ASSIGNMENT No- 1

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

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

Prerequisite : 1) Basic knowledge of Python.
2) Concept of preprocessing data.
3) Basic knowledge of Data Sciend & Big data analytics.

Theory :-

Data Preprocessing :

It is a process of preparing the raw data and making it suitable for a machine learning model. It is first and crucial step while creating a ML model.

While doing any operation with data, it is mandatory to clean it and put it in a formatted way.

Thus, we use data preprocessing.

Linear Regression :-

Linear regression is one of the easiest & most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.

It shows a linear relationship between dependant (y) and one or more independent (y) variables.

Since, linear regression shows the linear relationship, it finds how the value of dependant value of dependant variable is changing according to the value of independant variable.

Random Forest Regression Models :-

Random Forest is a popular machine learning algorithm that belongs to supervised learning tech. It can be used for both classification & regression problems in ML.

Random Forest is a classifier that contains a number of decision trees on various subsets of given dataset and takes average to improve the predictive accuracy of that dataset.

The greater number of trees in the forest leads to higher accuracy & prevents the problem of overfitting.

Boxplot :-

Boxplots are a measure of how well data is distributed across a data set. This divides the data set into three quartiles.

This graph represents minimum, maximum, average rst quartile, and the third quartile in the dataset.

R provides a boxplot() function to create a boxplot

There is foll syntax of boxplot() function:

boxplot(x, data, notch, varwidth, names, main)

S.N.	Parameter	Description.
1.	x	It is a vector or a formula
2.	data	It is the data frame
3.	notch	It is logical value set as true to a notch
4.	varwidth	It is also logical value set as true to draw
5.	names	It is group of labels that will be printed
6.	main	It is used to give a title to the graph

Outliers :-

Outliers refer to data points that exist outside of wha is to be expected. The major thing about the outliers is what you do with them. If you are going to analyze any tas to analyze datasets, you will always have some assumptions based on how this data is generated.

Matplotlib :-

Matplotlib is an amazing visualization library in Python on 2D plots of arrays. Matplotlib is a multi-platfor data visualization library built on Numpy arrays and designed to work with broader Scipy stack.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals.

Matplotlib consists of several plots like line, bar, scatter, histogram etc

Mean Squared Erros :-

The MSE of an estimator measures the average of erros squares i.e the average squared differences between the estimated values & true values

It is a risk function, corresponding to the expected value of the squared erros loss. It is always non-negativ and values close to zero are better

Conclusion :- In this way, We have explord concept correlation and implement linear regression and random forest regression models.

ASSIGNMENT - 2.

Title :-

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 neighbhbours & support vector Machine for classification.

Analyze their performance

Objectives :-

Students should be able to classify email using the binary classification & implement email spam detection technique by using K-nearest Neighbours & Support Vector Machine algorithm.

Prerequisites :-

- Basic knowledge of Python.
- Concept of K-Nearest Neighbhour & Support Vector machine for classification.

Theory :-

Data Processing.

Data preprocessing is a process of preparing the raw data & making it suitable for machine learning model.

It is first & crucial step while creating a machine learning model.

when creating a machine learning project, we don't always come across clean & formatted data. while formatted data is mandatory.
So for this, we use data preprocessing task.

Why do we need data preprocessing?
A real-world data generally contains meissing values, noises, & maybe in an unusable format which cannot be directly used for machine learning model which also increases the acuracy & efficiency of machine learning model. It involves below steps:-

- Getting the dataset.
- Importing libraries.
- Importing datasets.
- Finding missing data.
- Encoding categorial data.
- Splitting dataset into training & test set.
- Feature scaling.

Conclusion:-

Hence, we implemented knearest neighbours & support vector machine for classificatio for email spam detection.

ASSIGNMENT - 3.

Title:-

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

Objectives:-

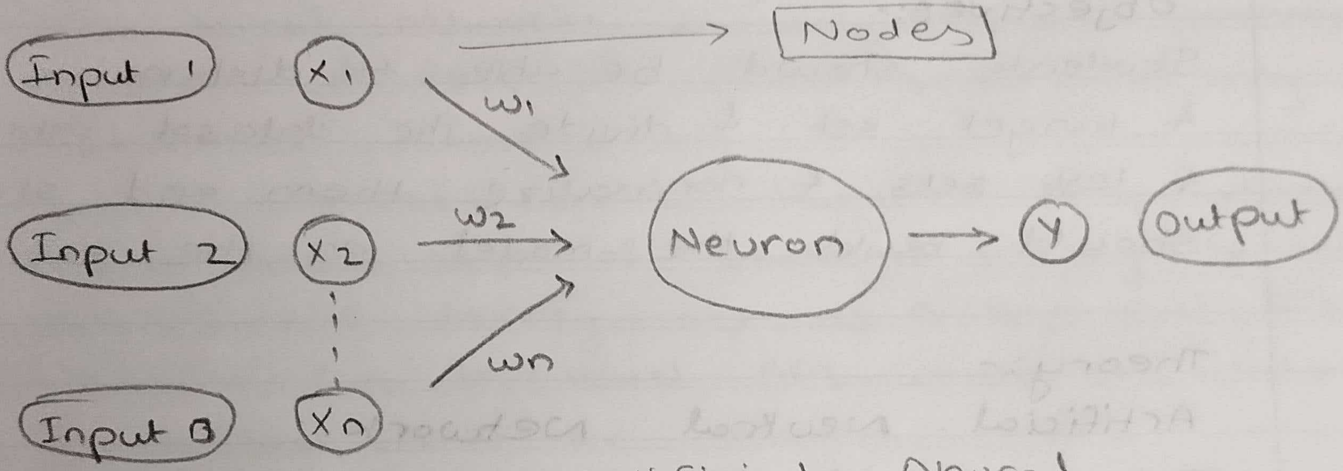
Students should be able to distinguish the feature & target set & divide the dataset into training & test sets & normalize them and students should build the model on the basis of that.

Theory:-

Artificial Neutral Network.

The term 'Artificial Neutral Network' is derived from biological neutral networks that develop the structure of human brain. Similar to human brain that has neurons interconnected to one another, artificial neutral networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes.

Biological Neural Network.	Artificial Neural Network
Dendrites	Inputs
Cell Nucleus	Nodes
Synapse	Weights
Axon.	Output.



Artificial Neural Network.

Keras:-

Keras is an open-source high-level Neural Network library, which is written in Python is capable enough to run on Theano, Tensor Flow or CNTK. It was developed by one of the Google engineers, Francois Chollet. It is made user-friendly, extensible & modular for facilitating faster experimentation with deep neural networks. It only supports Convolutional Networks & Recurrent Networks individually but also their combination.

It cannot handle low-level computations, so it makes use of Backend library to resolve it. Backend library act as a high-level API wrapper for low-level API, which lets it run on TensorFlow, CNTK or Theano.

·TensorFlow.

TensorFlow is a Google product, which is one of the most famous deep learning tools widely used in the research area of ML and deep neural network. It came into the market on 9th November 2015 under the Apache License 2.0. It is built in such a way that it can easily run on multiple CPUs & GPUs as well as on mobile operating systems. It consists of various wrappers in distinct languages such as Java, C++, Python.

Normalization

Normalization is a scaling technique in ML applied during data preperation to change the values of numeric columns in dataset to use a common scale. It is not necessary for all datasets in a model.

$$Xn = (X - X_{min}) / (X_{max} - X_{min})$$

Normalization techniques in ML.

Min-Max Scaling:- This technique is also referred as scaling. As we have already discussed above the Min-Max scaling method helps the dataset to shift & rescale the values of their attribute so they end up ranging between 0 ~~to~~ and 1.

Standardization Scaling:- It is also known as z-score normalization, in which values are centred around the mean with a unit standard deviation, which means the attribute becomes zero & the resultant distribution has a unit standard deviation.

$$x' = \frac{x - \mu}{\sigma}$$

Confusion Matrix.

Confusion matrix is a matrix used to determine the performance of a classification 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 be confusing

Since, its shows the errors in the model performance in the form of a matrix, hence also known as an error matrix. Some features of confusion matrix are given below:-

		Actual	Value.
		+ve(1)	-ve (0)
Predective Values.	+ve (1)	TP	FP
	-ve(0)	FN	TN

True Negative:- When models prediction and actual values is also No.

True Positive:- Model predicted yes, but actual value is No.

False Negative:- Model predicted No, but actual value is Yes. Type-II Error.

False Positive:- Model predicted Yes, but actual value is No. Type-I Error.

Calculations.

Accuracy:- It defines how often the model predicts the correct output.

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

Error Rate :- It defines how often model gives wrong prediction.

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

Precision :- It can be defined as no. of correct outputs provided by the model.

$$\text{precision} = \frac{TP}{TP + FP}.$$

Recall :- correct predictions out of all +ve values

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

F-measure :- Evaluate recall & precision at same time.

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}.$$

Conclusion :-

In this way we build a neural network based classifi~~cation~~er that can determine whether they will leave or not in next 6 month.