MACHINE LEARNING ASSIGNMENT

BY JASPER SHELDON M

Contents

CLASSIFICATION ON ELECTION SURVEY

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. DATA DESCRIPTION

Variable Name	Description
1. vote	Conservative or Labour (Party)
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 represents 'Eurosceptic' sentiment.
8. Political. Knowledge	Knowledge of parties' positions on European integration, 0 to 3.
9. gender	Female or male.

Table 1 Data Dictionary

2. DATA SAMPLE

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender
1	Labour	43	3	3	4	1	2	2	female
2	Labour	36	4	4	4	4	5	2	male
3	Labour	35	4	4	5	2	3	2	male
4	Labour	24	4	2	2	1	4	0	female
5	Labour	41	2	2	1	1	6	2	male

Table 2 Sample of the data

3. EXPLORATARY DATA ANALYSIS

1. ATTRIBUTES OF THE DATA:

```
Int64Index: 1525 entries, 1 to 1525
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic_cond_national  1525 non-null   int64
 3   economic_cond_household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political_knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
```

Table 3 Information of the data

The dataset has 1525 rows and 9 columns including the dependent variable from the above table we can infer that there are 7 numerical columns and 2 object columns. No missing values present in the dataset

Numerical columns: AGE, ECONOMIC_COND_NATIONAL, ECONOMIC_COND_HOUSEHOLD, BLAIR, HAGUE, EUROPE, POLITICAL_KNOWLEDGE

Categorical columns (ordinal): ECONOMIC_COND_NATIONAL, ECONOMIC_COND_HOUSEHOLD, BLAIR, HAGUE, EUROPE, POLITICAL_KNOWLEDGE

Object columns: VOTE, GENDER

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
mean	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315
min	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000
50%	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000
75%	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000
max	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000

Table 4 Summary of the data

OBSERVATIONS:

Column name consist of dot in between which is converted to hyphen for easy access

Age: It has a mean of 54, and standard deviation of 15 which indicates its spread wide. Minimum value is 24 and the maximum value is 93. Q3 has the value of 67 that indicates many old people haven't skipped the election.

Economic_cond_national: It has a mean of 3.2 and standard deviation of 0.9 which is close to 1. Maximum is 5 and minimum is 1. 50% of the data is above 3.

Economic_cond_household: Similar to economic_national it has mean of 3.1 and standard deviation of 0.9 which shows a minimal spread. Minimum value is 1 and maximum of 5, 50% the data has a value more than 4.

Blair: It has a mean of 3.3 and standard deviation of 1.1 and 50% the data has more than 4 the maximum value is 5. It has 5 levels.

Hague: Similar to Blair it has 5 levels, maximum is 5 minimum is 1 but standard deviation is 1.2.

Europe: It has a mean of 6.7 and standard deviation of 3.29 maximum is 11, it has 11 levels 50% of the data consists of 6.

Political Knowledge: It has mean of 1.54 and standard deviation of 1.1 indicating a narrow spread, has 3 levels, minimum is 1 and maximum is 3.75% of the data has the value 2.

Key Findings:

- It has no missing values or wrong entries in the dataset
- 8 duplicated records were present in the dataset; those records were removed before further analysis.

2. UNIVARIATE ANALYSIS

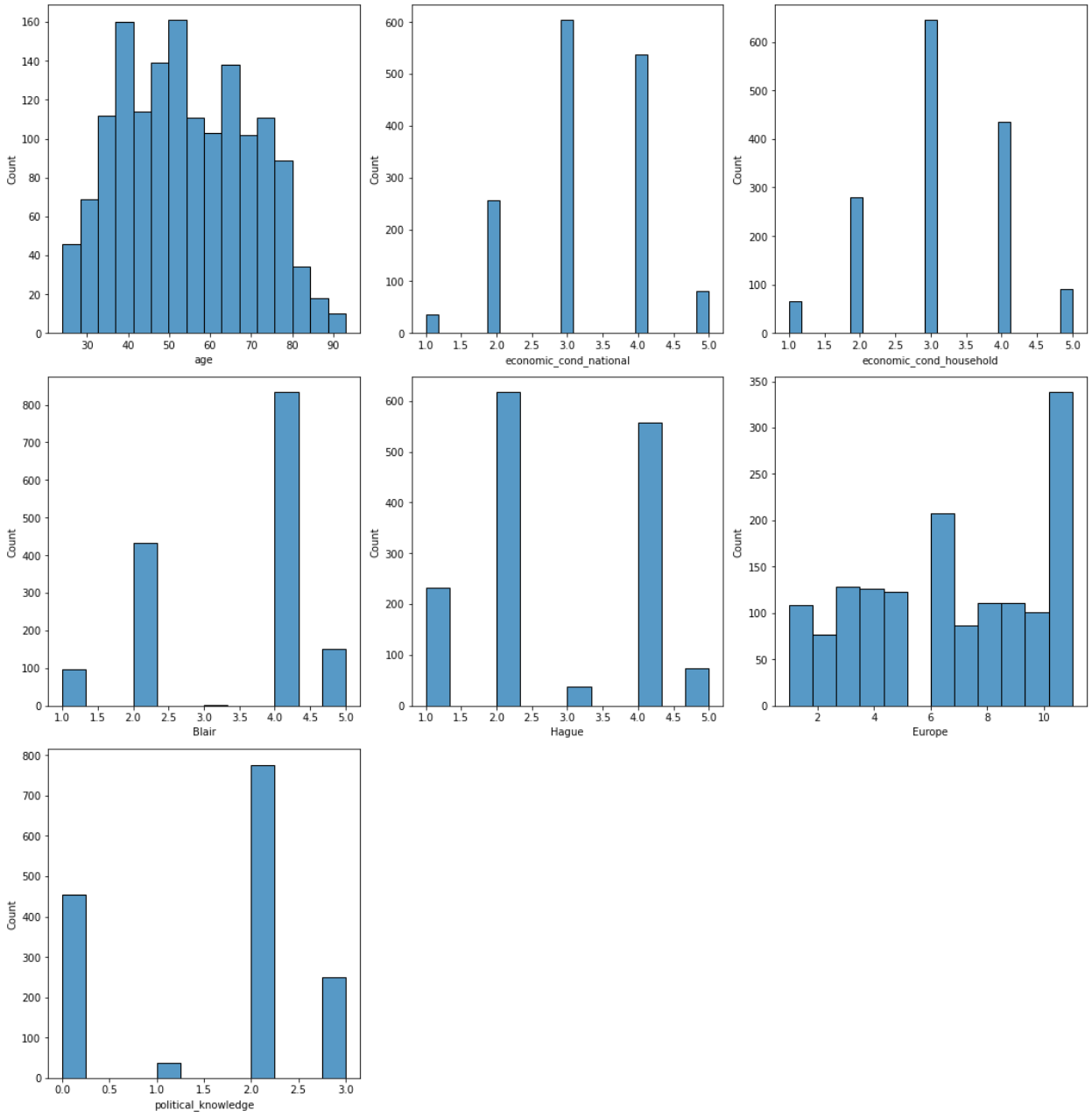


Figure 1 Histogram of the data

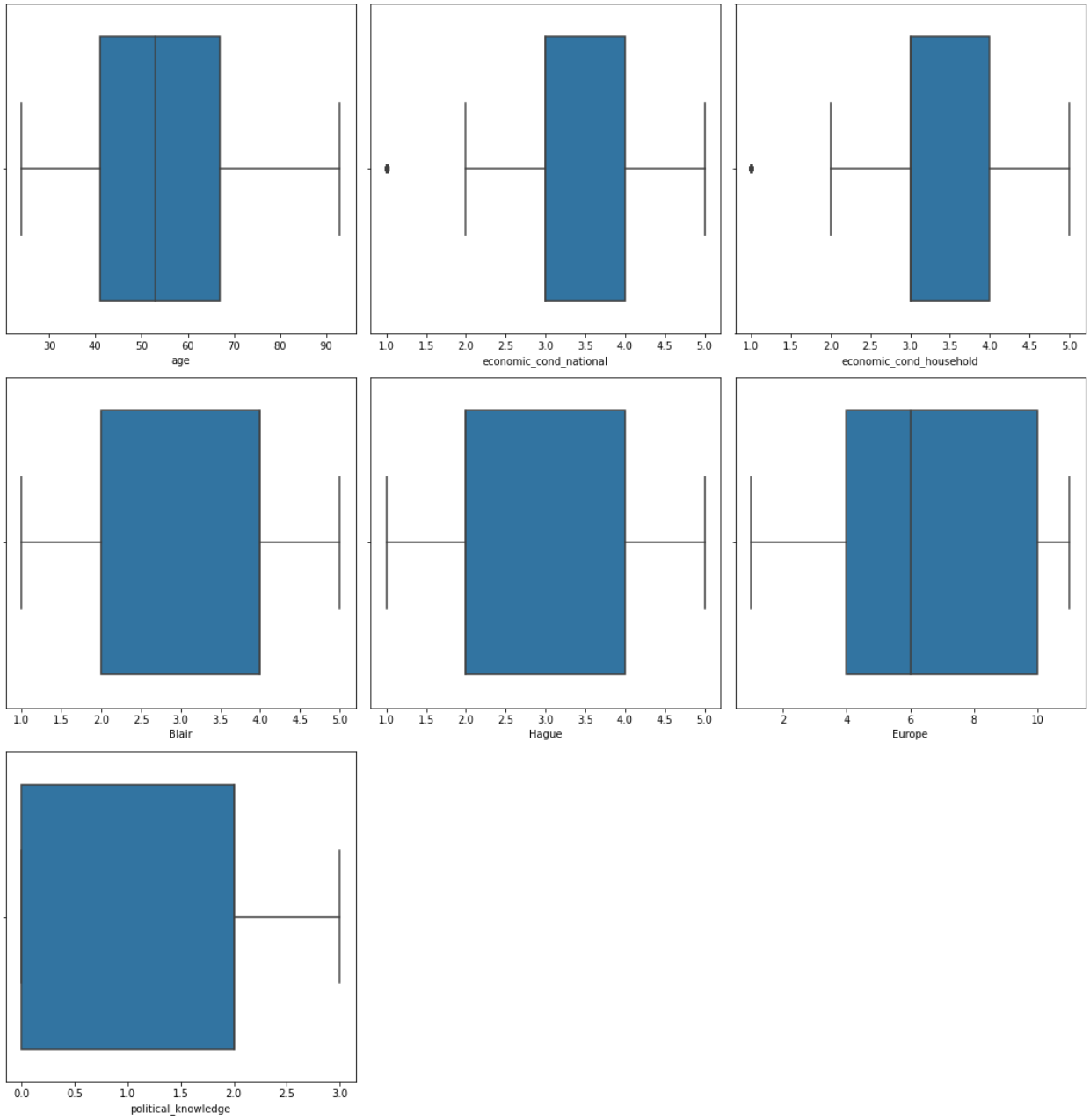


Figure 2 Boxplot of the data

	Lower_range	Upper range
age	0	0
economic_cond_national	37	0
economic_cond_household	65	0
Blair	0	0
Hague	0	0
Europe	0	0
political_knowledge	0	0

Figure 3 Outlier summary

```
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 5 Skewness of the data

OBSERVATIONS:

Age: It has peak value around 50 we can infer this from the histogram, it does not have any outliers and its right skewed with the value of 0.13.

Economic cond national: It has most frequent value as 3 and second most frequent value as 4 it has 37 outliers in the lower range and it left skewed with the value of- 0.23

Economic cond household: It has the most frequent value as 3 and second most value as 4, it has 65 outliers below the lower range and its left skewed with the value of -0.14.

Blair: it has most frequent value as 4 and the second most frequent value as 2, there are very few records present with the value 3.No outliers present and it's highly left skewed with the value of -0.53.

Hague: Contradicting to the Blair column it has the most frequent value as 2 and second most frequent value as 4. No outliers present and its right skewed with the value of 0.14

Europe: It has most frequent value as 11, and second frequent value as 6, there are no values present for the value 5. No outliers are present and left skewed with the value of-0.14

Political knowledge: Most frequent value is 2 and the second most frequent value is 0, it has no outliers and highly left skewed with the value of -0.42

3. BIVARIATE ANALYSIS:

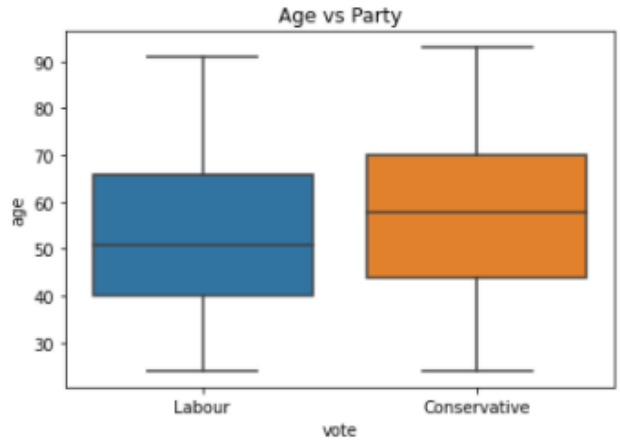


Figure 4 Boxplot of Age vs. Party

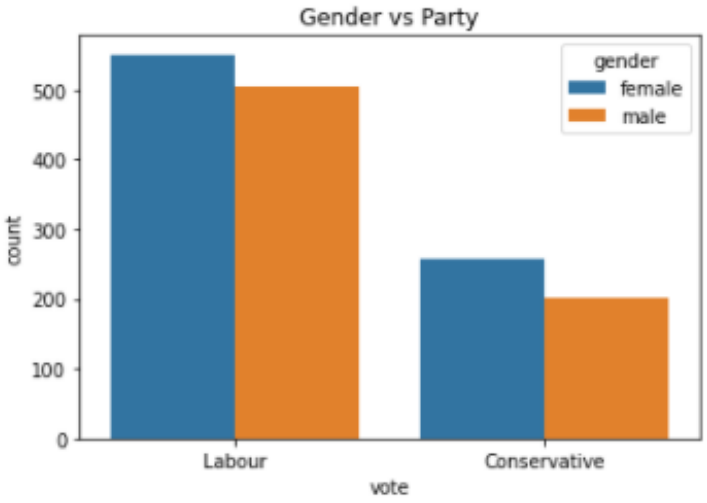


Figure 5 Count Plot of Gender vs. Party

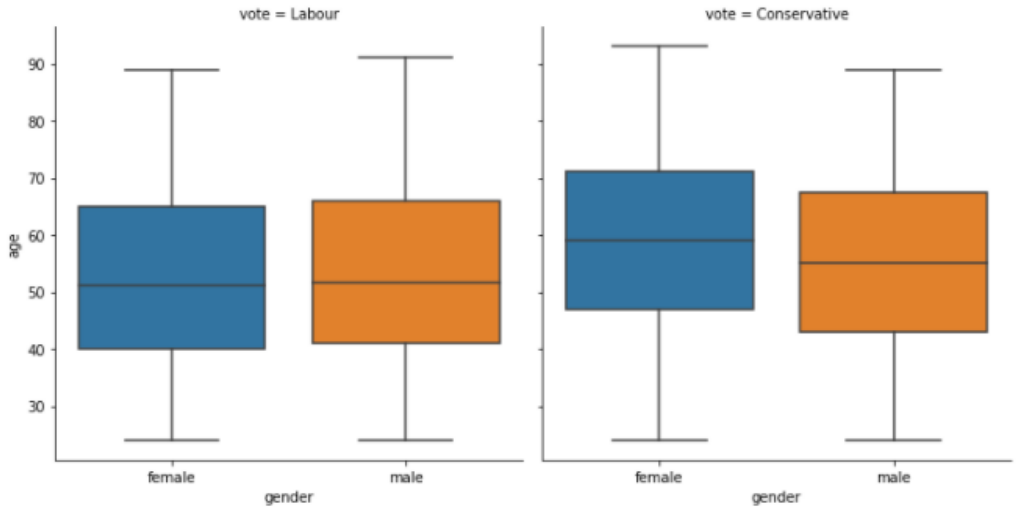


Figure 6 Boxplot of Age on Gender diff. by Party

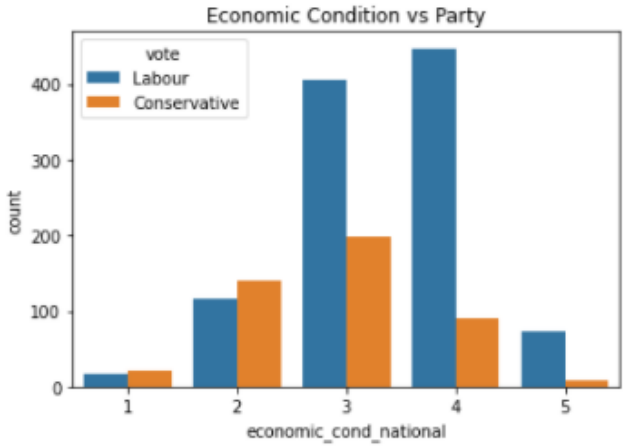


Figure 7 Count of National economic condition on Party

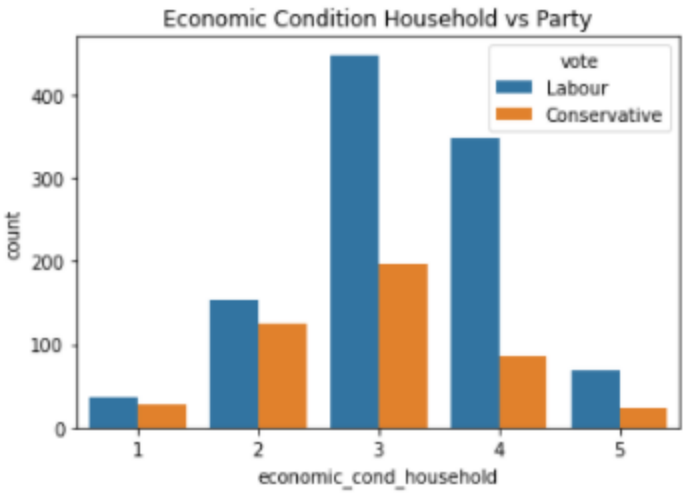


Figure 8 Count of Household economic condition on Party

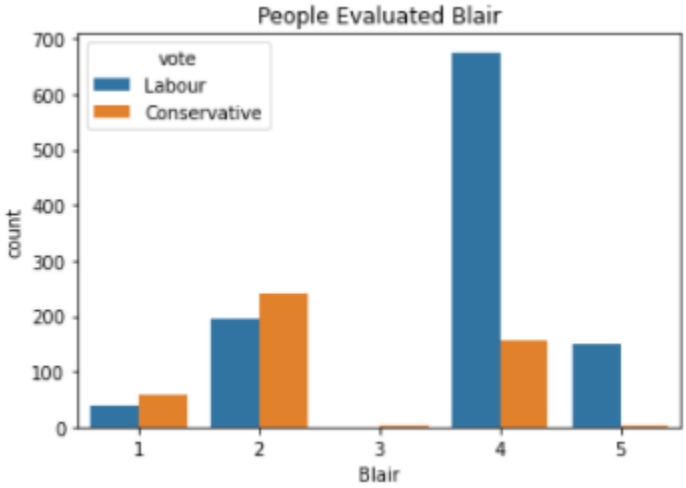


Figure 9 Count Plot of people evaluated Blair

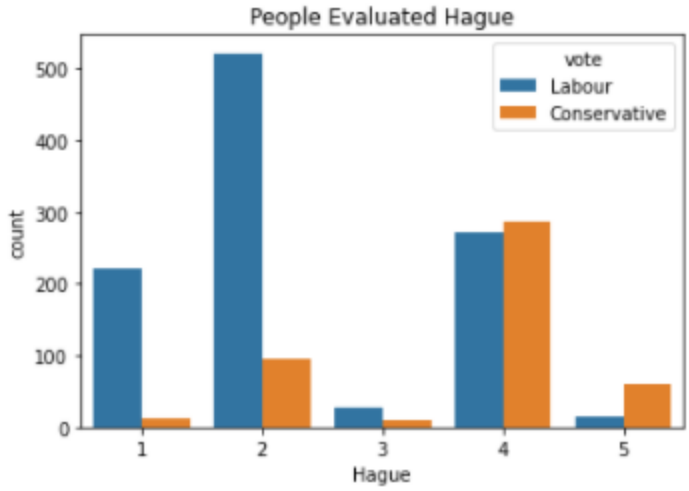


Figure 10 Count Plot of people evaluated Hague

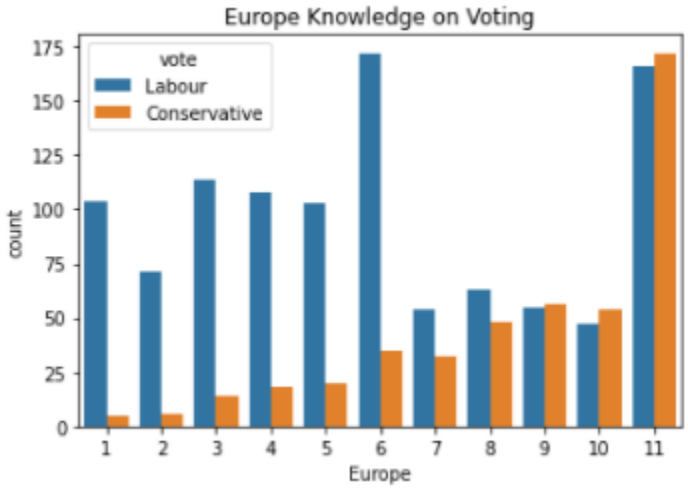


Figure 11 Count Plot of Europe knowledge on Party

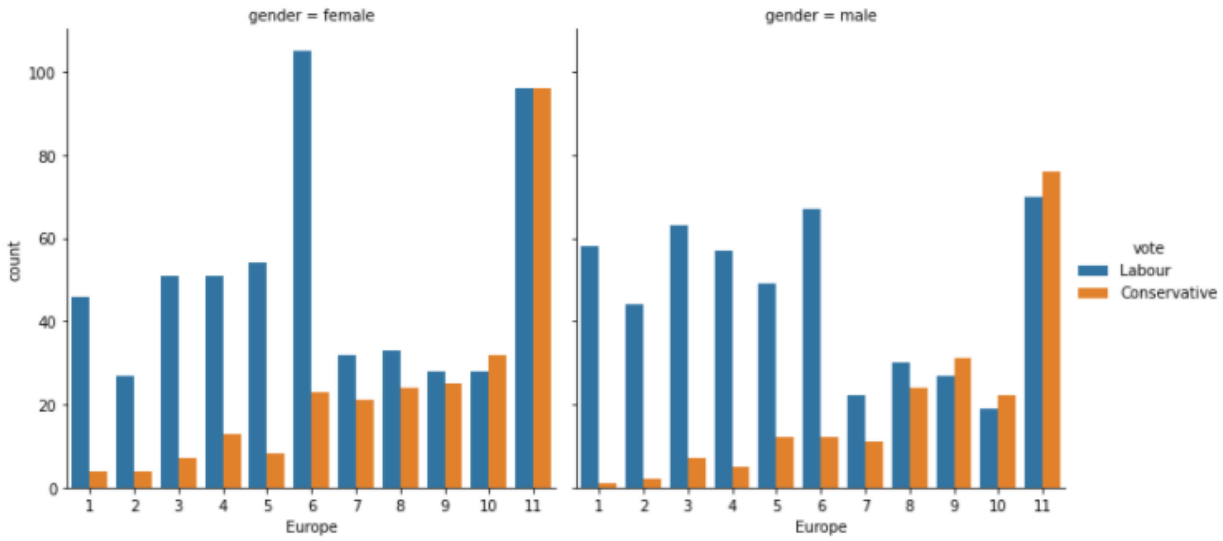


Figure 12 Europe knowledge on Party diff. Gender

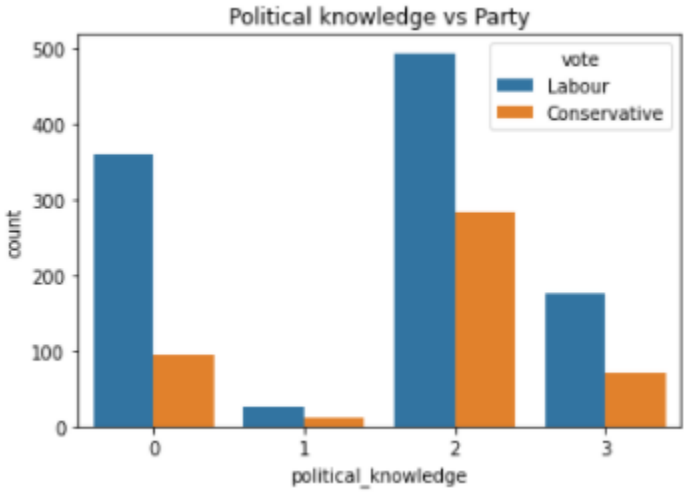


Figure 13 Political knowledge on Party

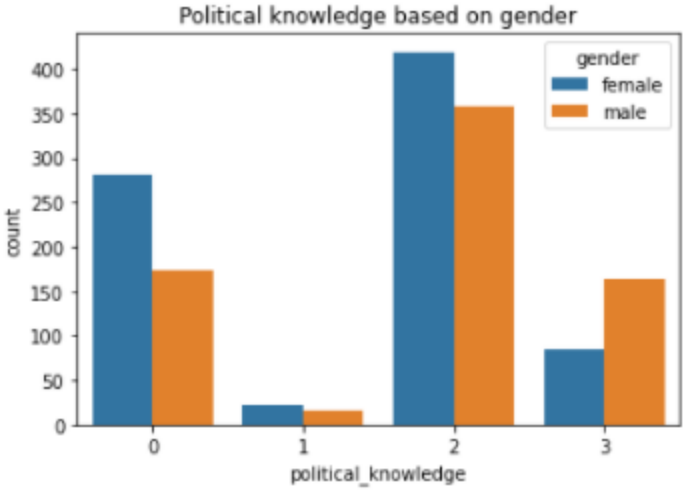


Figure 14 Count of People based on Political knowledge diff. Gender

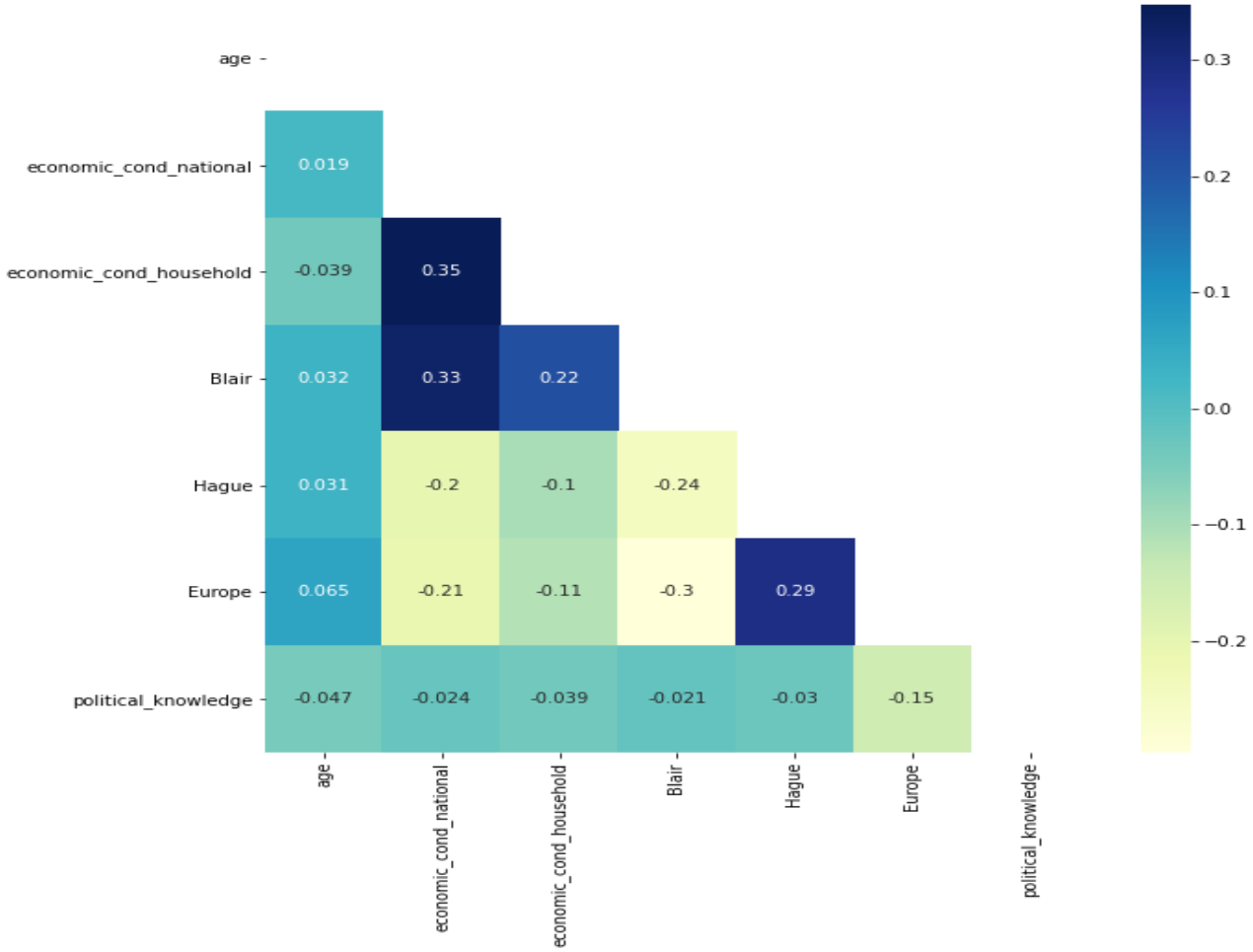


Figure 15 Heat map of Correlation

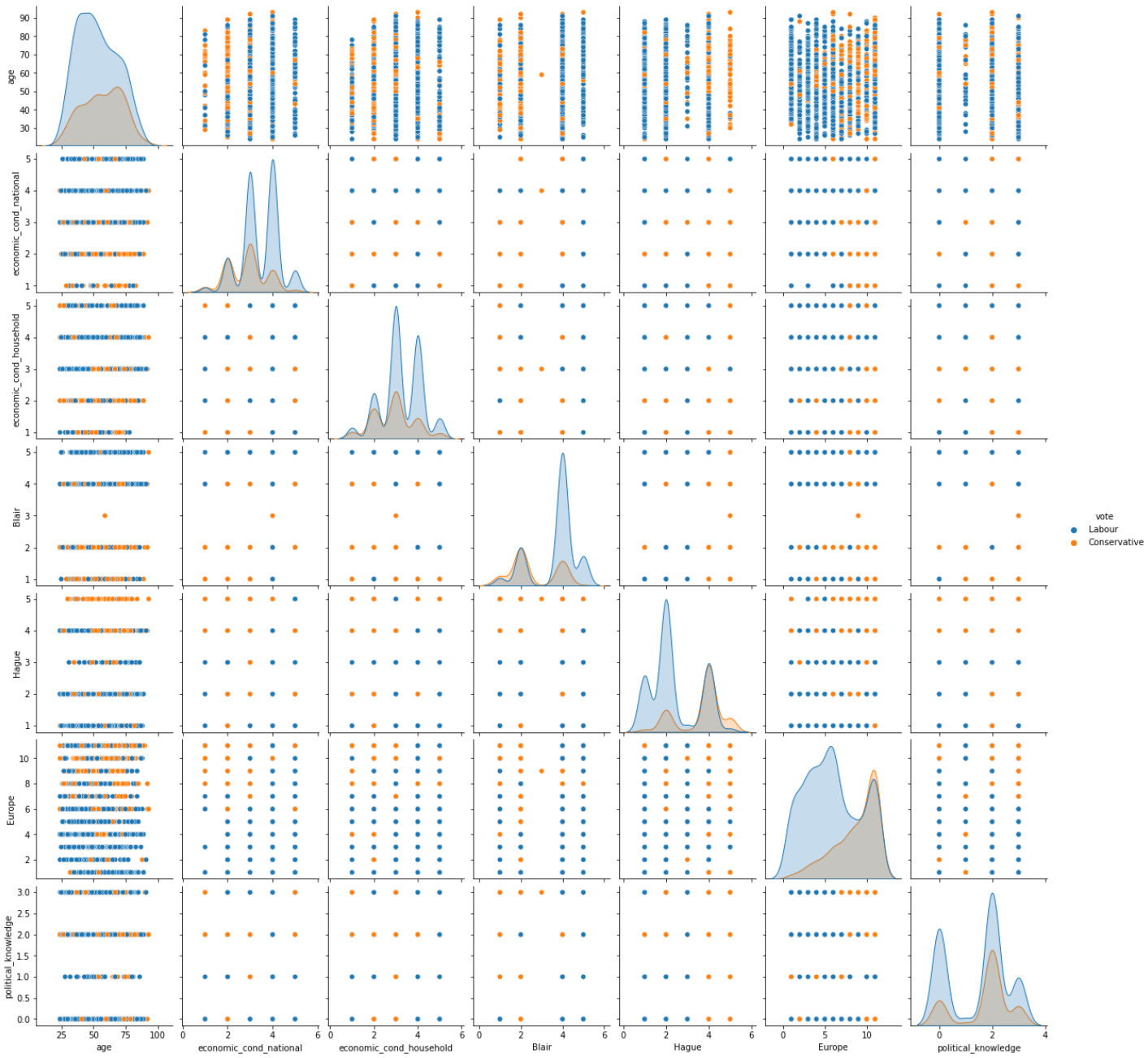


Figure 16 Pair plot

OBSERVATION:

- From Fig 4 we can see that people who vote for Conservative are comparatively older than people who vote for Labour Party
- The majority of voters in the sample are women; with respect to each parties the votes are casted by more women than men.
- From fig6 its evident that older women prefer Conservative than Labour party

- From fig 7 we can see the economic condition of people above 3 are preferring Labour than Conservative, whereas people with less economic condition are choosing Conservative.
- Doesn't show any significant insight from Household economic condition, people from all background are voting for Labour Party
- From fig 9 we can see people evaluating Blair given 4 and voted for the same party, and people who evaluated him and gave 2 chose the Conservative party which is obvious.
- Fig 10 more less indicate the previous insight , who evaluated Hague less have chosen Labour but significant insight is people who have evaluated Hague as 4 also voted for Labour Party equally which is suspicious further analysis is required to answer the question why.
- From fig 12, we can see people with high Eurosceptic sentiment are preferring Conservative party and people with low Eurosceptic sentiment are choosing Labour Party.
- From Fig 13 majority of the people voting for the Conservative party have a better political knowledge. People voting for Labour party have either moderate political knowledge or no political knowledge at all.
- Proportion of people having a better political knowledge are men(From fig 14)
- Fig 15. There is very weak correlation existing between the variables, the values range from -0.2 to 0.35. Maximum positive correlation is between national economic condition and household economic condition. The most negative correlation is between Blair and Europe. Hague as some correlation with Europe, Blair is positively correlated with national and household economic condition which explains the difference in the votes.
- From the pair plot we can see most of the scatter plot show random cloud format because of ordinal variables Along the diagonals we can see most of the variables doesn't show significant margin to separate the classes. In all the variables classes seem to overlap each other except for Europe. So this might be a good separator compared to the rest of the independent variables.

4. CHI-SQUARE TEST

Chi square test is done to find whether the categorical column has significant influence on the target variable so that we can choose them for model building.

H_0: Target Variable is independent of Categorical column

H_a: Target Variable is dependent of Categorical column

The P values after performing chi square test on 6 categorical columns are given below.

ECONOMIC_COND_NATIONAL= 1.18e-30

ECONOMIC_COND_HOUSEHOLD = 8.70E-12

BLAIR = 5.30E-60

HAGUE = 6.93E-73

EUROPE = 5.65e-47

POLITICAL_KNOWLEDGE = 1.96e-07

In all the cases the p value is less than 0.05 so we can reject the null hypothesis and accept the alternative hypothesis which indicates all the variables affect the target variable. So we can use them for predictions in the model.

4. DATA ENCODING:

From the information table above we can see we have 6 ordinal columns and 2 textual data and 1 numeric column. Since the categorical columns are ordinal in nature they don't need any encoding only Gender is binary encoded and the target variable (vote) and age are kept as such.

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender
1	Labour	43	3	3	4	1	2	2	0
2	Labour	36	4	4	4	4	5	2	1
3	Labour	35	4	4	5	2	3	2	1
4	Labour	24	4	2	2	1	4	0	0
5	Labour	41	2	2	1	1	6	2	1

Table 6 Sample of the encoded data

5. SCALING:

Of all the algorithms, algorithms which are distance based will only be affected by scaling. KNN classifier is going to be tested for prediction so for that model alone scaled version of the data is used. Scaling is necessary in this case as discussed in the summary above age is widely spread and lies in the range of 40 to 90 but the other columns lies in the range of 1-5 or 1-11 max, so if we bring them down to common scale KNN algorithm will work better. Scaling is mainly done to eliminate the weightage column might carry with the higher values. In case of other models they use Decision tree as base model so its performance won't be affected by scaling so it's not necessary.

6. DATA SPLITING:

Using the sklearn.train_test_split method the data is divided into 70:30 split, i.e. 70 % of the data is used for training purpose and 30 % of the date is used for testing purpose. 79: 30 is chosen because we need ample amount of data to train the model and equal proportion of it to check the model on unseen data.

7. MODEL DEVELOPMENT:

1. LOGISTIC CLASSIFIER (BASE MODEL):

For Train data:

SOLVER: newton-cg

Accuracy score : 0.8284637134778511

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

Area Under the curve : 0.8770371241983879

Table 7 Classification report of Logistic clf

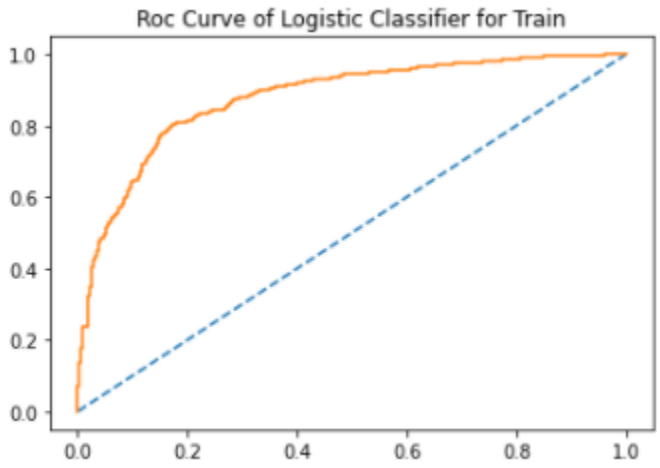


Figure 17 Roc curve

For test data:

Accuracy score : 0.8552631578947368

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

Area Under the curve : 0.9128611794731565

Table 8 Classification report Logistic clf

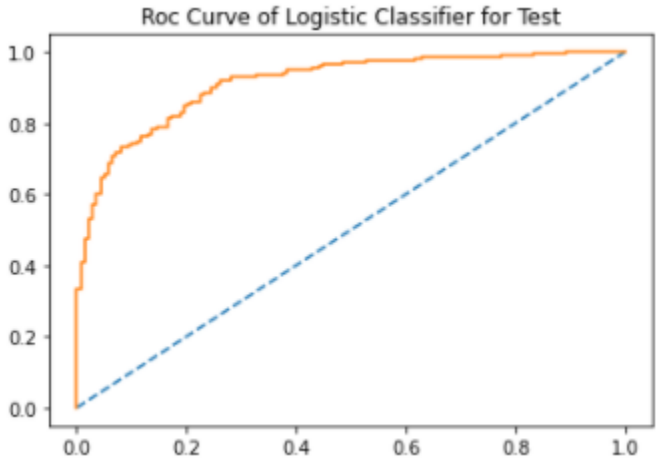


Figure 18 ROC curve

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.86 86% of the predictions regarding done to Labour is correct.

- Recall is 0.90, 90% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.74, 74% of the predictions are correct.
  - Recall is 0.66, 66% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.87 87% of the predictions regarding done to Labour is correct.
  - Recall is 0.93, 93% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.81, 81% of the predictions are correct.
  - Recall is 0.68, 68% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.85 which is good score.
- Area Under the cure value is 0.91 which indicates the model is performing better.

.

2. LDA (BASE MODEL):

For Training data:

Accuracy score : 0.822808671065033

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

Area Under the curve : 0.876932063641483

Table 9 Classification report LDA

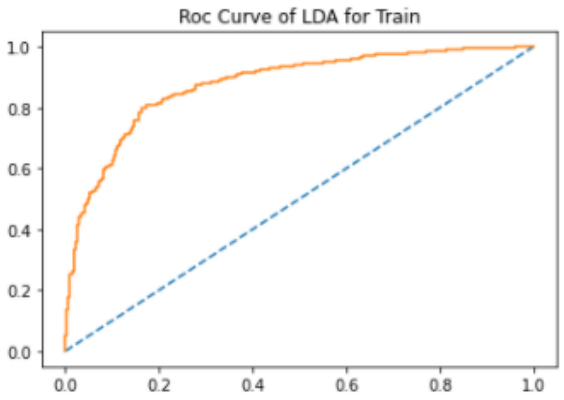


Figure 19 ROC curve

For Test Data:

Accuracy score : 0.8530701754385965

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

Area Under the curve : 0.9143651444717892

Table 10 Classification report

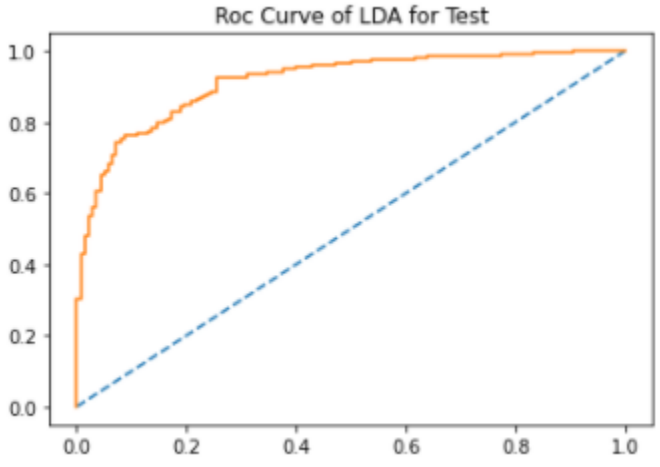


Figure 20 ROC curve

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
    - Precision is 0.86 86% of the predictions regarding done to Labour is correct.
    - Recall is 0.89, 89% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
    - Precision is 0.72, 72% of the predictions are correct.
    - Recall is 0.67, 67% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
    - Precision is 0.87 87% of the predictions regarding done to Labour is correct.
    - Recall is 0.92, 92% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
    - Precision is 0.80, 80% of the predictions are correct.
    - Recall is 0.69, 69% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.85 which is good score.
- Area Under the cure value is 0.91 which indicates the model is performing better.

3. NAÏVE BAYERS (BASE MODEL):

For Training Data:

Accuracy score : 0.8199811498586239

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

Area Under the curve : 0.8731624908597315

Table 11 Classification report Naive Bayer's

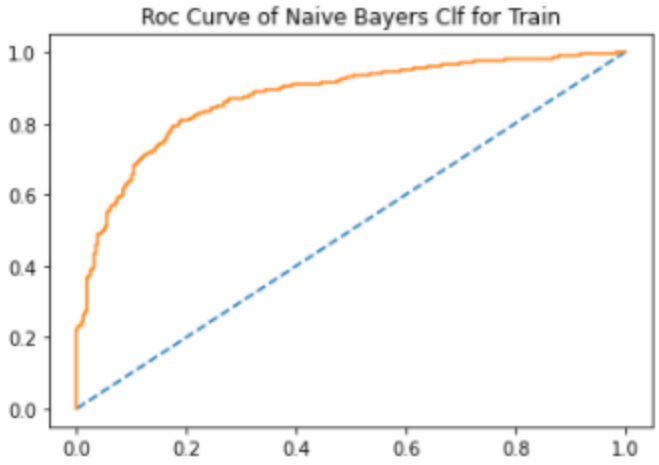


Figure 21 ROC curve of Naive Bayer's

For Test Data:

Accuracy score : 0.8574561403508771

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

Area Under the curve : 0.9124965818977304

Table 12 Classification Report of Naive Bayer's

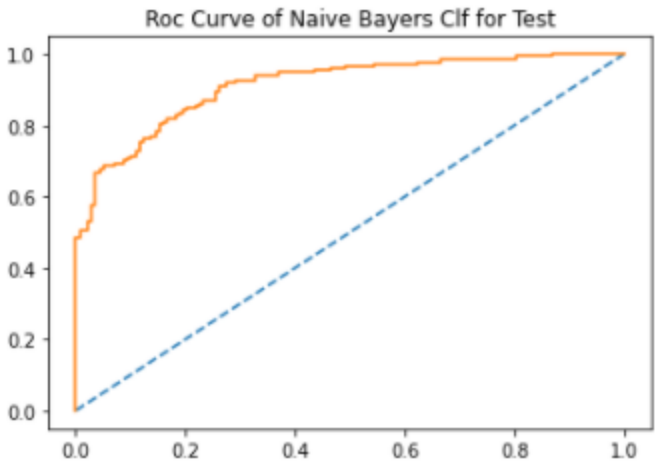


Figure 22 ROC curve of Naïve Bayer's

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.87 87% of the predictions regarding done to Labour is correct.
  - Recall is 0.87, 87% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.70, 70% of the predictions are correct.
  - Recall is 0.70, 70% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.88 88% of the predictions regarding done to Labour is correct.
  - Recall is 0.92, 92% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.79, 79% of the predictions are correct.
  - Recall is 0.72, 72% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.85 which is good score.

Area Under the cure value is 0.91 which indicates the model is performing better.

4. KNN (BASE MODEL):

For Train data:

Accuracy score : 0.8576814326107446

	precision	recall	f1-score	support
Conservative	0.79	0.73	0.76	322
Labour	0.89	0.91	0.90	739
accuracy			0.86	1061
macro avg	0.84	0.82	0.83	1061
weighted avg	0.86	0.86	0.86	1061

Area Under the curve : 0.923957168912161

Table 13 Classification Report KNN

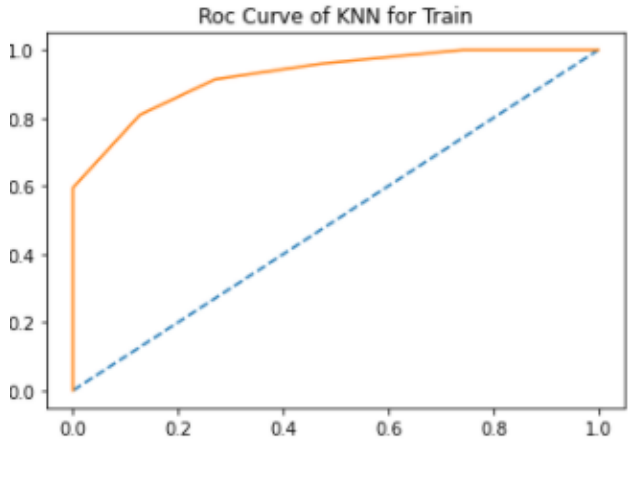


Figure 23 ROC curve of KNN

For Test Data:

```
Accuracy score : 0.8530701754385965
              precision    recall  f1-score   support

Conservative       0.77      0.74      0.75       138
      Labour       0.89      0.90      0.90       318

    accuracy                           0.85       456
   macro avg       0.83      0.82      0.82       456
weighted avg       0.85      0.85      0.85       456



Area Under the curve : 0.8733365235621184
```

Table 14 Classification report KNN

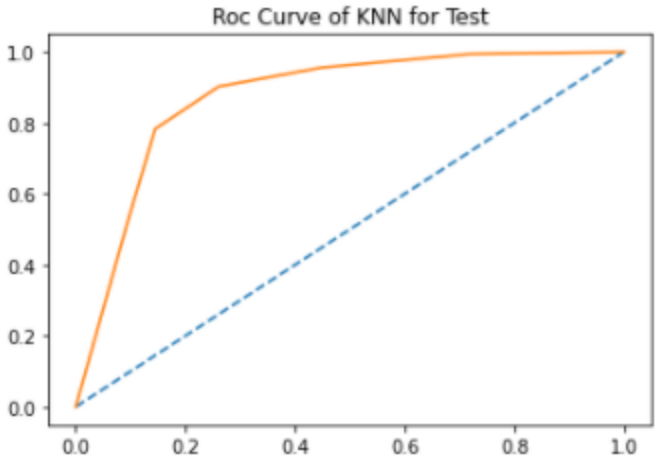


Figure 24 ROC curve of KNN

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.89 89% of the predictions regarding done to Labour is correct.
  - Recall is 0.91, 91% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.79, 79% of the predictions are correct.
  - Recall is 0.73, 73% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.89 89% of the predictions regarding done to Labour is correct.

- Recall is 0.90, 90% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.77, 77% of the predictions are correct.
  - Recall is 0.74, 74% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.85 which is good score.

Area Under the cure value is 0.87 which indicates the model is performing better.

5. RANDOM FOREST (BASE MODEL):

For Train Data:

Accuracy score : 1.0

	precision	recall	f1-score	support
Conservative	1.00	1.00	1.00	322
Labour	1.00	1.00	1.00	739
accuracy			1.00	1061
macro avg	1.00	1.00	1.00	1061
weighted avg	1.00	1.00	1.00	1061

Area Under the curve : 1.0

Table 15 Classification report on Random forest

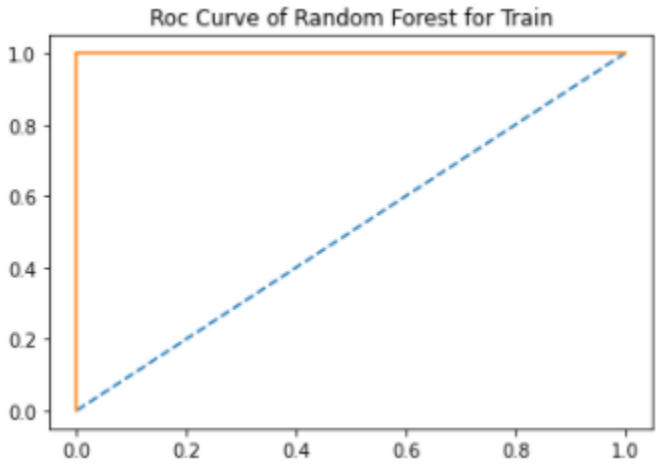


Figure 25 ROC curve of Random forest

For test data:

Accuracy score : 0.8421052631578947

	precision	recall	f1-score	support
Conservative	0.77	0.68	0.72	138
Labour	0.87	0.91	0.89	318
accuracy			0.84	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.84	0.84	0.84	456

Area Under the curve : 0.898869747516179

Table 16 Classification report on Random forest

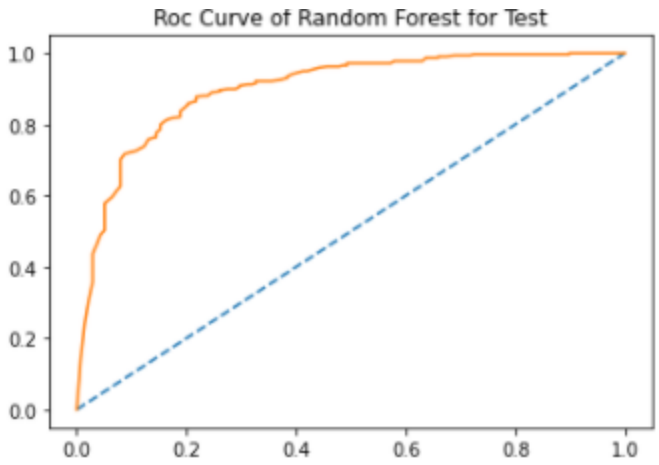


Figure 26 ROC curve of Random forest

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 1.0 100% of the predictions regarding done to Labour is correct.
  - Recall is 1.0, 100% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 1.0, 100% of the predictions are correct.
  - Recall is 1.0, 100% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.87 87% of the predictions regarding done to Labour is correct.

- Recall is 0.91, 91% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.77, 77% of the predictions are correct.
  - Recall is 0.68, 68% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.84 which is good score.

Area Under the cure value is 0.90 which indicates the model is performing better.

6. ADA BOOST CLASSIFIER (BASE MODEL):

For Train data:

Accuracy score : 0.8397737983034873

	precision	recall	f1-score	support
Conservative	0.75	0.70	0.73	322
Labour	0.87	0.90	0.89	739
accuracy			0.84	1061
macro avg	0.81	0.80	0.81	1061
weighted avg	0.84	0.84	0.84	1061

Area Under the curve : 0.9000033619378209

Table 17 Classification report Ada boost

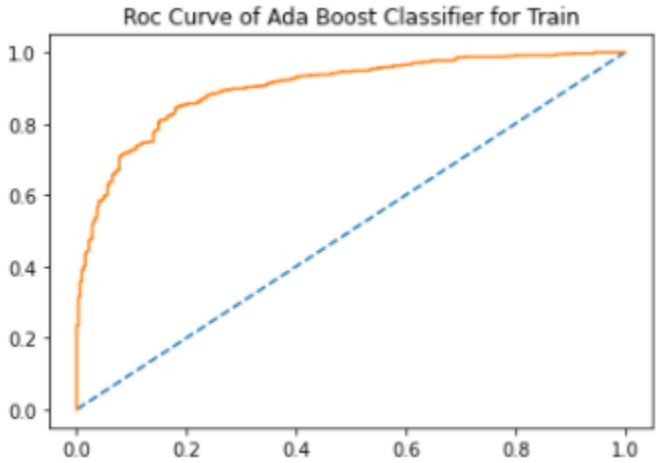


Figure 27 ROC curve of Ada boost

For Test data:

```
Accuracy score : 0.8355263157894737
              precision    recall  f1-score   support

Conservative       0.76      0.67      0.71       138
      Labour       0.86      0.91      0.88       318

    accuracy                           0.84       456
   macro avg       0.81      0.79      0.80       456
weighted avg       0.83      0.84      0.83       456


Area Under the curve : 0.9104571142101906
```

Table 18 Classification report Ada boost

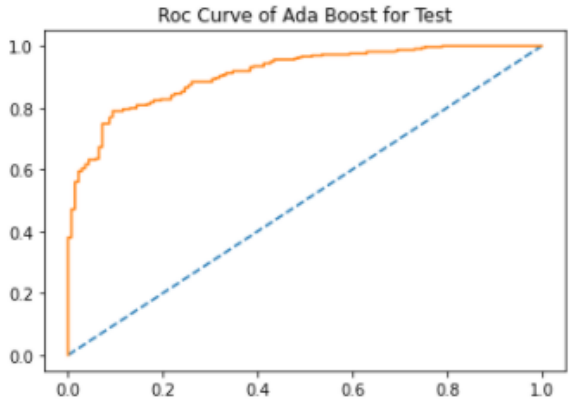


Figure 28 ROC curve of Ada boost

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.87 87% of the predictions regarding done to Labour is correct.
  - Recall is 0.9, 90% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.75, 75% of the predictions are correct.
  - Recall is 0.7, 70% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.86 86% of the predictions regarding done to Labour is correct.

- Recall is 0.91, 91% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.76, 76% of the predictions are correct.
  - Recall is 0.67, 67% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.84 which is good score.

Area Under the cure value is 0.90 which indicates the model is performing better.

7. GRADIENT BOSST (BASE MODEL):

For Train data:

Accuracy score : 0.885956644674835

	precision	recall	f1-score	support
Conservative	0.84	0.78	0.81	322
Labour	0.91	0.93	0.92	739
accuracy			0.89	1061
macro avg	0.87	0.86	0.86	1061
weighted avg	0.88	0.89	0.88	1061

Area Under the curve : 0.9470137587305323

Table 19 Classification report Gradient boost

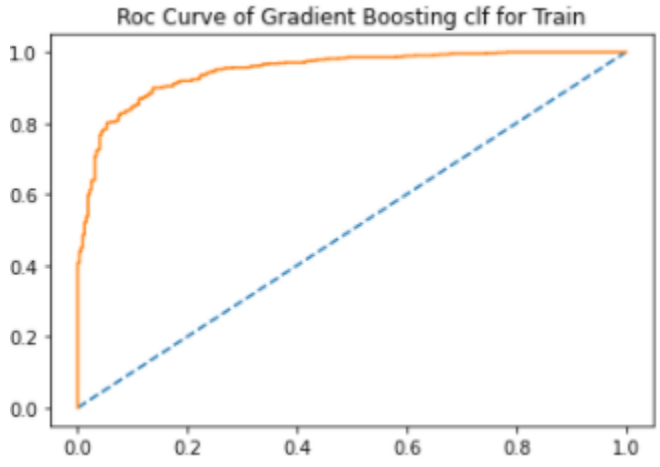


Figure 29 ROC curve gradient boost

For Test data:

```
Accuracy score : 0.8399122807017544
              precision    recall  f1-score   support

Conservative       0.76      0.68      0.72       138
      Labour       0.87      0.91      0.89       318

    accuracy                           0.84       456
   macro avg       0.82      0.79      0.80       456
weighted avg       0.84      0.84      0.84       456



Area Under the curve : 0.9042475617537145
```

Table 20 Classification report of Gradient Boost

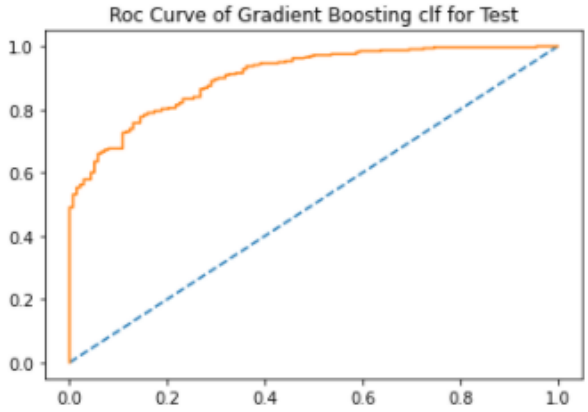


Figure 30 ROC curve of Gradient boost

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.91 91% of the predictions regarding done to Labour is correct.
  - Recall is 0.93, 93% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.84, 84% of the predictions are correct.
  - Recall is 0.78, 78% of the total people who vote for Conservative are predicted correctly.

TEST DATA

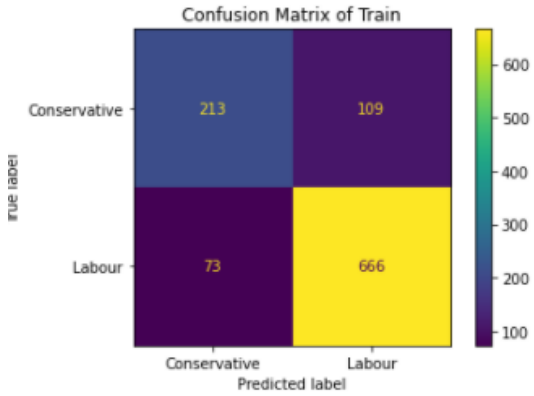
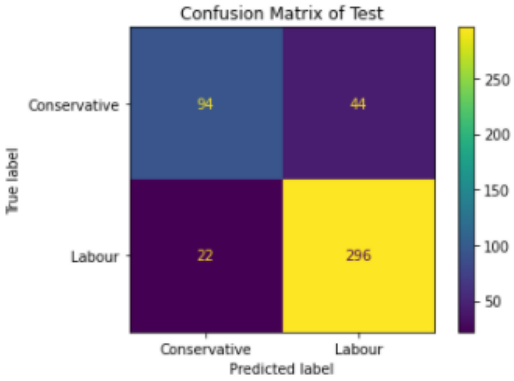
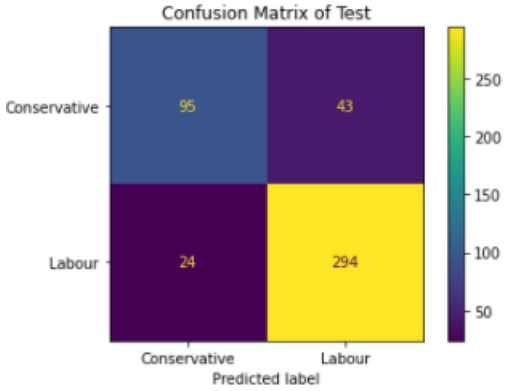
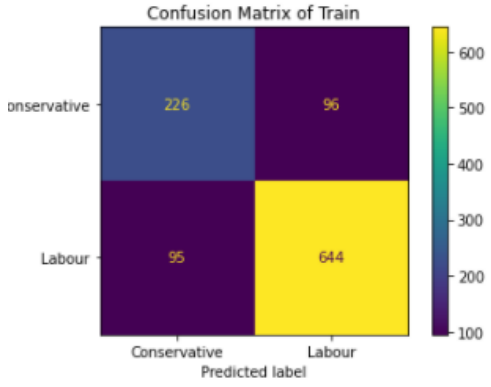
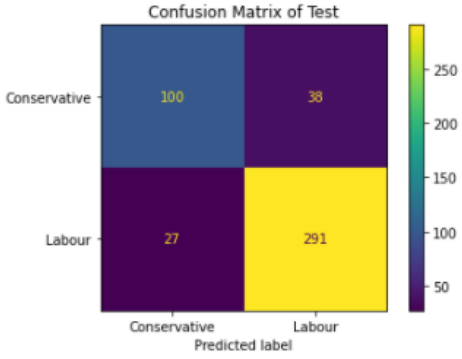
- FOR LABOUR
  - Precision is 0.87 87% of the predictions regarding done to Labour is correct.

    - Recall is 0.91, 91% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
    - Precision is 0.76, 76% of the predictions are correct.
    - Recall is 0.68, 68% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.84 which is good score.

Area Under the cure value is 0.90 which indicates the model is performing better.

CONFUSION MATRIX:

TRAIN DATA	TEST DATA
 Figure 31 Logistic clf	 Figure 32 Logistic clf
 Figure 33 LDA	 Figure 34 LDA
 Figure 35 Naive Bayer's	 Figure 36 Naive Bayer's

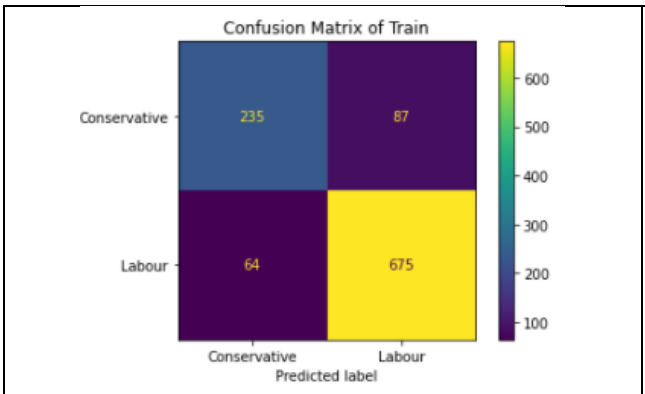


Figure 37 KNN

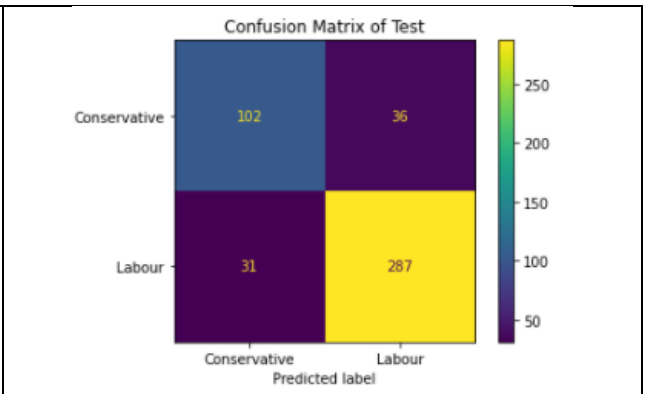


Figure 38 KNN

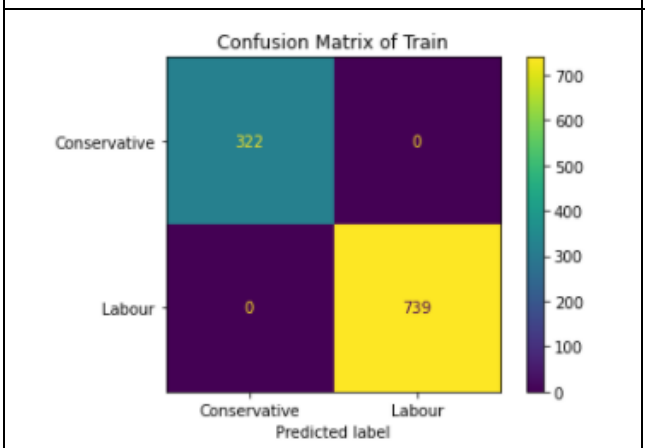


Figure 39 Random Forest

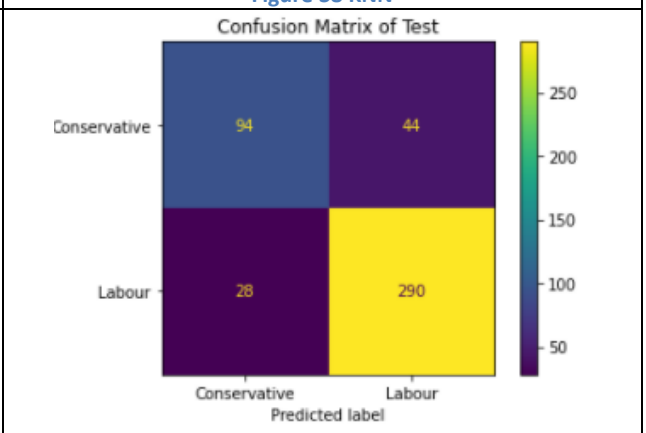


Figure 40 Random Forest

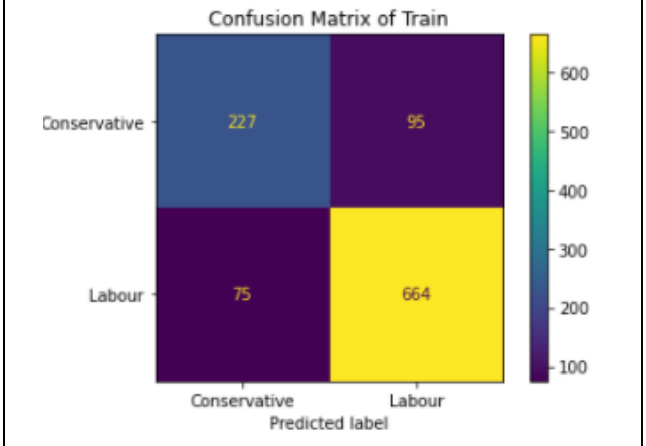


Figure 41 Ada Boost

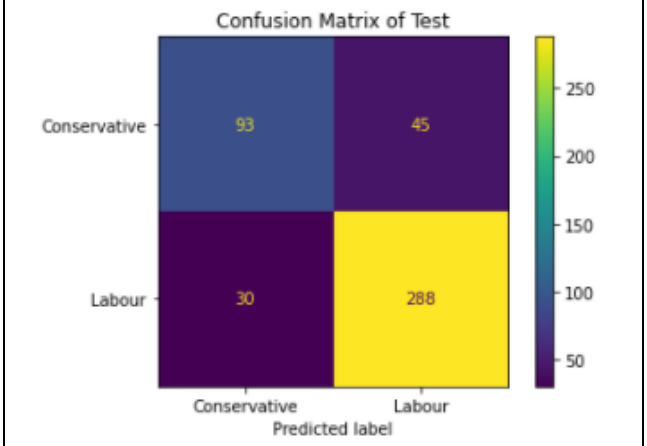


Figure 42 Ada Boost

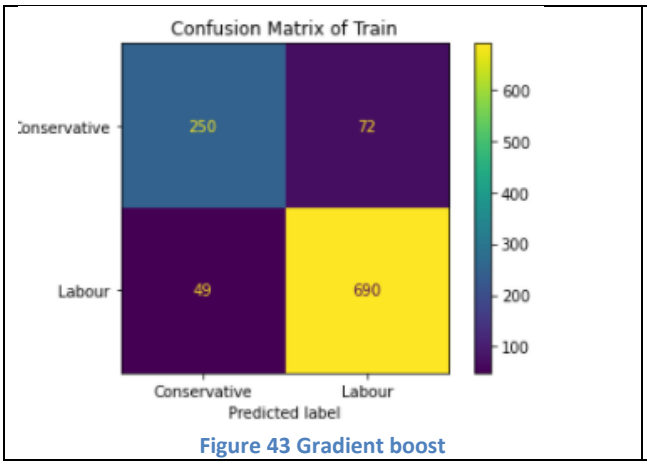


Figure 43 Gradient boost

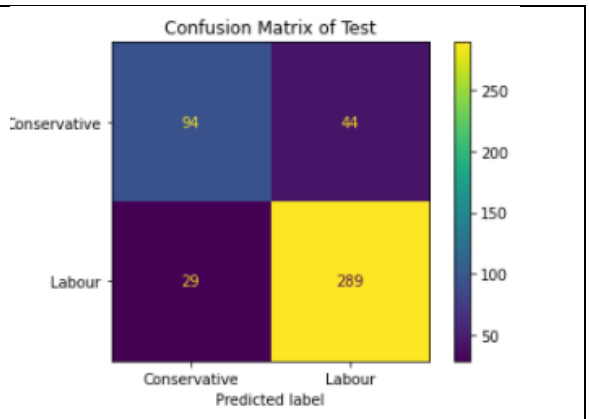


Figure 44 Gradient Boost

HYPER PARAMETER TUNING USING GRIDSEARCH:

1. LOGISTIC CLASSIFIER APPLYING GRIDSEARCH:

PARAMETERS USED:

- Penalty: l2, none terms are used, this is to add regularization term to the model to avoid over fitting of the data. None means no penalty added.
- Fit Intercept: Specifies if constant term must be added to the decision function, it acts like a base value for prediction when the input is 0.
- Solver: There are multiple optimization algorithms available; to choose the algorithm we use this parameter.
- Class weight : It uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data
- Tol: Tolerance for stopping criteria.

For Train data:

Accuracy score : 0.8133836003770029

	precision	recall	f1-score	support
Conservative	0.65	0.82	0.73	322
Labour	0.91	0.81	0.86	739
accuracy			0.81	1061
macro avg	0.78	0.82	0.79	1061
weighted avg	0.83	0.81	0.82	1061

Area Under the curve : 0.8775540221383606

Table 21 Classification report LogClf Grid Search Train

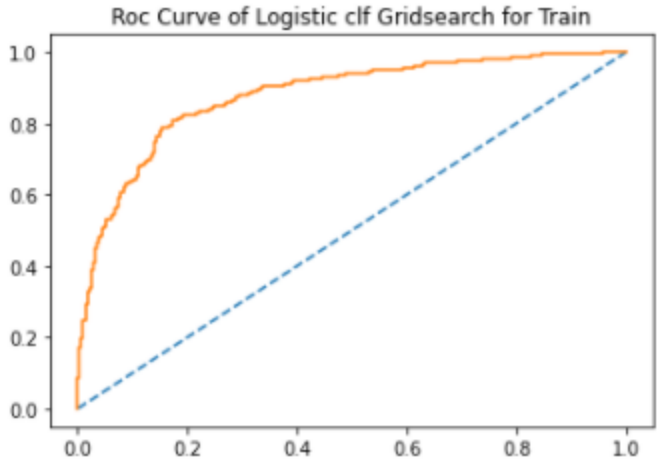


Figure 45 ROC curve LogClf Grid Search Train

For TEST data:

Accuracy score : 0.8289473684210527

	precision	recall	f1-score	support
Conservative	0.68	0.81	0.74	138
Labour	0.91	0.84	0.87	318
accuracy			0.83	456
macro avg	0.80	0.82	0.81	456
weighted avg	0.84	0.83	0.83	456

Area Under the curve : 0.9141372709871479

Table 22 Classification Report LogClf Grid Search Test

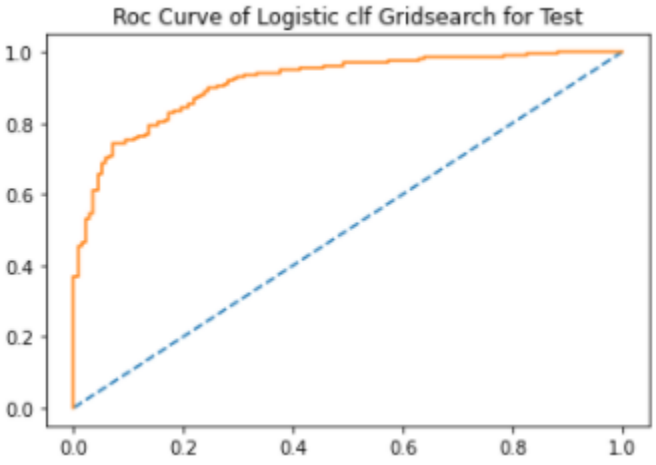


Figure 46 ROC curve LogClf Grid Search Test

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.91 91% of the predictions regarding done to Labour is correct.
  - Recall is 0.81, 81% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.65, 65% of the predictions are correct.
  - Recall is 0.82, 82% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.91 91% of the predictions regarding done to Labour is correct.
  - Recall is 0.81, 81% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.65, 65% of the predictions are correct.
  - Recall is 0.82, 82% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.81 which is good score.

Area Under the cure value is 0.87 which indicates the model is performing better.

Model seems to have good fit over the training and testing data.

2. LDA APPLYING GRID SEARCH:

PARAMETERS USED:

- Solver: [svd, lsqr, Eigen]; svd –Single value decomposition, lsqe-Least square solution, Eigen – Eigen value decomposition. Out of these solvers the best one is chosen.
- Store covariance: Use to calculate the weighted within class co variance matrix.
- Tol: Absolute threshold for a singular value of the input to be considered significant.

For Train data:

Accuracy score : 0.822808671065033

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

Area Under the curve : 0.876932063641483

Table 23 Classification Report LDA Grid Search Train

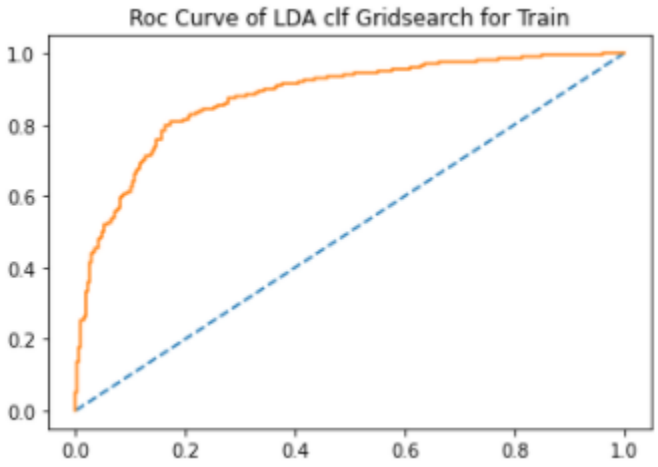


Figure 47 ROC curve LDA Grid Search Train

For Test data:

Accuracy score : 0.8530701754385965

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

Area Under the curve : 0.9143651444717892

Table 24 Classification Report LDA Grid Search Test

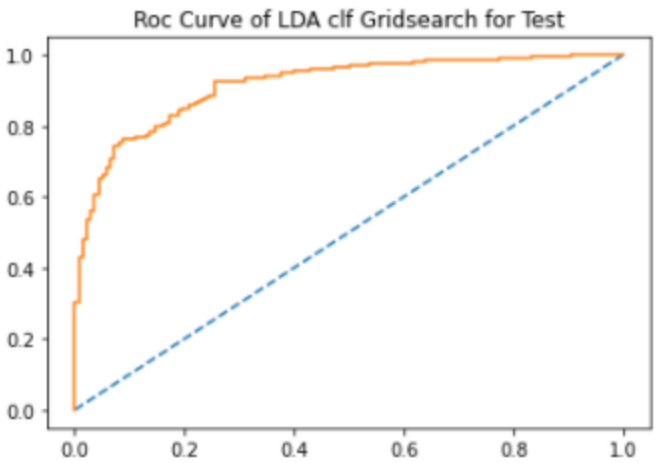


Figure 48 ROC curve LDA Grid search Test

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.86 86% of the predictions regarding done to Labour is correct.
  - Recall is 0.89, 89% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.72, 72% of the predictions are correct.
  - Recall is 0.67, 67% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.87 87% of the predictions regarding done to Labour is correct.
  - Recall is 0.92, 92% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.80, 80% of the predictions are correct.
  - Recall is 0.69, 69% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.85 which is good score.

Area Under the cure value is 0.91 which indicates the model is performing better.

Model seems to have good fit over the training and testing data.

3. NAÏVE BAYERS APPLYING GRID SEARCH:

PARAMETERS USED:

- Variable smoothing: Portion of the largest variance of all features that is added to variances for calculation stability.

For Training Data:

```
Accuracy score : 0.8067860508953817
              precision    recall  f1-score   support

Conservative       0.74      0.56      0.64       322
      Labour       0.83      0.91      0.87       739

    accuracy                           0.81      1061
   macro avg       0.78      0.74      0.75      1061
weighted avg       0.80      0.81      0.80      1061



Area Under the curve : 0.8538943847233545
```

Table 25 Classification report NB Grid Search Train

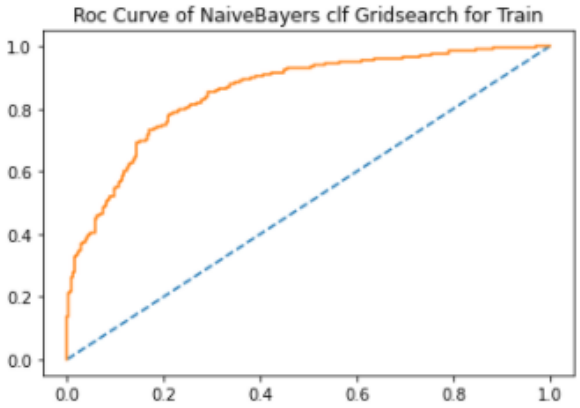


Figure 49 ROC curve NB Grid Search Train

For Test data:

```
Accuracy score : 0.8508771929824561
              precision    recall  f1-score   support

Conservative       0.82      0.65      0.73       138
      Labour       0.86      0.94      0.90       318

    accuracy                           0.85       456
   macro avg       0.84      0.79      0.81       456
weighted avg       0.85      0.85      0.85       456



Area Under the curve : 0.8953377085042384
```

Table 26 Classification report NB Grid Search Test

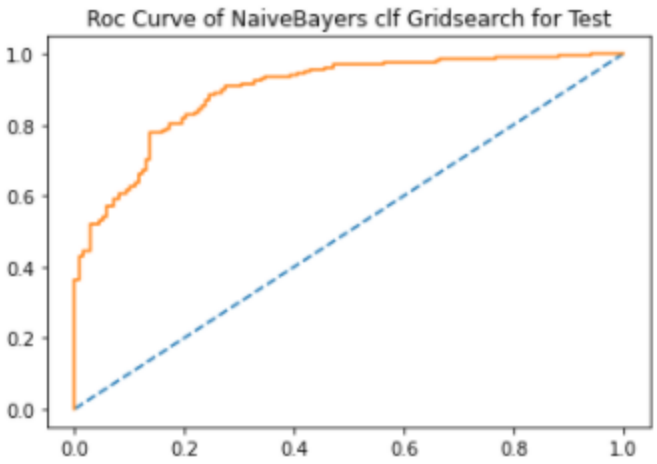


Figure 50 ROC curve NB Grid Search Test

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.83 83% of the predictions regarding done to Labour is correct.
  - Recall is 0.91, 91% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.74, 74% of the predictions are correct.
  - Recall is 0.56, 56% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.86 86% of the predictions regarding done to Labour is correct.
  - Recall is 0.94, 94% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.82, 82% of the predictions are correct.
  - Recall is 0.65, 65% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.85 which is good score.

Area Under the cure value is 0.91 which indicates the model is performing better.

Model seems to have good fit over the training and testing data.

4. KNN APPLYNG GRID SEARCH:

PARAMETERS USED:

- Neighbors : Number of neighbors to take into consideration for making the predictions
- Weights :[Uniform, Distance] Weights given to the neighbors based on the distance
- Algorithm: To choose from the available algorithm namely auto, ball tree, KD tree and brute force search.
- Leaf size : Number of leaves to consider for Ball tree and KD tree
- P: Power parameter, decides whether the distance used is Manhattan or Euclidean.

For Train data:

Accuracy score : 0.879359095193214

	precision	recall	f1-score	support
Conservative	0.81	0.80	0.80	322
Labour	0.91	0.92	0.91	739
accuracy			0.88	1061
macro avg	0.86	0.86	0.86	1061
weighted avg	0.88	0.88	0.88	1061

Area Under the curve : 0.9473142319232806

Table 27 Classification Report KNN Grid search Train

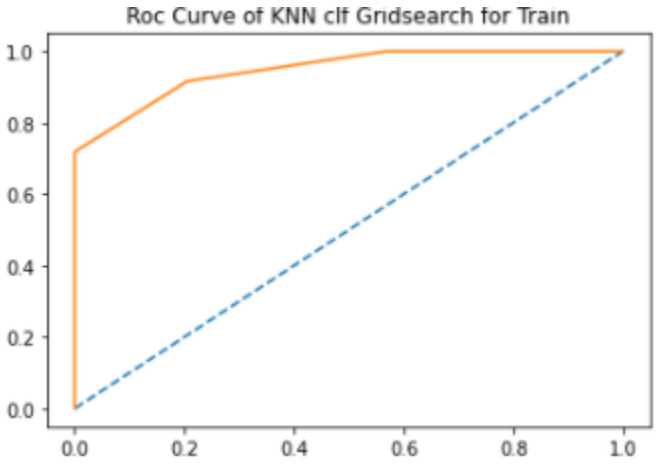


Figure 51 ROC curve KNN Grid search Train

For Test data:

Accuracy score : 0.5855263157894737

	precision	recall	f1-score	support
Conservative	0.34	0.38	0.36	138
Labour	0.72	0.67	0.69	318
accuracy			0.59	456
macro avg	0.53	0.53	0.53	456
weighted avg	0.60	0.59	0.59	456

Area Under the curve : 0.5563189317291041

Table 28 Classification Report KNN Grid search Test

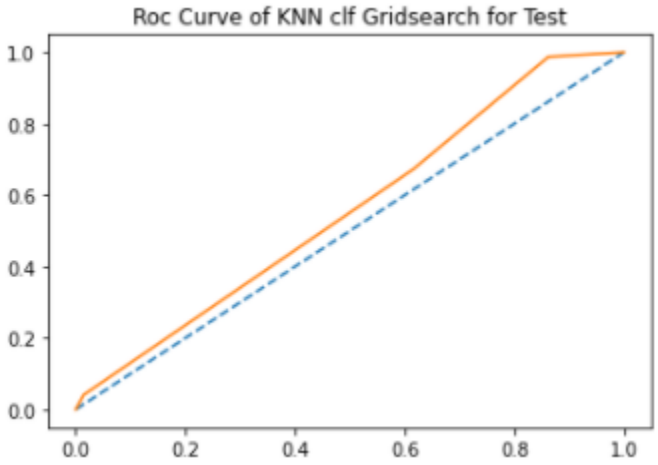


Figure 52 ROC curve KNN Grid search Test

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.91 91% of the predictions regarding done to Labour is correct.
  - Recall is 0.92, 92% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.81, 81% of the predictions are correct.
  - Recall is 0.80, 80% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.72 72% of the predictions regarding done to Labour is correct.
  - Recall is 0.67, 67% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.34, 34% of the predictions are correct.
  - Recall is 0.38, 38% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.55 which is not a good score.

Area Under the cure value is 0.55 which indicates the model is not performing well.

Model seems to have over fit the training hence the it performs very poorly over the test data.

5. RANDOM FOREST APPLYING GRID SEARCH:

PARAMETER USED:

- N_Estimators : Number of Decision Trees to consider for ensembling technique.
- Max features: Number of features to consider for each individual trees.
- Max Depth: Maximum depth allowed for each tree to grow.
- Min Sample split: Minimum sample required to make the split.
- Min Sample leaf: Minimum number of samples required at the leaf node.

For Train data:

Accuracy score : 0.825636192271442

	precision	recall	f1-score	support
Conservative	0.78	0.59	0.67	322
Labour	0.84	0.93	0.88	739
accuracy			0.83	1061
macro avg	0.81	0.76	0.78	1061
weighted avg	0.82	0.83	0.82	1061

Area Under the curve : 0.887837349448222

Table 29 Classification report for random forest Grid Search Train

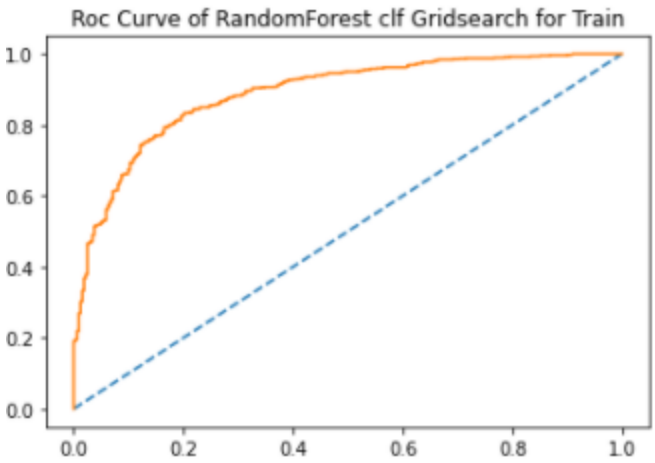


Figure 53 ROC curve for random forest Grid Search Train

For Test data:

Accuracy score : 0.8421052631578947

	precision	recall	f1-score	support
Conservative	0.83	0.60	0.70	138
Labour	0.85	0.95	0.89	318
accuracy			0.84	456
macro avg	0.84	0.77	0.80	456
weighted avg	0.84	0.84	0.83	456

Area Under the curve : 0.9100013672409079

Table 30 Classification report for random forest Grid Search Test

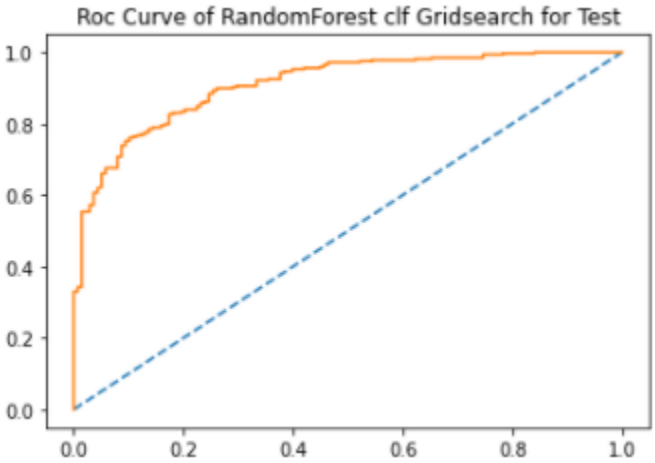


Figure 54 ROC curve for random forest Grid Search Test

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.84 84% of the predictions regarding done to Labour is correct.
  - Recall is 0.93 ,93% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.78, 78% of the predictions are correct.
  - Recall is 0.59,59% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.85 85% of the predictions regarding done to Labour is correct.
  - Recall is 0.95 ,95% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.83, 83% of the predictions are correct.
  - Recall is 0.60,60% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.84 which is good score.

Area Under the cure value is 0.91 which indicates the model is performing well.

Model seems to have over good fit over training and testing data.

6. ADA BOOST APPLYING GRID SEARCH

PARAMETERS USED:

- Base estimator: Base algorithm to make the first prediction and proceed with the boosting.
- N_estimator: Number of estimator to consider for ensembling technique.
- Learning rate: Weight applied to each classifier at each boosting iteration.
- Algorithm :[SAMME,SAMME.R] .

For Train Data:

Accuracy score : 0.7342130065975495

	precision	recall	f1-score	support
Conservative	0.54	0.76	0.63	322
Labour	0.87	0.72	0.79	739
accuracy			0.73	1061
macro avg	0.71	0.74	0.71	1061
weighted avg	0.77	0.73	0.74	1061

Area Under the curve : 0.7417338353827146

Table 31 Classification report Adaboost Grid Search Train

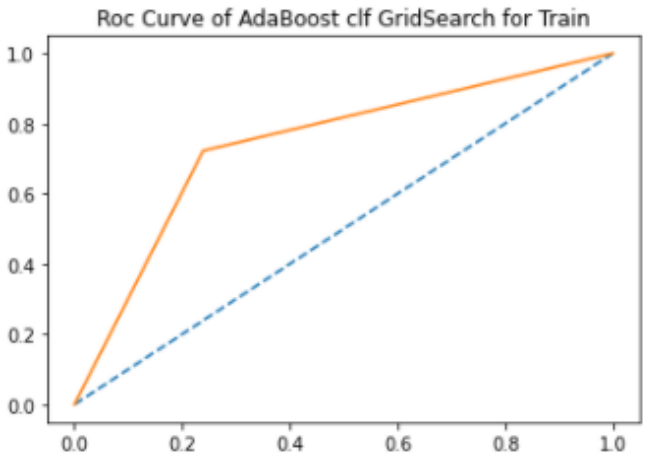


Figure 55 ROC curve Adaboost Grid Search Train

For Test data:

Accuracy score : 0.7412280701754386

	precision	recall	f1-score	support
Conservative	0.56	0.72	0.63	138
Labour	0.86	0.75	0.80	318
accuracy			0.74	456
macro avg	0.71	0.74	0.72	456
weighted avg	0.77	0.74	0.75	456

Area Under the curve : 0.7365326770576975

Table 32 Classification report Adaboost Grid Search Test

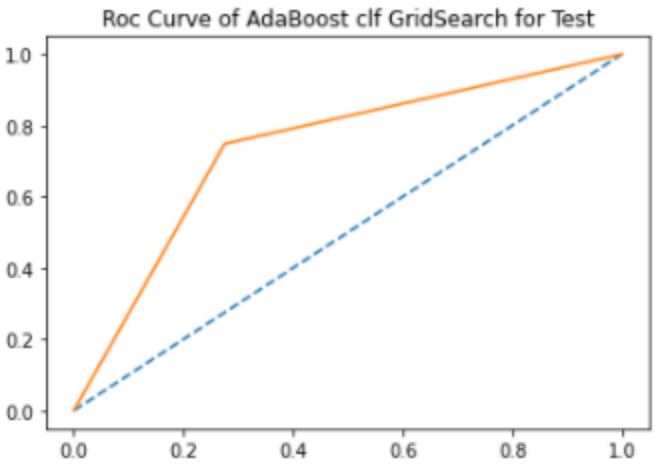


Figure 56 ROC curve Adaboost Grid Search Test

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.87 87% of the predictions regarding done to Labour is correct.
  - Recall is 0.72 ,72% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.54, 54% of the predictions are correct.
  - Recall is 0.76,76% of the total people who vote for Conservative are predicted correctly.

TEST DATA

- FOR LABOUR
  - Precision is 0.86 86% of the predictions regarding done to Labour is correct.
  - Recall is 0.75 ,75% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0.56, 56% of the predictions are correct.
  - Recall is 0.72,72% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.74 which is good score.

Area Under the cure value is 0.73 which indicates the model is average.

Model seems to have over good fit over training and testing data.

7. GRADIENT BOOST APPLYING GRID SEARCH

PARAMETERS USED:

- N_Estimators : Number of estimator to consider for ensembling technique.
- Learning rate: Weight applied to each classifier at each boosting iteration.
- Min Samples leaf: Minimum number of samples required at the leaf node.
- Min samples split: Minimum sample required to make the split.
- Max features: Number of features to consider for each individual trees.

For Train data:

Accuracy score : 0.6965127238454288

	precision	recall	f1-score	support
Conservative	0.00	0.00	0.00	322
Labour	0.70	1.00	0.82	739
accuracy			0.70	1061
macro avg	0.35	0.50	0.41	1061
weighted avg	0.49	0.70	0.57	1061

Area Under the curve : 0.8576345405491725

Table 33 Classification report Adaboost Grid Search Train

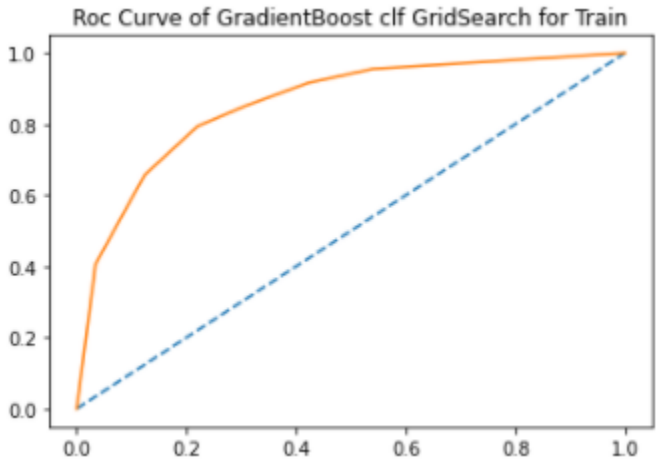


Figure 57 ROC curve of Gradient boost Grid Search Train

For Test data:

```
Accuracy score : 0.6973684210526315
              precision    recall  f1-score   support

Conservative       0.00      0.00      0.00       138
      Labour       0.70      1.00      0.82       318

    accuracy                           0.70       456
   macro avg       0.35      0.50      0.41       456
weighted avg       0.49      0.70      0.57       456



Area Under the curve : 0.8870431136632942
```

Table 34 Classification report Adaboost Grid Search Test

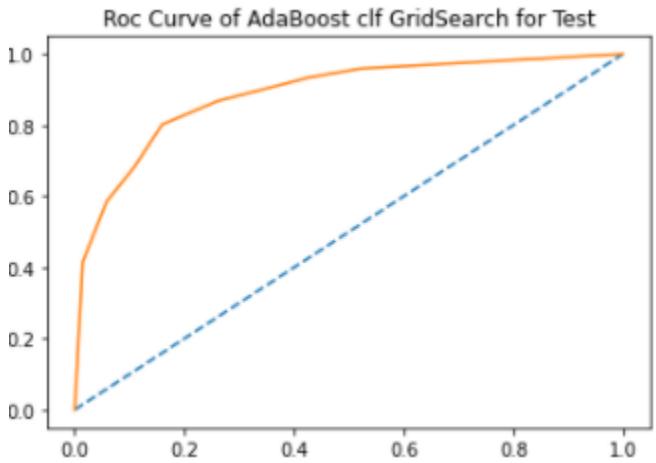


Figure 58 ROC curve of Gradient boost Grid Search Test

OBSERVATIONS:

TRAINING DATA

- FOR LABOUR
  - Precision is 0.70 70% of the predictions regarding done to Labour is correct.
  - Recall is 1.0 ,100% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0,0%
  - Recall is 0,0% of the total people who vote for Conservative is predicted

TEST DATA

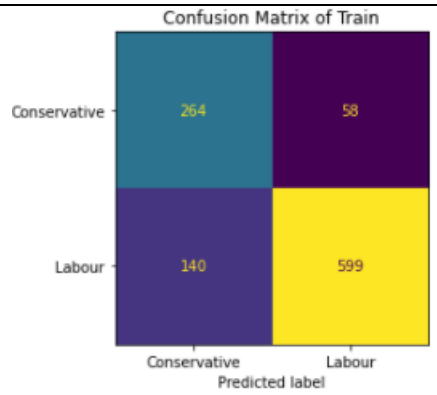
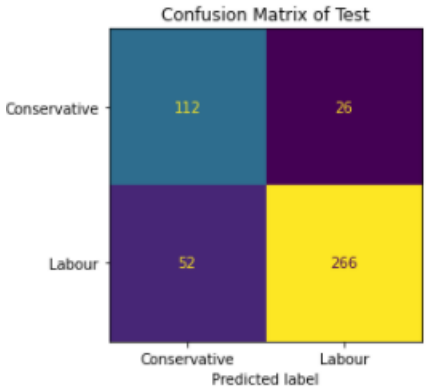
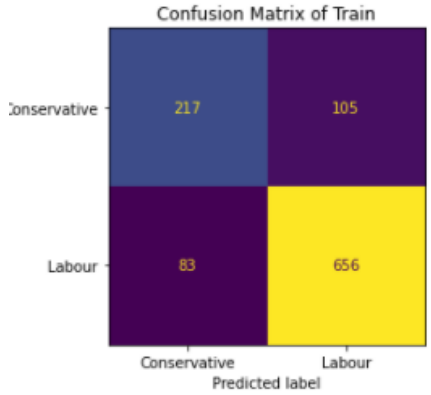
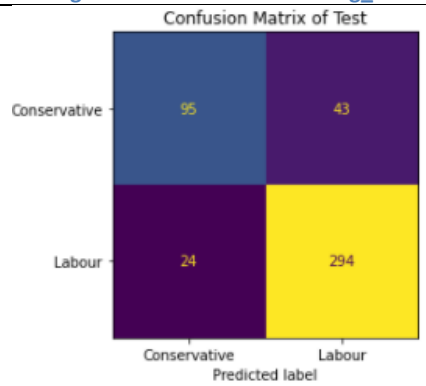
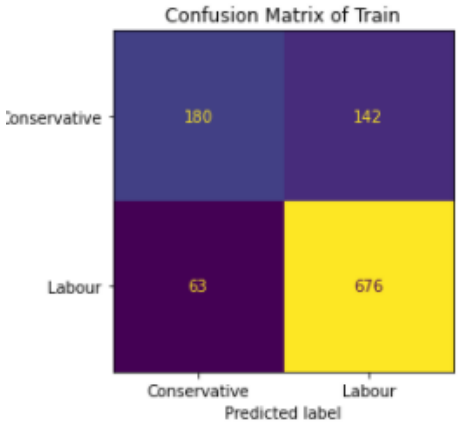
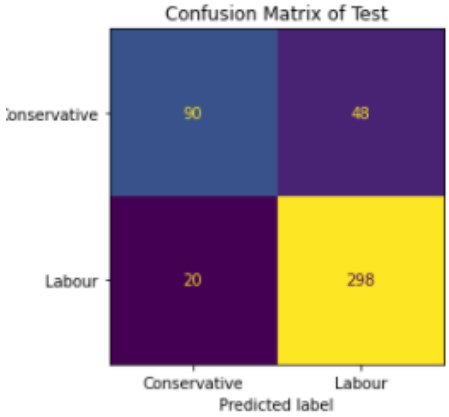
- FOR LABOUR
  - Precision is 0.70 70% of the predictions regarding done to Labour is correct.
  - Recall is 1.0 ,100% voters who vote for Labour is predicted correctly.

- FOR CONSERVATIVE
  - Precision is 0, 0% of the predictions are correct.
  - Recall is 0,0% of the total people who vote for Conservative are predicted correctly.
- Accuracy score is 0.69 which is Moderate.

Area Under the cure value is 0.88 which indicates the model is average because it predicts every data as Labour due to imbalance in the dataset most of the predictions appear to be correct and AUC is inflated because of the same reason.

Model seems to have poor fit over training and testing data.

CONFUSION MATRIX:

TRAINING DATA	TESTING DATA
 Figure 59 Confusion matrix Log_clf	 Figure 60 Confusion matrix Log_clf
 Figure 61 Confusion Matrix LDA	 Figure 62 Confusion matrix LDA
 Figure 63 Confusion matrix Naive bayers	 Figure 64 Confusion Matrix Naive Bayers

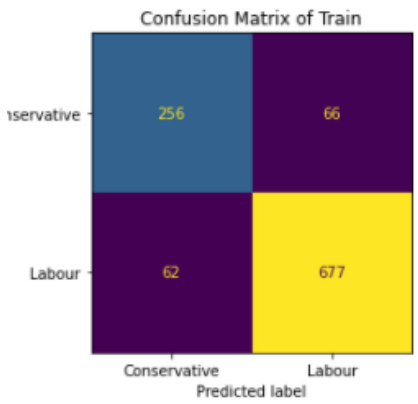


Figure 65 Confusion Matrix KNN

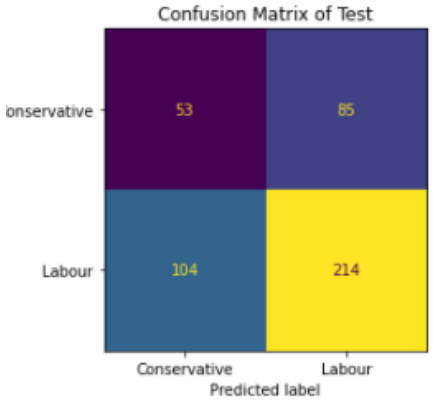


Figure 66 Confusion Matrix KNN

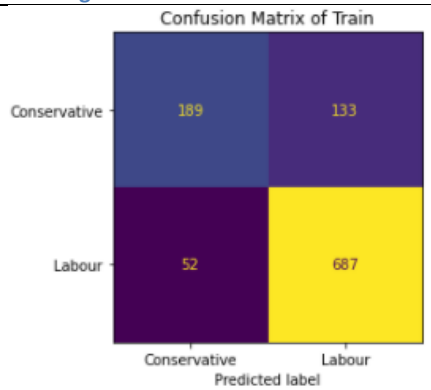


Figure 67 Confusion Matrix Random Forest

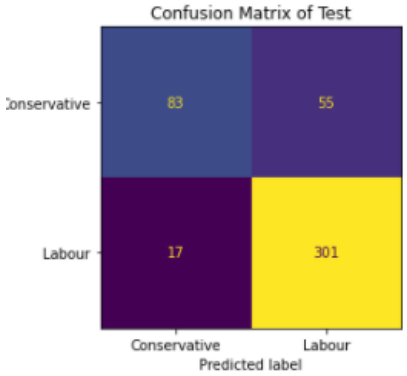


Figure 68 Confusion matrix Random Forest

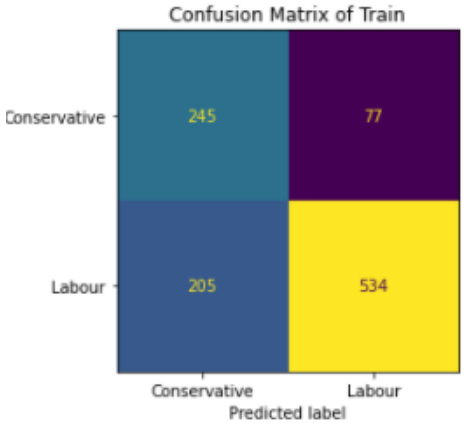


Figure 69 Confusion matrix Ada boost

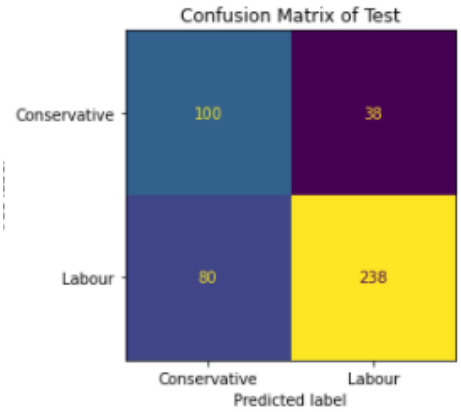


Figure 70 Confusion matrix of Ada boost

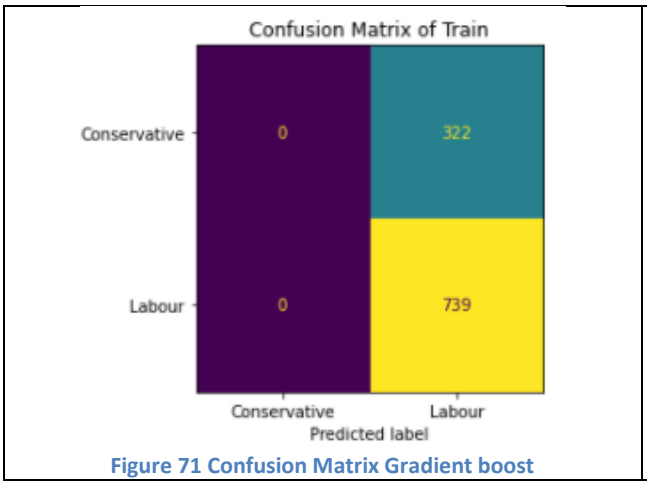


Figure 71 Confusion Matrix Gradient boost

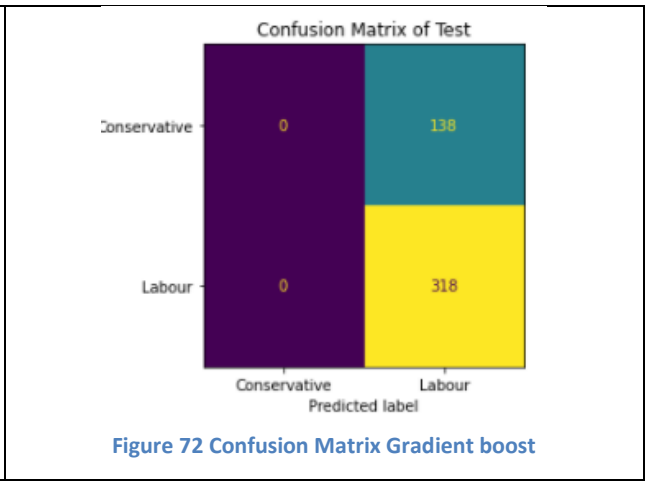


Figure 72 Confusion Matrix Gradient boost

MODEL SELECTION:

Precision = TP/ (TP+FP)

Recall= TP/ (TP+FN)

We are asked to find the votes casted by people to different Parties precisely so in that case we need to use Precision as the metric to choose the best model. Precision means how many selected items are actually positive, whereas recall is how many positive items are actually selected. The summary all the models with the summary are given below.

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_AUC	Test_AUC
0	log_clf	0.828	0.855	0.661	0.681	0.745	0.810	0.877	0.913
1	lda_clf	0.823	0.853	0.674	0.688	0.723	0.798	0.877	0.914
2	NB_clf	0.820	0.857	0.702	0.725	0.704	0.787	0.873	0.912
3	knn_clf	0.858	0.853	0.730	0.739	0.786	0.767	0.924	0.873
4	rforest	1.000	0.842	1.000	0.681	1.000	0.770	1.000	0.899
5	adaboost_clf	0.840	0.836	0.705	0.674	0.752	0.756	0.900	0.910
6	grad_boost	0.886	0.840	0.776	0.681	0.836	0.764	0.947	0.904

Table 35 Base model summary

Since the performance of all the models is moderate we tried to balance the data using SMOTE and tested these models again and the results are given below.

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_AUC	Test_AUC
0	log_clf	0.818	0.829	0.823	0.819	0.815	0.681	0.884	0.917
1	lda_clf	0.816	0.833	0.825	0.819	0.810	0.689	0.883	0.917
2	NB_clf	0.811	0.822	0.798	0.797	0.818	0.675	0.885	0.911
3	knn_clf	0.874	0.785	0.917	0.862	0.844	0.601	0.955	0.876
4	rforest	0.999	0.842	1.000	0.775	0.999	0.723	1.000	0.892
5	adaboost_clf	0.837	0.840	0.850	0.826	0.828	0.699	0.913	0.908
6	grad_boost	0.884	0.818	0.897	0.768	0.874	0.675	0.951	0.901

Table 36 Model summary after SMOTE

As we can see balancing the data has increased the recall for all the models but in return reduced the test precision. We are mainly focused on choosing the model based on precision so ,unbalanced dataset is further used for hyper parameter tuning.

After tuning all the models with Grid Search (result were given above) the results were:

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_AUC	Test_AUC
0	log_clf_tuned	0.813	0.829	0.820	0.812	0.653	0.683	0.878	0.914
1	lda_clf_tuned	0.823	0.853	0.674	0.688	0.723	0.798	0.877	0.914
2	NB_clf_tuned	0.807	0.851	0.559	0.652	0.741	0.818	0.854	0.895
3	knn_clf_tuned	0.555	0.586	0.311	0.384	0.286	0.338	0.510	0.556
4	rforest_clf_tuned	0.826	0.842	0.587	0.601	0.784	0.830	0.888	0.910
5	adaboost_clf_tuned	0.734	0.741	0.761	0.725	0.544	0.556	0.742	0.737
6	grad_boost_tuned	0.697	0.697	0.000	0.000	0.000	0.000	0.858	0.887

Table 37 Model summary of Grid Search

The final results show some improvement in the test precision of Random forest and Naïve Bayer's classifiers. The final model can be Random Forest with the maximum test precision of 0.830 , because it has a less train precision and more test precision indicating it has generalized on the data properly and classifies the voters properly. Overall accuracy is also in the acceptable range(0.826).

SUMMARY:

Analyzing the data we found out it had 9 columns in total in which we had 7 numerical columns which are ordinal 1 object column which is gender. The dataset had no missing values or incorrect entries in it so imputation was skipped. Although it had 8 duplicate records in this was removed from the dataset to proceed with the analysis. Once the data is cleaned EDA was done on it the following were the insights:

- ✓ People who vote for Conservative party are relatively older than people voting for Labour Party.
- ✓ Most of the voters are women.
- ✓ No party preference for each gender.
- ✓ Economic condition of the people plays a vital role in choosing the party, people above 3 are preferring Labour and people below 3 are choosing Conservative.

- ✓ People who evaluated Hague 4 still chose to vote for Labour party, this means people consider Hague to be a better a leader but still chose the other party for unknown reasons.
- ✓ Eurosceptic sentiment directly correlated to Hague or Conservative party because people with high Eurosceptic sentiment prefer conservative party and people with low Eurosceptic sentiment votes for Labour party.
- ✓ People with better understanding on Politics vote for Conservative party.
- ✓ People with minimum or no knowledge about Politics vote for Labour party.
- ✓ Very weak correlation exists in the dataset notable relationship is , Europe sentiment is positively correlated to Hague and economic condition is positively correlated to Blair.

Completing the EDA chi square test was done on all the 7 ordinal categorical columns ,the p value for all the columns were found to be insignificant which forces us to reject the null hypothesis and conclude all the categorical columns affects the dependent variable so all the columns are included in model building.

Before proceeding with the model development the data was encoded. The Gender column was one hot encoded all the ordinal columns were kept as such and the target variable was untouched, the labels were passed to functions calculating the metrics to work properly. After encoding, data was split in to 70,30 proportion for training and testing purpose. A copy of trained data is scaled and stored to work with KNN model.

Base models were created for the following algorithms: Logistic Classifier, Liner Discriminant Analysis, and Naïve Bayer's, K-Nearest neighbor, Random Forest, Ada Boost, and Gradient Boost Classifier. Following it up with the help of Grid Search tuned version of all the models were created and compared. By having precision has the selection metric based on the problem Random Forest was chosen has the best model.

RECOMMENDATION:

- ➢ Only old people seem to be participating in the election many youngsters haven't participated so Conservative party leader Hague can focus campaigning in the social media and other online platforms to reach the youngsters.
- ➢ Many men are not voting so a survey can be conducted to find the reason , and those running for the seats can attract men to vote in the elections by advertising properly or providing them with the voters slip which might encourage them to vote.
- ➢ Awareness programs can be conducted along with the campaign to make the people understand the importance of the elections and encourage more to cast their vote.
- ➢ Hague is behind Blair because people with better economic conditions prefer Blair, so Hague can try to focus his campaign on focusing the problems and solution needed for these people which may convince them to vote for him.

- People with less political understanding are the ones voting for Blair ,Hague can run small community welfare programs to reach these kind of people
- Hague can also try run his banner using digital marketing to people to convert many votes.
- Hague should focus the his campaign with plans that improve the economic growth which might convince the people. Like try to create more jobs, bring in more industries.
- In case of Blair he is already leading with a huge difference but to make it better he can focus on explain his solution for people who are economically backward.

TEXT ANALYSIS

ANALYSIS ON PRESIDENT SPEECHES

We will be analyzing following speeches of the Presidents of the United States of America:

- President Franklin D. Roosevelt in 1941
- President John F. Kennedy in 1961
- President Richard Nixon in 1973

The following speeches are retrieved from the inaugural package of NLTK. Once retrieved it was found out the inaugural package contains 5050 sentences and 149797 words in it.

After checking for the sentences and words these 3 particular speech were filtered and stored in a separate data frame. Number of character, words and sentences were calculated for these 3 speeches alone.

	Name	speech	sentences	words	characters
0	Roosevelt	on each national day of inauguration since 178...	69	1360	6174
1	Kennedy	vice president johnson, mr. speaker, mr. chief...	56	1390	6202
2	Nixon	mr. vice president, mr. speaker, mr. chief jus...	73	1819	8122

Table 38 Sentences, word and character count

DATA CLEANING:

Cleaning is necessary for the textual data as it contains numbers and punctuations which might not imply much of an information, in order to proceed with the cleaning the data is first converted to lower case. As python is case sensitive so same words with different cases will be considered as a separate word. Converting to lower case will make the cleaning easy. Numbers and punctuation were removed using the re package. The cleaned data is given below.

	Name	speech	sentences	words	characters
0	Roosevelt	on each national day of inauguration since th...	69	1360	6174
1	Kennedy	vice president johnson mr speaker mr chief jus...	56	1390	6202
2	Nixon	mr vice president mr speaker mr chief justice ...	73	1819	8122

Table 39 Summary of the clean data

Once the data is cleaned we can try to find the most frequent words used in each individual speeches and the following are the output.

```
array([list([('the', 114), ('of', 81), ('and', 46), ('to', 36), ('in', 35)]),
       list([('the', 86), ('of', 65), ('to', 42), ('and', 41), ('we', 30)]),
       list([('the', 83), ('of', 68), ('to', 65), ('in', 58), ('and', 50)])],
      dtype=object)
```

Table 40 Most Frequent words

From the above table we can see the most frequent words are common words like to, of, the etc. These prepositions don't bear any information to analysis or model building so these are called stop words. We must remove these stop words to find the useful most frequent words.

In order to proceed with further steps a process called lemmatization must be done in order to bring all the words to its root form to make the word list into a unique collections. For example words like punishing, punished, punishment are all denoting the meaning of punish. So to remove the 'ing','ed' and to find its root form lemmatization is done.

After lemmatization few more words were added to stop words list like 'mr',' mrs','u',''has','was','let','know'. As the speech is by the presidents of America, the word America is more popular in the speech. It includes the words like 'world',' nation ' so these were considered to be irrelevant and were included in the stop words list. So it was removed from the text data.

After cleaning the following is the most frequent words for all 3 speeches:

```
array([list([('life', 11), ('people', 9), ('spirit', 9), ('democracy', 9), ('year', 7)]),
       list([('side', 8), ('power', 7), ('new', 7), ('pledge', 7), ('citizen', 5)]),
       list([('peace', 19), ('responsibility', 16), ('new', 15), ('government', 10), ('great', 9)])],
      dtype=object)
```

Table 41 Most Frequent words after cleaning

	Name	speech	sentences	words	characters	clean_words	words_af_cleaned	characters_af_cleaned
0	Roosevelt	on each national day of inauguration since the...	69	1360	6174	[national, day, inauguration, since, people, r...	579	3655
1	Kennedy	vice president johnson mr speaker mr chief jus...	56	1390	6202	[vice, president, johnson, speaker, chief, jus...	640	3787
2	Nixon	mr vice president mr speaker mr chief justice ...	73	1819	8122	[vice, president, speaker, chief, justice, sen...	726	4545

Table 42 Count of Words & characters after cleaning

After removing all the unnecessary words we can see the word count has drastically reduced for all 3 speeches.

Once the speeches are pre-processed the word cloud is formed for 3 speeches individually.

FOR ROOSVELT:

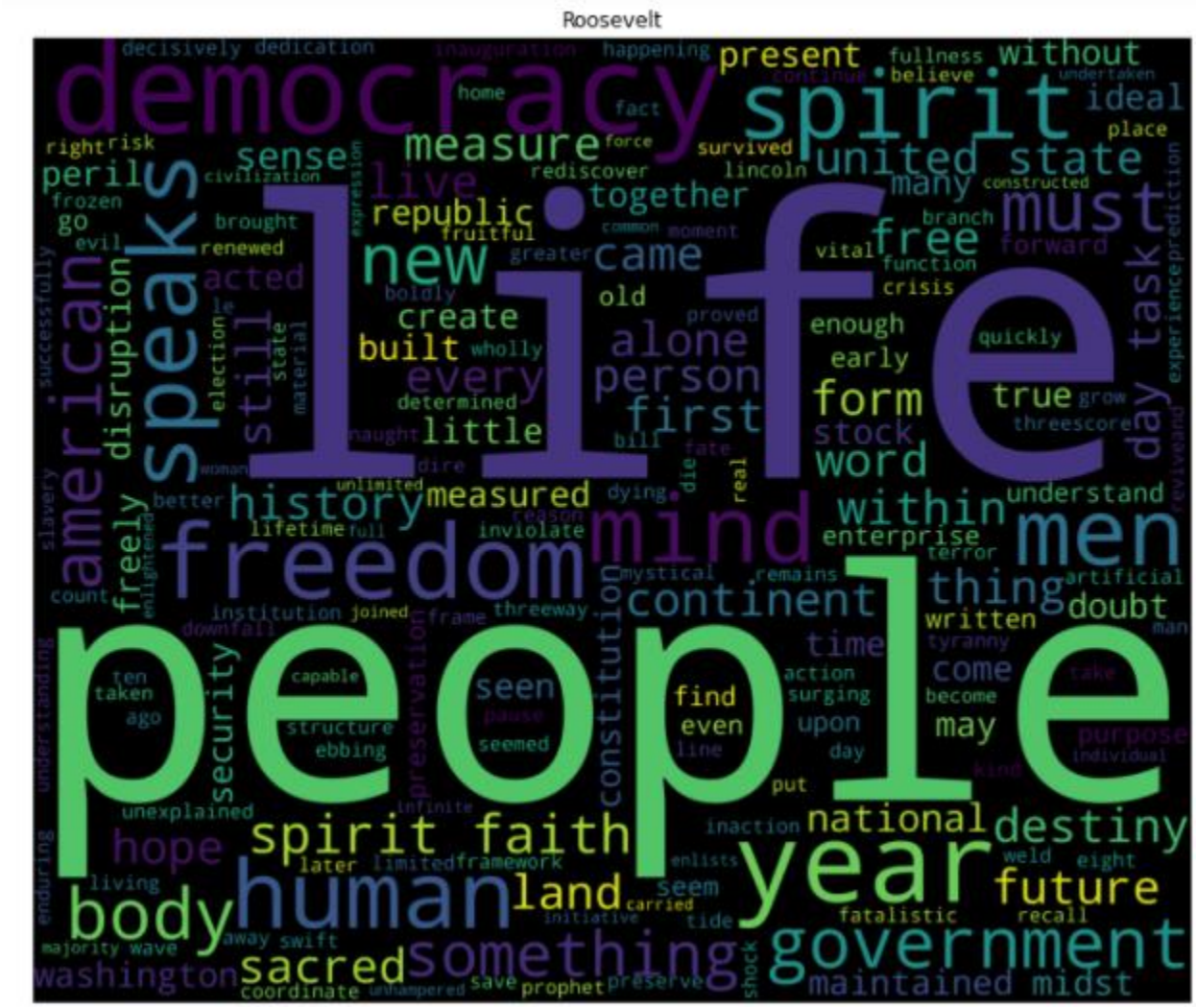


Figure 73 Word Cloud Roosevelt

Important words are :Life,People,human,freedom,democracy,speaks,government,body,year,spirit

FOR KENNEDY :

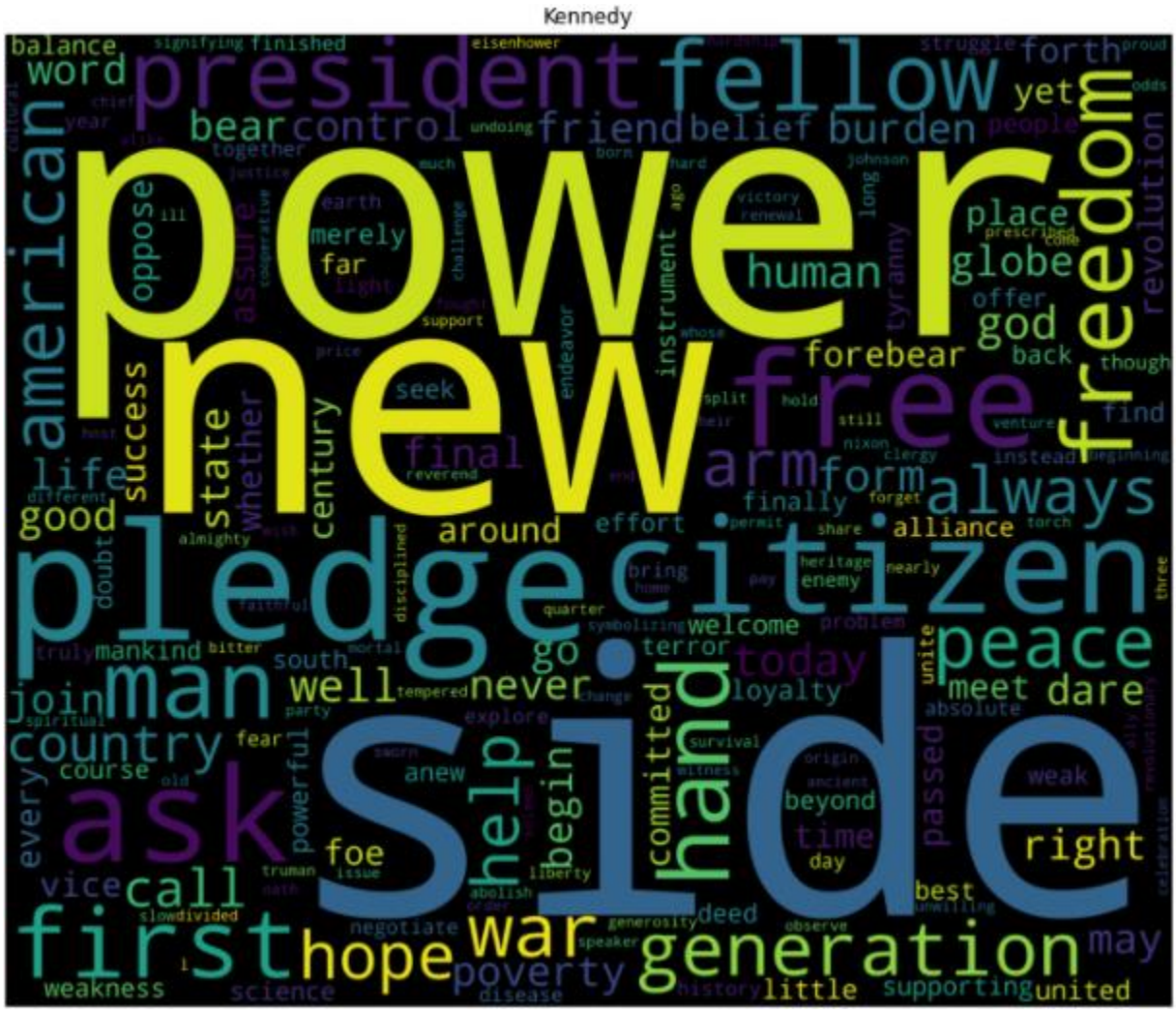


Figure 74 Word Cloud Kennedy

Important words :Power, pledge, side, free, freedom, new, control, free, war and poverty.

For Nixon:

Figure 75 Word Cloud Nixon

Important words: Peace, responsibility , government, abroad ,history , great , right and time.