Page No.
Date

Assignment No. 1

Kavita Sahu
BALO19109

Title:- Predict the price of Uber ride

Date of Completion:-

Objective:-
is to apply different pre-processing method, and Evaluate the model.

Problem Statement:-
Predict the price of the Uber ride from a given pickup point to the agreed drop-off Location
Perform the following task -
1. Pre-process the dataset.
2. Identify outliers
3. Check the correlation
4. Implement the linear regression and random forest regression models.
5. Evaluate the models and compare their respective scores like R2, RMSE etc.

Software and Hardware Requirement:-

Theory:-
Data Preprocessing can be done in four different ways:-

1) Data cleaning
Data in the real world is frequently incomplete, noisy, and inconsistent. Many bits of the data may be irrelevant or missing. Data cleaning is carried out to handle this aspect.

2) Data Integration:-
It is involved in a data analysis task that combines data from multiple sources into a coherent data store. These sources may include multiple databases.

3) Data Transformation:-
This stage is used to convert the data into a format that can be used in the mining process.

4) Data Reduction:
Because data mining is a methodology for dealing with large amount of data. When dealing with large amount of data, analysis becomes very difficult.

* Outliers -
This are the extreme values that differs from most other data points in a dataset. They can be big impact on

your statistical analysis and skew the result of any hypothesis test.

There are four ways to Identify outliers:-

1) Sorting method:
You can sort quantitative variable from low to high and scan for extremely low or extremely high values.

2) Using visualizations:-
You can use software to visualize your data with a box plot, or a box-and whisker plot so you can see the data distribution at a glance.

3) Statistical outlier detection -
Statistical outlier detection involves applying statistical tests or procedures to identify extreme values.

4) Using IQR - Inter Quartile Range.
$IQR = Q_3 - Q_1$.
Upper fence $= Q_3 + (1.5 * IQR)$
Lower fence $= Q_1 - (1.5 * IQR)$.
Firstly sort your data from low to high.
Identify the first quartile Q_1, the median and third quartile (Q_3).

- Correlation -
The correlation coefficient measures the relationship between two variables

The correlation coefficient can never be less than -1 or higher than 1.

1 = there is perfect linear relationship between the variable.

0 = there is no linear relationship between the variable.

-1 = there is a perfect negative linear relationship between the variable

a. Linear Regression -
is a machine learning algorithm based on supervised learning. Regression models a target prediction value based on independent variables.

b. Random forest is a supervised Machine learning Algorithm that is used widely in classification and Regression problems.

Conclusion -
Performed the different maintained task.

Page No.
Date

Assignment No:

Kavita Sahu
BAC019109.

Title:- Gradient Descent Algorithm.

Date of Completion:-

Objective:-
To find local minima of a function

Problem statement:-
Implement Gradient Descent Algorithm to find the local minima of a function for example,
Find the local minima of the function $y = (x+3)^2$ starting from the point $x = 2$.

Software and Hardware Requirements

Theory-
Gradient descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms.
It is basically used for updating the parameters of the learning model.

It is used to find a local minimum / maximum of a given function.
This method is commonly used in Machine Learning (ML) and deep Learning (DL) to minimize a cost / loss function (eg. in a linear regression).
Due to its importance and ease of implementation, this algorithm is usually taught at the beginning of almost all machine learning courses.

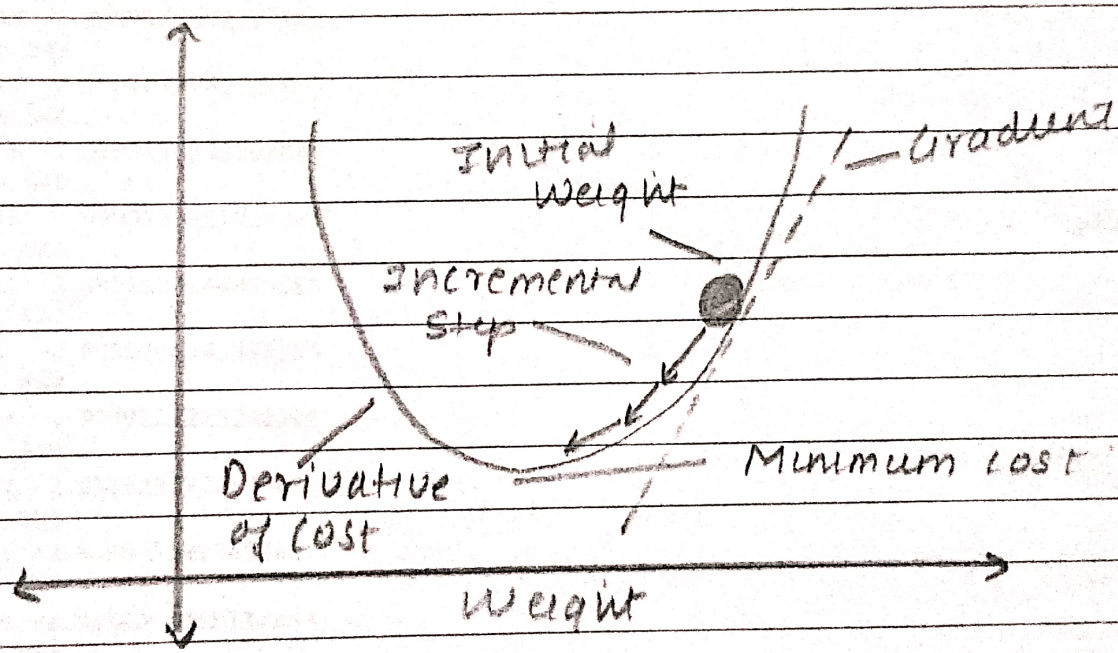
Gradient descent algorithm does not work for all functions.
There are two specific requirements.
A function has to be:-
1) differentiable
2) convex.

- If we move towards a negative gradient or away from the gradient of the function at the current point, it will be the local minimum of that function.

- Whenever we move towards a positive gradient or towards the gradient of the function at the.

current point, we will get the local maximums of that function.

The entire procedure is known as Gradient Ascent, which is also known as steepest descent.
The main objective of using a gradient descent algorithm is to minimize the cost function using iteration.



Conclusion:-
Implemented Gradient Descent Algorithm.

Assignment No. 3.

Kavita Sahu
BAC019109.

Title:- Neural Network classifier.

Date of completion:-

Objective:-
Is to normalize the test and train data and find the accuracy score and confusion matrix.

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.
The dataset contains 10,000 samples points with 14 distinct features such as CustomerId, Creditscore, Geography, Gender, Age, Tenure, Balance etc.
Perform following steps:-

1. Read the dataset.
2. Distinguish the feature and target set and divide the dataset 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 accuracy score and confusion matrix.

Page No.
Date

Software and Hardware Requirements:

Theory -
Neural Network consists of units (neurons), arranged in layers, which convert an input vector into some output.
Each unit takes an input, applies function to it and then passes the output on to the next layer.

Neural networks have found application in a wide variety of problems. These range from function representation to pattern recognition.

Accuracy -
It is defined as total correctly classified example divided by the total number of classified examples.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

This metric is very important when error in predicting all class is equally important.

Let take an example -

Email is spam or not spam. In this case if our model classify a email send by boss is spam and don't show it is more harmful than showing small amount of email as spam.

Confusion Matrix -
The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted and the other axis is the actual label.

		ACTUAL VALUES	
		Positive	Negative
PREDICATED VALUES	Positive	TP	FP
	Negative	FN	TN

Conclusion: -
Performed the above maintained steps in the given dataset.

Assignment No. 4.

Kavita Sahu
BA(0191109.

Title:- K-Nearest Neighbors.

Date of Completion:-

Objective -
Is to compute confusion matrix, accuracy etc on the given dataset.

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

Software and Hardware Requirements:-

Theory -
The K-Nearest neighbors algorithm also known as KNN or K-NN is a non parametric, supervised learning classifier which uses proximity to make classifications or predictions about the grouping of an individual data point.

The goal of K-nearest neighbor algorithm is to identify the nearest neighbor of a query point, so that we can assign a class label to that point.

Determine your distance metrics:-
Well there are several distance measures that you can choose from, this -
Euclidean distance ($p=2$):- This is the most commonly used distance measure and it is limited to real valued vectors.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Advantages of KNN Algorithm-
- Easy to implement
- Adapts easily
- Few Hyperparameters-
-

Confusion matrix -
It is a matrix of size 2x2 for binary classification with actual values on one axis and predicted on another.

Precision -

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

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

Out of the total positive, what percentage are predicted positive. It is same as TPR (true positive rate).

Conclusion -

Implemented KNN algorithm on diabetes.csv dataset.

Page No.
Date

Assignment No: 5

Kavita Sahu
BACO19109.

Title:- K Means Clustering

Date of Completion:-

Objective:-
To find groups in the data, with the number of groups represented by the variable K.

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

Software and Hardware Requirements-

Theory -
K-Means clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. On this, it groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be

created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

The K-means clustering algorithm mainly perform two tasks:-

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

The Elbow Method run K-means clustering on the dataset on the dataset for a range of values for K (say from 1-10) and then for each value of K computes an average score for all clusters.

Conclusion:-
Implemented K-Means clustering on Sales_data_sample.csv dataset.