

Business Report

MACHINE LEARNING
PGP-DSBA

TABLE OF CONTENTS

Problem Statement 1 – CNBE channel’s Exit poll	1
1.1 Read the dataset Do the descriptive statistics and do the null value condition check. Write an inference on it	2
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.	8
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).	12
1.4 Apply Logistic Regression and LDA (linear discriminant analysis).....	13
1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results	14
1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting	15
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.	17
1.8 Based on these predictions, what are the insights?	20
 Problem Statement 2 – Inaugural Corpora Presidents Speech.....	21
2.1 Find the number of words, sentences and characters of mentioned documents.....	21
2.2 Remove all the stop words from all three speeches.....	22
2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top 3 words (after removing the stop words)	22
2.4 Word clouds after cleaning texts for each president.....	23

Executive Summary:

CNBE one of the leading news channels who wants to analyze recent elections. Survey was conducted on 1525 voters with 9 different questionnaires. On the basis of the given information, voter will vote for which party, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Introduction:

We have to build a model, to predict which party do well in the election campaign and convey their ideology, plans for nation and overall performance on the ground and voters are ready to cast their vote.

And in perspective of voters, what things really matters to build their confidence in such party. Is it party ideology matters? Or leadership or anything else promises in the manifesto during the campaign.

Data Description:

The given dataset [Election Data.xlsx](#) is containing data of 1525 different voters with 9 different variables. Such as Which party choose to vote, what is the age of voter, their assessment regarding national economical condition as well as their owned households economic condition, assessing the leader of the parties labour and conservative, their attributes towards the European Integration, Knowledge about political parties and gender.

Shape of the dataset is (1525) rows and (9) columns.

- | | | |
|-----------------------------------|--|--------------------------|
| 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. | |

Data Ingestion:

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

The given dataset is load in jupyter notebook using pandas read_csv function and read first five entries of the dataset using head function. The first five entries in the dataset is looks as given below.

	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

(fig.1 – First Five Entries of the 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.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0            1525 non-null  int64
1   vote                  1525 non-null  object
2   age                  1525 non-null  int64
3   economic.cond.national 1525 non-null  int64
4   economic.cond.household 1525 non-null  int64
5   Blair                1525 non-null  int64
6   Hague                1525 non-null  int64
7   Europe               1525 non-null  int64
8   political.knowledge  1525 non-null  int64
9   gender               1525 non-null  object
dtypes: int64(8), object(2)
memory usage: 119.3+ KB
```

(fig.2[a] – Information of the dataset)

```
df.isnull().sum()

Unnamed: 0      0
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
```

(fig.2 [b]– null value condition check)

- As we seen in the fig.2[a] There are 10 variables the 9 variables discuss in the Data Description and Unnamed:0 is nothing but the indexing variable.

Hence, this cannot make any significance role in the process to getting insights we can drop it further.

- 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 shows the int64 datatype but they are also categorical in nature except age which is numerical in nature.
- There is no any null value present in the dataset.

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	1525.0	763.000000	440.373894	1.0	382.0	763.0	1144.0	1525.0
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

(fig 3[a] – Description of the numerical variable)

- From fig.3[a] we got the information about numerical variables as we said earlier Unnamed:0 is only for the purpose of indexing the min and max values of the Unnamed:0 proved it.
- The voters age ranges from 24 to 93 years. Mean and Median values of the Age shows the nearly same that means Age is normal distributed.
- Assessment or opinion about economic condition either national or household look like same.
- 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
- The voters have fairly knowledge about the party's position on European integration and the they show their European integration attitudes.

Target variable vote and gender is the object type of data. Description of these variable is showed by including include='object' attribute in the describe function.

	count	unique	top	freq
vote	1525	2	Labour	1063
gender	1525	2	female	812

(fig3[b] – Description of Object Datatype variable)

The above fig3[b] shows the description of the vote and gender variable.

- Vote has 2 unique values in the dataset of 1525 datapoints.. Labour class or party is on the top with the frequency of the 1063 vote.
- Gender also has 2 unique values male and female. Female is on the top with 812 entries.

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, as per the fig3[a] 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.

Duplicates Check

- There is no any duplicate data present in the dataset. When we keep 'Unnamed: 0' column as it is.

```
dups = df.duplicated()
print('Number of duplicate rows = %d' % (dups.sum()))
df[dups]
```

Number of duplicate rows = 0

Unnamed: 0 vote age economic.cond.national economic.cond.household Blair Hague Europe political.knowledge gender

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

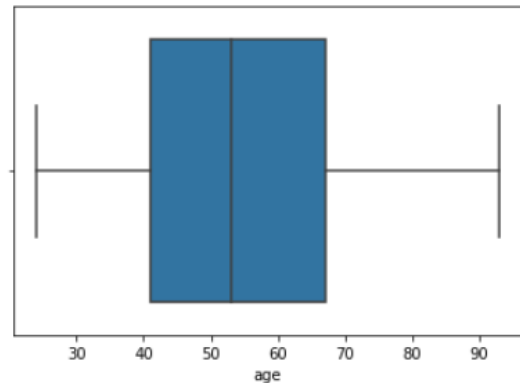
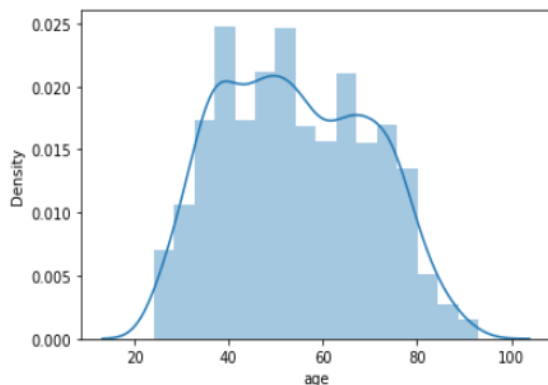
Number of duplicate rows = 8

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
67	Labour	35	4	4	5	2	3	2	male
626	Labour	39	3	4	4	2	5	2	male
870	Labour	38	2	4	2	2	4	3	male
983	Conservative	74	4	3	2	4	8	2	female
1154	Conservative	53	3	4	2	2	6	0	female
1236	Labour	36	3	3	2	2	6	2	female
1244	Labour	29	4	4	4	2	2	2	female
1438	Labour	40	4	3	4	2	2	2	male

After dropping these 8 entries we reset index and got the new shape of the dataset as 1517 rows and 9 columns.

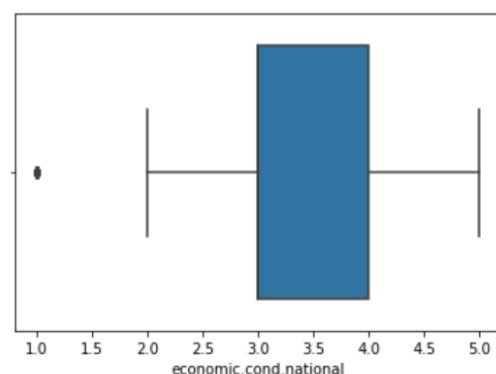
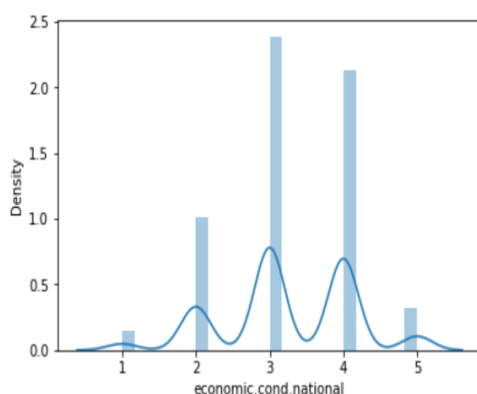
#1. Age:

- Total 1517 entries.
- Age ranges from 24 to 93
- Mean age of the voters in the dataset is 54
- Standard deviation of the age column is 15
- Median of the age column is 53
- 25 percentile of the age column 41 and 75 percentile is 67
- Data shows normally distributed.
- No outliers.



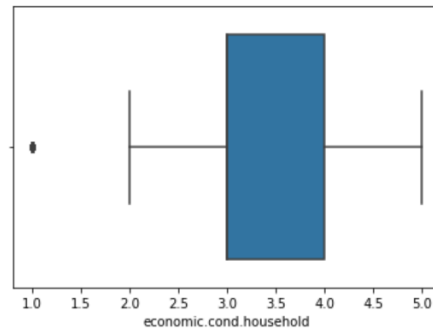
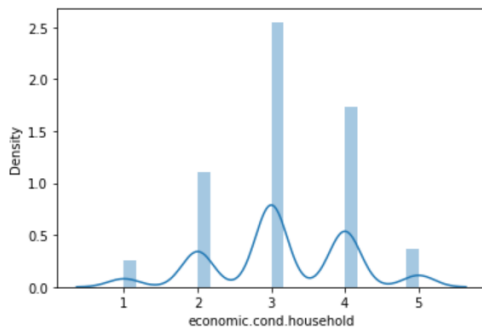
#2. Economic Condition National:

- Total 1517 data points.
- Ranges from 1 to 5 i.e. assessed ratings 1 as poor and 5 as best
- Mean of economic condition national is 3.245
- Standard deviation is 0.88 means 1,
- 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.
- Outliers shown as a data points 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:

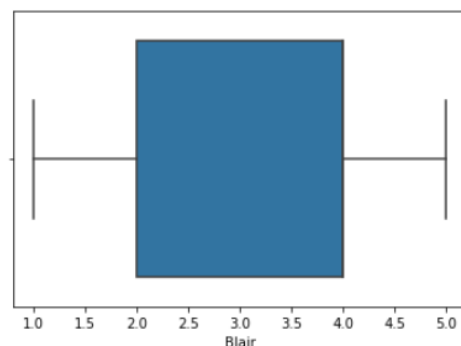
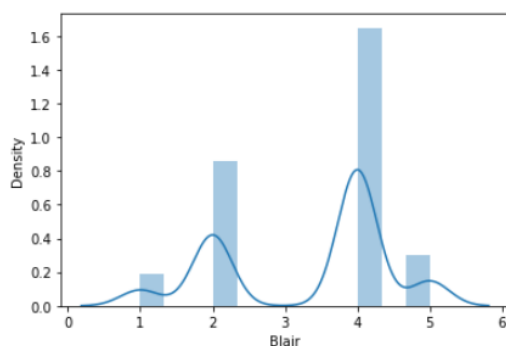
- Total 1517 data points.
- Ranges from 1 to 5 i.e. assessed ratings 1 as poor and 5 as best
- Mean of economic condition national is 3.13
- Standard deviation is 0.93 means 1,
- Median of the column is 3
- Outlier presents. Data shows the major of voters assessed economic condition is good and better 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

- Total 1517 data points.
- Ranges from 1 to 5 i.e. assessed ratings 1 as poor and 5 as best
- Mean of Blair is 3.33
- Standard deviation is 1.17
- Median of the column is 4
- No outliers.

The mean of Blair = 3.3355306526038233
The median of Blair = 4.0

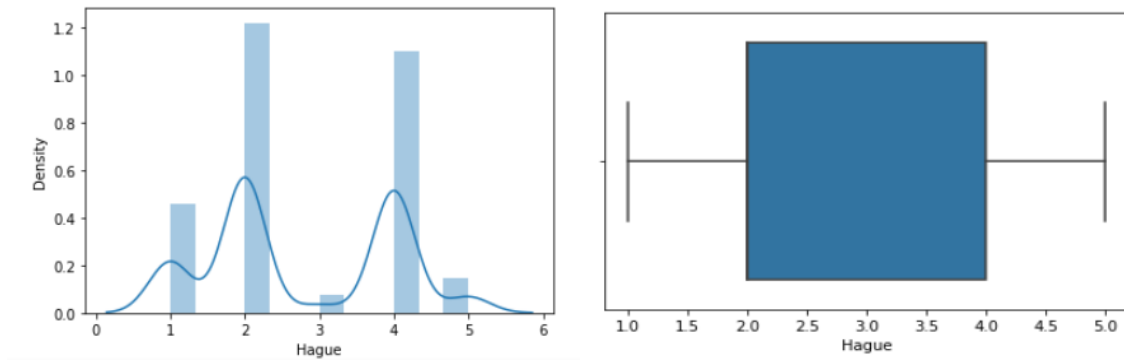


#5. Hague

- Total 1517 data points.
- Ranges from 1 to 5 i.e. assessed ratings 1 as poor and 5 as best
- Mean of Hague is 2.749
- Standard deviation is 1.23

- Median of the column is 2
- No outliers.

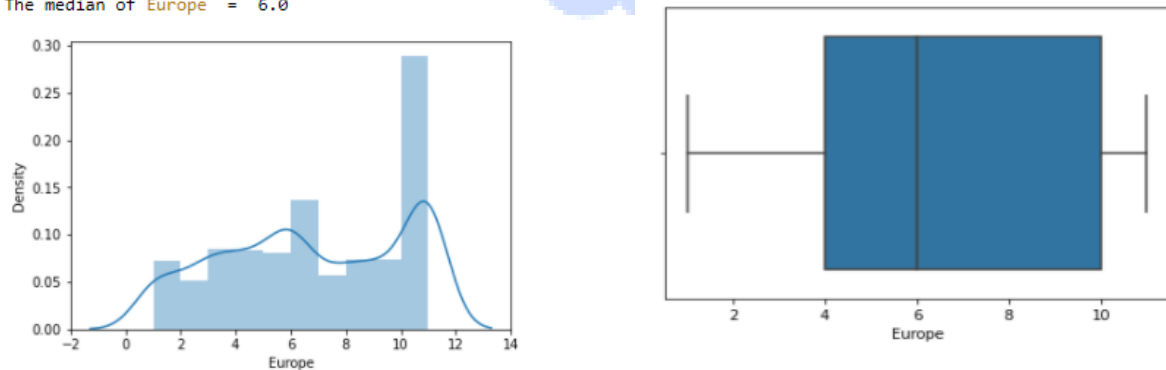
The mean of **Hague** = 2.7495056031641396
 The median of **Hague** = 2.0



#6. Europe

- Total 1517 data points.
- Ranges from 1 to 11 i.e. assessed ratings 1 as poor and 11 Highest 'Eurosceptic' sentiment.
- Mean of Europe is 6.740
- Standard deviation is 3.3
- Median = 6
- No outliers. But highest density of the peoples attitude is towards the 'Eurosceptic' sentiments.

The mean of **Europe** = 6.7402768622280815
 The median of **Europe** = 6.0

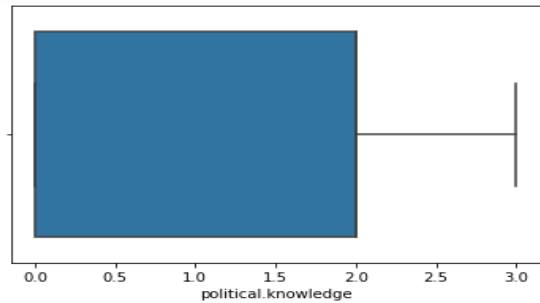
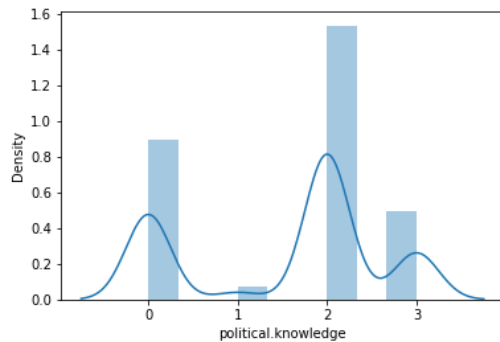


#7. Political Knowledge

- Total 1517 data points.
- Ranges from 1 to 3 i.e. knowledge of the party's position on European Integration.
- Mean of Political Knowledge is 1.54
- Median of the columns is 2
- Standard deviation is 1.08

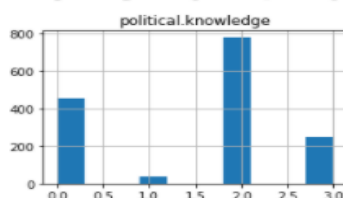
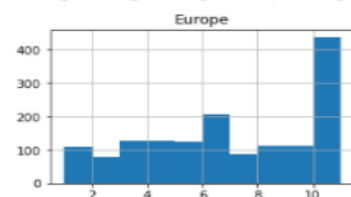
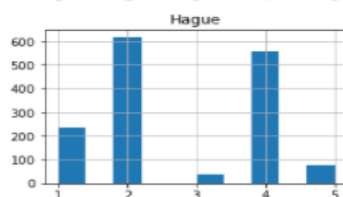
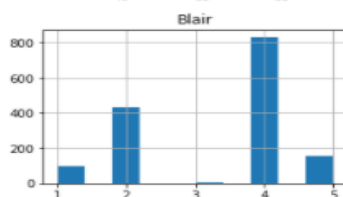
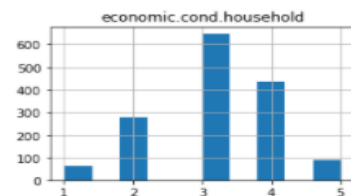
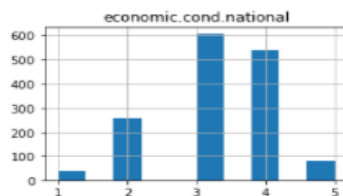
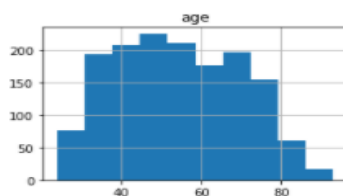
- No outliers. But most of the peoples have knowledge about the position on European integration of the labour and conservative party respectively.
- There is good amount of peoples that does not have any idea about the positions on the European integration of the parties.

The mean of `political.knowledge` = 1.5405405405405406
The median of `political.knowledge` = 2.0



Skewness of the	: age	0.14
Skewness of the	: economic.cond.national	-0.24
Skewness of the	: economic.cond.household	-0.14
Skewness of the	: Blair	-0.54
Skewness of the	: Hague	0.15
Skewness of the	: Europe	-0.14
Skewness of the	: political.knowledge	-0.42

Histograms

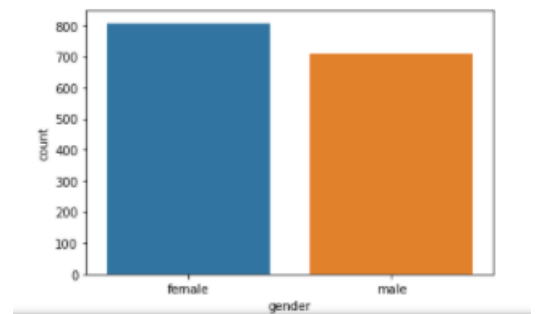


#8. Gender

- Total 1517 data points.
- Categorical in nature with two values male and female.
- female voters are greater than male with frequency of 812 out of 1517 data points.

GENDER : 2
male 709
female 808
Name: gender, dtype: int64

```
female 0.53263  
male 0.46737  
Name: gender, dtype: float64
```

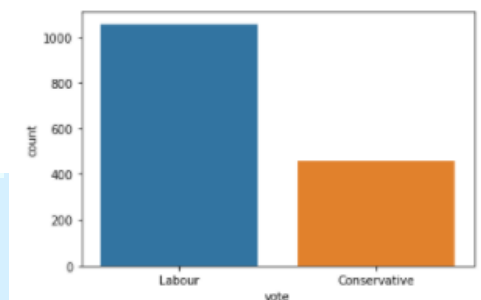


#9. Vote

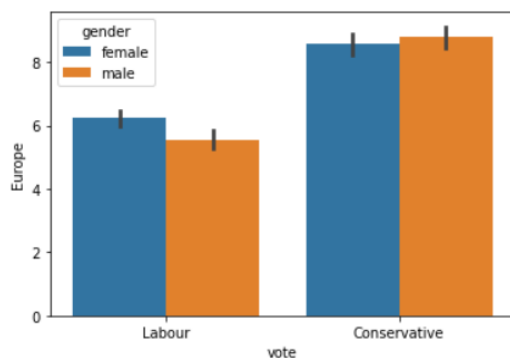
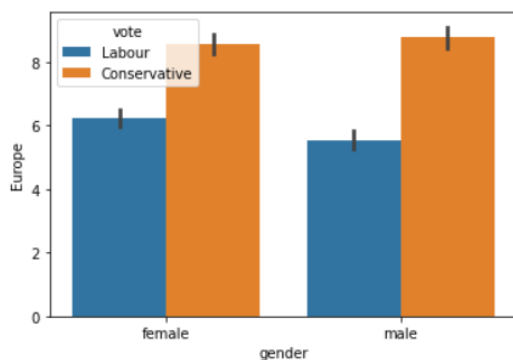
- This is target variable
- Total 1517 data points.
- Categorical in nature with two values Labour and Conservative.
- Labour party have maximum votes with total 69% of vote share.

VOTE : 2
Conservative 460
Labour 1057
Name: vote, dtype: int64

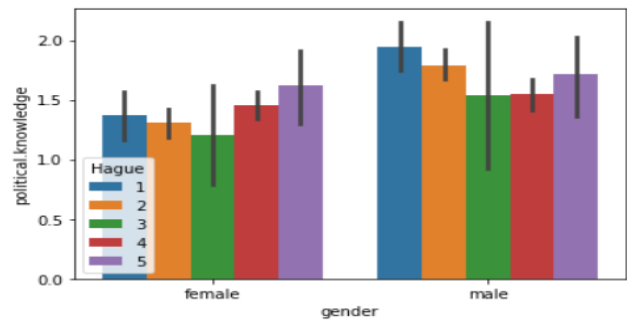
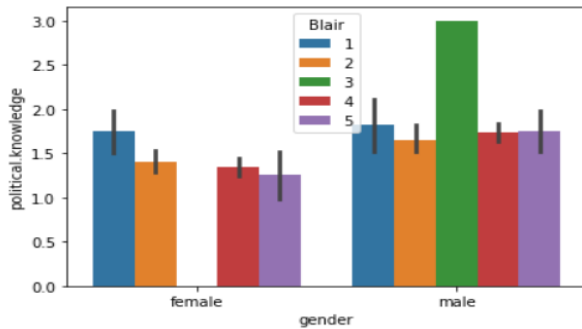
```
Labour 0.69677  
Conservative 0.30323  
Name: vote, dtype: float64
```



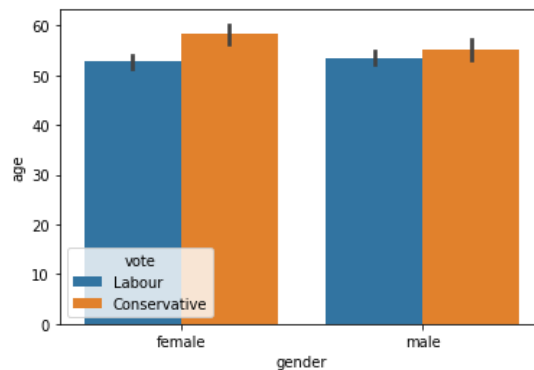
Bivariate:



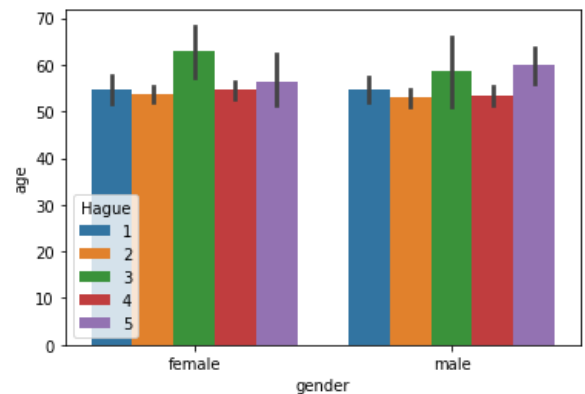
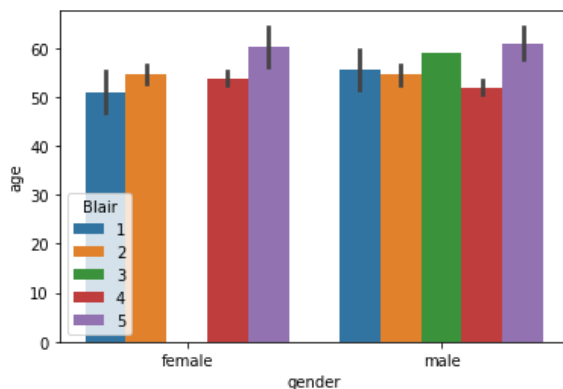
- As far as European integration concerned, male and female both with highest Eurosceptic sentiment vote for conservative party.
- Upto the range of 6 scale of sentiment male and female both choose labour party.



- Assessment of the political knowledge of Blairs stands on European Integration. Male voters have confident in the Blairs labour party's position. And on the other hand that much amount of female voters does not think same.
- Hague's party's position on European integration maximum number of male voters have not much confident about Hague but equal number of male and females shows the extreme confidence which shown by purple bar in above barplot..



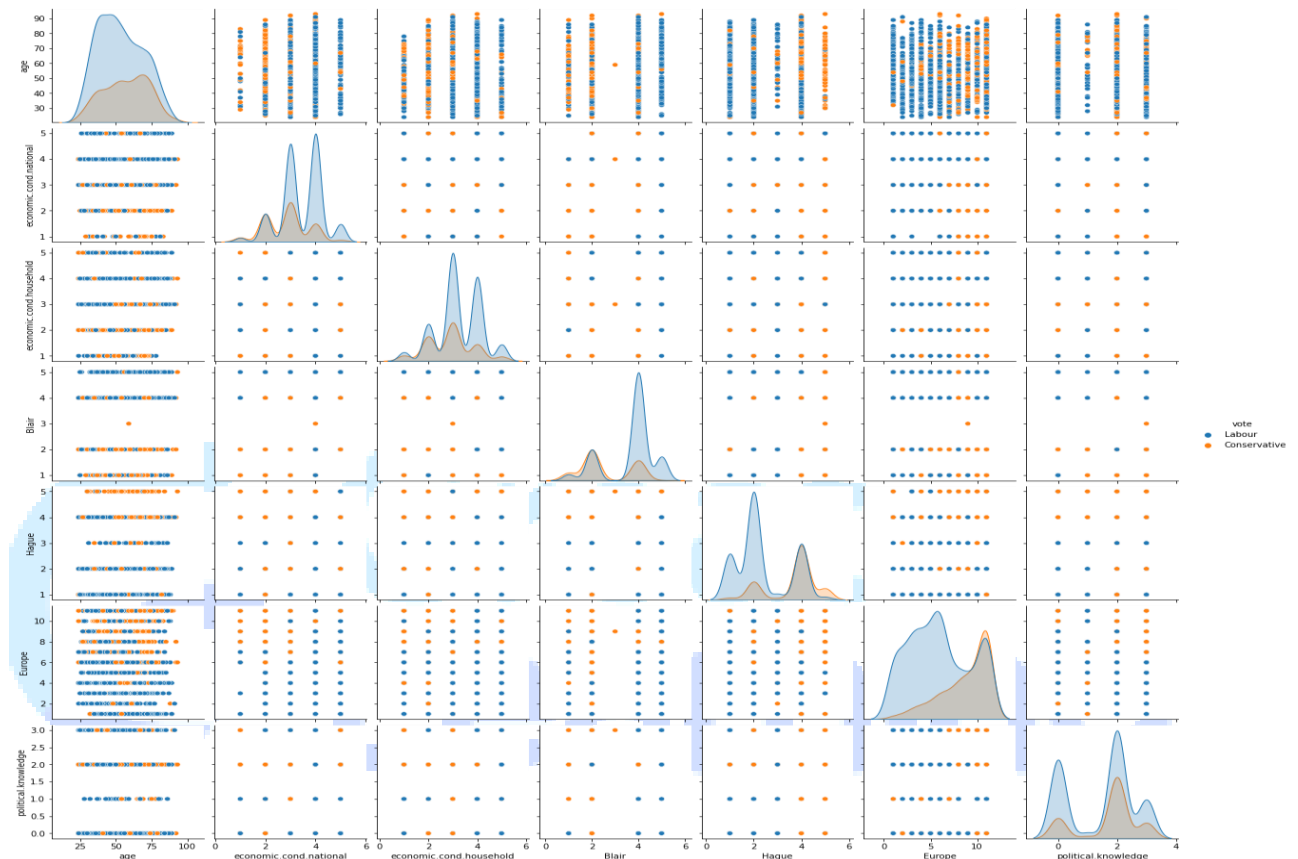
ing



- Old age voters either male or female choose conservative party
- Blair's performance assessment is better than Hague's. Females are more clear and assessed highest.
- Hague's assessment shows maximum middle value or neutral expect.

Pairplot

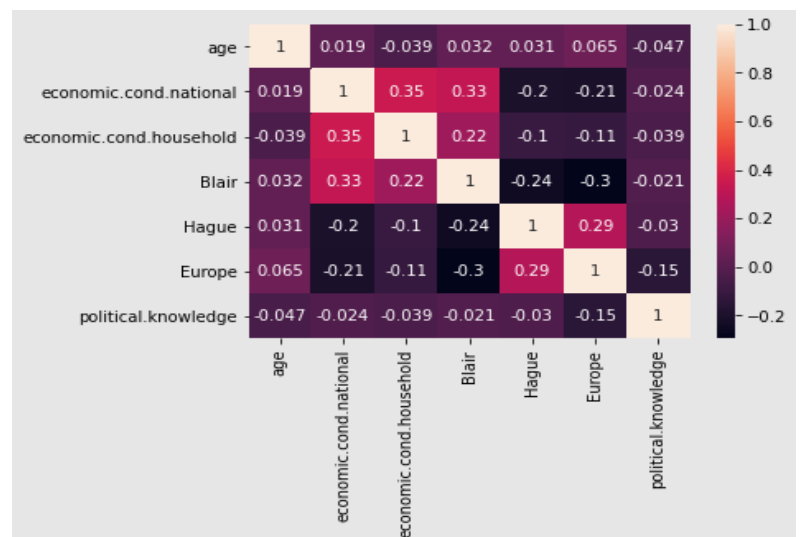
Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of kde plot. From the below graph, we can see that there is no any such distinctive factor who clearly distinguish in terms of votes hence, with hue vote columns diagonals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



Heatmap

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

- There is some positive relation between economic condition national and household with Blair. And Eurosceptic sentiment with Hague.
- That means voters choose Blair on the issues of economic condition while Hague is choose for European integration.



Data Preparation:

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 -

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1517 entries, 0 to 1524
Data columns (total 9 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   age                                   1517 non-null   int64
1   economic.cond.national                1517 non-null   int64
2   economic.cond.household               1517 non-null   int64
3   Blair                                 1517 non-null   int64
4   Hague                                 1517 non-null   int64
5   Europe                                1517 non-null   int64
6   political.knowledge                   1517 non-null   int64
7   vote_Labour                           1517 non-null   uint8
8   gender_male                           1517 non-null   uint8
dtypes: int64(7), uint8(2)
memory usage: 137.8 KB
```

- 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.
- 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 -

- The whole given dataset is split into 70:30 proportion using sklearn's train_test_split function.

- Before splitting process, we have to segregate target variable in y and predictors in x.
- 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.
- For this particular dataset we have x which contains all columns except vote_Labour and y contains vote_Labour as a target variable.
- 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 and test set is

```
: print('Train set size:', x_train.shape)
  print('Test set size:', x_test.shape)
```

```
Train set size: (1061, 8)
Test set size: (456, 8)
```

Modeling:

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 accuracy 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: 0.83

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 follows as mentioned in logistic regression and we gives 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.

The diagram illustrates Bayes' Theorem with the following components and labels:

- Likelihood:** Points to $P(x|c)$ in the numerator of the first equation.
- Class Prior Probability:** Points to $P(c)$ in the numerator of the first equation.
- Posterior Probability:** Points to $P(c|x)$ on the left side of the first equation.
- Predictor Prior Probability:** Points to $P(x)$ in the denominator of the first equation.

The first equation is:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

The second equation is:

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

scikit learn library helps here to build a Naive Bayes model in Python. There are three types of Naive Bayes model under the scikit-learn library:

1. Gaussian
2. Multinomial
3. Bernoulli

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: 0.84

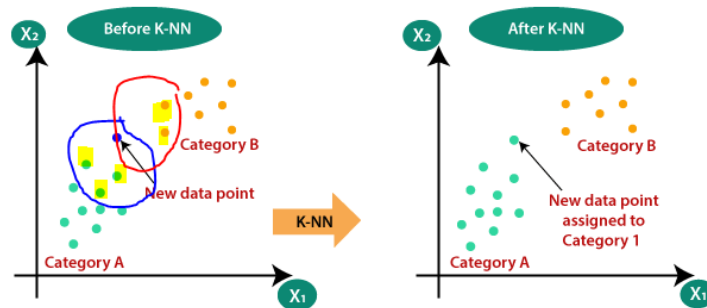
Accuracy on Test set Naive Bayes: 0.83

KNN Model: K-Nearest Neighbours is non parametric method as it do not compute coefficients a and b. Suppose we provide $k = 5$, 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: 0.87
Accuracy on Test set KNN: 0.82



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 overfitt 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 feeded to the multiple models parallel. Combining all parallel models prediction results the final output.

- Bagging classifier is available in the scikit learns ensemble package.
- 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.
- For the above feeded attributes classifier done great job on training set and provides 0.96 accuracy but on the test data it downs to 0.82 seems to be overfitted model.

Accuracy on Train set Bagging Classifier : 0.96
Accuracy on Test set Bagging Classifier : 0.82

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 also overfitted the model.

Accuracy on Train set Boosting Classifier : 1.00

Accuracy on Test set Boosting Classifier : 0.81

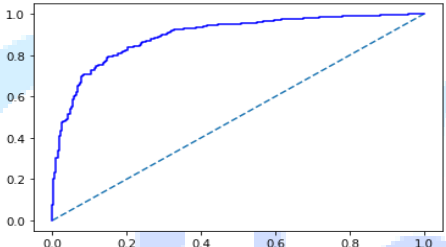
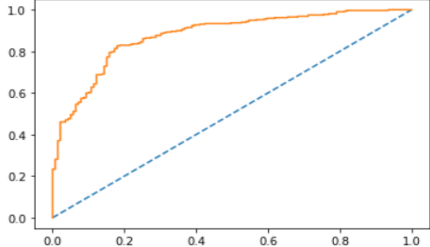
Insights.

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 84% 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 gives satisfactory results both methods train the model well but not performing well on the test data and model over fitted.
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.

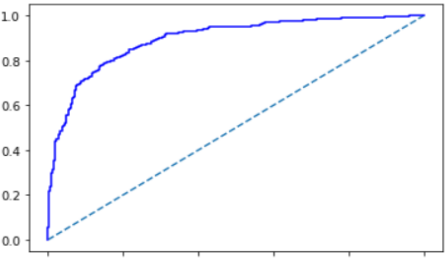
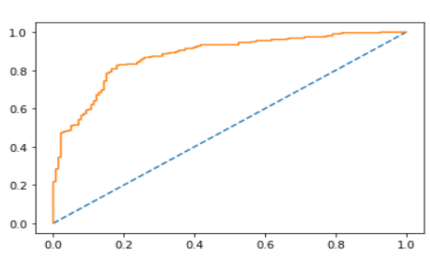
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

Model Evaluation

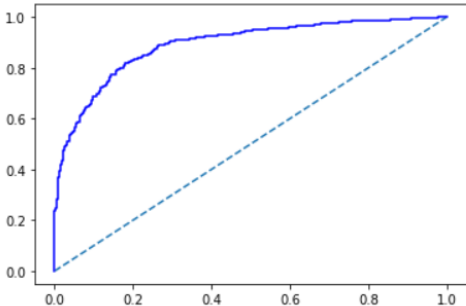
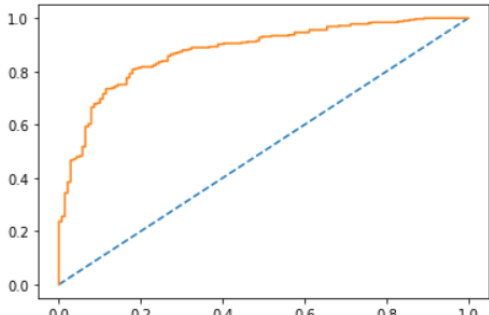
Logistic Regression

Train Evaluation	Test Evaluation
1. Accuracy : 0.84 2. Precision : 0.87 3. Recall : 0.92 4. F1 Score : 0.89	1. Accuracy : 0.83 2. Precision : 0.86 3. Recall : 0.90 4. F1 Score : 0.88
<ul style="list-style-type: none"> AUC : 0.89 ROC Curve : 	<ul style="list-style-type: none"> AUC : 0.88 ROC Curve : 

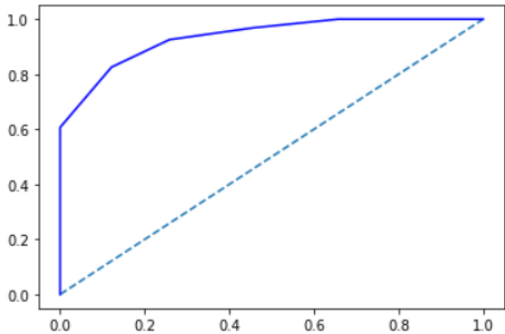
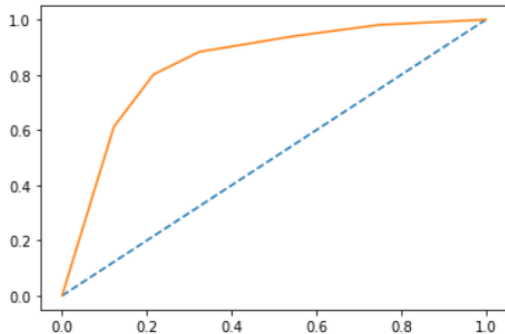
Linear Discriminant Analysis

Train Evaluation	Test Evaluation
1. Accuracy : 0.84 2. Precision : 0.87 3. Recall : 0.90 4. F1 Score : 0.89	1. Accuracy : 0.82 2. Precision : 0.86 3. Recall : 0.89 4. F1 Score : 0.88
<ul style="list-style-type: none"> AUC : 0.89 ROC Curve : 	<ul style="list-style-type: none"> AUC : 0.88 ROC Curve : 

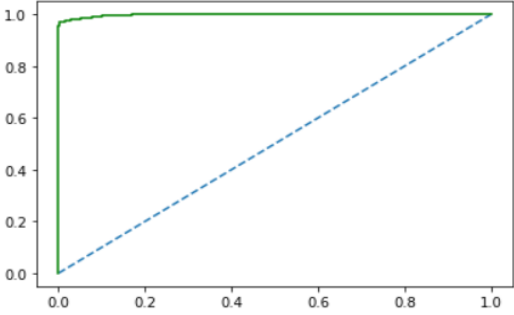
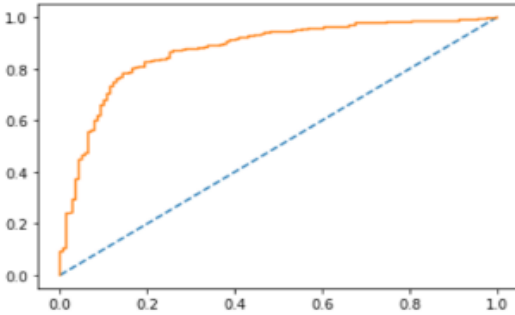
Naïve Bayes

Train Evaluation	Test Evaluation
<ol style="list-style-type: none"> 1. Accuracy : 0.84 2. Precision : 0.88 3. Recall : 0.90 4. F1 Score : 0.89 	<ol style="list-style-type: none"> 1. Accuracy : 0.82 2. Precision : 0.86 3. Recall : 0.88 4. F1 Score : 0.87
<ul style="list-style-type: none"> • AUC : 0.84 • ROC Curve : 	<ul style="list-style-type: none"> • AUC : 0.87 • ROC Curve : 

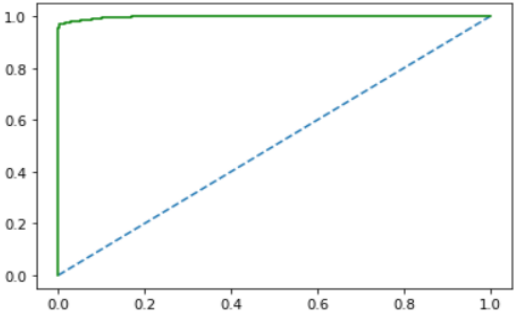
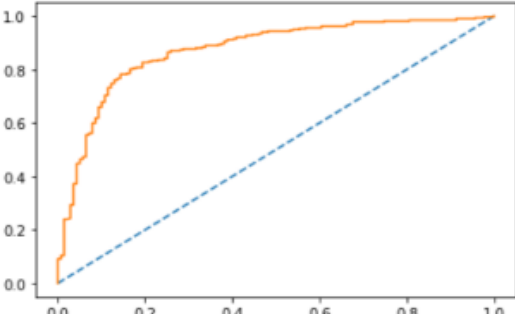
KNN

Train Evaluation	Test Evaluation
<ol style="list-style-type: none"> 1. Accuracy : 0.87 2. Precision : 0.89 3. Recall : 0.93 4. F1 Score : 0.91 	<ol style="list-style-type: none"> 1. Accuracy : 0.82 2. Precision : 0.86 3. Recall : 0.88 4. F1 Score : 0.87
<ul style="list-style-type: none"> • AUC : 0.93 • ROC Curve : 	<ul style="list-style-type: none"> • AUC : 0.84 • ROC Curve : 

Bagging

Train Evaluation	Test Evaluation
<ol style="list-style-type: none"> 1. Accuracy : 0.97 2. Precision : 0.96 3. Recall : 0.99 4. F1 Score : 0.98 	<ol style="list-style-type: none"> 1. Accuracy : 0.82 2. Precision : 0.84 3. Recall : 0.90 4. F1 Score : 0.87
<ul style="list-style-type: none"> • AUC : 0.93 • ROC Curve : 	<ul style="list-style-type: none"> • AUC : 0.87 • ROC Curve : 

Boosting

Train Evaluation	Test Evaluation
<ol style="list-style-type: none"> 1. Accuracy : 1.00 2. Precision : 1.00 3. Recall : 1.00 4. F1 Score : 1.00 	<ol style="list-style-type: none"> 1. Accuracy : 0.81 2. Precision : 0.84 3. Recall : 0.90 4. F1 Score : 0.87
<ul style="list-style-type: none"> • AUC : 0.93 • ROC Curve : 	<ul style="list-style-type: none"> • AUC : 0.87 • ROC Curve : 

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. Euroscepticism, 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 behaviour 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

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

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 Roosevelt's 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.

```
print('Number of Character in Roosevelt Speech :', len(inaugural.raw('1941-Roosevelt.txt')))  
print('Number of Character in Kennedy Speech :', len(inaugural.raw('1961-Kennedy.txt')))  
print('Number of Character in Kennedy Speech :', len(inaugural.raw('1973-Nixon.txt')))
```

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

- **Words() :**

```
print('Roosevelts word count = ', len(inaugural.words('1941-Roosevelt.txt')))  
print('Kennedy word count = ', len(inaugural.words('1961-Kennedy.txt')))  
print('Nixon word count = ', len(inaugural.words('1973-Nixon.txt')))
```

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

- **Sents() :**

```
print('Number of Sentences in Roosevelt Speech :', len(inaugural.sents('1941-Roosevelt.txt')))  
print('Number of Sentences in Kennedy Speech :', len(inaugural.sents('1961-Kennedy.txt')))  
print('Number of Sentences in Kennedy Speech :', len(inaugural.sents('1973-Nixon.txt')))
```

```
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 stopwords 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 stopwords 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 stopwords 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

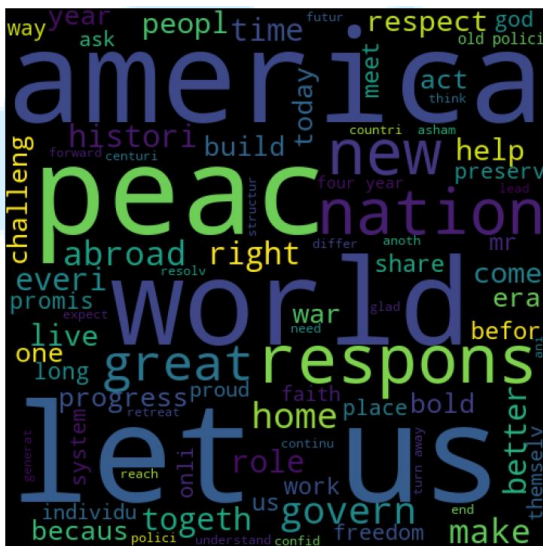
2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)

Roosevelt	Kennedy	Nixon
[('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)]	[('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)]	[('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)]

[illegible]

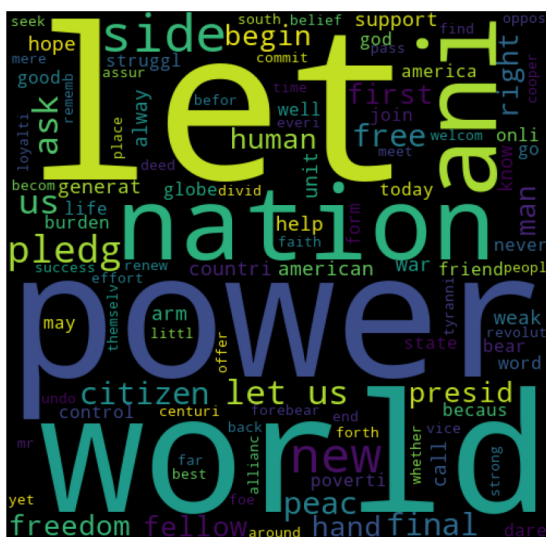
Talked about

- Nation
- People
- Spirit
- Life
- Democracy
- American



Talked about

- let
- nation
- power
- world
- pledge
- peace



Talked about

- America
- Peace
- World
- Let us
- response
- new nation