**Business Report**

**Machine Learning**

Contents

[**Problem 01:** 4](#_Toc121675095)

[Data Dictionary: 4](#_Toc121675096)

[Project Goal: 4](#_Toc121675097)

[Exploratory Data Analysis 4](#_Toc121675098)

[Converting Object variables to categorical variables 14](#_Toc121675099)

[Scaling of Data 15](#_Toc121675100)

[Train-Test Data Split 15](#_Toc121675101)

[Logistic Regression 15](#_Toc121675102)

[Linear Discriminant Analysis 18](#_Toc121675103)

[KNN Model Analysis 20](#_Toc121675104)

[Naïve Bayes Model Analysis 23](#_Toc121675105)

[Bagging with Random Forest Model: 25](#_Toc121675106)

[Gradient Boosting Model: 28](#_Toc121675107)

[Ada Boosting Model: 30](#_Toc121675108)

[Final Model: Comparing all the models 32](#_Toc121675109)

[Overall Best/Optimized model: 32](#_Toc121675110)

[Problem 02: 33](#_Toc121675111)

[2.1 Find the number of characters, words, and sentences for the mentioned documents. 33](#_Toc121675112)

[2.2 Remove all the stop words from all three speeches 33](#_Toc121675113)

[Libraries used: 33](#_Toc121675114)

[Stemming: 33](#_Toc121675115)

[Most common words in Roosevelt speech after removing stopwords 33](#_Toc121675116)

[Most common words in Kennedy speech after removing stopwords 34](#_Toc121675117)

[Most common words in Nixon speech after removing stopwords 34](#_Toc121675118)

[Top three words in Roosevelt's speech(after removing the stopwords) 34](#_Toc121675119)

[Top three words in Kennedy's speech(after removing the stopwords 34](#_Toc121675120)

[Top three words in Nixon's speech(after removing the stopwords 34](#_Toc121675121)

[Roosevelt speech 35](#_Toc121675122)

[Kennedy speech 36](#_Toc121675123)

[Nixon speech 37](#_Toc121675124)

[Conclusion: 37](#_Toc121675125)

List of all the tables and figures in the project:

|  |
| --- |
| List of all figures and tables: |
| Exploratory Data Analysis |
| Dataset (first 5) Sample |
| Dataset (last 5) Sample |
| Shape of the dataset |
| Datatypes in the dataset |
| Statistics of the dataset |
| Missing Values |
| Duplicated Values |
| Data Skewness Values |
| Unique values and their count in dataset |
| Boxplot and frequency distribution of other variables |
| Count plot of Gender |
| Count plot of vote |
| Hague versus political knowledge |
| Blair versus political knowledge |
| Vote versus age with gender |
| Boxplot and Strip plot of Hauge and Blair versus age |
| Before Outliers Treatment |
| After Outliers Treatment |
| Correlation Heatmap |
| Pair Plot |
| Logistic Regression |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Linear Discriminant Analysis |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| KNN Model |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Naïve Bayes Model |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Bagging with Random Forest |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Gradient Boosting Model |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Ada Boosting |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| WordCloud |
| WordCloud with Roosevelt Speech |
| WordCloud with Kennedy Speech |
| WordCloud with Nixon Speech |

# **Problem 01:**

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.

## Data Dictionary:

1. vote: Party choice: Conservative or Labour

2. age: in years

3. economic.cond.national: Assessment of current national economic conditions, 1 to 5.

4. economic.cond.household: Assessment of current household economic conditions, 1 to 5.

5. Blair: Assessment of the Labour leader, 1 to 5.

6. Hague: Assessment of the Conservative leader, 1 to 5.

7. Europe: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment.

8. political.knowledge: Knowledge of parties' positions on European integration, 0 to 3.

9. gender: female or male.

## Project Goal:

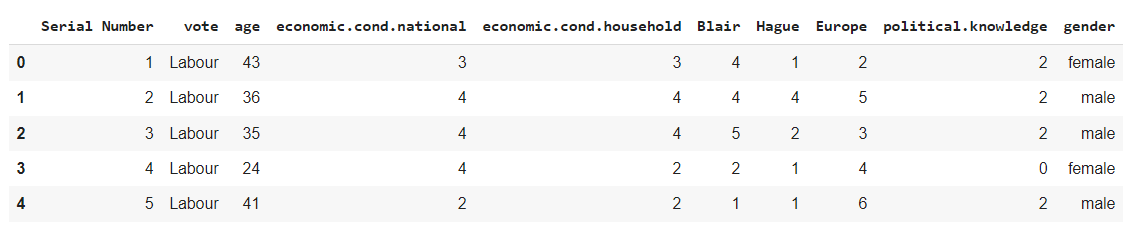
This report's goal is to use Python to analyse the dataset "Election Data.xlsx" and eventually produce insights about it. The core elements of this exploration summary report will be organized as follows:

* Using the dataset as input in a Jupyter notebook.
* Recognizing the dataset's structure.
* Exploratory Data Analysis.
* Graphical Interpretation; and
* Machine-Learning Models for Prediction
* Insights or Inferences from the data

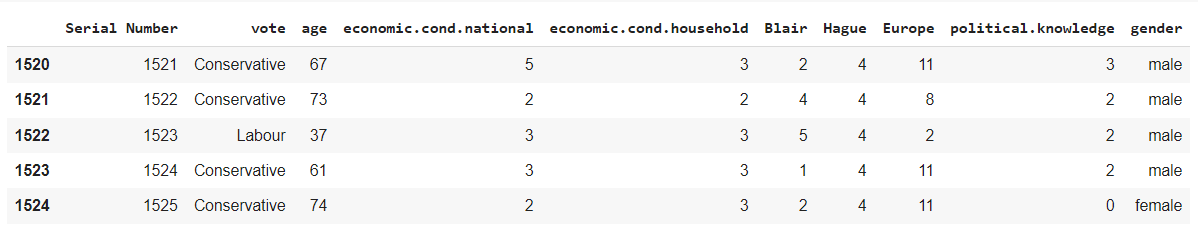
1.1 Read the dataset. Do the descriptive statistics and do the null value condition check.

### Exploratory Data Analysis

#### Sample of the dataset:



Dataset (first 5) Sample



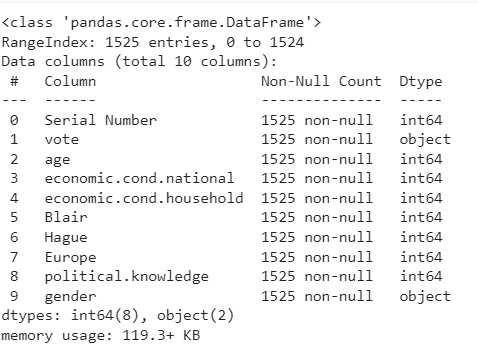
Dataset (last 5) Sample

* The first column was renamed as Serial number for convenience but eventually from the dataset for further analysis since it was insignificant.
* Th dataset has 1525 rows and 10 columns (inclusive of Serial Number)



Shape of the dataset

#### Data Types in the dataset:



Datatypes in the dataset

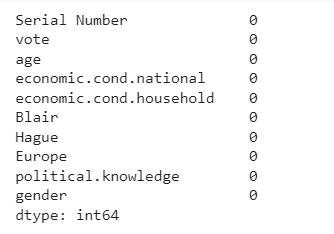
There are total of 10 columns and 1525 rows in the dataset. Out of 10, 2 columns are of object type, and all others are int.

#### Statistics of the dataset:



Statistics of the dataset

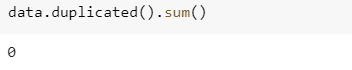
#### Missing Values:



Missing Values

From the above results we can see that there is missing value only in the wchar and rchar columns present in the dataset.

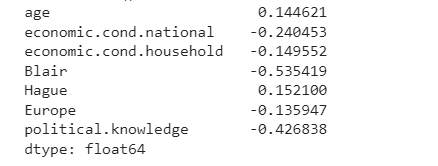
#### Duplicates in the dataset:



Duplicated Values

#### Data Skewness:

Skewness is a metric for the asymmetry of a real-valued random variable's probability distribution with respect to its mean.



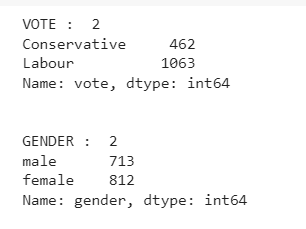
Data Skewness Values

##### Inference from Initial Data Analysis:

1. The column “Serial Number : 0” is removed from the dataset before proceeding further as its insignificant for the analysis.
2. There are 1525 rows and 9 columns
3. Numerical Columns : age, economical\_cond\_national, economical\_cond\_household, Blair, Hague, Europe and political\_knowledge.
4. Non-Numerical Columns : vote and gender.
5. There are no null or duplicated values in the given dataset
6. Age has positive skewness whereas other variables have negative skewness
7. Blair is the column with maximum skewness out of all the columns

##### Inference from Statistical Analysis of the data:

1. Vote and age both have two different values.
2. The Labour vote total is 1036, which is more than the Conservative vote total.
3. Age is a continuous variable, there are both male and female voters, and female voters outnumber male votes.



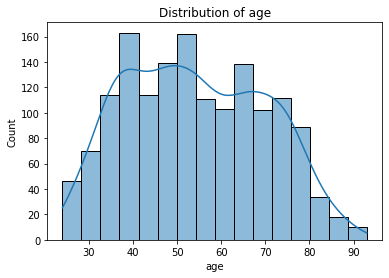
Unique values and their count in dataset

1. Minimum voter age is 24, maximum voter age is 93, average voter age is 54,
2. Maximum assessment of the nation's economic situation is 5,
3. Average assessment of the current state of household finances is 3
4. Minimum assessment of the Labour leader is 1
5. Average evaluation of the Conservative leader is 2.7, and average evaluation of the Labour leader is 3.3, although according to Hague.

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

#### Univariate Analysis:

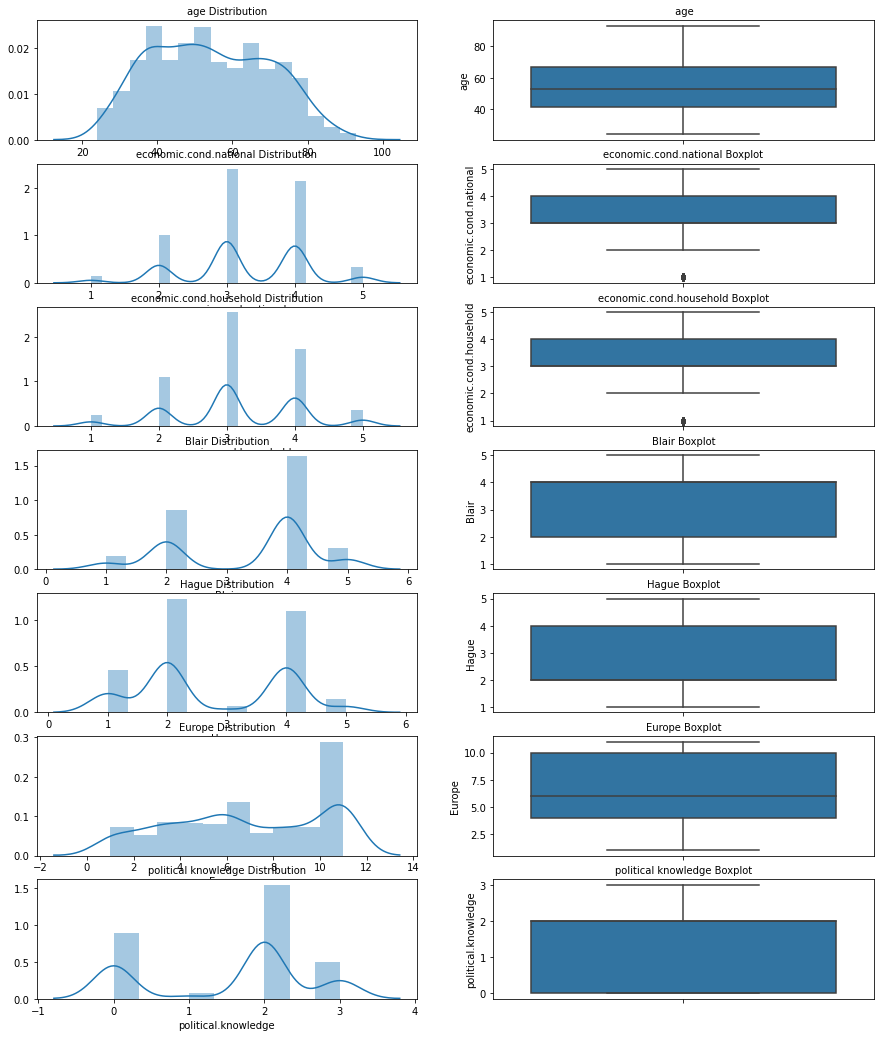
##### Frequency Distribution of Age



###### Inference:

In this dataset, age is the continuous variable and follows a normal distribution unlike the other variables.

##### Boxplot and frequency distribution of other variables



Boxplot and frequency distribution of other variables

###### Inference:

1. Age is the only normally distributed others are multimodal skewness seen
2. only economic.cond.national and economic.cond.household have outliers

##### Count plot of Gender



Count plot of Gender

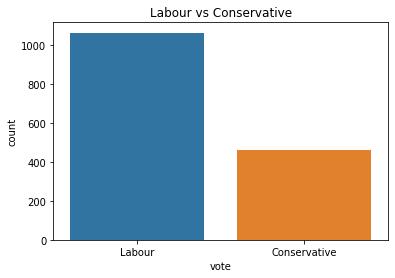
###### Inference:

The number of female voters is greater compared to male voters in the dataset.

female 0.532459 (53%)

male 0.467541 (46%)

##### Count plot of vote:



Count plot of vote

###### Inference:

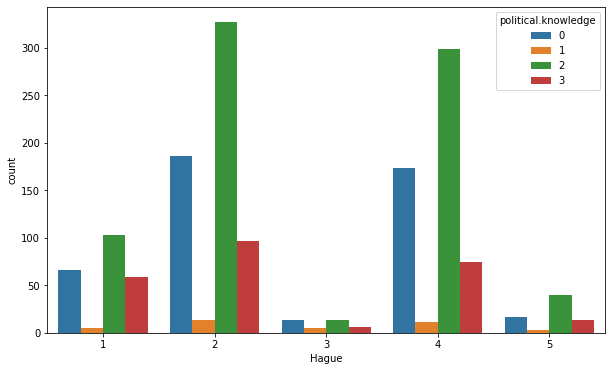
The Labour party easily outnumbers Conservative party.

Labour 0.697049 (69%)

Conservative 0.302951 (30%)

#### Bivariate Analysis:

##### Hague versus political knowledge

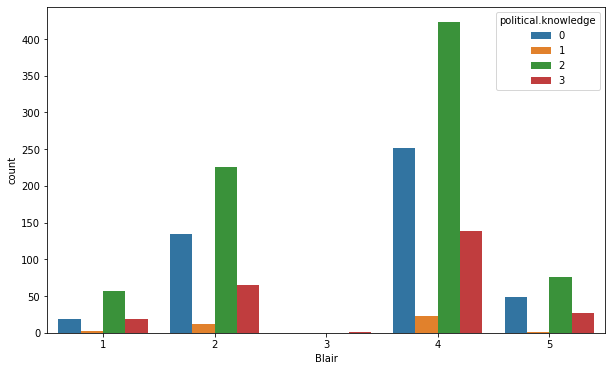


Hague versus political knowledge

###### Inference:

Assessment of Hague with political knowledge shows no 3 cluster have very less distribution whereas no 2 cluster have more distribution

##### Blair versus political knowledge

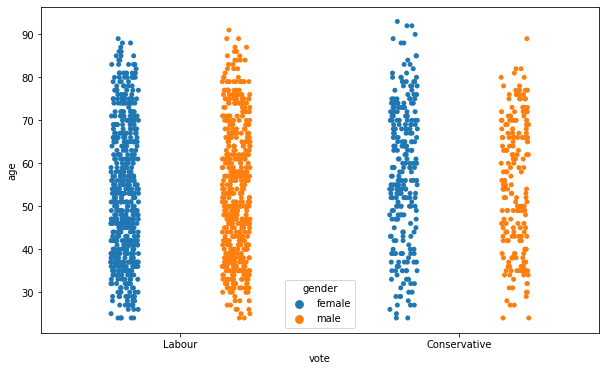


Blair versus political knowledge

###### Inference:

Assessment of Blair with political knowledge shows no 3 cluster have very less or almost no distribution whereas no. 4 cluster have more distribution

##### Vote versus age with gender

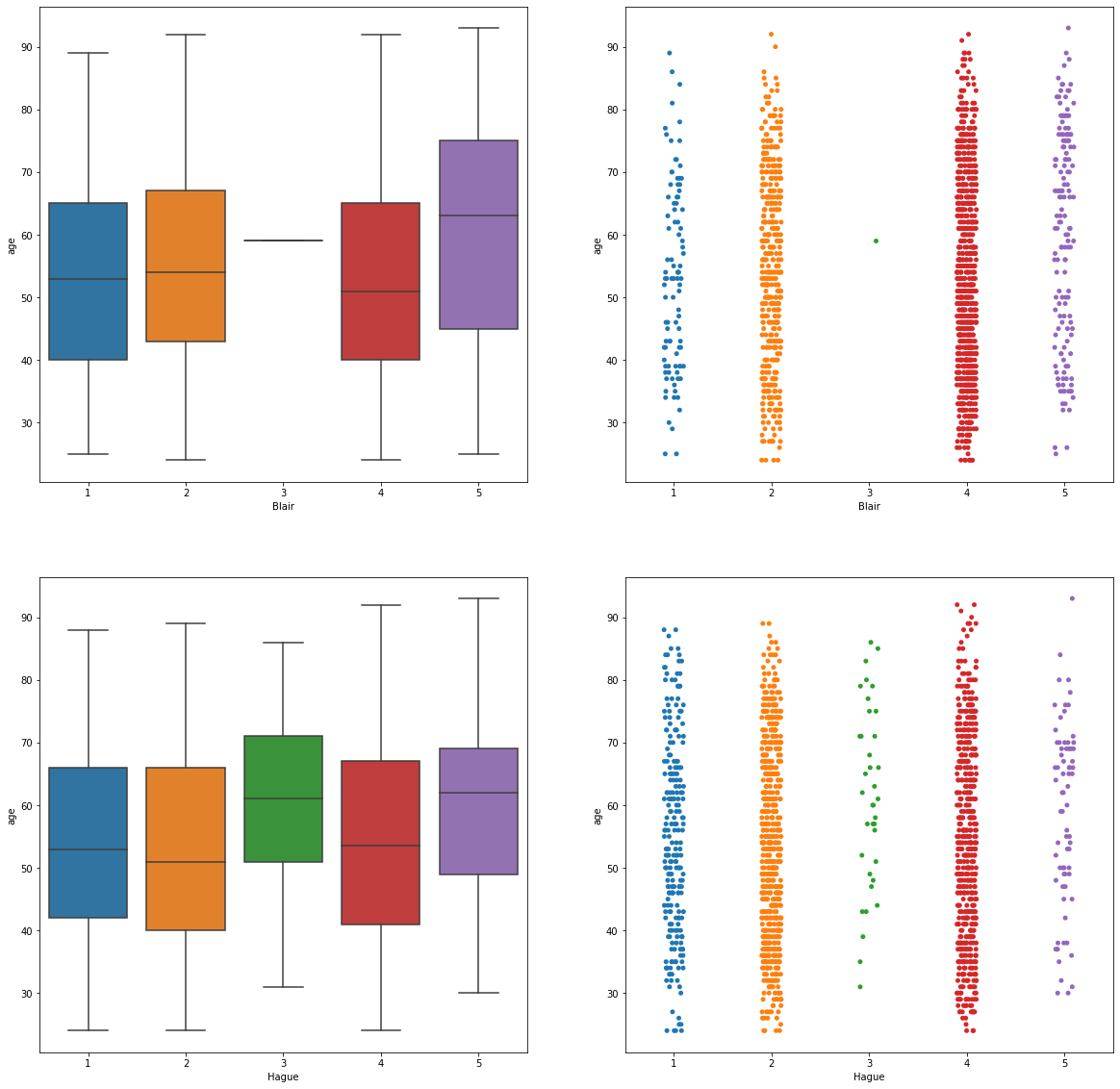


Vote versus age with gender

###### Inference:

The labour party is densely voted when compared to Conservative party.

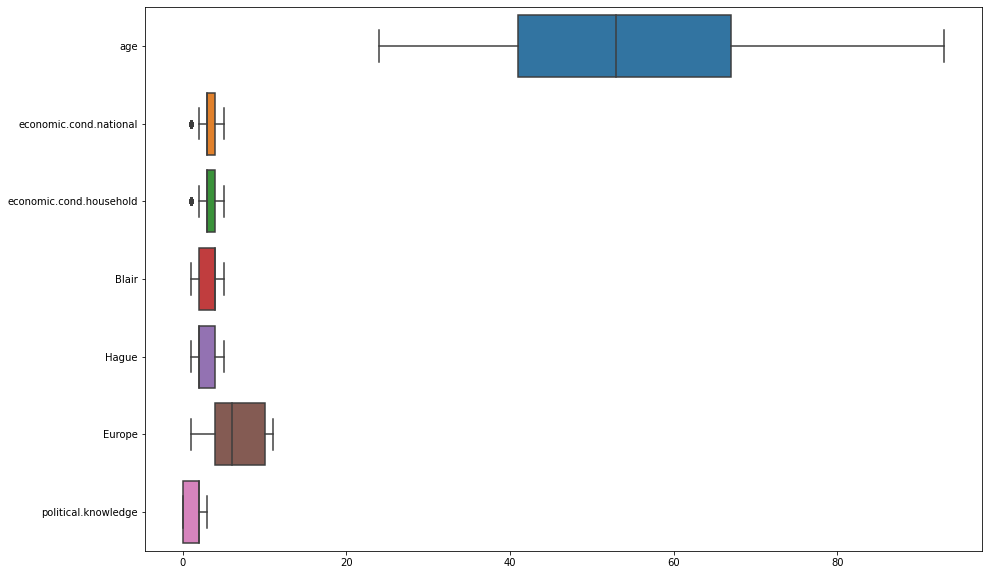
##### Boxplot and Strip plot of Hauge and Blair versus age



###### Inference:

We can see in both the plots that the distribution in Class 3 is almost null or extremely less

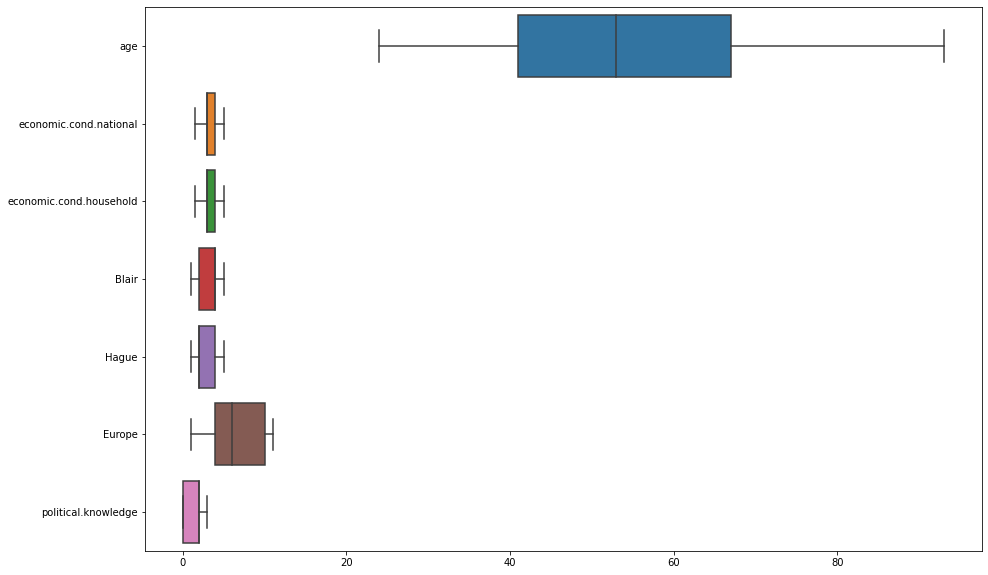
#### Detection of Outliers:



Before Outliers Treatment

##### Treating the Outliers:

It’s important to treat the Outliers before we proceed to any ML model as the algorithms are highly sensitive to the outliers and will significantly cause errors in the prediction, we are about to make using the linear equation.



After Outliers Treatment

#### Correlation Plot

 Correlation Heatmap

###### Understanding from Heatmap

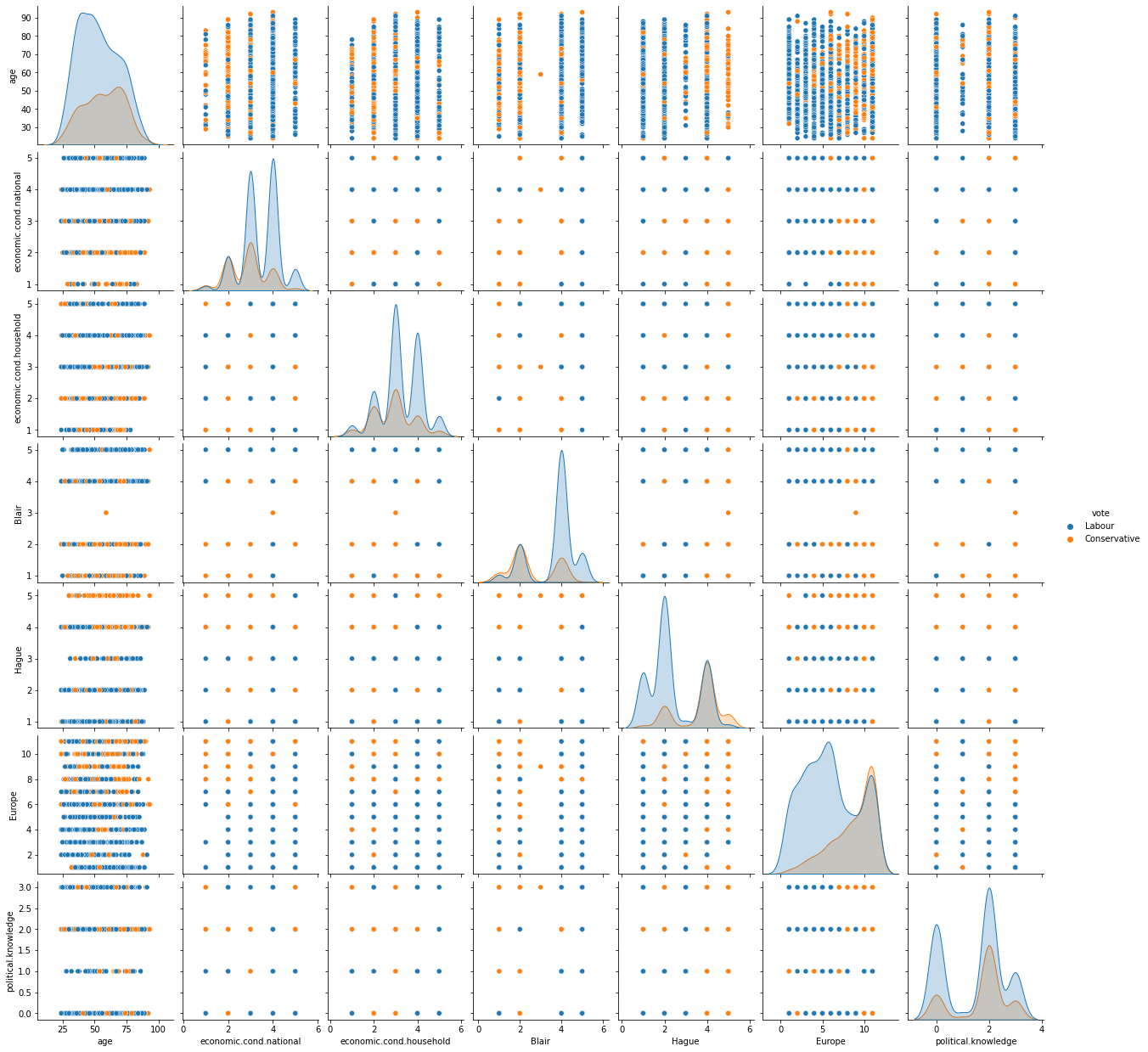
* There is very less correlation between the variables
* The highest positive correlation is seen between “economic\_cond\_national” and “economic\_cond\_household” (35%) with nearly similar results seen from “Blair” and “economic\_cond\_national” (33%)
* The highest negative correlation is seen between “Blair” and “Europe” (30%) with nearly similar results seen from “Blair” and “Hague”(24%)
* There is less or no chance of multicollinearity

#### Pair Plot

Pair plot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

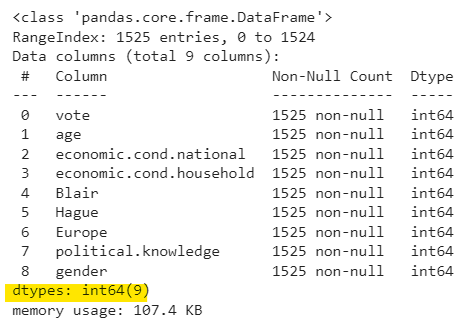
###### Insights from the below pair plot:

1. The relationships between the variables are not linear.
2. Some of the attributes appear to adhere to an exponential distribution.
3. Conservative Party:  parties' stances on European integration seems to be unknown

 Pair Plot

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)

### Converting Object variables to categorical variables



* In this case, the gender and vote are categorical variables that have been encoded as integers.
* All of the variables in the data frame are integers.

### Scaling of Data

* + In most cases, data sets contain features that vary widely in size, units, and range. However, this is a problem because most machine learning algorithms use the Euclidean distance between two data points in their calculations.
  + Problems can be difficult to model due to different scales of input variables.
  + This means transforming the data to fit a certain scale. 0-100 or 0-1
  + Usually distance-based methods (e.g., KNN) are sensitive to extreme differences and can cause distortion, so scaling is required.
  + Tree-based methods use split methods (e.g.
  + decision trees) are not required, so scaling is generally not required.
  + In this dataset, age is the only continuous variable, and the remaining variables are 1-5. The age variable is a continuous variable, so it's just scaled
  + The scaling method done only for the variable 'age' is z-score scaling.
  + Z-score scaling is the most common form of scaling starting with the formula (x – mean) / standard deviation).
  + All ML models are done with scaled data.

### Train-Test Data Split

Separating independent (train) and dependent (test)variables for the linear regression model

X = independent (train) variables

Y = dependent (test)variables

#### Shape of Test and Train data:

The training set for the independent variables: (1067, 9)

The training set for the dependent variable: (1067, 1)

The test set for the independent variables: (458, 9)

The test set for the dependent variable: (458, 1)

1.4 Apply Logistic Regression and LDA (linear discriminant analysis)

### Logistic Regression

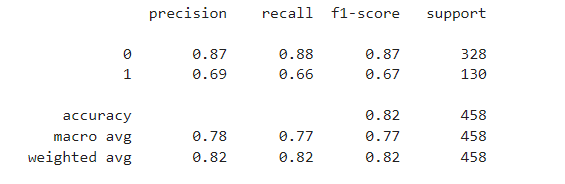
The dataset is now treated for outliers and also split the data into test and train data. In this case the target variable is “vote”

#### Logistic Regression for Train Data:

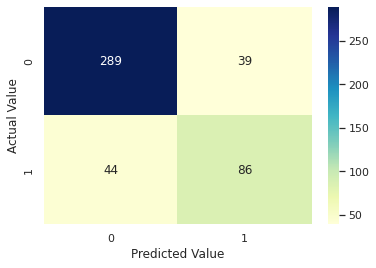
###### Accuracy for Train Data:



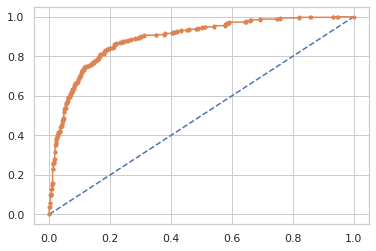
###### Classification Report for Train Data:



###### Confusion Matrix for Train Data:



###### AUC Curve for Train Data:

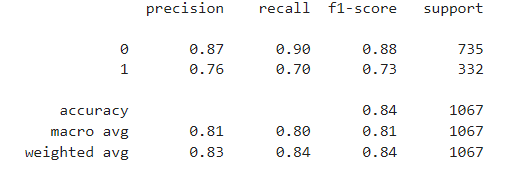


#### Logistic Regression for Test Data:

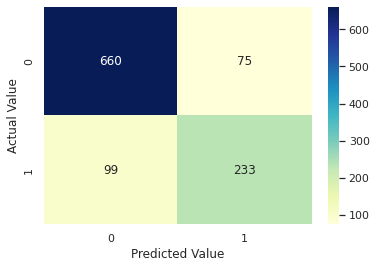
###### Accuracy for Test Data:



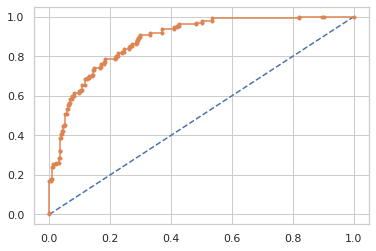
###### Classification Report for Test Data:



###### Confusion Matrix for Test Data:



###### AUC Curve for Test Data:



#### Validity of the Logistic Regression Model:

• The model is neither overfitted nor underfitted.

• Errors on test data are slightly larger than on training data. This is perfectly fine as the error margin is low and the inaccuracy in both the training and test data are not too enormous. Therefore, the model is neither overfitted nor underfitted.

### Linear Discriminant Analysis

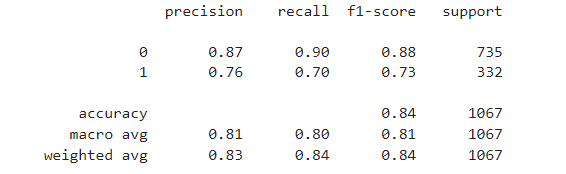
1. Linear discriminant analysis is not just a dimension reduction tool, but also a robust classification method.
2. With or without data normality assumption, we can arrive at the same LDA features, which explains its robustness.

#### Linear Discriminant Analysis for Train Data:

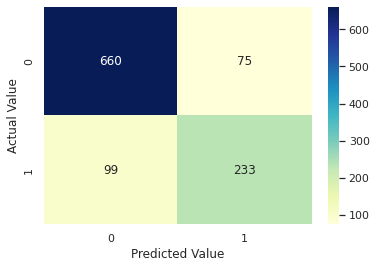
###### Accuracy for Train Data:



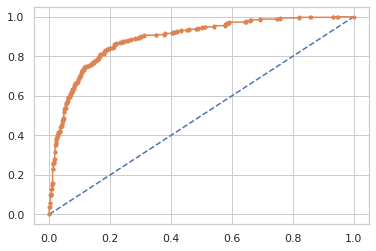
###### Classification Report for Train Data:



###### Confusion Matrix for Train Data:



###### AUC Curve for Train Data:

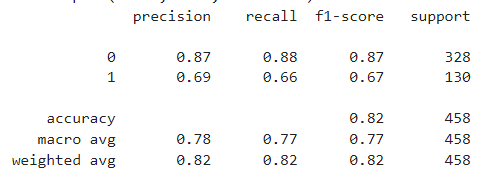


#### Linear Discriminant Analysis for Test Data:

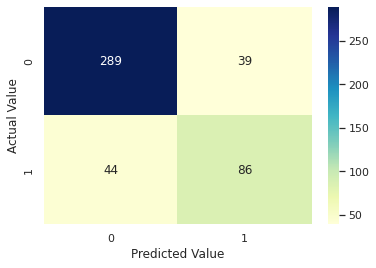
###### Accuracy for Test Data:



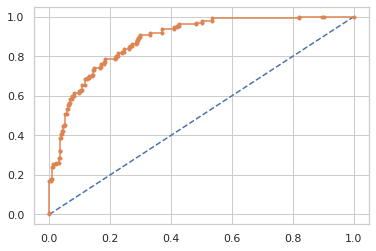
###### Classification Report for Test Data:



###### Confusion Matrix for Test Data:



###### AUC Curve for Test Data:



#### Validity of the Linear Discriminant Analysis Model:

• The model doesn’t seem to be overfitted or underfitted

• But the accuracy is decreased by 2% from training data to test data. The performance of the model is considered good but not optimal.

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results

### KNN Model Analysis

* KNN is a nonparametric, delayed learning algorithm. Nonparametric means that there are no assumptions about the underlying data distribution.
* For KNN, K is the number of nearest neighbours. The number of neighbours is the central determining criterion

KNN has the following basic steps:

* Calculate distance
* Find nearest neighbour
* Create characteristic and target variables.
* Split the data into training data and test data.
* Generate a k-NN model using neighbouring values.
* Train or fit the data to the model.
* Predict the future.

KNN classifier model

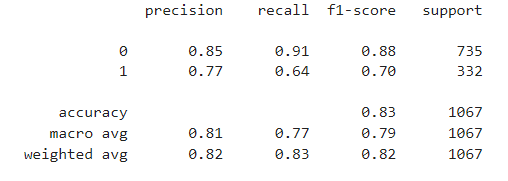
* Create a KNN classifier object by first importing the KNeighboursClassifier module and passing the neighbour argument number in the KNeighboursClassifier() function.
* Then fit the model to the training set with fit() and run predictions on the test set with Predict().
* Let's build an ANN classifier model with k=15.

#### KNN for Train Data:

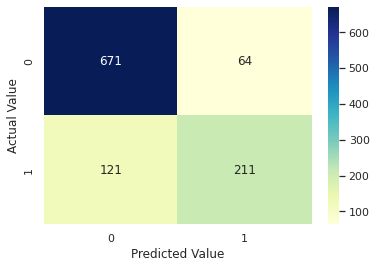
###### Accuracy for Train Data:



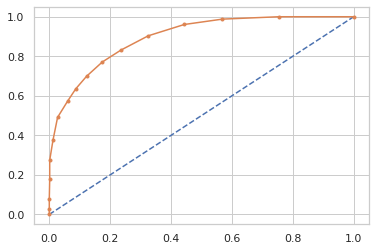
###### Classification Report for Train Data:



###### Confusion Matrix for Train Data:



###### AUC Curve for Train Data:

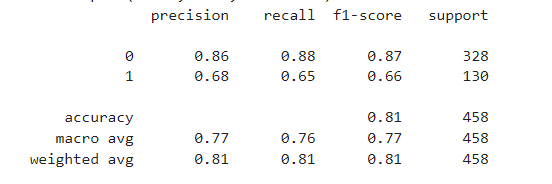


#### KNN for Test Data:

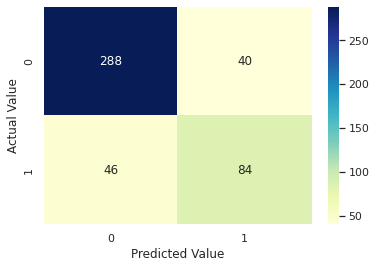
###### Accuracy for Test Data:



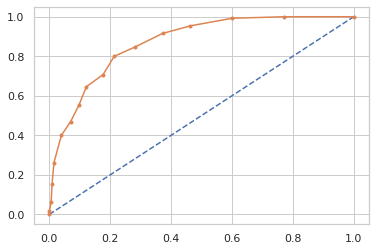
###### Classification Report for Test Data:



###### Confusion Matrix for Test Data:



###### AUC Curve for Test Data:



#### Validity of the KNN Model:

• The model doesn’t seem to be overfitted or underfitted

• Overall, the model is a good performer.

### Naïve Bayes Model Analysis

Naive Bayes is a statistical classification technique based on the Bayes Theorem and one of the simplest Supervised Learning algorithms. The Naive Bayes classifier is a quick, accurate, and trustworthy method, especially on large datasets

When to use the Naive Bayes Classifier?

Naive Bayes classifiers tend to perform especially well in any of the following situations:

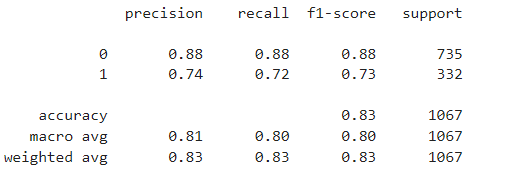
* When the naive assumptions actually match the data.
* For very well-separated categories, when model complexity is less important.
* And for very high-dimensional data when model complexity is again less important.

#### Naive Bayes for Train Data:

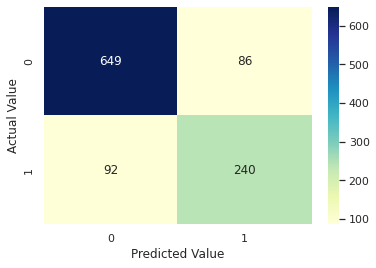
###### Accuracy for Train Data:



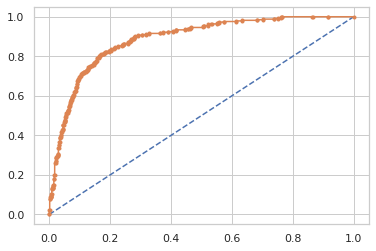
###### Classification Report for Train Data:



###### Confusion Matrix for Train Data:



###### AUC Curve for Train Data:

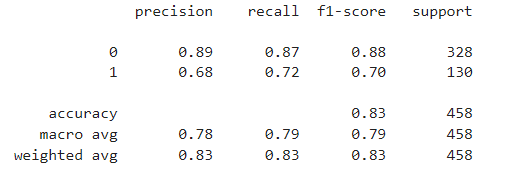


#### Naïve Bayes for Test Data:

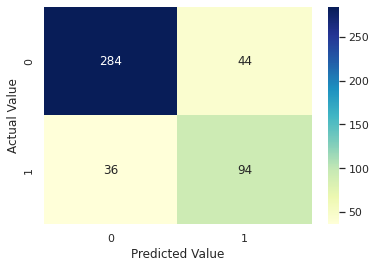
###### Accuracy for Test Data:



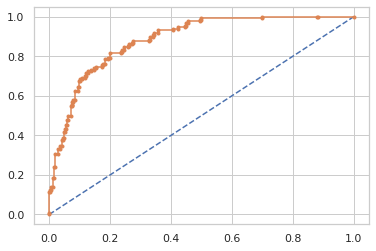
###### Classification Report for Test Data:



###### Confusion Matrix for Test Data:



###### AUC Curve for Test Data:



1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging) and Boosting.

### Bagging with Random Forest Model:

* Tuning is the process of maximizing a model’s performance without overfitting or creating too high of a variance. In machine learning, this is accomplished by selecting appropriate “hyperparameters.”
* Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset.
* In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

#### Parameters:

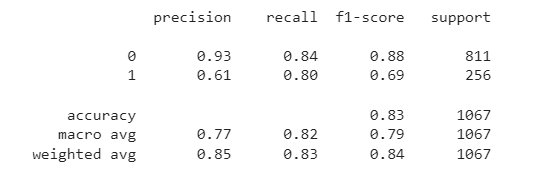
* N\_estimators (only used in Random Forests) is the number of decision trees used in making the forest (105).
* Max\_depth is an integer that sets the maximum depth of the tree. The default is None, which means the nodes are expanded until all the leaves are pure
* Min\_samples\_split is the minimum number of samples required to split an internal node.
* Min\_samples\_leaf defines the minimum number of samples needed at each leaf.
* We now fit random forest classifier model to the bagging model by training the model with our independent variable and dependent variables. At this point, you have the classification model defined.

#### Bagging for Train Data:

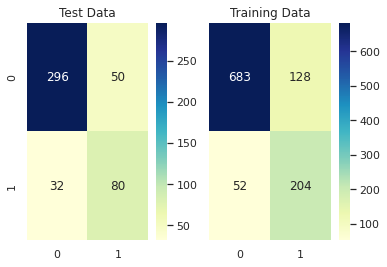
###### Accuracy for Train Data:



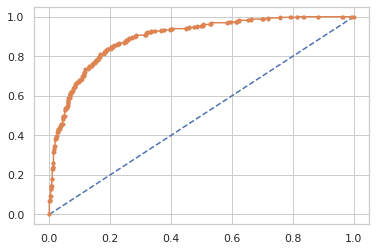
###### Classification Report for Train Data:



###### Confusion Matrix for Train Data:



###### AUC Curve for Train Data:

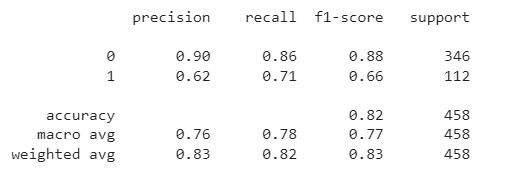


#### Bagging for Test Data:

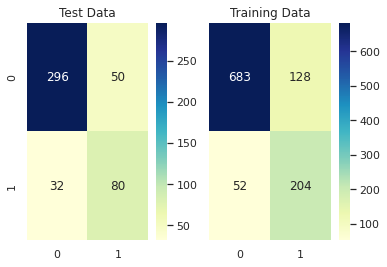
###### Accuracy for Test Data:



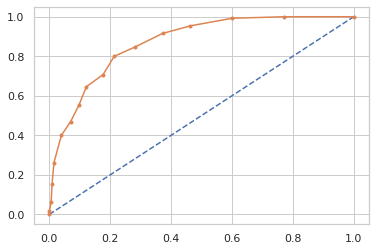
###### Classification Report for Test Data:



###### Confusion Matrix for Test Data:



###### AUC Curve for Test Data:



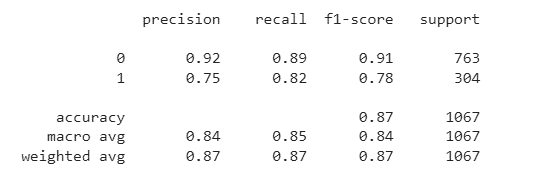
### Gradient Boosting Model:

#### Gradient Boosting for Train Data:

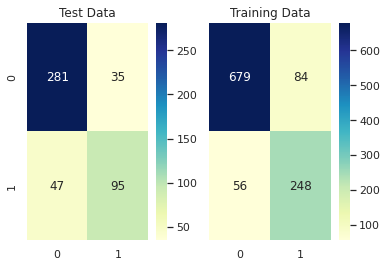
###### Accuracy for Train Data:



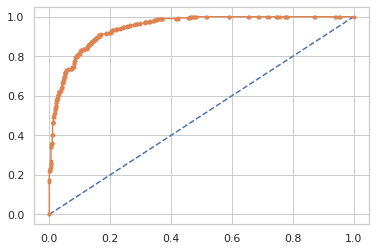
###### Classification Report for Train Data:



###### Confusion Matrix for Train Data:



###### AUC Curve for Train Data:

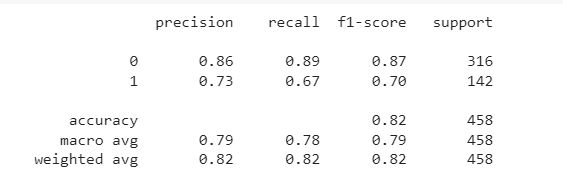


#### Gradient Boosting for Test Data:

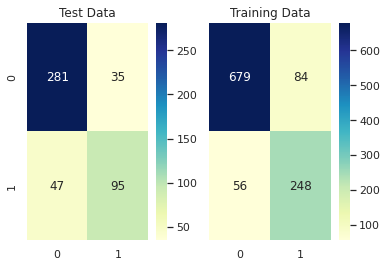
###### Accuracy for Test Data:



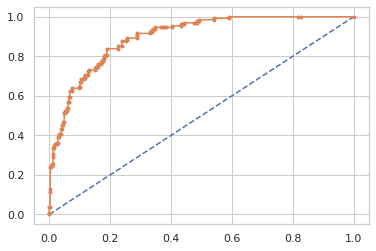
###### Classification Report for Test Data:



###### Confusion Matrix for Test Data:



###### AUC Curve for Test Data:



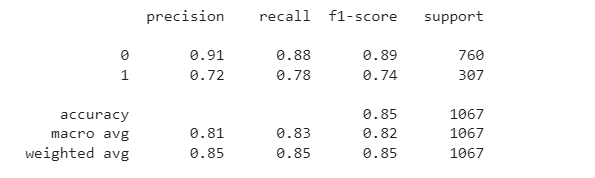
### Ada Boosting Model:

#### Ada Boosting for Train Data:

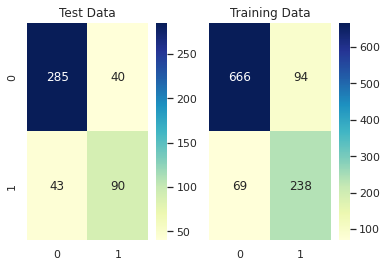
###### Accuracy for Train Data:



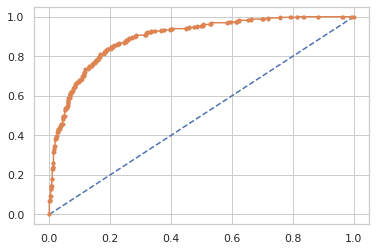
###### Classification Report for Train Data:



###### Confusion Matrix for Train Data:



###### AUC Curve for Train Data:

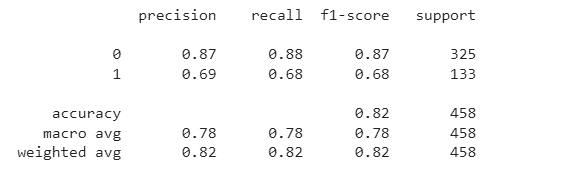


#### Ada Boosting for Test Data:

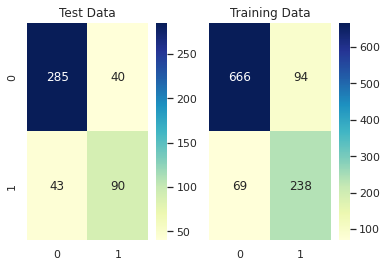
###### Accuracy for Test Data:



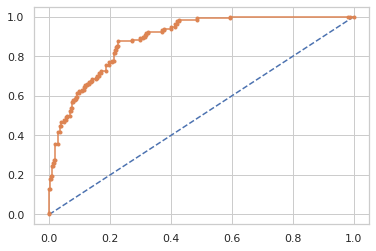
###### Classification Report for Test Data:



###### Confusion Matrix for Test Data:



###### AUC Curve for Test Data:



1.7  Performance Metrics:

* The accuracy, AUC, confusion matrix is all plotted and mentioned model wise in the above sections.
* Let us now compare the metrics of all the models to see which is the optimal model.

### Final Model: Comparing all the models

|  | **LR Train** | **LR Test** | **LDA Train** | **LDA Test** | **KNN Train** | **KNN Test** | **NB Train** | **NB Test** | **BAGGING Train** | **BAGGING Test** | **ADA Train** | **ADA Test** | **Gradient Train** | **Gradient Test** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Accuracy** | 0.84 | 0.82 | 0.84 | 0.82 | 0.83 | 0.81 | 0.83 | 0.83 | 0.83 | 0.82 | 0.85 | 0.82 | 0.87 | 0.82 |
| **AUC** | 0.89 | 0.88 | 0.89 | 0.88 | 0.90 | 0.87 | 0.89 | 0.88 | 0.89 | 0.89 | 0.91 | 0.88 | 0.94 | 0.90 |
| **Recall** | 0.66 | 0.66 | 0.72 | 0.66 | 0.64 | 0.66 | 0.72 | 0.72 | 0.61 | 0.62 | 0.72 | 0.69 | 0.75 | 0.73 |
| **Precision** | 0.69 | 0.69 | 0.74 | 0.69 | 0.77 | 0.69 | 0.74 | 0.68 | 0.80 | 0.71 | 0.78 | 0.68 | 0.82 | 0.67 |
| **F1 Score** | 0.67 | 0.67 | 0.73 | 0.67 | 0.70 | 0.67 | 0.73 | 0.70 | 0.69 | 0.66 | 0.74 | 0.68 | 0.78 | 0.70 |

#### Conclusion:

* Almost all the models performed well with accuracy between 82% to 84%. Gradient boosting improved the accuracy to 87% so it is better model for to predict which party a voter will vote.
* Comparing all the model ,Gradient boosting model is best model for this dataset with accuracy of 87% in both training and test set
* Accuracy ,AUC, Precision, Recall for test data are almost in line with training data in Gradient boosting model. This indicates no overfitting or underfitting in the model
* Gradient boosting improved the accuracy to 87% so it is better model for to predict which party a voter will vote

### Overall Best/Optimized model:

Almost all the models performed well with accuracy between 82% to 84% with scaled data. But Gradient boosting is best and optimised model with accuracy of 87% and also best AUC,Precision,f1 score, Recall

# Problem 02:

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

(Hint: use .words(), .raw(), .sent() for extracting counts)

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

The number of characters in Roosevelt speech are: 7571

The number of characters in Kennedy speech are: 7618

The number of characters in Nixon speech are: 9991

The number of words in Roosevelt speech are: 1536

The number of words in Kennedy speech are: 1546

The number of words in Nixon speech are: 2028

The number of sentences in Roosevelt speech are: 68

The number of sentences in Kennedy speech are: 52

The number of sentences in Nixon speech are: 69

## 2.2 Remove all the stop words from all three speeches

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words to take up space in our database or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider stopping words.

### Libraries used:

from nltk.corpus import stopwords

nltk.download('stopwords')

Stopwords removal is done through “from nltk.corpus import stopwords”.

### Stemming:

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.

#### Library used:

from nltk.stem.porter import PorterStemmer

## Most common words in Roosevelt speech after removing stopwords

Most common words in Roosevelt speech after removing stopwords

['nation', 'know', 'peopl', 'spirit', 'life', 'democraci', 'us', 'america', 'live', 'year', 'human', 'freedom', 'measur', 'men', 'govern', 'new', 'bodi', 'mind', 'speak', 'da y', 'state', 'american', 'must', 'someth', 'faith', 'unit', 'task', 'preserv', 'within' , 'histori', 'three', 'form', 'futur', 'seem', 'hope', 'understand', 'thing', 'free', ' alon', 'still', 'everi', 'contin', 'like', 'person', 'world', 'sacr', 'word', 'came', ' land', 'first']

## Most common words in Kennedy speech after removing stopwords

['let', 'us', 'power', 'world', 'nation', 'side', 'new', 'pledg', 'ask', 'citizen', 'peac', 'shall', 'free', 'final', 'presid', 'fellow', 'freedom', 'begin', 'man', 'h and', 'human', 'first', 'gener', 'american', 'war', 'alway', 'know', 'support', 'un it', 'cannot', 'hope', 'help', 'weak', 'arm', 'countri', 'call', 'today', 'well', ' god', 'form', 'poverti', 'life', 'globe', 'right', 'state', 'dare', 'word', 'go', ' friend', 'bear']

## Most common words in Nixon speech after removing stopwords

Most common words in Nixon speech after removing stopwords

['us', 'let', 'america', 'peac', 'world', 'respons', 'new', 'nation', 'govern', 'gr eat', 'year', 'home', 'abroad', 'make', 'togeth', 'shall', 'time', 'polici', 'role' , 'right', 'everi', 'histori', 'better', 'come', 'respect', 'peopl', 'live', 'help' , 'four', 'war', 'today', 'era', 'progress', 'other', 'build', 'act', 'challeng', ' one', 'mr', 'share', 'meet', 'promis', 'long', 'work', 'preserv', 'freedom', 'place ', 'system', 'god', 'way']

* 1. Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)

## Top three words in Roosevelt's speech(after removing the stopwords)

[('nation', 17), ('know', 10), ('peopl', 9)]

## Top three words in Kennedy's speech(after removing the stopwords

[('let', 16), ('us', 12), ('power', 9)]

## Top three words in Nixon's speech(after removing the stopwords

[('us', 26), ('let', 22), ('america', 21)]

2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords)

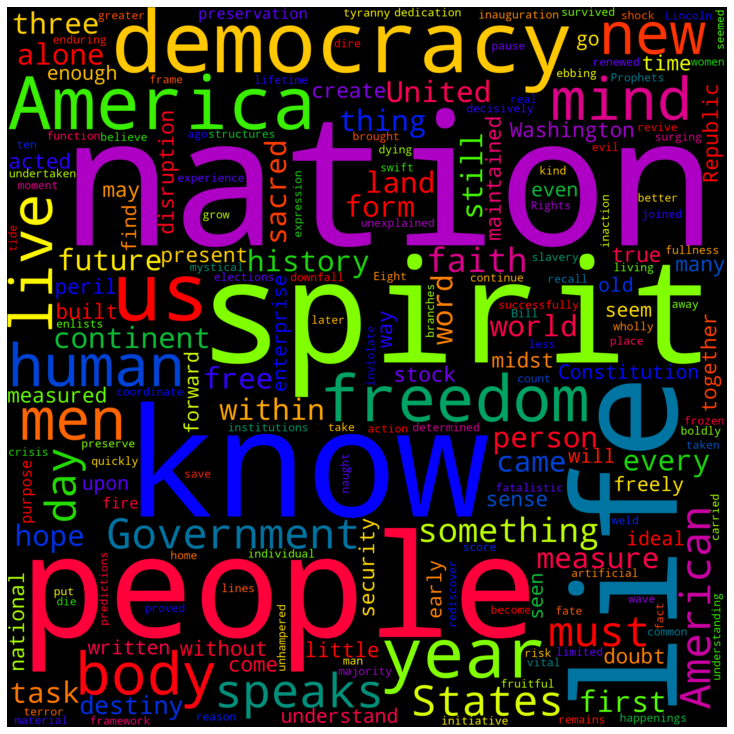
Wordcloud

Many times, you might have seen a cloud filled with lots of words in different sizes, which represent the frequency or the importance of each word. This is called WordCloud.

Library used

from wordcloud import WordCloud

## Roosevelt speech



WordCloud of Roosevelt Speech

Inference:

Most Frequent words are nation, people, spirit, life

Less Frequent words are weld, kind, women, moment

## Kennedy speech



WordCloud of Kennedy Speech

Inference:

Most Frequent words are let, world, slides, power

Less Frequent words are best, wishes, slow

## Nixon speech



WordCloud of Nixon Speech

Inference:

Most Frequent words are America, let, us, nation

Less Frequent words are flimsy, adopted, saw

## Conclusion:

This project data presented from '1941-Roosevelt.txt', '1961-Kennedy.txt' and '1973-Nixon.txt', we analysed some interesting insights like the number of characters, words, and sentences from the speeches. To Identify the strength and the sentiment of these presidential speeches the stop words were removed (punctuation and lowering the characters were removed) along with stemming. We analysed some of the common words from their speeches which inspired many Americans. word cloud method which visually show most common words to least common words