**Project Name:** Machine Learning

**Student Name:** Arun Sivaji

**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.

Dataset for Problem: Election_Data.xlsx

### 1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

---

| | Unnamed: 0 | vote | age | economic.cond.national | economic.cond.household | Blair | Hague | Europe | political.knowledge | gender |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | Labour | 43 | 3 | 3 | 4 | 1 | 2 | 2 | female |
| **1** | 2 | Labour | 36 | 4 | 4 | 4 | 4 | 5 | 2 | male |
| **2** | 3 | Labour | 35 | 4 | 4 | 5 | 2 | 3 | 2 | male |
| **3** | 4 | Labour | 24 | 4 | 2 | 2 | 1 | 4 | 0 | female |
| **4** | 5 | Labour | 41 | 2 | 2 | 1 | 1 | 6 | 2 | male |

```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   vote                   1525 non-null   object
 1   age                    1525 non-null   int64
 2   economic.cond.national 1525 non-null   int64
 3   economic.cond.household 1525 non-null  int64
 4   Blair                  1525 non-null   int64
 5   Hague                  1525 non-null   int64
 6   Europe                 1525 non-null   int64
 7   political.knowledge    1525 non-null   int64
 8   gender                 1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

**Information about Dataset**

```
vote                    0
age                     0
economic.cond.national  0
economic.cond.household 0
Blair                   0
Hague                   0
Europe                  0
political.knowledge     0
gender                  0
dtype: int64
```

**Null Values in Dataset**

Information of the dataset derived by using info function and description of the dataset using describe function. Following images shows the result info function and describe function. By default, describe function generates the results for the numerical variable only.

1.All the variables have int64 datatype only differs in the vote and gender as they are categorical variable with object datatypes. Although even all other variables show the int64 datatype but they are also categorical in nature except age which is numerical in nature.

2. There is no any null value present in the dataset.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **age** | 1525.0 | 54.182295 | 15.711209 | 24.0 | 41.0 | 53.0 | 67.0 | 93.0 |
| **economic.cond.national** | 1525.0 | 3.245902 | 0.880969 | 1.0 | 3.0 | 3.0 | 4.0 | 5.0 |
| **economic.cond.household** | 1525.0 | 3.140328 | 0.929951 | 1.0 | 3.0 | 3.0 | 4.0 | 5.0 |
| **Blair** | 1525.0 | 3.334426 | 1.174824 | 1.0 | 2.0 | 4.0 | 4.0 | 5.0 |
| **Hague** | 1525.0 | 2.746885 | 1.230703 | 1.0 | 2.0 | 2.0 | 4.0 | 5.0 |
| **Europe** | 1525.0 | 6.728525 | 3.297538 | 1.0 | 4.0 | 6.0 | 10.0 | 11.0 |
| **political.knowledge** | 1525.0 | 1.542295 | 1.083315 | 0.0 | 0.0 | 2.0 | 2.0 | 3.0 |

1. The voters age ranges from 24 to 93 years. Mean and Median values of the Age show the nearly same that means Age is normal distributed.

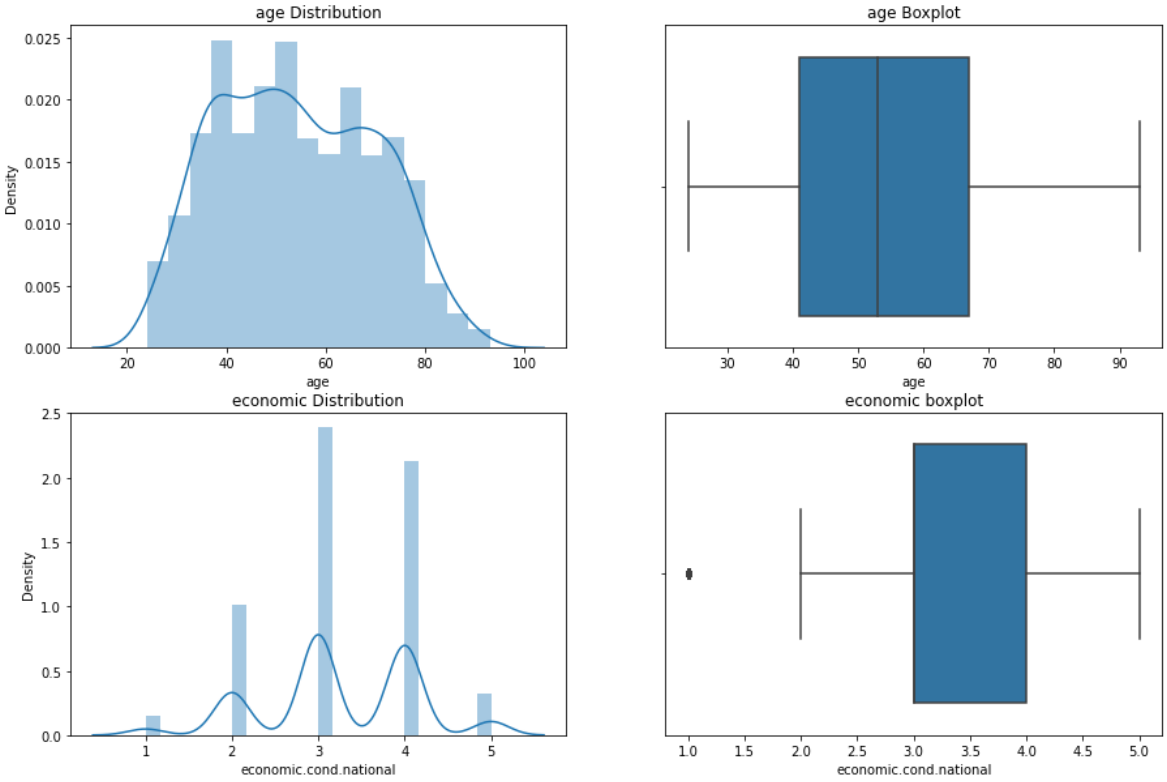
2. Assessment or opinion about economic condition either national or household look like same.
3. As the leader's assessment concerned, Blair- the leader of the labour party shows the better popularity than Hague – the leader of conservative party. On a very first look voters have confidence in Blair

4.The voters have fairly knowledge about the party's position on European integration and the they show their European integration attitudes

5. That means, there is 53.25% of voters are female and 69.70% voters choose the labour party to cast their vote. And as of now, Blair the leader of the labour party is the key factor for the increment in the vote share.

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

But there are duplicates present in the dataset and for the overfitting issues we can drop these duplicate entries.




**#1. Age:**

1. Total 1517 entries.

2.Age ranges from 24 to 93

3.Mean age of the voters in the dataset is 54

4.Standard deviation of the age column is 15

5.Median of the age column is 53

6.25 percentile of the age column 41 and 75 percentile is 67

7.Data shows normally distributed.  No outliers

**#2. Economic Condition National**:

1. Total 1517 data points.

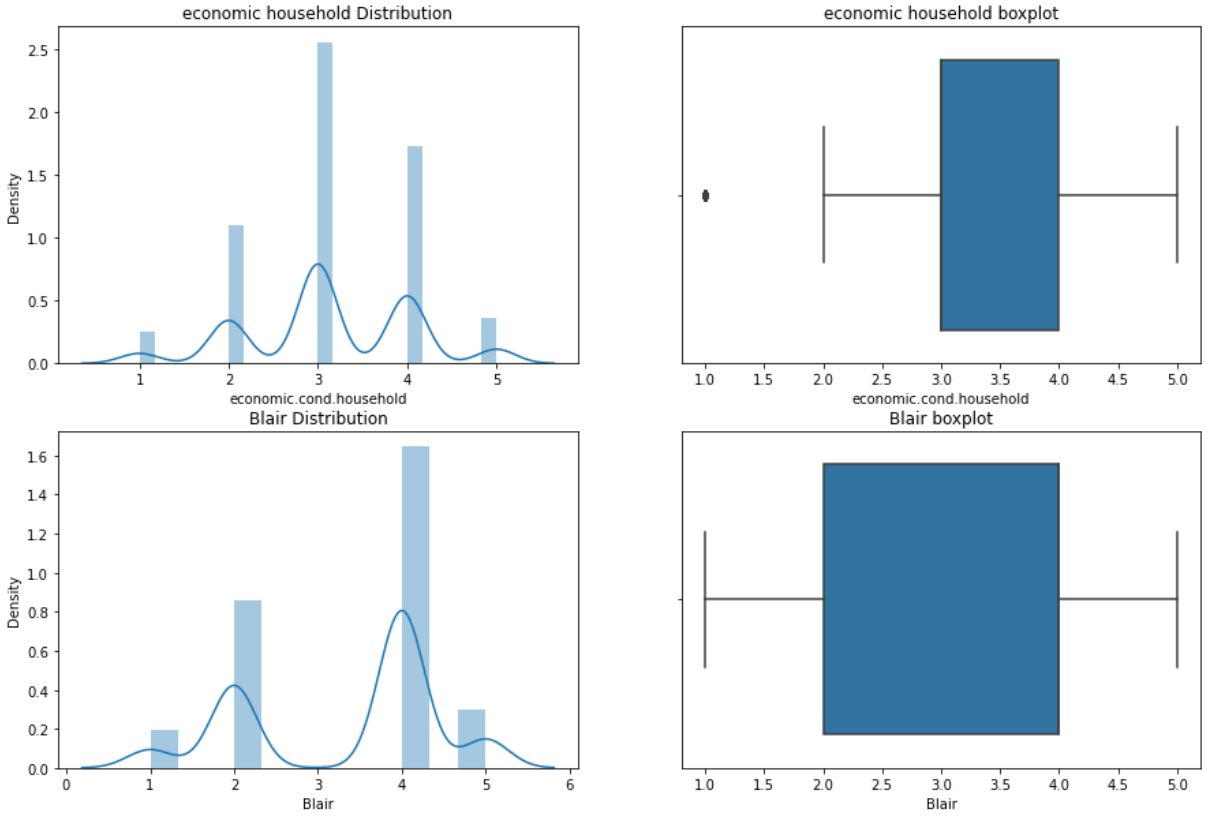
2. Ranges from 1 to 5 i.e., assessed ratings 1 as poor and 5 as best

3.Mean of economic condition national is 3.245

4. Standard deviation is 0.88 means 1,

5. Median of the column is 3 Data shows the major of voters assessed economic condition is good and better i.e., lies in the category of 3 and 4. Very few people thinks that national economic condition is very poor or very rich.

6. Outliers shown as a data point of 1. There are 37 voters who thinks economic condition of nation is very poor while 257 voter's thinks same but not extreme poor and rated as 2 so we can have capped these 37 entries to the 2 for better model performance.



**#3. Economic Condition Household:**

1. Total 1517 data points.

2.Ranges from 1 to 5 i.e., assessed ratings 1 as poor and 5 as best

3.Mean of economic condition national is 3.13

4.Standard deviation is 0.93 means 1,

5.Median of the column is 3

6.Outlier presents. Data shows the major of voters assessed economic condition is good and better i.e., lies in the category of 3 and 4. Very few people thinks that economic condition is very poor or very rich.

**#4. Blair**

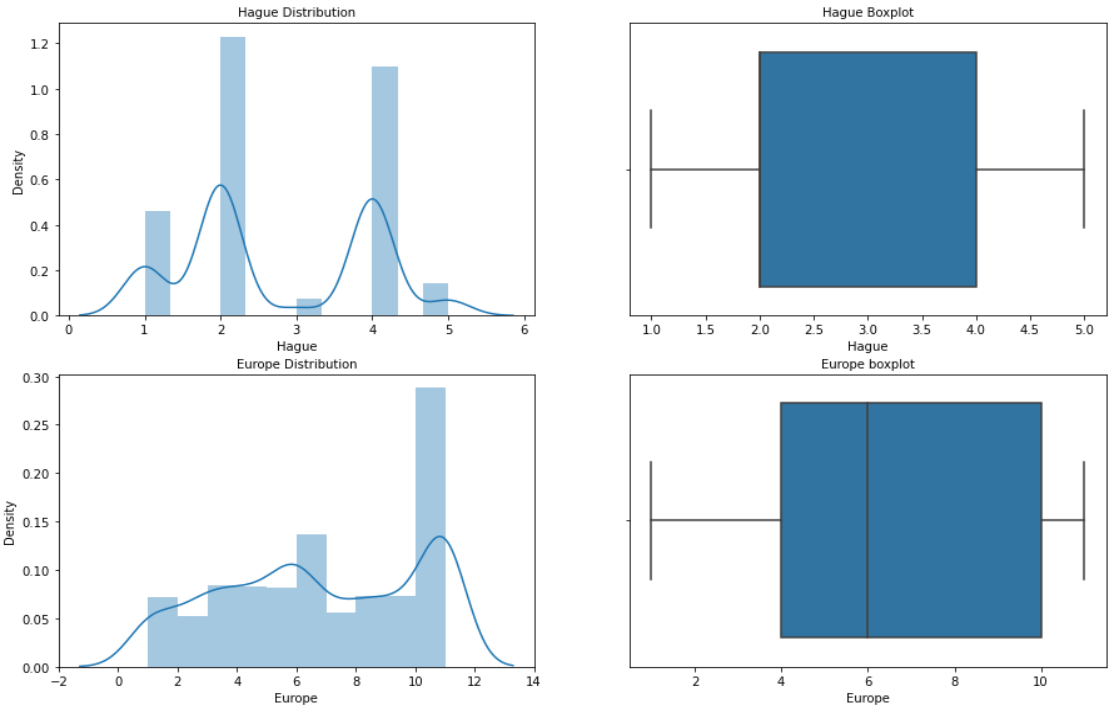
1. Total 1517 data points.

2. Ranges from 1 to 5 i.e., assessed ratings 1 as poor and 5 as best

3. Mean of Blair is 3.33 ? Standard deviation is 1.17

4. Median of the column is 4

5.No outliers.



**#5. Hague**

1.Total 1517 data points.

2. Ranges from 1 to 5 i.e., assessed ratings 1 as poor and 5 as best

3. Mean of Hague is 2.749

4 Standard deviation is 1.23

5.Median of the column is 2

6. No outliers.

**#6. Europe**

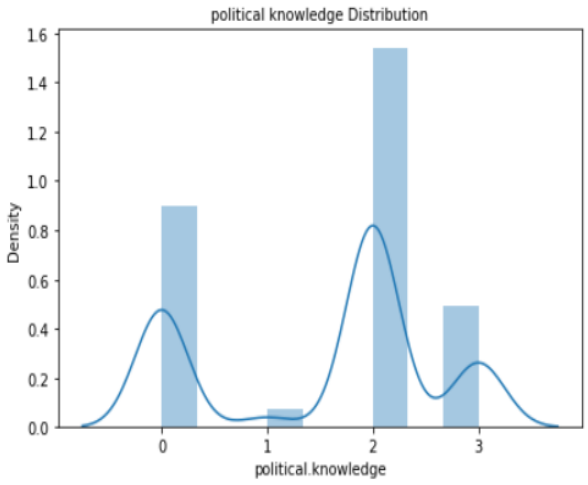
1.Total 1517 data points.

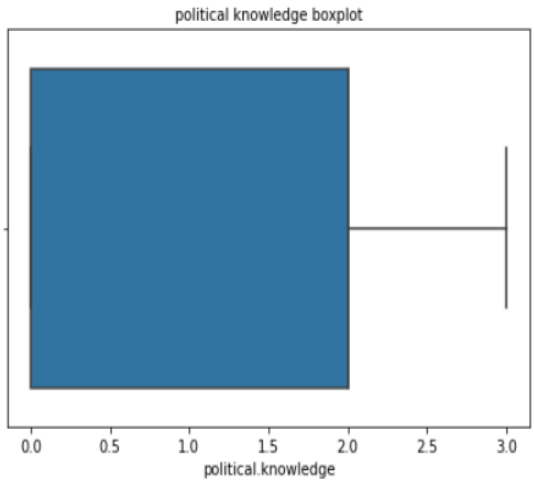
2. Ranges from 1 to 11 i.e., assessed ratings 1 as poor and 11 Highest 'Eurosceptic' sentiment.

3. Mean of Europe is 6.740 Standard deviation is 3.3

4.Median = 6

5. No outliers. But highest density of the people's attitude is towards the 'Eurosceptic' sentiments.



**#7. Political Knowledge**

1. Total 1517 data points.

2. Ranges from 1 to 3 i.e., knowledge of the party's position on European Integration.

3. Mean of Political Knowledge is 1.54

4. Median of the columns is 2

5.Standard deviation is 1.08

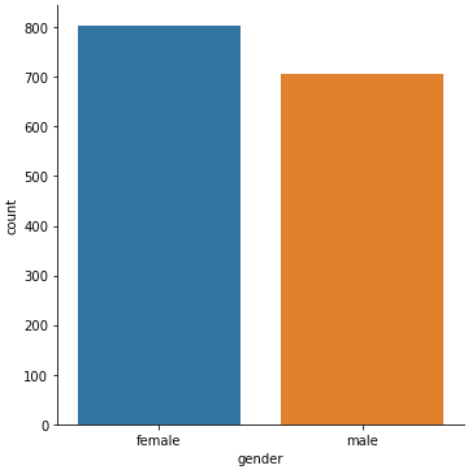
6.No outliers. But most of the peoples have knowledge about the position on European i integration of the labour and conservative party respectively.

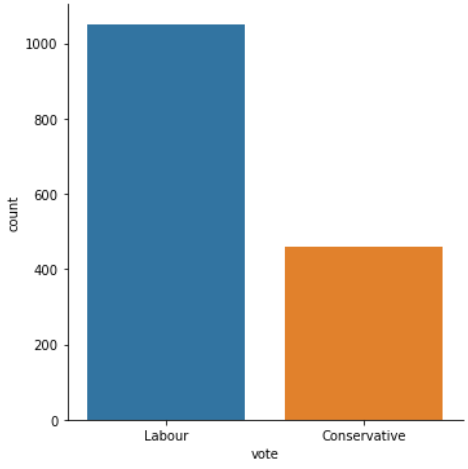
7.There is good number of peoples that does not have any idea about the positions on the European integration of the parties.

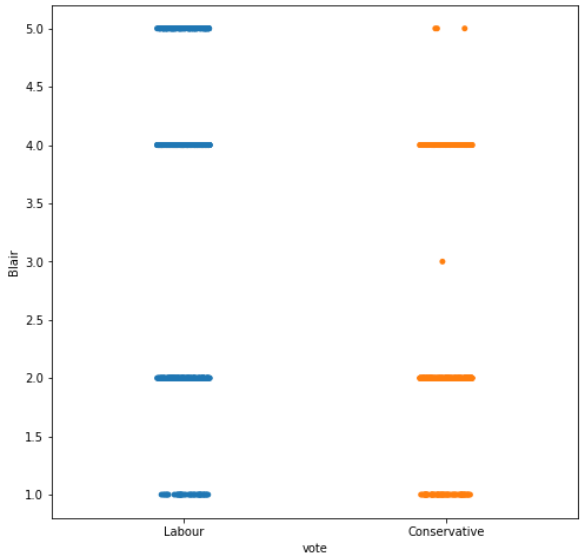
**SKEWNESS OF THE DATASET: -**

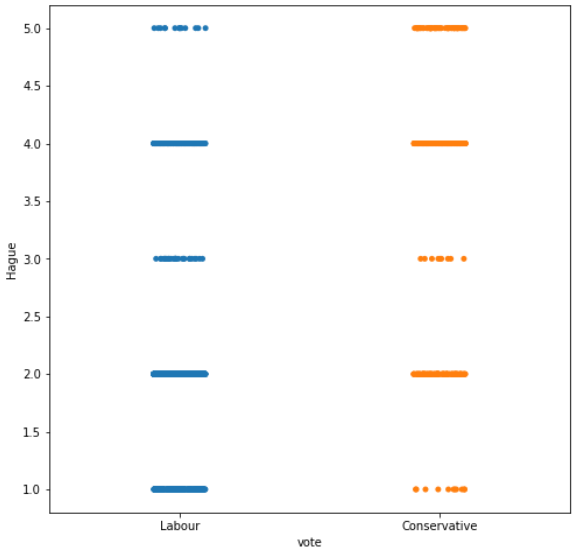
| | |
|---|---|
| age | 0.134944 |
| economic.cond.national | -0.236477 |
| economic.cond.household | -0.138688 |
| Blair | -0.543655 |
| Hague | 0.140236 |
| Europe | -0.147907 |
| political.knowledge | -0.418982 |

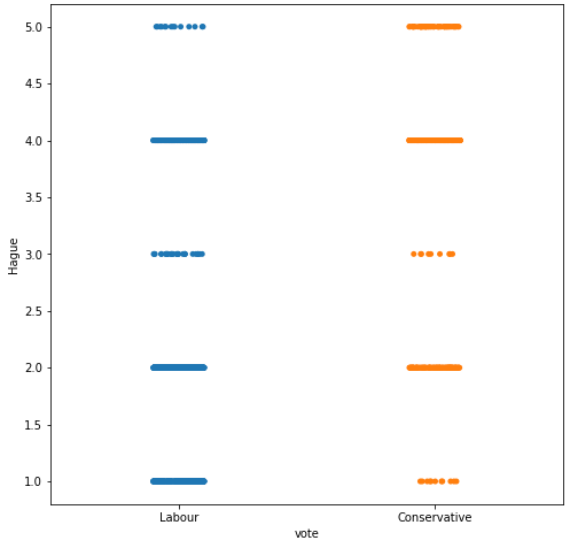
**Bivariate Analysis: -**

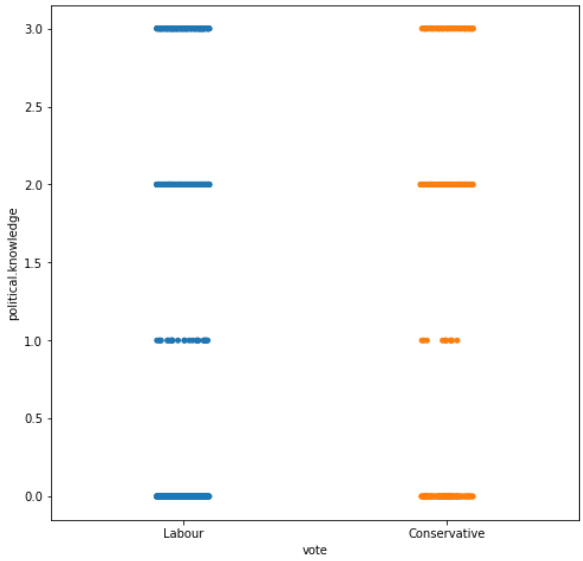




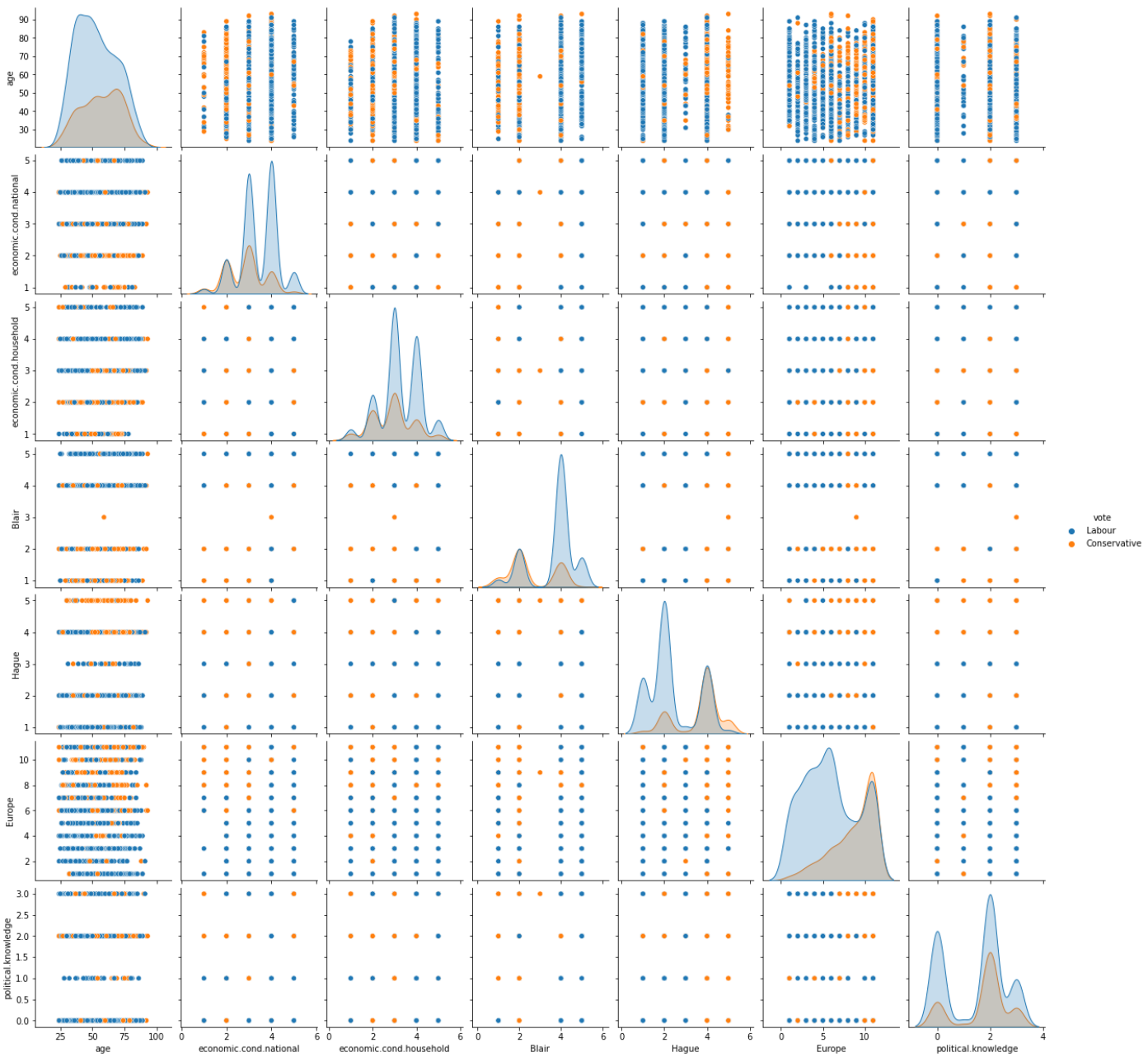



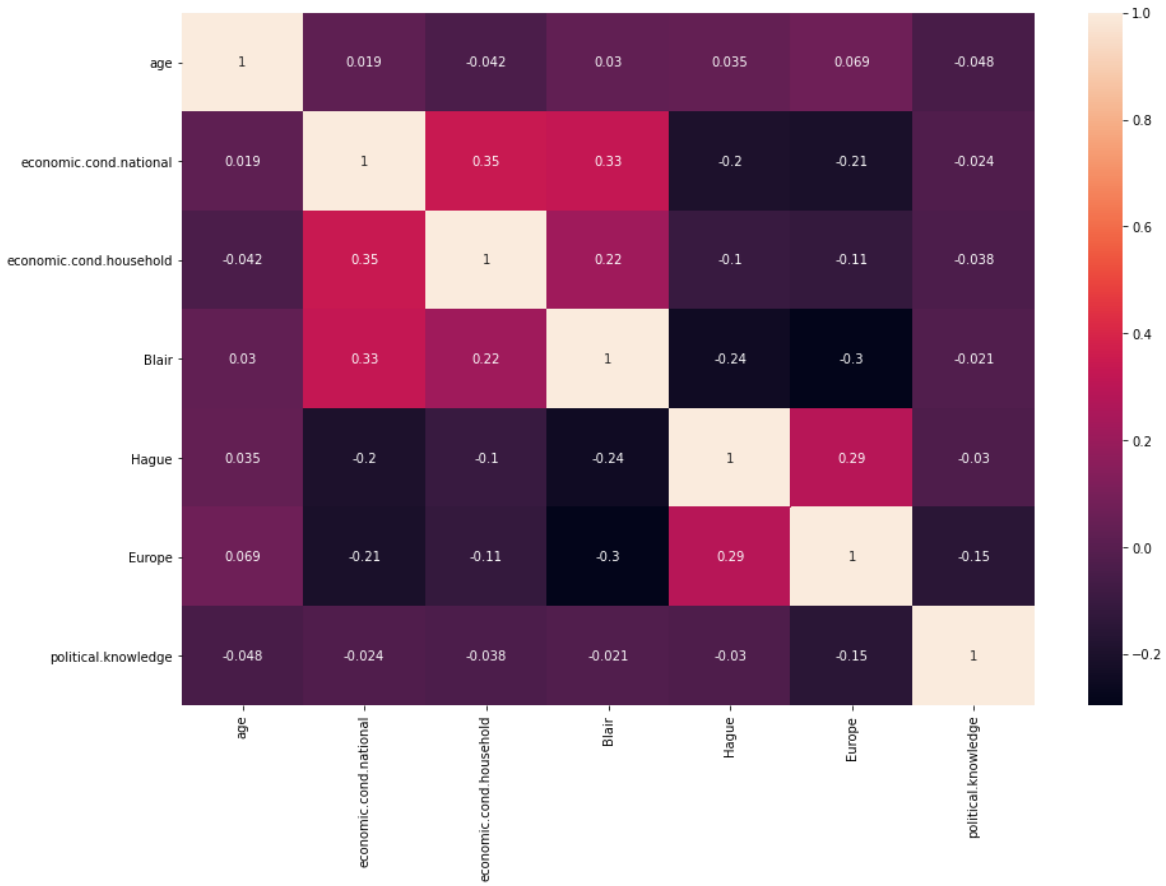

**Pair plot: -**

Pair plot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of kde plot. From the below graph, we can see that there is no any such distinctive factor who clearly distinguish in terms of votes hence, with hue vote columns diagonals overlap each other. Attitude towards the European Integration plays major role. Highest the attitude results vote to the conservative party.


age
economic.cond.national
economic.cond.household
Blair
Hague
Europe
political.knowledge
vote
Labour
Conservative

**Heatmap: -**

From below heat map we can see that there is no positive collinearity between variables.




1.There is some positive relation between economic condition national and household with Blair. And Eurosceptic sentiment with Hague.

2. That means voters choose Blair on the issues of economic condition while Hague is choose for European integration.

**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).**

Encoding -

1.We have to encode vote and gender column as they have string values. One

2.hot encoding turns our categorical data into a binary vector representation. Pandas get dummies makes this very easy! This means that for each unique value in a column, a new column is created. But we have only 2 unique values that's why we use drop=first attribute while running pandas get_dummies function.

| | age | economic.cond.national | economic.cond.household | Blair | Hague | Europe | political.knowledge | vote_Labour | gender_male |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 43 | 3 | 3 | 4 | 1 | 2 | 2 | 1 | 0 |
| **1** | 36 | 4 | 4 | 4 | 4 | 5 | 2 | 1 | 1 |
| **2** | 35 | 4 | 4 | 5 | 2 | 3 | 2 | 1 | 1 |
| **3** | 24 | 4 | 2 | 2 | 1 | 4 | 0 | 1 | 0 |
| **4** | 41 | 2 | 2 | 1 | 1 | 6 | 2 | 1 | 1 |

1.After encoding categorical columns two new columns generated namely vote_Labour and gender_male and as we passed drop_first attribute true our original columns gone.

2.Now our dataset is with 1517 datapoint with all 9 numeric features.

**Scaling -**

Different variables are on different measures means like Age in years continuous variables all others are numeric but categorical in nature. Some of on the scale of 5 while Eurosceptic sentiments on the scale of 11 and Political knowledge is on the scale of 3. Hence, scaling is required for better performance. After scaling each column magnitude is reduced uniformly.

Data Split -

1. The whole given dataset is split into 70:30 proportion using sklearns train_test_split function.

2.Before splitting process, we have to segregate target variable in y and predictors in x.

3. The values of x and y are passed through the train test split function along with test size attribute 0.30 which results splitting x and y into 70: 30 proportions as 70% of the data goes into train set and 30% of remaining data is for test set x_train, x_test, y_train and y_test for the model building.

4. For this particular dataset we have x which contains all columns except vote_Labour and y contains vote_Labour as a target variable.

5. After passing these values into train_test_split function we got the results x_train, x_test and y_train , y_test the shape of train.

**1.4 Apply Logistic Regression and LDA (linear discriminant analysis).**

**Logistic Regression:**

Logistic Regression is white box algorithm which predicts probability values and corresponding cut-offs. Logit function uses linear model to predicting probability of data point belonging to a class. Internally first creates linear regression which gives raw number. Raw number further converted into probability using sigmoid function 1 / (1+e-z )

1. Logistic Regression is available scikit learn's linear model package. We have to import this function and passing derived x_train and y_train to fit the model.

2. After fitting the model we check the model accuracy score using model.score() function. Which gives 0.84.

3. Then test this fitted model on unseen data which is in the x_test and predict the same.

4. For model evaluation, we have actual y_test and predicted values of the x_test from which we can evaluate the model

5. Before that, we check the model accuaracy score again but on the test set. It gives 0.83. It seems model performs better and theres is no overfit/underfit issue with the model.

Accuracy on Train set Logistic Regression: 0.84

Accuracy on Test set Logistic Regression: 0.82

**Linear Discriminant Analysis:**

LDA is a classification technique which is based on logit function. It is used for classification as well as dimensionality reduction practices.  It assumes all independent variables are normally distributed and there is equal variance / covariance for classes.

It classifies observation in to two or more classes where classes are fixed. For applying this model to given dataset the same steps follow as mentioned in logistic regression and we give model accuracy score on train and test set.

Accuracy on Train set LDA Model: 0.84

Accuracy on Test set LDA Model: 0.82

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

**Naïve Bayes Model:**

It based on Bayes theorem to predict and follows Naïve assumption regarding predictors that they are mutually independent. It is simple to implement and fast processing. It works well with small size dataset.

We used Gaussian type of model which assumes that features follow a normal distribution. After successfully following same steps above mentioned we get the model accuracy on train and test set as follows.

Accuracy on Train set Naive Bayes: 0.83

Accuracy on Test set Naive Bayes: 0.82

**KNN Model:**

K-Nearest Neighbours is non parametric method as it do not compute coefficients a and b. Suppose we provide k = 7, it picks random data point and based on Euclidian or Manhattan or whatever distance defined randomly choose 5 nearest data points If 3 or more data points is associated with same class, model predicts the tested data point is associated with majority of class.

Always choose k odd number. If we choose even number, there is rise of tie in the decision.
Accuracy on Train set KNN: 0.85

Accuracy on Test set KNN: 0.83

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

**Bagging:**

If we provide the dataset to the model and using non ensemble method there is chances to overfit the model. For this particular dataset Random Forest overfitted, to overcome this we use bagging classifier which splits dataset random rows and random columns and feuded to the multiple models parallel. Combining all parallel model's prediction results the final output.

1. Bagging classifier is available in the scikit learns ensemble package.

2.We proved Random Forest as a base estimator and n_estimator as 10. Bagging classifier aggregates 10 base estimators' predication by means of voting for classification and provides the final prediction.

3. For the above feeder attributes classifier done great job on training set and provides 0.98 accuracy but on the test data it downs to 0.81 seems to be overfitted model.

Accuracy on Train set Bagging Classifier: 0.98

Accuracy on Test set Bagging Classifier: 0.81

**Boosting:**

Another ensemble method is boosting which is in sequential manner. It takes train set and make prediction and bunch of wrong prediction take as a new subset which is feeded to another model which gives another prediction and, in this models, wrong prediction again feeded to next model. Boosting classifier benefits having simple model. If row gets predicted wrong again and again its weight starts to increase. For the current dataset boosting no overfit or underfit model.

**ADA Boosting:**

Accuracy on Train set Boosting Classifier: 0.84

Accuracy on Test set Boosting Classifier: 0.81

**Gradient Boosting:**

Accuracy on Train set Boosting Classifier: 0.89

Accuracy on Test set Boosting Classifier: 0.83

**Insights.**

1. For the given dataset Logistic Regression model, LDA, Naïve Bayes performs well on train as well as test set. These models are enough tune to capture 89% of the data

2. KNN also good model but as the AUC concerned it is slightly going down to 87% on unseen data while other performance metrics like accuracy, recall, precision and f1 score remains same.

3. Bagging and Boosting with Random Forest as base classifier do not give satisfactory results both methods train the model well but not performing well on the test data.

4. Overall Logistic Regression and LDA are the best models to create an exit poll that will help in predicting overall win and seats covered by a particular party.

5. Comparison of models based on performance metrics is shown below

**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.**

**<u>Logistic Regression</u>**

**Train Evaluation**

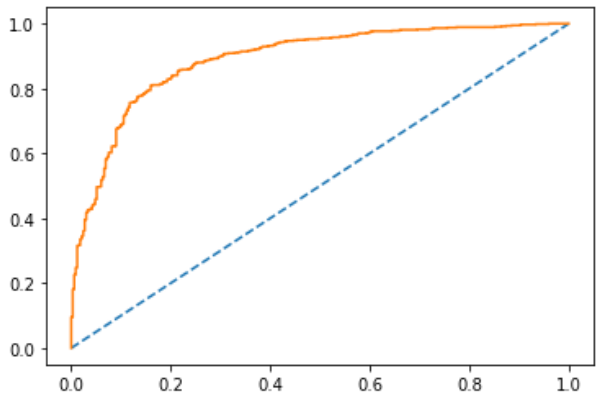
1. Accuracy: 0.84

2. Precision: 0.87

3. Recall: 0.91.

4. F1 Score: 0.89

5. AUC: 0.88



**Test Evaluation**

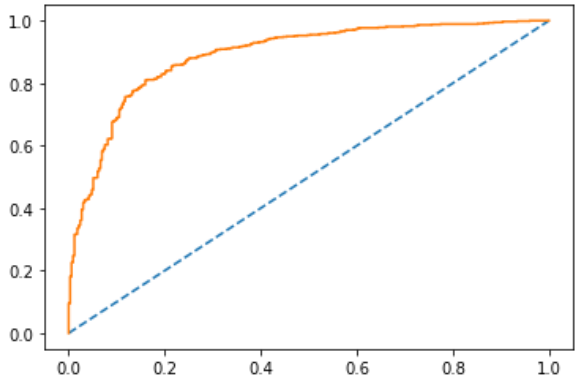
1. Accuracy: 0.82

2. Precision: 0.87

3. Recall: 0.89.

F1 Score: 0.88

5. AUC Score: 0.88




## Linear Discriminant Analysis

**Train Evaluation**

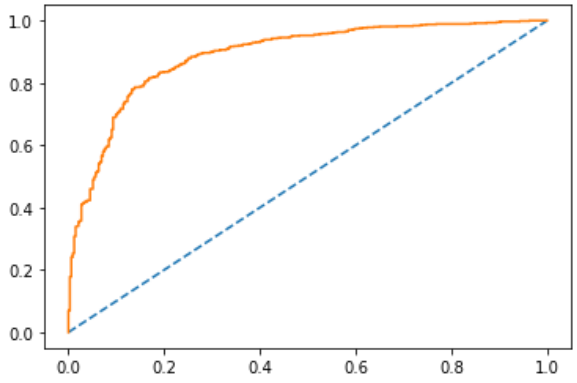
1. Accuracy: 0.84

2. Precision: 0.87

3. Recall: 0.90

4. F1 Score: 0.88

5. AUC: 0.88




**Test Evaluation**

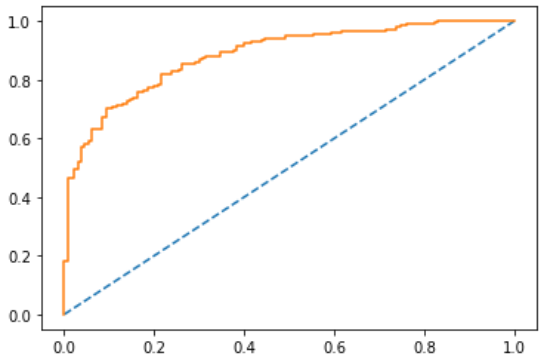
1. Accuracy: 0.82

2. Precision: 0.87

3. Recall: 0.89.

4.F1 Score: 0.87

5. AUC Score: 0.88




**<u>Naïve Bayes</u>**

**Train Evaluation**

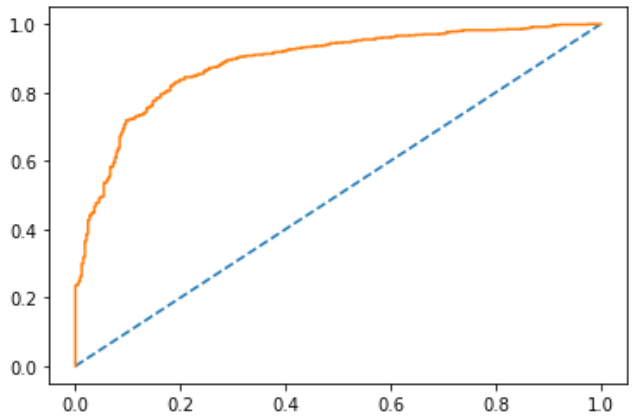
1. Accuracy: 0.83

2. Precision: 0.88

3. Recall: 0.88

4. F1 Score: 0.88

5. AUC: 0.886




**Test Evaluation**

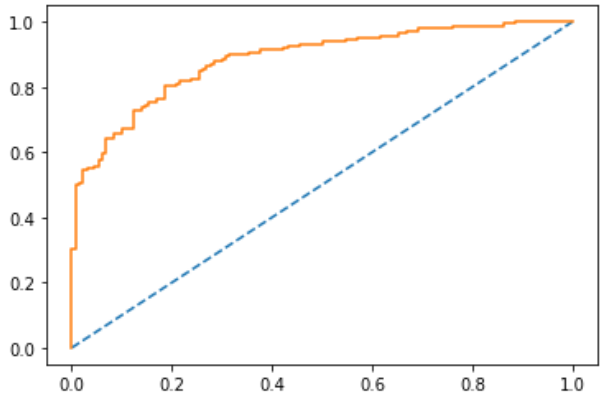
1. Accuracy: 0.82

2. Precision: 0.89

3. Recall: 0.87

4. F1 Score: 0.88

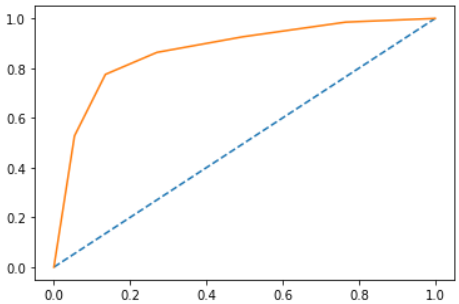
5. AUC: 0.885



**KNN: -**

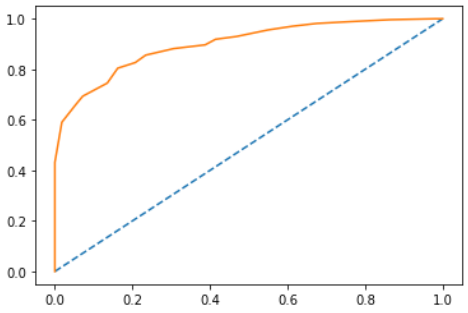
**Train Evaluation**

1. Accuracy: 0.85
2. Precision: 0.88
3. Recall: 0.91
4. F1 Score: 0.90
5. AUC: 0.904




**Test Evaluation**

1. Accuracy: 0.83
2. Precision: 0.90
3. Recall: 0.87
4. F1 Score: 0.88
5. AUC: 0.900



## Bagging: -

**Train Evaluation**

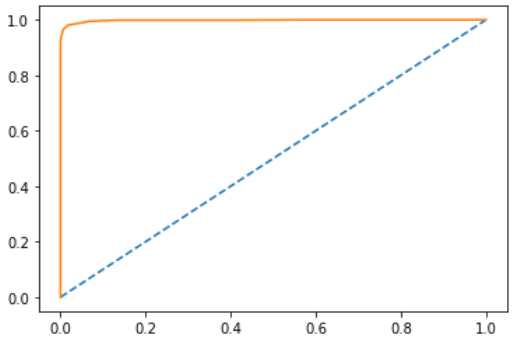
1. Accuracy: 0.98

2. Precision: 0.99

3. Recall: 0.98

4. F1 Score: 0.99

5. AUC: 0.998




**Test Evaluation**

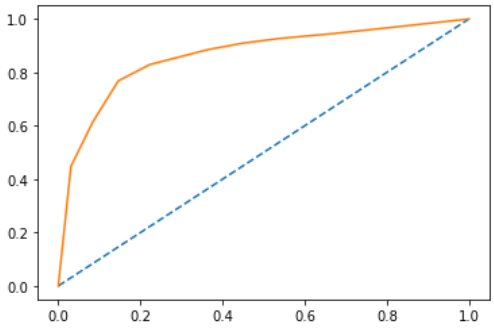
1. Accuracy: 0.81

2. Precision: 0.88

3. Recall: 0.86

4. F1 Score: 0.87

5. AUC: 0.863



## Boosting: -

### Train Evaluation:

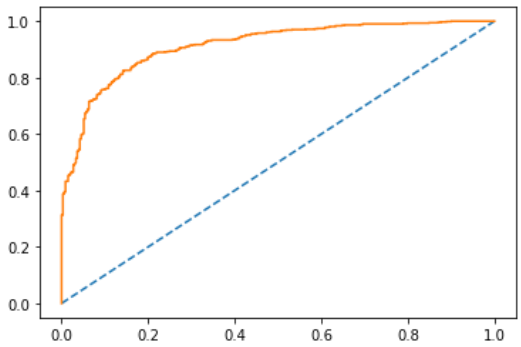
1. Accuracy: 0.85

2. Precision: 0.88

3. Recall: 0.91

4. F1 Score: 0.89

5. AUC: 0.913




### Test Evaluation:

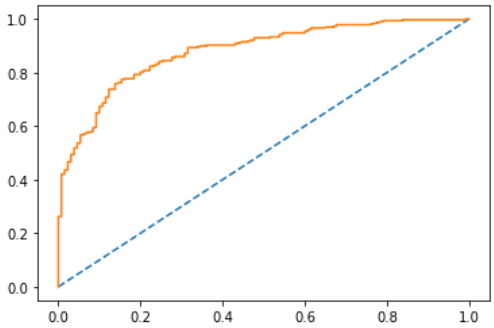
1. Accuracy: 0.82

2. Precision: 0.89

3. Recall: 0.87

4. F1 Score: 0.88

5. AUC: 0.879



**1.8 Based on these predictions, what are the insights?**

1. Most people's political views and party allegiances are fairly constant.

2. In Europe but not Run by Europe i.e., Eurosceptism, Conservative party's stand accepted majorly by old aged and female voters but not sure about the leader Hague.

3. For the labour party, Blair's performance recognized by most of the people. But European Integration with new policies can't captures females vote, which is larger in proportion than male.

4. Within a constituency, there can be very wide socio-economic variation between the voters to vote that swing plays major roles because there is only two-party politics. The vote swing is clearly between them. That Labour party's vote count decrease results vote increase in the conservative party.

5. Voters who vote by post are not included in exit polls. This is potentially a source of bias, if the pattern of vote-changing among postal voters differs from the vote

changing behavior of those who use a polling station.

**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.**

After downloading inaugural corpora in NLTK we get several fileids. From which for this particular problem statement we consider only three text files mainly, President Roosevelts speech in 1941. President Kennedy's in 1961 and President Nixon's in 1973.

Simply applying len function on raw file of each president's speech we get number of characters including spaces the results are below.

```
Number of Character in Roosevelt Speech : 7571
Number of Character in Kennedy Speech : 7618
Number of Character in Kennedy Speech: 9991
```

**Words () :**

```
Roosevelts word count =  1536
Kennedy word count = 1546
Nixon word count = 2028
```

**Sents ():**

```
Number of Sentences in Roosevelt Speech : 68
Number of Sentences in Kennedy Speech : 52
Number of Sentences in Kennedy Speech: 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”) 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 default there are 179 stopwords in the nltk corpora of stopwords for English language.

Taking each word in the text using word_tokenizer function and check it with stop words list

Before that making lower case of each of the token words for uniformity, case sensitivity

After these steps text clean and counts of words before and after removal of stop words is

Before Stemming and Stopwords removal wordcount of Roosevelt : 1536
After Stemming and Stopwords removal wordcount of Roosevelt : 635

Before Stemming and Stopwords removal wordcount of Kennedy : 1546
After Stemming and Stopwords removal wordcount of Kennedy : 708

Before Stemming and Stopwords removal wordcount of Nixon : 2028
After Stemming and Stopwords removal wordcount of Nixon: 864

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

**Roosevelt**

```
[('nation', 17),
 ('know', 10),
 ('peopl', 9),
 ('spirit', 9),
 ('life', 9),
 ('democraci', 9),
 ('becaus', 9),
 ('us', 8),
 ('america', 8),
 ('live', 7),
 ('year', 7),
 ('human', 6),
 ('freedom', 6),
 ("'s", 5),
 ('measur', 5)]
```

**Kennedy:-**

```
[('let', 16),
 ('us', 12),
 ('power', 9),
 ('world', 8),
 ('nation', 8),
 ('ani', 8),
 ('side', 8),
 ('new', 7),
 ('pledg', 7),
 ('ask', 6),
 ('citizen', 5),
 ('peac', 5),
 ('shall', 5),
 ('free', 5),
 ('final', 5)].
```

Nixon: -

```
[('us', 26),
 ('let', 22),
 ('america', 21),
 ('peac', 19),
 ('world', 18),
 ('respons', 17),
 ('new', 15),
 ('nation', 15),
 ("'s", 14),
 ('great', 12),
 ('govern', 10),
 ('year', 9),
 ('home', 9),
 ('abroad', 8),
 ('make', 8)]
```

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

1941 – Roosevelt's Speech Talked about

Nation

People

Spirit

Life

Democracy

American

**1961 – Kennedy's Speech**

Talked about

let

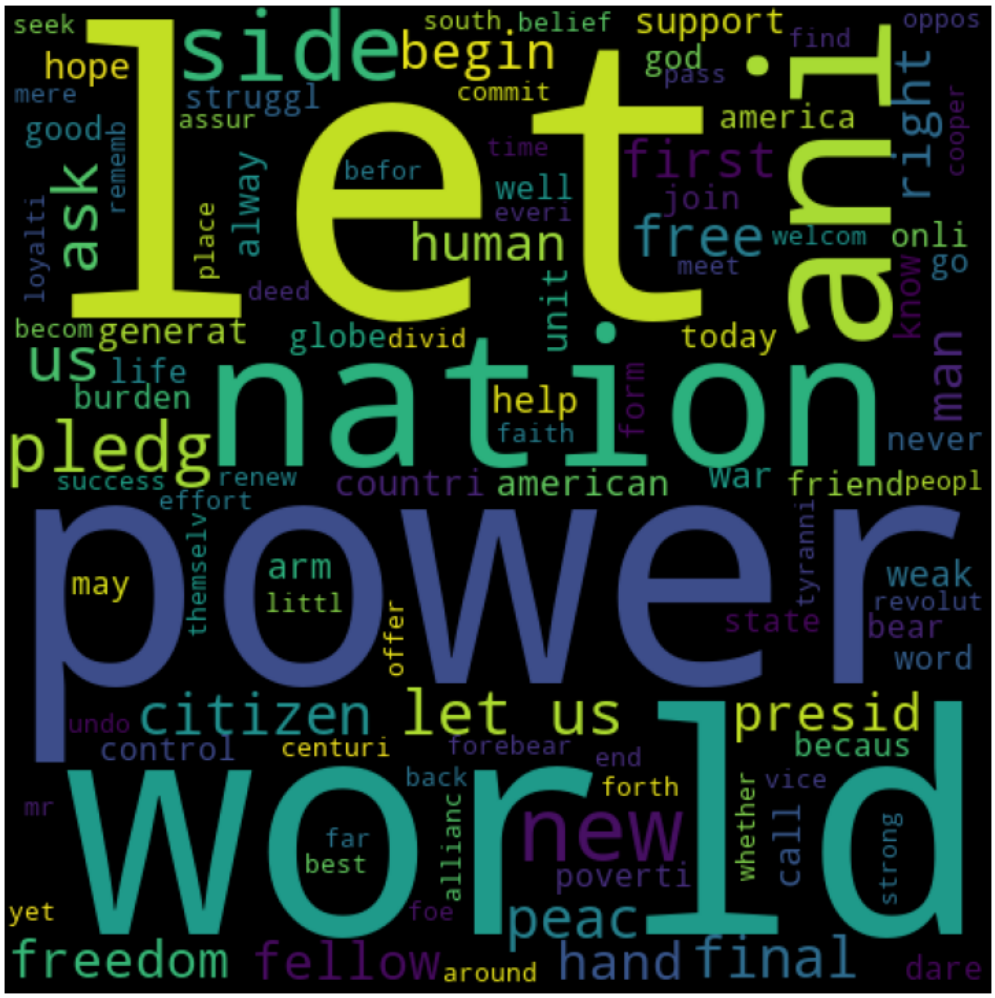
nation

power

world

pledge

peace



**1973 – Nixon's Speech**

Talked about

America

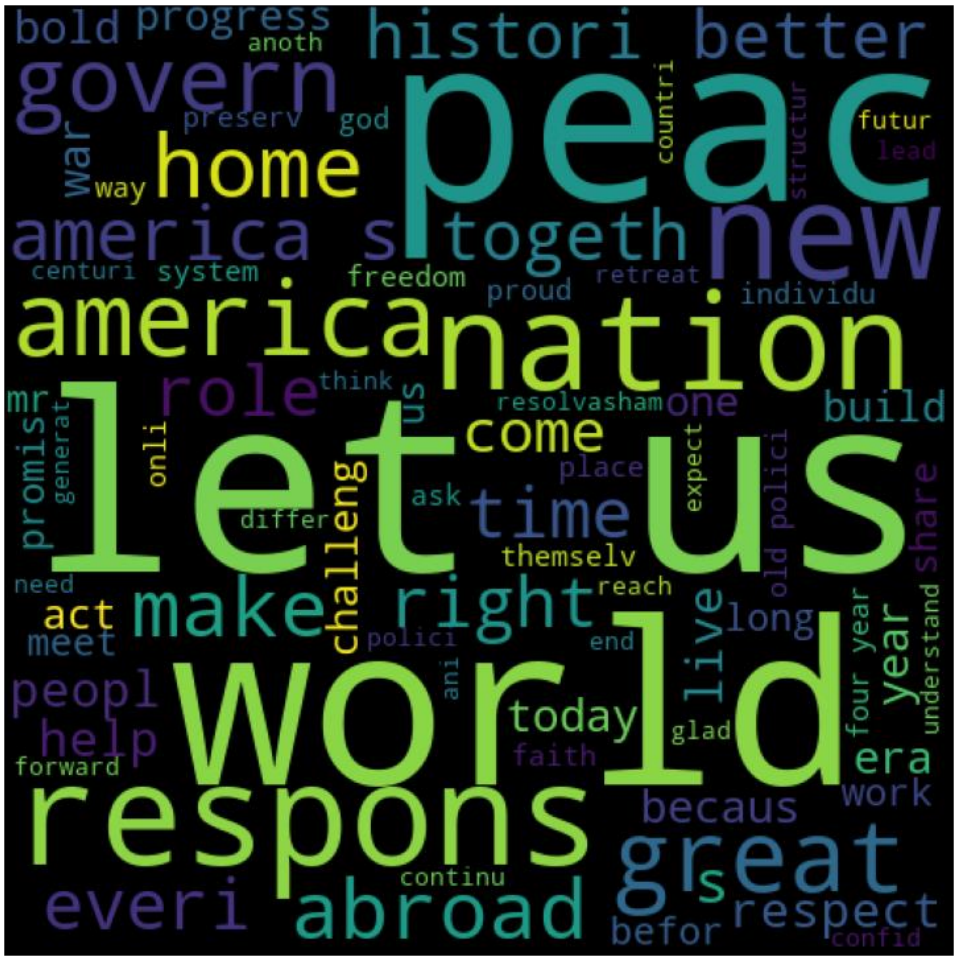
Peace

World

Let us

response

new nation




----------END OF THE REPORT----------