Python Programming

CS5590 APL

Done by

Naga Srividya Varanasi (Class ID-22)

Chandra Sekhar Pentakota (class ID-27)

Pujitha Koppanathi (class ID-12)

Github Link:

Youtube Link:

**Introduction:**

This report is about python programming using several concepts in python and machine learning.

**Objective:**

The main goal of the project is to get hands on experience on machine learning concepts like performing exploratory data analysis on the dataset like handling null values, along with the classification algorithm like Naïve Bayes, SVM and KNN, and also multiple regression techniques and finally evaluating model using RMSE and R2 score.

**Question1:**

**Approach:**

a)Picking a dataset and performing exploratory data analysis.

b) Applying the Naïve Bayes, SVM and KNN classification algorithms.

**Workflow**: It consists of source code along with output and graph.

**Source Code:**

1)Reading data from text file.

A screenshot of a social media post

Description automatically generated

2) displaying first few rows of data.

A screenshot of a social media post

Description automatically generated

**3)**plotting the graph.

A screenshot of a social media post

Description automatically generated

4) Applying different classification techniques.

A screenshot of a cell phone

Description automatically generated

**5)** Applying Kneighbours and guassian NB.

A screenshot of a social media post

Description automatically generated

6) Comparison between KNN, Naïve Bayes and SVM

A screenshot of a cell phone

Description automatically generated

**Comparison:**

As compared to KNN which is giving 50%, Naïve Bayes and SVM are better as they give better accuracy score of almost 60%.

**Question2:**

**Approach:**

a) Performing EDA on the dataset and applying K means on the dataset.

b) And evaluating silhouette score.

**Workflow:** It consists of source code and the output.

**Source Code**:

1)Reading dataset and describing.

A screenshot of a cell phone

Description automatically generated

A close up of a piece of paper

Description automatically generated

2)plotting heatmap to find correlation

A screenshot of a social media post

Description automatically generated

A close up of text on a white background

Description automatically generated

3)plotting different graphs

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

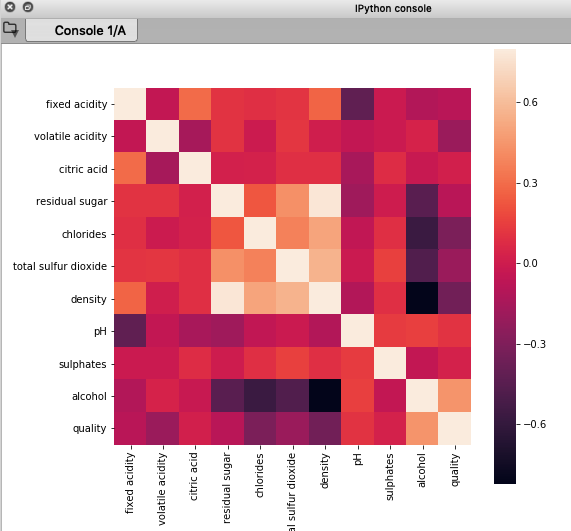
Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated



A screenshot of a cell phone

Description automatically generated

4) Silhouette score

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a piece of paper

Description automatically generated

**Question3:**

**Approach**:

1. Read the data from file

2. Tokenize text into words and apply lemmatization technique word.

3. Find all trigrams for words.

4. Extract top 10 of the most repeated trigrams based on count.

5. Go through text in file

6. Find all sentences with most repeated tri-grams

7. Extract those sentences and concatenate

8. Print concatenated result

**Workflow**: It consists of source code as well as output.

**Source Code:**

**A screenshot of a social media post

Description automatically generated**

A screenshot of a social media post

Description automatically generated

A close up of a newspaper

Description automatically generated

Lemmatizing:

A close up of a newspaper

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Trigrams frequency in ascending order.

A close up of a newspaper

Description automatically generated

The concatinated result:

A picture containing text, newspaper, building

Description automatically generated

**Question4:**

**Approach:**

**1.**To create multiple regression on the dataset

2. evaluate the model using RMSE and R2

**Workflow**: It consists of source code as well as output screenshots.

**Source Code**:

1. Reading the data and finding the skewness of the data.

A screenshot of a cell phone

Description automatically generated

b)Before and after applying log on skewness

A screenshot of a cell phone

Description automatically generated

c)Finding nulls and correlation on the dataset

A screenshot of a social media post

Description automatically generated

d)finding outliers

A screenshot of a cell phone

Description automatically generated

e)evaluating using RMSE nad R2 after performing EDA

A screenshot of a cell phone

Description automatically generated

f)Before performing EDA

A screenshot of a cell phone

Description automatically generated

**Analysis:**

RMSE is good after applying EDA on the dataset, and R2 is better before applying EDA.