Machine Learning

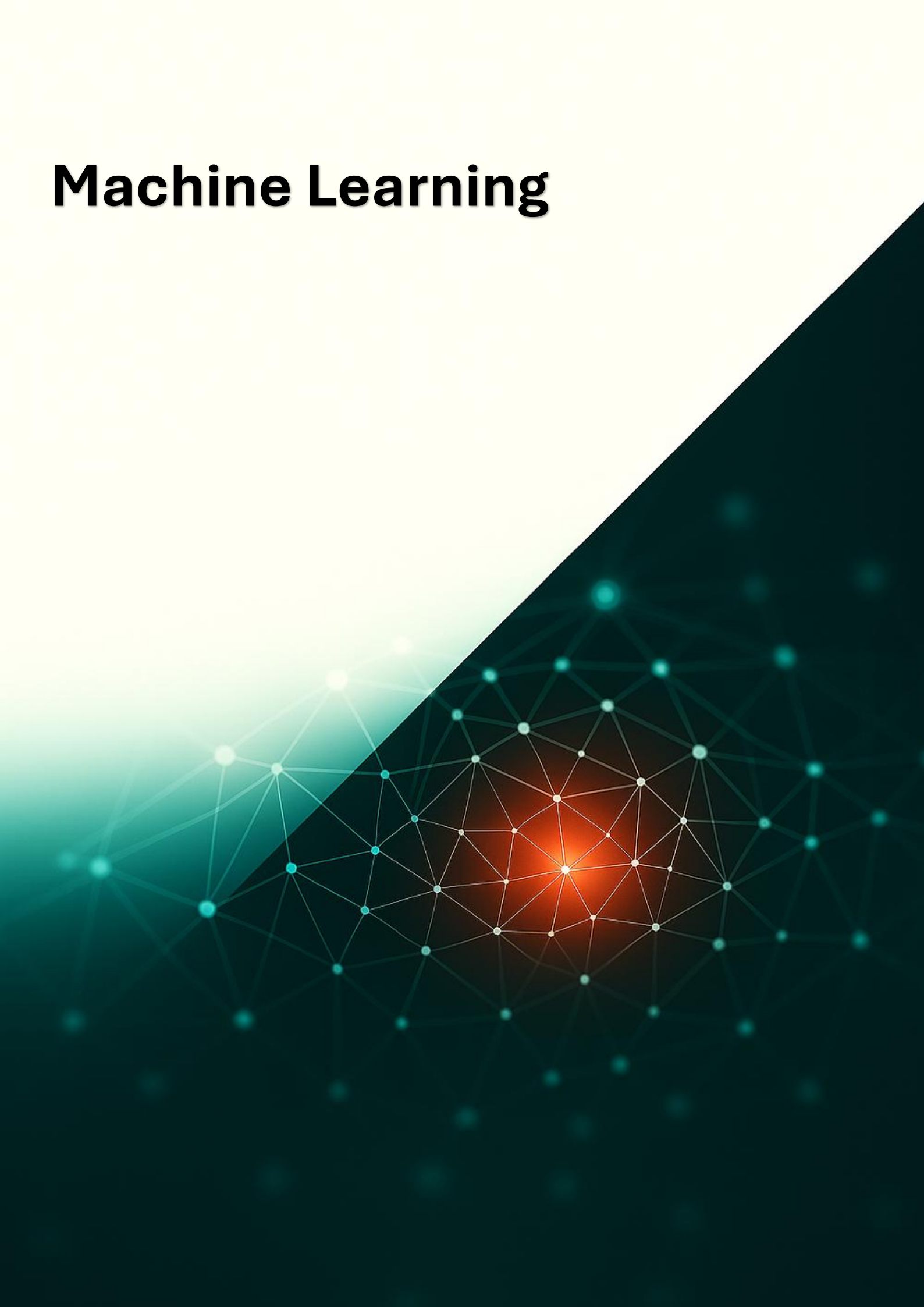

Table of Contents:

List of Figures:

List of Tables:

Problem 1:

You are hired by one of the leading news channels CNBE who wants to analyse 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.

Data Dictionary

Sl.no	Variable Name	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
8	**political.knowledge**	Knowledge of parties' positions on European integration, 0 to 3
9	**gender**	female or male

Table 1. Data Dictionary

Objective of the report:

The purpose of the report is to create an exit poll for the news channel CNBE that will aid in predicting overall win and seats covered by a particular political party: "Conservative" or "Labour." The classification will be performed using several classification models, which will then be compared.

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

Read and view data after dropping 'Unnamed: 0' variable:

Reading the dataset from the .csv file and checking the head (10) of the dataset i.e., the first 10 rows of the dataset.

Checking for number of rows and columns:

- Number of rows: 1525
- Number of columns: 9

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	Labour	43	3	3	4	1	2	2	female
1	Labour	36	4	4	4	4	5	2	male
2	Labour	35	4	4	5	2	3	2	male
3	Labour	24	4	2	2	1	4	0	female
4	Labour	41	2	2	1	1	6	2	male
5	Labour	47	3	4	4	4	4	2	male
6	Labour	57	2	2	4	4	11	2	male
7	Labour	77	3	4	4	1	1	0	male
8	Labour	39	3	3	4	4	11	0	female
9	Labour	70	3	2	5	1	11	2	male

Figure 1. Dataset of Problem 1.

Brief Introduction of Dataset:

According to the dataset, the election will be contested by two parties: Labour and Conservative. Other variables in the dataset provide information about the respondents 'gender' and 'age'. The assessment of current national economic conditions and household economic conditions from 1 to 5 rating of the respondents. There are two more variables that talks about the assessment of the Labour leader and Conservative leader – Blair and Hague respectively. They are also assessed on a scale of 1 to 5. Another variable called Europe is found in the dataset which measures the respondents' attitude toward European Integration on a 11-point scale. Higher scores represent Eurosceptic sentiment. Lastly, the variable talks about the knowledge of the parties' position on European Integration on a scale of 0 to 3.

In order to be able to create an exit poll for the news channel CNBE that will aid in predicting overall win and seats covered by a particular political party: "Conservative" or "Labour." The classification will be performed using several classification models, which will then be compared.

Checking for the information of features and Null values:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
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: 107.4+ KB
```

Figure 2. Information of features for Problem 1.

Checking the Skewness and Kurtosis:

Variable	Skewness	Kurtosis
age	0.144621	-0.946897
economic.cond.national	-0.240453	-0.255994
economic.cond.household	-0.149552	-0.206302
Blair	-0.535419	-1.065582
Hague	0.152100	-1.391743
Europe	-0.135947	-1.237841
political.knowledge	-0.426838	-1.216646

Table 2. Skewness & Kurtosis of features.

Skewness and Kurtosis is also calculated for each column, Data with high skewness indicates lack of symmetry and high value of kurtosis indicates heavily tailed data.

Checking the description of dataset:

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

Figure 3. Summary of the dataset for Problem 1.

Observations:

- Dataset has 9 columns and 1525 rows excluding the 'unnamed:0' column.
- The first column "Unnamed: 0" has only serial numbers, so we can drop it as it is not useful.
- There are 9 variables out of which the 'vote' variable is the dependent variable and the rest are independent variables.
- Among the independent variables, except for 'age' variable which is continuous, all the other 7 variables are categorical type.
- Among the categorical independent variables, except for 'gender' all the other are ordinal categorical variables
- Vote will be target variable.
- The dataset is used for predicting 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.
- There are 8 Duplicated instances. These are voters with the all the attributes same does not seem to be logical and do not add any value to the study hence we will drop the duplicates so avoid any overlapping.
- The target variable 'vote' shows that 69.68% voters are in favour of the 'Labour' Party and 32.32% in favour of 'Conservative' Party. There is a class imbalance noticed from a modelling perspective. We'll look at the classification model outputs to determine whether this has an effect and if any treatment is necessary.

- The percentage of male and female voters are nearly same with 53.26% female and 46.74% male. As a result, all genders appear to be equally represented.
- Most of the voters surveyed assessed the current national economic conditions on average scale with around 75% voters giving a rating of 3 or 4. Only around 8% of surveyed voters gave extreme ratings of 5 or 1.
- Most of the voters surveyed assessed the current household economic conditions on average scale with around 70% voters giving a rating of 3 or 4. Only around 10% of surveyed voters gave extreme ratings of 5 or 1.
- The Labour Party Leader 'Blair' received ratings of 4 or above by around 65% of the surveyed voters.
- The Conservative Party Leader 'Hague' received ratings of 4 or below by around 40% of the surveyed voters.
- Around 22% of the surveyed voters highly disregard closer links between Britain and European Union, i.e., they rated the maximum (11) on the 'Eurosceptic' sentiment scale.
- Around 67% of the surveyed voters are familiar with the viewpoints of the 'Labour' and 'Conservative' parties on European integration. The rest of the of the surveyed voters have no notion or had just a rudimentary understanding of this front.

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

Checking for data types in the dataset:

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

Figure 4. Datatypes of Features.

Checking for duplicates in this dataset:

```
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

Figure 5. Output of duplicates instances in Problem 1.

There are 8 Duplicated instances. These are two or more voters with the all the attributes same does not seem to be logical and do not add any value to the study hence we will drop the duplicates so avoid any overlapping.

- Number of Rows and Columns before dropping duplicates (1525, 9)
- Number of Rows and Columns after dropping duplicates (1517, 9)

Checking for Null values in the dataset:

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

Figure 6. Null values of Problem 1.

When we check for null values in the given dataset using isnull() function, we get the above output, which clearly shows that **there are no null values present.**

Data Visualization:

Let us define a function 'Univariate Analysis numeric' to display information as part of univariate analysis of numeric variables. The function will accept column name and number of bins as arguments. The function will display the statistical description of the numeric variable, histogram or distplot to view the distribution and the box plot to view 5-point summary and outliers if any.

Univariate Analysis for Numeric Variable - 'age'

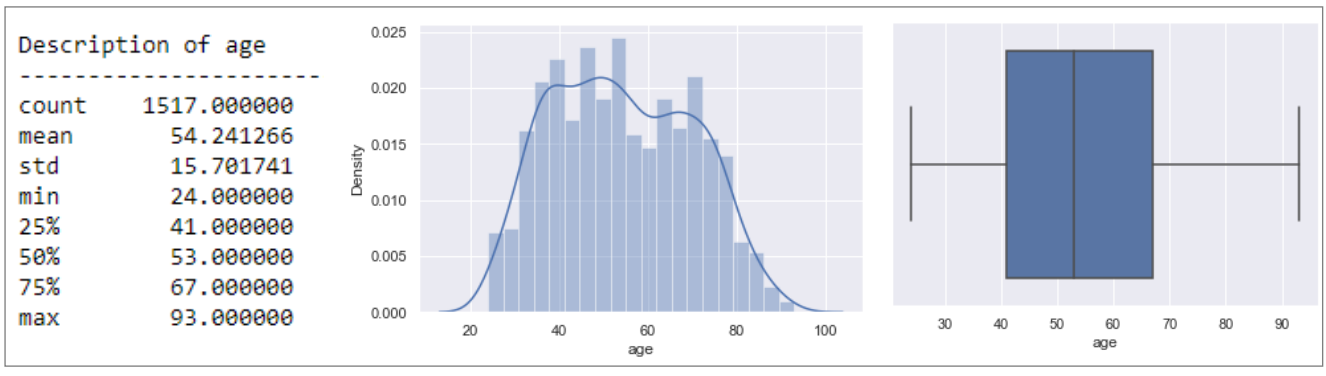

Figure 7. Boxplot and Distplot of 'Age' in years

- From the above graphs, we can infer that mean 'age' of the surveyed voters is around 54yrs with the minimum of 24yrs and maximum of 93yrs.
- The distribution of 'age' is slightly right skewed with skewness value of 0.14462.
- The distribution is almost normally distributed.
- The distplot shows the distribution of most of data from 25 to 80.
- The box plot of the 'age' variable shows no outliers.

Univariate Analysis for Categorical Variables:

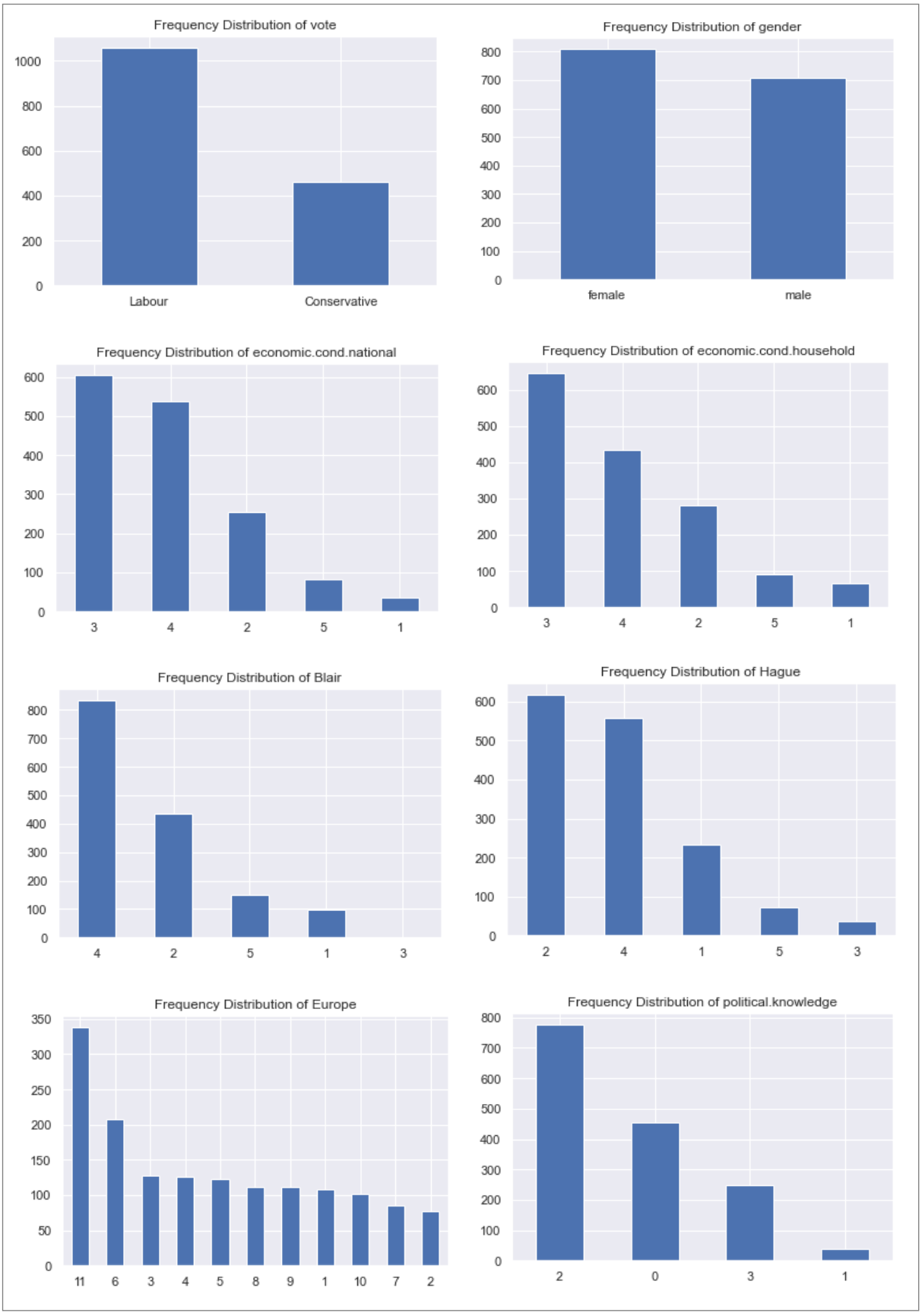

Figure 8. Count plots for categorical variables of Problem 1.

Observations:

- The target variable 'vote' shows that 69.68% voters are in favour of the 'Labour' Party and 32.32% in favour of 'Conservative' Party
- From 'gender' variable it is observed that male and female are nearly same.
- The conservative party does seem to be lesser favourable party of choice for respondents.
- Most voters rated 3 or above for the national economic condition and economic condition of household.
- For 'Blair' of the Labour Party most surveyed voters gave a rating of 4 and above and for 'Hague' of the Conservative party most voters gave a rating of 2 or below.
- One explanation for this might be because the Labour Party's leader has higher assessment ratings than the Conservative Party's leader. People believe Blair is a better leader for taking the country ahead.
- Although over 22% of surveyed voters are strongly Eurosceptic, with the rating of 11, the rest of the surveyed voters appear to be evenly spread between low and high Eurosceptic emotions, with just a slightly higher number of voters giving an average rating of 6.
- People are either extremely Eurosceptic or in the centre when it comes to European integration, as shown in the graph above. We may make the assumption that the party that does not favour the European Integration might be able to get the votes of the people.
- Respondents either seem to have a fair knowledge (49%) of the parties' position on the European Integration or no idea at all (30%).

Bivariate Analysis

Boxplot of Vote vs Economic conditions of national and household:

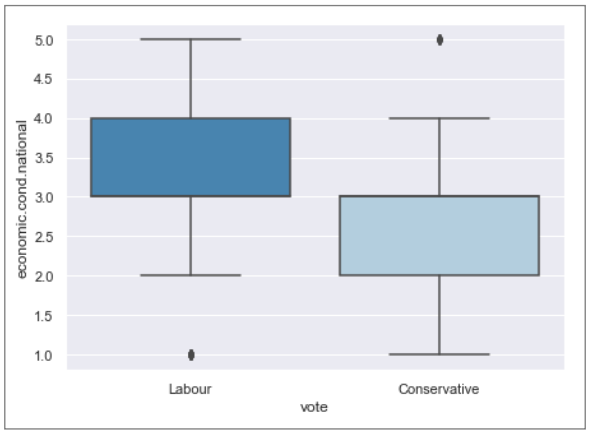

Figure 9. Boxplot of vote vs economic.cond.national.

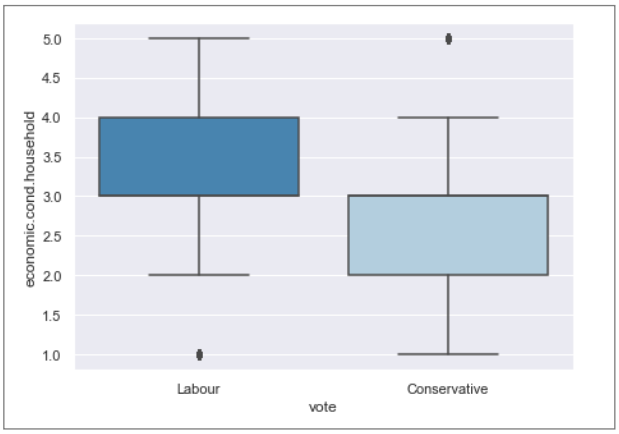

Figure 10. Boxplot of vote vs economic.cond.household

- The assessment rating of current national economic conditions and current household economic conditions for Labour party is better than the conservative party.
- More than 50% of the voters have rated 3 and 4 to 'Labour' party, while we can see that the 'Conservative' party has most of the ratings of 2 and 3.

Boxplot of Vote vs Blair and Hague:

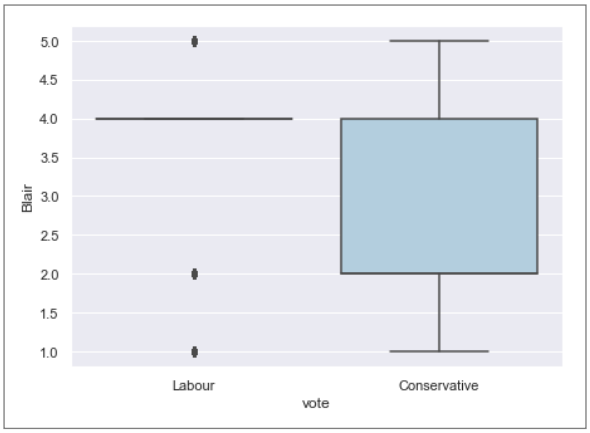

Figure 12. Boxplot of vote vs Blair

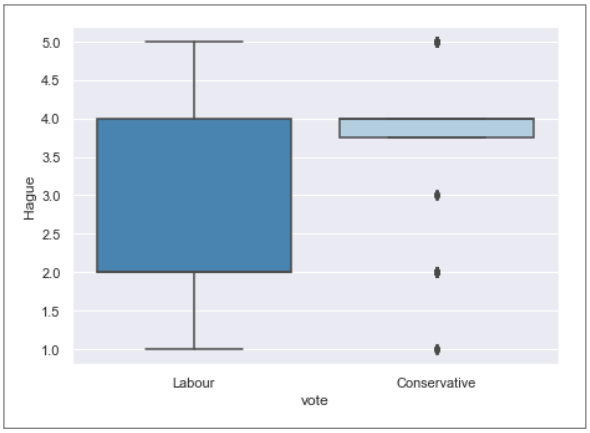

Figure 11. Boxplot of vote vs Hague

- The assessment rating of the Labour leader and Conservative leader – Blair and Hague respectively.
- For 'Blair' of the Labour Party most respondents gave a rating of 4 and above and for 'Hague' of the Conservative party most respondents gave a rating of 2 or below.

Boxplot of Vote vs Europe and Political knowledge:

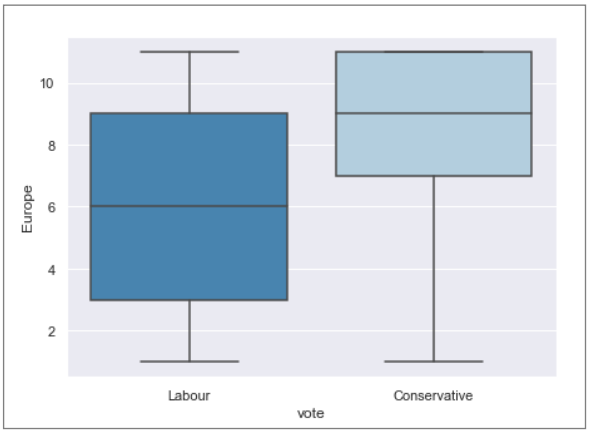

Figure 14. Boxplot of vote vs Europe.

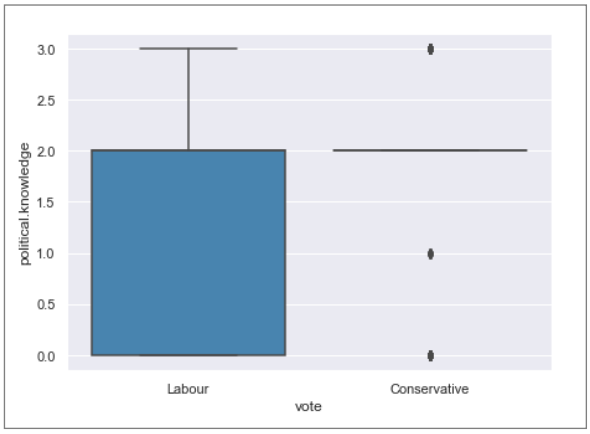

Figure 13. Boxplot of vote vs political knowledge.

- From the above plot we can see that the 'Conservative party' has high disregard on European union and the median ratings on 'Eurosceptic' sentiment scale is around 9, while the 'Labour party' ratings is 6.
- We could make the assumption that the party that are not against in favour of the European Integration might be able to get more votes of the people.
- An interesting observation to be made here is that respondents with a strong negative attitude towards European Integration are highly likely to vote for Conservative Party and respondents who are likely to vote for Labour Party seem to have evenly distributed attitude towards the Integration.
- People with less Knowledge of parties ‘positions on European integration and who show less Eurosceptic sentiment are more inclined towards Labour party.
- People with Knowledge of parties ‘positions on European integration and who show more Eurosceptic sentiment are more inclined towards Conservative party.

Boxplot of Age against Europe and Political knowledge with vote as hue:

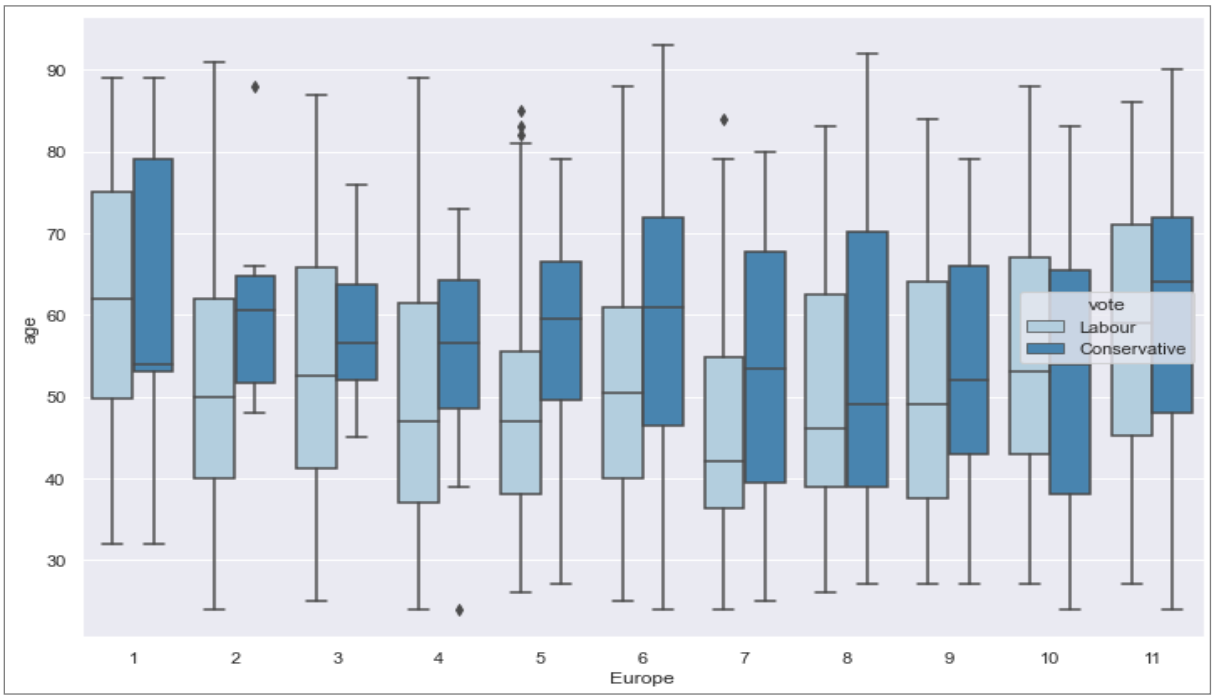

Figure 15. Boxplot of Age vs Europe with vote as hue.

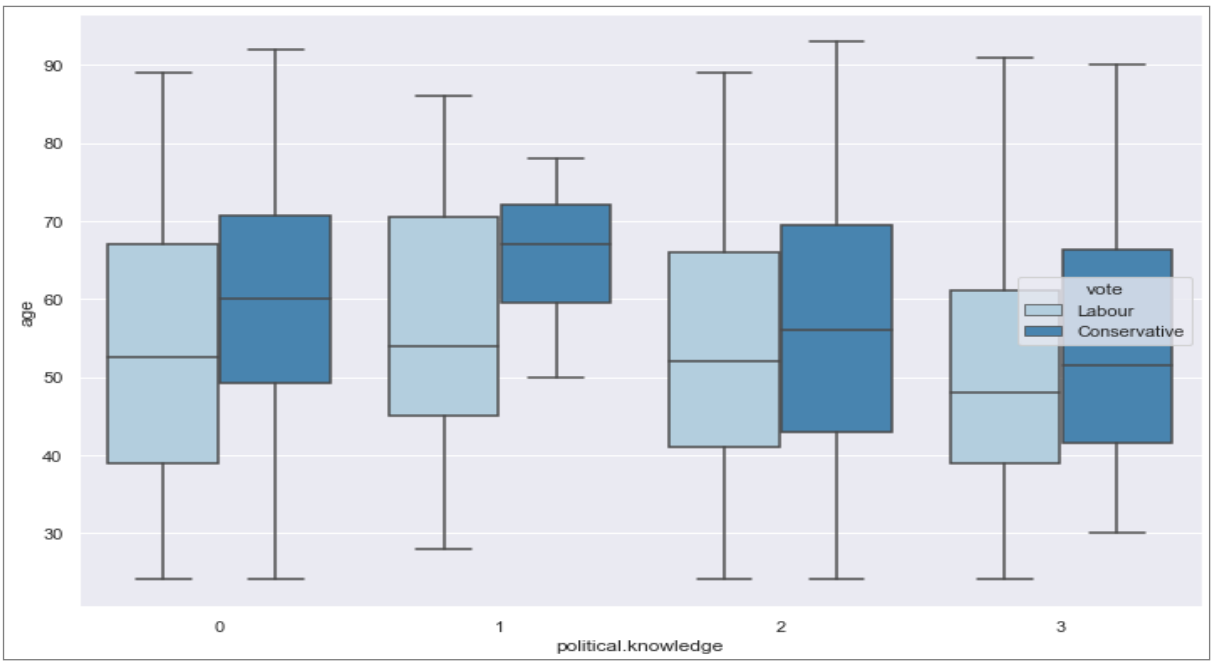

Figure 16. Boxplot of Age vs Political knowledge with vote as hue.

- The median age of people who voted for conservative party has slightly more people of older age group, than the people voting Labour party.
- People with more Knowledge of parties' positions on European integration and who show more Eurosceptic sentiment mostly belong to the older age group favouring Conservative party.

Count plot of categorical variables with vote as hue:

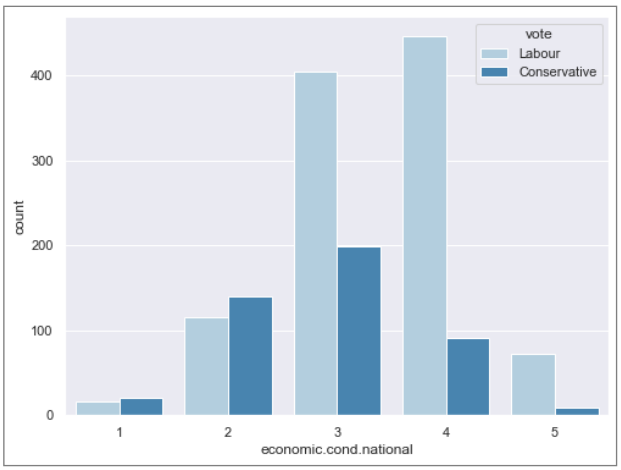

Figure 17. Count plot of Economic condition of national.

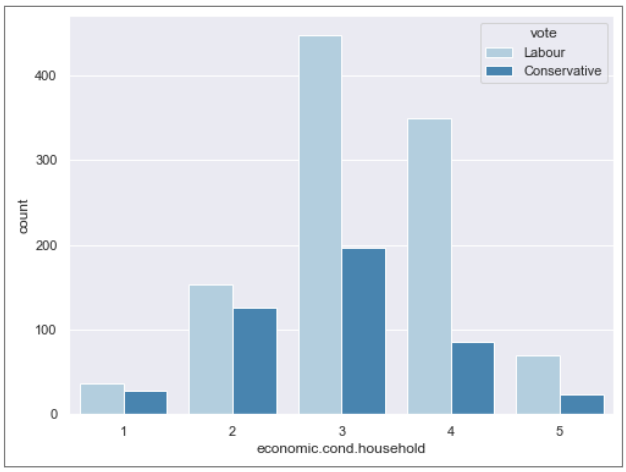

Figure 18. Count plot of Economic condition of household.

- Most of the voters surveyed assessed the current national economic conditions and current household conditions as 3 and 4 on the scale from 1 to 5.
- Many voters assessed the ratings of economic conditions and household conditions favouring to 'Labour' Party which is almost 50% (approximately) more than the people voting 'Conservative' Party.

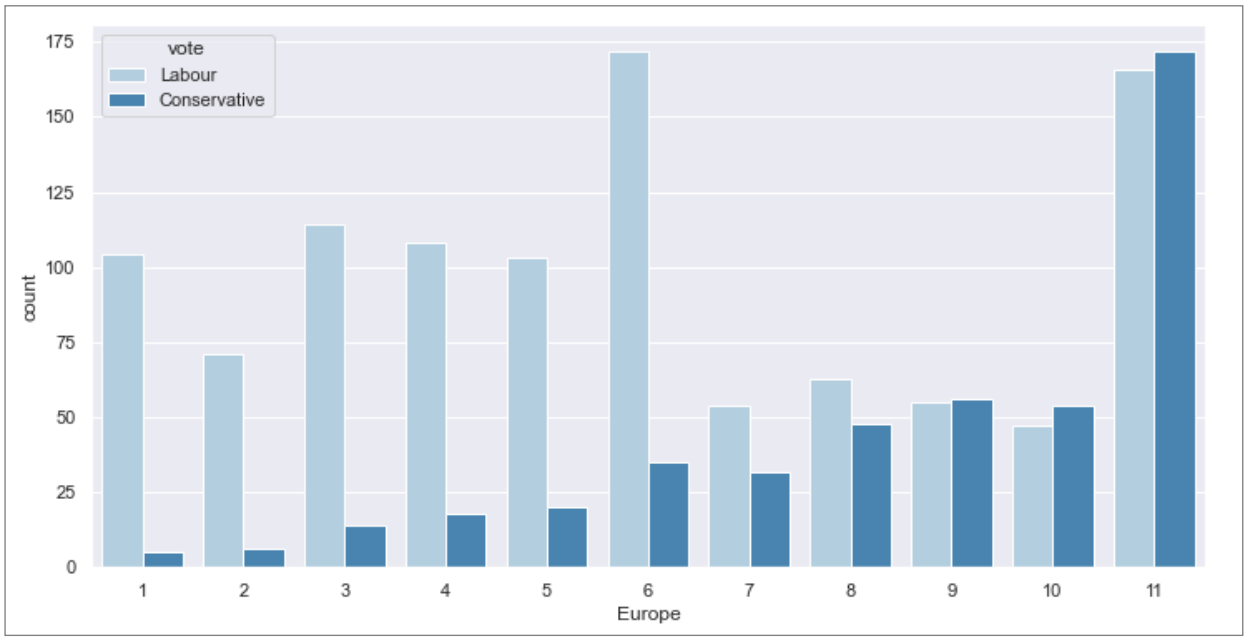

Figure 19. Count plot of Europe with vote as hue.

- We can observe that as respondents with strong negative attitude toward European integration are more likely to vote Conservative party and respondents who are likely to vote Labour party seems to have even distribution.
- As the scores towards 'Eurosceptic' sentiment scores get higher, we can see the trend in graph of linearly increase in respondents voting Conservative party.

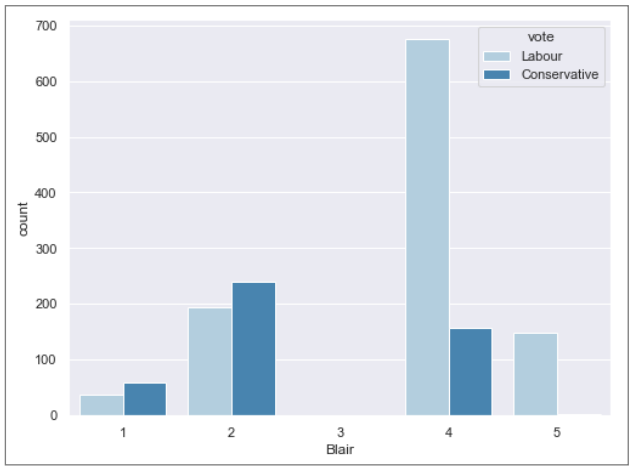

Figure 20. Count plot of Blair with vote as hue.

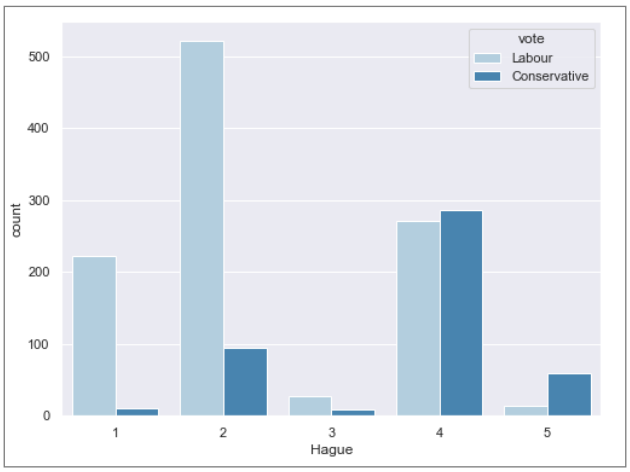

Figure 21. Count plot of Hague with vote as hue.

- For 'Blair' of the Labour Party most surveyed voters gave a rating of 4 and above and for 'Hague' of the Conservative party most voters gave a rating of 2 or below.
- One plausible explanation we can assume is, the voters voting Blair a higher rating might have rated Hague with lower ratings. The rating of 4 which is around 500 for Blair is almost equal to the rating of 2 for Hague.

Stacked Bar graph of Political knowledge and Europe:

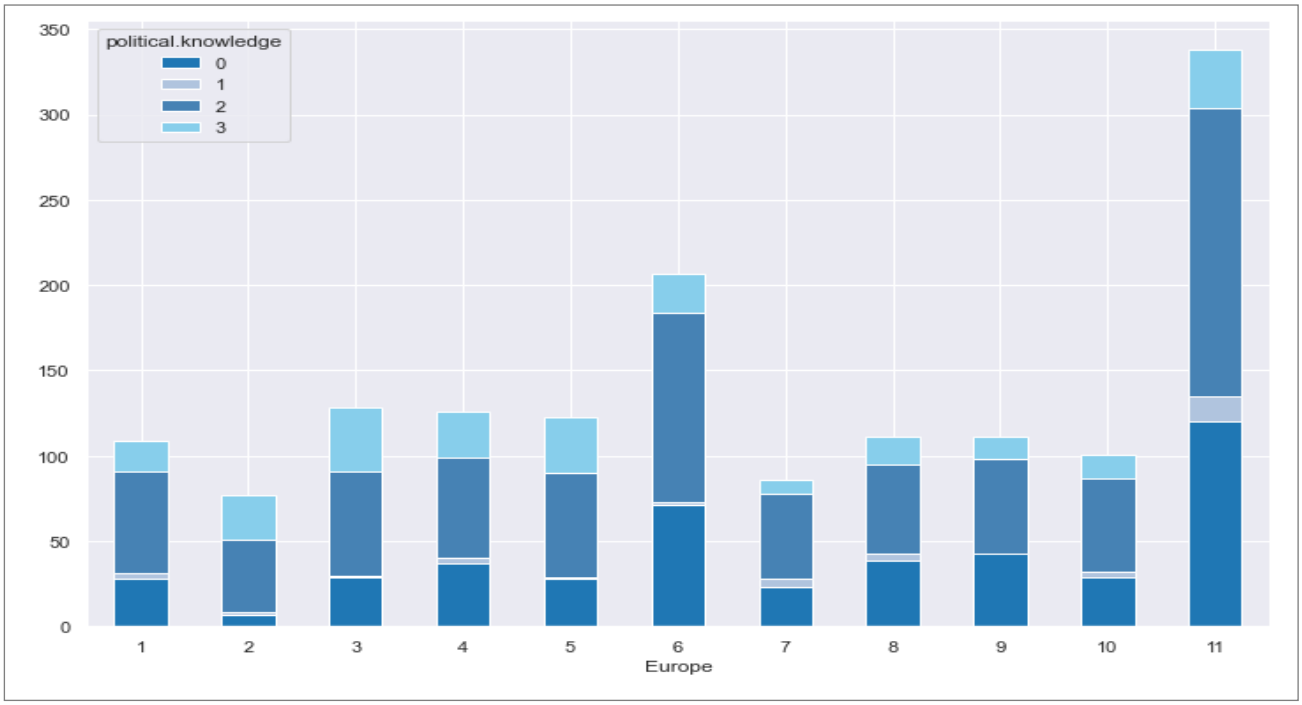

Figure 22.Stacked Bar graph of Political knowledge and Europe

- Respondents either seem to have a fair knowledge of the parties' position on the European Integration or no idea at all.
- We can see that, around 50-60% of the people with knowledge of level 2 of parties position on European integration is almost equally distributed in each level of Eurosceptic' sentiment scale.

Pair plot:

A pair plot helps to understand the distribution of the variables across classes. In the given dataset, most of the values are discrete ordinal values except age which is continuous in nature. Hence the scatter plot will depict points in straight lines

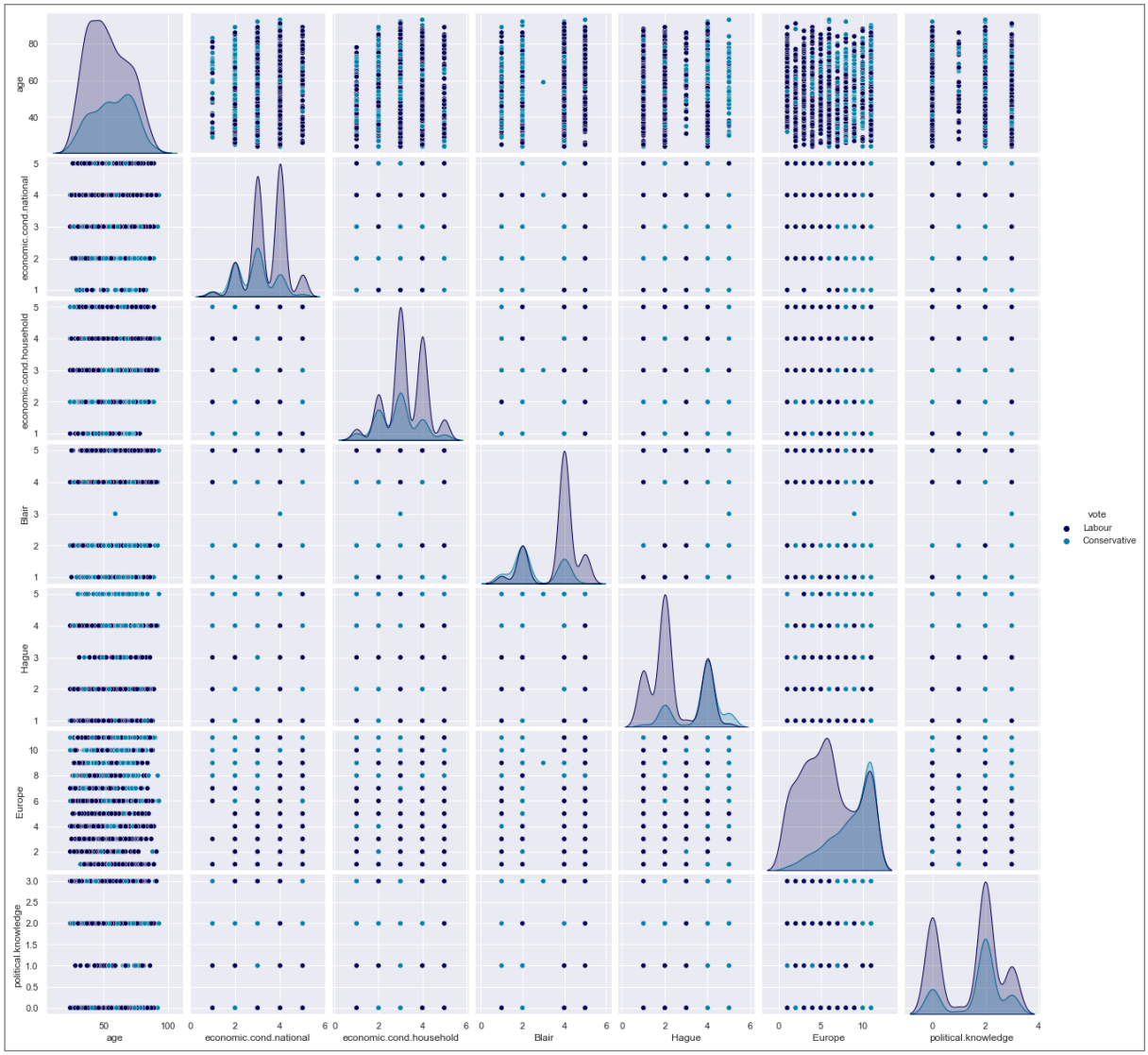

Figure 23. Pair plot for Problem 1.

Observations:

- The density distribution curve changes for Labour Party and Conservative Party were detected for all variables, suggesting that these variables are predicted to be beneficial in the categorization of the target variable vote.
- There is no defined pattern in the above graphs depicting the relation between the variables.
- Ratings of 0, 2 & 3 on Knowledge of parties' positions on European integration has not been influenced by different age groups.
- The Eurosceptic sentiments have spread across the complete spectrum of age groups.
- Participants Eurosceptic sentiment has not influenced their assessments on national and household economic conditions.

Correlation Matrix:

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
age	1.000000	0.018687	-0.038868	0.032084	0.031144	0.064562	-0.046598
economic.cond.national	0.018687	1.000000	0.347687	0.326141	-0.200790	-0.209150	-0.023510
economic.cond.household	-0.038868	0.347687	1.000000	0.215822	-0.100392	-0.112897	-0.038528
Blair	0.032084	0.326141	0.215822	1.000000	-0.243508	-0.295944	-0.021299
Hague	0.031144	-0.200790	-0.100392	-0.243508	1.000000	0.285738	-0.029906
Europe	0.064562	-0.209150	-0.112897	-0.295944	0.285738	1.000000	-0.151197
political.knowledge	-0.046598	-0.023510	-0.038528	-0.021299	-0.029906	-0.151197	1.000000

Figure 24. Correlation matrix of Problem 1.

Heatmap:

A heatmap gives us the correlation between numerical variables. If the correlation value is tending to 1, the variables are highly positively correlated whereas if the correlation value is close to 0, the variables are not correlated. Also, if the value is negative, the correlation is negative. That means, higher the value of one variable, the lower is the value of another variable and vice-versa.

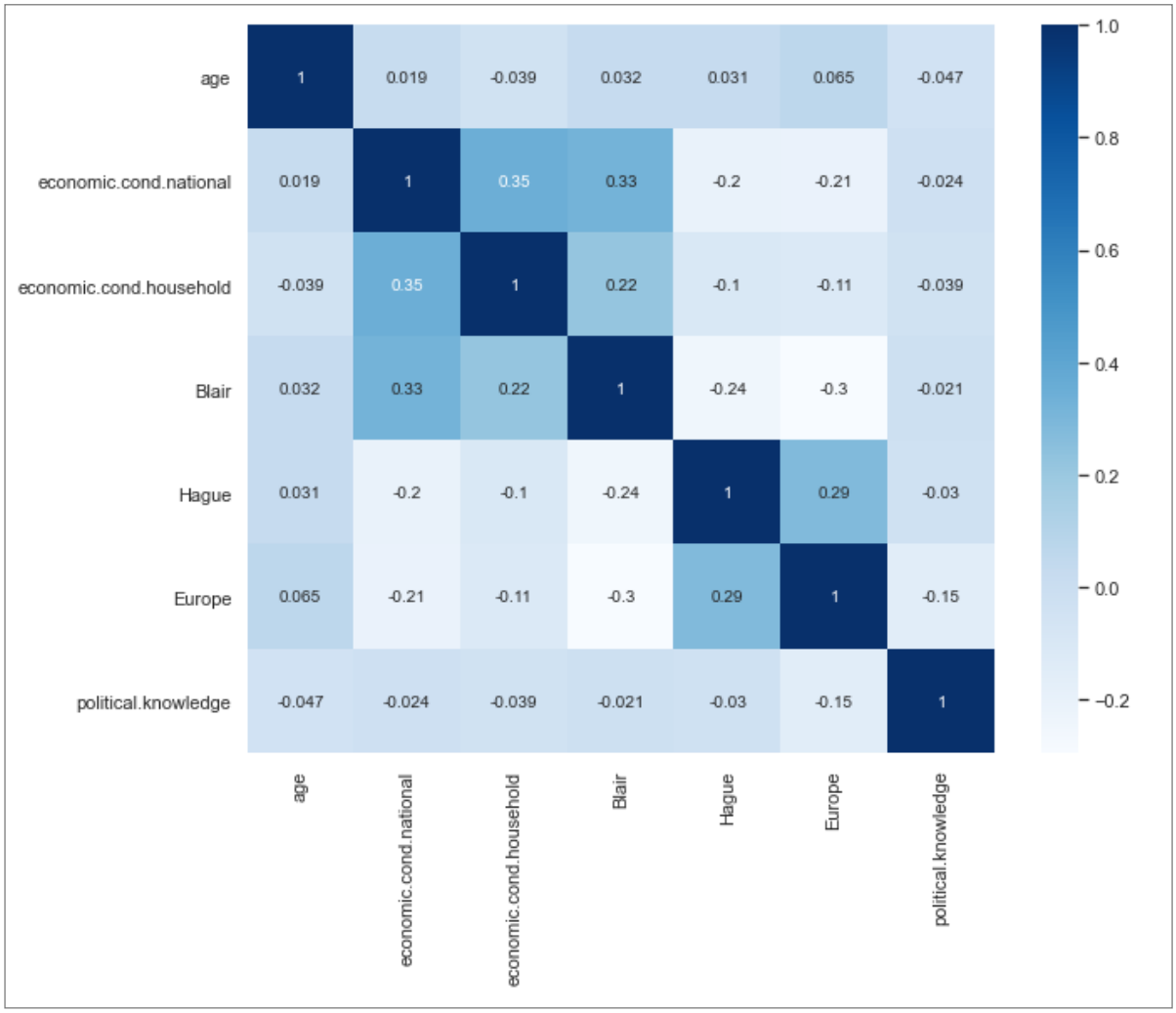

Figure 25. Heatmap of Problem 1.

Observations:

- Negative Correlation is an indication that mentioned variables move in the opposite direction people who are voting for Blair are obviously not voting for Hague. Hence there is a negative correlation between the two indicating cause and effect relationship between the variables.
- Ratings of household economic condition national economic condition have some significant positive correlation of 0.35.
- Respondents giving high rating to conservative party have certain extent of correlation with Europe with positive correlation of 0.29. That means voters who are against Europe integration with high score of Eurosceptic' sentiment is more likely to vote Hague.
- Participants giving high rating to national economic condition are supporters of labour party with positive correlation of 0.33
- There is negative correlation between age and & political knowledge.
- All other variables have a weak correlation. Hence, there is no strong multicollinearity among variables

Checking for Outliers:

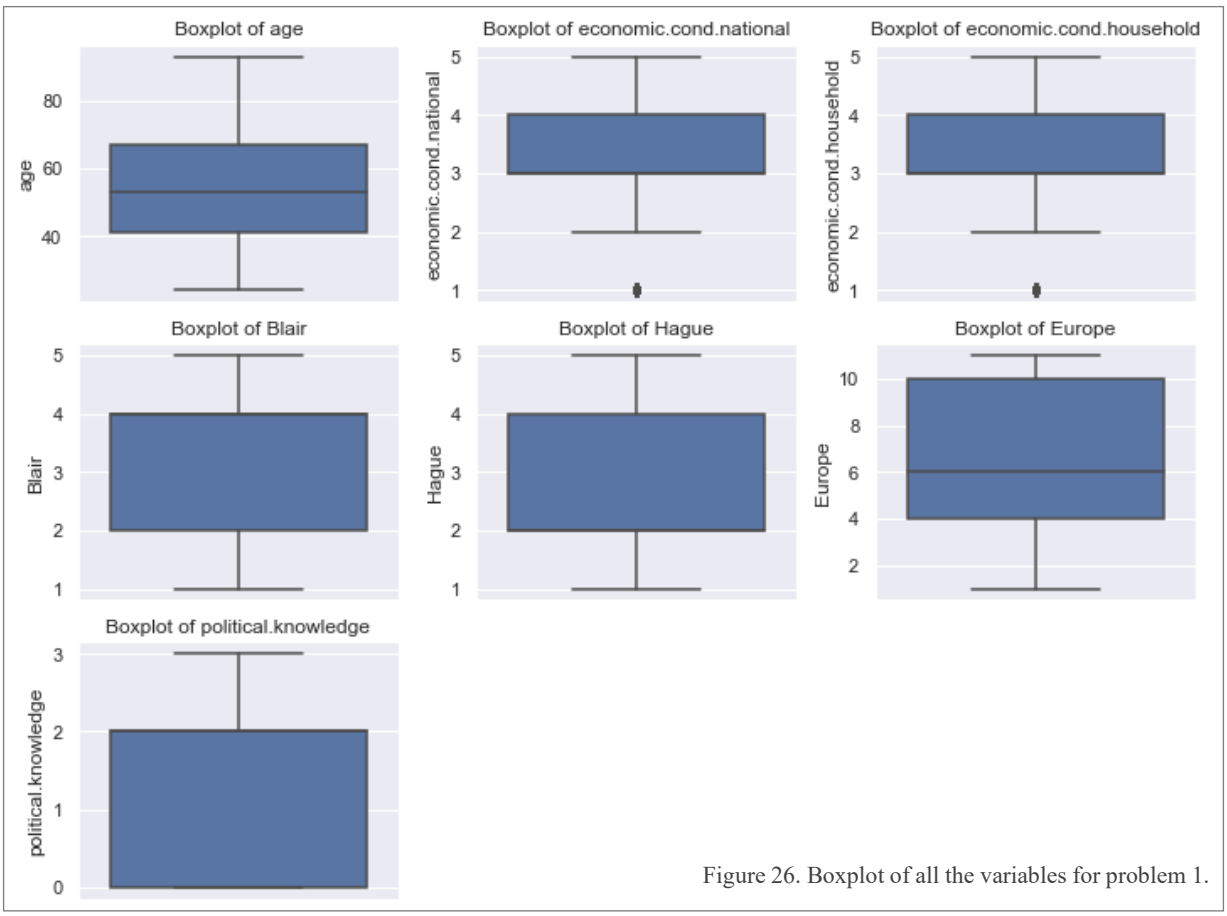

Figure 26. Boxplot of all the variables for problem 1.

Outliers are present in economic condition household and national other than that none of the variables have any outliers, since 'Economic Condition Household' and 'Economic condition National' are of ordinal type, i.e. they follow a certain order or degree of magnitude and the outliers are treated only for continuous variables and not the ordinal categorical variables.

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 the categorical variables:

Encoding is a process of converting categorical data variables so they can be provided to machine learning algorithms for predictions. Encoding is a crucial part of feature engineering for machine learning.

Since 'vote' and 'gender' are categorical data, we encode the data using categorical coding. Rest of the variables are already in ordinal integer data type. The following table shows data which take values 0 and 1 for column 'Vote' and 'Gender'.

Checking the head of Dataset after encoding the Categorical variables:

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

Figure 27. Encoded data for model building.

We can see that the count of class 0 and class 1 in vote is 460 and 1057 respectively, where as in Gender, female voters (808) are more in number than male voters (709).

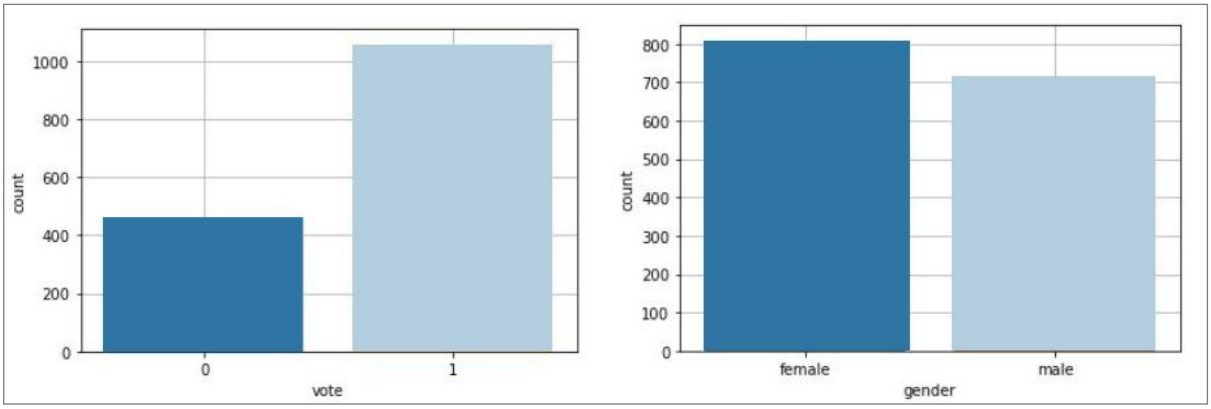

Figure 28. Count plot of vote and gender.

Vote (Target variable)			
Class 0	Conservative party	460	30.323%
Class 1	Labour party	1057	69.677%

Table 4. Proportion and class details of 'vote' variable after encoding.

Gender			
Class 0	Female voter	808	53.263%
Class 1	Male voter	709	46.737%

Table 3.Proportion and class details of 'gender' variable after encoding.

The object data types are converted to numerical. Now the data types of all the variables are in acceptable format for Modelling.

Checking the data types of variables:

```
age                        int64
economic.cond.national     int32
economic.cond.household    int32
Blair                      int32
Hague                      int32
Europe                     int32
political.knowledge        int32
vote                        int8
gender                      int8
```

Figure 29. Datatypes after Encoding data.

Now the dataset is cleaned, encoded and ready to use for model building. The multiple classification models are built and the models are compared to find best performing model.

Is Scaling necessary here or not?

We are building logistic Regression, LDA, KNN, Naïve bayes, Bagging and Boosting models in our project.

Logistic Regression and Linear Discriminant Analysis (LDA) finds its coefficients using the variation between the classes so the scaling doesn't matter. Hence, Scaling is not required while modelling trees, Naive Bayes are by designed to equipped which gives weights to the features accordingly. Performing a feature scaling in these algorithms may not have much effect.

On the other hand, KNN requires scaling of data because KNN uses the Euclidean distance between two data points to find nearest neighbours. Distance & Gradient descent algorithms are sensitive to magnitudes. The features with high magnitudes will weigh more than features with low magnitudes. Role of Scaling is mostly important these algorithms.

Random Forest is a tree-based model and hence does not require feature scaling. This algorithm requires partitioning, even if you apply Normalization then also the result would be the same.

Most of the variables in the given dataset take ordinal values ranging from 0 to 11. However, the values in variable age range from 24 to 93. We thus notice that variables with very high/ low values cannot be compared for analysis and may have a high variance in modelling exercise which effects the KNN model, so we can copy the data and scale the dataset and save it on different variable, which can be used for KNN model building and for rest other models the unscaled dataset can be used.

Scaling the dataset for KNN model and Gradient Boosting:

Scaling the data using Standard Scaler and therefore normalise the values where the **means will be 0 and standard deviation 1.** Scaling of data is done using importing a package called StandardScaler from sklearn. pre-processing.

Plots prior to scaling and after scaling dataset:

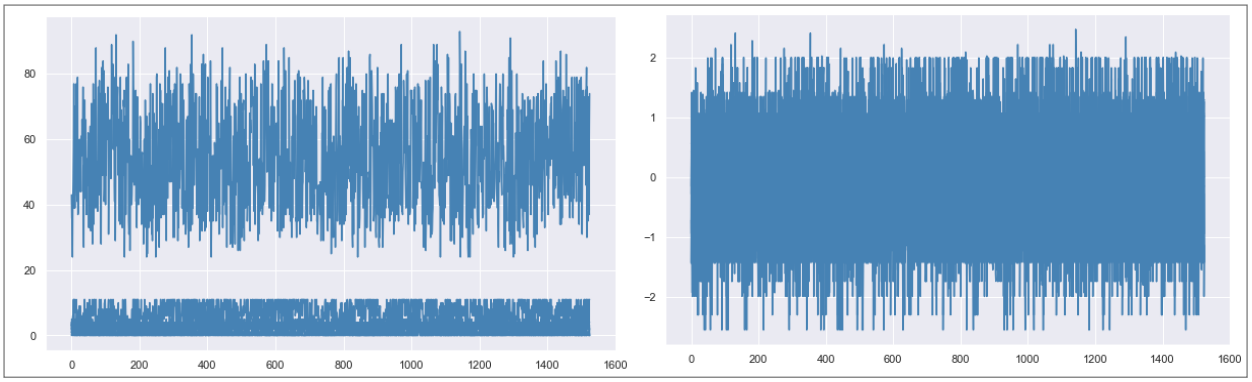

Figure 30. Plots to compare prior and after scaling data

Before Scaling:
Before scaling our dataset is distributed as shown in the figure above. Some variables are plotted higher as they must be having higher absolute values while some are near the zero line as their absolute values are lower than the other variables.

Even if the values for these variables are lower, they play significant role in the dataset, we cannot ignore them. Therefore, scaling is necessary.

For this dataset, we would be performing Standard Scaler function to scale the dataset. Standard Scaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance.

After Scaling:
We can see that all variables are scaled now and the values are close to each other. If we now check the plot of the scaled dataset, we would find that all variables are distributed similar to each other and all variables would be significant.

We can see now that all variables are scaled to have a mean tending to 0 and standard deviation to one. Therefore, scaling is very important for this dataset.

Checking the output of scaled dataset:

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	-0.716161	-0.278185	-0.148020	0.565802	-1.419969	-1.437338	0.423832	-0.936736
1	-1.162118	0.856242	0.926367	0.565802	1.014951	-0.527684	0.423832	1.067536
2	-1.225827	0.856242	0.926367	1.417312	-0.608329	-1.134120	0.423832	1.067536
3	-1.926617	0.856242	-1.222408	-1.137217	-1.419969	-0.830902	-1.421084	-0.936736
4	-0.843577	-1.412613	-1.222408	-1.988727	-1.419969	-0.224465	0.423832	1.067536

Figure 31. Output of scaled dataset for KNN.

Here, we are building a model, to predict which party a voter will vote for on the basis of the given information and to create an exit poll that will help in predicting overall win and seats covered by a particular party. In order to do our analysis, we are expected to build model using Logistic Regression, LDA, KNN Model and Naïve Bayes Model. For now, we are not scaling the data and will do the scaling based on the models we will run ahead. Hence, as mentioned scaling might be necessary for two models and might not be necessary for the other two.

Data Split: Split the data into train and test (70:30).

1. Capture the target column into separate vectors for training set and test set.
We split the data into train and test set in the ratio 70:30 where 70% of our data (i.e., 1061 observations) will be used for training purposes and 30% (i.e., 456 observations) will be used for testing purposes.

X = all independent variables [‘age', 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political. Knowledge', 'gender']

Y = dependent variable [‘vote’]

Splitting the dataset in to train and test in the ratio of 70:30 using train test split from sklearn, keeping the random state as 1.
After splitting the X and y into 70:30, we can check the number of observations each one takes:

- X_train: (1061, 8)
- X_test: (456, 8)
- y_train: (1061,)
- y_test: (456,)

Total Observations: 1517
The data is now read to fit the models on train and check the performance of test data. The data is divided in 70% of train and 30% of test

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

Logistic Regression Model:

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. The classification algorithm Logistic Regression is used to predict the likelihood of a categorical dependent variable.

To build a Logistic Regression model:

- Fitting the Logistic Regression model which is imported from Sklearn linear model with solver 'liblinear'.
- Predicting on Training and Testing dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.
- **Model evaluation through Accuracy, Confusion Matrix, Classification report, AUC, ROC curve.**

Classification report:

	precision	recall	f1-score	support
0	0.74	0.66	0.70	322
1	0.86	0.90	0.88	739
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.82	0.83	0.82	1061

	precision	recall	f1-score	support
0	0.81	0.69	0.75	138
1	0.87	0.93	0.90	318
accuracy			0.86	456
macro avg	0.84	0.81	0.82	456
weighted avg	0.85	0.86	0.85	456

Figure 32. Classification report of Logistic Regression model of train (left) & test (right).

Confusion Matrix for training and testing data:

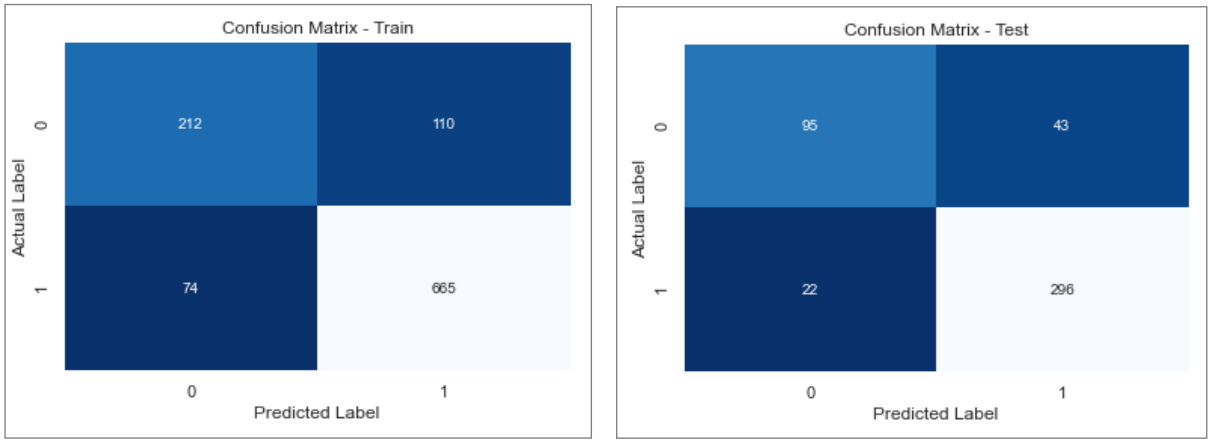

Figure 33. Confusion matrix of Logistic Regression model of train (left) & test (right).

ROC Curve and ROC_AUC score:

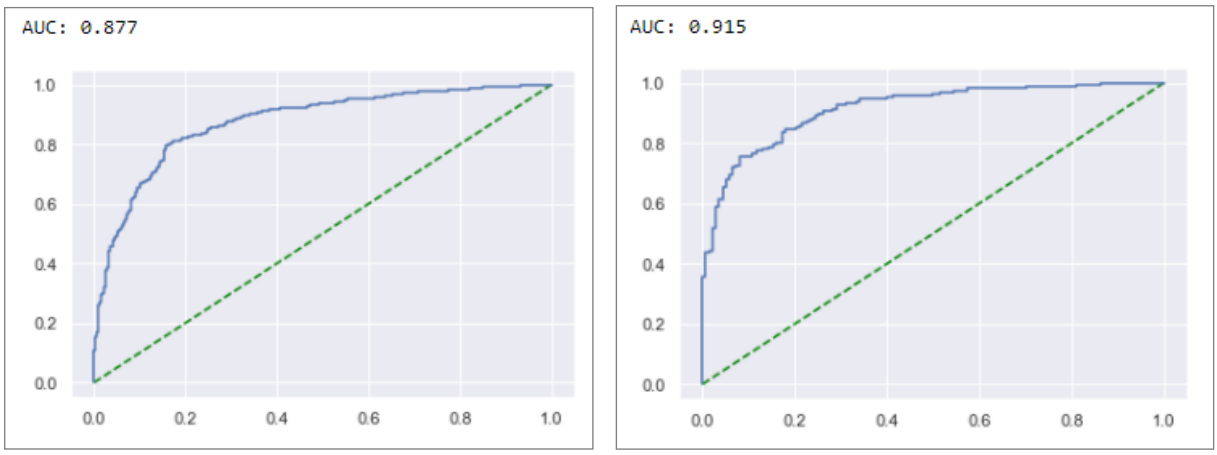

Figure 34. ROC curve and AUC score of Logistic Regression model of train (left) & test (right).

Logistic Regression Model					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	665		296	
2.	True Negative	212		95	
3.	False Positive	110		43	
4.	False Negative	74		22	
5.	AUC score	87.7%		91.5%%	
6.	Accuracy	83%		86%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	74%	86%	81%	87%
8.	Recall	66%	90%	69%	93%
9.	F1 score	70%	88%	75%	90%

Table 5. Model performance for logistic regression model.

Inferences:

- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Model performance metrics i.e., Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, and that the model is a good classification model overall. The test is slightly performing better than train.
- **Overall, the metrics are good fit.**
- Further the model improved using feature engineering, hyper parameter tuning (including combination of various parameters).

Linear discriminant analysis model:

Linear Discriminant Analysis uses linear combination of independent variables to predict the class in the response variable of a given observation. The prediction is made simply by the use of Bayes' Theorem which estimated the probability of the output class given the input. It also makes use of the probability of each class and also the data belonging to the class. The class which has the highest probability is considered as the output class and the model makes the prediction. The LDA model is built using the sklearn. discriminant analysis package and then fit in the training data. Using this fitted model, the predictions are made on the testing data.

LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis. On the train data set, we fit our Linear Discriminant model. By default, LDA uses a cut-off probability of 0.5. So, initially, we'll create our LDA model with a default probability of 0.5 and see how it performs, then we'll see how it performs with multiple cut-off probabilities to see which one performs the best.

To build a Linear discriminant analysis model:

- Fitting the linear discriminant analysis model from Sklearn discriminant analysis.
- Predicting on Training and Testing dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.

Model performance of LDA:

Classification report:

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.72	0.67	0.70	322
1	0.86	0.89	0.87	739
accuracy			0.82	1061
macro avg	0.79	0.78	0.79	1061
weighted avg	0.82	0.82	0.82	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.80	0.69	0.74	138
1	0.87	0.92	0.90	318
accuracy			0.85	456
macro avg	0.84	0.81	0.82	456
weighted avg	0.85	0.85	0.85	456

Figure 35. Classification report for LDA model

Confusion Matrix for training and testing data:

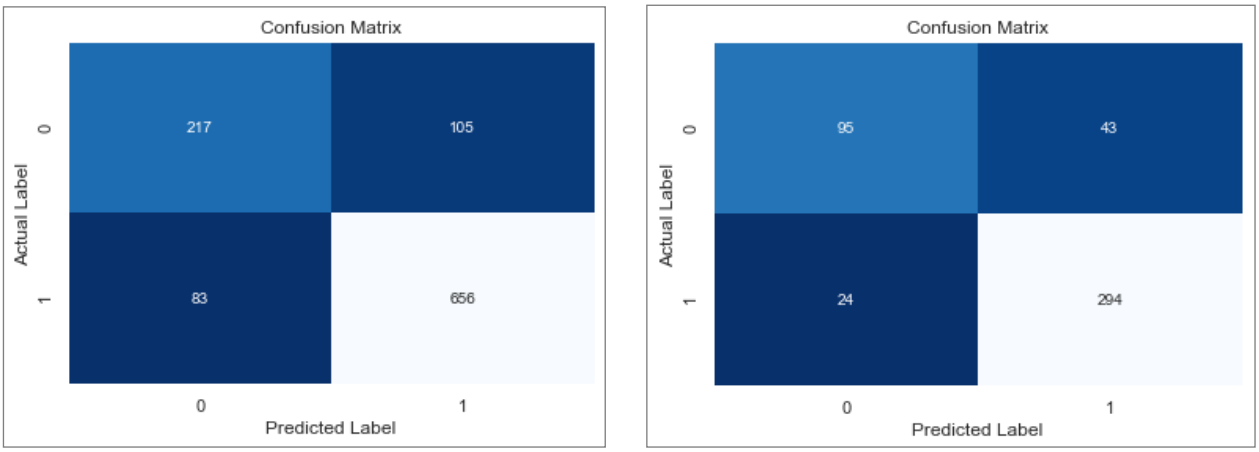

Figure 36. Confusion matrix of LDA model of train (left) & test (right).

ROC Curve and ROC_AUC score:

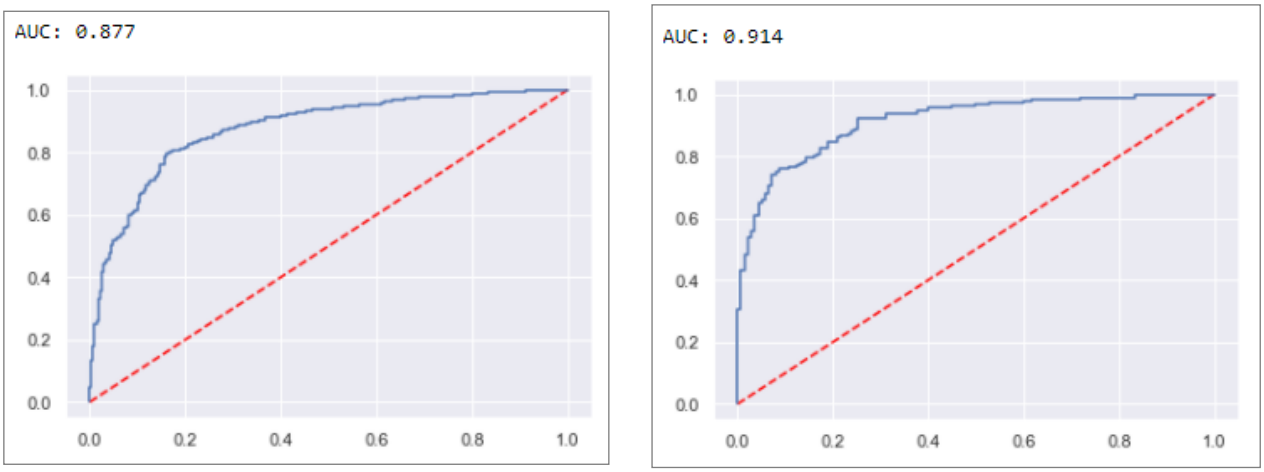

Figure 37. ROC curve and AUC score of LDA model of train (left) & test (right).

Linear Discriminant Analysis Model					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	656		294	
2.	True Negative	217		95	
3.	False Positive	105		43	
4.	False Negative	83		24	
5.	AUC score	87.7%		91.4%	
6.	Accuracy	82%		85%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	72%	86%	80%	87%
8.	Recall	67%	89%	69%	92%
9.	F1 score	70%	87%	74%	90%

Table 6. Model performance for LDA model.

Inferences:

- Logistic Regression is performing slightly better than LDA model.
- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Model performance metrics i.e., Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
- **Overall, the metrics are good fit.**
- Further the model improved using feature engineering, hyper parameter tuning including combination of various parameters and changing the custom probability for classification where all the metrics are better.

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

K-Nearest Neighbors Model:

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm. KNN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. Generally, good KNN performance usually requires pre-processing of data to make all variables similarly scaled and cantered.

To build a K-Nearest Neighbors model:

- Scaled dataset is used to build KNN model, as it is distance-based algorithm.
- Fitting the KNN model which is imported from Sklearn neighbors model which considers default n_neighbors (k=5)
- Predicting on Training and Testing scaled dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.
- Checking the Model performance of train and test data.

Model performance of KNN:

Classification report:

	precision	recall	f1-score	support
0	0.77	0.71	0.74	307
1	0.89	0.92	0.90	754
accuracy			0.86	1061
macro avg	0.83	0.81	0.82	1061
weighted avg	0.85	0.86	0.85	1061

	precision	recall	f1-score	support
0	0.77	0.69	0.72	153
1	0.85	0.89	0.87	303
accuracy			0.82	456
macro avg	0.81	0.79	0.80	456
weighted avg	0.82	0.82	0.82	456

Figure 38. Classification report of KNN model of train (left) & test (right).

Confusion Matrix for training and testing data:

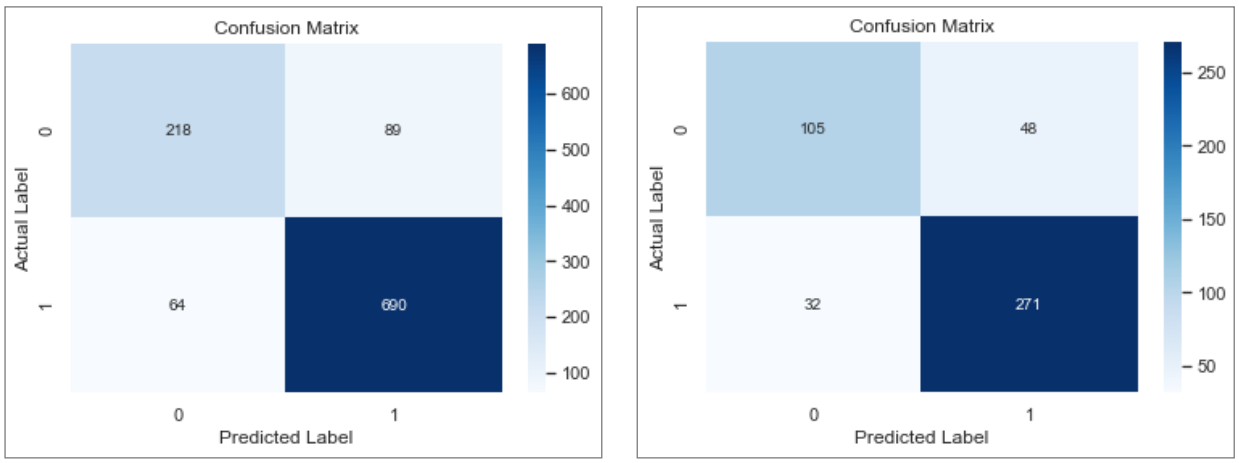

Figure 39. Confusion matrix of KNN model of train (left) & test (right).

ROC Curve and ROC_AUC score:

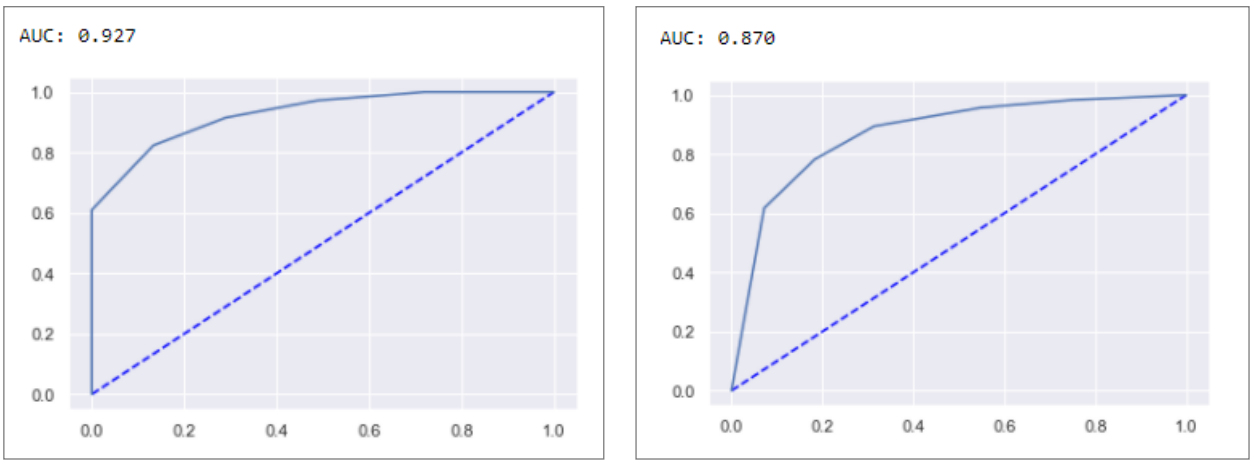

Figure 40. ROC curve and AUC score of KNN model of train (left) & test (right).

KNN Model					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	690		271	
2.	True Negative	218		105	
3.	False Positive	89		48	
4.	False Negative	64		32	
5.	AUC score	92.7%		87%	
6.	Accuracy	86%		82%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	77%	89%	77%	85%
8.	Recall	71%	92%	69%	89%
9.	F1 score	74%	92%	72%	87%

Table 7. Model performance for KNN model.

Inferences:

- Logistic Regression and LDA is performing slightly better than KNN model.
- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Model performance metrics i.e., Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, and that the model is a good classification model overall.
- **Overall, the metrics are good fit.**
- Further the model improved by finding the model performance for different K-values and plot the graph to check at which K value the mis classification is least.

Naïve Bayes Model

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.

The algorithm while calculating likelihoods of numerical features it assumes the feature to be normally distributed and then we calculate probability using mean and variance of that feature only and also it assumes that all the predictors are independent to each other. Scale doesn't matter. Performing a feature scaling in this algorithm may not have much effect.

In a supervised learning situation, Naive Bayes Classifiers are trained very efficiently. Naive Bayed classifiers need a small training data to estimate the parameters needed for classification. Naive Bayes Classifiers have simple design and implementation and they can applied to many real life situations. **Gaussian Naive Bayes** is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.

To build a Naïve Bayes model:

- Fitting the Gaussian Naïve Bayes model which is imported from Sklearn naïve bayes.
- Predicting on Training and Testing scaled dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.
- Checking the Model performance of train and test data.

Model performance of Naïve Bayes:

Classification report:

	precision	recall	f1-score	support
0	0.70	0.70	0.70	322
1	0.87	0.87	0.87	739
accuracy			0.82	1061
macro avg	0.79	0.79	0.79	1061
weighted avg	0.82	0.82	0.82	1061

	precision	recall	f1-score	support
0	0.79	0.72	0.75	138
1	0.88	0.92	0.90	318
accuracy			0.86	456
macro avg	0.84	0.82	0.83	456
weighted avg	0.86	0.86	0.86	456

Figure 41.Classification report Naive Bayes model of train (left) & test (right).

Confusion Matrix for training and testing data:

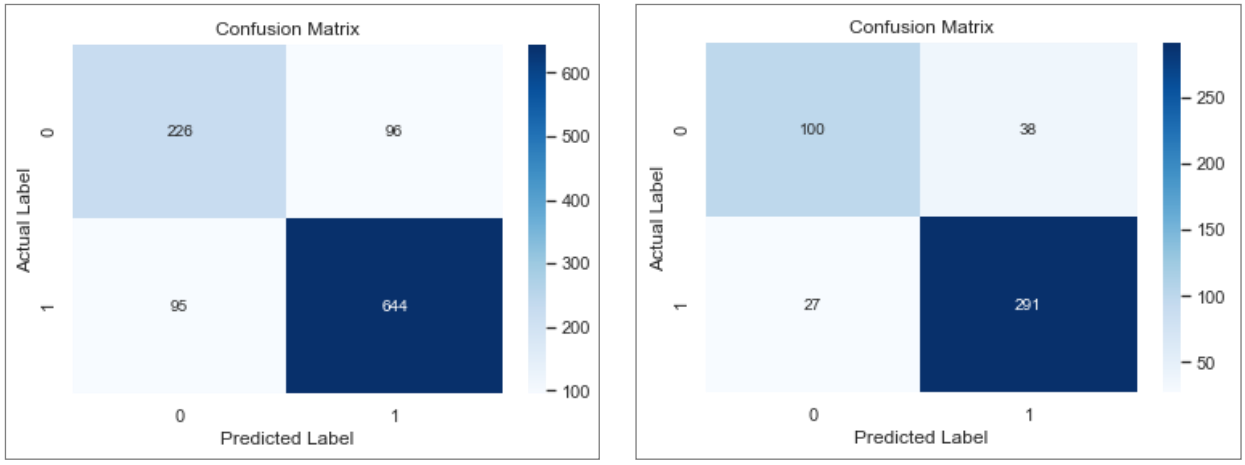

Figure 42. Confusion matrix of Naive Bayes model of train (left) & test (right).

ROC Curve and ROC_AUC score:

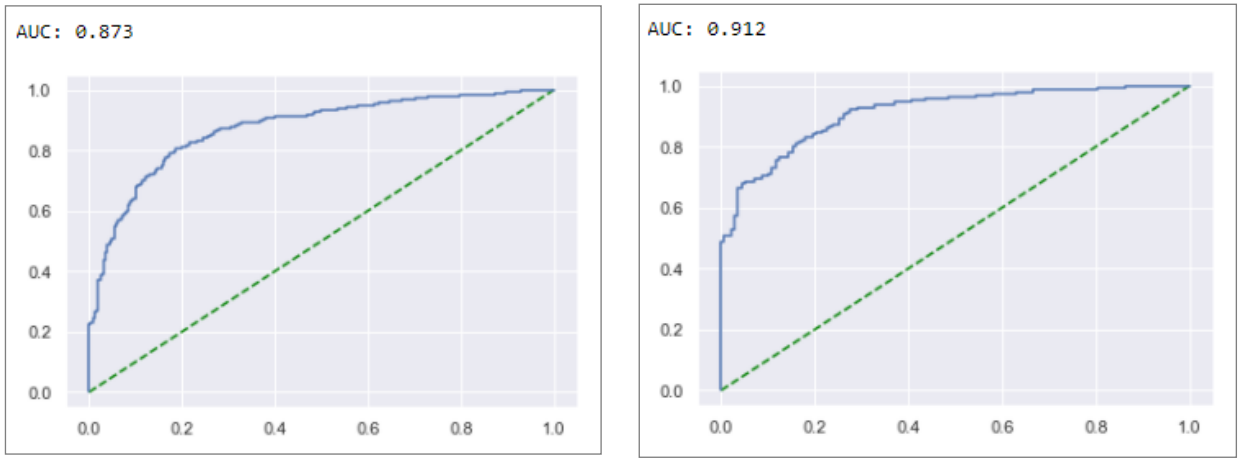

Figure 43. ROC curve and AUC score of Naive Bayes model of train (left) & test (right).

Naïve Bayes Model					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	644		291	
2.	True Negative	226		100	
3.	False Positive	96		38	
4.	False Negative	95		27	
5.	AUC score	87.3%		91.2%	
6.	Accuracy	82%		86%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	70%	87%	79%	88%
8.	Recall	70%	87%	72%	92%
9.	F1 score	70%	87%	75%	90%

Table 8. Model performance for Naive Bayes model.

Inferences:

- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Model performance metrics i.e., Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
- **Overall, the metrics are good fit.**
- Further the model improved using feature engineering, hyper parameter tuning including combination of various parameters or technique of SMOTE and check the performance.

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

Model Tuning of Logistic Regression Model:

Initially, we fit the train data and labels in the Logistic Regression model, based on the model performance the model is tunned using Grid search, the best parameters are used and the model is re-built and model performance is calculated which includes Classification report of accuracy, recall, precision and F1 score for both train and test data.

Grid Search: Grid search divides the hyperparameter domain into distinct grids. Then, using cross-validation, it attempts every possible combination of values in this grid, computing some performance measures. The ideal combination of values for the hyperparameters is the point on the grid that maximizes the average value in cross-validation. Grid search is a comprehensive technique that considers all possible combinations in order to locate the best point in the domain.

Hyperparameter Tuning:

- ***'penalty':*** *['l2', 'none', 'l1', 'elasticnet'],*
- ***'solver'****: ['liblinear' ,'lbfgs', 'newton-cg'],*
- ***'tol':*** *[0.0001, 0.00001],*
- ***'Max_iter':*** *[10000, 5000, 15000]*
- ***Cross validation (cv)****:3*
- ***Scoring****: 'f1'*

Penalized logistic regression imposes a penalty to the logistic model for having too many variables. This results in shrinking the coefficients of the less contribute variables toward zero. This is also known as regularization. In our grid search, we take 'L2', 'none', 'L1' and 'elasticnet' as our arguments and check which is preferred by grid search.

The solver is the process that runs for the optimization of the weights in the model. The solver uses a Coordinate Descent (CD) algorithm that solves optimization problems by successively performing approximate minimization along coordinate directions or coordinate hyperplanes. Different solvers take a different approach to get the best fit model. In our case, we have taken 'lbfgs', 'liblinear' and 'newton-cg' as our arguments. We will check which is preferred by grid search.

Tol is the tolerance of optimization. When the training loss is not improved by at least the given tol on consecutive iterations, convergence is considered to be reached and the training stops. We will be checking for tolerance of 0.0001 and 0.00001.

The logistic regression uses an iterative maximum likelihood algorithm to fit the data. There are no set criteria for **maximum iterations**. The solver will run the model till it reaches convergence or till the max iterations, you have provided. In this case, we have given 5000, 10000 and 15000 as inputs. We will see which fits better.

We have taken cross-validation as 3 and scoring as F1 for our grid search.

The final best parameters are:

- ***Max_iter*** *is '10000'*
- ***Penalty*** *is 'l1'*
- ***Solver*** *used is 'liblinear'*
- ***Tol*** *is 0.0001*

Our new model, which is based on the grid search algorithm's best parameters and the model's performance is tested using these parameters is then saved in a distinct variable as best_model. This model is used to predict the values of the target variable, and then the model's performance is evaluated using these parameters.

Checking the Coefficients:

- The coefficient for age is -0.011538787749405897
- The coefficient for economic.cond.national is 0.36832115035821333
- The coefficient for economic.cond.household is 0.0477179571139086
- The coefficient for Blair is 0.5594458430681535
- The coefficient for Hague is -0.8318050470115125
- The coefficient for Europe is -0.19960933084913768
- The coefficient for political.knowledge is -0.38121549749054096
- The coefficient for gender is 0.027541872719356368

The features 'Hague' and 'Blair' contributes largely in the model building. The assessment rating of the voters for the Conservative and Labour party. Followed by the current economic conditions and political knowledge. 'Age' is the least attribute affecting the model building.

Model performance of tuned Logistic Regression Model:

Classification report:

	precision	recall	f1-score	support
0	0.74	0.66	0.70	322
1	0.86	0.90	0.88	739
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.82	0.83	0.82	1061

	precision	recall	f1-score	support
0	0.82	0.68	0.74	138
1	0.87	0.93	0.90	318
accuracy			0.86	456
macro avg	0.84	0.81	0.82	456
weighted avg	0.85	0.86	0.85	456

Figure 44. Classification report of tuned Logistic Regression model of train (left) & test (right).

Confusion Matrix for training and testing data:

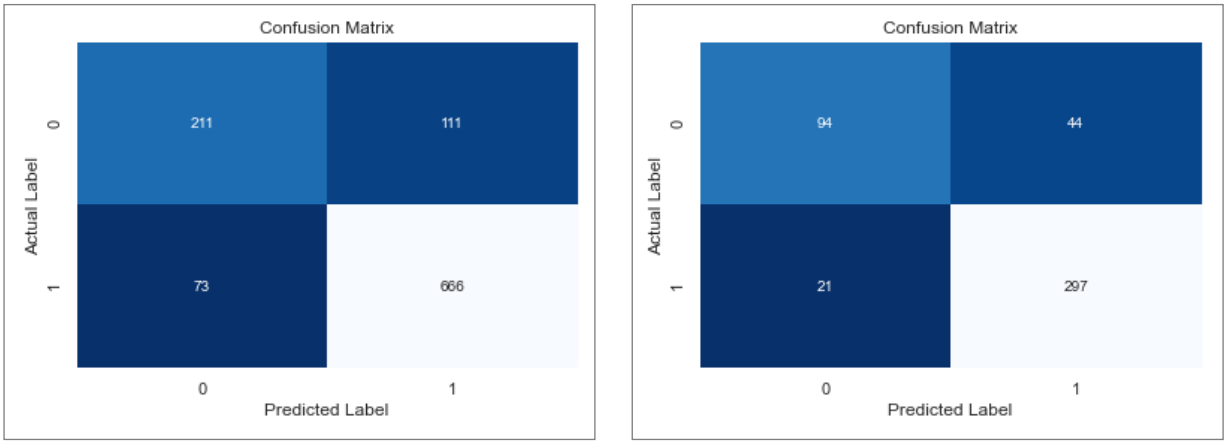

Figure 45. Confusion matrix of tuned Logistic Regression model of train (left) & test (right).

ROC Curve and ROC_AUC score:

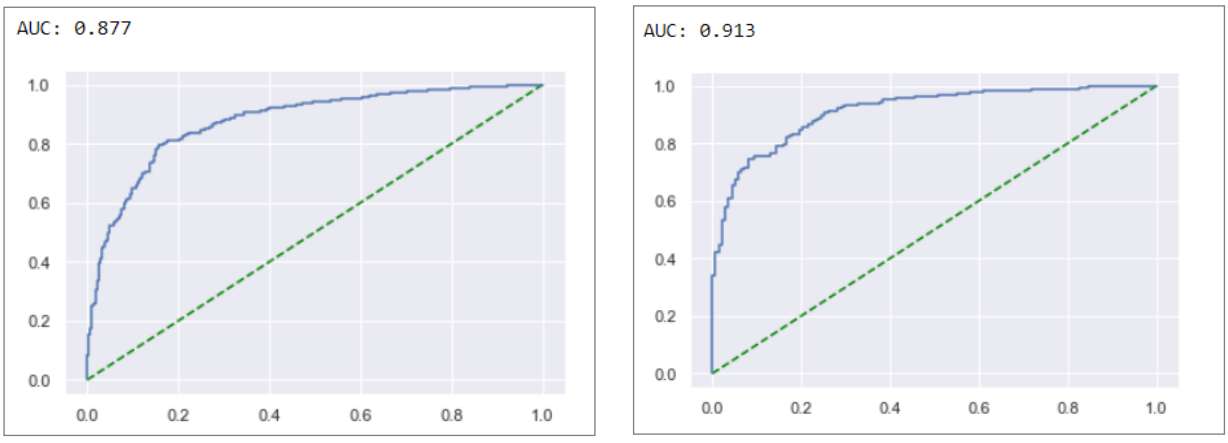

Figure 46. ROC curve and AUC score of tuned Logistic Regression model of train (left) & test (right).

Tuned Logistic Regression Model					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	666		297	
2.	True Negative	211		94	
3.	False Positive	111		44	
4.	False Negative	73		21	
5.	AUC score	87.7%		91.3%	
6.	Accuracy	83%		86%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	74%	86%	82%	87%
8.	Recall	66%	90%	68%	93%
9.	F1 score	70%	88%	74%	90%

Table 9. Model performance of tuned Logistic Regression model

Inferences:

- The model performance even after applying grid search with hyper parameters is almost similar to normal Logistic regression model. There is slight increase in performance in the conservative class.
- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Model performance metrics i.e., Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
- **Overall, the metrics are good fit.**

Model Tuning of Linear Discriminant Analysis Model:

Initially, we fit the train data and labels in the LDA model, based on the model performance the model is tunned using Grid search, the best parameters are used and the model is re-built and model performance is calculated.

Grid Search: Grid search divides the hyperparameter domain into distinct grids. Then, using cross-validation, it attempts every possible combination of values in this grid, computing some performance measures. The ideal combination of values for the hyperparameters is the point on the grid that maximizes the average value in cross-validation. Grid search is a comprehensive technique that considers all possible combinations in order to locate the best point in the domain. LDA grid search with hyperparameters taking the 'solver': ['svd' . 'lsqr', 'eigen'] there is no much difference in model performance.

Using custom probability cut-off technique for tunning LDA model:
We obtain an LDA model based on a default custom cut-off probability (i.e., 0.5). To get the best results, we'll need to test our model with several cut-off probabilities and choose the one that produces the greatest results. To do so, we'll start with probability 0.1 and work our way

up to 0.9 with a 1 interval, checking each probability recall and F1 score value along the way. We will use the likelihood that we will get the best recall and F1 score balance as our final probability value.

Cut off probability	Recall	F1 Score	Precision
0.1	0.9959	0.8307	0.7125
0.2	0.9783	0.8511	0.7531
0.3	0.9540	0.8608	0.7842
0.4	0.9296	0.8713	0.8198
0.5	0.8999	0.8785	0.8581
0.6	0.8566	0.8707	0.8853
0.7	0.8038	0.8559	0.9153
0.8	0.6901	0.7913	0.9273
0.9	0.5115	0.6661	0.9545

Table 10. LDA cut off probability performance table

From the above table we can see that the 'Recall' is decreasing from 0.1 to 0.9 and F1 score is increasing till 0.6 and dropping after that and the precision is increasing from 0.4. The cut off probability **0.4** provides the optimal balance of recall, precision and F1 score. As a result, we'll see the performance of 0.4 cut off probability metrics.

Model performance of LDA model with custom probability of 0.4:

Classification report:

Classification Report of the custom cut-off train data:

	precision	recall	f1-score	support
0	0.77	0.53	0.63	322
1	0.82	0.93	0.87	739
accuracy			0.81	1061
macro avg	0.79	0.73	0.75	1061
weighted avg	0.80	0.81	0.80	1061

Classification Report of the custom cut-off test data:

	precision	recall	f1-score	support
0	0.85	0.60	0.70	138
1	0.85	0.95	0.90	318
accuracy			0.85	456
macro avg	0.85	0.78	0.80	456
weighted avg	0.85	0.85	0.84	456

Figure 47. Classification report of tuned LDA model

Confusion Matrix for training and testing data:

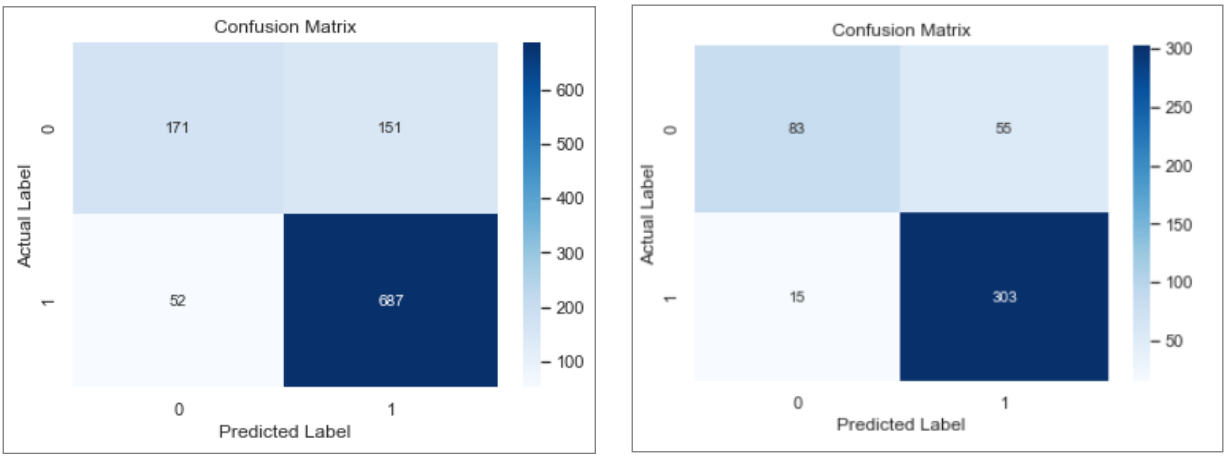

Figure 48. Confusion matrix of tuned LDA model of train (left) & test (right).

ROC Curve and ROC_AUC score:

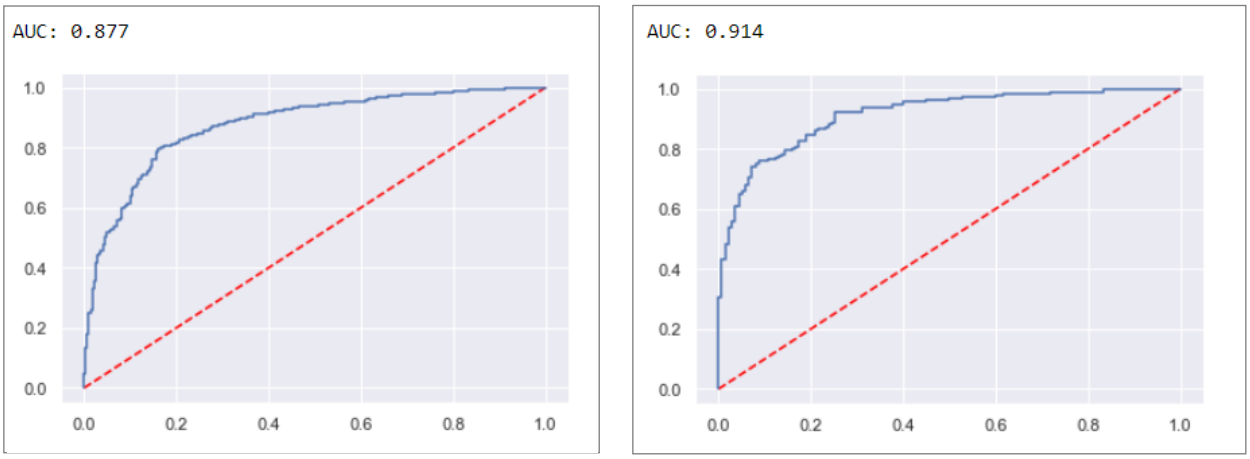

Figure 49. ROC curve and AUC score of tuned LDA model of train (left) & test (right).

LDA model with custom probability of 0.4					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	687		297	
2.	True Negative	171		94	
3.	False Positive	151		44	
4.	False Negative	52		21	
5.	AUC score	87.7%		91.3%	
6.	Accuracy	81%		85%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	77%	82%	85%	85%
8.	Recall	53%	93%	60%	95%
9.	F1 score	63%	87%	70%	90%

Table 11. Model performance of tuned LDA model.

Inferences:

- The model performance at custom probability of 0.4 is almost similar to normal LDA. However, this model increases slightly the performance metrics of conservative class compared to default model.
- In this model also it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- The recall and F1score metrics of training is overfitting compared to performance of test data.
- **Overall, the model is good fit.**

Model Tuning of KNN Model:

Initially, we fit the train data and labels in the KNN model which uses the default k = 5 to build the model. The model is re-built and model performance is calculated which includes

Classification report of accuracy, recall, precision and F1 score for both train and test data for which the miss classification of target class is least. We can say that lesser the mis-classification greater the performance of model.

Running a loop for K= 1 to 19 odd numbers and find MSE:
Run the KNN with no of neighbours to be 1,3,5..19 and Find the optimal number of neighbours from *K=1,3,5,7....19** using the Mis classification error

Hint: Misclassification error (MCE) = 1 - Test accuracy score. Calculated MCE for each model with neighbours = 1,3,5...19 and find the model with lowest MCE.

MCE = [1 - x for x in ac_scores]

```
[0.2171052631578947,
 0.1907894736842105,
 0.17543859649122806,
 0.18201754385964908,
 0.17763157894736847,
 0.17105263157894735,
 0.17763157894736847,
 0.17324561403508776,
 0.16666666666666663,
 0.16666666666666663]
```

Figure 50. MSE for odd K values from 1 to 9

Plot misclassification error vs k (with k value on X-axis) :

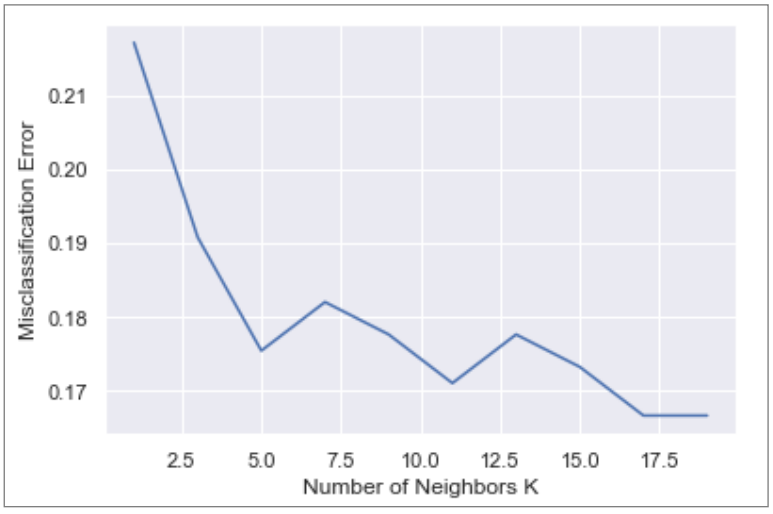

Figure 51. Plot of misclassification error vs K

From the above graph, we can see that from k= 17 the error is least and constant, however at k=11 the MSE is least with the higher balance of the Accuracy, Precision, Recall and F1 score.

To build a tuned K-Nearest Neighbors model at K=11:

- Scaled dataset is used to build KNN model, as it is distance-based algorithm.
- Fitting the KNN model which is imported from Sklearn neighbors model which considers with n_neighbors (k=11)
- Predicting on Training and Testing scaled dataset.

- Getting the Predicted Classes and Probabilities and creating a data frame.
- Checking the Model performance of train and test data in the next question.

Model performance of tuned KNN model at K = 11:

Classification report:

	precision	recall	f1-score	support
0	0.75	0.68	0.71	307
1	0.87	0.91	0.89	754
accuracy			0.84	1061
macro avg	0.81	0.79	0.80	1061
weighted avg	0.84	0.84	0.84	1061

	precision	recall	f1-score	support
0	0.78	0.69	0.73	153
1	0.85	0.90	0.88	303
accuracy			0.83	456
macro avg	0.81	0.79	0.80	456
weighted avg	0.83	0.83	0.83	456

Figure 52. Classification report of tuned KNN model of train (left) & test (right).

Confusion Matrix for training and testing data:

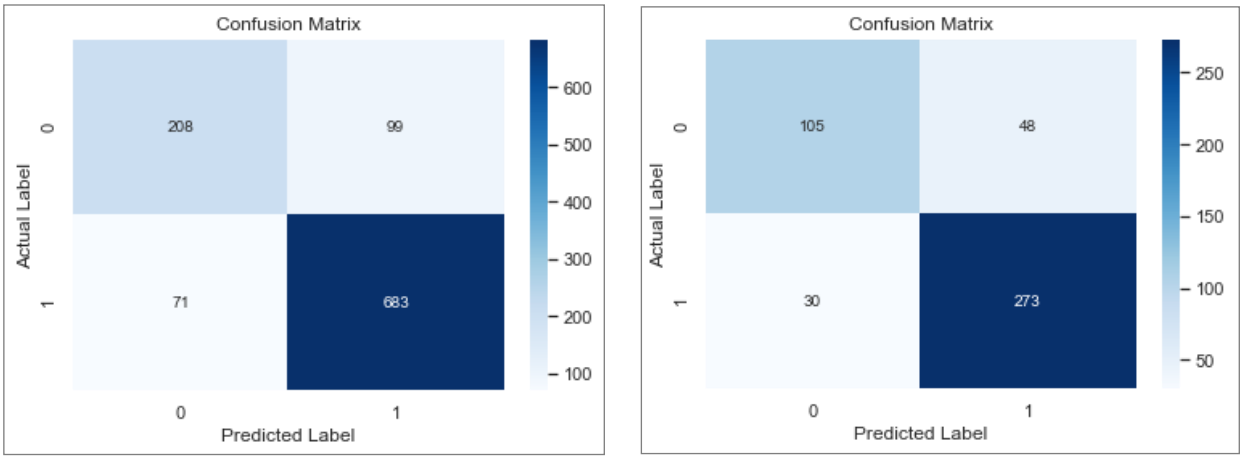

Figure 53. Confusion matrix of tuned KNN model of train (left) & test (right).

ROC Curve and ROC_AUC score:

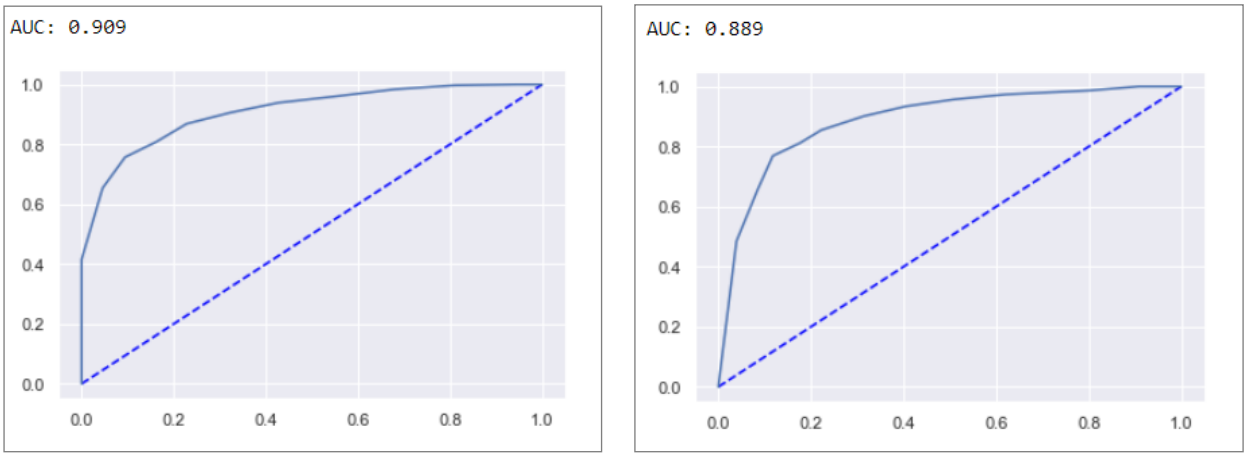

Figure 54. ROC curve and AUC score of tuned KNN model of train (left) & test (right).

Tuned KNN Model at k = 11					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	683		273	
2.	True Negative	208		105	
3.	False Positive	99		48	
4.	False Negative	71		30	
5.	AUC score	90.9%		88.9%	
6.	Accuracy	84%		83%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	75%	87%	78%	85%
8.	Recall	68%	91%	69%	90%
9.	F1 score	71%	89%	73%	88%

Table 12. Model performance of tuned KNN model

Inferences:

- The model performance at k = 11 is almost similar to default K=5. However, this model increases slightly the performance metrics of conservative class compared to default model.
- In this model also it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- The recall and F1score metrics of training is overfitting compared to performance of test data.
- **Overall, the model is good fit.**

Model Tuning of Naïve Bayes model with SMOTE technique:

For optimal Model performance we can apply SMOTE as a technique to remove class imbalance and check if the performance of the model improves for Naïve bayes model. SMOTE (Synthetic Minority Oversampling Technique) – Oversampling. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class.

To build a tuned Naïve bayes model with smote technique:

- Fitting the SMOTE which is imported from Sklearn imblearn sampling.
- The technique of SMOTE balances the minority class by replicating the samples.
- The balanced data is further fit into Gaussian naïve bayes model.
- Predicting on Training and Testing scaled dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.
- Checking the Model performance of train and test data in the next question.

Model performance of tuned Naïve Bayes with SMOTE technique.:

Classification report:

	precision	recall	f1-score	support
0	0.82	0.80	0.81	739
1	0.80	0.82	0.81	739
accuracy			0.81	1478
macro avg	0.81	0.81	0.81	1478
weighted avg	0.81	0.81	0.81	1478

	precision	recall	f1-score	support
0	0.67	0.80	0.73	138
1	0.90	0.83	0.87	318
accuracy			0.82	456
macro avg	0.79	0.82	0.80	456
weighted avg	0.83	0.82	0.83	456

Figure 55. Classification report of tuned Naive Bayes model of train (left) & test (right).

Confusion Matrix for training and testing data:

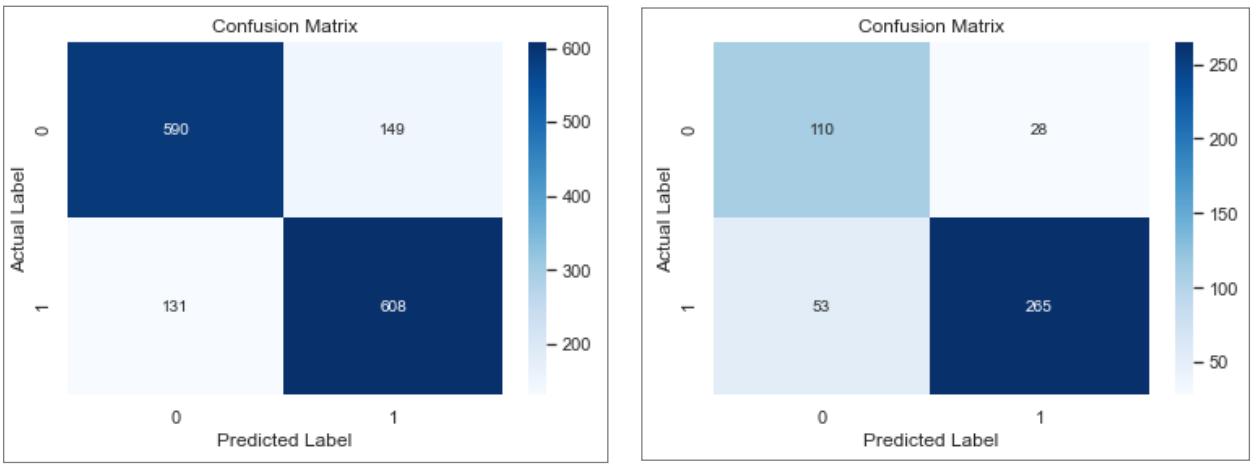

Figure 56. Confusion matrix of tuned Naive Bayes model of train (left) & test (right).

ROC Curve and ROC_AUC score:

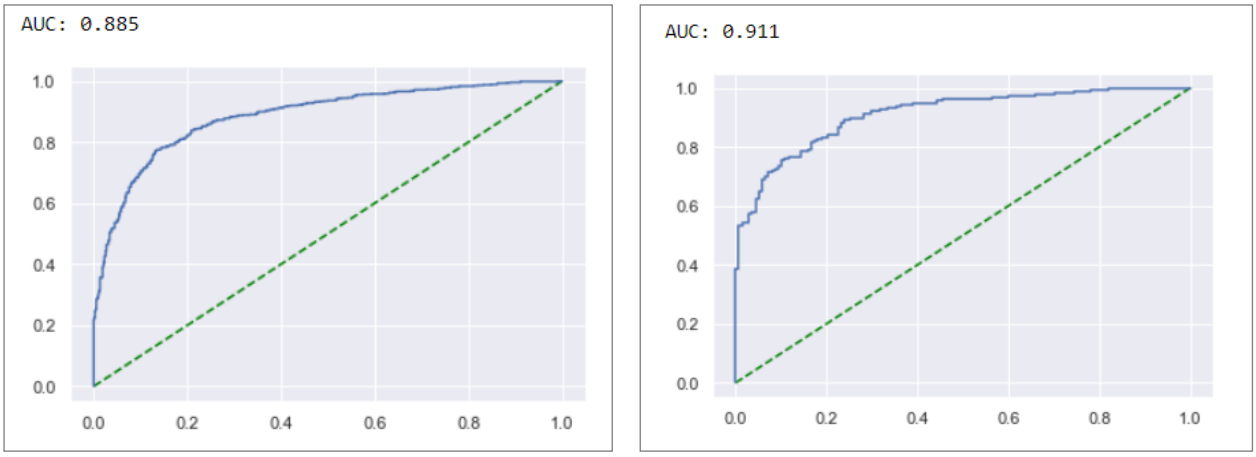

Figure 57. ROC curve and AUC score of tuned Naive Bayes model of train (left) & test (right).

Naïve Bayes model with SMOTE					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	608		265	
2.	True Negative	590		110	
3.	False Positive	149		28	
4.	False Negative	132		53	
5.	AUC score	88.5%		91.1%	
6.	Accuracy	81%		82%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	82%	80%	67%	90%
8.	Recall	82%	82%	80%	83%
9.	F1 score	81%	81%	73%	87%

Table 13. Model performance of tuned Naive bayes model

Inferences:

- The naïve bayes model with smote is performing slightly better after oversampling the minority class.
- In this model also it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Model performance metrics i.e., Accuracy, AUC, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
- **Overall, the metrics are good fit.**

Bagging (Random Forest should be applied for Bagging):

Bagging, also known as Bootstrap aggregating, is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model.

Random Forest applied to Bagging changes the algorithm the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation.

It is a simple tweak. In CART, when selecting a split point, the learning algorithm is allowed to look through all variables and all variable values in order to select the most optimal split-point. The random forest algorithm changes this procedure so that the learning algorithm is limited to a random sample of features of which to search.

To build a Bagging (Random Forest should be applied for Bagging):

- Fitting the train data in Random Forest model which is imported from Sklearn ensemble with n_estimators = 100 and random state =1.
- Building the bagging model using bagging classifier imported from Sklearn ensemble.
- Bagging classifier is fit to training data with Random Forest as the base estimator.

- Predicting on Training and Testing scaled dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.
- Checking the Model performance of train and test data in the next question.

Model performance of Bagging:

Classification report:

	precision	recall	f1-score	support
0	0.97	0.93	0.95	322
1	0.97	0.99	0.98	739
accuracy			0.97	1061
macro avg	0.97	0.96	0.96	1061
weighted avg	0.97	0.97	0.97	1061

	precision	recall	f1-score	support
0	0.79	0.69	0.74	138
1	0.87	0.92	0.90	318
accuracy			0.85	456
macro avg	0.83	0.80	0.82	456
weighted avg	0.85	0.85	0.85	456

Figure 58. Classification report of Bagging model of train (left) & test (right).

Confusion Matrix for training and testing data:

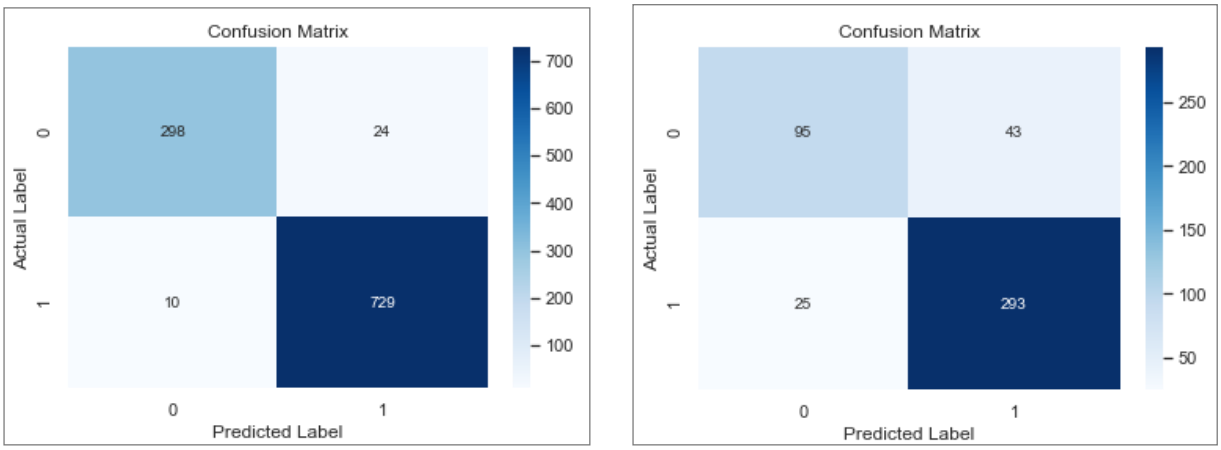

Figure 59. Confusion matrix of Bagging model of train (left) & test (right).

ROC Curve and ROC_AUC score:

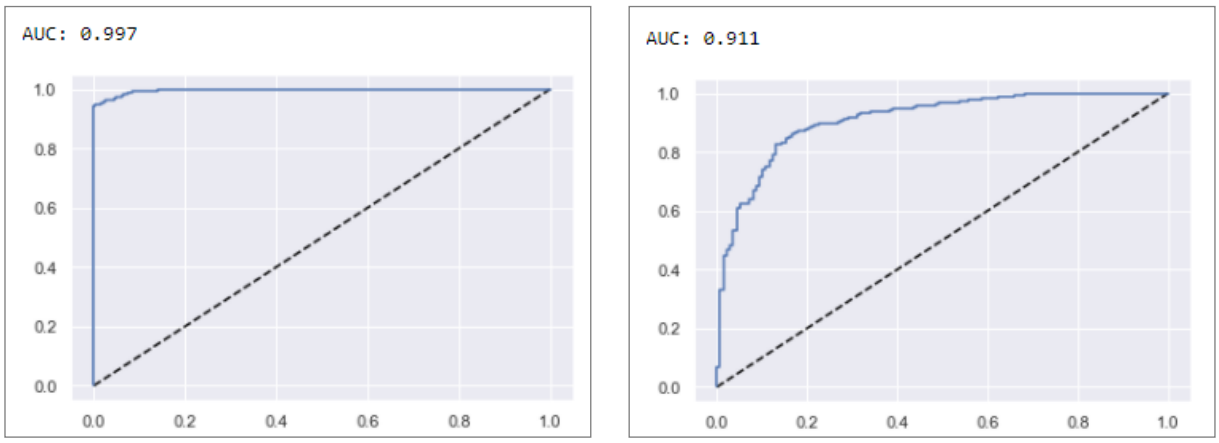

Figure 60. ROC curve and AUC score of Bagging model of train (left) & test (right).

Bagging (Random Forest should be applied for Bagging)					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	729		293	
2.	True Negative	298		95	
3.	False Positive	24		43	
4.	False Negative	10		25	
5.	AUC score	97%		91.1%	
6.	Accuracy	82%		85%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	97%	97%	79%	87%
8.	Recall	93%	99%	69%	92%
9.	F1 score	95%	98%	74%	90%

Table 14. Model performance for Bagging model.

Inferences:

- From the analysis we can see that the train performance is better and the test is not performing that better compared to the train data, there is more than 10% variation range compared to train.
- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters and also Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Bagging with Random Forest has performed exceptionally well on the train data for both the classes 0 and 1 with an accuracy score of 1. Recall, Precision, F1 score is also 1 on Train Data for both the classes. However, a closer look at the metrics of the Testing Dataset shows that this model is overfitted for conservative class.
- **Overall, the metrics are good fit.**

Boosting:

The term 'Boosting' refers to a family of algorithms which converts weak learner to strong learners Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree.

For choosing the right distribution, here are the following steps:

- Step 1: The base learner takes all the distributions and assign equal weight or attention to each observation.
- Step 2: If there is any prediction error caused by first base learning algorithm, then we pay higher attention to observations having prediction error. Then, we apply the next base learning algorithm.
- Step 3: Iterate Step 2 till the limit of base learning algorithm is reached or higher accuracy is achieved.

Finally, it combines the outputs from weak learner and creates a strong learner which eventually improves the prediction power of the model. Boosting pays higher focus on examples which are mis-classified or have higher errors by preceding weak rules.

AdaBoost (Adaptive Boosting)

The AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique in Machine Learning used as an Ensemble Method. In Adaptive Boosting, all the weights are re-assigned to each instance where higher weights are given to the incorrectly classified models, and it fits the sequence of weak learners on different weights.

Adaboost starts by making predictions on the original dataset in easy language, and then it gives equal weights to each observation. If the prediction made using the first learner is incorrect, it allocates the higher importance to the incorrectly predicted statement and an iterative process. It goes on to add new learners until the limit is reached in the model.

To build a AdaBoost Model:

- Fitting the train data in AdaBoost Classifier model which is imported from Sklearn ensemble with n_estimators = 100 and random state =1 where n_estimators parameter is used to control the number of weak learners, learning_rate parameter controls the contribution of all the vulnerable learners in the final output, base_estimator parameter helps to specify different machine learning algorithms.
- Predicting on Training and Testing scaled dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.
- Checking the Model performance of train and test data in the next question.

Model performance of AdaBoost model:

Classification report:

	precision	recall	f1-score	support
0	0.78	0.71	0.74	322
1	0.88	0.91	0.89	739
accuracy			0.85	1061
macro avg	0.83	0.81	0.82	1061
weighted avg	0.85	0.85	0.85	1061

	precision	recall	f1-score	support
0	0.75	0.69	0.72	138
1	0.87	0.90	0.88	318
accuracy			0.84	456
macro avg	0.81	0.79	0.80	456
weighted avg	0.83	0.84	0.83	456

Figure 61. Classification report of AdaBoost model of train (left) & test (right).

Confusion Matrix for training and testing data:

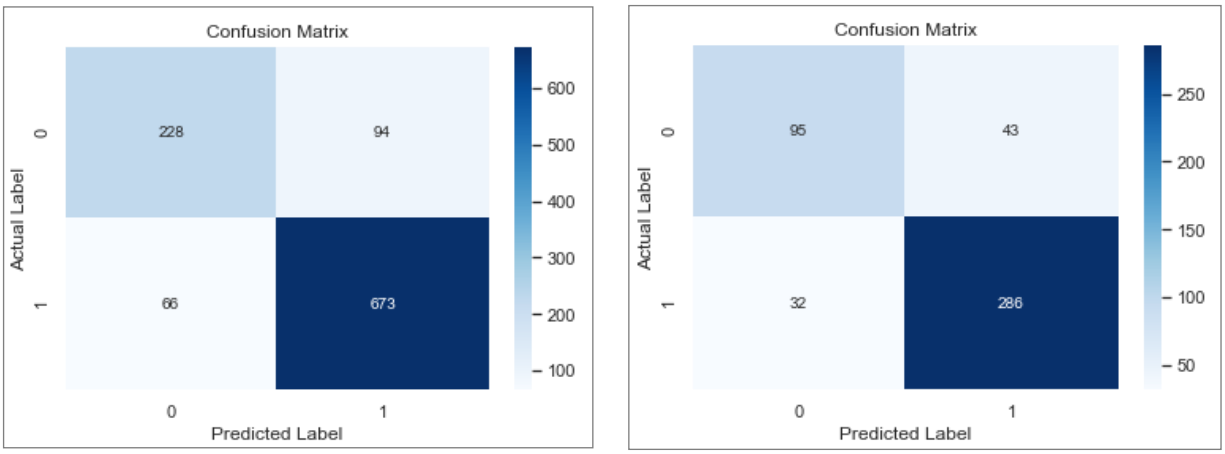

Figure 62. Confusion matrix of AdaBoost model of train (left) & test (right).

ROC Curve and ROC_AUC score:

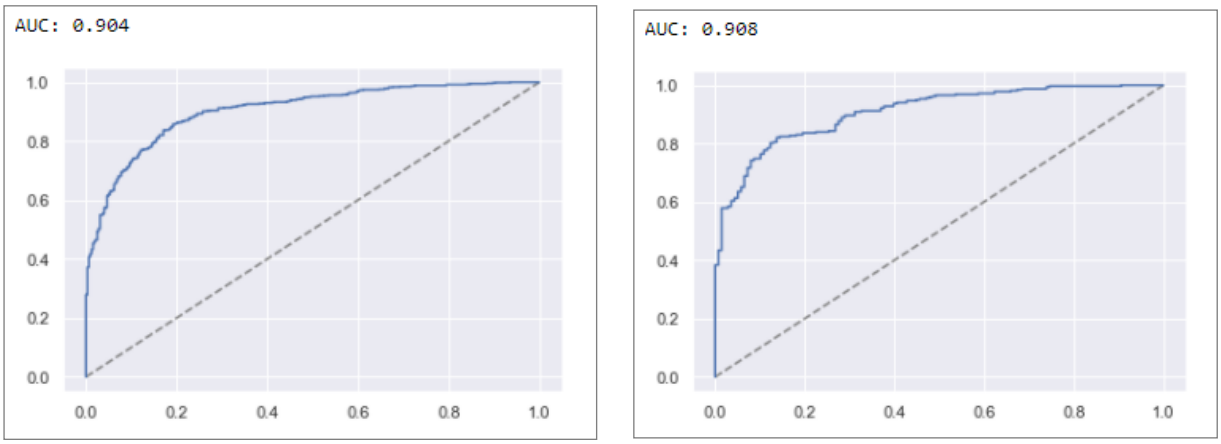

Figure 63. ROC curve and AUC score of AdaBoost model of train (left) & test (right).

AdaBoost Model					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	673		286	
2.	True Negative	228		95	
3.	False Positive	94		43	
4.	False Negative	66		32	
5.	AUC score	90.4%		90.8%	
6.	Accuracy	85%		84%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	78%	88%	75%	87%
8.	Recall	71%	91%	69%	90%
9.	F1 score	74%	89%	72%	88%

Table 15. Model performance for AdaBoost model.

Inferences:

- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well.
- Model performance metrics i.e., Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
- **Overall, the metrics are good fit.**

Gradient Boosting Model:

In the gradient boosting algorithm, we train multiple models sequentially, and for each new model, the model gradually minimizes the loss function using the Gradient Descent method. The Gradient Tree Boosting algorithm takes decision trees as the weak leaners because the nodes in a decision tree consider a different branch of features for selecting the best split, which means all the trees are not the same. Hence, they can capture different outputs from the data all the time.

The gradient tree boosting algorithm is sequentially built because, for each new tree, the model considers the errors of the last tree, and the decision of every successive tree is built on the mistakes made by the previous tree.

To build a AdaBoost Model:

- Scaled dataset is used to build Gradient boosting model.
- Fitting the train data in Gradient Boosting Classifier model which is imported from Sklearn ensemble with random state = 1.
- Predicting on Training and Testing scaled dataset.
- Getting the Predicted Classes and Probabilities and creating a data frame.
- Checking the Model performance of train and test data in the next question.

Model performance of Gradient Boosting model:

Classification report

	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

	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

Figure 64. Classification report of Gradient Boosting model of train (left) & test (right).

Confusion Matrix for training and testing data:

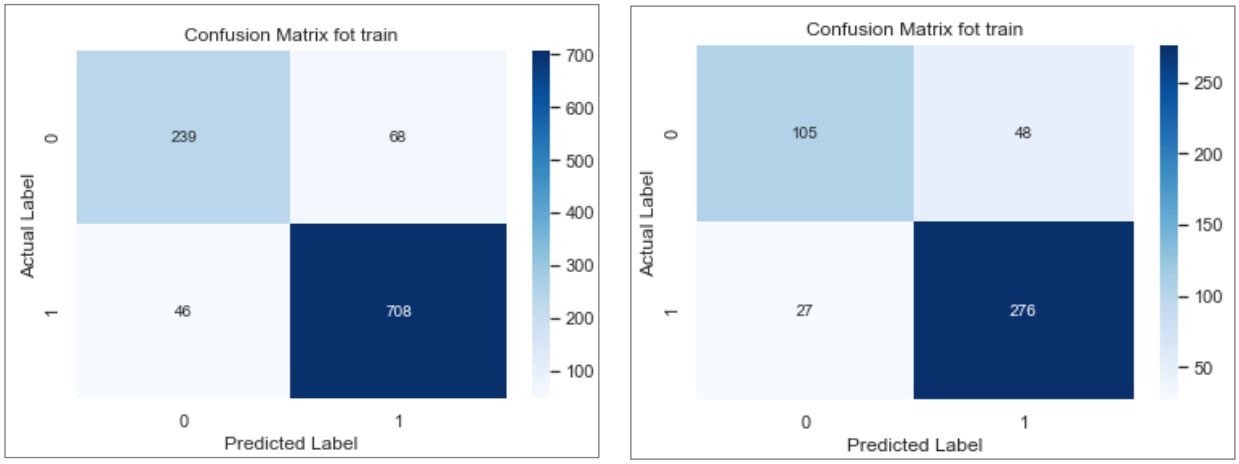

Figure 65. Confusion matrix of Gradient Boosting model of train (left) & test (right).

ROC Curve and ROC_AUC score:

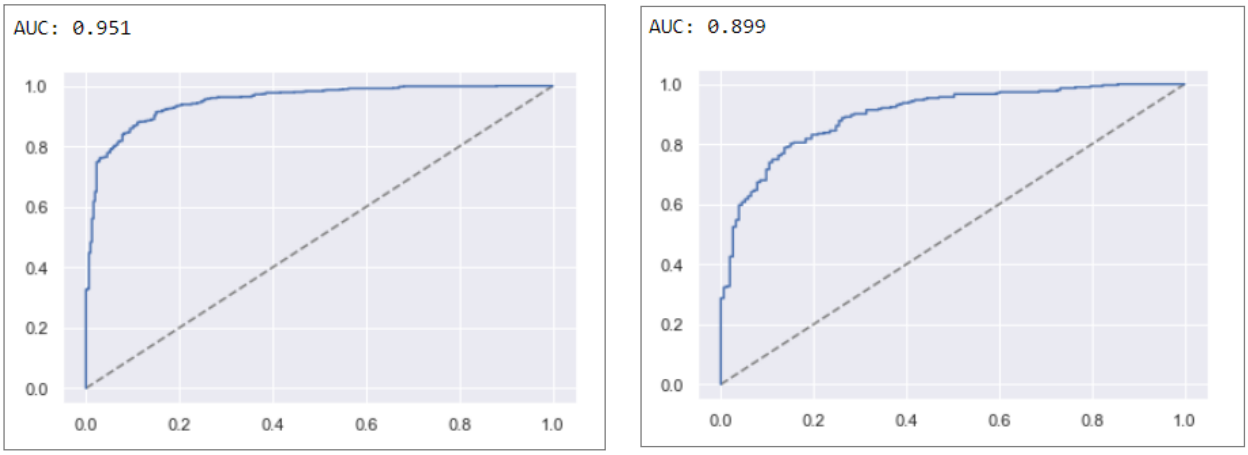

Figure 66. ROC curve and AUC score of Gradient Boosting model of train (left) & test (right).

Model performance table of Gradient Boosting Model:

Gradient model is performing better than the AdaBoost model. The classification for both classes is pretty good for both test and train data. It can be best suitable model, lets see the comparison od all model metrices in later part.

Gradient Boosting Model					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	708		276	
2.	True Negative	239		105	
3.	False Positive	68		48	
4.	False Negative	46		27	
5.	AUC score	95.1%		89.9%	
6.	Accuracy	89%		84%	
		Conservative	**Labour**	**Conservative**	**Labour**
7.	Precision	84%	91%	80%	85%
8.	Recall	78%	94%	69%	91%
9.	F1 score	81%	93%	74%	88%

Table 16. Model performance for Gradient Boosting model

Inferences:

- From the analysis it can be said that the model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The high precisions, recall and f1 scores indicate that overall, the model seems to classify the respondents well compared to other models as well.
- Model performance metrics i.e., Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.
- This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
- **Overall, the metrics are good fit.**

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion.

Model performance helps to understand how good the model that we have trained using the dataset is so that we have confidence in the performance of the model for future predictions.

We evaluate our models' performance on **train and test datasets of the tuned models**. We try to determine if the model is underfitting or overfitting by checking for accuracy, precision, and other factors. We have specific scores and matrices for our model's performance. Following are the methods used to evaluate the model performance:

1. Confusion Matrix
2. Classification Report
   - Accuracy
   - Precision
   - Recall
   - F1 Score
3. ROC curve
4. AUC score

1. Confusion Matrix:

This gives us how many zeros (0s) i.e. (class = No claim) and ones (1s) i.e. (class = Yes claim) were correctly predicted by our model and how many were wrongly predicted.

		Predicted Class	
		Class = No	Class = Yes
Actual class	Class = No	True Negative	False Positive
	Class = yes	False Negative	True Positive

Table 17. Confusion matrix

I.Accuracy:

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

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

II.Precision:

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/(TP + FP)

III.**Recall** (Sensitivity):

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

Recall = TP/(TP + FN)

IV.F1 Score:

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. That is, a good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1 , while the model is a total failure when it's 0

F1 score = 2 x [(Precision x Recall) / (Precision + Recall)]

2. ROC Curve:

ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

3. AUC Score:

AUC score gives the area under the ROC curve built. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative.

The Labour party is denoted as 1 and the Conservative party as 0. To create an exit poll for news channel CNBE predicting classification of both parties. In this scenario Accuracy followed other metrics plays the major in comparing the model performance. We have to consider the model performance for both the classes is good with the good fit of the model with Accuracy, AUC, precision, and recall for training data and test data are almost nearly in the general norm of +/- 10% of each other.

Model Comparison Summary:

Model	Model tuning approach	Output
Logistic Regression	Applying grid search CV using hyperparameters: penalty, solver, tolerance, max iteration.	Accuracy, Recall and Precision Scores remained same even after implementing Grid search CV, suggesting base model adapted is good enough.
LDA	1. Applying Grid search CV using multiple solver: 'svd', 'lsqr', 'eigen'. 2. Identifying different threshold probability for best possible performance score.	At threshold probability 0.4 we have best possible scores for recall, precision, accuracy and F1 score for training data and similar result can be observed in test data for probability 0.4 and 0.5 against base model assumption of 0.5, however test data performance better than train data suggest under sampling/under fit model in general.
Naïve Bayes	SMOTE for class imbalance, although classes where in ratio of approx. 70:30.	By Applying SMOTE we were able to improve recall, precision and F1 score for the conservative class, However the Accuracy score of base model is better.
KNN	Identifying appropriate K nearest neighbour where mis classification is minimum	At k = 11, we were able to get better test score compared to base model score of 83%. However, the other performance metrics are almost similar to base model.
Random Forest	Basic model run	It was over fit model with Train score at 100% and Test score at 84%

Bagging with Random Forest	To counter fit issue by identifying appropriate n_estimator to increased stability and accuracy of model.	RF with bagging helped minimize fit issue of overfitting by taking train accuracy score to 97% and Test score accuracy to 85%. The model is overfitting.
Ada Boost	Adaptive Boosting to correct any incorrectly classified instance which can counter fit issues	The model does a better job of correctly classifying the Labour Party voters than the Conservative voters. The train and test accuracy are 85% and 84% respectively.
Gradient Boost	It Target the model to follow to have minimal variation issues or classification issues	Gradient Boost Model has minimal fit issue with Train and test accuracy of 89% and 84% respectively. And also, the other metric scores are also pretty much balanced between the train and test for both the classes.

Table 18. Model comparison summary.

Model Evaluation metrics for Class 1 – Labour party						
Models	**Train/Test**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
Logistic Regression	Train	83%	86%	90%	88%	87.7%
	Test	86%	87%	93%	90%	91.5%
Tuned Logistic Regression	Train	83%	86%	90%	88%	87.7%
	Test	86%	87%	93%	90%	91.3%
LDA	Train	82%	86%	89%	87%	87.7%
	Test	85%	87%	92%	90%	91.4%
Tuned LDA	Train	81%	82%	93%	87%	87.7%
	Test	85%	85%	95%	90%	91.3%
Naïve Bayes	Train	82%	87%	87%	87%	87.3%
	Test	86%	88%	92%	90%	91.2%
Naïve Bayes with SMOTE	Train	81%	80%	82%	81%	88.5%
	Test	82%	90%	83%	87%	91.1%
KNN	Train	86%	89%	92%	92%	92.7%
	Test	82%	85%	89%	87%	87.0%
Tuned KNN	Train	84%	87%	91%	89%	90.9%
	Test	83%	85%	90%	88%	88.9%
Bagging with Random Forest	Train	82%	97%	99%	98%	97.0%
	Test	85%	87%	92%	90%	91.1%
Ada Boost	Train	85%	88%	91%	89%	90.4%
	Test	84%	87%	90%	88%	90.8%
Gradient Boosting	Train	89%	91%	94%	93%	95.1%
	Test	84%	85%	91%	88%	89.9%

Table 19. Model Evaluation metrics for Class 1 – Labour party

Model Evaluation metrics for Class 0 – Conservative party						
Models	**Train/Test**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
Logistic Regression	Train	83%	74%	66%	70%	87.7%
	Test	86%	81%	69%	75%	91.5%
Tuned Logistic Regression	Train	83%	74%	66%	70%	87.7%
	Test	86%	82%	68%	74%	91.3%
LDA	Train	82%	72%	67%	70%	87.7%
	Test	85%	80%	69%	74%	91.4%
Tuned LDA	Train	81%	77%	53%	63%	87.7%
	Test	85%	85%	60%	70%	91.3%
Naïve Bayes	Train	82%	70%	70%	70%	87.3%
	Test	86%	79%	72%	75%	91.2%
Naïve Bayes with SMOTE	Train	81%	82%	82%	81%	88.5%
	Test	82%	67%	80%	73%	91.1%
KNN	Train	86%	77%	71%	74%	92.7%
	Test	82%	77%	69%	72%	87.0%
Tuned KNN	Train	84%	75%	68%	71%	90.9%
	Test	83%	78%	69%	73%	88.9%
Bagging with Random Forest	Train	82%	97%	93%	95%	97.0%
	Test	85%	79%	69%	74%	91.1%
Ada Boost	Train	85%	78%	71%	74%	90.4%
	Test	84%	75%	69%	72%	90.8%
Gradient Boosting	Train	89%	84%	78%	81%	95.1%
	Test	84%	80%	69%	74%	89.9%

Table 20. Model Evaluation metrics for Class 0 – Conservative party

All the models are so close in their performance. There are only slight differences in terms of accuracy and precision in the classification. All the models have performed well based on their F1 scores. Few models performed really well on the training set like Bagging model and Boosting model. Few models did better on the testing than training and they were the Logistic Regression Model and the Linear Discriminant Analysis model. Of all the models, the best model built that classifies the respondents well for the purpose of creating an exit poll in predicting the seats that will be won by the particular parties is based on following parameters:

We can compare the models on the following parameters:

1. Difference in performance between Train and Test Data: We have already observed that all the model performance metrics are well within the general norm of +/- 10% between train and test.

2. Difference between minority and majority class:

• **For Training Data**, it is observed that Gradient Boosting model has low and similar difference between majority class and minority class scores. This is also observed only in KNN model but the Accuracy is high in Gradient Boosting model.

• **For Testing Data**, it is also observed that Gradient Boosting model has low and similar difference between majority class and minority class scores. Moreover, the f1-score difference between majority and minority class is minimum for Naïve bayes with smote followed by Gradient Boosting. Hence according to this logic, we can consider the Gradient Boosting model better than other models.

3. Overfitting: Certain models have values of performance metrics above 90% but the performance in the test is less than the +/- 10% rule. Even though the accuracy of bagging model is good fit, we can see that the precision and recall for the test is less than more than 10% difference from the train.

4. Overall Higher Performance Metrics: Best model is one which has performed well on both Test and train data set and has high accuracy and AUC score. Gradient Boosting Model seems to fit well, where the variance in scores for test and train data is not too high as also the Model Score and AUC is more than to 90% for both the classes – 0 and 1.

Thus, it observed that Gradient Boosting model has performed very well in all the performance metrics both for training and testing data.

Moreover, for the given case study we want to correctly predict votes in favour of both Conservative Party and Labour Party. Thus type I error and type II error are both equally important for us. Hence 'f1 score' (note: here we are considering 'f1 score ' because it calculates the non-weighted average of minority and majority class) is the most important performance metric here. And since as observed 'f1 score' for Gradient Boosting Model is higher than all other models. Thus, according the performance metrics, **Gradient Boosting Model is overall better than the other models.**

Gradient Boosting Model is best optimized model to create an exit poll for the news channel CNBE that will aid in predicting overall win and seats covered by a particular political party: "Conservative" or "Labour."

1.8 Based on these predictions, what are the insights?

Business Insights:

An **election exit poll** is a poll of voters taken immediately after they have exited the polling stations. In order to anticipate the election outcome, voters are asked who they voted for. Exit polls are also used to gather demographic information on voters and to learn why they voted the way they did. Because real votes are cast anonymously, polling is the only way to gather this data. Politicians mostly rely on public opinion data to determine their positions on numerous political matters**. Political parties build election rallies and campaigns on this data, which highlights certain significant social, economic, and cultural feelings among the general people.**

- The attributes ‘Hague’ and ‘Blair’ important features in predicting the dependent variable
- Overall, all of the models developed are better at classifying Labour Party voters than Conservative Party voters.
- The exit poll should be able to predict that if the sample is a true reflection of the population than most of the voters would prefer the Labour Party
- The parties should keep a close eye on the public's perception of European integration and educate the public about their perspective. This appears to have a significant effect in their decision to vote for one of the two parties.
- The Labour Party and their leader seem to have positive ratings in the public’s eye.

Recommendations:

1. To check which party appears to have a better chance of winning:
When using an exit poll to effectively predict the outcome of a current election, a model with a better accuracy score should be considered. This will aid in properly predicting results 90% of the time, representing the genuine circumstance in real elections.

2. To build a new election campaign:
If the key objective of the business challenge is to create a new marketing campaign for a political party, it is essential to understand what are the sentiments of 80% of the population. In this situation, a model that accurately predicts on either class 0 or class 1 will be able to effectively help detect the masses' sentiments. As a result, any model with a better F1 score in either the Labour or Conservative parties can be considered.

3. To find out whether there are any fraudulent activities going on:
Election booths are frequently manipulated to obtain a specific result. To detect fraudulent behaviour within an area, exit polls might be conducted across several regions or seats. If the exit poll results differ from the results of the real elections, it is possible to pinpoint the source of the fraud, resulting in re-elections. In this situation, a model with more than 90% scores on either class 1 or class 0 on Recall or Precision can be considered. Precision is a measure of the relevancy of the results, whereas recall is a measure of the number of actually relevant results returned.

4. To assess the success or failure of a certain political campaign:
A successful campaign is one that causes the general public's preference to shift from one political party to another. In this situation, an exit poll will be done to ascertain the public's general feelings. The campaigning activity will then be carefully carried out in order to influence the voters' mindsets and ideas who voted for the other party. The next exit poll will be performed to see if people who previously voted for the opposing party have now switched their votes to the concerned party. This will aid in determining the campaign's effectiveness. The model with the greatest score for either class 0 or class 1 may be used to predict whether or not a voter would vote for the particular party.

Problem 2:

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

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

After importing all the necessary libraries, we download the inaugural data from NLTK corpus. There are multiple speeches given by various leaders and this can be seen by calling. fileids( ) function.

For our analysis, we are going to focus on the following three speeches:

1. 1941-Roosevelt.txt
2. 1961-Kennedy.txt
3. 1973-Nixon.txt

We can use the. raw () function to depict the raw file before processing.

Snapshots of all the raw files are given below:

1. 1941 - Roosevelt's Speech:


```
'On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn
Washington\'s day the task of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people
was to preserve that Nation from disruption from within.\n\nIn this day the task of the people is to save that Nation and its i
nstitutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a mome
nt and take stock -- to recall what our place in history has been, and to rediscover what we are and what we may be. If we do n
ot, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of th
e human spirit. The life of a man is three-score years and ten: a little more, a little less. The life of a nation is the fulln
ess of the measure of its will to live.\n\nThere are men who doubt this. There are men who believe that democracy, as a form of
Government and a frame of life, is limited or measured by a kind of mystical and artificial fate that, for some unexplained rea
son, tyranny and slavery have become the surging wave of the future -- and that freedom is an ebbing tide.\n\nBut we Americans
know that this is not true.\n\nEight years ago, when the life of this Republic seemed frozen by a fatalistic terror, we proved
that this is not true. We were in the midst of shock -- but we acted. We acted quickly, boldly, decisively.\n\nThese later year
s have been living years -- fruitful years for the people of this democracy. For they have brought to us greater security and,
I hope, a better understanding that life\'s ideals are to be measured in other than material things.\n\nMost vital to our prese
nt and our future is this experience of a democracy which successfully survived crisis at home; put away many evil things; buil
t new structures on enduring lines; and, through it all, maintained the fact of its democracy.\n\nFor action has been taken wit
hin the three-way framework of the Constitution of the United States. The coordinate branches of the Government continue freely
to function. The Bill of Rights remains inviolate. The freedom of elections is wholly maintained. Prophets of the downfall of A
merican democracy have seen their dire predictions come to naught.\n\nDemocracy is not dying.\n\nWe know it because we have see
n it revive--and grow.\n\nWe know it cannot die -- because it is built on the unhampered initiative of individual men and women
joined together in a common enterprise -- an enterprise undertaken and carried through by the free expression of a free majorit
y.\n\nWe know it because democracy alone, of all forms of government, enlists the full force of men\'s enlightened will.\n\nWe
know it because democracy alone has constructed an unlimited civilization capable of infinite progress in the improvement of hu
man life.\n\nWe know it because, if we look below the surface, we sense it still spreading on every continent -- for it is the
most humane, the most advanced, and in the end the most unconquerable of all forms of human society.\n\nA nation, like a perso
n, has a body--a body that must be fed and clothed and housed, invigorated and rested, in a manner that measures up to the obje
ctives of our time.\n\nA nation, like a person, has a mind -- a mind that must be kept informed and alert, that must know itsel
f, that understands the hopes and the needs of its neighbors -- all the other nations that live within the narrowing circle of
```


Figure 67. Snapshot of 1941- Roosevelt's speech

2. 1961 – Kennedy's Speech:

'Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well a s a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our forebears l prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for w hich our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity of the state, but from the hand of God.\n\nWe dare not forget today that we are the heirs of that first revolution. Let the wor d go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans - - born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage -- and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world.\n\nLet every nation know, whether it wishes us well or ill, that we shall pay any price, bear any burden, meet any hardship, support any friend, oppose any foe, in order to assure the survival and the success of liberty.\n\nThis much we pledge -- and more.\n\nTo those old allies whose cultural and spiritual origins we share, we pledge the loyalty of faithful friends. United, there is little we cannot do in a host of cooperative ventures. Divided, there is litt le we can do -- for we dare not meet a powerful challenge at odds and split asunder.\n\nTo those new States whom we welcome to the ranks of the free, we pledge our word that one form of colonial control shall not have passed away merely to be replaced by a far more iron tyranny. We shall not always expect to find them supporting our view. But we shall always hope to find them str ongly supporting their own freedom -- and to remember that, in the past, those who foolishly sought power by riding the back of the tiger ended up inside.\n\nTo those peoples in the huts and villages across the globe struggling to break the bonds of mass misery, we pledge our best efforts to help them help themselves, for whatever period is required -- not because the Communists may be doing it, not because we seek their votes, but because it is right. If a free society cannot help the many who are poor, it cannot save the few who are rich.\n\nTo our sister republics south of our border, we offer a special pledge -- to convert ou r good words into good deeds -- in a new alliance for progress -- to assist free men and free governments in casting off the ch ains of poverty. But this peaceful revolution of hope cannot become the prey of hostile powers. Let all our neighbors know that we shall join with them to oppose aggression or subversion anywhere in the Americas. And let every other power know that this H emisphere intends to remain the master of its own house.\n\nTo that world assembly of sovereign states, the United Nations, our last best hope in an age where the instruments of war have far outpaced the instruments of peace, we renew our pledge of suppor t--to prevent it from becoming merely a forum for invective -- to strengthen its shield of the new and the weak -- and to enlar ge the area in which its writ may run.\n\nFinally, to those nations who would make themselves our adversary, we offer not a ple

Figure 68. Snapshot of 1961-Kennedy's speech

3. 1973 – Nixon's Speech:

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and go od country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of see mingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are a bout to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnati on at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilit ies greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis p ast year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and by our missions to Peking and to Moscow, we were able to establish the base for a new and more durable pattern of relationships among the nations of the world. Because of America\'s bold initiatives, 1972 will be long remembered as the year of the greates t progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flim sy peace which is merely an interlude between wars, but a peace which can endure for generations to come.\n\nIt is important th at we understand both the necessity and the limitations of America\'s role in maintaining that peace.\n\nUnless we in America w ork to preserve the peace, there will be no peace.\n\nUnless we in America work to preserve freedom, there will be no freedo m.\n\nBut let us clearly understand the new nature of America\'s role, as a result of the new policies we have adopted over the se past four years.\n\nWe shall respect our treaty commitments.\n\nWe shall support vigorously the principle that no country ha s the right to impose its will or rule on another by force.\n\nWe shall continue, in this era of negotiation, to work for the l imitation of nuclear arms, and to reduce the danger of confrontation between the great powers.\n\nWe shall do our share in defe nding peace and freedom in the world. But we shall expect others to do their share.\n\nThe time has passed when America will ma ke every other nation\'s conflict our own, or make every other nation\'s future our responsibility, or presume to tell the peop le of other nations how to manage their own affairs.\n\nJust as we respect the right of each nation to determine its own futur e, we also recognize the responsibility of each nation to secure its own future.\n\nJust as America\'s role is indispensable in preserving the world\'s peace, so is each nation\'s role indispensable in preserving its own peace.\n\nTogether with the rest o f the world, let us resolve to move forward from the beginnings we have made. Let us continue to bring down the walls of hostil ity which have divided the world for too long, and to build in their place bridges of understanding -- so that despite profound differences between systems of government, the people of the world can be friends.\n\nLet us build a structure of peace in the world in which the weak are as safe as the strong -- in which each respects the right of the other to live by a different syste m -- in which those who would influence others will do so by the strength of their ideas, and not by the force of their arms.\n \nLet us accept that high responsibility not as a burden, but gladly -- gladly because the chance to build such a peace is the

Figure 69. Snapshot of Nixon's speech

Checking the Number of Characters in each speech by using. len ( ) function:

1. President Franklin D. Roosevelt's speech have **7571 Characters**.
2. President John F. Kennedy's Speech have **7618 Characters.**
3. President Richard Nixon's Speech have **9991 Characters.**

Checking the Number of Words in each speech by using .len ( ) function on the list all words in speech:

1. There are **1536 words** in Roosevelt's speech
2. There are **1546 words** in Kennedy's speech.
3. There are **2028 words** in Nixon's speech

Check the Number of Sentences in each speech by using .len( ) function on the sents( ) on each speech:

1. There are **68 sentences** in Roosevelt's speech.
2. There are **52 sentences** in Kennedy's speech.
3. There are **69 sentences** in Nixon's speech.

2.2 Remove all the stopwords from all three speeches.

Stop Words: A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages. You can find them in the nltk_data directory.

Following are the list of stopwords in NLTK directory:

{'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which', 'those', 'i', 'after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'} + Punctuations

- Defining a variable 'stopwords' which contains the list of punctuations from the string library and the English stopwords from nltk
- Extended ['--'] to stopwords list
- Converting all the words to lower case
- Only keeping the words which are not the 'stopwords'

1- 1941 - Roosevelt's Speech:

- **The word count before removal of stopwords for Roosevelt's Speech is 1536 words.**
- **The word count after removal of stopwords for Roosevelts' Speech is 632 words.**

Below are the Snapshots of part of list of words before and after removal of stopwords:

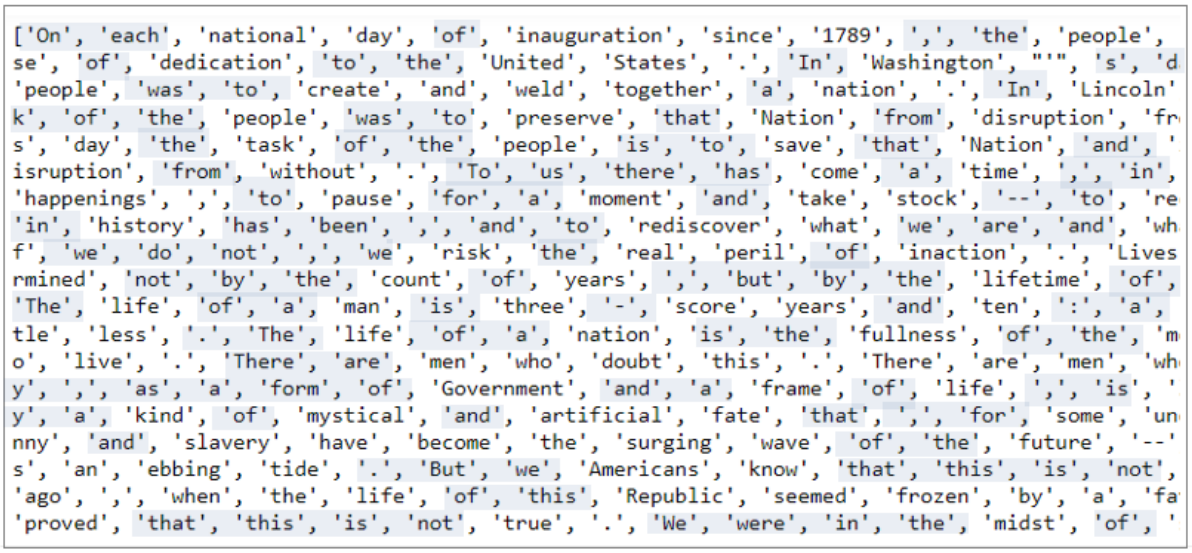

Figure 70. Snapshot of list of words of Roosevelt's speech before removal of stopwords

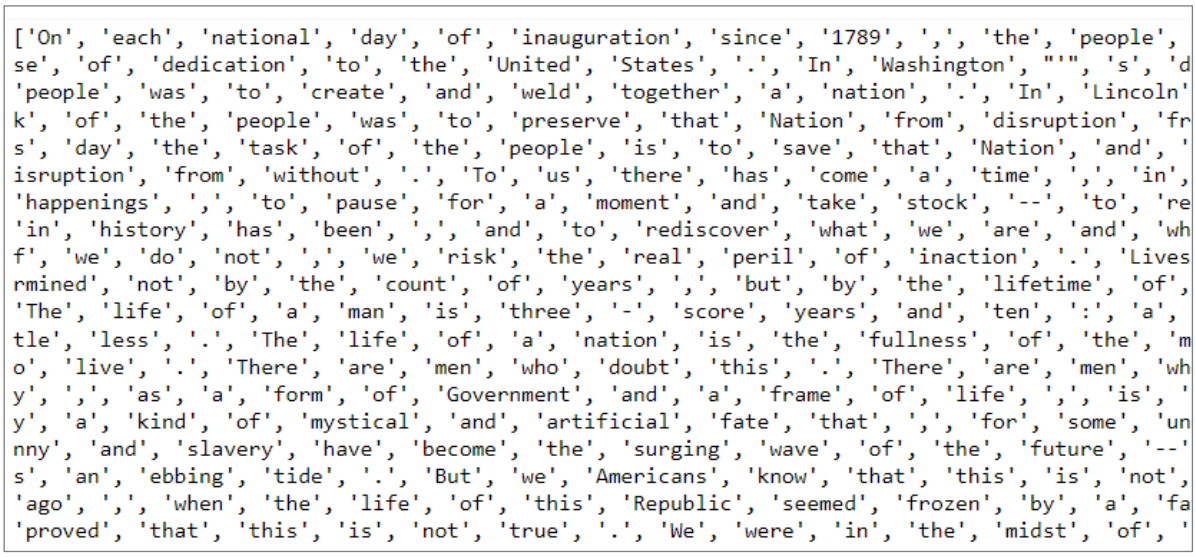

Figure 71. Snapshot of list of words of Roosevelt's speech after removal of stopwords

We can see from the above snapshots that the blue highlighted words are the stopwords and the new cleaned speech doesn't contain stopwords and all are in lower case. The words count is reduced from 1536 to 632 words after removal of stopwords which are not useful for further analysis.

2- 1961 – Kennedy's Speech:

- **The word count before removal of stopwords for Kennedy's Speech is 1546 words.**
- **The word count after removal of stopwords for Kennedy' Speech is 697 words.**

Below are the Snapshots of part of list of words before and after removal of stopwords:

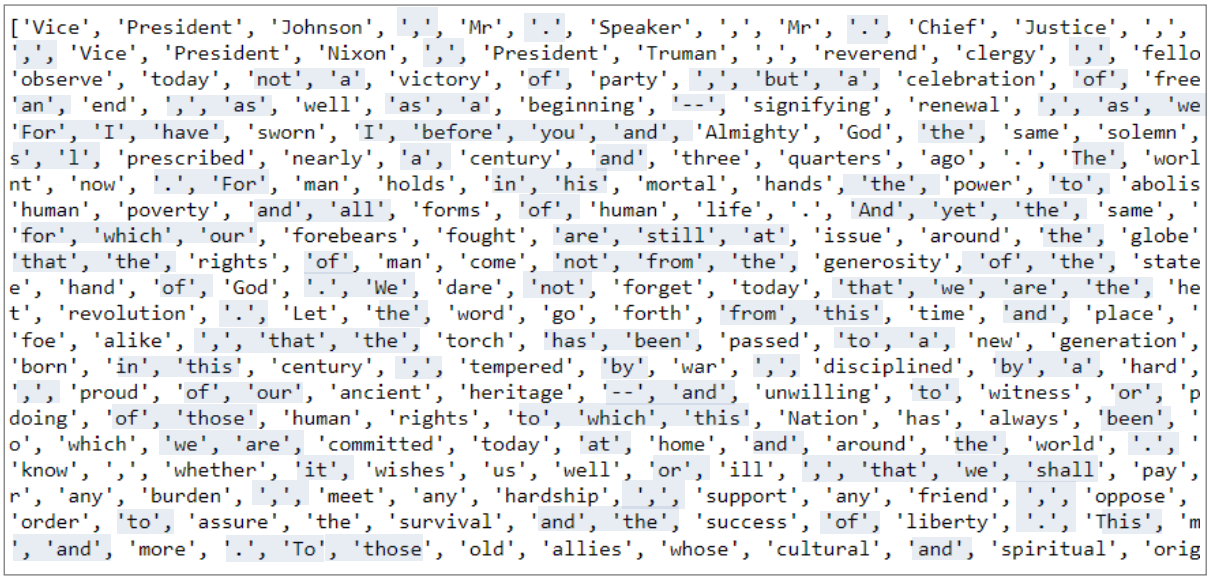

Figure 72. Snapshot of list of words of Kennedy's speech before removal of stopwords

Figure 73. Snapshot of list of words of Kennedy's speech after removal of stopwords

We can see from the above snapshots that the blue highlighted words are the stopwords and the new cleaned speech doesn't contain stopwords and all are in lower case. The words count is reduced from 1546 to 697 words after removal of stopwords which are not useful for further analysis.

3. 1973 – Nixon's Speech:

- **The word count before removal of stopwords for Nixon's Speech is 2028 words.**
- **The word count after removal of stopwords for Nixon's Speech is 836 words.**

Below are the Snapshots of part of list of words before and after removal of stopwords:

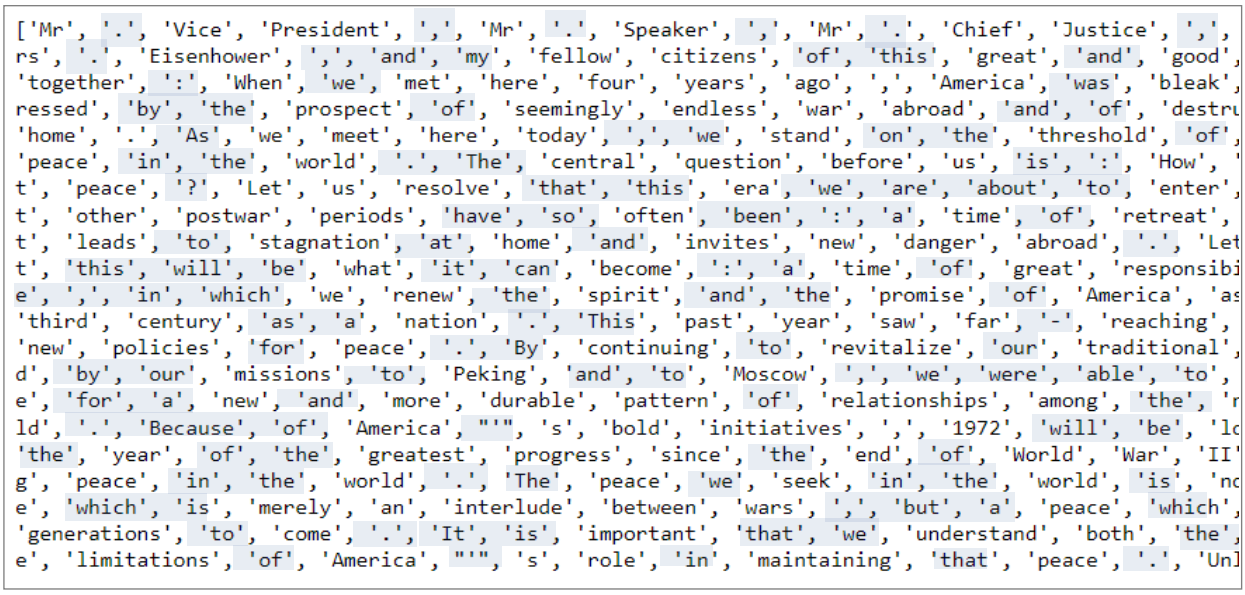

Figure 75. Snapshot of list of words of Nixon's speech before removal of stopwords

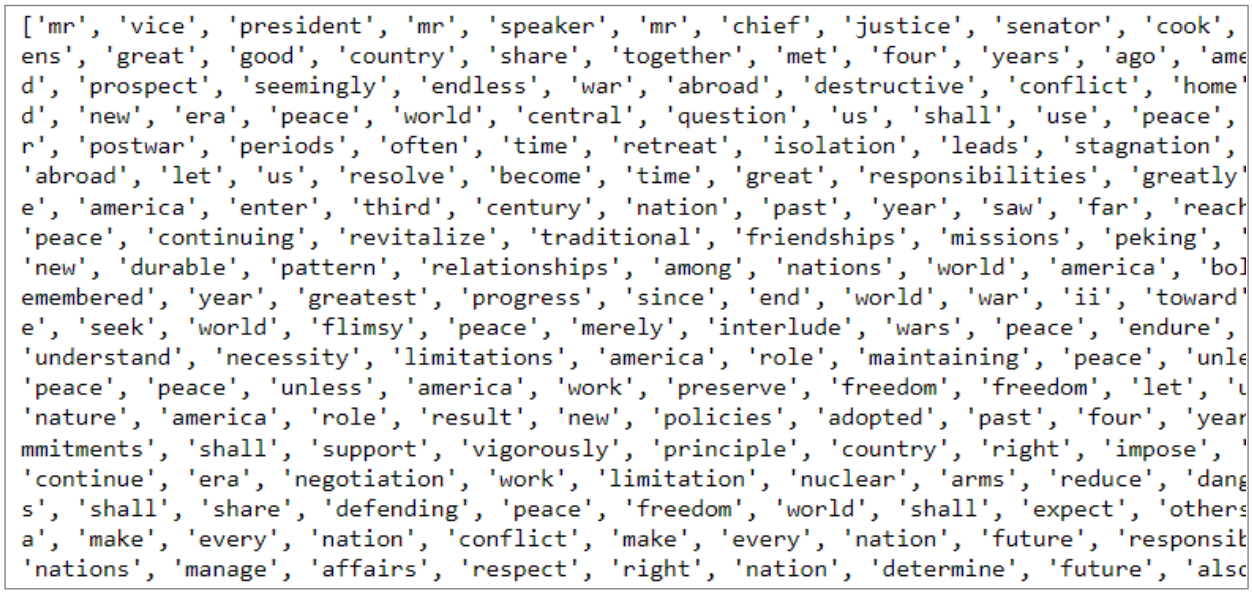

Figure 74.. Snapshot of list of words of Nixon's speech after removal of stopwords

We can see from the above snapshots that the blue highlighted words are the stopwords and the new cleaned speech doesn't contain stopwords and all are in lower case. The words count is reduced from 2028 to 836 words after removal of stopwords which are not useful for further analysis.

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)

The Frequency of each of the words for all the three speeches are found:

Roosevelt's speech

```
FreqDist({'the': 104, 'of': 81, ',': 77, '.': 67, 'and': 44, 'to': 35, 'in': 30, 'a': 29, '--': 25, 'is': 24, ...})
```

Kennedy's speech

```
FreqDist({',': 85, 'the': 83, 'of': 65, '.': 51, 'to': 38, 'and': 37, 'a': 29, 'we': 27, '--': 25, 'in': 24, ...})
```

Nixon's speech

```
FreqDist({',': 96, 'the': 80, '.': 68, 'of': 68, 'to': 65, 'in': 54, 'and': 47, 'we': 38, 'a': 34, 'that': 32, ...})
```

After cleaning the data i.e., removing all the stopwords, the top 15 most common words for each of the Speeches are given below:

Roosevelt's	Kennedy's	Nixon's
['nation', 'know', 'spirit', 'life', 'democracy', 'us', 'people', 'america', 'years', 'freedom', 'human', 'men', 'new', 'body', 'mind']	['let', 'us', 'world', 'sides', 'new', 'pledge', 'citizens', 'power', 'shall', 'free', 'nations', 'ask', 'president', 'fellow', 'freedom']	['us', 'let', 'america', 'peace', 'world', 'new', 'nation', 'responsibility', 'government', 'great', 'home', 'abroad', 'together', 'years', 'shall']

Table 21. Top 15 most common words for each of the speeches

Frequently used words in Roosevelt Speech: 'nation': 12, 'know': 10, 'spirit': 9

Frequently used words in Kennedy Speech: 'let': 16, 'us': 12, 'world': 8

Frequently used words in Nixon Speech: 'us': 26, 'let': 22, 'america': 21

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

Word Cloud for Roosevelt's Speech (after cleaning)!!

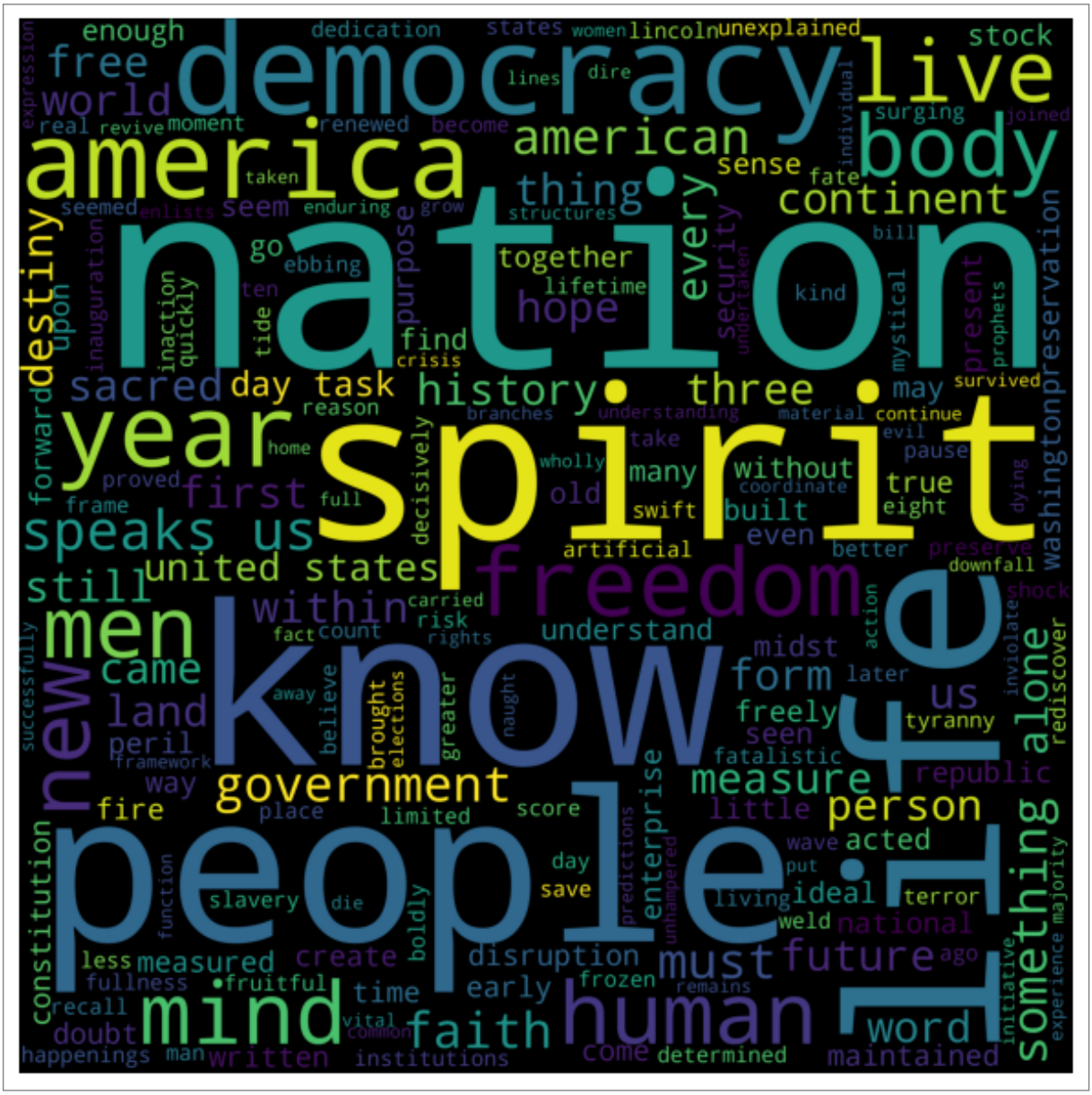

Figure 76. Word Cloud for Roosevelt's Speech (after cleaning)

Top words:

'nation', 'know',' spirit', 'life', 'democracy', 'us', 'people', 'america', 'years', 'freedom', 'human',' men', 'new', 'body', 'mind'

Word Cloud for Kennedy's Speech (after cleaning)!!

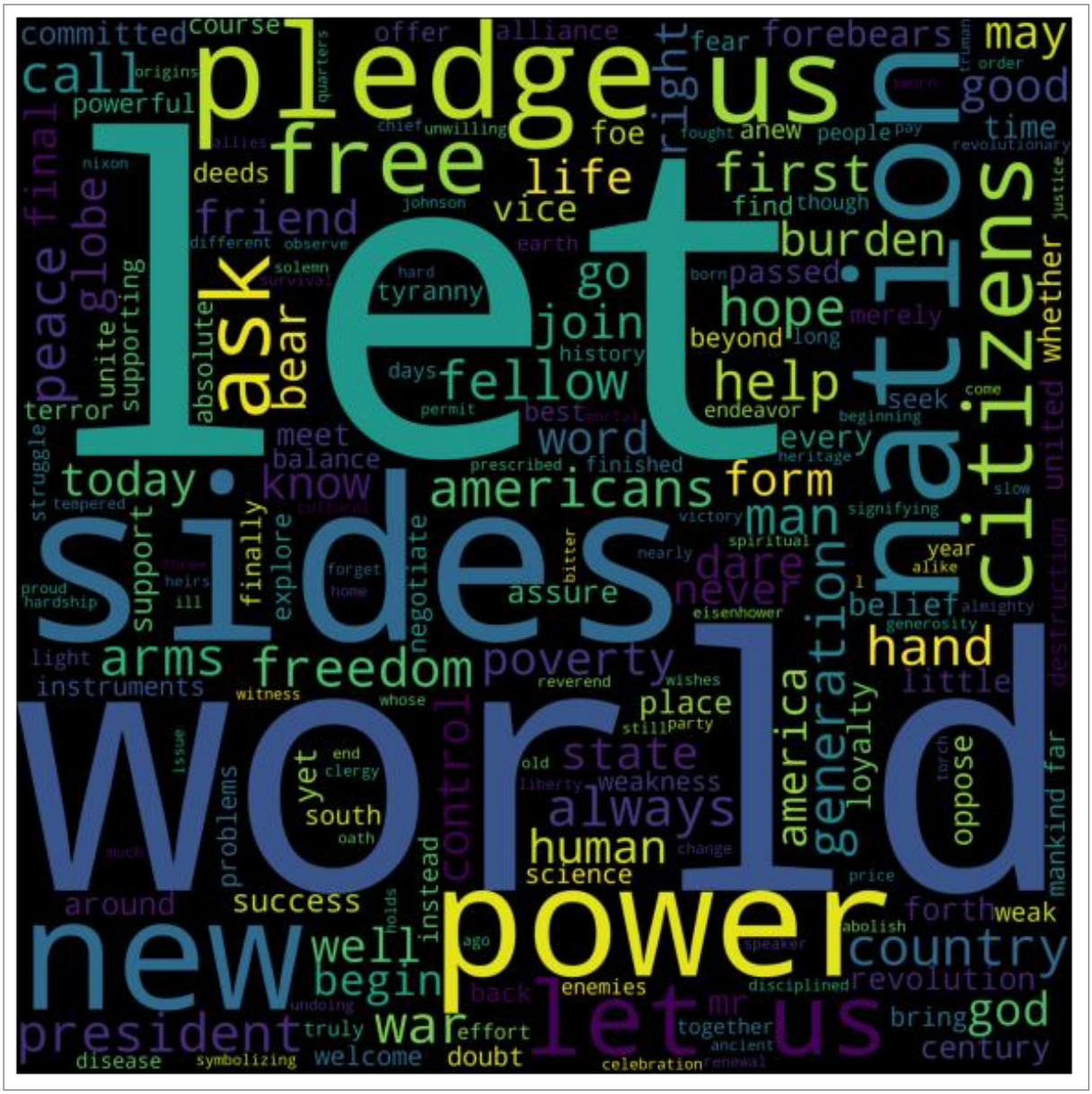

Figure 77. Word Cloud for Kennedy's Speech (after cleaning)

Top words:

'let', 'us', 'world', 'sides', 'new', 'pledge', 'citizens', 'power', 'shall', 'free', 'nations', 'ask', 'president', 'fellow', 'freedom'

Word Cloud for Nixon's Speech (after cleaning)!!

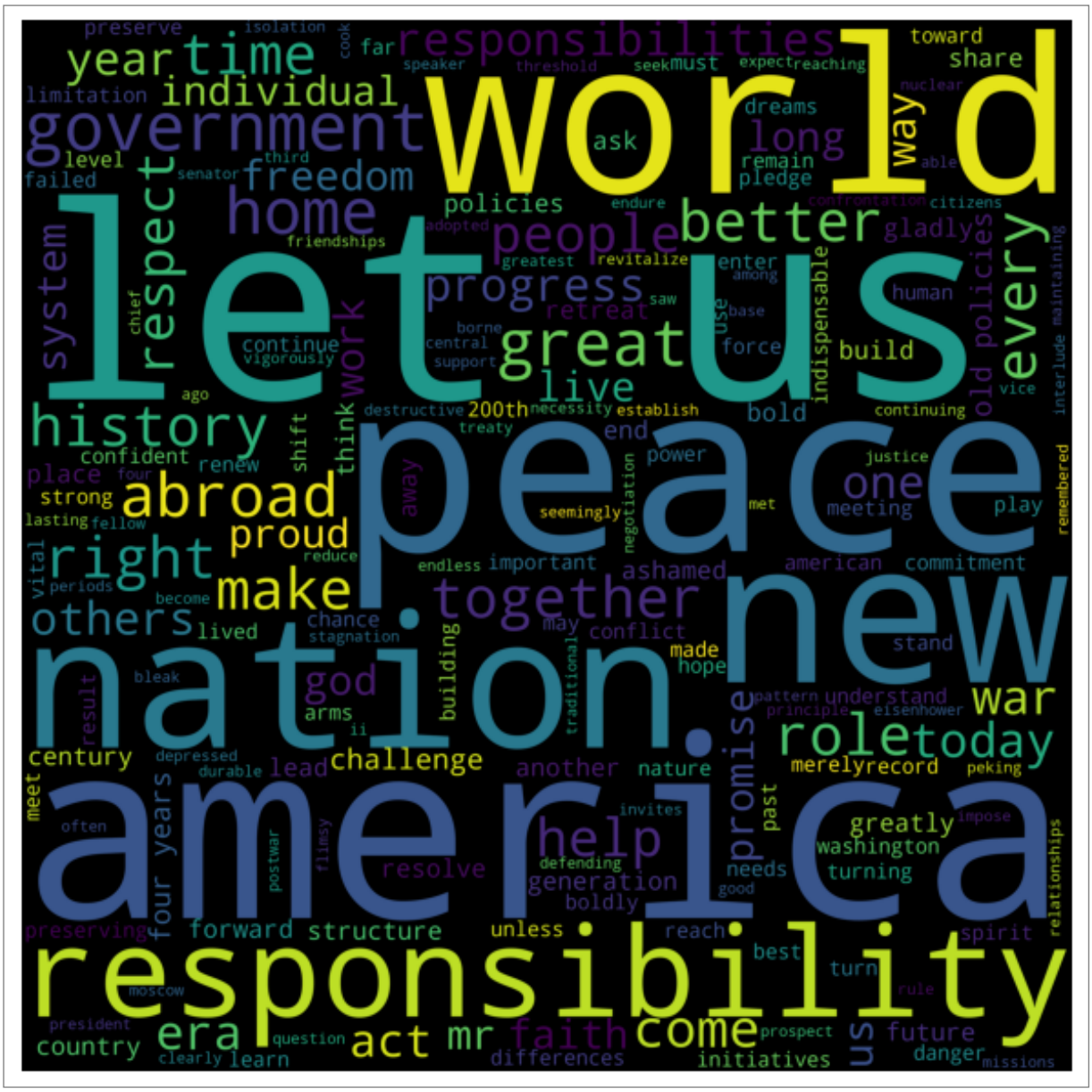

Figure 78. Cloud for Nixon's Speech (after cleaning)

Top words:

'us', 'let', 'america', 'peace', 'world', 'new', 'nation', 'responsibility', 'government', 'great', 'home', 'abroad', 'together', 'years', 'shall'