MACHINE LEARNING

BUSINESS REPORT

SUMMARY OF THE DATASETS:

A leading news channels CNBE wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. It wants 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.

This project, is based on the inaugural corpora from the nltk in Python. It will be based on the speeches of the Presidents of the United States of America

Yashveer Kothari. A
POST GRADUATE PROGRAMME IN DATA SCIENCE AND BUSINESS ANALYTICS.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABOUT MACHINE LEARNING

- Machine Learning (ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e., without any human assistance. The process starts with feeding good quality data and then training our machines(computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate.

ABOUT NAÏVE BAYES CLASSIFIER

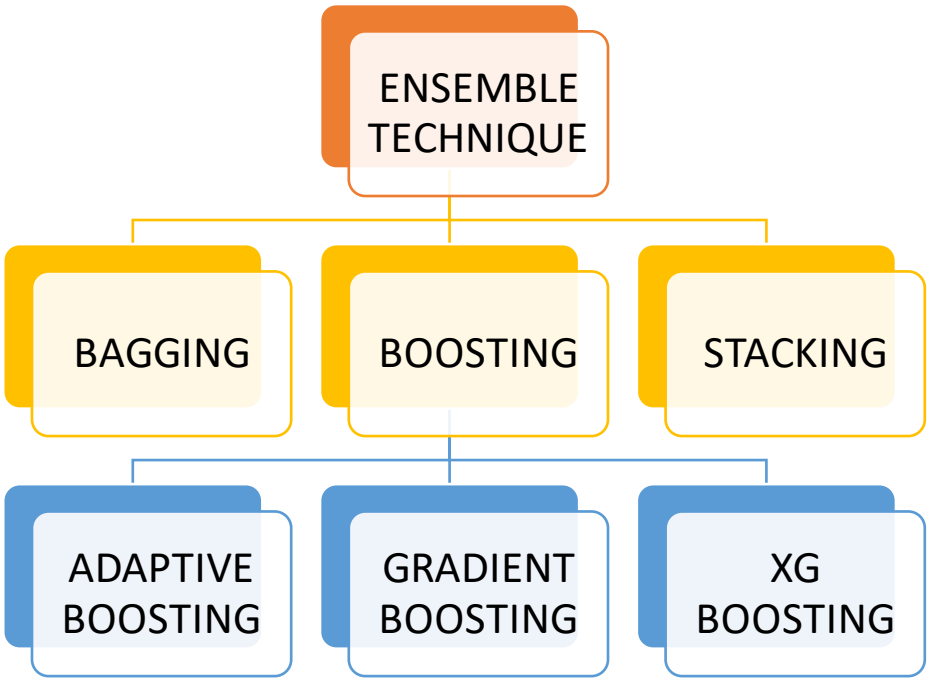
- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.
- In **Gaussian Naive Bayes,** continuous values associated with each feature (or independent variable) are assumed to be distributed according to a Gaussian distribution. All we would have to do is estimate the mean and standard deviation of the continuous variable.

ABOUT K-NEAREST NEIGHBOURS

- The k-nearest neighbours algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.
- KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

ABOUT ENSEMBLE MODEL

- Ensemble modelling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications.



1. BAGGING:

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

2. BOOSTING:

Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model and then trained sequentially—that is, each model tries to compensate for the weaknesses of its predecessor. With each iteration, the weak rules from each individual classifier are combined to form one, strong prediction rule.

2.1 ADAPTIVE BOOSTING:

This method operates iteratively, identifying misclassified data points and adjusting their weights to minimize the training error. The model continues optimize in a sequential fashion until it yields the strongest predictor.

2.3 GRADIENT BOOSTING:

his works by sequentially adding predictors to an ensemble with each one correcting for the errors of its predecessor. However, instead of changing weights of data points like AdaBoost, the gradient boosting trains on the residual errors of the previous predictor. The name, gradient boosting, is used since it combines the gradient descent algorithm and boosting method.

2.2 XG BOOSTING:

Extreme gradient Boosting / XG Boost is an implementation of gradient boosting that's designed for computational speed and scale. XG Boost leverages multiple cores on the CPU, allowing for learning to occur in parallel during training.

- **UNDERFITTING:**

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data.

Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model.

- **OVERFITTING:**

A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance.

A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

ABOUT TEXT MINING

- Text mining is the process of exploring and analysing large amounts of unstructured text data aided by software that can identify concepts, patterns, topics, keywords and other attributes in the data.
- Structured: Data is organized into pre-defined structure like a table of database -with rows and columns.
- Unstructured Data: Data does not have a pre-defined structure. Think of a collection of emails, a bunch of satellite images or the entire text of speeches from the British parliament since 1803.
- Bag of words: Documents simply represented by the words in the document and their frequencies. Disregards grammar and word order
- Semantic: Mapping natural language rules to get a formal representation of the meaning of the text
- Stop words: Common words that are not useful in providing value or context. E.g.: 'the', 'an', 'in' etc.
- Stemming: Returning words to their original stem. E.g.: 'Chopping', 'Chopped' are all replaced with 'Chop'
- A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

SENTIMENT ANALYSIS

- Sentiment analysis (opinion mining) is a text mining technique that uses machine learning and natural language processing (nlp) to automatically analyse text for the sentiment of the writer (positive, negative, neutral, and beyond).

PROBLEM 1

INTRODUCTION

The dataset contains data relating to voters in different regions and categories. A leading news channels CNBE wants to analyse recent elections. There are 1525 records with 9 variables. The channel wants us to build a model, to predict which party a voter will vote for on the basis of the dataset, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

TABLE 1: DATA DICTIONARY

VARIABLE NAME	DESCRIPTION
Vote	Party choice: Conservative or Labour
Age	In Years
economic.cond.national	Assessment of current national economic conditions, 1 to 5.
economic.cond.household:	Assessment of current household economic conditions, 1 to 5.
Blair	Assessment of the Labour leader, 1 to 5.
Hague	Assessment of the Labour leader, 1 to 5.
Europe	an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
political. knowledge	Knowledge of parties' positions on European integration, 0 to 3.
gender	Female or Male.

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

TABLE 2: TOP 5 DATA SAMPLE

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

TABLE 3: DATASET INFORMATION

```
Int64Index: 1525 entries, 1 to 1525
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
```

TABLE 4: DATASET DESCRIPTION

	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

TABLE 5: MISSING VALUE

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

TABLE 6: SKEWNESS OF THE DATASET

vote	0.858449
age	0.144621
economic.cond.national	-0.240453
economic.cond.household	-0.149552
Blair	-0.535419
Hague	0.152100
Europe	-0.135947
political.knowledge	-0.426838
gender	0.130239

TABLE 7: EXTRACTING DUPLICATES

Number of duplicate rows = 9

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
68	35	4	4	5	2	3	2	1
627	39	3	4	4	2	5	2	1
871	38	2	4	2	2	4	3	1
984	74	4	3	2	4	8	2	0
1155	53	3	4	2	2	6	0	0
1185	61	3	3	4	2	6	0	0
1237	36	3	3	2	2	6	2	0
1245	29	4	4	4	2	2	2	0
1439	40	4	3	4	2	2	2	1

OUTPUT: REMOVING DUPLICATES

```
Number of duplicate rows = 0
shape after removing: (1516, 8)
```

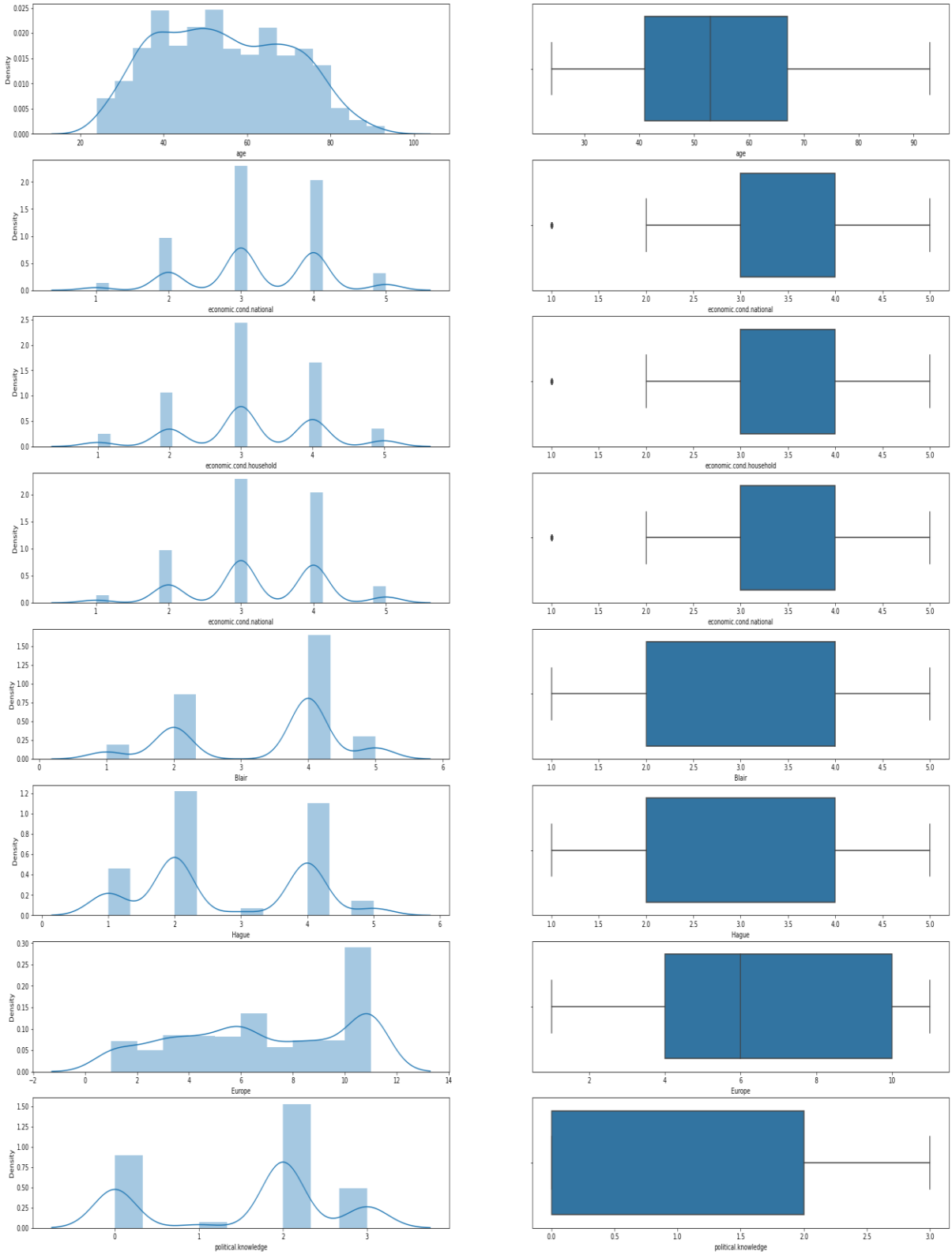
INFERENCE FOR Q.1.1

1. There are 1525 rows and 9 columns, which is the shape of the dataset.
2. There are 7 columns which are integer data type and 2 object type.
3. The maximum age in the dataset is 93 and lowest is 24 years.
4. There are no missing values in the dataset.
5. There are 8 duplicate rows in the dataset. Which has been removed as it does not make much of a difference to the dataset and duplicates in voting can affect the results of the elections.
6. Rules of Skewness:

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.
- If the skewness is less than -1 or greater than 1, the data are highly skewed.
- As per the above, there is no skewness in the data. All the value are between -0.5 and 0.5
- The value of Blair is little higher
- Overall, the data is symmetrical.

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

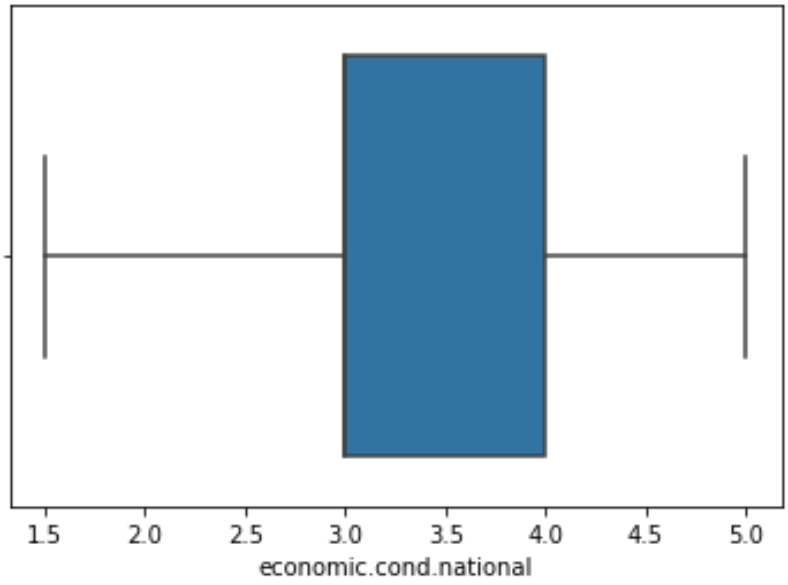
FIGURE 1: DISTRIBUTION AND BOXPLOT



INFERENCE FOR FIGURE 1

1. Age:
    1.1. The data is distributed normally
    1.2. There are no outliers present.
2. All the variables are continuously distributed with no skewness.
3. There are outliers in variables “economic cond national and “economic cond household”, which are again very few.
4. Other values do not have any outliers.

FIGURE 2: REMOVING OUTLIERS



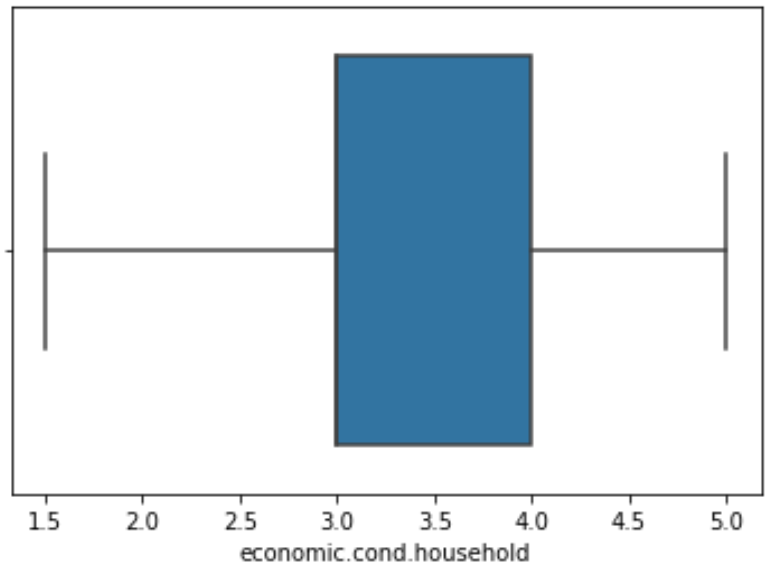


FIGURE 3: BAR GRAPH FOR VOTE

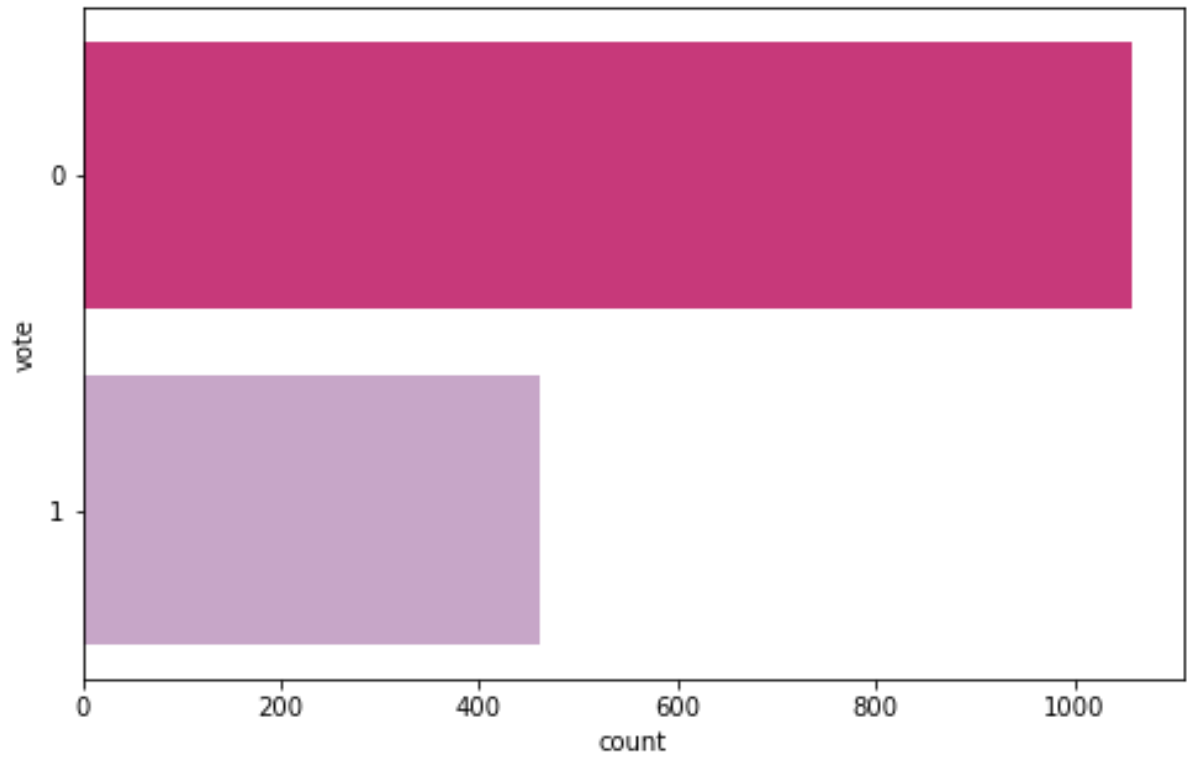


TABLE 8: VALUE COUNTS FOR VOTE

	vote
0	1057
1	460

INFERENCE FOR VOTE

1. Labour party has higher number of votes. It has more than double the votes of conservative party.
2. Labour party has 1057 votes.
3. Conservative party has 460 votes.

FIGURE 4: BAR GRAPH FOR GENDER

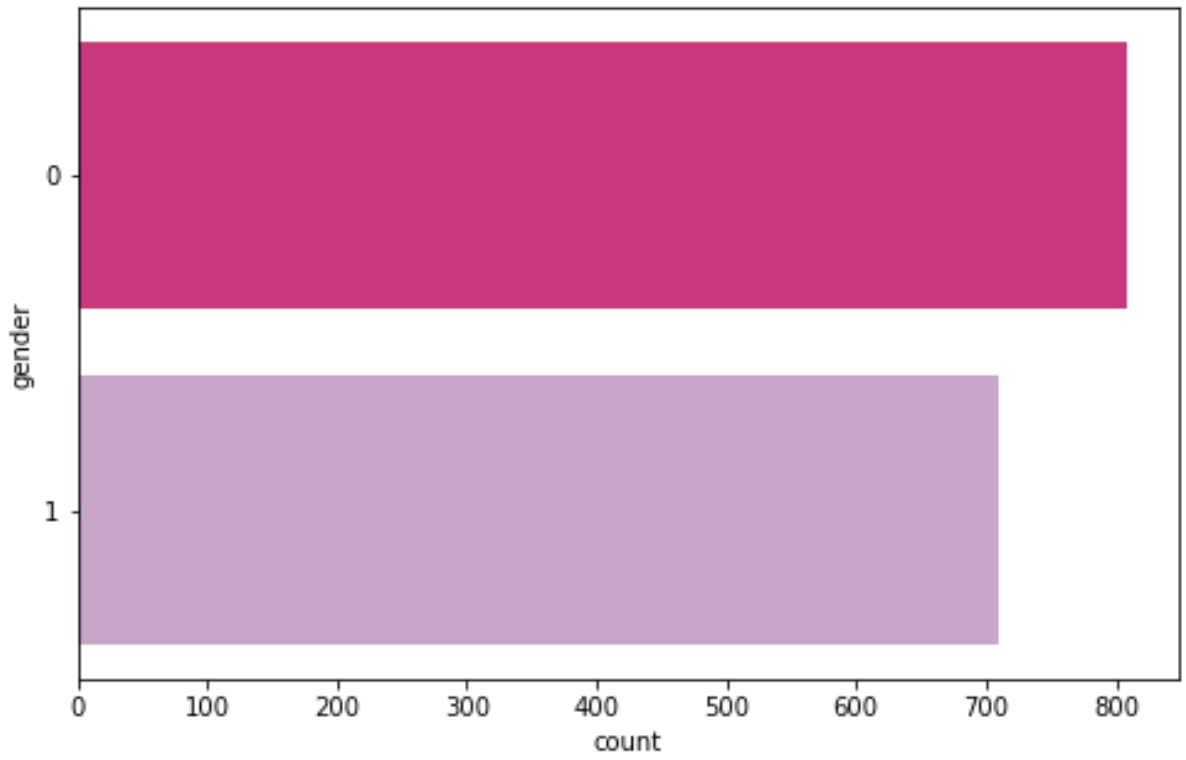


TABLE 9: VALUE COUNTS FOR GENDER

	gender
0	808
1	709

INFERENCE FOR GENDER

1. As per the Gender Bar graph, we can see that there are more males (0) than females (1)
2. There are 808 males and 709 females.

FIGURE 5: BAR GRAPH FOR ECONOMIC COND NATIONAL

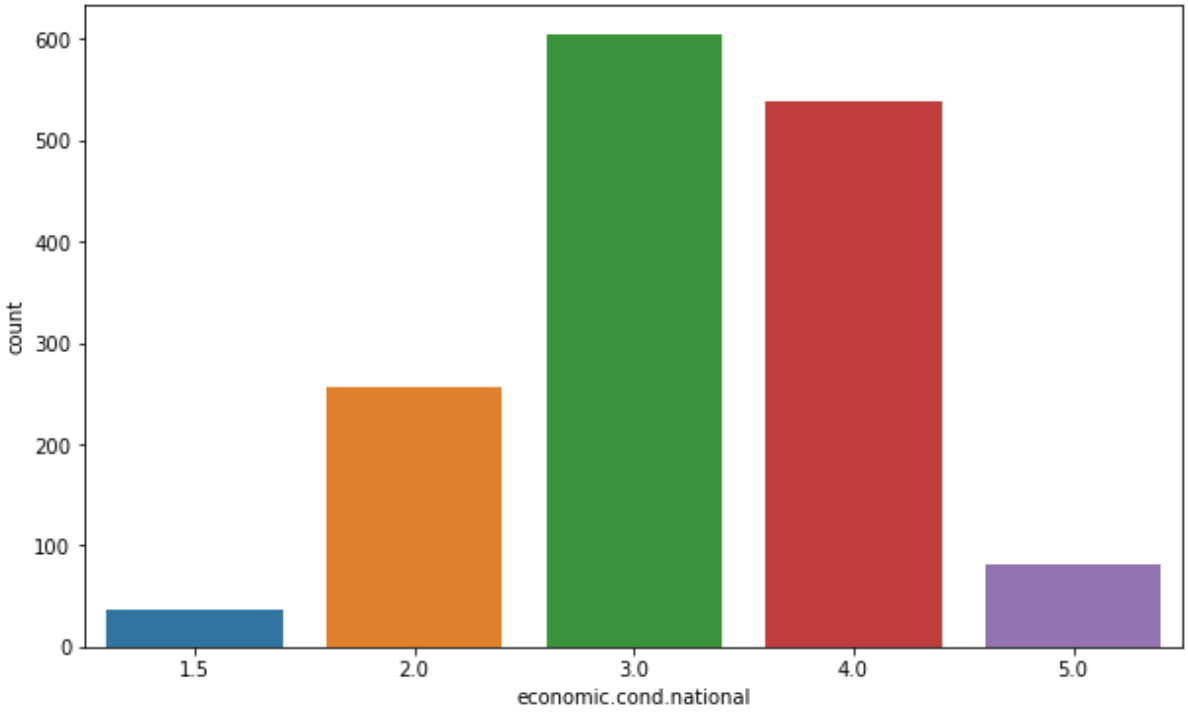


TABLE 10: VALUE COUNTS FOR ECONOMIC COND NATIONAL

	economic.cond.national
3.0	604
4.0	538
2.0	256
5.0	82
1.5	37

INFERENCE FOR ECONOMIC COND NATIONAL

1. The top 2 variables are 3 and 4.
2. 1 has the least value which is 37.
3. 3 has the highest value which is 604.
4. 3 is slightly higher than the 2nd highest variable 4 whose value is 538.
5. The average score of 'economic. cond. national' is 3.24 (Table 4)

FIGURE 6: BAR GRAPH FOR ECONOMIC COND HOUSEHOLD

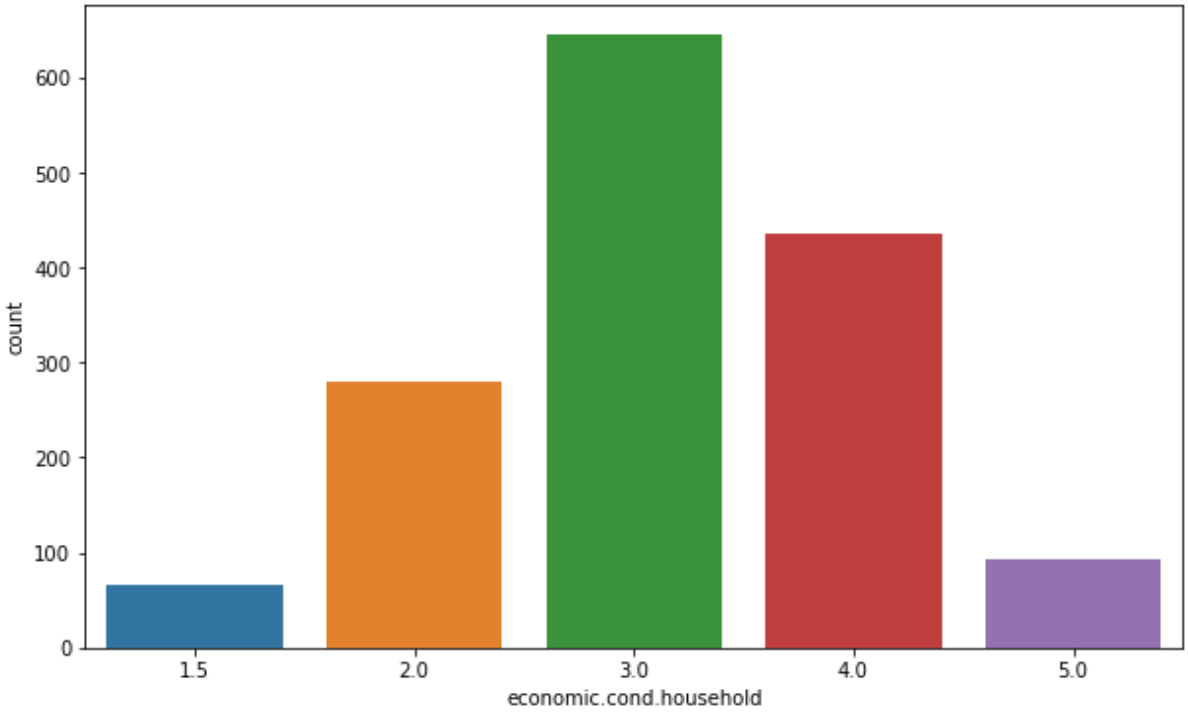


TABLE 11: VALUE COUNTS FOR ECONOMIC COND HOUSEHOLD

	economic.cond.household
3.0	645
4.0	435
2.0	280
5.0	92
1.5	65

INFERENCE FOR ECONOMIC COND HOUSEHOLD

1. The top 2 variables are 3 and 4.
2. 1 has the least value which is 65.
3. 3 has the highest value which is 645.
4. 3 is moderately higher than the 2nd highest variable 4 whose value is 435.
5. The average score of 'economic. cond. household' is 3.13 (Table 4)

FIGURE 7: BAR GRAPH FOR BLAIR

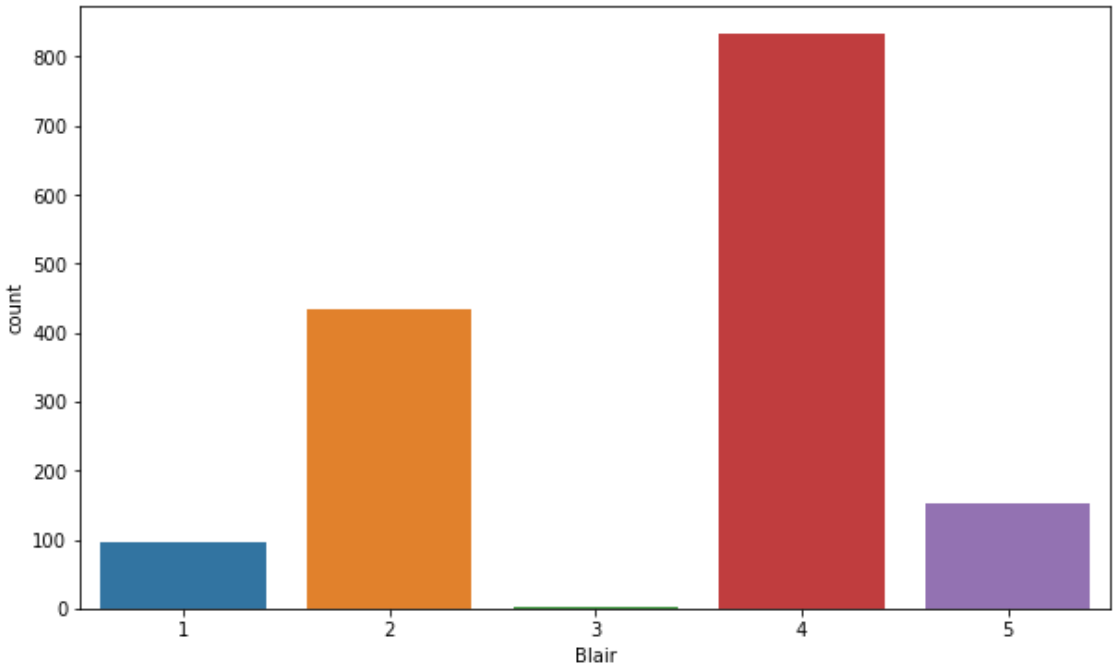


TABLE 12: VALUE COUNTS FOR BLAIR

	Blair
4	833
2	434
5	152
1	97
3	1

INFERENCE FOR BLAIR

1. The top 2 variables are 2 and 4.
2. 3 has the least value which is 1.
3. 4 has the highest value which is 833.
4. 4 is much higher than the 2nd highest variable 2 whose value is 434
5. The average score of 'Blair' is 3.33. (Table 4)

FIGURE 8: BAR GRAPH FOR HAGUE

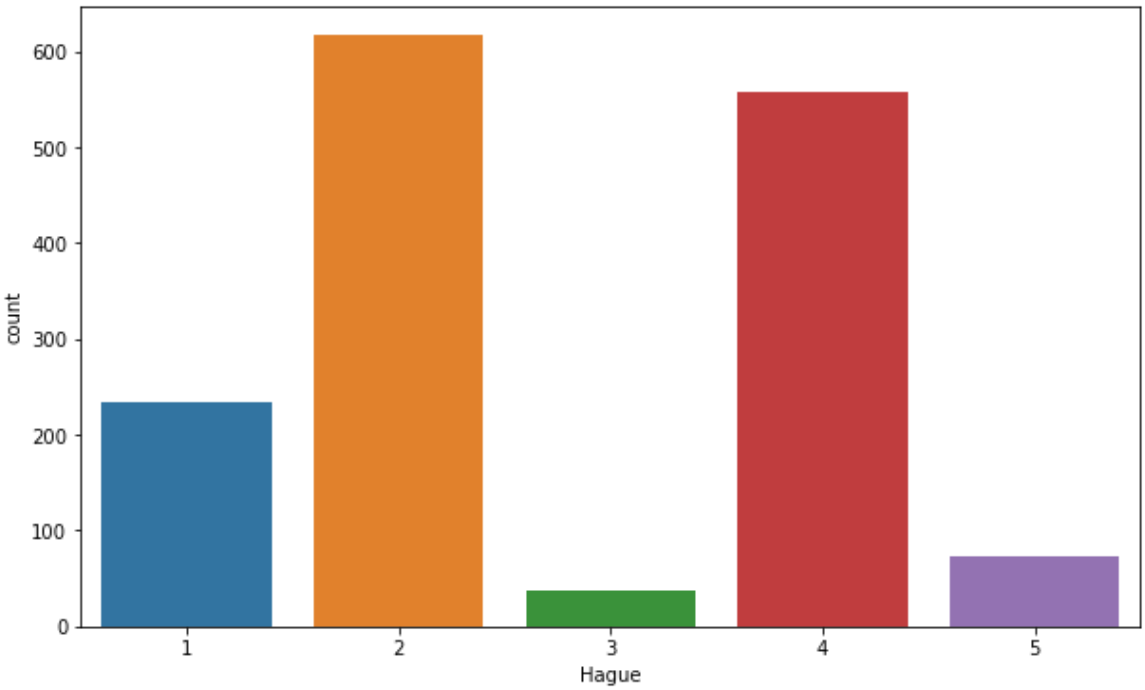


TABLE 13: VALUE COUNTS FOR HAGUE

	Hague
2	617
4	557
1	233
5	73
3	37

INFERENCE FOR HAGUE

1. The top 2 variables are 2 and 4.
2. 3 has the least value which is 37.
3. 2 has the highest value which is 617.
4. 2 is slightly higher than the 2nd highest variable 4 whose value is 557.

5. The average score of 'Blair' is 2.74 (Table 4)

FIGURE 9: BAR GRAPH FOR EUROPE

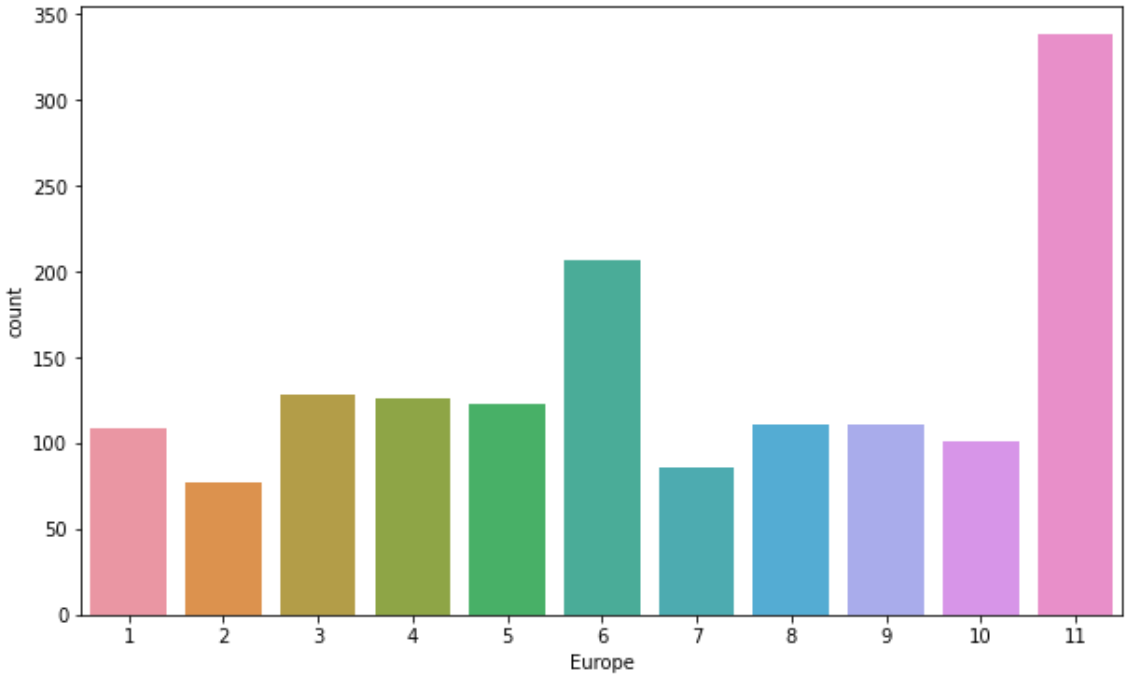


TABLE 14: VALUE COUNTS FOR EUROPE

	Europe
11	338
6	207
3	128
4	126
5	123
9	111
8	111
1	109
10	101
7	86
2	77

INFERENCE FOR EUROPE

1. The top 2 variables are 11 and 6.
2. 2 has the least value which is 77.
3. 11 has the highest value which is 338.
4. 11 is moderately higher than the 2nd highest variable 6 whose value is 207.
5. The average score of 'Europe' is 6.74. (Table 4).

FIGURE 10: BAR GRAPH FOR POLITICAL KNOWLEDGE

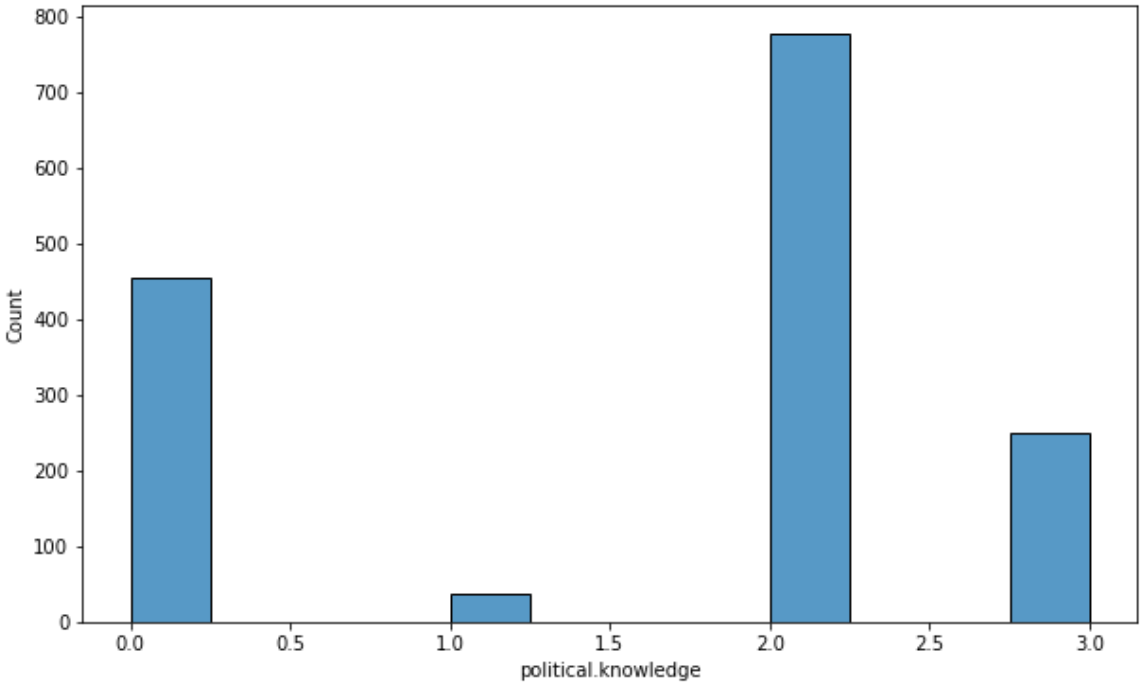


TABLE 15: VALUE COUNTS FOR POLITICAL KNOWLEDGE

	political.knowledge
2	776
0	454
3	249
1	38

INFERENCE FOR POLITICAL KNOWLEDGE

1. The top 2 variables are 2 and 0.
2. 1 has the least value which is 38.
3. 2 has the highest value which is 776.
4. 2 is much higher than the 2nd highest variable 0 whose value is 454.
5. We can see that, 454 out of 1517 people do not have any knowledge of parties' positions on European integration which is 29.93% of the total population.
6. The average score of 'Europe' is 6.74.(Table 4).

FIGURE 11: STRIP PLOT FOR VOTE & AGE

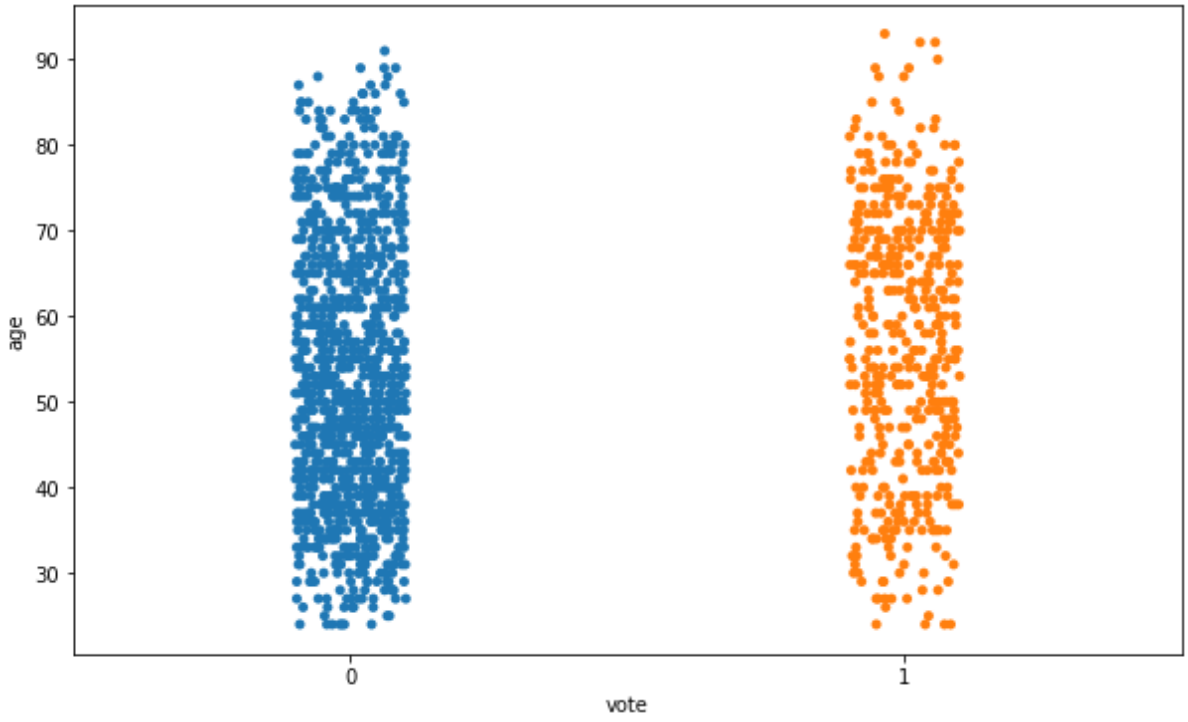


TABLE 16: COUNT FOR VOTE & GENDER

vote	gender	0
1	1	203
	0	257
0	1	506
	0	551

INFERENCE FOR TABLE 16 AND FIGURE 11

1. We can clearly see that; the labour party has got more votes than the conservative party. (Figure 11)
2. In every age group, the labour party has got more votes than the conservative party (1). (Table 16)
3. Female votes are considerably higher than the male votes in both parties as per the above output. (Table 16)
4. In both genders, the labour party has got more votes than the conservative party. (Table 16).

FIGURE 12: STRIP PLOT FOR VOTE & ECONOMIC COND NATIONAL

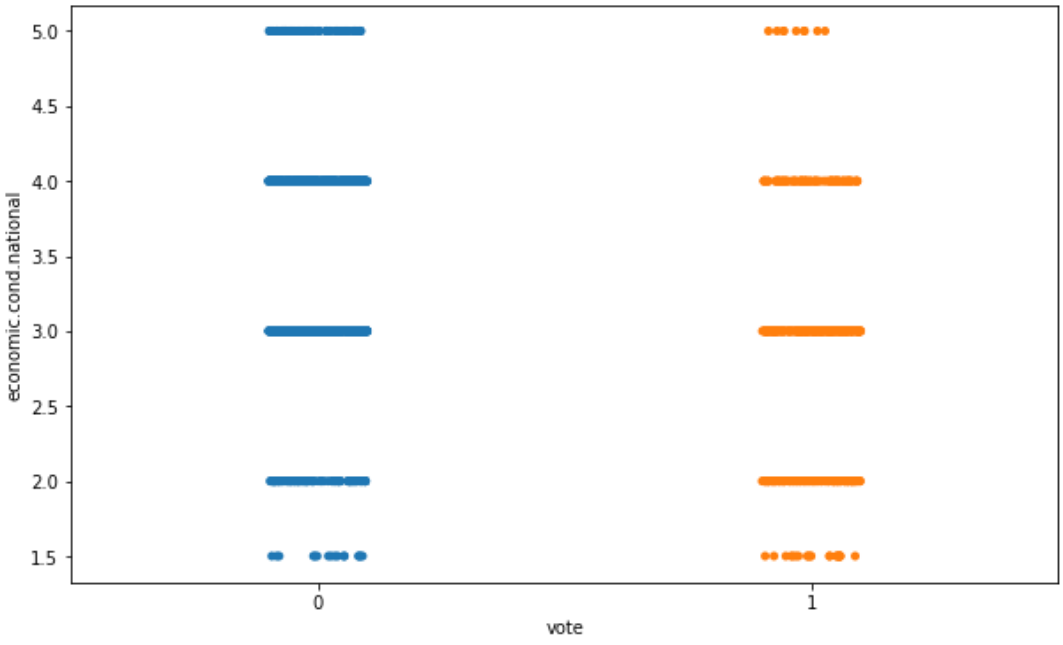


TABLE 17: COUNT FOR VOTE & ECONOMIC COND NATIONAL

vote	economic.cond.national	0
1	5.0	9
0	1.5	16
1	1.5	21
0	5.0	73
1	4.0	91
0	2.0	116
1	2.0	140
	3.0	199
0	3.0	405
	4.0	447

INFERENCE FOR TABLE 17 AND FIGURE 12

1. Labour party has higher votes overall.
2. Out of 82 people who gave a score of 5, 73 people have voted for the labour party.
3. Out of 538 people who gave a score of 4, 447 people have voted for the labour party. This is the highest set of people in the labour party.
4. Out of 604 people who gave a score of 3, 405 people have voted for the labour party. This is the 2nd highest set of people in the labour party. The remaining 199 people who have voted for the conservative party is the highest set of people in that party.
5. Out of 256 people who gave a score of 2, 116 people have voted for the labour party. 140 people have voted for the conservative party. This is the instance where the conservative party has got more votes than the labour party.
6. Out of 37 people who gave a score of 1, 16 people have voted for the labour party. 21 people have voted for the conservative party.
7. The score of 3, 4 and 5 have more votes in the labour party.
8. The score of 1 and 2 have more votes in the conservative party.

FIGURE 13: STRIP PLOT FOR VOTE & ECONOMIC COND HOUSEHOLD

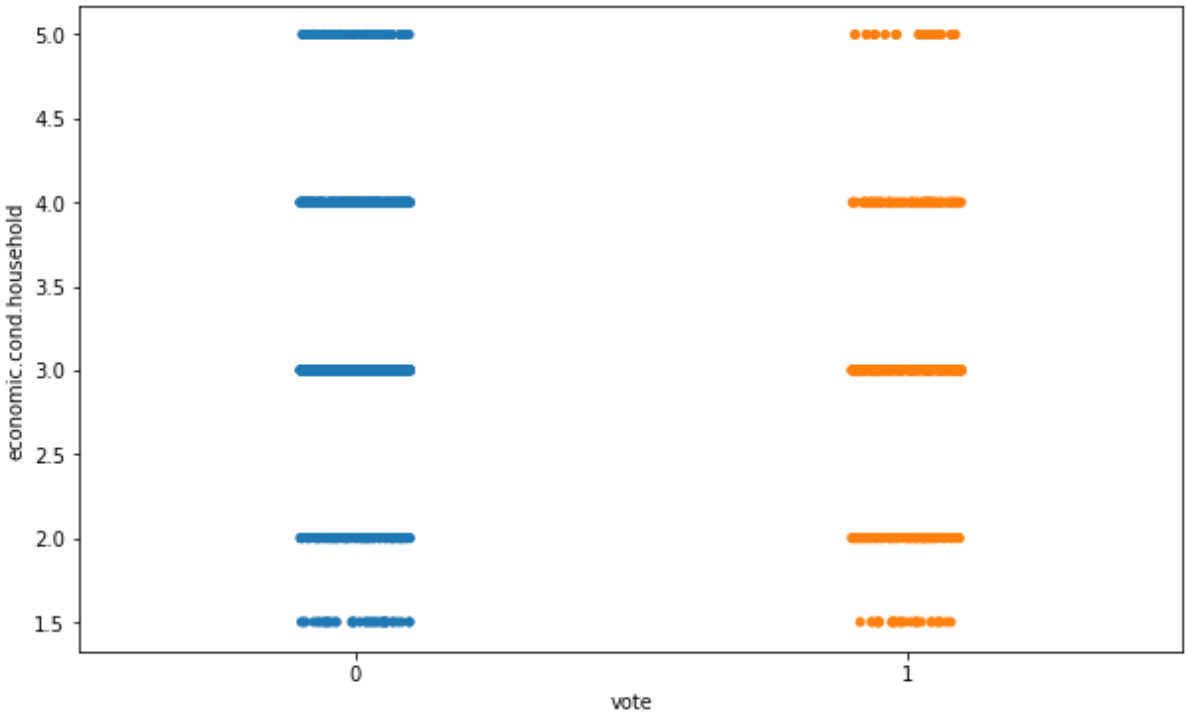


TABLE 18: COUNT FOR VOTE & ECONOMIC COND HOUSEHOLD

		0
vote	economic.cond.household	
1	5.0	23
	1.5	28
0	1.5	37
	5.0	69
1	4.0	86
	2.0	126
0	2.0	154
1	3.0	197
0	4.0	349
	3.0	448

INFERENCE FOR TABLE 18 AND FIGURE 13

1. Labour party has higher votes overall.
2. Out of 92 people who gave a score of 5, 69 people have voted for the labour party.
3. Out of 435 people who gave a score of 4, 349 people have voted for the labour party. This is the 2nd highest set of people in the labour party.
4. Out of 645 people who gave a score of 3, 448 people have voted for the labour party. This is the highest set of people in the labour party. The remaining 197 people who have voted for the conservative party is the highest set of people in that party.
5. Out of 280 people who gave a score of 2, 154 people have voted for the labour party. 126 people have voted for the conservative party.
6. Out of 65 people who gave a score of 137 people have voted for the labour party. 28 people have voted for the conservative party.
7. In all the instances, the labour party have more votes than the conservative party.

FIGURE 14: STRIP PLOT FOR VOTE & BLAIR

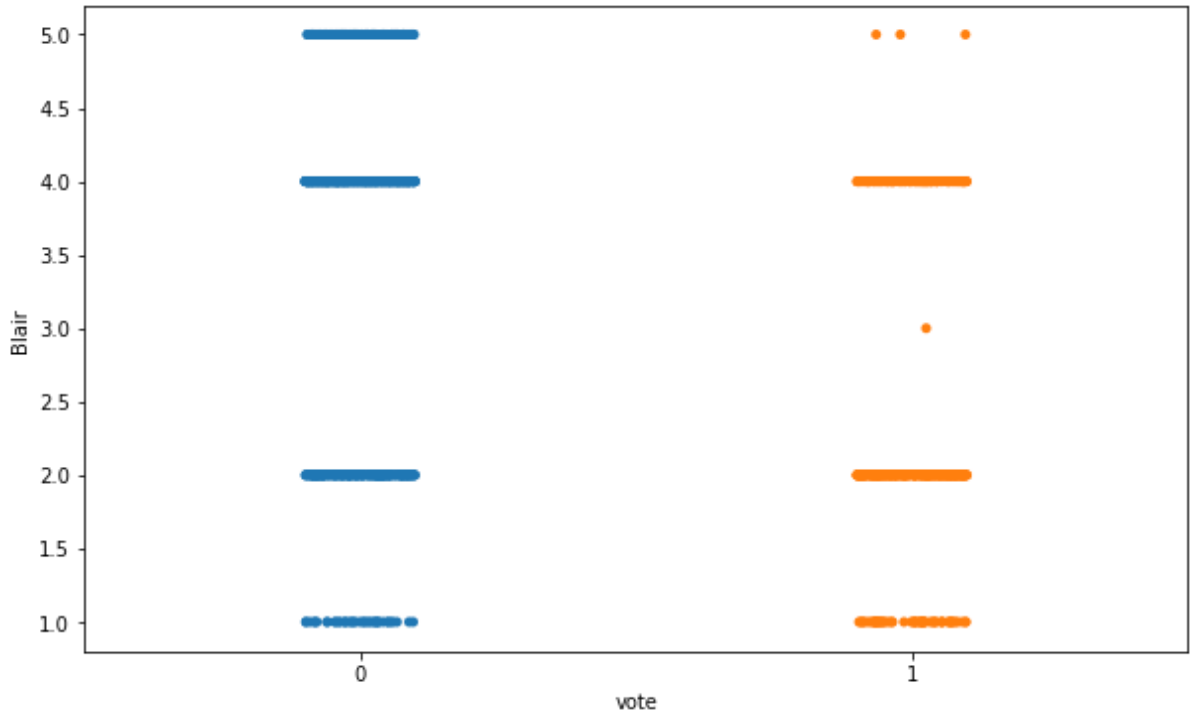


TABLE 19: COUNT FOR VOTE & BLAIR

vote	Blair	0
1	3	1
	5	3
0	1	38
1	1	59
0	5	149
1	4	157
0	2	194
1	2	240
0	4	676

INFERENCE FOR TABLE 19 AND FIGURE 14

1. Labour party has higher votes overall.
2. Out of 152 people who gave a score of 5,149 people have voted for the labour party. The remaining 3 people, despite giving a score of 5 to the labour leader, have chosen to vote for the conservative party.
3. Out of 833 people who gave a score of 4,676 people have voted for the labour party. The remaining 157 people, despite giving a score of 4 to the labour leader, have chosen to vote for the conservative party.
4. Only 1 person has given a score of 3 and that person has voted for the conservative party.
5. Out of 434 people who gave a score of 2,240 people have voted for the conservative party. The remaining 194 people, despite giving an unsatisfactory score of 2 to the labour leader, have chosen to vote for the labour party.
6. Out of 97 people who gave a score of 159 people have voted for the conservative party. The remaining 38 people, despite giving the lowest score of 1 to the labour leader, have chosen to vote for the labour party.
7. The score of 4 and 5 have more votes in the labour party.
8. The score of 1, 2 and 3 have more votes in the conservative party.

FIGURE 15: STRIP PLOT FOR VOTE & HAGUE

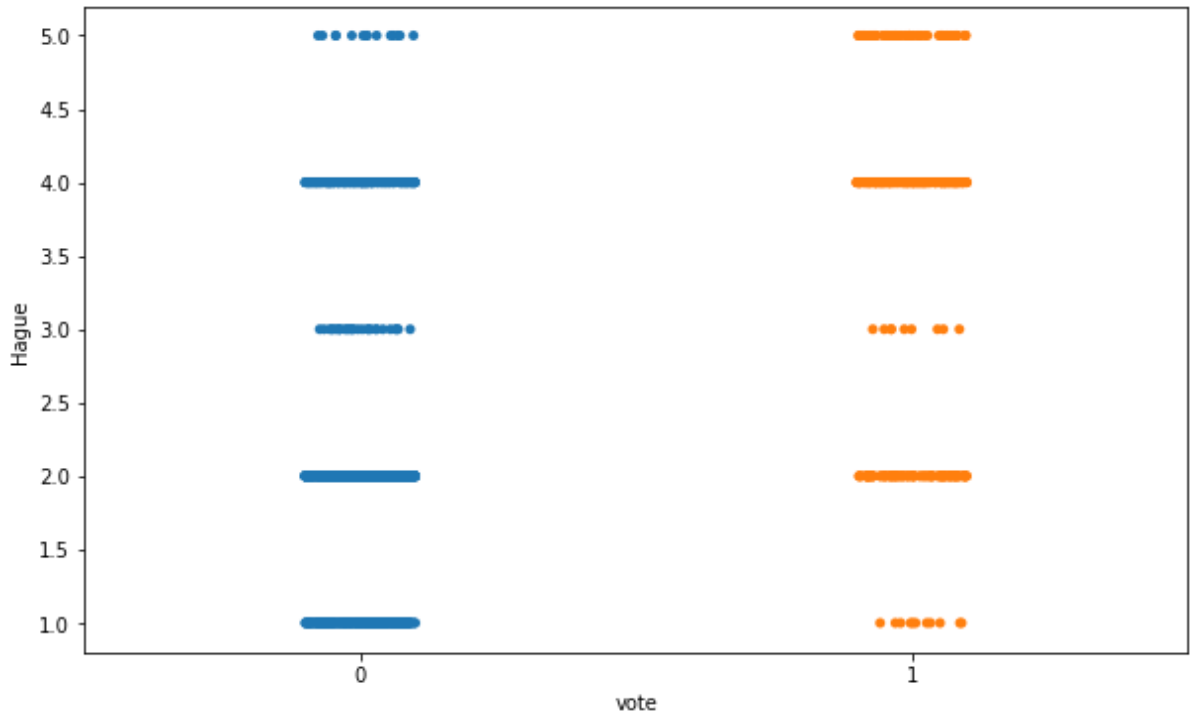


TABLE 20: COUNT FOR VOTE & HAGUE

		0
vote	Hague	
1	3	9
	1	11
0	5	14
	3	28
1	5	59
	2	95
0	1	222
	4	271
1	4	286
0	2	522

INFERENCE FOR TABLE 20 AND FIGURE 15

1. Labour party has higher votes overall.
2. Out of 73 people who gave a score of 5,59 people have voted for the conservative party. The remaining 14 people, despite giving a score of 5 to the conservative leader, have chosen to vote for the labour party.
3. people, despite giving a score of 4 to the conservative leader, have chosen to vote for the labour party.
4. Out of 37 people who gave a score of 328 have voted for the labour party. The remaining 9, despite giving an average score of 3 to the conservative party, have chosen to vote for the conservative party.
5. Out of 617 people who gave a score of 2,522 people have voted for the labour party. The remaining 95 people, despite giving an unsatisfactory score of 2 to the conservative leader, have chosen to vote for the conservative party.
6. Out of 233 people who gave a score of 1,222 people have voted for the labour party. The remaining 11 people, despite giving the lowest score of 1 to the conservative leader, have chosen to vote for the conservative party.
7. The score of 4 and 5 have more votes in the conservative party, although in 4, the votes are almost equal in both the parties. Conservative party gets slightly higher.
8. The score of 1, 2 and 3 have more votes in the labour party. Still, a significant percentage of people who gave a bad score to the conservative leader still chose to vote for 'Hague'.

FIGURE 16: STRIP PLOT FOR VOTE & EUROPE

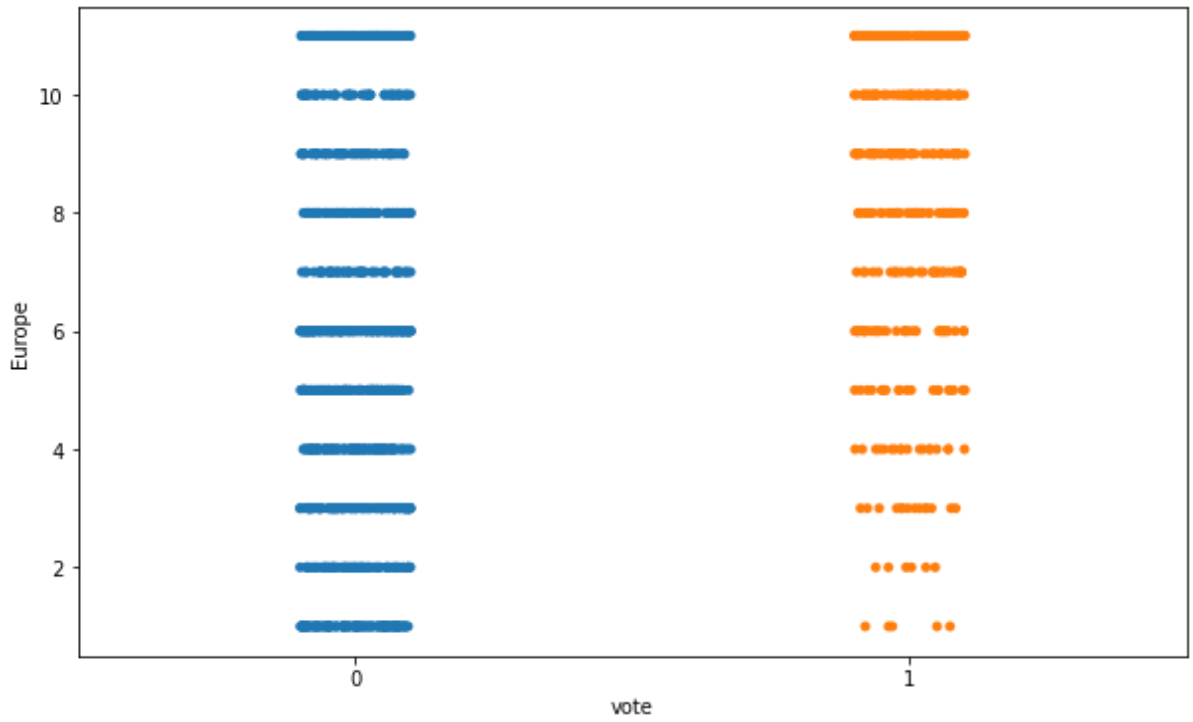


TABLE 21: COUNT FOR VOTE & EUROPE

vote	Europe	0
1	1	5
	2	6
	3	14
	4	18
	5	20
	7	32
	6	35
0	10	47
1	8	48
	10	54
0	7	54
	9	55
1	9	56
0	8	63
	2	71
	5	103
	1	104
	4	108
	3	114
	11	166
	6	172
1	11	172

INFERENCE FOR TABLE 21 AND FIGURE 16

1. Out of 338 people who gave a score of 11,166 people have voted for the labour party and 172 people have voted for the conservative party.
2. People who gave score of 7 to 10 have voted for labour and conservative almost equally. Conservative party seem to be slightly higher in these instances.
3. Out of 207 people who gave a score of 6,172 people have voted for the labour party and 35 people have voted for the conservative party.
4. People who gave a score of 1 to 6 have predominantly voted for the labour party. As we can see, there are a total of 770 people who have given scores from 1 to 6. Out of 770 people, 672 people have voted for the labour party. So, 87.28% of the people have chosen labour party.
5. So, we can infer that lower the 'Eurosceptic' sentiment, higher the votes for labour party.

FIGURE 17: STRIP PLOT FOR VOTE & POLITICAL KNOWLEDGE

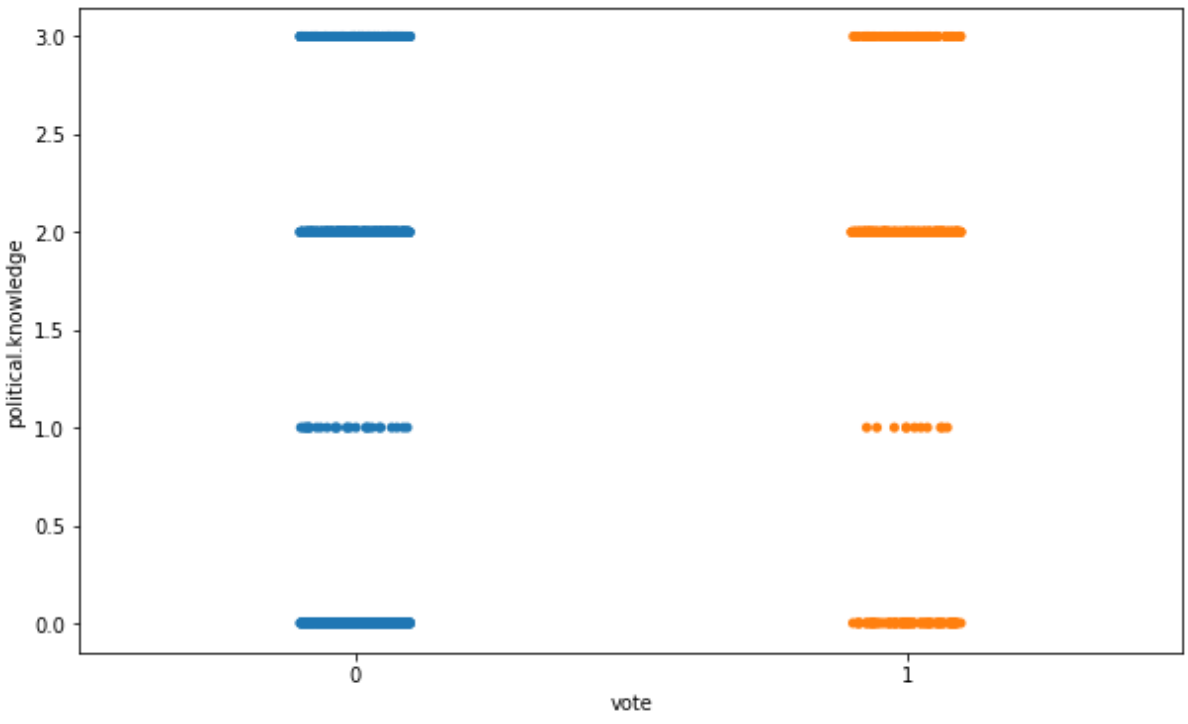


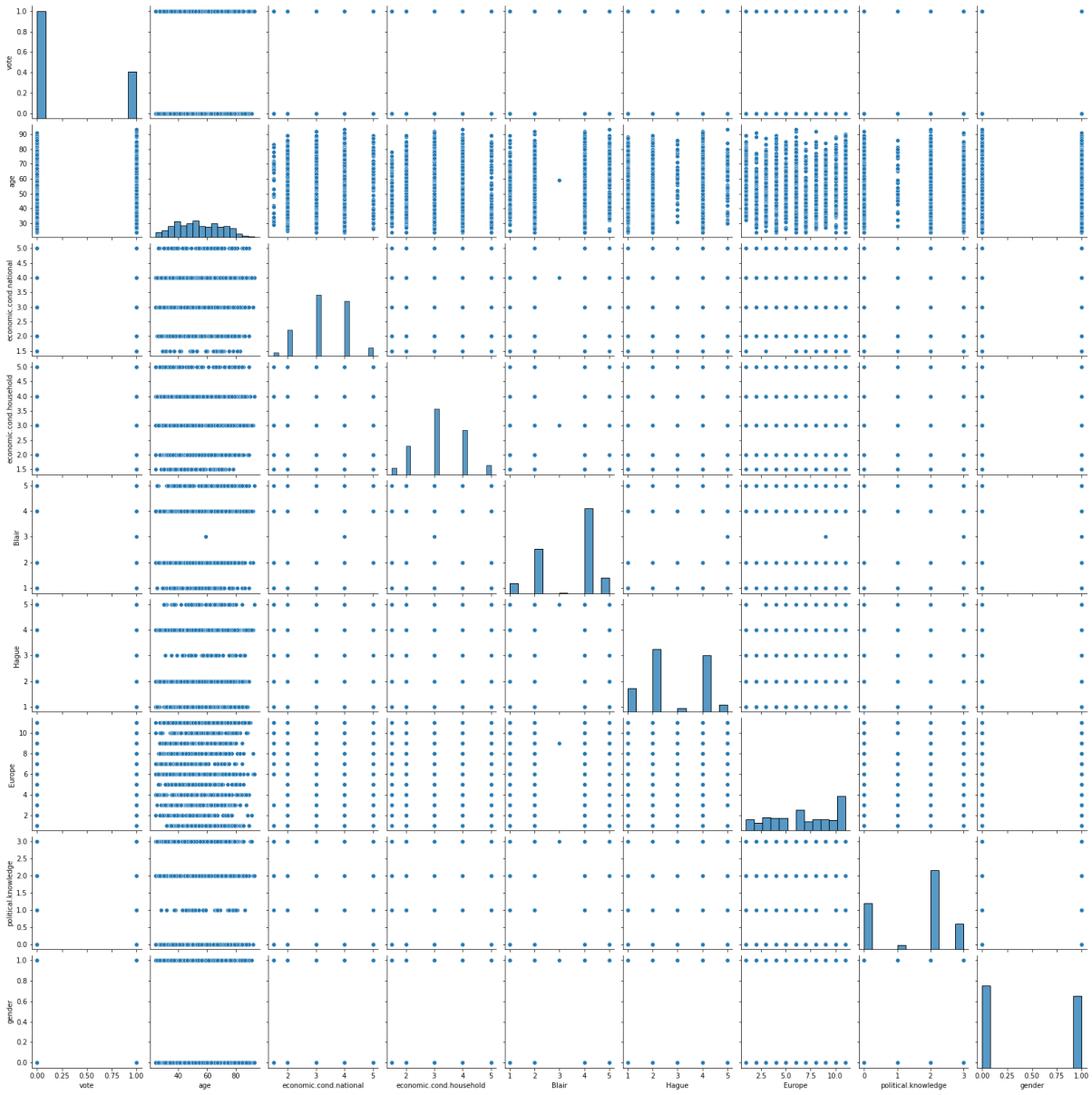
TABLE 22: COUNT FOR VOTE & POLITICAL KNOWLEDGE

vote	political.knowledge	0
1	1	11
0	1	27
1	3	72
	0	94
0	3	177
1	2	283
0	0	360
	2	493

INFERENCE FOR TABLE 22 AND FIGURE 17

1. Out of 249 people who gave a score of 3,177 people have voted for the labour party and 72 people have voted for the conservative party.
2. Out of 776 people who gave a score of 2,493 people have voted for the labour party and 283 people have voted for the conservative party.
3. Out of 38 people who gave a score of 1,27 people have voted for the labour party and 11 people have voted for the conservative party.
4. Out of 454 people who gave a score of 0, 360 people have voted for the labour party and 94 people have voted for the conservative party.
5. We can see that, in all instances, labour party gets the higher number of votes.
6. Out of 1,517 people, 454 people gave a score of 0. So, this means that, 29.93% of the people are casting their votes without any political knowledge.

FIGURE 18: PAIR PLOT



INFERENCE FOR FIGURE 18

1. Pair plot is a combination of histograms and scatter plots.
2. From the histogram, we can see that, the 'Blair', 'Europe' and 'political. knowledge' variables are slightly left skewed.
3. All other variables seem to be normally distributed.
4. From the scatter plots, we can see that, there is mostly no correlation between the variables.

TABLE 23: CORRELATION TABLE

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
vote	1.000000	0.109274	-0.304028	-0.176807	-0.426606	0.468186	0.384612	0.111589	-0.034464
age	0.109274	1.000000	0.022283	-0.044403	0.032084	0.031144	0.064562	-0.046598	-0.017933
economic.cond.national	-0.304028	0.022283	1.000000	0.344462	0.323603	-0.199175	-0.206605	-0.029273	0.054950
economic.cond.household	-0.176807	-0.044403	0.344462	1.000000	0.216653	-0.099644	-0.112186	-0.040521	0.033001
Blair	-0.426606	0.032084	0.323603	0.216653	1.000000	-0.243508	-0.295944	-0.021299	0.067624
Hague	0.468186	0.031144	-0.199175	-0.099644	-0.243508	1.000000	0.285738	-0.029906	-0.028309
Europe	0.384612	0.064562	-0.206605	-0.112186	-0.295944	0.285738	1.000000	-0.151197	-0.076059
political.knowledge	0.111589	-0.046598	-0.029273	-0.040521	-0.021299	-0.029906	-0.151197	1.000000	0.156923
gender	-0.034464	-0.017933	0.054950	0.033001	0.067624	-0.028309	-0.076059	0.156923	1.000000

FIGURE 19: CORRELATION PLOT/ HEATMAP

INFERENCE FOR FIGURE 19

1. We can see that, mostly there is no correlation in the dataset through this matrix. There are some variables that are moderately positively correlated and some that are slightly negatively correlated.
2. ‘economic. cond. national’ with ‘economic. cond. household’ have moderate positive correlation.
3. ‘Blair' with 'economic. cond.national' and ‘economic.cond.household’ have moderate positive correlation.
4. ‘Europe’ with ‘Hague’ have moderate positive correlation.
5. 'Hague' with 'economic. cond. national' and 'Blair' have moderate negative correlation.
6. 'Europe' with 'economic. cond.national' and 'Blair' have moderate negative correlation.

Q.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.

1. The dataset contains features highly varying in magnitudes, units and range between the 'age' column and other columns.
2. But since, most of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem.
3. If left alone, these algorithms only take in the magnitude of features neglecting the units.
4. The results would vary greatly between different units, 1 km and 1000 metres.
5. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.
6. To supress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.
7. In this case, we have a lot of encoded, ordinal, categorical and continuous variables. So, we use the Min Max technique to scale the data.

TABLE 24: ENCODING THE DATASET

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
1	43	3.0	3.0	4	1	2	2	1	0
2	36	4.0	4.0	4	4	5	2	1	1
3	35	4.0	4.0	5	2	3	2	1	1
4	24	4.0	2.0	2	1	4	0	1	0
5	41	2.0	2.0	1	1	6	2	1	1

TABLE 25: DATASET AFTER SCALING

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
1	0.275362	0.428571	0.428571	0.75	0.00	0.1	0.666667	1	0
2	0.173913	0.714286	0.714286	0.75	0.75	0.4	0.666667	1	1
3	0.159420	0.714286	0.714286	1.00	0.25	0.2	0.666667	1	1
4	0.000000	0.714286	0.142857	0.25	0.00	0.3	0.000000	1	0
5	0.246377	0.142857	0.142857	0.00	0.00	0.5	0.666667	1	1

ABOUT TRAIN AND TEST SPLIT

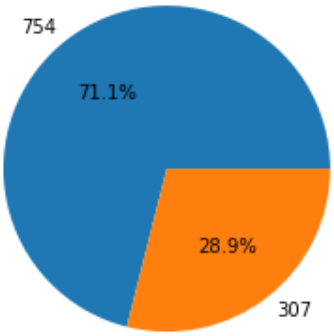
1. **Train Dataset:** It is used to fit the machine learning model.

2. **Test Dataset:** It is used to evaluate the fit machine learning model.

- The train-test split is a technique for evaluating the performance of a machine learning algorithm. The procedure involves taking a dataset and dividing it into two subsets..
- The data is divided into 2 subsets, training and testing set. Earlier, we have extracted the target variable ‘vote’ in a separate vector for subsets. Random state chosen as 1.

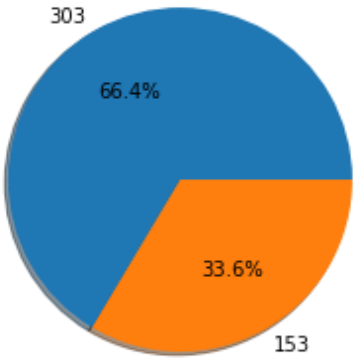
FIGURE 20: SPLITTING THE DATA

TRAIN DATA:




```
0    71.065033
1    28.934967
```

TEST DATA:




```
0    66.447368
1    33.552632
```

Q.1.4. Apply Logistic Regression and LDA (Linear Discriminant Analysis) Interpret the inferences of both model's. Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting).

LOGISTIC REGRESSION

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.

TABLE 26: CLASSIFICATION REPORT FOR TRAIN DATA

```
[[197 110]
 [ 66 688]]
              precision    recall  f1-score   support

           0       0.75      0.64      0.69       307
           1       0.86      0.91      0.89       754

    accuracy                           0.83      1061
   macro avg       0.81      0.78      0.79      1061
weighted avg       0.83      0.83      0.83      1061
```

FIGURE 21: CONFUSION MATRIX FOR TRAIN DATA

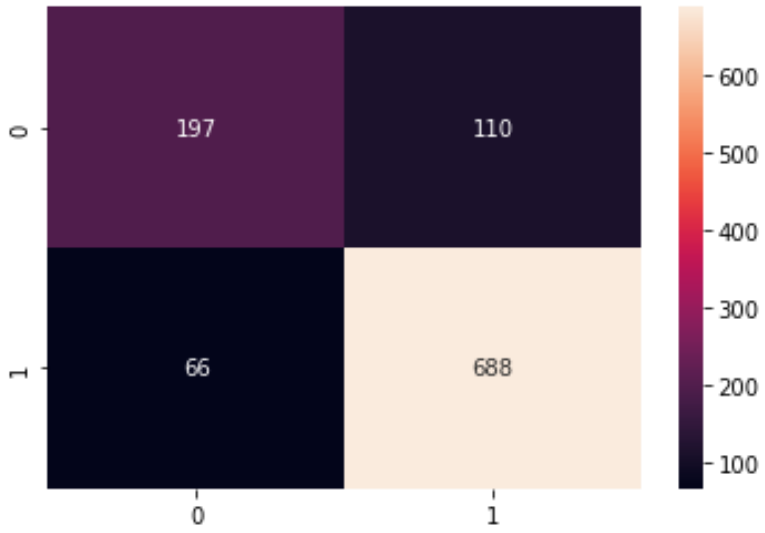


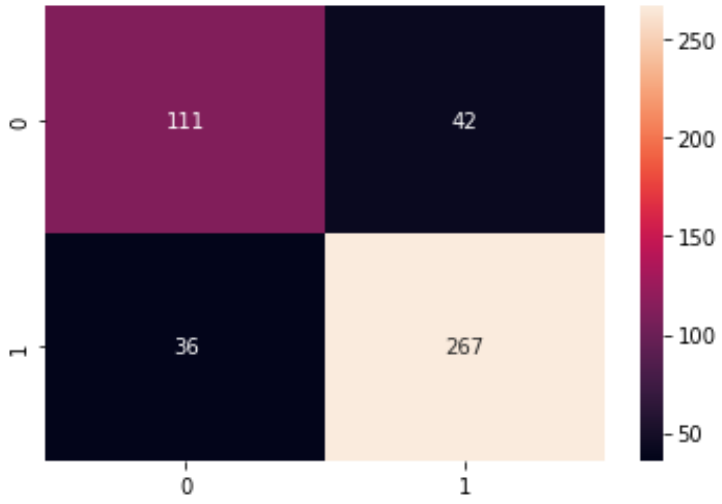
TABLE 27: CLASSIFICATION REPORT FOR TEST DATA

```
[[111  42]
 [ 36 267]]
              precision    recall  f1-score   support

           0       0.76      0.73      0.74       153
           1       0.86      0.88      0.87       303

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

FIGURE 22: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR LOGISTIC REGRESSION

1. 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.
2. 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.
3. This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
4. Overall, the metrics are good fit.

LOGISTIC REGRESSION					
#		Train Data		Test Data	
1	True Positive	197		111	
2	True Negative	688		267	
3	False Positive	110		42	
4	False Negative	66		36	
5	AUC score	89%		88.30%	
6	Accuracy	83%		83%	
		Conservative	Labour	Conservative	Labour
7	Precision	75%	86%	76%	86%
8	Recall	64%	91%	73%	88%
9	F1 score	69%	89%	74%	87%

LINEAR DISCRIMINANT ANALYSIS

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.

TABLE 28: CLASSIFICATION REPORT FOR TRAIN DATA

```
[[200  69]
 [107 685]]
              precision    recall  f1-score   support

           0       0.65      0.74      0.69       269
           1       0.91      0.86      0.89       792

    accuracy                           0.83      1061
   macro avg       0.78      0.80      0.79      1061
weighted avg       0.84      0.83      0.84      1061
```

FIGURE 23: CONFUSION MATRIX FOR TRAIN DATA

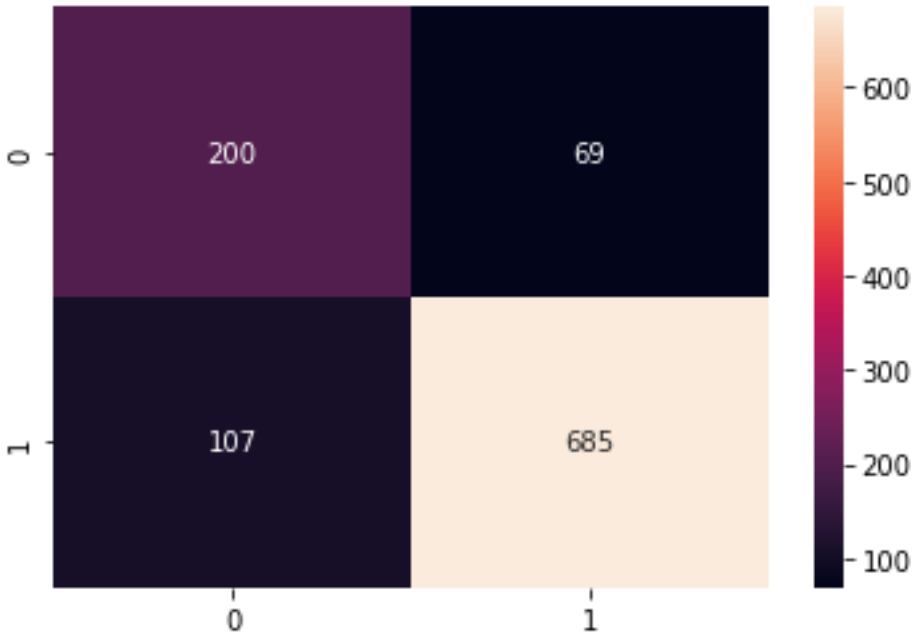


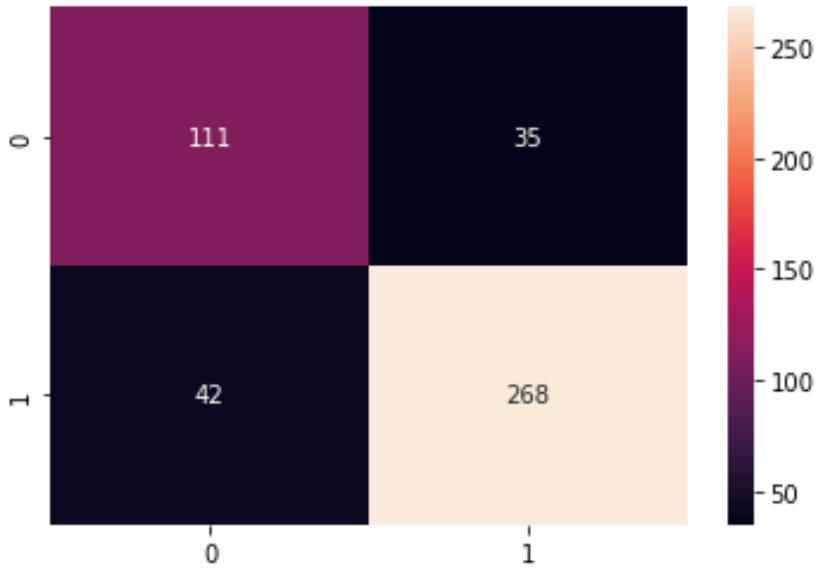
TABLE 29: CLASSIFICATION REPORT FOR TEST DATA

```
[[111  35]
 [ 42 268]]
              precision    recall  f1-score   support

           0       0.73      0.76      0.74       146
           1       0.88      0.86      0.87       310

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

FIGURE 24: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR LDA

LINEAR DISCRIMINANT ANALYSIS					
#		Train Data		Test Data	
1	True Positive	200		111	
2	True Negative	685		268	
3	False Positive	69		35	
4	False Negative	107		42	
5	AUC score	89%		88%	
6	Accuracy	83%		83%	
		0 (Conservative)	1 (Labour)	0 (Conservative)	1 (Labour)
7	Precision	65%	91%	73%	88%
8	Recall	74%	86%	76%	86%
9	F1 score	69%	89%	74%	87%

1. Logistic Regression is performing slightly better than LDA model.
2. 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.
3. This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
4. Overall, the metrics are good fit.

Q.1.5. Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model. Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting).

K-NEAREST NEIGHBOURS MODEL

The k-nearest neighbours (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 Neighbours model:

• Scaled dataset is used to build KNN model, as it is distance-based algorithm.

• Fitting the KNN model which is imported from Sklearn neighbours model which considers default neighbours (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.

TABLE 30: CLASSIFICATION REPORT FOR TRAIN DATA

```
[[221  65]
 [ 86 689]]
              precision    recall  f1-score   support

           0       0.72      0.77      0.75       286
           1       0.91      0.89      0.90       775

    accuracy                           0.86      1061
   macro avg       0.82      0.83      0.82      1061
weighted avg       0.86      0.86      0.86      1061
```

FIGURE 25: CONFUSION MATRIX FOR TRAIN DATA

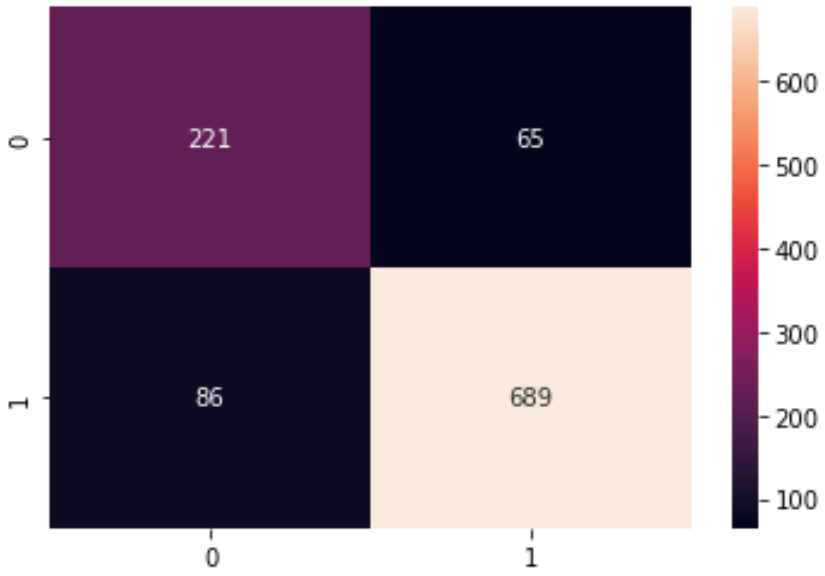


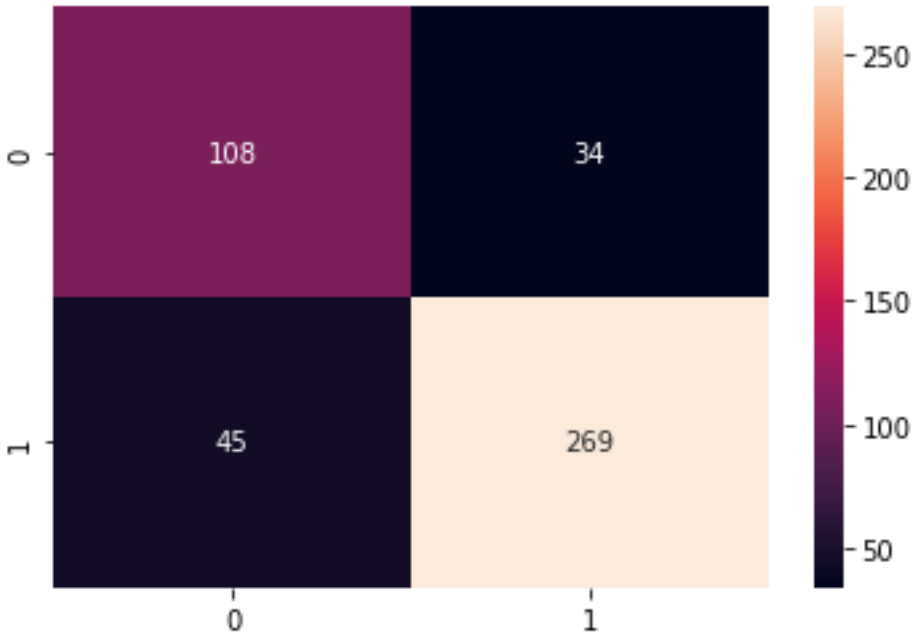
TABLE 31: CLASSIFICATION REPORT FOR TEST DATA

```
[[108  34]
 [ 45 269]]
              precision    recall  f1-score   support

           0       0.71      0.76      0.73       142
           1       0.89      0.86      0.87       314

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

FIGURE 26: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR KNN MODEL

K-NEAREST NEIGHBOURS					
#		Train Data		Test Data	
1	True Positive	221		104	
2	True Negative	689		269	
3	False Positive	65		34	
4	False Negative	86		45	
5	AUC score	89%		89%	
6	Accuracy	86%		83%	
		0 (Conservative)	1 (Labour)	0 (Conservative)	1 (Labour)
7	Precision	72%	91%	71%	89%
8	Recall	77%	89%	76%	86%
9	F1 score	75%	90%	73%	87%

• 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 CLASSIFIER

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.

TABLE 32: CLASSIFICATION REPORT FOR TRAIN DATA

```
[[212  81]
 [ 95 673]]
```

	precision	recall	f1-score	support
0	0.69	0.72	0.71	293
1	0.89	0.88	0.88	768
accuracy			0.83	1061
macro avg	0.79	0.80	0.80	1061
weighted avg	0.84	0.83	0.84	1061

FIGURE 27: CONFUSION MATRIX FOR TRAIN DATA

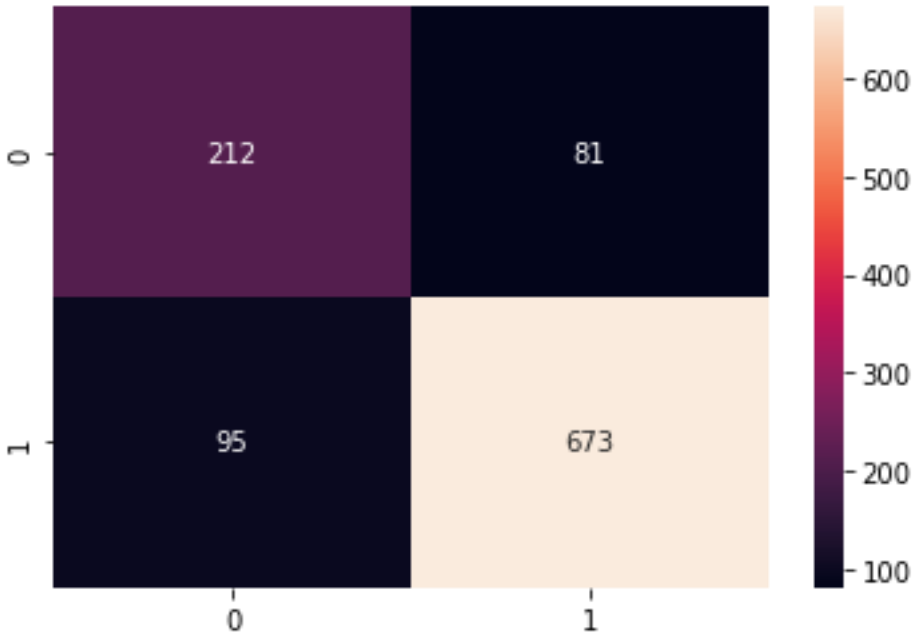


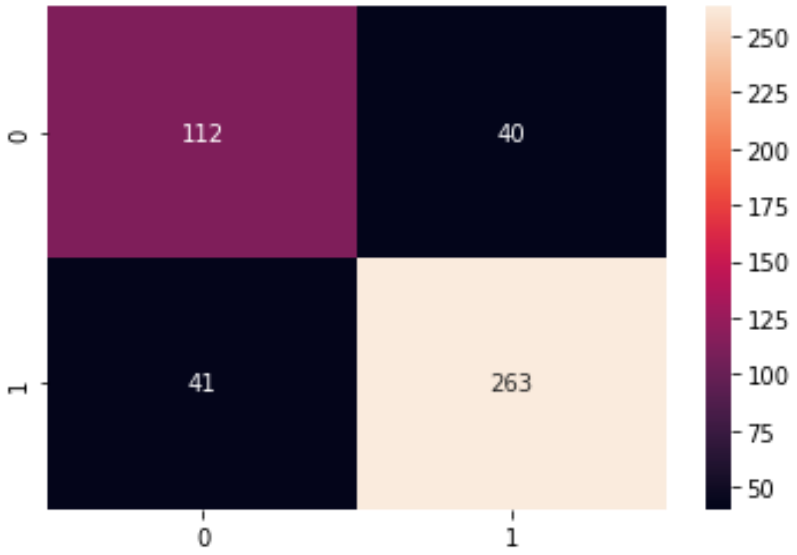
TABLE 33: CLASSIFICATION REPORT FOR TEST DATA

```
[[112  40]
 [ 41 263]]
              precision    recall  f1-score   support

           0       0.73      0.74      0.73       152
           1       0.87      0.87      0.87       304

    accuracy                           0.82       456
   macro avg       0.80      0.80      0.80       456
weighted avg       0.82      0.82      0.82       456
```

FIGURE 28: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR NAÏVE BAYES CLASSIFIER

NAÏVE BAYES CLASSIFIER					
#		Train Data		Test Data	
1	True Positive	212		112	
2	True Negative	673		263	
3	False Positive	81		40	
4	False Negative	95		41	
5	AUC score	89%		87%	
6	Accuracy	83%		82%	
		Conservative	Labour	Conservative	Labour
7	Precision	69%	89%	73%	87%
8	Recall	72%	88%	74%	87%
9	F1 score	71%	88%	73%	87%

- 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.

Q.1.6. Model Tuning, Bagging and Boosting. Apply grid search on each model (include all models) and make models on best params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances

MODEL TUNNING OF LOGISTIC REGRESSION

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':** *['elasticnet','l2','none','l1'],*
- **'solver':** *['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],*
- **'tol':** *[0.0001,0.00001,0.000001],*
- **'verbose':** *[True, False]*

1. **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 'elastic net' as our arguments and check which is preferred by grid search.
2. **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.
3. **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.
4. We have taken cross-validation as 3 and scoring as F1 for our grid search.

The final best parameters are:

```
{'penalty': 'l2', 'solver': 'liblinear', 'tol': 0.0001, 'verbose': True}

LogisticRegression(max_iter=10000, n_jobs=2, solver='liblinear', verbose=True)
```

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.

TABLE 34: CLASSIFICATION REPORT OF TUNED LOGISTIC REGRESSION TRAIN DATA

	precision	recall	f1-score	support
0	0.77	0.63	0.69	307
1	0.86	0.92	0.89	754
accuracy			0.84	1061
macro avg	0.81	0.78	0.79	1061
eighted avg	0.83	0.84	0.83	1061

FIGURE 29: CONFUSION MATRIX OF TUNED LOGISTIC REGRESSION TRAIN DATA

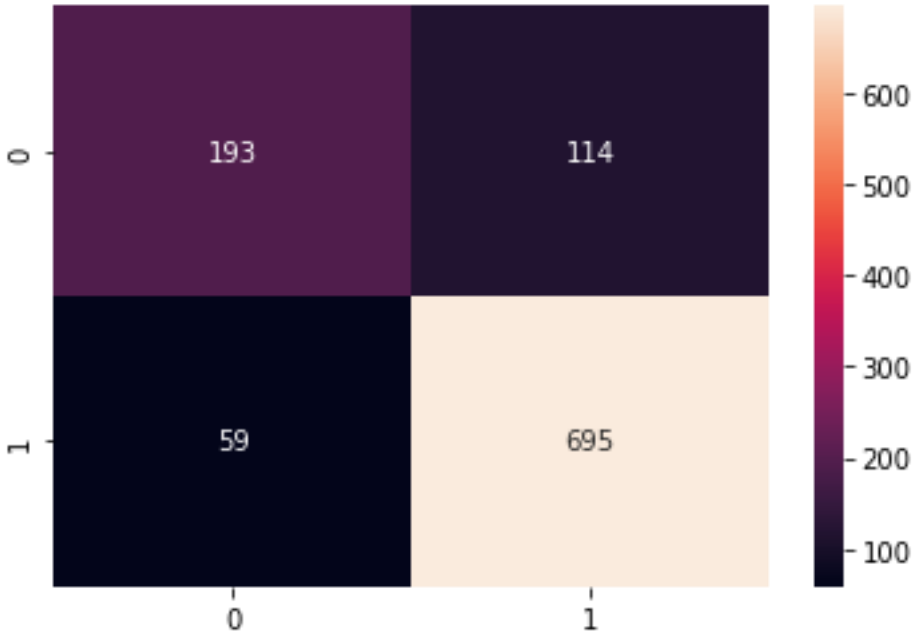
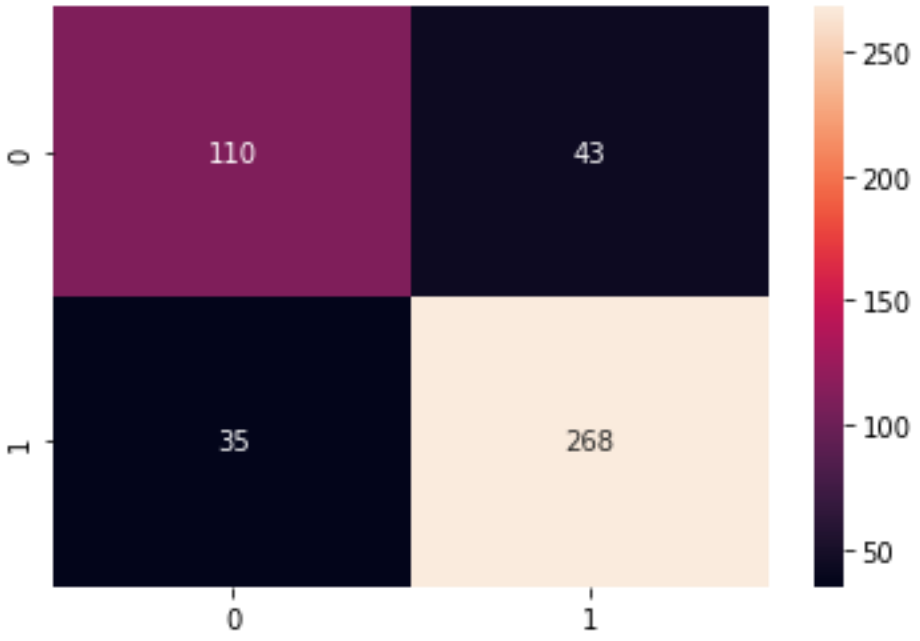


TABLE 35: CLASSIFICATION REPORT OF TUNED LOGISTIC REGRESSION TEST DATA

	precision	recall	f1-score	support
0	0.76	0.73	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

FIGURE 30: CONFUSION MATRIX OF TUNED LOGISTIC REGRESSION TEST DATA



INFERENCE OF TUNED LOGISTIC REGRESSION

TUNED LOGISTIC REGRESSION					
#		Train Data		Test Data	
1	True Positive	193		110	
2	True Negative	695		268	
3	False Positive	114		43	
4	False Negative	59		35	
5	AUC score	89%		87%	
6	Accuracy	84%		83%	
		Conservative	Labour	Conservative	Labour
7	Precision	77%	86%	76%	86%
8	Recall	63%	92%	73%	88%
9	F1 score	69%	89%	74%	87%

1. 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.
2. 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.
3. 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.
4. This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
5. Overall, the metrics are good fit.

MODEL TUNNING OF LDA

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.

TABLE 36: CUT OFF PROBABILITY

CUT-OFF PROBABLITY	RECALL	F1 SCORE	PRECISION
0.1	0.98	0.85	0.75
0.2	0.96	0.79	0.79
0.3	0.95	0.87	0.81
0.4	0.94	0.89	0.85
0.5	0.9	0.88	0.86
0.6	0.87	0.87	0.88
0.7	0.83	0.87	0.92
0.8	0.75	0.83	0.94
0.9	0.6	0.73	0.96

From the above table we can see that recall is going downwards, precision is going upwards and F1 score is also slowly going downwards, therefore, to maintain a balance we shall opt for **cut-off probability of 0.4** where there is a balance.

TABLE 37: CLASSIFICATION REPORT OF TUNED LDA WITH CUT-OFF 0.4 FOR TRAIN DATA

Classification Report of the custom cut-off train data:

	precision	recall	f1-score	support
0	0.79	0.58	0.67	307
1	0.85	0.94	0.89	754
accuracy			0.84	1061
macro avg	0.82	0.76	0.78	1061
weighted avg	0.83	0.84	0.83	1061

FIGURE 31: CONFUSION MATRIX OF TUNED LDA WITH CUT-OFF 0.4 FOR TRAIN DATA

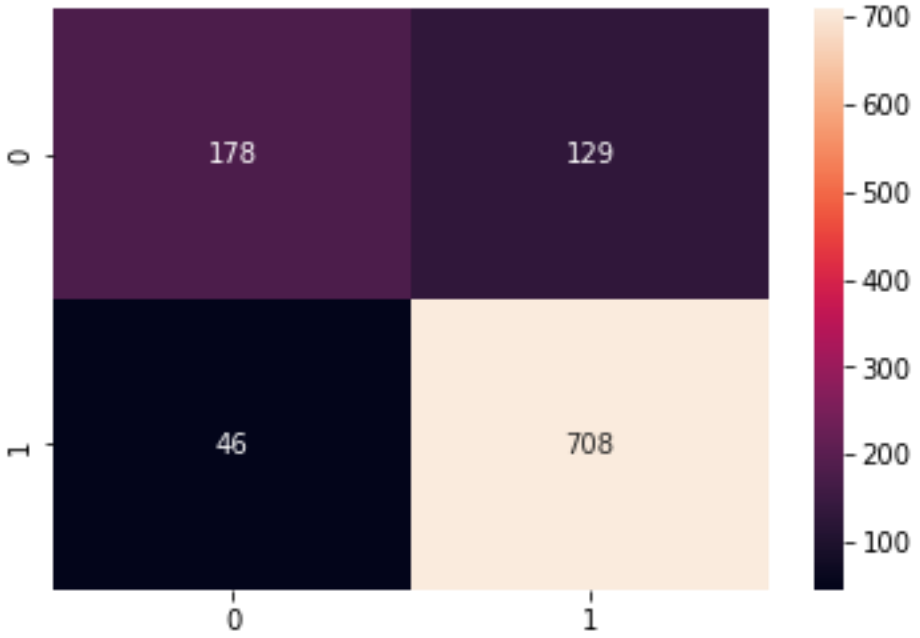
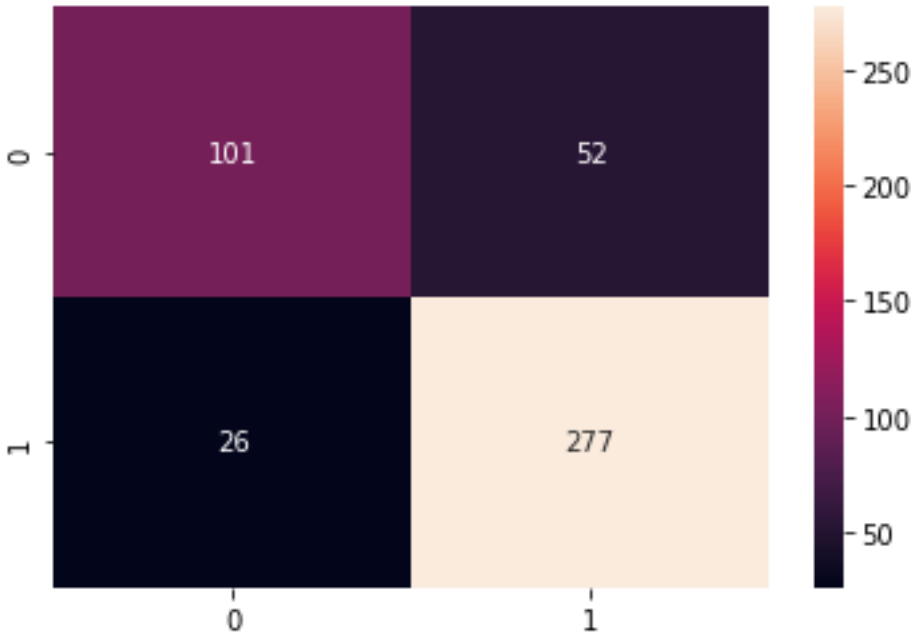


TABLE 38: CLASSIFICATION REPORT OF TUNED LDA WITH CUT-OFF 0.4 FOR TEST DATA

Classification Report of the custom cut-off test data:

	precision	recall	f1-score	support
0	0.80	0.66	0.72	153
1	0.84	0.91	0.88	303
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.82	456

FIGURE: CONFUSION MATRIX OF TUNED LDA WITH CUT-OFF 0.4 FOR TEST DATA



INFERENCE OF TUNED LDA

TUNED LINEAR DISCRIMINANT ANALYSIS					
#		Train Data		Test Data	
1	True Positive	178		101	
2	True Negative	708		277	
3	False Positive	129		52	
4	False Negative	46		26	
5	AUC score	89%		88%	
6	Accuracy	84%		83%	
		Conservative	Labour	Conservative	Labour
7	Precision	79%	85%	80%	84%
8	Recall	58%	94%	66%	91%
9	F1 score	67%	89%	72%	88%

- 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 TUNNING OF LDA USING GRID SEARCH CV

As mentioned above, LDA model tunning using Grid Search CV does not make any difference the regular model.

The Best Parameters are:

```
{'penalty': 'l2', 'solver': 'liblinear', 'tol': 0.0001, 'verbose': True}

LogisticRegression(max_iter=10000, n_jobs=2, solver='liblinear', verbose=True)
```

TABLE 39: CLASSIFICATION REPORT OF TUNED LDA WITH GRID SEARCH CV FOR TRAIN DATA

	precision	recall	f1-score	support
0	0.74	0.65	0.70	307
1	0.87	0.91	0.89	754
accuracy			0.84	1061
macro avg	0.81	0.78	0.79	1061
weighted avg	0.83	0.84	0.83	1061

FIGURE 33: CONFUSION MATRIX OF TUNED LDA WITH GRID SEARCH CV FOR TRAIN DATA

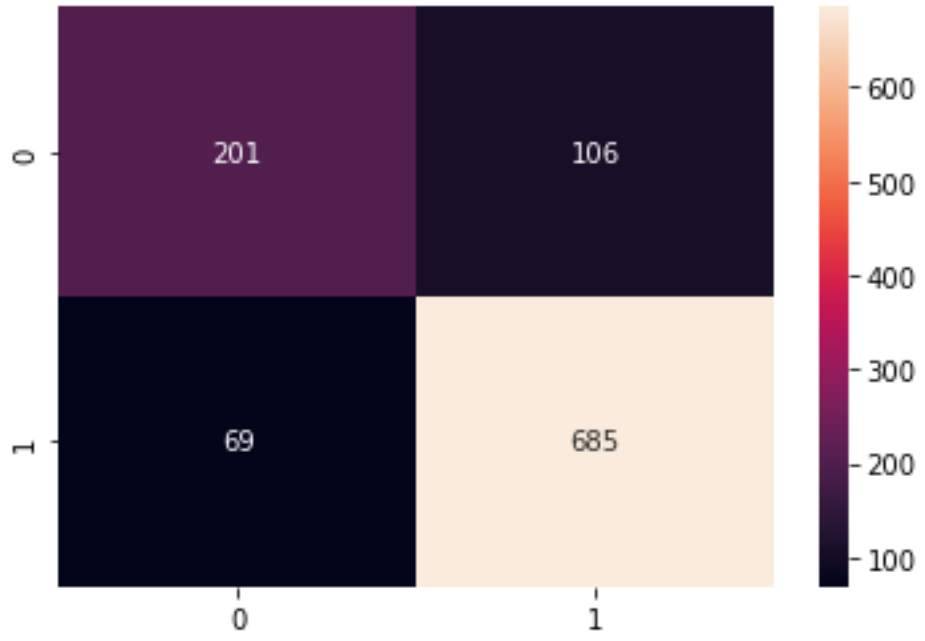
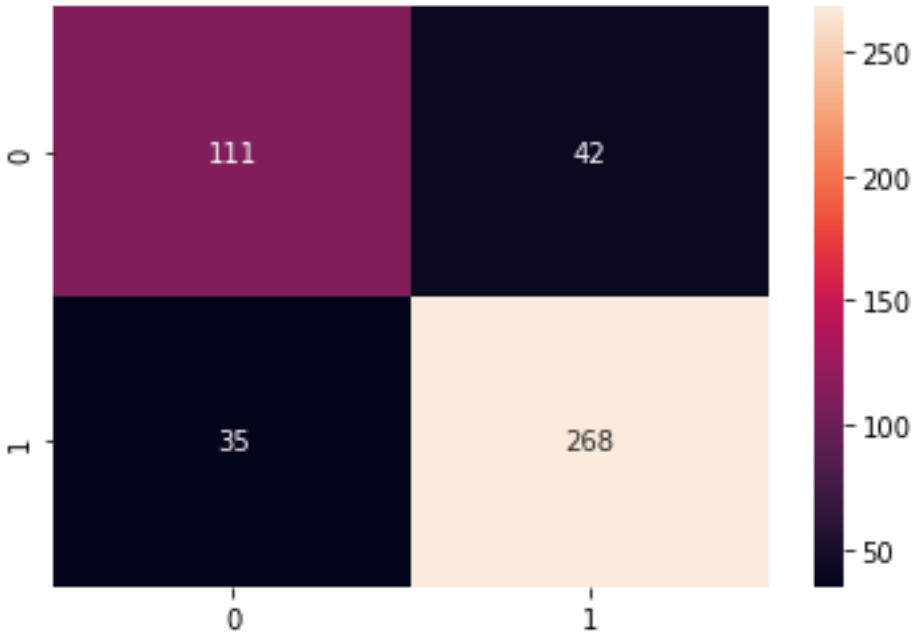


TABLE 40: CLASSIFICATION REPORT OF TUNED LDA WITH GRID SEARCH CV FOR TEST DATA

	precision	recall	f1-score	support
0	0.76	0.73	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

FIGURE 34: CONFUSION MATRIX OF TUNED LDA WITH GRID SEARCH CV FOR TEST DATA



INFERENCE OF TUNED LDA USING GRID SEARCH CV

TUNED LDA USING GRID SEARCH CV					
#		Train Data		Test Data	
1	True Positive	201		111	
2	True Negative	685		268	
3	False Positive	106		42	
4	False Negative	69		35	
5	AUC score	89%		88%	
6	Accuracy	84%		83%	
		Conservative	Labour	Conservative	Labour
7	Precision	74%	87%	76%	86%
8	Recall	65%	91%	73%	88%
9	F1 score	70%	89%	74%	87%

- The above tuned LDA model using Grid Search CV does not make much difference than the regular model.

MODEL TUNNING 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:

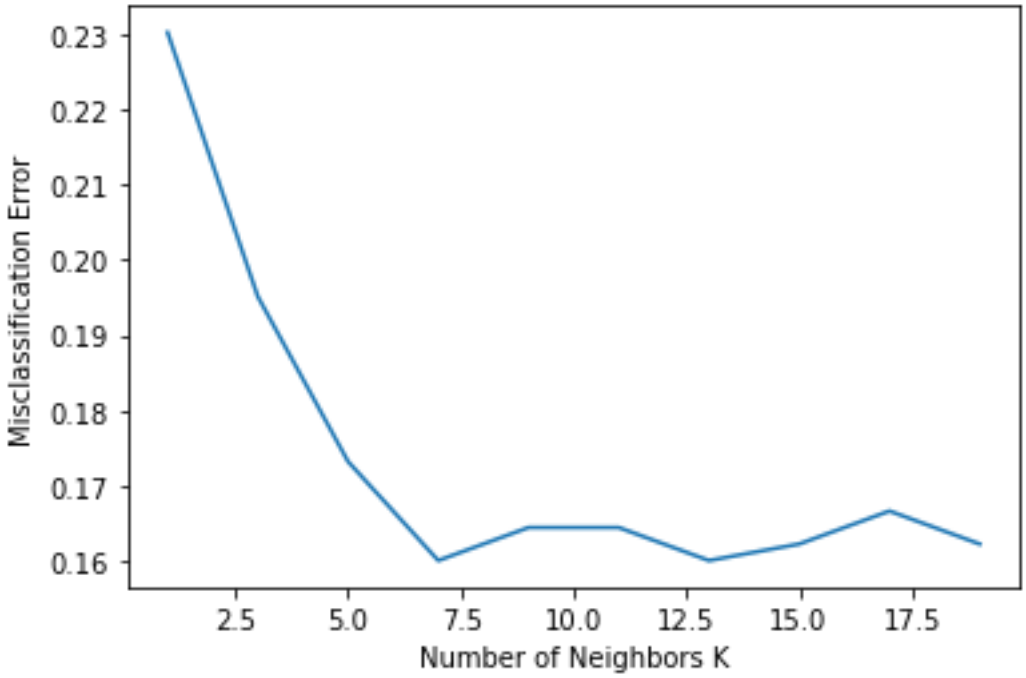
Run the KNN with no of neighbours to be 1,3, 5..19 and find the optimal number of neighbours from *K=1,3,5,7....19** using the Mis classification error

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.

TABLE 41: MISCLASSIFICATION ERROR

```
[0.23026315789473684,
 0.19517543859649122,
 0.17324561403508776,
 0.1600877192982456,
 0.16447368421052633,
 0.16447368421052633,
 0.1600877192982456,
 0.16228070175438591,
 0.16666666666666663,
 0.16228070175438591]
```

FIGURE: MISCLASSIFICATION ERROR



- 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.

TABLE 42: CLASSIFICATION REPORT OF TUNED KNN FOR TRAIN DATA

```
0.8416588124410933
[[213  94]
 [ 74 680]]
              precision    recall  f1-score   support

           0       0.81      0.66      0.73       307
           1       0.87      0.94      0.90       754

    accuracy                           0.86      1061
   macro avg       0.84      0.80      0.82      1061
weighted avg       0.85      0.86      0.85      1061
```

FIGURE 36: CONFUSION MATRIX OF TUNED KNN FOR TRAIN DATA

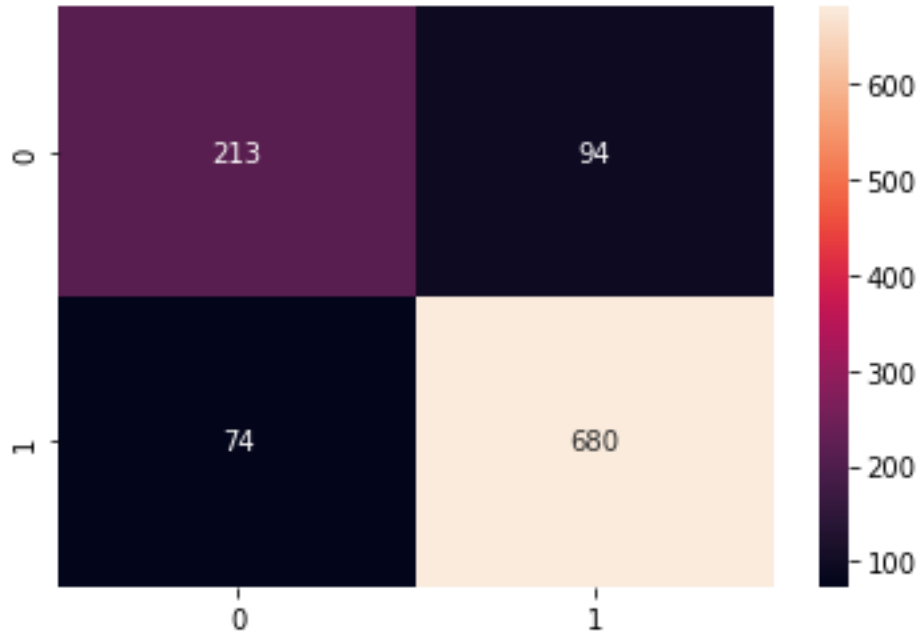


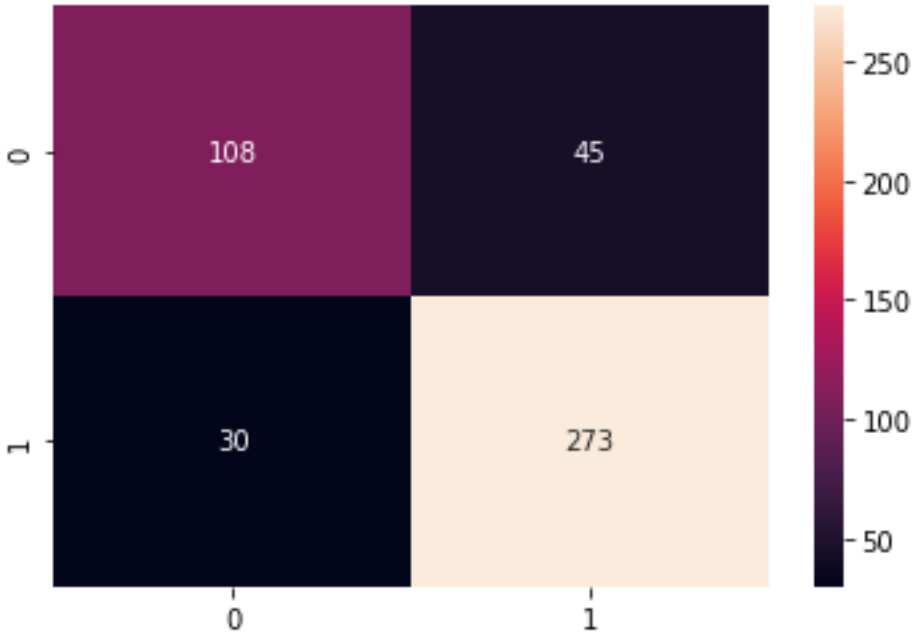
TABLE 43: CLASSIFICATION REPORT OF TUNED KNN FOR TEST DATA

```
0.8355263157894737
[[108  45]
 [ 30 273]]
              precision    recall  f1-score   support

           0       0.78      0.71      0.74       153
           1       0.86      0.90      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 37: CONFUSION MATRIX OF TUNED KNN FOR TEST



INFERENCE OF TUNED KNN MODEL

TUNED K-NEAREST NEIGHBOURS					
#		Train Data		Test Data	
1	True Positive	213		108	
2	True Negative	680		273	
3	False Positive	94		45	
4	False Negative	74		30	
5	AUC score	89%		88%	
6	Accuracy	86%		84%	
		Conservative	Labour	Conservative	Labour
7	Precision	81%	87%	78%	86%
8	Recall	66%	94%	71%	90%
9	F1 score	73%	90%	74%	88%

1. 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.
2. 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.
3. The recall and F1score metrics of training is overfitting compared to performance of test data.
4. **Overall, the model is good fit.**

NAIVE BAYES TUNED USING SMOTE

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.

TABLE 44: CLASSIFICATION REPORT FOR TRAIN DATA

```
0.8269230769230769
[[623 131]
 [130 624]]
              precision    recall  f1-score   support

           0       0.83      0.83      0.83       754
           1       0.83      0.83      0.83       754

    accuracy                           0.83      1508
   macro avg       0.83      0.83      0.83      1508
weighted avg       0.83      0.83      0.83      1508
```

FIGURE 38: CONFUSION MATRIX FOR TRAIN DATA

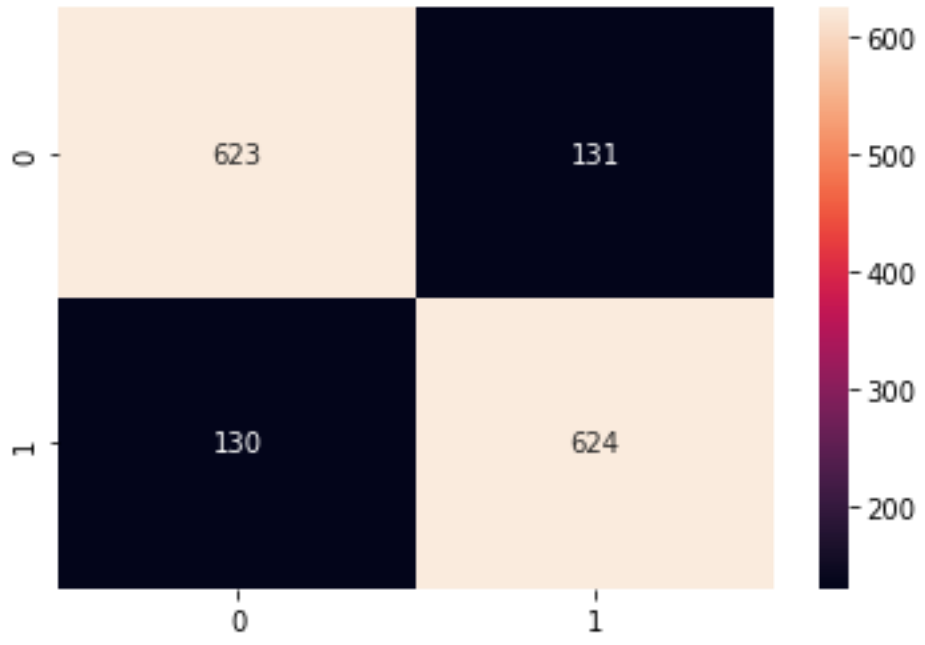


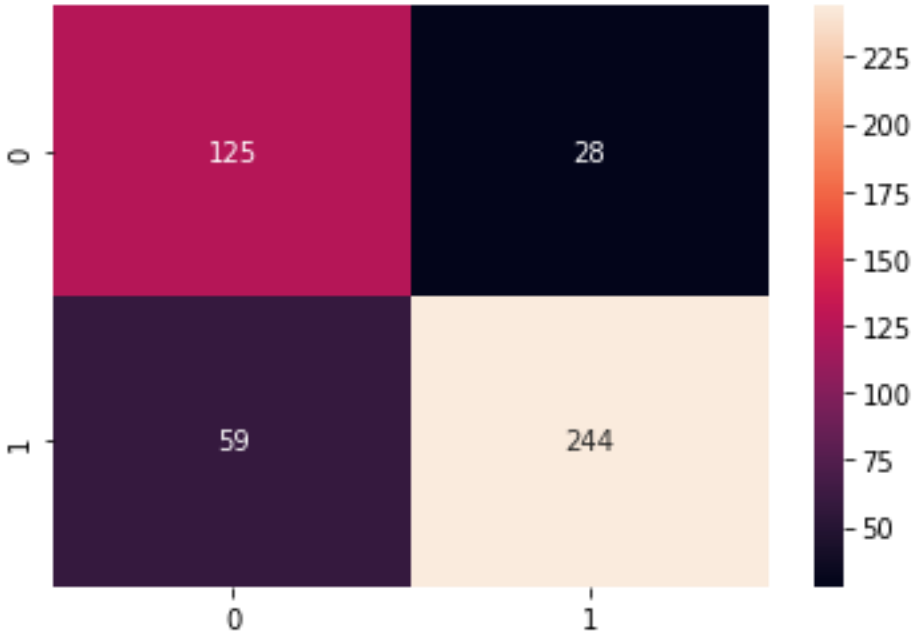
TABLE 45: CLASSIFICATION REPORT FOR TEST DATA

```
0.8092105263157895
[[125  28]
 [ 59 244]]
              precision    recall  f1-score   support

           0       0.68      0.82      0.74       153
           1       0.90      0.81      0.85       303

    accuracy                           0.81       456
   macro avg       0.79      0.81      0.80       456
weighted avg       0.82      0.81      0.81       456
```

TABLE 39: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR NAÏVE BAYES USING SMOTE

TUNED NAÏVE BAYES USING SMOTE					
#		Train Data		Test Data	
1	True Positive	623		125	
2	True Negative	624		244	
3	False Positive	131		28	
4	False Negative	130		59	
5	AUC score	90%		88%	
6	Accuracy	83%		81%	
		Conservative	Labour	Conservative	Labour
7	Precision	83%	83%	68%	90%
8	Recall	83%	83%	82%	81%
9	F1 score	83%	83%	74%	85%

1. The naïve bayes model with smote is performing slightly better after oversampling the minority class.
2. 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.
3. 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.
4. This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
5. Overall, the metrics are good fit.

BAGGING USING RANDOM FOREST

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.

TABLE 46: CLASSIFICATION REPORT FOR TRAIN DATA

```
0.8444863336475024
[[185 122]
 [ 43 711]]
              precision    recall  f1-score   support

           0       0.81      0.60      0.69       307
           1       0.85      0.94      0.90       754

    accuracy                           0.84      1061
   macro avg       0.83      0.77      0.79      1061
weighted avg       0.84      0.84      0.84      1061
```

FIGURE 40: CONFUSION MATRIX FOR TRAIN DATA

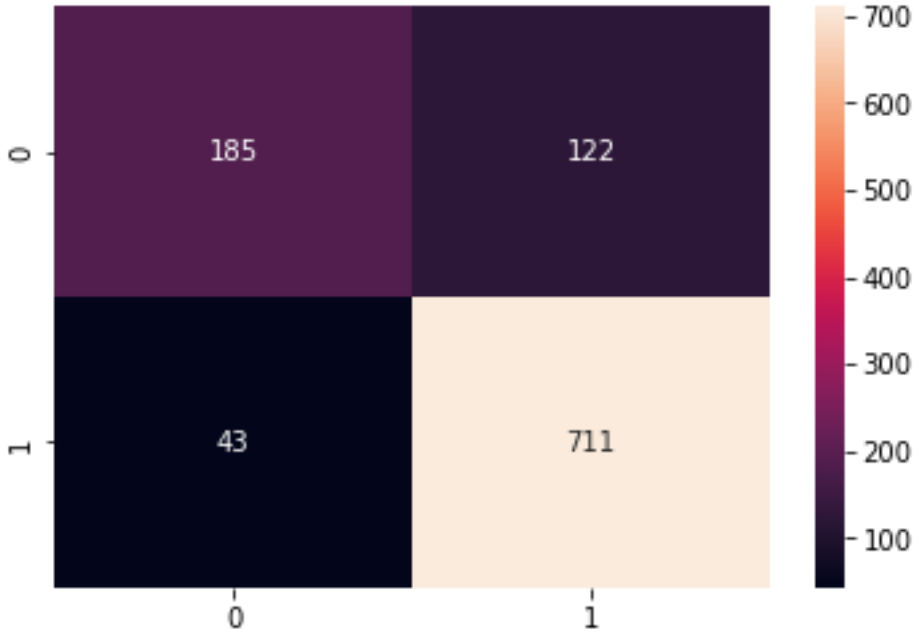


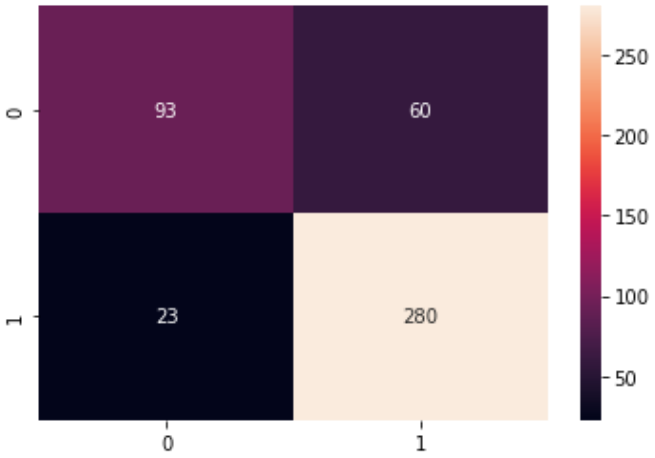
TABLE 47: CLASSIFICATION REPORT FOR TEST DATA

```
0.8179824561403509
[[ 93  60]
 [ 23 280]]
              precision    recall  f1-score   support

           0       0.80      0.61      0.69       153
           1       0.82      0.92      0.87       303

    accuracy                           0.82       456
   macro avg       0.81      0.77      0.78       456
weighted avg       0.82      0.82      0.81       456
```

FIGURE 41: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR BAGGING

BAGGING					
#		Train Data		Test Data	
1	True Positive	185		93	
2	True Negative	711		280	
3	False Positive	122		60	
4	False Negative	43		23	
5	AUC score	91%		88%	
6	Accuracy	84%		82%	
		Conservative	Labour	Conservative	Labour
7	Precision	81%	85%	80%	82%
8	Recall	60%	94%	61%	92%
9	F1 score	69%	90%	69%	87%

1. 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.
2. 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
3. Bagging with Random Forest has performed exceptionally well on the train data than test data.
4. Overall, the metrics are good fit.

ADAPTIVE BOOSTING

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.

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.

TABLE 48: CLASSIFICATION REPORT FOR TRAIN DATA

```
0.8369462770970783
[[186 121]
 [ 52 702]]
              precision    recall  f1-score   support

           0       0.78      0.61      0.68       307
           1       0.85      0.93      0.89       754

    accuracy                           0.84      1061
   macro avg       0.82      0.77      0.79      1061
weighted avg       0.83      0.84      0.83      1061
```

FIGURE 42: CONFUSION MATRIX FOR TRAIN

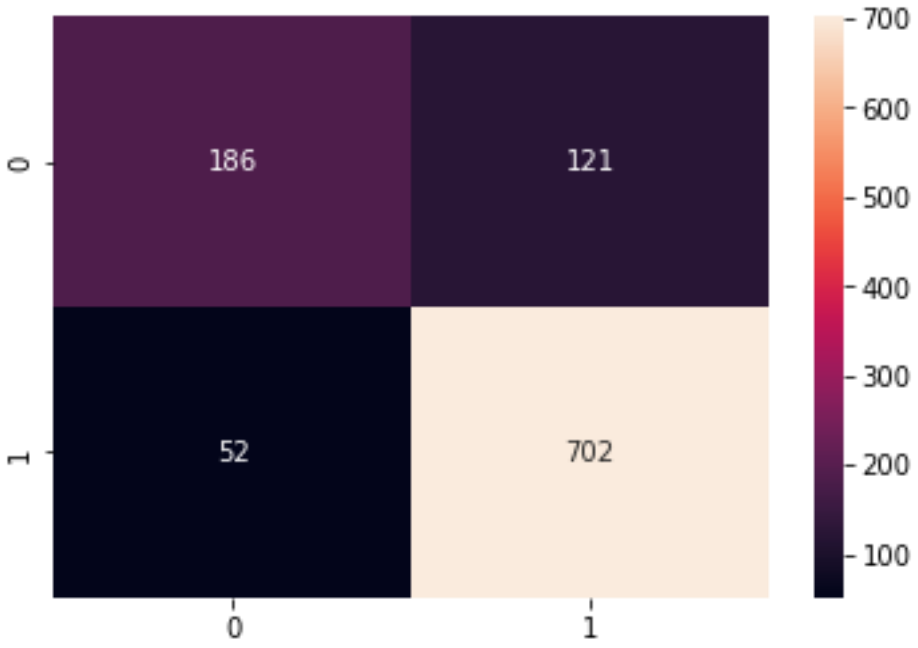


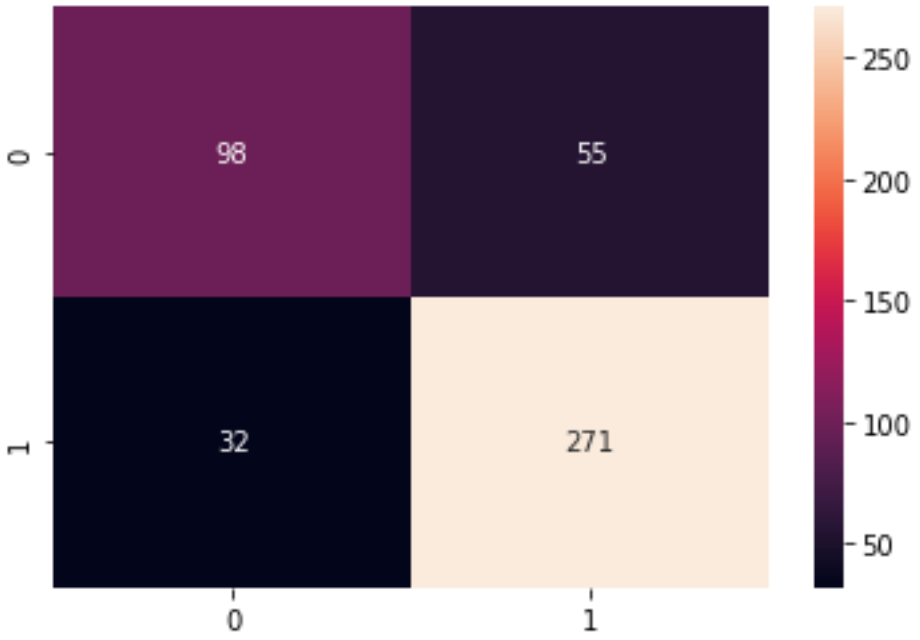
TABLE 49: CLASSIFICATION REPORT FOR TEST DATA

```
0.8092105263157895
[[ 98  55]
 [ 32 271]]
              precision    recall  f1-score   support

           0       0.75      0.64      0.69       153
           1       0.83      0.89      0.86       303

    accuracy                           0.81       456
   macro avg       0.79      0.77      0.78       456
weighted avg       0.81      0.81      0.80       456
```

FIGURE 43: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR ADA BOOSTING

ADA BOOSTING					
#		Train Data		Test Data	
1	True Positive	186		98	
2	True Negative	702		271	
3	False Positive	121		55	
4	False Negative	52		32	
5	AUC score	91%		88%	
6	Accuracy	84%		81%	
		Conservative	Labour	Conservative	Labour
7	Precision	78%	85%	75%	83%
8	Recall	61%	93%	64%	89%
9	F1 score	68%	89%	69%	86%

1. 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.
2. 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.
3. This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
4. Overall, the metrics are good fit.

GRADIENT BOOSTING

To build a Gradient Boost 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.

TABLE 50: CLASSIFICATION REPORT FOR TRAIN DATA

```
0.8925541941564562
[[239  68]
 [ 46 708]]
              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
```

FIGURE 44: CONFUSION MATRIX FOR TRAIN

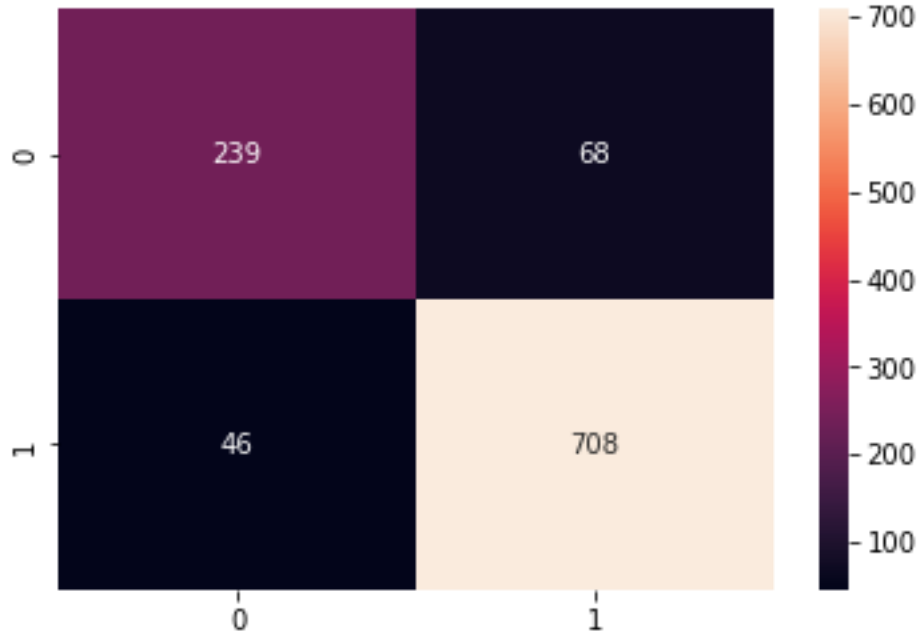


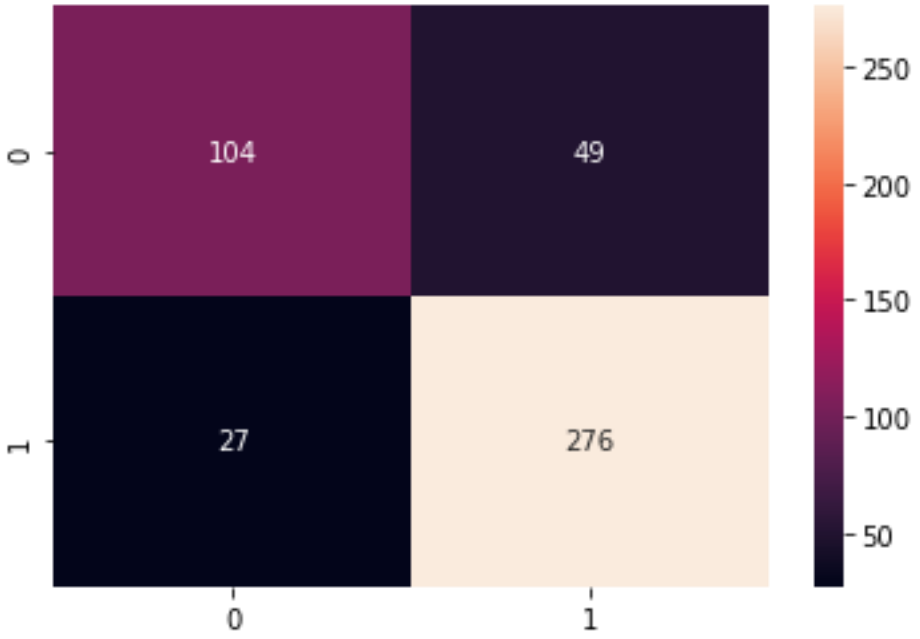
TABLE 51: CLASSIFICATION REPORT FOR TEST DATA

```
0.8333333333333334
[[104  49]
 [ 27 276]]
              precision    recall  f1-score   support

           0       0.79      0.68      0.73       153
           1       0.85      0.91      0.88       303

    accuracy                           0.83       456
   macro avg       0.82      0.80      0.81       456
weighted avg       0.83      0.83      0.83       456
```

FIGURE 45: CONFUSION MATRIX FOR TEST DATA



INFERENCE FOR GRADIENT BOOSTING

#		Train Data		Test Data	
1	True Positive	239		104	
2	True Negative	708		276	
3	False Positive	68		49	
4	False Negative	46		27	
5	AUC score	95%		89%	
6	Accuracy	89%		83%	
		Conservative	Labour	Conservative	Labour
7	Precision	84%	91%	79%	85%
8	Recall	78%	94%	68%	91%
9	F1 score	81%	93%	73%	88%

1. 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.

2. 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.
3. 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.
4. This shows that there was neither overfitting or underfitting, & the model performance of test is slightly better than train dataset.
5. Overall, the metrics are good fit.

Q.1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model, classification report. Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, after comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.

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.

- **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)

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

Precision = TP/ (TP + FP)

- **Recall** (Sensitivity): Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

Recall = TP/ (TP + FN)

- **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)]

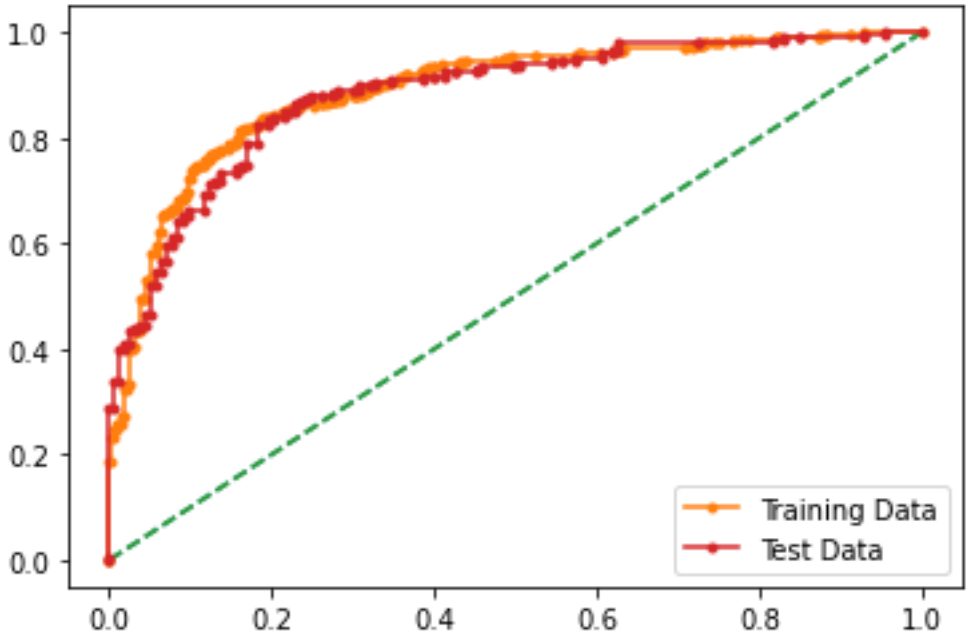
2.ROC Curve:

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:

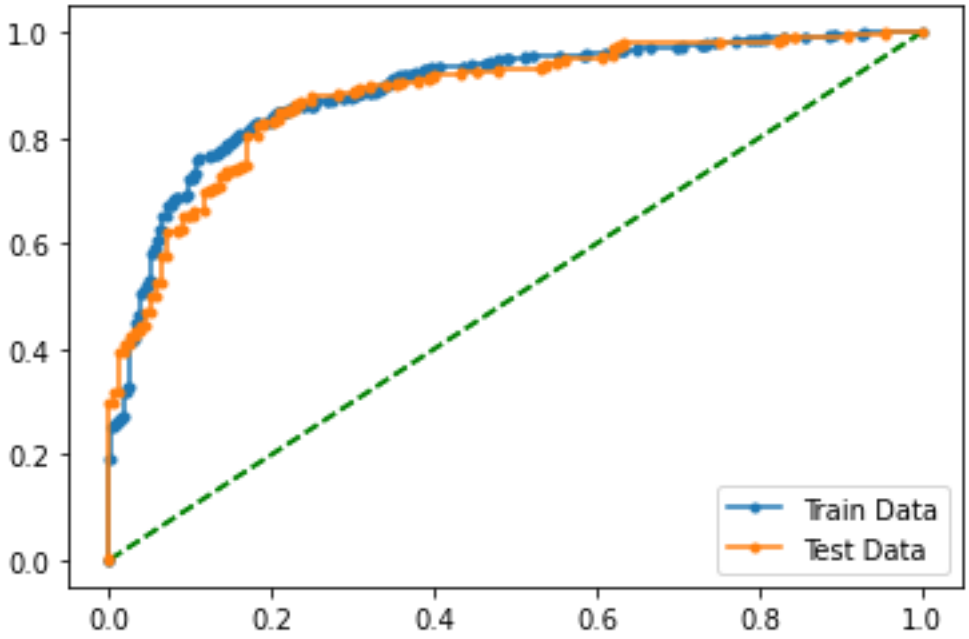
AUC score gives the area under the ROC curve built. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative.

FIGURE 46: AUC & ROC CURVE FOR LOGISTIC REGRESSION



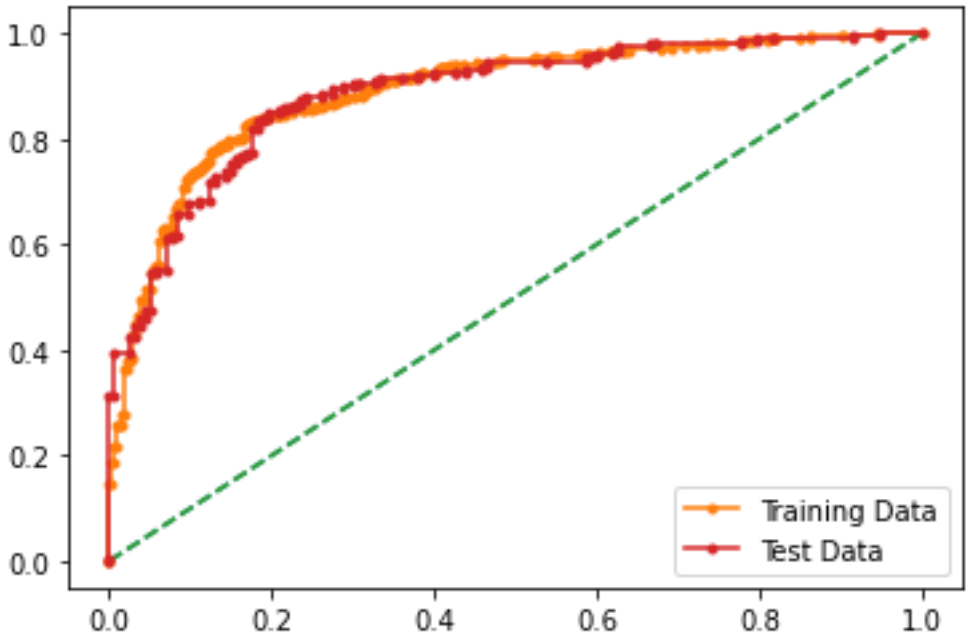
- **AUC for the Training Data: 0.890 & AUC for the Test Data: 0.883**

FIGURE 47: AUC & ROC CURVE FOR TUNED LOGISTIC REGRESSION



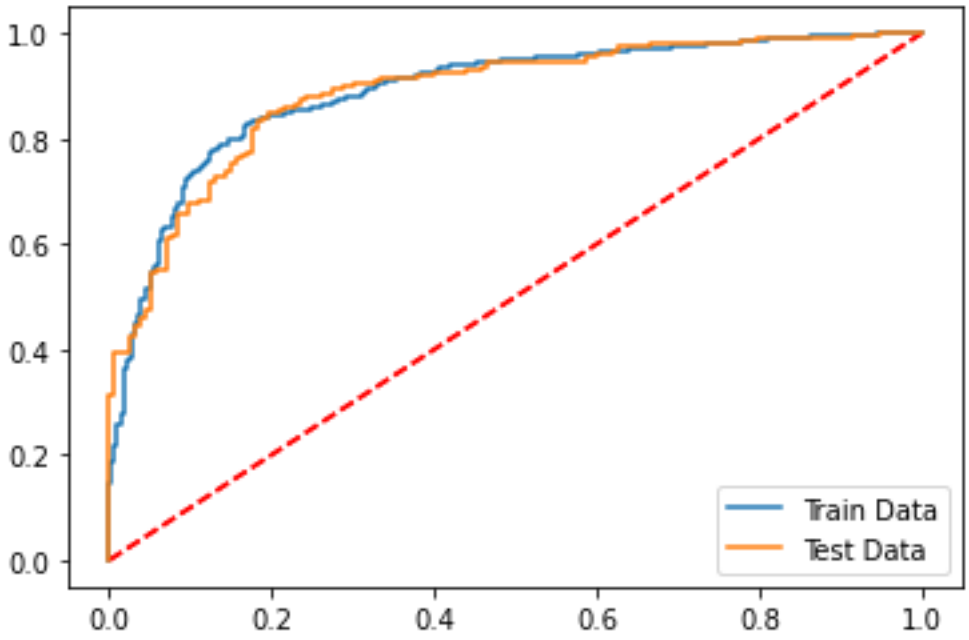
- **AUC for Train Data: 0.891 & AUC for Test Data: 0.881**

FIGURE 48: AUC & ROC CURVE FOR LDA



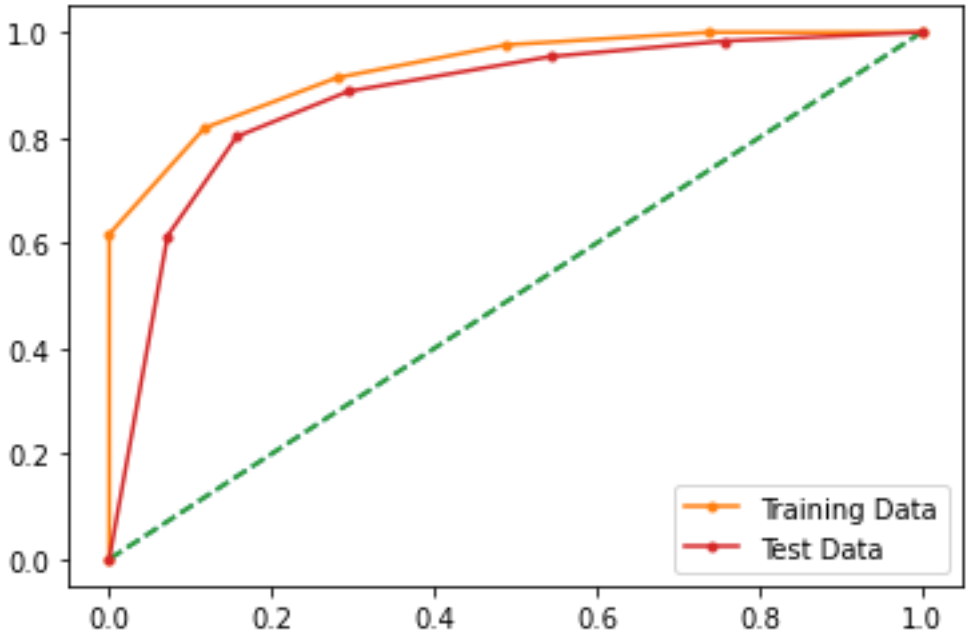
- **Training Data AUC: 0.890 & Test Data AUC: 0.888**

FIGURE 49: AUC & ROC CURVE FOR TUNED LDA



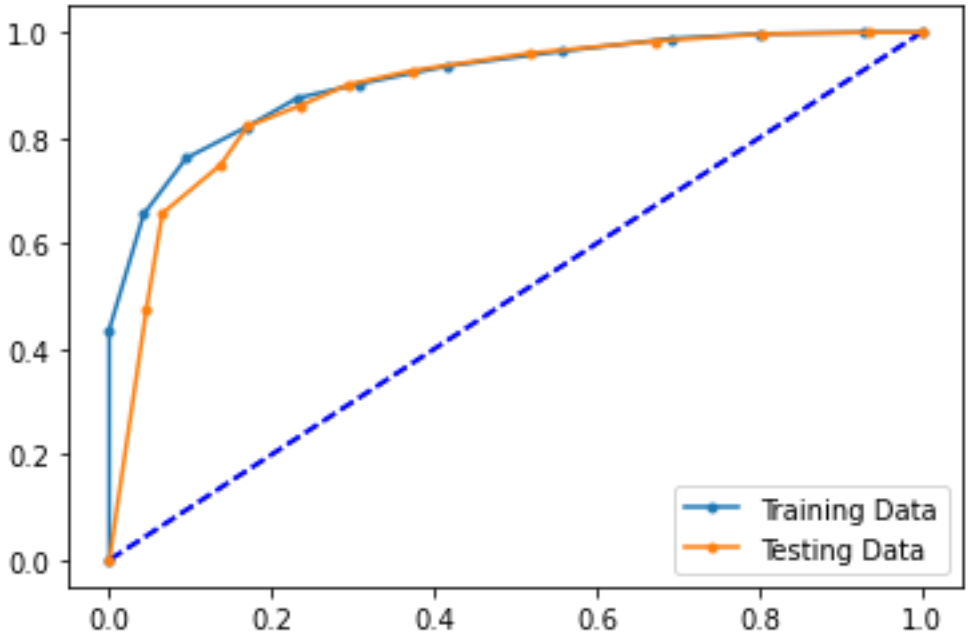
- **AUC for train data: 0.890 & AUC for test data: 0.888**

FIGURE 50: AUC & ROC CURVE FOR KNN MODEL



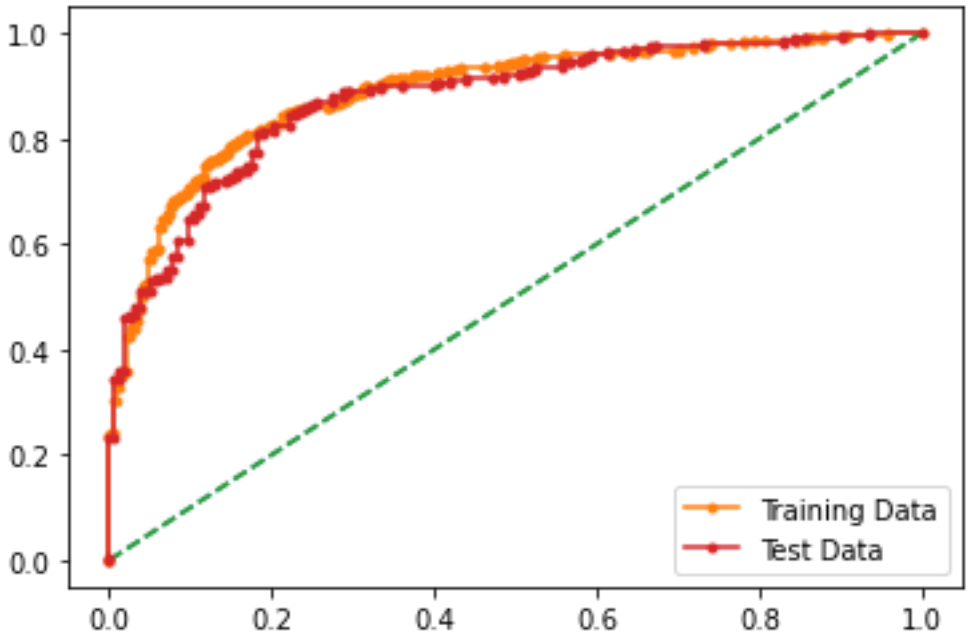
- **Training Data AUC: 0.931 & Test Data AUC: 0.876**

FIGURE 51: AUC & ROC CURVE FOR TUNED KNN MODEL



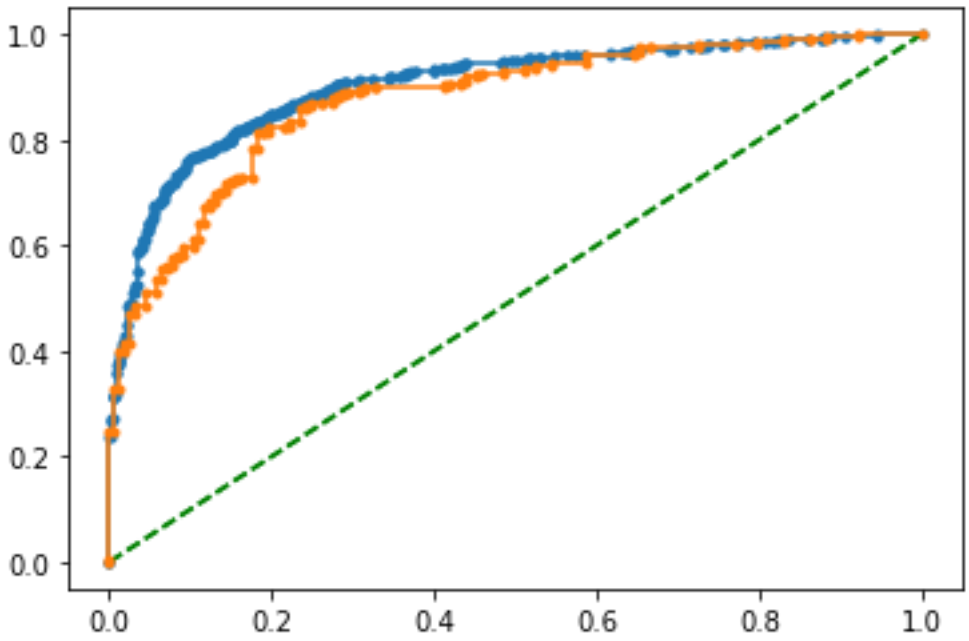
- **AUC for train data: 0.912 & AUC for test data: 0.890**

FIGURE 52: AUC & ROC CURVE FOR NAÏVE BAYES CLASSIFIER



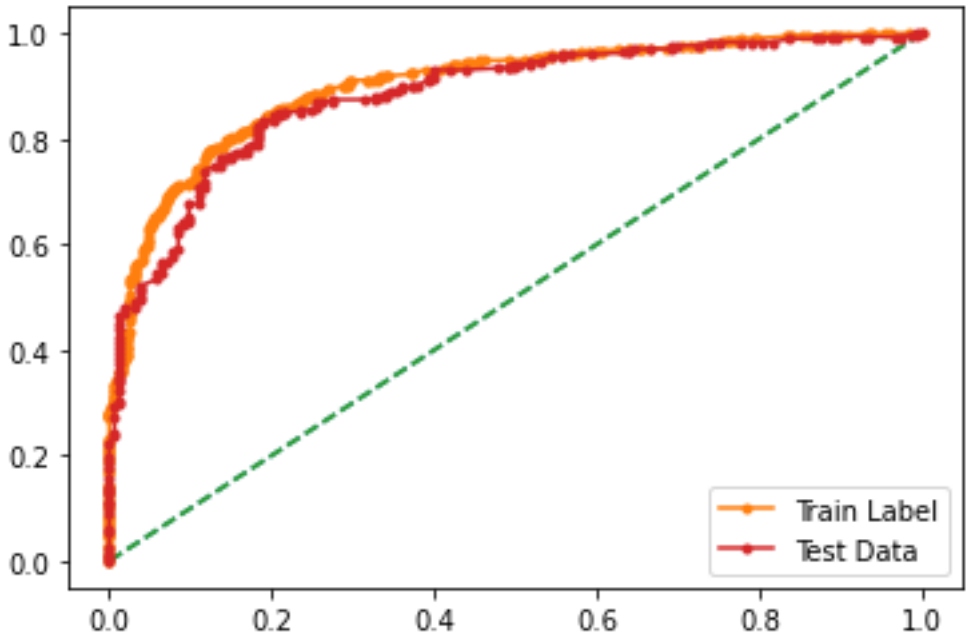
- **Train AUC: 0.889 & Test AUC: 0.876**

FIGURE 53: AUC & ROC CURVE FOR NAÏVE BAYES CLASSIFIER WITH SMOTE



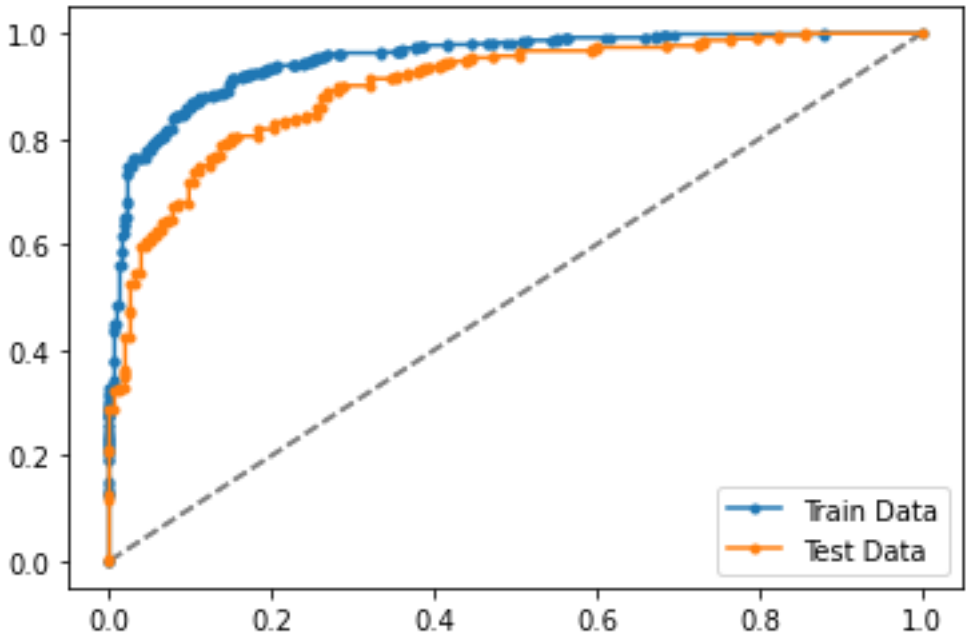
- **AUC for train data: 0.904 & AUC for test data: 0.876**

FIGURE 54: AUC & ROC CURVE FOR ADA BOOSTING



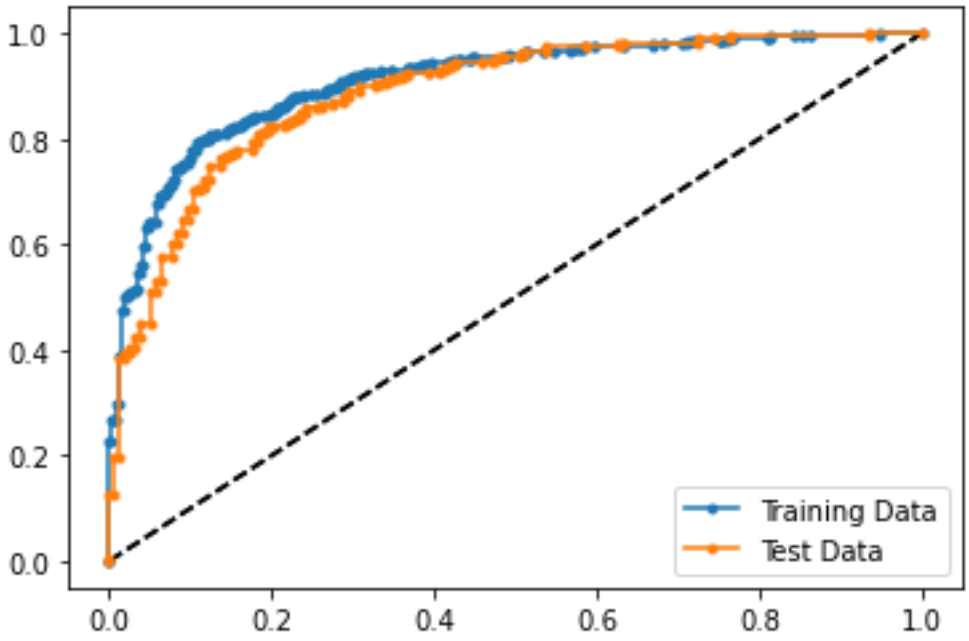
- **Train Data AUC: 0.902 & Test Data AUC: 0.884**

FIGURE 55: AUC & ROC CURVE FOR GRADIENT BOOSTING



- **AUC for train data: 0.951 & AUC for test data: 0.899**

FIGURE 56: AUC & ROC CURVE FOR BAGGING WITH RANDOM FOREST



- **AUC for train data: 0.910 & AUC for test data: 0.887**

TABLE 52: MODEL COMPARISION 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 53: MODEL EVALUATION METRICS FOR LABOUR PARTY						
Models	Train/Test	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression	Train	83%	86%	91%	89%	89%
	Test	83%	86%	88%	87%	88%
Tuned Logistic Regression	Train	84%	86%	92%	89%	89%
	Test	83%	86%	88%	87%	87%
LDA	Train	83%	91%	86%	89%	89%
	Test	83%	88%	86%	87%	88%
Tuned LDA	Train	84%	85%	94%	89%	89%
	Test	83%	84%	91%	88%	88%
KNN	Train	86%	91%	89%	90%	89%
	Test	83%	89%	86%	87%	89%
Tuned KNN	Train	86%	87%	94%	90%	89%
	Test	84%	86%	90%	88%	88%
Naïve Bayes Classifier	Train	83%	89%	88%	88%	89%
	Test	82%	87%	87%	87%	87%
Naïve Bayes Classifier with SMOTE	Train	83%	83%	83%	83%	90%
	Test	81%	90%	81%	85%	88%
Bagging with Random Forest	Train	84%	85%	94%	90%	91%
	Test	82%	82%	92%	87%	88%
ADA Boosting	Train	84%	85%	93%	89%	91%
	Test	81%	83%	89%	86%	88%
Gradient Boosting	Train	89%	91%	94%	93%	95%
	Test	83%	85%	91%	88%	89%

TABLE 54: MODEL EVALUATION METRICS FOR CONSERVATIVE PARTY						
Models	Train/Test	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression	Train	83%	75%	64%	69%	89%
	Test	83%	76%	73%	74%	88%
Tuned Logistic Regression	Train	84%	77%	63%	69%	89%
	Test	83%	76%	73%	74%	87%
LDA	Train	83%	65%	74%	69%	89%
	Test	83%	73%	76%	74%	88%
Tuned LDA	Train	84%	79%	58%	67%	89%
	Test	83%	80%	66%	72%	88%
KNN	Train	86%	72%	77%	75%	89%
	Test	83%	71%	76%	73%	89%
Tuned KNN	Train	86%	81%	66%	73%	89%
	Test	84%	78%	71%	74%	88%
Naïve Bayes Classifier	Train	83%	69%	72%	71%	89%
	Test	82%	73%	74%	73%	87%
Naïve Bayes Classifier with SMOTE	Train	83%	83%	83%	83%	90%
	Test	81%	68%	82%	74%	88%
Bagging with Random Forest	Train	84%	80%	60%	69%	91%
	Test	82%	81%	61%	69%	88%
ADA Boosting	Train	84%	78%	68%	68%	91%
	Test	81%	75%	74%	69%	88%
Gradient Boosting	Train	89%	84%	78%	81%	95%
	Test	83%	79%	91%	88%	89%

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 your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

BUSINESS INSIGHTS

1. The Labour Party and their leader seem to have positive ratings in the public's eye, if this had to be a real exit poll then Labour Party would have the majority population's vote.
2. The attributes 'Hague' and 'Blair' important features in predicting the dependent variable.
3. Models developed are better at classifying Labour Party voters than Conservative Party voters.
4. Educating the public on the what is the parties perspective towards them and same way the parties should also understand the people's perception/perspective.
5. On a scale of 0 to 3, about 30% of the total population has zero knowledge about politics/parties.

6. People who gave low scores to a certain party, still decided to vote for the same party instead of voting for the other party. This can be because of lack of political knowledge among the people.

BUSINESS RECOMMENDATIONS

1. **To assess the success or failure of a certain political campaign:**

Exit polls are conducted to understand whether the people are switching from one party to another based on the campaign done by the party. When people shift from supporting the existing or already supporting party to an opposite party then the election campaign of a particular party is considered as successful one. Therefore, the political parties should take the exit polls seriously and conduct the election campaigns as per the exit polls.

2. Build / Conduct new election campaigns:

In the case of a business, it is important for the company to create new marketing campaigns as per the consumers interests to attract sales, like wise the parties should create new election campaigns as per the peoples interest or perception or sentiments. In this case a better accuracy or a higher F1 score will let us understand the sentiments/ perception of the public and campaign accordingly.

3. To assess the chances of winning the polls:

To check the chances of winning through the exit polls as a model, accuracy of the model should be considered, which will help in knowing the chances of winning the real polls.

4. Fraudulent activities:

Fraudulent activities occur during elections and to if there is any difference between the exit polls and real polls we can find out through the models precision and recall, where precision measures the relevancy of the results and recall measures the actual relevant results returned.

PROBLEM 2

INTRODUCTION

Here, 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

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 institutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a moment 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 not, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of the 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 fulness 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 reason, 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 years 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 present and our future is this experience of a democracy which successfully survived crisis at home; put away many evil things; built new structures on enduring lines; and, through it all, maintained the fact of its democracy.\n\nFor action has been taken within 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 American democracy have seen their dire predictions come to naught.\n\nDemocracy is not dying.\n\nWe know it because we have seen 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 majority.\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 human 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 person, has a body--a body that must be fed and clothed and housed, invigorated and rested, in a manner that measures up to the objectives of our time.\n\nA nation, like a person, has a mind -- a mind that must be kept informed and alert, that must know itself, that understands the hopes and the needs of its neighbors -- all the other nations that live within the narrowing circle of

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

1973-NIXON'S SPEECH

Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and g ood country we share together:

When we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and o f destructive conflict at home.

As we meet here today, we stand on the threshold of a new era of peace in the world.

The central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not b e what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.

Let us resolve that this will be what it can become: a time of great responsibilities greatly borne, in which we renew the sp irit and the promise of America as we enter our third century as a nation.

This past year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendsh ips, and by our missions to Peking and to Moscow, we were able to establish the base for a new and more durable pattern of re lationships among the nations of the world. Because of America's bold initiatives, 1972 will be long remembered as the year o f the greatest progress since the end of World War II toward a lasting peace in the world.

- **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.

Q.2.2. Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.

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.

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'
- **Following is 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.

SPEECH BEFORE STOPWORDS: 1941 ROOSEVELT’S SPEECH

```
['On', 'each', 'national', 'day', 'of', 'inauguration', 'since', '1789', ',', 'the', 'people', 'have', 'renewed', 'their', 'sen
se', 'of', 'dedication', 'to', 'the', 'United', 'States', '.', 'In', 'Washington', "'", 's', 'day', 'the', 'task', 'of', 'the',
'people', 'was', 'to', 'create', 'and', 'weld', 'together', 'a', 'nation', '.', 'In', 'Lincoln', "'", 's', 'day', 'the', 'tas
k', 'of', 'the', 'people', 'was', 'to', 'preserve', 'that', 'Nation', 'from', 'disruption', 'from', 'within', '.', 'In', 'thi
s', 'day', 'the', 'task', 'of', 'the', 'people', 'is', 'to', 'save', 'that', 'Nation', 'and', 'its', 'institutions', 'from', 'd
isruption', 'from', 'without', '.', 'To', 'us', 'there', 'has', 'come', 'a', 'time', ',', 'in', 'the', 'midst', 'of', 'swift',
'happenings', ',', 'to', 'pause', 'for', 'a', 'moment', 'and', 'take', 'stock', '--', 'to', 'recall', 'what', 'our', 'place',
'in', 'history', 'has', 'been', ',', 'and', 'to', 'rediscover', 'what', 'we', 'are', 'and', 'what', 'we', 'may', 'be', '.', 'I
f', 'we', 'do', 'not', ',', 'we', 'risk', 'the', 'real', 'peril', 'of', 'inaction', '.', 'Lives', 'of', 'nations', 'are', 'dete
rmined', 'not', 'by', 'the', 'count', 'of', 'years', ',', 'but', 'by', 'the', 'lifetime', 'of', 'the', 'human', 'spirit', '.',
'The', 'life', 'of', 'a', 'man', 'is', 'three', '-', 'score', 'years', 'and', 'ten', ':', 'a', 'little', 'more', ',', 'a', 'lit
tle', 'less', '.', 'The', 'life', 'of', 'a', 'nation', 'is', 'the', 'fullness', 'of', 'the', 'measure', 'of', 'its', 'will', 't
o', 'live', '.', 'There', 'are', 'men', 'who', 'doubt', 'this', '.', 'There', 'are', 'men', 'who', 'believe', 'that', 'democrac
y', ',', 'as', 'a', 'form', 'of', 'Government', 'and', 'a', 'frame', 'of', 'life', ',', 'is', 'limited', 'or', 'measured', 'b
y', 'a', 'kind', 'of', 'mystical', 'and', 'artificial', 'fate', 'that', ',', 'for', 'some', 'unexplained', 'reason', ',', 'tyra
nny', 'and', 'slavery', 'have', 'become', 'the', 'surging', 'wave', 'of', 'the', 'future', '--', 'and', 'that', 'freedom', 'i
s', 'an', 'ebbing', 'tide', '.', 'But', 'we', 'Americans', 'know', 'that', 'this', 'is', 'not', 'true', '.', 'Eight', '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', 'acte
d', '.', 'We', 'acted', 'quickly', ',', 'boldly', ',', 'decisively', '.', 'These', 'later', 'years', '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', 'ide
als', 'are', 'to', 'be', 'measured', 'in', 'other', 'than', 'material', 'things', '.', 'Most', 'vital', 'to', 'our', 'present',
```

SPEECH AFTERSTOPWORDS: 1941 ROOSEVELT'S SPEECH

```
No. of words after removing stop words in Roosevelts speech>> 630
['national', 'day', 'inauguration', 'since', 'people', 'renewed', 'sense', 'dedication', 'united', 'states', 'washington', 'da
y', 'task', 'people', 'create', 'weld', 'together', 'nation', 'lincoln', 'day', 'task', 'people', 'preserve', 'nation', 'disrup
tion', 'within', 'day', 'task', 'people', 'save', 'nation', 'institutions', 'disruption', 'without', 'us', 'come', 'time', 'mid
st', 'swift', 'happenings', 'pause', 'moment', 'take', 'stock', 'recall', 'place', 'history', 'rediscover', 'may', 'risk', 'rea
l', 'peril', 'inaction', 'lives', 'nations', 'determined', 'count', 'years', 'lifetime', 'human', 'spirit', 'life', 'man', 'thr
ee', 'score', 'years', 'ten', 'little', 'little', 'less', 'life', 'nation', 'fullness', 'measure', 'live', 'men', 'doubt', 'me
n', 'believe', 'democracy', 'form', 'government', 'frame', 'life', 'limited', 'measured', 'kind', 'mystical', 'artificial', 'fa
te', 'unexplained', 'reason', 'tyranny', 'slavery', 'become', 'surging', 'wave', 'future', 'freedom', 'ebbing', 'tide', 'americ
ans', 'know', 'true', 'eight', 'years', 'ago', 'life', 'republic', 'seemed', 'frozen', 'fatalistic', 'terror', 'proved', 'tru
e', 'midst', 'shock', 'acted', 'acted', 'quickly', 'boldly', 'decisively', 'later', 'years', 'living', 'years', 'fruitful', 'ye
ars', 'people', 'democracy', 'brought', 'us', 'greater', 'security', 'hope', 'better', 'understanding', 'life', 'ideals', 'meas
ured', 'material', 'things', 'vital', 'present', 'future', 'experience', 'democracy', 'successfully', 'survived', 'crisis', 'ho
me', 'put', 'away', 'many', 'evil', 'things', 'built', 'new', 'structures', 'enduring', 'lines', 'maintained', 'fact', 'democra
cy', 'action', 'taken', 'within', 'three', 'way', 'framework', 'constitution', 'united', 'states', 'coordinate', 'branches', 'g
overnment', 'continue', 'freely', 'function', 'bill', 'rights', 'remains', 'inviolate', 'freedom', 'elections', 'wholly', 'main
tained', 'prophets', 'downfall', 'american', 'democracy', 'seen', 'dire', 'predictions', 'come', 'naught', 'democracy', 'dyin
g', 'know', 'seen', 'revive', 'grow', 'know', 'cannot', 'die', 'built', 'unhampered', 'initiative', 'individual', 'men', 'wome
n', 'joined', 'together', 'common', 'enterprise', 'enterprise', 'undertaken', 'carried', 'free', 'expression', 'free', 'majorit
y', 'know', 'democracy', 'alone', 'forms', 'government', 'enlists', 'full', 'force', 'men', 'enlightened', 'know', 'democracy',
'alone', 'constructed', 'unlimited', 'civilization', 'capable', 'infinite', 'progress', 'improvement', 'human', 'life', 'know',
'look', 'surface', 'sense', 'still', 'spreading', 'every', 'continent', 'humane', 'advanced', 'end', 'unconquerable', 'forms',
'human', 'society', 'nation', 'like', 'person', 'body', 'body', 'must', 'fed', 'clothed', 'housed', 'invigorated', 'rested', 'm
anner', 'measures', 'objectives', 'time', 'nation', 'like', 'person', 'mind', 'mind', 'must', 'kept', 'informed', 'alert', 'mus
t', 'know', 'understands', 'hopes', 'needs', 'neighbors', 'nations', 'live', 'within', 'narrowing', 'circle', 'world', 'natio
n', 'like', 'person', 'something', 'deeper', 'something', 'permanent', 'something', 'larger', 'sum', 'parts', 'something', 'mat
ters', 'future', 'calls', 'forth', 'sacred', 'guarding', 'present', 'thing', 'find', 'difficult', 'even', 'impossible', 'hit',
'upon', 'single', 'simple', 'word', 'yet', 'understand', 'spirit', 'faith', 'america', 'product', 'centuries', 'born', 'multitu
des', 'came', 'many', 'lands', 'high', 'degree', 'mostly', 'plain', 'people', 'sought', 'early', 'late', 'find', 'freedom', 'fr
eely', 'democratic', 'aspiration', 'mere', 'recent', 'phase', 'human', 'history', 'human', 'history', 'permeated', 'ancient',
```

- 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
- 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 630 words after removal of stopwords which are not useful for further analysis.

SPEECH BEFORE STOPWORDS: 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', 'as', 'a', 'beginning', '--', 'signifying', 'renewal', ',', 'as', 'well', 'as', 'change', '.', 'For', 'I', 'have', 'sworn', 'I', 'before', 'you', 'and', 'Almighty', 'God', 'the', 'same', 'solemn', 'oath', 'our', 'forebear s', 'l', 'prescribed', 'nearly', 'a', 'century', 'and', 'three', 'quarters', 'ago', '.', 'The', 'world', 'is', 'very', 'differe nt', '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', 'which', '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', 'th e', 'hand', 'of', 'God', '.', 'We', 'dare', 'not', 'forget', 'today', 'that', 'we', 'are', 'the', 'heirs', 'of', 'that', 'firs t', 'revolution', '.', 'Let', 'the', 'word', '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', 'un doing', 'of', 'those', 'human', 'rights', 'to', 'which', 'this', 'Nation', 'has', 'always', 'been', 'committed', ',', 'and', 't o', 'which', 'we', 'are', 'committed', 'today', 'at', 'home', 'and', 'around', 'the', 'world', '.', 'Let', 'every', 'nation', 'know', ',', 'whether', 'it', 'wishes', 'us', 'well', 'or', 'ill', ',', 'that', 'we', 'shall', 'pay', 'any', 'price', ',', 'bea r', 'any', 'burden', ',', 'meet', 'any', 'hardship', ',', 'support', 'any', 'friend', ',', 'oppose', 'any', 'foe', ',', 'in', 'order', 'to', 'assure', 'the', 'survival', 'and', 'the', 'success', 'of', 'liberty', '.', 'This', 'much', 'we', 'pledge', '-- ', 'and', 'more', '.', 'To', 'those', 'old', 'allies', 'whose', 'cultural', 'and', 'spiritual', 'origins', 'we', 'share', ',', 'we', 'pledge', 'the', 'loyalty', 'of', 'faithful', 'friends', '.', 'United', ',', 'there', 'is', 'little', 'we', 'cannot', 'd o', 'in', 'a', 'host', 'of', 'cooperative', 'ventures', '.', 'Divided', ',', 'there', 'is', 'little', 'we', 'can', 'do', '--', 'for', 'we', 'dare', 'not', 'meet', 'a', 'powerful', 'challenge', 'at', 'odds', 'and', 'split', 'asunder', '.', 'To', 'those', 'new', 'States', 'whom', 'we', 'welcome', 'to', 'the', 'ranks', 'of', 'the', 'free', ',', 'we', 'pledge', 'our', 'word', 'tha t', 'one', 'form', 'of', 'colonial', 'control', 'shall', 'not', 'have', 'passed', 'away', 'merely', 'to', 'be', 'replaced', 'b y', 'a', 'far', 'more', 'iron', 'tyranny', '.', 'We', 'shall', 'not', 'always', 'expect', 'to', 'find', 'them', 'supporting',

SPEECH AFTER STOPWORDS: 1961 KENNEDY'S SPEECH

```
The number of words after after removing stopwords is>> 697
['vice', 'president', 'johnson', 'mr', 'speaker', 'mr', 'chief', 'justice', 'president', 'eisenhower', 'vice', 'president', 'ni
xon', 'president', 'truman', 'reverend', 'clergy', 'fellow', 'citizens', 'observe', 'today', 'victory', 'party', 'celebration',
'freedom', 'symbolizing', 'end', 'well', 'beginning', 'signifying', 'renewal', 'well', 'change', 'sworn', 'almighty', 'god', 's
olemn', 'oath', 'forebears', 'l', 'prescribed', 'nearly', 'century', 'three', 'quarters', 'ago', 'world', 'different', 'man',
'holds', 'mortal', 'hands', 'power', 'abolish', 'forms', 'human', 'poverty', 'forms', 'human', 'life', 'yet', 'revolutionary',
'beliefs', 'forebears', 'fought', 'still', 'issue', 'around', 'globe', 'belief', 'rights', 'man', 'come', 'generosity', 'stat
e', 'hand', 'god', 'dare', 'forget', 'today', 'heirs', 'first', 'revolution', 'let', 'word', 'go', 'forth', 'time', 'place', 'f
riend', 'foe', 'alike', 'torch', 'passed', 'new', 'generation', 'americans', 'born', 'century', 'tempered', 'war', 'discipline
d', 'hard', 'bitter', 'peace', 'proud', 'ancient', 'heritage', 'unwilling', 'witness', 'permit', 'slow', 'undoing', 'human', 'r
ights', 'nation', 'always', 'committed', 'committed', 'today', 'home', 'around', 'world', 'let', 'every', 'nation', 'know', 'wh
ether', 'wishes', 'us', 'well', 'ill', 'shall', 'pay', 'price', 'bear', 'burden', 'meet', 'hardship', 'support', 'friend', 'opp
ose', 'foe', 'order', 'assure', 'survival', 'success', 'liberty', 'much', 'pledge', 'old', 'allies', 'whose', 'cultural', 'spir
itual', 'origins', 'share', 'pledge', 'loyalty', 'faithful', 'friends', 'united', 'little', 'cannot', 'host', 'cooperative', 'v
entures', 'divided', 'little', 'dare', 'meet', 'powerful', 'challenge', 'odds', 'split', 'asunder', 'new', 'states', 'welcome',
'ranks', 'free', 'pledge', 'word', 'one', 'form', 'colonial', 'control', 'shall', 'passed', 'away', 'merely', 'replaced', 'fa
r', 'iron', 'tyranny', 'shall', 'always', 'expect', 'find', 'supporting', 'view', 'shall', 'always', 'hope', 'find', 'strongl
y', 'supporting', 'freedom', 'remember', 'past', 'foolishly', 'sought', 'power', 'riding', 'back', 'tiger', 'ended', 'inside',
'peoples', 'huts', 'villages', 'across', 'globe', 'struggling', 'break', 'bonds', 'mass', 'misery', 'pledge', 'best', 'effort
s', 'help', 'help', 'whatever', 'period', 'required', 'communists', 'may', 'seek', 'votes', 'right', 'free', 'society', 'canno
t', 'help', 'many', 'poor', 'cannot', 'save', 'rich', 'sister', 'republics', 'south', 'border', 'offer', 'special', 'pledge',
'convert', 'good', 'words', 'good', 'deeds', 'new', 'alliance', 'progress', 'assist', 'free', 'men', 'free', 'governments', 'ca
sting', 'chains', 'poverty', 'peaceful', 'revolution', 'hope', 'cannot', 'become', 'prey', 'hostile', 'powers', 'let', 'neighbo
rs', 'know', 'shall', 'join', 'oppose', 'aggression', 'subversion', 'anywhere', 'americas', 'let', 'every', 'power', 'know', 'h
emisphere', 'intends', 'remain', 'master', 'house', 'world', 'assembly', 'sovereign', 'states', 'united', 'nations', 'last', 'b
est', 'hope', 'age', 'instruments', 'war', 'far', 'outpaced', 'instruments', 'peace', 'renew', 'pledge', 'support', 'prevent',
'becoming', 'merely', 'forum', 'invective', 'strengthen', 'shield', 'new', 'weak', 'enlarge', 'area', 'writ', 'may', 'run', 'fi
```

- 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.
- 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.

SPEECH BEFORE STOPWORDS: 1973 NIXON'S SPEECH

```
['Mr', '.', 'Vice', 'President', ',', 'Mr', '.', 'Speaker', ',', 'Mr', '.', 'Chief', 'Justice', ',', 'Senator', 'Cook', ',',
'Mrs', '.', 'Eisenhower', ',', 'and', 'my', 'fellow', 'citizens', 'of', 'this', 'great', 'and', 'good', 'country', 'we', 'sha
re', 'together', ':', 'When', 'we', 'met', 'here', 'four', 'years', 'ago', ',', 'America', 'was', 'bleak', 'in', 'spirit',
',', 'depressed', 'by', 'the', 'prospect', 'of', 'seemingly', 'endless', 'war', 'abroad', 'and', 'of', 'destructive', 'confli
ct', 'at', 'home', '.', 'As', 'we', 'meet', 'here', 'today', ',', 'we', 'stand', 'on', 'the', 'threshold', 'of', 'a', 'new',
'era', 'of', 'peace', 'in', 'the', 'world', '.', 'The', 'central', 'question', 'before', 'us', 'is', ':', 'How', 'shall', 'w
e', 'use', 'that', 'peace', '?', 'Let', 'us', 'resolve', 'that', 'this', 'era', 'we', 'are', 'about', 'to', 'enter', 'will',
'not', 'be', 'what', 'other', 'postwar', 'periods', 'have', 'so', 'often', 'been', ':', 'a', 'time', 'of', 'retreat', 'and',
'isolation', 'that', 'leads', 'to', 'stagnation', 'at', 'home', 'and', 'invites', 'new', 'danger', 'abroad', '.', 'Let', 'u
s', 'resolve', 'that', 'this', 'will', 'be', 'what', 'it', 'can', 'become', ':', 'a', 'time', 'of', 'great', 'responsibilitie
s', 'greatly', 'borne', ',', 'in', 'which', 'we', 'renew', 'the', 'spirit', 'and', 'the', 'promise', 'of', 'America', 'as',
'we', 'enter', 'our', 'third', 'century', 'as', 'a', 'nation', '.', 'This', 'past', 'year', 'saw', 'far', '-', 'reaching', 'r
esults', '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', 'n
ations', 'of', 'the', 'world', '.', 'Because', 'of', 'America', "'", 's', 'bold', 'initiatives', ',', '1972', 'will', 'be',
'long', 'remembered', 'as', 'the', 'year', 'of', 'the', 'greatest', 'progress', 'since', 'the', 'end', 'of', 'World', 'War',
'II', 'toward', 'a', 'lasting', 'peace', 'in', 'the', 'world', '.', 'The', 'peace', 'we', 'seek', 'in', 'the', 'world', 'is',
'not', 'the', 'flimsy', 'peace', 'which', 'is', 'merely', 'an', 'interlude', 'between', 'wars', ',', 'but', 'a', 'peace', 'wh
```

SPEECH AFTER STOPWORDS: 1973 NIXON'S SPEECH

```
The number of words after after removing stopwords is>> 836
['mr', 'vice', 'president', 'mr', 'speaker', 'mr', 'chief', 'justice', 'senator', 'cook', 'mrs', 'eisenhower', 'fellow', 'citiz
ens', 'great', 'good', 'country', 'share', 'together', 'met', 'four', 'years', 'ago', 'america', 'bleak', 'spirit', 'depresse
d', 'prospect', 'seemingly', 'endless', 'war', 'abroad', 'destructive', 'conflict', 'home', 'meet', 'today', 'stand', 'threshol
d', 'new', 'era', 'peace', 'world', 'central', 'question', 'us', 'shall', 'use', 'peace', 'let', 'us', 'resolve', 'era', 'ente
r', 'postwar', 'periods', 'often', 'time', 'retreat', 'isolation', 'leads', 'stagnation', 'home', 'invites', 'new', 'danger',
'abroad', 'let', 'us', 'resolve', 'become', 'time', 'great', 'responsibilities', 'greatly', 'borne', 'renew', 'spirit', 'promis
e', 'america', 'enter', 'third', 'century', 'nation', 'past', 'year', 'saw', 'far', 'reaching', 'results', 'new', 'policies',
'peace', 'continuing', 'revitalize', 'traditional', 'friendships', 'missions', 'peking', 'moscow', 'able', 'establish', 'base',
'new', 'durable', 'pattern', 'relationships', 'among', 'nations', 'world', 'america', 'bold', 'initiatives', '1972', 'long', 'r
emembered', 'year', 'greatest', 'progress', 'since', 'end', 'world', 'war', 'ii', 'toward', 'lasting', 'peace', 'world', 'peac
e', 'seek', 'world', 'flimsy', 'peace', 'merely', 'interlude', 'wars', 'peace', 'endure', 'generations', 'come', 'important',
'understand', 'necessity', 'limitations', 'america', 'role', 'maintaining', 'peace', 'unless', 'america', 'work', 'preserve',
'peace', 'peace', 'unless', 'america', 'work', 'preserve', 'freedom', 'freedom', 'let', 'us', 'clearly', 'understand', 'new',
'nature', 'america', 'role', 'result', 'new', 'policies', 'adopted', 'past', 'four', 'years', 'shall', 'respect', 'treaty', 'co
mmitments', 'shall', 'support', 'vigorously', 'principle', 'country', 'right', 'impose', 'rule', 'another', 'force', 'shall',
'continue', 'era', 'negotiation', 'work', 'limitation', 'nuclear', 'arms', 'reduce', 'danger', 'confrontation', 'great', 'power
s', 'shall', 'share', 'defending', 'peace', 'freedom', 'world', 'shall', 'expect', 'others', 'share', 'time', 'passed', 'americ
a', 'make', 'every', 'nation', 'conflict', 'make', 'every', 'nation', 'future', 'responsibility', 'presume', 'tell', 'people',
'nations', 'manage', 'affairs', 'respect', 'right', 'nation', 'determine', 'future', 'also', 'recognize', 'responsibility', 'na
tion', 'secure', 'future', 'america', 'role', 'indispensable', 'preserving', 'world', 'peace', 'nation', 'role', 'indispensabl
e', 'preserving', 'peace', 'together', 'rest', 'world', 'let', 'us', 'resolve', 'move', 'forward', 'beginnings', 'made', 'let',
'us', 'continue', 'bring', 'walls', 'hostility', 'divided', 'world', 'long', 'build', 'place', 'bridges', 'understanding', 'des
pite', 'profound', 'differences', 'systems', 'government', 'people', 'world', 'friends', 'let', 'us', 'build', 'structure', 'pe
ace', 'world', 'weak', 'safe', 'strong', 'respects', 'right', 'live', 'different', 'system', 'would', 'influence', 'others', 's
trength', 'ideas', 'force', 'arms', 'let', 'us', 'accept', 'high', 'responsibility', 'burden', 'gladly', 'gladly', 'chance', 'b
uild', 'peace', 'noblest', 'endeavor', 'nation', 'engage', 'gladly', 'also', 'act', 'greatly', 'meeting', 'responsibilities',
```

- 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.

- 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.

ROOSEVELT'S SAMPLE SPEECH AFTER STOPWORDS

'national day inauguration since people renewed sense dedication united states washington day task people create weld together nation lincoln day task people preserve nation disruption within day task people save nation institutions disruption without us come time midst swift happenings pause moment take stock recall place history rediscover may risk real peril inaction lives nat ions determined count years lifetime human spirit life man three score years ten little little less life nation fullness measur e live men doubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplained re ason tyranny slavery become surging wave future freedom ebbing tide americans know true eight years ago life republic seemed fr ozen fatalistic terror proved true midst shock acted acted quickly boldly decisively later years living years fruitful years pe ople democracy brought us greater security hope better understanding life ideals measured material things vital present future experience democracy successfully survived crisis home put away many evil things built new structures enduring lines maintained

KENNEDY'S SAMPLE SPEECH AFTER STOPWORDS

'vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change s worn almighty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands p ower abolish forms human poverty forms human life yet revolutionary beliefs forebears fought still issue around globe belief ri ghts man come generosity state hand god dare forget today heirs first revolution let word go forth time place friend foe alike torch passed new generation americans born century tempered war disciplined hard bitter peace proud ancient heritage unwilling witness permit slow undoing human rights nation always committed committed today home around world let every nation know whethe r wishes us well ill shall pay price bear burden meet hardship support friend oppose foe order assure survival success liberty much pledge old allies whose cultural spiritual origins share pledge loyalty faithful friends united little cannot host coopera tive ventures divided little dare meet powerful challenge odds split asunder new states welcome ranks free pledge word one form colonial control shall passed away merely replaced far iron tyranny shall always expect find supporting view shall always hope find strongly supporting freedom remember past foolishly sought power riding back tiger ended inside peoples huts villages acro ss globe struggling break bonds mass misery pledge best efforts help help whatever period required communists may seek votes ri ght free society cannot help many poor cannot save rich sister republics south border offer special pledge convert good words g ood deeds new alliance progress assist free men free governments casting chains poverty peaceful revolution hope cannot become prey hostile powers let neighbors know shall join oppose aggression subversion anywhere americas let every power know hemispher e intends remain master house world assembly sovereign states united nations last best hope age instruments war far outpaced in

NIXON'S SAMPLE SPEECH AFTER STOPWORDS

'mr vice president mr speaker mr chief justice senator cook mrs eisenhower fellow citizens great good country share together me t four years ago america bleak spirit depressed prospect seemingly endless war abroad destructive conflict home meet today stan d threshold new era peace world central question us shall use peace let us resolve era enter postwar periods often time retreat isolation leads stagnation home invites new danger abroad let us resolve become time great responsibilities greatly borne renew spirit promise america enter third century nation past year saw far reaching results new policies peace continuing revitalize t raditional friendships missions peking moscow able establish base new durable pattern relationships among nations world america bold initiatives 1972 long remembered year greatest progress since end world war ii toward lasting peace world peace seek world flimsy peace merely interlude wars peace endure generations come important understand necessity limitations america role mainta ining peace unless america work preserve peace peace unless america work preserve freedom freedom let us clearly understand new nature america role result new policies adopted past four years shall respect treaty commitments shall support vigorously princ iple country right impose rule another force shall continue era negotiation work limitation nuclear arms reduce danger confront ation great powers shall share defending peace freedom world shall expect others share time passed america make every nation co nflict make every nation future responsibility presume tell people nations manage affairs respect right nation determine future also recognize responsibility nation secure future america role indispensable preserving world peace nation role indispensable preserving peace together rest world let us resolve move forward beginnings made let us continue bring walls hostility divided world long build place bridges understanding despite profound differences systems government people world friends let us build structure peace world weak safe strong respects right live different system would influence others strength ideas force arms le t us accept high responsibility burden gladly gladly chance build peace noblest endeavor nation engage gladly also act greatly meeting responsibilities abroad remain great nation remain great nation act greatly meeting challenges home chance today ever h istory make life better america ensure better education better health better housing better transportation cleaner environment

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

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

FREQUENCY OF WORDS IN ROOSEVELT'S SPEECH

```
FreqDist({'nation': 12, 'know': 10, 'spirit': 9, 'life': 9, 'democracy': 9, 'us': 8, 'people': 7, 'america': 7, 'years': 6, 'fr
eedom': 6, ...})
```

TOP 3 WORDS USED IN ROOSEVELT'S SPEECH

```
['nation', 'know', 'spirit']
```

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

FREQUENCY OF WORDS IN KENNEDY'S SPEECH

```
FreqDist({'let': 16, 'us': 12, 'world': 8, 'sides': 8, 'new': 7, 'pledge': 7, 'citizens': 5, 'power': 5, 'shall': 5, 'free': 5,
...})
```

TOP 3 WORDS IN KENNEDY'S SPEECH

```
['let', 'us', 'world']
```

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

FREQUENCY OF WORDS IN NIXON'S SPEECH

```
FreqDist({'us': 26, 'let': 22, 'america': 21, 'peace': 19, 'world': 18, 'new': 15, 'nation': 11, 'responsibility': 11, 'governm
ent': 10, 'great': 9, ...})
```

TOP 3 WORDS IN NIXON'S SPEECH

```
['us', 'let', 'america']
```

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

Q.2.4. Plot the word cloud of each of the three speeches. (After removing the stopwords).

FIGURE 57: WORD CLOUD - ROOSEVELT 1941

FIGURE 58: WORD CLOUD - KENNEDY 1961

>>>>>THE END<<<<<

FIGURE 59: WORD CLOUD - NIXON 1973