

[MACHINE LEARNING]

**There are two case studies , one is U.S Election Prediction and another is Election Speech analysis of three American President Franklin Roosevelt , John F Kennedy and Richard Nixon**

TABLE OF CONTENTS

[Problem-1 3](#_Toc14464)

[1.1 Read the Data set. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks) 3](#_Toc27951)

[1.2 Perform Uni-variate and Bi-variate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks) 6](#_Toc3839)

[1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (4 Marks) 15](#_Toc9701)

[1.4 Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks) 16](#_Toc30818)

[1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results. (4 marks) 18](#_Toc22219)

[1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks) 19](#_Toc10144)

[Ada Boost Classifier 19](#_Toc26591)

[Classification report on train data 19](#_Toc27162)

[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. Final Model: Compare the models and write inference which model is best/optimized. (7 marks) 21](#_Toc18671)

[1.8 Based on these predictions, what are the insights? (5 marks) 23](#_Toc12433)

[Problem-2 24](#_Toc3900)

[2.1 Find the number of characters, words, and sentences for the mentioned documents. – 3 Marks 24](#_Toc3255)

[2.2 Remove all the stopwords from all three speeches. – 3 Marks 24](#_Toc31038)

[2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks 25](#_Toc12403)

[2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks [ refer to the End-to-End Case Study done in the Mentored Learning Session ] 25](#_Toc20838)

[Table 1 - Dataframe Head 5](#_Toc21128)

[Table 2 -Dataframe Description 5](#_Toc28302)

[Table 3 -Party wise vote count 8](#_Toc4889)

[Table 4 -Gender wise voters 9](#_Toc25995)

[Table 5 -voters view on economic condition 9](#_Toc13341)

[Table 6 -voters view on house hold economic cond 10](#_Toc19115)

[Table 7 -Voters rating on Blair 10](#_Toc24313)

[Table 8 -voters view on Hague 11](#_Toc22550)

[Table 9 -voters attitude towards EU 11](#_Toc4540)

[Table 10 -Voters estimation of their political knowledge 12](#_Toc1000)

[Table 11 -Distribution curve 13](#_Toc14335)

[Table 12 -Box plot with outliers 13](#_Toc52)

[Table 13 -Box plot without outlier 14](#_Toc19216)

[Table 14 -Pairplot 15](#_Toc31817)

[Table 15 -Correlation plot 16](#_Toc15695)

[Table 16 -Encoded dataframe 16](#_Toc2867)

[Table 17 24](#_Toc13527)

[Table 18 25](#_Toc21381)

[Table 19 -ROC curve of Adaboost-train 27](#_Toc22881)

[Table 20 -ROC curve of Adaboost-test 27](#_Toc27734)

[Table 21 -ROC curve of Gradient Boosting-train 28](#_Toc14694)

[Table 22 -ROC curve of Gradient boosting -test 28](#_Toc2954)

[Table 23 -KNN ROC curve-train 29](#_Toc2692)

[Table 24 -KNN ROC Curve test 29](#_Toc30048)

[Table 25 -LDA ROC curve -train 30](#_Toc22345)

[Table 26 -LDA ROC curve-test 30](#_Toc5697)

[Table 27 -LR ROC curve -train 31](#_Toc2399)

[Table 28 -LR ROC curve -test 31](#_Toc14734)

[Table 29 -Bagging ROC curve-train 32](#_Toc22718)

[Table 30 -Bagging ROC curve-test 32](#_Toc14255)

[Table 31 32](#_Toc22306)

# Problem-1

You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

## Read the Data set. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks)

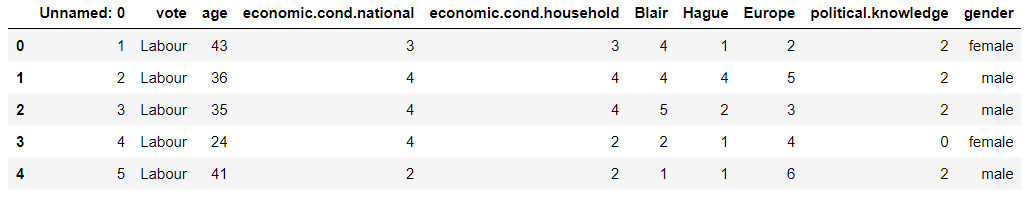


Table 1 - Dataframe Head

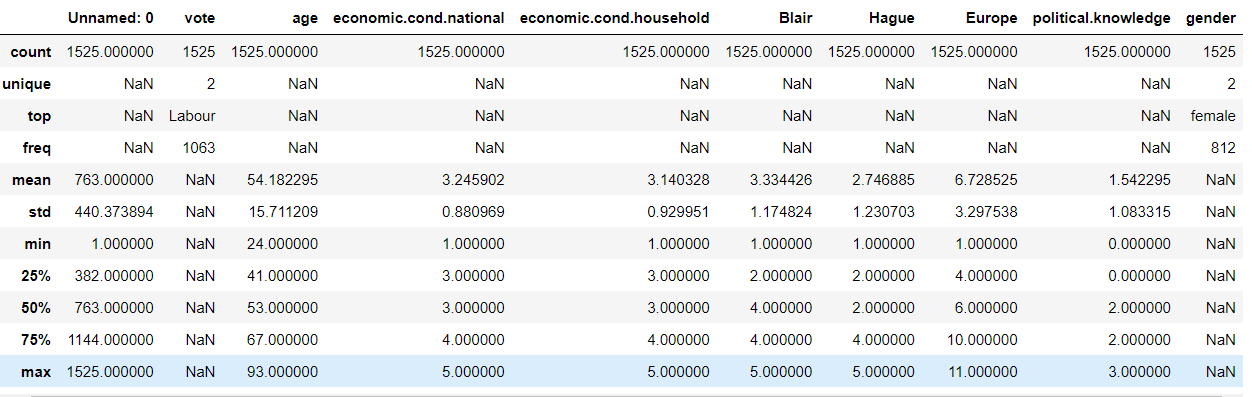


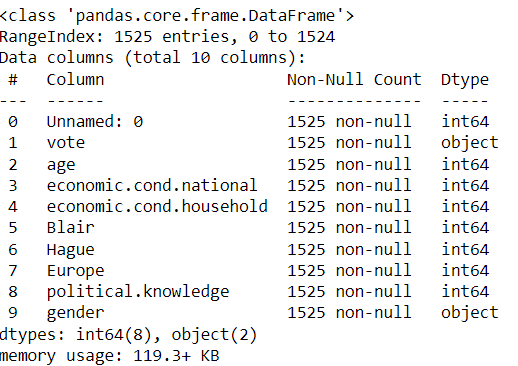
Table 2-Dataframe Description

The following assumptions can be made from the description of the Election data

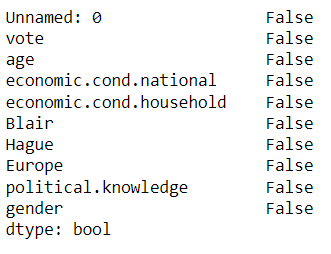
The data set gives the demographic , national and economical condition view points, and view points on respective party leaders of 1525 voters.

There are two political parties , "Labour" and "Conservative" and Blair and Hague are the respective leaders of these parties.

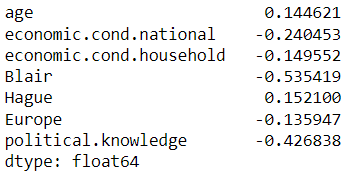
The age of the voters ranges between 24yrs to 93yrs. 75% voters are in the age of up to 67yrs.



There are two categorical variables one is "vote" and other is "gender" the rest are numerical.



There are no missing values in the data set.



Skew ness is a statistical term and it is a way to estimate or measure the shape of a distribution. It is an important statistical methodology that is used to estimate the asymmetrical behavior rather than computing frequency distribution. Skewness can be two types:

Symmetrical: A distribution can be called symmetric if it appears the same from the left and right from the center point.

Asymmetrical: A distribution can be called asymmetric if it doesn’t appear the same from the left and right from the center point.

Distribution on the basis of skewness value:

Skewness = 0: Then normally distributed.

Skewness > 0: Then more weight in the left tail of the distribution.

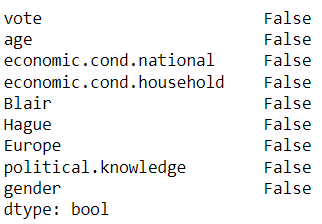
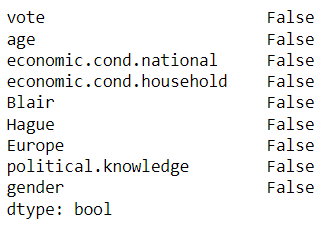
Skewness < 0: Then more weight in the right tail of the distribution.

After looking at the above skewness values of each variables in Election data following observations have been made:

1. All the variables are asymmetrically distributed
2. Only "Age" is positively skewed, which means that majority of the data distribution is on the left side of the mean. So most of the population age is more than the average age, so we are dealing with a comparatively older age population here.
3. All other variables are negatively skewed which means majority of the data distribution is on the right side of the mean.

## Perform Uni-variate and Bi-variate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

There are 1525 observations on 9 variables.



There are no NaN or missing values in the data set.

### Uni-variate Analysis

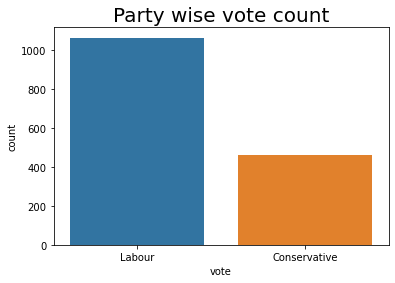


Table 3-Party wise vote count

Here as we can see there is a huge difference between the Labour party votes and conservative party votes. Our target variable is "Vote" and seeing this kind of difference between the two classes, we can say that the data set is having a class imbalance problem.



Table 4-Gender wise voters

Number of female voters is more than the male voters.

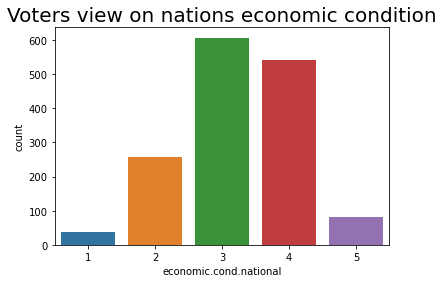


Table 5-voters view on economic condition

Most of the voters have rated 3 to nation's economic condition

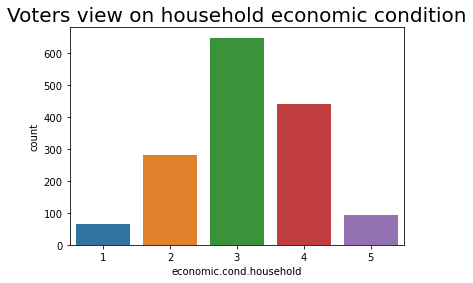


Table 6-voters view on house hold economic cond

Most of the voters have given 3 rating to household economic condition.

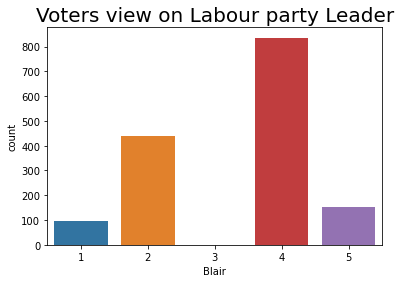


Table 7-Voters rating on Blair

Most of the people have given 4 rating to the leader of Labour party

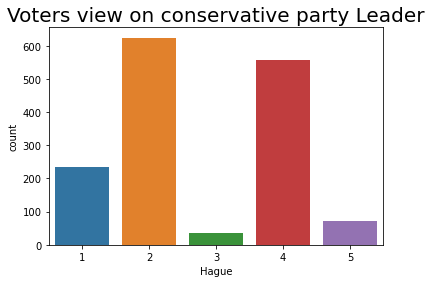


Table 8-voters view on Hague

Most of the voters have given 2 rating to the leader of conservative party leader.

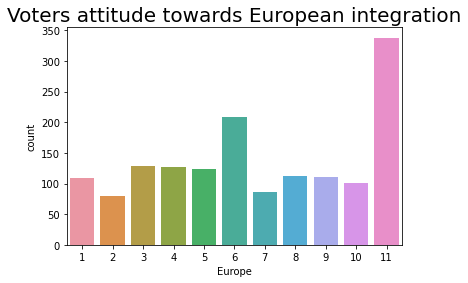


Table 9-voters attitude towards EU

Here higher the score, higher is the voters eurosceptic sentiment i.e they oppose increasing power of european union in their nation.

So, as it can be seen from the above graphs, most of the voters have given 11 rating which means European union is a boiling issue in this election and most voters have opposed the increasing power of EU in their nation.

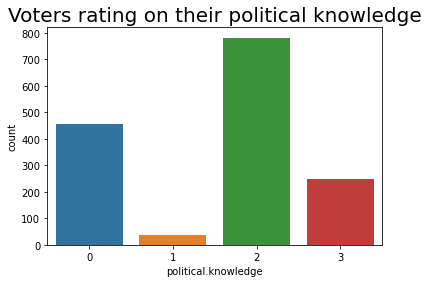


Table 10-Voters estimation of their political knowledge

Most of the voters have rated themselves 2 in political knowledge.

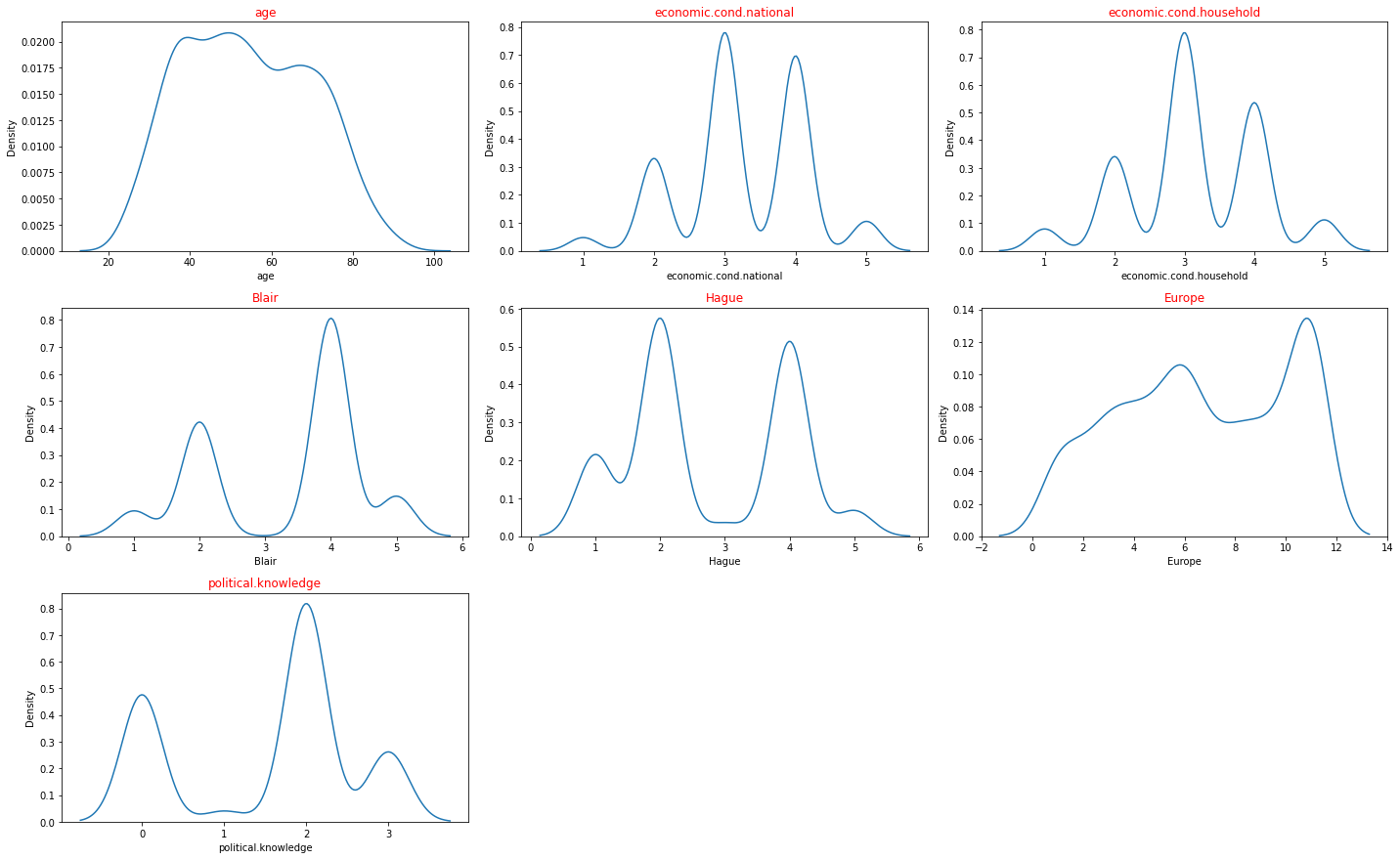


Table 11-Distribution curve

What we have interpreted from count plot, can also be seen here in distribution plots of each variable.

Here the distribution of "Age" shows that most of the voters are in the age bracket of 24 to 67 yrs.

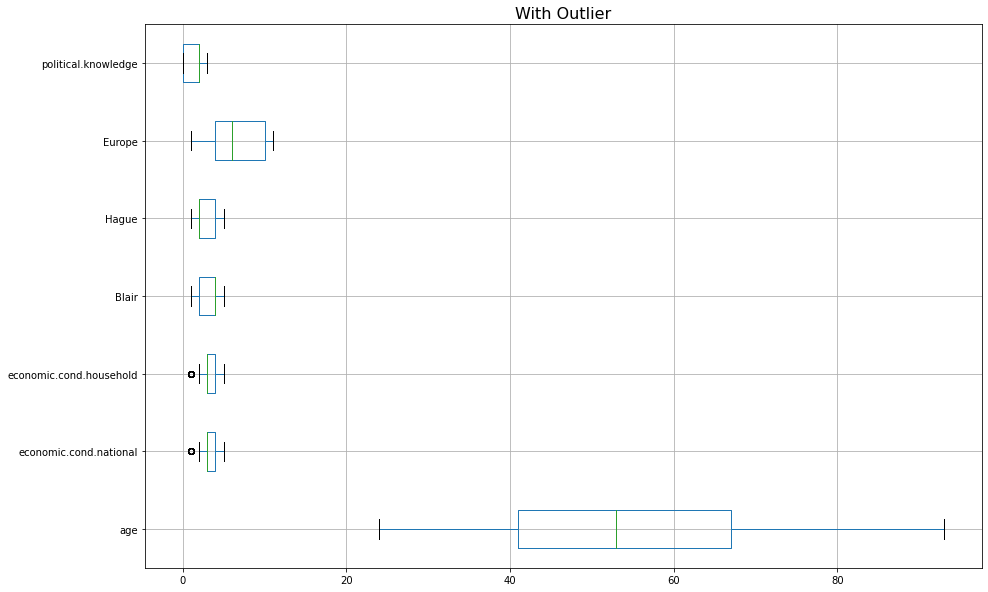


Table 12-Box plot with outliers

There are outliers observed in "economic condition household" and "economic condition national" variables.

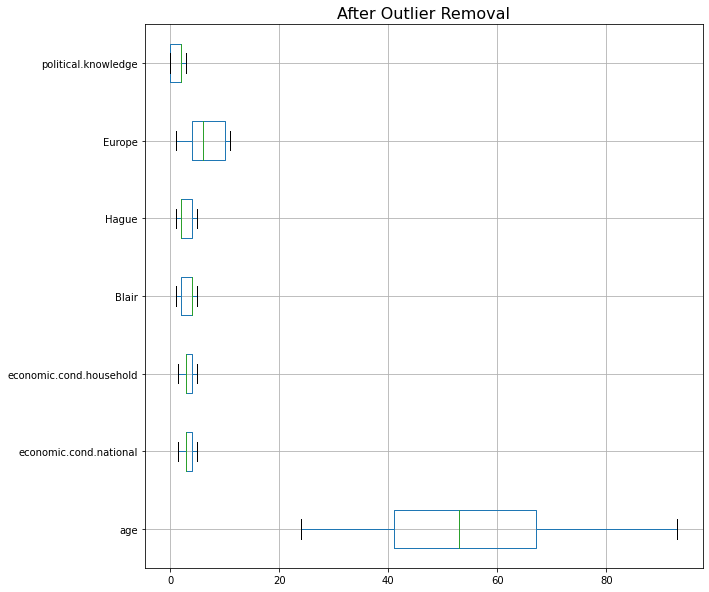


Table 13-Box plot without outlier

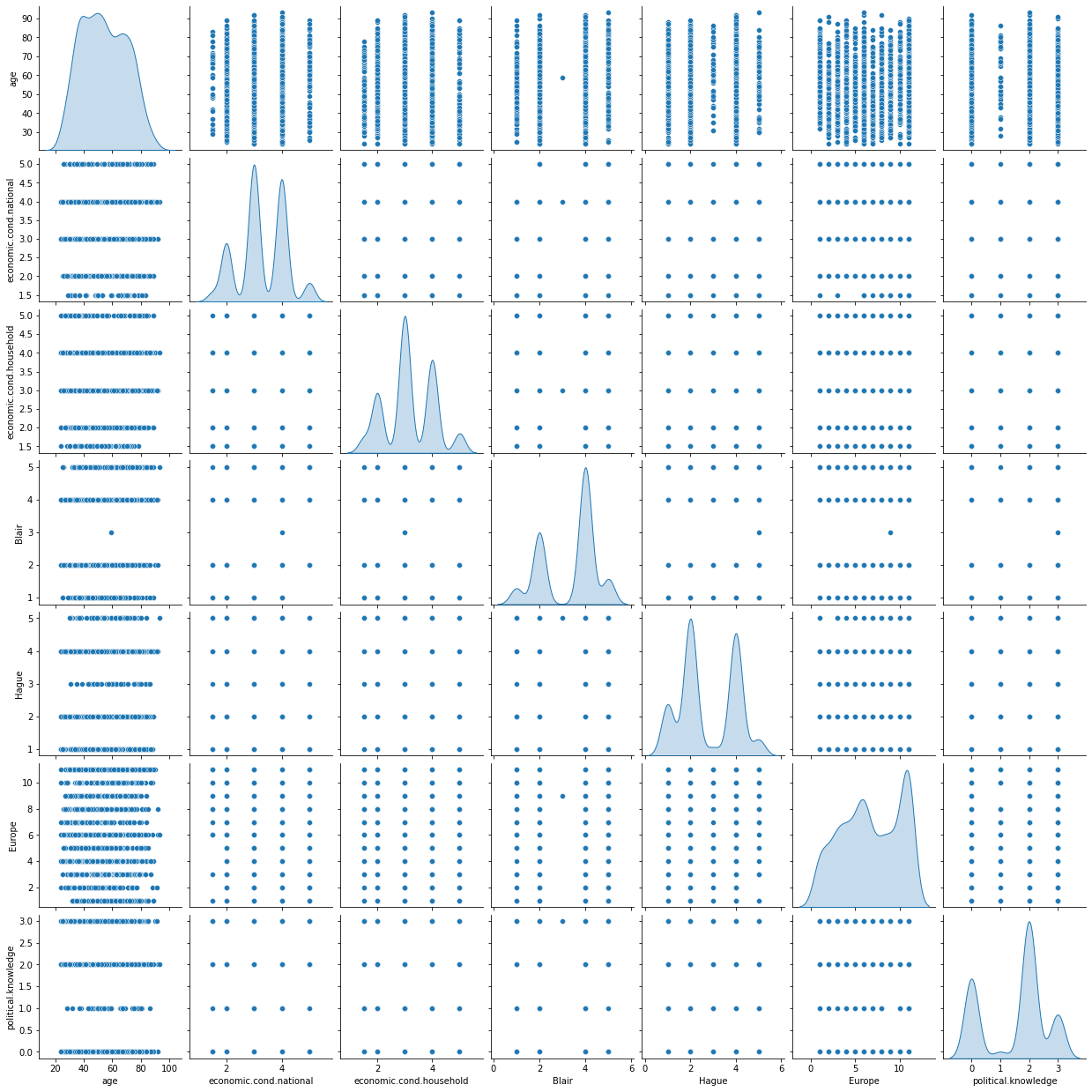


Table 14-Pairplot

from the above pair plot it can be observed that there are no correlations between the variables.

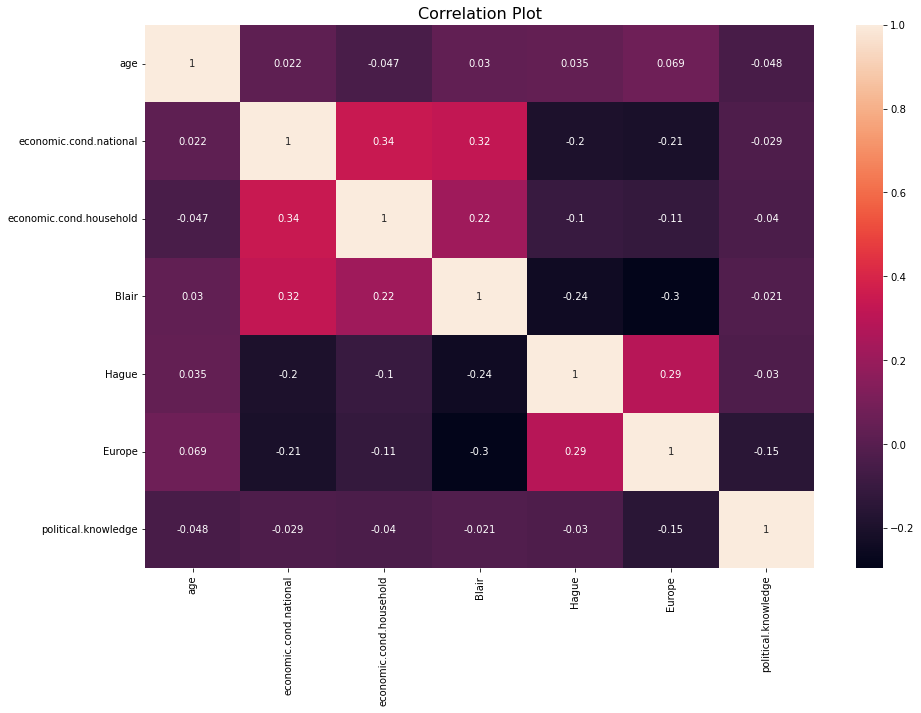


Table 15-Correlation plot

There are no major correlations observed.

## Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (4 Marks)

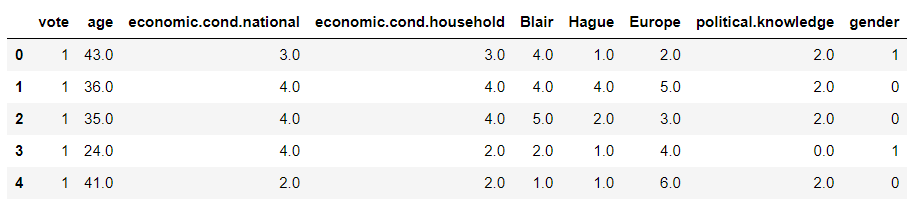
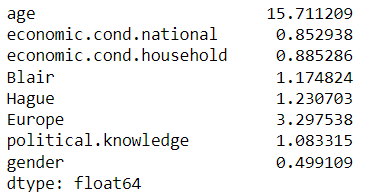


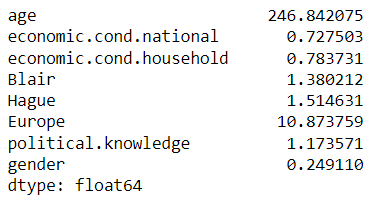
Table 16-Encoded dataframe

“Gender” has been encoded male 0 and female 1 and Vote as Labour : 1 and Conservative: 0



The Standard Deviation is a statistic that indicates how much variance or dispersion there is in a group of statistics. A low Standard Deviation means that the value is close to the mean of the set (also known as the expected value), and a high Standard Deviation means that the value is spread over a wider area.

Looking at the above std values we can say that the "age" variable is having highest std and also not in the same scale so there is a need of scaling.

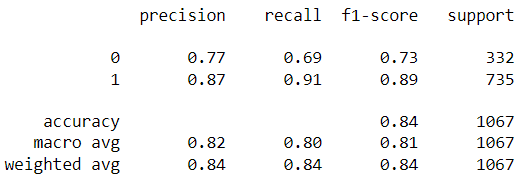


The variance is a numerical value that represents how broadly individuals in a group may change. The variance will be larger if the individual observations change largely from the group mean and vice versa.

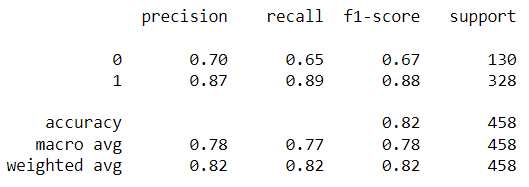
Here looking at the above variances of each variables we can say that some of the variables like "Age" and "Europe" are having high variance values , so we have to scale the data to bring all the variables on the same scale.

## Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks)

**Logistics Regression**

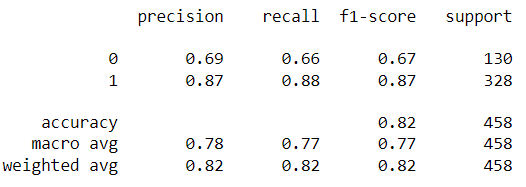


Logistics regression applied and we got the above classification report on train data.

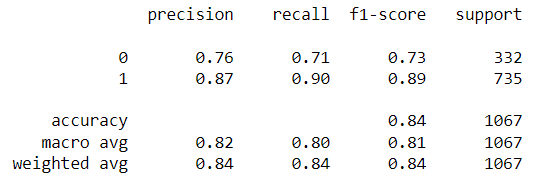


Test data classification report of Logistics regression is plotted above.

**Linear Discriminant Analysis**



LDA classification report on test data is plotted above.

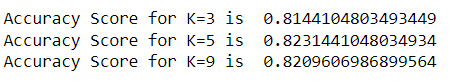


LDA classification report on train data is plotted above.

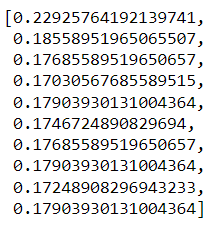
## Apply KNN Model and Naive Bayes Model. Interpret the results. (4 marks)

**KNN Model**

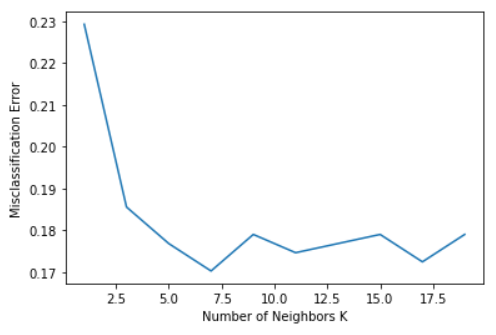
Different K values have been taken and accuracy score has been calculated.



Miss-classification error w.r.t different K values have been plotted below.

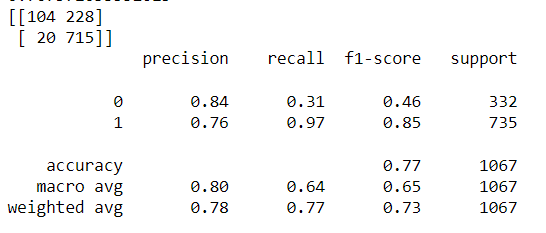


We can plot a graph taking above values of MCE and K

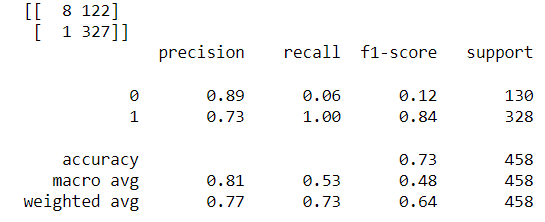


K=3 is best.

Classification report of KNN using train data

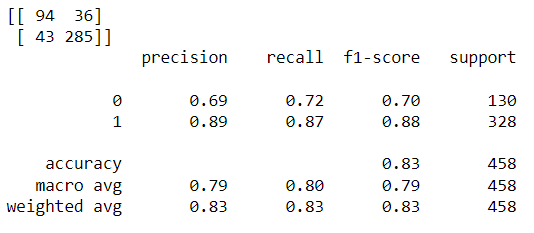


Classification report of KNN using test data

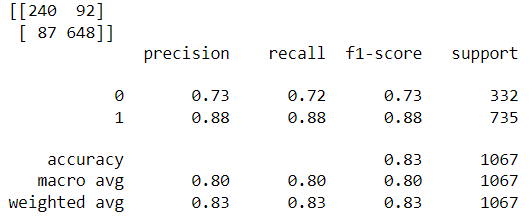


**Naive Bayes**

Naive Bayes applied and the classification report on train data is plotted below.



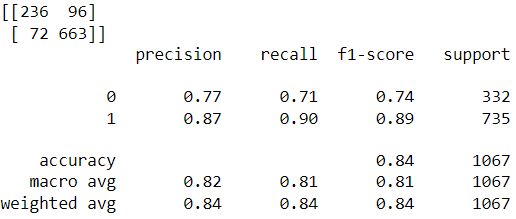
Naive bayes classification report on test data.



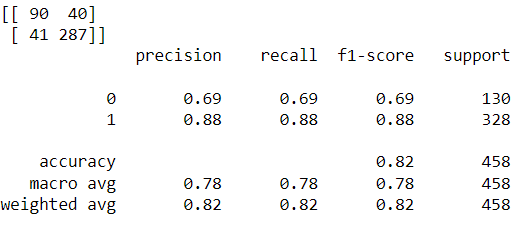
## Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks)

### Ada Boost Classifier

Classification report on train data

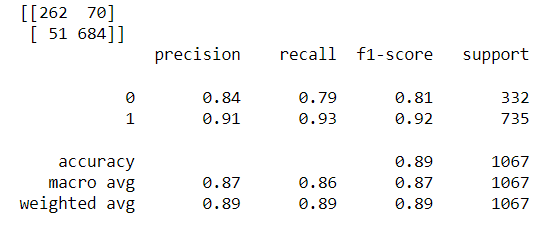


AdaBoost Classification report on test data.

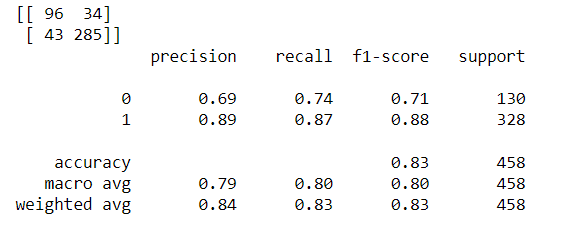


### Gradient Boosting

Classification report on train data.

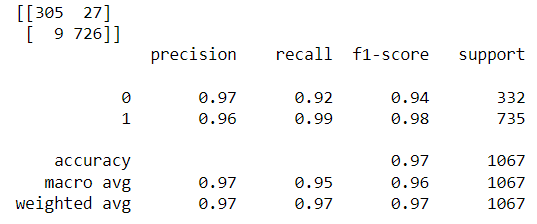


Classification report on test data.

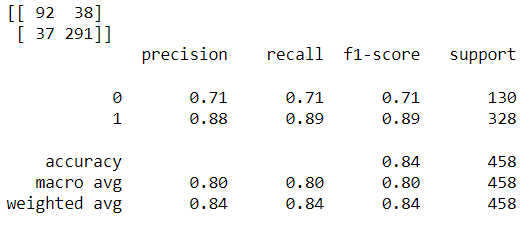


### Bagging

Classification report of Bagging classifier on train data.



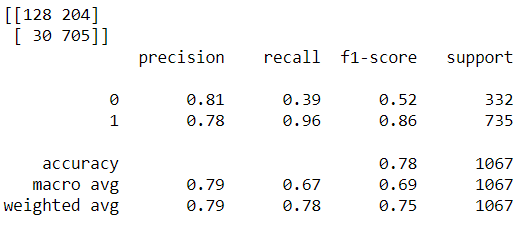
Classification report on test data.



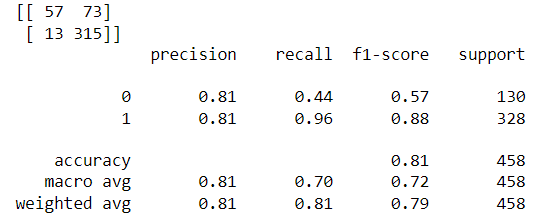
### Hyper-parameter Tuning

#### AdaBoost Classifier

Classification report of Hyper tuned model with train data.

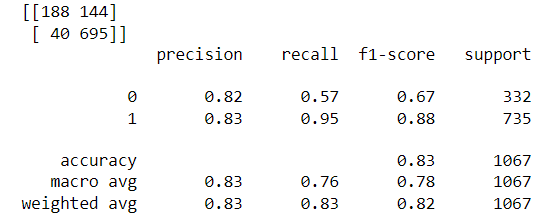


Classification report of Hyper tuned model with test data.



#### Gradient Boosting

Classification report of hyper tuned model using train data.



Classification report of hyper tuned model using test data.

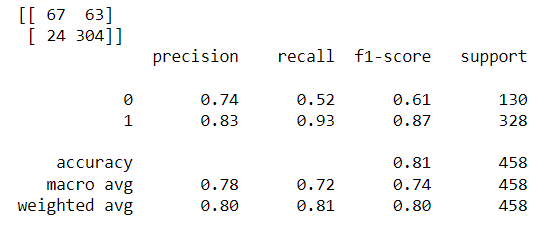
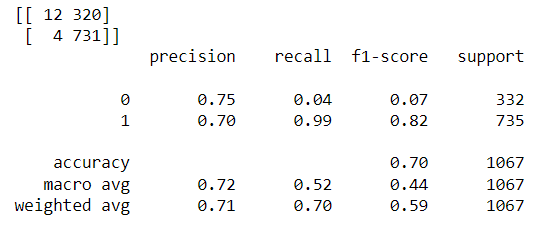


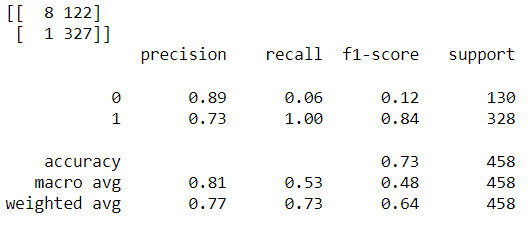
Table 17

#### KNN

Classification report of KNN model after hyper tuning using train data



Classification report of KNN model after hyper tuning using test data



#### Linear Discriminant Analysis

Classification report after hyper tuned LDA model using train data

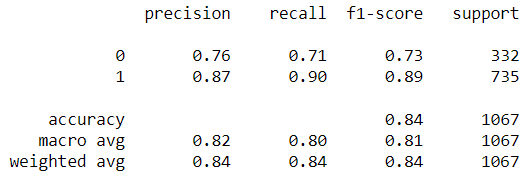
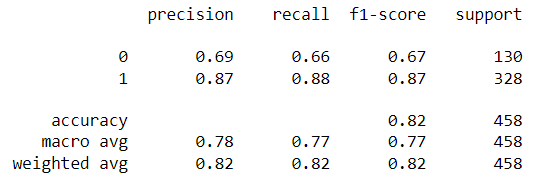


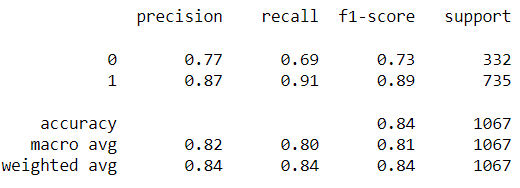
Table 18

Classification report after hyper tuned LDA model using test data

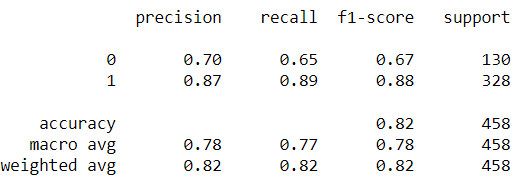


#### Logistics Regression

Classification report of hyper tuned LR model on train data.



Classification report of hyper tuned LR model on test data.



## 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. Final Model: Compare the models and write inference which model is best/optimized. (7 marks)

### All hyper tuned model AUC score and ROC curve

#### AdaBoost

AUC Score - 0.865

ROC curve in train data

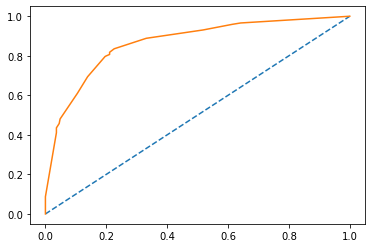


Table 19-ROC curve of Adaboost-train

**ROC curve in test data**

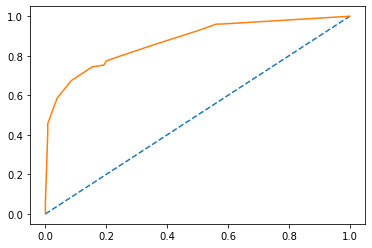


Table 20-ROC curve of Adaboost-test

#### Gradient Boosting

AUC score : 0.90

ROC Curve in train data

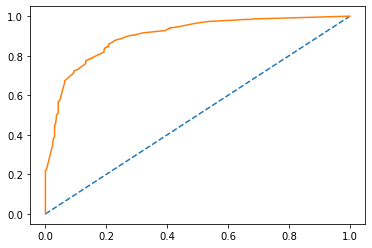


Table 21-ROC curve of Gradient Boosting-train

ROC Curve in test data

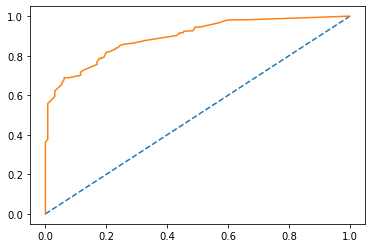


Table 22-ROC curve of Gradient boosting -test

#### KNN

AUC score : 0.792

ROC Curve using train data

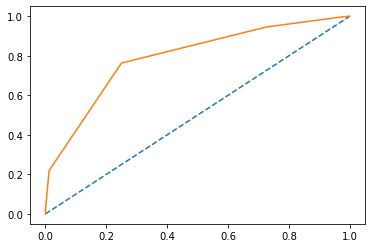


Table 23-KNN ROC curve-train

ROC Curve using test data

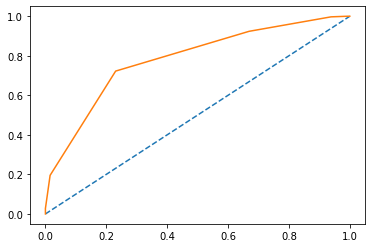


Table 24-KNN ROC Curve test

**LDA**

**AUC score : 0.749**

**ROC Curve using train data**

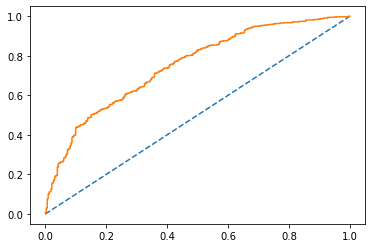


Table 25-LDA ROC curve -train

ROC Curve using test data

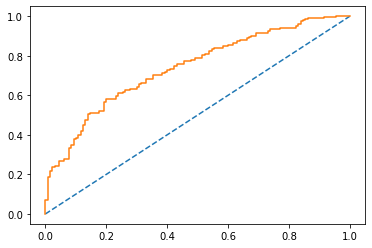


Table 26-LDA ROC curve-test

#### Logistics Regression

AUC Score : 0.76

ROC Curve using train data

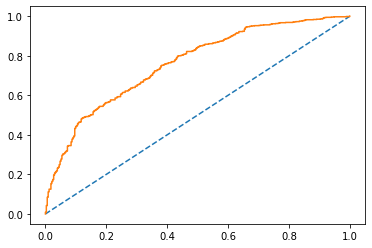


Table 27-LR ROC curve -train

ROC Curve using test data

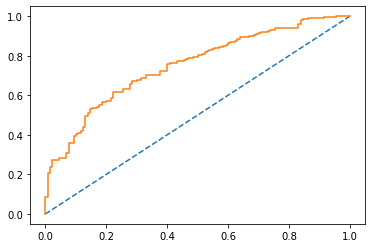


Table 28-LR ROC curve -test

#### Bagging

AUC Score : 0.997

ROC Curve using train data

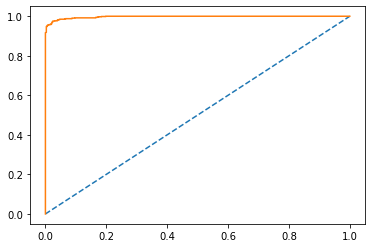


Table 29-Bagging ROC curve-train

ROC Curve using test data

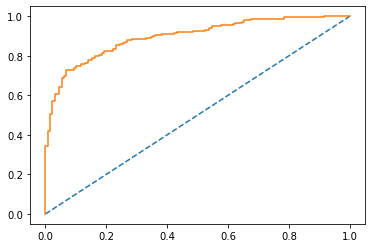
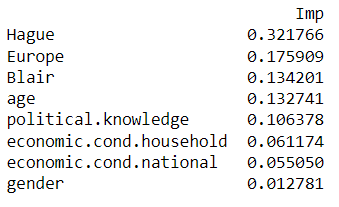


Table 30-Bagging ROC curve-test

Table 31

**Feature Importance**



## Comparison of All model performance

Class of interest is 1 (I.e Labour=1)

Recall refers to the percentage of total relevant results correctly classified by the algorithm and hence we will compare Recall of class "1" for all models.

F1-score metric is to find an equal balance between precision and recall, which is extremely useful in most scenarios when we are working with imbalanced data sets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model (Regular) | Model  (Hyper-tuned) | Recall | | F1 Score | | Remarks |
| Train | Test | Train | Test |  |
| Logistics Regression |  | 0.91 | 0.89 | 0.89 | 0.88 |  |
| Linear Discriminant Analysis |  | 0.90 | 0.88 | 0.87 | 0.89 |  |
| KNN |  | 0.97 | 0.97 | 0.85 | 0.86 |  |
| Naive Bayes |  | 0.88 | 0.87 | 0.88 | 0.88 |  |
| Bagging |  | 0.99 | 0.89 | 0.98 | 0.89 |  |
| AdaBoost Classifier |  | 0.90 | 0.88 | 0.89 | 0.88 |  |
| Gradient Boosting |  | 0.93 | 0.87 | 0.92 | 0.88 | Best model |
|  | AdaBoost Classifier | 0.96 | 0.96 | 0.86 | 0.88 |  |
|  | Gradient Boosting | 0.95 | 0.93 | 0.88 | 0.87 |  |
|  | KNN | 0.99 | 1 | 0.82 | 0.84 | Over fitting problem |
|  | Linear Discriminant Analysis | 0.90 | 0.88 | 0.89 | 0.87 |  |
|  | Logistics Regression | 0.91 | 0.89 | 0.89 | 0.88 |  |

Gradient boosting is the best model considering the recall score and F1 score values.

## Based on these predictions, what are the insights? (5 marks)

Insights :

1. There are more supporters of Labour party than conservative party.
2. Voters view on Labour party leader Blair is very positive.
3. There are more female voters than male voters.
4. European union integration can be an issue in the election so the party which is supporting this issue can have better chances of winning.
5. The party which has more voters base in females are likely to win.
6. Issues related to women safety and female LFPR can be given importance.
7. Voters don’t give economic condition a priority.

# 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:

President Franklin D. Roosevelt in 1941

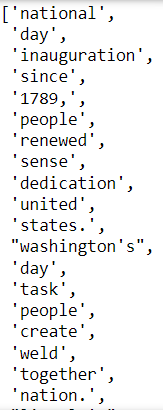
President John F. Kennedy in 1961

President Richard Nixon in 1973

## 2.1 Find the number of characters, words, and sentences for the mentioned documents. – 3 Marks

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total No of Characters | Total No of Words | Total No of Sentences |
| Roosevelt | 7571 | 1536 | 67 |
| Kennedy | 7618 | 1546 | 52 |
| Nixon | 9991 | 2028 | 68 |

## 2.2 Remove all the stopwords from all three speeches. – 3 Marks



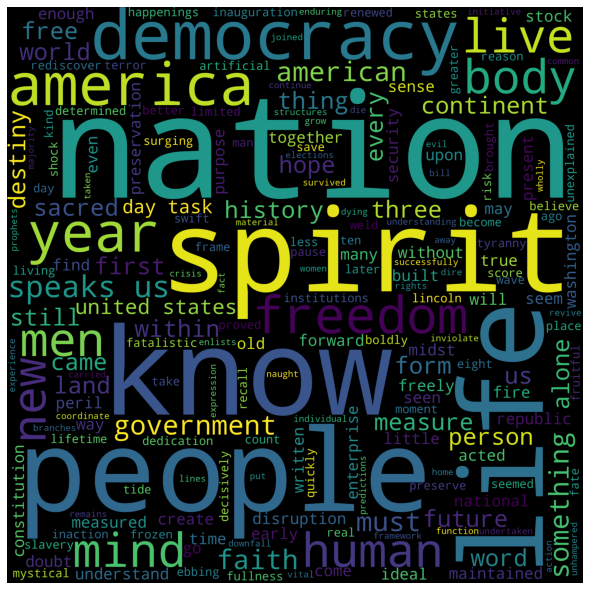
All stopwords removed

## 2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks

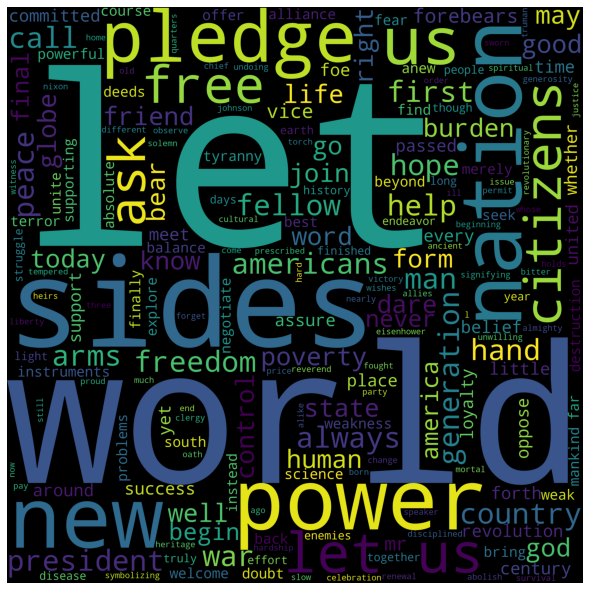
|  |  |  |
| --- | --- | --- |
|  | Most frequent word | Frequency |
| Roosevelt | Nation | 12 |
| Kennedy | Let | 16 |
| Nixon | us | 26 |

## 2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks [ refer to the End-to-End Case Study done in the Mentored Learning Session ]

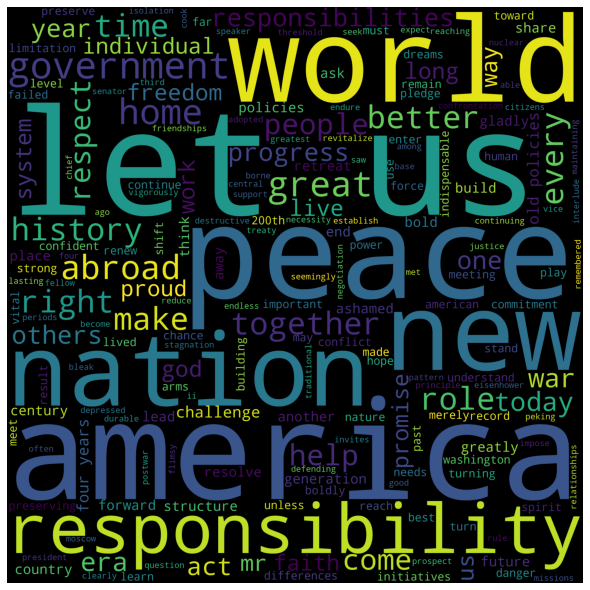
**Word Cloud for Roosevelt**



**Word cloud of Kennedy**

****

**Word cloud of Nixon**

****