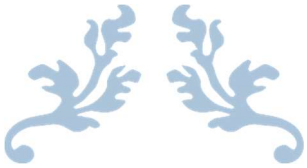


PROJECT REPORT

MACHINE LEARNING

DECEMBER 10, 2023

SHUBHAM KUMAR

Contents

Problem 1

Problem 2

Problem 1:

You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

1.1. Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.

Solution: -

- **Head (First Five Rows of the Dataset)**

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

Table 1.1A
First Five Rows of the Dataset

- **Tail (Last Five Rows of the Dataset)**

	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1520	1521	Conservative	67	5	3	2	4	11	3	male
1521	1522	Conservative	73	2	2	4	4	8	2	male
1522	1523	Labour	37	3	3	5	4	2	2	male
1523	1524	Conservative	61	3	3	1	4	11	2	male
1524	1525	Conservative	74	2	3	2	4	11	0	female

Table 1.1B
Last Five Rows of the Dataset

- **Datatypes of all the features in a dataset**

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

Table 1.1C
Datatypes of the features in a dataset

From the above Table 1.1C –

- There 10 variables/attributes in a dataset.
  1. int64: - There are 8 Integer datatype
     (Unnamed: 0, age, economic.cond.national, economic.cond.household, Blair, Hague, Europe, political.knowledge)

  2. object: - There are 2 Object datatype
     (vote, gender)

- **Shape of the Dataset**

```
Number of Rows in a Dataset =  1525
Number of Columns in a Dataset =  10
```

Fig. 1.1A
Shape of the Dataset

- **Descriptive Statistics of the Dataset**

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	1525.0	NaN	NaN	NaN	763.0	440.373894	1.0	382.0	763.0	1144.0	1525.0
vote	1525	2	Labour	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	1525.0	NaN	NaN	NaN	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	NaN	NaN	NaN	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	NaN	NaN	NaN	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	NaN	NaN	NaN	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	NaN	NaN	NaN	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	NaN	NaN	NaN	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	NaN	NaN	NaN	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0
gender	1525	2	female	812	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 1.1D
Descriptive Statistics of the Dataset

From Table 1.1D we can see the Descriptive stats or 5-point summary of the dataset. The 5-point summary includes min, 25%, 50%, 75% and max. It also includes mean and standard deviation.

For Categorical variables, Descriptive stats includes unique (number of unique values present in that column), freq (it tells us about the number of times for a category which occurred most in the column)

Also column “Unamed: 0” signifies nothing significant but the index or numbers i.e., 1 to 1525 which signifies nothing important and adds an extra feature to the dataset so it can be dropped.

- **Duplicated data in a Dataset**

```
Total number of duplicated data/ rows in a dataset =  8
```

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender
67	Labour	35	4	4	5	2	3	2	male
626	Labour	39	3	4	4	2	5	2	male
870	Labour	38	2	4	2	2	4	3	male
983	Conservative	74	4	3	2	4	8	2	female
1154	Conservative	53	3	4	2	2	6	0	female
1236	Labour	36	3	3	2	2	6	2	female
1244	Labour	29	4	4	4	2	2	2	female
1438	Labour	40	4	3	4	2	2	2	male

Table 1.1E
Duplicated data in a Dataset

From above Table 1.1E, we can see there 8 duplicated rows in a dataset, which are to be dropped.

```
Total number of duplicated data/ rows in a dataset =  0
```

vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender

Fig. 1.1B
Duplicated data after dropping duplicates

- **<u>Check for Null Values in a Dataset</u>**

```
Total number of Null Entries in each columns -

vote                    0
age                     0
economic_cond_national  0
economic_cond_household 0
Blair                   0
Hague                   0
Europe                  0
political_knowledge     0
gender                  0
dtype: int64
```

Fig. 1.1C
Null Values in a Dataset

- **<u>Check for Skewness</u>**

```
age                     0.139800
economic_cond_national -0.238474
economic_cond_household -0.144148
Blair                  -0.539514
Hague                   0.146191
Europe                 -0.141891
political_knowledge    -0.422928
dtype: float64
```

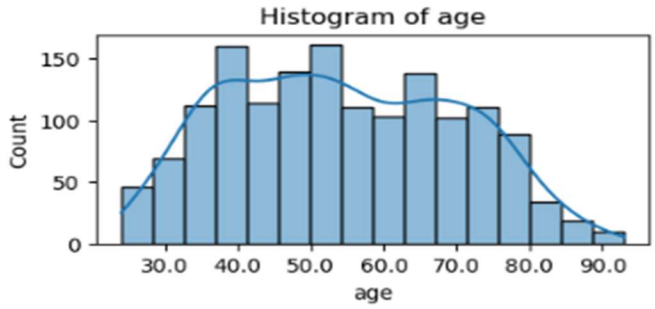
Table 1.1F
Skewness of the numeric variables

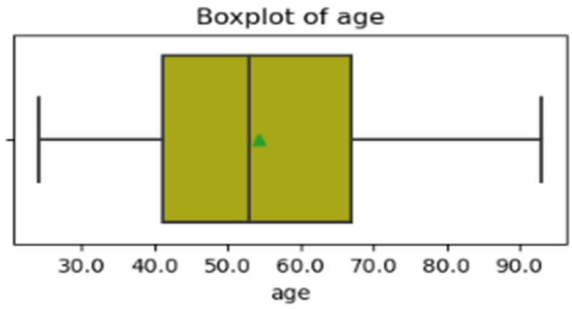
- A **negative value** for skewness indicates that the tail is on the left side of the distribution, which extends towards more negative values.
- A **positive value** for skewness indicates that the tail is on the right side of the distribution, which extends towards more positive values.
- So, except “age” and “Hague” all the variables are left skewed

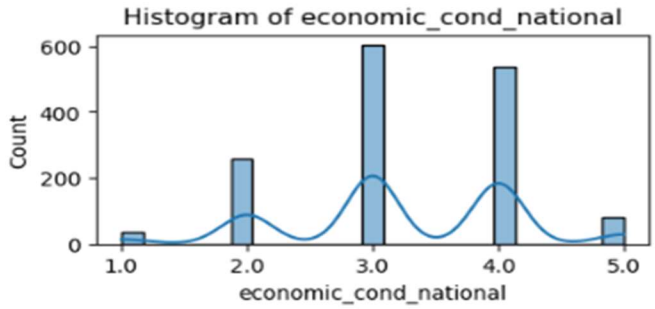
1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

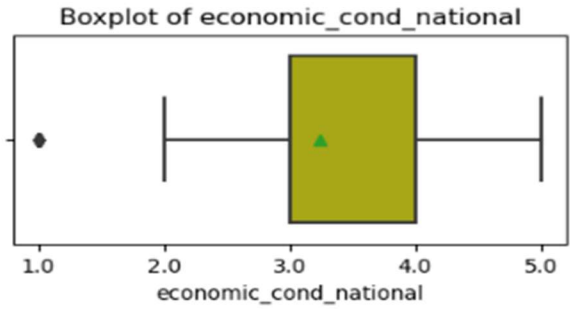
Solution: -

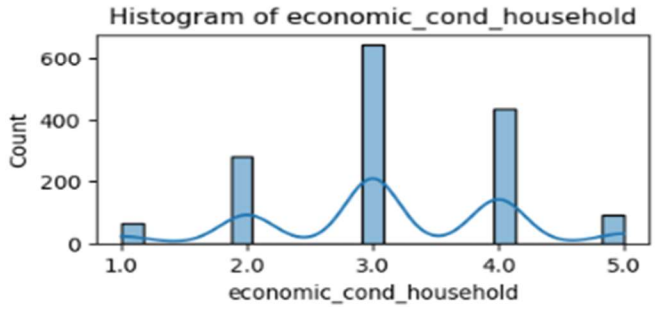
UNIVARIATE ANALYSIS OF NUMERIC VARIABLES/ATTRIBUTES

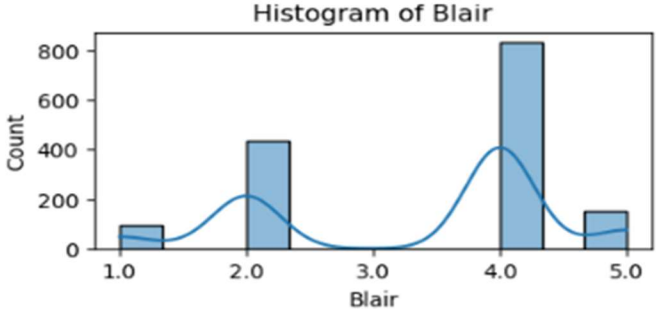


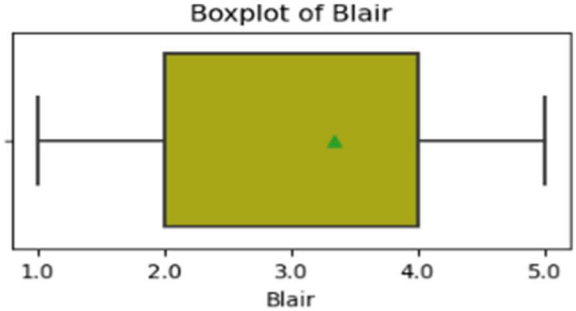












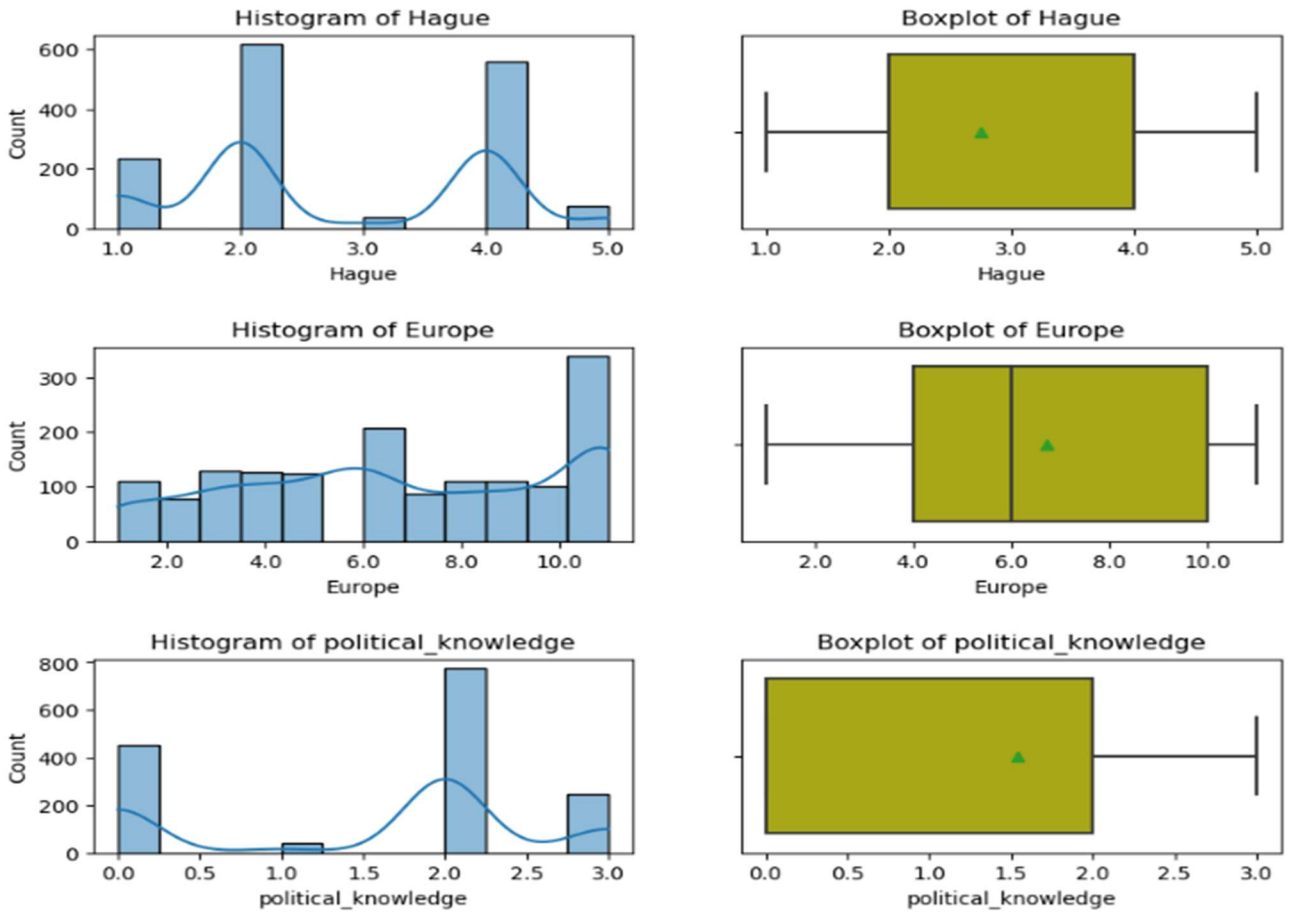


Fig 1.2A

Histogram and Boxplot of Numeric Variables in the Dataset

Inferences from Fig 1.2A: -

- **Age -** The data distribution is nearly normally distributed. The distribution has negligible skewness as mean is greater than median and no outliers found.
- **National economic condition -** The data distribution cannot be determinate, due to categorical in nature. Also we can see outliers present on lower side.
- **Economic household condition- -** The data distribution cannot be determinate, due to categorical in nature. Also we can see outliers present on lower side.
- **Blair -** The data distribution cannot be fully determined, due to categorical in nature. There are no outliers present in this column
- **Hague --** The data distribution cannot be fully determined, due to categorical in nature. There are no outliers present in this column
- **Europe -** No outliers found. The data distribution cannot be fully ascertained, due to categorical in nature.
- **Political Knowledge -** No outliers found. Due to categorical in nature, the data distribution cannot be fully ascertained.

Age: -

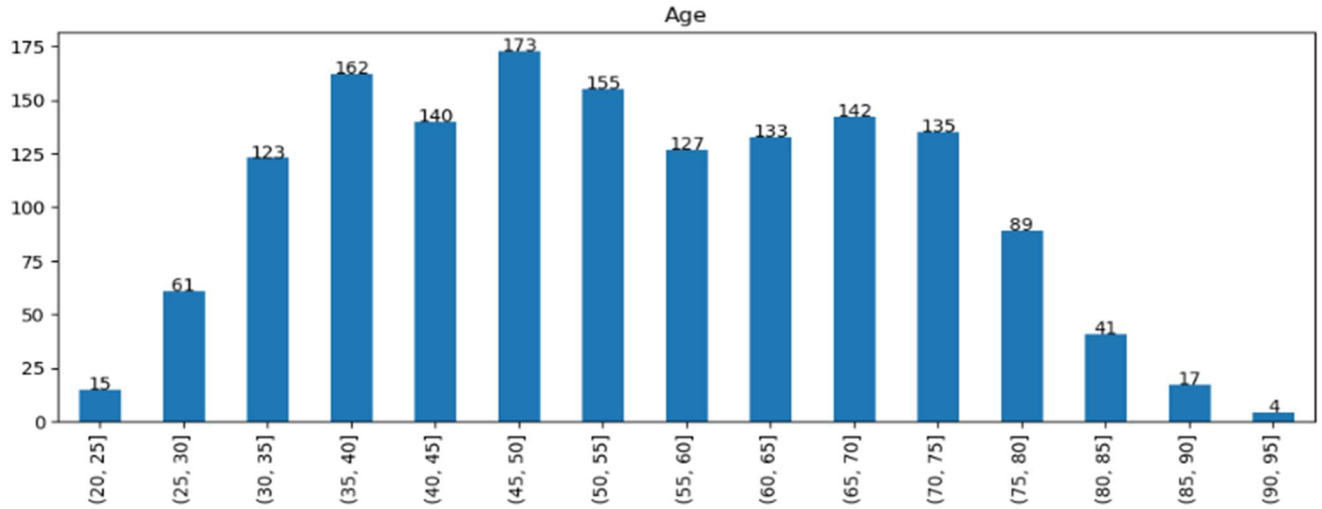


Fig. 1.2B

Count of people in Age Categories

Inferences from Age columns: -

- Maximum number of people are of age group (40-55) i.e., 468 almost 31 % of total
- There are very few young age people (20-25) i.e., 15 approx. 1% of total
- Also there are very few extremely old people (85-95) i.e., 21 approx. 1% of total
- Maximum number of people are in age group (30-75) i.e., 1290 approx. 85% of total

National Economic Condition: -

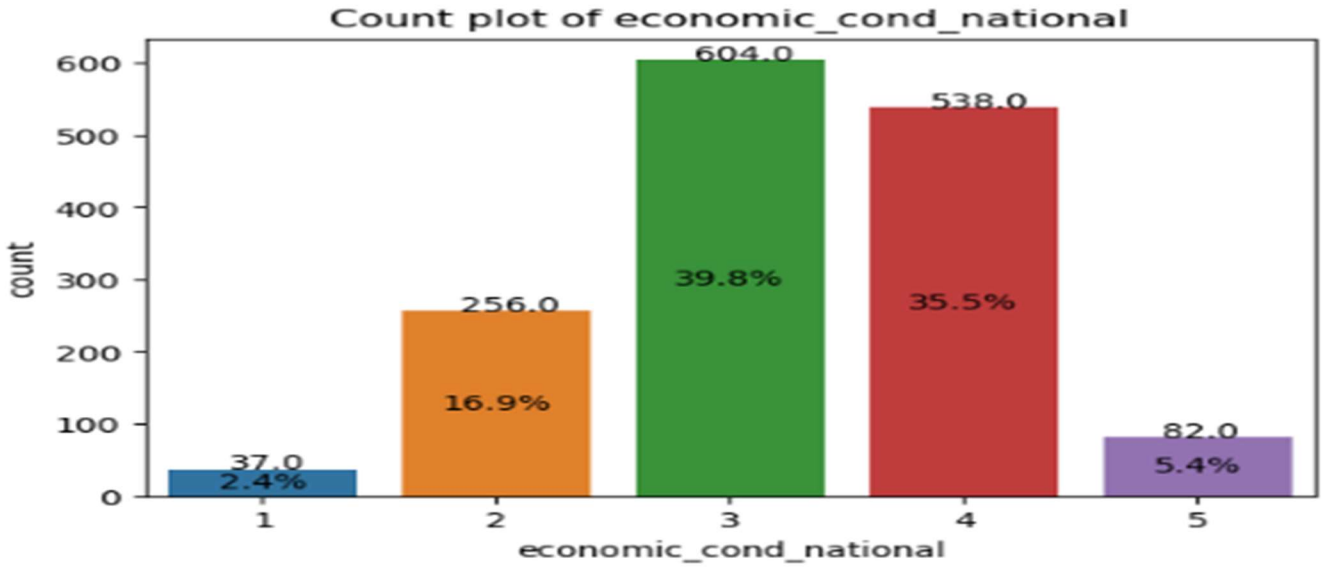


Fig 1.2C

Inferences from Fig 1.2C: -

- The above graphical illustrations show, that a significant percentage of voters have rated National Economic Condition Medium i.e., 604 or 39.8%
- Very few percentage of voters have rated National Economic condition very low or very high i.e 2.4% (37) or 5.4% (82) respectively.

Economic Condition Household: -

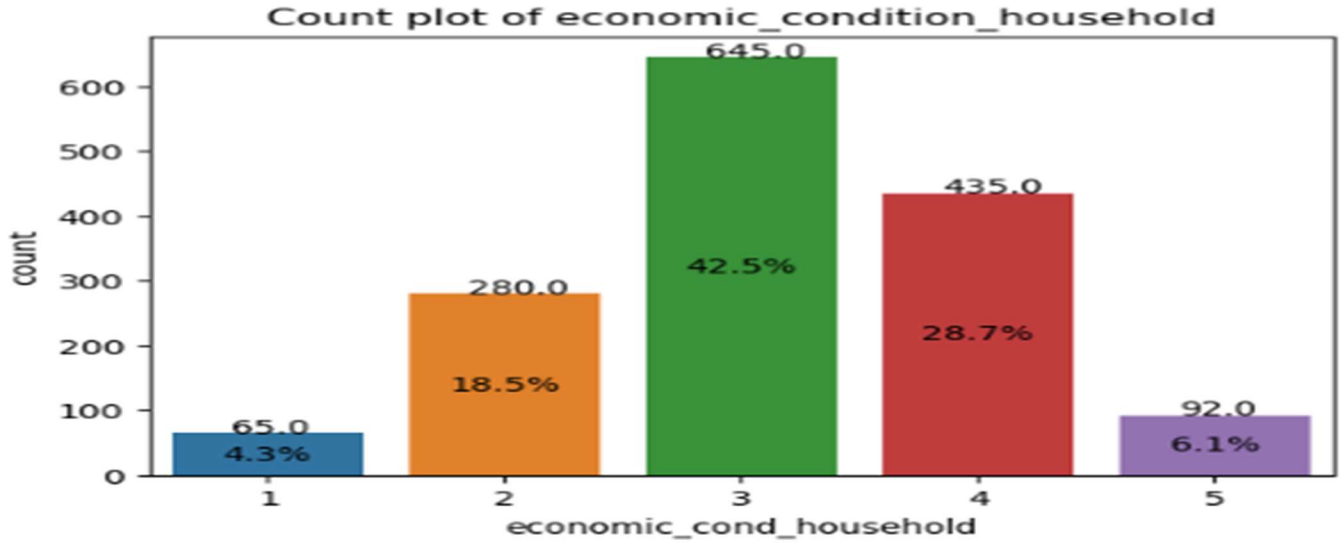


Fig. 1.2D

Inferences from Fig 1.2D: -

- The above graphical illustrations show, that a significant percentage of voters have rated Economic Condition Household Medium i.e., 645 or 42.5%
- Very few percentage of voters have rated Economic Condition Household very low or very high i.e 4.3% (65) or 6.1% (92) respectively.

Blair: -

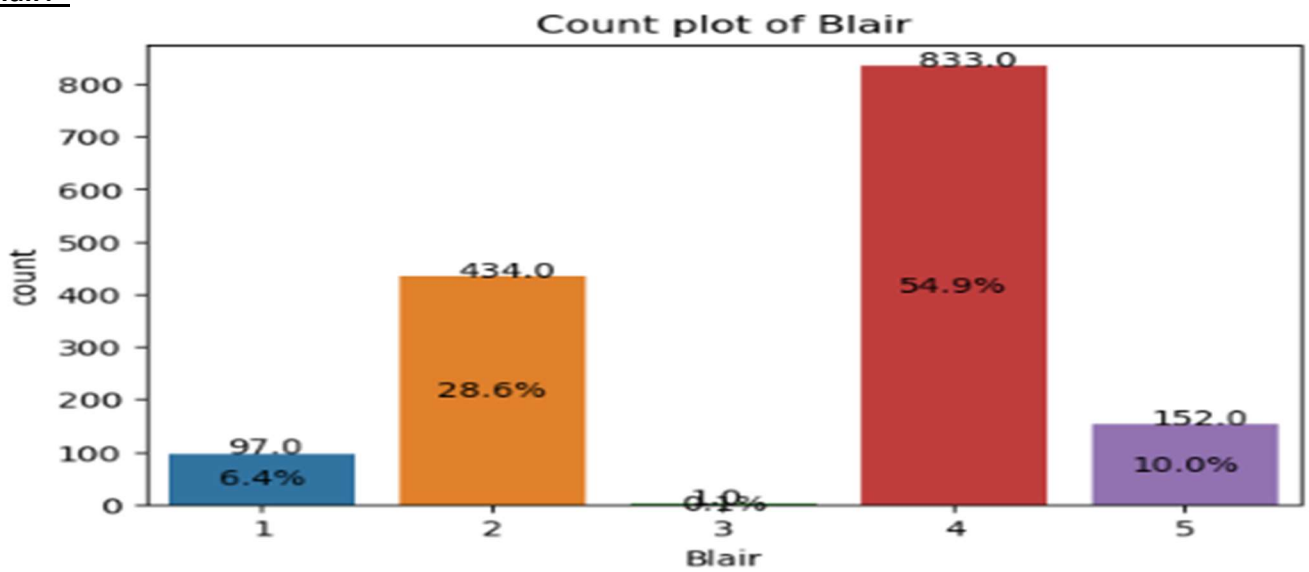


Fig 1.2E

Inferences from Fig 1.2E: -

- The above count plot show, that a significant percentage of voters have rated Blair on higher side (4) i.e., 833 or 54.9% of total voters
- It can be clearly seen that Blair did not receive any average score of 3 as only 1 people who voted has given score 3 i.e., 0.000065915% which is negligible.

Blair: -

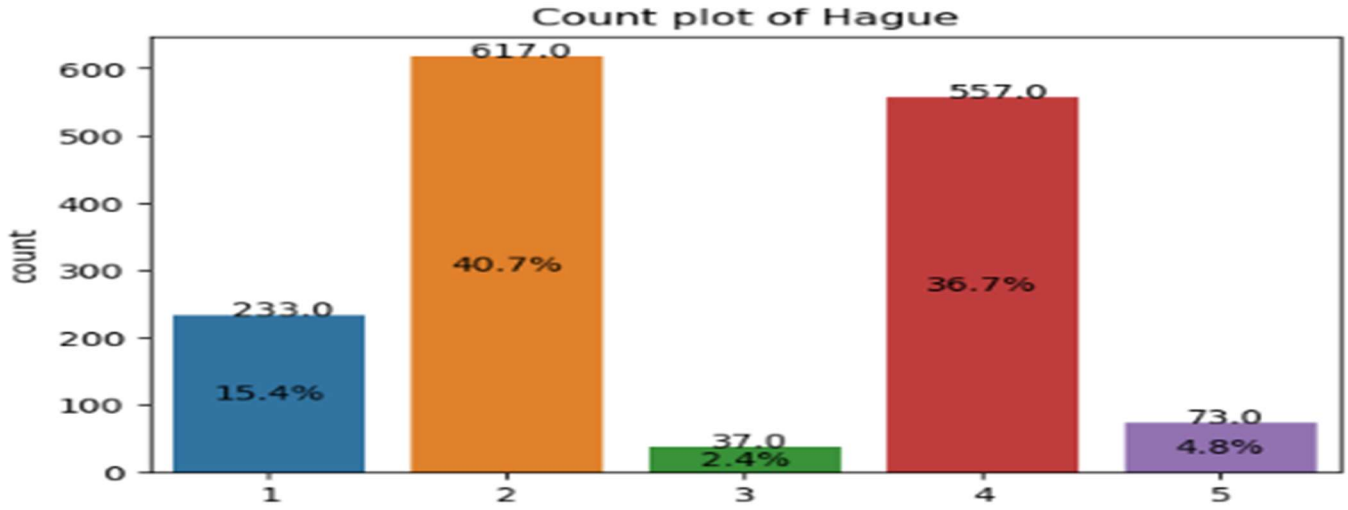


Fig. 1.2F

Inferences from Fig. 1.2F: -

- The above count plot show, that a significant percentage of voters have rated Hague as 2 or 4 i.e., 617 (40.7%) or 557 (36.7%) of total voters
- It can be seen clearly that Hague received very few average score of 3 i.e., 37 or 2.4% of total voters

Europe: -

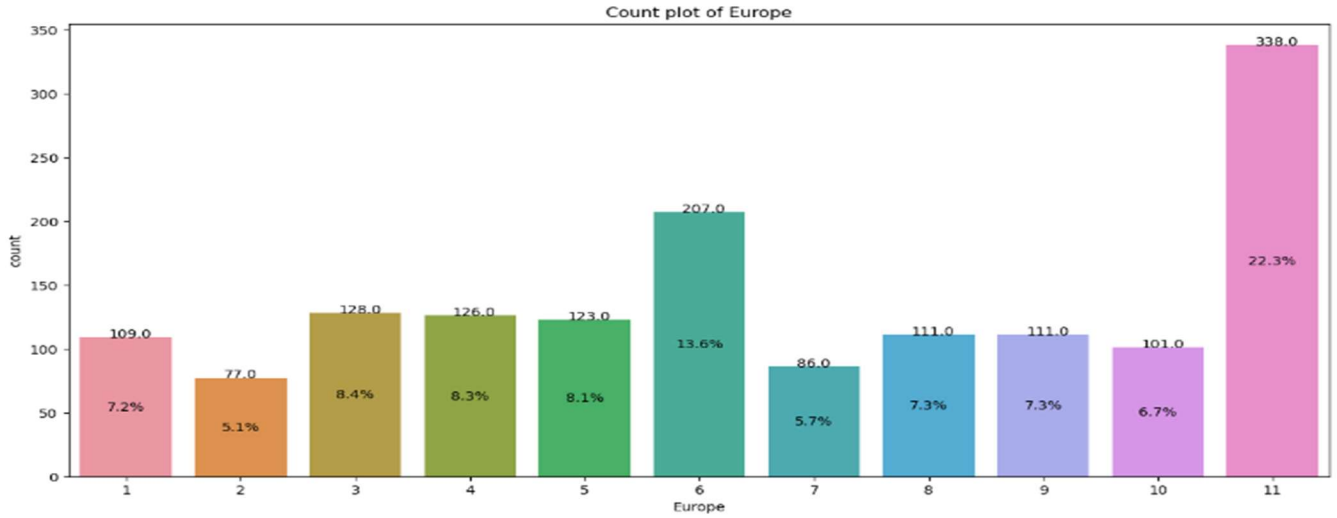


Fig. 1.2G

Inferences from Fig. 1.2G: -

- The above count plot show, that a significant percentage of voters have rated Europe as 11 i.e., 338 (22.3%) of total voters.
- Least percentage of voters who have rated Europe as 2 i.e., 77 or 5.1% of total voters.

Political Knowledge: -

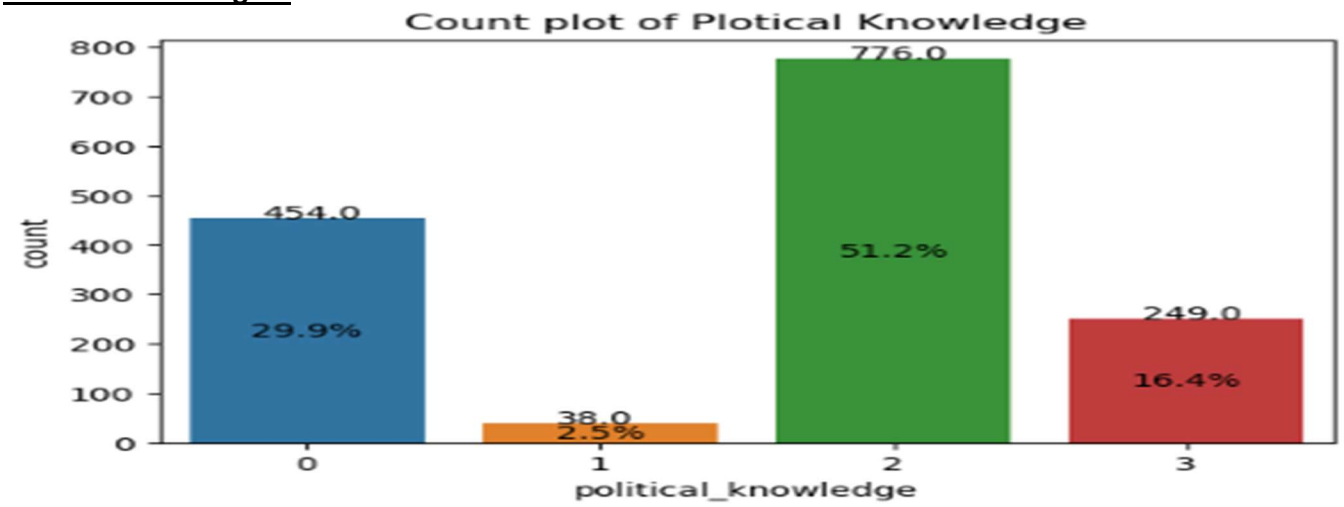


Fig. 1.2H

Inferences from Fig. 1.2H: -

- The above count plot show, that a significant percentage of voters have political knowledge as 2 on the scale of (0-3) i.e., 776 (51.2%) of total voters.
- Least percentage of voters who have political knowledge of 1 on the scale of (0-3) i.e., 38 or 2.5% of total voters.

Univariate Analysis of Categorical Variables

Gender: -

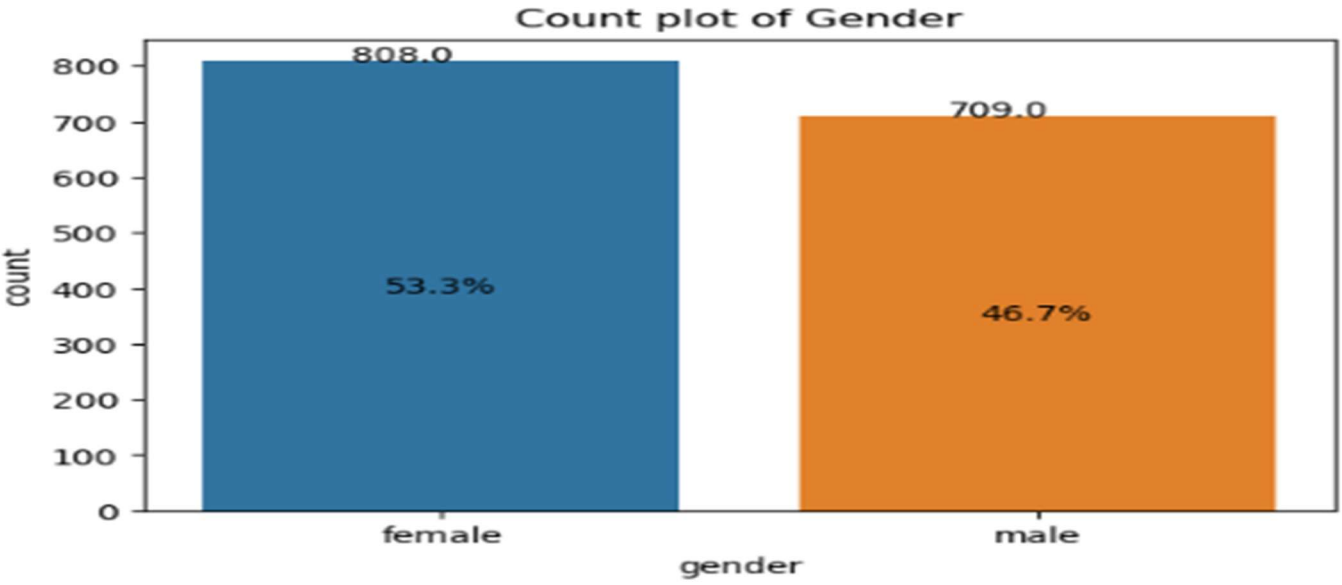


Fig. 1.2I

Inferences from Fig. 1.2I: -

- The above count plot show, that a significant percentage of voters are Female i.e., 808 or 53.3% of total

Vote: -

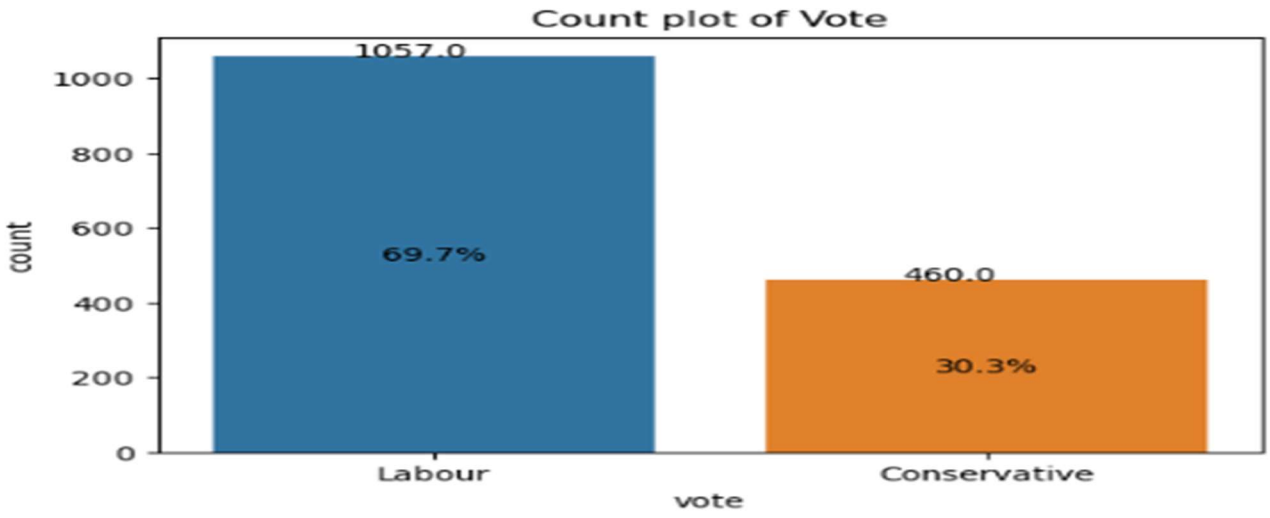


Fig. 1.2J

Inferences from Fig. 1.2J: -

- The above count plot show, that a significant percentage of voters have given vote to Labour party i.e., 1057 or 69.7% of total.

Bi-VARIATE ANALYSIS OF VARAIBLES

Bi-Variate Analysis Between Gender and Vote

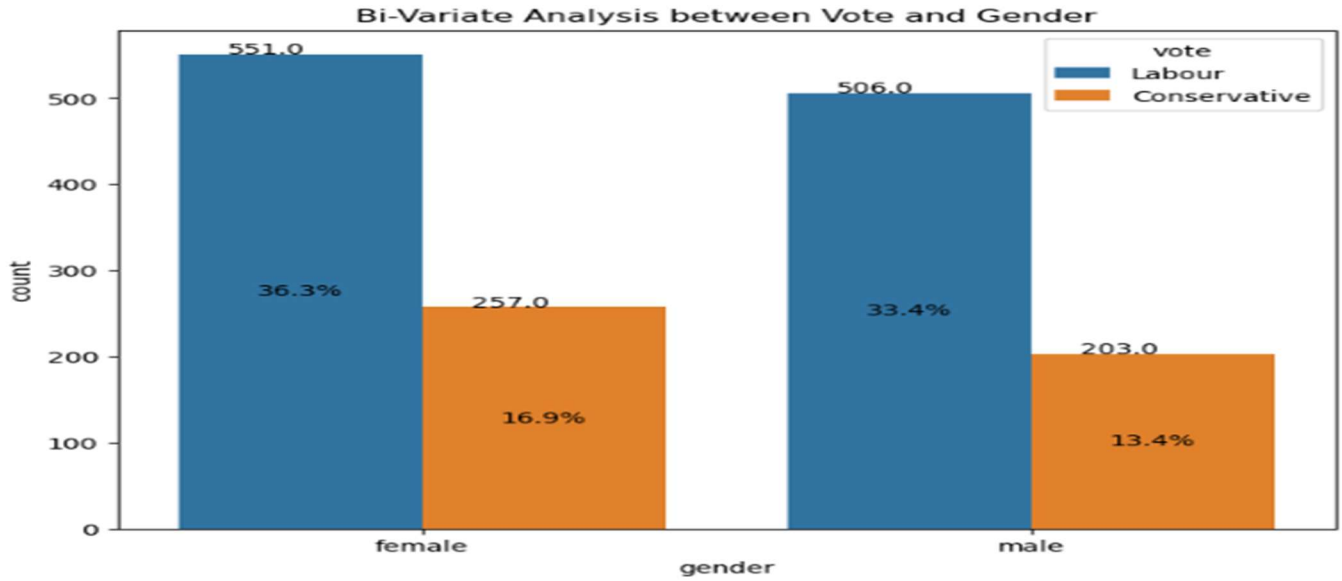


Fig. 1.2K

Inferences from Fig. 1.2K: -

- Almost 36.3% of females and 33.4% of males voted for Labour party
- 16.9% of females and 13.4% of males voted for Conservative party
- Out of 1057 votes to Labour party 551, were female and 506 were male
- Out of 460 votes to Conservative party, 257 were female and 203 were male

Bi-Variate Analysis between Economic Condition National and Vote/Gender

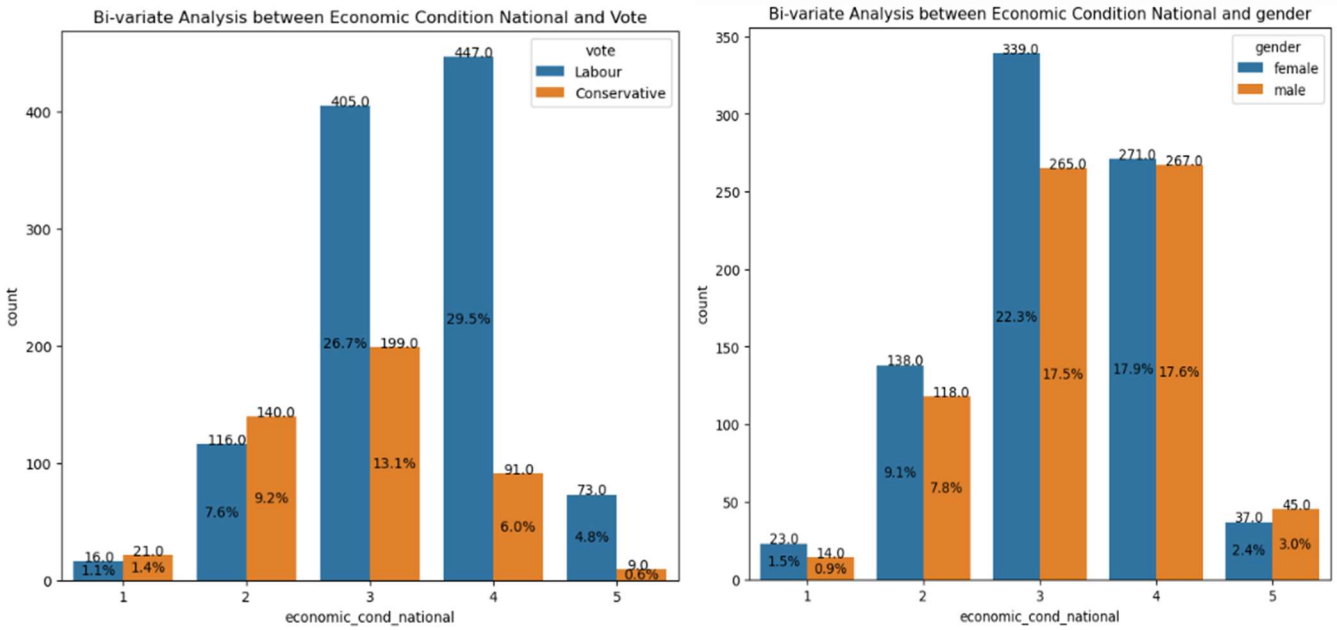


Fig. 1.2L

Inferences from Fig. 1.2L: -

- From the above plot we can see majority voters who voted for Labour party have rated Current Economic Condition Nation as 4 out of 5
- The majority voters who voted for Conservative party have rated Current Economic Condition Nation as 3 out of 5
- From the above plot we can see majority Female voter have rated Current Economic Condition National as 3 out of 5
- The Majority Male voter have rated Current Economic Condition National as 3 or 4 out of 5

Bi-Variate Analysis between Economic Condition Household and Vote/Gender

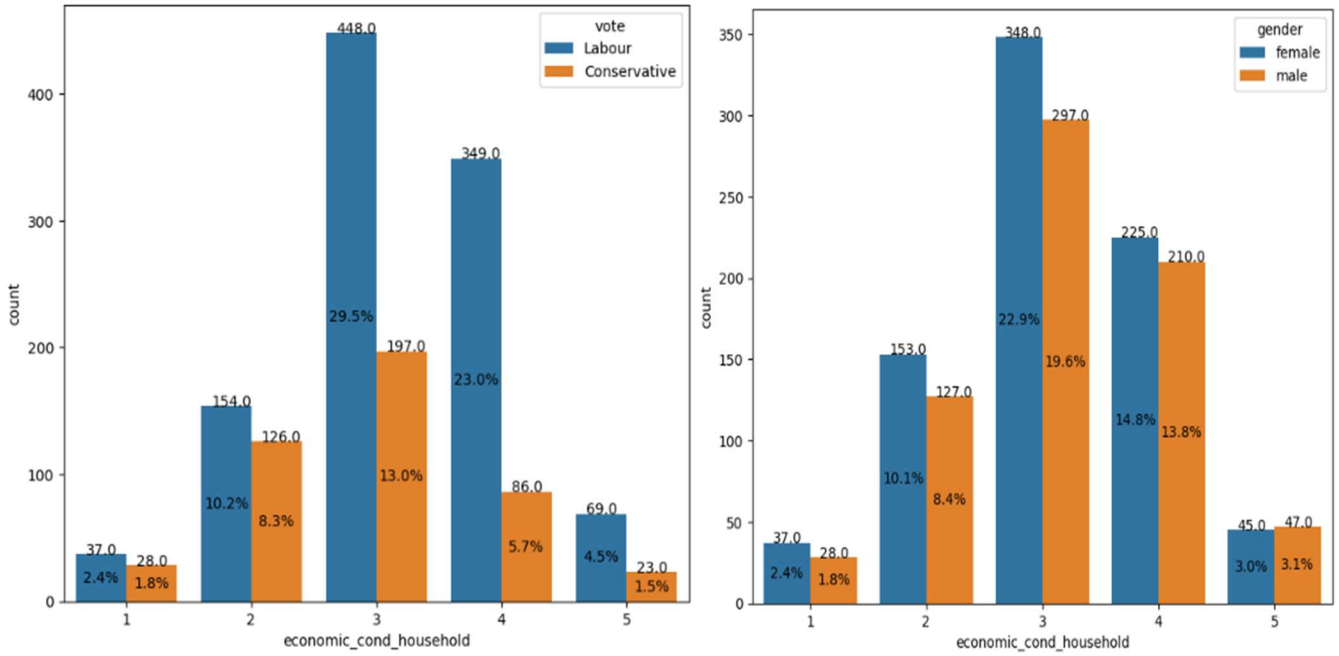


Fig. 1.2M

Inferences from Fig. 1.2M: -

- From the above plot we can see majority voters who voted for Labour party have rated Economic Condition Household Nation as 3 out of 5 i.e., 29.5 % of total voters.
- The majority voters who voted for Conservative party have rated Current Economic Condition Nation as 3 out of 5 i.e., 13 % of total voters.
- From the above plot we can see majority Female and Male voter have rated Current Economic Condition National as 3 out of 5 i.e., 22.9% and 19.6% for each female and male of total voters.

Bi-Variate Analysis between Blair(Assessment of Labour Leader) and Gender

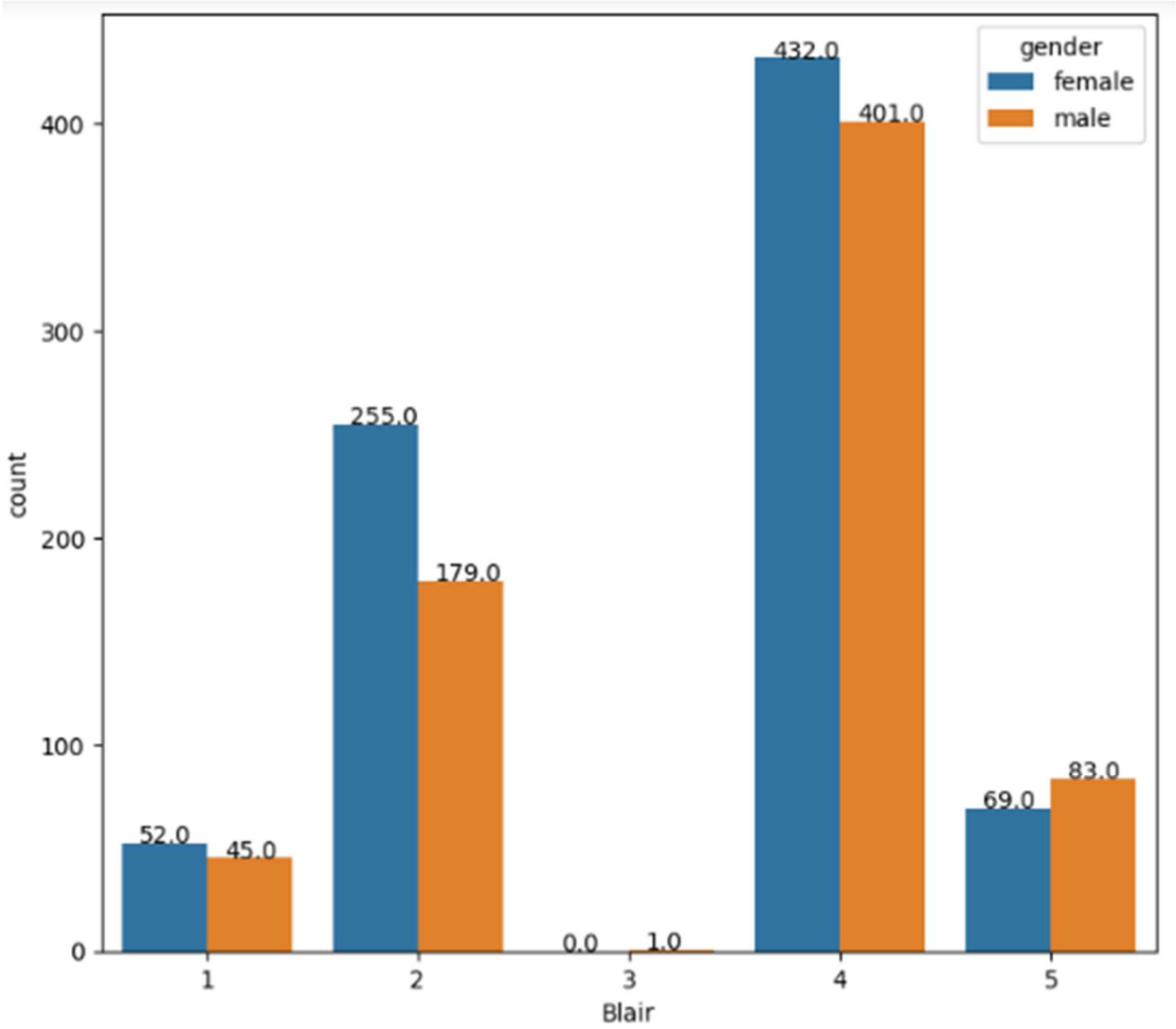


Fig. 1.2N

Inferences from Fig. 1.2N: -

- From the above plot we can see few voters have rated Blair (Assessment of the Labour leader) very low. 3.4% females and 2.9% males have rated him with lowest score of 1 out of 5
- We can see significant number of voters have rated Blair (Assessment of the Labour leader) low. 16.5% of females and 11.7% of males have rated him with highest score of 2 out of 5
- We can see none/negligible voters have rated Blair (Assessment of the Labour leader) neutral. No females and 0.06% males have rated him with highest score of 3 out of 5
- We can see majority voters have rated Blair (Assessment of the Labour leader) high. 44.5% of females and 26.4% of males have rated him with high score of 4 out of 5
- We can see few voters have rated Blair (Assessment of the Labour leader) very high. 4.5% of females and 5.4% of males have rated him with highest score of 5 out of 5

Bi-Variate Analysis between Hague(Assessment of Conservative Leader) and Gender

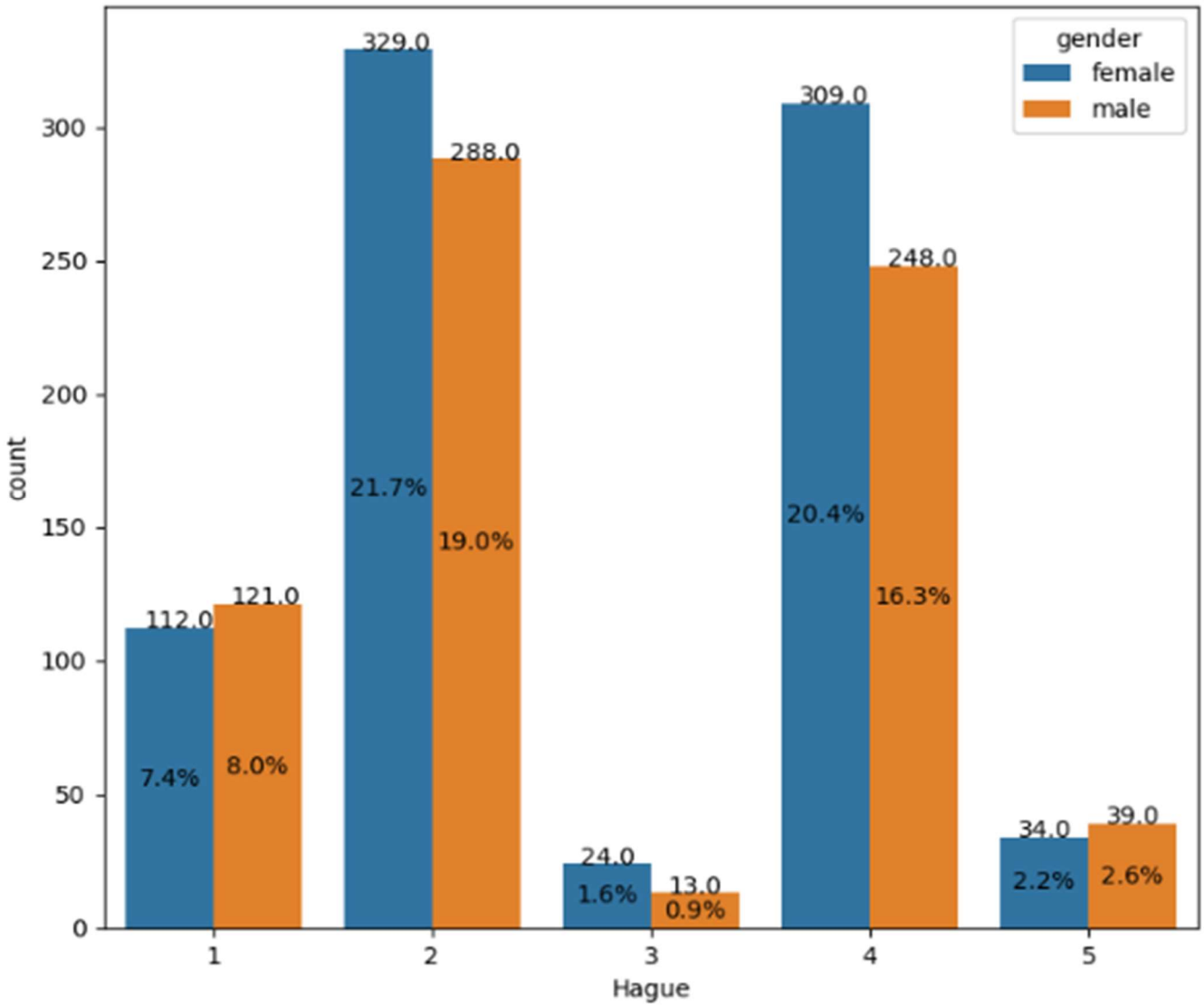


Fig. 1.2 O

Inferences from Fig. 1.2O: -

- From the above plot we can see few voters have rated Hague (Assessment of the Conservative leader) very low. 7.4% females and 8% males have rated him with lowest score of 1 out of 5
- We can see significant number of voters have rated Hague (Assessment of the Conservative leader) low. 21.7% of females and 19% of males have rated him with highest score of 2 out of 5
- We can see none/negligible voters have rated Hague (Assessment of the Conservative leader) neutral. 1.6% of females and 0.9% males have rated him with highest score of 3 out of 5
- We can see majority voters have rated Hague (Assessment of the Conservative leader) high. 20.4% of females and 16.4% of males have rated him with high score of 4 out of 5
- We can see few voters have rated Hague (Assessment of the Conservative leader) very high. 2.2 % of females and 2.6% of males have rated him with highest score of 5 out of 5

Bi-Variate Analysis between Political Knowledge and Gender

Political Knowledge represents Knowledge of parties' positions on European integration, scale 0 - 3

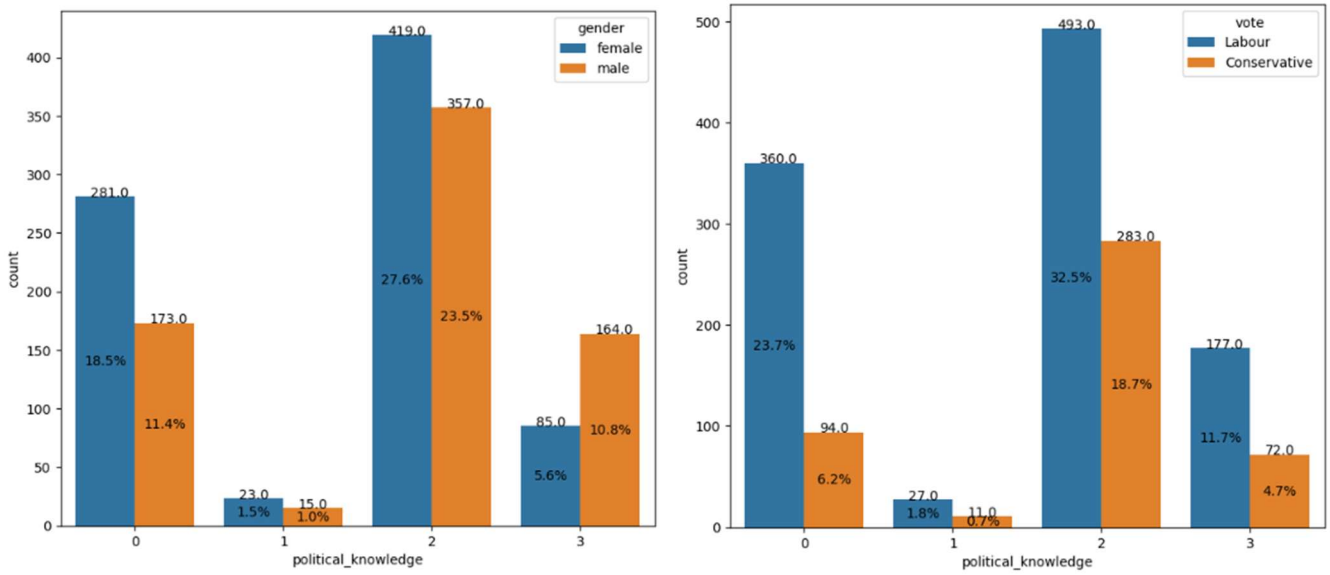


Fig. 1.2P

Inferences from Fig. 1.2P: -

- From the above plot we can see majority voters who voted for Labour party have rated Political Knowledge of 2 out of 3 i.e., 32.5 % of total voters.
- The majority voters who voted for Conservative party have Political Knowledge rated as 3 out of 5 i.e., 18.7 % of total voters.
- From the above plot we can see majority Female and Male voter have Political Knowledge of 2 out of 5 i.e., 27.6% and 23.5% for each female and male of total voters.

Multi-Variate Analysis

Heat Map of Continuous Variables

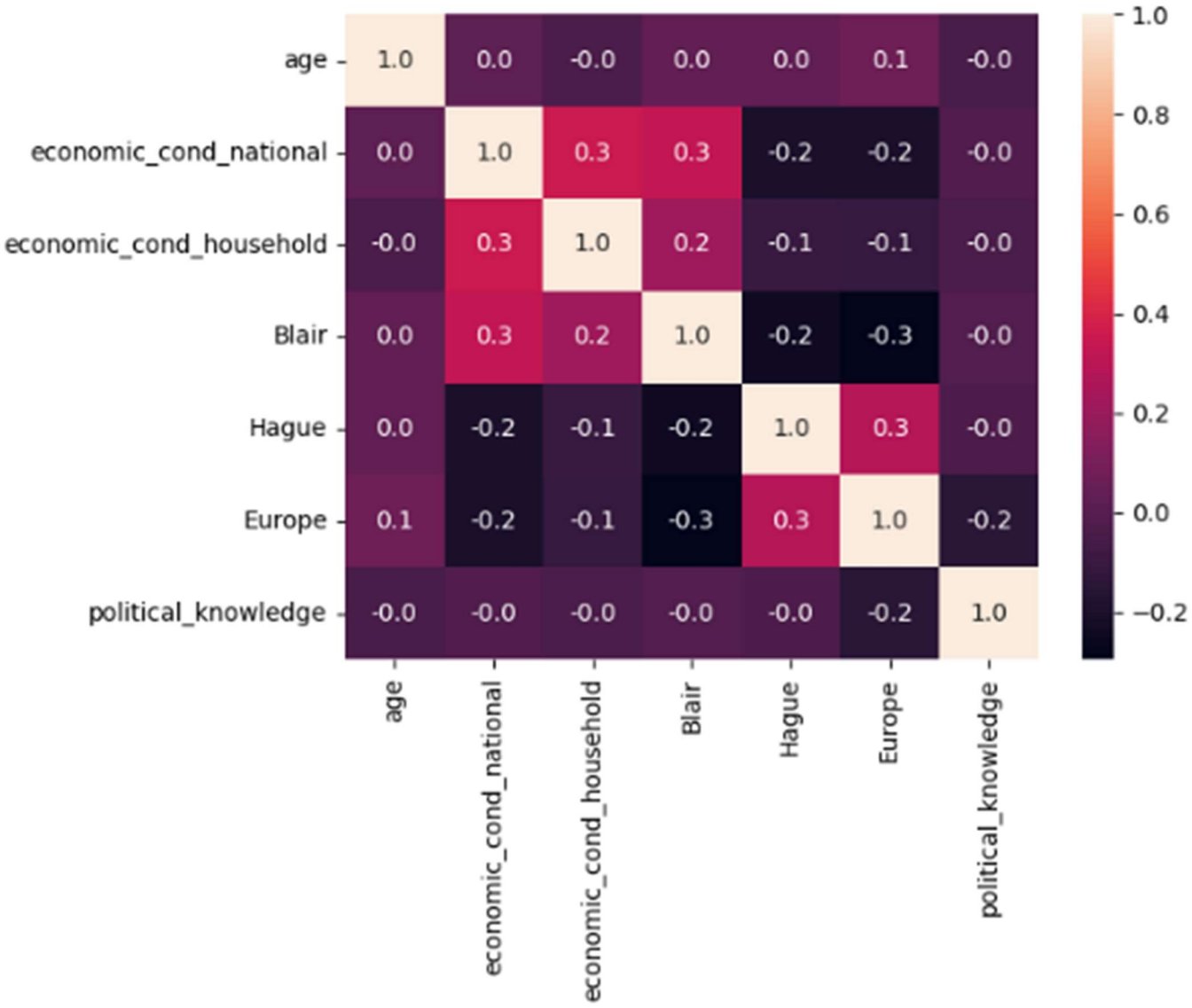


Fig. 1.2Q

Inferences from Fig. 1.2Q: -

- From the above plot Fig. 1.2Q, we can see that none of the correlation value is higher than 0.5, which means none of the continuous variables are related to each other and out dataset is free from multi-collinearity.
- Also there are some variables which are slightly inversely related such as “Europe” and “Economic Condition National” or “Europe” and “Blair” etc. Since the correlation value is not much higher so it will not affect much.

Pair Plot of Variables

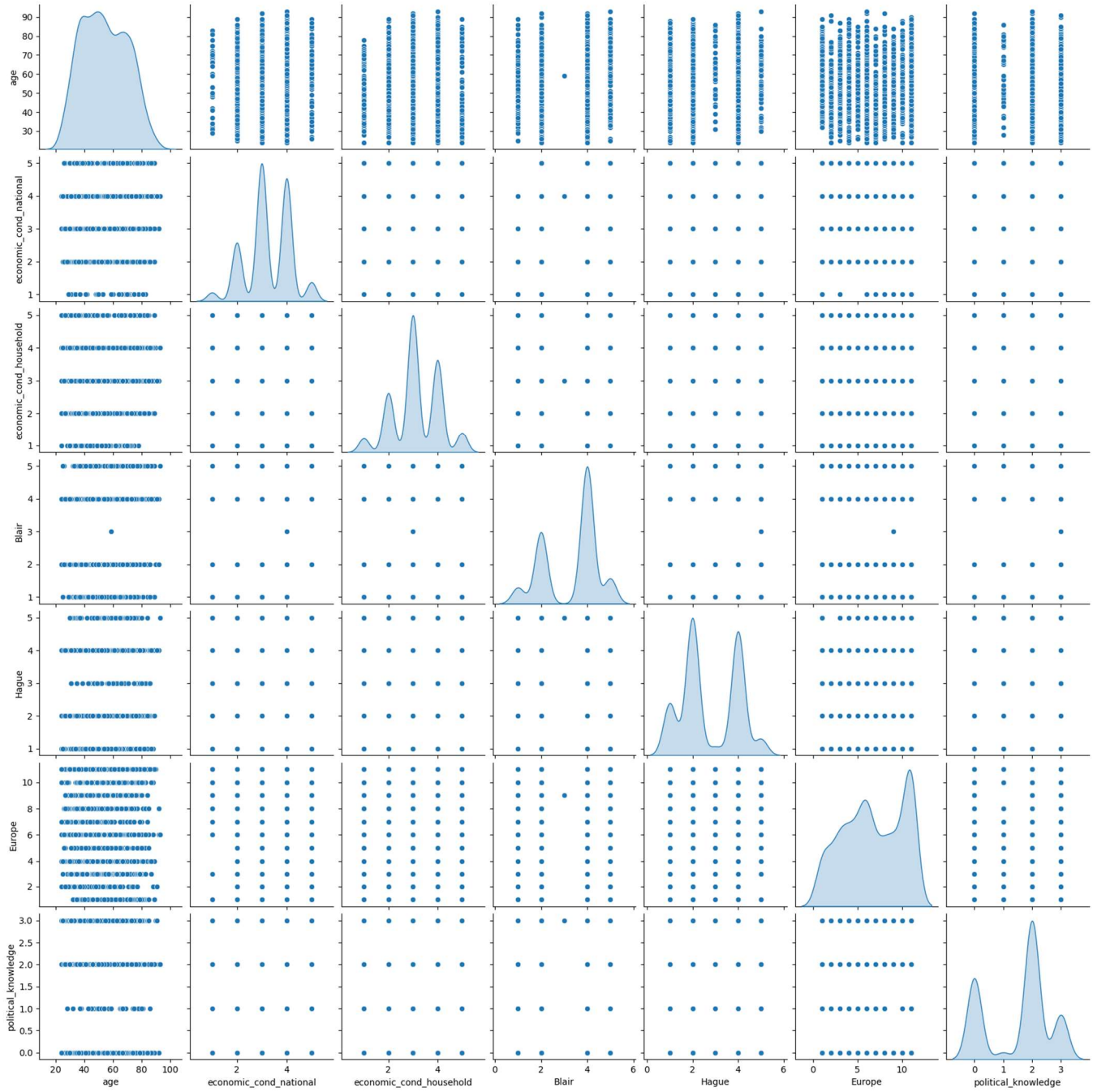


Fig. 1.2R

Check for Outliers: -

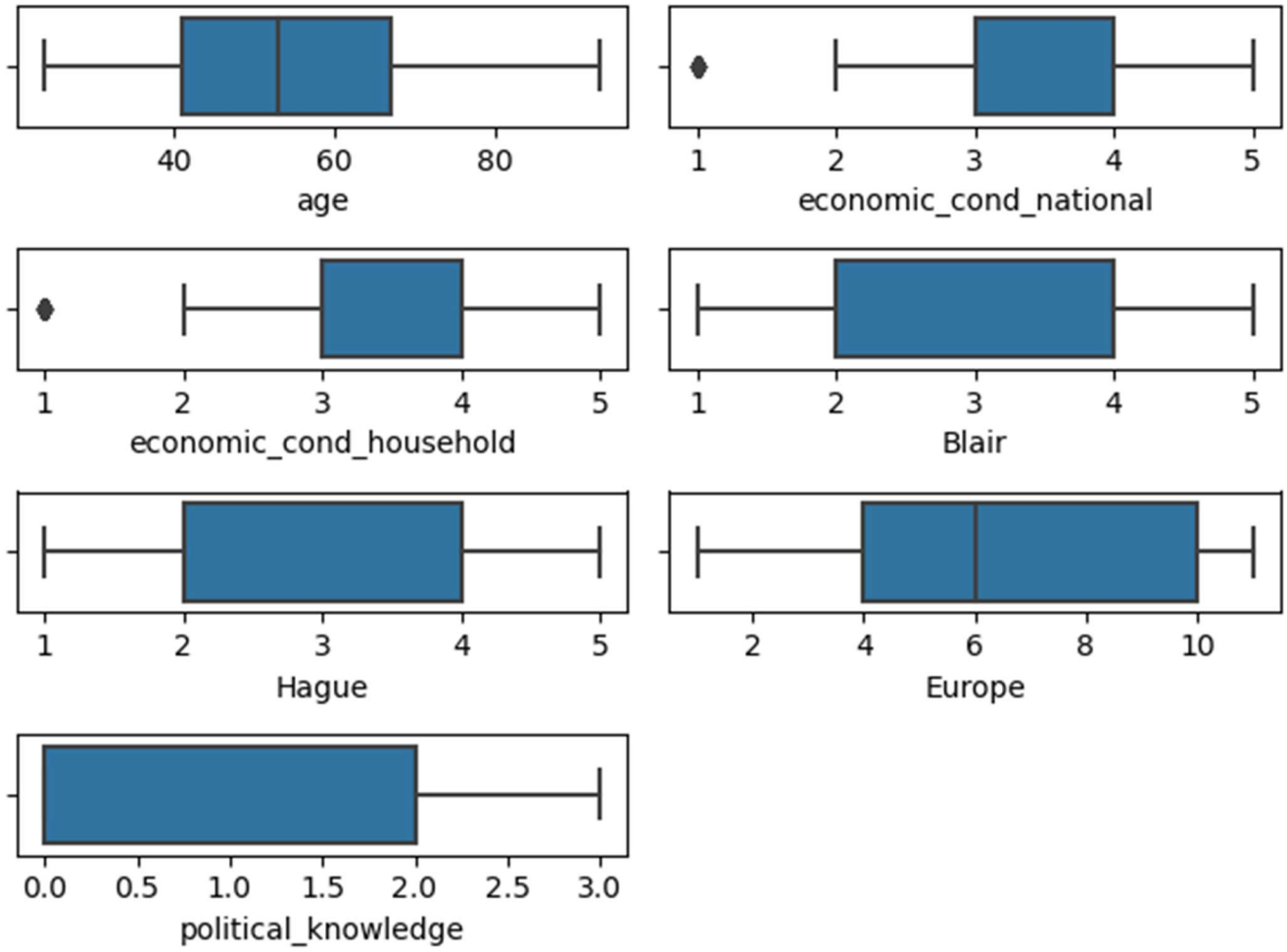


Fig 1.2 S

Inferences from the above plot Fig 1.2 S: -

- From above boxplot of continuous variables, we can see there are outliers present in two features i.e., Economic Condition National and Economic Condition Household and
- All other variables i.e., Age, Blair, Hague, Europe, Political Knowledge doesn't have any outliers

Lower and Upper Fence of Different Continuous Variables

```
Lower Fence and Upper Fence of age  =  2.0 & 106.0
Lower Fence and Upper Fence of economic_cond_national  =  1.5 & 5.5
Lower Fence and Upper Fence of economic_cond_household  =  1.5 & 5.5
Lower Fence and Upper Fence of Blair  =  -1.0 & 7.0
Lower Fence and Upper Fence of Hague  =  -1.0 & 7.0
Lower Fence and Upper Fence of Europe  =  -5.0 & 19.0
Lower Fence and Upper Fence of political_knowledge  =  -3.0 & 5.0
```

Table 1.2A
Lower and Upper Fence of Continuous Variables

Below you can see the dataset where Economic Condition National is less than Lower Limit :-

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge
19	37	3	1	1	1	5	2
21	53	2	1	2	4	5	2
35	41	3	1	4	4	6	3
42	66	1	1	4	2	8	0
62	28	4	1	4	2	6	2
...	...	...	...	...	...	...	...
1469	70	1	1	2	5	11	2
1480	55	2	1	4	4	7	2
1493	34	3	1	4	2	6	2
1501	44	3	1	4	2	9	3
1507	52	2	1	1	4	8	2

65 rows × 7 columns

Table 1.2B

Below you can see the dataset where Economic Condition Household is less than Lower Limit :-

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge
39	72	1	3	2	2	11	2
42	66	1	1	4	2	8	0
57	32	1	2	1	4	1	2
91	49	1	1	2	4	8	3
105	60	1	3	1	4	8	0
109	31	1	1	2	4	11	0
159	67	1	3	2	4	11	0
165	53	1	3	1	4	6	2
215	34	1	1	1	4	11	3
291	60	1	2	2	5	11	2
343	37	1	4	1	5	11	2
385	68	1	1	2	4	10	2
394	83	1	2	2	4	6	0
435	49	1	1	2	2	6	0
441	78	1	1	4	4	1	0
446	66	1	1	2	4	9	2
449	48	1	1	4	4	11	2
493	53	1	1	1	4	7	2
508	29	1	2	1	4	11	2
510	53	1	2	1	2	10	2
541	75	1	5	2	4	11	2
572	71	1	4	2	4	8	0
629	81	1	3	4	2	11	1
748	69	1	1	4	2	11	0
818	66	1	2	1	4	11	3
838	75	1	3	2	2	11	2
930	64	1	3	2	1	3	0
938	59	1	4	2	4	11	0
956	49	1	2	4	2	11	0
992	49	1	2	2	4	11	3
1018	42	1	1	1	4	6	2
1045	37	1	2	4	4	11	0
1046	50	1	1	4	4	6	0
1182	71	1	2	2	3	9	0
1223	71	1	1	2	4	9	0
1243	41	1	4	5	1	11	0
1469	70	1	1	2	5	11	2

Table 1.2 C

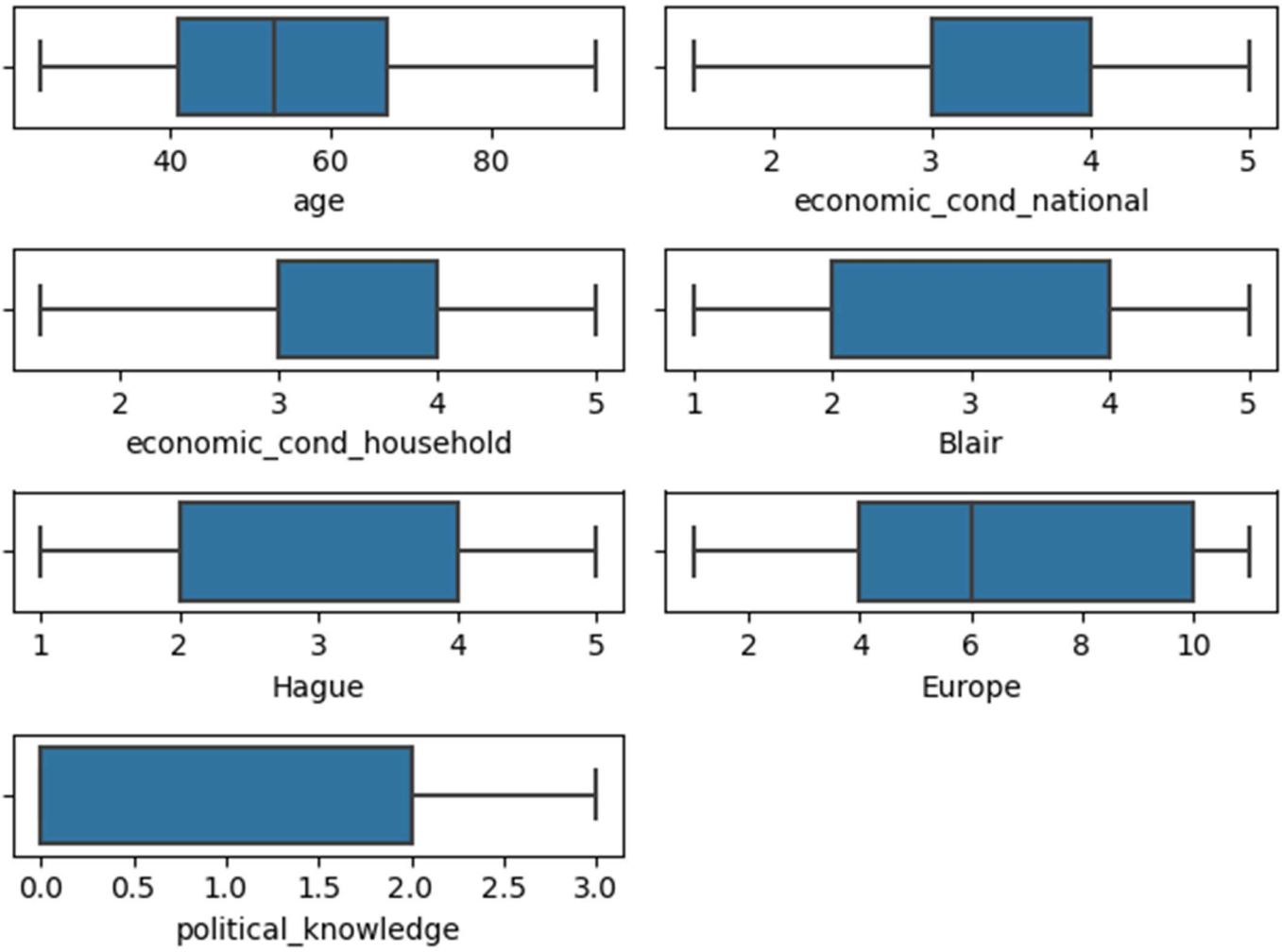
Inferences from Table 1.2B and 1.2C :-

- Total datasets which are having Economic Condition National less than 1.5 or 1 = 65 rows
- Total datasets which are having Economic Condition Household less than 1.5 or 1 = 39 rows

Values in both these columns are Categorical/Ordinal in nature i.e., it has discrete values in the scale of 1 to 5. So even if we treat the outlier's values to the lower limit it won't change our dataset i.e., treating the outliers means we will replace all the 1's in both these columns with 1.5 which won't affect our dataset.

So Here we have treated the Outliers which the lower limit i.e., replace all the 1 in both the columns with 1.5.

Boxplot after treating the Outliers: -



1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not? (2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed.

Solution: -

Encode the data for modelling: - There are two string type variables present- gender and vote. For the purpose of transforming the categorical data we have used pd.Categorical().codes and pd.get_dummies().

1. Gender: - For Gender, we have used pd.get_dummies(). In pd.get_dummies() it create dummy variables for each categories in the variables and by defining "drop_first = True". It will drop the first dummy variables and adds rest of the variables instead of original variable.
2. Vote: - For Vote, we have used "pd.Categorical().codes". It assigns codes to all the categories. Since here we have only 2 categories "Labour" And "Conservative". So it assigns 1 to Labour and 0 to Conservative.

Scaling of the data: - Scaling od a feature is done in order to scale down the magnitude of the feature. Some of the models/algorithm are sensitive to the magnitude. Higher the magnitude, higher would be the impact of that feature while preparing the model which is not good for algorithm which are based on Magnitude/Distance.

Most commonly used scaling technique: -

1. Z-score or Standardization: - It scale down all the values in the variables from -1 to +1 such that the variable mean =0 and standard deviation = 1 .

Formula for Zscore: -

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

2. MinMax or Standardization: - It scales down all the values in the variable from 0 to 1.

Formula for MinMax Scaling : -

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

It is recommended to do feature scaling when we are dealing with distance based model/algorithms (KNN, Regression) since they are sensitive to the range of data points. It is very useful in checking and reducing the multi-collinearity in data.

But the tree based methods would not require scaling in general because it uses split method.

Thus before KNN, we did the appropriate scaling

Since most of the variables are in range of 0-11 except “age”. we will scale only the “age” variable using MinMax method. We have used MinMax scaling as the distribution is not normal and doesn’t have outliers.

Encoded and Unscaled Dataset: -

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	vote	gender_male
0	43.0	3.0	3.0	4.0	1.0	2.0	2.0	1	0
1	36.0	4.0	4.0	4.0	4.0	5.0	2.0	1	1
2	35.0	4.0	4.0	5.0	2.0	3.0	2.0	1	1
3	24.0	4.0	2.0	2.0	1.0	4.0	0.0	1	0
4	41.0	2.0	2.0	1.0	1.0	6.0	2.0	1	1
...	...	...	...	...	...	...	...	...	...
1520	67.0	5.0	3.0	2.0	4.0	11.0	3.0	0	1
1521	73.0	2.0	2.0	4.0	4.0	8.0	2.0	0	1
1522	37.0	3.0	3.0	5.0	4.0	2.0	2.0	1	1
1523	61.0	3.0	3.0	1.0	4.0	11.0	2.0	0	1
1524	74.0	2.0	3.0	2.0	4.0	11.0	0.0	0	0

1517 rows × 9 columns

Table 1.3A

Encoded and Scaled Dataset: -

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	vote	gender_male
0	0.275362	3.0	3.0	4.0	1.0	2.0	2.0	1	0
1	0.173913	4.0	4.0	4.0	4.0	5.0	2.0	1	1
2	0.159420	4.0	4.0	5.0	2.0	3.0	2.0	1	1
3	0.000000	4.0	2.0	2.0	1.0	4.0	0.0	1	0
4	0.246377	2.0	2.0	1.0	1.0	6.0	2.0	1	1
...	...	...	...	...	...	...	...	...	...
1520	0.623188	5.0	3.0	2.0	4.0	11.0	3.0	0	1
1521	0.710145	2.0	2.0	4.0	4.0	8.0	2.0	0	1
1522	0.188406	3.0	3.0	5.0	4.0	2.0	2.0	1	1
1523	0.536232	3.0	3.0	1.0	4.0	11.0	2.0	0	1
1524	0.724638	2.0	3.0	2.0	4.0	11.0	0.0	0	0

Table 1.3B

Dataset split: - Splitting the dataset into Train and test

We need to split the data set into train and test so that we can compare the model performance between train and test model. Here we have split the dataset into 70:30 ratios i.e., 70 % of data into train and 30% in test.

Train Dataset: - X = (1061, 8) Test Dataset: - X = (456, 8)

Y = (1061, 1) Y = (456, 1)

1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both models (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting).

Solution: -

1. **LOGISTICS REGRESSION: -**

This is a regression analysis that should be performed when the dependent variable is binary in nature. Like all regression analysis, logistic regression analysis is predictive analysis. Logistic regression is used to describe data and to describe the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or independent relationship level variables.

Since Logistics Regression doesn't require scaling so we will not use scaled data for logistic regression: -

We have used GridSearchCV () in order to find the best/optimum parameters for creating a model: -

Best parameters for our model =

```
LogisticRegression
LogisticRegression(max_iter=50, penalty='l1', random_state=1,
                   solver='liblinear', tol=0.01)
```

Importance / Coefficient of all the variables after creating Logistic Regression Model: -

	Import/Coeff
age	-0.011221
economic_cond_national	0.698725
economic_cond_household	0.140960
Blair	0.621018
Hague	-0.790731
Europe	-0.199595
political_knowledge	-0.287398
gender_male	0.189402

Fig. 1.4A

Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Train Data

	precision	recall	f1-score	support
0	0.75	0.64	0.69	307
1	0.86	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

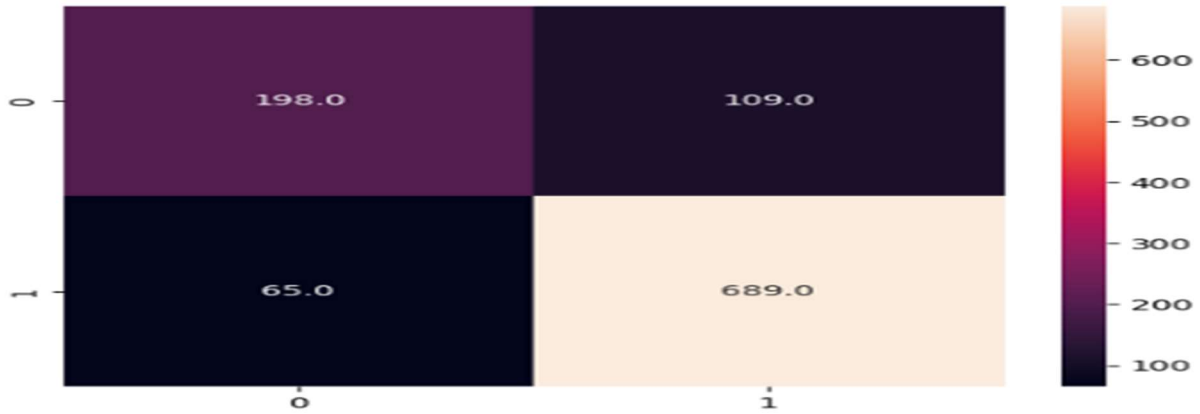


Fig 1.4B

Inferences from fig 1.4B: -

- Recall of 1's(Labour) = 0.91 Recall of 0's(Conservative) = 0.64
- Precision of 1's (Labour) = 0.86 Precision of 0's(Conservative) = 0.75
- Accuracy = 0.84

Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Test Data

	precision	recall	f1-score	support
0	0.75	0.72	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.83	0.83	0.83	456

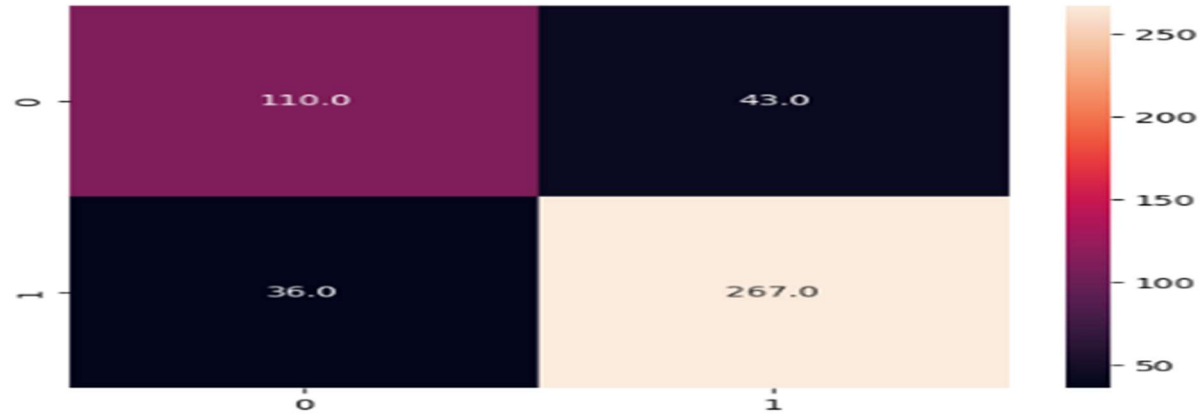


Fig 1.4B

Inferences from fig 1.4C: -

- Recall of 1's(Labour) = 0.88 Recall of 0's(Conservative) = 0.72
- Precision of 1's (Labour) = 0.86 Precision of 0's(Conservative) = 0.75
- Accuracy = 0.83

From above Accuracy of the Train and Test model i.e., 0.84 and 0.83 respectively, our model equally better in both train and test. Hence it neither Over fitted nor Under fitted.

2. Linear Discriminant Analysis: -

LDA as a name suggest is linear model for classification and dimension reductionality. Most commonly used for feature extraction in pattern recognition problem.

	Importance
age	0.356167
economic_cond_national	-0.306658
economic_cond_household	-0.118894
Blair	-0.853074
Hague	1.026248
Europe	0.238921
political_knowledge	0.555771
gender_male	0.074105

Fig 1.4C

Linear Discriminant Analysis (LDA) Equation: -

LDA Equation: Vote = [2.488] +(-0.02)* age +(0.633)* economic_cond_national +(0.067)* economic_cond_household +(0.741)* Blair +(-0.927)* Hague +(-0.224)* Europe +(-0.429)* political_knowledge +(0.148)* gender_male

Fig 1.4D

Linear Discriminant Analysis(LDA) Classification Report and Confusion matrix on Train Data

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

Fig. 1.4E

Inferences from fig 1.4E: -

- Recall of 1's (Labour) = 0.91 Recall of 0's(Conservative) = 0.65 for train data
- Precision of 1's (Labour) = 0.86 Precision of 0's(Conservative) = 0.74 for train data
- Accuracy of train data = 0.83

Linear Discriminant Analysis(LDA) Classification Report and Confusion matrix on 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

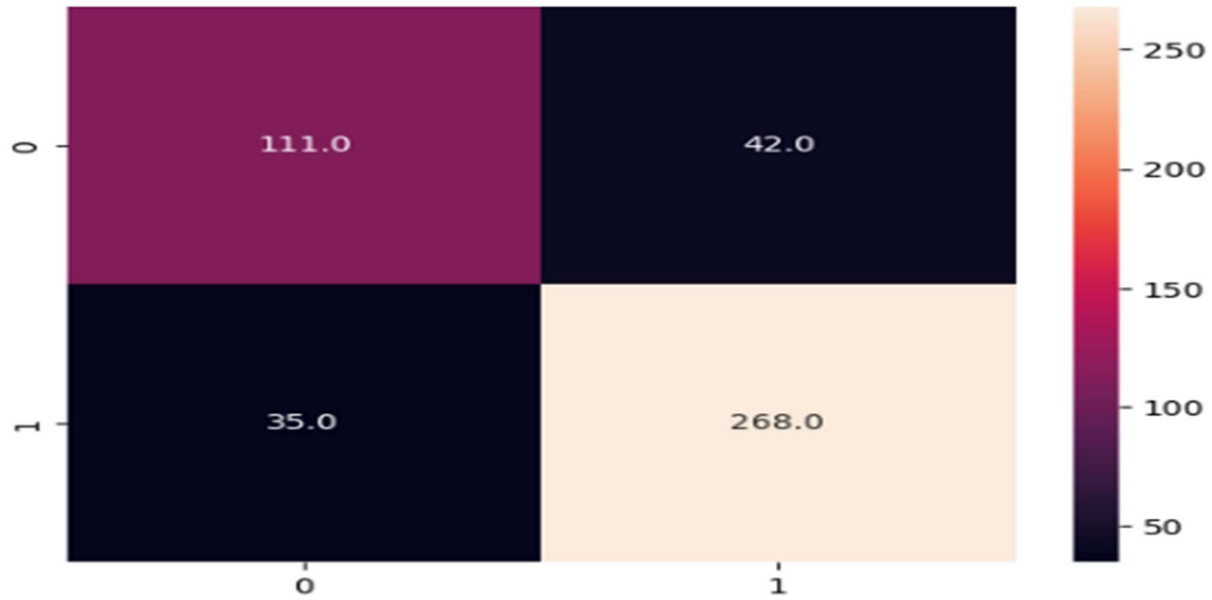


Fig 1.4F

Inferences from fig 1.4F: -

- Recall of 1's (Labour) = 0.88 Recall of 0's(Conservative) = 0.73 for test data
- Precision of 1's (Labour) = 0.86 Precision of 0's(Conservative) = 0.76 for test data
- Accuracy of test data = 0.83

From above Accuracy of the Train and Test model i.e., 0.83 and 0.83 respectively, our model equally better in both train and test. Hence it neither Over fitted nor Under fitted.

1.5) Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Solution: -

1. **KNN Model: -**

- A k-nearest-neighbor algorithm, often abbreviated KNN, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.
- The k-nearest-neighbor is an example of a "lazy learner" algorithm, meaning that it does not build a model using the training set until a query of the data set is performed
- Since it is a distance based algorithm, hence for KNN model we will use scaled data.

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	vote	gender_male
0	0.275362	3.0	3.0	4.0	1.0	2.0	2.0	1	0
1	0.173913	4.0	4.0	4.0	4.0	5.0	2.0	1	1
2	0.159420	4.0	4.0	5.0	2.0	3.0	2.0	1	1
3	0.000000	4.0	2.0	2.0	1.0	4.0	0.0	1	0
4	0.246377	2.0	2.0	1.0	1.0	6.0	2.0	1	1
...	...	...	...	...	...	...	...	...	...
1520	0.623188	5.0	3.0	2.0	4.0	11.0	3.0	0	1
1521	0.710145	2.0	2.0	4.0	4.0	8.0	2.0	0	1
1522	0.188406	3.0	3.0	5.0	4.0	2.0	2.0	1	1
1523	0.536232	3.0	3.0	1.0	4.0	11.0	2.0	0	1
1524	0.724638	2.0	3.0	2.0	4.0	11.0	0.0	0	0

1517 rows × 9 columns

Table 1.5A

KNN Model Classification Report and Confusion matrix on Train Data

	precision	recall	f1-score	support
0	0.76	0.71	0.73	307
1	0.89	0.91	0.90	754
accuracy			0.85	1061
macro avg	0.82	0.81	0.82	1061
weighted avg	0.85	0.85	0.85	1061

Fig. 1.5B

Inferences from fig 1.5B: -

- Recall of 1's (Labour) = 0.91 Recall of 0's(Conservative) = 0.71 for train data
- Precision of 1's (Labour) = 0.89 Precision of 0's(Conservative) = 0.76 for train data
- Accuracy of train data = 0.85

KNN Model Classification Report and Confusion matrix on Test Data

	precision	recall	f1-score	support
0	0.72	0.70	0.71	153
1	0.85	0.86	0.86	303
accuracy			0.81	456
macro avg	0.79	0.78	0.78	456
weighted avg	0.81	0.81	0.81	456

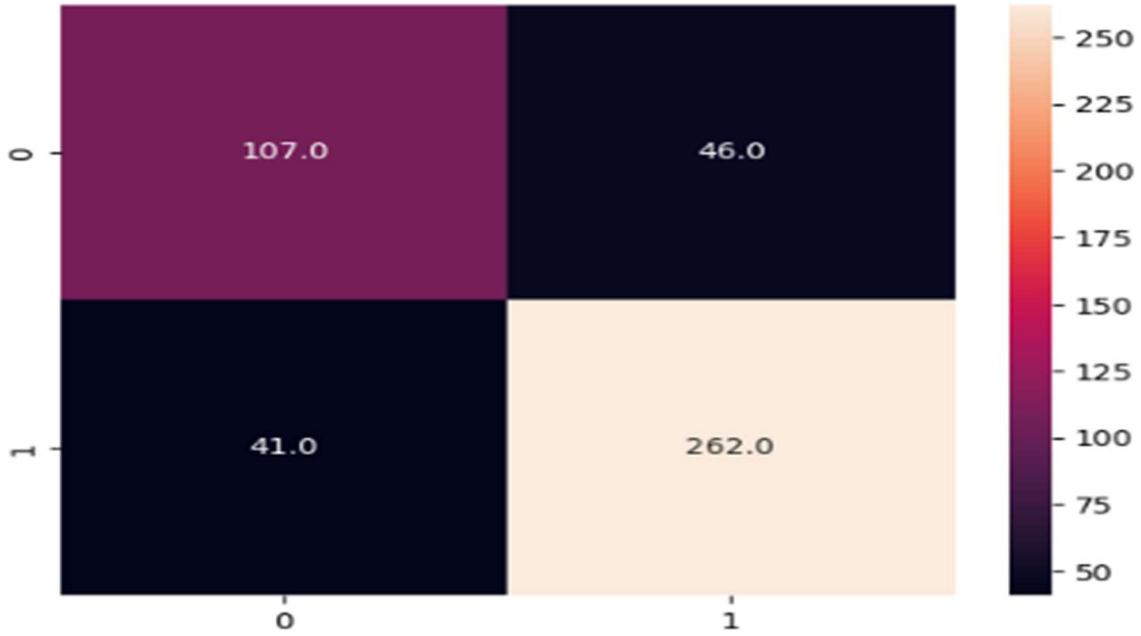


Fig 1.5C

Inferences from fig 1.5C: -

- Recall of 1's (Labour) = 0.86 Recall of 0's(Conservative) = 0.70 for test data
- Precision of 1's (Labour) = 0.85 Precision of 0's(Conservative) = 0.72 for test data
- Accuracy of test data = 0.81

From above Accuracy of the Train and Test model i.e., 0.85 and 0.81 respectively, our model performed better in both train data but didn't performed well with test data. Hence it neither Over fitted model.

2. **Naïve's Bayes Algorithm: -**

- A naive Bayes classifier is a calculation that utilizes Bayes' hypothesis to arrange objects.
- Naïve Bayes classifiers expect to be strong, or naive, autonomy between attributes of data points.
- Formula for Baye's Theorem is given as: -

Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- A, B =events
- P (B|A) =Probability of A Given B is true
- P (A)= Probability of B given A is true
- P(B) = the independent probabilities of A and B

Naïve's Bayes Classification Report and Confusion matrix on Train Data

	precision	recall	f1-score	support
0	0.72	0.69	0.71	307
1	0.88	0.89	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.83	0.83	1061

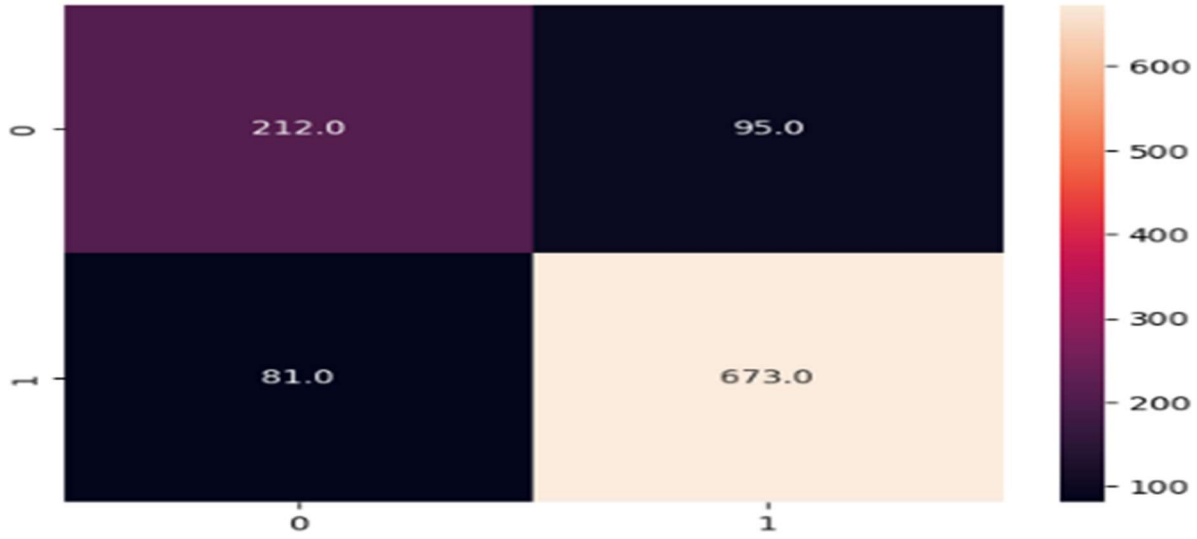


Fig. 1.5D

Inferences from fig 1.5B: -

- Recall of 1's (Labour) = 0.89 Recall of 0's(Conservative) = 0.69 for train data
- Precision of 1's (Labour) = 0.88 Precision of 0's(Conservative) = 0.72 for train data
- Accuracy of train data = 0.83

Naïve's Bayes Classification Report and Confusion matrix on Test Data

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

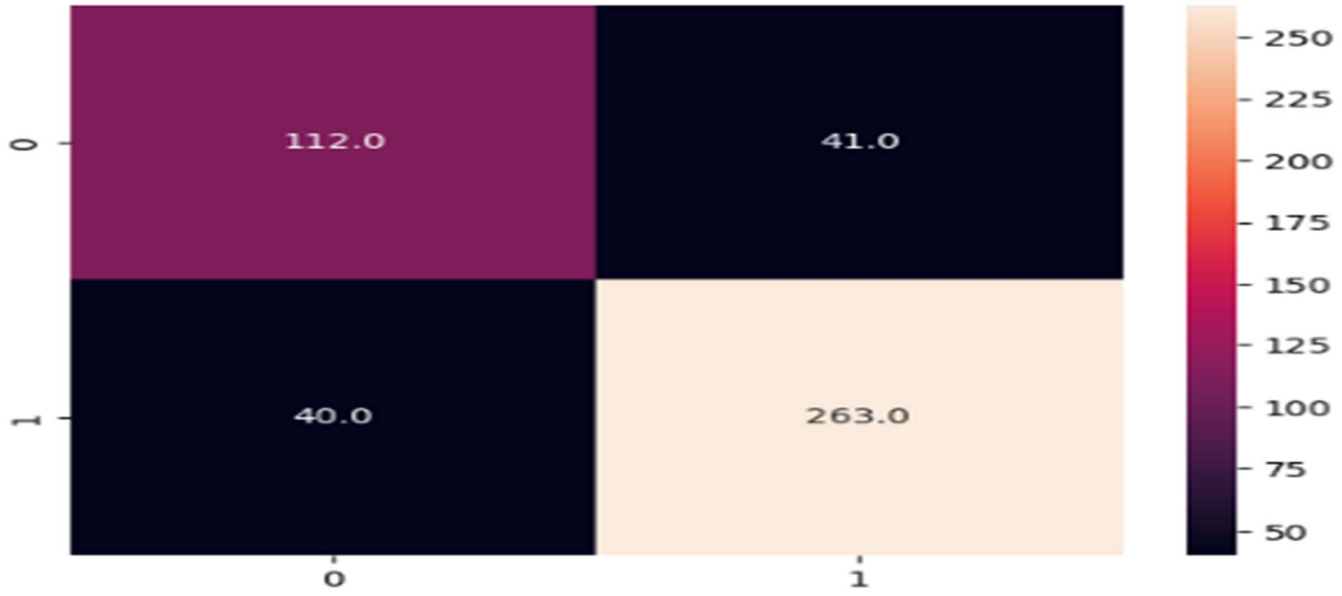


Fig 1.5E

Inferences from fig 1.5E: -

- Recall of 1's (Labour) = 0.87 Recall of 0's(Conservative) = 0.73 for test data
- Precision of 1's (Labour) = 0.87 Precision of 0's(Conservative) = 0.74 for test data
- Accuracy of test data = 0.82

From above Accuracy of the Train and Test model i.e., 0.83 and 0.82 respectively, our model equally better in both train and test. Hence it neither Over fitted nor Under fitted.

1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. 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

Solution: -

Bagging: -

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the first dataset and afterward total their singular forecasts (either by casting a ballot or by averaging) to shape a final prediction. Such a meta-estimator can regularly be utilized as a method for diminishing the variance of a black-box estimator (e.g., a decision tree), by bringing randomization into its development technique and afterward making a group out of it

The Model is built with parameters with base_estimator as Random Forest classifier and n_estimators of 100.

```
BaggingClassifier
BaggingClassifier(estimator=RandomForestClassifier(random_state=1),
                  random_state=1)
  estimator: RandomForestClassifier
  RandomForestClassifier(random_state=1)
    RandomForestClassifier
    RandomForestClassifier(random_state=1)
```

Bagging Model Classification Report and Confusion matrix on Train Data

	precision	recall	f1-score	support
0	0.97	0.89	0.93	307
1	0.96	0.99	0.97	754
accuracy			0.96	1061
macro avg	0.96	0.94	0.95	1061
weighted avg	0.96	0.96	0.96	1061

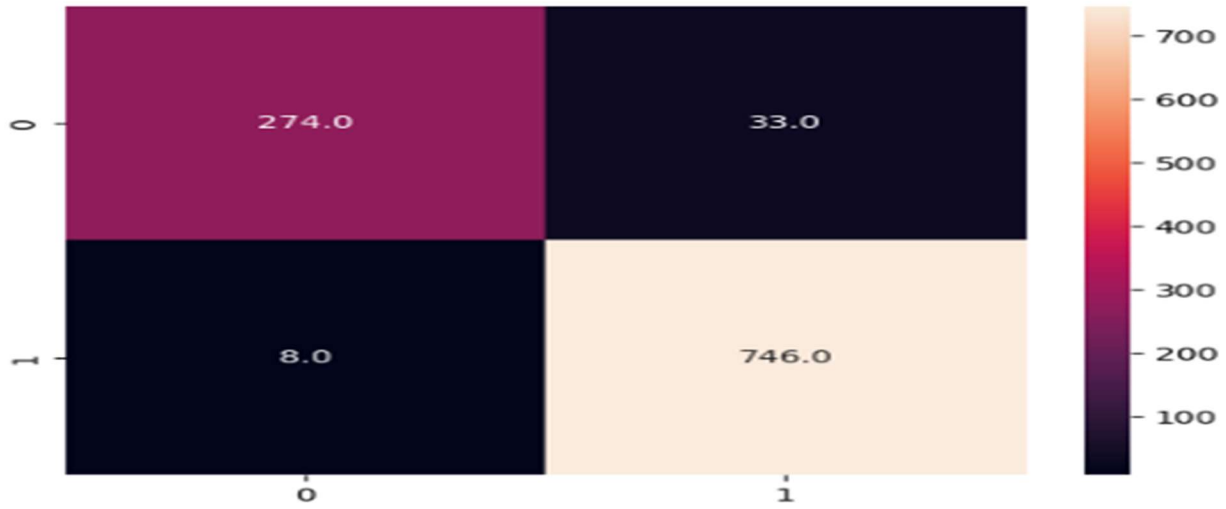


Fig. 1.6A

Inferences from fig 1.6A: -

- Recall of 1's (Labour) = 0.99 Recall of 0's(Conservative) = 0.89 for train data
- Precision of 1's (Labour) = 0.96 Precision of 0's(Conservative) = 0.97 for train data
- Accuracy of train data = 0.96

Bagging Model Classification Report and Confusion matrix on Test Data

	precision	recall	f1-score	support
0	0.80	0.67	0.73	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	0.83	456

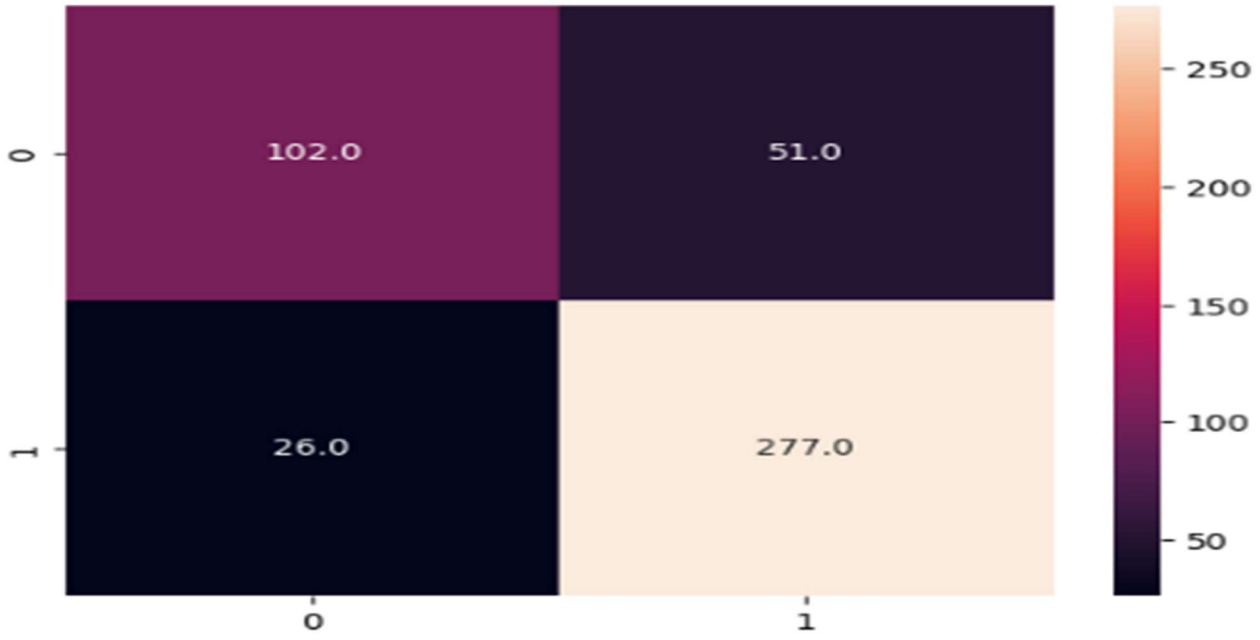


Fig. 1.6B

Inferences from fig 1.6B: -

- Recall of 1's (Labour) = 0.91 Recall of 0's(Conservative) = 0.67 for test data
- Precision of 1's (Labour) = 0.84 Precision of 0's(Conservative) = 0.80 for test data
- Accuracy of test data = 0.83

From above Accuracy of the Train and Test model i.e., 0.96 and 0.83 respectively. The model performed better in both train data but didn't performed well with test data. Hence it Over fitted model

Boosting: -

There are two type of boosting Technique: -

1. **Adaptive Boosting** **2. Gradient Boosting**

Adaptive Boosting(AdaBoosting)

- ADA Boost algorithm, short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble Method in Machine Learning.
- It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights to incorrectly classified instances.
- Boosting is used to reduce bias as well as the variance for supervised learning. It works on the principle where learners are grown sequentially. Except for the first, each subsequent learner is grown from previously grown learners

First we create a basic model and later using GridSearchCv() we find the best parameters for our model: -

Best Parameters

```
AdaBoostClassifier
AdaBoostClassifier(algorithm='SAMME', learning_rate=1, n_estimators=30,
                   random_state=1)
```

Adaptive Boosting with GridSearchCV() Classification Report and Confusion matrix on Train Data

	precision	recall	f1-score	support
0	0.76	0.67	0.72	307
1	0.87	0.91	0.89	754
accuracy			0.84	1061
macro avg	0.82	0.79	0.80	1061
weighted avg	0.84	0.84	0.84	1061

Fig. 1.6C

Inferences from fig 1.6C: -

- Recall of 1's (Labour) = 0.91 Recall of 0's(Conservative) = 0.67 for train data
- Precision of 1's (Labour) = 0.87 Precision of 0's(Conservative) = 0.76 for train data
- Accuracy of train data = 0.84

Adaptive Boosting with GridSearchCV() Classification Report and Confusion matrix on Test Data

	precision	recall	f1-score	support
0	0.74	0.71	0.73	153
1	0.86	0.87	0.87	303
accuracy			0.82	456
macro avg	0.80	0.79	0.80	456
weighted avg	0.82	0.82	0.82	456

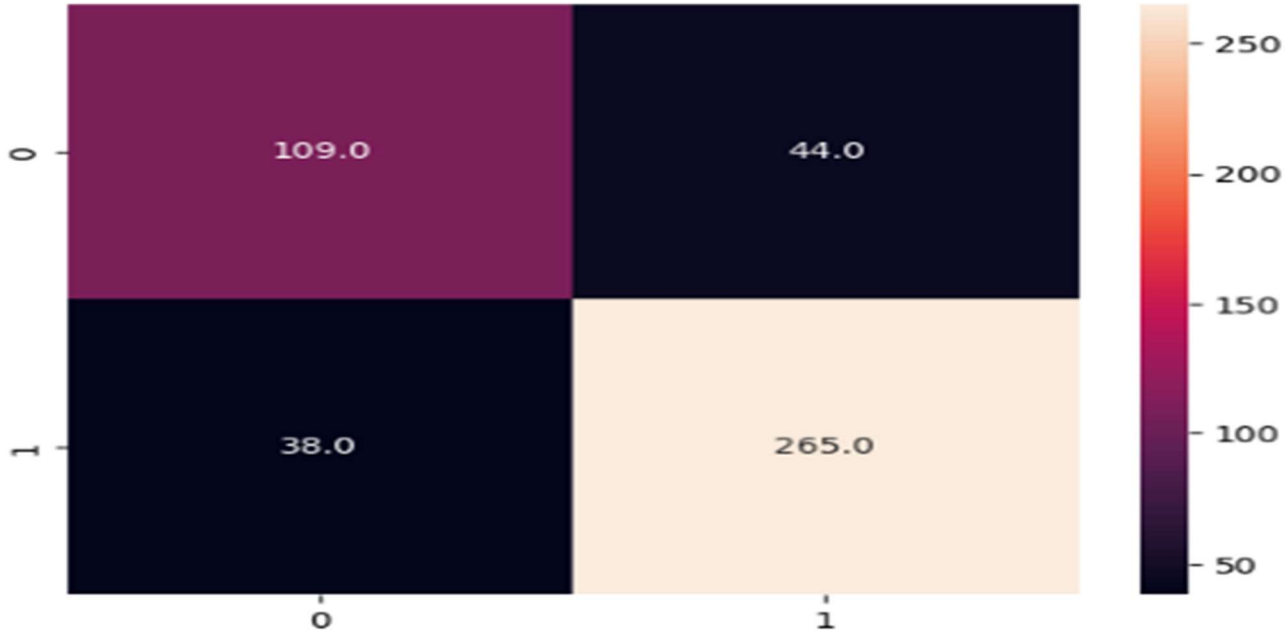


Fig 1.6D

Inferences from fig 1.6D: -

- Recall of 1's (Labour) = 0.87 Recall of 0's(Conservative) = 0.71 for test data
- Precision of 1's (Labour) = 0.86 Precision of 0's(Conservative) = 0.74 for test data
- Accuracy of test data = 0.82

From above Accuracy of the Train and Test model i.e., 0.84 and 0.82 respectively, our model equally better in both train and test. Hence it neither Over fitted nor Under fitted.

Gradient Boosting: -

- Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.
- Decision trees are usually used when doing gradient boosting.

First we create a basic model and later using GridSearchCv() we find the best parameters for our model: -

Best Parameters: -

```
GradientBoostingClassifier
GradientBoostingClassifier(ccp_alpha=0, min_samples_leaf=5, n_estimators=50,
                           random_state=1)
```

Gradient Boosting with GridSearchCV() Classification Report and Confusion matrix on Train Data

	precision	recall	f1-score	support
0	0.82	0.72	0.77	307
1	0.89	0.94	0.91	754
accuracy			0.87	1061
macro avg	0.86	0.83	0.84	1061
weighted avg	0.87	0.87	0.87	1061

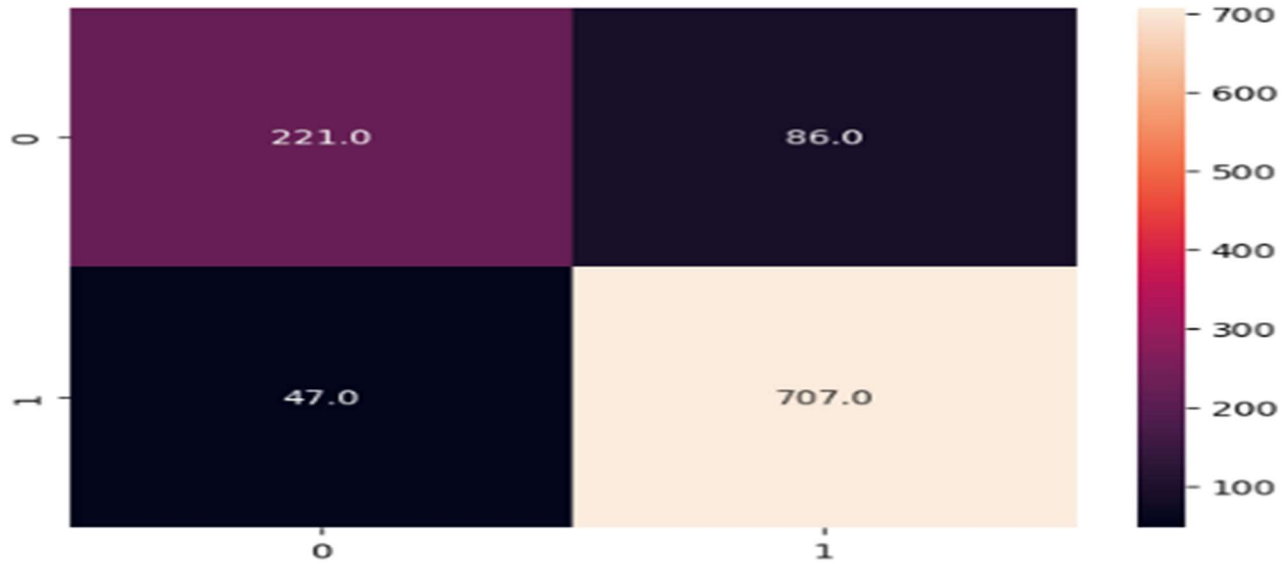


Fig. 1.6E

Inferences from fig 1.6E: -

- Recall of 1's (Labour) = 0.94 Recall of 0's(Conservative) = 0.72 for train data
- Precision of 1's (Labour) = 0.89 Precision of 0's(Conservative) = 0.82 for train data
- Accuracy of train data = 0.87

Gradient Boosting with GridSearchCV() Classification Report and Confusion matrix on Test Data

	precision	recall	f1-score	support
0	0.80	0.67	0.73	153
1	0.84	0.91	0.88	303
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
eighted avg	0.83	0.83	0.83	456

Fig. 1.6F

Inferences from fig 1.6F: -

- Recall of 1's (Labour) = 0.91 Recall of 0's(Conservative) = 0.67 for train data
- Precision of 1's (Labour) = 0.84 Precision of 0's(Conservative) = 0.80 for train data
- Accuracy of train data = 0.83

From above Accuracy of the Train and Test model i.e., 0.87 and 0.83 respectively. The model performed better in both train data but didn't performed well with test data. Hence it slightly Over fitted model

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model, classification report (4 pts) 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. (3 pts)

Solution: -

Logistics Regression Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

TRAIN DATA

	precision	recall	f1-score	support
0	0.75	0.64	0.69	307
1	0.86	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

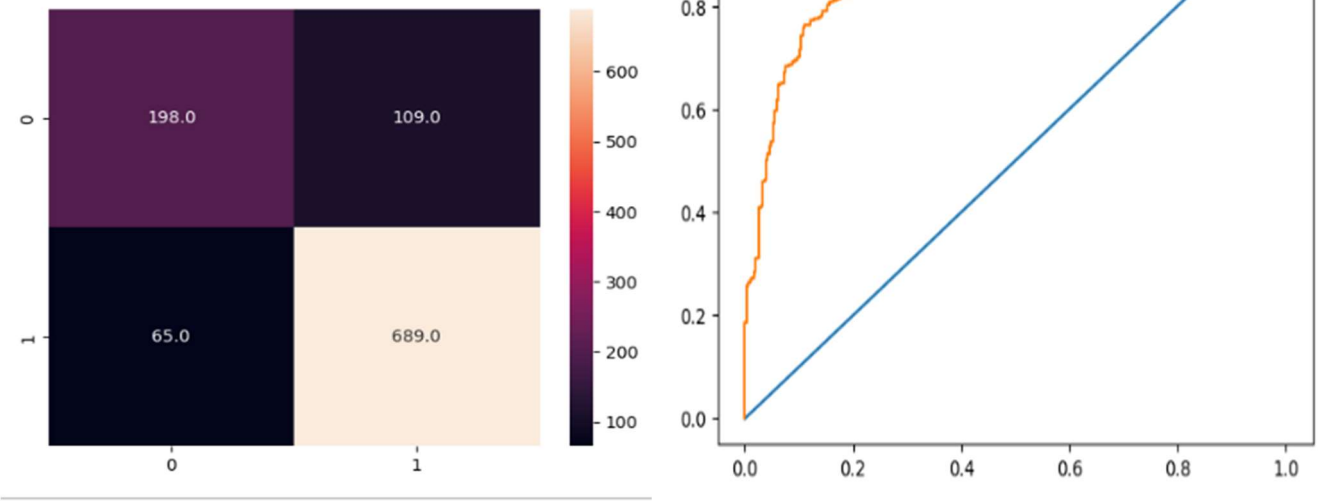


Fig 1.7A
Log Reg Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Train Data

Inferences from fig 1.7A: -

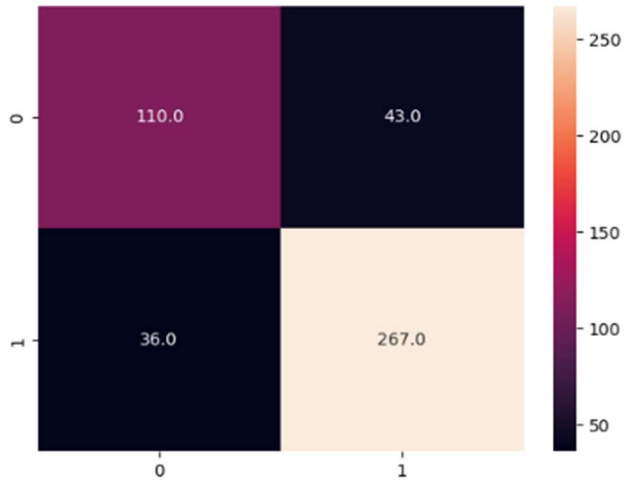
- Recall of 1's (Labour) = 0.91
- Precision of 1's (Labour) = 0.86
- Accuracy of train data = 0.84
- ROC AUC SCORE of train data= 0.89

Recall of 0's(Conservative) = 0.64 for train data

Precision of 0's(Conservative) = 0.75 for train data

Test Data

	precision	recall	f1-score	support
0	0.75	0.72	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.83	0.83	0.83	456



ROC AUC SCORE is 0.8791388942815851

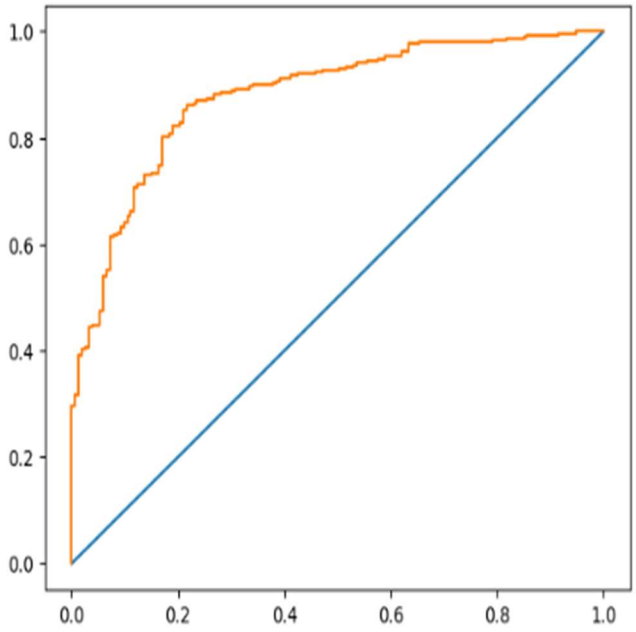


Fig 1.7B
Log Reg Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Test Data

Inferences from fig 1.4C: -

- Recall of 1's(Labour) = 0.88 Recall of 0's(Conservative) = 0.72
- Precision of 1's (Labour) = 0.86 Precision of 0's(Conservative) = 0.75
- Accuracy = 0.83
- ROC AUC SCORE =0.88

Linear Discriminant Analysis Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

TRAIN DATA

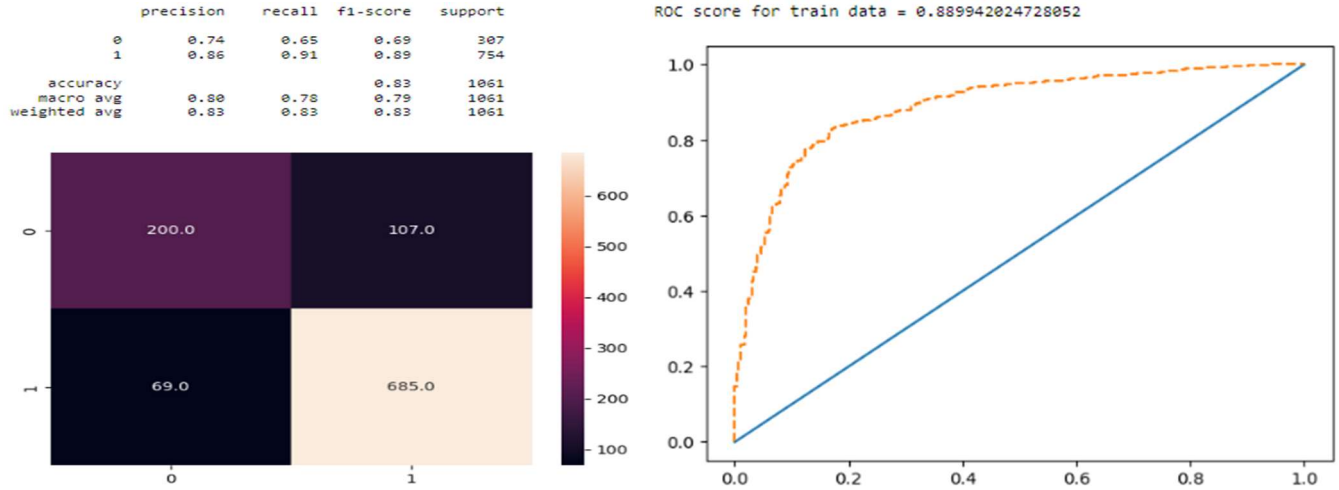


Fig 1.7C

LDA Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Train Data

Inferences from fig 1.7C: -

- Recall of 1's (Labour) = 0.91
- Precision of 1's (Labour) = 0.86
- Accuracy of train data = 0.83
- ROC AUC SCORE of train data= 0.8899

Recall of 0's(Conservative) = 0.65 for train data

Precision of 0's(Conservative) = 0.74 for train data

TEST DATA

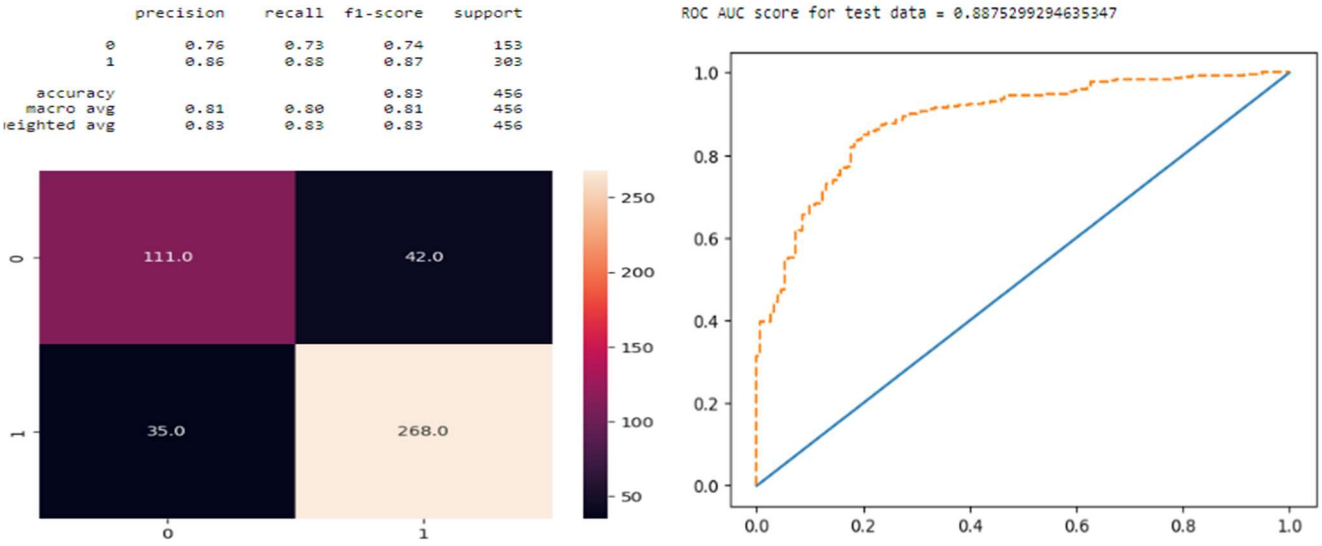


Fig 1.7D

LDA Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Test Data

Inferences from fig 1.7D: -

- Recall of 1's (Labour) = 0.88
- Precision of 1's (Labour) = 0.86
- Accuracy of test data = 0.83
- ROC AUC SCORE of test data= 0.8875

Recall of 0's(Conservative) = 0.73 for test data

Precision of 0's(Conservative) = 0.76 for test data

NAÏVE Bayes Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

TRAIN DATA

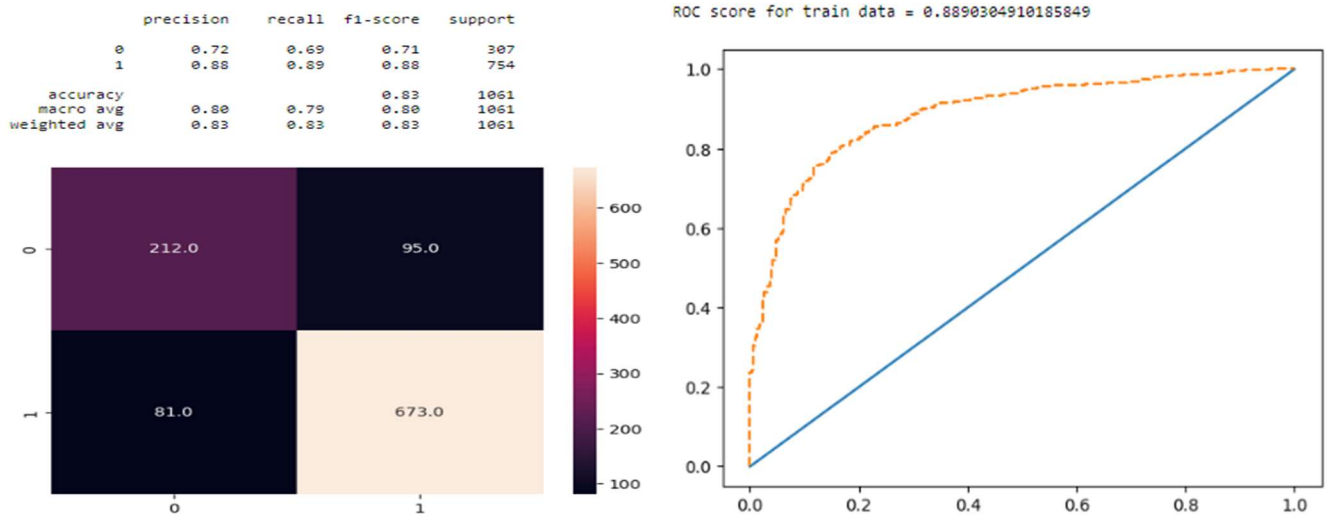


Fig 1.7E

Naïve Bayes Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Train Data

Inferences from fig 1.7E: -

- Recall of 1's (Labour) = 0.89
- Precision of 1's (Labour) = 0.88
- Accuracy of train data = 0.83
- ROC AUC SCORE of train data= 0.8890

Recall of 0's(Conservative) = 0.69 for train data

Precision of 0's(Conservative) = 0.72 for train data

TEST DATA

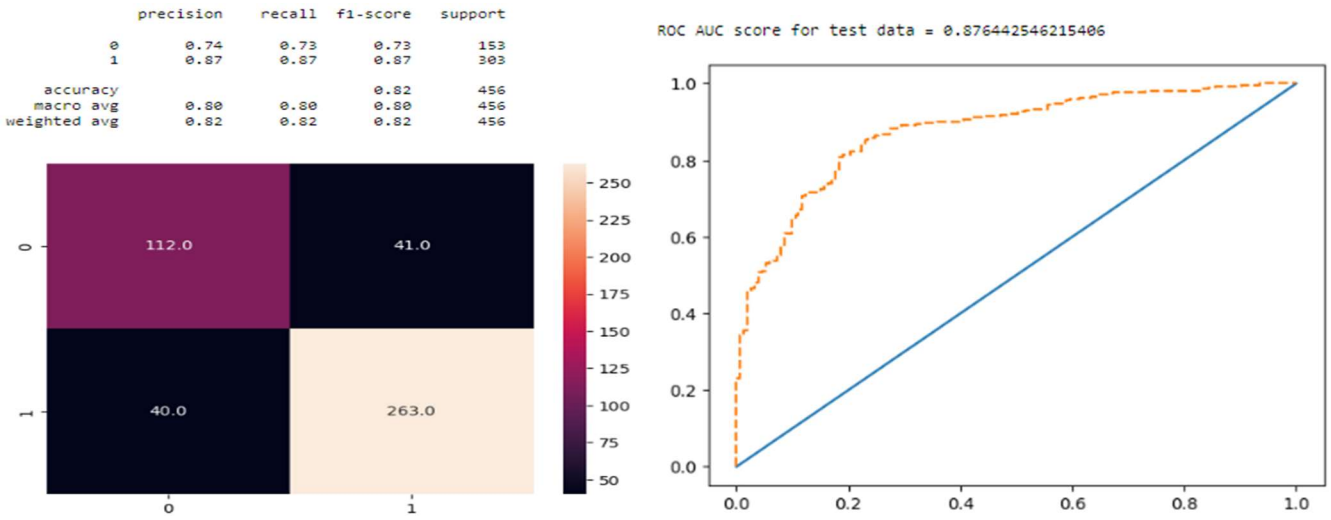


Fig 1.7F

Naïve Bayes Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Test Data

Inferences from fig 1.7D: -

- Recall of 1's (Labour) = 0.87
- Precision of 1's (Labour) = 0.87
- Accuracy of test data = 0.82
- ROC AUC SCORE of test data= 0.8875

Recall of 0's(Conservative) = 0.73 for test data

Precision of 0's(Conservative) = 0.74 for test data

KNN Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

TRAIN DATA

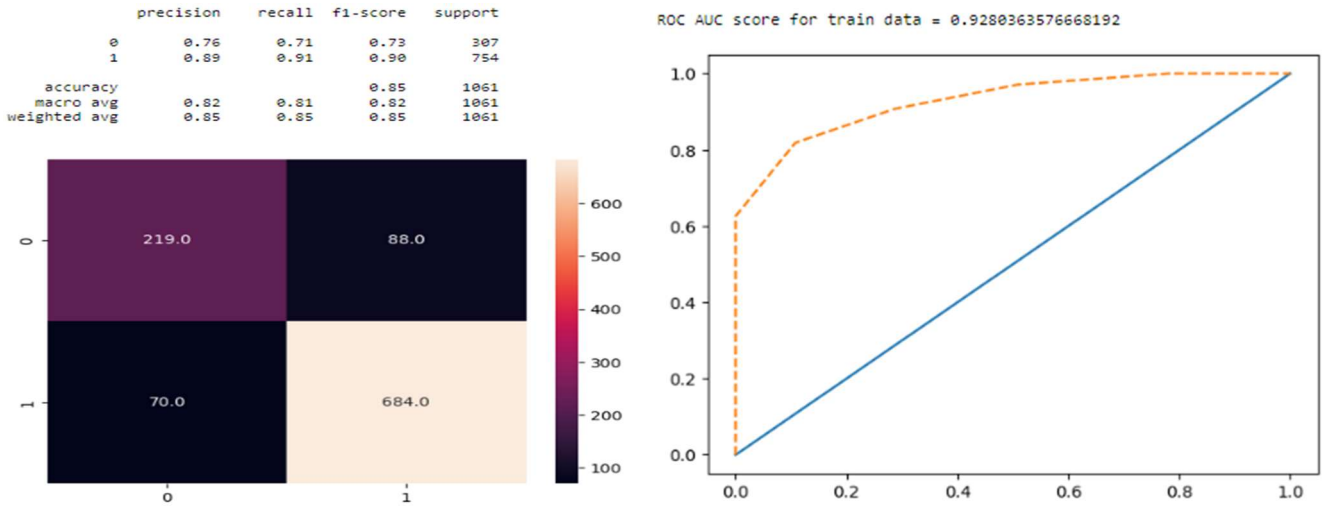


Fig 1.7G

KNN Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Train Data

Inferences from fig 1.7G: -

- Recall of 1's (Labour) = 0.91 — Recall of 0's(Conservative) = 0.71 for train data
- Precision of 1's (Labour) = 0.89 — Precision of 0's(Conservative) = 0.76 for train data
- Accuracy of train data = 0.85
- ROC AUC SCORE of train data= 0.928

TEST DATA

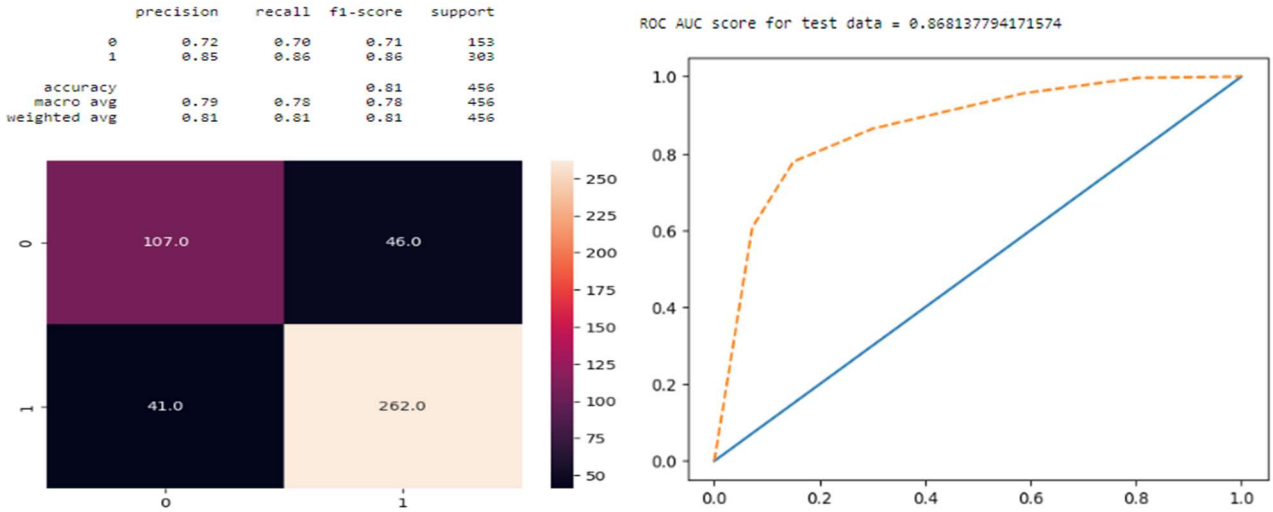


Fig 1.7H

KNN Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Test Data

Inferences from fig 1.7H: -

- Recall of 1's (Labour) = 0.86 — Recall of 0's(Conservative) = 0.70 for test data
- Precision of 1's (Labour) = 0.85 — Precision of 0's(Conservative) = 0.72 for test data
- Accuracy of test data = 0.81
- ROC AUC SCORE of test data= 0.868

Bagging Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

TRAIN DATA

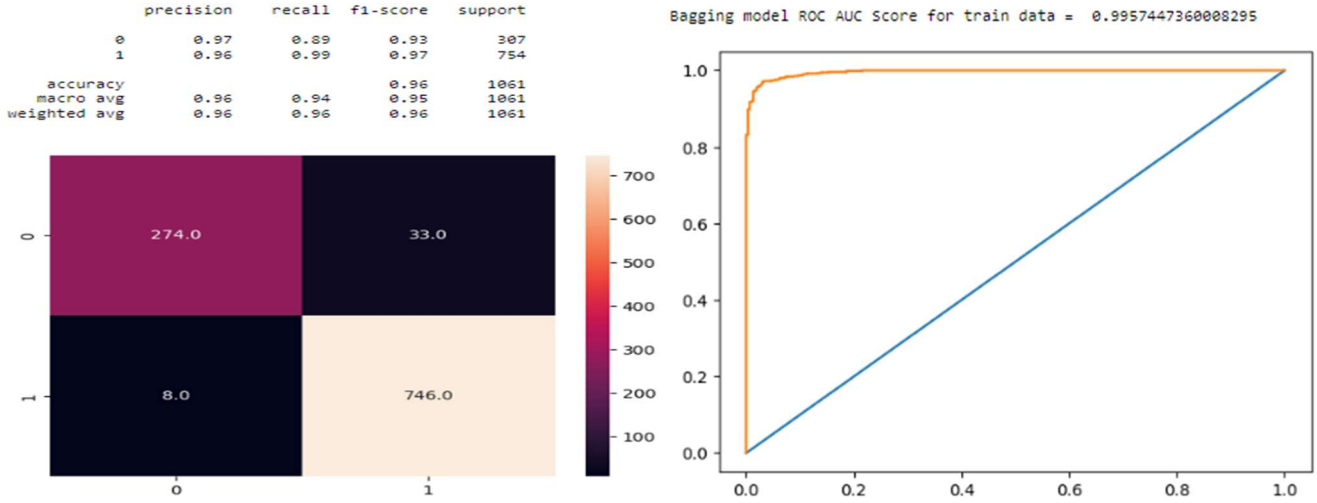


Fig 1.7I

Bagging Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Train Data

Inferences from fig 1.7I: -

- Recall of 1's (Labour) = 0.99 — Recall of 0's(Conservative) = 0.89 for train data
- Precision of 1's (Labour) = 0.96 — Precision of 0's(Conservative) = 0.97 for train data
- Accuracy of train data = 0.96
- ROC AUC SCORE of train data= 0.996

TEST DATA

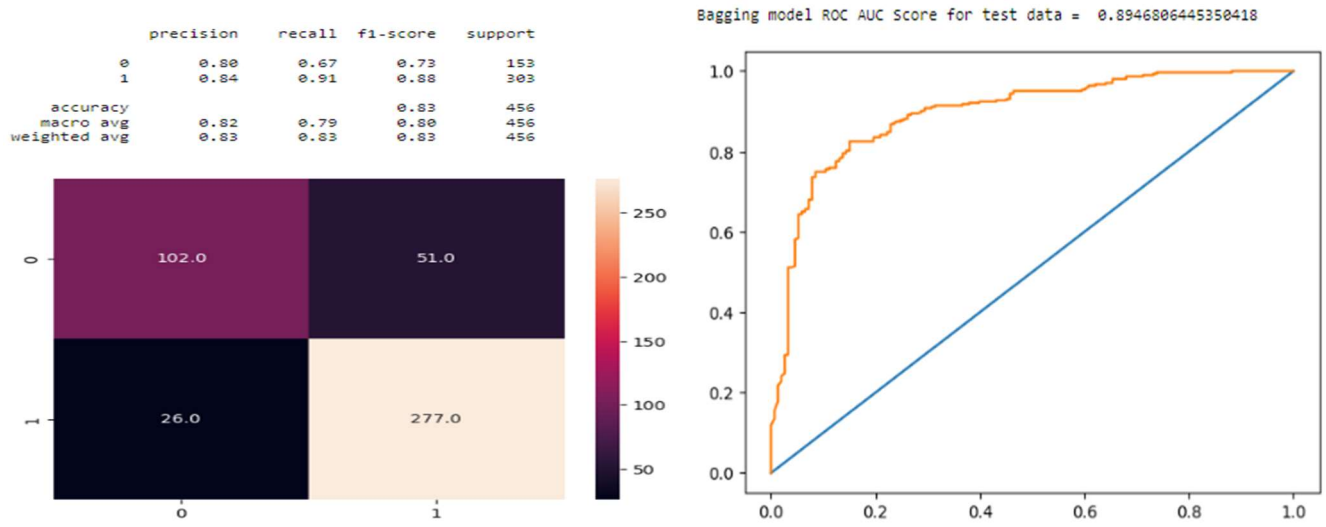


Fig 1.7J

Bagging Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Test Data

Inferences from fig 1.7H: -

- Recall of 1's (Labour) = 0.91 — Recall of 0's(Conservative) = 0.67 for test data
- Precision of 1's (Labour) = 0.84 — Precision of 0's(Conservative) = 0.80 for test data
- Accuracy of test data = 0.83
- ROC AUC SCORE of test data= 0.868

ADABoosting Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

TRAIN DATA

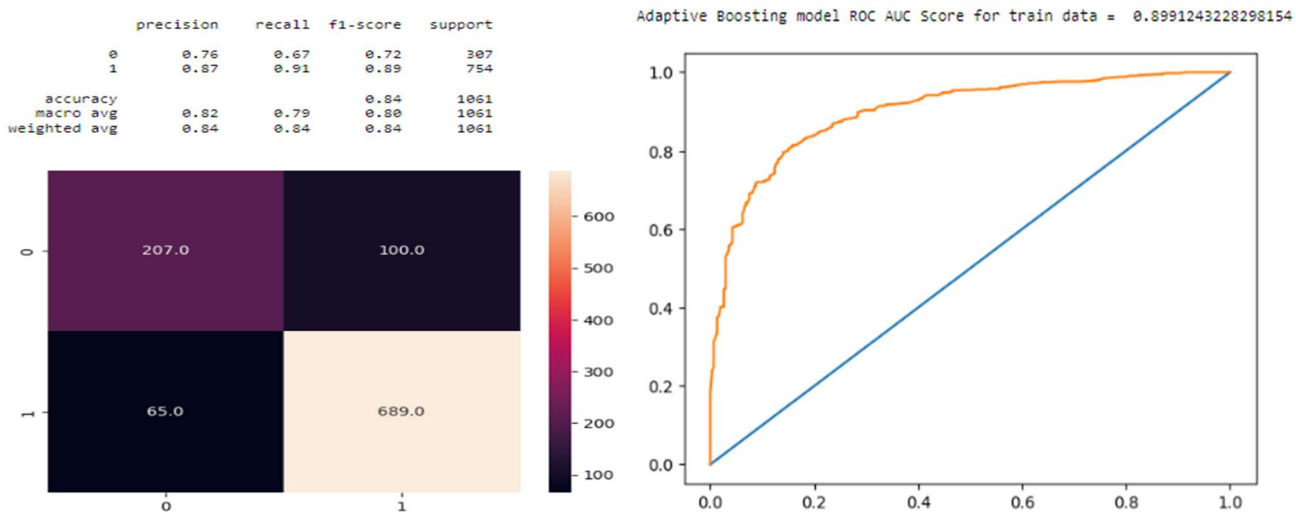


Fig 1.7K

ADABoosting Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Train Data

Inferences from fig 1.7K: -

- Recall of 1's (Labour) = 0.91
- Precision of 1's (Labour) = 0.87
- Accuracy of train data = 0.84
- ROC AUC SCORE of train data= 0.899

Recall of 0's(Conservative) = 0.67 for train data

Precision of 0's(Conservative) = 0.76 for train data

TEST DATA

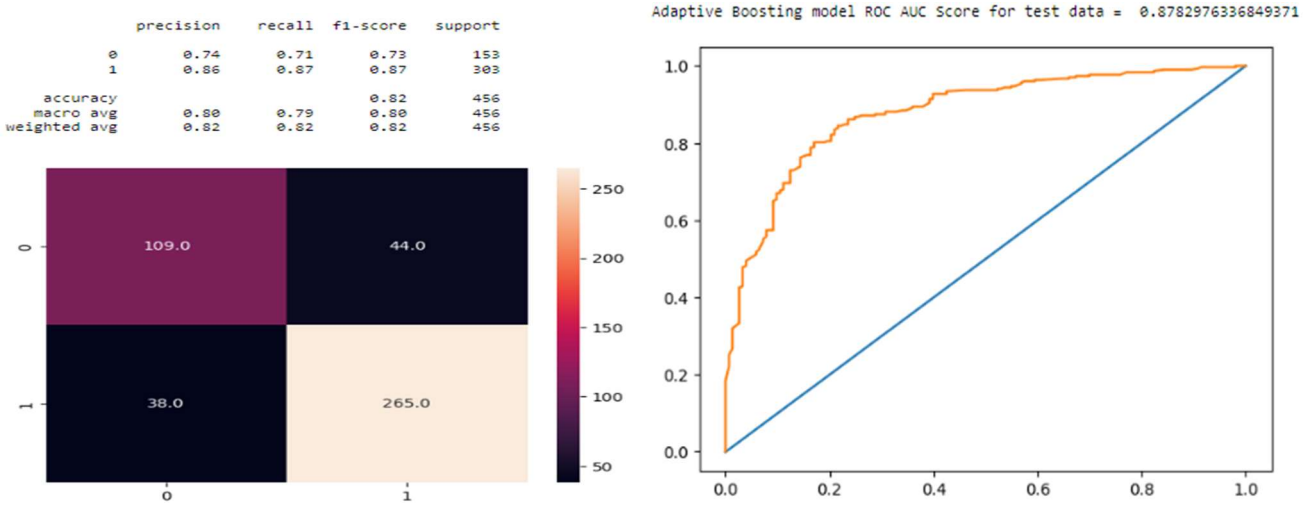


Fig 1.7L

ADABoosting Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Test Data

Inferences from fig 1.7L: -

- Recall of 1's (Labour) = 0.87
- Precision of 1's (Labour) = 0.86
- Accuracy of test data = 0.82
- ROC AUC SCORE of test data= 0.878

Recall of 0's(Conservative) = 0.71 for test data

Precision of 0's(Conservative) = 0.74 for test data

Gradient Boosting Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

TRAIN DATA

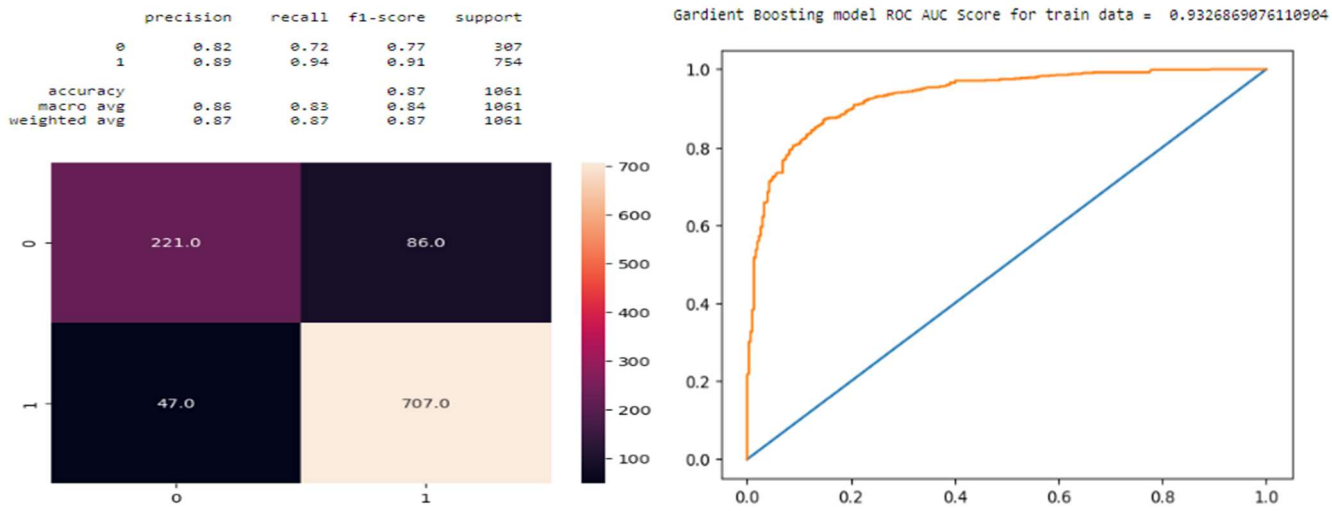


Fig 1.7M

Grad Boosting Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Train Data

Inferences from fig 1.7K: -

- Recall of 1's (Labour) = 0.94 — Recall of 0's(Conservative) = 0.72 for train data
- Precision of 1's (Labour) = 0.89 — Precision of 0's(Conservative) = 0.82 for train data
- Accuracy of train data = 0.87
- ROC AUC SCORE of train data= 0.933

TEST DATA

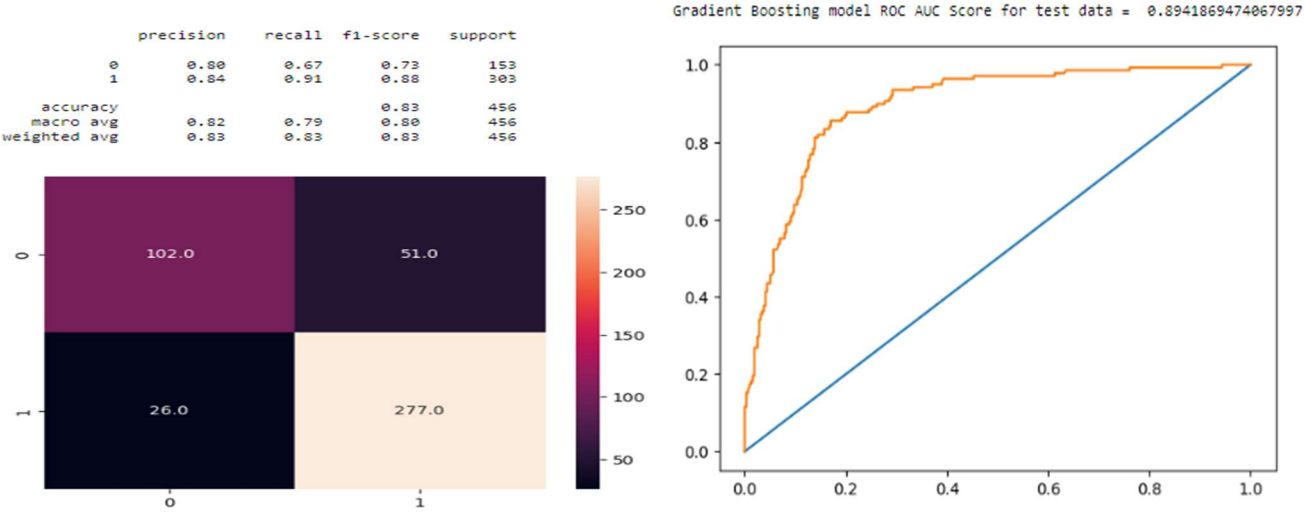


Fig 1.7N

Grad Boosting Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve of Test Data

Inferences from fig 1.7N: -

- Recall of 1's (Labour) = 0.91 — Recall of 0's(Conservative) = 0.67 for test data
- Precision of 1's (Labour) = 0.84 — Precision of 0's(Conservative) = 0.80 for test data
- Accuracy of test data = 0.83
- ROC AUC SCORE of test data= 0.894

Comparison Between Different Classification Model					
Performance Parameters	**Accuracy**	**Recall**	**Precision**	**F1 score**	**ROC AUC score**
Logistic Reg. Train Data	**0.84**	**0.91**	**0.86**	**0.89**	**0.89**
Logistic Reg. Test Data	**0.83**	**0.88**	**0.86**	**0.87**	**0.88**
LDA Train Data	**0.83**	**0.91**	**0.86**	**0.89**	**0.89**
LDA Test Data	**0.83**	**0.88**	**0.86**	**0.87**	**0.89**
KNN Train data	**0.85**	**0.91**	**0.89**	**0.9**	**0.93**
KNN Test Data	**0.81**	**0.86**	**0.85**	**0.86**	**0.86**
Naives Train Data	**0.83**	**0.89**	**0.88**	**0.88**	**0.89**
Naives Test Data	**0.82**	**0.87**	**0.87**	**0.87**	**0.89**
Bagging Train Data	**0.96**	**0.94**	**0.96**	**0.97**	**0.99**
Bagging Test Data	**0.83**	**0.91**	**0.84**	**0.88**	**0.87**
ADA Boost Train Data	**0.84**	**0.91**	**0.87**	**0.89**	**0.9**
ADA Boost Test Data	**0.82**	**0.87**	**0.86**	**0.87**	**0.88**
Grad Boost Train Data	**0.87**	**0.94**	**0.89**	**0.91**	**0.93**
Grad Boost Test Data	**0.83**	**0.91**	**0.84**	**0.88**	**0.89**

Table 1.7A

Inferences from Table 1.7A: -

From the above table if we have to choose the best optimized model than it would be gradient Boosting as it performed well with all parameters with slightest of Over fitting

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.

Solution: -

Inferences

The business issue essentially spun around fostering a model to anticipate which party a citizen would vote in favor of depending on the data about the citizens. The model will in this way be utilized to make an exit poll that will help in predicting the overall win and seats covered by a specific party.

For this to achieve, the analyses assumed CNBE wish to focus more on accurately predicting the Labor's win and hence that has been the class of choice for prediction. The analysis and building of Machine Learning models based on a restricted dataset of 1525 citizens with specific details of the electors.

INSIGHTS: -

- Majority of the Voters are between the age group (35-75) . There are no votes in the age group (18-23)
- More 50% of the voters are of age above 53 years and only the bottom 25% voters are aged less than 41 years;
- There are more Female voters than Male voters
- Conservative voters have better political knowledge of political parties' position on European integration than their Labour counterpart
- Majority of people (75% of voters) think that household and national economic condition is satisfactory as most have ranked them in 3 or 4 out of 5
- Labour leader Blair is more popular among people than Conservative leader Hague as Blair has received a rating of 4 on average, whereas Hague has received a mixed rating of 2 and 4;
- 43% approx Conservative voters have rated the national economic condition average with score of 3, further indicating, the overall assessment to be between poor and average.
- Most people find Blair to be a better leader and if the Conservative party wants to win then they have to focus in improving Hague's image among people, or go with a different candidate
- The general population does not seem to be very eurosceptic as cumulative frequency of non-eurosceptic people (who opted for 6 or less) seem to be higher than the cumulative frequency of eurosceptic people (who opted for 7 or higher

RECOMMENDATION: -

- CNBE must gather data of voters aged between 18 and 24 so as to make the predictions more accurate.
- It needs to be addressed that, the larger the number of voters, better the MachineLearning models can be optimized
- The dataset must also include additional assessment ratings such as Health Policy, Employment Generation, Relationship with USA or other Countries, etc
- Even larger dataset could help us predicting the result more accurately.

2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)

Solution: -

Total Words in each Speeches: -

	Name	Speech	Totalwords
0	Roosevelt	On each national day of inauguration since 178...	1323
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769

Fig. 2.1A

Insights from the Fig. 2.1A: -

- Total words in Roosevelt Speech = 1323
- Total words in Kennedy Speech = 1364
- Total words in Nixon Speech = 1769

Total words in all speech = 4456 words

Total Character in each Speeches: -

	Name	Speech	character_count
0	Roosevelt	On each national day of inauguration since 178...	7651
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	7673
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10106

Fig. 2.1B

Insights from the Fig. 2.1B: -

- Total Characters in Roosevelt Speech = 7651
- Total Characters in Kennedy Speech = 7673
- Total Characters in Nixon Speech = 10106

Total Characters in all speech = 25430 Characters

Total Sentences in each Speeches: -

	Name	Speech	Totalwords
0	Roosevelt	On each national day of inauguration since 178...	38
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	27
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	51

Fig. 2.1C

Insights from the Fig. 2.1C: -

- Total Sentence in Roosevelt Speech = 38
- Total Sentence in Kennedy Speech = 27
- Total Sentence in Nixon Speech = 51

Total Sentences in all speech = 116

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.

Solution: -

Before removal of stopwords there are certains steps which needs to be taken: -

Step 1. Converting Upper case in Lower case: -

The text must be converted to lowercase in order to reduce the redundant words such as 'The' and 'the'. Here, these are two separate words in the speech which however for the purpose of building models, word clouds make it inaccurate. In order to mitigate the issue of double counting of words, the text of 3 speeches have been converted to lowercase

Step 2. Removal of the Punctuation marks: -

After converting Upper case into Lower case , another important pre-processing step involves removal of punctuations which if not removed will cause incorrect building of models and word clouds. Thus the text contains punctuations like commas, full stop, apostrophe, etc have been removed. The text also contains some special characters such as '--' and '\', they too are removed.

Step3. Removal of Stopwords: -

	Name	Speech	speech_stopword_removed	Totalwords	speech_stopword_removed_totalwords
0	Roosevelt	On each national day of inauguration since 178...	national day inauguration since 1789 people re...	1323	624
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	vice president johnson mr speaker mr chief jus...	1364	689
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	mr vice president mr speaker mr chief justice ...	1769	819

Fig. 2.2A

Insights from the Fig. 2.2A: -

Speech	Speech with stopwords	Speech without Stopwords
Roosevelt	**1323**	**624**
Kennedy	**1364**	**689**
Nixon	**1769**	**819**

	Speech	speech_stopword_removed
0	On each national day of inauguration since 178...	national day inauguration since 1789 people re...
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	vice president johnson mr. speaker mr. chief j...
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	mr. vice president mr. speaker mr. chief justi...

Sample Sentence after removal of stopwords

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

Solution: -

As we have removed all the stopwords, an important text pre-processing step is taken to reduce the words to their root words, called **Stemming**. It is a rule-based approach because it slices the inflected words from prefix or suffix as per the need using a set of commonly under used prefix and suffix, like "-ing", "-ed", "-es", "-pre", etc. It results in a word that is actually not a word. Using the Porter Stemmer method available in then ltk package, the texts of 3 speeches are stemmed to their root word

Top three words in Roosevelt Speech: -

	Words	Frequency
0	nation	16
82	know	10
24	us	8

Top three words in Kennedy Speech: -

	Words	Frequency
70	let	11
108	us	11
46	power	9

Top three words in Nixon Speech: -

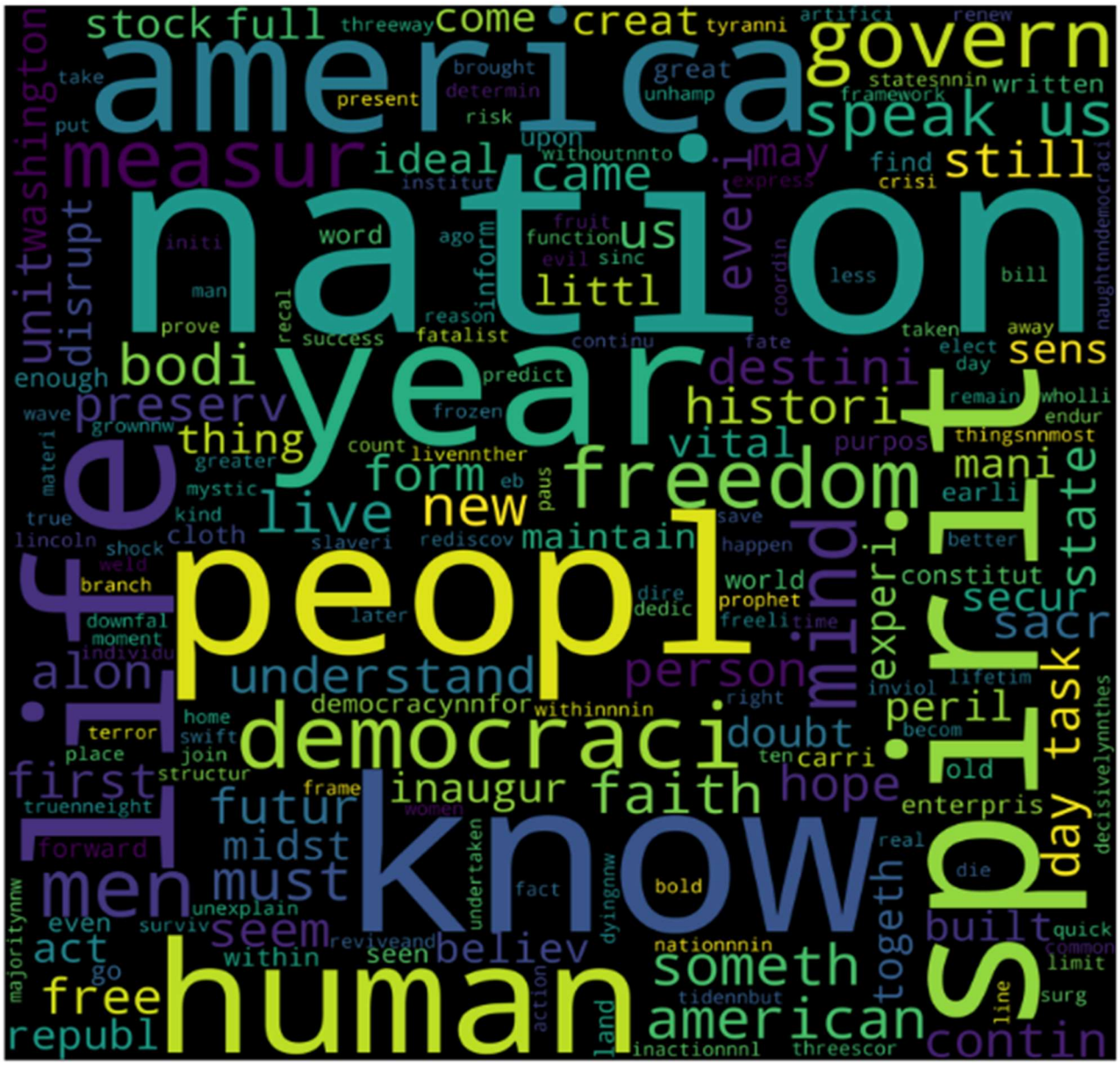
	Words	Frequency
43	us	26
21	america	19
62	respons	16

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

Solution: -

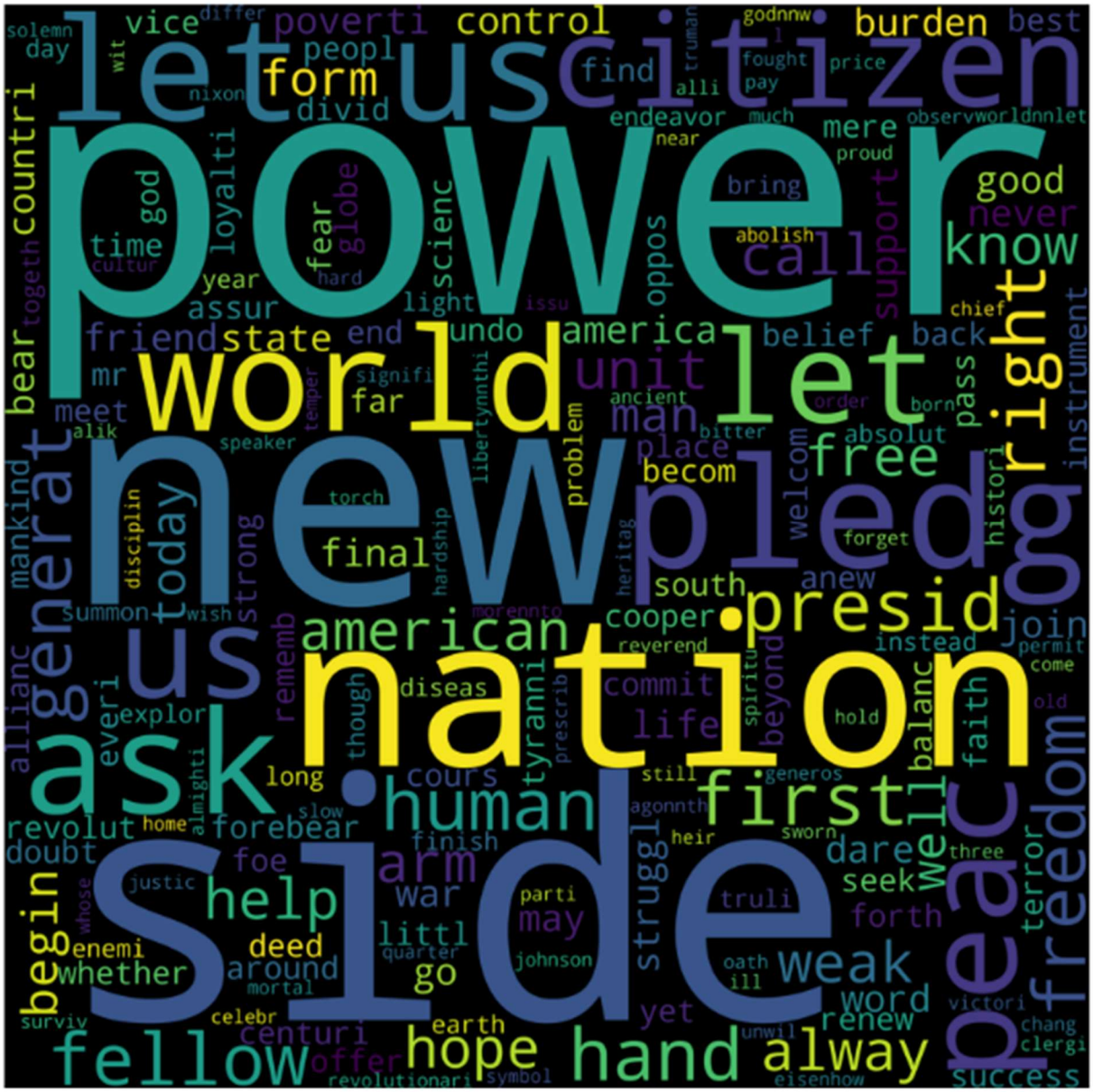
1. Word cloud for Roosevelt Speech

Word Cloud for Roosevelt Speech (after cleaning)!!



2. Word cloud for Kennedy Speech

Word Cloud for Kennedy Speech (after cleaning)!!



3. Word cloud for Nixon Speech

Word Cloud for Nixon Speech (after cleaning)!!