Problem 1

Problem Statement:

You are hired by one of the leading news channel 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

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

Reading the data

The data is read, and the **top five records** of the dataset are observed to get an overview.

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

Descriptive Statistics

The descriptive statistics of the numerical variables of the dataset is first analysed.

	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

Then, the categorical variables are analysed.

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

Below are the inferences:

- Most of the voters (75% of the observations) appear to be elderly, as the average age is 67 years, and highest is 93 years.
- Most of the voters have a score of 4 for the variables - economic.cond.national, economic.cond.household, Blair and Hague.
- A majority of the voters (75%) have a high score of 10 for variable 'Europe' i.e. they are highly Eurosceptic.
- Most of the voters have voted for Labour party
- A large number of the voters is female.

Null value check

There are no null values

1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts)

Exploratory Data Analysis

The column properties are observed

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1525 entries, 1 to 1525
Data columns (total 9 columns):
 #   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: 119.1+ KB
```

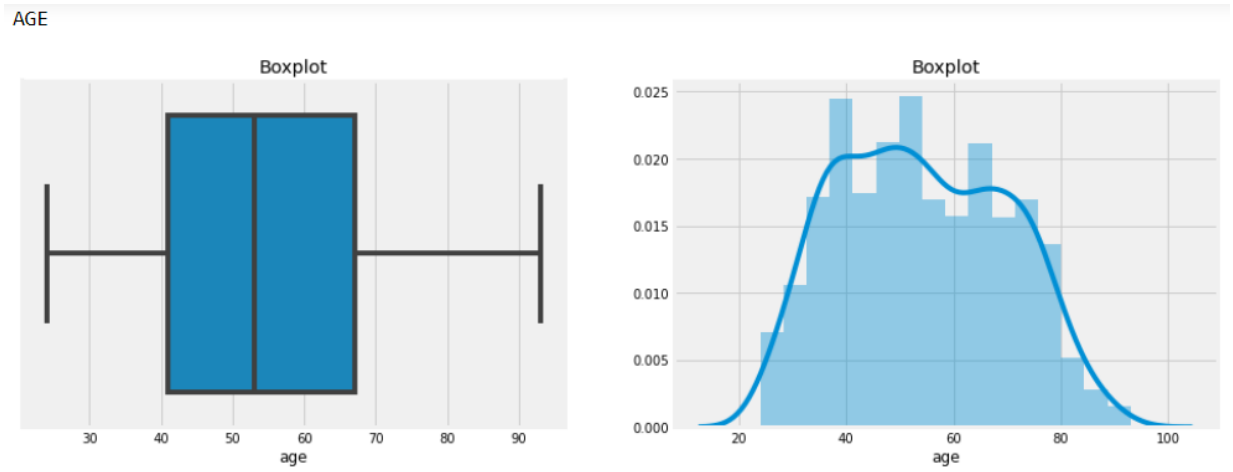
- There are 1525 observations, 9 columns and no null values.
- There are 8 duplicate observations in the dataset which are removed.

UNIVARIATE ANALYSIS

Univariate Analysis of Numerical variables:

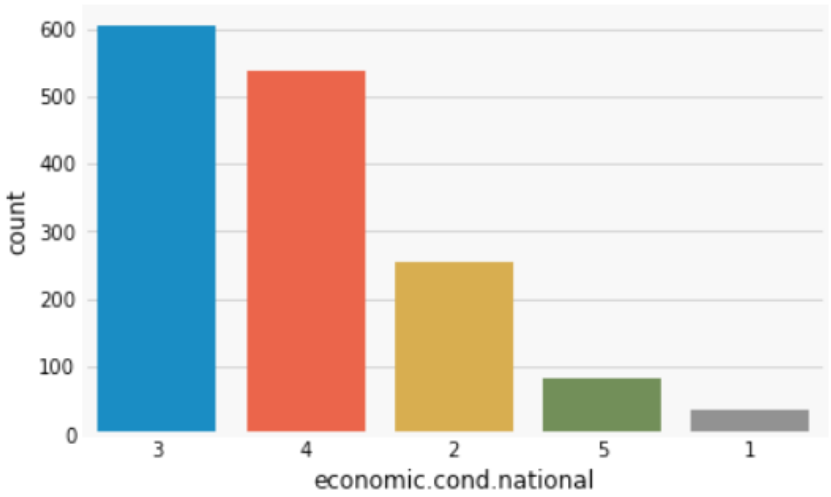
AGE:

- The boxplot shows that there are no outliers
- The distribution plot shows that the data distribution is somewhat normal.



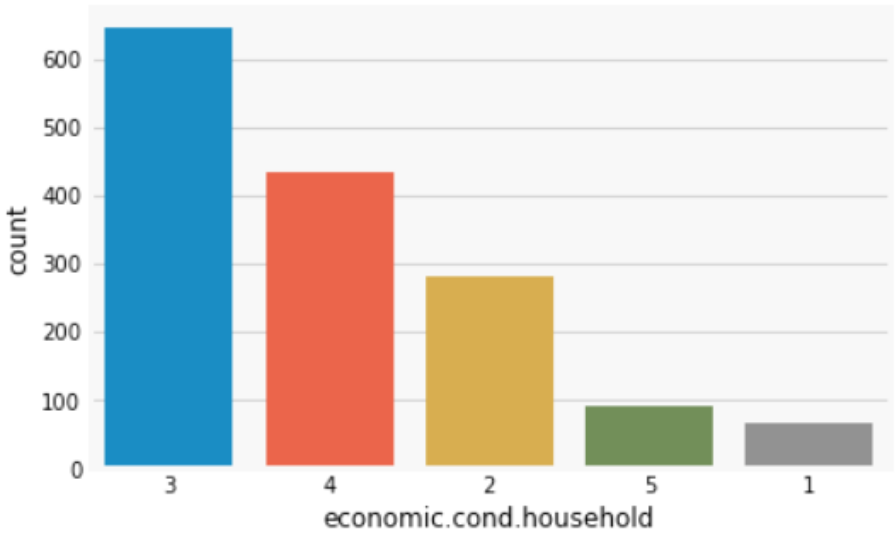
ECONOMIC.COND.NATIONAL:

The countplot shows that there are 604 observations that have a score of 3 on the national economic conditions assessment, and 538 observations have a score of 4.



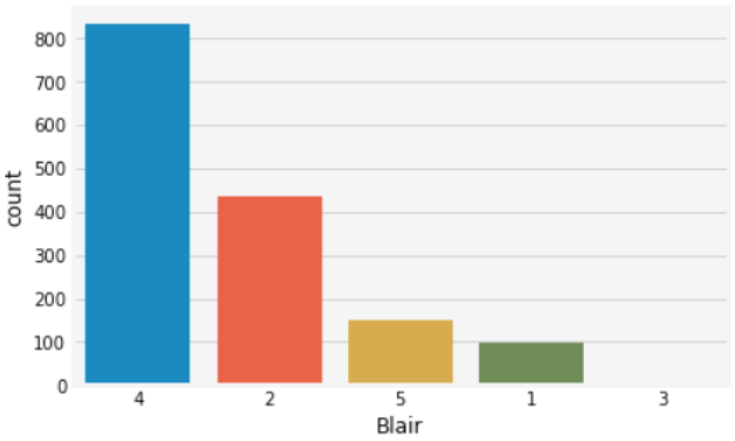
ECONOMIC.COND.HOUSEHOLD:

The countplot shows that 645 observations have a score of 3 on the household economic conditions assessment. The number of those who score above 3 are less.



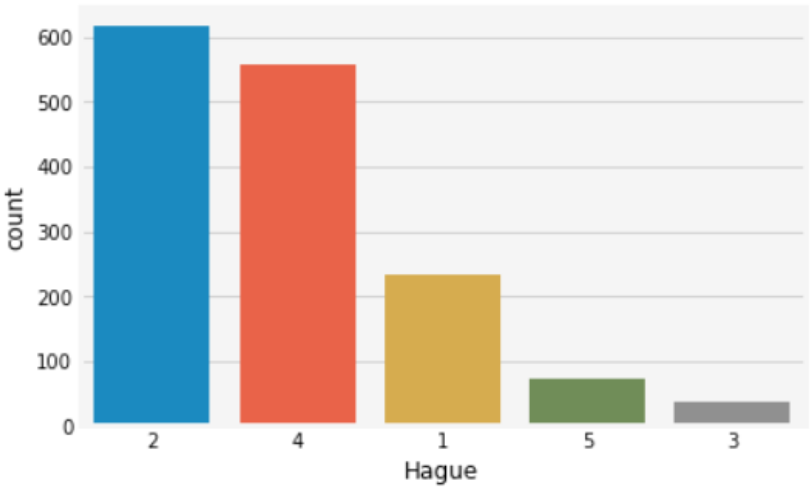
BLAIR:

A majority of the observations (833) have a score of 4 for the assessment of labour leader.



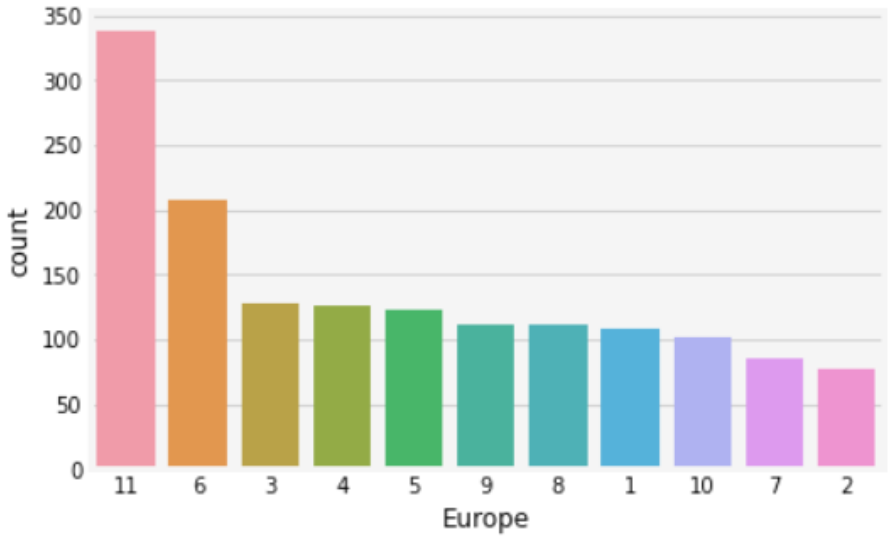
HAGUE:

A majority of the observations (617) have a score of 4 for the assessment of conservative leader.



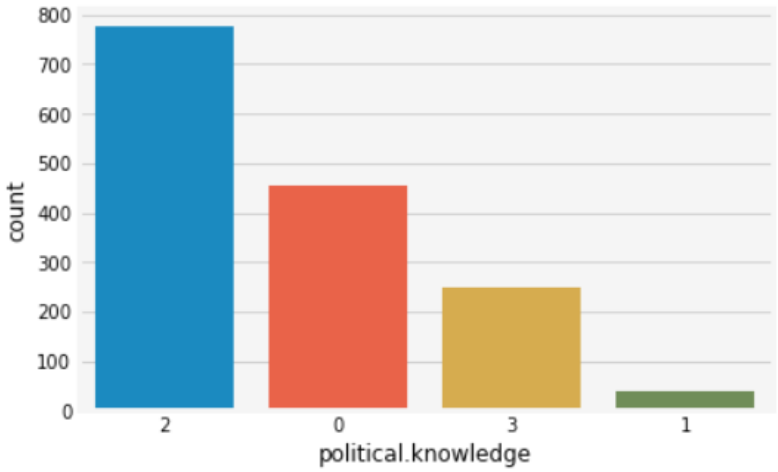
EUROPE:

Around 338 voters have a high score of 11, indicating that they're highly 'Eurosceptic'



POLITICAL.KNOWLEDGE:

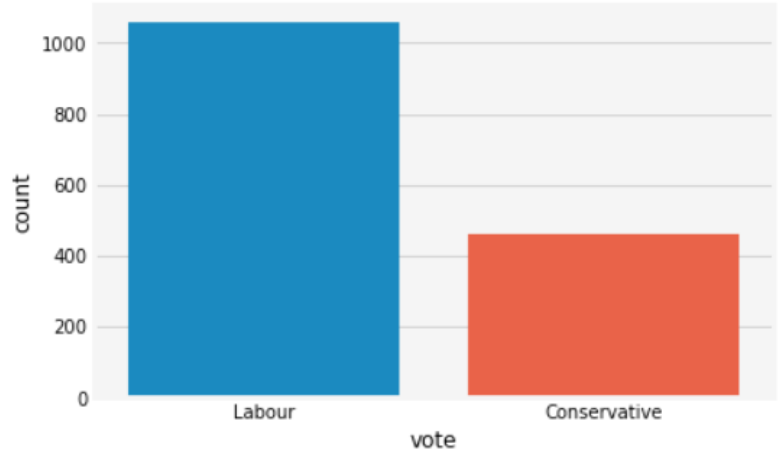
A large number of the observations have a score of 2 regarding the knowledge of parties' positions on European integration. Very few (249) have knowledge of 3.



Univariate Analysis of Categorical variables:

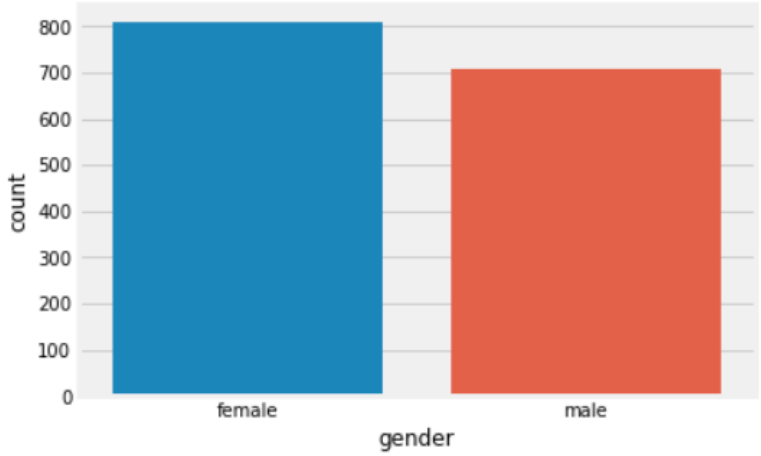
VOTE:

A majority of the voters (1057) have voted for the Labour party



GENDER:

The number of male and female voters appears to be almost equal.

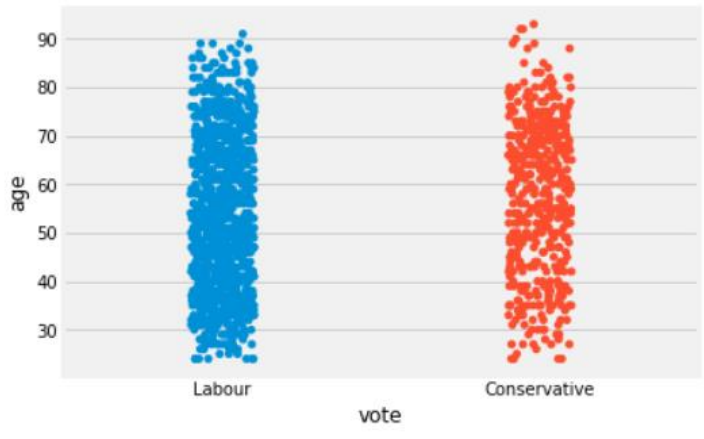


BIVARIATE ANALYSIS

VOTE vs. AGE:

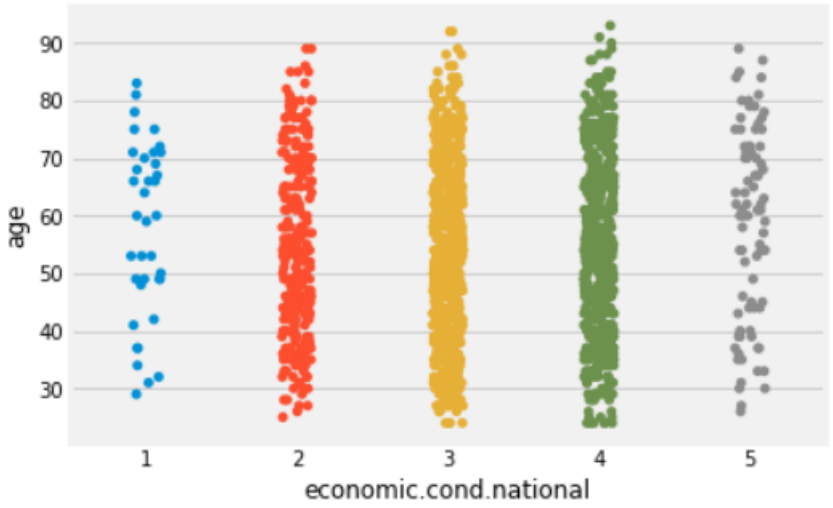
The strip plot shows that there isn't much of a correlation between vote and age variables.

VOTE vs. AGE



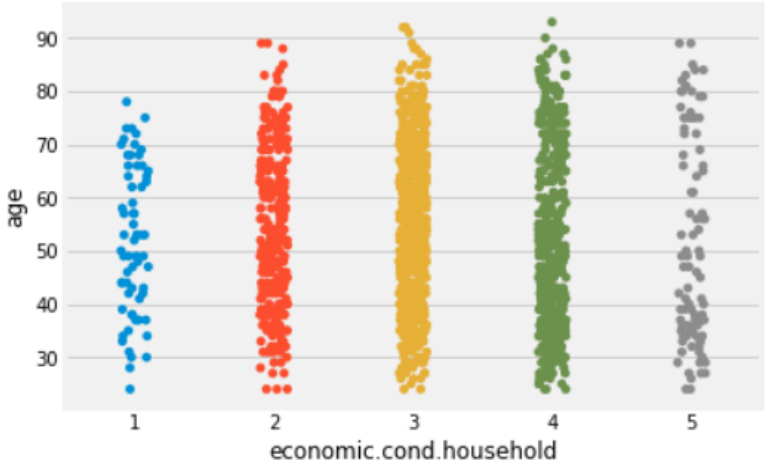
ECONOMIC.COND.NATIONAL vs. AGE:

- Very few observations have scores 1 and 5. There are even people over 60 years, who have a score of 1.
- Most people in the age range of 30 to 60 years have a score of 2
- A high concentration of observations in the age range of 30 to 80 years have a score of 3



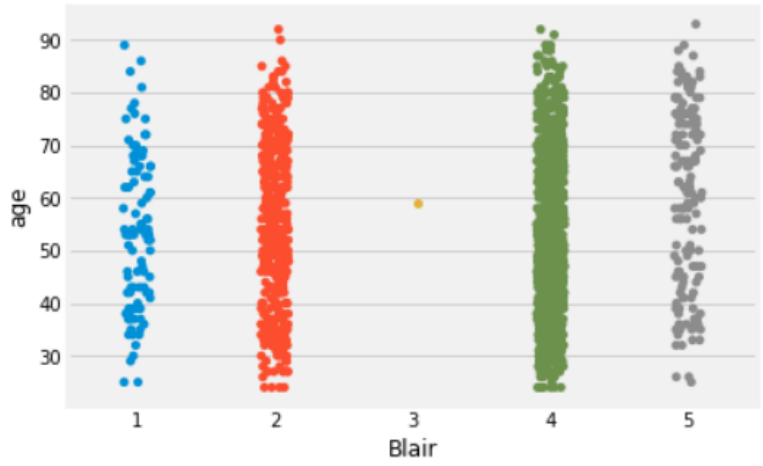
ECONOMIC.COND.HOUSEHOLD vs. AGE:

The household economic conditions appear to be pretty much the same as for the national economic conditions.



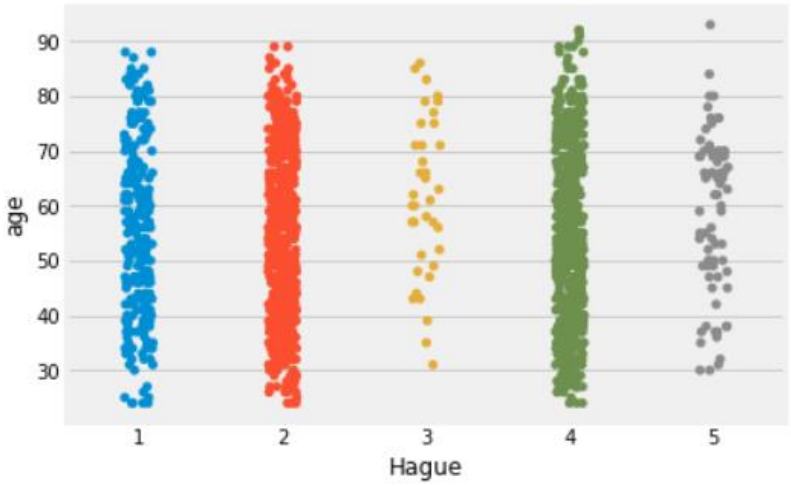
BLAIR vs. AGE:

- The assessment of labour leaders has scores of 2 and 4 for most of the observations
- There appears to be one observations at the age of 60 years that has a score of 3



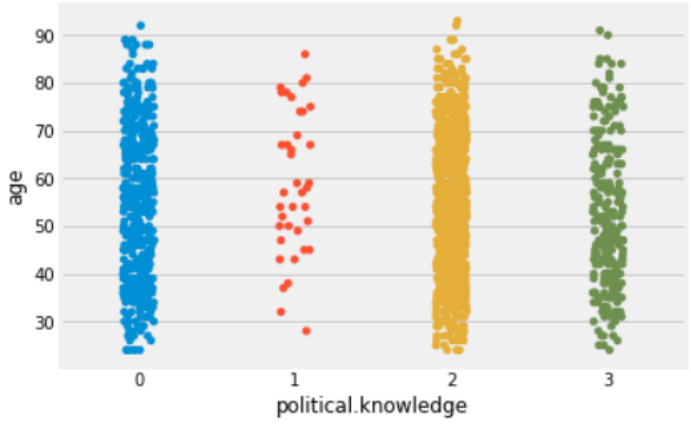
HAGUE vs. AGE:

A high concentration of voters between the age range of 30 to 80 years have scores of 2 and 4 for assessment of conservative leader



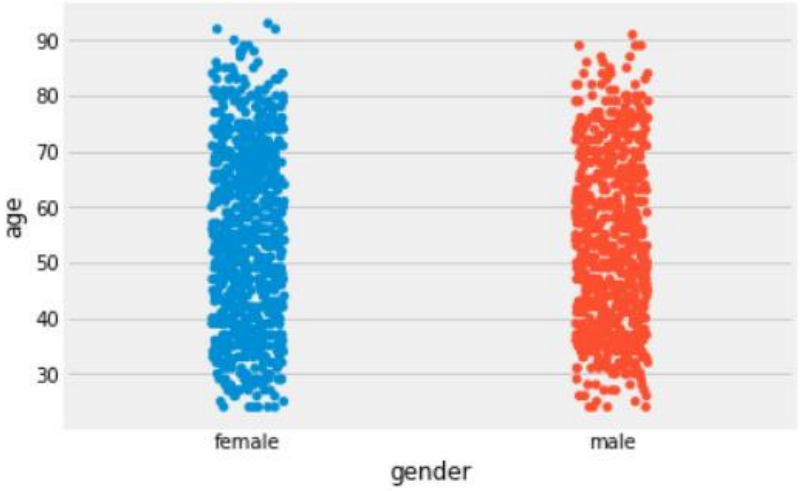
POLITICAL.KNOWLEDGE vs. AGE:

Very few people at all ages have little to no knowledge of parties' positions on European integration.



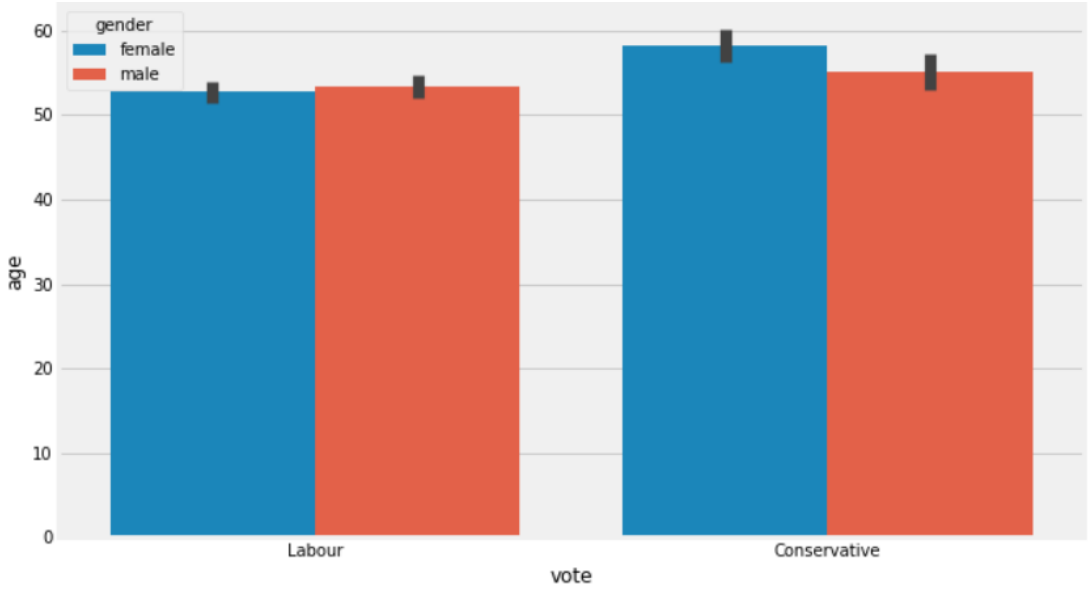
GENDER vs. AGE:

There doesn't appear to be any correlation.



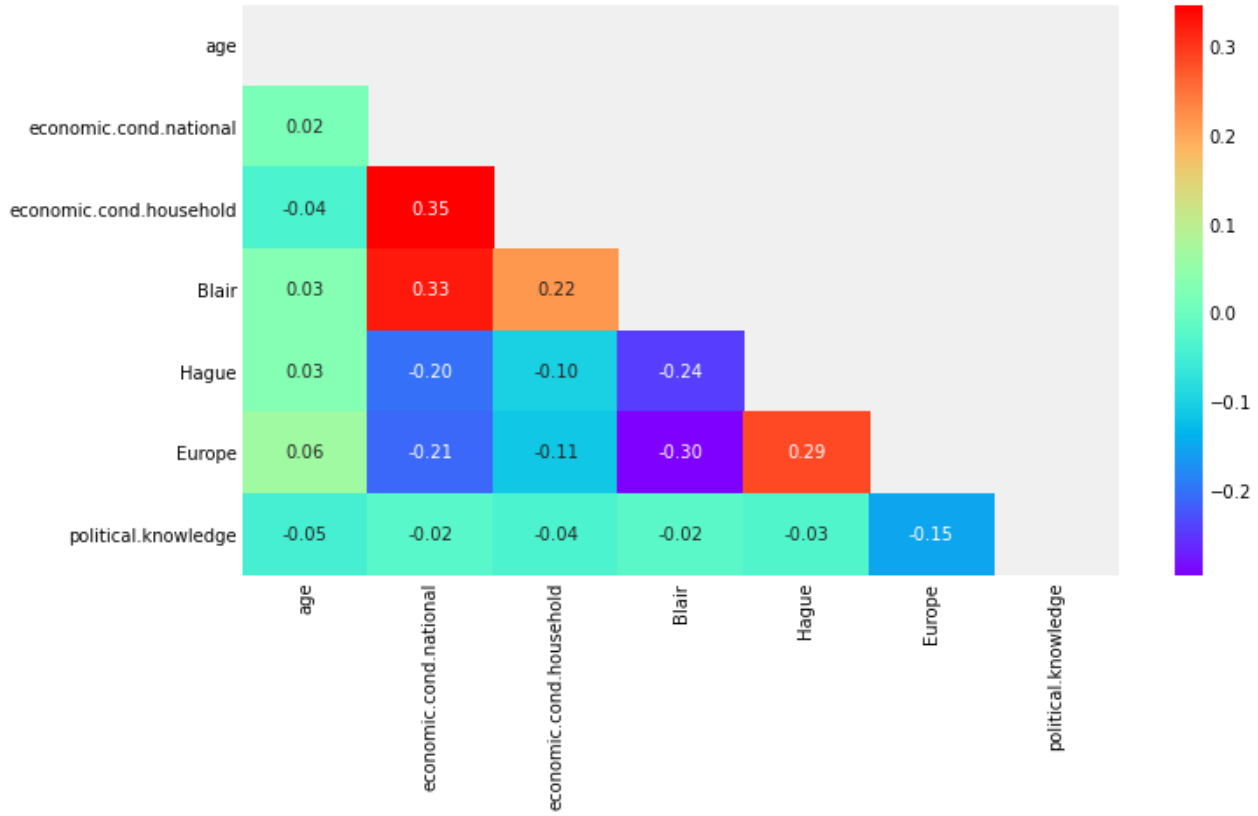
Barplot of vote vs. age in terms of gender:

There is a slightly higher count of females above 55 years who have voted for Conservative party.



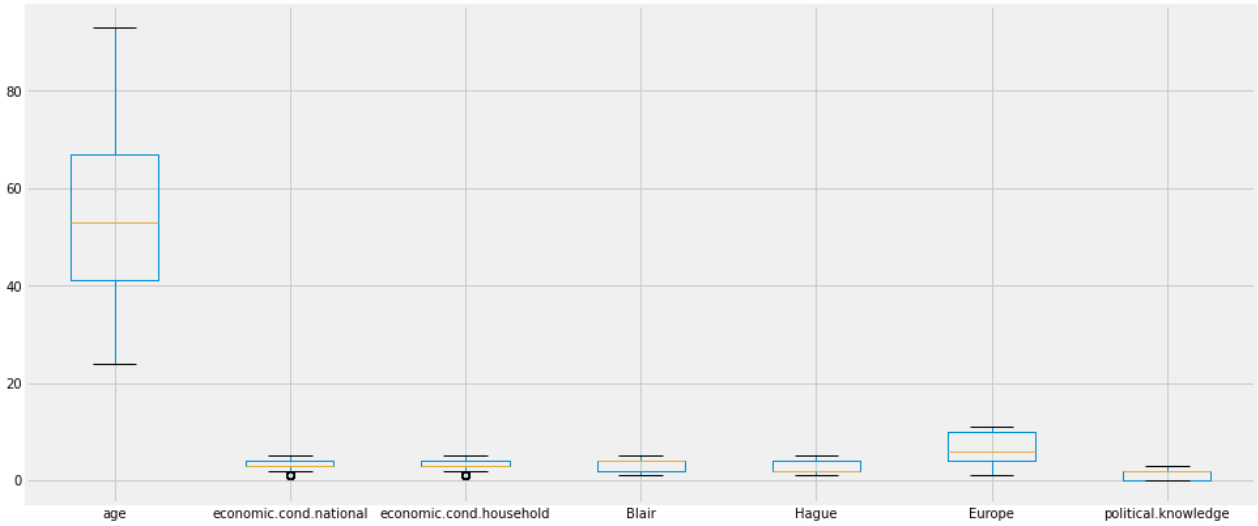
Multivariate analysis

The below heatmap (using Pearson's correlation) displays no significant correlations between the variaibles.



Outlier Check:

The below boxplot shows that there are no outliers for continuous variable 'age'. The remaining variables are related to assessment scores.



1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(3 pts), Data Split: Split the data into train and test (70:30) (2 pts).

Encoding object variables:

The variables having string values (i.e. vote and gender) are encoded into integer values as machine learning models will not accept string data.

- The pd.get_dummies() is used for 'gender'.
- The 'vote' variable is encoded using LabelEncoder (1 represents 'Labour' and 0 represents 'Conservative').

Is Scaling necessary:

Model	Reason	Yes/No
Naïve Bayes	The main condition is the features need to be independent of each other, so scale doesn't matter.	No
K Nearest Neighbours (KNN)	Distance measures like Euclidean distance, Manhattan distance...etc. are involved to calculate the distances from a given data point.	Yes
Logistic Regression	It is also a distance-based algorithm and uses gradient descent	Yes
Linear Discriminant Analysis (LDA)	It uses a linear combination of predictors, where the coefficients are determined by finding the variance between the classes. So, scaling will have no influence.	No
Ensemble Techniques (Bagging, Boosting, Random Forest)	They are decision-tree based models, i.e. they are not distance-based, so they can handle data of different value ranges.	No

Data split:

The data having all the features (i.e. except 'vote') is taken as 'X' and the 'vote' variable is taken as 'y'.

The 'X' and 'y' dataframes are split into train and test sets. Around 70% of the data is taken for training and 30% is taken for testing.

Scaling is applied to the train and test dependent data (on continuous variable 'age' since the rest represent scores or binary values) using 'zscore'.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_male
388	0.557746	2	2	4	4	5	3	1
237	-1.475167	3	3	2	3	6	0	0
124	0.176575	4	4	4	2	11	2	0
97	-0.839882	3	3	4	4	5	2	0
212	-0.585768	3	3	2	2	2	1	0

1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (3 pts). Interpret the inferences of both models (2 pts)

The accuracy of the Logistic Regression and LDA models are checked first the default parameters. The hyperparameter tuning will be done at a later stage.

Logistic Regression:

It is a supervised machine learning model used for classification of the output variable. The logistic regression classifier is initialized and then fit to the training data, using only the default parameters.

Training Accuracy	**84.07 %**
Testing Accuracy	**84.43 %**

Linear Discriminant Analysis (LDA):

It is a supervised machine learning classification model, where linear combination of the predictors is used for determining the class of the target. The LDA classifier is initialized and then fit to the training data, using only the default parameters.

Applying the LDA model to both the scaled and un-scaled data yielded the same results

Training Accuracy	**83.79 %**
Testing Accuracy	**84.87 %**

Inferences:

- Both Linear Regression and LDA models appear to do well on the training data.
- LDA performs better on the testing data.
- Scaling is not needed for LDA.
- Out of both the models, LDA is the better one.

1.5) Apply KNN Model and Naïve Bayes Model (5 pts). Interpret the inferences of each model (2 pts)

The accuracy of the KNN and Naïve Bayes models are checked first the default parameters. The hyperparameter tuning will be done at a later stage.

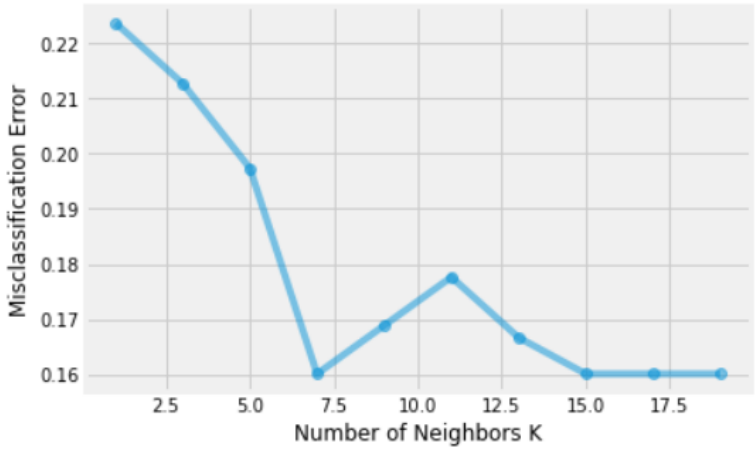
K-Nearest Neighbors (KNN):

It is a supervised machine learning model used for classification of the output variable. It classifies the given data point by analysing the data points that are in close proximity. The KNN classifier is initialized and then fit to the training data, using only the default parameters.

Training Accuracy	**86.80 %**
Testing Accuracy	**84.65 %**

The KNN model usually works best if 'k' is given an odd value, which will ensure that there are no ties in the voting mechanism. On experimenting with all odd values of 'k' from 1 to 19, it appears the 'k' = 7, 15, 17 and 19 have the lowest misclassification errors

Misclassification plot for all values of k:



K value	Training Accuracy	Testing Accuracy
7	85.77	83.99
15	84.45	84.21

Naïve Bayes:

It is a supervised machine learning model, which applies Bayes' Conditional Probability in order to classify the target. Naïve Bayes model assumes that there is independence among the predictor variables.

Training Accuracy	**83.13 %**
Testing Accuracy	**82.89 %**

Inferences:

- The KNN model appears to do well with the default parameters, since it has a good testing accuracy of 84.65%.
- The KNN models for k = 7 and 15 also seem to do reasonably well on both the training and testing data.
- The Naïve Bayes model has a slightly lower accuracy on both train and test sets
- Out of both the models, KNN is the better one.

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

In the model tuning i.e. hyperparameter-tuning process, the parameters in the classifiers are given different values in order to select the best combination of hyperparamters for each of the mentioned machine learning model.

The ensemble techniques – Bagging, Boosting and Random Forest classifiers are also applied.

The RandomizedSearchCV is applied instead of GridSearchCV since it takes less time (due to randomly selecting a set of hyperparameter values) and also finds the best tuned parameters most often than not.

Thus, the different sets of parameters are fit to the training data in the RandomizedSearchCV with a cross-fold of 10, and below are the results of fitting to the train data.

Models	Best Parameters	Best Score
Random_Forest	{'n_estimators': 1200, 'min_samples_split': 40, 'min_samples_leaf': 5, 'max_depth': 5, 'criterion': 'entropy', 'bootstrap': False}	0.835046729
Logistic_Regression	{'tol': 1e-17, 'solver': 'liblinear', 'penalty': 'l2', 'n_jobs': 6, 'max_iter': 10000}	0.835029095
AdaBoosting	{'n_estimators': 100, 'learning_rate': 0.1}	0.83221654
LDA	{'tol': 1e-17, 'solver': 'lsqr'}	0.830320931
Naive_Bayes	{'var_smoothing': 1e-10, 'priors': [0.25, 0.75]}	0.827481926
KNN	{'n_neighbors': 19, 'metric': 'minkowski'}	0.822782578
Bagging	{'n_jobs': 4, 'n_estimators': 200, 'bootstrap': True}	0.81805678

Inferences:

- It appears that all the models do reasonably well on the train set.
- Random Forest is the best model followed by Logistic Regression.

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model (4 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized (3 pts)

PERFORMANCE ON TRAIN SET:

1. LOGISTIC REGRESSION

1.1 The Classification Report displays the below details:

- The **Precision** metric tells us how many of the classified data points are true positives. For party 'Labour' (represented by 1), precision is 87%, and for party 'Conservative', it is 76%.

- The **Recall** metric tells us how many of the classified data points are true positives out of the true positives and false negatives. For party 'Labour', Recall is 91%, however, it is low for party 'Conservative' – 66%.

- The **Accuracy Score** is 84.07 %. This is also shown in the below classification report for the field "accuracy".
- The **F1 Score** for the party 'Labour' – which takes the harmonic mean of Precision and Recall (i.e. it checks for Type 1 and Type 2 errors) – is good (89%). It is also reasonably well for 'Conservative' – 71%.

LOGISTIC_REGRESSION

Classification Report

Accuracy: 84.07 %

	precision	recall	f1-score	support
0	0.76	0.66	0.71	311
1	0.87	0.91	0.89	750
accuracy			0.84	1061
macro avg	0.82	0.79	0.80	1061
weighted avg	0.84	0.84	0.84	1061

1.2 Confusion Matrix

It displays the number of correct and incorrect predictions made in the model in a tabular format. Below is the data understood from the confusion matrix.

TN (True Negatives)	= 206	TP (True Positives)	= 686
FN (False Negatives)	= 64	FP (False Positives)	= 105

The Accuracy can also be calculated from the Confusion Matrix.

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

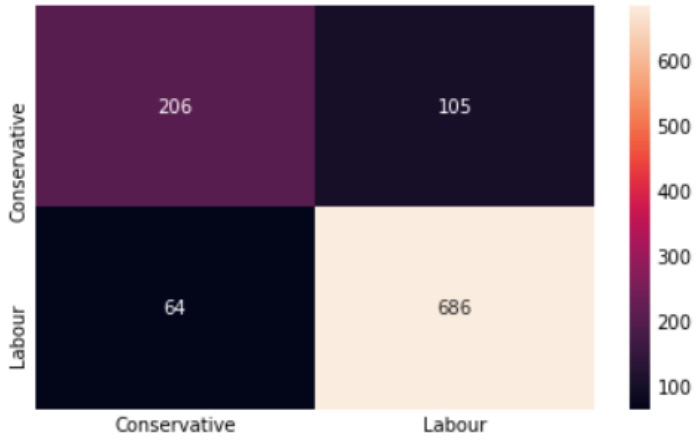
= (686 + 206) / (686 + 206 + 105 + 64)

= 0.8407163

i.e. **84.07 %**

There are large number of True Negatives and True Positives which is good.

Confusion Matrix



2. LINEAR DISCRIMINANT ANALYSIS:

2.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 87% and 75% respectively.
- The **Recall** for party 'Labour' is good – 91%, however, it is low for party 'Conservative' – 68%.
- The **Accuracy Score** is 83.79 %.
- The **F1 Score** for the party 'Labour' –is good (89%). It is also somewhat well for 'Conservative' – 71%.

LDA

Classification Report

Accuracy: 83.79 %

	precision	recall	f1-score	support
0	0.75	0.68	0.71	311
1	0.87	0.91	0.89	750
accuracy			0.84	1061
macro avg	0.81	0.79	0.80	1061
weighted avg	0.83	0.84	0.84	1061

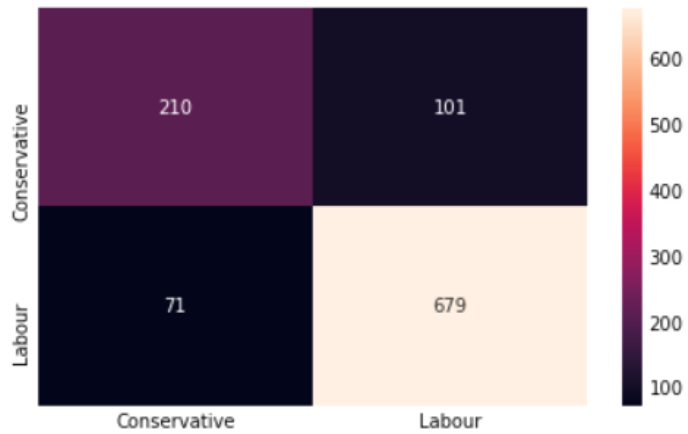
2.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 210	TP (True Positives)	= 679
FN (False Negatives)	= 71	FP (False Positives)	= 101

There are large number of True Negatives and True Positives which is good.

Confusion Matrix



3. NAÏVE BAYES:

3.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 87% and 75% respectively.
- The **Recall** for party 'Labour' is good – 91%, however, it is low for party 'Conservative' – 67%.
- The **Accuracy Score** is 83.79 %.
- The **F1 Score** for the party 'Labour' –is good (89%). It is also somewhat well for 'Conservative' – 71%.

```
NAIVE_BAYES

Classification Report

Accuracy:        83.79 %

              precision    recall  f1-score   support

           0       0.75      0.67      0.71       311
           1       0.87      0.91      0.89       750

    accuracy                           0.84      1061
   macro avg       0.81      0.79      0.80      1061
weighted avg       0.83      0.84      0.83      1061
```

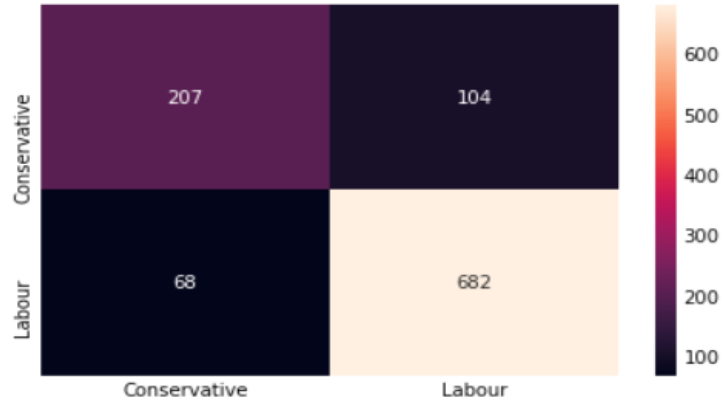
3.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 207	TP (True Positives)	= 682
FN (False Negatives)	= 68	FP (False Positives)	= 104

The number of True Negatives and True Positives are large, especially the True Positives (682).

Confusion Matrix



4. K-NEAREST NEIGHBORS:

4.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 87% and 82% respectively.
- The **Recall** for party 'Labour' is good – 94%, however, it is low for party 'Conservative' – 66%.
- The **Accuracy Score** is 85.86 %.
- The **F1 Score** for the party 'Labour' –is good (90%). It is also somewhat well for 'Conservative' – 73%.

KNN

Classification Report

Accuracy: 85.86 %

	precision	recall	f1-score	support
0	0.82	0.66	0.73	311
1	0.87	0.94	0.90	750
accuracy			0.86	1061
macro avg	0.85	0.80	0.82	1061
weighted avg	0.86	0.86	0.85	1061

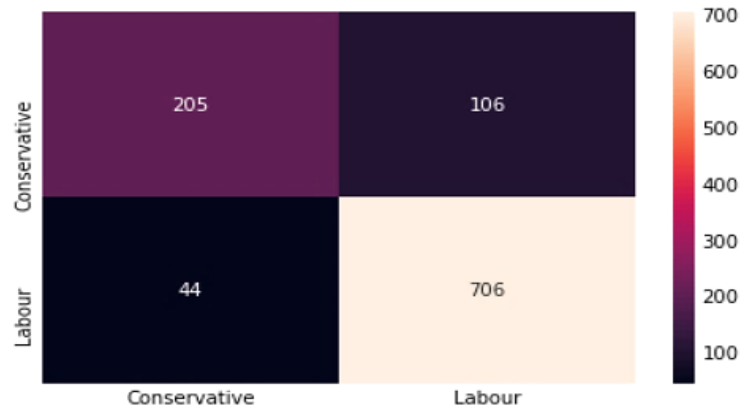
4.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 205	TP (True Positives)	= 706
FN (False Negatives)	= 44	FP (False Positives)	= 106

The number of falsely classified data points is less, especially the False Negatives.

Confusion Matrix



5. RANDOM FOREST:

5.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 86% and 73% respectively.
- The **Recall** for party 'Labour' is good – 90%, however, it is low for party 'Conservative' – 66%.
- The **Accuracy Score** is 83.03 %.
- The **F1 Score** for the party 'Labour' –is good (88%). It is also somewhat well for 'Conservative' – 69%.

```
RANDOM_FOREST

Classification Report

Accuracy:        83.03 %

              precision    recall  f1-score   support

           0       0.73      0.66      0.69       311
           1       0.86      0.90      0.88       750

    accuracy                           0.83      1061
   macro avg       0.80      0.78      0.79      1061
weighted avg       0.83      0.83      0.83      1061
```

5.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 205	TP (True Positives)	= 676
FN (False Negatives)	= 74	FP (False Positives)	= 106

There are large number of correctly classified data points (i.e. TP and TN).

Confusion Matrix

6. BAGGING:

6.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is 100%.
- The **Recall** for both parties is also 100%.

- The **Accuracy Score** is 100 %.
- The **F1 Score** for both parties is also 100%.

However, this indicates over-fitting and the model may not do well on the test data.

BAGGING

Classification Report

Accuracy: 100.0 %

	precision	recall	f1-score	support
0	1.00	1.00	1.00	311
1	1.00	1.00	1.00	750
accuracy			1.00	1061
macro avg	1.00	1.00	1.00	1061
weighted avg	1.00	1.00	1.00	1061

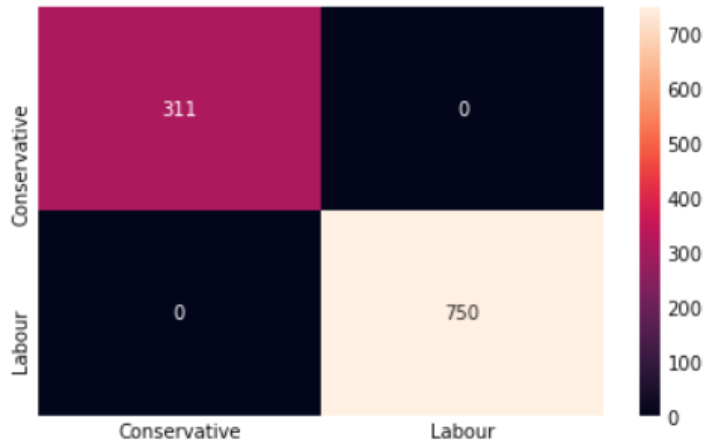
6.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 311	TP (True Positives)	= 350
FN (False Negatives)	**= 0**	**FP (False Positives)**	**= 0**

There are zero number of falsely classified data points which means that this model with have low bias and high variance (i.e. it overfits the training data)

Confusion Matrix



7. ADABOOSTING:

7.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 85% and 77% respectively.
- The **Recall** for party 'Labour' is good – 93%, however, it is low for party 'Conservative' – 61%.
- The **Accuracy Score** is 83.41 %.
- The **F1 Score** for the party 'Labour' –is good (89%). It is also somewhat well for 'Conservative' – 68%.

ADABOOSTING

Classification Report

Accuracy: 83.41 %

	precision	recall	f1-score	support
0	0.77	0.61	0.68	311
1	0.85	0.93	0.89	750
accuracy			0.83	1061
macro avg	0.81	0.77	0.79	1061
weighted avg	0.83	0.83	0.83	1061

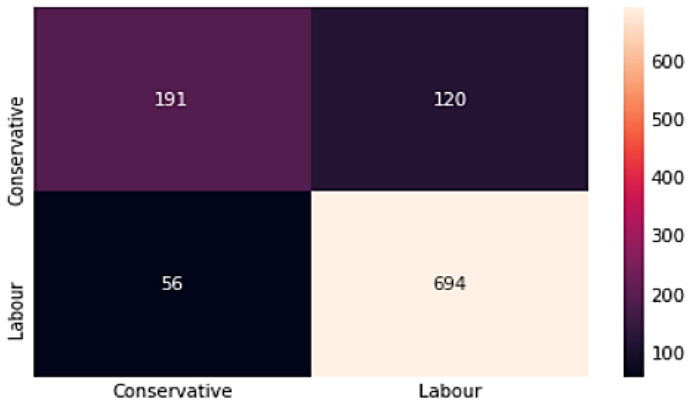
7.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 191	TP (True Positives)	= 694
FN (False Negatives)	= 56	FP (False Positives)	= 120

The number of True Negatives are less compared to the number of True Positives.

Confusion Matrix



ROC AUC Score for the train data set:

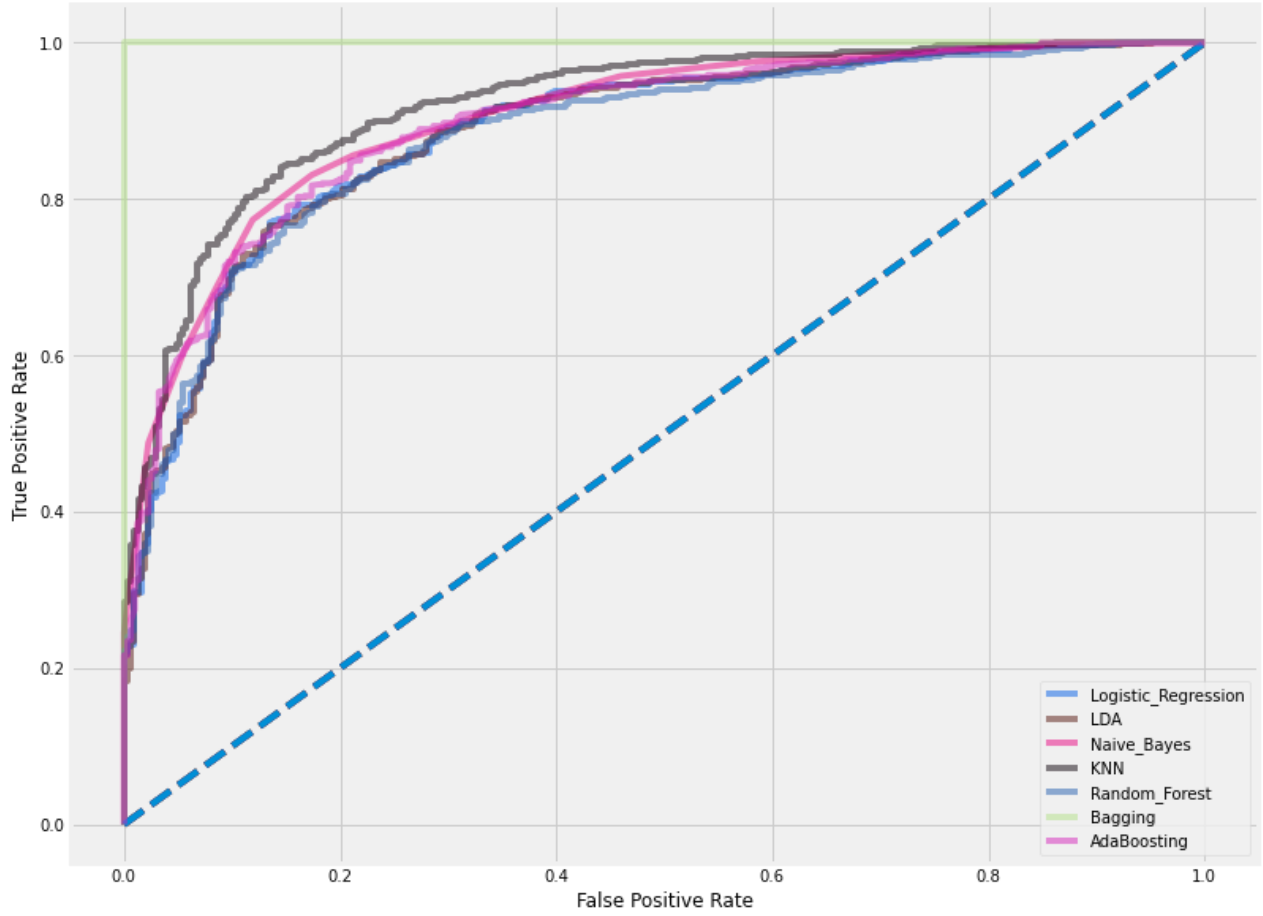
ROC stands for Receiver Operating Characteristic Curve and AUC stands for Area Under Curve.

True Positive Rate (TPR) is a synonym for Recall - TP / (TP + FN)

False Positive Rate (FPR) is - FP / (FP + TN)

The ROC curve plots the TPR (True Positive Rate) versus the FPR (False Positive Rate) at different classification threshold values. Generally, if the curve is very steep, then the model is good.

From the below graph, it appears that the KNN model has done a better job in training the data compared to the other models.



The ROC AUC scores for the models are as follows:

AUC Score Logistic_Regression	**0.8861779206859592**
AUC Score LDA	**0.8863751339764201**
AUC Score Naive_Bayes	**0.9035241157556271**
AUC Score KNN	**0.9199657020364416**
AUC Score Random_Forest	**0.8815777063236871**

AUC Score Bagging **1.0**

AUC Score AdaBoosting **0.8969560557341907**

PERFORMANCE ON TEST SET:

1. LOGISTIC REGRESSION

1.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 85% and 77% respectively.
- The **Recall** for party 'Labour' is good – 90%, however, it is low for party 'Conservative' – 68%.
- The **Accuracy Score** is 82.89 %.
- The **F1 Score** for the party 'Labour' –is good (88%). It is also somewhat good for 'Conservative' – 72%.

```
LOGISTIC_REGRESSION

Classification Report

Accuracy:        82.89 %

              precision    recall  f1-score   support

           0       0.77      0.68      0.72       149
           1       0.85      0.90      0.88       307

    accuracy                           0.83       456
   macro avg       0.81      0.79      0.80       456
weighted avg       0.83      0.83      0.83       456
```

1.2 Confusion Matrix

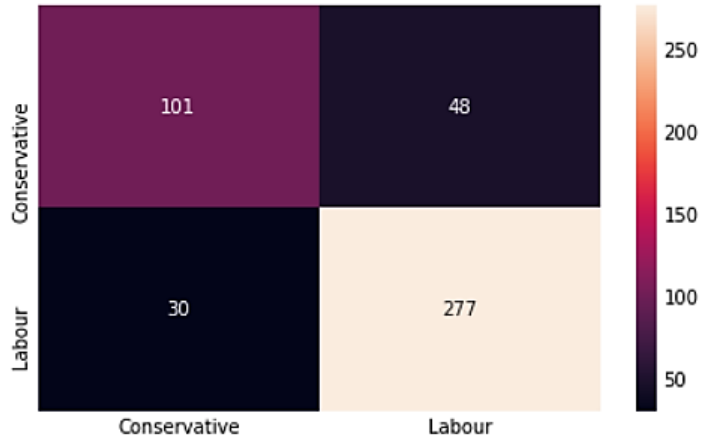
It displays the number of correct and incorrect predictions made in the model in a tabular format. Below is the data understood from the confusion matrix.

TN (True Negatives)	= 101	TP (True Positives)	= 277
FN (False Negatives)	= 30	FP (False Positives)	= 48

There appears to be a good number of data points classified as True Positives – 277.

The number of false positives and false negatives is less.

Confusion Matrix



2. LINEAR DISCRIMINANT ANALYSIS:

2.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 86% and 76% respectively.
- The **Recall** for party 'Labour' is good – 89%, and it is also reasonably well for party 'Conservative' – 70%.
- The **Accuracy Score** is 82.89 %.
- The **F1 Score** for the party 'Labour' –is good (88%). It is also somewhat well for 'Conservative' – 73%.

LDA

Classification Report

Accuracy: 82.89 %

	precision	recall	f1-score	support
0	0.76	0.70	0.73	149
1	0.86	0.89	0.88	307
accuracy			0.83	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.83	0.83	0.83	456

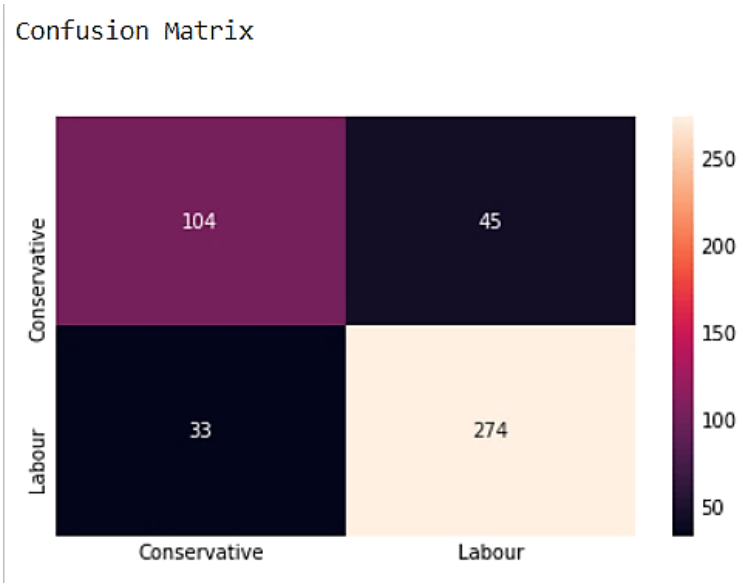
2.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives) = 104 TP (True Positives) = 274

FN (False Negatives) = 33 FP (False Positives) = 45

The number of True Positives and True Negatives is large which is good.



3. NAÏVE BAYES:

3.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 86% and 79% respectively.
- The **Recall** for party 'Labour' is good – 91%, however, it is low for party 'Conservative' – 69%.
- The **Accuracy Score** is 83.99 %.
- The **F1 Score** for the party 'Labour' –is good (88%). It is also somewhat well for 'Conservative' – 74%.

NAIVE_BAYES

Classification Report

Accuracy: 83.99 %

	precision	recall	f1-score	support
0	0.79	0.69	0.74	149
1	0.86	0.91	0.88	307
accuracy			0.84	456
macro avg	0.83	0.80	0.81	456
weighted avg	0.84	0.84	0.84	456

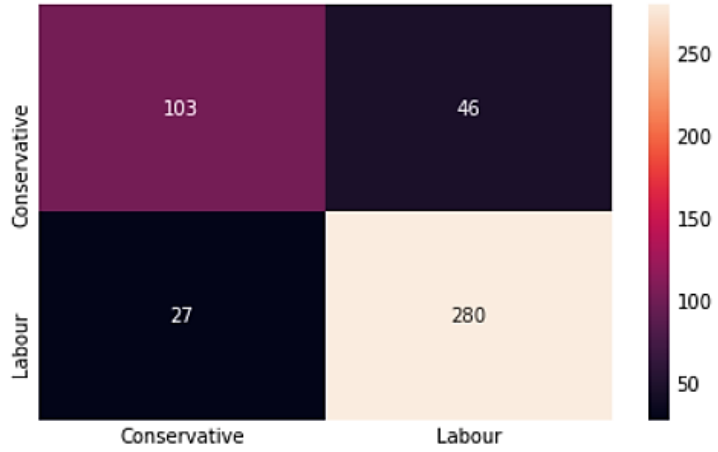
3.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 103	TP (True Positives)	= 280
FN (False Negatives)	= 27	FP (False Positives)	= 46

The number of False Positives and False Negatives is less.

Confusion Matrix



4. K-NEAREST NEIGHBORS:

4.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 84% and 80% respectively.
- The **Recall** for party 'Labour' is good – 92%, however, it is low for party 'Conservative' – 64%.
- The **Accuracy Score** is 82.89 %.
- The **F1 Score** for the party 'Labour' –is good (88%). It is also fine for 'Conservative' – 71%.

KNN

Classification Report

Accuracy: 82.89 %

	precision	recall	f1-score	support
0	0.80	0.64	0.71	149
1	0.84	0.92	0.88	307
accuracy			0.83	456
macro avg	0.82	0.78	0.79	456
weighted avg	0.83	0.83	0.82	456

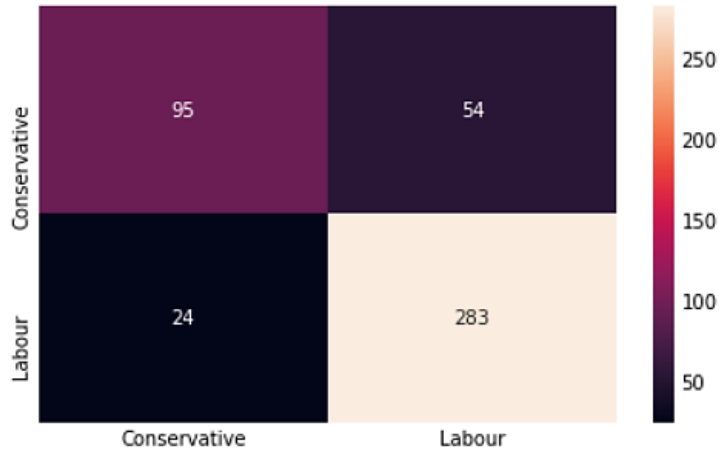
4.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 95	TP (True Positives)	= 283
FN (False Negatives)	= 24	FP (False Positives)	= 54

The number of False Positives and False Negatives is a less.

Confusion Matrix



5. RANDOM FOREST:

5.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 87% and 76% respectively.
- The **Recall** for party 'Labour' is good – 89%, and it is also fine for party 'Conservative' – 72%.
- The **Accuracy Score** is 83.33 %.
- The **F1 Score** for the party 'Labour' –is good (88%). It is also somewhat well for 'Conservative' – 74%.

RANDOM_FOREST

Classification Report

Accuracy: 83.33 %

	precision	recall	f1-score	support
0	0.76	0.72	0.74	149
1	0.87	0.89	0.88	307
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

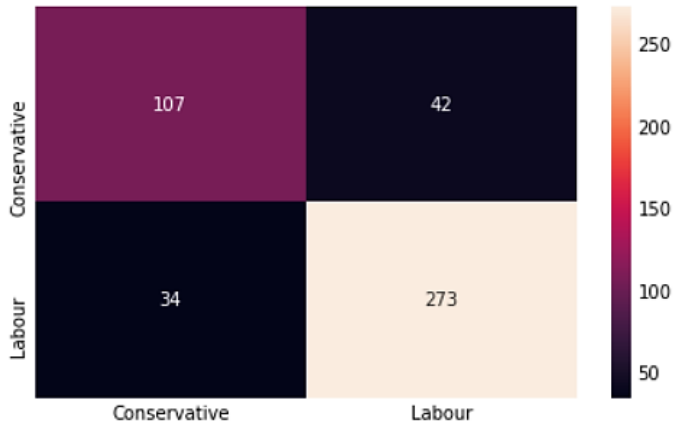
5.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 107	TP (True Positives)	= 273
FN (False Negatives)	= 34	FP (False Positives)	= 42

The number of False Positives and False Negatives is less, which is good as large number of data points are classified correctly.

Confusion Matrix



6. BAGGING:

6.1 The Classification Report displays the below details:

- The **Precision** for parties 'Labour' (represented by 1) and 'Conservative' (0) is 83% and 71% respectively.
- The **Recall** for party 'Labour' is good – 88%, however, it is a bit less for party 'Conservative' – 63%.

- The **Accuracy Score** is 79.61 %.
- The **F1 Score** for the party 'Labour' –is good (85%). It is also less for 'Conservative' – 67%.

However, this indicates over-fitting and the model may not do well on the test data.

BAGGING

Classification Report

Accuracy: 79.61 %

	precision	recall	f1-score	support
0	0.71	0.63	0.67	149
1	0.83	0.88	0.85	307
accuracy			0.80	456
macro avg	0.77	0.75	0.76	456
weighted avg	0.79	0.80	0.79	456

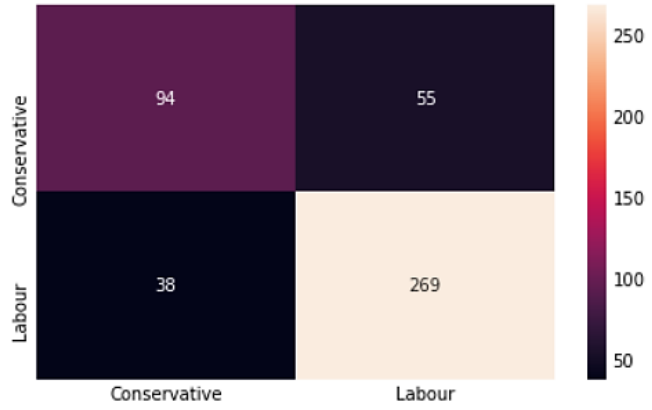
6.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 94	TP (True Positives)	= 269
FN (False Negatives)	= 38	FP (False Positives)	= 55

The number of True Positives is higher than True Negatives.

Confusion Matrix



7. ADABOOSTING:

7.1 The Classification Report displays the below details:

- The **Precision** for both parties 'Labour' (represented by 1) and 'Conservative' (0) is good – 84% and 79% respectively.
- The **Recall** for party 'Labour' is good – 92%, however, it is a bit less for party 'Conservative' – 64%.
- The **Accuracy Score** is 82.89 %.
- The **F1 Score** for the party 'Labour' –is good (88%). It is also somewhat well for 'Conservative' – 71%.

ADABOOSTING

Classification Report

Accuracy: 82.89 %

	precision	recall	f1-score	support
0	0.79	0.64	0.71	149
1	0.84	0.92	0.88	307
accuracy			0.83	456
macro avg	0.82	0.78	0.79	456
weighted avg	0.83	0.83	0.82	456

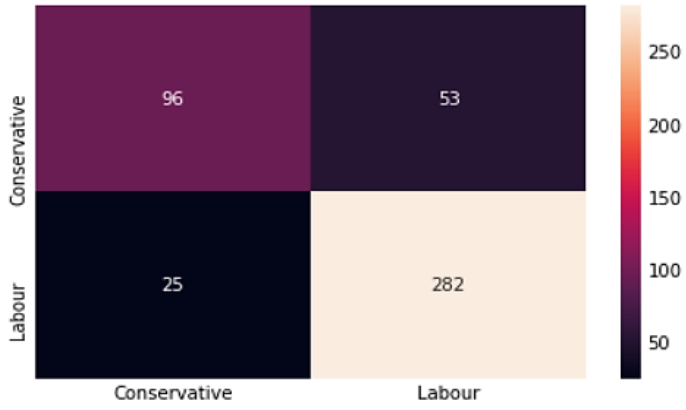
7.2 Confusion Matrix

Below is the data understood from the confusion matrix.

TN (True Negatives)	= 96	TP (True Positives)	= 282
FN (False Negatives)	= 25	FP (False Positives)	= 53

The number of False Positives and False Negatives is less, which is good.

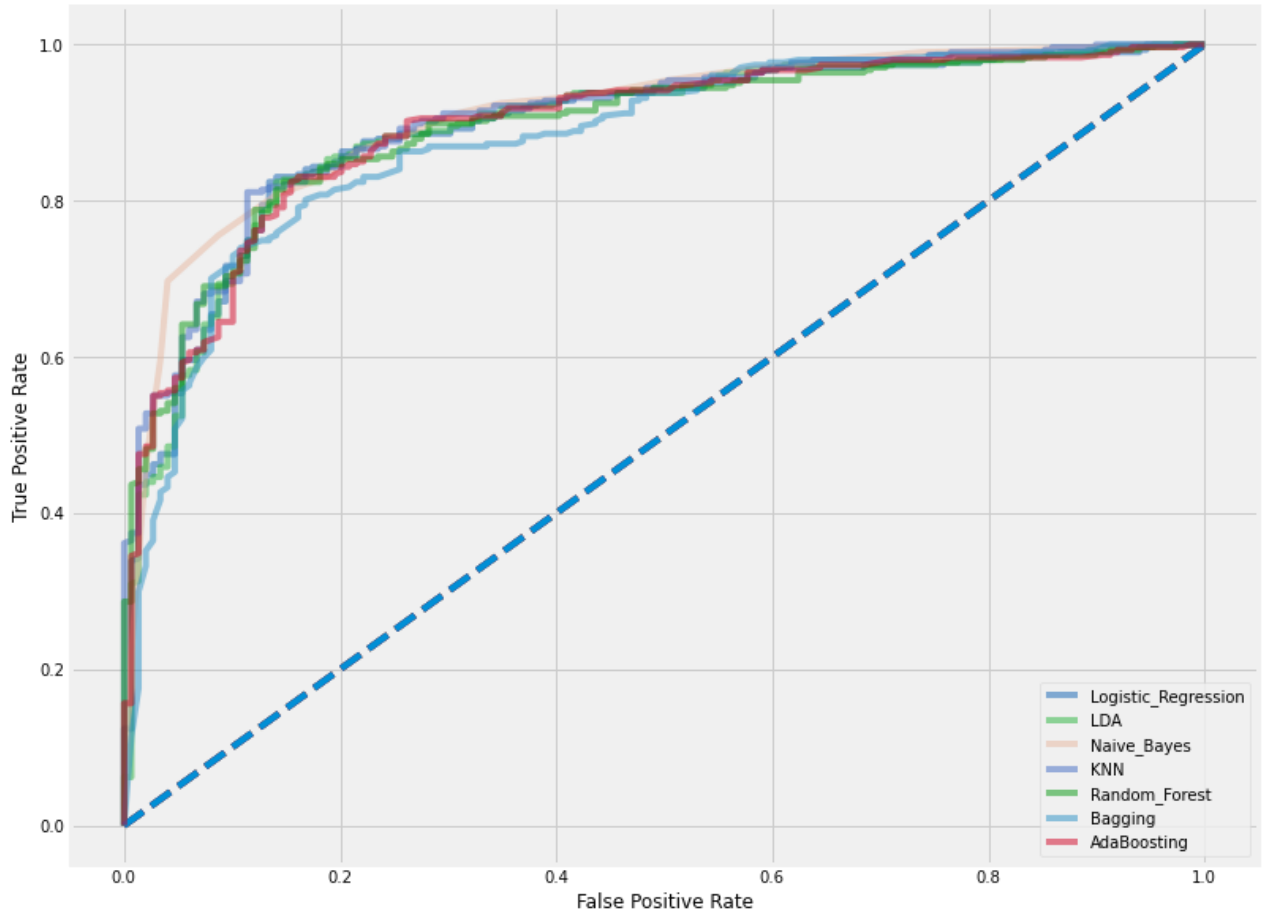
Confusion Matrix



ROC AUC Score for the test data set:

From the below graph, it seems that the Naïve Bayes model has done a better job, compared to the other models.

The 2nd best model is KNN.



The ROC AUC scores for the models are as follows:

AUC Score Logistic_Regression	**0.8932404958135671**
AUC Score LDA	**0.892606519030234**
AUC Score Naive_Bayes	**0.9085324530529262**
AUC Score KNN	**0.9036683208359749**
AUC Score Random_Forest	**0.8939619176704632**
AUC Score Bagging	**0.8789869488227708**
AUC Score AdaBoosting	**0.8946396169905777**

Comparison of the model performance

The performance metrics for each model are summarized in the below table.

Recall is also considered in the comparison, as it is important to know out of the identified true data points, how many are actually true data points (i.e. True Positives). In this case, recall is taken (how many have actually voted for 'Labour' party)

	Logistic_Regression	LDA	Naive_Bayes	KNN	Random_Forest	Bagging	AdaBoosting
Recall_Train	91.47	90.53	90.93	94.13	90.13	100.00	92.53
Recall_Test	90.23	89.25	91.21	92.18	88.93	87.62	91.86
F1_Train	89.03	88.76	88.80	90.40	88.25	100.00	88.75
F1_Test	87.66	87.54	88.47	87.89	87.78	85.26	87.85
Accuracy_Train	84.07	83.79	83.79	85.86	83.03	100.00	83.41
Accuracy_Test	82.89	82.89	83.99	82.89	83.33	79.61	82.89
AUC_Score_Train	88.62	88.64	90.35	92.00	88.16	100.00	89.70
AUC_Score_Test	89.32	89.26	90.85	90.37	89.40	87.90	89.46

Inferences:

- All of the models appear to do well as they have accuracy above 80%.
- The **best model is the Gaussian Naïve Bayes classifier** as the AUC scores for training and testing are 90.35% and 90.85% respectively.
    - In terms of Recall as well, it does a little better on the testing data (91.21%) than on the training data (90.93%) compared to other models.
- Based on the AUC scores for train and test, the 2nd best model is KNN – 92% for train and 90.37% for test.
    - The Recall for KNN is also very good for both training (94.13%) and testing (92.18%).
- The model that does worse than all the other models is Bagging, as the accuracy is 100% for the train data (signifying over-fitting), however accuracy is almost 20% less for test data.

1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

Insights:

1. In the given dataset, the distribution of voters who have voted for 'Labour' and 'Conservative' is 70% and 30% respectively.
2. A slightly higher number of females have voted for 'Conservative' party.
3. From the given sample, a majority of the observations have a high score of 11 for 'Europe' – indicating that they are highly 'Eurosceptic'.
4. Around 617 people have scored only 2 on the assessment of the Conservative leader, and around 833 people have scored 4 on the assessment of the Labour leader.
5. A large number of voters (776) have a score of 2 (out of 3) in the assessment of knowledge of parties' positions on European integration, which is good.

6. A majority of the observations (645) have a score of 3 on the assessment of the household economic conditions. Less number of observations have a score above 3.
7. The Gaussian Naïve Bayes model is the best model for predicting the party the people will vote for.
    - ***From the Gaussian Naïve Bayes model test data confusion matrix (totally 456 data points), 103 people will vote for Conservative and 280 people will vote for Labour.***
8. The optimized Random Forest model was used to find the best features relevant for making predictions, and it is observed that 'Hague', 'Blair' and 'Europe' have high feature importance.

	Imp
Hague	0.330848
Blair	0.230758
Europe	0.201656
economic.cond.national	0.078300
political.knowledge	0.067236
age	0.055423
economic.cond.household	0.031617
gender_male	0.004162

<u>Recommendations:</u>

1. A survey can be taken among the female voters for checking if there is any specific reason for voting for the Conservative party.
2. The variables 'Hague', 'Blair' and 'Europe' should be paid more attention to, in the future for predicting the party, the people will vote for.
3. More samples can be taken to verify the findings as the data distribution for this sample is 30% for the Conservative party.
4. The Gaussian Naïve Bayes model can be used for further making predictions to new data.
    Hyperparameters: *'var_smoothing': 1e-10, 'priors': [0.25, 0.75]*