Machine Learning-2

2024

PGP DSBA PROGRAM

by: ABHISHEK K HIREMATH

Great Learning
POWER AHEAD

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Problem 1:

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

Dataset for Problem: Election_Data.xlsx

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

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

Solution:

- Loaded the required packages and read the dataset.
- Remove the unwanted column from the original dataset has it is a serial number.
- Dataset has 1525 observations and 9 features including the target variable.

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

Fig: -1: Dataset head

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

Fig: -2: Dataset tail

- To understand the data, need to look into data structure, summary statistics, skewness and missing values of the data features:

Information about the Dataset:

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

Fig: -3: Dataset Info

The dataset has 9 Features: 7 numerical type & 2 object type, 1525 records.

Inference:

- The column "Unnamed: 0" is removed from the dataset before proceeding further as its insignificant for the analysis.
- There are 1525 rows and 9 columns
- Numerical Columns: age, economical_cond_national, economical_cond_household, Blair, Hague, Europe and political_knowledge.
- Non-Numerical Columns: vote and gender.
- There are no null values Summary of the dataset.

Missing values check: There are no missing values in the dataset.

```
vote                        0
age                         0
economic.cond.national      0
economic.cond.household     0
Blair                       0
Hague                       0
Europe                      0
political.knowledge         0
gender                      0
dtype: int64
```

Fig: -4: Check missing values

Duplicate records check: There are 8 duplicate records in the dataset which have been dropped.

Summary of the Dataset:

	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

Fig: -5: 5-point Summary

- Data quality**:** No **instances of "bad" or corrupt data were found.**
- Economic condition ratings**: Both national and household levels** lean positively towards Tony Blair **(ratings >3), indicating public approval of his economic stewardship.**
- Leadership perception**: William Hague's rating** falls below **the neutral score,** signifying less public confidence **in his leadership.**
- Skewness**: All variables show** skewness values within |0.5|, **suggesting** minimal data asymmetry**, implying data can be treated as** essentially unscrewed.

	vote	gender
count	1517	1517
unique	2	2
top	Labour	female
freq	1057	808

- Categorical variables**: Two levels present –** Labour and gender**.**
- Labour frequency**:** Higher **representation with** 1057 **occurrences out of** 1517 **in the target variable.**

1.2 Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers.

Interpret the inferences for each Distribution plots(histogram) or similar plots for the continuous columns. Box plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

For univariate & bivariate analysis we treat the rating variables from survey (economic.cond.national, economic.cond.household, Blair, Hague, Europe, political.knowledge) as ordinal categorical variables or discrete numerical variables as when necessary.

Univariate Analysis:

We first take a look at the numerical variables.

1. **age: Inyears**

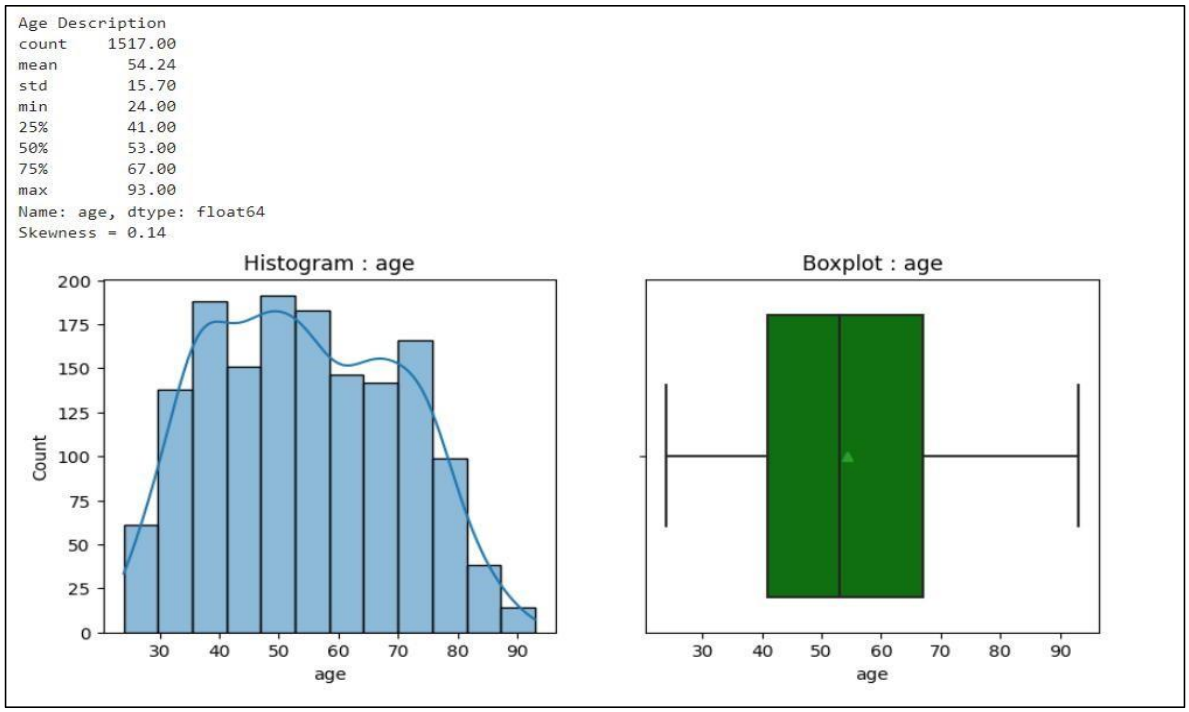


Fig: -6: Univariate Analysis – age

Age, ranging from 24 to 93, has a slight right skew but can be considered normally distributed for analysis due to minimal skewness and a wavy peak.

Its mean (54.24) slightly exceeds the median (53), indicating this minor skewness. No outliers are present.

2. **economic.cond.national: Assessment of current national economic conditions, 1 to 5 (1 being the lowest rating, 5 being the highest rating).**

```
economic.cond.national %age Count
3     39.82
4     35.46
2     16.88
5      5.41
1      2.44
Name: economic.cond.national, dtype: float64
Average Score = 3.25
```

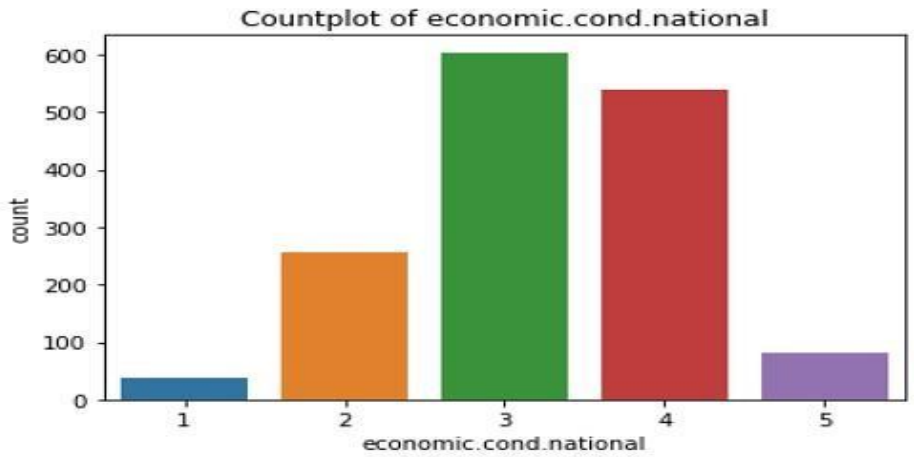


Fig: -7: Univariate Analysis – economic.cond.national

3. **economic.cond.household:** Assessment of current household economic conditions, 1 to 5 (1 being the lowest rating, 5 being the highest rating).

```
economic.cond.household %age Count
3    42.52
4    28.68
2    18.46
5     6.06
1     4.28
Name: economic.cond.household, dtype: float64
Average Score = 3.14
```

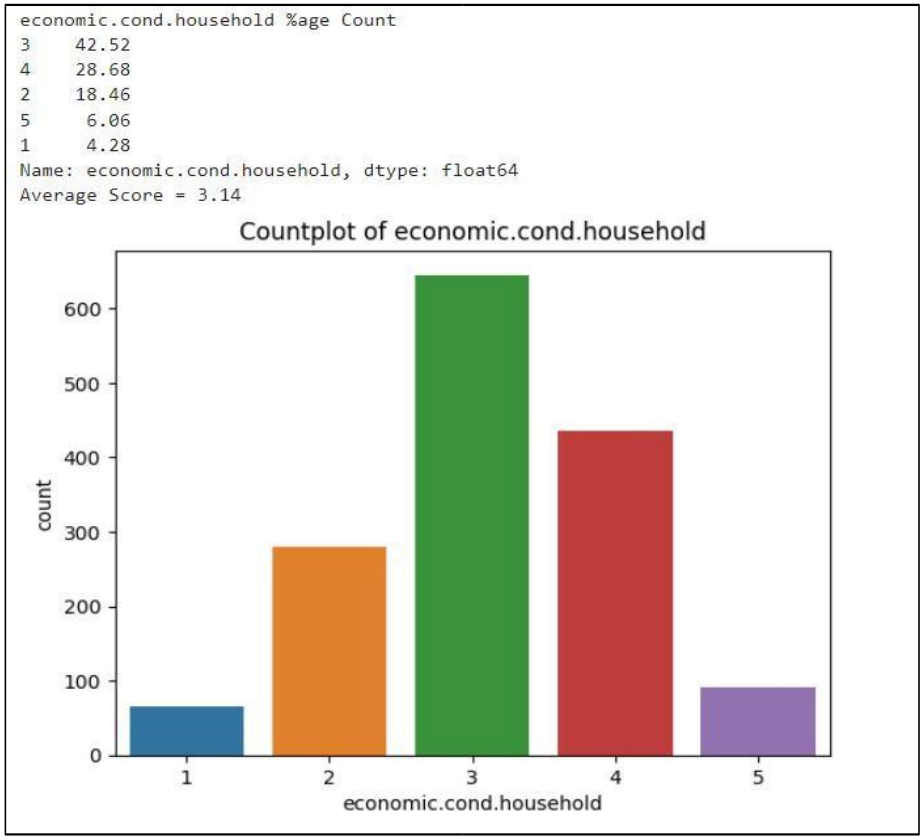


Fig: -8: Univariate Analysis – economic.cond.household.

4. Blair: Assessment of the Labour leader Tony Blair, 1 to 5.

```
Blair %age Count
4    54.91
2    28.61
5    10.02
1     6.39
3     0.07
Name: Blair, dtype: float64
Average Score = 3.34
```

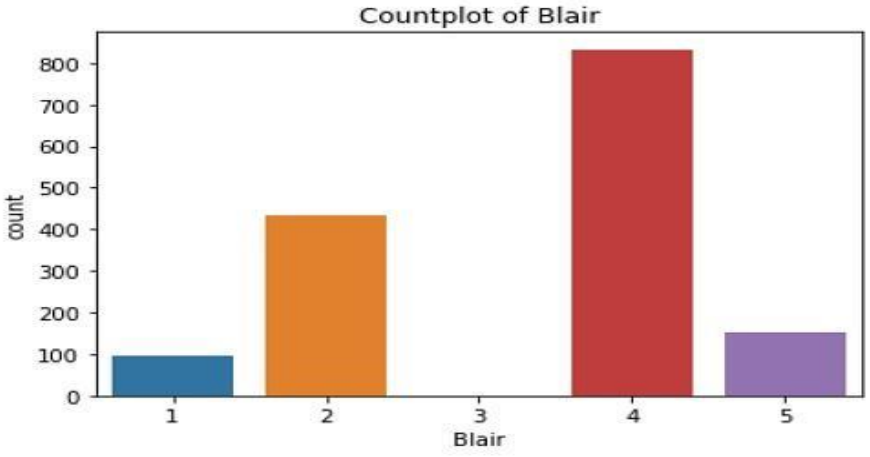


Fig: -9: Univariate Analysis – Blair

5. **Hague:** Assessment of the Conservative leader William Hague, 1 to 5.

```
Hague %age Count
2    40.67
4    36.72
1    15.36
5     4.81
3     2.44
Name: Hague, dtype: float64
Average Score = 2.75
```

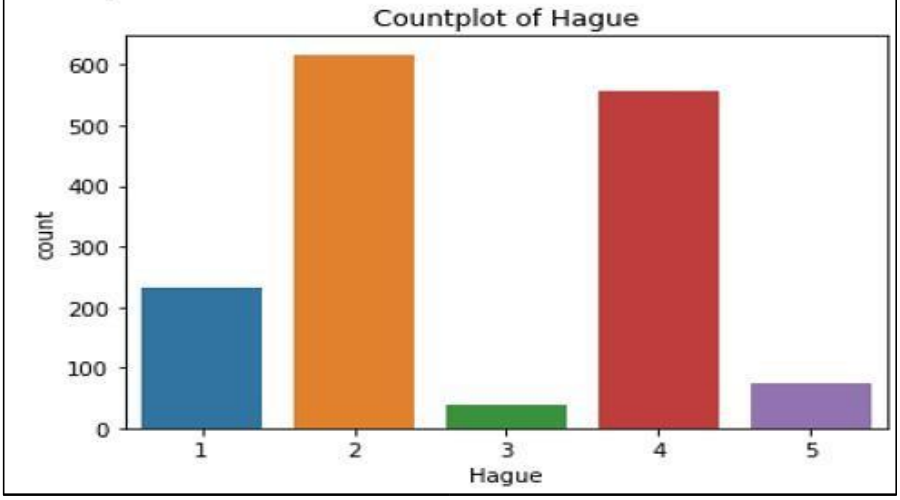


Fig: -10: Univariate Analysis – Hague

6. **Europe:** An 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment

```
Europe %age Count
11     22.28
6      13.65
3       8.44
4       8.31
5       8.11
9       7.32
8       7.32
1       7.19
10      6.66
7       5.67
2       5.08
Name: Europe, dtype: float64
Average Score = 6.74
```

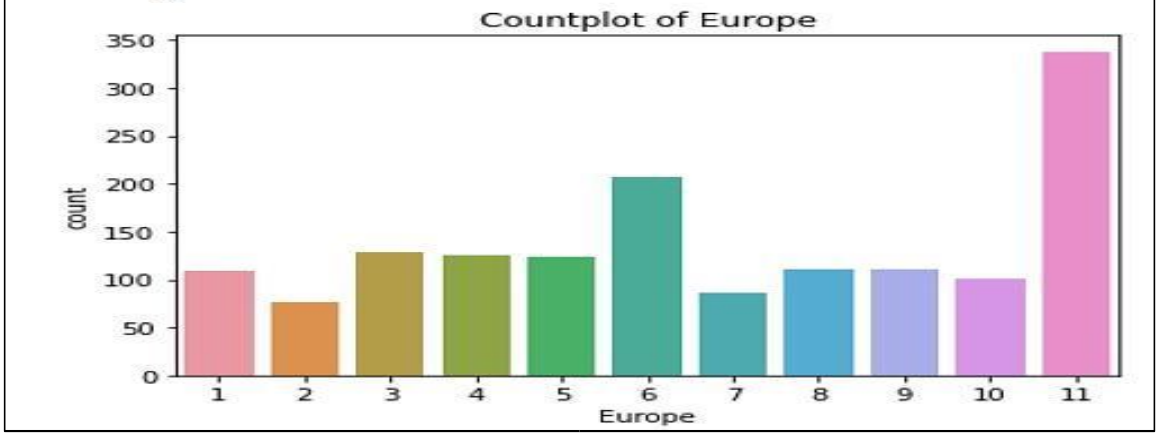


Fig: -11: Univariate Analysis – Europe

7. **Political.knowledge:** Knowledge of parties' positions on European integration, 0 to 3.

```
political.knowledge %age Count
2    51.15
0    29.93
3    16.41
1     2.50
Name: political.knowledge, dtype: float64
Average Score = 1.54
```

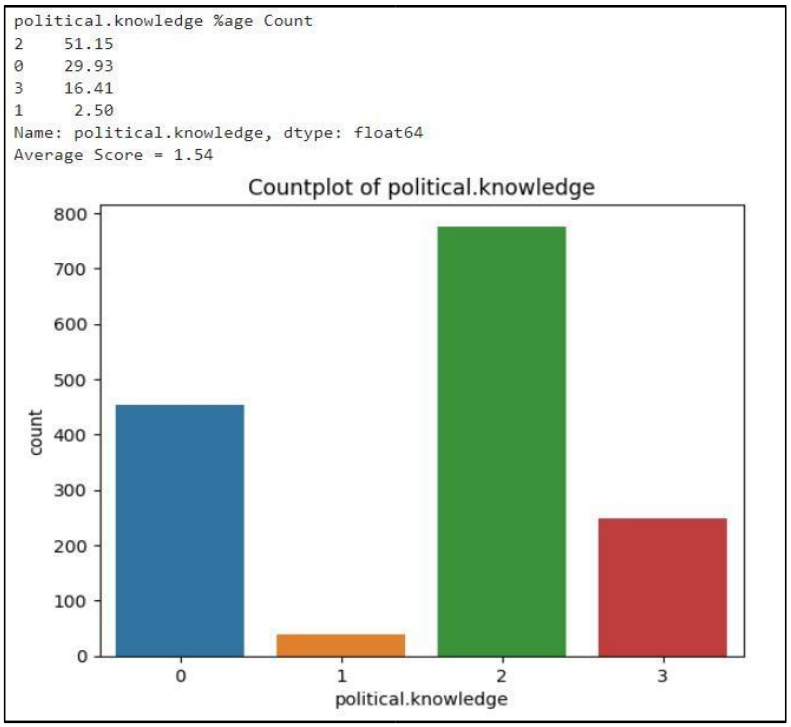


Fig: -12: Univariate Analysis – political.knowledge

8. **vote:** Party choice - Conservative or Labour [Target Variable].

```
vote %age Count
vote
Labour          69.68
Conservative    30.32
Name: proportion, dtype: float64
```

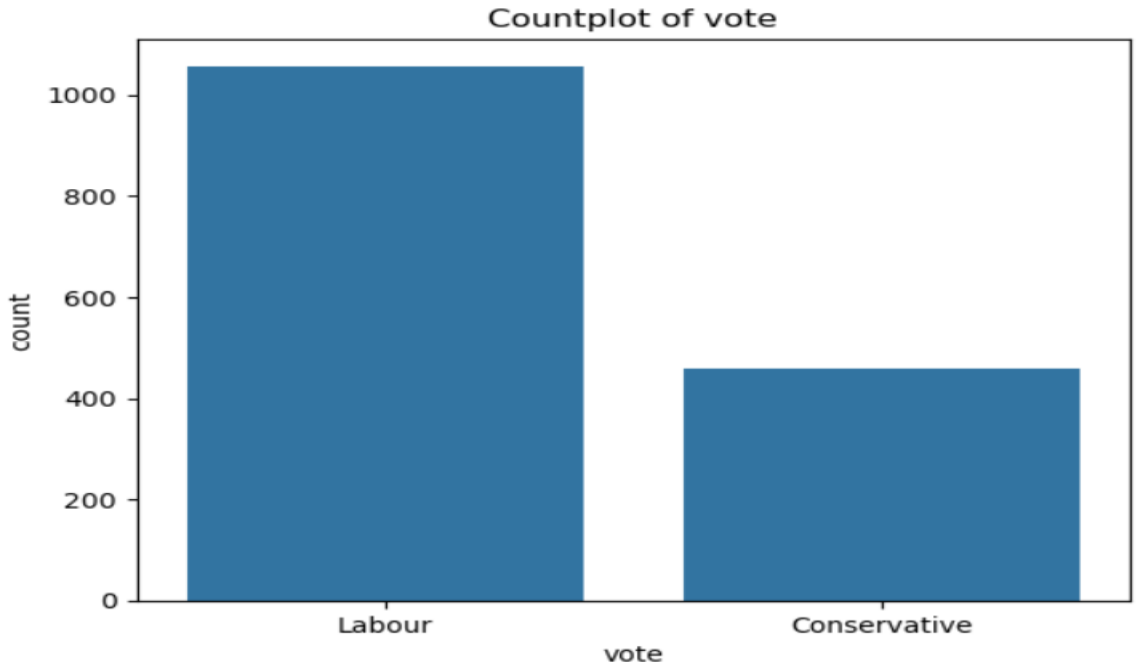


Fig: -13: Univariate Analysis – Vote

9. **Gender:** Male/Female.

```
gender %age Count
gender
female    53.26
male      46.74
Name: proportion, dtype: float64
```

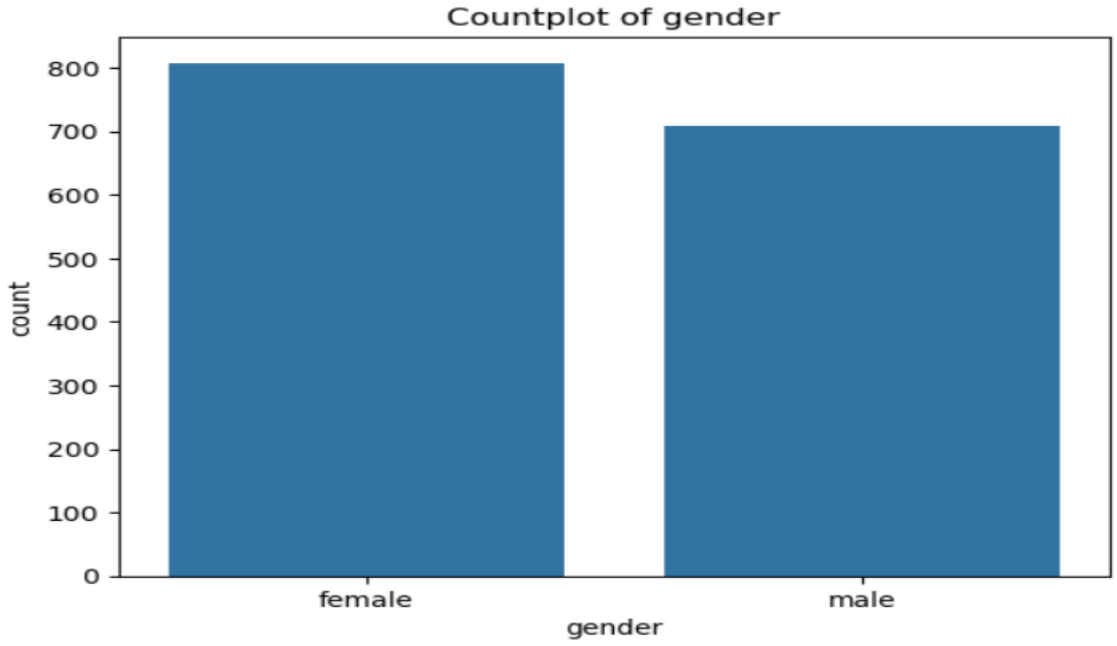


Fig: -14: Univariate Analysis – gender

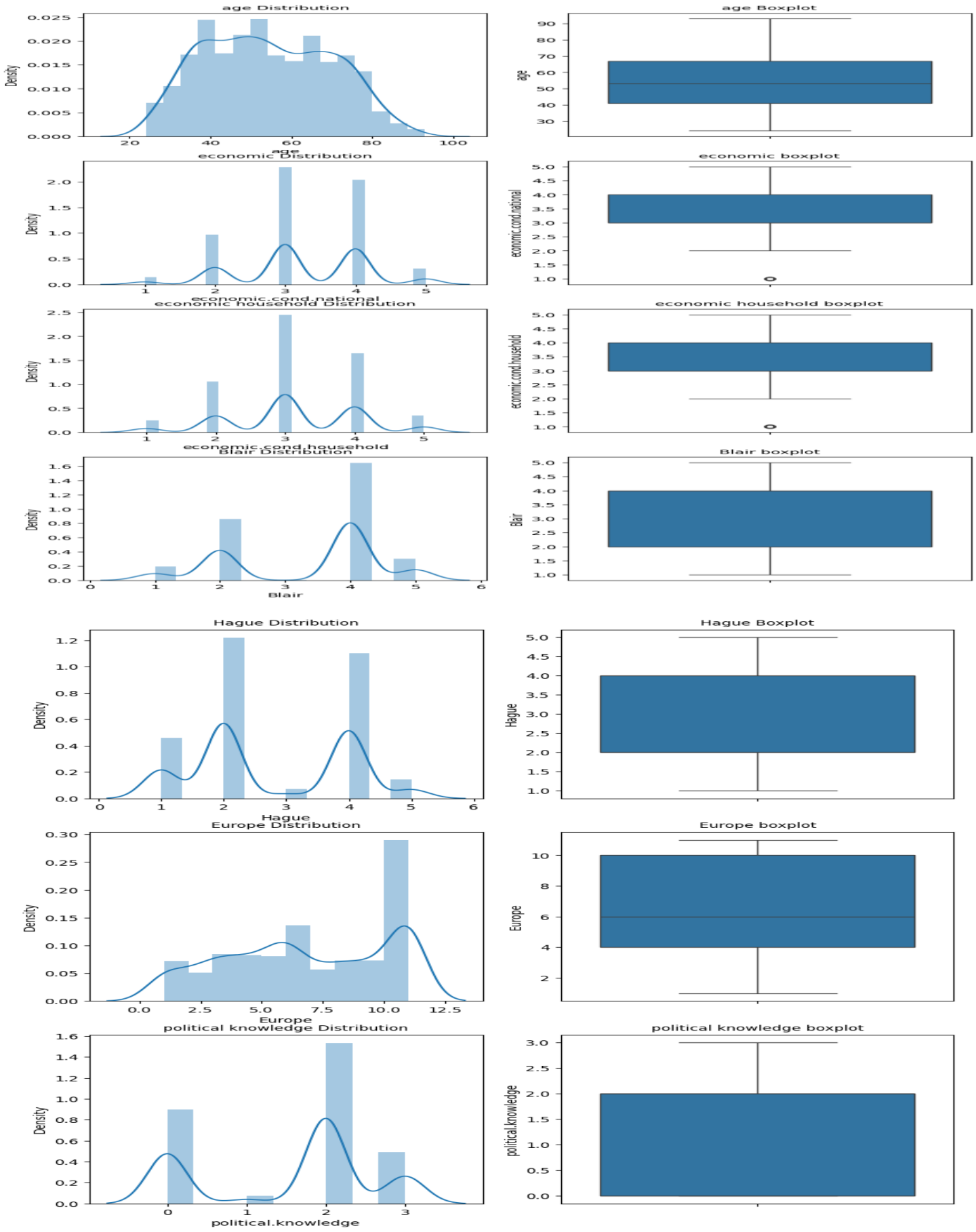


Fig: -15: Univariate Analysis (boxplot And Histplot)

Inferences from above:

- age - Data is normally distributed and the mean and median are almost same. The mean for age is 54.18 and the median is 53. The minimum age of the labour leader is 24 and the maximum age of the labour leader is 93. No outlier is detected.
- National economic condition **received a rating of** 3 or 4 from ~75% **respondents,** average score: 3.25.
- Household economic condition **rated** 3 or 4 **by** ~70% of individuals**,** average score: 3.14**.**
- Blair's **favorable rating by** ~65% respondents**,** average score: 3.34**, suggests satisfaction with** Labour party.
- Hague rated poorly **by** ~55% **respondents,** average score: 2.75**.**
- Average score of 6.74 **indicates** majority **leaning towards** Brexit**, with** 22% strongly favoring**.**
- 30% respondents **unaware of their party's stance on** European Integration**, while** 50% are well-informed**.**
- **Survey of 1500 individuals show** ~70% support for Labour Party**, ~30% for Conservative Party.**
- Slight class imbalance noted in target variable, but no drastic underrepresentation of Conservative class. Over/under sampling techniques not required.
- Europe - Data is normally distributed and the mean and median are almost same. The mean is 6.72 and the median is 6. The rating of respondents' attitudes toward European integration lies from 1 to 11. No outlier is detected.
- political.knowledge- Data is normally distributed and the mean and median are almost same. There is no left skew. The mean is 1.54 and the median is 2. The rating of knowledge of parties' positions on European integration lies from 0 to 3. No outlier is detected.is detected.

Bivariate Analysis:

To analyze different variable types, we use specific techniques:

Numerical Variables:

- Pair plot: Visualize the relationship between two numerical variables.
- Correlation matrix heatmap: Assess the correlation between numerical variables.

Categorical Variables:

- Cross tables: Examine the relationship between two categorical variables.
- Bar plots: Visualize the distribution and frequencies of categories in each variable.

Numerical vs. Categorical Variables:

- Boxplots, violin plots, or bar plots: Compare the distribution of a numerical variable across categories of a categorical variable.

Numerical VS Numerical Variable:

Pair Plot: -

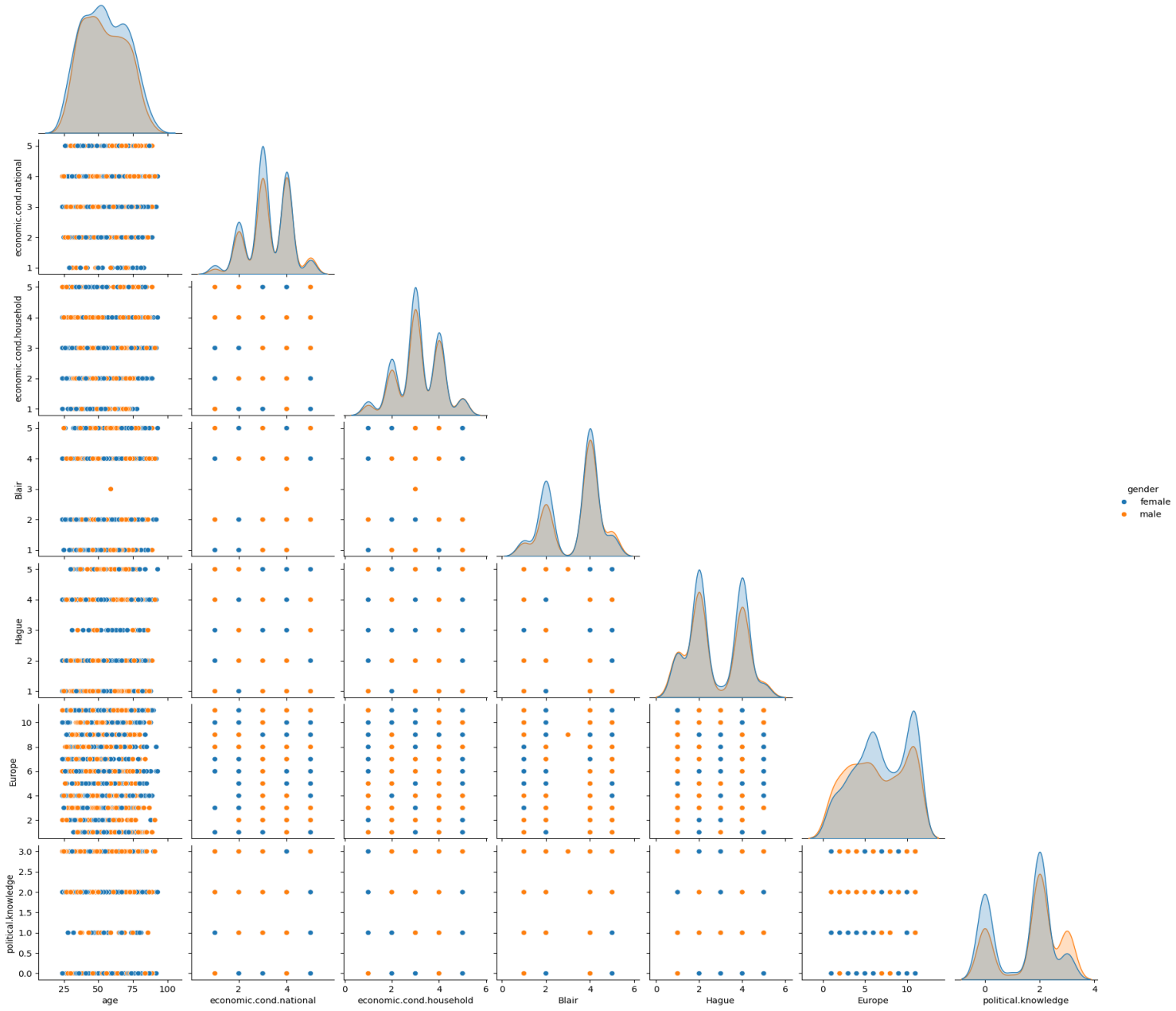


Fig: -16: Pairplot

Inference:

- There is no linear relationship between variables
- Some of the attributes look like they may have an exponential distribution
- Conservative party: Knowledge of parties' positions on European integration is unknown

Orange represents the male distribution and blue represent the female distribution. We can see that most of the observations are from male gender. Let us check for any imbalance problem in the dataset. Let us use the vote variable to find out the number of labour and conservative counts in the dataset.

Correlation Matrix Heatmap:

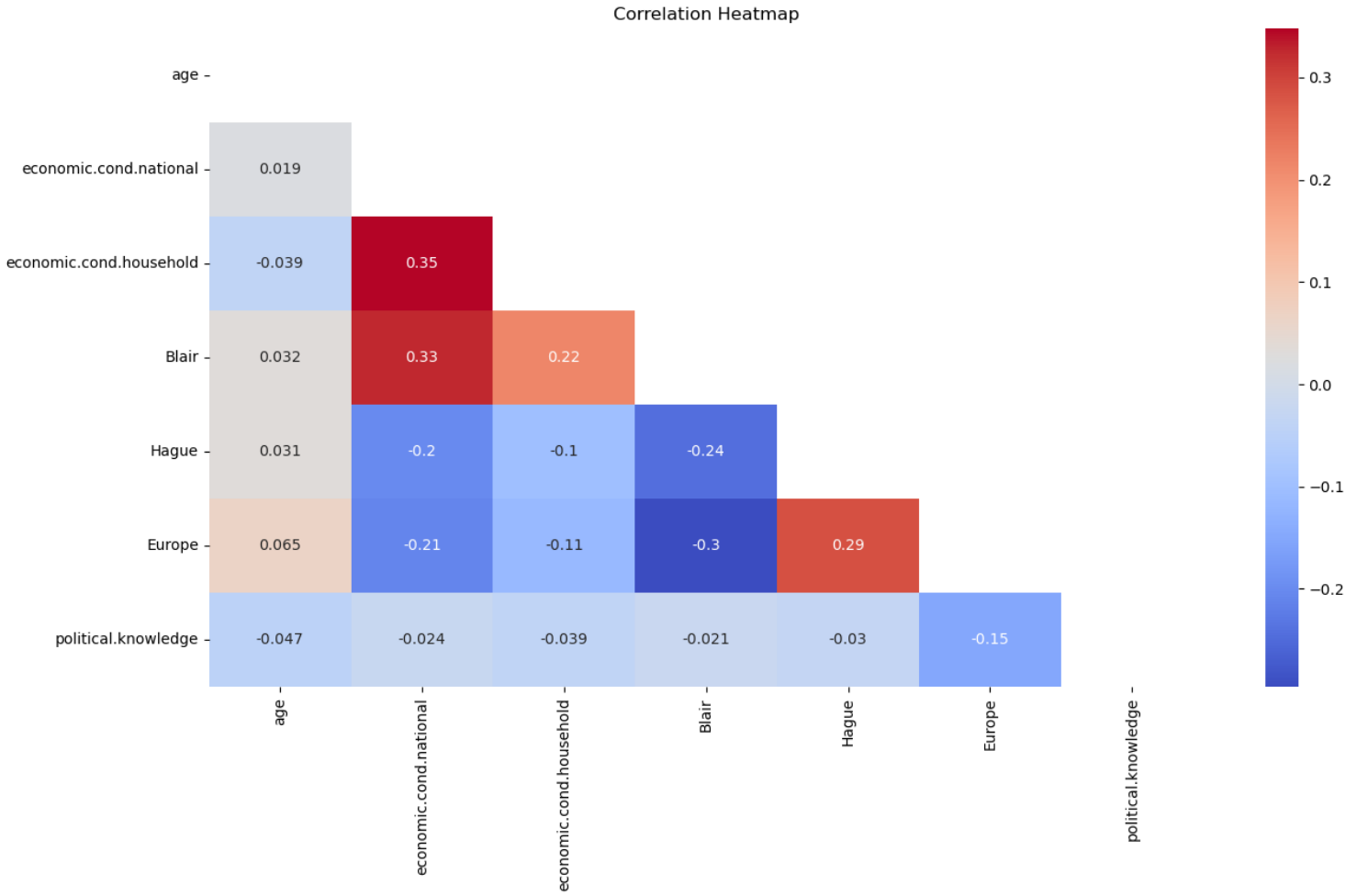
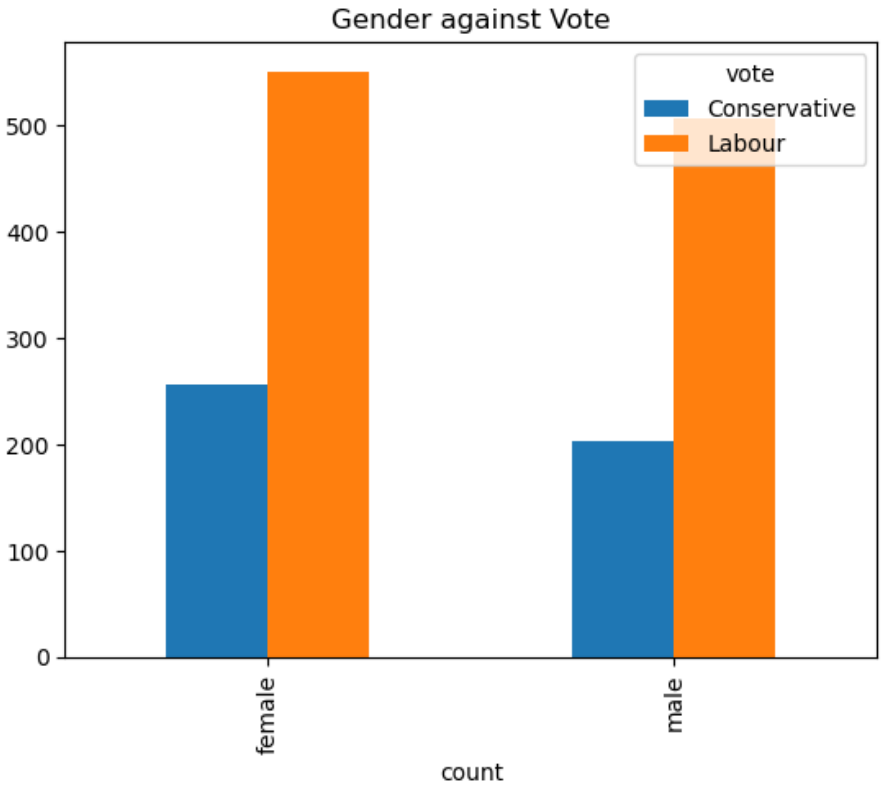


Fig: -17: Heatmap

- Mild positive correlation **exists between** national and household economic condition ratings**, and also with Labour Party leader Tony Blair's ratings. Conversely,** slight negative correlation **with** Conservative Party leader William Hague's ratings**, suggesting** general satisfaction **with current economy and preference for Labour.**
- Mild negative correlation **between** Brexit sentiments and Blair's ratings**, mild positive correlation with Hague's ratings, suggesting** Brexit supporters are discontent with Labour's EU stance **and** prefer Conservatives.
- Blair and Hague's **ratings exhibit** weak negative correlation**, as expected from opposing election candidates.**

Categorical VS Categorical Variable:



gender	female	male
vote		
Conservative	31.81	28.63
Labour	68.19	71.37

Fig: -18: barplot and crosstab for Gender against Vote count.

Gender alone seems to have no significant impact on the votes. Although one can say that the female percentage that have voted for the conservatives is slightly more when compared with the labour party.

Numerical VS Categorical Variable:

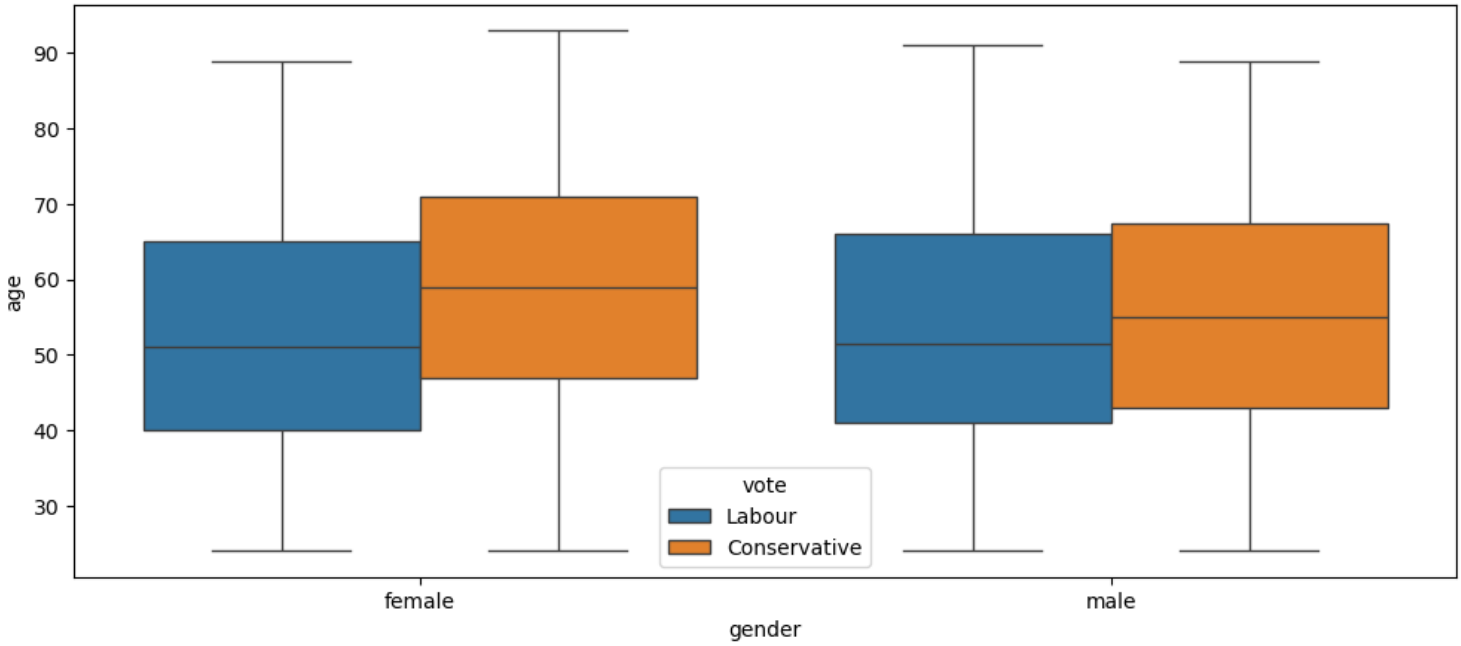
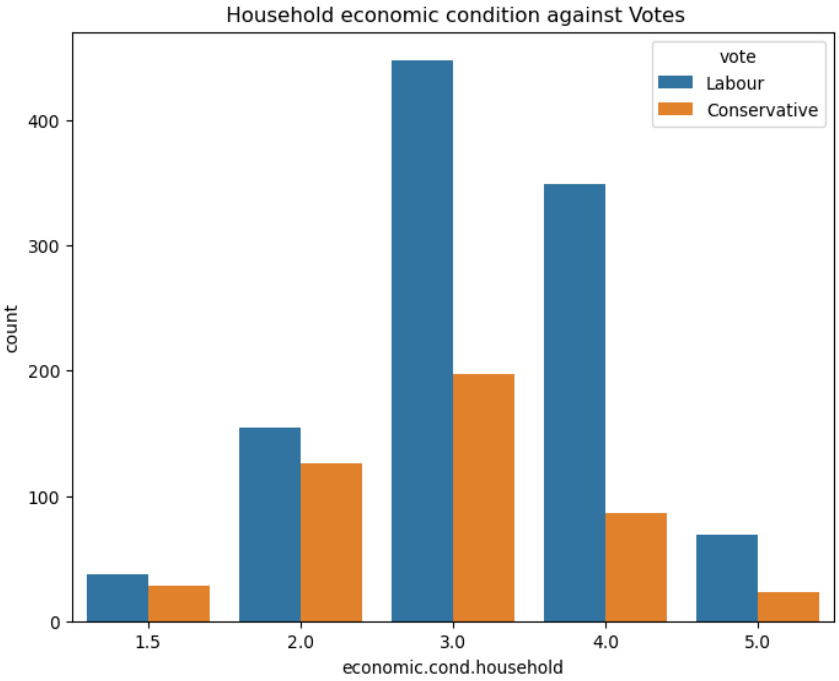


Fig: -19: Boxplot for Gender, Age against Vote.

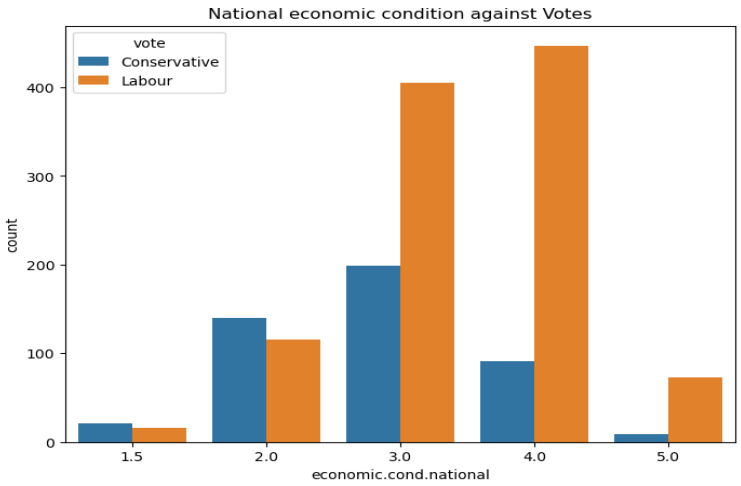
- Avg age group near to 60 choose to elect conservative party whereas avg age group near to 50 choose to elect Labour party.
- Middle 50% of the female people who fall under the age group within the range of 40 to 65 choose to elect Labour party.
- Middle 50% of the female people who fall under the age group within the range of 50 to 70 choose to elect conservative party.
- So, we can see from the boxplot that has the age increases people choose to elect conservative party.



economic.cond.household	1.5	2.0	3.0	4.0	5.0
vote					
Conservative	43.08	45.0	30.54	19.77	25.0
Labour	56.92	55.0	69.46	80.23	75.0

Fig: -20: Barplot and Crosstab for Household vs Vote.

* The Majority of voters rating their household economic conditions as average or above prefer Labour**.** This noticeable vote difference for these ratings between Labour and Conservatives reinforces Labour's appeal.



economic.cond.national	1.5	2.0	3.0	4.0	5.0
vote					
Conservative	56.76	54.69	32.95	16.91	10.98
Labour	43.24	45.31	67.05	83.09	89.02

Fig: -21: Bar plot and Crosstab for National

The plot confirms Labour's majority vote **and highlights that** voters rating national economic conditions as average or above **lean towards Labour. This echoes the previously observed positive correlation with Labour leader's ratings**

Blair	1	2	3	4	5
vote					
Conservative	60.82	55.3	100.0	18.85	1.97
Labour	39.18	44.7	0.0	81.15	98.03

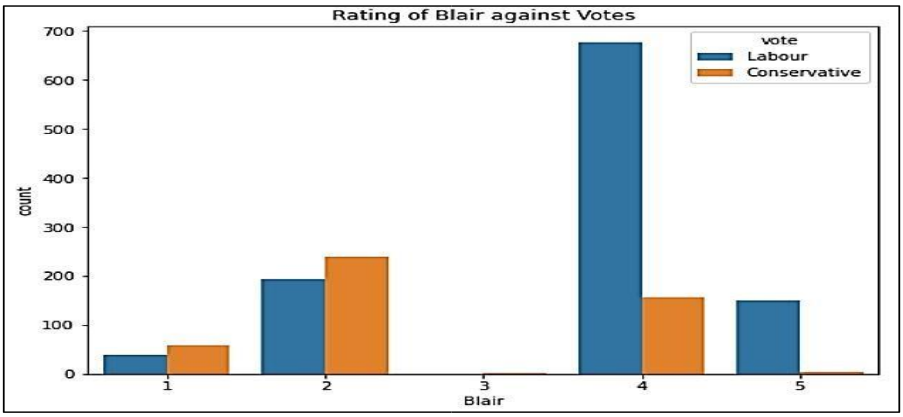
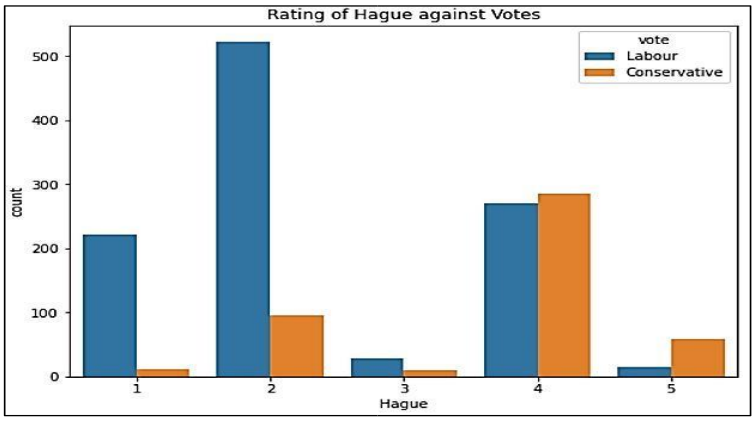


Fig: -21: Rating of Labour leader Blair against Votes.

Voters giving high ratings to Tony Blair typically support Labour**,** while those rating him low typically vote Conservative, as expected.

Hague	1	2	3	4	5
vote					
Conservative	4.72	15.4	24.32	51.35	80.82
Labour	95.28	84.6	75.68	48.65	19.18



Voters who rated William Hague highly tend to vote Conservative**,** while those giving him low ratings generally support Labour.

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

vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
Conservative	462	462	462	462	462	462	462	462
Labour	1063	1063	1063	1063	1063	1063	1063	1063

- Distribution of Target variable is 70:30.
- Most of the voters elected to labour party. The ratio is almost 1:2
- The model's ability to predict class labour will be more compared to conservative.

Encode the Data:

- For a given dataset, we have 8 categorical variables, 6 variables which are of ordinal data type already integer data type, remaining 2 nominal variables gender and the target variable vote with the type as object, needs to be converted into categorical data type.

```
feature: vote
['Labour', 'Conservative']
Categories (2, object): ['Conservative', 'Labour']
[1 0]


feature: gender
['female', 'male']
Categories (2, object): ['female', 'male']
[0 1]
```

Fig: -22: Encoded Data

- After encoding the data, the target variable 'vote' was captured in to separate vector for training and test data set.
- Then, the data was split into train and test in the ratio of 70:30.

Scaling:

- We have feature age in years with different unit weight and the remaining is of ratings ranging from 1 to 5; 1 to 11 and 0 to 3. Hence scaling is required for certain models to get accurate results.
- Scaled data can be done only for training set, for test set scaling is not required, because in the real-world data will not be scaled, need to be passed has it is through the model with whatever measurements they come in.

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

Apply Logistic Regression:

Feature scaling doesn't make a much effect for logistic regression test data accuracy; hence scaling is not performed for logistic regression.

Build a Logistic regression model using a grid search cross validation to get best parameter/ estimators for a given dataset.

Accuracy of the logistic regression model for train set is 82.658% and accuracy of the logistic regression model for test set is 85.95%. The model is a valid since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply Linear Discriminant Analysis:

Feature scaling doesn't make a much effect for LDA test data accuracy; hence scaling is not performed for LDA model.

Linear Discriminant analysis was built without any specific parameter setting and then values were predicted on both training and test dataset.
Accuracy of the LDA model for train set is 82.3% and accuracy of the LDA model for test set is 86%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of both the models LR and LDA models, Test data accuracy of logistic regression is quite good compared to LDA model. Hence Logistic regression perform well for predicting Labour or conservative party.

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

Apply Gaussian Naïve Bayes:

For Naive Bayes algorithm while calculating likelihoods of numerical features it assumes that the feature to be normally distributed and then we calculate probability using mean and variance of that feature only and also it assumes that all the predictors are independent to each other. Feature scaling doesn't matter. Performing a features scaling in this algorithms may not have much effect.

Now build a GaussianNB classifier. Then the classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method with test set features as its parameters.

Accuracy of GaussianNB model for train set is 82% and accuracy of GaussianNB model for test set is 85.5%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply K-nearest neighbour [KNN-model]:

In order to have good KNN model, data requires pre-processing to make all independent variables similarly scaled and centered. Hence need to perform z-score on all numeric attributes in models that calculate distance and see the performance for KNN.

By default, value of n_neighbors=5, in order to get best KNN model need to try for different K values and find out for the corresponding k-value which is the least Misclassification error.

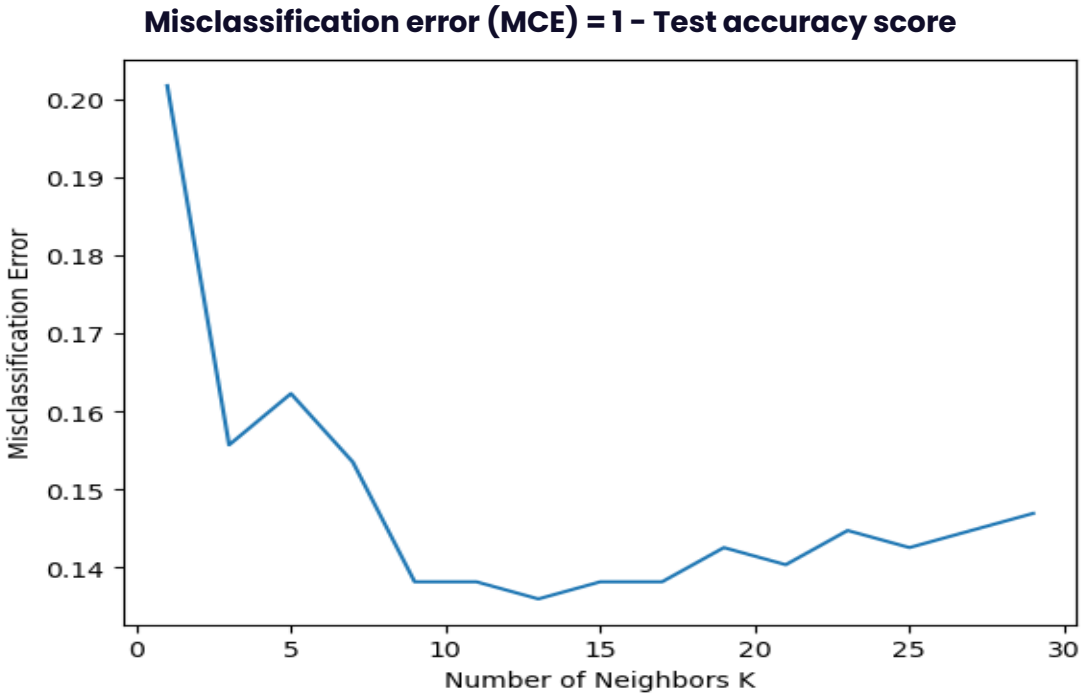


Fig: -23: Plot misclassification error vs K.

From the above plot the optimum number of neighbour or k-value is found to be 23.

Now build a K-NeighborsClassifier using the value of n_neighbors=13 and metric= 'Euclidean'. Then the classifier is trained using scaled training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method with test set features as its parameters.

Accuracy of KNN model for train set is 83.7% and accuracy of KNN model for test set is 86.4%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of both the models NB and KNN models, Test data accuracy of KNN model for n_neighbour=13 is quite good compared to NB model. Hence KNN model perform well for predicting Labour or conservative party.

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

Apply Ensemble Random Forest model:

Feature scaling is not required for random forest model, hence scaling is not performed.

Build a Random Forest model using a grid search cross validation to get best parameter/ estimators for a given dataset.

```
{'max_depth': 7, 'max_features': 4, 'min_samples_leaf': 40, 'min_samples_split': 100, 'n_estimators': 501}

RandomForestClassifier(max_depth=7, max_features=4, min_samples_leaf=40,
                       min_samples_split=100, n_estimators=501,
                       random_state=27)
```

Fig: -24: Best estimator for Random Forest.

With the above parameter setting values were predicted for both training and test dataset.

Accuracy of the random forest model for train set is 81.5% and test set is 83.5%. The model is a valid since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply Bagging using base estimator has Random Forest:

Bagging was built using the above tuned random forest has a base estimator and n_estimators=100. Then the classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method with test set features as its parameters.

```
BaggingClassifier(base_estimator=RandomForestClassifier(max_depth=7,
                                                        max_features=4,
                                                        min_samples_leaf=40,
                                                        min_samples_split=100,
                                                        n_estimators=501,
                                                        random_state=27),
                  n_estimators=100, random_state=27)
```

Accuracy of the Bagging model for train set is 80.7% and test set is 83.97%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of Random Forest and Bagging models, Test data accuracy of Random Forest model quite good compared to Bagging model. Hence Random Forest model performs well for predicting Labour or conservative party.

Apply Gradient Boosting:

Gradient Boosting was built with n_estimators=100 and then classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method.

Accuracy of the Gradient boosting model for train set is 88.6% and accuracy of the Gradient model for test set is 84.2%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply Ada Boosting:

AdaBoost was built with n_estimators=100 and then classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method.

Accuracy of the Adaboost model for train set is 84.9% and test set is 83.6%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of Gradient Boosting and AdaBoost models, Test data accuracy of both the models are almost equal, since the difference between the train and test set is very low for Ada boost compared to Gradient Boost model, Ada Boost model performs well for predicting Labour or conservative party.

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.

1. Model Evaluation-Logistic Regression:

- Accuracy on training set is 83% and on testing set is 85.4%.
- Classification report:

Classification report for Logistic Regression model on Training set is

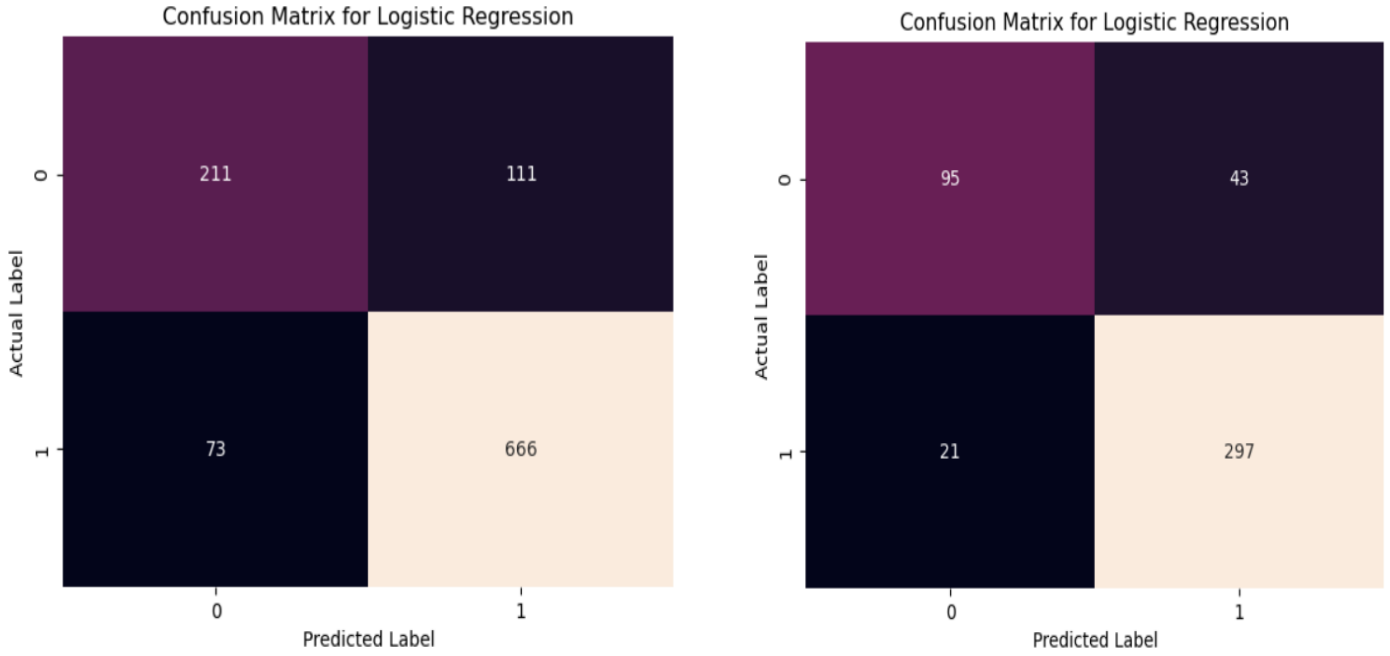
	precision	recall	f1-score	support
0	0.74	0.66	0.70	322
1	0.86	0.90	0.88	739
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.82	0.83	0.82	1061

Classification report for Logistic Regression model on Testing set is

	precision	recall	f1-score	support
0	0.82	0.69	0.75	138
1	0.87	0.93	0.90	318
accuracy			0.86	456
macro avg	0.85	0.81	0.83	456
weighted avg	0.86	0.86	0.86	456

Fig: -25: Classification report on Train and Test set for LR

Confusion Matrix:



LR_train_precision 0.86
LR_train_recall 0.9
LR_train_f1 0.88

LR_test_precision 0.87
LR_test_recall 0.93
LR_test_f1 0.9

Fig: -26: Confusion matrix on Train and Test set of LR

- AUC score on the train dataset is 86% and on test dataset is 91.5%.

ROC Curve:

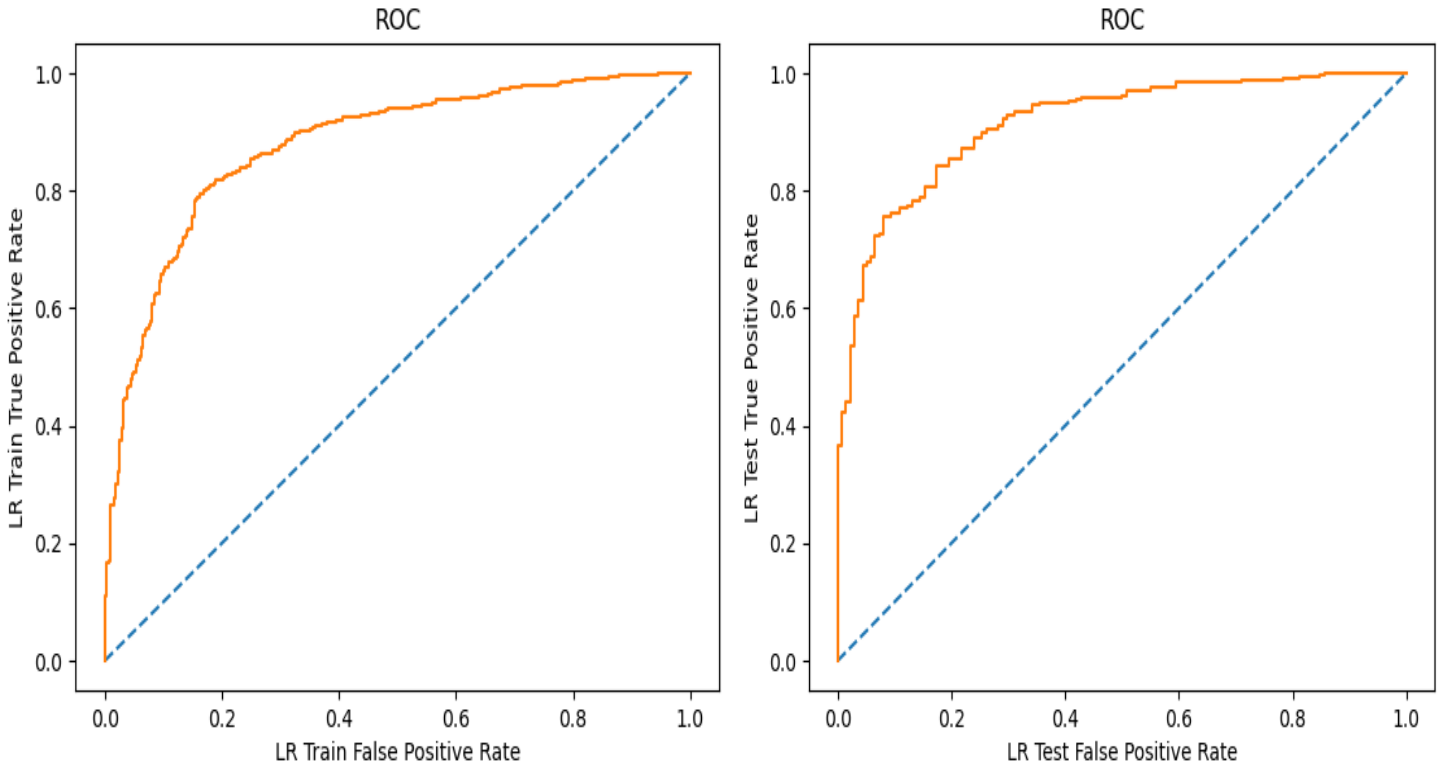


Fig: - 27: ROC curve on Train and Test set for LR

Logistic Regression Output:

Logistic Regression Conclusion:	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	82.3	87.5	85	90	88
Test set	85.1	91.4	87	92	90

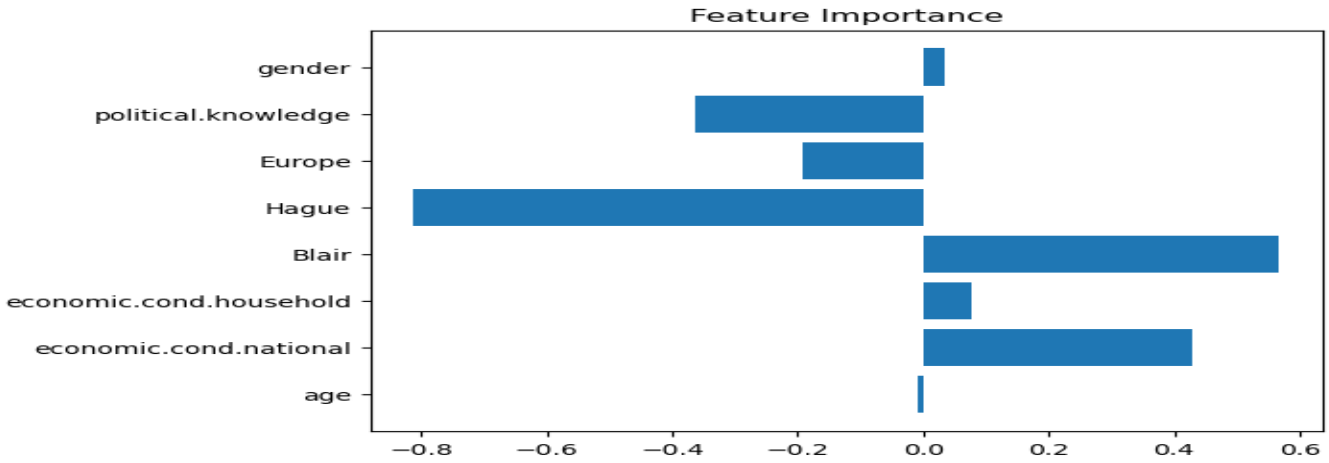


Fig: -28: Free Importance and Performance metrics.

2. Model Evaluation-Linear Discriminant Analysis:

- Accuracy on train set is 82.3% and on test set is 86%.

Classification report:

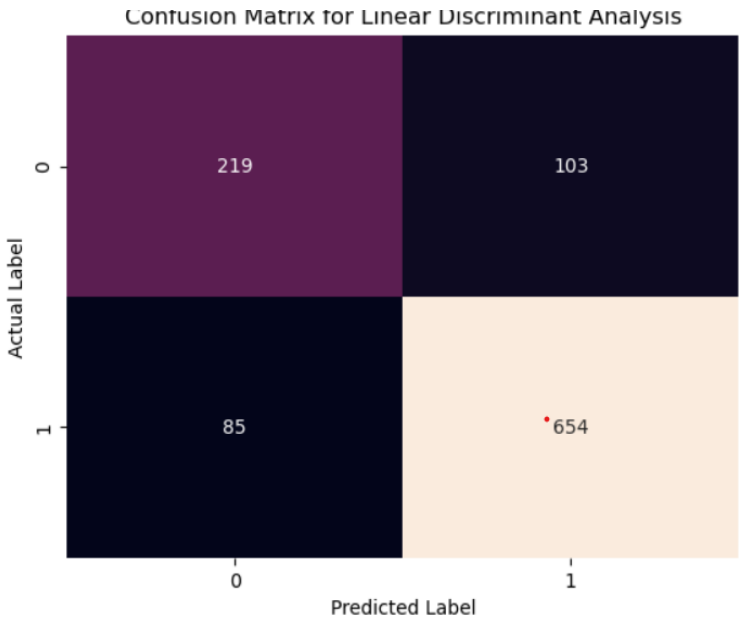
Classification report for Linear Discriminant Analysis model on Training set is

	precision	recall	f1-score	support
0	0.72	0.68	0.70	322
1	0.86	0.88	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

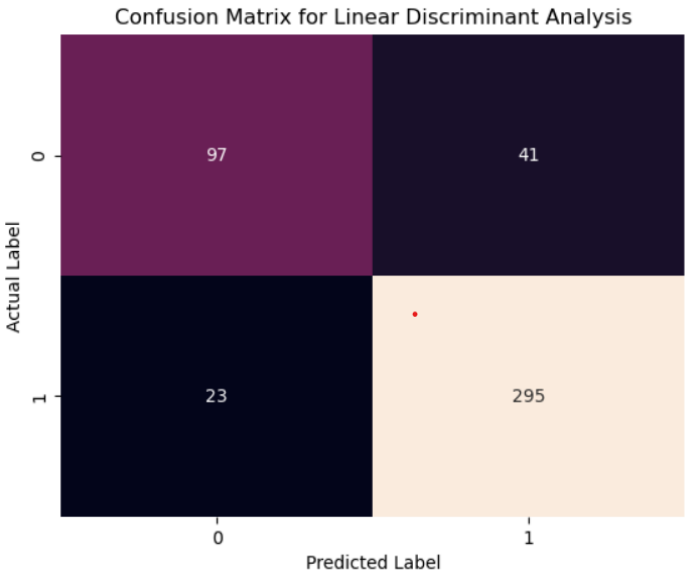
Classification report for Linear Discriminant Analysis model on Testing set is

	precision	recall	f1-score	support
0	0.81	0.70	0.75	138
1	0.88	0.93	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

Fig: -29: Classification Report on Train and Test set for LDA.



Confusion Matrix for Linear Discriminant Analysis model on Testing set is



LDA_train_precision 0.86
LDA_train_recall 0.88
LDA_train_f1 0.87

LDA_test_precision 0.88
LDA_test_recall 0.93
LDA_test_f1 0.9

Fig: - 30: Confusion matrix on Train and Test set for LDA.

- AUC score on the train dataset is 87.7% and on test dataset is 91.5%.

ROC Curve:

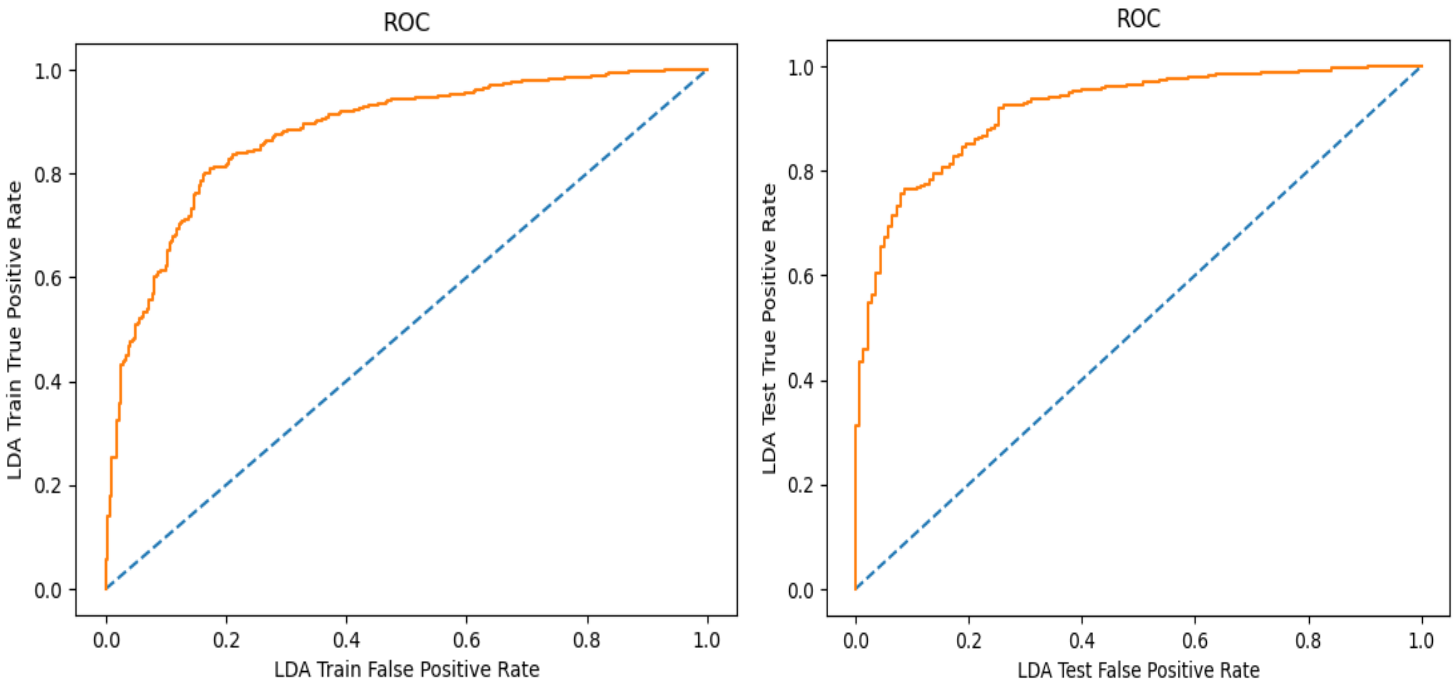
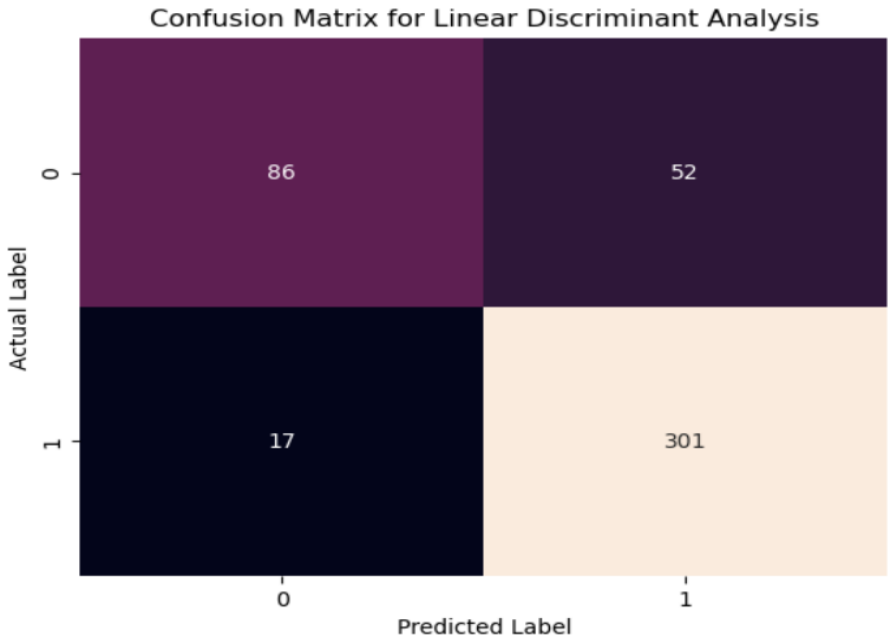


Fig: -31: ROC curve for Train and Test set for LDA.

By changing the probability cut-off values; we see that 0.4 and 0.5 gives better accuracy than the rest of the custom cut-off values. But 0.4 cut-off gives us the best 'f1-score'. Hence, we will take the cut-off as 0.4 to get the optimum 'f1' score in order to improve the test set results.

Accuracy for LDA model on testing set with cut-off value 0.4 is 82.9




```
LDA_test_precision_new  0.85
LDA_test_recall_new  0.95
LDA_test_f1_new  0.9
```

Fig: - 32: Confusion matrix for LDA model cutoff value 0.4

LDA Output:

	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Linear Discriminant Analysis Conclusion:					
Train set	82.3	87.7	86	88	87
Test set	86.0	91.5	88	93	90
Improved Test set	84.9	NaN	85	95	90

Fig: -33: Performance metrics Output LDA.

- From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
- By changing the probability cut-off value from default 0.5 to 0.4, we could see that the precision has improved from 91% to 94% on the test set and the model accuracy is of 84.7%
- Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

3. Model Evaluation-Gaussian Naïve Bayes:

- Accuracy on train set is 82% and on test set is 85.5%.

Classification Report:

Classification report for Naive Bayes model on Training set is

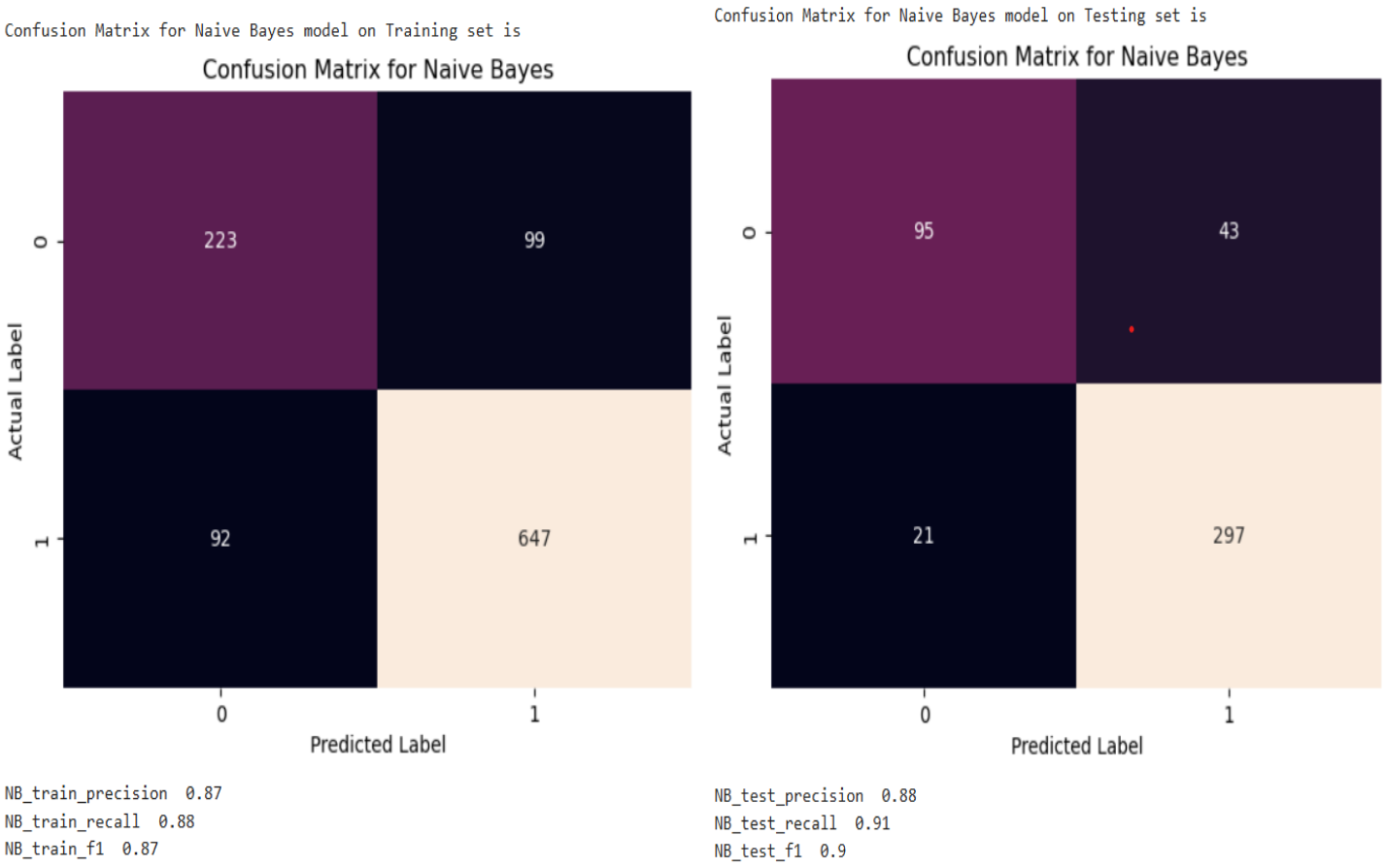
	precision	recall	f1-score	support
0	0.71	0.69	0.70	322
1	0.87	0.88	0.87	739
accuracy			0.82	1061
macro avg	0.79	0.78	0.79	1061
weighted avg	0.82	0.82	0.82	1061

Classification report for Naive Bayes model on Testing set is

	precision	recall	f1-score	support
0	0.78	0.72	0.75	138
1	0.88	0.91	0.90	318
accuracy			0.86	456
macro avg	0.83	0.82	0.82	456
weighted avg	0.85	0.86	0.85	456

Fig: -34: Classification report On Train and Test set of GNB

Confusion Matrix:



NB_train_precision 0.87
NB_train_recall 0.88
NB_train_f1 0.87

NB_test_precision 0.88
NB_test_recall 0.91
NB_test_f1 0.9

Fig: -35: Confusion Matrix on Train and Test set GNB.

- AUC score on the train dataset is 87.4% and on test dataset is 91.3%.

ROC Curve:

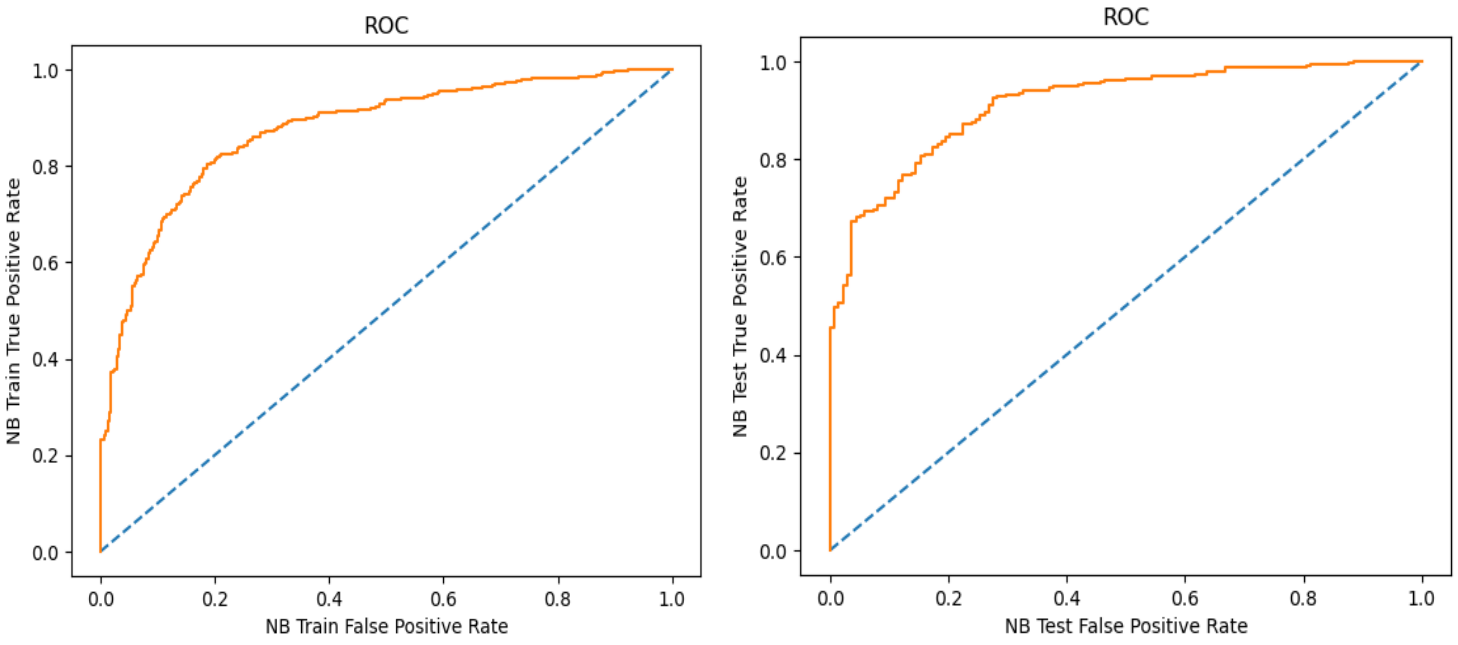


Fig: -36: ROC curve for Train and Test GNB.

By changing the probability cut-off values; we see that 0.4 and 0.5 gives better accuracy than the rest of the custom cut-off values. But 0.4 cut-off gives us the best 'f1-score'. Hence, we will take the cut-off as 0.4 to get the optimum 'f1' score in order to improve the test set results.

Accuracy for LDA model on testing set with cut-off value 0.4 is 85.5.

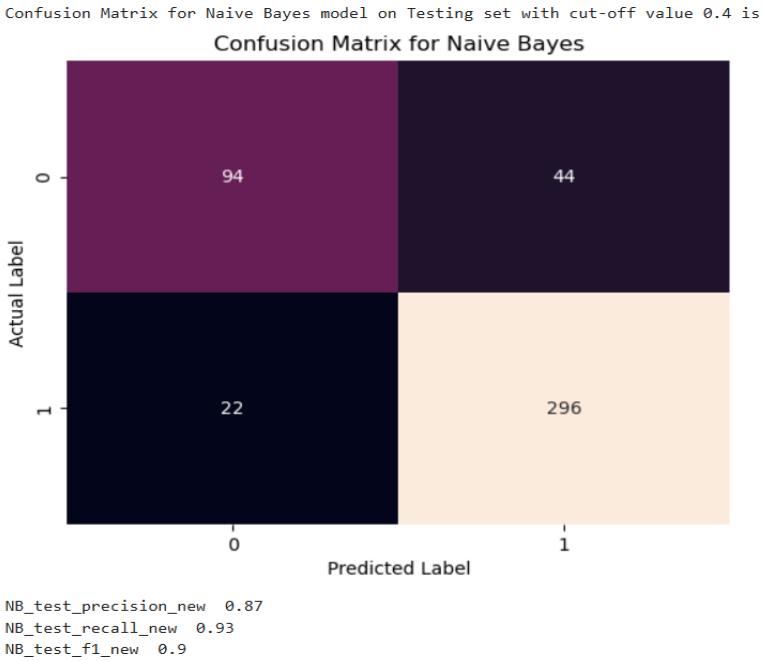


Fig: -37: Confusion matrix for NB model on testing set cutoff of 0.4

Gaussian Naïve Bayes Output:

Naive Bayes Analysis Conclusion:	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	82.0	87.4	87	88	87
Test set	85.5	91.3	88	91	90
Improved Test set	85.5	NaN	87	91	90

Fig: -38: Performance Metrics Output GNB.

- From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
- By changing the probability cut-off value from default 0.5 to 0.4, we could see that the precision has improved from 87% to 88% on the test set and the model accuracy is of 85.5%
- Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

4. Model Evaluation- KNN:

- Accuracy on train set is 83.8% and on test set is 84.6%.

Classification report:

Classification report for K-nearest neighbour model on Training set is

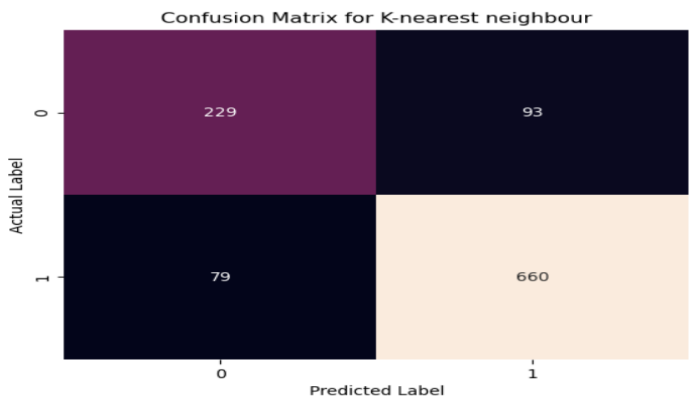
	precision	recall	f1-score	support
0	0.74	0.71	0.73	322
1	0.88	0.89	0.88	739
accuracy			0.84	1061
macro avg	0.81	0.80	0.81	1061
weighted avg	0.84	0.84	0.84	1061

Classification report for K-nearest neighbour model on Test set is

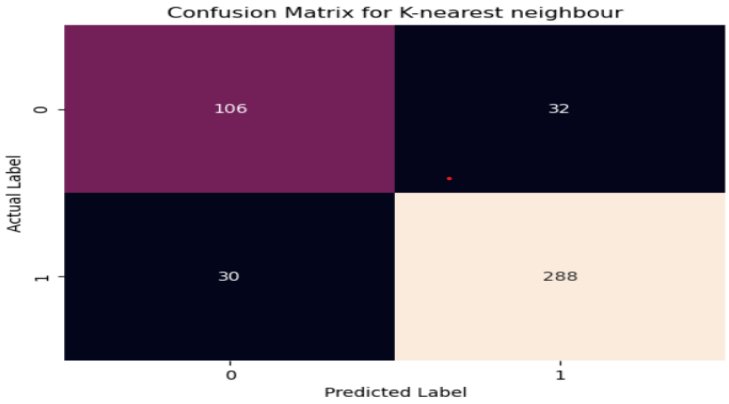
	precision	recall	f1-score	support
0	0.78	0.77	0.77	138
1	0.90	0.91	0.90	318
accuracy			0.86	456
macro avg	0.84	0.84	0.84	456
weighted avg	0.86	0.86	0.86	456

Fig: -39: Classification Report on Train and Test set.

Confusion Matrix:



KNN_train_precision 0.88
KNN_train_recall 0.89
KNN_train_f1 0.88



KNN_test_precision 0.9
KNN_test_recall 0.91
KNN_test_f1 0.9

Fig: -40: Confusion matrix on Train and Test set.

- AUC score on the train dataset is 90.5% and on test dataset is 89.2%.

ROC Curve:

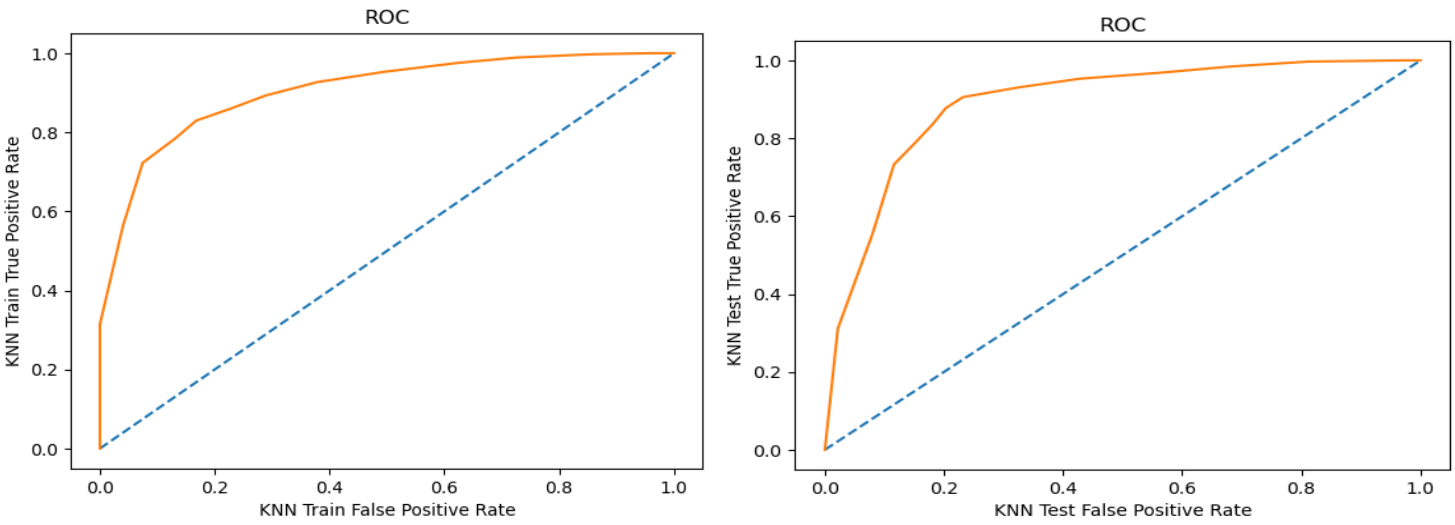


Fig: -41: ROC curve for Train and Test set.

KNN Output:

	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
K-nearest neighbors kNN Conclusion:					
Train set	83.8	90.5	88	89	88
Test set	86.4	89.2	90	91	90

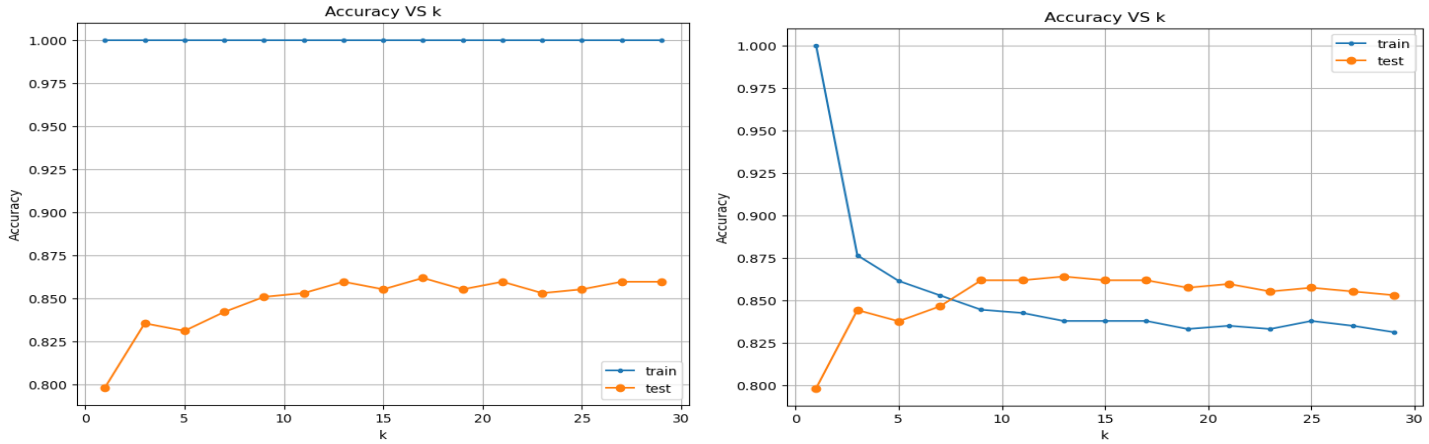


Fig: - 42: Performance metrics output.

- From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
- Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

5. Model Evaluation- Random Forest:

- Accuracy on train set is 81.5% and on test set is 83.6%.

Classification report:

Classification report for RandomForestClassifier model on Training set is

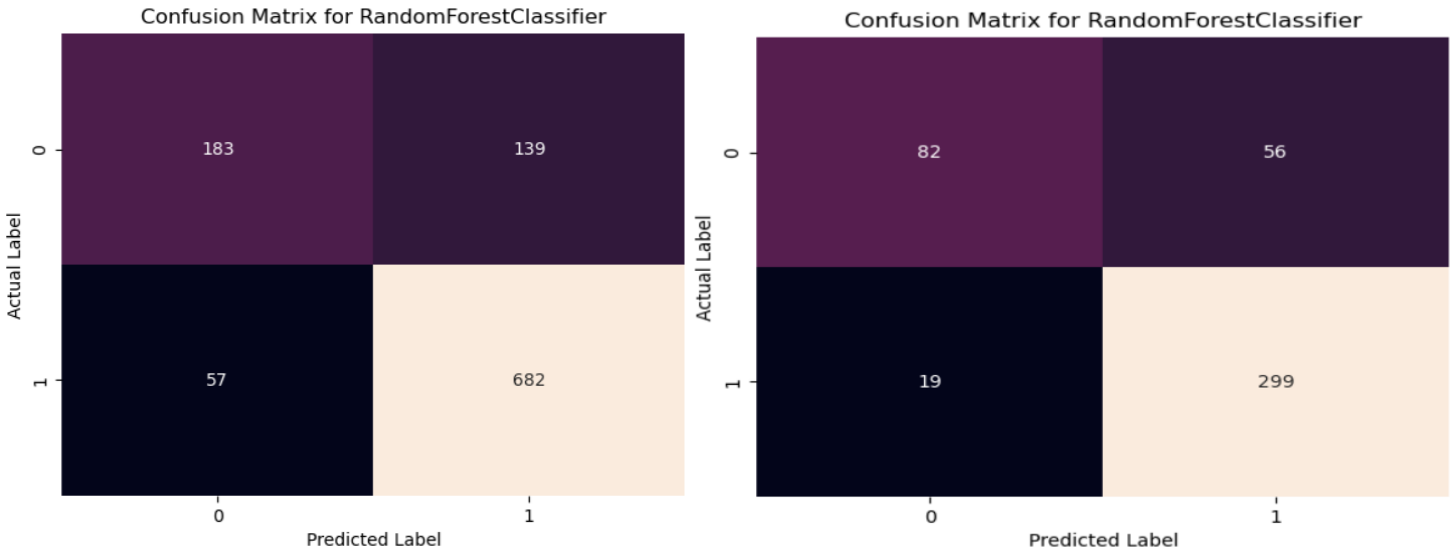
	precision	recall	f1-score	support
0	0.76	0.57	0.65	322
1	0.83	0.92	0.87	739
accuracy			0.82	1061
macro avg	0.80	0.75	0.76	1061
weighted avg	0.81	0.82	0.81	1061

Classification report for RandomForestClassifier model on Test set is

	precision	recall	f1-score	support
0	0.81	0.59	0.69	138
1	0.84	0.94	0.89	318
accuracy			0.84	456
macro avg	0.83	0.77	0.79	456
weighted avg	0.83	0.84	0.83	456

Fig: -43: Classification Report for Train and Test set.

Confusion matrix:



RF_train_precision 0.83
RF_train_recall 0.92
RF_train_f1 0.87

RF_test_precision 0.84
RF_test_recall 0.94
RF_test_f1 0.89

Fig: -44: Confusion matrix on Train and Test set.

- AUC score on the train dataset is 88.4% and on test dataset is 90.4%.

ROC Curve:

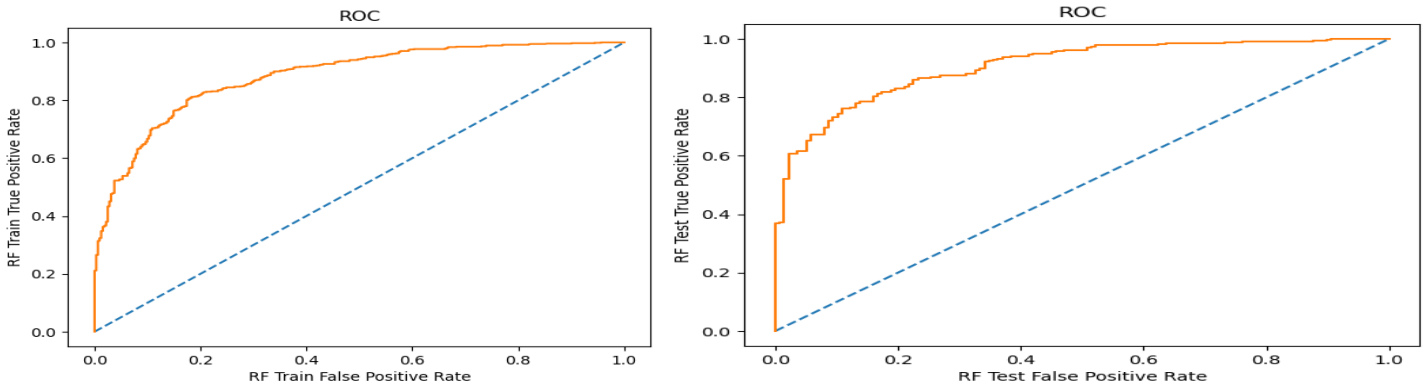


Fig: -45: ROC curve for Train and Test set.

Random Forest Output:

	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Random Forest Conclusion:					
Train set	81.5	88.4	83	92	87
Test set	83.6	90.6	84	94	89

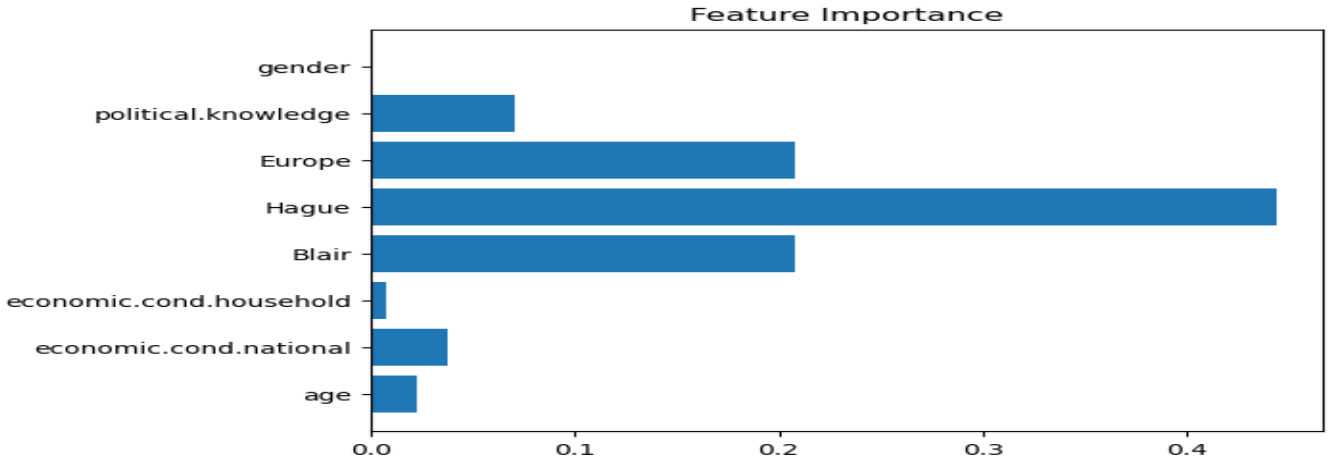


Fig: -46: Performance metrics Output RF.

- From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
- Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

6. <u>Model Evaluation- Bagging:</u>

- Accuracy on train set is 80.8% and on test set is 84%.

Classification report:

Classification report for BaggingClassifier model on Training set is

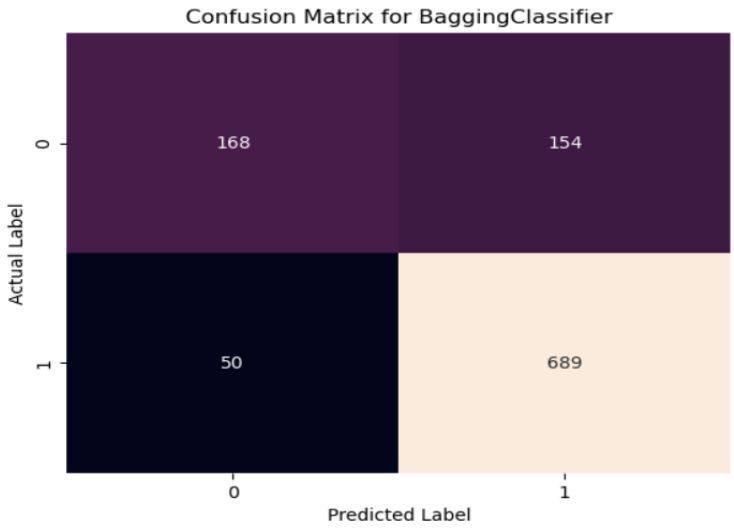
	precision	recall	f1-score	support
0	0.77	0.52	0.62	322
1	0.82	0.93	0.87	739
accuracy			0.81	1061
macro avg	0.79	0.73	0.75	1061
weighted avg	0.80	0.81	0.80	1061

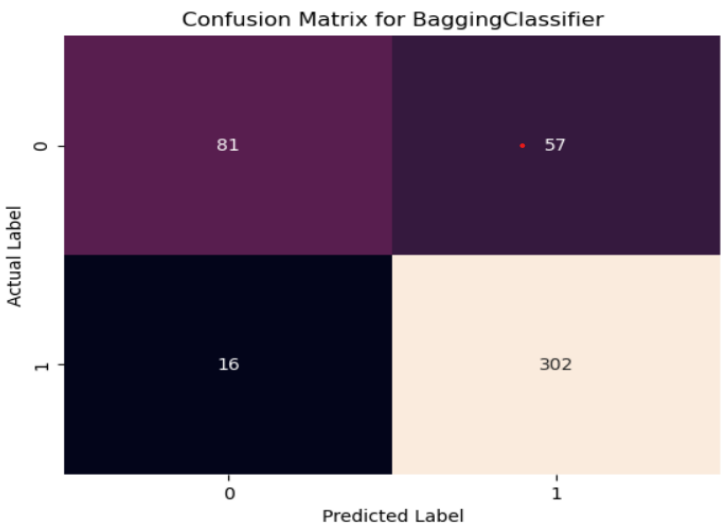
Classification report for BaggingClassifier model on Test set is

	precision	recall	f1-score	support
0	0.84	0.59	0.69	138
1	0.84	0.95	0.89	318
accuracy			0.84	456
macro avg	0.84	0.77	0.79	456
weighted avg	0.84	0.84	0.83	456

Fig: -47: Classification Report for Train and Test set BG.

Confusion Matrix:





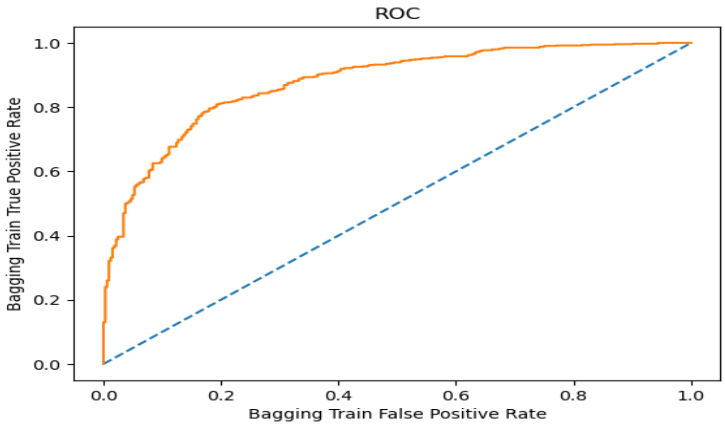

```
BG_train_precision  0.82
BG_train_recall  0.93
BG_train_f1  0.87
```

```
BG_test_precision  0.84
BG_test_recall  0.95
BG_test_f1  0.89
```

Fig: -49: Confusion matrix for Train and Test set BG.

- AUC score on the train dataset is 87.8% and on test dataset is 90.3%.

ROC Curve:



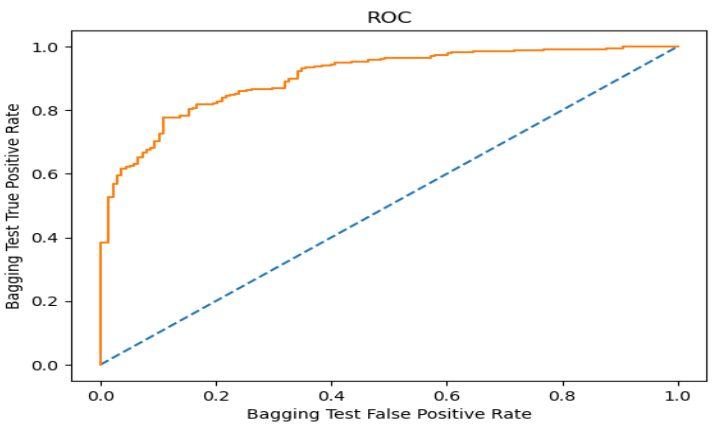


Fig: -50: ROC curve for Train and Test set BG.

Bagging Output:

	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
BaggingClassifier Conclusion:					
Train set	80.8	87.8	82	93	87
Test set	84.0	90.3	84	95	89

Fig: -51: performance output

- From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
- Also, model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

7. Model Evaluation- Gradient Boosting:

- Accuracy on train set is 88.6% and on test set is 84.2%.

Classification report:

Classification report for GradientBoostingClassifier model on Training set is

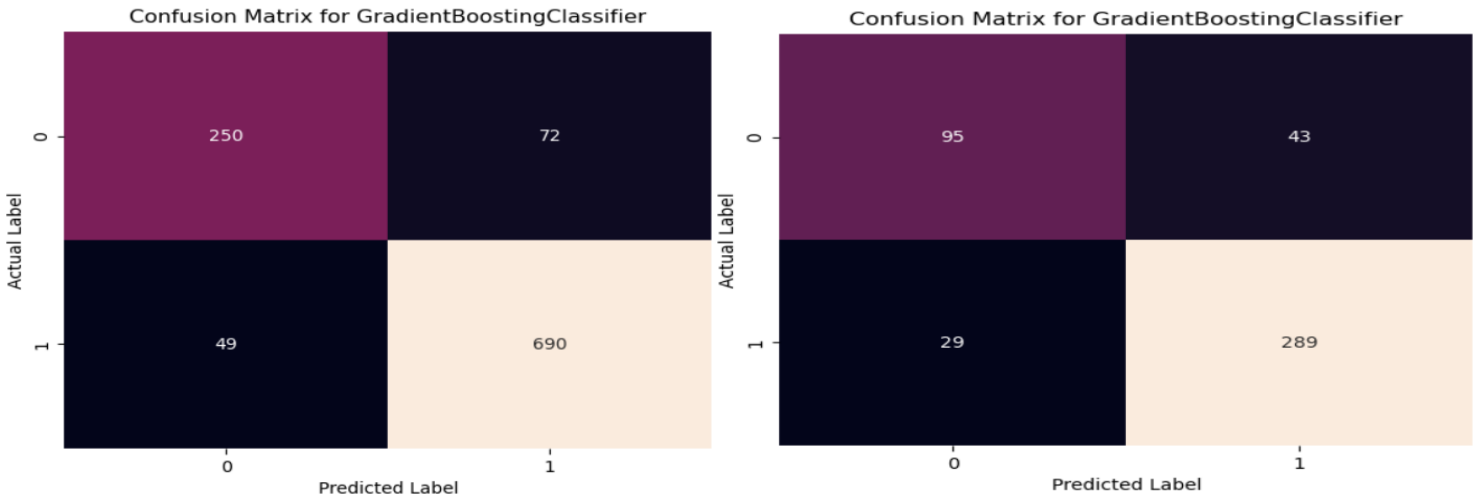
	precision	recall	f1-score	support
0	0.84	0.78	0.81	322
1	0.91	0.93	0.92	739
accuracy			0.89	1061
macro avg	0.87	0.86	0.86	1061
weighted avg	0.88	0.89	0.88	1061

Classification report for GradientBoostingClassifier model on Test set is

	precision	recall	f1-score	support
0	0.77	0.69	0.73	138
1	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

Fig: -51: Classification report for train Test set GB.

Confusion Matrix:



GB_train_precision 0.91
GB_train_recall 0.93
GB train f1 0.92

GB_test_precision 0.87
GB_test_recall 0.91
GB_test_f1 0.89

Fig: -52: Confusion matrix for Train and Test set GB.

- AUC score on the train dataset is 94.7% and on test dataset is 90.4%.

ROC curve:

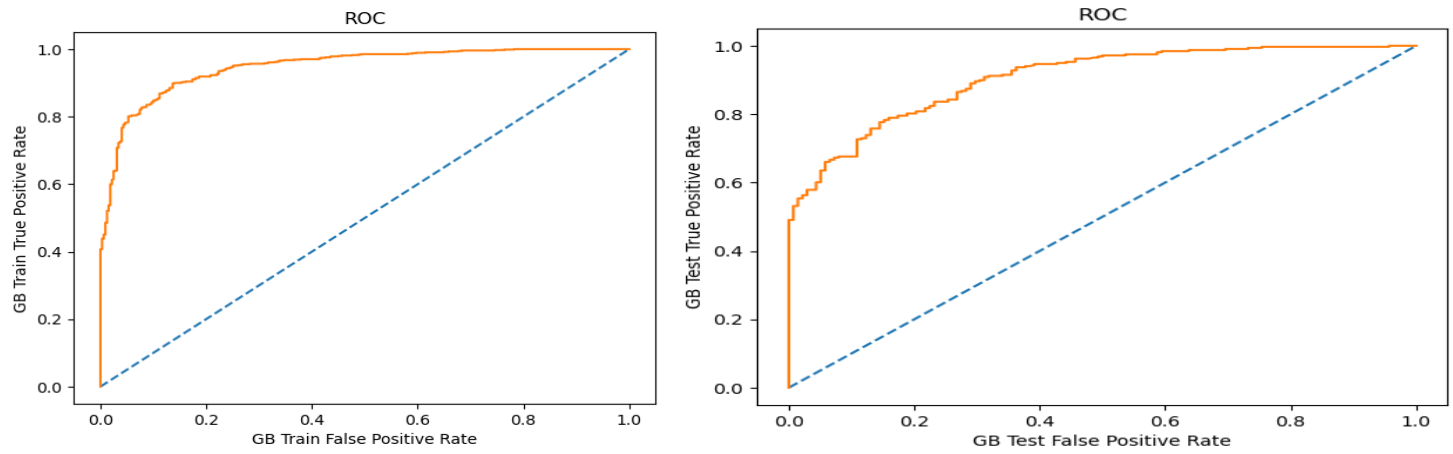


Fig: - 53: ROC curve for Train and Test set GB.

Gradient Boosting Output:

Gradient Boost Conclusion:	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	88.6	94.7	91	93	92
Test set	84.2	90.4	87	91	89

Fig: -55: Performance Metrics Output GB.

- From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
- Also, model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

8. Model Evaluation- Ada Boosting:

- Accuracy on train set is 84.9% and on test set is 83.6%.

Classification report:

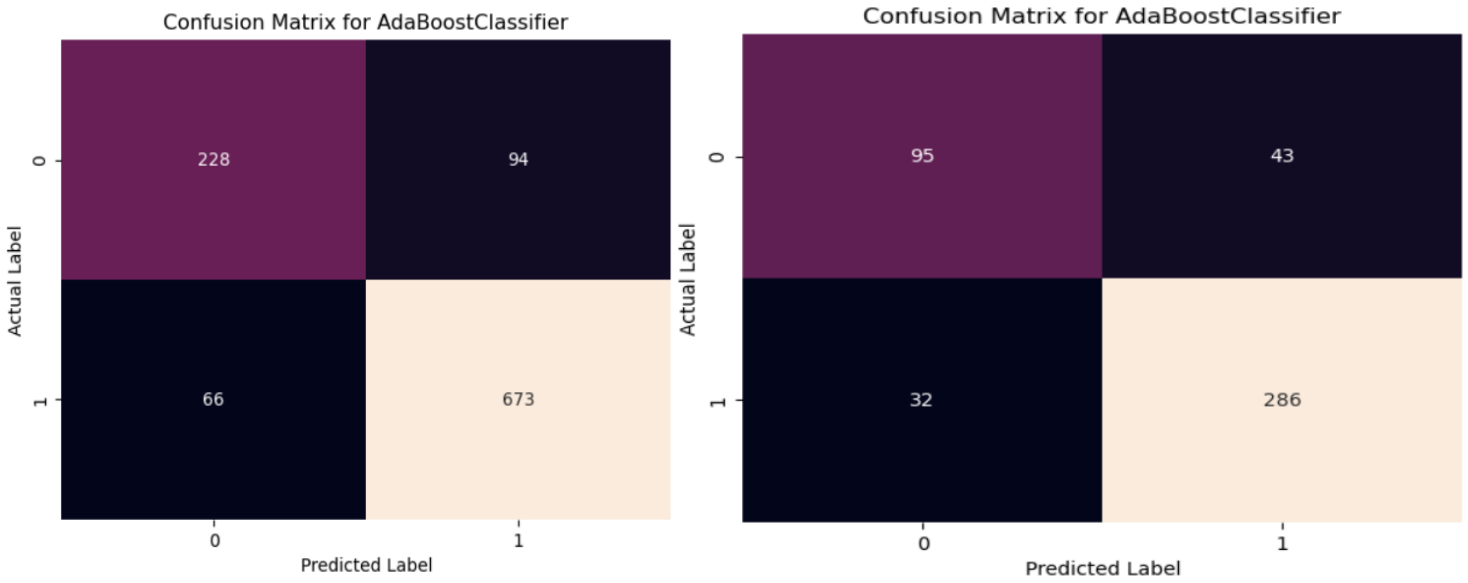
Classification report for AdaBoostClassifier model on Training set is

	precision	recall	f1-score	support
0	0.78	0.71	0.74	322
1	0.88	0.91	0.89	739
accuracy			0.85	1061
macro avg	0.83	0.81	0.82	1061
weighted avg	0.85	0.85	0.85	1061

Classification report for AdaBoostClassifier model on Test set is

	precision	recall	f1-score	support
0	0.75	0.69	0.72	138
1	0.87	0.90	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

Fig: -56: Classification report for train and test set AB.



AB_train_precision 0.88
AB_train_recall 0.91
AB_train_f1 0.89

AB_test_precision 0.87
AB_test_recall 0.9
AB test f1 0.88

Fig: -57: Confusion Matrix for Train and Test AB.

- AUC score on the train dataset is 90.4% and on test dataset is 90.8%.

ROC Curve:

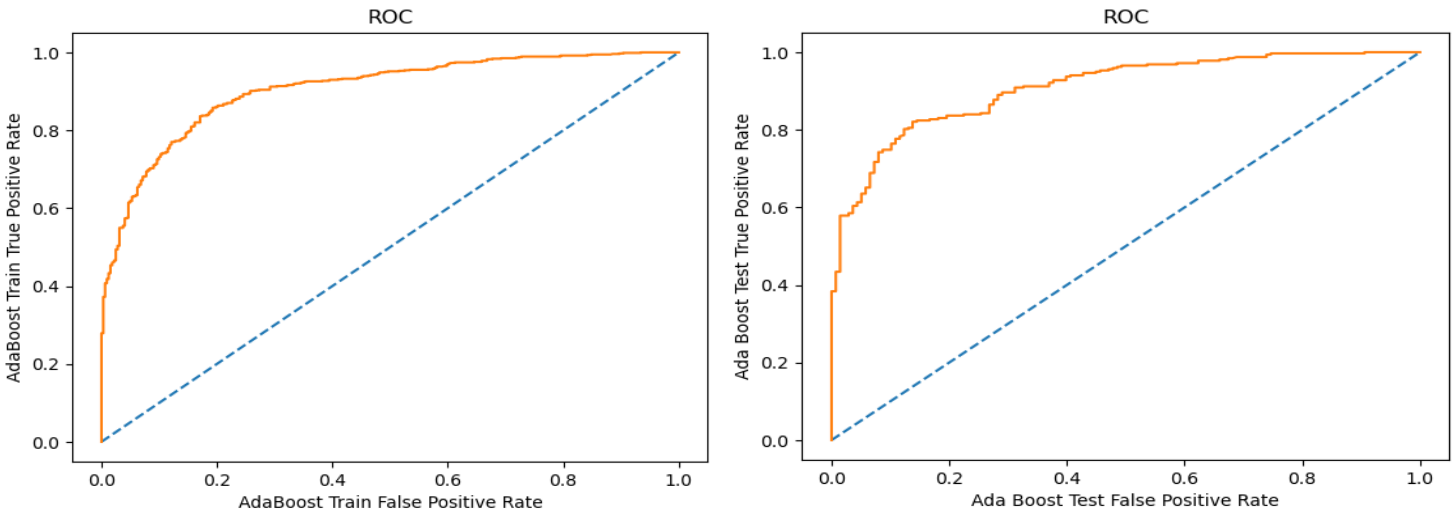


Fig: -58: ROC curve for Train and Test set AB.

AdaBoost Output:

	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
AdaBoostClassifier Conclusion:					
Train set	84.9	90.4	88	91	99
Test set	83.6	90.8	87	90	88

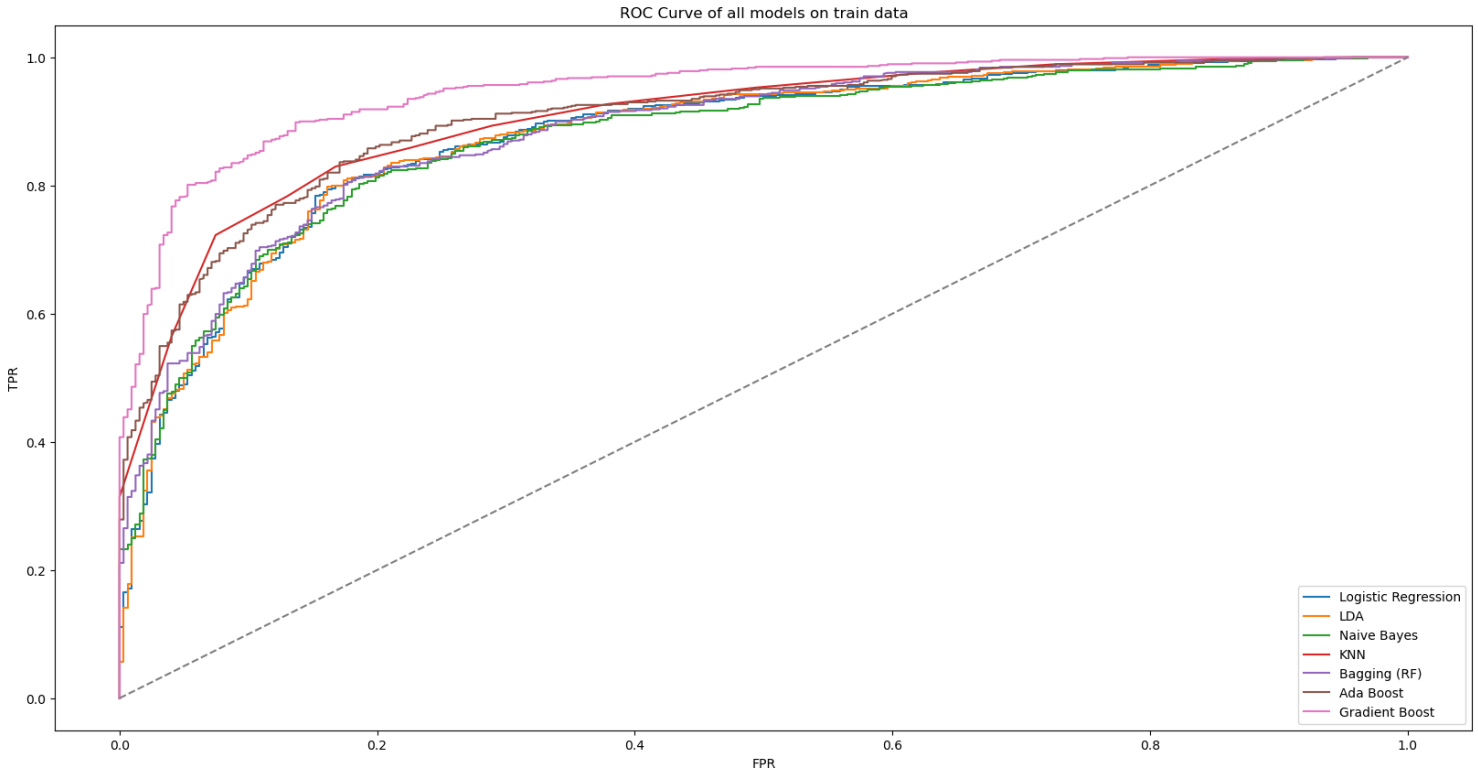
Fig: -59: Performance metrics output AB.

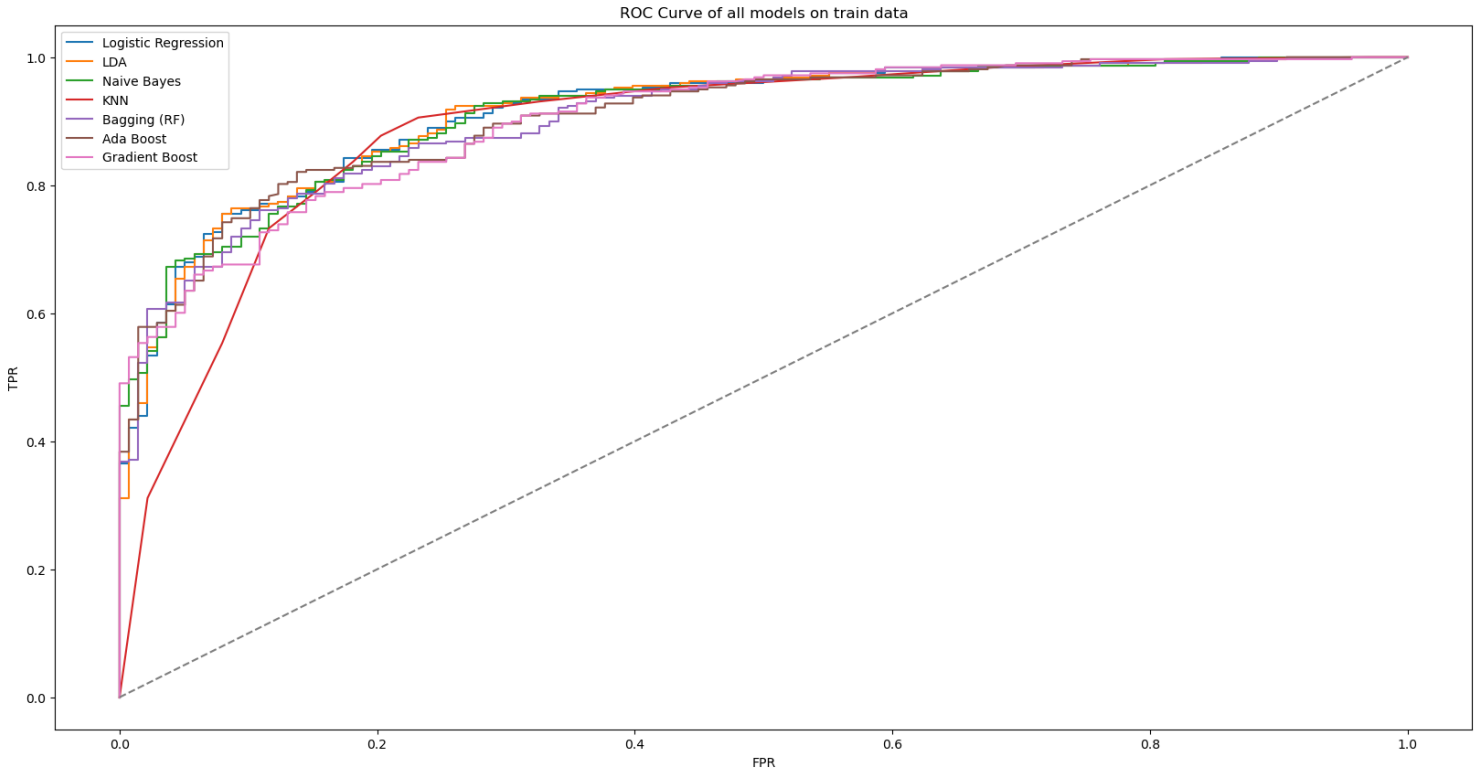
- From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
- Also, model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

1.8 Based on these predictions, what are the insights?

	LR Train	LR Test	LDA Train	LDA Test[0.5]	LDA Test[0.4]	NB Train	NB Test[0.5]	NB Test[0.4]	KNN Train	KNN Test	RF Train	RF Test	Bagging Train	Bagging Test	Gradient-Boost Train	Gradient-Boost Test	AdaBoost Train	AdaBoost Test
Accuracy	0.827	0.860	0.823	0.860	0.849	0.820	0.855	0.855	0.838	0.864	0.815	0.836	0.808	0.840	0.886	0.842	0.849	0.836
AUC	0.877	0.915	0.877	0.915	NaN	0.874	0.913	NaN	0.905	0.892	0.884	0.906	0.878	0.903	0.947	0.904	0.904	0.908
Recall	0.900	0.930	0.880	0.930	0.950	0.880	0.910	0.930	0.890	0.910	0.920	0.940	0.930	0.950	0.930	0.910	0.910	0.900
Precision	0.860	0.870	0.860	0.880	0.850	0.870	0.880	0.870	0.880	0.900	0.830	0.840	0.820	0.840	0.910	0.870	0.880	0.870
F1 Score	0.880	0.900	0.870	0.900	0.900	0.870	0.900	0.900	0.880	0.900	0.870	0.890	0.870	0.890	0.920	0.890	0.890	0.880

Fig: -60: Performance metrics comparison model.

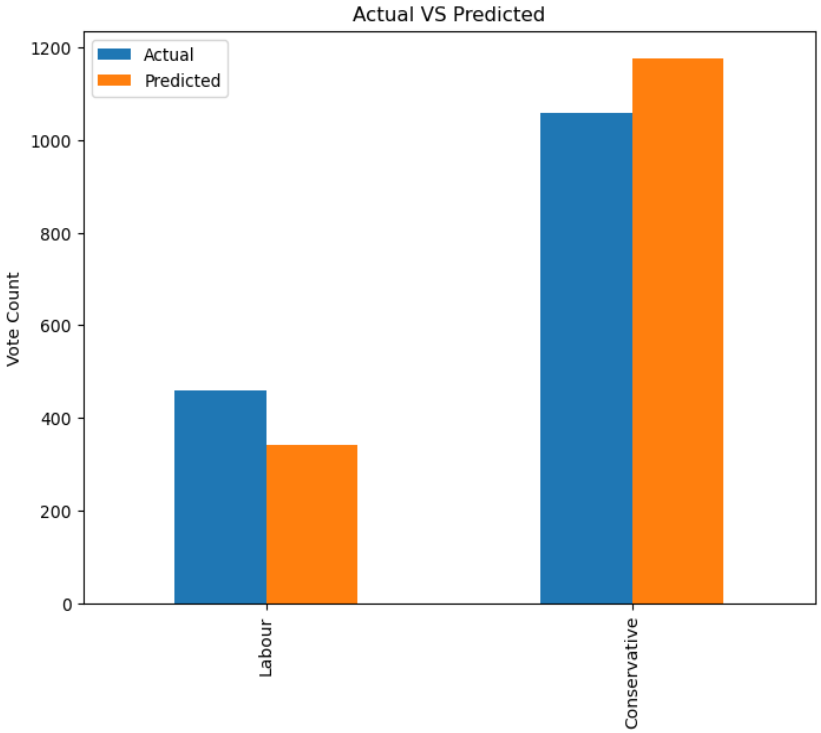




Insights:

* Performance metrics of all models are similar.
* High accuracy and AUC scores **are observed for the** majority class (Labour).
* Slightly lower performance **is observed for the** minority class (Conservative) across all models**.**
* Random Forest model **shows** slightly better performance **in predicting the** minority class**.**

- By comparing recall score and precision score of all the models, Gaussian Naïve Bayes model performs well in predicting labour or conservative party.
- The test data recall of Naïve Bayes is 90% i.e only 10% of the people who is in favour of labour party, automatically he or she will be voting against the labour party.
- The test data precision of Naïve Bayes is 88% i.e only 12% of the people were made false predictions that the votes to be in favour were actually predicted against the labour party.



Model	Prediction
Logistic Regression	conservative party
Linear Discriminant Analaysis	conservative party
K-Nearest Neighbour	labour party
Naive Bayes	labour party
Bagging(with Random Forest)	labour party
Adaptive Boosting	labour party
Gradient Boosting	conservative party

- All the models **have** performed fairly well **and have approximately similar performance metrics after tuning.**
- Random Forest is slightly better **as it** performs slightly better in predicting **the** minority class **as compared to other models. The** Random Forest model predicts a slightly higher number of votes for conservatives **&** slightly lower **number of votes for** Labour**. This might help in** adjusting for the bias **in sampling using class weights parameter.**
- Labour is getting twice the number of votes as compared to conservatives**. Thus,** Labour is most likely to come back to power.
- **The** most important features **in classifying are -: The** ratings of each of the candidate**,** Eurosceptic sentiments**, followed by the rating of** national economic condition**,** age **&** political knowledge **of their parties position.**
- Gender plays no significant role **in the classification process but is important to ascertain whether the survey is unbiased.**
- **People who have** higher Eurosceptic sentiment**, have** voted for the conservative party**.**

- Young to middle aged **voters (< 50 yrs of age) seem** more inclined to vote for Labour party**, whereas people** beyond 60 yrs **of age prefer to vote for the** conservative party**.**
- **People are** generally happy **with the** national & household economic conditions **with ~**80% voting **it fair to very good and naturally have chosen labour to continue.**

- Recommendations:
- **We can** drop the gender variable **since it has the least feature importance and after ensuring the survey is not biased.**
- Consider adding additional variables **such as** constituency/region, level of education, religion, race, immigrant status**, etc., to** enhance **the predictive power of the model.**
- Including constituency/region data **can help in clustering votes based on** geographical areas **and** enable predicting the range **(confidence interval) of the number of seats each** party is likely to win**. * If possible,** gather more data points **to further evaluate and improve the model's performance.**

Problem 2

Problem Statement:

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

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

Out of the 59 speeches of presidents in the corpus, we focus on three specific speeches: President Roosevelt's speech in 1941, President JFK's speech in 1961, and President Nixon's speech in 1973.

Our primary objective is to analyze Franklin Roosevelt's speech to gain insights into the usage of symbols and punctuation.

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

No. of Characters:

Franklin D. Roosevelt's **speech in 1941 :**7571**.**

John F. Kennedy's **speech in 1961 :** 7618**.**

Richard Nixon's **speech in 1973 :** 9991**.**

No. of Words:

Franklin D. Roosevelt's **speech in 1941 :** 1536**.**

John F. Kennedy's **speech in 1961 :** 1546**.**

Richard Nixon's **speech in 1973 :**2028**.**

No. of Sentences:

Franklin D. Roosevelt's **speech in 1941 :** 68

John F. Kennedy's **speech in 1961 :** 52

Richard Nixon's **speech in 1973 :** 69

Loaded the required packages and extract three speeches using the given code snippet.

Also import stop words and punctuation (special characters) from nltk library for data- cleaning process.

'On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn Washington\'s day the tas
k of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people was to preserve that Nation from disruption from wi
thin.\n\nIn this day the task of the people is to save that Nation and its institutions from disruption from without.\n\nTo us there has come a time, in
the midst of swift happenings, to pause for a moment and take stock -- to recall what our place in history has been, and to rediscover what we are and wh
at we may be. If we do not, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of the
human spirit. The life of a man is three-score years and ten: a little more, a little less. The life of a nation is the fullness of the measure of its wi
ll to live.\n\nThere are men who doubt this. There are men who believe that democracy, as a form of Government and a frame of life, is limited or measure
d by a kind of mystical and artificial fate that, for some unexplained reason, tyranny and slavery have become the surging wave of the future -- and that
freedom is an ebbing tide.\n\nBut we Americans know that this is not true.\n\nEight years ago, when the life of this Republic seemed frozen by a fatalist
ic terror, we proved that this is not true. We were in the midst of shock -- but we acted. We acted quickly, boldly, decisively.\n\nThese later years hav
e been living years -- fruitful years for the people of this democracy. For they have brought to us greater security and, I hope, a better understanding
that life\'s ideals are to be measured in other than material things.\n\nMost vital to our present and our future is this experience of a democracy which
successfully survived crisis at home; put away many evil things; built new structures on enduring lines; and, through it all, maintained the fact of its
democracy.\n\nFor action has been taken within the three-way framework of the Constitution of the United States. The coordinate branches of the Governmen
t continue freely to function. The Bill of Rights remains inviolate. The freedom of elections is wholly maintained. Prophets of the downfall of American
democracy have seen their dire predictions come to naught.\n\nDemocracy is not dying.\n\nWe know it because we have seen it revive--and grow.\n\nWe know
it cannot die -- because it is built on the unhampered initiative of individual men and women joined together in a common enterprise -- an enterprise und
ertaken and carried through by the free expression of a free majority.\n\nWe know it because democracy alone, of all forms of government, enlists the ful
l force of men\'s enlightened will.\n\nWe know it because democracy alone has constructed an unlimited civilization capable of infinite progress in the i
mprovement of human life.\n\nWe know it because, if we look below the surface, we sense it still spreading on every continent -- for it is the most human
e, the most advanced, and in the end the most unconquerable of all forms of human society.\n\nA nation, like a person, has a body--a body that must be fe
d and clothed and housed, invigorated and rested, in a manner that measures up to the objectives of our time.\n\nA nation, like a person, has a mind -- a
mind that must be kept informed and alert, that must know itself, that understands the hopes and the needs of its neighbors -- all the other nations that
live within the narrowing circle of the world.\n\nAnd a nation, like a person, has something deeper, something more permanent, something larger than the
sum of all its parts. It is that something which matters most to its future -- which calls forth the most sacred guarding of its present.\n\nIt is a thin
g for which we find it difficult -- even impossible -- to hit upon a single, simple word.\n\nAnd yet we all understand what it is -- the spirit -- the fa
ith of America. It is the product of centuries. It was born in the multitudes of those who came from many lands -- some of high degree, but mostly plain
people, who sought here, early and late, to find freedom more freely.\n\nThe democratic aspiration is no mere recent phase in human history. It is human
history. It permeated the ancient life of early peoples. It blazed anew in the middle ages. It was written in Magna Charta.\n\nIn the Americas its impact
has been irresistible. America has been the New World in all tongues, to all peoples, not because this continent was a new-found land, but because all th
ose who came here believed they could create upon this continent a new life -- a life that should be new in freedom.\n\nIts vitality was written into our
own Mayflower Compact, into the Declaration of Independence, into the Constitution of the United States, into the Gettysburg Address.\n\nThose who first
came here to carry out the longings of their spirit, and the millions who followed, and the stock that sprang from them -- all have moved forward constan
tly and consistently toward an ideal which in itself has gained stature and clarity with each generation.\n\nThe hopes of the Republic cannot forever tol
erate either undeserved poverty or self-serving wealth.\n\nWe know that we still have far to go; that we must more greatly build the security and the opp

Fig: -61: Sample speech by Franklin D. Roosevelt.

'Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well as a beginning -- signifying renewal, as well as cha nge. For I have sworn I before you and Almighty God the same solemn oath our forebears l prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same re volutionary beliefs for which our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity of the state, but from the hand of God.\n\nWe dare not forget today that we are the heirs of that first revolution. Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans -- born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage -- and unwilling to witness or permit the slow undoing of those human rights to which this Nati on has always been committed, and to which we are committed today at home and around the world.\n\nLet every nation know, whether it wishes us well or il l, that we shall pay any price, bear any burden, meet any hardship, support any friend, oppose any foe, in order to assure the survival and the success o f liberty.\n\nThis much we pledge -- and more.\n\nTo those old allies whose cultural and spiritual origins we share, we pledge the loyalty of faithful fr iends. United, there is little we cannot do in a host of cooperative ventures. Divided, there is little we can do -- for we dare not meet a powerful chal lenge at odds and split asunder.\n\nTo those new States whom we welcome to the ranks of the free, we pledge our word that one form of colonial control sh all not have passed away merely to be replaced by a far more iron tyranny. We shall not always expect to find them supporting our view. But we shall alwa ys hope to find them strongly supporting their own freedom -- and to remember that, in the past, those who foolishly sought power by riding the back of t he tiger ended up inside.\n\nTo those peoples in the huts and villages across the globe struggling to break the bonds of mass misery, we pledge our best efforts to help them help themselves, for whatever period is required -- not because the Communists may be doing it, not because we seek their votes, but because it is right. If a free society cannot help the many who are poor, it cannot save the few who are rich.\n\nTo our sister republics south of our bo rder, we offer a special pledge -- to convert our good words into good deeds -- in a new alliance for progress -- to assist free men and free governments in casting off the chains of poverty. But this peaceful revolution of hope cannot become the prey of hostile powers. Let all our neighbors know that we s hall join with them to oppose aggression or subversion anywhere in the Americas. And let every other power know that this Hemisphere intends to remain th e master of its own house.\n\nTo that world assembly of sovereign states, the United Nations, our last best hope in an age where the instruments of war h ave far outpaced the instruments of peace, we renew our pledge of support--to prevent it from becoming merely a forum for invective -- to strengthen its shield of the new and the weak -- and to enlarge the area in which its writ may run.\n\nFinally, to those nations who would make themselves our adversar y, we offer not a pledge but a request: that both sides begin anew the quest for peace, before the dark powers of destruction unleashed by science engulf all humanity in planned or accidental self-destruction.\n\nWe dare not tempt them with weakness. For only when our arms are sufficient beyond doubt can w e be certain beyond doubt that they will never be employed.\n\nBut neither can two great and powerful groups of nations take comfort from our present cou rse -- both sides overburdened by the cost of modern weapons, both rightly alarmed by the steady spread of the deadly atom, yet both racing to alter that uncertain balance of terror that stays the hand of mankind\'s final war.\n\nSo let us begin anew -- remembering on both sides that civility is not a sign of weakness, and sincerity is always subject to proof. Let us never negotiate out of fear. But let us never fear to negotiate.\n\nLet both sides explore what problems unite us instead of belaboring those problems which divide us.\n\nLet both sides, for the first time, formulate serious and precise proposa ls for the inspection and control of arms -- and bring the absolute power to destroy other nations under the absolute control of all nations.\n\nLet both sides seek to invoke the wonders of science instead of its terrors. Together let us explore the stars, conquer the deserts, eradicate disease, tap the oc ean depths, and encourage the arts and commerce.\n\nLet both sides unite to heed in all corners of the earth the command of Isaiah -- to "undo the heavy burdens ... and to let the oppressed go free."\n\nAnd if a beachhead of cooperation may push back the jungle of suspicion, let both sides join in creatin

Fig: -62: Sample Speech by John F. Kennedy's.

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we share togeth er:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and of destructive conflic t at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilities greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis past year saw far-reaching resu lts from our new policies for peace. By continuing to revitalize our traditional friendships, and by our missions to Peking and to Moscow, we were able t o establish the base for a new and more durable pattern of relationships among the nations of the world. Because of America\'s bold initiatives, 1972 wil l be long remembered as the year of the greatest progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flimsy peace which is merely an interlude between wars, but a peace which can endure for generations to come.\n\nIt is important that we understand both the necessity and the limitations of America\'s role in maintaining that peace.\n\nUnless we in America work to preserve the peace, there will be no peace.\n\nUnless we in America work to preserve freedom, there will be no freedom.\n\nBut let us clearly understand the new nature of America \'s role, as a result of the new policies we have adopted over these past four years.\n\nWe shall respect our treaty commitments.\n\nWe shall support vig orously the principle that no country has the right to impose its will or rule on another by force.\n\nWe shall continue, in this era of negotiation, to work for the limitation of nuclear arms, and to reduce the danger of confrontation between the great powers.\n\nWe shall do our share in defending peace and freedom in the world. But we shall expect others to do their share.\n\nThe time has passed when America will make every other nation\'s conflict our own, or make every other nation\'s future our responsibility, or presume to tell the people of other nations how to manage their own affairs.\n\nJust as we respect the right of each nation to determine its own future, we also recognize the responsibility of each nation to secure its own future.\n\nJust as America\'s role is indispensable in preserving the world\'s peace, so is each nation\'s role indispensable in preserving its own peace.\n\nTogether with the rest of the world, let us resolve to move forward from the beginnings we have made. Let us continue to bring down the walls of hostility which have d ivided the world for too long, and to build in their place bridges of understanding -- so that despite profound differences between systems of governmen t, the people of the world can be friends.\n\nLet us build a structure of peace in the world in which the weak are as safe as the strong -- in which each respects the right of the other to live by a different system -- in which those who would influence others will do so by the strength of their ideas, and not by the force of their arms.\n\nLet us accept that high responsibility not as a burden, but gladly -- gladly because the chance to build such a peace is the noblest endeavor in which a nation can engage; gladly, also, because only if we act greatly in meeting our responsibilities abroad will we remain a great Nation, and only if we remain a great Nation will we act greatly in meeting our challenges at home.\n\nWe have the chance today to do more than e ver before in our history to make life better in America -- to ensure better education, better health, better housing, better transportation, a cleaner e nvironment -- to restore respect for law, to make our communities more livable -- and to insure the God-given right of every American to full and equal o pportunity.\n\nBecause the range of our needs is so great -- because the reach of our opportunities is so great -- let us be bold in our determination to meet those needs in new ways.\n\nJust as building a structure of peace abroad has required turning away from old policies that failed, so building a new era of progress at home requires turning away from old policies that have failed.\n\nAbroad, the shift from old policies to new has not been a retreat fr om our responsibilities, but a better way to peace.\n\nAnd at home, the shift from old policies to new will not be a retreat from our responsibilities, b ut a better way to progress.\n\nAbroad and at home, the key to those new responsibilities lies in the placing and the division of responsibility. We have lived too long with the consequences of attempting to gather all power and responsibility in Washington.\n\nAbroad and at home, the time has come to turn away from the condescending policies of paternalism -- of "Washington knows best."\n\nA person can be expected to act responsibly only if he has responsi

Fig: -63: Sample speech by Richard Nixon's

2.2 Remove all the stop words from all three speeches.

Forming a data frame with the President's speeches with the President's names as index.

Text Pre-processing:

Before analyzing text data, we perform essential pre-processing steps to extract valuable insights. These steps include: Converting the text to a consistent case (lower or upper).

- Removing special symbols and punctuations.
- Removing extra white spaces (not applicable in any speech).
- Removing stop words.
- Stemming words to their original form.
- Retaining numbers (as they may be relevant, such as the year mentioned in a president's speech).

Lower case conversion:

All the speeches are converted to a lower case. The idea is to bring all of the text to one case either lower or upper because python is case sensitive and it views 'cat' & 'Cat' as different.

Special Characters & Punctuations Removal:

We utilize the regular expressions library to replace non-word and non-space characters with null. By using the pattern [^\w\s], where '\w' represents letters, numbers, and underscores, and '\s' denotes spaces, we effectively exclude special characters and punctuations (except underscores) from the text data.

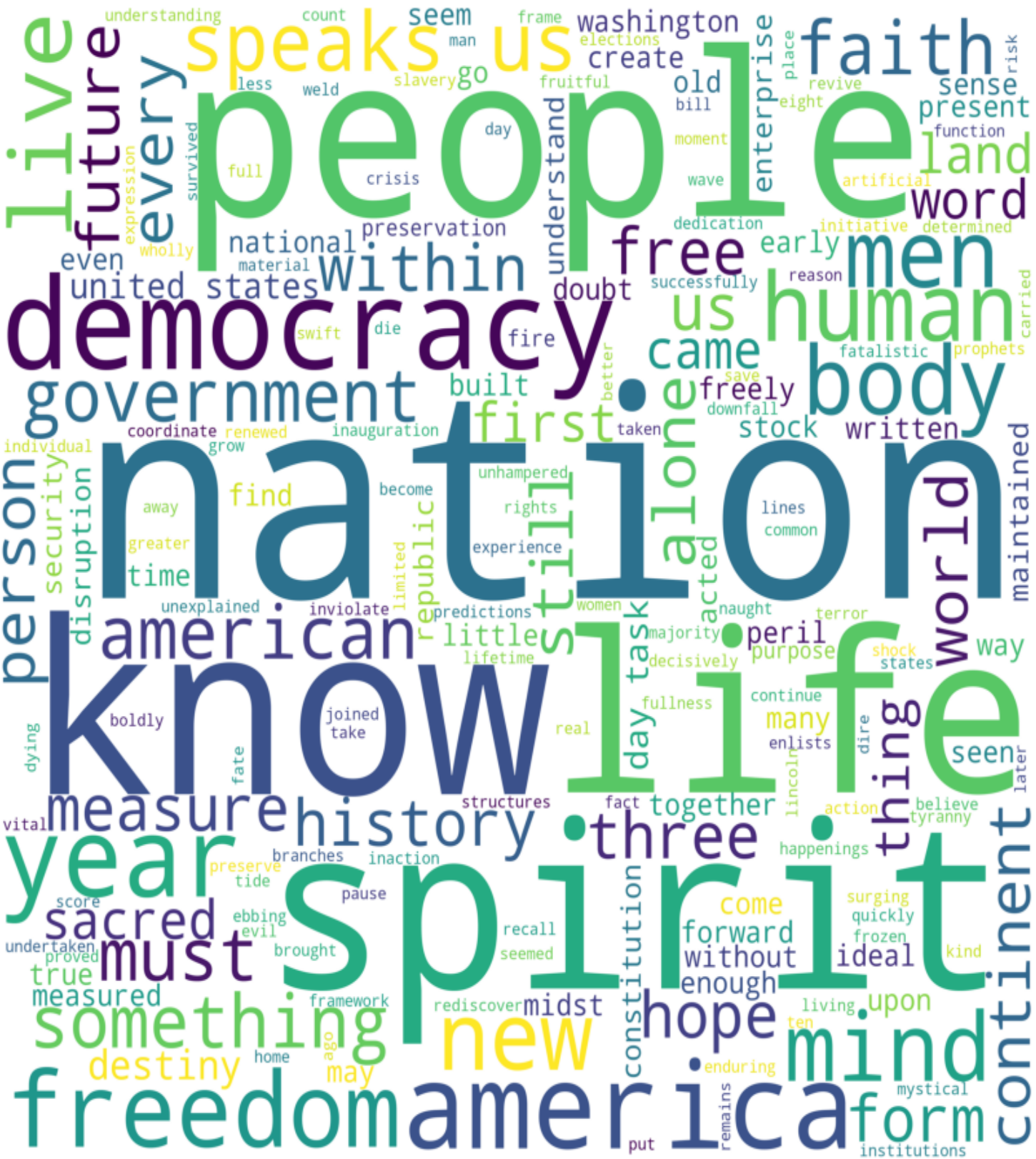
2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)

Year	President's Name	Sl. No.	Top words	Count
1941	**Franklin D. Roosevelt**	**1**	**nation**	**11**
		2	**know**	**10**
		3	**spirit**	**9**
1961	**John F. Kennedy**	**1**	**world**	**8**
		2	**sides**	**8**
		3	**new**	**7**
1973	**Richard Nixon**	**1**	**peace**	**19**
		2	**world**	**16**
		3	**new**	**15**

Table 2.6 Most frequent words of each President's speech – before stemming

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

- **Word Cloud for 1941-Roosevelt Speech:**



- **Word Cloud for 1961-Kennedy Speech:**

- **Word Cloud for 1973-Nixon Speech:**

Inference:

- Most Frequent words are America, let, us, nation
- Less Frequent words are flimsy, adopted, saw

Conclusion:

This project data presented from '1941-Roosevelt.txt', '1961-Kennedy.txt' and '1973-Nixon.txt', we analyzed some interesting insights like the number of characters, words, and sentences from the speeches. To Identify the strength and the sentiment of these presidential speeches the stop words were removed (punctuation and lowering the characters were removed) along with stemming. We analyzed some of the common words from their speeches which inspired many Americans. word cloud method which visually show most common words to least common words.