Assignment No : 1

Page No :
Date : / /

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

1. Pre-process the dataset.
2. Identify outliners.
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 : Students should be able to process dataset and identify outliers, to check correlation and implement linear regression and random forest regression models.

- Prerequisite : 1. Basic knowledge of python
2. Concept of pre-processing data.
3. Basic knowledge of data science & big data analytics.

- Theory :

1. Data Preprocessing :

It is a process of preparing the raw data & making it suitable for a machine learing model. It is an imp. and crucial step while creating a machine learning model.

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

Why do we need Data Preprocessing ?

A real-world data generally contains noise, missing values & maybe in an unusable format which cannot be directly used for machine learing models. Hence

It is required tasks for cleaning the data and making it suitable for the model.
- It Involves below steps :
1. Getting the dataset 2. Importing libraries
3. Importing datasets 4. Finding missing data
5. Encoding categorical data 6. Split dataset (train & test)
7. Feature scaling.

2. Linear Regression :
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
- It is a statistical method that is used for predictive analysis.
- It makes predictions for continuous / real or numeric variables such as sales, age, product etc.
- It provides a sloped straight line representing the relationship between variables.

3. Random Forest Regression models :
- It is used for both classification and regression problems in ML.
- It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Random forest is a classifier that contains a no. of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

- The greater no. of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

4. Boxplot :

- Boxplots are a measure of how well data is distributed across a data set. This data divides the data set into three quartiles.
- This graph represents the minimum, max, average & third quartile in the data set.
- Syntax of boxplot() function :

Sr. No.	Parameter	Description
1.	x	It is a vector or a formula.
2.	data	It is the data frame.
3.	Notch	Logical value set as true to draw a notch.
4.	main	It is used to give a title to graph.

5. Outliers :

-It refers to the data points that exist outside of what is to be expected.

- Types : 1] Global outliers
  2] Collective outliers
  3] Contextual outliers.

6. Haversine :

- The Haversine formula calculates the shortest distance between two points on a sphere using their latitudes & longitudes measured along the surface. It is important for use in navigation.

7. Matplotlib :

- It is a visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on Numpy arrays and designed to work with the broader Scipy Stack.

8. Mean Squared Error :

- The MSE or MSD (Deviation) of an estimator measures the average of error squares. i.e. the average squared diff. between the estimated values and true value.

- Conclusion : In this way we have explored concept correlation and implement linear regression and random forest regression models.

Assignment No:

Page No:
Date: / /

- Title : Classify the email using the binary classification method. Email spam detection has 2 states :
  a] Normal state - Not spam
  b] Abnormal state - Spam
  c] Use K-nearest neighbors and support vector machine for classification. Analyze their performance.

- Dataset Description : The csv file contains 5172 rows, each row for email. There are 3002 columns. The first column indicates email name. The last column has the labels for predictions. Thus info. regarding all 5172 emails are stored in a compact dataframe rather than as separate text files.

- Objective : Students should be able to classify the binary classification and implement email spam detection technique by using K-nearest neighbors and support vector machine algorithm.

- Prerequisite : 1. Basic knowledge of python.
  2. Concept of K-nearest Neighbors & support vector machine for classification.

- Theory :

1. Data Prepmcessing :
   - It is a process of preparing the raw data and making it suitable for a machine learning model. It is imp and crucial step while creating a machine learning model.

2. Binary classification :
   - It is a type of supervised learning where the target

variables has only two possible outcomes or classes.
- The goal is to train a model that can predict one of the two classes with high accuracy.

3. K-nearest neighbors :
- It is a popular supervised learning algorithm used for classification and regression tasks.
- It handles non-linear relationships and it is easy to implement.

4. Support Vector Machine :
- Prepare your data by scaling / normalizing the features.
- Choose a kernel function to transform the data into a higher-dimensional space.
- Find the optimal hyperlane that maximally separtes the classes.
- Maximize the margin between the hyperlane and the data points.

5. Train-test-split procedure :
- Use the training set to train the model.
- Use the validation set to tune hyperparameters and evaluate the model's performance.
- Use the testing set to evaluate the final performance of the train model.

- Conclusion : Hence we have studied the binary classification method.

Assignment No:

Page No :
Date : / /

- Title : Given a bank customer, build a neural-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 CustomerId, CreditScore, Geography, Gender, Age, Balance etc.

- Objective : Students should be able to distinguish the feature and target set and divide the data set into training and tests sets and normalize them & students should build the model on the basic of that.

- Prerequisite : 1. Basic knowledge of Python
  2. Concept of confusion Matrix

- Theory :

1. Artificial Neural Network :

- The name is dervied from biological neural networks that develop the structure of a humain 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.

Biological Neural Network	Artifical Neural Network
Dendrites	Inputs
Cell nucleus	Nodes
Synapse	Weights
Axon	Output

- Architecture of an artifical neural network :
- There are mainly 3 layers
1. Input layer : It accepts inputs in several different formats provided by the programmer.
2. Hidden layer : It is in-between input & output layers. It performs all the calculations and hidden features & patterns.
3. Output layer : The input goes through a series of transformations using the hidden layer, which results in output that is conveyed using this layer.

2. Keras :
- It is an open-source high-level neural network library which is written in Python is capable enough to run on Theano, TensorFlow.
- It was developed by one of the google engineers, Francois Chollet.
- It is made user-friendly, extensible and modular for facilitating faster experimentation with deep neural network

3. TensorFlow :
- It is a Google product, which is one of the most famous deep learning tools. widely used in the research area of machine learning and deep neural network.
- It is built in a such way that it can easily run on multiple CPUs and GPUs as well as on mobile OS.
- It consists of various wrappers in distinct languages such as Java, C++ or python.

4. Normalization :
- It is a scaling tech. in ML applied during data preparation to change the values of numeric columns

in the dataset to use a common scale.
Mathematically, $X_n = (X - X_{min}) / (X_{max} - X_{min})$

- Min-Max Scaling : This tech. is also refered to as scaling. As we have already discussed, this method helps the dataset to shift and rescale the values of their attributes, so they end up ranging between 0 & 1.

- Standardization scaling : It is also known as Z-score normalization, in which values are centered around the mean with a unit standard deviation, which means the attributes becomes zero and the resultant distribution has a unit SD.
  - It is expressed as follows:

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

5. Confusion Matrix : It is used to determine the performance of the classification models for a given set of test data. It shows the errors in the model performance in the form of a matrix, hence also known as error matrix.
   - It looks like the below table :

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted values	(1)	True Positive (TP)	False Positive (FP)
	(0)	False Negative (FN)	True Negative (TN)

- TN : Model has given prediction No, and the real or actual value was also No.
- TP : Model has predicted yes, and the actual value was also true.

- FN : Model has predicted no, but the actual value was Yes, it is also called as Type-II error.
- FP : Model has predicted yes, but the actual value was No, it is also called as Type-I error.

- Calculations using Confusion Matrix :

1] Accuracy : It is the ratio of the no. of correct predictions made to all the no. of predictions.

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

2] Error rate : Ratio of the wrong predictions made to all the no. of predictions.

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

3] Precision : $\frac{TP}{TP + FP}$

4] Recall : $\frac{TP}{TP + FN}$

5] F-measure : $\frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$

6] ROC Curve : The ROC is a graph displaying a classifier performance for all the possible thresholds. The graph is plotted between the true positive rate & false positive rate.

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

Assignment No :

- Title : Implement K-nearest Neighbors algorithm on diabets.csv dataset. Compute confusion matrix, accuracy, error rate, precision & recall.

- Dataset Description : We will try to build a ML model to accurately predict whether or not the patients in the dataset have diabetes or not ?
The datasets consists of several medical predictor variables and one target variable, outcome. Predictor variables includes the number of pregancies the patient has had, their BMI, insulin level, age and so on.

- Objective : Students should be able to preprocess dataset and identify outliers, to check correlation and implement KNN algorithm and random forest classification models.

- Prerequisite : 1. Basic Knowledge of Python.
2. Concept of confusion matrix
3. Concept of Roc-auc curve
4. Concept of Random forest & KNN algo.

- Theory :
KNN : K-Nearest Neighbours is a supervised ML model. Supervised learning is when a model learns from data that is already labeled.
A supervisied learning model takes in a set of input objects and output values.
The model then trains on that data to learn how to map the inputs to the desired output so it can learn to make predictions on unseen data.
KNN models work by taking a data point and looking at the 'K' closest labeled data points.

The data point is then assigned the label of the majority of the 'k' closest points.
Scikit-learn is a ML library for python.

For our K-NN model, the first step is to read in the data we will use as input. For this example, we are using diabetes dataset. To start, we will use Pandas to read in the data.

- Split up the dataset into inputs and targets
- Split the dataset into train and test data.
- Building and training the model
- Testing the model

- K-Fold Cross-Validation :
Cross-validation is when the dataset is randomly split up into 'k' groups.
One of the groups is used as the test set and the rest are used as the training set.
The model is trained on the training set and scored on the test set.
Then the process is repeated until each unique group as been used as the test set.

For example, for 5-Fold cross validation, the dataset would be split into 5 groups, and the model would be trained and tested 5 separate times so each group would get a chance to be the test set.

The AUC-ROC curve helps you to evaluate how good a model is at distinguishing between positive and negative classes.
We can use this curve to decide whether to implement a ML model.

- 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 :

Page No :
Date : / /

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

- Dataset Description : 1. Customer ID 2. Gender 3. Age 4. Annual income 5. Spending score of the customer.

- Objective : Students should able to understand how to use unsupervised learning to segment diff. clusters or groups and used to them to train your model to predict future things.

- Prerequisite : 1. Unsupervisied learning
  2. Clustering
  3. Elbow method.

- Theory :
  Clustering algorithms try to find natural clusters in data, the various aspects of how the algorithms to cluster data can be tuned and modified.
  - Clustering is based on the principle that items within the same cluster must be similar to each other.
  - The data is grouped in such a way that related elements are close to each other.

- Uses of clustering :

1. Marketing :
   In the field of marketing, clustering can be used to identify various customer groups with existing customer data. Based on that, customers can be

provided with discounts, offers, codes etc.

2. Real Estate :
Clustering can be used to understand and divide various property locations based on value and importance. Clustering algo. can process through the data and identify various groups of property on the basis of portable price.

3. BookStore and Library management :
Libraries and book stores can use clustering to better manage the book database. With proper book ordering better operations can be implemented.

4. Document analysis :
Often, we need to group together various research texts and documents according to similarity. Using clustering the algo. can process the text and group it into different themes.

- K-Means clustering :
  - It is an unsupervisied ML algorithm that divides the given data into the given no. of clusters.
  - Here, the 'K' is the given number of predefined clusters, that need to be created.
  - It is a centroid based algorithm in which each cluster is associated with a centroid.
  - The main idea is to reduce the distance between the data points and their respective cluster centroid.

- The algo. takes raw unlabelled data as an input and divides the dataset into clusters and the process is repeated until the best clusters are found.
- K-means is very easy and simple to implement.
- It is highly scalable, can be applied to both small and large datasets.

Conclusion : In this way, we have used K-means clustering on the given dataset.