Machine Learning 2 Project - Business Report

By: Aaryani Kadiyala

PGP-Data Science and Business Analytics

(PGPDSBA.O.JAN24.A)

Table of Contents

List of Tables:

List of Figures:

Problem 1:

Context:

CNBE, a prominent news channel, is gearing up to provide insightful coverage of recent elections, recognizing the importance of data-driven analysis. A comprehensive survey has been conducted, capturing the perspectives of 1525 voters across various demographic and socio-economic factors. This dataset encompasses 9 variables, offering a rich source of information regarding voters' characteristics and preferences.

Objective:

The primary objective is to leverage machine learning to build a predictive model capable of forecasting which political party a voter is likely to support. This predictive model, developed based on the provided information, will serve as the foundation for creating an exit poll. The exit poll aims to contribute to the accurate prediction of the overall election outcomes, including determining which party is likely to secure the majority of seats.

Data Description:

System measures used:

1. **vote**: Party choice: Conservative or Labour
2. **age**: in years
3. **economic.cond.national**: Assessment of current national economic conditions, 1 to 5.
4. **economic.cond.household**: Assessment of current household economic conditions, 1 to 5.
5. **Blair**: Assessment of the Labour leader, 1 to 5.
6. **Hague**: Assessment of the Conservative leader, 1 to 5.
7. **Europe**: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
8. **political.knowledge**: Knowledge of parties' positions on European integration, 0 to 3.
9. **gender:** female or male.

Solution:-

	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

Figure 1: Dataset Sample

Dropped Unnamed:0 from the data set as it does not add any significance to the data set.

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

Figure 2: Dataset Sample after dropping columns

- The data contains 1525 rows and 9 columns.
- Data consists of 2 object & 7 int64 data type columns.
- 2 object type columns are vote & gender.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

Figure 3: Data types of the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
vote	1525	2	Labour	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	1525.0	NaN	NaN	NaN	54.18	15.71	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	NaN	NaN	NaN	3.25	0.88	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	NaN	NaN	NaN	3.14	0.93	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	NaN	NaN	NaN	3.33	1.17	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	NaN	NaN	NaN	2.75	1.23	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	NaN	NaN	NaN	6.73	3.3	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	NaN	NaN	NaN	1.54	1.08	0.0	0.0	2.0	2.0	3.0
gender	1525	2	female	812	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 4: Statistical summary of the dataset

- There are no null values in the dataset.
- There are 8 duplicate values in the dataset.
- After removing the duplicates, dataset has 1517 rows & 9 columns.
- Converting the necessary variables to object as it is meant to be. Because these variables have values that are numeric but are a categorical column.

Checking the descriptive dataset again,

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
vote	1517	2	Labour	1057	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	1517.0	NaN	NaN	NaN	54.241266	15.701741	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1517.0	5.0	3.0	604.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
economic.cond.household	1517.0	5.0	3.0	645.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Blair	1517.0	5.0	4.0	833.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Hague	1517.0	5.0	2.0	617.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Europe	1517.0	11.0	11.0	338.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
political.knowledge	1517.0	4.0	2.0	776.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gender	1517	2	female	808	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 5: Statistical summary after removing duplicates

- From the above snippet we can come to a conclusion that the dataset has only one integer column which is 'age'
- The mean and median for the only integer column 'age' is almost same indicating the column is normally distributed.
- 'vote' has two unique values Labour and Conservative, which is also a dependent variable. 'gender' has two unique values male and female.
- Rest all the columns has object variables with 'Europe' being highest having 11 unique values.
- Replacing Male with 1 & Female with 0, Conservative with 1 & Labour with 0.

```
gender
0    808
1    709
Name: count, dtype: int64
```

```
vote
0    1057
1     460
Name: count, dtype: int64
```

Figure 6: Value counts after replacing with 0's & 1's

Univariant, Bivariant & Multivariant Analysis: -

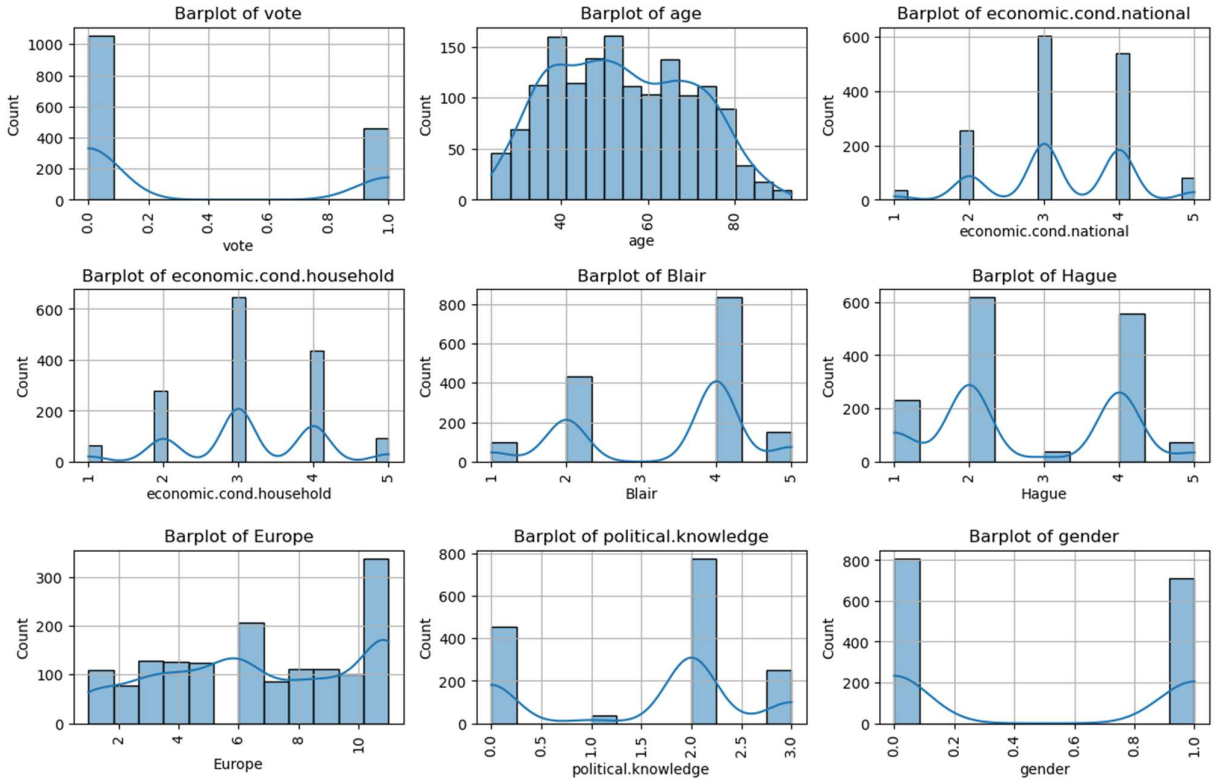

Figure 7: Univariant analysis of the dataset

From the above analysis it can summarized as follows,

- There are more female voters than male.
- Maximum no. of voters are around 40 years old and minimum no. of voters are around 95 years old.
- Most of the people are supporters of labour party.
- National economic and household assessment of conditions which vary from 1 to 5 are maximum at 3 in both assessments.
- Blair's assessment is maximum at 4 whereas, Hague's assessment is maximum at 2.
- Euroseptic sentiment in Europe maximum is 11.

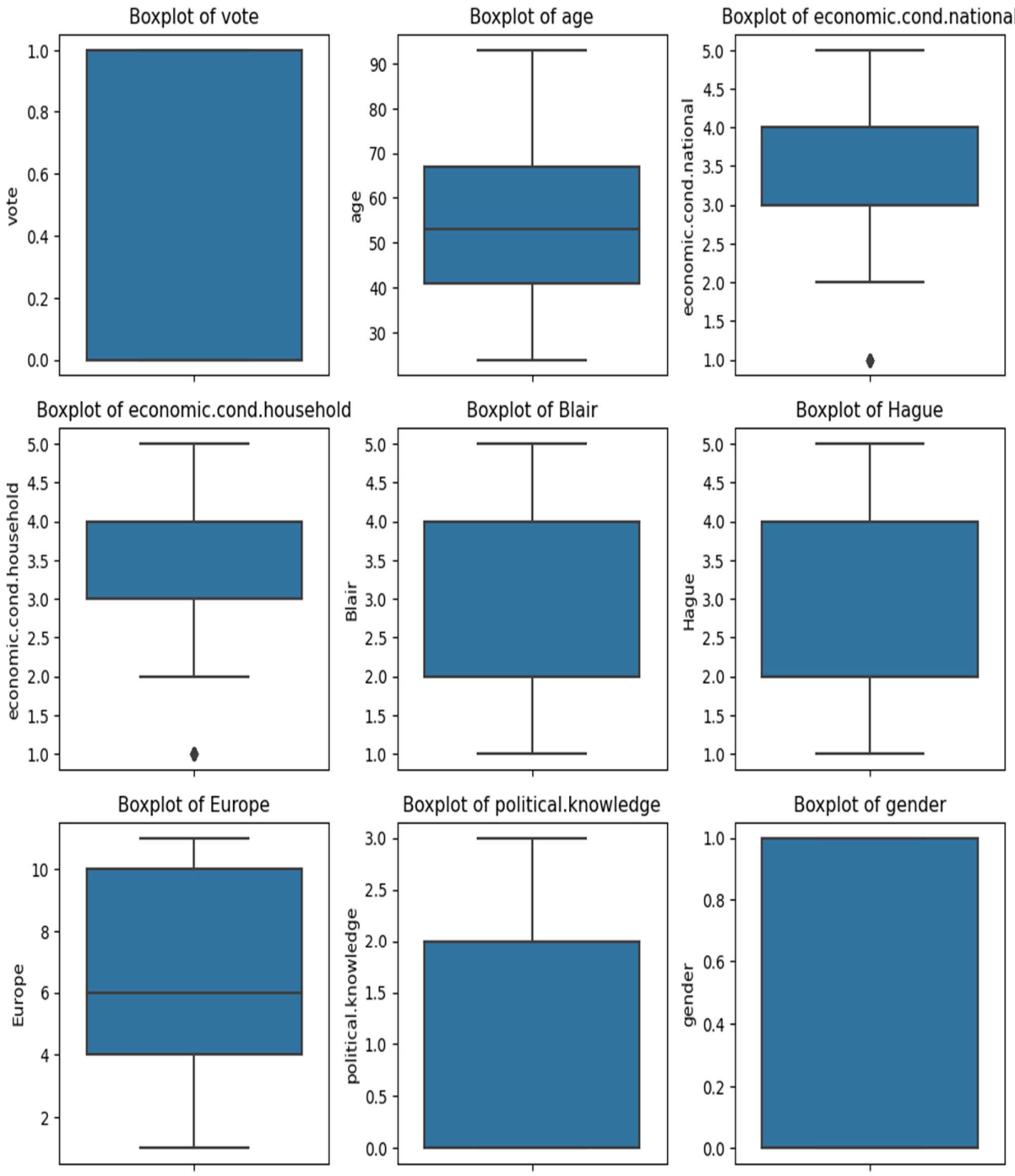

Figure 8: Box Plot of the dataset

From the above analysis it can summarized as follows,

- None of the columns have outliers except for the columns 'economic.cond.national' & 'economic.cond.household'.

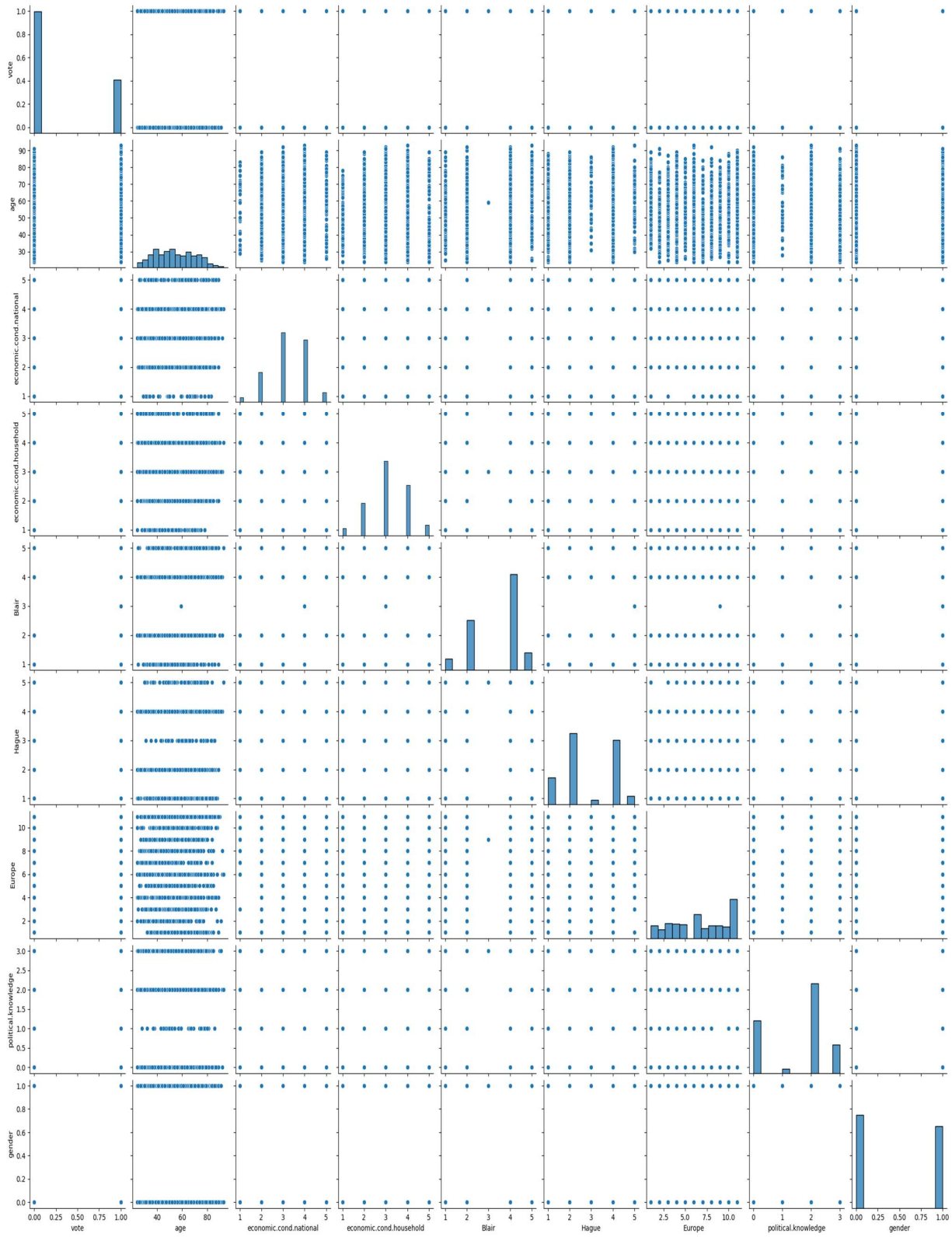

Figure 9: Pair Plot of the dataset

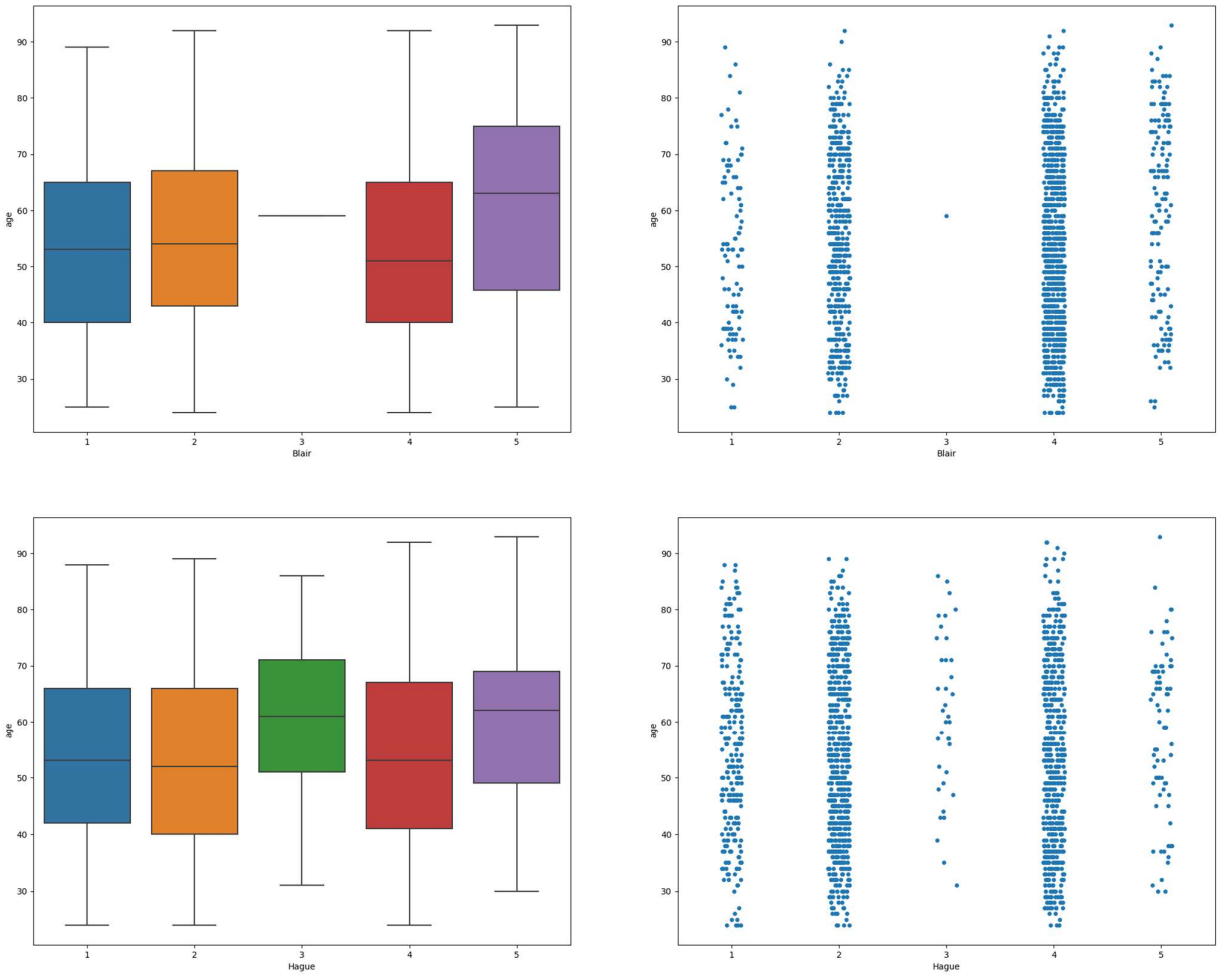

Figure 10: Box Plot & Scatter plot of Blair vs Age & Hague vs Age

From the above analysis it can summarized as follows,

- Maximum no. of voters for Blair, with 4 assessment and minimum no. of voters for Blair are with 1 as the assessment.
- Blair doesn't have any voters with 3 as the assessment.
- Maximum no. of voters for Hague, with 4 assessment and minimum no. of voters for Hague are with 3 as the assessment.
- Most of the old age voters prefer to vote for conservative.
- Most of the young and mid-age population prefer to vote for Labour.

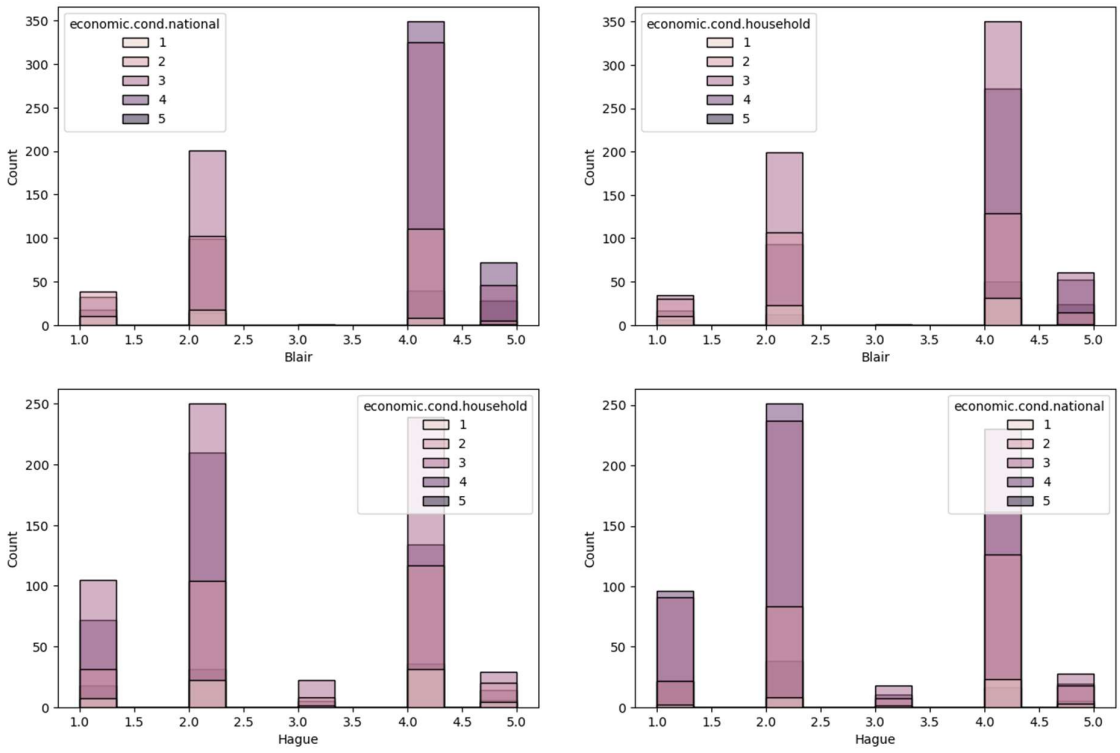

Figure 11: Histogram of Blair & Hague vs economic.cond.national & economic.cond.household

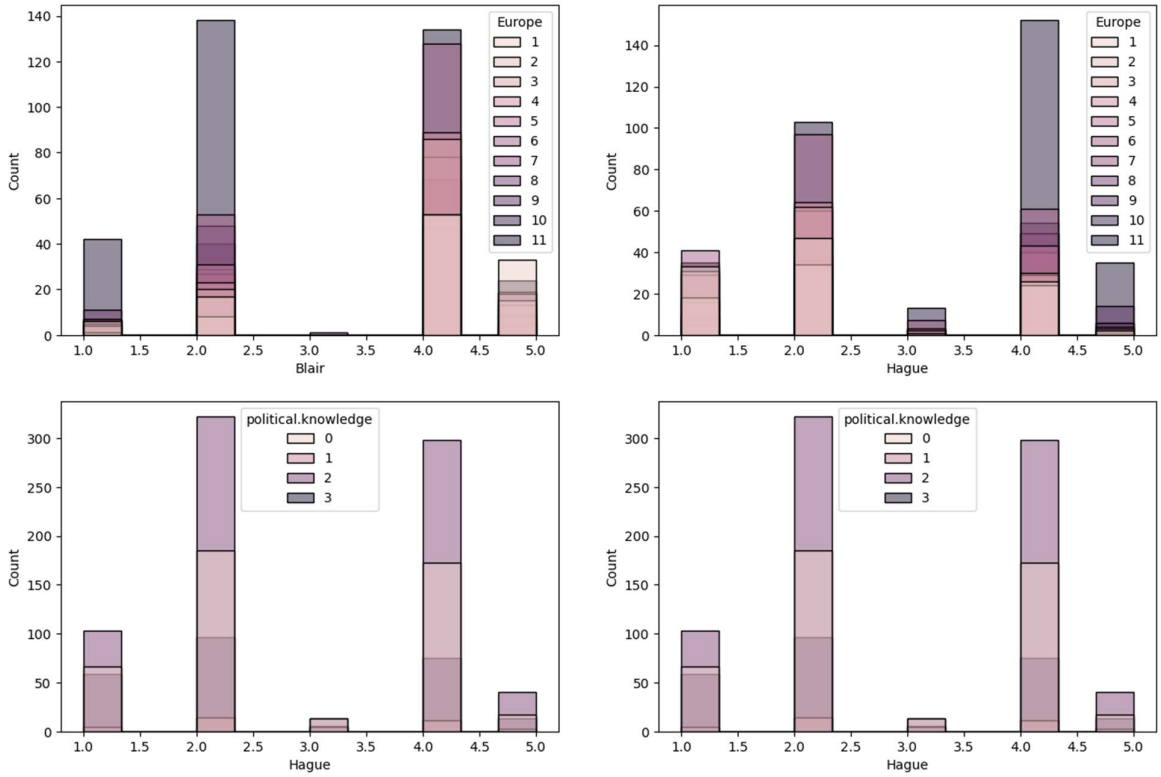

Figure 12: Histogram of Blair & Hague vs Europe & political.knowledge

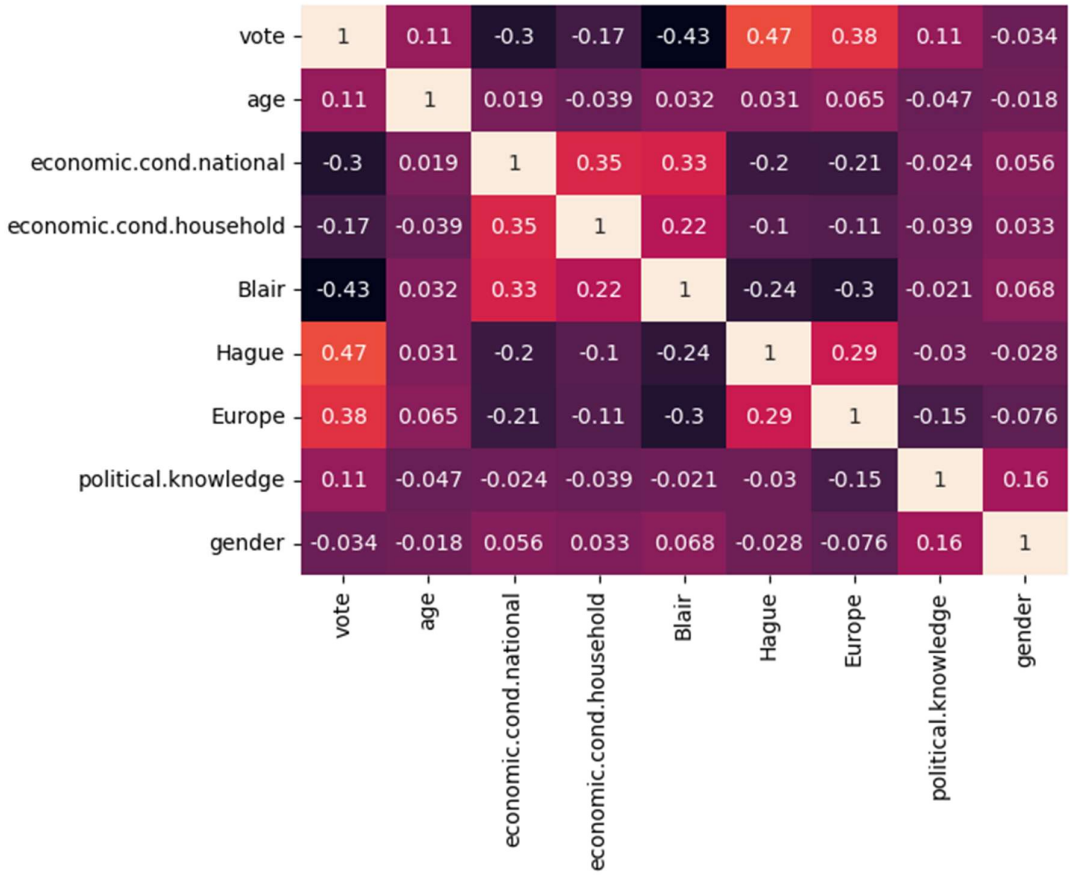

Figure 13: Heatmap of the dataset

From the above analysis it can summarized as follows,

- 68% Corelation between Blair & gender.
- -76% Corelation between Europe & gender.
- 56% Corelation between economic.cond.national & gender.
- 47% Corelation between Hauge & vote.
- 38% Corelation between Europe & vote.
- 65% Corelation between Blair & Age.

Model Building, Model Performance evaluation & Model Performance improvement: -

'Vote' variable has taken has a y variable (dependent variable) and all other variables are taken has x variable (independent variable).

The dataset has been split into training & testing dataset with 70:30 ratio. 70% as training dataset & 30% as testing dataset.

```
Training Data Shape: (1061, 8)
Testing Data Shape: (456, 8)
```

Figure 14: Shape of training & testing dataset

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
991	34	2	4	1	4	11	2	0
1274	40	4	3	4	4	6	0	1
649	61	4	3	4	4	7	2	0
677	47	3	3	4	2	11	0	1
538	44	5	3	4	2	8	0	1

Figure 15: Training Dataset

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
504	71	3	3	2	2	8	2	0
369	43	3	2	4	2	8	3	1
1075	89	5	5	5	2	1	2	1
1031	47	2	3	2	4	8	2	0
1329	33	5	4	4	4	8	0	1

Figure 16: Testing Dataset

Models: -

Logistic Regression: -

Using logistic regression we are trying to predict the dependent variable; logistic regression is used in predicting the categorical dependent variable. To perform the regression, model the data set has to be all numeric, to achieve this we have encoded all the object data in the dataset to numeric.

```
Classification Report of Training Dataset

              precision    recall  f1-score   support

           0       0.86      0.91      0.88       754
           1       0.74      0.64      0.69       307

    accuracy                           0.83      1061
   macro avg       0.80      0.77      0.79      1061
weighted avg       0.83      0.83      0.83      1061
```

Figure 17: Classification Report of Training Dataset

Classification Report of Testing Dataset

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

Figure 18: Classification Report of Testing Dataset

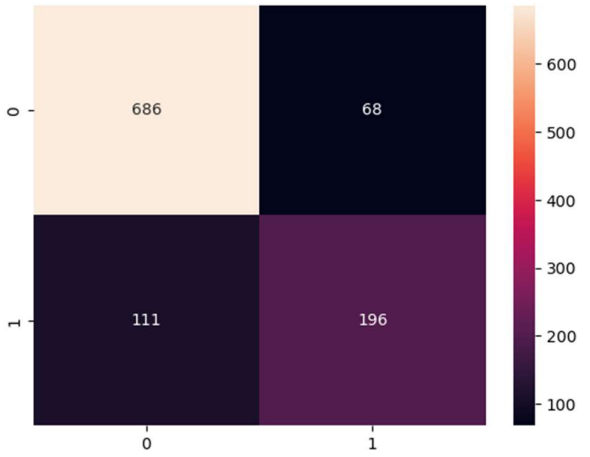

Figure 19: Confusion Matrix of training Data Set

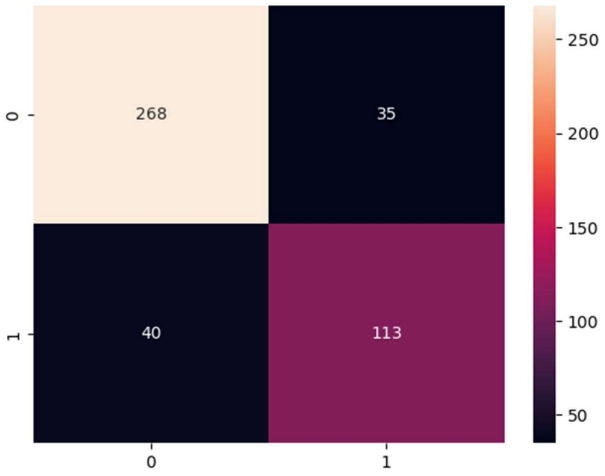

Figure 20: Confusion Matrix of testing Data Set

Inference from Train data:

- 686 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 196 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 111 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 68 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 268 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 113 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 40 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)

- 35 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

AUC of training Data Set : 0.890

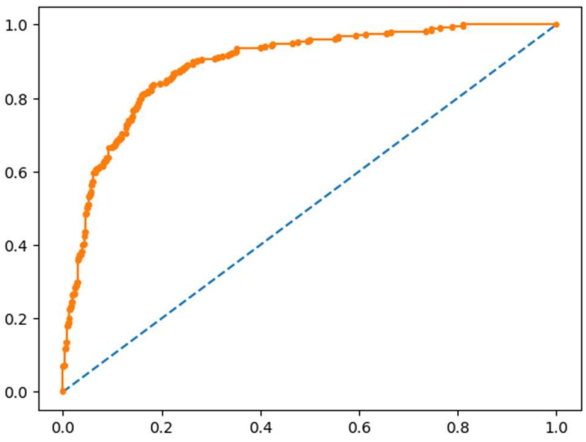

Figure 21: AUC of training Data Set

AUC of testing Data Set : 0.883

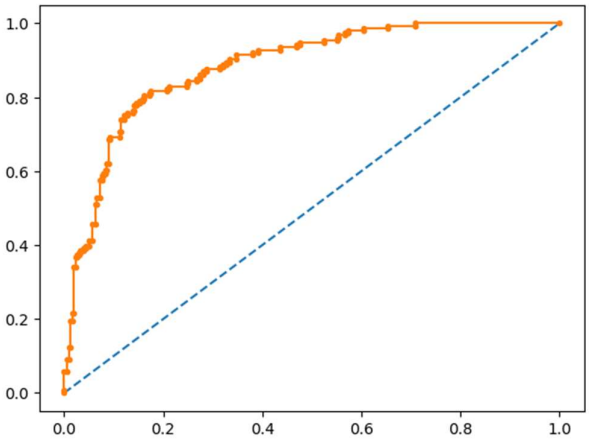

Figure 22: AUC of testing Data Set

LDA (Linear Discriminant Analysis):

Classification Report of Training Dataset

	precision	recall	f1-score	support
0	0.91	0.86	0.89	792
1	0.65	0.74	0.69	269
accuracy			0.83	1061
macro avg	0.78	0.80	0.79	1061
weighted avg	0.84	0.83	0.84	1061

Classification Report of Testing Dataset

	precision	recall	f1-score	support
0	0.89	0.86	0.88	311
1	0.73	0.77	0.74	145
accuracy			0.83	456
macro avg	0.81	0.82	0.81	456
weighted avg	0.84	0.83	0.83	456

Figure 23: Classification Report of Training & Testing Dataset

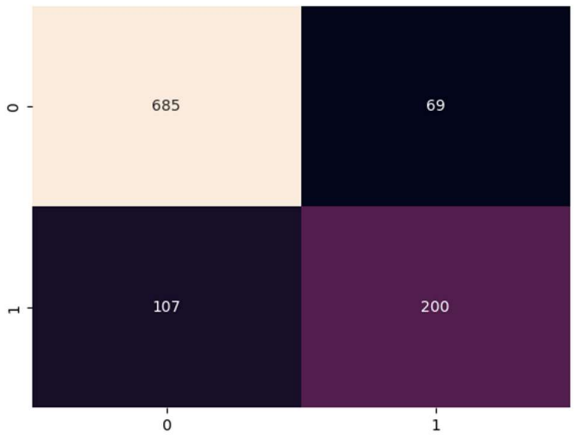

Figure 24: Confusion Matrix of training Data Set

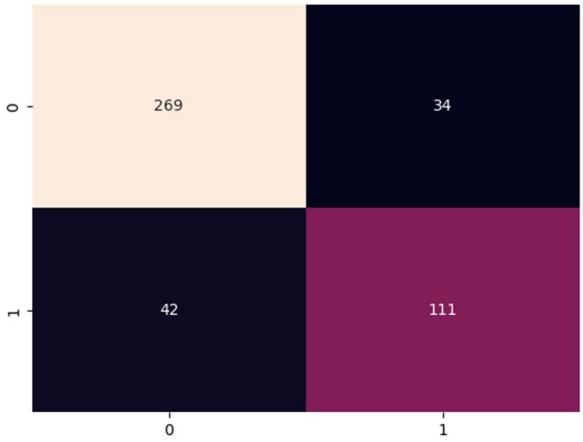

Figure 25: Confusion Matrix of testing Data Set

Inference from Train data:

- 685 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 200 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 107 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 69 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 269 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 111 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 42 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 34 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

AUC of training Data Set : 0.889

AUC of testing Data Set : 0.888

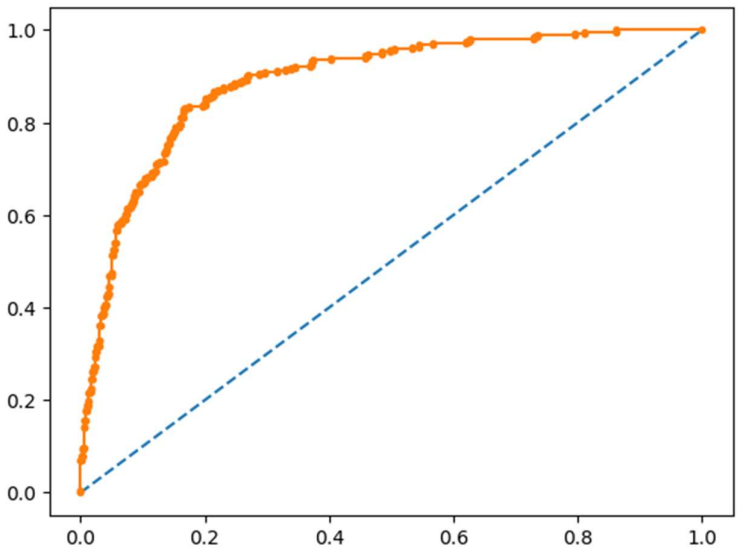

Figure 26: AUC of training Data Set

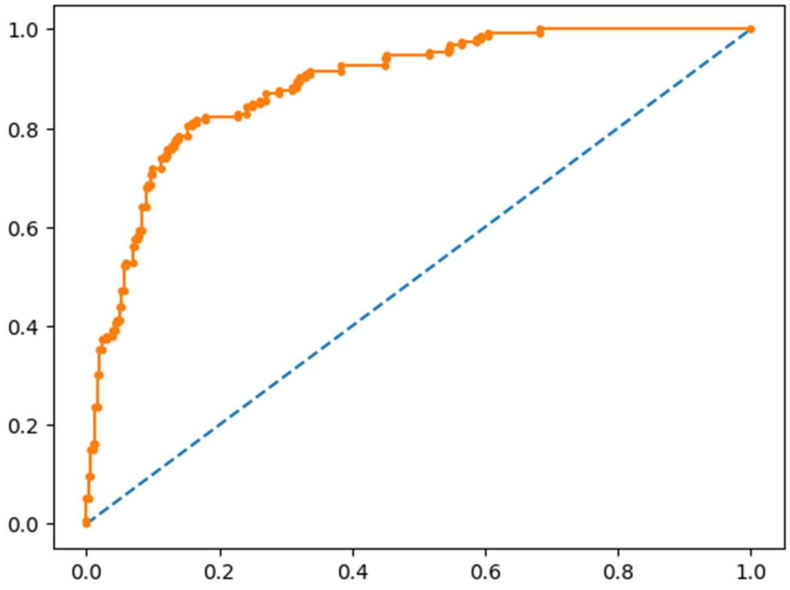

Figure 27: AUC of testing Data Set

KNN (K-Nearest Neighbour):

KNN Model Score: 0.8245614035087719

As shown above we have obtained 82.4 as a KNN Model Score.

AUC of training Data Set : 0.924

AUC of testing Data Set : 0.861

Classification Report of Training Dataset

	precision	recall	f1-score	support
0	0.93	0.88	0.90	797
1	0.68	0.79	0.73	264
accuracy			0.86	1061
macro avg	0.80	0.83	0.82	1061
weighted avg	0.87	0.86	0.86	1061

Classification Report of Testing Dataset

	precision	recall	f1-score	support
0	0.91	0.84	0.87	327
1	0.66	0.78	0.72	129
accuracy			0.82	456
macro avg	0.78	0.81	0.79	456
weighted avg	0.84	0.82	0.83	456

Figure 28: Classification Report of Training & Testing Dataset

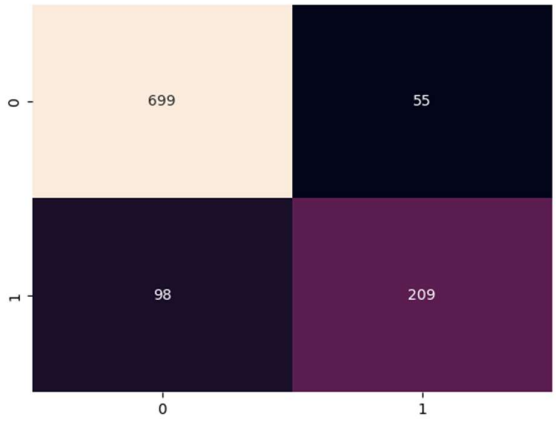

Figure 29: Confusion Matrix of training Data Set

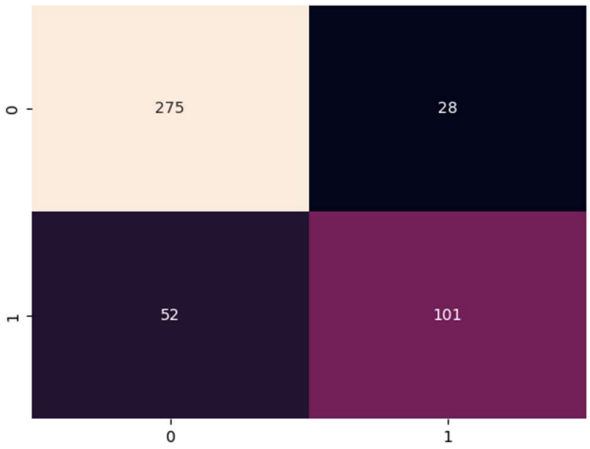

Figure 30: Confusion Matrix of testing Data Set

Inference from Train data:

- 699 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 209 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 98 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 55 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 275 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 101 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 52 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 28 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

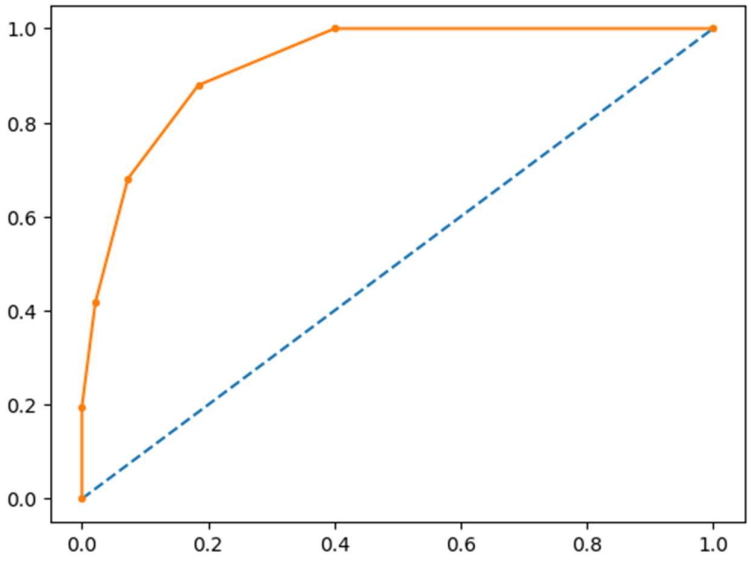

Figure 31: AUC of training Data Set

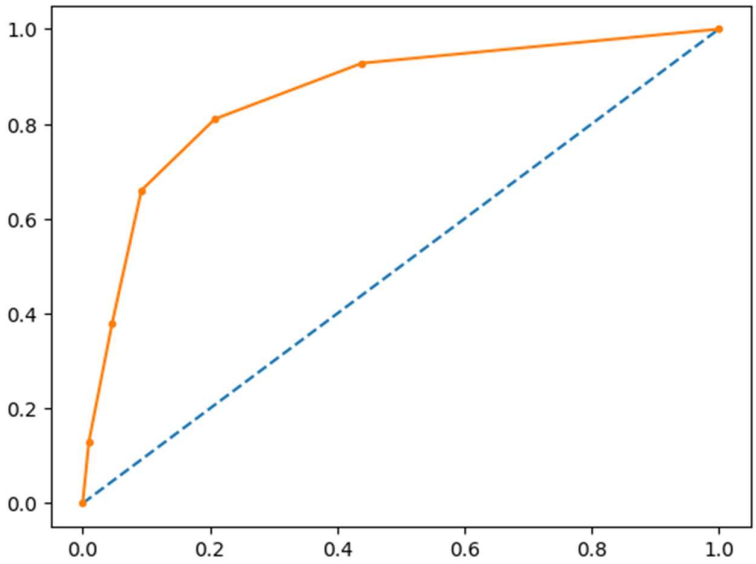

Figure 32: AUC of testing Data Set

Naïve Bayes: -

Classification Report of Training Dataset

	precision	recall	f1-score	support
0	0.90	0.88	0.89	771
1	0.69	0.73	0.71	290
accuracy			0.84	1061
macro avg	0.79	0.80	0.80	1061
weighted avg	0.84	0.84	0.84	1061

Classification Report of Testing Dataset

	precision	recall	f1-score	support
0	0.87	0.87	0.87	304
1	0.73	0.74	0.73	152
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

Figure 33: Classification Report of Training & Testing Dataset

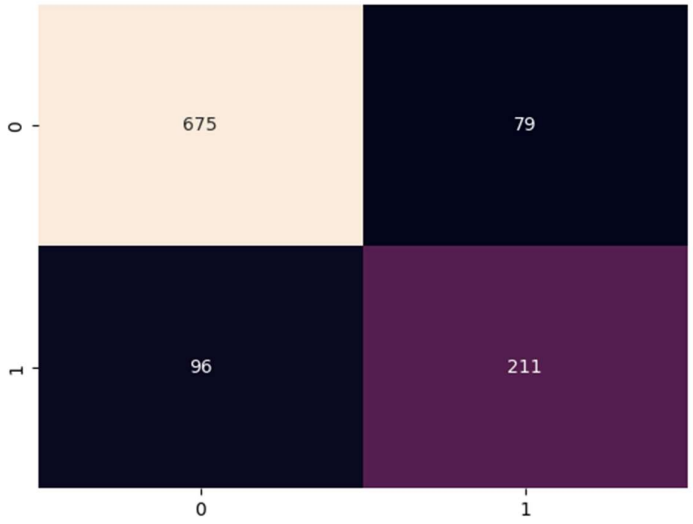

Figure 34: Confusion Matrix of training Data Set

Inference from Train data:

- 675 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 211 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 96 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 79 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

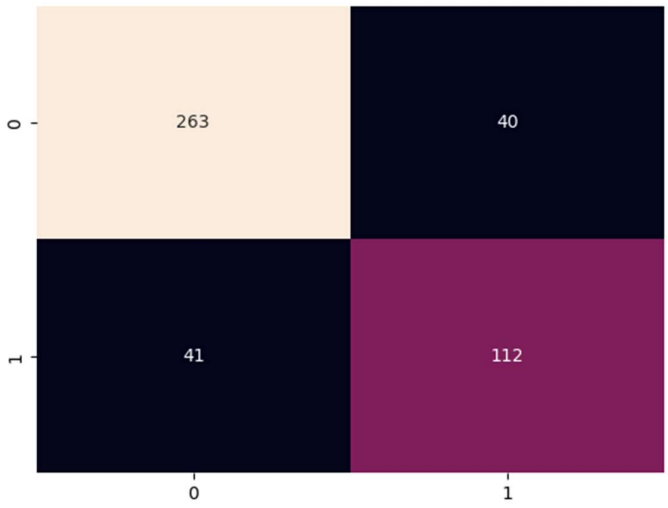

Figure 35: Confusion Matrix of testing Data Set

Inference from Test data:

- 263 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 112 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 41 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 40 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

AUC of training Data Set : 0.888

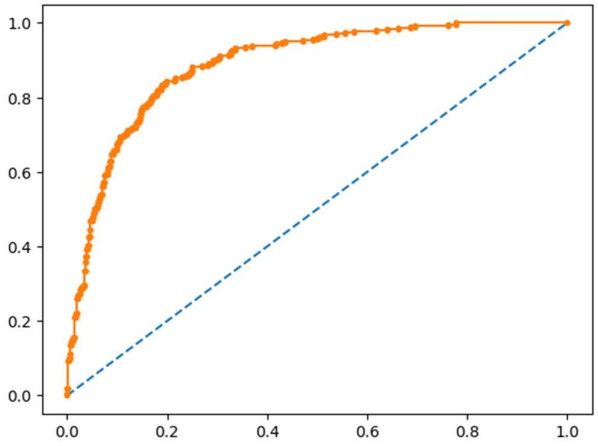

Figure 36: AUC of training Data Set

AUC of testing Data Set : 0.876

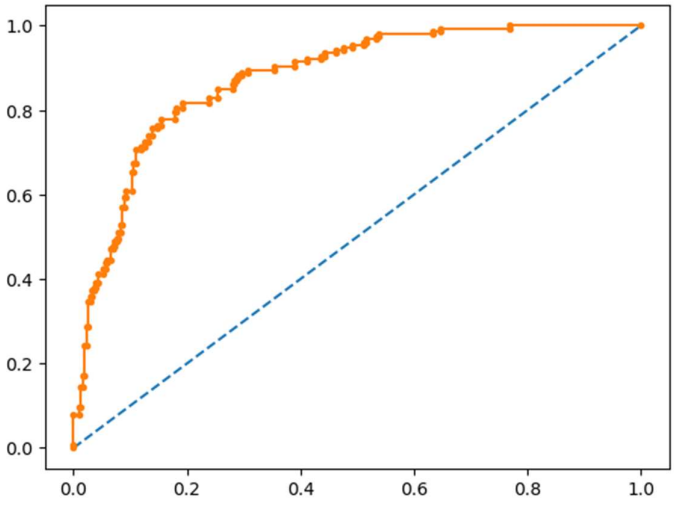

Figure 37: AUC of testing Data Set

Bagging and Boosting: -

ABD Model Score: 0.8369462770970783

As shown above we have obtained 83.6 as a ABD Model Score.

Classification Report of Training Dataset

	precision	recall	f1-score	support
0	0.93	0.85	0.89	823
1	0.61	0.78	0.68	238
accuracy			0.84	1061
macro avg	0.77	0.82	0.79	1061
weighted avg	0.86	0.84	0.84	1061

Classification Report of Testing Dataset

	precision	recall	f1-score	support
0	0.89	0.83	0.86	326
1	0.64	0.75	0.69	130
accuracy			0.81	456
macro avg	0.77	0.79	0.78	456
weighted avg	0.82	0.81	0.81	456

Figure 38: Classification Report of Training & Testing Dataset

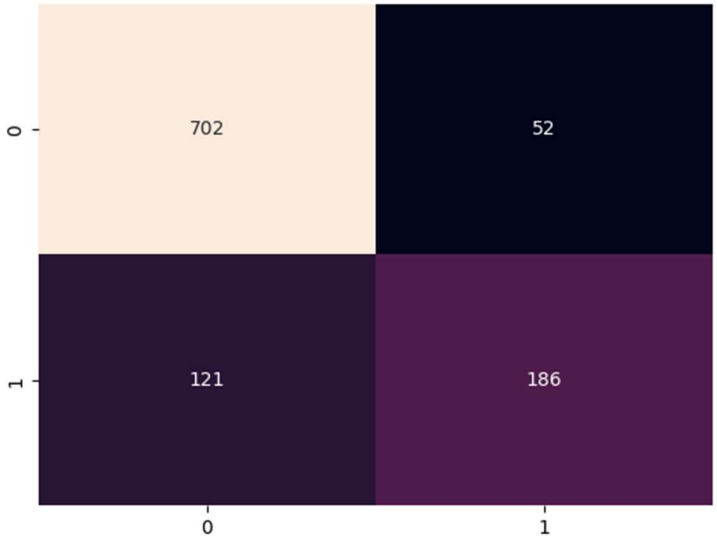

Figure 39: Confusion Matrix of training Data Set

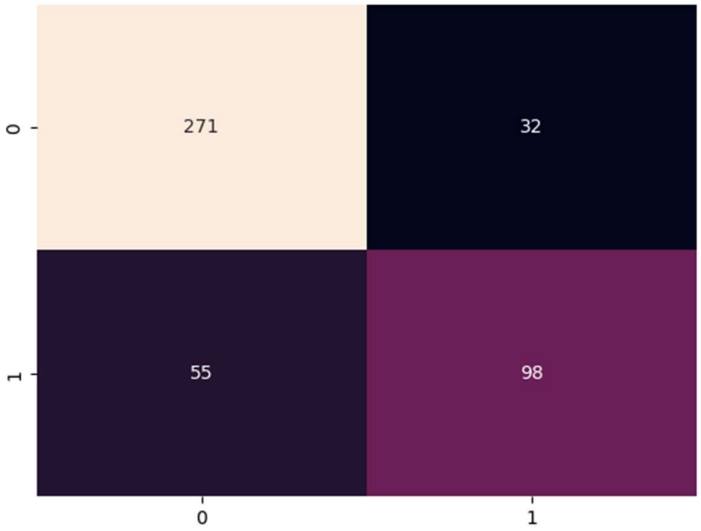

Figure 40: Confusion Matrix of testing Data Set

Inference from Train data:

- 702 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 186 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 121 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 52 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0).

Inference from Test data:

- 271 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 98 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 55 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 32 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

AUC of training Data Set : 0.902

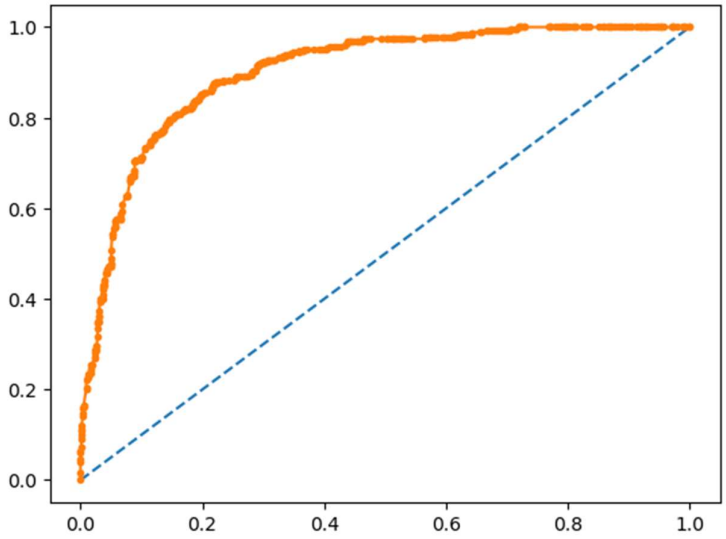

Figure 41: AUC of training Data Set

AUC of testing Data Set : 0.884

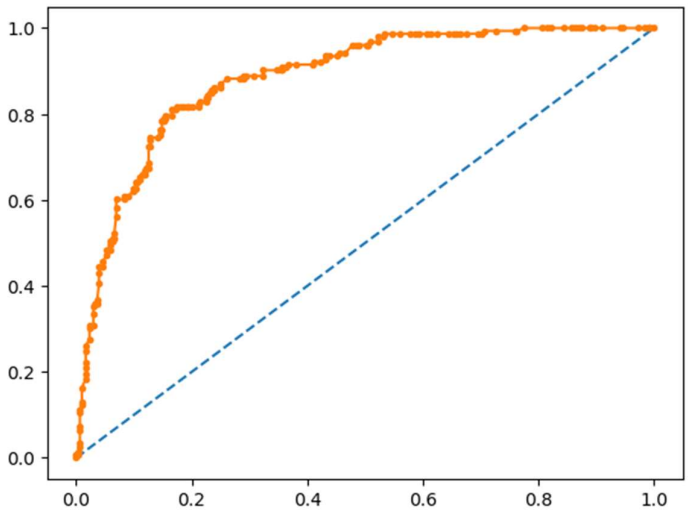

Figure 42: AUC of testing Data Set

Design Tree: -

Classification Report of Training Dataset

	precision	recall	f1-score	support
0	0.87	0.90	0.89	730
1	0.76	0.71	0.73	331
accuracy			0.84	1061
macro avg	0.82	0.80	0.81	1061
weighted avg	0.84	0.84	0.84	1061

Classification Report of Testing Dataset

	precision	recall	f1-score	support
0	0.83	0.86	0.84	290
1	0.74	0.68	0.71	166
accuracy			0.80	456
macro avg	0.78	0.77	0.78	456
weighted avg	0.79	0.80	0.79	456

Figure 43: Classification Report of Training & Testing Dataset

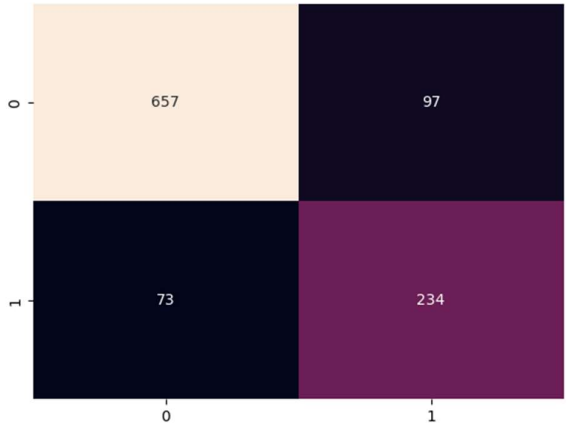

Figure 44: Confusion Matrix of training Data Set

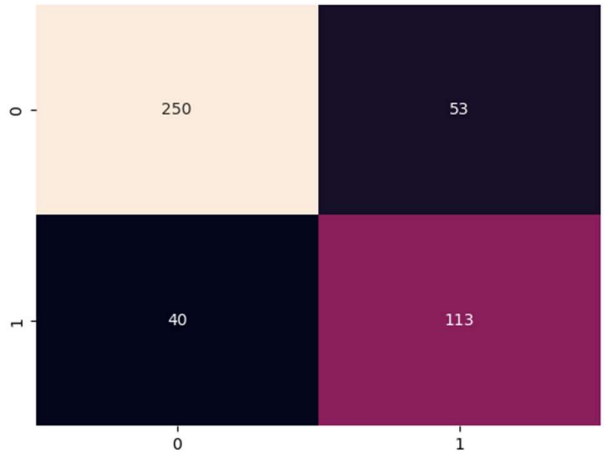

Figure 45: Confusion Matrix of testing Data Set

Inference from Train data:

- 657 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 234 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 97 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 73 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0).

Inference from Test data:

- 250 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 113 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 53 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 40 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

AUC of training Data Set : 0.907

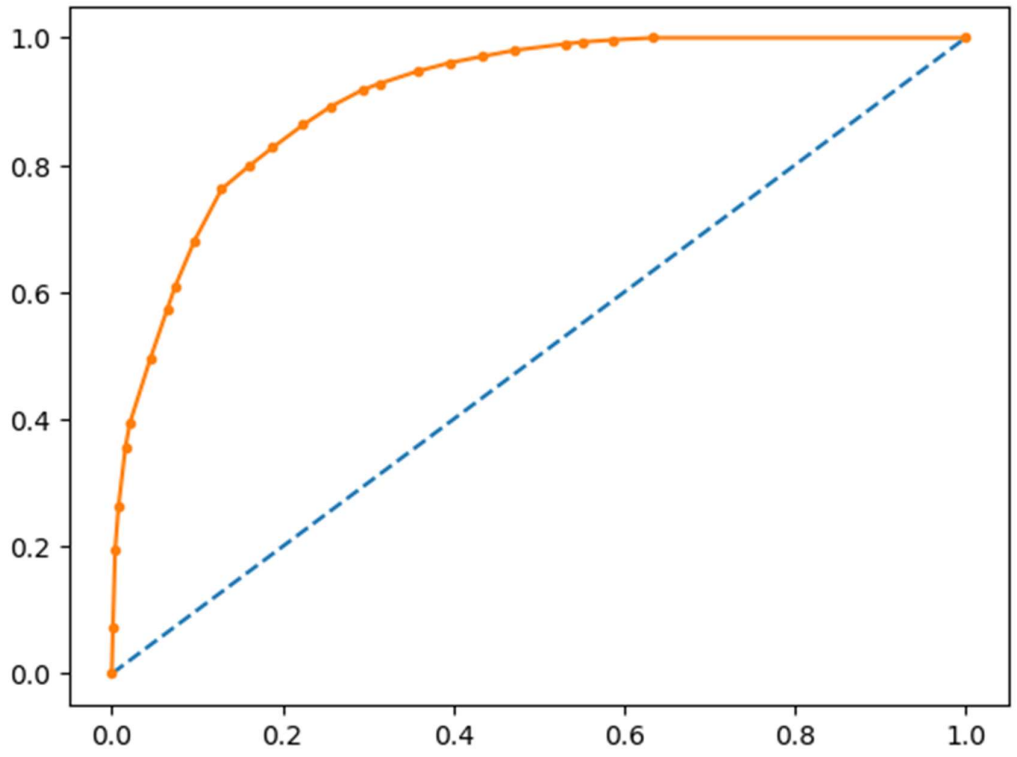

Figure 46: AUC of training Data Set

AUC of testing Data Set : 0.856

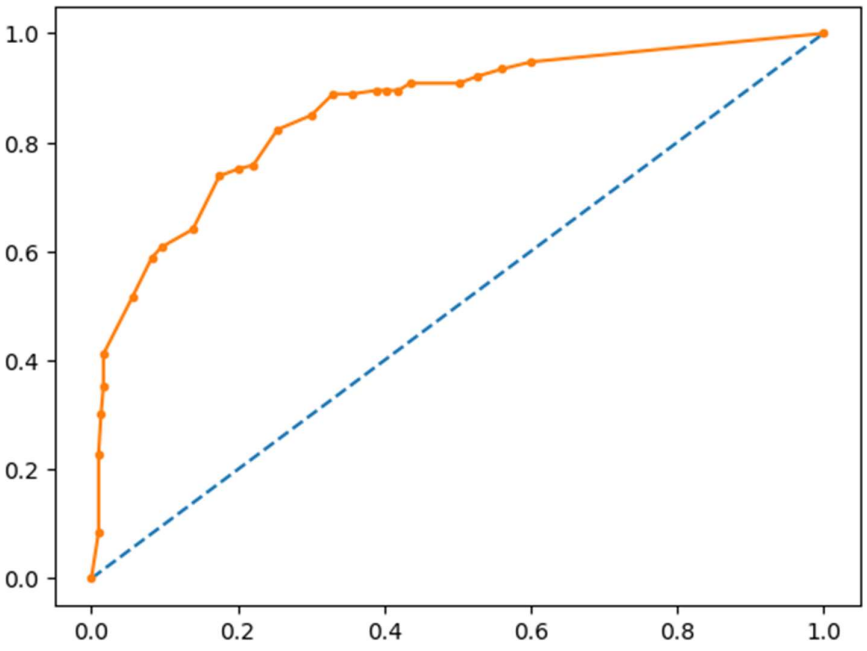

Figure 47: AUC of testing Data Set

Random Forest: -

Classification Report of Training Dataset

	precision	recall	f1-score	support
0	0.94	0.87	0.90	813
1	0.66	0.81	0.73	248
accuracy			0.86	1061
macro avg	0.80	0.84	0.82	1061
weighted avg	0.87	0.86	0.86	1061

Classification Report of Testing Dataset

	precision	recall	f1-score	support
0	0.92	0.82	0.87	340
1	0.61	0.80	0.69	116
accuracy			0.82	456
macro avg	0.77	0.81	0.78	456
weighted avg	0.84	0.82	0.83	456

Figure 48: Classification Report of Training & Testing Dataset

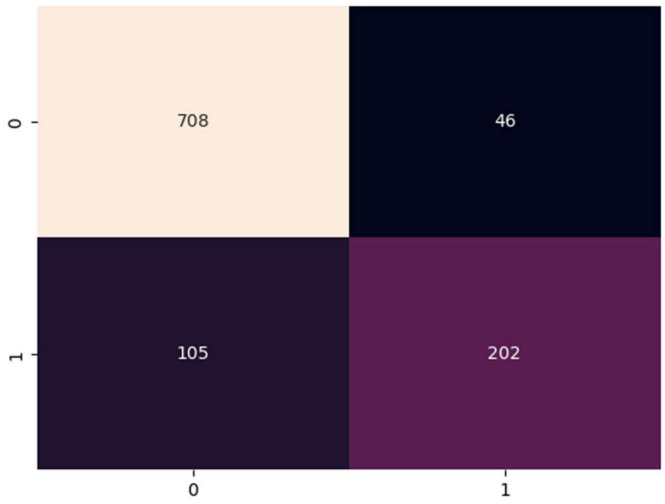

Figure 49: Confusion Matrix of training Data Set

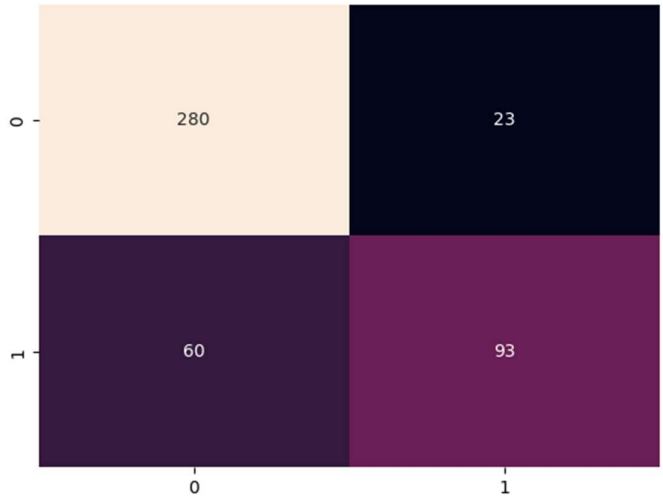

Figure 50: Confusion Matrix of testing Data Set

Inference from Train data:

- 708 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour
- 202 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 105 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 46 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0).

Inference from Test data:

- 280 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Labour is predicted as labour

- 93 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; Vote to Conservative is predicted as conservative.
- 60 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 23 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

AUC of training Data Set : 0.918

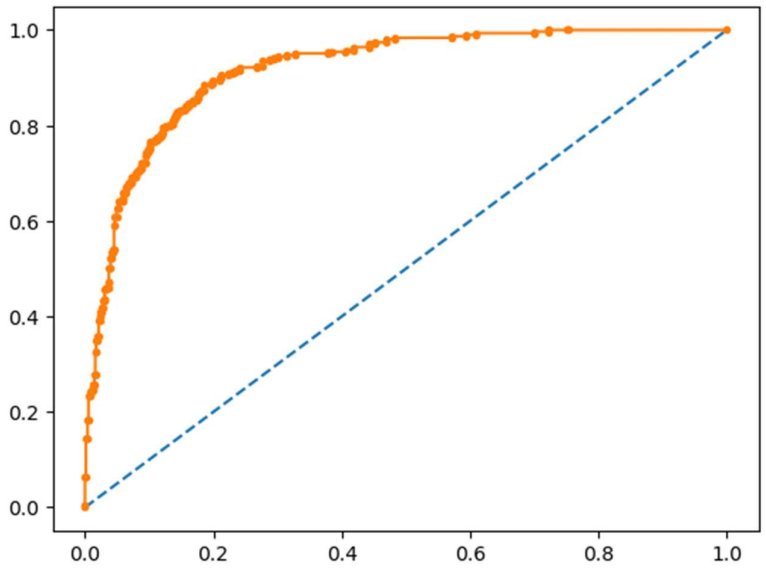

Figure 51: AUC of training Data Set

AUC of testing Data Set : 0.891

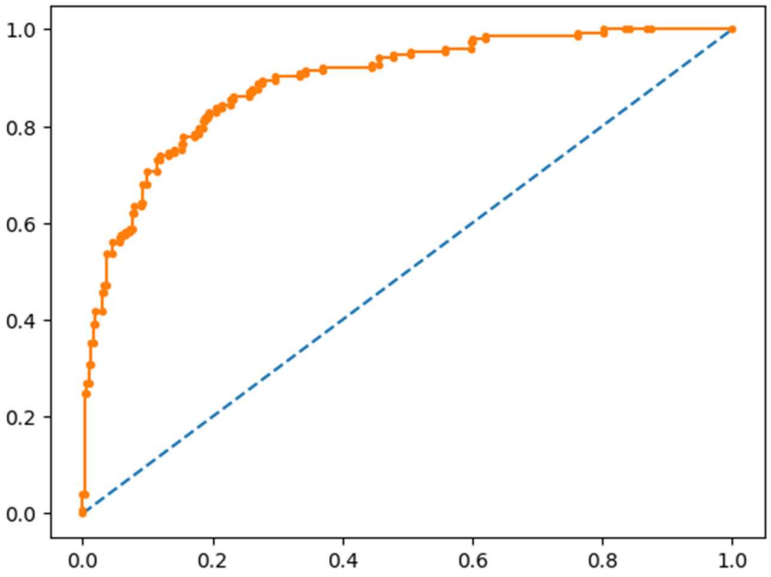

Figure 52: AUC of testing Data Set

Model Comparison:

Logistic regression, LDA, KNN & Naïve Bayes models are thoroughly explained in the before sections. We are here to compare the all 4 models and identify which make more sense with respect you predicting dependent variable (Vote).

		AUC	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
Logistic Regression	**Train**	0.890	0.86	0.74	0.91	0.64	0.88	0.69
	Test	0.883	0.87	0.76	0.88	0.74	0.88	0.75
LDA	**Train**	0.889	0.91	0.65	0.86	0.74	0.89	0.69
	Test	0.888	0.89	0.73	0.86	0.77	0.88	0.74
KNN	**Train**	0.924	0.93	0.68	0.88	0.79	0.9	0.73
	Test	0.861	0.91	0.66	0.84	0.78	0.87	0.72
Navies Bayes	**Train**	0.888	0.9	0.69	0.88	0.73	0.89	0.71
	Test	0.876	0.87	0.73	0.87	0.74	0.87	0.73
Bagging & Boosting	**Train**	0.902	0.93	0.61	0.85	0.78	0.89	0.68
	Test	0.884	0.89	0.64	0.83	0.75	0.86	0.69
Design Tree	**Train**	0.907	0.87	0.76	0.9	0.71	0.89	0.73
	Test	0.856	0.83	0.74	0.86	0.68	0.84	0.71
Random Forest	**Train**	0.918	0.94	0.66	0.87	0.81	0.9	0.73
	Test	0.891	0.92	0.61	0.82	0.8	0.87	0.69

Table 1 : Model Comparison Chart

Table 1 helps us to understand how each models came out with the important component like AUC, Accuracy, precision, recall, f1-score. Logistic regression, LDA, Navies Bayes, Bagging & Boosting, Design Tree & Random Forest performed on a same level predicting the dependent variable, but when it is compared with KNN it shows lesser performance in both train and test data.

KNN performed well than other models.

- KNN has highest values in most of the criteria
- Highest Accuracy score 0.924
- Top score 0.93 in precision 0
- Top score 0.88 in recall 0
- Top score 0.90 in f1-score 10

Observations

Comparing all the performance measure, Naïve Bayes model from second iteration is performing best. Although there are some other models such as SVM and Extreme Boosting which is performing almost same as that of Naïve Bayes. But Naïve Bayes model is very consistent when train and test results are compared with each other. Along with other parameters such as Recall value, AUC_SCORE and AUC_ROC_Curve, those results were pretty good is this model.

- Labour party is performing better than Conservative from huge margin.
- Female voters turn out is greater than the male voters.
- Those who have better national economic conditions are preferring to vote for Labour party.

- Persons having higher Eurosceptic sentiments conservative party are preferring to vote for Conservative party.
- Those who have higher political knowledge have voted for Conservative party
- Looking at the assessment for both the leaders, Labour Leader is performing well as he has got better ratings in assessment.

Problem 2: -

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

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

Solution:-

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

Figure 53: Total no. of words in the speeches

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

Figure 54: Total no. of characters including blank spaces

	Speech	avg_word
0	On each national day of inauguration since 178...	4.783825
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	4.626100
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	4.713397

Figure 55: Average word length

- Franklin D. Roosevelt has 68 sentences in his speech.
- John F. Kennedy has 52 sentenses in his speech.
- Richard Nixon has 68 sentences in his speech.
- President Franklin D. Roosevelt's speech have 7671 Characters (including spaces) and 1323 words.
- President John F. Kennedy's Speech have 7673 Characters (including spaces) and 1364 words.

- President Richard Nixon's Speech have 10106 Characters (including spaces) and 1769 words.

	Speech	stopwords
0	On each national day of inauguration since 178...	632
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	618
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	899

Figure 56: Total no. of stopwords

	Speech	numerics
0	On each national day of inauguration since 178...	14
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	7
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10

Figure 57: Total no. of numbers

	Speech	UpperCase
0	On each national day of inauguration since 178...	1
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	5
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	13

Figure 58: Total no. of Uppercase words

	Speech	upper_letter
0	On each national day of inauguration since 178...	119
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	94
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	132

Figure 59: Total no. of uppercase letters

```
--        22
It        12
know       9
us         8
life       6
The        6
freedom    5
years      5
speaks     5
people     5
Name: count, dtype: int64
```

```
--        24
us        11
let        8
sides      7
new        7
pledge     7
shall      5
ask        5
I          5
fellow     4
Name: count, dtype: int64
```

```
us        25
--        17
new       15
I         12
peace     11
let        9
America    9
great      9
better     7
policies   7
Name: count, dtype: int64
```

Figure 60: Common words in Franklin D. Roosevelt, John F. Kennedy & Richard Nixon Speech

After Removing additional stop words,

	Name	Speech	Totalwords	char_count	avg_word	stopwords	numerics	UpperCase	upper_letter	Processed_Speech
0	Roosevelt	On each national day of inauguration since 178...	1323	7651	4.783825	632	14	1	119	On national day inauguration since 1789, peopl...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364	7673	4.626100	618	7	5	94	Vice President Johnson, Mr. Speaker, Mr. Chief...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769	10106	4.713397	899	10	13	132	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

Figure 61: Speeches after removing stop words

Most frequently used words from President Franklin D. Roosevelt's speech are

- Know
- The
- Life

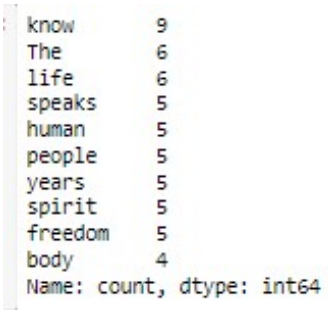

Figure 62: words in Franklin D. Roosevelt Speech

Most frequently used words from President John F. Kennedy's speech are

- Let
- new
- sides

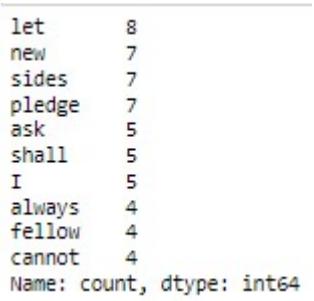

Figure 63: words in John F. Kennedy Speech

Most frequently used words from President Richard Nixon's speech are

- new
- I
- Peace

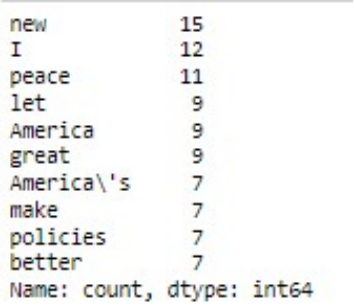

Figure 64: words in Richard Nixon Speech

Word Cloud: -

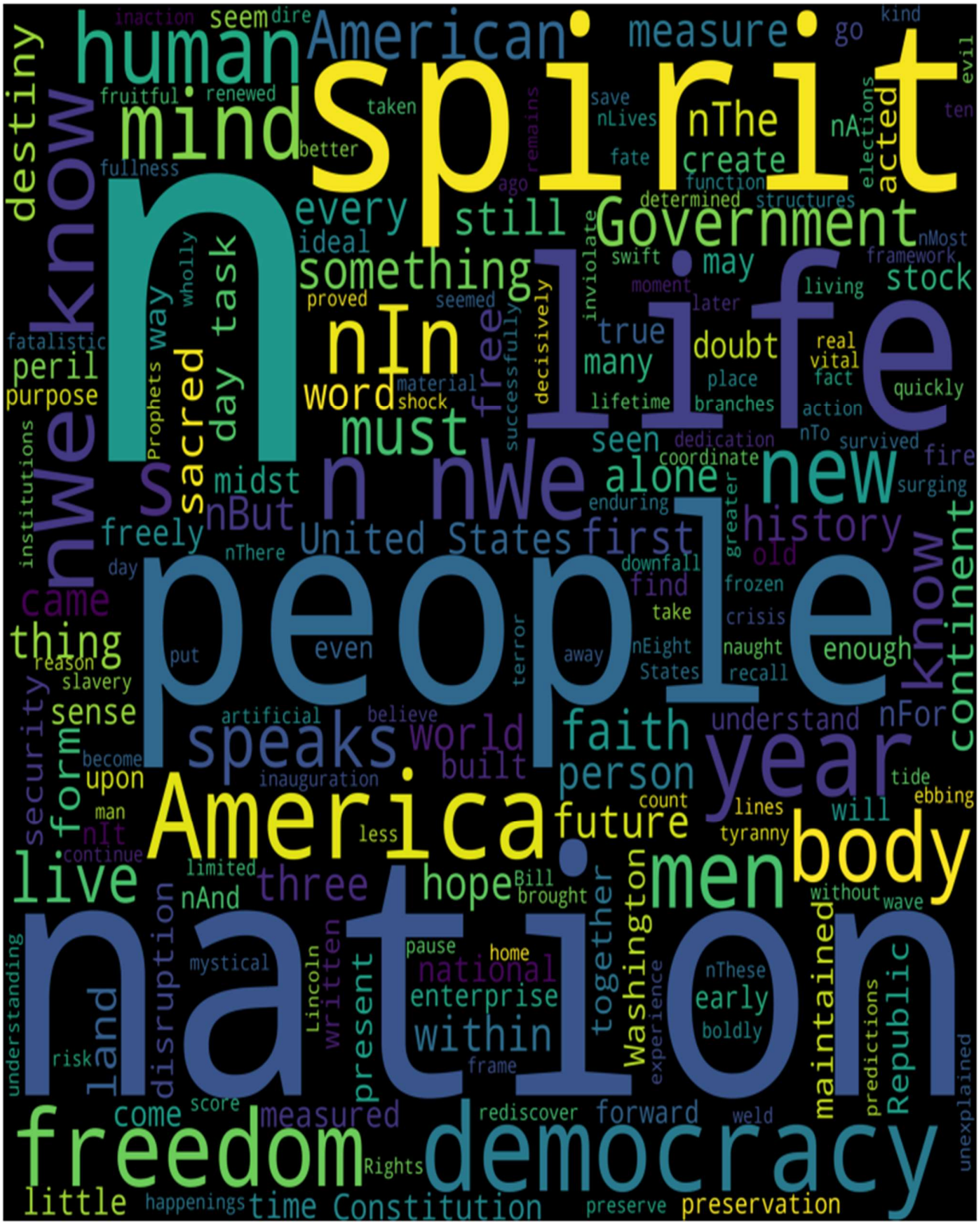

Figure 65: Word Cloud for President Franklin D. Roosevelt's speech (after cleaning)!!

Figure 66: Word Cloud for President John F. Kennedy's Speech (after cleaning)!!

Figure 67: Word Cloud for President Richard Nixon's Speech (after cleaning)!!