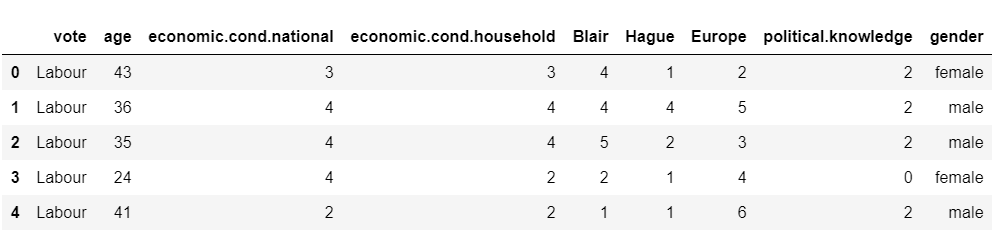
**Business Report**

**Machine Learning Project**

Problem 1:

1.1) Read the dataset. Do the descriptive statistics and do null value condition check.

First 5 columns of imported dataset



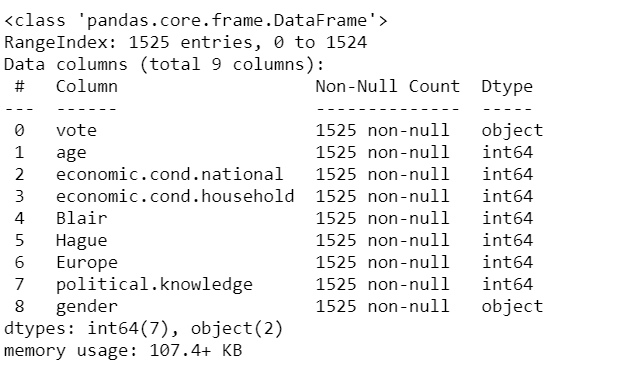
Decriptive Statistics

1)Shape

no. of rows: 1525

no. of columns: 9

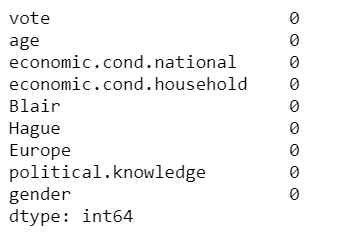
2)Data type of data features



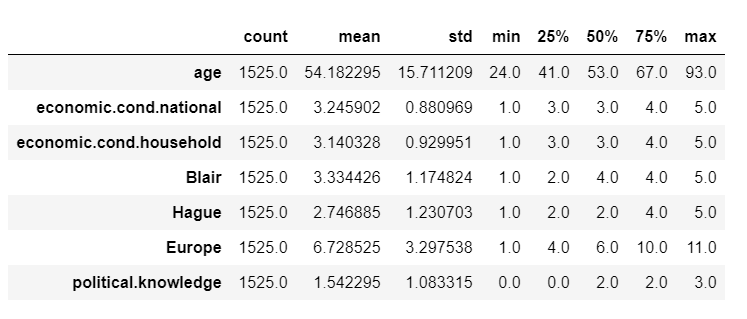
7 numerical columns and 2 categorical columns

3)Null value check

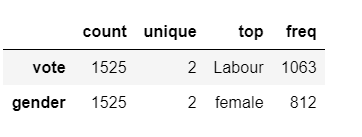
No null values



4)Decription of continuous variables



5)Description of categorical variables



6) Unique values for categorical variables

VOTE : 2

Conservative 462

Labour 1063

Name: vote, dtype: int64

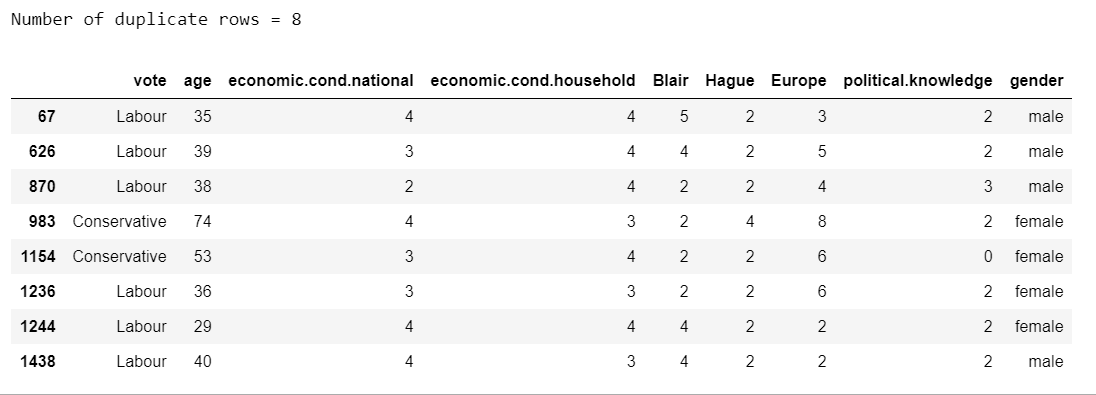
GENDER : 2

male 713

female 812

1.2)Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each.

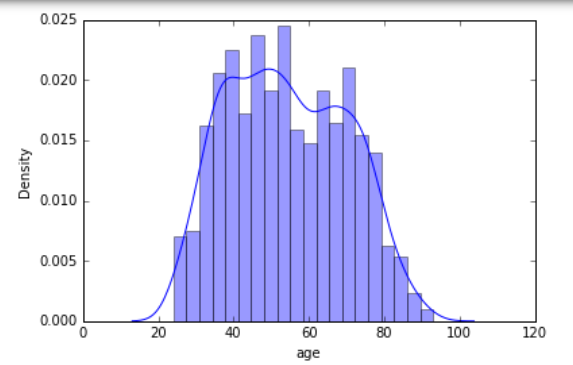
7) Check for duplicate data

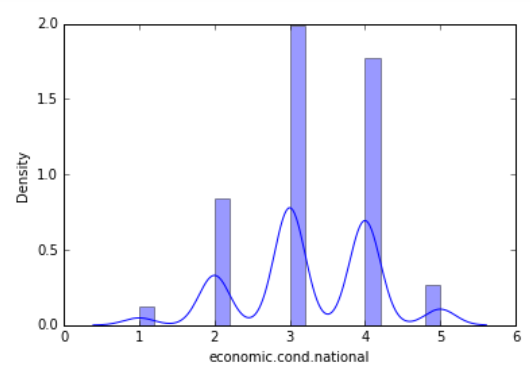


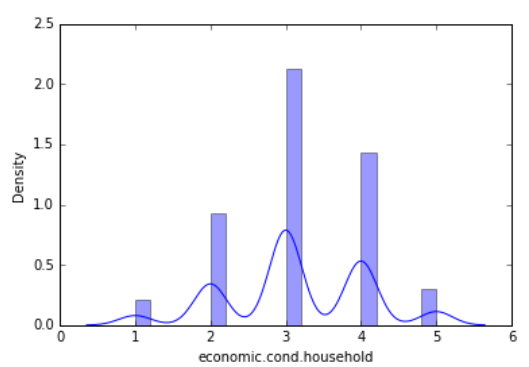
Removal of duplicates data

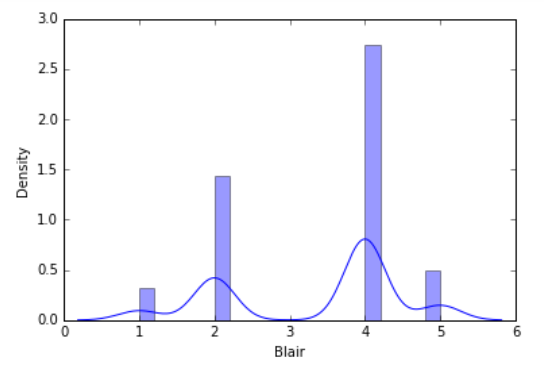


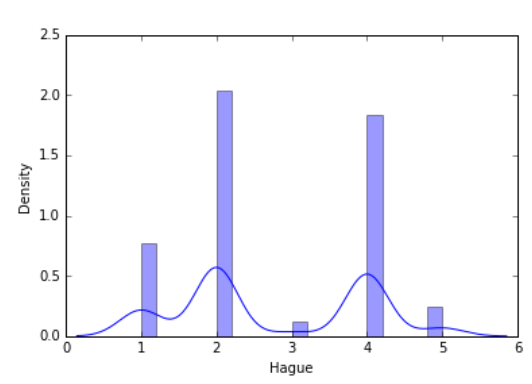
8)Univariate Analysis

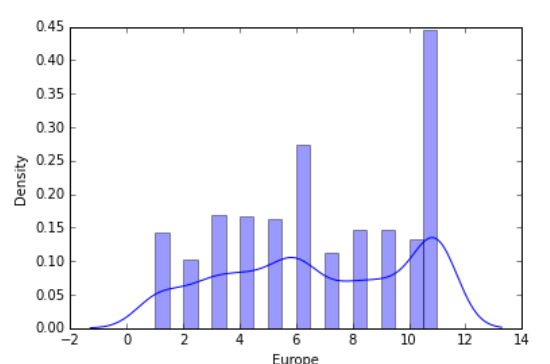


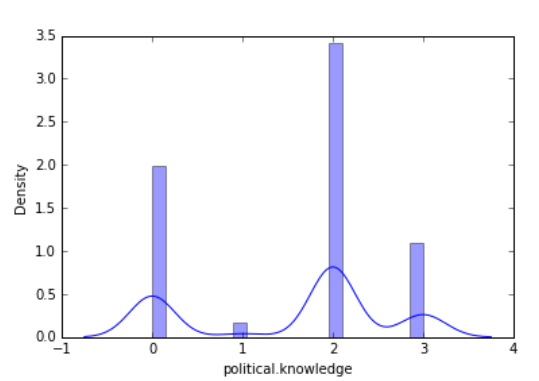




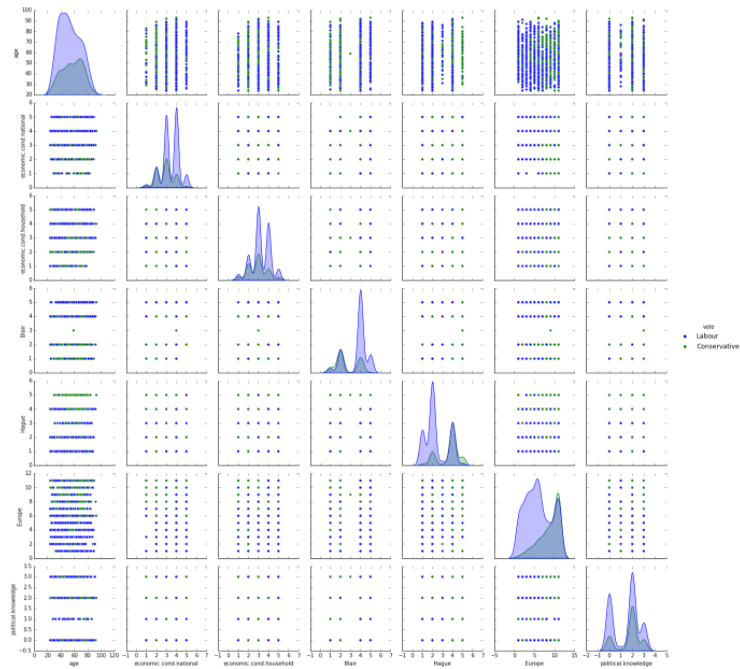




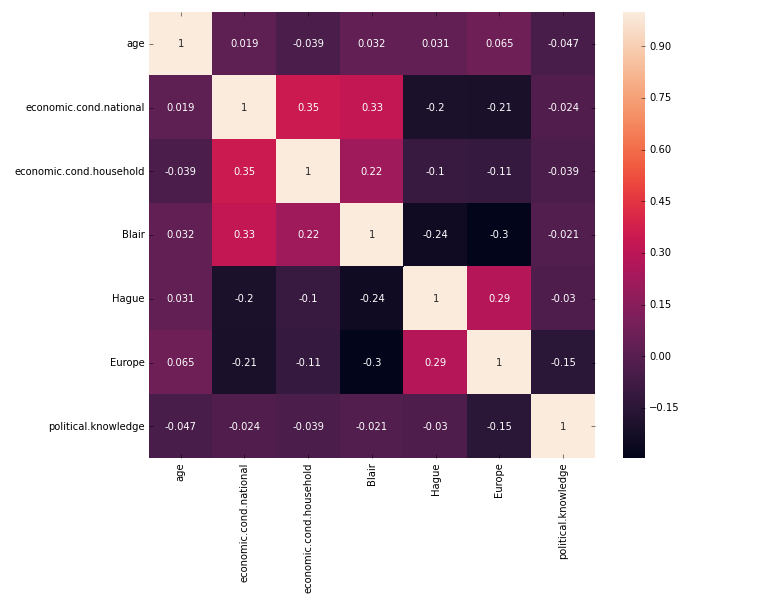




9)Bivariate Analysis

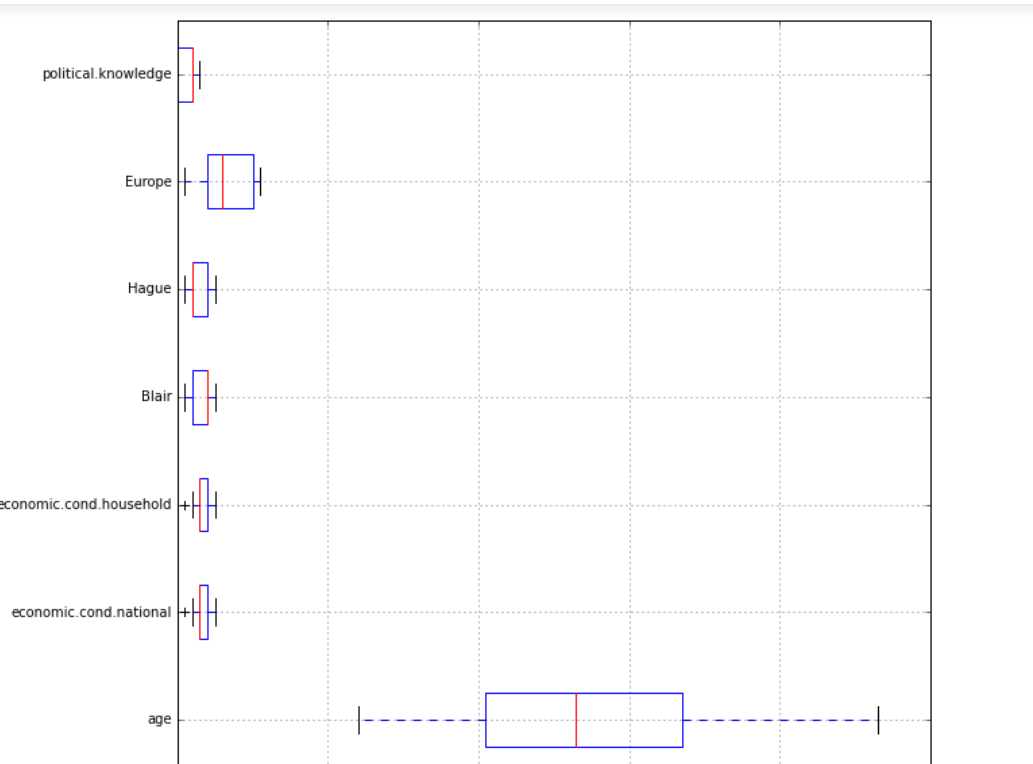


Heatmap

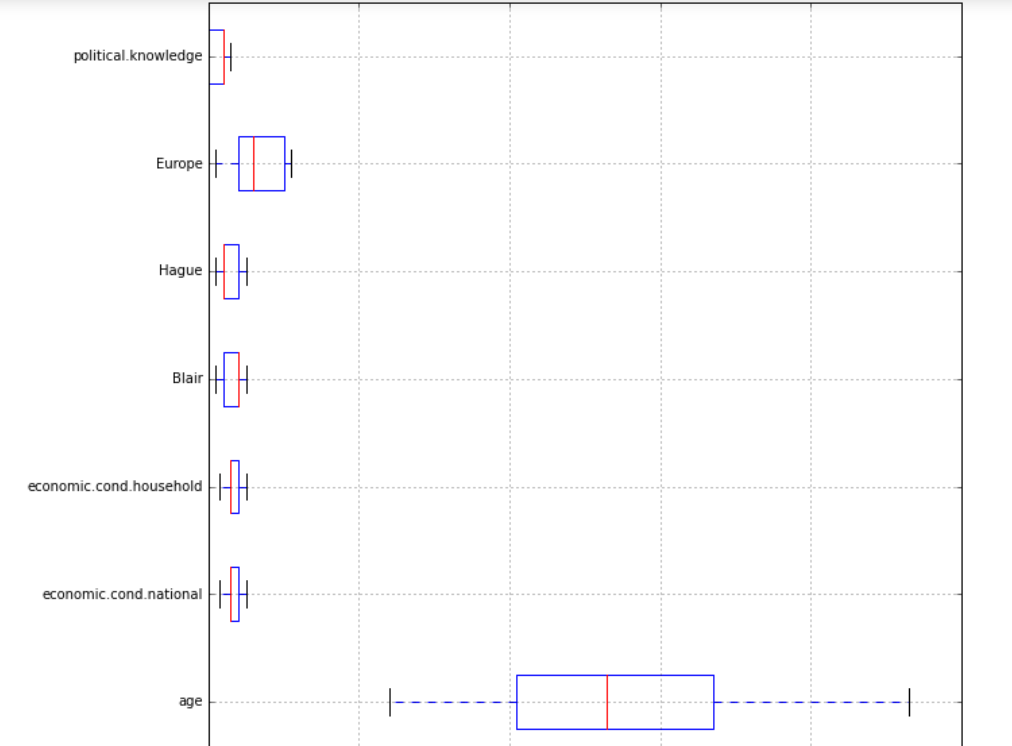


10) Outliers treatment

“economic.cond.national” and “economic.cond.household” have outliers.

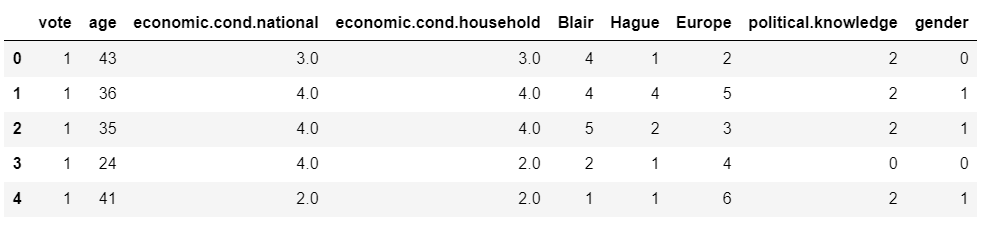


After outlier treatment



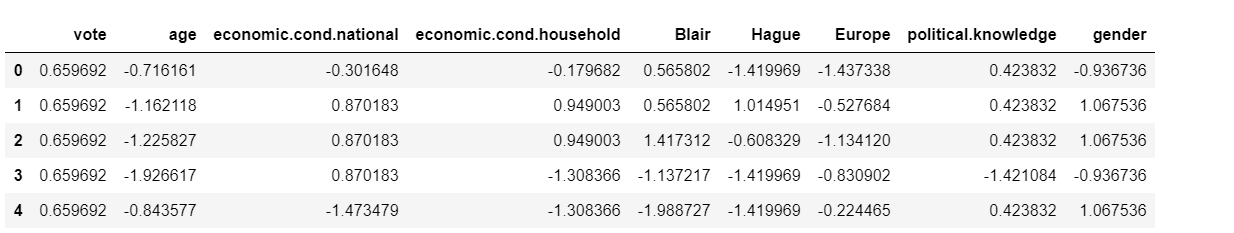
1.3)Encode the data (having string values) for Modelling. Is Scaling necessary here or not?( 3 pts), Data Split: Split the data into train and test (70:30)

Encoded data

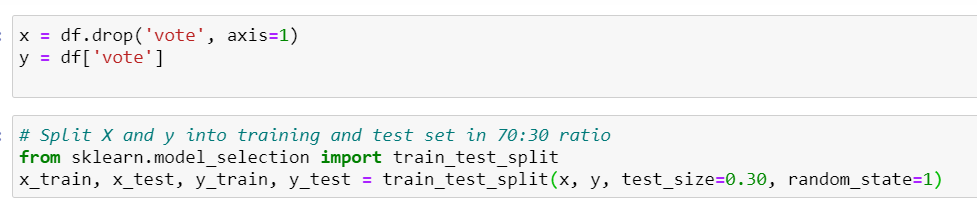


Scaled data

Scaling is done as age variable is not in sync with other variables

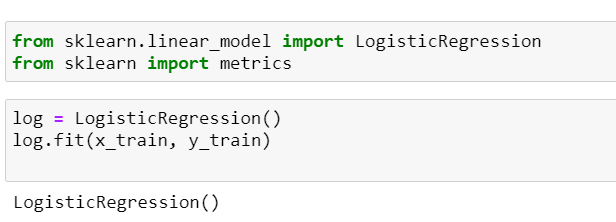


Train Test split

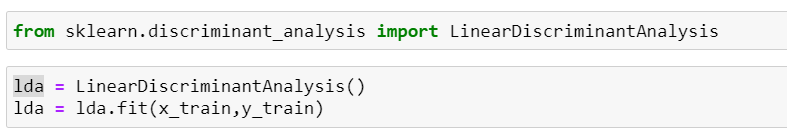


1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (3 pts). Interpret the inferences of both models

Logistic Regression

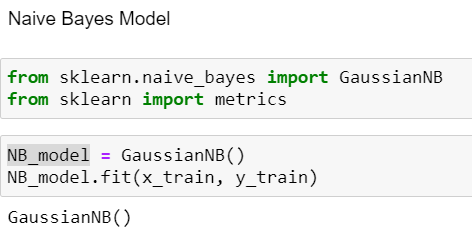


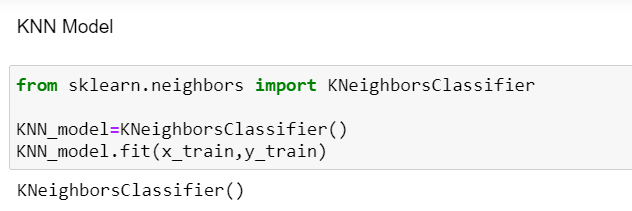
LDA



Accuracy, Precision, Recall and F1 score of Logistic regression and LDA are almost same.

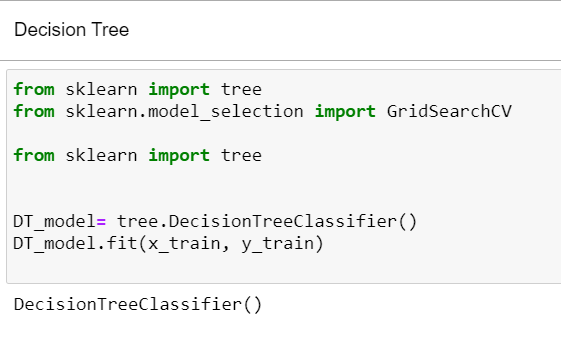
1.5) Apply KNN Model and Naïve Bayes Model (5 pts). Interpret the inferences of each model

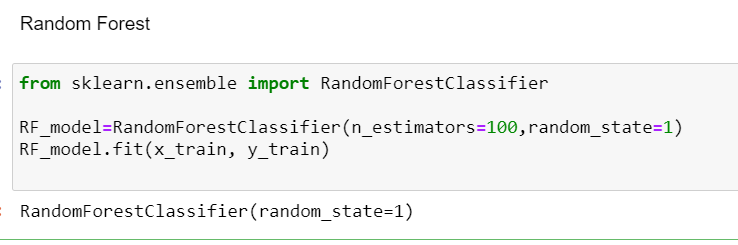


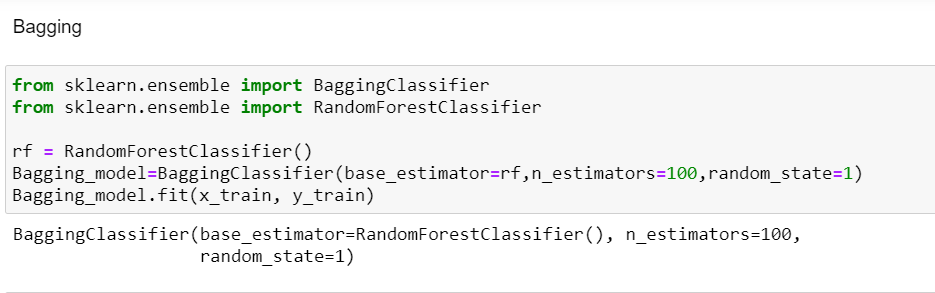


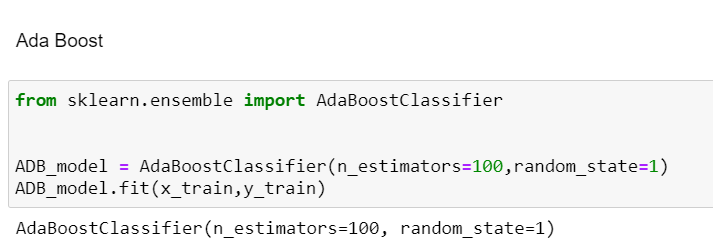
KNN model is better than Naïve bayes as accuracy, recall , precision and f1 score of both training and testing data is greater in KNN.

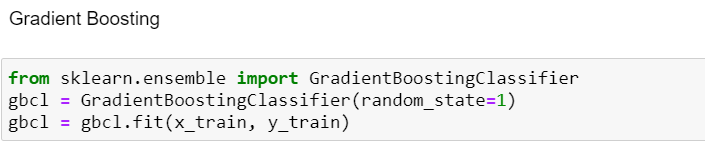
1.6) Model Tuning (2 pts) , Bagging ( 2.5 pts) and Boosting



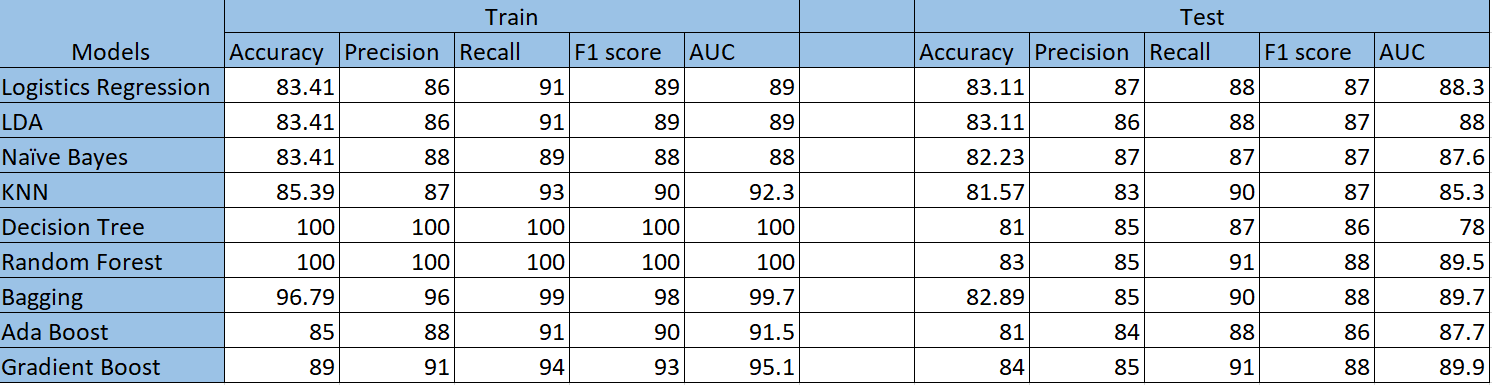








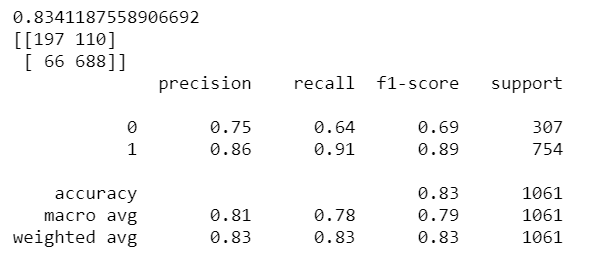
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model (4 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized

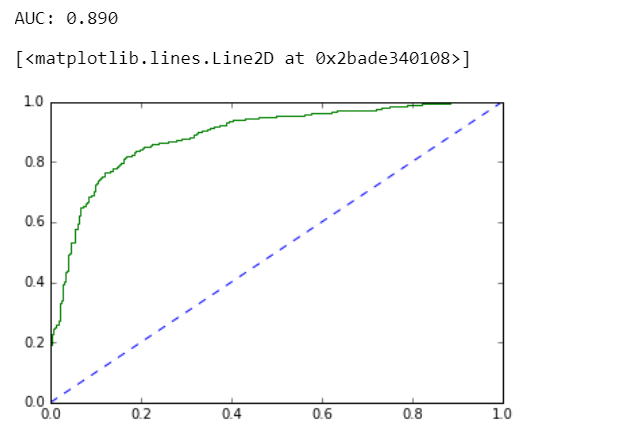


Gradient Boosting model is better as it has acceptable training scores and testing scores.

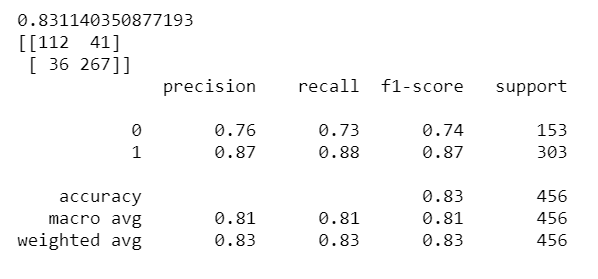
Logistics regression

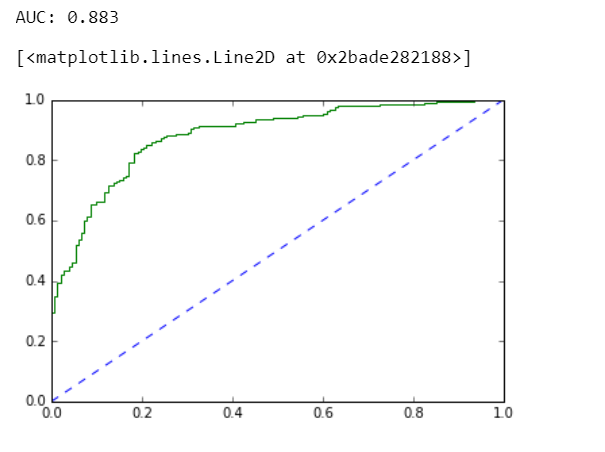
Training data metrics





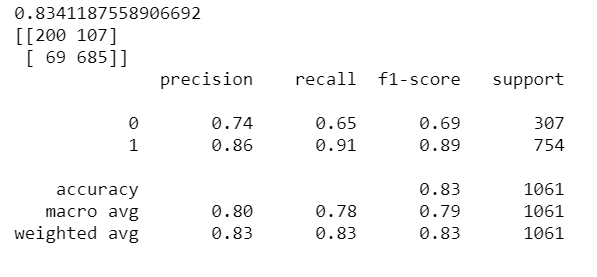
Testing data metrics

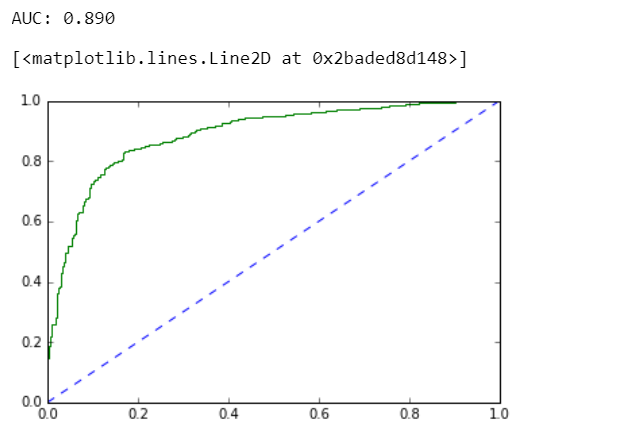




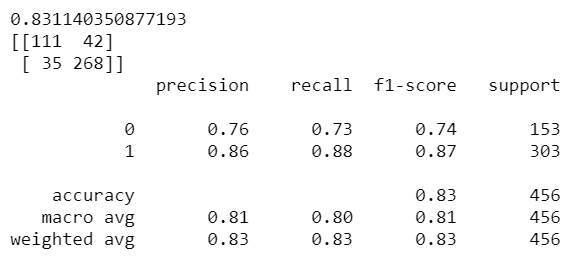
LDA

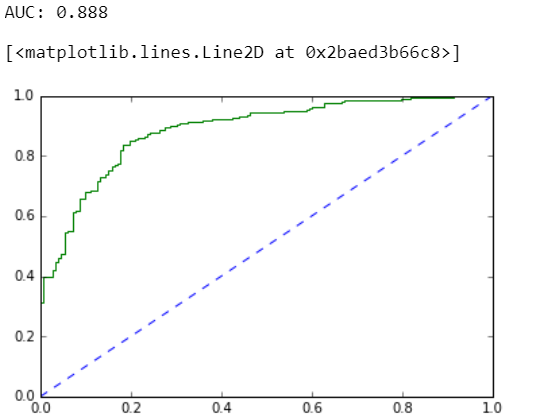
Training data metrics





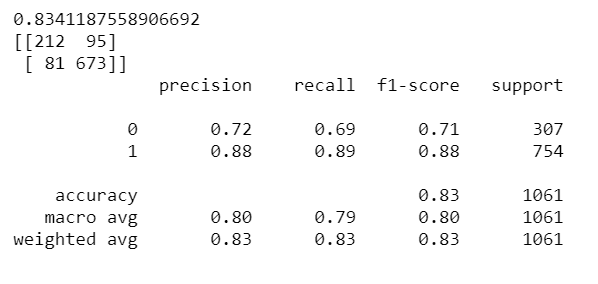
Testing data metrics

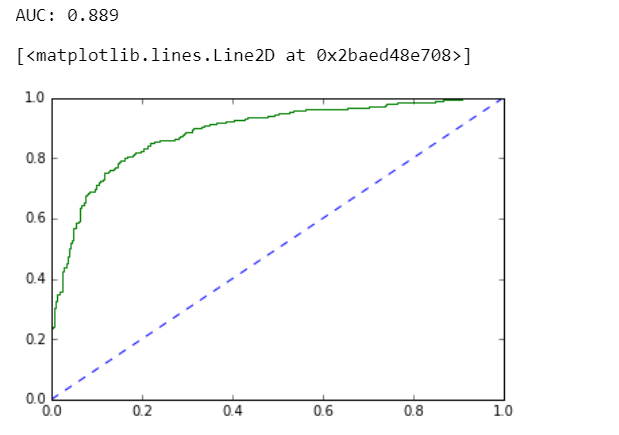




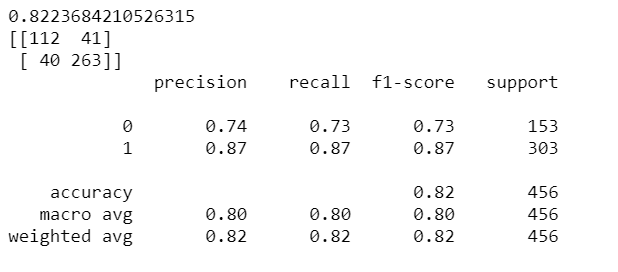
Naïve bayes

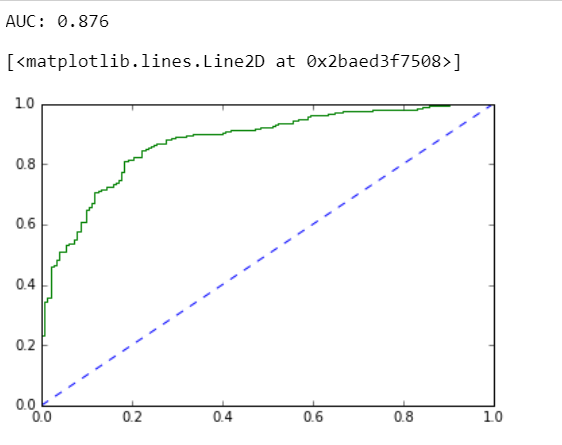
Training data metrics





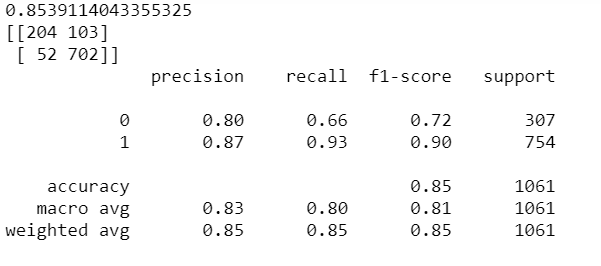
Testing data metrics

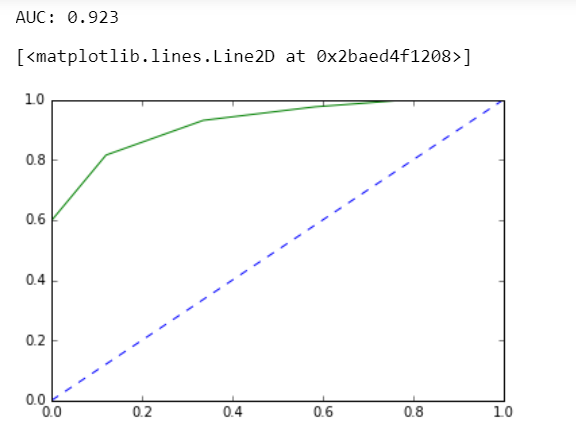




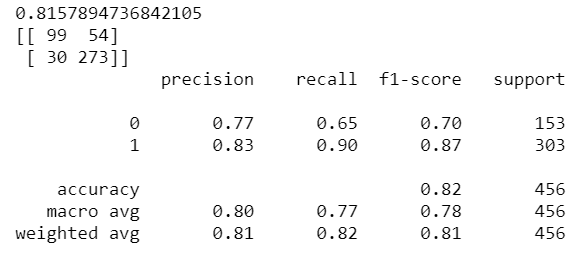
KNN

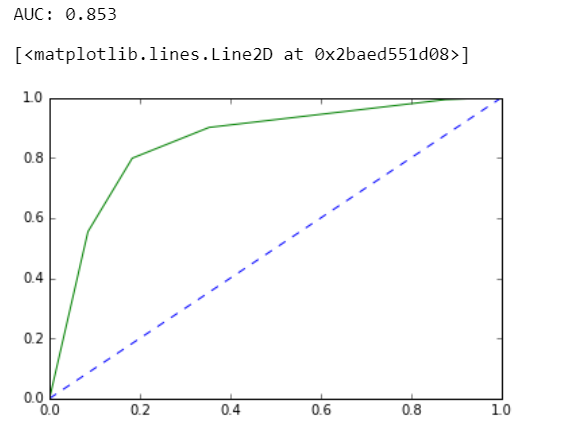
Training data metrics





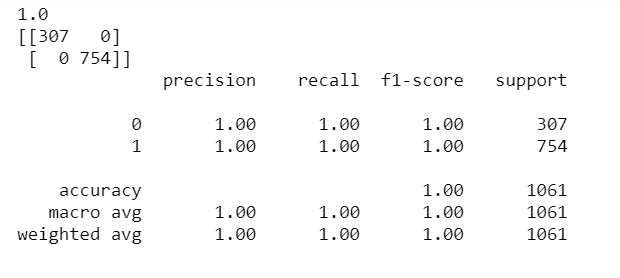
Testing data metrics

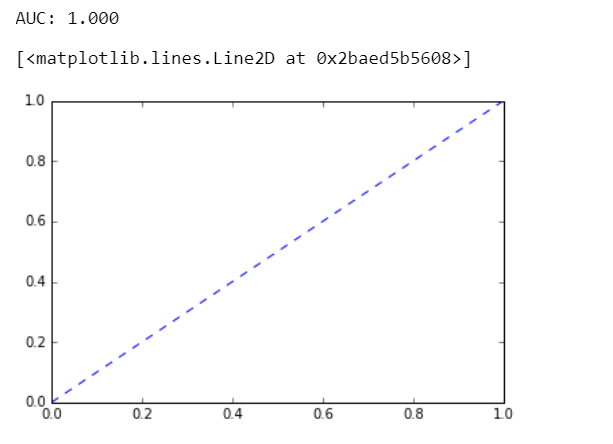




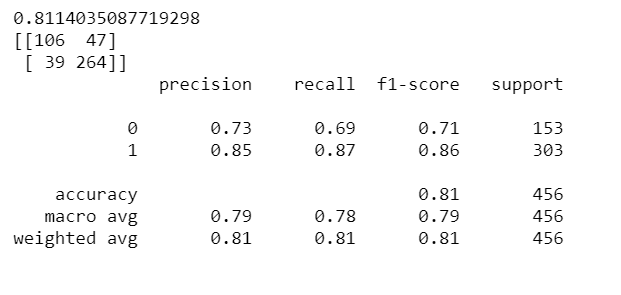
Decision Tree

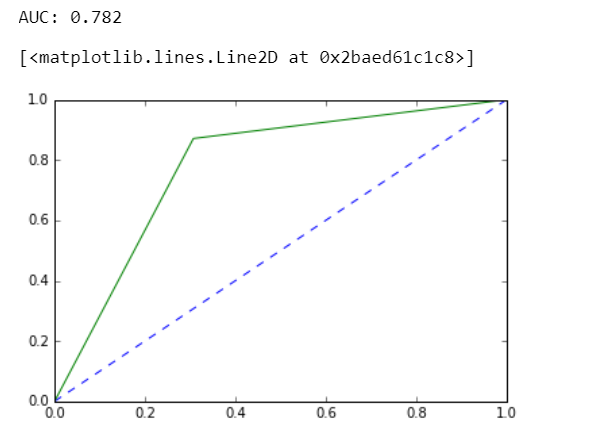
Training data metrics





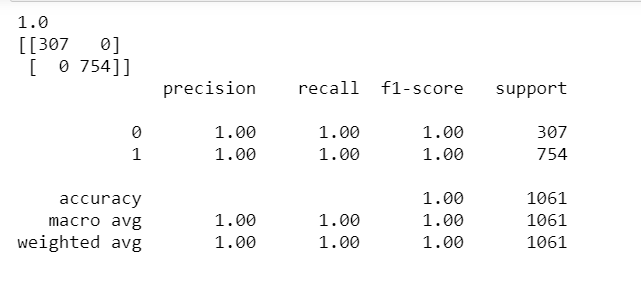
Testing data metrics

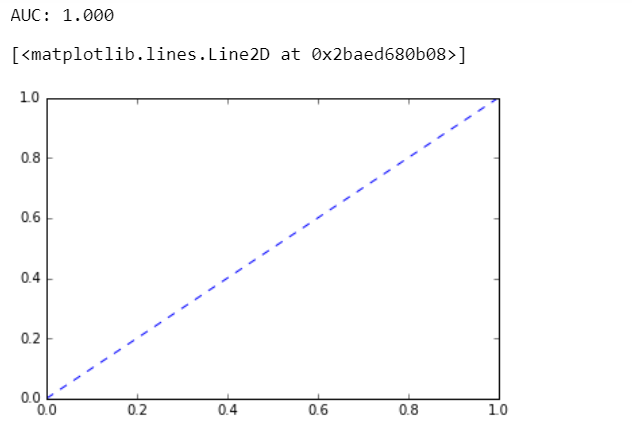




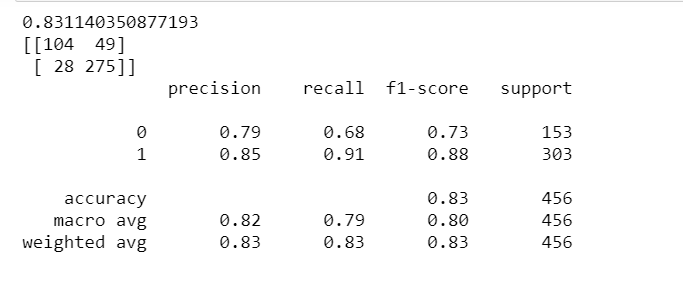
Random forest

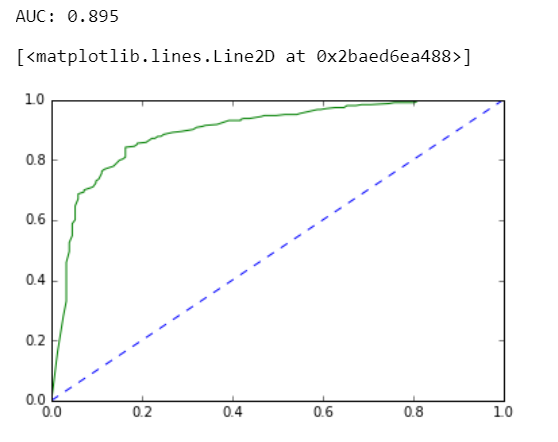
Training data metrics





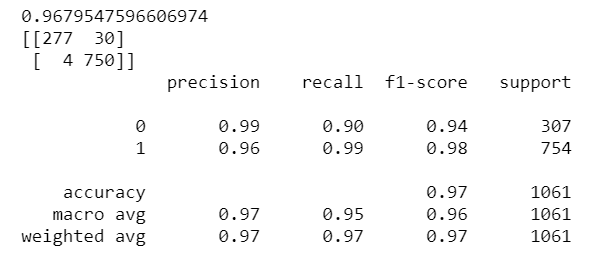
Testing data metrics

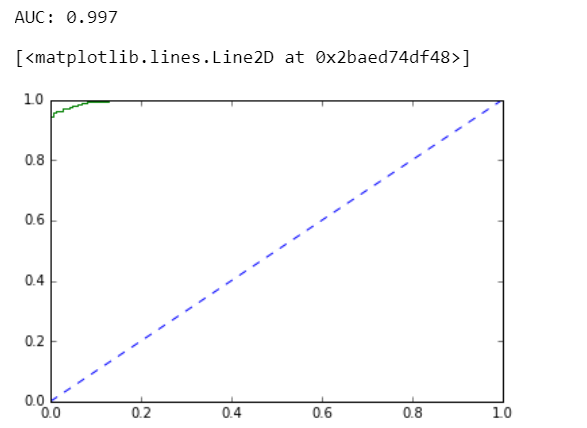




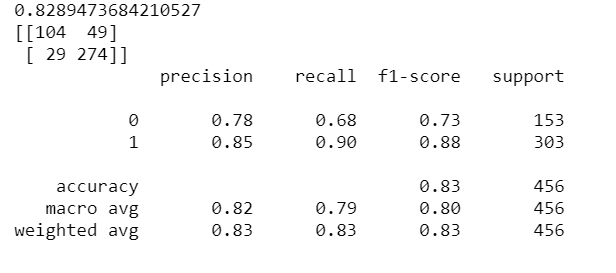
Bagging

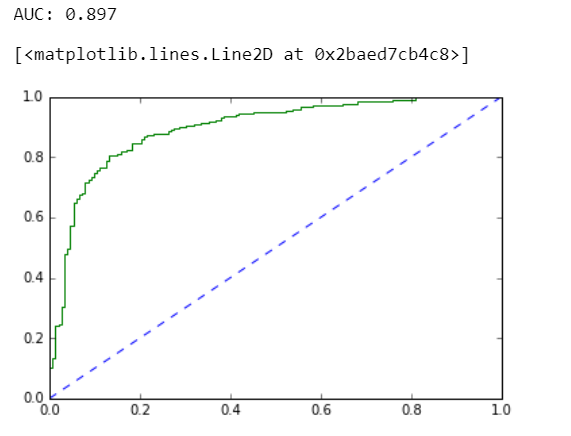
Training data metrics





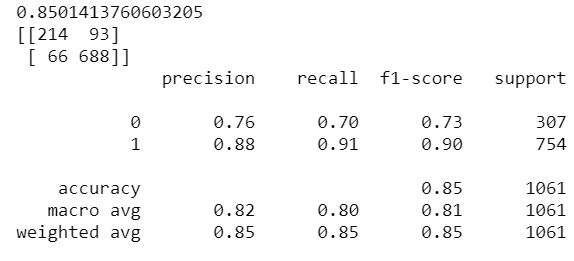
Testing data metrics

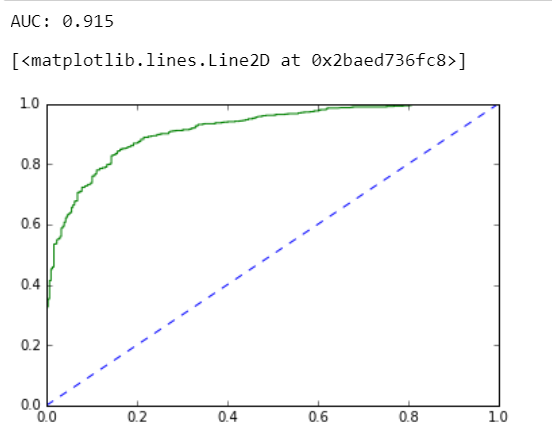




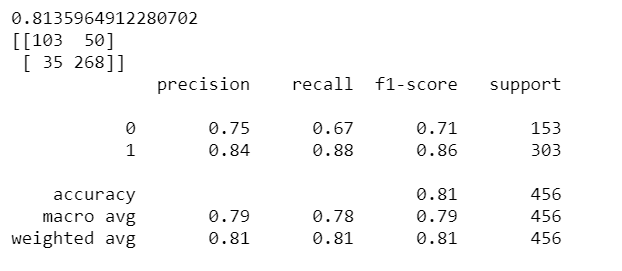
Ada Boost

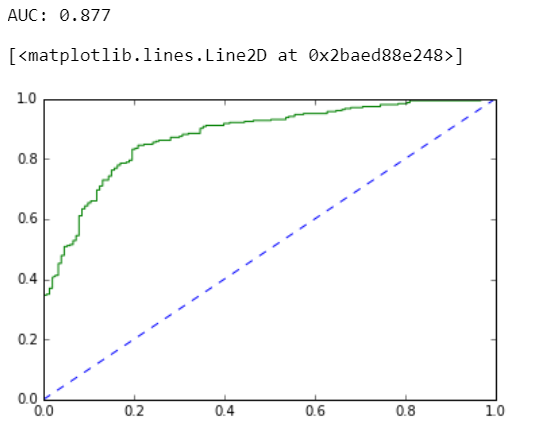
Training data metrics





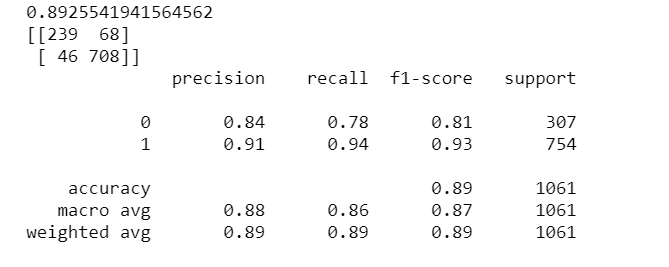
Testing data metrics

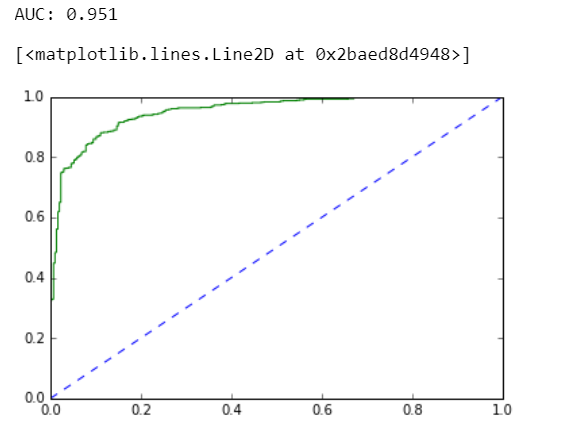




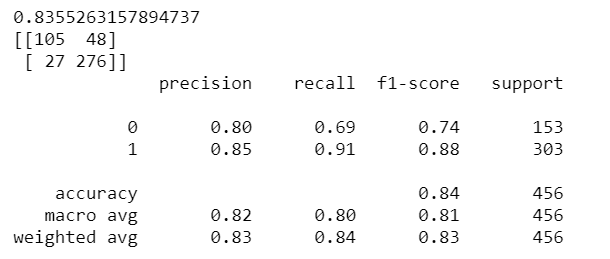
Gradient boost

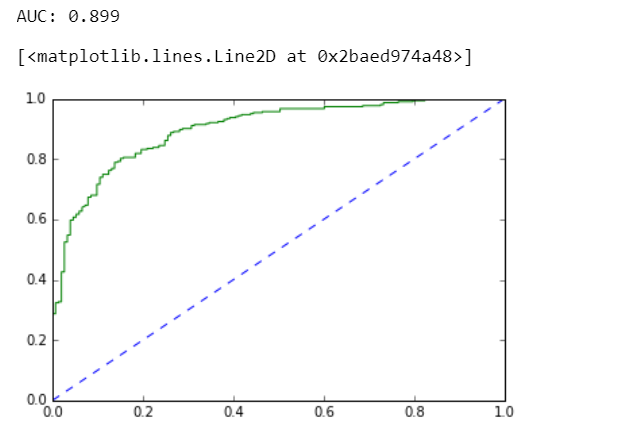
Training data metrics





Testing data metrics



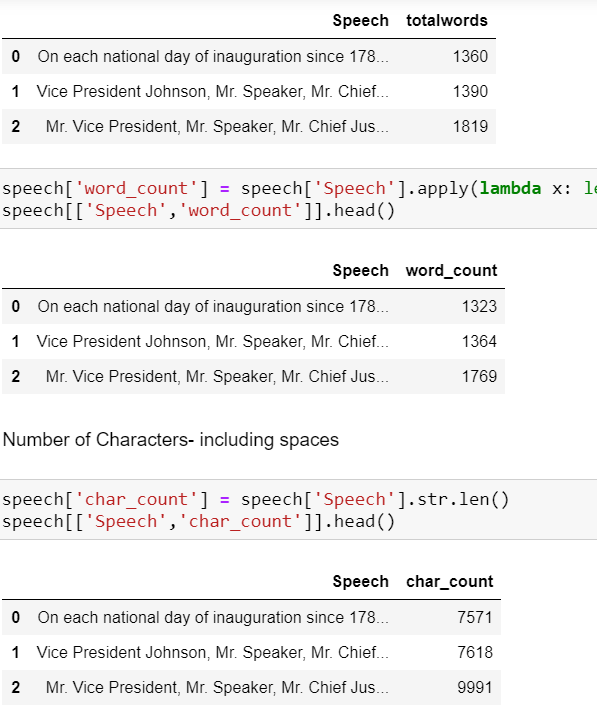


Problem 2 :

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

2.1) Find the number of characters, words and sentences for the mentioned documents.



2.2) Remove all the stopwords from the three speeches.



