Machine Learning

PROJECT REPORT

DSBA

Prepared By : Renjith K P

Contents

List of Tables	
Table 1	Dataset Sample
Table 2	DataTypes
Table 3	Five Point Summary
Table 4	Classification Report Logistic Regression
Table 5	Feature Importance Logistic Regression
Table 6	Classification Report LDA
Table 7	Feature importance LDA
Table 8	Classification Report KNN Model
Table 9	Classification Report Naïve Bayes Model
Table 10	Classification Report Boosting
Table 11	Classification Report of Toned Logistic Regression Model
Table 12	Classification Report of Toned LDA Model
Table 13	Classification Report of Toned LDA Model
Table 14	Classification Report of Toned Boosting Model
Table 15	Model Comparison
Table 16	Count of words
Table 17	Top 3 words

Problem 1:

Linear Regression

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.

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

Data loaded and checked following are the observations.

Executive Summary

Here the dataset have collection of survey information about voters who will cast their vote for recent elections. We will build a model to predict which party a voter will vote for on the basis of the given information. In this problem we will understand the data in depth then build machine learning models following this we will be using different bagging and boosting techniques to improve the performance of the model

Introduction

Purpose of this whole exercise is to build a model that can be used to predict which party a voter will vote for on the basis of the given information. Data consist of information about 1525 voters with 10 variables. This assignment will help in improving the understanding of the student in exploring summary statistics, exploratory data analysis, data cleaning, different model building for predictions, model performance improvement using bagging and boosting

Data Description

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.rchar - Number of characters transferred per second by system read calls
8. political.knowledge: Knowledge of parties' positions on European integration, 0 to 3.
9. gender: female or male.
10. usr - Portion of time (%) that cpus run in user mode

Sample of the Dataset

	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

Table 1 : Dataset Sample

Check the types of variables in the data frame.

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

Table 2 : Datatypes

Five Point Summary

	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

Table 3 : Five Point Summary

Inferences :

- Total 1525 voters' information's are available
- Variable names 'Unnamed:0' doesn't contain any use full information we can drop this variable, except this total 9 variables are there
- Out of the 9 variables 7 variables are of int64 data type and 2 variables are of object data type
- There were no null values present in this dataset, 8 duplicated entries observed and all of those removed. After removing duplicated entries 1517 voter information's are available in dataset
- Political.knowledge of 455 voters is 0, which is about 30% of total voters. Abot 30% of voters in this survey doesn't have any information about parties' positions on European integration
- In age mean is greater than median indicating the distribution is right skewed slightly
- Age range of voters in between 24 to 93 years, since 18 is the minimum eligible age for voting, people who is aged between 18 to 24 and people aged above 93 is not considered in this survay

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

Exploratory Data Analysis

Univariate Analysis

Boxplot and Histogram of Age

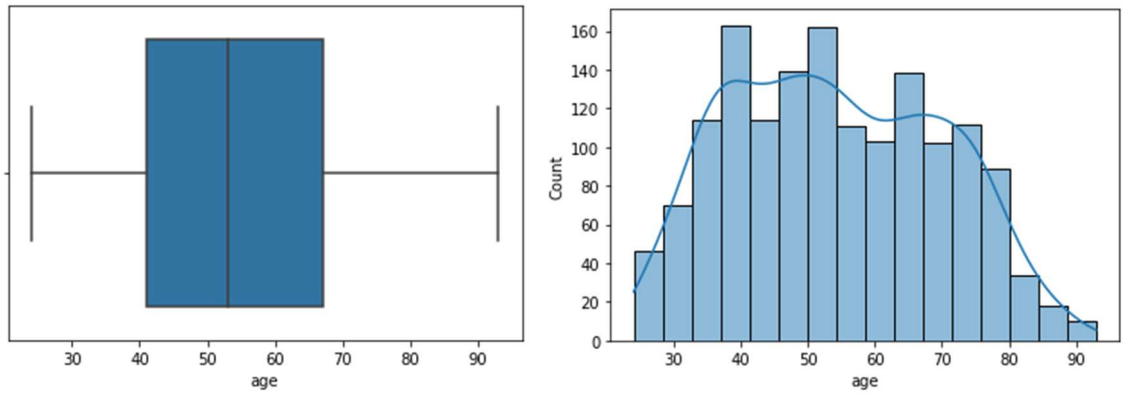


Figure 1 : age histogram and boxplot

Counterplots of Categorical Variables

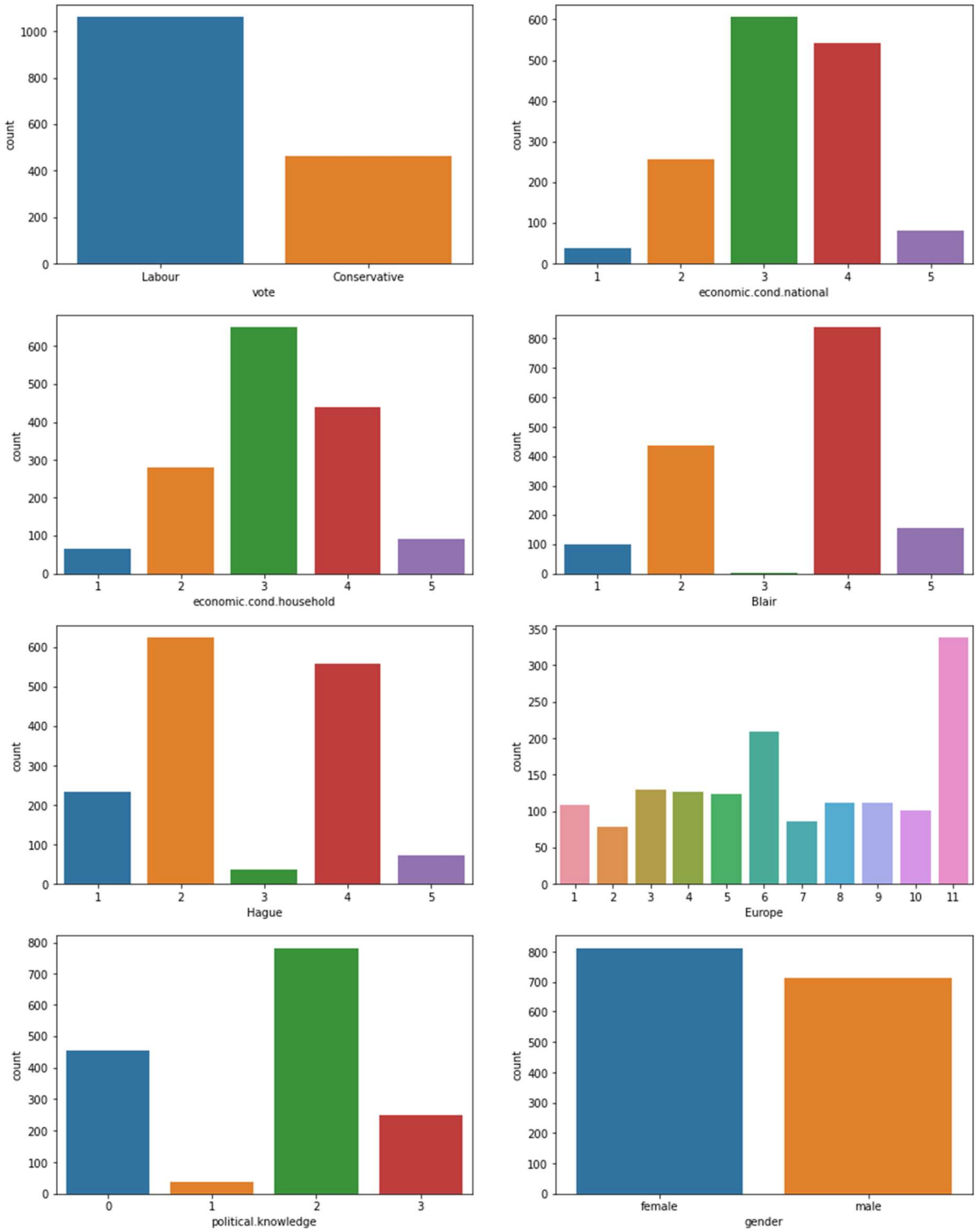
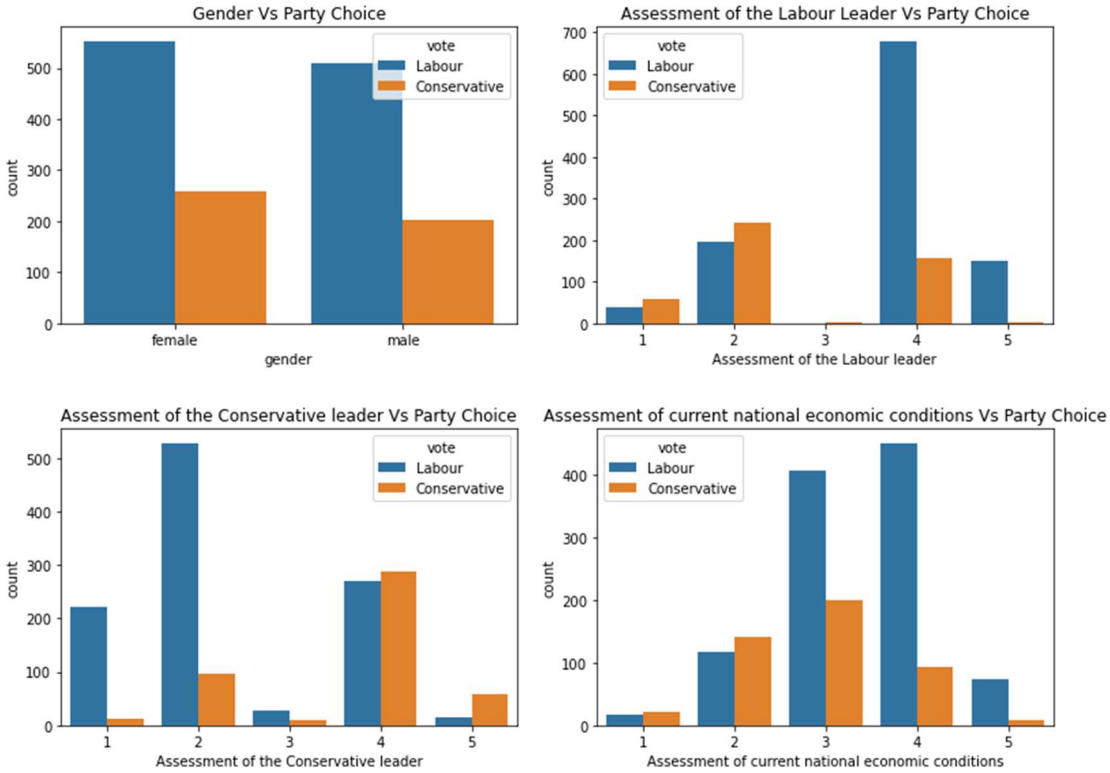


Figure 2 : Count plots for categorical variables

Inferences

- Survey covered people in age range 24 to 93. Distribution of age is slightly skewed to right.
- Majority of people in survey voted for Labour party. Out of 1525 peoples, 1063 peoples voted for labour party.
- Majority of the people assess current economic conditions and household conditions as 3 & 4 out of 5.
- More than 50% people believe in labour reader, they rated 4 out of 5
- Survey covers almost equal proportion of males (47%) and females (53%)

Bivariate Analysis



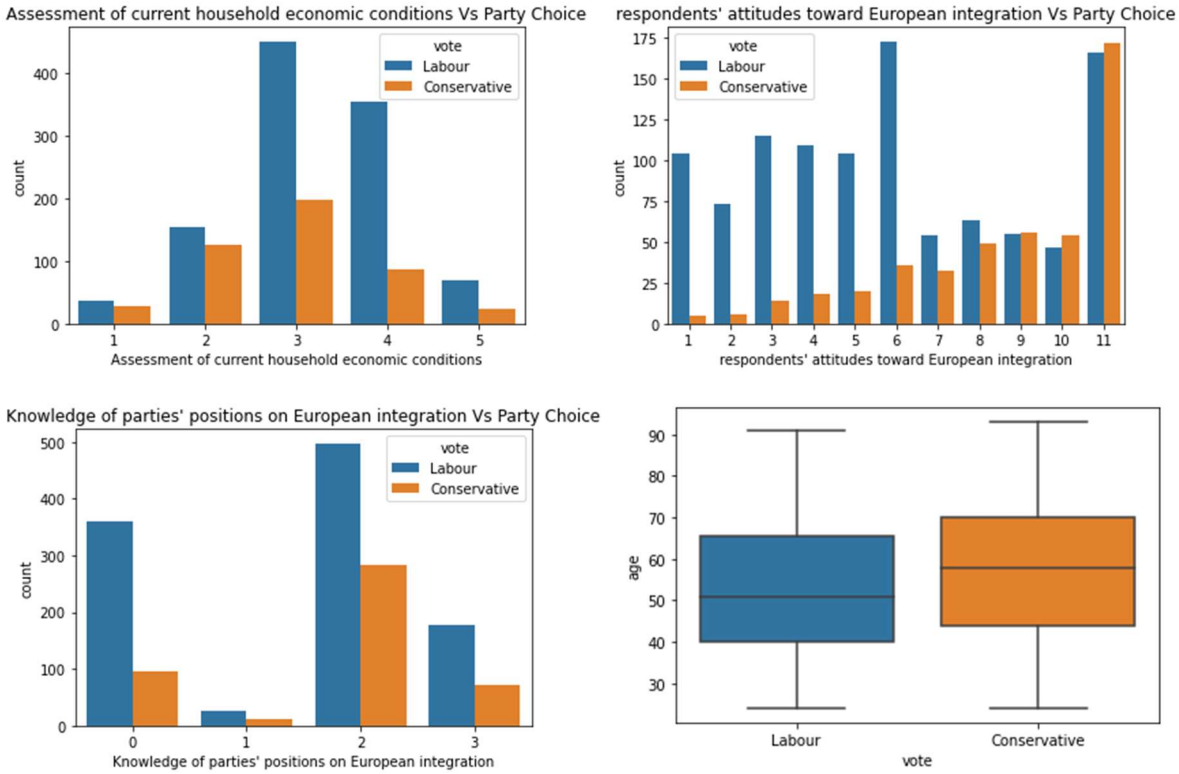


Figure 3 : Bivariate Plots for Continuous and Categorical Variables

Inferences

- More votes for labour party is casted by people assessed labour party leader as 4 out of 5
- More votes for conservative party is casted by people assessed conservative party leader as 4 out of 5
- Majority of peoples casted votes for conservative party rated their leader 4 out of 5
- Majority of peoples casted votes for labour party rated conservative party leader 2 out of 5
- People voted for conservative party aged slightly older than people voted for labour party

Multivariate Analysis

Pair Plot

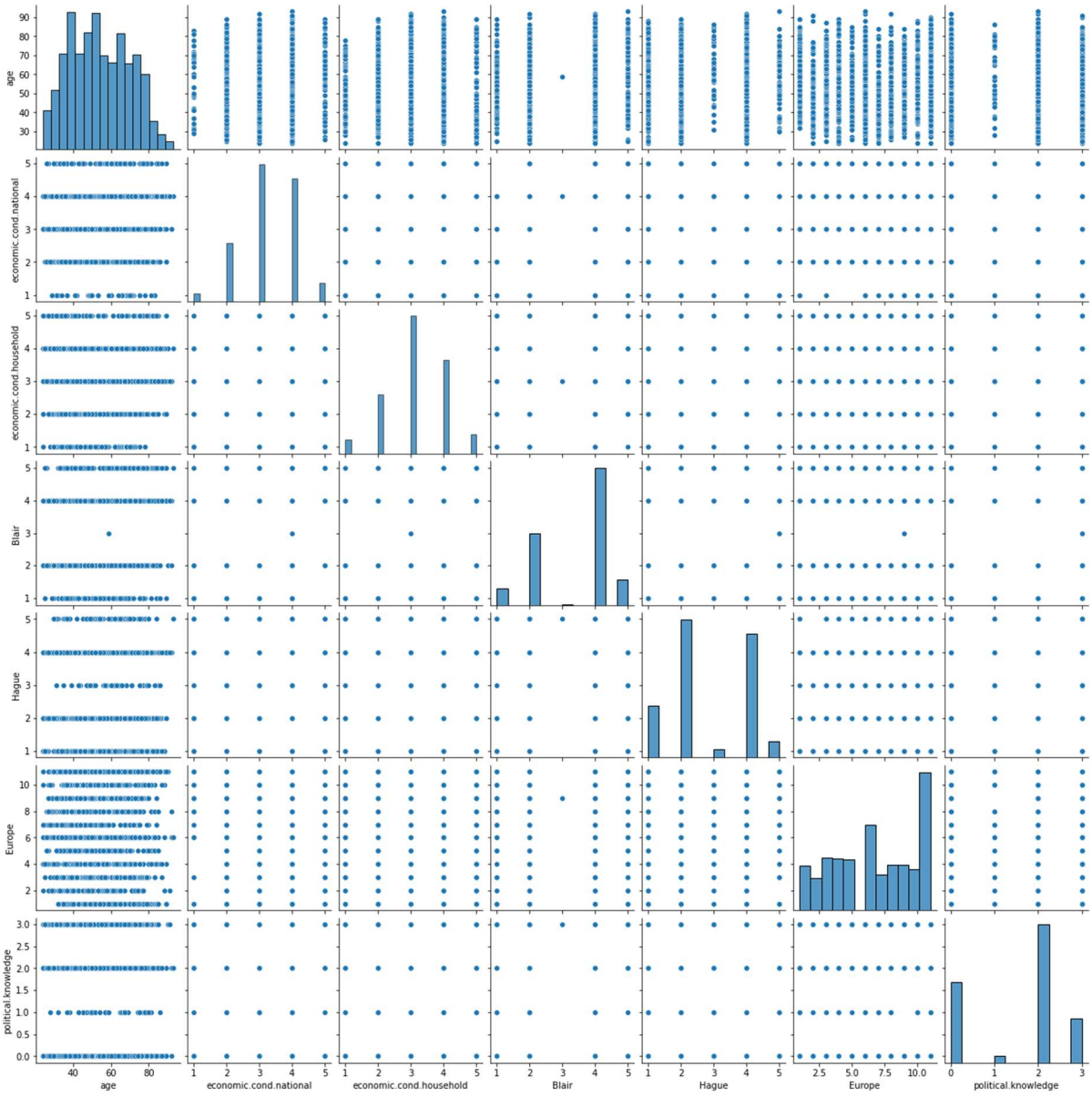


Figure 4 : Pairplot

Heat Map

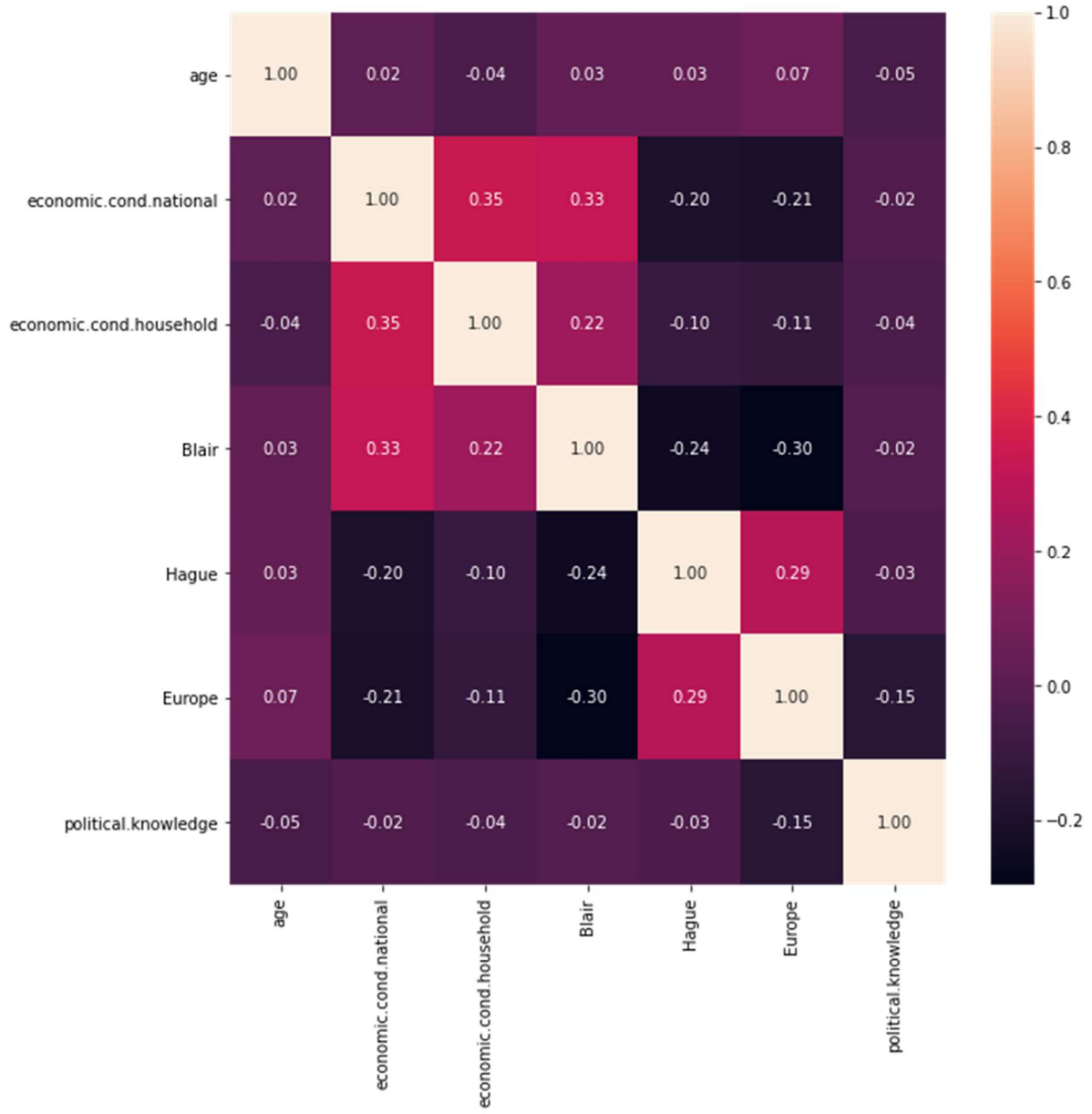


Figure 5 : Heatmap

Inferences

- There were no multicollinearity present in the data

Outlier Check and Treatment

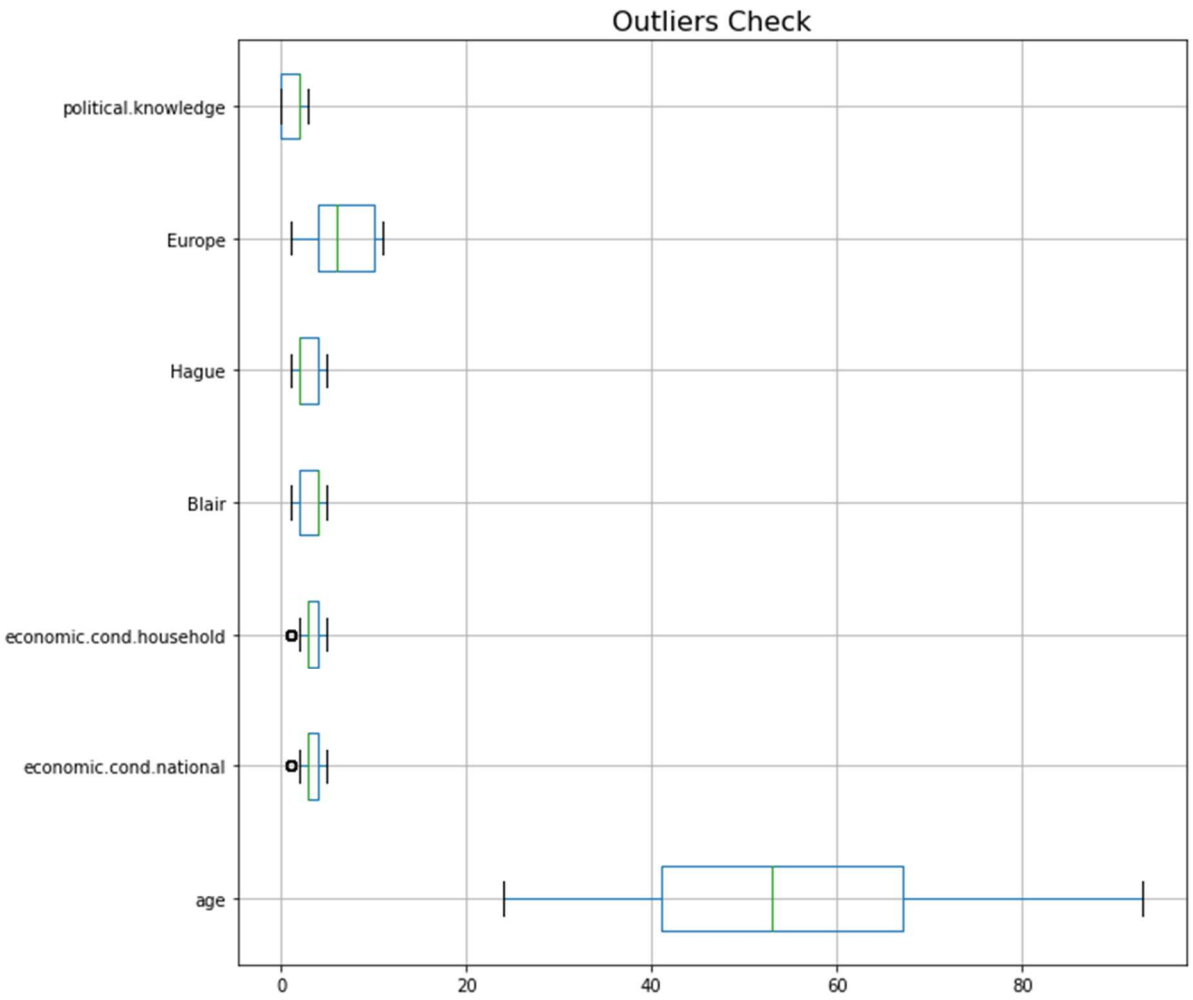


Figure 6 : Boxplot to Check Outliers

From box plots we could see some outliers present in some variables. Except age other variables are categorical and in age there were no outliers present. The outliers will be the values that are out of the (1.5*IQR) from the 25 or 75 percentile. There were no values present in age. Outliers present in categorical variables here cannot be considered as outliers since they are true values. Hence we can conclude there were no outliers present in the data.

Data types present were int64 and object, we need to convert object data types to numerical by encoding them. All features except age is converted to categorical data type.

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).

Data encoded using cat.codes and one hot encoding. All features except age and gender is converted to categorical data type since they all are ordinal.

Variable age is having standard deviation 15.7, which is very high compared to other variables. Europe is also having higher standard deviation (3.3) compared to other variables where standard deviation is around 1. Mean of variables are different from one another. Since we are dealing with most categorical variables these measures were not that significant to discuss, these were applicable for continuous numerical variables.

Some modelling techniques like KNN requires scaling, hence we will created models with both scaled and normal data and check performance.

We created two datasets one with scaling and another without scaling and performance of these compared, we couldn't find any improvement in model performance with scaling the data.

Data is split into train and test in ration of 70% train and 30% test as per the instructions. Train dataset contains 1067 voters details and test data set contains 456 voters details. Test_size is given 0.3 to consider 30% as test data. random_state is assigned to 1, we get the same train and test sets across different executions by this.

1.4 Apply Logistic Regression and LDA (linear discriminant analysis)

Logistic Regression

Some important hyper parameters for logistic regression are chosen as below

- solver is chosen 'newton-cg', 'newton-cg' can handle multinominal loss for multiclass problems
- penalty is choses defaule 'l2'
- The C parameter controls the penalty strength, default 1 is chosen
- maximum iterations for the solver to converge is given 100, default value
- Random_state is given as 1, throughout this project we will be using the same to control shuffling

Most of the parameters are chosen as default since those doesn't have any significant impact on the model.

Confusion Matrix

Figure 7 : Confusion Matrix Logistic Regression

Classification Reports

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.64	0.69	307
1	0.86	0.91	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.77	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.76	0.74	0.75	153
1	0.87	0.88	0.88	303
accuracy			0.84	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.84	0.83	456

Table 4 : Classification Report Logistic Regression

Feature Importance

	Attribute	Importance
1	economic.cond.national	0.628713
3	Blair	0.600866
7	gender_male	0.191826
2	economic.cond.household	0.063089
0	age	-0.014948
5	Europe	-0.211608
6	political.knowledge	-0.322012
4	Hague	-0.823236

Table 5 : Feature Importance Logistic Regression

- Accuracy of the model in train and test data set are 83% and 84% respectively, model is having good accuracy
- In the test data 74% of people voted for conservative party is identified correctly and 88% of the people voted for labour party predicted correctly. High recall for labour party voters is due to the higher number of labour party supporters in train and test datasets
- Recall is the measure which can be chosen as most important to assess the performance of model here. Because in this case how many people voted actually for the parties predicted as they will vote is important here. In the test dataset recall is 74% and 88% for conservative party and labour party respectively
- Accuracy, Precision and Recall for test data is almost in line with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification. But since the number of conservative party voters are less in the available data, model doesn't perform well in predicting conservative party voters
- Out of people classified as labour party voters 87% were correct in the test dataset and 76% were correct out of people classified as conservative party voters
- Assessment of the labour leader is the most important feature followed by assessment of current national economic conditions

Linear Discriminant Analysis (LDA)

Default hyper parameters chosen while building LDA model. Solver is chosen as default 'svd'. There were no requirement of dimensionality reduction hence n_components set as default 1. Changing other parameters doesn't give any performance improvements hence all are set as default.

Confusion Matrix

Figure 8 : Confusion Matrix LDA

Classification Report

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.65	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.77	0.73	0.74	153
1	0.86	0.89	0.88	303
accuracy			0.83	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 6 : Classification Report LDA

Feature Importance

	Attribute	Importance
3	Blair	0.742400
1	economic.cond.national	0.604920
7	gender_male	0.149080
2	economic.cond.household	0.050069
0	age	-0.020037
5	Europe	-0.223612
6	political.knowledge	-0.430335
4	Hague	-0.926634

Table 7 : Feature importance LDA

- Accuracy of the model in train and test data set are 83%, model is having good accuracy
- In the test data 73% of people voted for conservative party is identified correctly and 89% of the people voted for labour party predicted correctly. High recall for labour party voters is due to the higher number of labour party supporters in train and test datasets
- Recall is the measure which can be chosen as most important to assess the performance of model here. Because in this case how many people voted actually for the parties predicted as they will vote is important here.
- Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification. But since the number of conservative party voters are less in the available data, model doesn't perform well in predicting conservative party voters.
- In test data 77% prediction of conservative party voters are correct and 86% predictions of labour party voters are correct
- Assessment of the labour leader is the most important feature followed by assessment of current national economic conditions

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

KNN Model

Most important hyperparameters for KNN model are

N_neighbors : Number of neighbors to use by default for kneighbors queries. Default value 5 is given here.

metric : 'minkowski distance is chosen as default

weights : default uniform is chosen, all points in each neighborhood are weighted equally by this.

KNN model is built after doing z score transformation since KNN requires scaling of data.

When we checked misclassification errors with different values of n_neighbors, its found that number of neighbors K equals to 18,17 having lowest error followed by 7. We made models with different values of K to check whether if we get any model having accuracy difference less than 10% between train and test dataset.

The best model we got is with n_neighbors = 18. Its having a test accuracy of 84% and train accuracy of 100%

Confusion Matrix

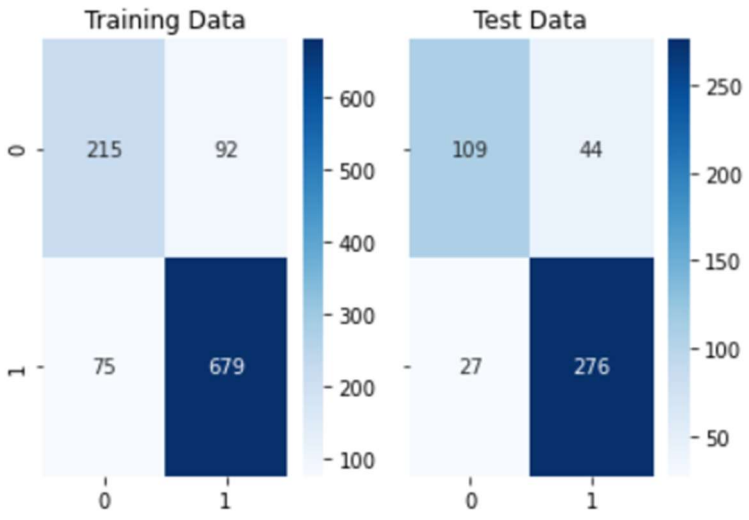


Figure 9 : Confusion Matrix KNN Model

Classification Report

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.70	0.72	307
1	0.88	0.90	0.89	754
accuracy			0.84	1061
macro avg	0.81	0.80	0.81	1061
weighted avg	0.84	0.84	0.84	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.80	0.71	0.75	153
1	0.86	0.91	0.89	303
accuracy			0.84	456
macro avg	0.83	0.81	0.82	456
weighted avg	0.84	0.84	0.84	456

Table 8 : Classification Report KNN Model

- Accuracy of the model in train and test data set are 84%, model is having good accuracy
- In the test data 71% of people voted for conservative party is identified correctly and 89% of the people voted for labour party predicted correctly. High recall for labour party voters is due to the higher number of labour party supporters in train and test datasets
- Recall is the measure which can be chosen as most important to assess the performance of model here. Because in this case how many people voted actually for the parties predicted as they will vote is important here.
- Accuracy, Precision and Recall for test data is almost inline with training data. This proves no overfitting or under fitting has happened, and overall the model is a good

model for classification. But since the number of conservative party voters are less in the available data, model doesn't perform well in predicting conservative party voters.

- In test data 80% prediction of conservative party voters are correct and 86% predictions of labour party voters are correct

Naïve Bayes Model

Hyper parameters in Naïve Bayes Model are var_smoothing : Portion of the largest variance of all features that is added to variances for calculation stability. This chosen as default=1e-9.

Another hyper parameter is priors, which means prior probabilities of classes, here also default 'None' is chosen.

Confusion Matrix

Figure 10 : Confusion Matrix Naïve Bayes Model

Classification Report

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.73	0.69	0.71	307
1	0.88	0.90	0.89	754
accuracy			0.84	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.84	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.74	0.73	0.73	153
1	0.87	0.87	0.87	303
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

Table 9 : Classification Report Naïve Bayes Model

- Accuracy of the model in train and test data set are 84% and 82% respectively, model is having good accuracy
- In the test data 73% of people voted for conservative party is identified correctly and 87% of the people voted for labour party predicted correctly. High recall for labour party voters is due to the higher number of labour party supporters in train and test datasets
- Recall is the measure which can be chosen as most important to assess the performance of model here. Because in this case how many people voted actually for the parties predicted as they will vote is important here.
- Accuracy, Precision and Recall for test data is almost inline with training data. This proves no overfitting or under fitting has happened, and overall the model is a good model for classification. But since the number of conservative party voters are less in the available data, model doesn't perform well in predicting conservative party voters.

- In test data 74% prediction of conservative party voters are correct and 87% predictions of labour party voters are correct. Precision for conservative party prediction is low

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

Bagging

The most important hyper parameter for bagging is *n_estimators which is chosen'50.' It means the number of base estimators in the ensemble*

Estimator is given as the random forest classifier

Random_state is given 1, which controls the shuffling of data

base_estimator chosen is RandomForestClassifier()

max_features : The number of features to draw from X to train each base estimator is chosen default ' 1'

Confusion Matrix

Figure 11 : Confusion Matrix Bagging

Classification Report

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.98	0.90	0.94	307
1	0.96	0.99	0.98	754
accuracy			0.96	1061
macro avg	0.97	0.94	0.96	1061
weighted avg	0.96	0.96	0.96	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.79	0.67	0.73	153
1	0.85	0.91	0.88	303
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.83	456

Table 10 : Classification Report Bagging

- Accuracy of the model in train and test data set are 96% and 83% respectively, From train dataset to test dataset accuracy drop is 13% which indicates model is overfitting.
- In the test data 67% of people voted for conservative party is identified correctly and 91% of the people voted for labour party predicted correctly. High recall for labour party voters is due to the higher number of labour party supporters in train and test datasets
- Recall is the measure which can be chosen as most important to assess the performance of model here. Because in this case how many people voted actually for the parties predicted as they will vote is important here.
- Accuracy, Precision and Recall for test data is significantly lower than of train dataset. This indicates model is overfitted.
- In test data 79% prediction of conservative party voters are correct and 85% predictions of labour party voters are correct. Precision is good.

Boosting

There are some parameter pairings that are important to consider for gradient boosting. The first is the learning rate, also called shrinkage or eta (*learning_rate*) and the number of trees in the model (*n_estimators*). Both could be considered on a log scale, although in different directions.

All parameters chosen as default. Random_state chosen as 1.

Confusion Matrix

Figure 12 : Confusion Matrix Boosting

Classification Report

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.84	0.78	0.81	307
1	0.91	0.94	0.93	754
accuracy			0.89	1061
macro avg	0.88	0.86	0.87	1061
weighted avg	0.89	0.89	0.89	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.80	0.69	0.74	153
1	0.85	0.91	0.88	303
accuracy			0.84	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.83	0.84	0.83	456

Table 10 : Classification Report Boosting

- Accuracy of the model in train and test data set are 89% and 84% respectively, Accuracy of the model is good.
- In the test data 69% of people voted for conservative party is identified correctly and 91% of the people voted for labour party predicted correctly. High recall for labour party voters is due to the higher number of labour party supporters in train and test datasets
- Recall is the measure which can be chosen as most important to assess the performance of model here. Because in this case how many people voted actually for the parties predicted as they will vote is important here.
- Precision,Accuracy and Recall for test data is lower than of train dataset. The model is slightly overfitted.
- In test data 80% prediction of conservative party voters are correct and 85% predictions of labour party voters are correct. Precision is good.

Grid Search

Grid search toning in Logistic Regression Model

Confusion Matrix

Figure 13 : Confusion Matrix Toned Logistic Regression Model

Classification Report

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.63	0.68	307
1	0.86	0.91	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.77	0.78	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.76	0.72	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 11 : Classification Report of Toned Logistic Regression Model

Grid search toning on LDA model

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.65	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.77	0.73	0.74	153
1	0.86	0.89	0.88	303
accuracy			0.83	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 12 : Classification Report of Toned LDA Model

Grid Search Toning KNN Model

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.63	0.68	307
1	0.86	0.91	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.77	0.78	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.76	0.72	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 13 : Classification Report of Toned KNN Model

Grid Search Toning for Boosting

Classification Report

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.84	0.78	0.81	307
1	0.91	0.94	0.93	754
accuracy			0.89	1061
macro avg	0.88	0.86	0.87	1061
weighted avg	0.89	0.89	0.89	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.80	0.69	0.74	153
1	0.85	0.91	0.88	303
accuracy			0.84	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.83	0.84	0.83	456

Table 14 : Classification Report of Toned Boosting Model

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 Comparison Table													
Model No	Description	Accuracy		Precision				Recall				AUC Score	
		Train	Test	Train (0)	Train (1)	Test (0)	Test (1)	Train (0)	Train (1)	Test (0)	Test (1)	Train	Test
1	Logistic Regression	83%	84%	74%	86%	76%	87%	64%	91%	74%	88%		
2	LDA	83%	83%	74%	86%	77%	86%	65%	91%	73%	89%		
3	KNN	84%	84%	74%	88%	80%	86%	70%	90%	71%	91%		
4	Naïve Bayes	84%	82%	73%	88%	74%	87%	69%	90%	73%	87%		
5	Bagging	96%	83%	98%	96%	79%	85%	90%	99%	67%	91%		
6	Boosting	89%	84%	84%	91%	80%	85%	78%	94%	69%	91%		
7	Logistic Regression Toned	83%	83%	74%	86%	76%	86%	63%	91%	72%	88%		
8	LDA Toned	83%	83%	74%	86%	77%	86%	65%	91%	73%	89%		
9	KNN Model Toned	83%	83%	74%	86%	76%	86%	63%	91%	72%	88%		
10	Boosting Model Toned	89%	84%	84%	91%	80%	85%	78%	94%	69%	91%		

Table 15 : Model Comparison

KNN Model is the best model here

1.8 Based on these predictions, what are the insights?

- KNN Model is the best model for prediction in this case
- Labour party can consider this exit poll survey prediction since recall for labour party class is 91%
- In prediction of conservative party voters recall is low due to sample is biased.