**Python**

**Project-1**

**Team-2**

**Introduction:**

To analyse how machine learning helps in predicting the occurrence of events. To understand the Concepts of regression,clustering,classification,NLTK features.

**Objective:**

To design the models and to calculate their efficiencies by calculating the scores and to perform the comparison of which model suits the best.

**Approaches/Methods**

1.The different models or approaches used are

a.SMOTE library to balance the unbalanced dataset.

b.SVC,KNN,Naïve Bayes prediction scores and obtaining calculation reports

c.Silhouette\_score calculation.

d.Plotting or visualising the graphs using pandas or seaborn.

e.Analysing the SVM models using linear and non linear kernels.

d.NLTK features like tokenization,vectorization,DecisiontreeClassifier, MLPClassifier

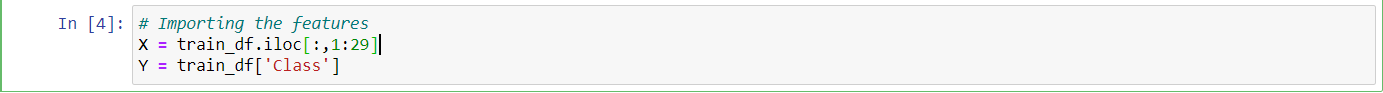
**Workflow:**

**Question 1:**

**a.Analysis of Data**

1.Import the creditcard dataset.

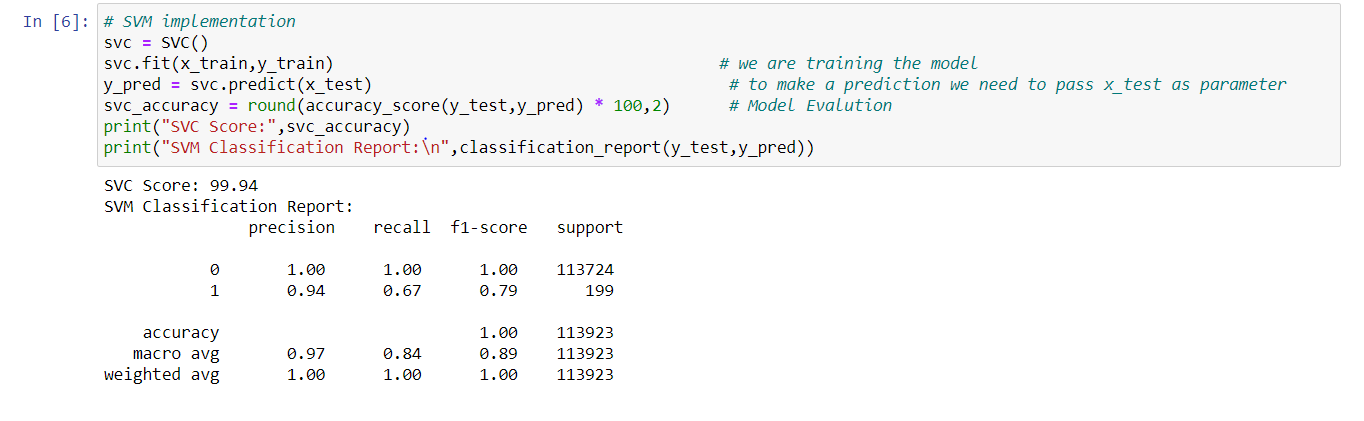
2.Loading the features to x and y.



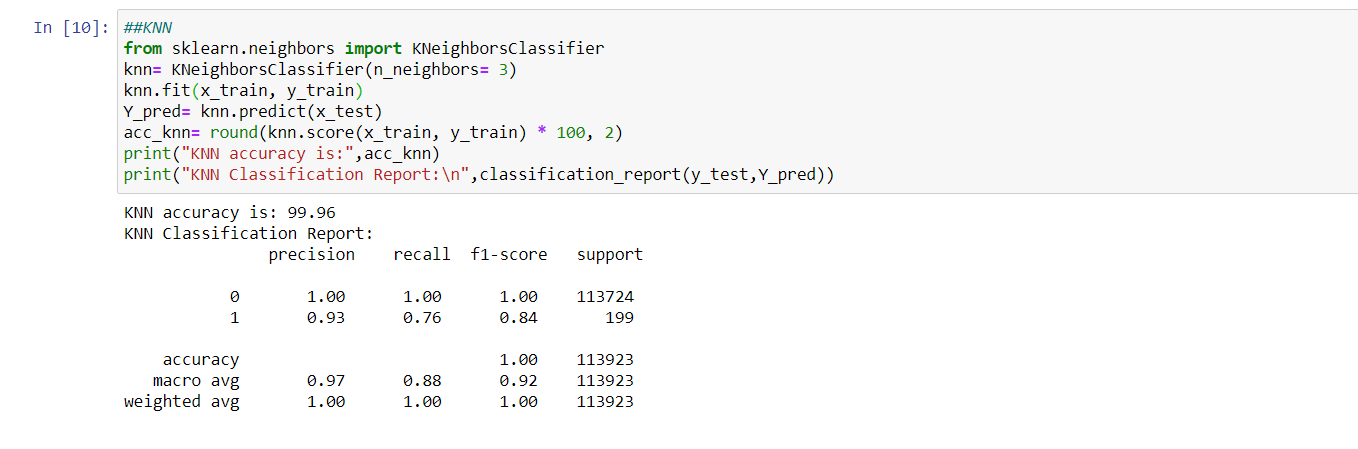
3.Creating train test split



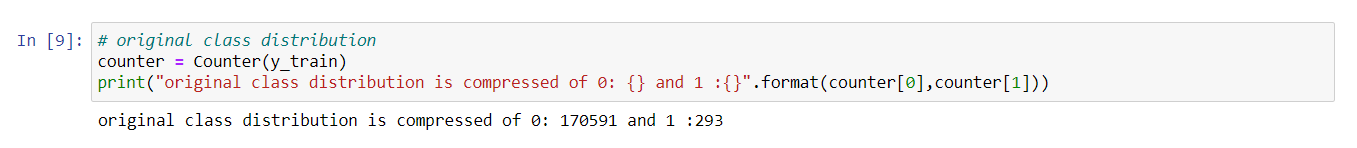
4.SVC analysis



5. KNN analysis



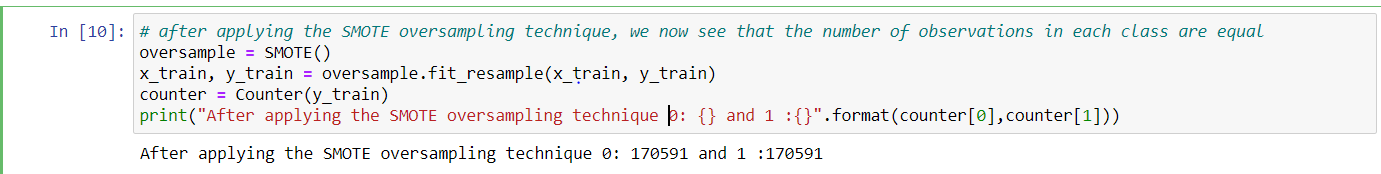
**b) visualising the samples of data.**



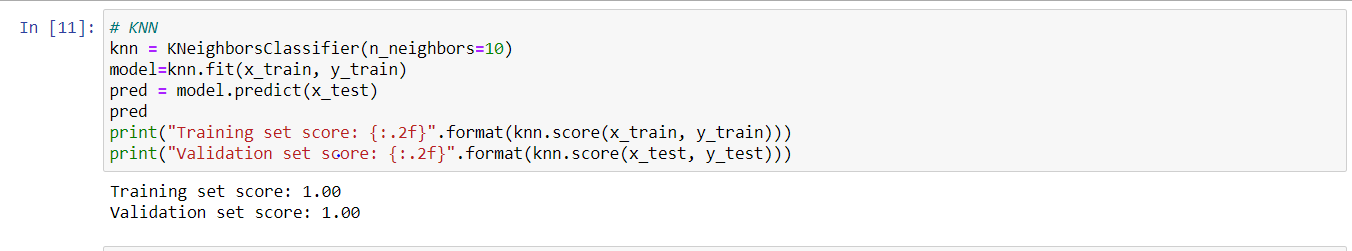
**Inference:**

Here the number of samples of 1 are negligible when compared to zero hence converting them with the help of SMOTE balancing technique.

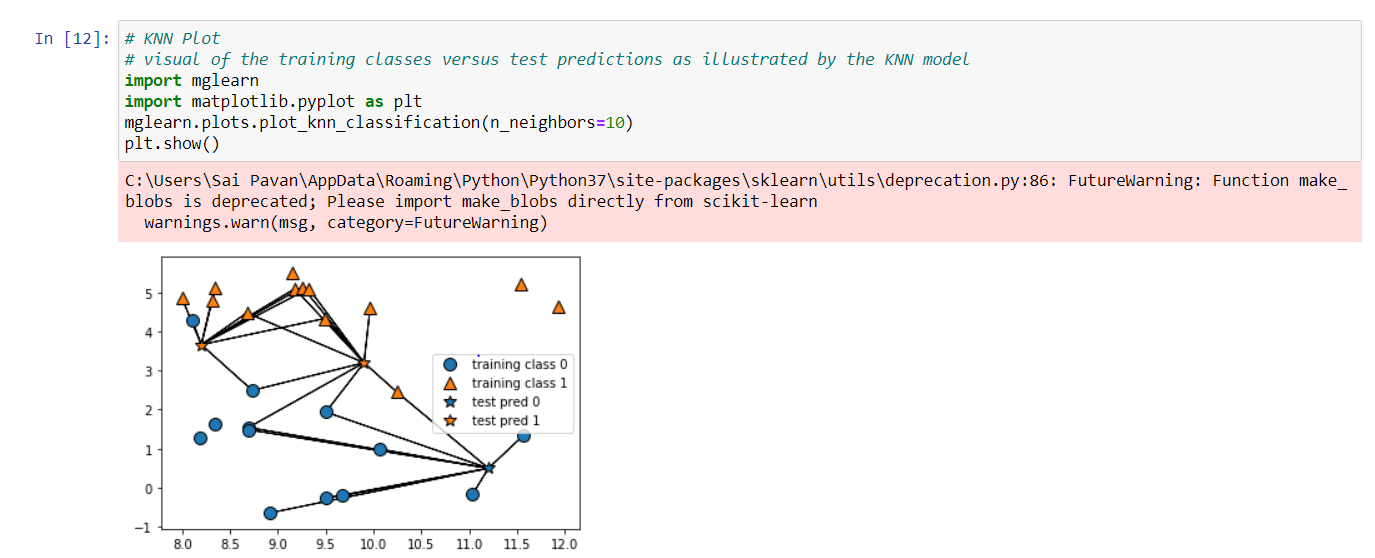
6.Here there are very few values for 1 class. Hence applying SMOTE oversampling technique.to visualize the samples.



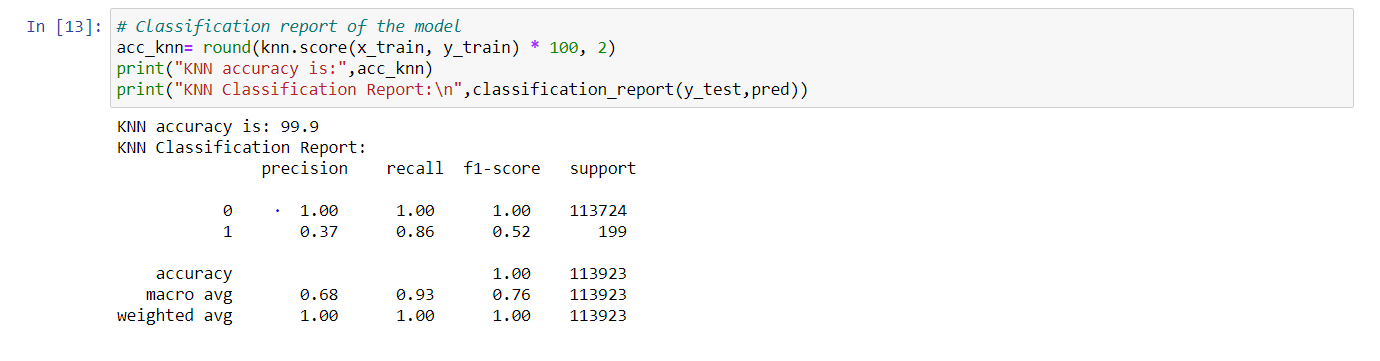
7.Creating new model



8.Plotting training vs test predictions



9.Calculating KNN accuracy



**Inference:**

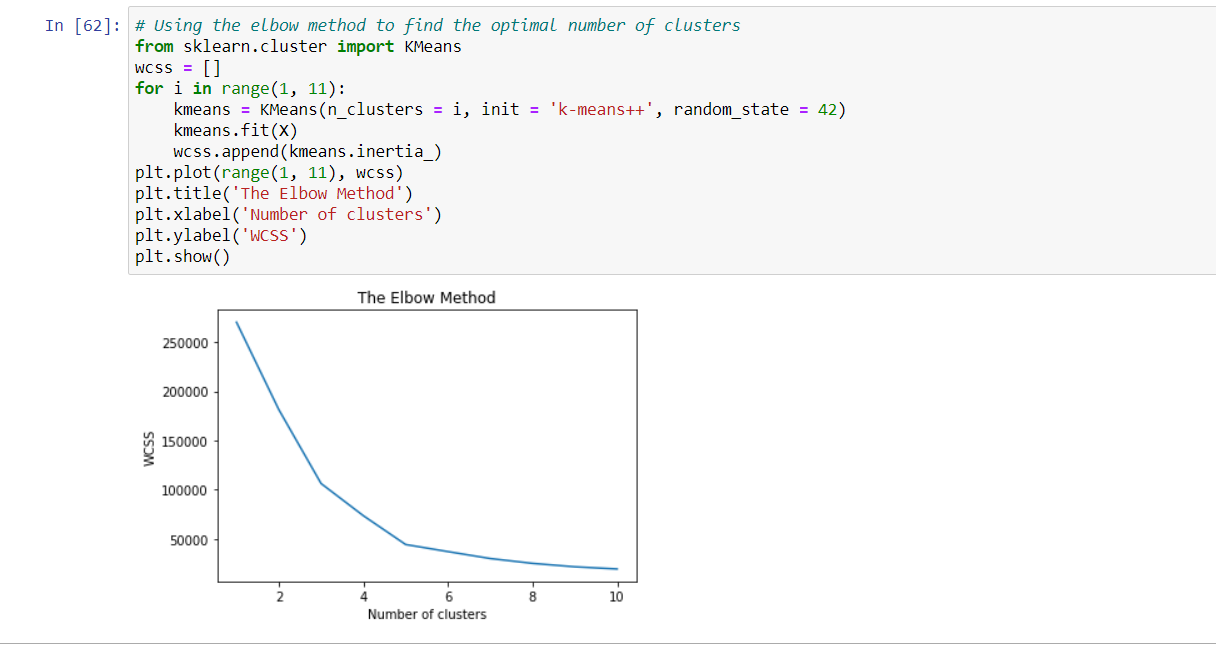
After balancing the dataset and training the same the accuracy can be improved.

**Question 2:**

1.Import the customer dataset.

**a.Report which K is the best using the elbow method.**

2.Calculate the k using the elbow method.

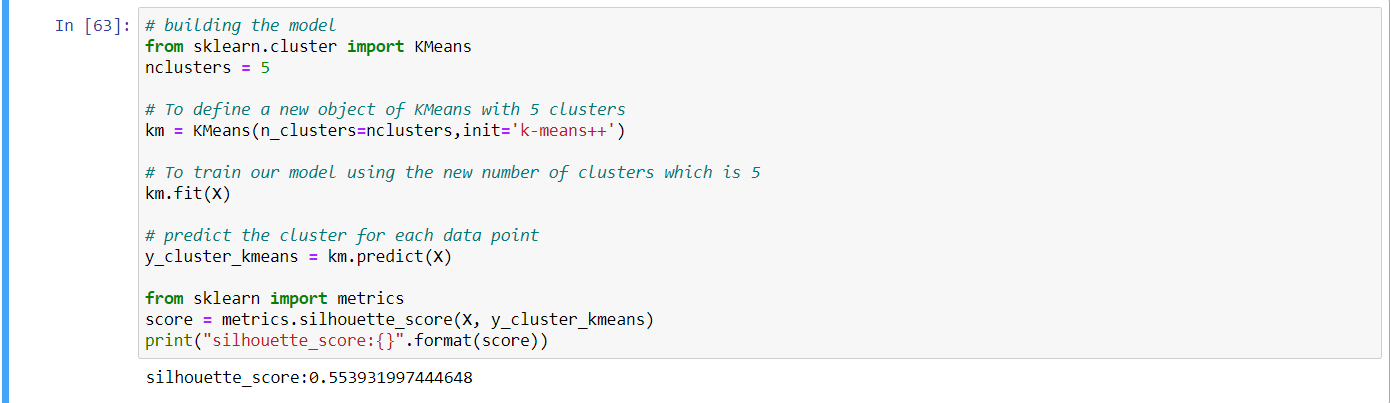


**Inference:**

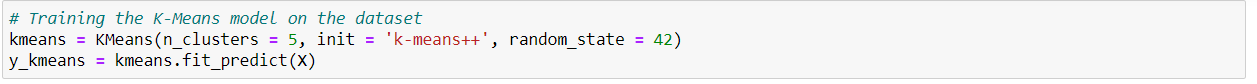
From the elbow method k value is identified as 5.

In case if the k value is taken as 3, the score is calculated as 0.4.As the K value changes to 5 the score increases which indicates that the model is near to the fit line.

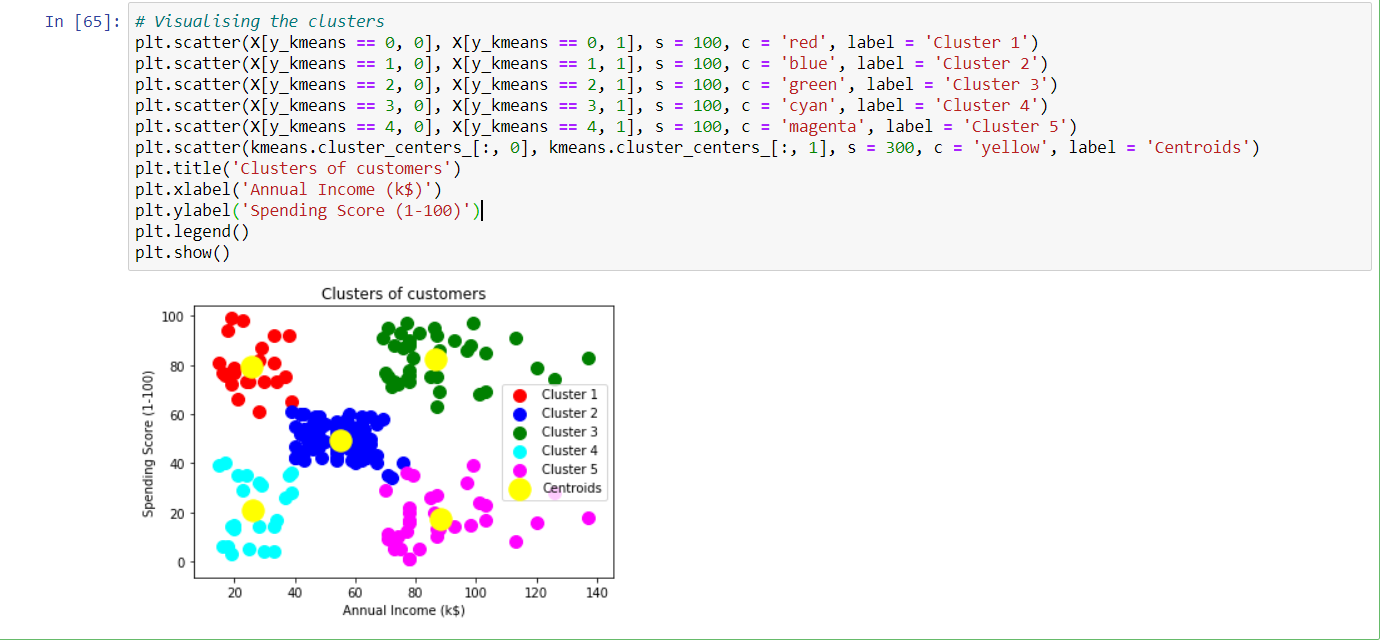
**b. silhouette\_score calculation**



**4.Training the model**



**5.Visualising the clusters**

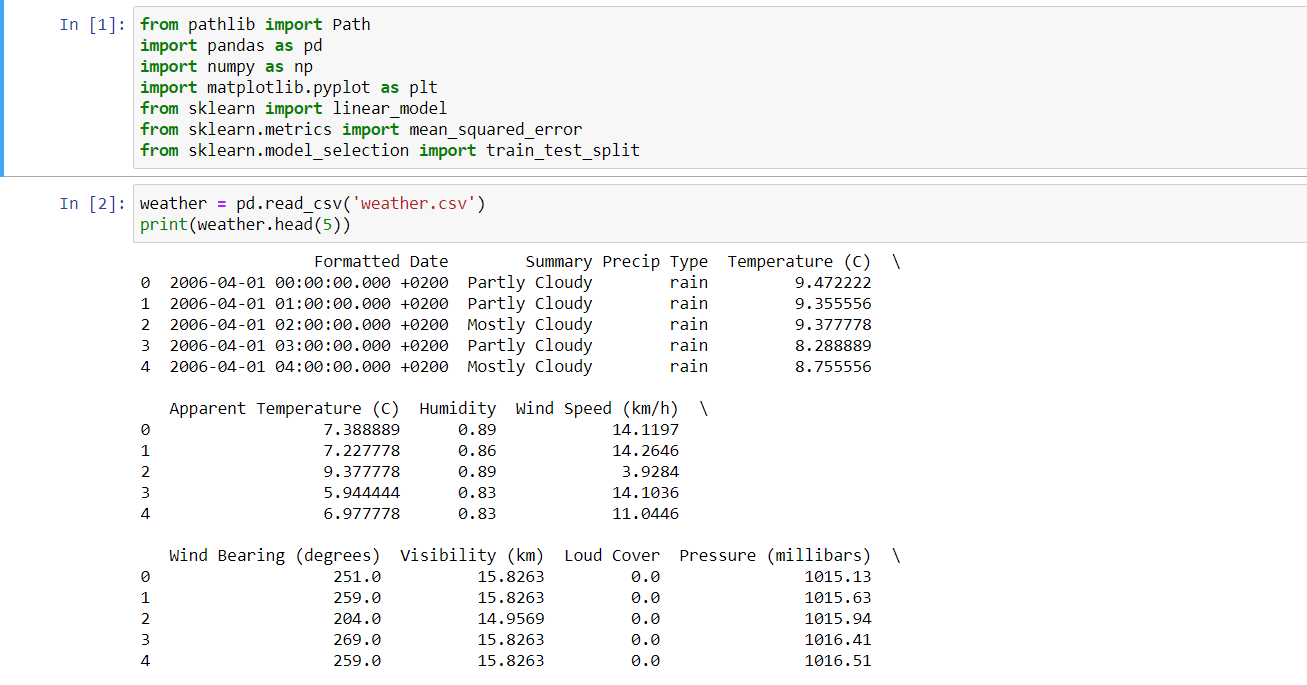


**Question 3:**

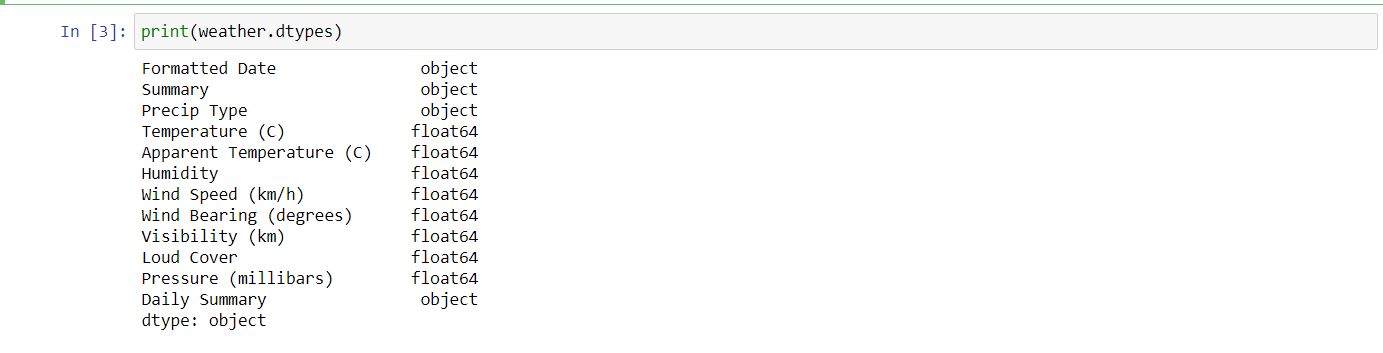
1. **Apply some Exploratory Data Analysis to draw some insight from the data,**

**Visualize the data and draw the model line**

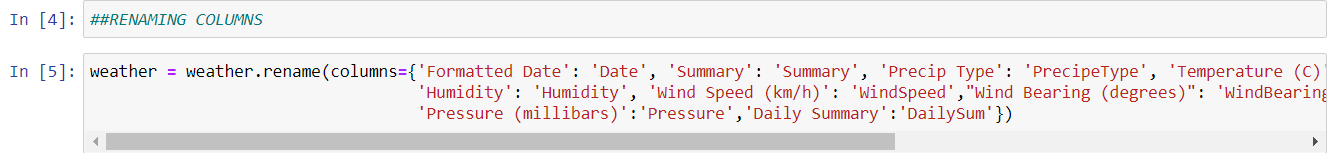
1.Make the necessary imports and load the data.



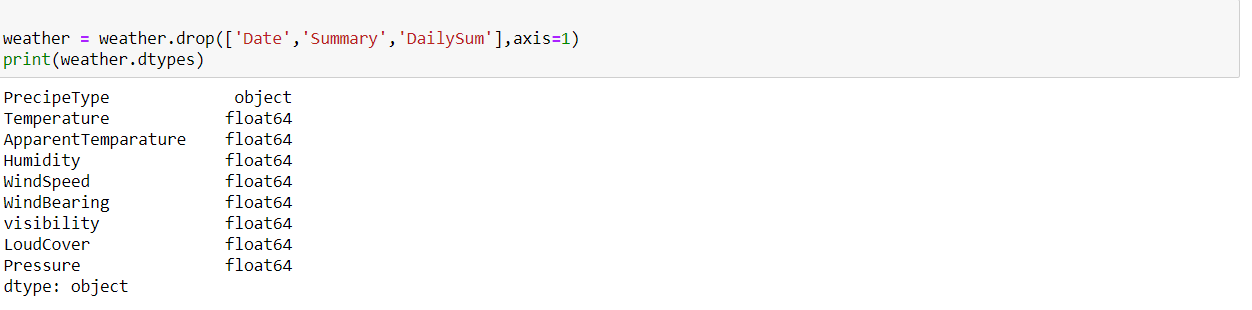
2.Checking the types.



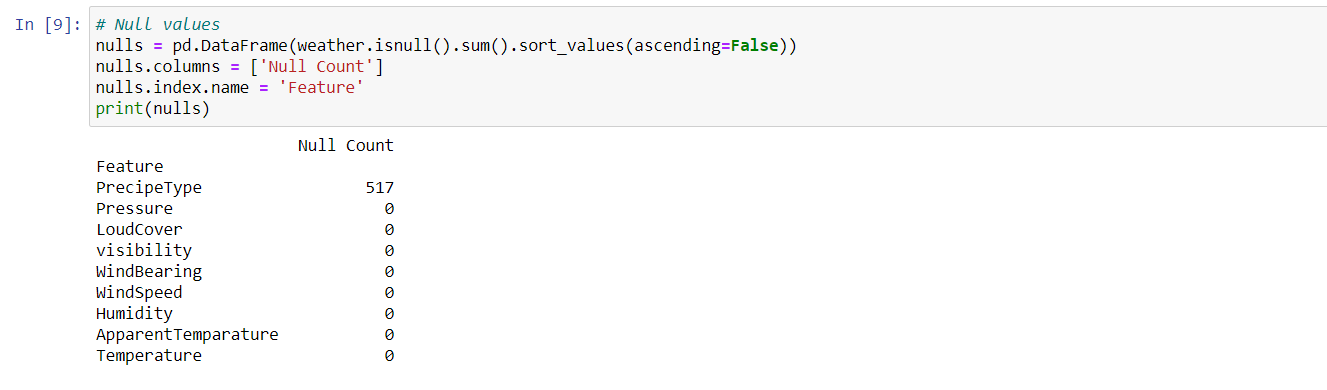
3.Renaming the columns.



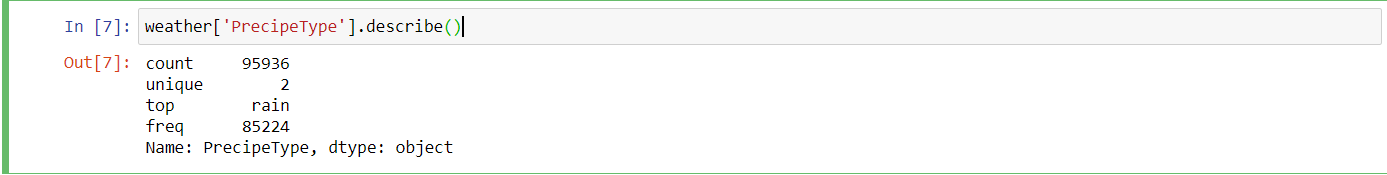
4.Dropping the unnecessary columns.



Printing null values



5.Checking the top value in precipetype.



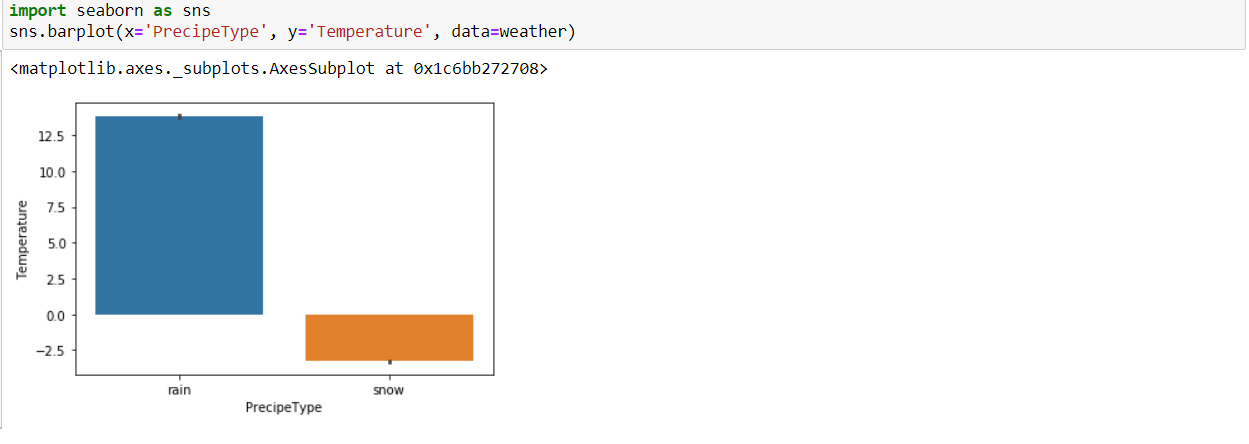
**Inference:**

As the top value is rain and there are 95936 counts and only 517 nulls.The nulls will be replaced by rain.

6.Handling null values



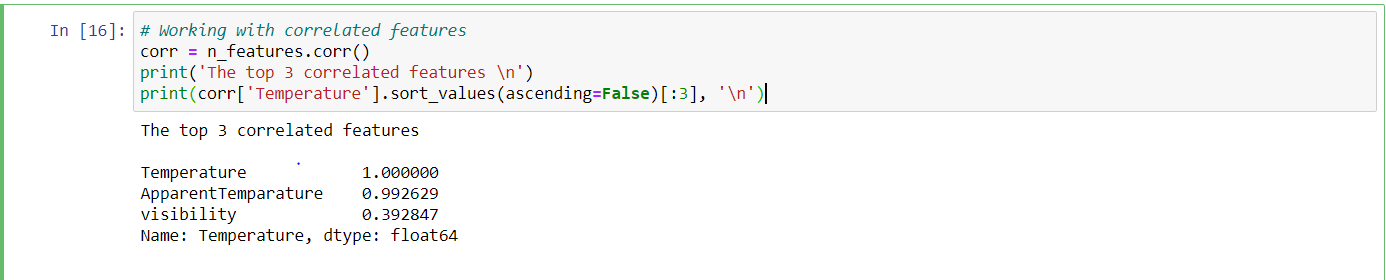
7.Impact of precipe type vs Temperature



8.Convert the precipe type to int values.

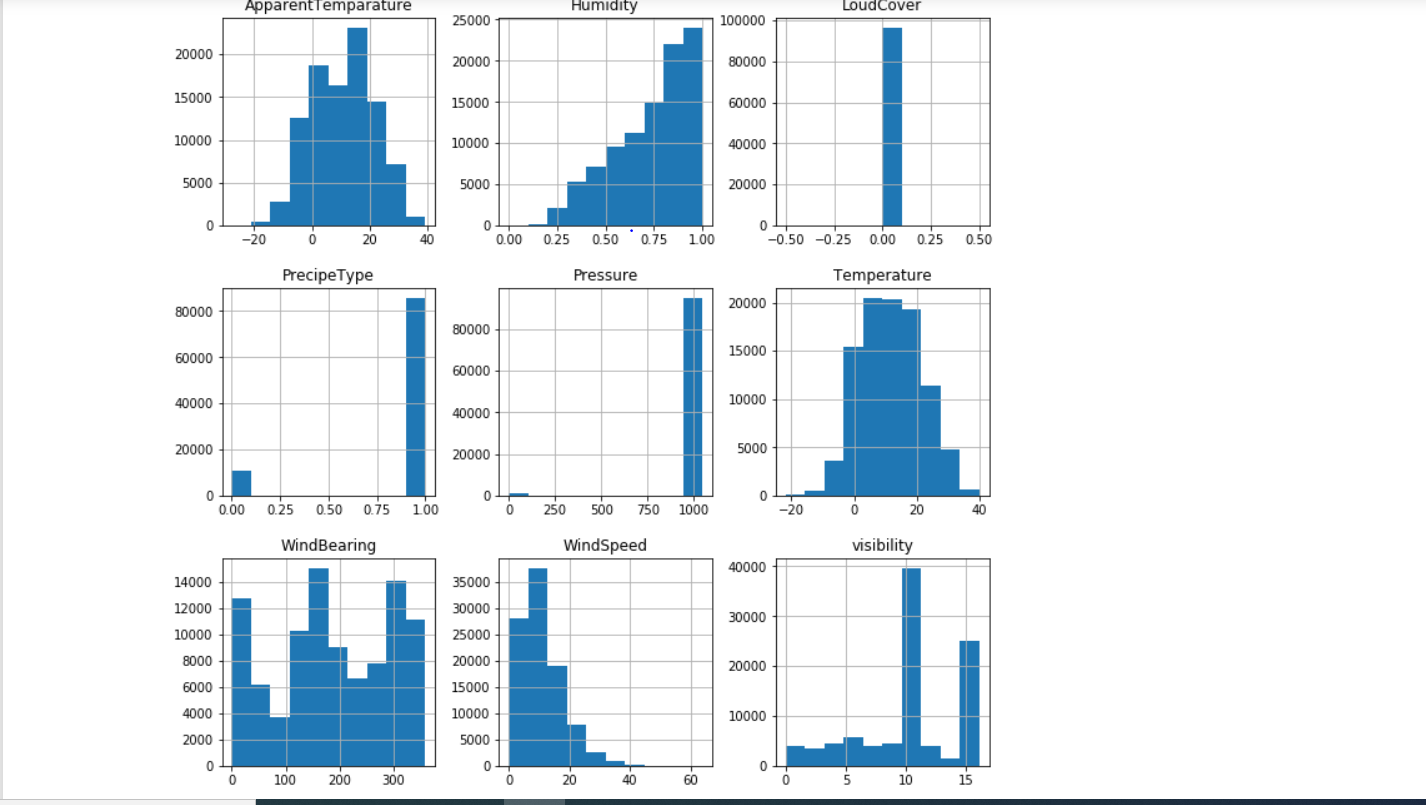


9.Top correlated features



10.Individual variations

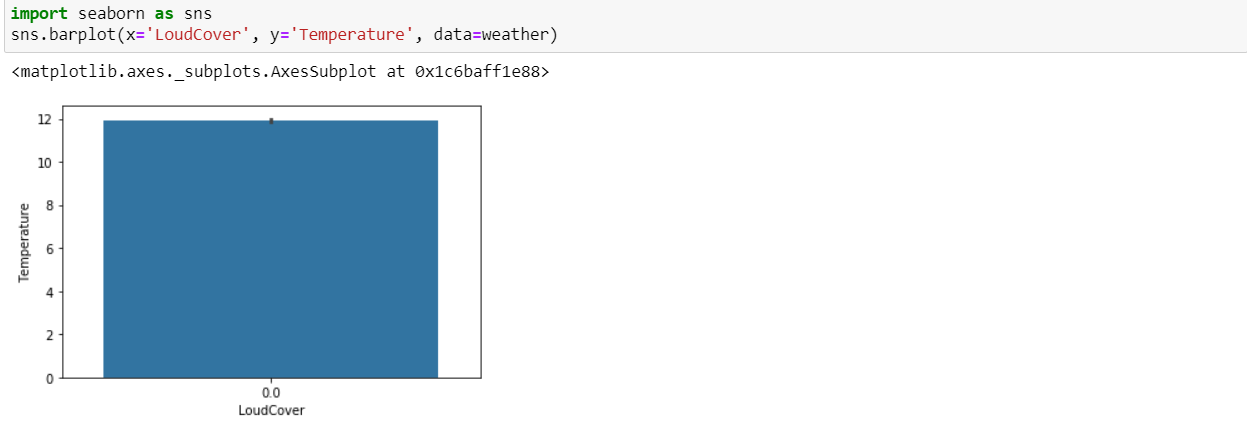




**Inference:**

From the graph we can infer that loud cover is steady. Understanding the impact of loudcover with temperature.

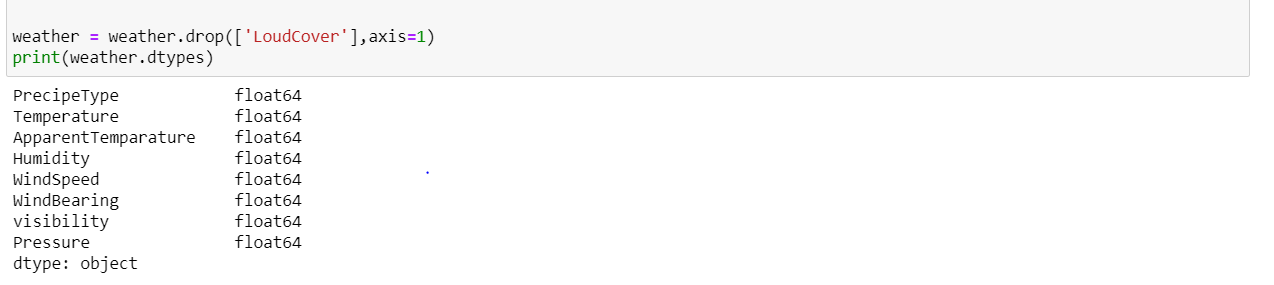
11.From the graphs plotted above we can see there is no impact of loud cover with temperature. So, plotting loud cover.



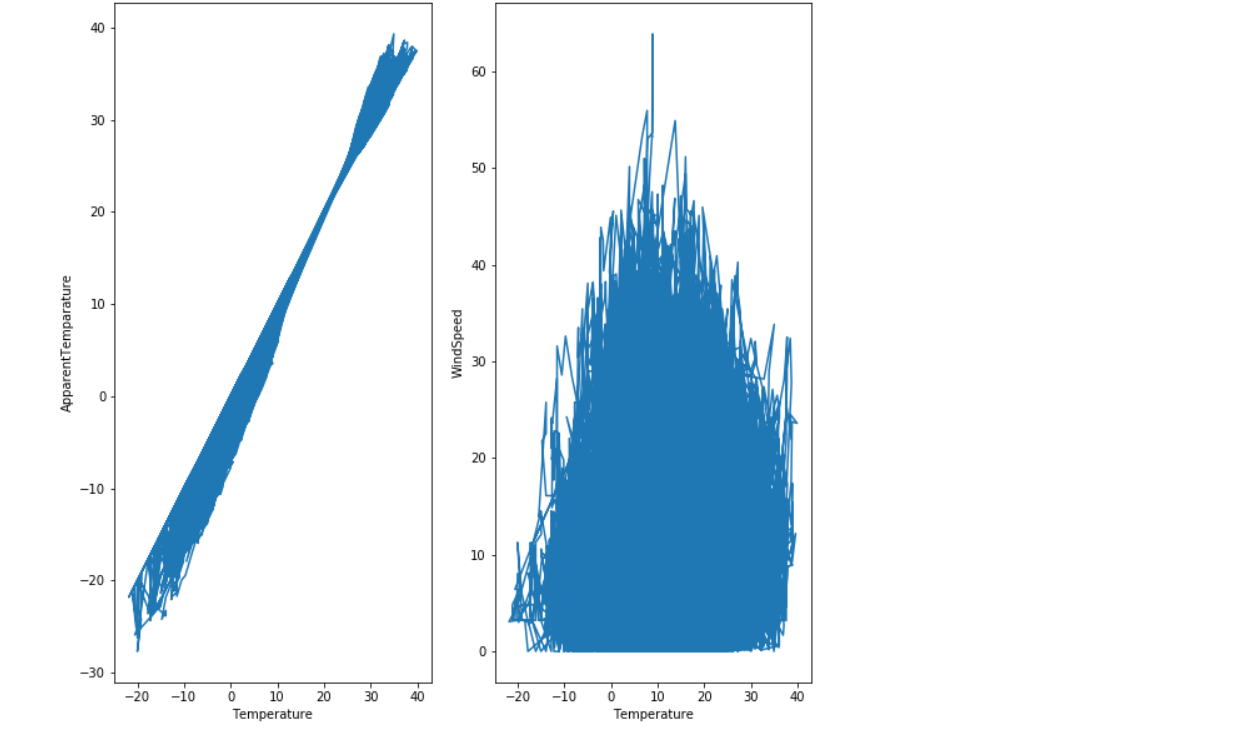
**Inference:**

From the graph we can infer that loud cover and temperature are steady and there is no change of loudcover with temperature. Hence dropping temperature.

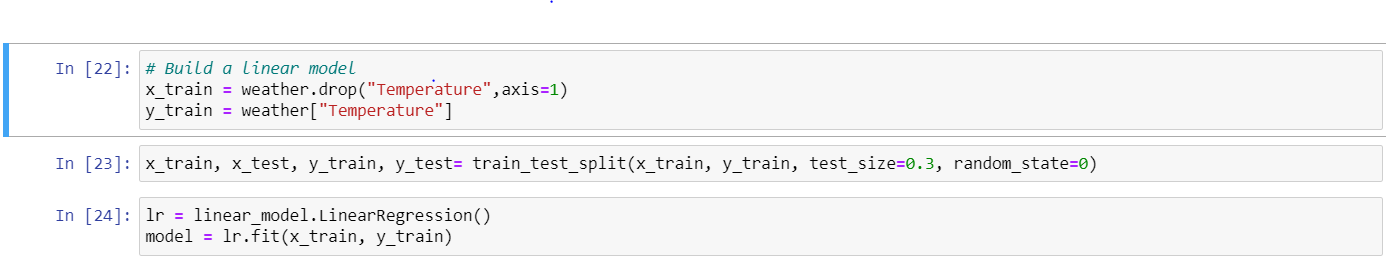
12.Dropping loudcover.



13.Temperature vs windspeed and Temperature vs Apparent temperature.

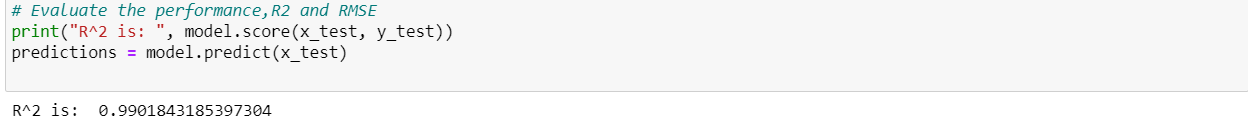


14.Building a linear model with all the features.



**b) Evaluate the model**

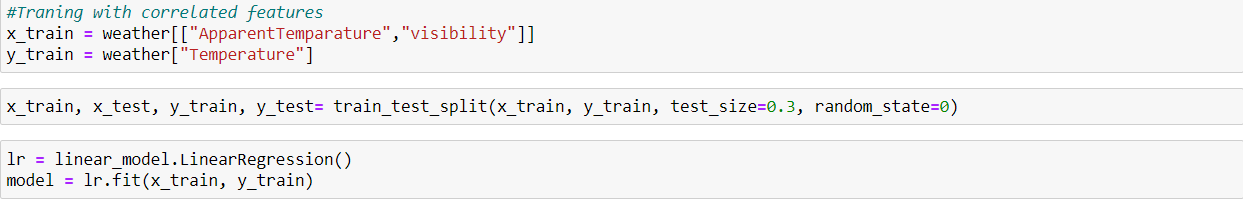
**Calculating R2 score**



**Calculating RMSE**



14.Building a linear model with correlated features

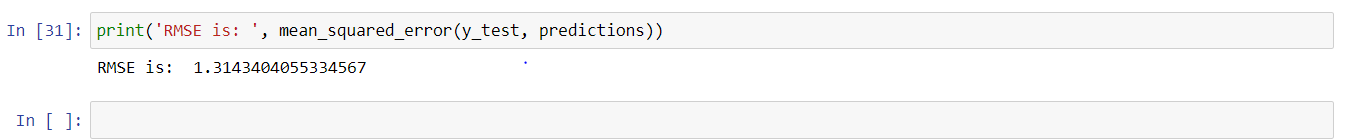


**b) Evaluate the model**

**Calculating R2 score**



**Calculating RMSE**

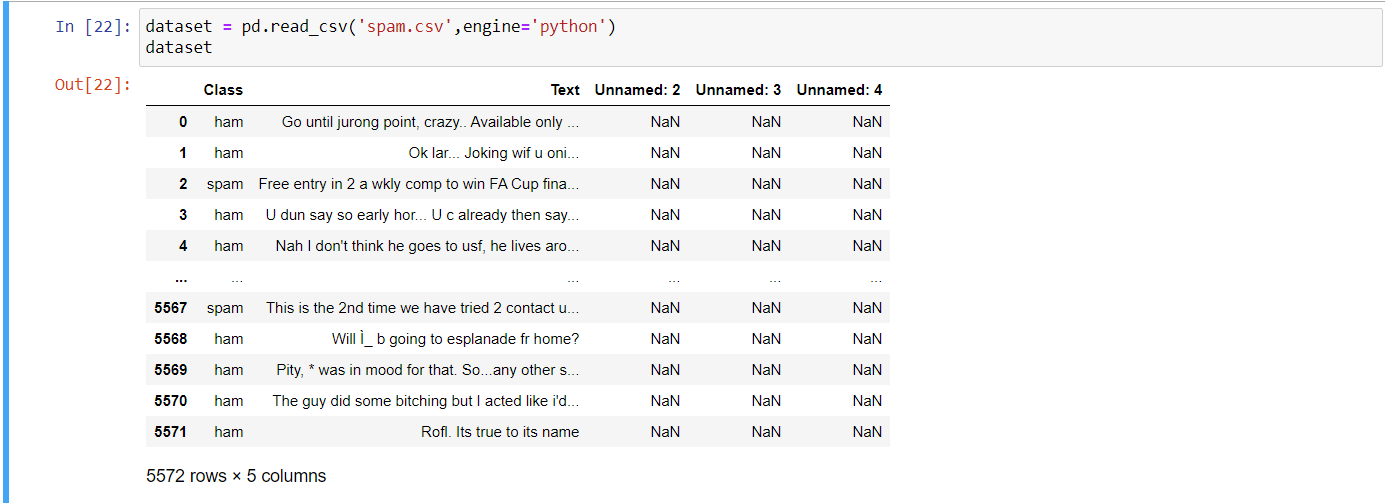


**Inference:**

The model with r2 value between 0.3 and 0.5 then the model is weak. If it lies between 0.5 and 0.7 then it is moderate and if it is greater than 0.7 then it is strong model. That is, it fits the best. From the above we can observe an increase in the R2 value which indicates that when the correlated features are considered the value increased

**Question4:**

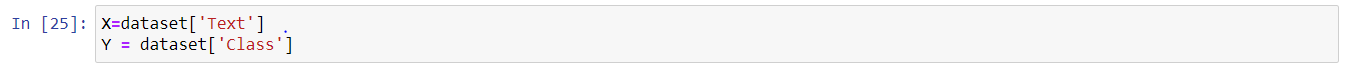
1.Do the necessary imports and load the dataset.



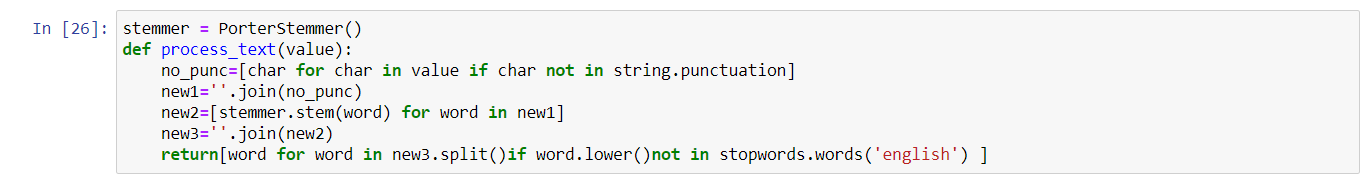
2.Checking null values.



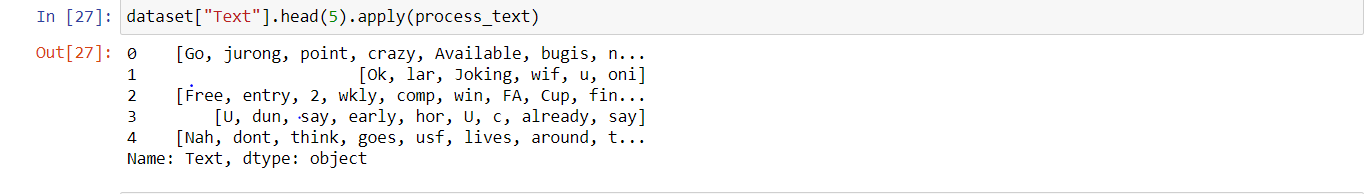
3.Loading dataset



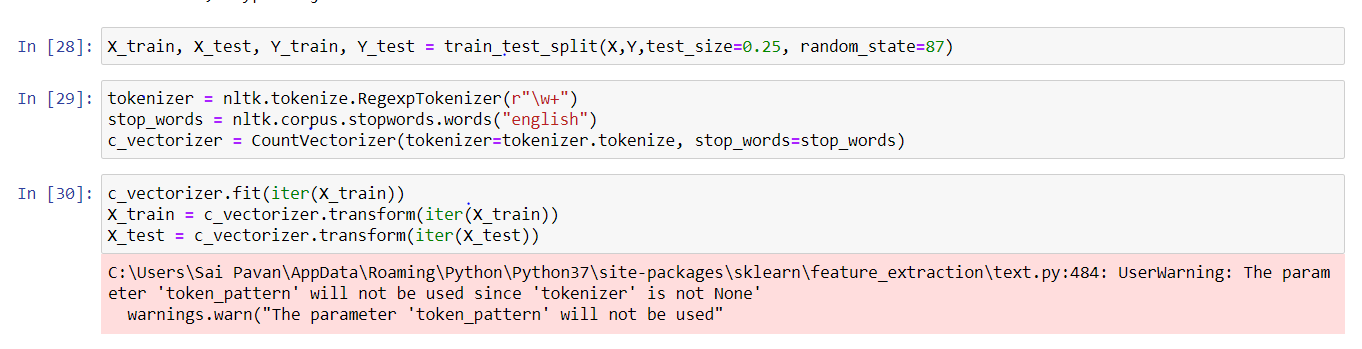
4.Creating function inorder to split the words in the text.



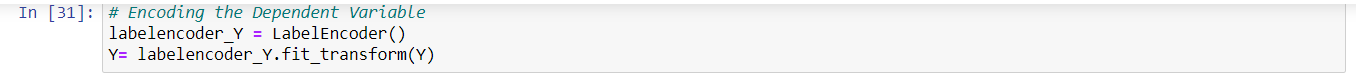
5.Words after splitting



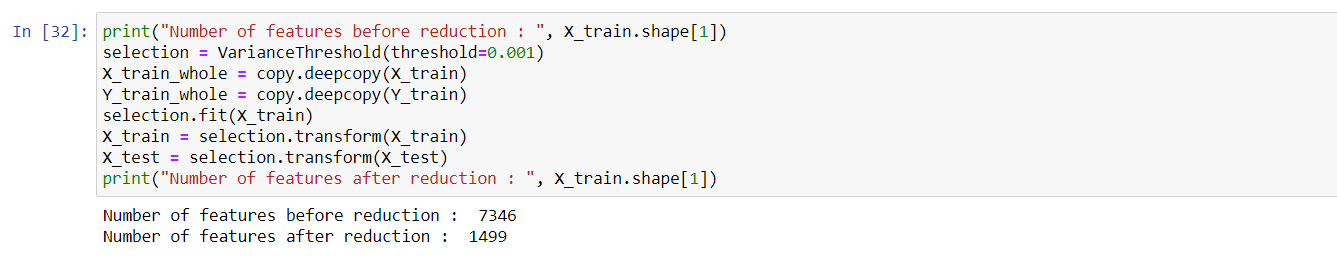
6.Applying the count tokenizer



7.Applying label encoding



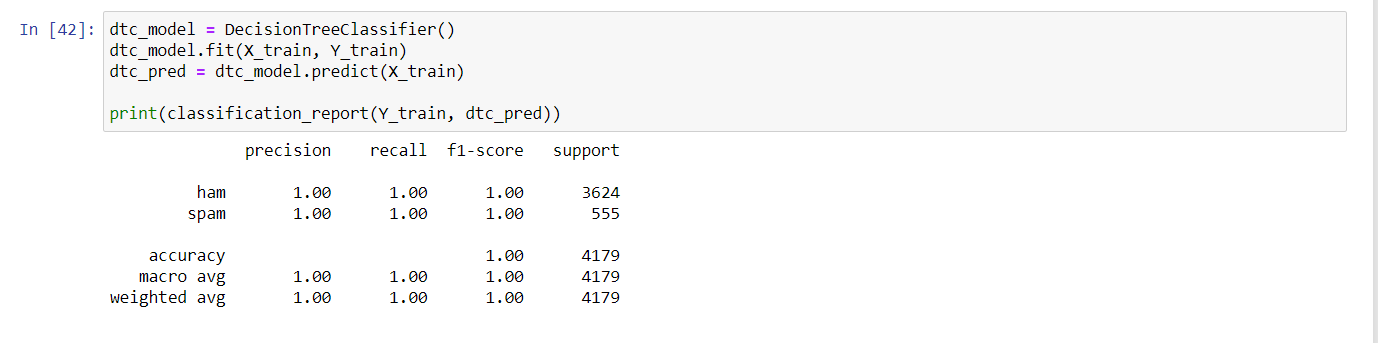
8.Analysis of features



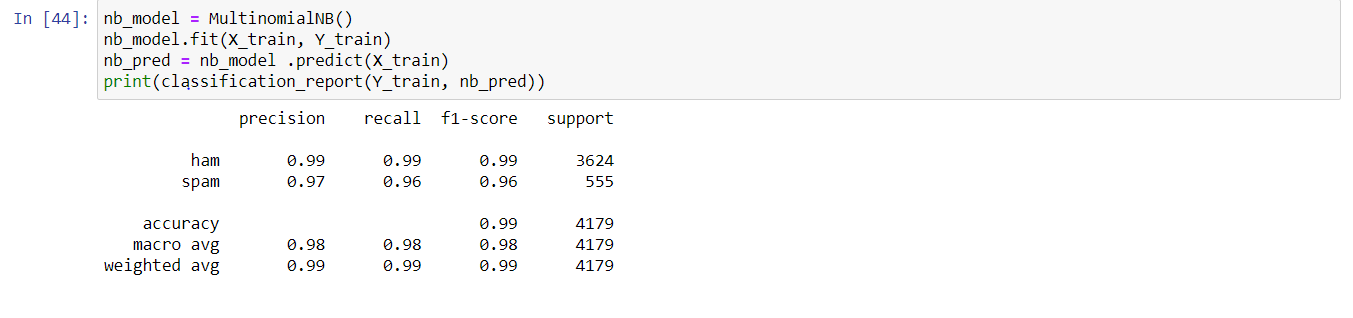
**Inference:**

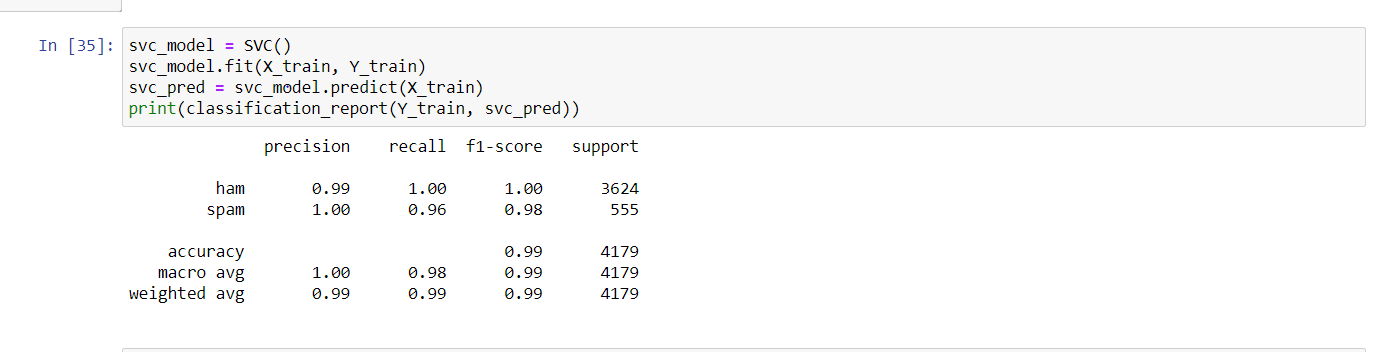
From the above it can be inferred that there are 1499 samples in the 0.001 variant data.

9.Design tree classifier



10.Multinomial NB



11. SVC model analysis

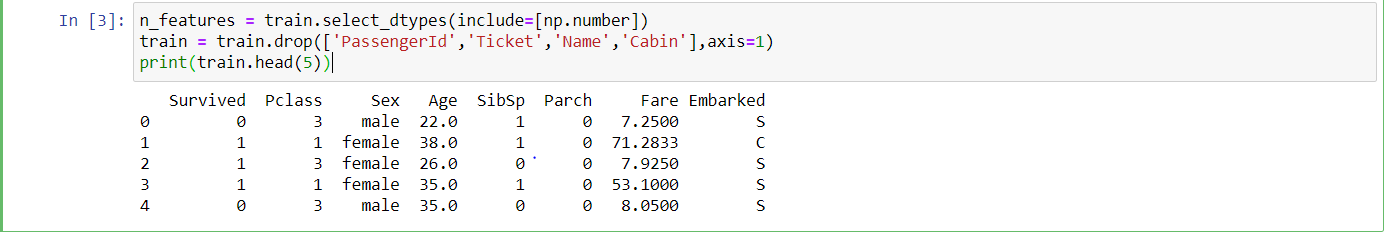
**Inference:**

From the above it can be inferred that the accuracy in decision tree is best.

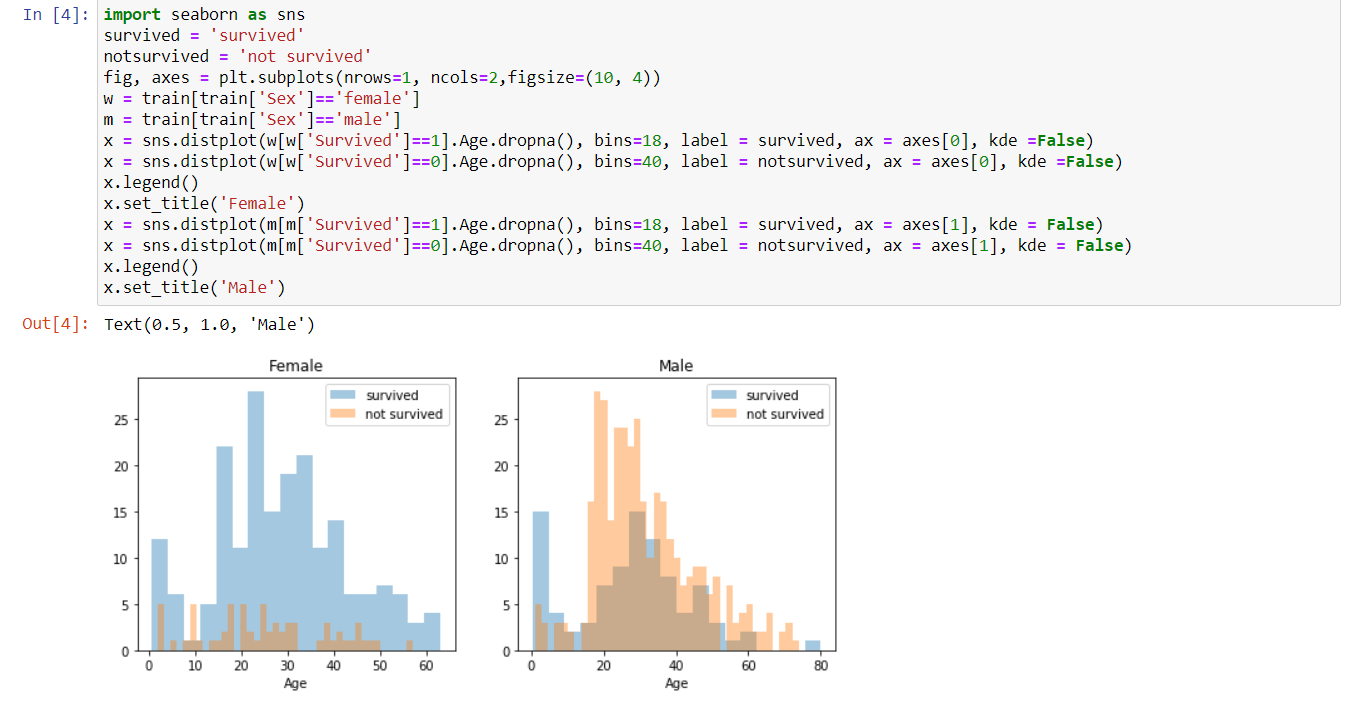
**Question 5:**

1.Make the necessary imports and load the data.

2.Dropping the unncessary features



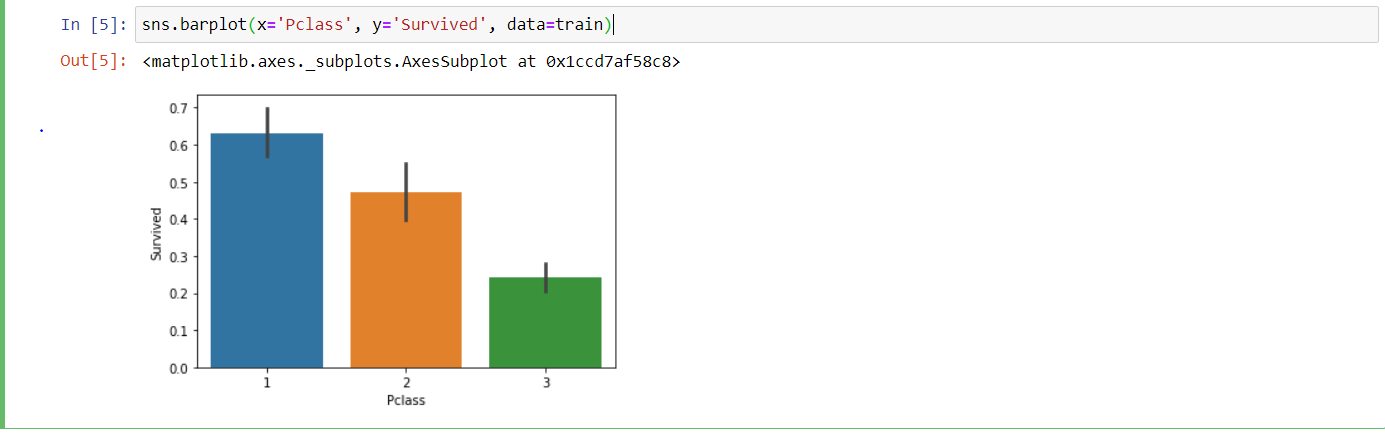
3.Analysis of Gender vs age vs survival



**Inference:**

The probability is highest when the age is between 18 and 30 years old in men where as in women it is between 14 to 40 years.And infants also have some high survival rate.

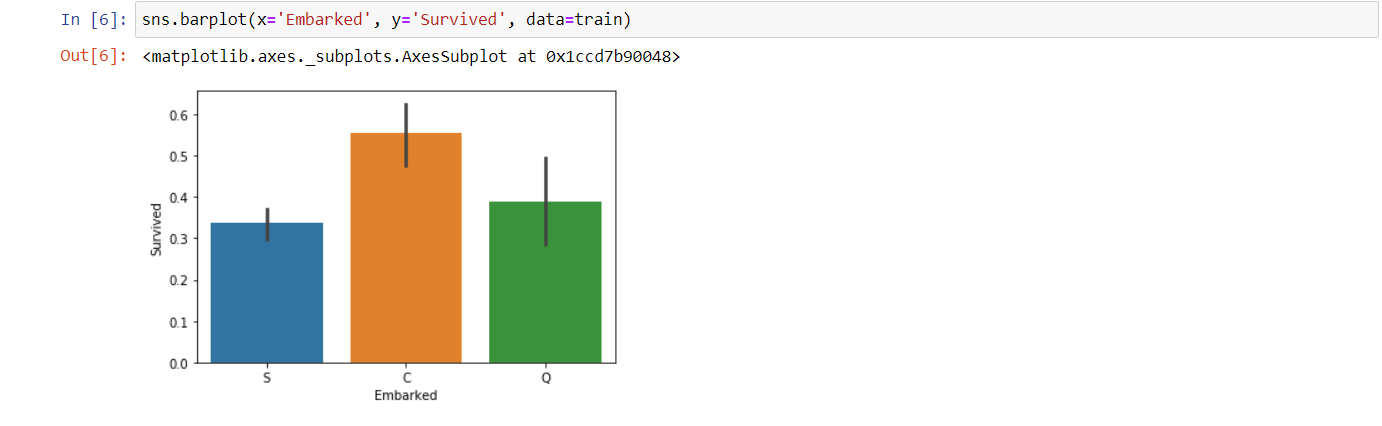
4.Comparision of pclass vs survived



**Inference:**

From this graph we can infer that pclass is contributing to the survival rate.

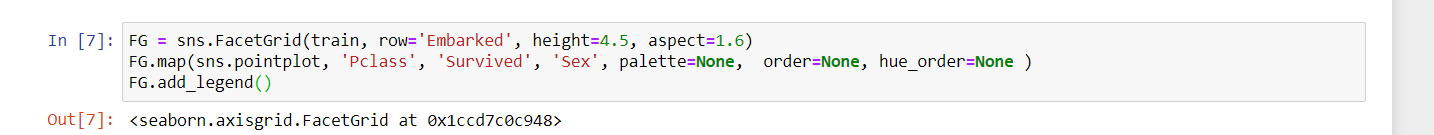
5.Comparision of embarked vs survived.

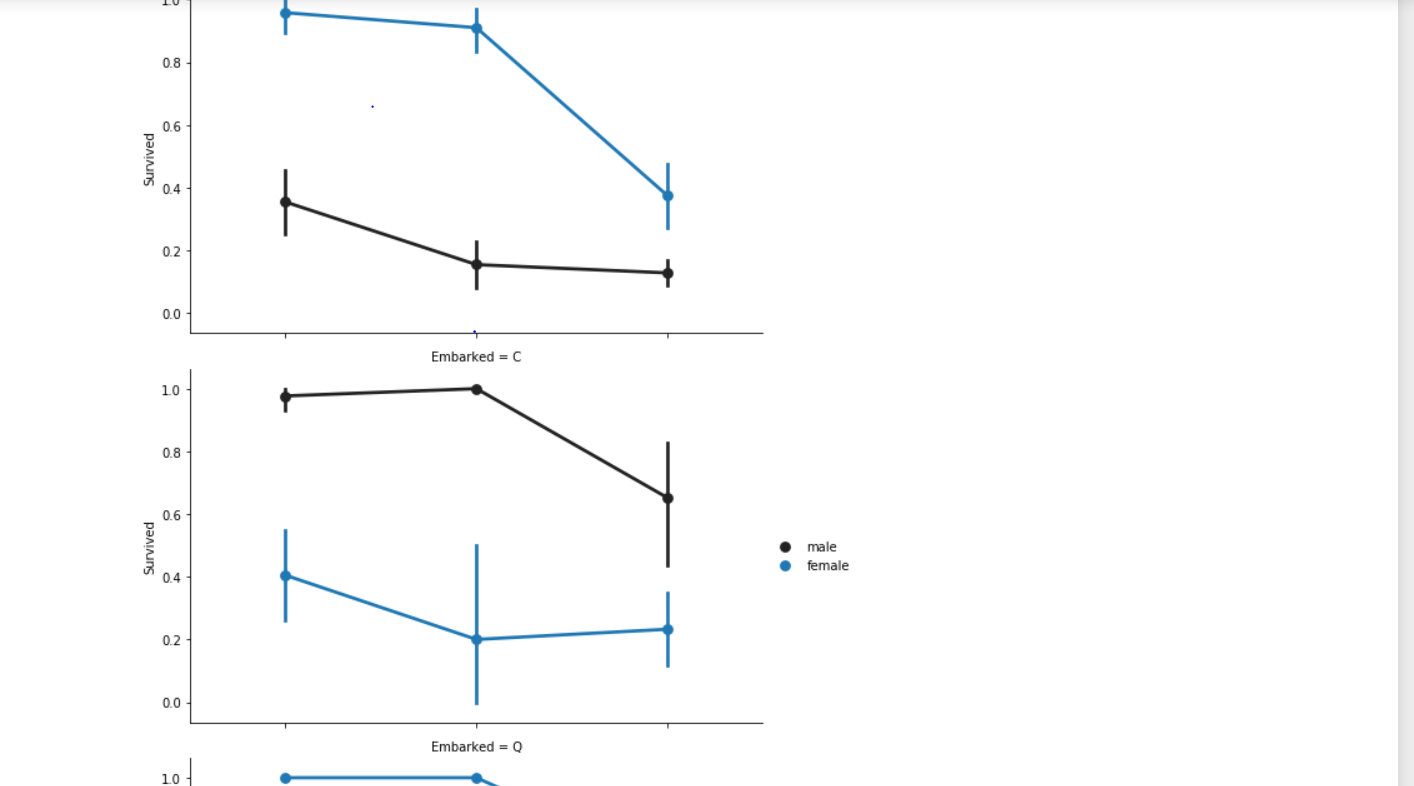


**Inference:**

From this graph we can infer that embarked also contributes to the survival where C has highest survival rate.

6.Comparsion of Pclass,survived and gender

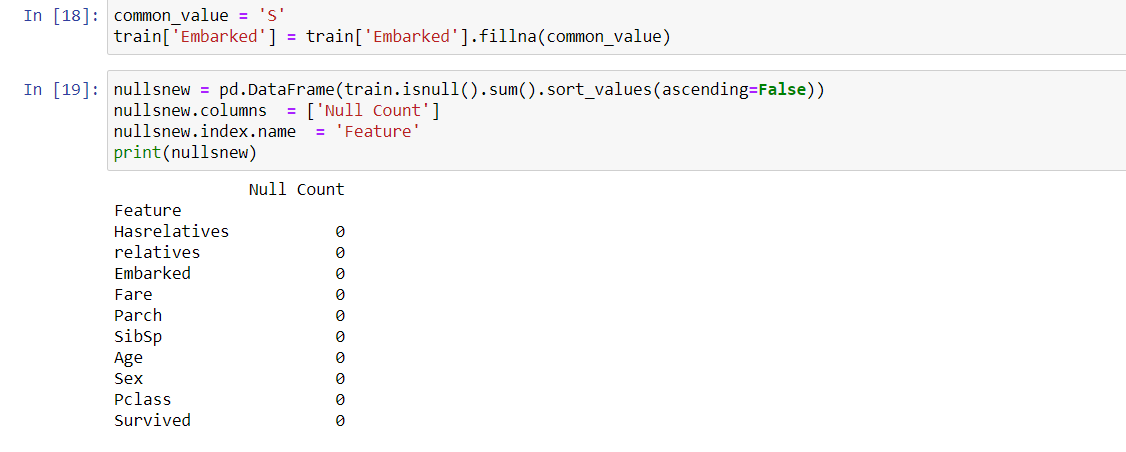


7. 

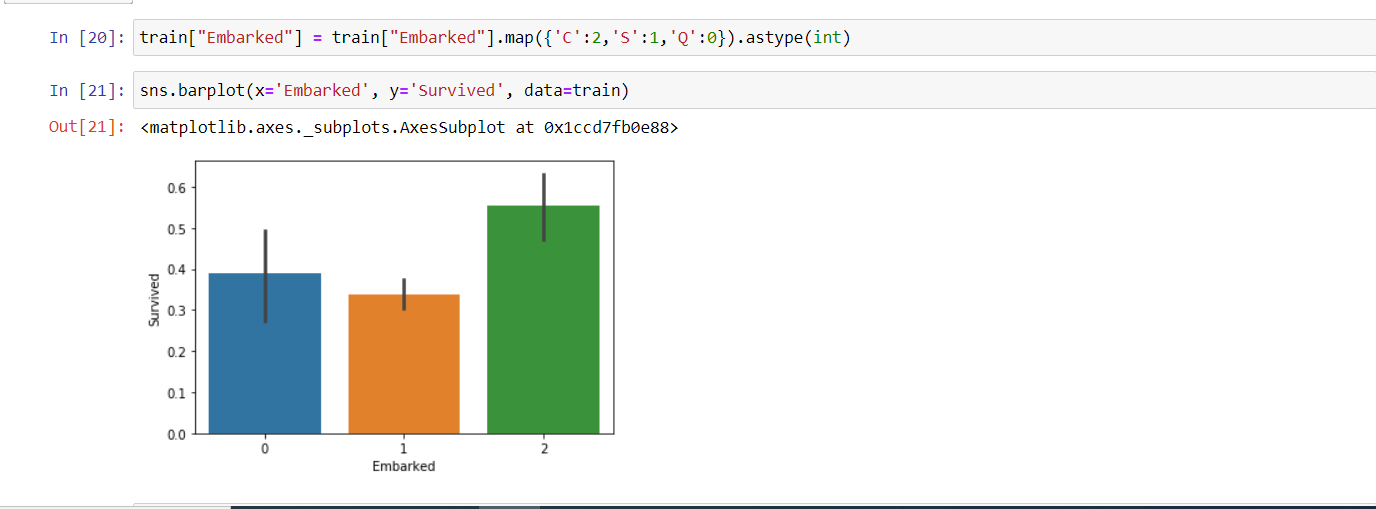
**Inference:**

From the graph we can infer that women at Q and S high chance of survival.Men have high probability on C than on Q and S.

8.Removing null values.



9.Converting numeric to discrete



**Inference:**

After handling the null values the values are mapped to 0,1,2.

10.Correalted features and data types.

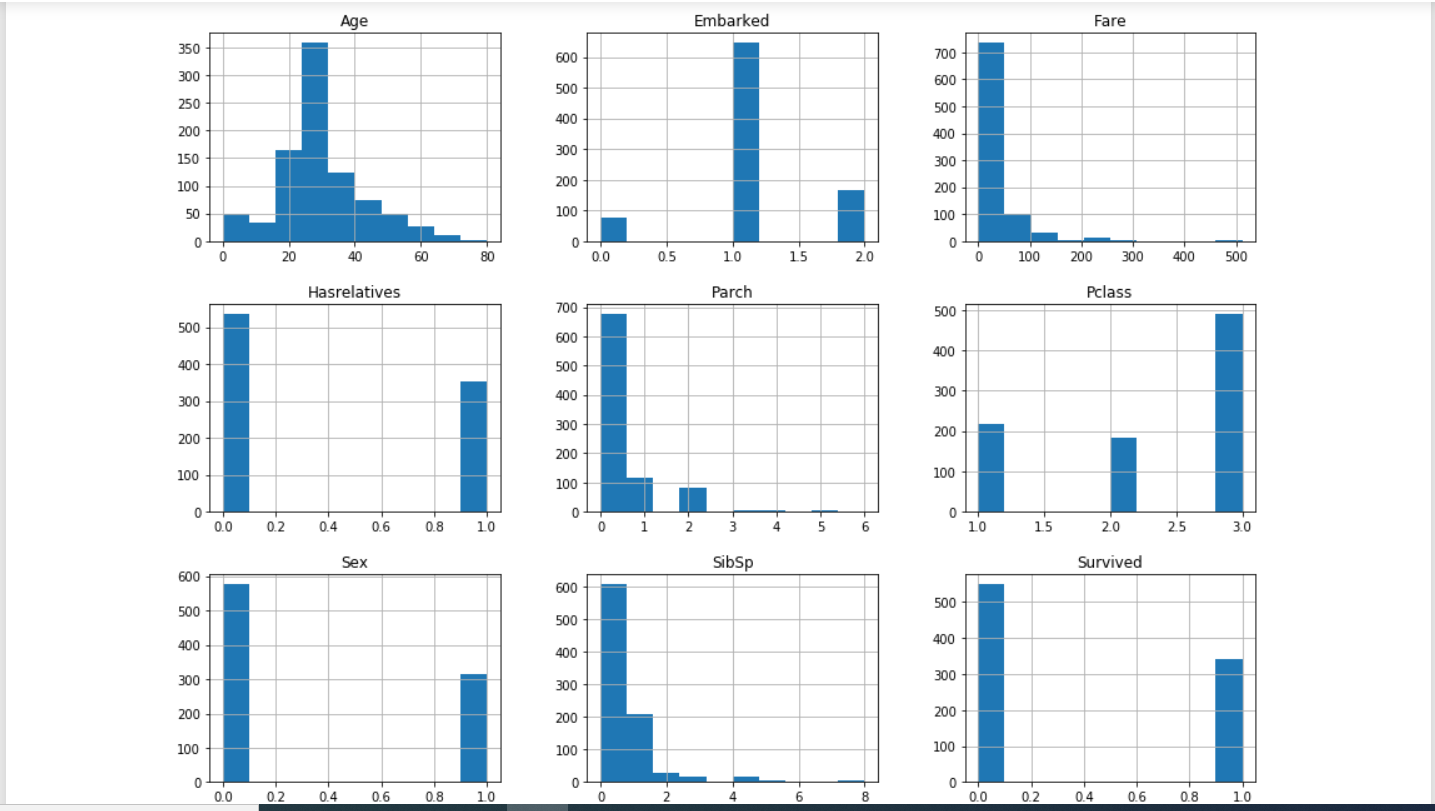


11.Converting the features to same unit



12.Plotting all the features to see the variation.

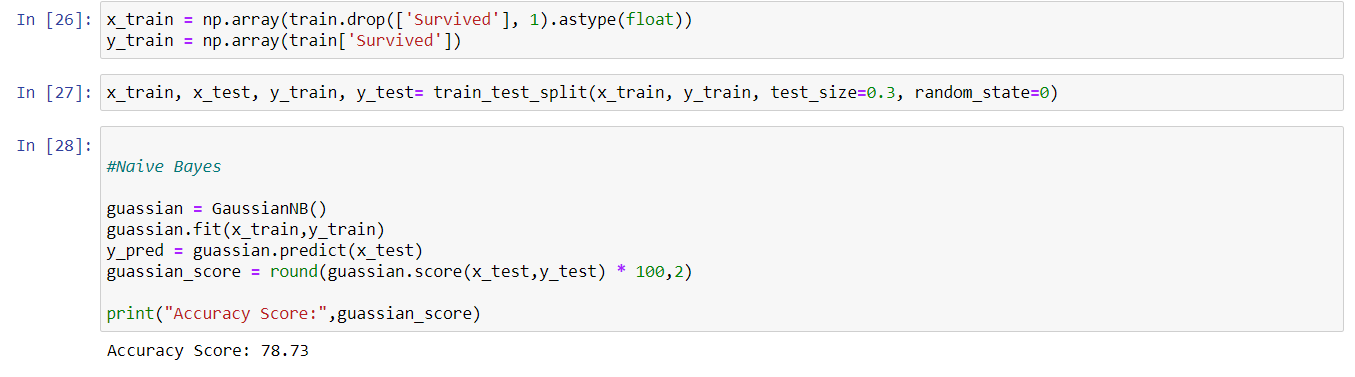




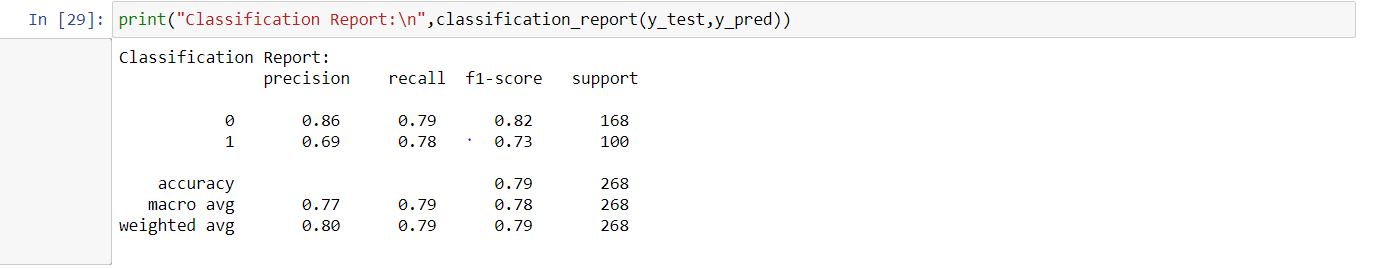
**Inference:**

From the graph we can infer that all the plots have variance. Hence all the above features are considered while building the model.

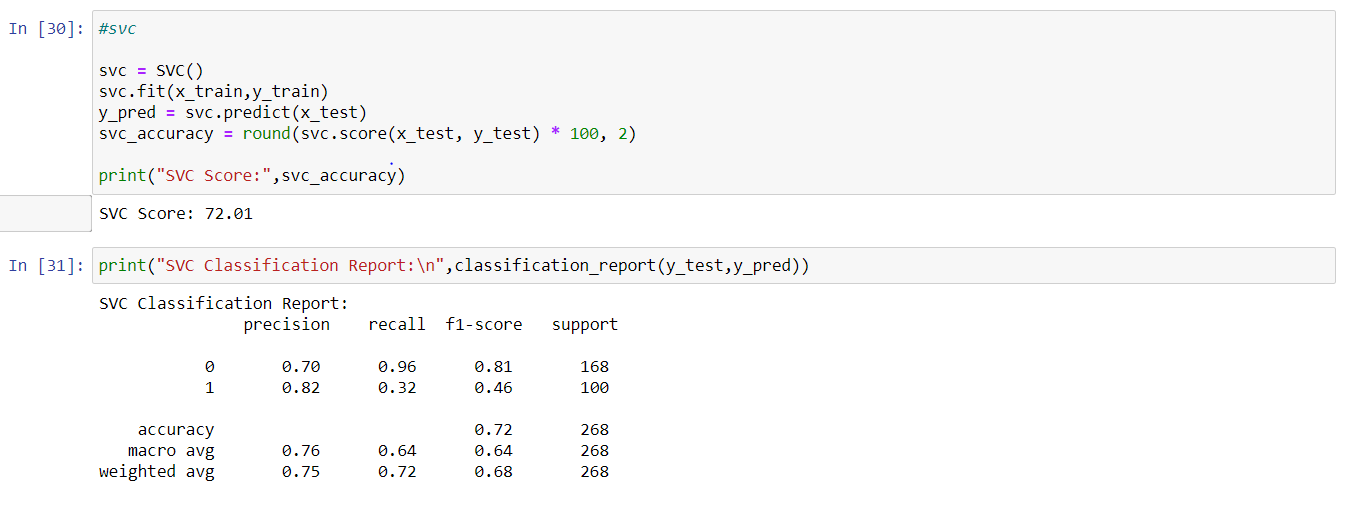
13.Guassian analysis



14.Classification report



15.SVC score and classification



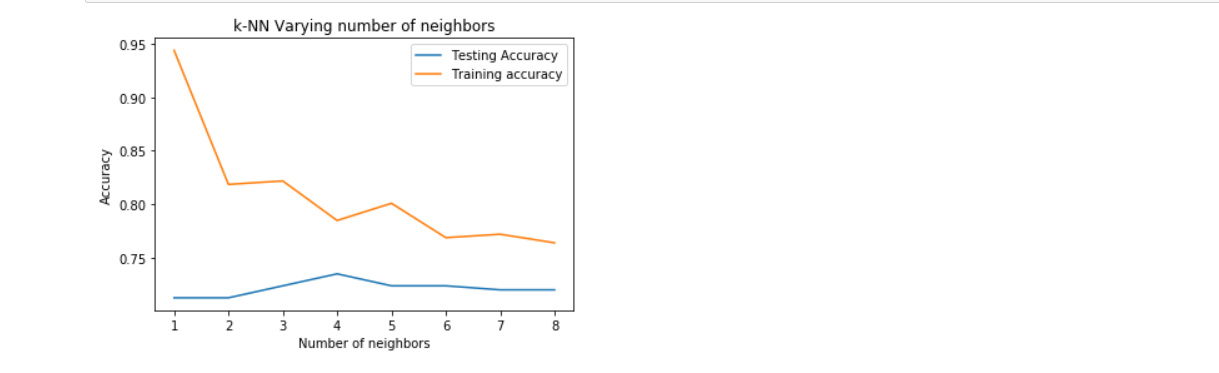
**Inference:**

Here we can observe that guassian distribution is to best than the svc.

16.KNN analysis



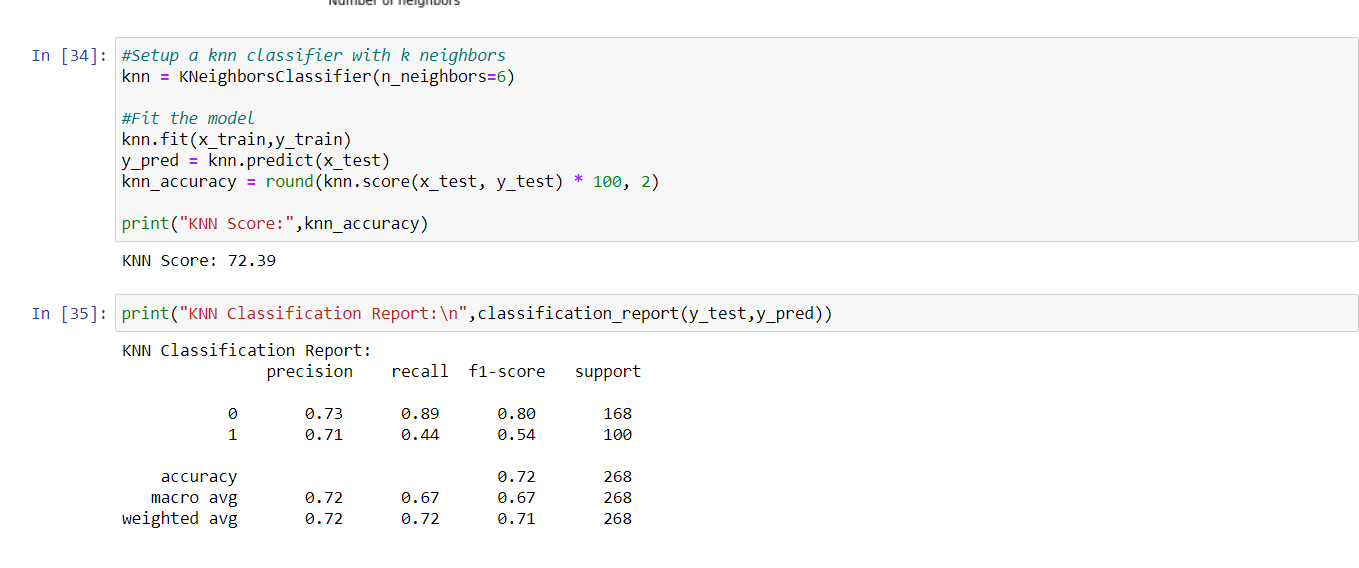
15.Accuracy graph



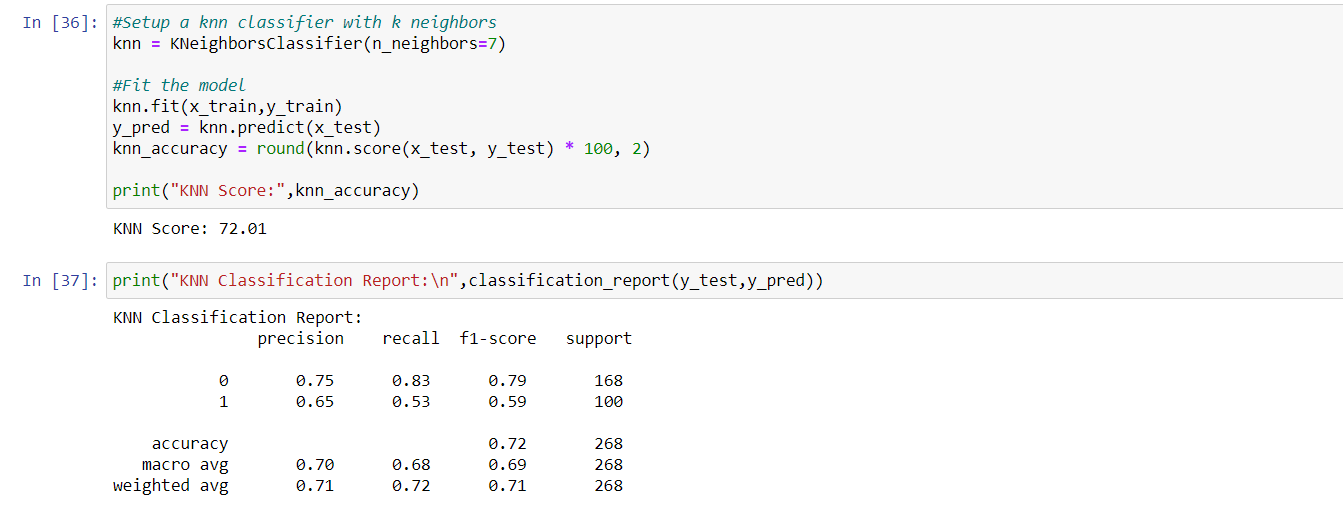
**Inference:**

From the above we can infer that at k=6 we can observe that there is maximum efficiency or score.Since the testing and the training data sets are almost close at k=6.

16.KNN score and report

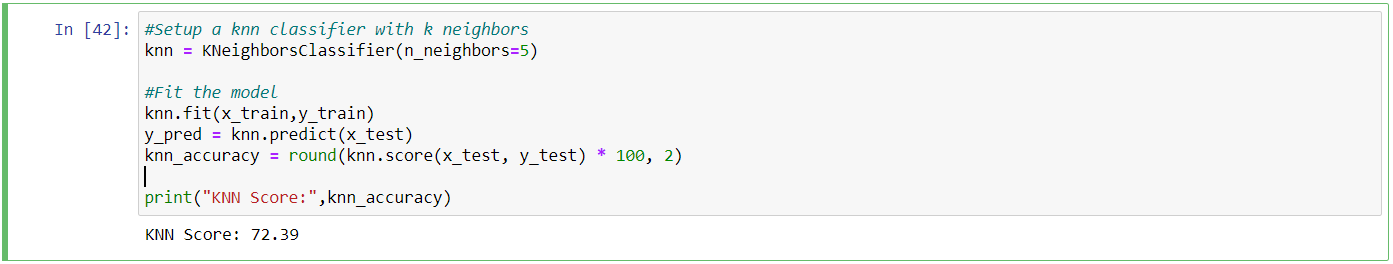


17.KNN score for n=7

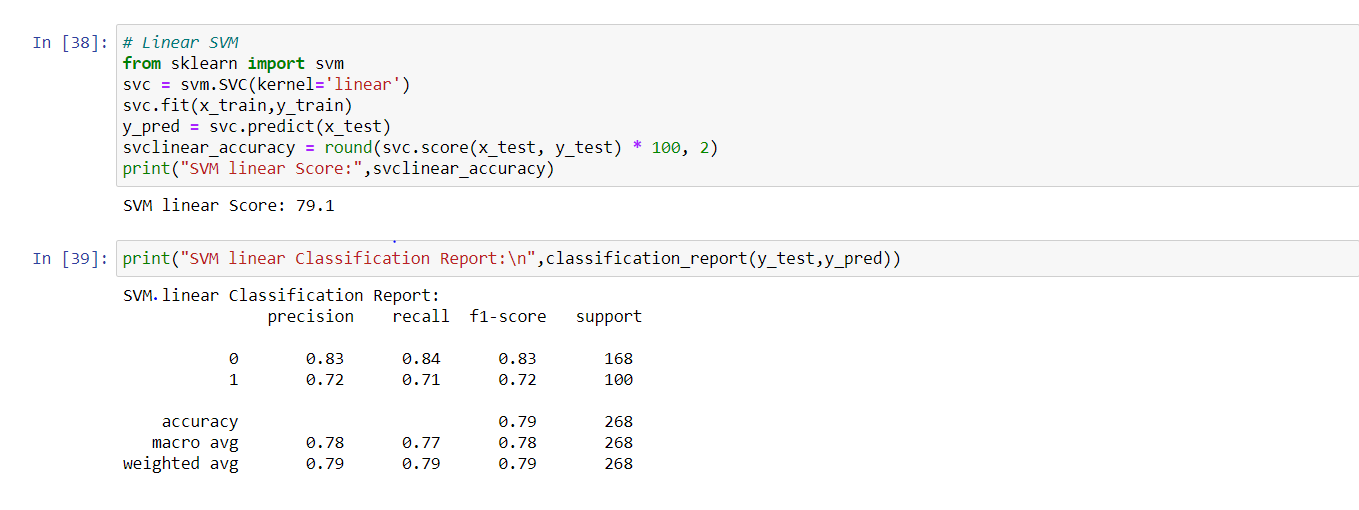


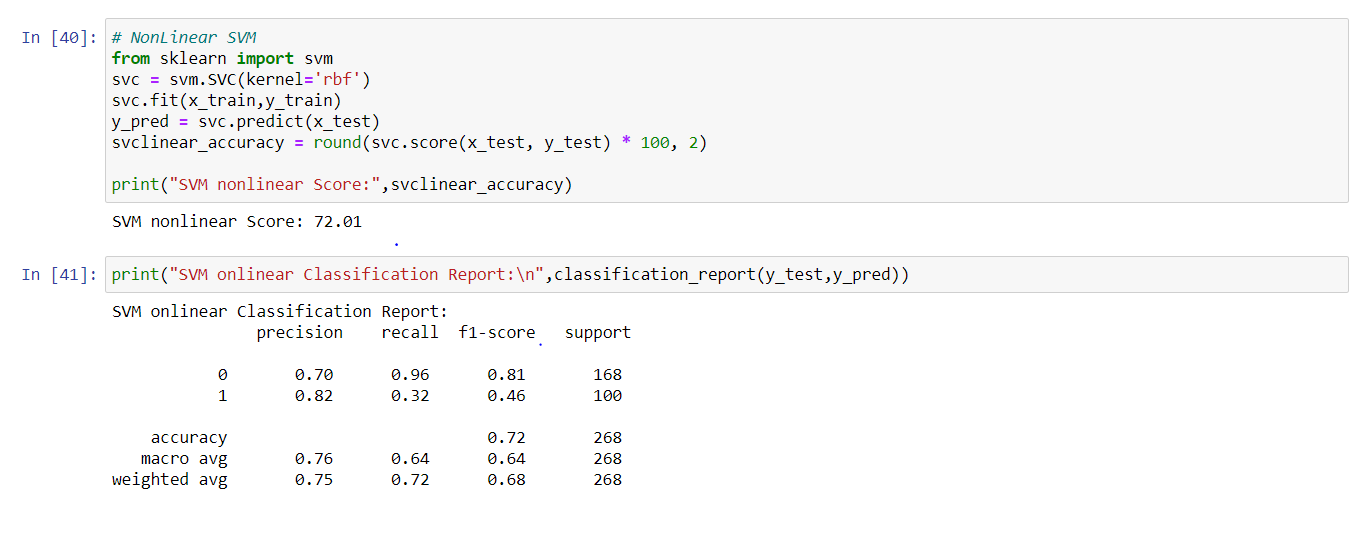
**Inference:**

From the above we can infer that at k=6 we can observe that there is maximum efficiency or score. Since the testing and the training data sets are almost close at k=6 or k=5.



18.Linear SVM

19.Non linear SVM



**Inference:**

From the above we can infer that the accuracy of the linear kernel is greater than the non linear kernel.

**Datasets:**

1. <https://drive.google.com/file/d/1qmnoNF1BLJ6Kv8MJLRGczh0tCJHsh94Q/view?usp=sharing>

2. <https://drive.google.com/file/d/1nlPh6Zx-_xb9j-eNSfoJdMJqF-_d1xUW/view?usp=sharing>

3. <https://drive.google.com/file/d/1L30rLh4grwhgKiZZpdR8KPVAZrxl2OX9/view?usp=sharing>

4. <https://drive.google.com/file/d/1nRtiEnFWXWVqlqBAn2nxy2i7BsrkTxeg/view?usp=sharing>

5. <https://drive.google.com/file/d/1vrJ117E6DYq9_3mBPnlAz56Qkr-sH9Az/view?usp=sharing>

**Parameters:**

1.In the credit card dataset all the others are features except time and class is the target.

2.Here the comparision is taken between Annual Income and spending score.Here the clustering is used.

3.In the weather dataset all the features except summary,dailysummary,Date,loudcover and temperature is the target.Here the model created is a linear regression model.

4.Here the classes like text and class are taken whereas the other features are ignored.

5.Here in the train data set survived is taken as the target and all the others are taken as features except passengerid,Ticket ,name and cabin.

**Evaluation and Discussion:**

1.With the help of the smote balancing technique the unbalanced data can be balanced.

After balancing the dataset with the help of the above technique the accuracy score can be improved.

2.After calculating the value of k using elbow method the k value can be identified as the second drop from the left hand side. When the first drop is considered the value of the k is 3 and the accuracy is less when the k value is considered as 5.

3.The model shows a high accuracy score when correlated features are taken when compared to all the features.

4.The decision tree is considered as a best technique for the model compared to SVC and

Multinomial NB

5.Here the guassian distribution is better than SVC and the from the graph between the test data and trained data the k value is identified to be 6.And the linear kernel model is better than non linear kernel.

**Conclusion:**

From this project we are able to analyse the different classification, clustering and regression models along with NLTK features. We are able to calculate the accuracy scores and find out the efficiency model.