**PROJECT-EXAM 1 REPORT**

**Introduction**

This Project Exam consists of 5 programs that covers working with basic Python concepts such as sets, dictionaries, loops, list, tuples, OOPS concepts, along with Machine Learning concepts such as classification (k-NN, Naïve Bayes, SVM,..), clustering(K-Means,..), regression models (linear and multiple), PCA, Natural language processing and also deals with packages such as Numpy, BeautifulSoup, Scikit-learn, Nltk, matplotlib, Seaborn, pandas etc.

**Objective**

The objective of this project is to gain basic understanding of Python and Machine Learning concepts in both supervised and unsupervised techniques and apply this knowledge so as to get optimal solution for each program that requires to do the following:

1. The objective of first task is dataset manipulation and a supervised method classification of our choice.
2. The objective of second task is of applying a K-means unsupervised method to a given set of data.
3. The objective of this task is to analyse the dataset and model, analyse it.
4. Cleaning the text data of the given dataset and applying the techniques for transforming the text data into numeric format using Count Vectorization and Tfidf Vectorization followed by evaluating the model and interpreting the result.
5. The objective of the task is to perform exploratory data analysis on any dataset. Applying Naïve Bayes, SVM and KNN classification methods and find the accuracy. Also applying linear SVM and non-linear SVM and find the accuracy.

**Workflow**

The workflow is common across all the programs:

* Importing the required Libraries
* Reading the Dataset
* Data Pre-processing
* Model Creation
* Model Execution
* Model Evaluation

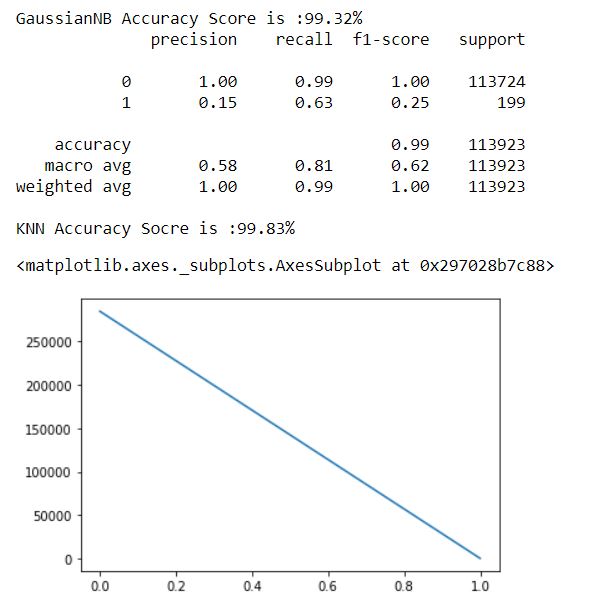
**Approaches**

**Task-1:**

**Implementation:**

* In this program, initially we have loaded the given data set and created data frame by using pandas library.
* Then sliced the target column i.e., ‘class’ of dataset into “y” and remaining other columns into “x”.
* Now we have split the dataset into train and test by using train\_test\_spilt method with test size 0.4 and with the random\_state=0.
* For implementing Naive Bayes algorithm created GaussianNB() object.
* By using x\_test data we found predictions.
* Calculated the accuracy score for the test data, evaluated the model.
* Generated the classification report for the predicated data and for test data.
* Then applied KNN classification algorithm.
* Now calculated the score for the test data.
* By using matplot library we have visualized the data using line graph.





**Problems with Imbalanced Datasheet:**

When dealing with imbalanced dataset, the core problem is that ML algorithms generate inaccurate information. That's because these other algorithms show preference and toward the class of the large percentage. ML algorithms are not going to find minority class because they are in the dataset much less.   Understand that we have 1%  minority class but also 99 % majority class data then ML algorithms classify them all correspond to majority class.

**Data approach to Handle Imbalanced Datasheet:**

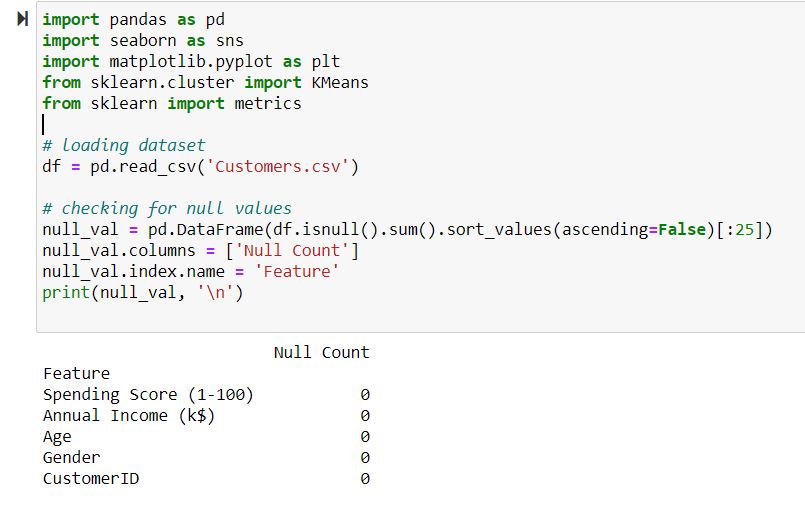
It comprises of resampling of the data to minimize the impacts of class imbalance. The method to data has received tremendous acknowledgement among practitioners because it's much more flexible to allow the use of the genetic algorithm.  Over-sampling and under-sampling are the two most popular techniques.

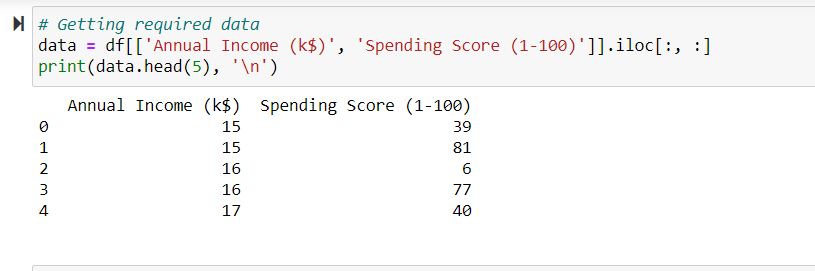
* Over-sampling raises the number of members of the minority class in the course sample. The benefit of over-sampling is that no data from the training samples is lost, since all minority and significant percentage class observations are maintained. By contrast, it's prone to overfitting.
* Under-sampling, unlike over-sampling, seeks to reduce the number of predominant samples to match the allocation of the groups. Since it eliminates findings from the original data set, valuable information can be discarded.

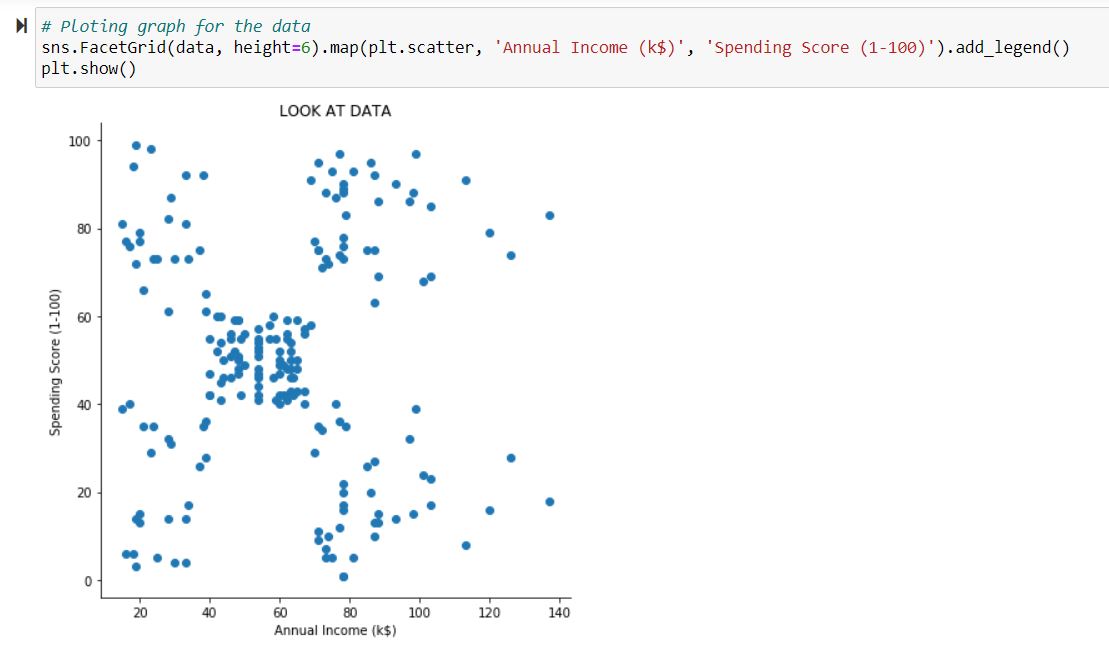
**Task-2:**

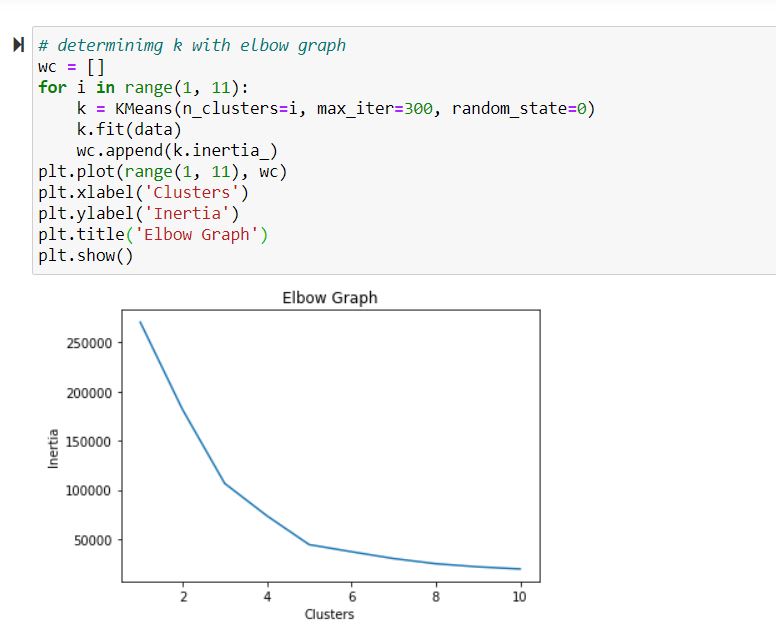
**Implementation:**

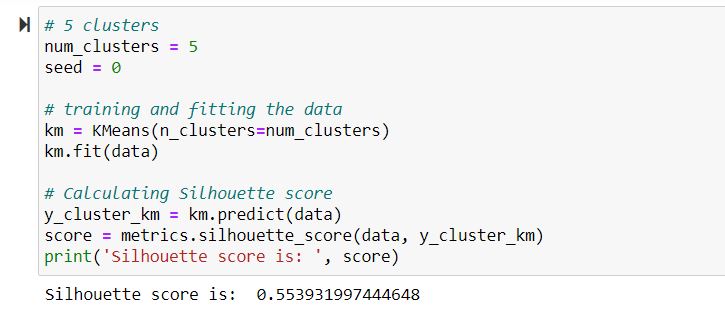
* Initially we have loaded the given data set and created data frame by using pandas library.
* Then checked the data for null values of which there were no null values.
* After that required data is taken from the entire dataset.
* Then plotted a graph for visualizing the data which was created in the above step.
* For determine the number of clusters elbow graph was created and displayed.
* The data showed 5 clusters would be optimal so then the data was trained with 5 clusters.
* Then the silhouette score was calculated as well as a final scatter plot with the clustered data is displayed.

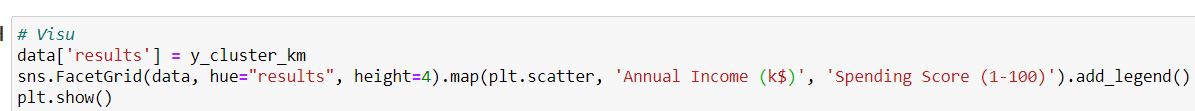


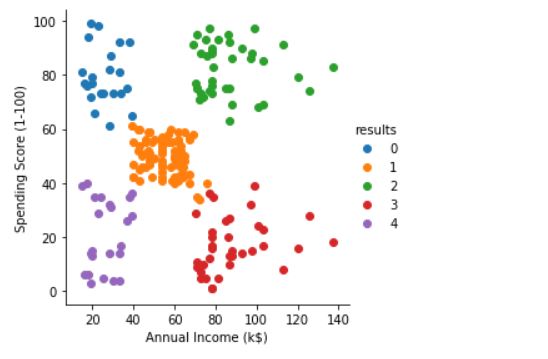








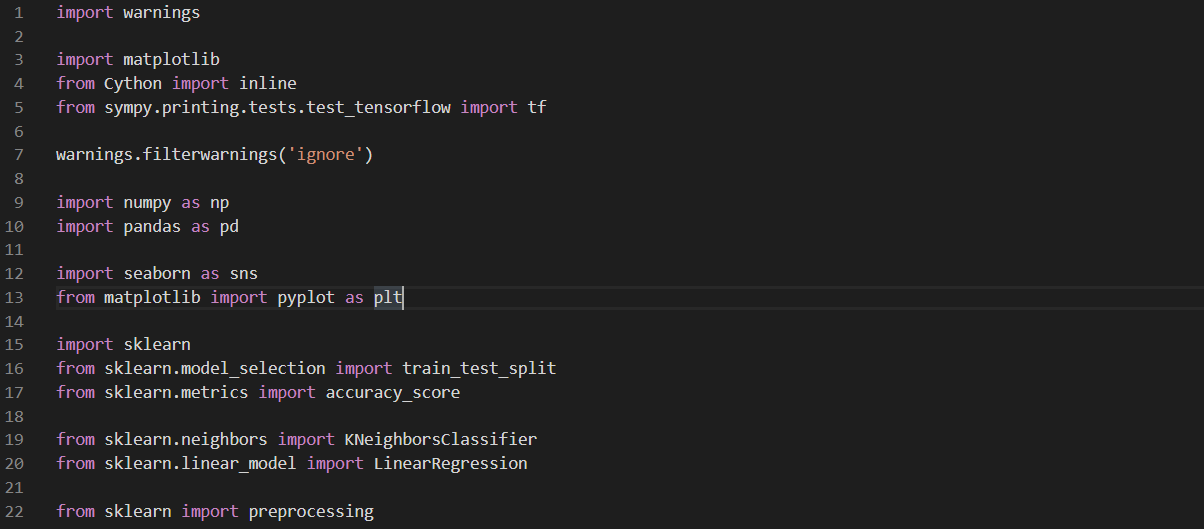


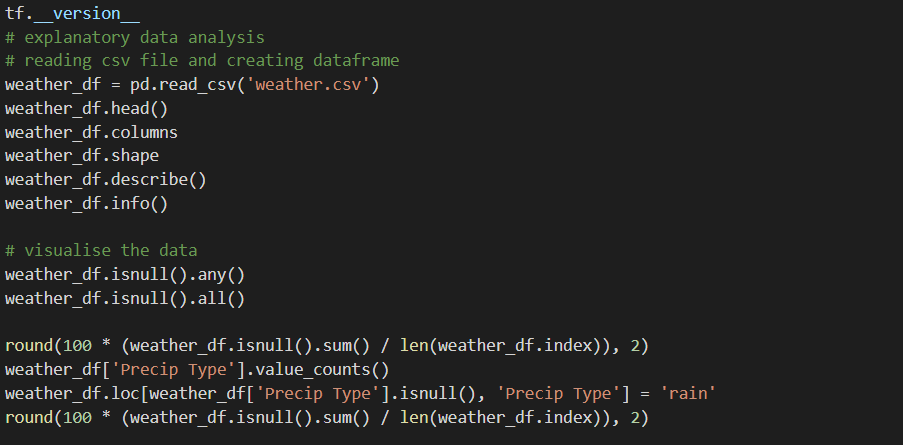


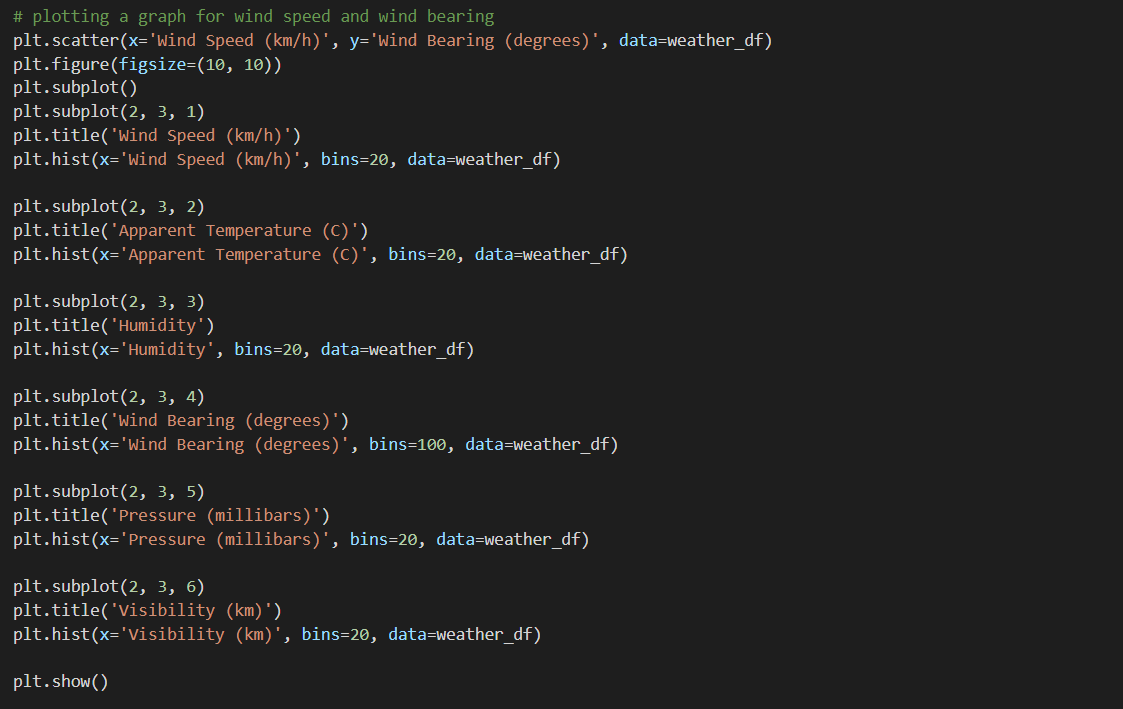
**Task-3:**

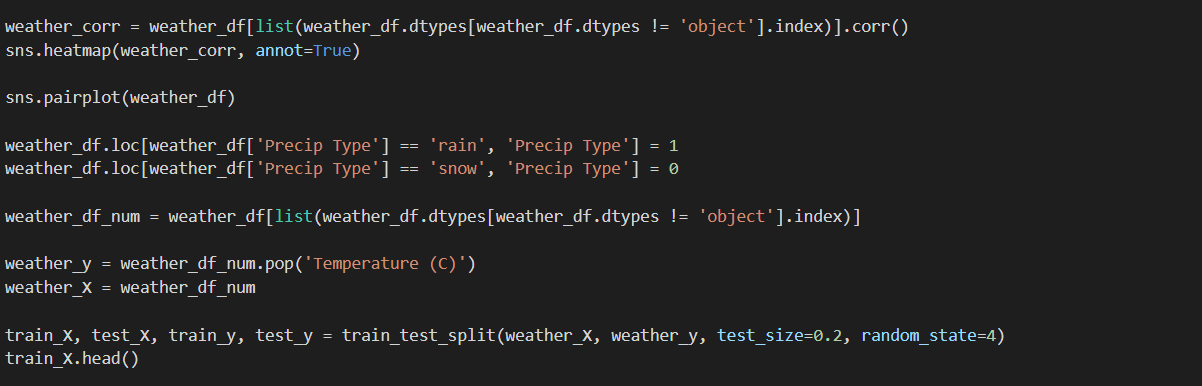
**Implementation:**

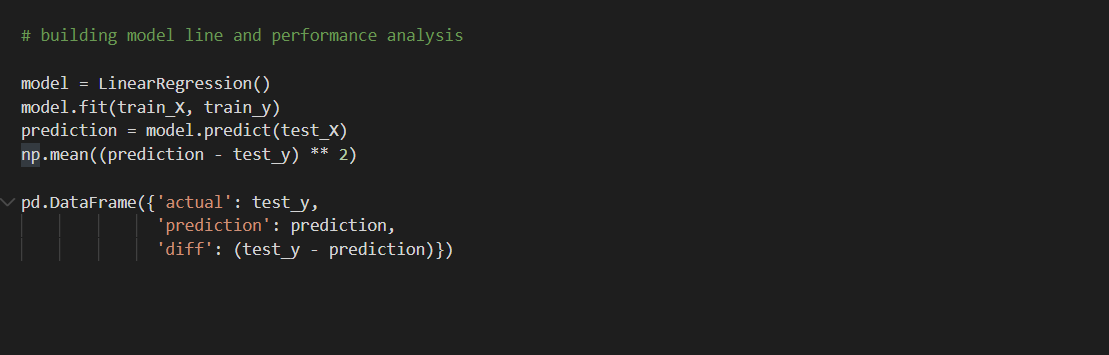
The objective of this task is to analyse the given weather dataset, hourly temperature in that data. And also, build the model and performance analysis is done. Here we have plotted the graph with different parameters.

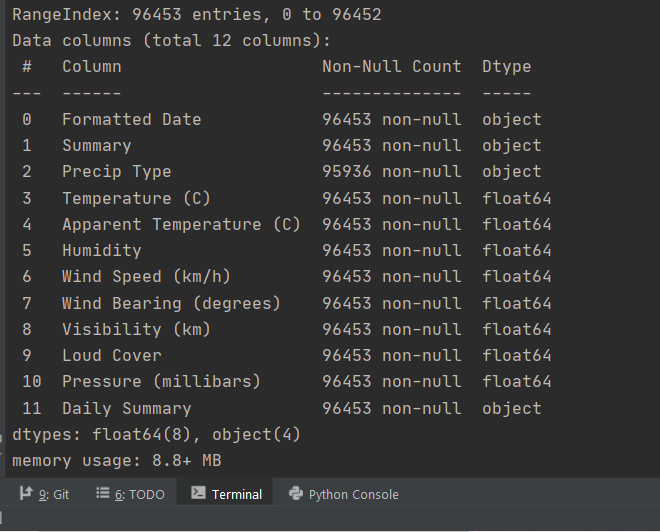


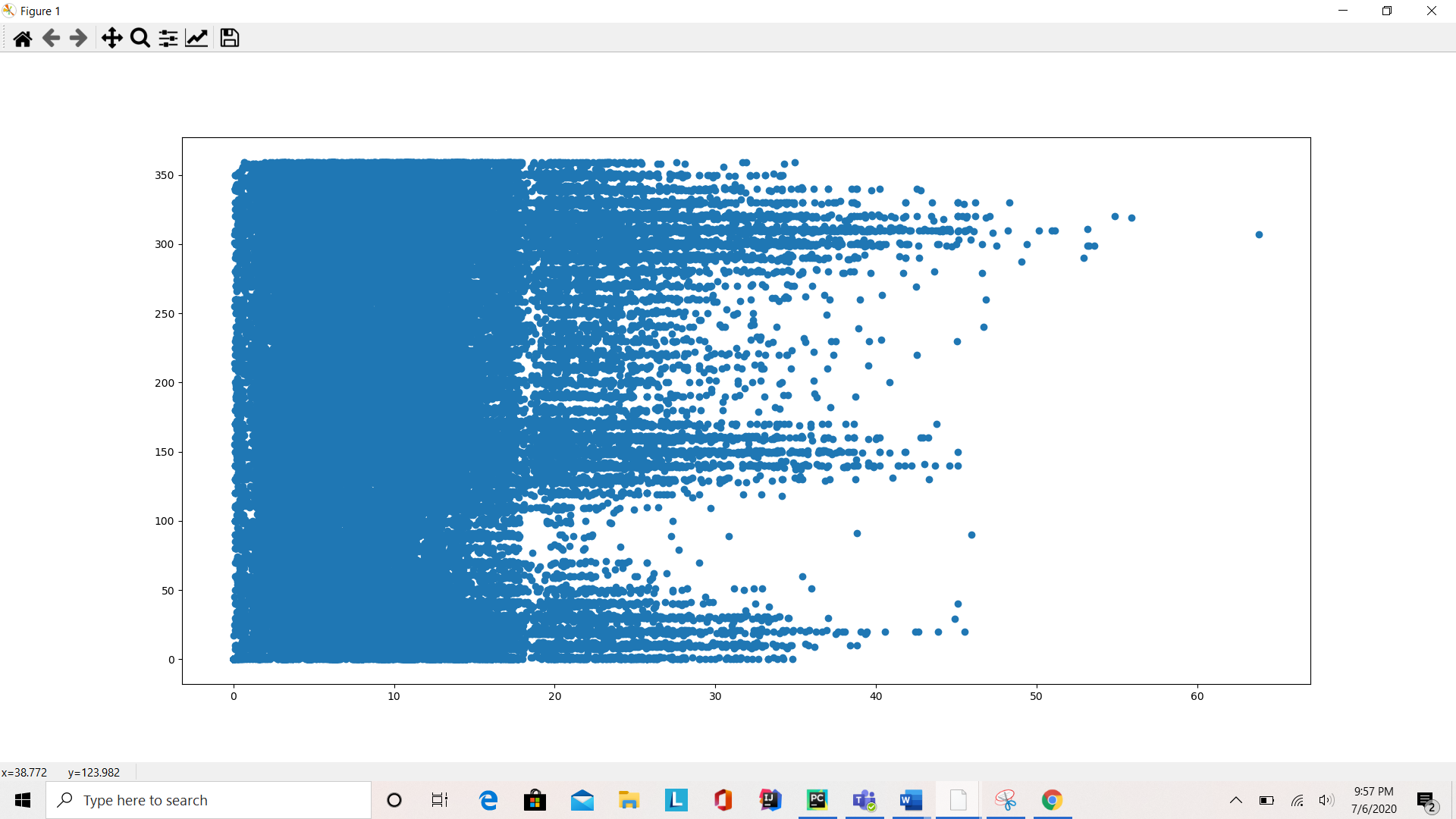


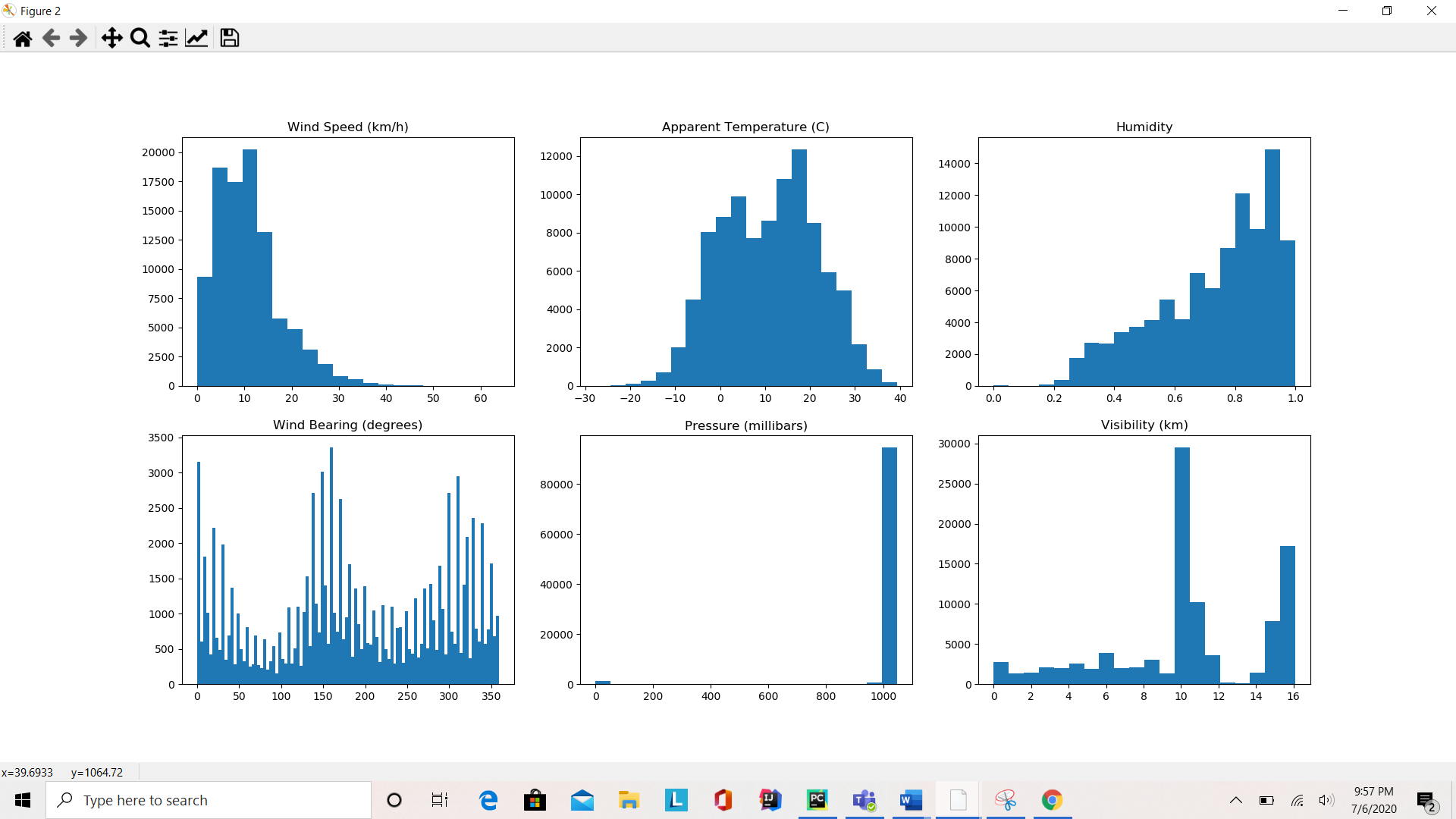












**Task-4:**

**Implementation:**

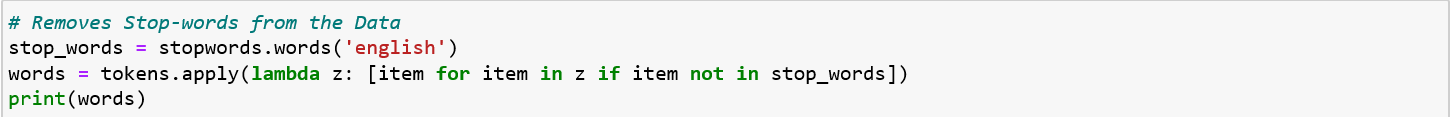
* In this program, first we read the dataset and split it accordingly with target feature in ‘y’.

A screenshot of a cell phone

Description automatically generated

* For data cleaning or pre-processing, word tokenisation using Regular Expression is performed which also removes punctuations that is followed by removing stop words and lemmatization. The output is as shown below:



A screenshot of a social media post

Description automatically generated

* Next, we apply the Tfidf and Count Vectorization techniques individually to the train set and test set that been achieved from cleaned data and y. The model is then evaluated and the accuracy scores are interpreted.

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

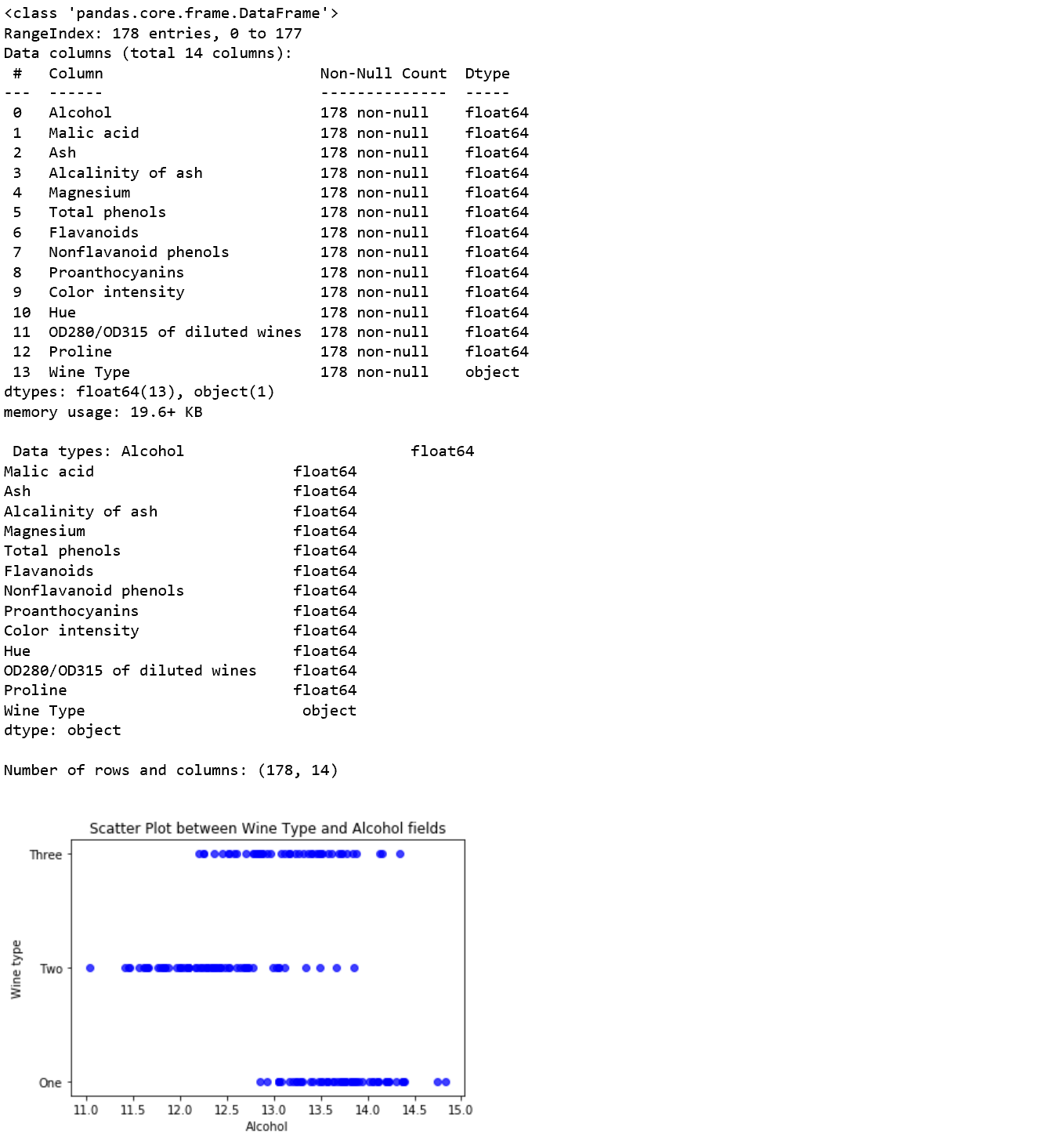
Description automatically generated

**Task-5:**

**Implementation:**

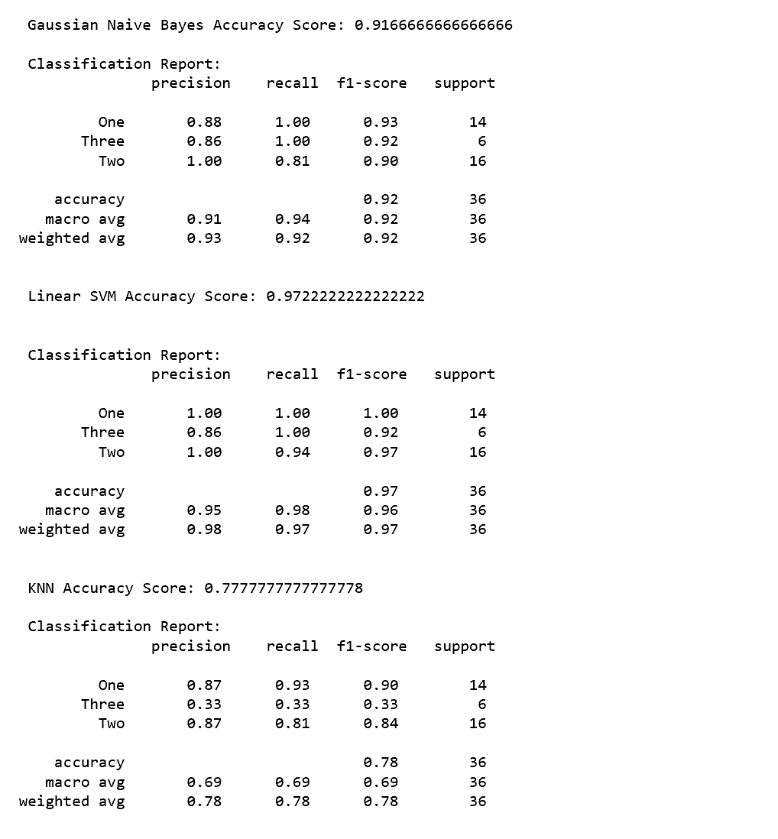
* In task (a), first loading the dataset using pandas library. As part of Exploratory data analysis,
  + Displaying the csv dataset information using info().
  + Finding the total number of rows and columns in the dataset using shape().
  + Assigning alcohol column to x frame and wine type column to y frame and applying the scatter plot and show the graph to analyze on this two columns.





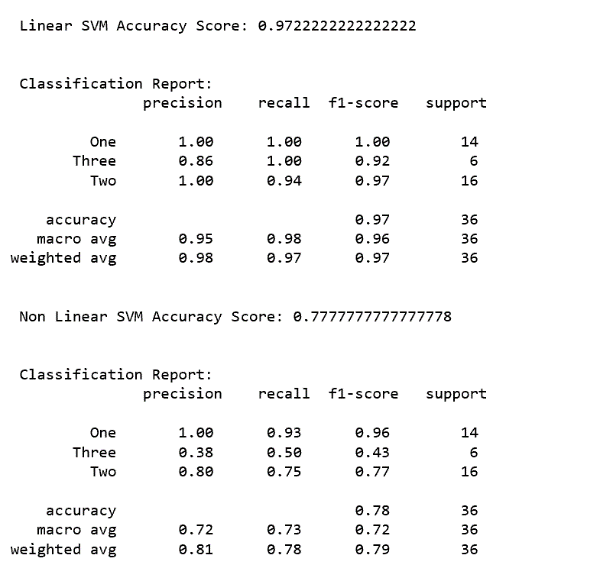
* In task(b), loading the data set using pandas library.
* Check the null values against the columns in the data set and remove the null values.
* Pre-processing the data into x and y sets.
* Splitting data into training data and testing data.
* From Scikit learn library, creating Gaussian Naïve Bayes model and fitting the data to the model. Predicting the data. Using metrics, finding the Naives Bayes classifier accuracy and the classification report.
* From Scikit learn library, creating linear SVM model and fitting the data to the model. Predicting the data. Using metrics, calculating the SVM classifier accuracy and the classification report.
* From Scikit learn library, creating KNN model (n\_neighbours = 5) and fitting the data to the model. Predicting the data. Using metrics, calculating the KNN classifier accuracy and the classification report.





* In task (c), loading the data set using pandas library.
* Checking the null values against the colums in the data set and removed the null values.
* Pre-processing the data into x and y sets.
* Splitting data into training data and testing data.
* From Scikit learn library, creating linear SVM model(kernel = linear) and fitting the data to the model. Predicting the data. Using metrics, finding the linear SVM classifier accuracy and the classification report.
* From Scikit learn library, creating non- linear SVM model(kernel = rbf) and fitting the data to the model. Predicting the data. Using metrics, calculating the non linear SVM classifier accuracy and the classification report.





**Datasets**

1. <https://umkc.app.box.com/s/6cnan5zesgntmsxzlgbjwjoiii6fywmj>
2. <https://umkc.app.box.com/s/lwr3s70prbe3tdifzx6fblq7wqln54xq>
3. <https://umkc.app.box.com/s/de9bdscv3ys9ff1sjogbfzppzy1o477a>
4. <https://umkc.app.box.com/s/mlmyznn0667tbgui5urnzbos1i1c48qf>
5. <https://www.mldata.io/dataset-details/wine/>

**Evaluation**

1. From Task 1, it is observed that the accuracy score of KNN algorithm is higher than Naive Bayes Algorithm.
2. In Task 2, we have calculated silhouette score i.e., 0.55 and cluster our customers into buying groups based off of their Annual Income and Spending Scores
3. For Task 3, We will be seeing a gradual increase in the graphs.
4. From Task 4, it is observed that the accuracy score of Count Vectorization is higher than that of Tfdif Vectorization.
5. From Task 5, In first part for the dataset we have shown the data, displayed the data types for each column and applied scatter plot. In second part, we have found that accuracy of SVM is greater than KNN, KNN is greater than Naïve Bayes i.e SVM has the highest accuracy and Naïve Bayes has the least accuracy. In third part, linear SVM accuracy is higher than non linear SVM accuracy.

**Conclusion**

We have successfully applied the concepts and obtained the required accurate outputs to all the programs. This project Exam has enabled us to gain deeper understanding on the working of these Python and Machine Learning concepts in practice.

**Submission By:**

**Team 4-**

1. **Pooja Reddy Gopu – 17**
2. **Tejaswi Guturu – 18**
3. **Mounika Rachamadugu – 37**
4. **Pujitha Sakhavarapu - 40**