Team Members:

- Poonam Kankariya (Class ID: 10)
- Sai Bhavani Nikita Rayapareddy (Class ID: 33)

Introduction:

The given lab assignment is associated with programming with Python. This explores the various machine and deep learning concepts of handling data in terms of classification and/or regression using supervised and unsupervised approaches.

Objectives:

The lab work involves a collection of tasks to be accomplished listed as below. * Creation of dictionary using keys and their corresponding values using a list of tuples containing information of students * Evaluation of a string to identify the longest sub-string existing within without the repetition of characters resulting in the derivation of the sub-string with new revised length * Development of an airline booking system that enables the user to reserve an airline ticket per user needs * Performing multiple regression of a data set and evaluating its performance (accuracy) using Root Mean Squared Error (RMSE) and R Squared (R2) methods of performance evaluation * Using a data set containing both qualitative and quantitative data, perform an exploratory analysis to identify the most correlated features associated with the target, to remove null values associated with the features, if any, and to convert any categorical features into numerical features * Perform classification on data set comprising of both qualitative and quantitative data using the algorithms: Naive Bayes, Support Vector Machines and K-Nearest Neighbors * Cluster analysis on a data set of our choice, computing their performance using the silhouette score and visualizing their results, along with application of elbow method to identify the ideal number of clusters

Concepts Incorporated:

The following concepts have been explored and utilized to work on data of different forms to accomplish different objectives. * Dictionary, list and tuples * String and sub-string concepts * Class and their corresponding functions and approaches * Multiple Regression along with RSME and R2 methods of performance evaluation * Data Analysis * Classification - Naive Bayes, Support Vector Machine (Linear SVM) and K-Nearest Neighbors * K-Means Clustering along with silhouette score approach of performance evaluation and elbow method to determine optimal clusters

Data Sets Used:

Multiple data sets have been to accomplish different objectives. * Diabetes (Clustering) * Boston Housing (Multiple Regression) * Bank (Classification)

Question 1:

Workflow:

To accomplish the goal of creating a dictionary of keys and their corresponding values using a list of tuples containing information of students: * Identify the list of tuples containing student information in an array form * Evaluate if the dictionary created contains the student information * If student name exists in the dictionary, update their subject name and corresponding scores earned * If student name does not exist in the dictionary, append it to the same and then add its associated information of subject and score * Display the updated dictionary with all the student information that is available in the list of tuples to the user

Parameters:

List of tuples containing student information is: [("akash",('Physics',90)), ("gaurav",('Arts', 92)), (("anand", 14)), ("suraj",('History',20)), ("akhil", ('Chemistry',25)),("akash",('Chemistry',95)), ("ashish",('Maths',30))]

Evaluation:

The dictionary displayed to the user is: {'akash': [('Physics', 90), ('Chemistry', 95)], 'gaurav': [('Arts', 92)], 'anand': [14], 'suraj': [('History', 20)], 'akhil': [('Chemistry', 25)], 'ashish': [('Maths', 30)]}

Question 2:

Workflow:

To accomplish the evaluation of sub-string and its length derived from the string made available to the user: * Obtain a string as an input from the user

- Assign placeholders to hold characters of the string made available and the sub-string derived by identifying the non-repetitive characters

- Using a set established, identify if the character is present in the set
  - If yes, update the sub-string derived be the longest
  - If no, add the character to the sub-string and update its length

- Once all the characters of the original string are compared, display the resulting longest sub-string derived

Parameters:

The string provided by the user is 'poonam' of length 6.

Evaluation:

The longest sub-string obtained from the string is 'onam' of length 4.

Question 3:

Workflow:

To accomplish the task of creating an airline booking reservation system: * Create multiple super and sub-classes pertaining to Person, Airline, Flight, Passenger, Employee etc. * Define variables and functions within each of the class and identify them as public or private based on them belonging to a super-class or within the same class * Create a main function that calls for this airline booking reservation system to present the desired output to the customer

Evaluation:

Question 4:

Workflow:

To accomplish the goal of multiple regression: * Import appropriate libraries required to perform the task of multiple regression

- Identify a data set and read the same in an appropriate format

- Compute the skewness of the original distribution of the data and normalize the same

- Identify the features that are positively and negatively correlated to the target variable, leading to elimination of any features that may not be correlated or useful for the process, if needed

- Identify the pivotal feature associated with the target variable, the one with correlation rate exceeding 50% (0.5)

- Identify any null values associated with the features and either assign a specific value (for instance, 0) or remove them from the data set as considered appropriate

- Split the data set into training and testing sets to be used appropriately

- Build the linear regression model involving multiple predictor features

- Compute the R2 and RMSE scores to evaluate the performance of the model obtained

- Display the model obtained in a graphical format for simple visual interpretation

Parameters:

Boston housing data set has the following predictor features used to determine the median valuation of owner-occupied homes. * crim * zn * indus * chas * nox * rm * age * dis * rad * tax * ptratio * black * lstat

Evaluation:

Target variable 'medv' (median valuation) of the owner-occupied homes is spread well using the multiple regression with an acceptable level of performance based on the scores obtained as follows. * R2 = 0.83 (closer to 1 results in a good performing model) * RMSE = 0.02

(closer to 0 results in a good performing model)

Discussion:

The model fit using linear regression involving multiple features is a good fit with very low margin of error ensuring an optimum performance of the same on test data set. This model neither under fits nor over fits the data set making it ideal for further use on newer data set.

Question 5:

Workflow:

To accomplish the tasks of exploratory data analysis and classification: * Import appropriate libraries required to perform the task of data analysis and classification

- Identify a data set containing both quantitative and qualitative features, and read the same in an appropriate format

- Identify any null values associated with the features and either assign a specific value (for instance, 0) or remove them from the data set as considered appropriate

- Convert the non-numerical (categorical) features into numerical features

- Identify the features that are positively and negatively correlated to the target variable

- Identify the predictor features (x) and target variables from the update data containing all numerical values
- Remove any of the features that may not be correlated or useful for the process of classification, as considered appropriate
- Split the data set into training and testing sets to be used appropriately

- Fit the model using Naive Bayes algorithm, predict its result using the test data set and evaluate the performance (accuracy) of the same using score metrics

- Fit the model using Linear SVM algorithm, predict its result using the test data set and evaluate the performance (accuracy) of the same using score metrics

- Fit the model using K-nearest neighbor algorithm, predict its result using the test data set and evaluate the performance (accuracy) of the same using score metrics

Parameters:

The parameters used for exploratory data analysis and classification from the bank data set are: * Age of the client/customer * Job * Marital status * Education qualification * Default on credit status * Balance on an yearly average * Housing loan * Loan of personal nature * Contact mode * Day of the month last contacted * Month of the year last contacted * Duration of last contact * Campaign * pdays - days passed since last contact * previous - number of contacts before the campaign

Evaluation:

The client's (customer's) subscription towards a term deposit is identified using the model with the performance of the same evaluated using score metrics resulting in the following. * Naive Bayes: 1.0 * Linear SVM: 100.0 * K-Nearest Neighbor: 0.875

Discussion:

Based on the various algorithms used for the classification, we have identified that Naive Bayes and Lienar SVM seem to be perfect fit for our data set, while their is minor error involved when using K-Nearest Neighbor algorithm.

Question 6:

Workflow:

To accomplish the goal of perform K-means clustering: * Import appropriate libraries required to perform clustering of data

- Identify a data set and read the same in appropriate format

- Identify the features that are positively and negatively correlated to the target variable, leading to elimination of any features that may not be correlated or useful for the process, if needed

- Identify any null values associated with the features and either assign a specific value (for instance, 0) or remove them from the data set as considered appropriate

- Determine the clusters using graphical plots with the help of the correlated features identified

- Compute the optimal number of clusters to derive best performance of the model using Elbow method and present the same in a graphical form

- Pre-process the data set used to standardize (normalize) its distribution

- Perform K-means clustering using a set value of clusters and evaluate its performance (accuracy) using silhouette score

Parameters:

The key features used to build the clusters are: * Pregnancies * Glucose * Blood Pressure * Skin Thickness * Insulin * BMI * Diabetes Pedigree Function * Age

Evaluation:

The elbow method identified that 3 is the ideal number of clusters to be used to derive a model that performs the best. The performance of K-means clustering provides a score of 0.538 with a sample size of 40 is at its best to identify if the patient has diabetes or not.

Discussion:

When no more than 3 clusters are involved with a size of the data set being 40 provides the best result for the diabetes data set used.

Code Execution Video:

Code explained by Poonam

Code explained by Nikita Reddy

Conclusion:

The lab provided an excellent means to learn and master the basics of data analysis through classification and regression using both supervised and unsupervised approaches, along with the learning associated with several features such as dictionary, list and tuples associated with strings.