Assignment No:01

Problem Statement:

Predict the price of the Uber ride from a given pickup point to the agreed drop off location. Perform following tasks:

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

Objective :

The objective of a linear regression model is to find a relationship between one or more features(independent variables) and a continuous target variable(dependent variable).

Theory :

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.
Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Conclusion :

We learn linear regression and random forest regression models and
Linear regression makes predictions for continuous/real or numeric variables.

Assignment No: 02

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.

Objectives:

1) To detect and filter out spam and phishing emails with about 99.9 percent accuracy.
2) The implication of this is that one out of a thousand messages succeed in evading their email spam filter.

Theory:

a) Normal State- Not Spam :

Understanding the problem is a crucial first step in solving any machine learning problem. In this practical, we will explore and understand the process of classifying emails as spam or not spam. This is called Spam Detection, and it is a binary classification problem.

b) Abnormal State – Spam use KNN and SVM:

In supervised learning, a set of input variables, such as blood metabolite or gene expression levels, are used to predict a quantitative response variable like hormone level or a qualitative one such as healthy versus diseased individuals.

Both have been successfully applied to challenging pattern-recognition problems in biology and medicine. SVM and KNN exemplify several important trade-offs in machine learning (ML). SVM is less computationally demanding than KNN and is easier to interpret but can identify only a limited set of patterns.

Conclusion:

In this way we studied about Email classification using binary classification method and email spam detection.

Assignment No:03 (According to Syllabus)

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 Pre-process the dataset.

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 Customerld, CreditScore, Geography, Gender, Age, Tenure, Balance, etc.

Objective :

Students should be able to distinguish the feature and target set and divide the data set into training and test sets and normalize them and students should build the model on the basis of that.

Theory :

The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes

Normalization

Normalization is a scaling technique in Machine Learning applied during data preparation tochange the values of numeric columns in the dataset to use a common scale. It is not necessary for all datasets in a model. It is required only when features of machine learning models have different ranges.

Confusion Matrix

The confusion matrix is a matrix used to determine the performance of the classi cation models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may beconfusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as anerror matrix

Conclusion :

In this way we build a neural network-based classifier that can determine whether they will leave or not in the next 6 months

ASSIGNMENT No: 3

Problem statement-

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

Objective-

To use a database in which the data points are separated into several classes to predict the classification of a new sample point.

Procedure-

K-Nearest Neighbor (KNN) Algorithm

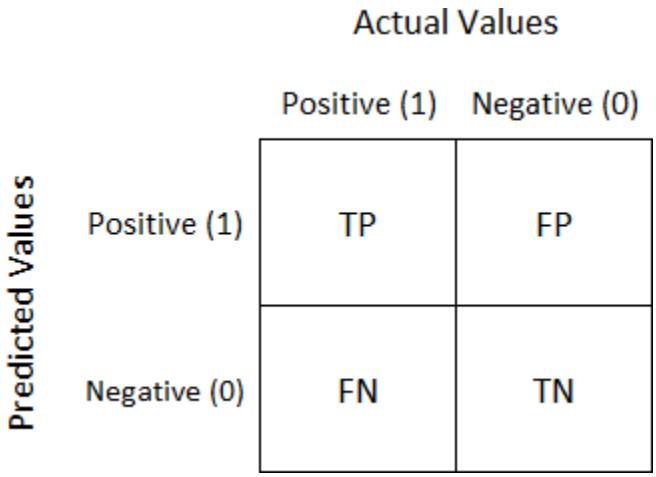
- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- 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.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new

By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B.

$$\text{Euclidean Distance between } A_1 \text{ and } B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$$

Confusion matrix

It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.



Conclusion –

In this way we find the nearest neighbor data using a K-Nearest neighbor algorithm and by using confusion matrix we find accuracy, error, precision and recall.

ASSIGNMENT N0: 4

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 minimize the sum of distances between the points and their respective cluster centroid.

Procedure-

K-Means Clustering

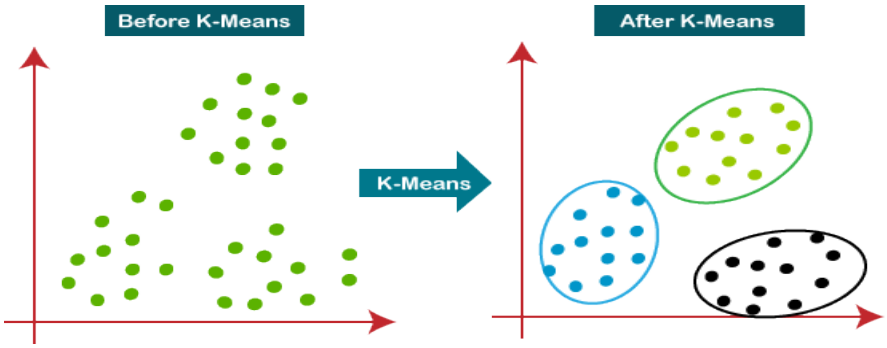
K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has data points with some commonalities, and it is away from other clusters.



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 to calculate the value of WCSS (for 3 clusters) is given below:

$$WCSS= \sum_{P_i \text{ in Cluster1}} distance(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} distance(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} distance(P_i\ C_3)^2$$

In the above formula of WCSS,

$\sum_{Pi \text{ in Cluster1}}$ distance(P_i C_1)2: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

Conclusion –

In this way we find the similar group of cluster using K means clustering and find the number of groups using elbow method.