MACHINE LEARNING

A solution document for the rubric given with quality

Gudla Sai Srinivas

Table of Contents

List of Tables

Table of Figures

Problem 1:

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

Data set: Election Data

Data Dictionary:

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

Table 1 Data Dictionary

Data Ingestion:

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

- The required packages were loaded.
- The data is loaded using pandas.
- After loading the dataset, it is observed that one column in the dataset is unnamed and has no significance in model building, hence we drop the column before proceeding in the EDA phase.
- The Dataset has 1525 rows and 9 features.
- The data type of the variables are as follows:

Data Type	Count of Columns
int64	7
object	2
Grand Total	**9**

Table 2. Count of Data Types

➢ Data Exploration was performed using the following functions:

- Head

vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
Conservative	67	5	3	2	4	11	3	male
Conservative	73	2	2	4	4	8	2	male
Labour	37	3	3	5	4	2	2	male
Conservative	61	3	3	1	4	11	2	male
Conservative	74	2	3	2	4	11	0	female

Table 3. First 5 rows of Data Set

- Tail

vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
Conservative	67	5	3	2	4	11	3	male
Conservative	73	2	2	4	4	8	2	male
Labour	37	3	3	5	4	2	2	male
Conservative	61	3	3	1	4	11	2	male
Conservative	74	2	3	2	4	11	0	female

Table 4. Last 5 Rows of Data set

- Shape

  The dataset has 1525 rows and 9 variables (After removing the unnamed variable).

  From the given dataset we can observe that there are total 9 variables in which 6 of the independent variables are categorical with some coding. One continuous variable and one nominal categorical variable.

- Summary

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

Table 5. Summary of the Data

Mean age is quite high which 54.182 and minimum age for voting is 24 can be observed from the summary

- Check Duplicates

  After checking duplicates, we found that there are 8 duplicates values which were removed from the dataset.

- Null Values

```
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
vote                        1525 non-null object
age                         1525 non-null int64
economic.cond.national      1525 non-null int64
economic.cond.household     1525 non-null int64
Blair                       1525 non-null int64
Hague                       1525 non-null int64
Europe                      1525 non-null int64
political.knowledge         1525 non-null int64
gender                      1525 non-null object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

Fig. 1. Information on data set

  No null values were observed in the dataset.

- Skewness

Skewness of age = 0.13979987012068112

Skewness of economic.cond.national = -0.23847421478161793

Skewness of economic.cond.household = -0.14414766882077137

Skewness of Blair = -0.5395141989831328

Skewness of Hague = 0.1461913444629453

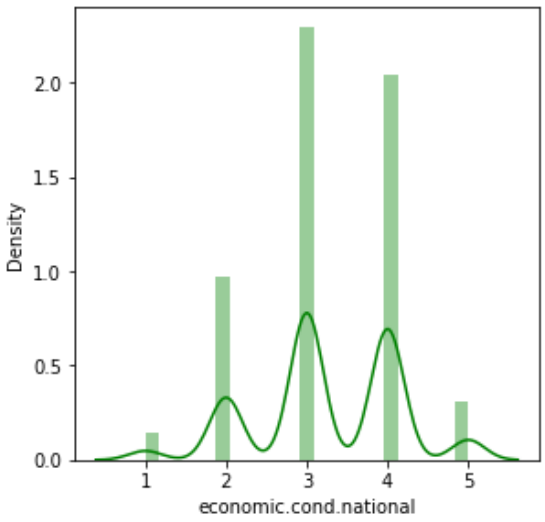
Skewness of Europe = -0.14189094981032258

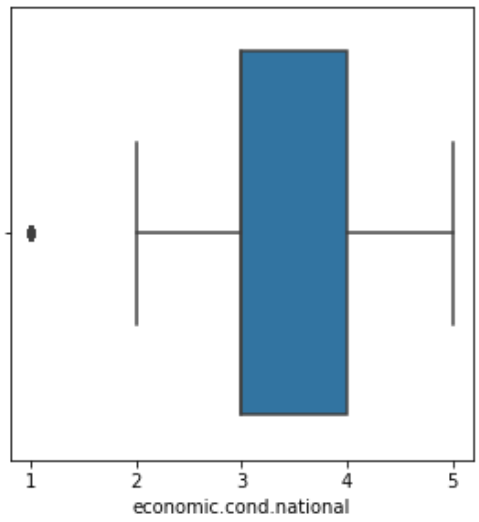
Skewness of political.knowledge = -0.4229276205374301

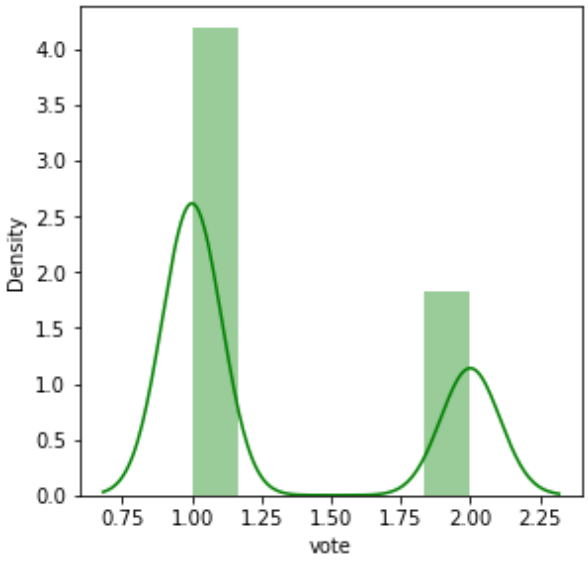
As per the values of skewness some are negatively skewed but ‘age’ and ‘Hague’ are positively skewed variables.

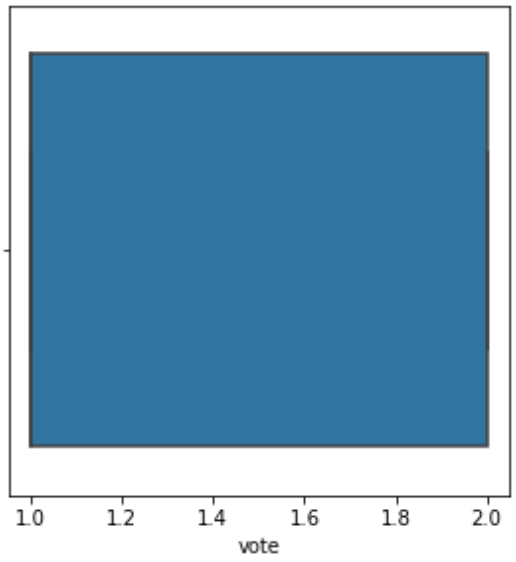
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

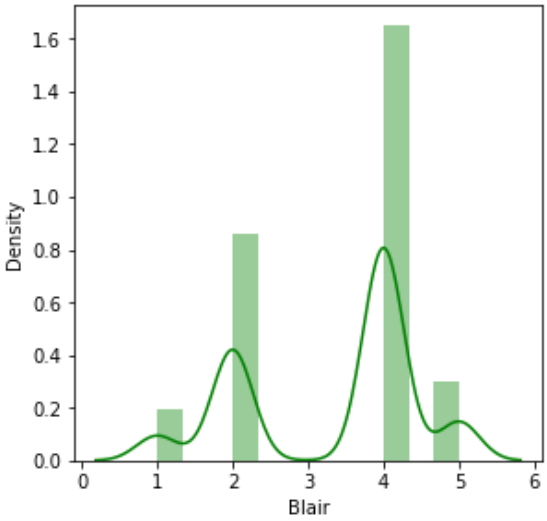
Univariate Analysis:

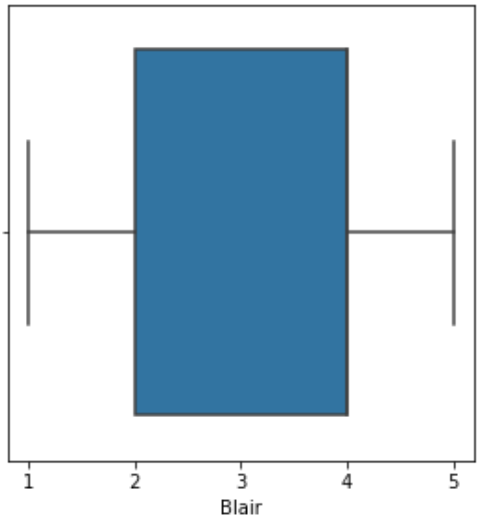


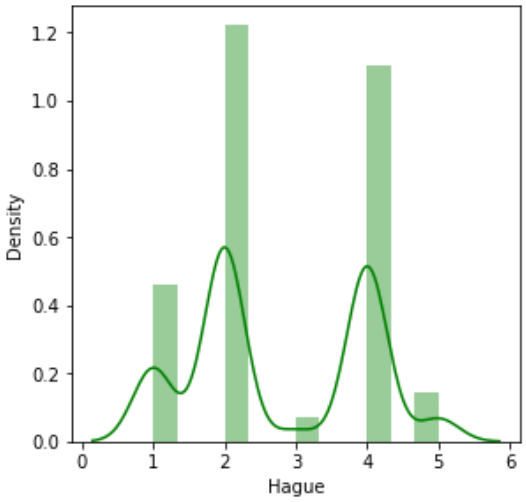


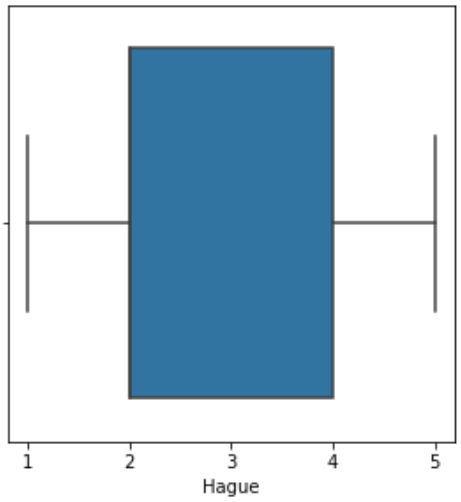


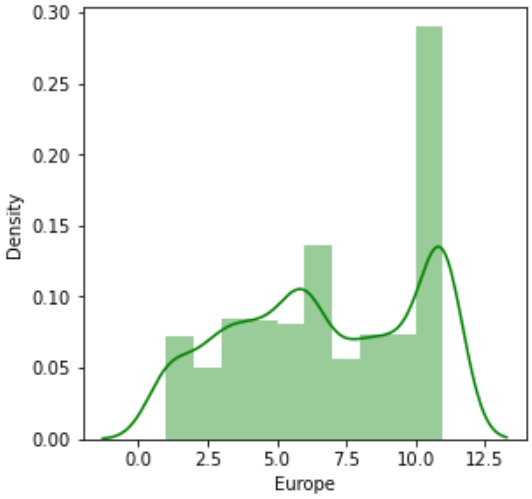


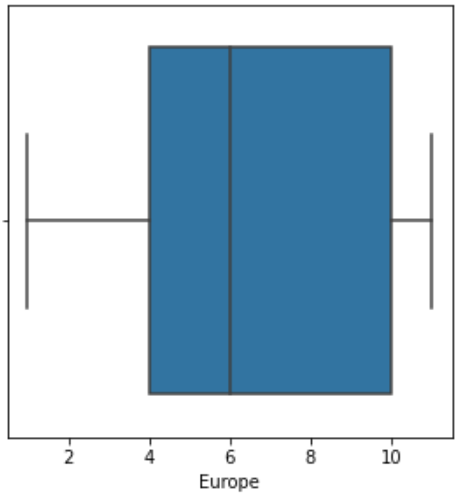

Density
Blair

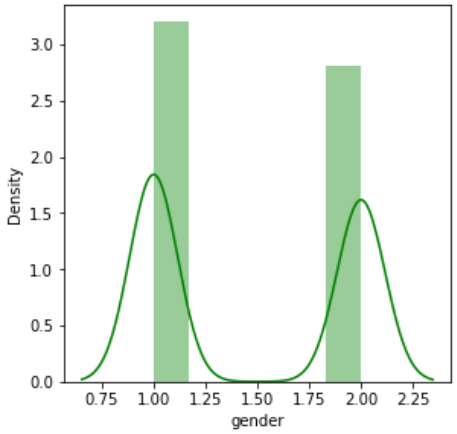

Blair

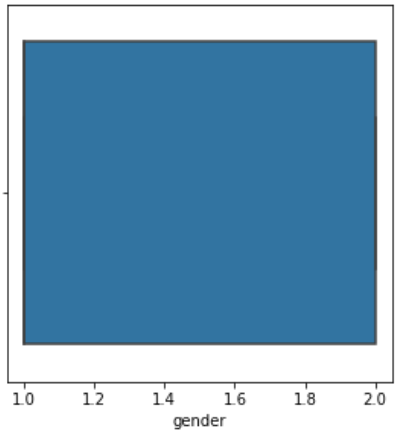

Density
Hague

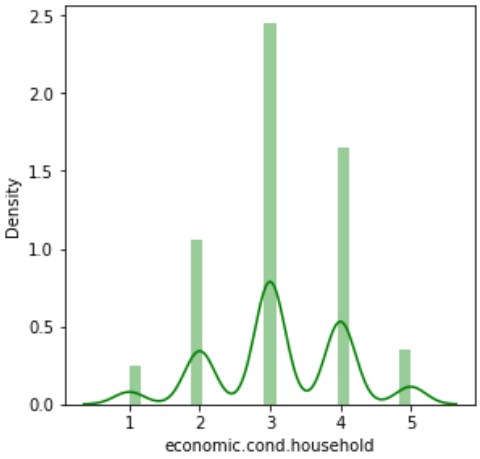

Hague

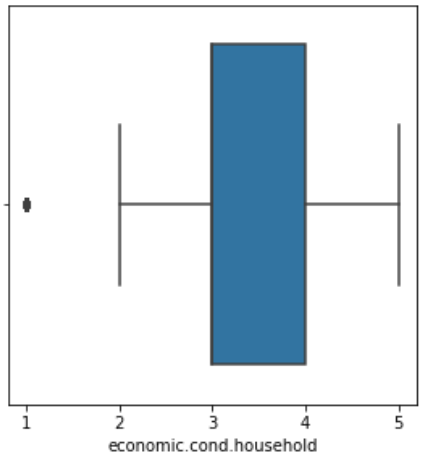

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0.0
2.5
5.0
7.5
10.0
12.5
Europe

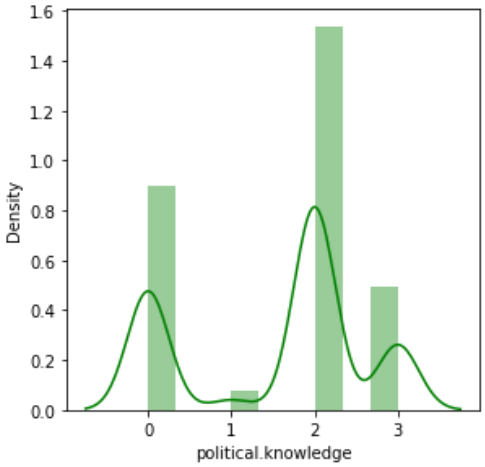

2
4
6
8
10
Europe

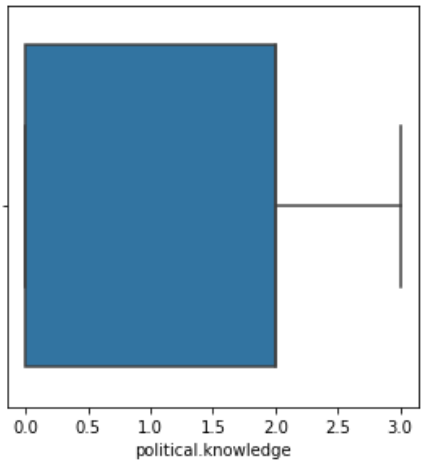

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
0.75
1.00
1.25
1.50
1.75
2.00
2.25
gender


1.0
1.2
1.4
1.6
1.8
2.0
gender


Density
economic.cond.household


economic.cond.household


Density
political.knowledge


political.knowledge

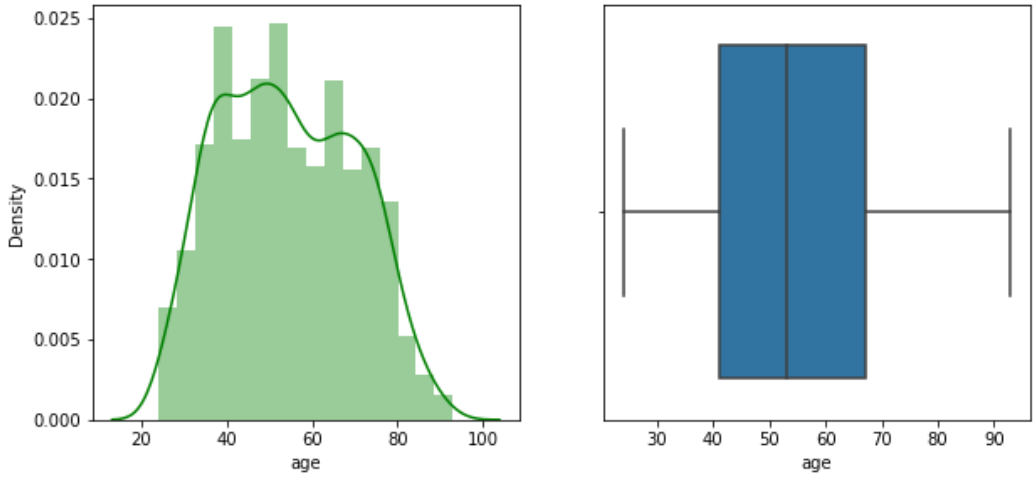


Fig. 3. Univariate Analysis

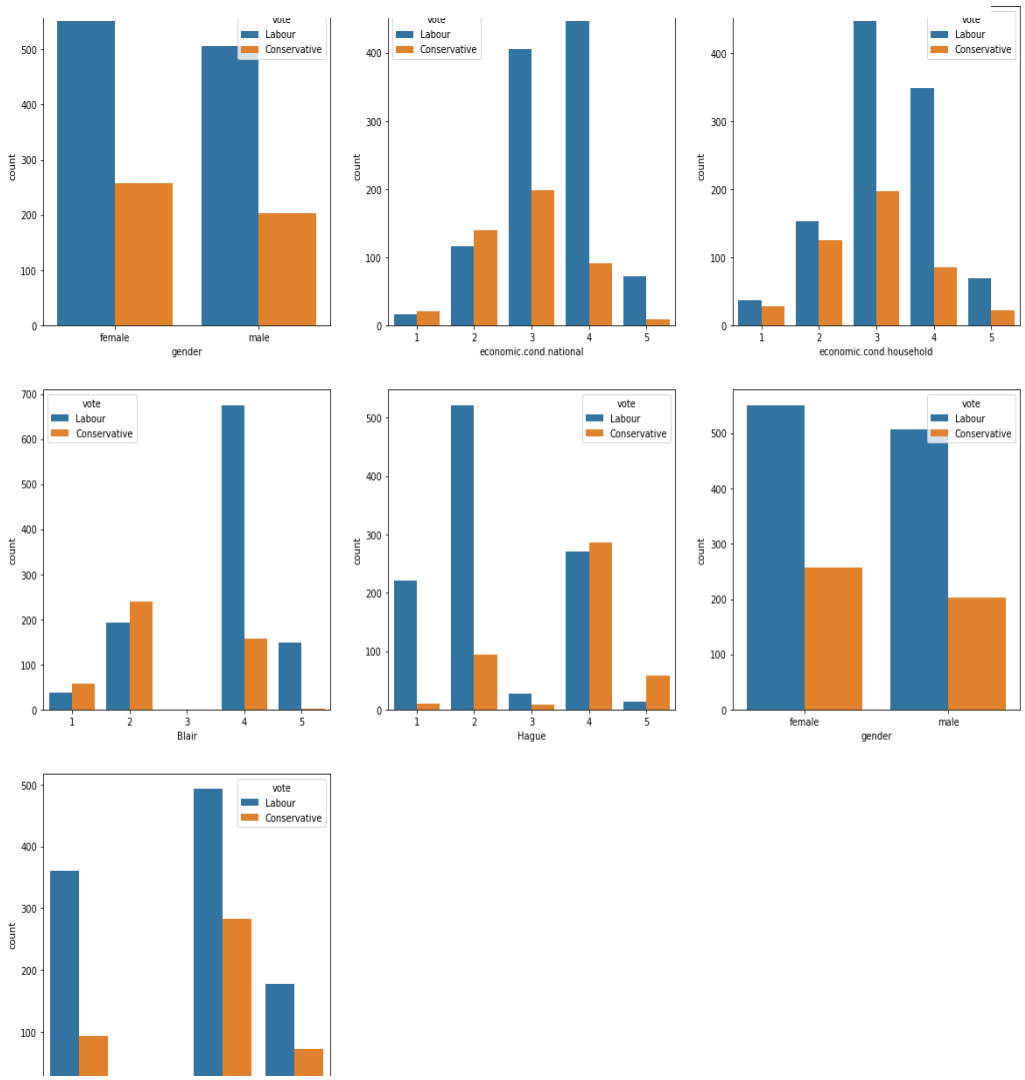


Fig. 2. Bivariate analysis on individual Variables

Inferences:

- Outlier is observed in' economic.cond.national' and the survey states there is a majority of neutral response.
- The survey is taken of majorly people within the age group 40-75.
- 'Blair' & 'Hague' states with very less people have neutral response, 'Blair' has a majority of positive responses, however 'Hague' has a majority of negative responses.
- After conversion of the categoric data 'Vote' which had two levels- 'Conservative' and 'Labour', it is observed that 'Labour' group has higher count.
- Almost equal contribution is observed from the gender factor.
- When comparing the parties with all the variables, labour party has high count.

Bivariate Analysis:

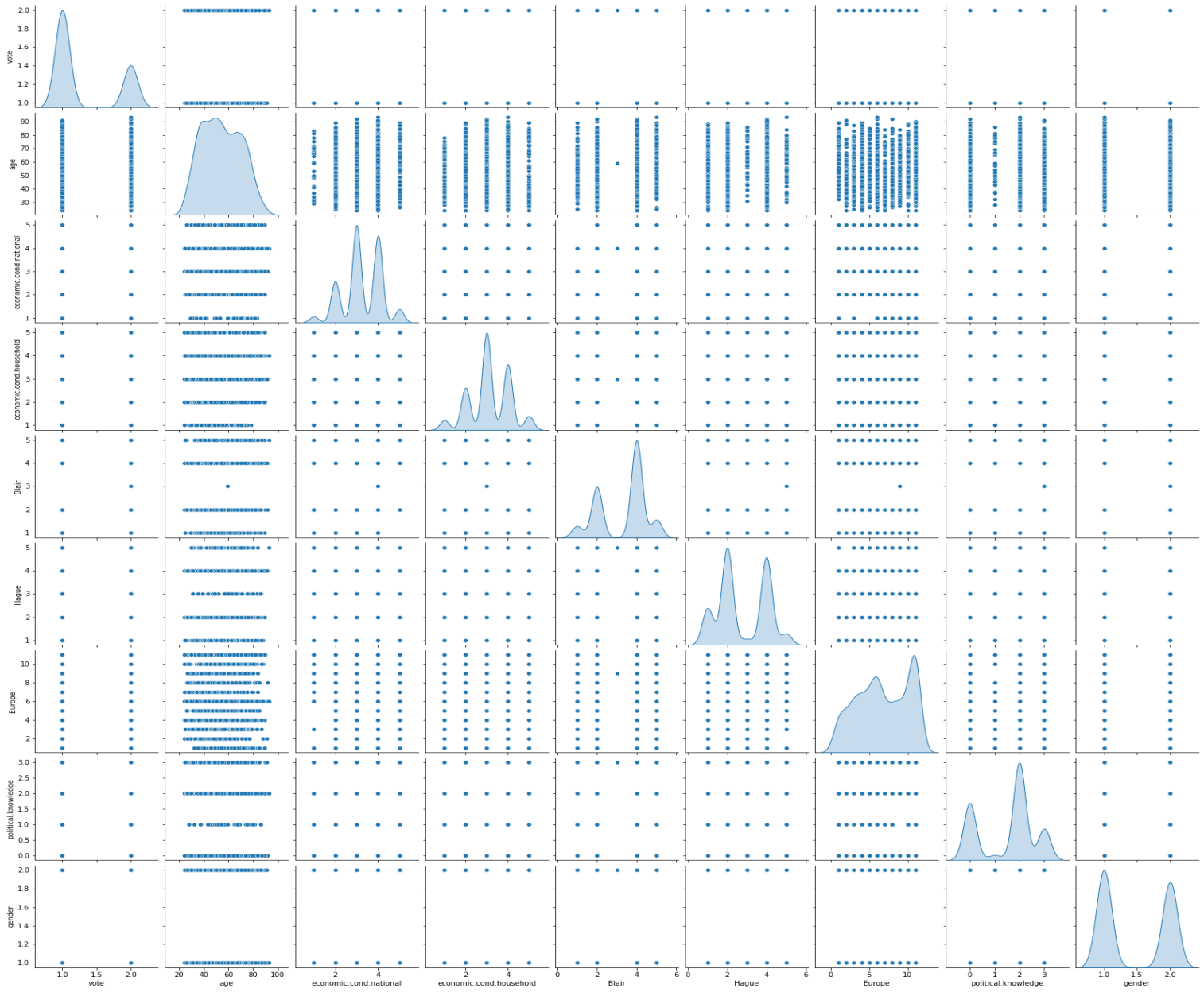


Fig. 4. Pair Plot

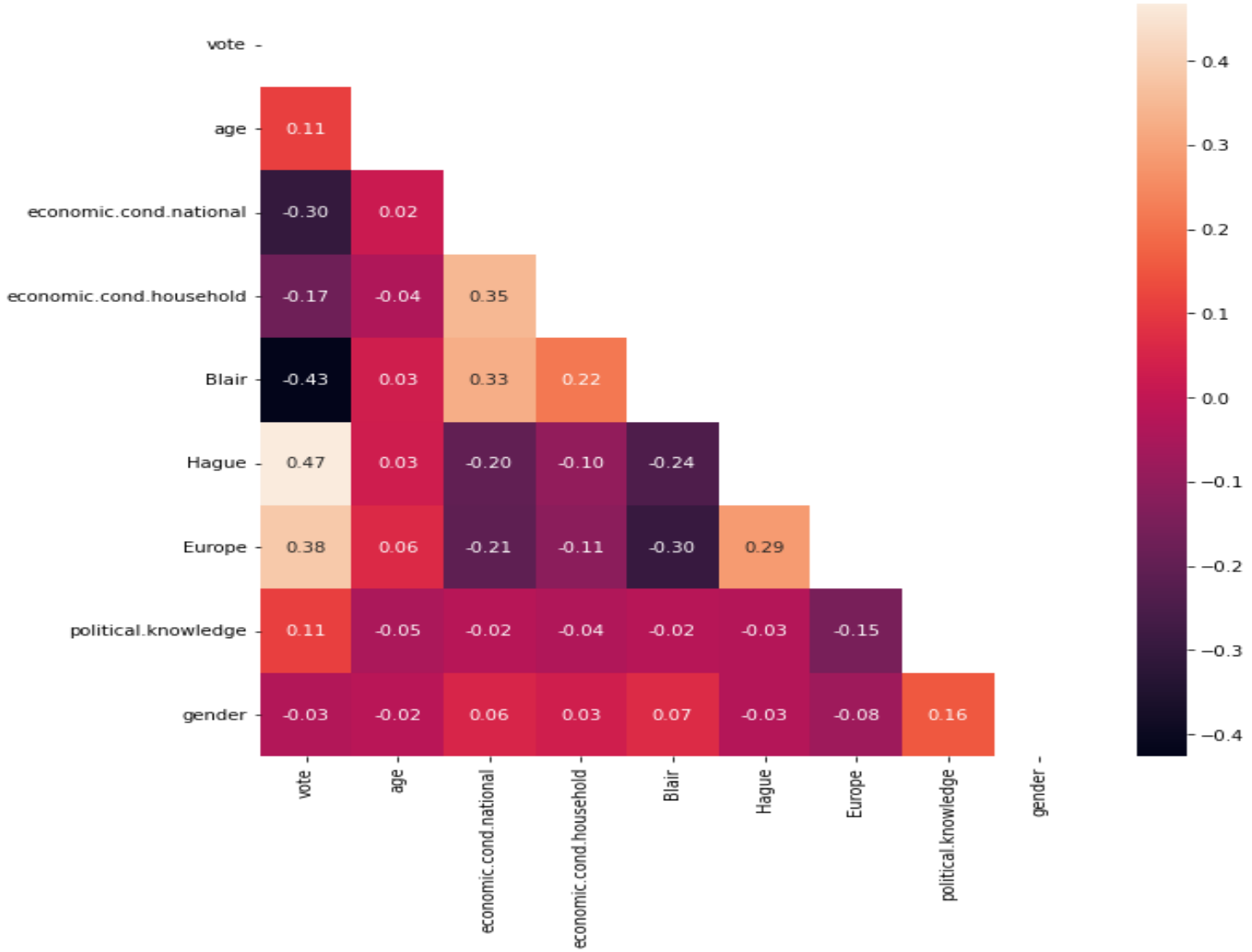


Fig. 5. Heat Map

- Outliers have been observed in 'economic.cond.national' &' economic.cond.household'.
- No correlation is observed between the variables.

1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? (2 pts), Data Split: Split the data into train and test (70:30) (2 pts).

Encoded the data by converting the object variables into categorical and then coded with levels using code function of python.

- Scaling is optional for models like Linear regression model, LDA &Logistic regression. However, for distance-based models like KNN scaling is required. Have scaled the data using Z score.
- Scaled data is used for only KNN model only
- After scaling the data is first divided into two variables 'X' & 'y' which includes the independent and dependant variables respectively. The dependant variable is our target variable.
- After the identification of the target variable the dataset is divided into train-test split with 70:30 proportion with random state 100.
- The training data set has 1061 rows with 8 features.

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

Logistic Regression:

Train Data Set results

	precision	recall	f1-score	support
0	0.75	0.68	0.71	322
1	0.87	0.90	0.88	739
accuracy			0.83	1061
macro avg	0.81	0.79	0.80	1061
weighted avg	0.83	0.83	0.83	1061

Fig. 6. Logistic Reression Train Result

Accuracy = 0.83

Results of logistic Regression for test data set as follows

	precision	recall	f1-score	support
0	0.76	0.66	0.71	138
1	0.86	0.91	0.88	318
accuracy			0.83	456
macro avg	0.81	0.78	0.79	456
weighted avg	0.83	0.83	0.83	456

Fig. 7. Test Result

Accuracy = 0.83

Confusion Matrix:

Fig. 8. Train vs Test Confusion Matrix

Inferences:

For the train and test, there is no difference between the accuracy. Model gives good result for predicting.

There is no overfitting or underfitting observed in the data.

Linear Discriminant Analysis:

Train:

	precision	recall	f1-score	support
0	0.74	0.68	0.71	322
1	0.87	0.90	0.88	739
accuracy			0.83	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.83	0.83	1061

Fig. 9. LDA Train Result

Accuracy = 0.83

Test:

	precision	recall	f1-score	support
0	0.75	0.70	0.72	138
1	0.87	0.90	0.89	318
accuracy			0.84	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.84	0.84	0.84	456

Fig. 10. LDA Test result

Accuracy = 0.84

Confusion Matrices:

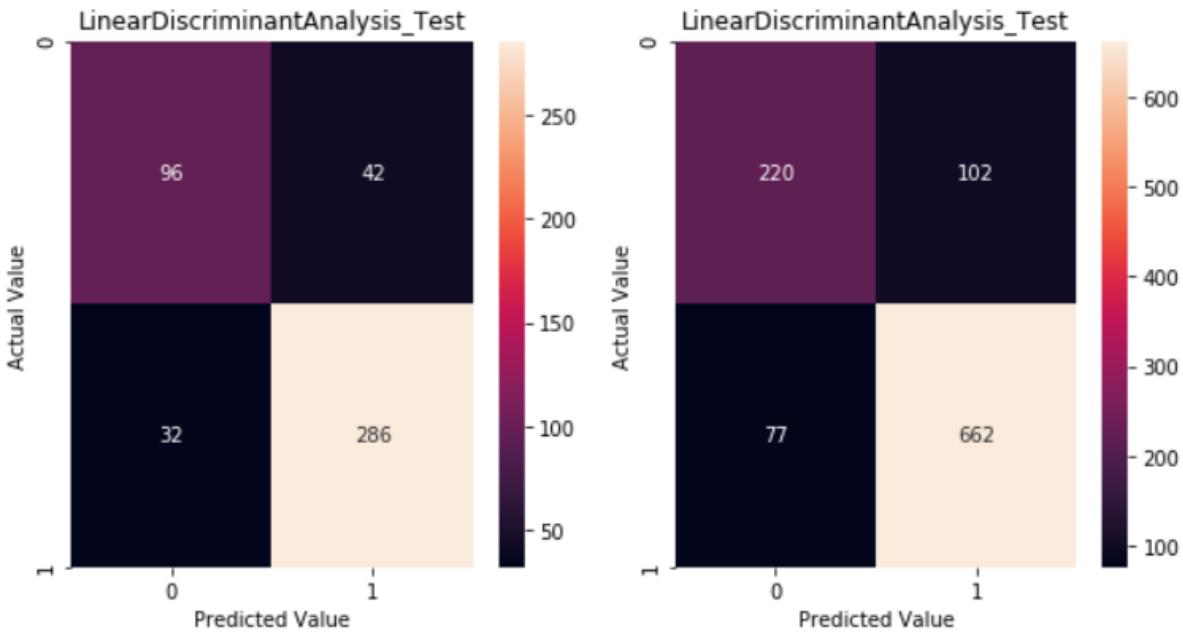


Fig. 11. LDA Confusion Matrix

Inferences:

The accuracy value of test is little higher than the train values. There might be a chance of underfitting for the given model.

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

Naïve Bayes Model:

Classification report for Train data, Accuracy = 0.84

	precision	recall	f1-score	support
0	0.73	0.73	0.73	322
1	0.88	0.88	0.88	739
accuracy			0.84	1061
macro avg	0.81	0.81	0.81	1061
weighted avg	0.84	0.84	0.84	1061

Fig. 12. Naive Base train Result

For Test Data:

Accuracy = 0.81

	precision	recall	f1-score	support
0	0.70	0.67	0.69	138
1	0.86	0.87	0.87	318
accuracy			0.81	456
macro avg	0.78	0.77	0.78	456
weighted avg	0.81	0.81	0.81	456

Fig. 13. Navie Bayes Test Result

Inferences:

There is a slight reduction in the accuracy values of Test compared to train,

Very low over fitting can be observed on the model.

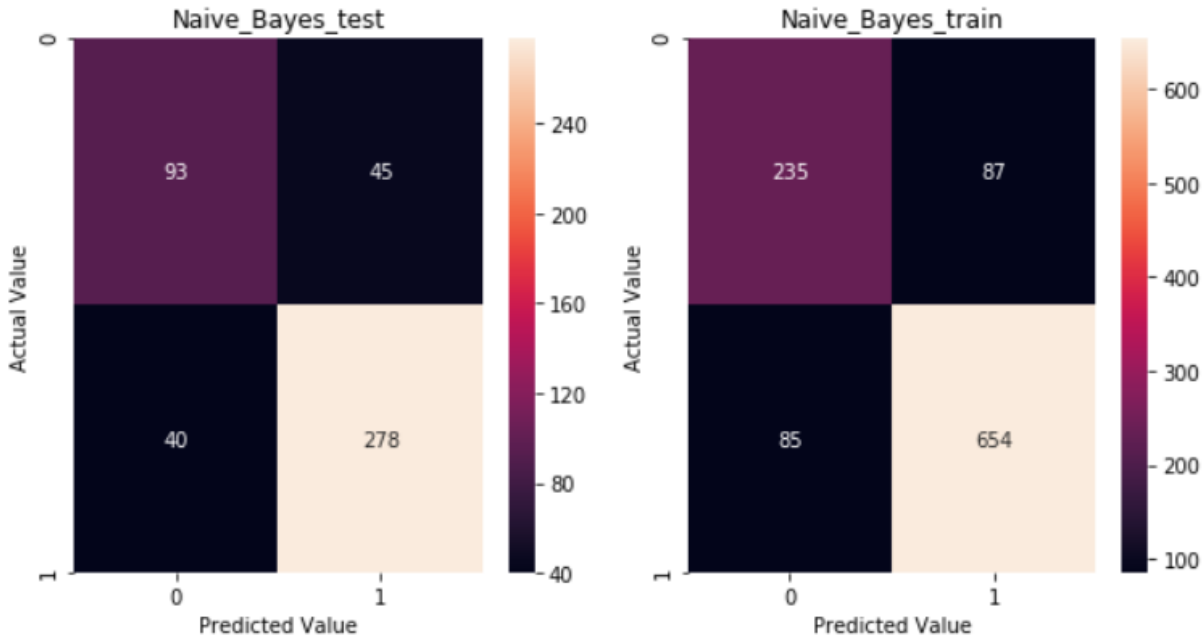


Fig. 14. Naive Bayes Confustion Matricess

KNN Model:

Train Accuracy = 1

Test Accuracy = 0.82

Train data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	322
1	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

Fig. 15. KNN Train

Test data:

	precision	recall	f1-score	support
0	0.71	0.68	0.69	138
1	0.86	0.88	0.87	318
accuracy			0.82	456
macro avg	0.79	0.78	0.78	456
weighted avg	0.82	0.82	0.82	456

Fig. 16. KNN Test

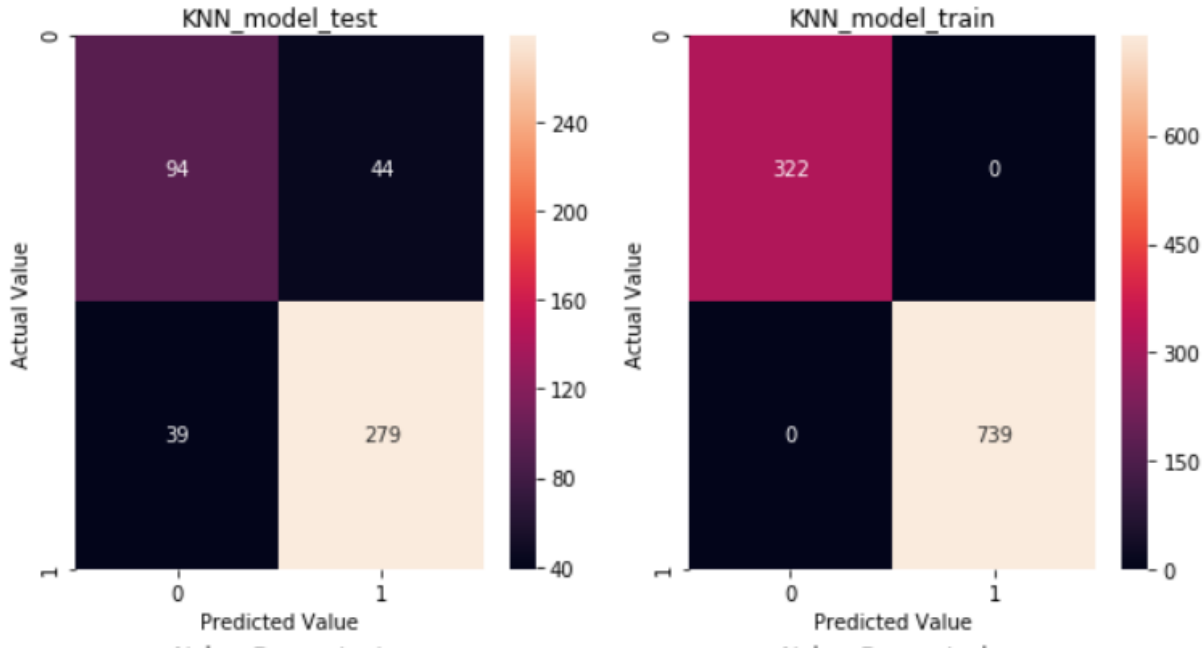


Fig. 17. KNN Confusion Matrices

There is a reduction in Accuracy values of Test data, there is some over fitting in the model can be observed.

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

Bagging:

Train Data

	precision	recall	f1-score	support
0	1.00	1.00	1.00	322
1	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

Fig. 18. Train result of Bagging

Accuracy Train = 1

Test Data

	precision	recall	f1-score	support
0	0.66	0.72	0.69	138
1	0.87	0.84	0.86	318
accuracy			0.80	456
macro avg	0.77	0.78	0.77	456
weighted avg	0.81	0.80	0.80	456

Fig. 19. Test Result of Bagging

Accuracy test = 0.8

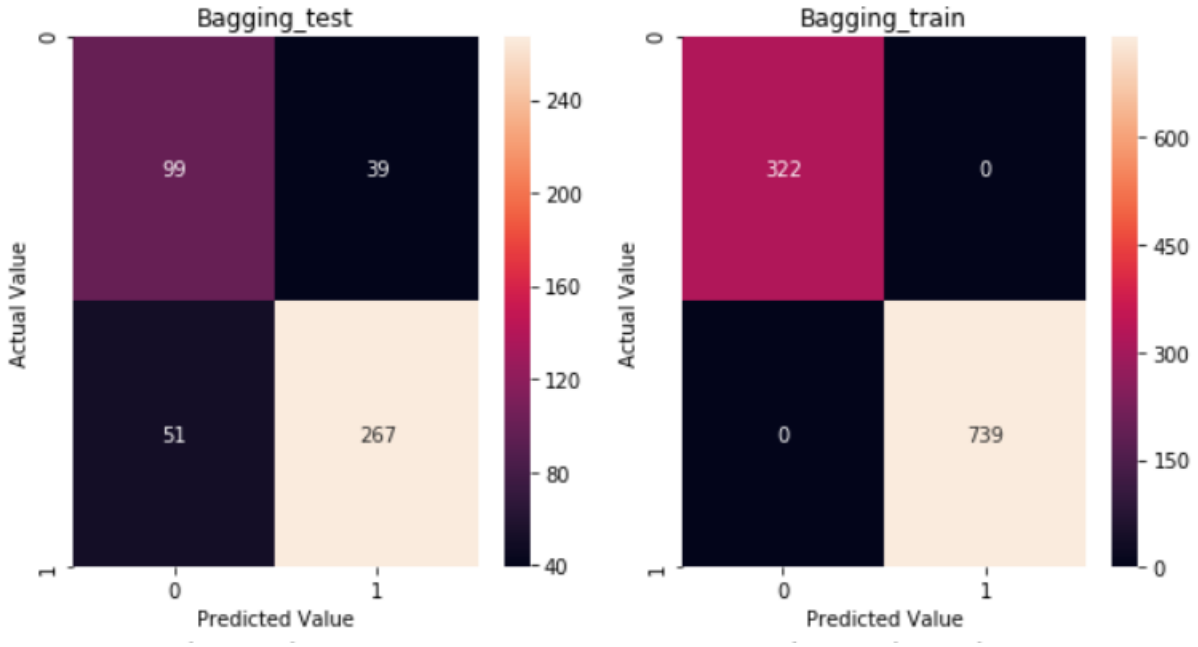


Fig. 20. Confusion matrix of Bagging

There is a reduction in the Accuracy value from train to test. There are chances for over fitting.

Boosting:

There are two types of boosting observed, One is Ada boosting, Gradient Boosting is the other one.

Ada Boosting:

Train data

	precision	recall	f1-score	support
0	0.77	0.70	0.74	322
1	0.88	0.91	0.89	739
accuracy			0.85	1061
macro avg	0.82	0.81	0.81	1061
weighted avg	0.84	0.85	0.84	1061

Fig. 21. Train Report

Test Data

	precision	recall	f1-score	support
0	0.71	0.66	0.68	138
1	0.86	0.88	0.87	318
accuracy			0.82	456
macro avg	0.78	0.77	0.78	456
weighted avg	0.81	0.82	0.81	456

Fig. 22. Test Report

Test Accuracy = 0.82

Train Accuracy = 0.85

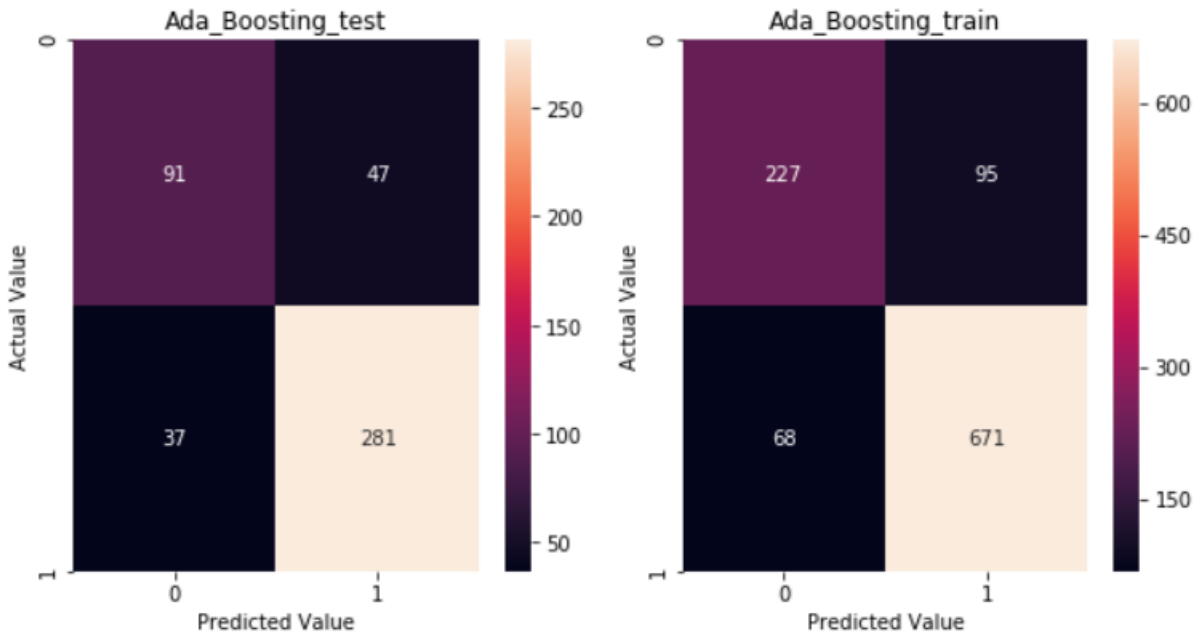


Fig. 23. Confusion Matrix

A very less over fitting is observed as per accuracy values.

Gradient Boosting:

Train Data

	precision	recall	f1-score	support
0	0.85	0.79	0.82	322
1	0.91	0.94	0.92	739
accuracy			0.89	1061
macro avg	0.88	0.86	0.87	1061
weighted avg	0.89	0.89	0.89	1061

Fig. 24. Train Result

Test Data

	precision	recall	f1-score	support
0	0.72	0.68	0.70	138
1	0.86	0.88	0.87	318
accuracy			0.82	456
macro avg	0.79	0.78	0.79	456
weighted avg	0.82	0.82	0.82	456

Fig. 25. Test Result

Accuracy Test = 0.82

Accuracy Train = 0.89

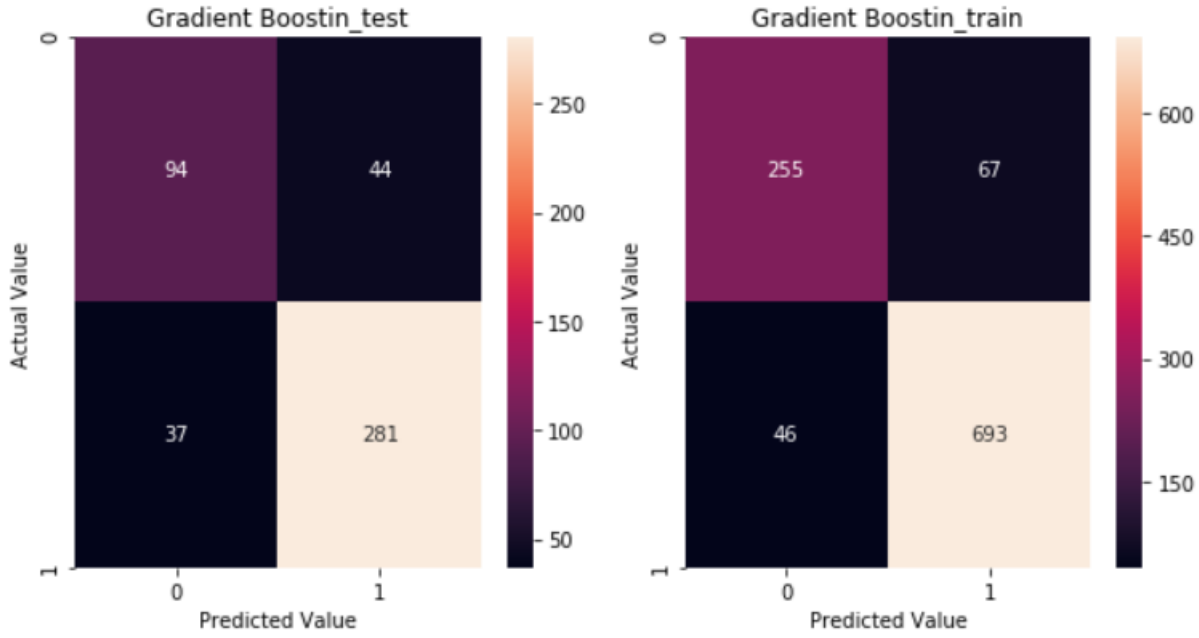


Fig. 26. Confusion Matrix

A slight reduction in accuracy values is observed between train and test.

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

ROC curves:

Logistic Regression:

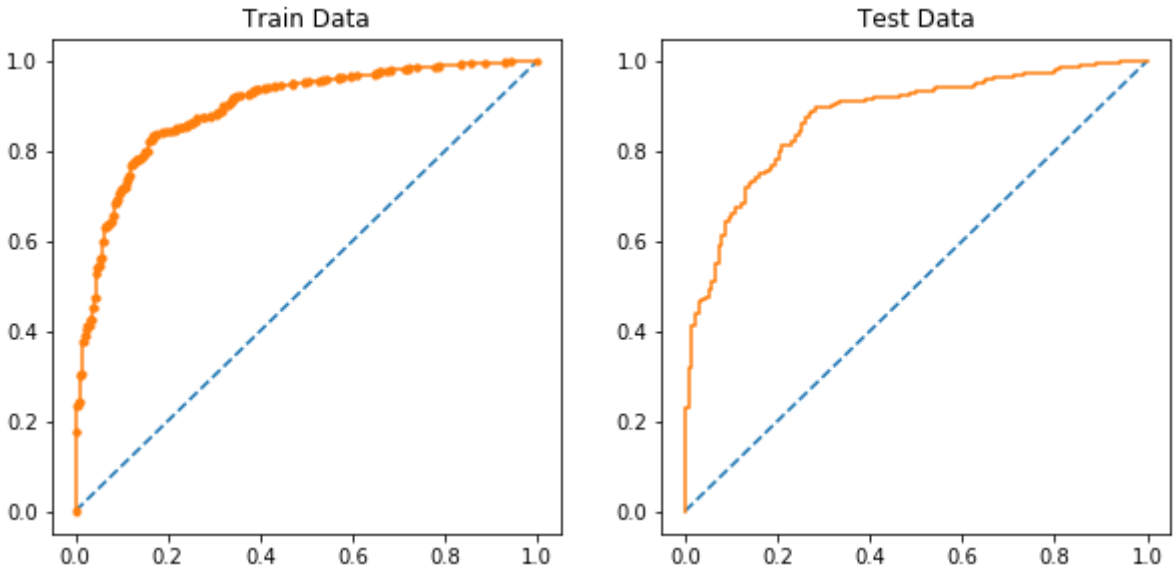


Fig. 27. ROC_Logistic Regression

ROC AUC, Test = 0.87 ; Train = 0.89

Linear Discriminant Analysis:

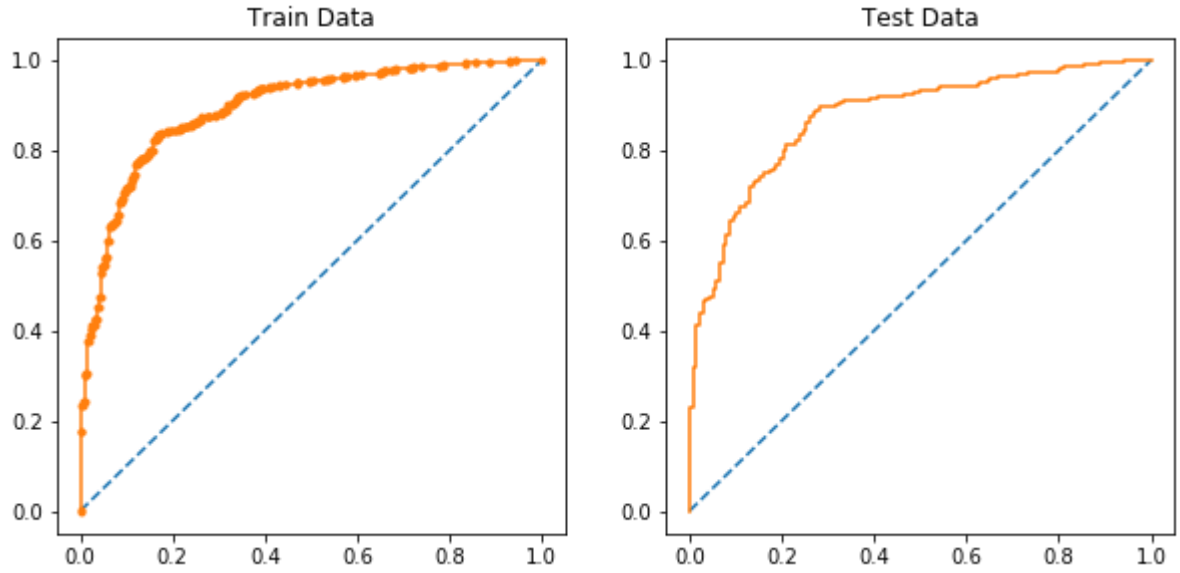


Fig. 28. ROC_ LDA

ROC AUC, Train = 0.89

Test = 0.87

KNN:

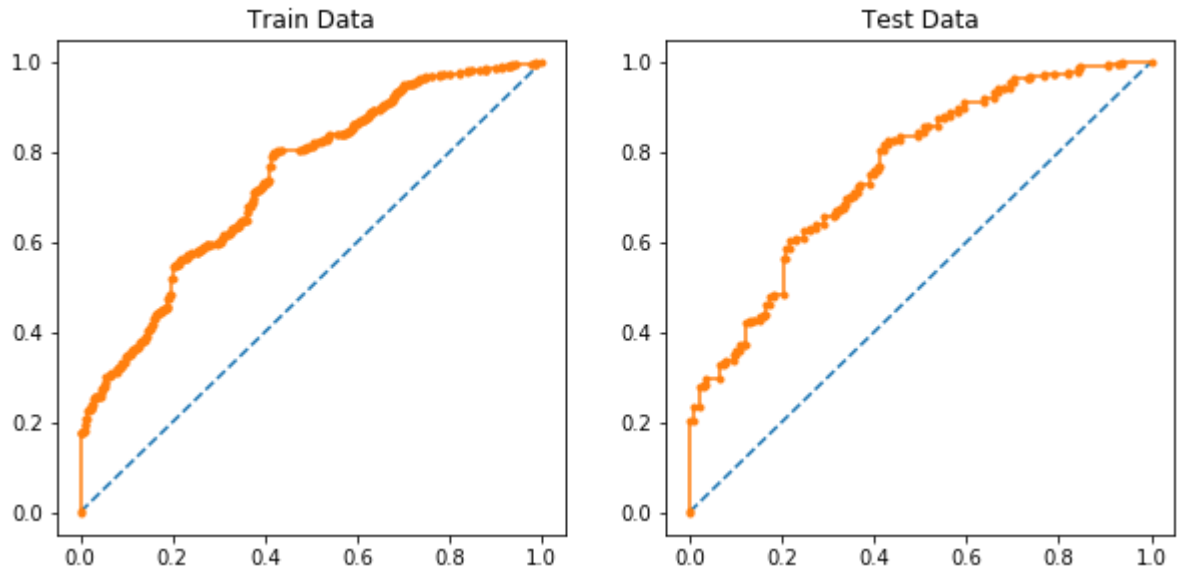


Fig. 29. ROC_KNN

ROC AUC, Train = 0.73 ; Test = 0.76

Naïve Bayes:

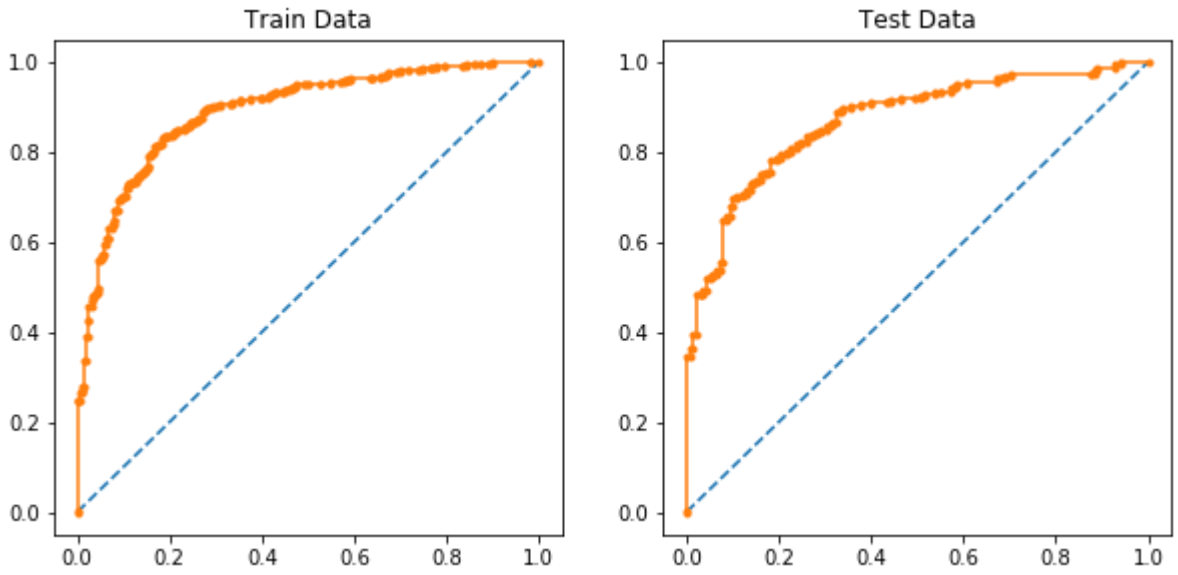


Fig. 30. ROC_Naive Bayes

ROC AUC, Train = 0.89 ; Test = 0.87

Bagging:

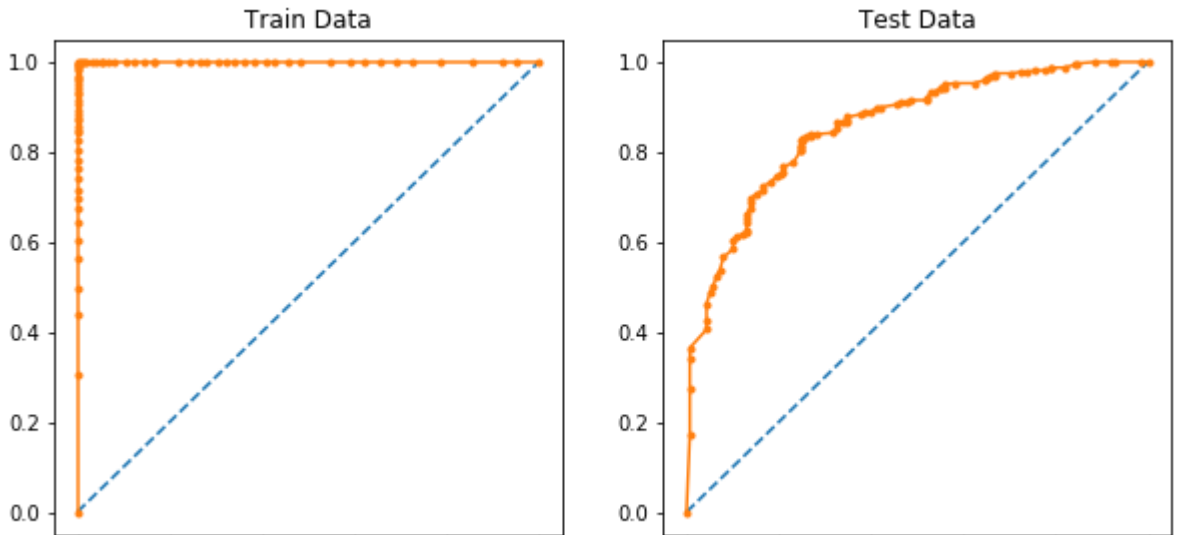


Fig. 31. ROC_Bagging

ROC AUC, Train = 1, Test = 0.86

Ada Boosting:

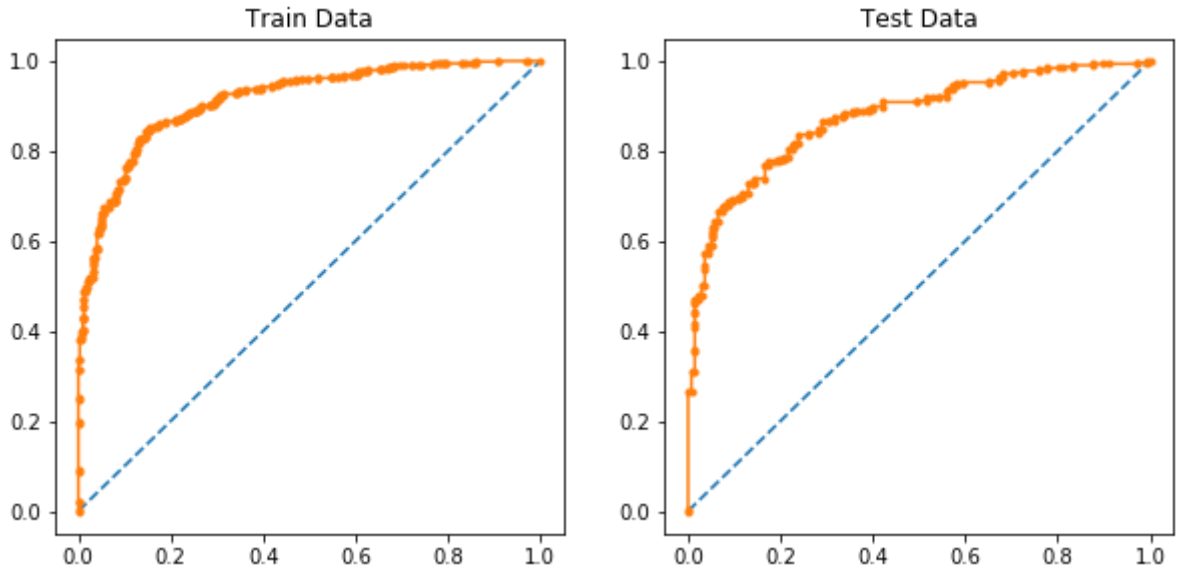


Fig. 32. ROC_Ada Boosting

ROC AUC, Train = 0.91 ; Test = 0.88

Gradient Boosting:

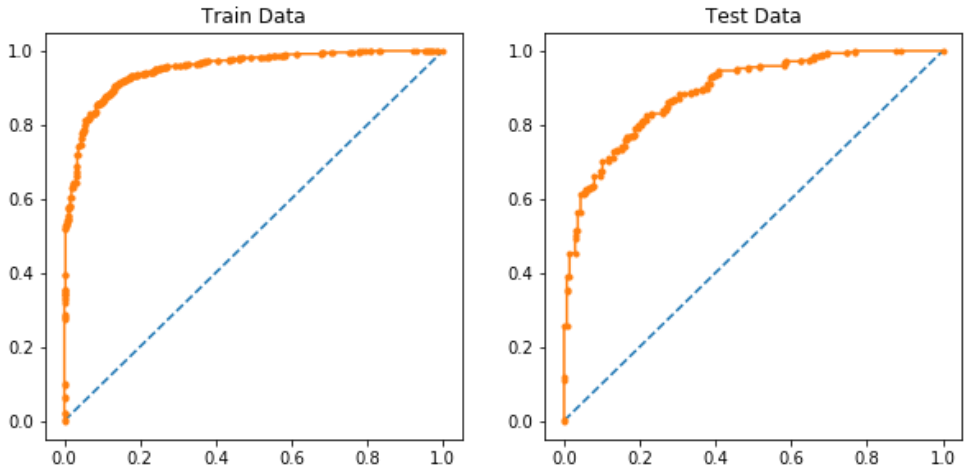


Fig. 33. ROC_Gradient Boosting

ROC AUC, Train = 0.95, Test = 0.89

	AUC_TEST	AUC_TRAIN
Logistic Regression	0.873849	0.894263
LDA	0.874054	0.894557
KNN	0.760642	0.738630
Naive Bayes	0.870499	0.890829
Bagging	0.859060	1.000000
Ada Boosting	0.876379	0.913777
Gradient Boostin	0.891282	0.951065
Random Forest	0.870796	0.889151

Fig. 34. AUC values of test and Train w.r.t Models

1.8 Based on these predictions, what are the insights?

> The model performance depends on the input data and their distributions.

> Bagging and Boosting models are performed well as per the accuracy values seen in the model results.

Problem 2:

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

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

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

- To find the characters we have used the raw() function.
- To find the words in the documents we have used the words() function.
- To find the sentences in the documents we have used the sents() function.
- The output of the same is as follows:

	char_count	word_count	sent_count
1941-Roosevelt	7571	1350	68
1961-Kennedy	7618	1370	52
1973-Nixon	9991	1819	69

Table 6. Count of Characters, Words, Sentences

2.2 Remove all the stop words from all three speeches.

- Stop words are the words which occur most frequently and have no significance in the result. Hence we have extracted the stop words from all the three texts.
- Before removing the stop words we have collected a list of stopwords from nltk and extracted them from the main text.

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

- After removing of the stop words the words which occur most no of times for each president is as follows:

President	Most Occurring Word
Roosevelt	'nation': 12, 'know': 10, 'spirit': 9,
Kennedy	'let': 16, 'us': 12, 'world': 8
Nixon	'us': 26, 'let': 22, 'america': 21

Table 7. Most occuring words

The above values we got from one of the function of nltk i.e.., FreqDist.

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

Word Cloud of the speeches of the variables after removing the stopwords are as follows:

Fig. 35. Roosevelt Speech's words

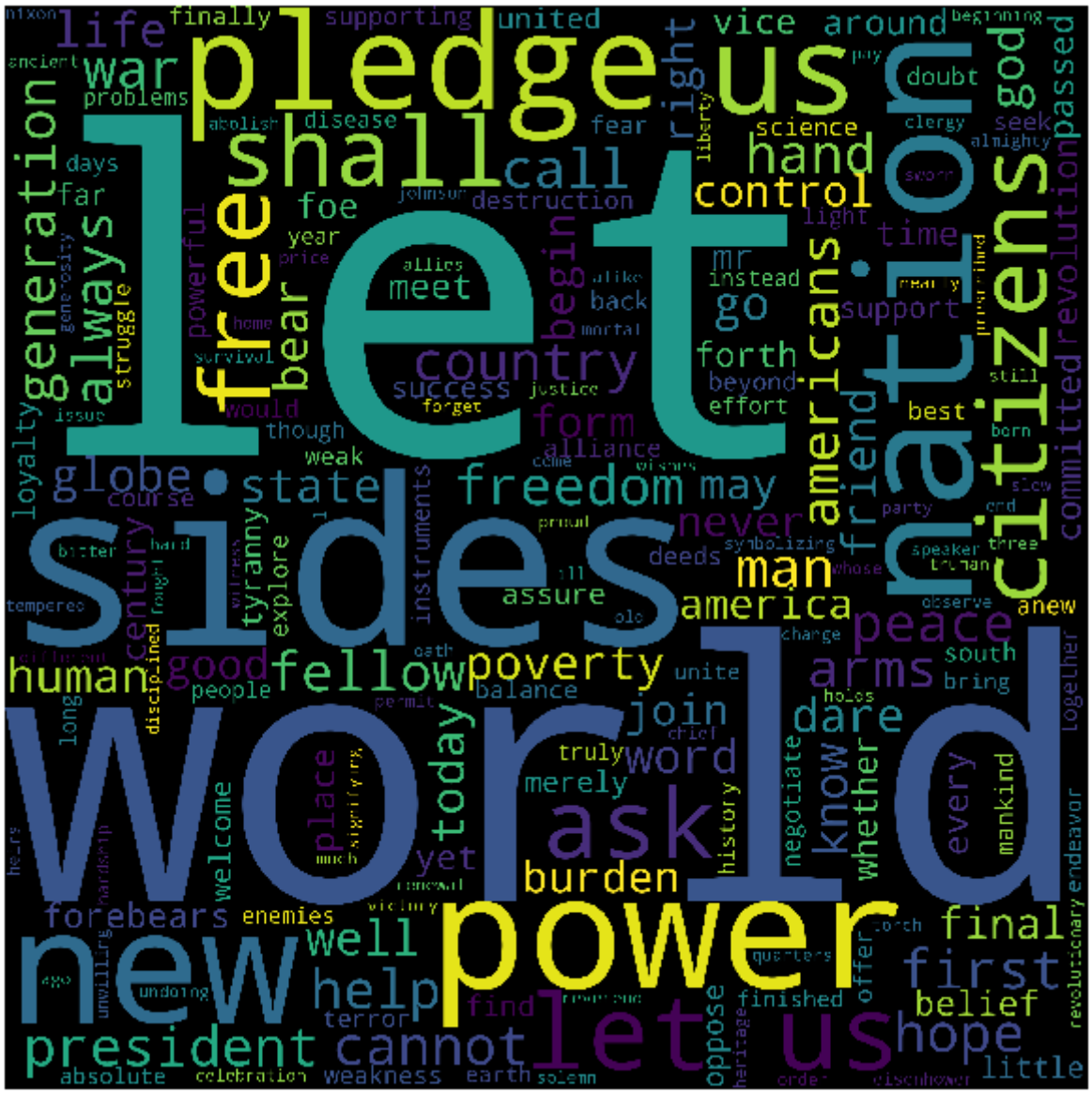


Fig. 36. Kennedy's Speech's words

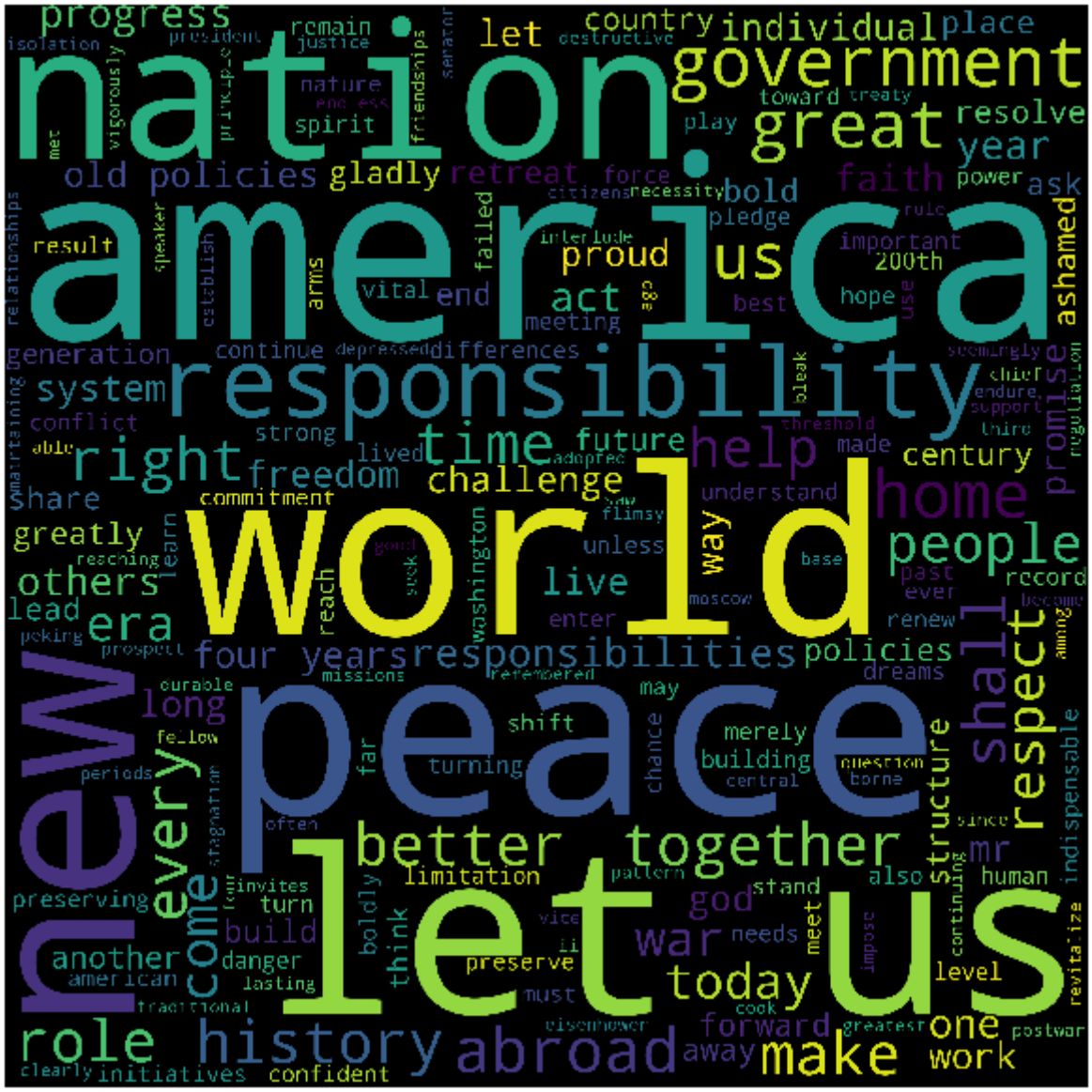


Fig. 37. Nixon Words