement the K-NN algorithm:**

1. Data Pre-processing step
2. Fitting the K-NN algorithm to the Training set
3. Predicting the test result
4. Test the accuracy of the result (Creation of Confusion matrix)
5. Visualizing the test set result.

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

- **What is Confusion Matrix and why do you need it?**
  The Confusion Matrix is a performance measurement for machine learning classification problems where the output can be two or more classes. It is a table with four different combinations of predicted and actual values.
  The confusion matrix is as follows.



- **Let's understand TP, FP, FN, TN.**

1. **True Positive:**
   Interpretation: You predicted positive and it's true. You predicted that a Man is a terrorist, and he is.
2. **True Negative:**
   Interpretation: You predicted negative and it's true. You predicted that a man is not a terrorist, and he is not.
3. **False Positive: (Type 1 Error)**
   Interpretation: You predicted positive and it's false. You predicted that a man is a terrorist, but he is not.
4. **False Negative: (Type 2 Error)**
   Interpretation: You predicted negative and it's false. You predicted that a man is not a terrorist, but he is.

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

- **Accuracy:**
  Accuracy represents the number of correctly classified data instances over the total number of data instances.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

- **Precision**
  For a good classifier, precision should ideally be 1 (high). Precision becomes 1 only when the numerator and denominator are equal.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**
  The recall is also known as sensitivity or true positive rate and is defined as follows

$$Recall = \frac{TP}{TP + FN}$$

Conclusion:
Thus, we implemented the K-Nearest Neighbors algorithm on a given data set and to evaluate the performance we Computed the confusion matrix, accuracy, error rate, precision and recall on the given dataset.

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

Group-B: Assignment No. 4

Title: Build a Neural Network-Based Classifier for Customer Churn

Problem Statement:

Given a bank customer, build a neural network-based classifier that can determine whether they will leave or not in the next 6 months.

Dataset Description: The case study is from an open-source dataset from Kaggle. The dataset contains 10,000 sample points with 14 distinct features such as Customer ID, Credit Score, Geography, Gender, Age, Tenure, Balance, etc.

Perform the following steps:

1. Read the dataset.
2. Distinguish the feature and target set and divide the data set into training and test sets.
3. Normalize the train and test data.
4. Initialize and build the model. Identify the points of improvement and implement the same.
5. Print the accuracy score and confusion matrix (5 points).

Objective:

- To build a neural network-based classifier
- Compute the confusion matrix, and accuracy score on the given dataset.

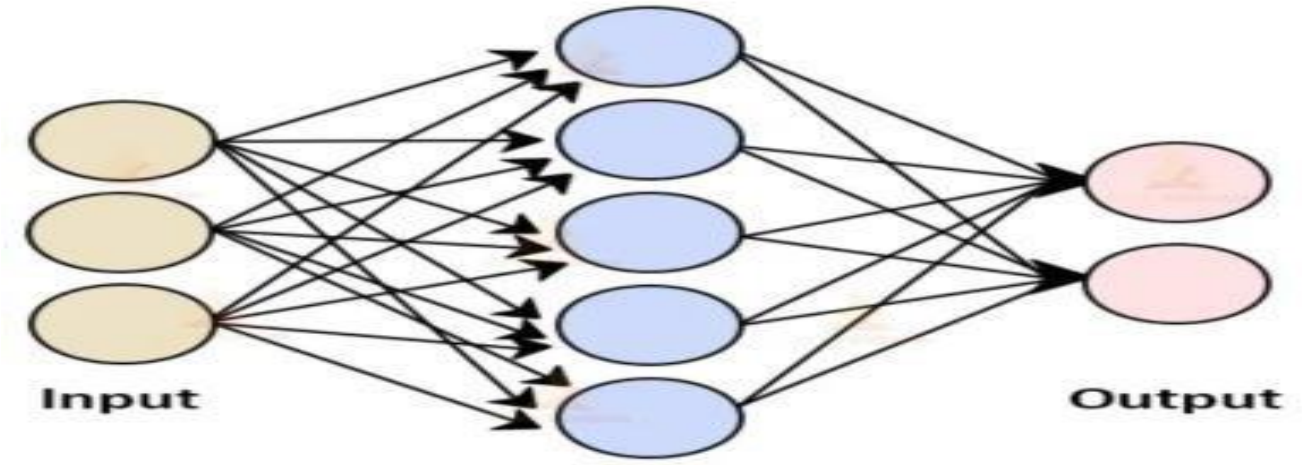
Theory:

- **What is a Neural Network?**

  A neural Network is a series of algorithms that try to mimic the human brain and find the relationship between the sets of data. It is being used in various use cases like regression, classification, Image Recognition and many more.

  The Artificial Neural Network does former processing is slower while in the latter processing is faster.

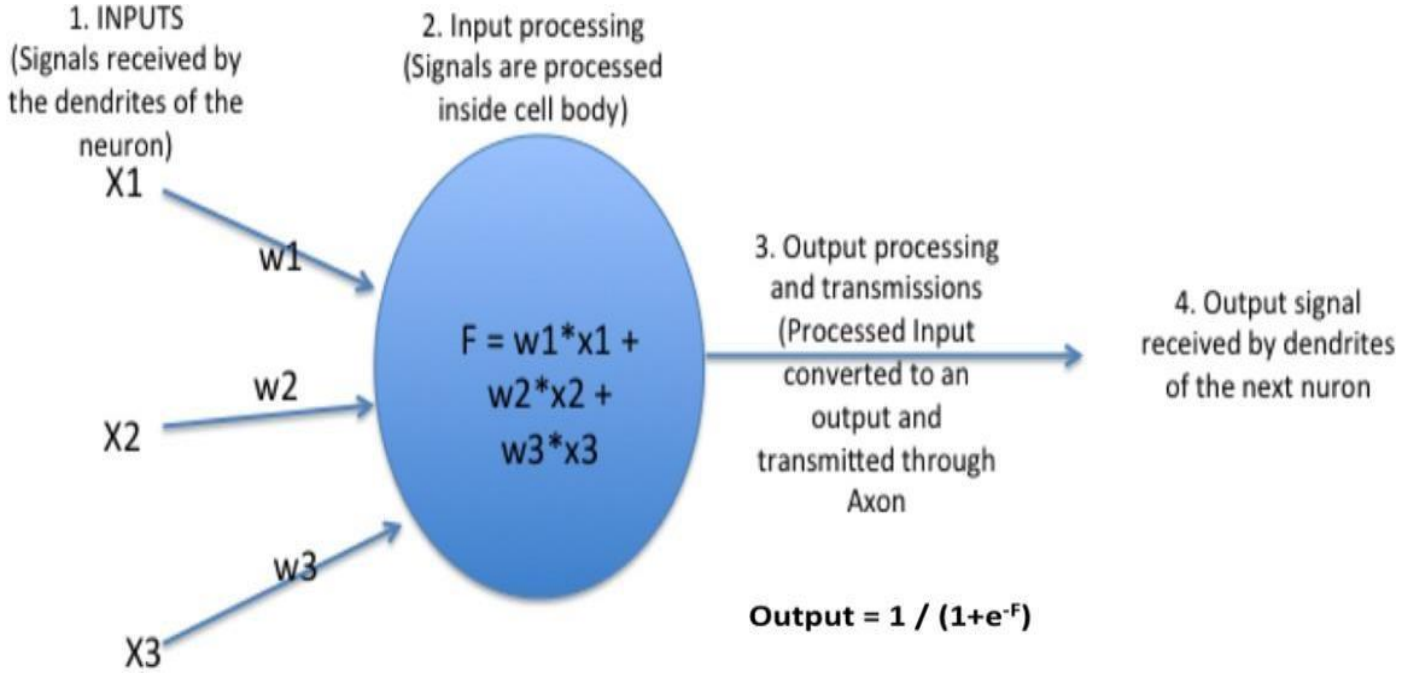
- **Architecture of ANN**



A neural network has many layers, and each layer performs a specific function, as the complexity of the model increases, the number of layers also increases that why it is known as the multi-layer perceptron.

- **Perceptron**

  As discussed above multi-layered perceptrons are the hidden or dense layers. They are made up of many neurons. Neurons are the primary unit that works together to form perceptron.



Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

- **Working With ANN**

  At First, information is fed into the input layer which then transfers it to the hidden layers, and interconnection between these two layers assigns weights to each input randomly at the initial point.

  Then bias is added to each input neuron after this, the weighted sum which is a combination of weights and bias is passed through the activation function.

  The activation Function has the responsibility of which node to fire for feature extraction and finally output is calculated.

- **Advantages**

1. ANN can learn and model non-linear and complex relationships as many relationships between input and output are non-linear.
2. After training, ANN can infer unseen relationships from unseen data, and hence it is generalized.

- **Applications**

  There are many applications of ANN. Some of them are:

1. Image Preprocessing and Character Recognition.
2. Forecasting.
3. Credit rating.
4. Fraud Detection.
5. Portfolio Management

Conclusion:

We have learned how to build a neural network-based classifier that can determine whether a bank customer will leave or not in the next 6 months.

Laboratory Practice-3: [DAA + ML + BCT]
Group-B (ML)

Group-B: Assignment No. 5

Title: Classify the Email using the Binary Classification Method

Problem Statement:

Classify the email using the binary classification method. Email Spam detection has two states:

a) Normal State – Not Spam

b) Abnormal State – Spam.

Use K-Nearest Neighbors and Support Vector Machine for classification. Analyze their performance.

Objective:

3. To read the dataset and classify whether Email is Spam or Not Spam using K-NN
4. To read the dataset and classify whether Email is Spam or Not Spam using SVM
5. Compute the confusion matrix, and accuracy score on the given dataset.

Theory:

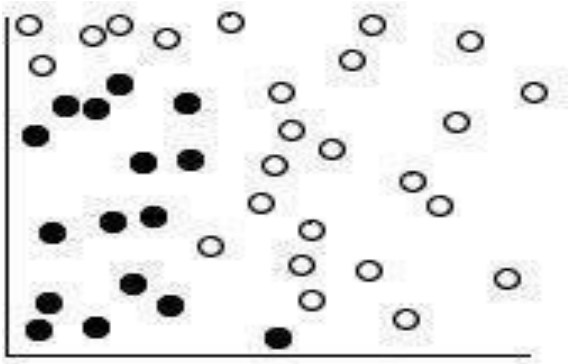
- **What is SVM?**

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data.

SVM has applications in many disciplines, including customer relationship management (CRM), facial and other image recognition, bioinformatics, text mining and voice and speech recognition.

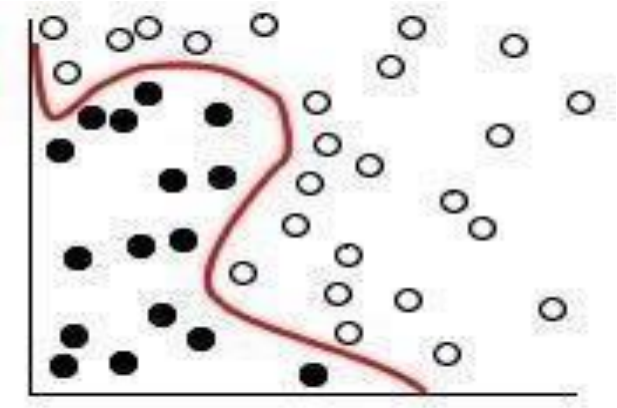
For example, consider the following figure, in which the data points fall into two different categories.

Figure 1. Original dataset



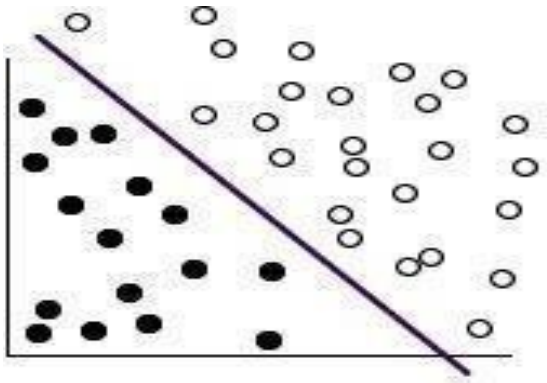
The two categories can be separated with a curve, as shown in the following figure.

Figure 2. Data with separator added



After the transformation, the boundary between the two categories can be defined by a hyperplane, as shown in the following figure.

Figure 3. Transformed data



The mathematical function used for the transformation is known as the kernel function.

SVM in Modeler supports the following kernel types:

1. Linear
2. Polynomial
3. Radial basis function (RBF)
4. Sigmoid

- **Tuning Hyperparameters**

- **Kernel**: The main function of the kernel is to transform the given dataset input data into the required form. There are various types of functions such as linear, polynomial, and radial basis functions (RBF). Polynomial and RBF are useful for non-

linear hyperplanes.

- **Regularization**: Regularization parameter in Python's Scikit-learn C parameter used to maintain regularization. Here C is the penalty parameter, which represents misclassification or error term.
- **Gamma**: A lower value of Gamma will loosely fit the training dataset, whereas a higher value of Gamma will exactly fit the training dataset, which causes over-fitting.

- **Advantages of SVM:**
- Effective in high-dimensional cases
- It is memory efficient as it uses a subset of training points
- SVM Classifiers offer good accuracy and perform faster prediction

- **Disadvantages of SVM:**
- SVMs do not directly provide probability estimates
- SVM is not suitable for large datasets because of its high training time

- **SVM Kernel:**

The SVM kernel is a function that takes low-dimensional input space and transforms it into higher-dimensional space, i.e., it converts not separable problem to a separable problem. It is mostly useful in non-linear separation problems. Simply put the kernel, does some extremely complex data transformations and then finds out the process to separate the data based on the labels or outputs defined.

Conclusion:

We understood what SVM is and, the different kernels used for hyperparameter tuning.

We also implemented K-NN and SVM for Email classification and found that SVM gives good results compared to K-NN for the given data set.