***Lab Assignment-1***

**Team Mates:**

Haritha Bheemineni - 8

Satya Anusha Prattipati - 18

Yamini Reddy Dontireddy- 28

**Overview:**

The assignments main agenda is to cover and learn the core python programming, machine learning concepts and also to use python in- built libraries such as Numpy, Pandas and sklearn.

**Moto***:* In order, to learn and cover the below topics:

1. Python Core concepts – Loops, Tuples, Dicts, String concepts
2. OOPS concepts (Inheritance, Web scraping)
3. Machine learning concepts (K-Means, SVM, KNN and NLP implementation etc)

**Tools/Software’s used:**

1. PyCharm,
2. Python3
3. Interpreter - Anaconda
4. Google Cloud

**Tasks:**

1)Return all possible subsets of a group that contains duplicates, numbers. Do now not encompass null subset.

**Method:**

1.Retrieve the input from the console and store the value as var(variable) and assign it to a list.

2. In a sequential order , check each and every character with the previous character right from the second element of string.

3. For the matching characters – add the result to the list.

4. Follow the above step, till you finish covering the entire string and finally display the subsets without null subset.

**Program Snippet :**

A screenshot of a computer

Description automatically generated

**O/P Snippet:**

A close up of a logo

Description automatically generated

2)Concatenate two dictionaries and sort the concatenated dictionary by value.

**Method:**

1**.**Take the static values in the program.

2. Dictionaries concept comes into picture - Ensure to create an empty Dictionary

3.Then concatenate the given input values.

4. Depending on the values, sort the dictionaries.

**Program Snippet :**

A screenshot of a computer

Description automatically generated

**O/P Snippet:**

A screenshot of a computer

Description automatically generated

3)Creation of an Airline booking Reservation system

**Method:**

Created the below classes:

***“Flight ”*** – Using display\_flight function, displayed the airline details.

***“Employee”*** - Using display\_employee function, displayed the employee details.

***“Passenger”*** – To input the passenger data (dynamic data).

***“Baggage”-*** To input the baggage data (dynamic data).

***“Fare”*** – Inheritance concept comes into picture here in order to access the Parent classes Passenger and Baggage. The fares has been calculated and finally the ticket has been displayed.

**Program Snippet:**

A screenshot of a computer

Description automatically generated

**O/P Snippet:**

A screenshot of a computer

Description automatically generated

4) Fetch the clustering algorithm comparisons from the below link.

<https://catalog.umkc.edu/course-offerings/graduate/comp-sci/>

**Method:**

1. Import all the required libraries such as (Beautiful soup)
2. Web scrapping concept comes into picture.
3. Load the extracted output into .csv file.

**Program Snippet:**

A screenshot of a computer

Description automatically generated

**O/P Snippet:**

A screenshot of a computer

Description automatically generated



*5)*Apply 3 classification algorithms Naïve Baye’s, SVM and KNN on the particular datasets post EDA operations and show the comparisons among classifiers the chosen data set after performing EDA and report which classifier gives a better result

**Method:**

1. Import pythonlibraries for data cleaning - Pandas and Numpy.
2. Data is analyzed and handled the null values.
3. Now, apply various classification algorithms in order to build a model and to predict the output.
4. You can now calculate the score of all the algorithms and looking at the output values, you can come to a conclusion of the good classifier.

**Program Snippet:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**O/P Snippet:**

A screenshot of a cell phone

Description automatically generatedA screenshot of a computer

Description automatically generated

**Final analysis of the output:**

The model that’s built on SVM holds the high accuracy score than the rest of all algorithms.

6) Visualize the clusters using matplotlib or seaborn post K-Means method implemation on the dataset.

a. Using the elbow method , show which K is best.

b. Evaluate with silhouette score or other scores relevant for unsupervised approaches (before applying clustering clean the data set with the EDA.

**Method:**

1. Import the Python libraries- Pandas and Numpy.
2. Perform Exploratory Data Analysis .
3. In order to use the elbow method, Identify the number of clusters.
4. Use Kmeans algorithm to build model.
5. Finally, calculate the silhouette score evaluate the model.

**Program Snippet:**

*A screenshot of a computer

Description automatically generated*

**O/P Snippet:**

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated

**Final Analysis:**

Using elbow method , the K value is identified as 3 for the given dataset.

*7)A*pply tokenization, lemmatization techniques, find all the trigrams and Extract the top 10 of the most repeated trigrams based on their count of the given dataset.

Go through the text in the file, find all the sentences with the most repeated tri-grams, extract those sentences and concatenate those and print the concatenated result.

**Program Snippet:**

A screenshot of a computer

Description automatically generated

**O/P Snippet:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a social media post

Description automatically generated

*8)* For a given dataset, create a mutile regression and evaluate the model RMSE and R2 . Also, report the observations(PRE and Post EDA improvements)

**Method:**

1. Import the python libraries – Numpy and Pandas
2. Then, a model is built using multiple regression and calculate the RMSE and R2 score on application of EDA.
3. Report Pre and post EDA improvements.

**Program Snippet:**

A screenshot of a computer

Description automatically generated

**O/P Snippet:**

A screenshot of a social media post

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Final Analysis:**

Post Null values removal and dropping the non-correlated features, the RMSE score is decreased . But, the R2 score is slightly elevated(increased).