Greate Learning

Machine Learning (ML) Project Report

By

Name: Abhishek Pradhan

BATCH: PGPDSBA.O.FEB23.B

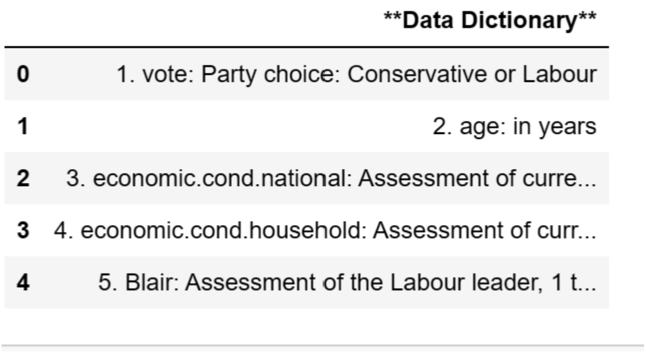
1.1) Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.	4
1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.	7
1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed.	4
1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)	4
1.5) Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)	4
1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.	7

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)	7
1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.	5
2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)	3
2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.	3
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)	3
2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)	

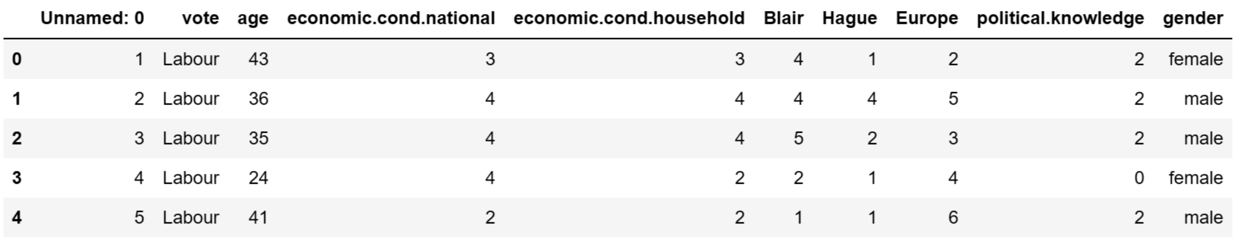
Problem 1: You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that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

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

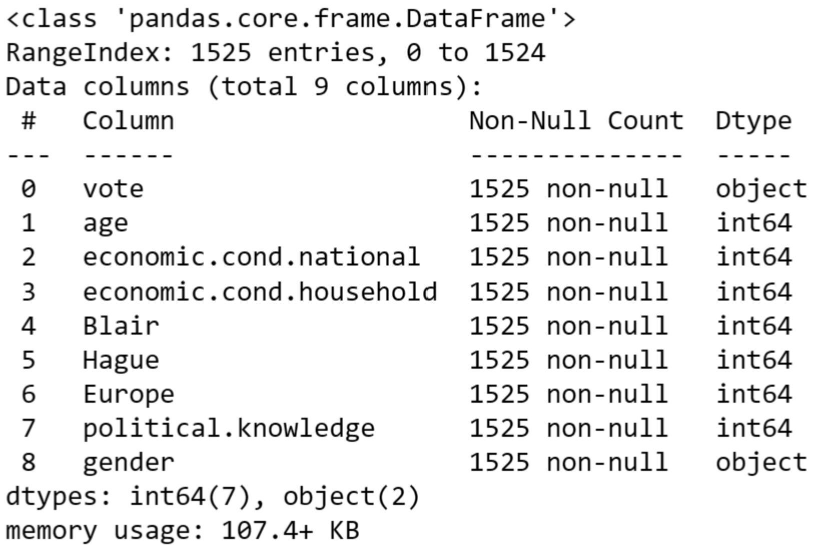
Ans:



	Data Dictionary
0	1. vote: Party choice: Conservative or Labour
1	2. age: in years
2	3. economic.cond.national: Assessment of curre...
3	4. economic.cond.household: Assessment of curr...
4	5. Blair: Assessment of the Labour leader, 1 t...



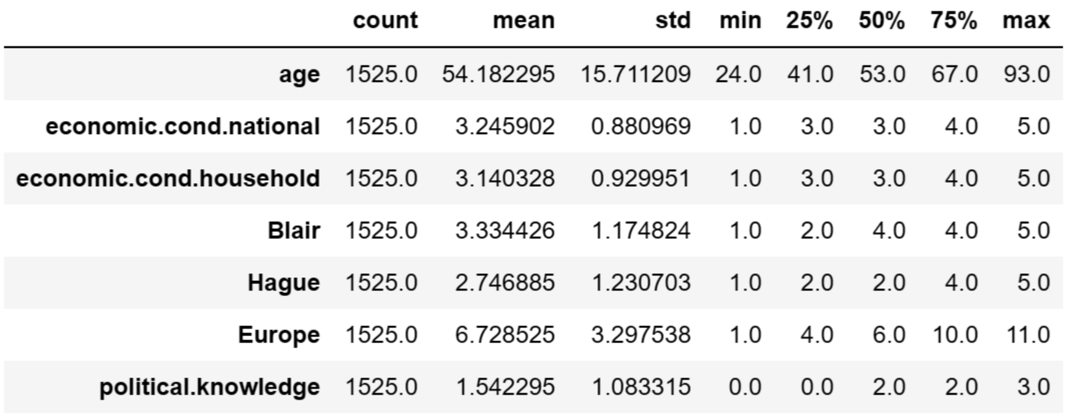
	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	1	Labour	43	3	3	4	1	2	2	female
1	2	Labour	36	4	4	4	4	5	2	male
2	3	Labour	35	4	4	5	2	3	2	male
3	4	Labour	24	4	2	2	1	4	0	female
4	5	Labour	41	2	2	1	1	6	2	male



```
<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
```

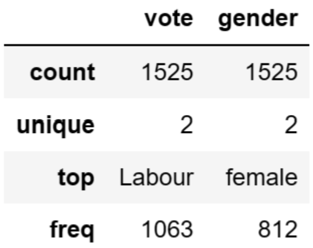
The column “Unnamed : 0” is removed from the dataset. There are no null values in the dataset.

There are 1525 rows and 9 columns

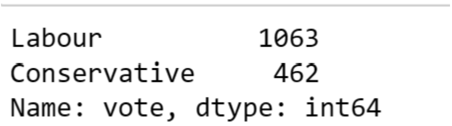


	count	mean	std	min	25%	50%	75%	max
age	1525.0	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	2.746885	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

Here is the description of the numerical columns from the dataset.



	vote	gender
count	1525	1525
unique	2	2
top	Labour	female
freq	1063	812

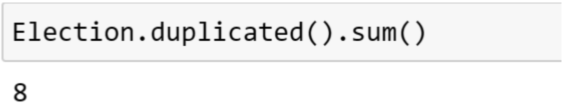


```
Labour          1063
Conservative     462
Name: vote, dtype: int64
```

In the analyzed dataset, two primary unique values emerged for the 'vote' variable: Conservative and Labour. Notably, the Labour vote counts substantially higher at 1036 compared to the Conservative count. Moreover, in terms of gender, the data demonstrates that female voters outnumber male voters.

Moving to the 'age' variable, this continuous parameter ranges from a minimum of 24 to a maximum of 93. With an average age of 54, it becomes evident that voters span a diverse range of age groups.

Shifting focus to economic assessments, the highest possible assessment of current national economic conditions is 5. Meanwhile, the average assessment for current household economic conditions hovers around 3.



```
Election.duplicated().sum()
```

```
8
```

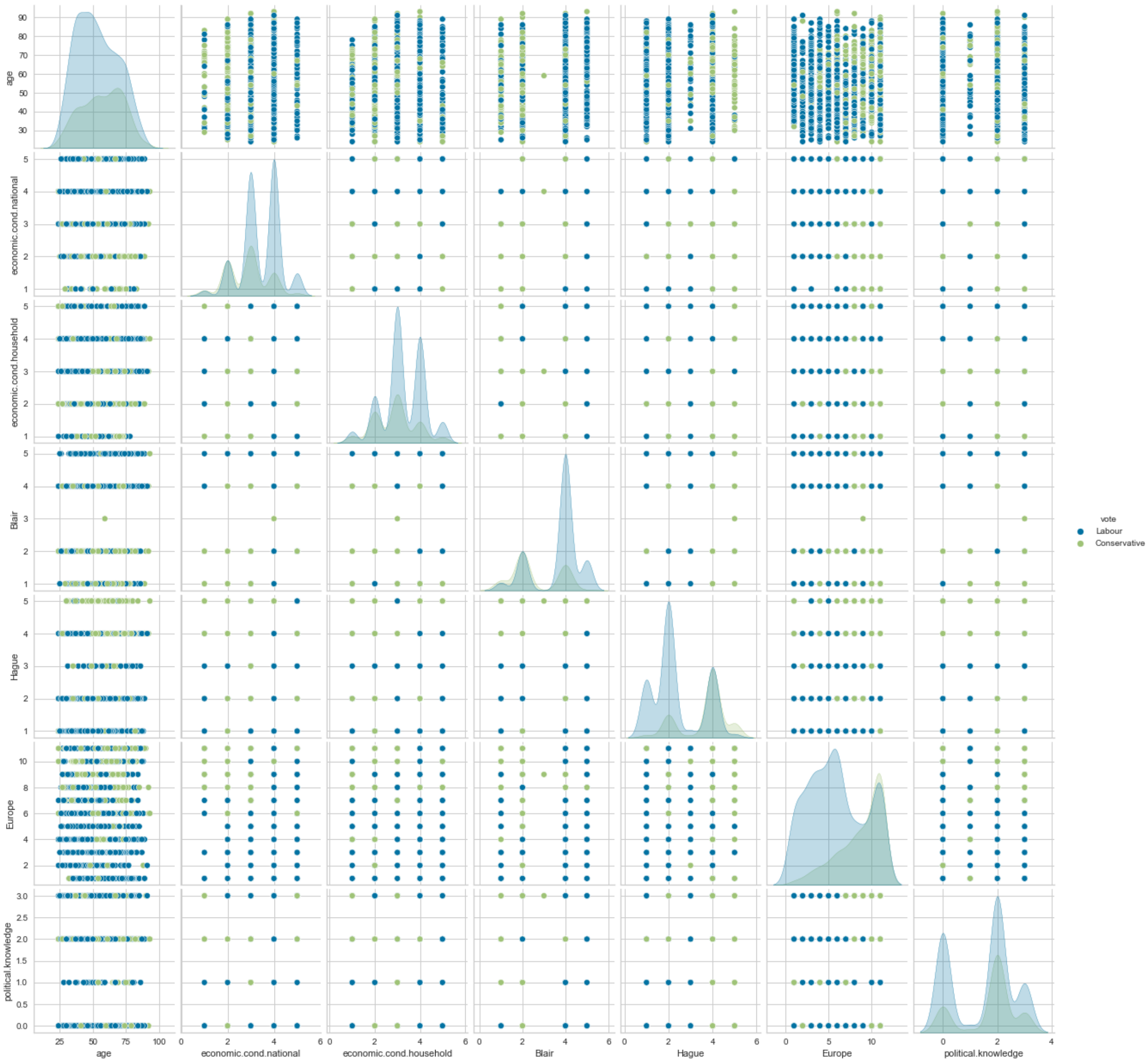
Number of duplicate rows = 8

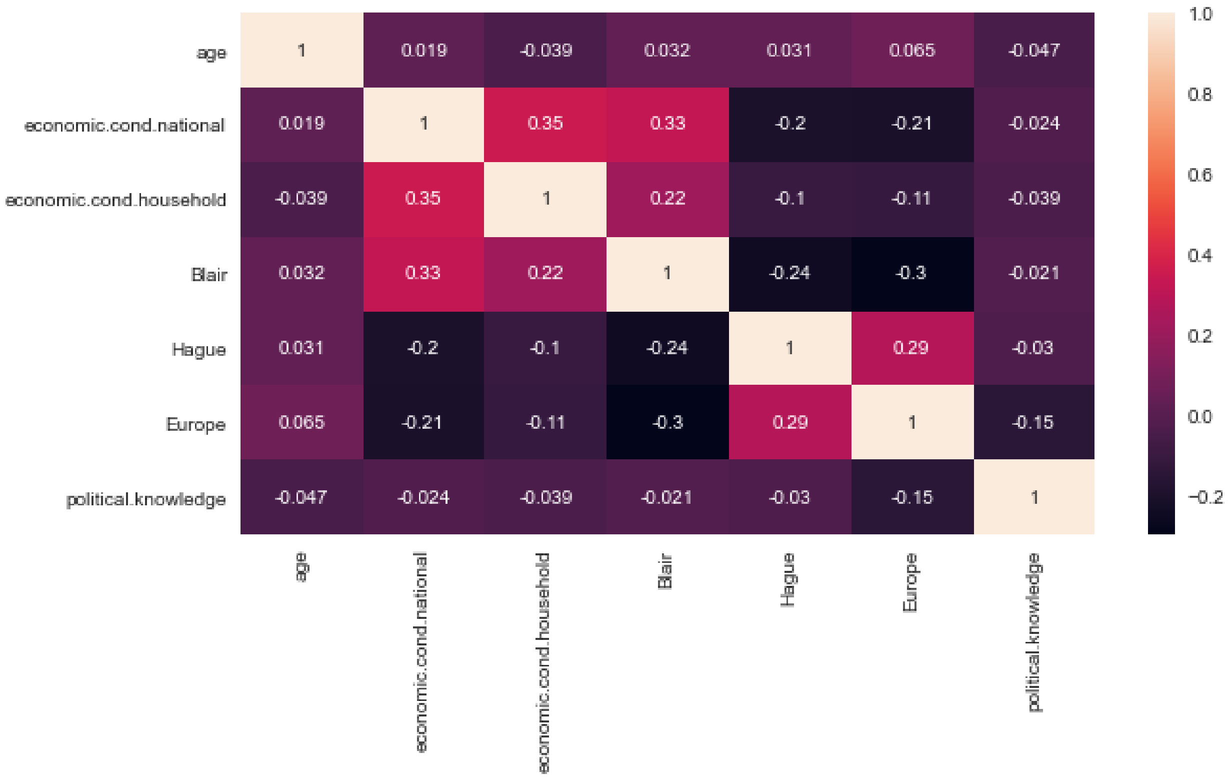
Before (1525, 9)

After (1517, 9)

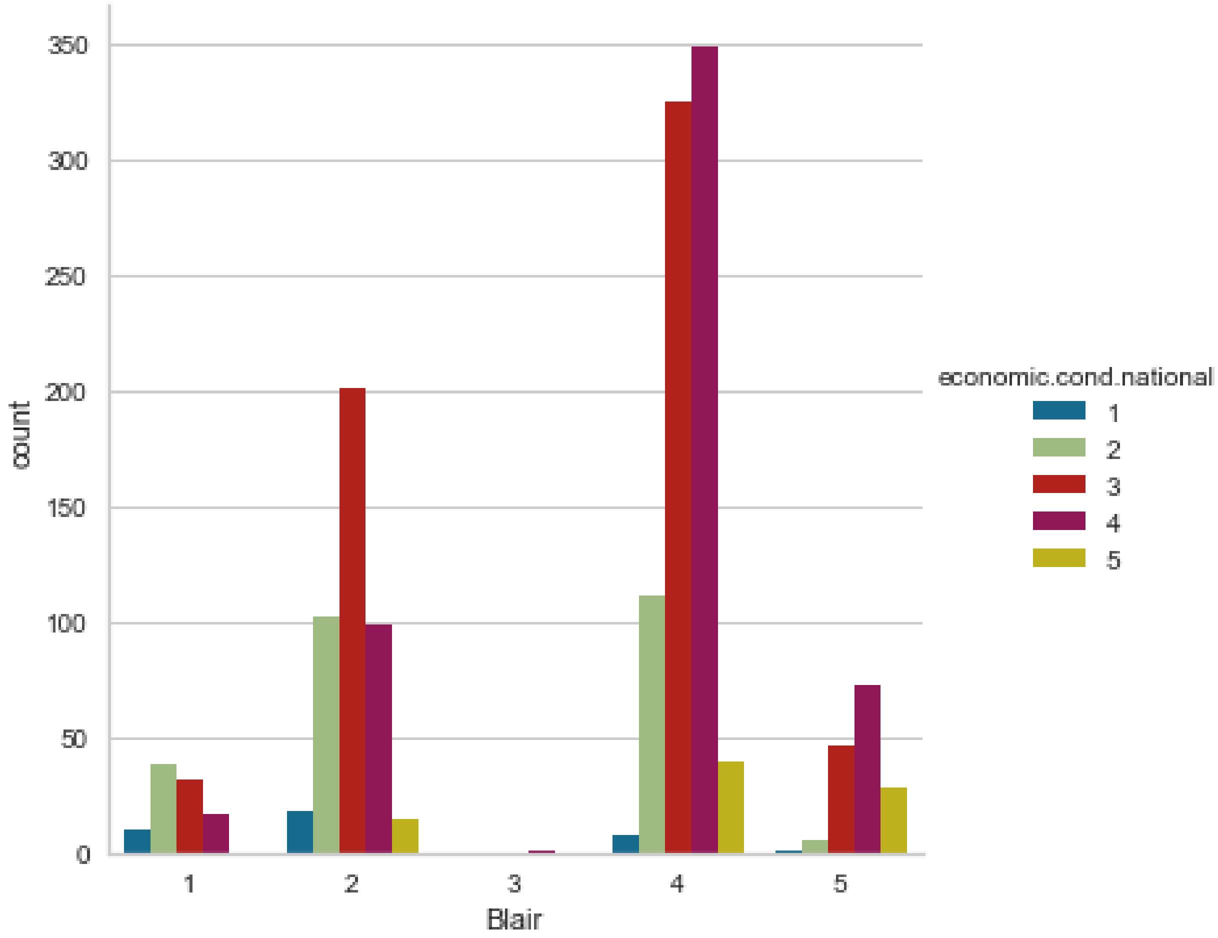
1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots. Appropriate plots for categorical variables. Inferences on each

plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

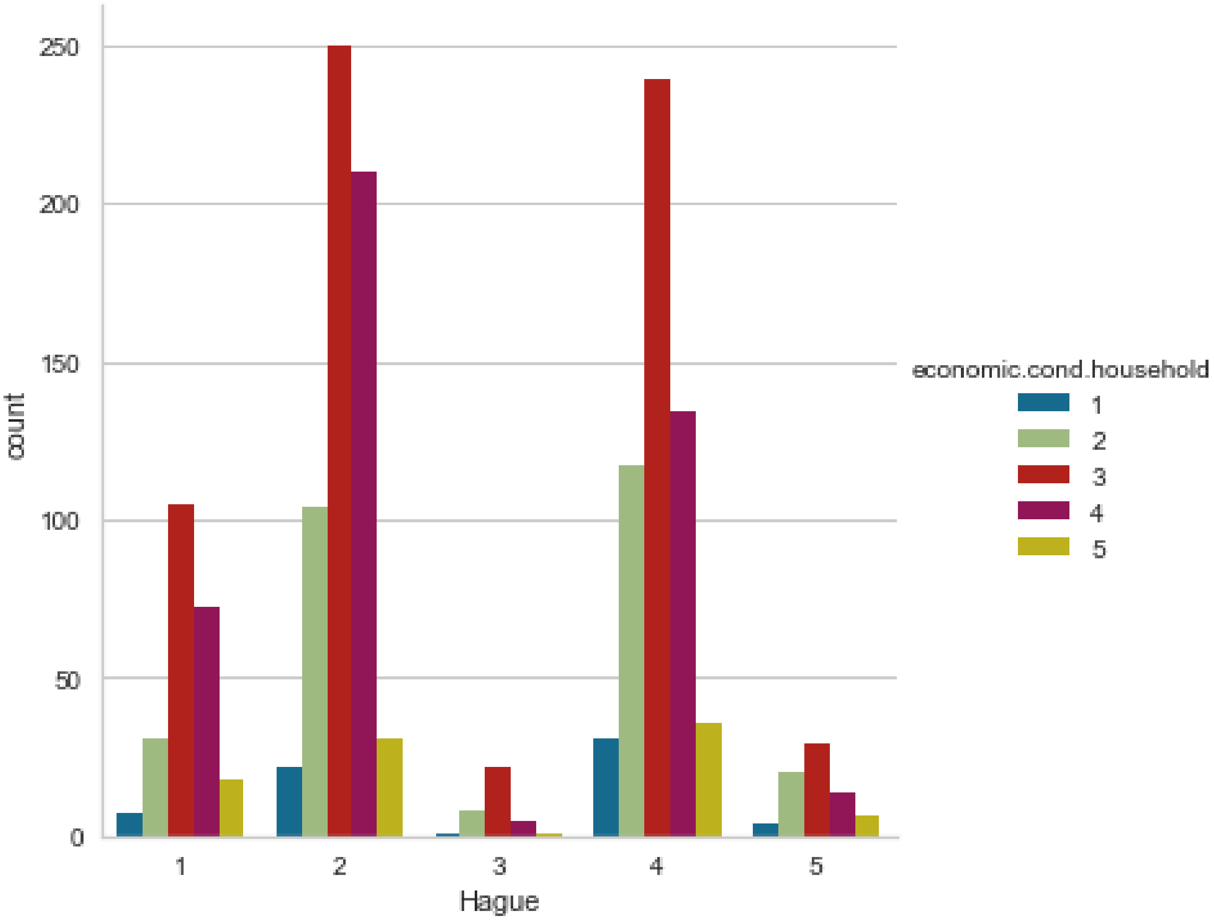


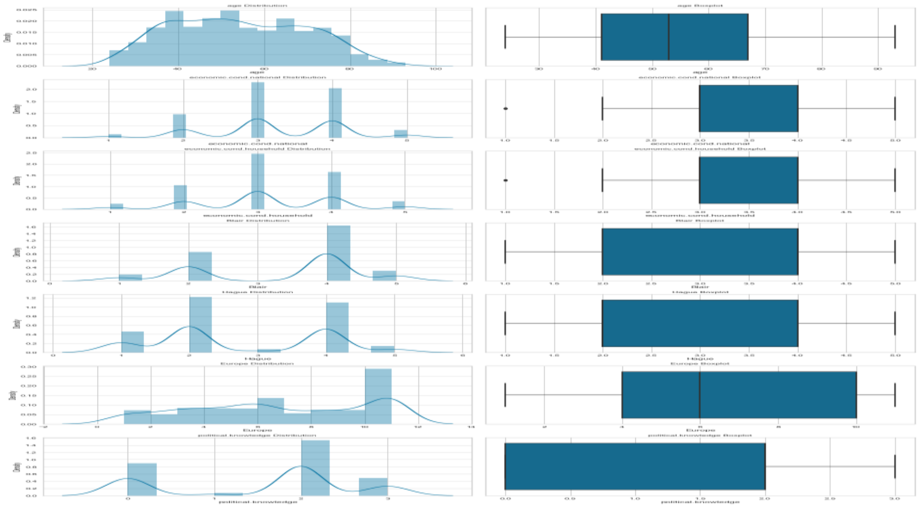


There is very less correlation between the variables. There is a positive correlation between Blair with both economic condition household and economic condition national.
Same with Hague and Europe but very weak.

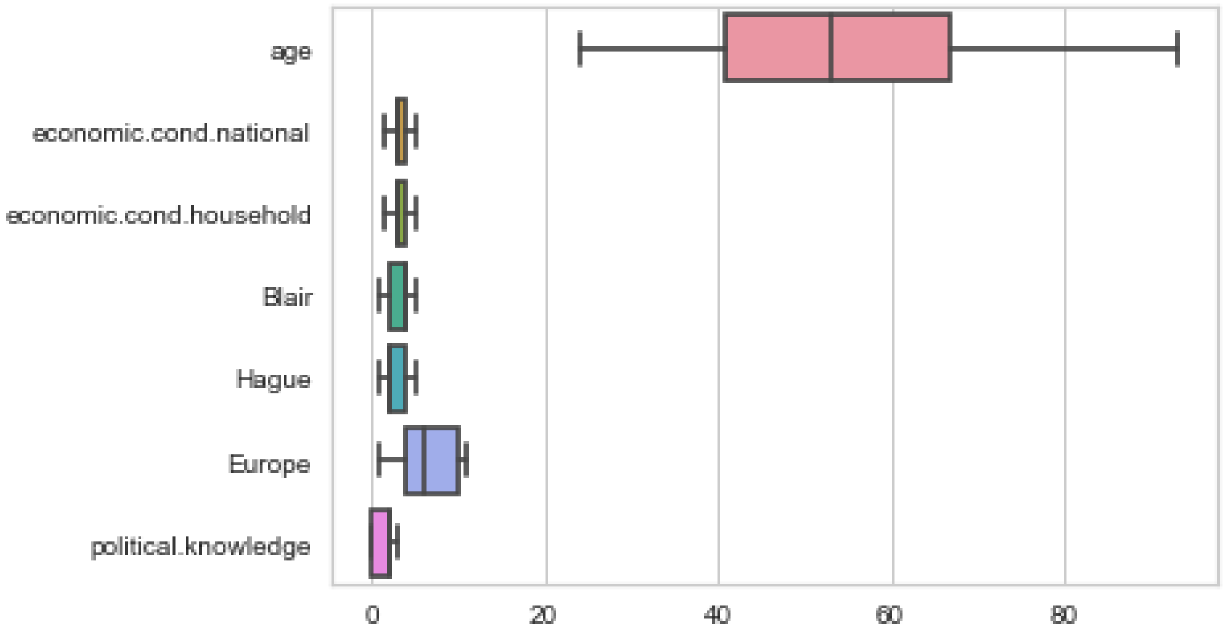


current national economic conditions with Blair shows no 3 cluster have very less distributio n whereas no 4 cluster have more distribution





Here we could see some outliers on some feature .So need to treat them



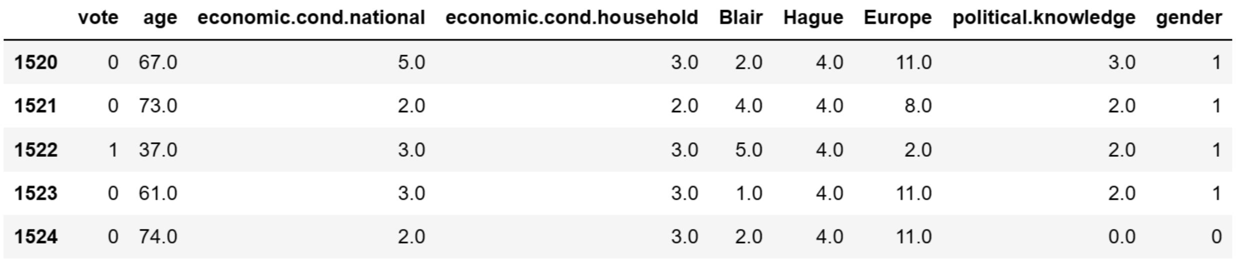
Outliers after the treatment.

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

Ans:

Need to convert the Obeject data to categorical variables.

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


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

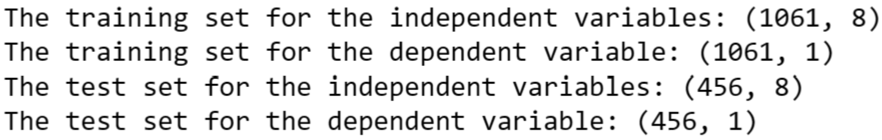


	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1520	0	67.0	5.0	3.0	2.0	4.0	11.0	3.0	1
1521	0	73.0	2.0	2.0	4.0	4.0	8.0	2.0	1
1522	1	37.0	3.0	3.0	5.0	4.0	2.0	2.0	1
1523	0	61.0	3.0	3.0	1.0	4.0	11.0	2.0	1
1524	0	74.0	2.0	3.0	2.0	4.0	11.0	0.0	0

.Then again change float to int type.Then the method of scaling performed only on the ‘age’ variable is the Z-score scaling.
All the model prediction are done with scaled data

Train-Test

Split Separating independent (train) and dependent (test) variables for the linear regression model
X = independent (train) variables
Y = dependent (test) variables



```
The training set for the independent variables: (1061, 8)
The training set for the dependent variable: (1061, 1)
The test set for the independent variables: (456, 8)
The test set for the dependent variable: (456, 1)
```

To construct predictive models for the given dataset, it's essential to divide the data into distinct train and test sets. This process, known as data splitting, is crucial for assessing the models' performance accurately. In this case, a 70:30 split has been chosen, whereby 70% of the data is designated for training and the remaining 30% for testing.

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

Ans:

After performing the **Logistic Regression** could find the Accuracy for the Train dat a is 83.12%

And for test it is 83.55%.

The coefficient for age is -0.23271304779231494
The coefficient for economic.cond.national is 0.6285844283087529
The coefficient for economic.cond.household is 0.06324265467366959
The coefficient for Blair is 0.6007162474217334
The coefficient for Hague is -0.8231886776961427
The coefficient for Europe is -0.2116674456980624
The coefficient for political.knowledge is -0.32193521834814426
The coefficient for gender is 0.19200247767643372

The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.
Economic.cond.national have more positive coeffiecient . A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase the vote

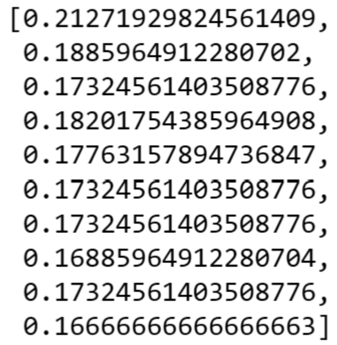
LDA:
Accuracy score of training data:83.4%
Accuracy score of test data:83.3%

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

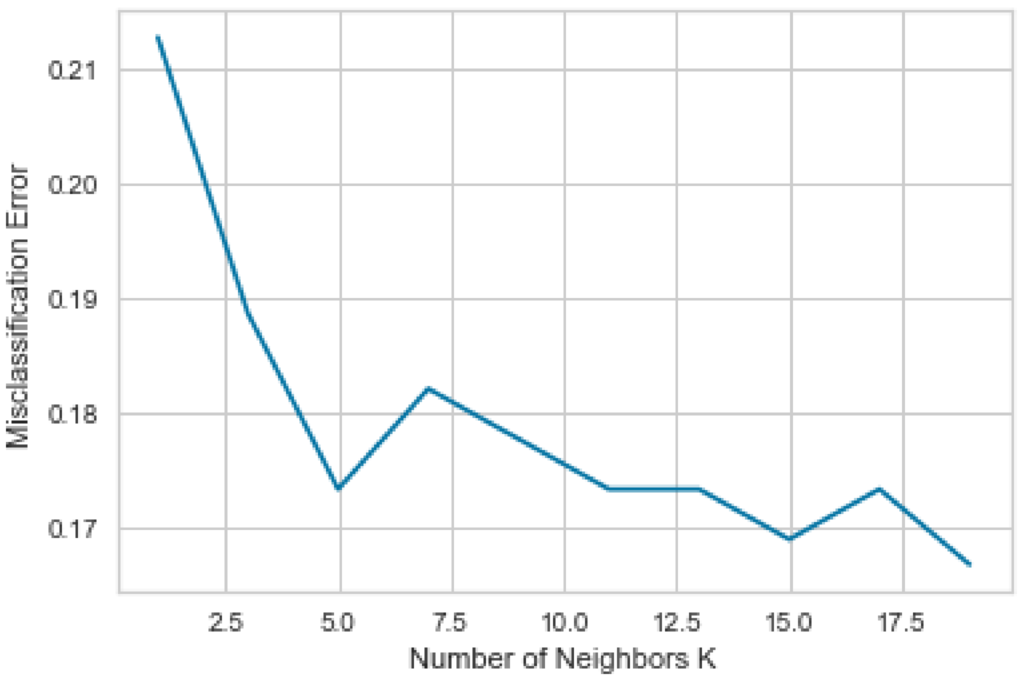
Ans:

KNN

misclassification error:



```
[0.21271929824561409,
 0.1885964912280702,
 0.17324561403508776,
 0.18201754385964908,
 0.17763157894736847,
 0.17324561403508776,
 0.17324561403508776,
 0.16885964912280704,
 0.17324561403508776,
 0.16666666666666663]
```



The number of neighbors(K) in KNN is a hyperparameter that you need choose at t he time of model buildin g. You can think of K as a controlling variable for the predic tion model. n_neighbors = 15
Accuracy score of training data:84.6%
Accuracy score of test data:83.1%

Gaussian Naive Bayes
Accuracy score of training data:83.5%
Accuracy score of test data:82.2%

1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.

Ans:
Bagging with randomforest:
A Bagging classifier
Accuracy score of training data:83.69%
Accuracy score of test data:82.23%

AdaBoost (Adaptive Boosting):
Accuracy score of training data:85.01%
Accuracy score of test data:81.35%

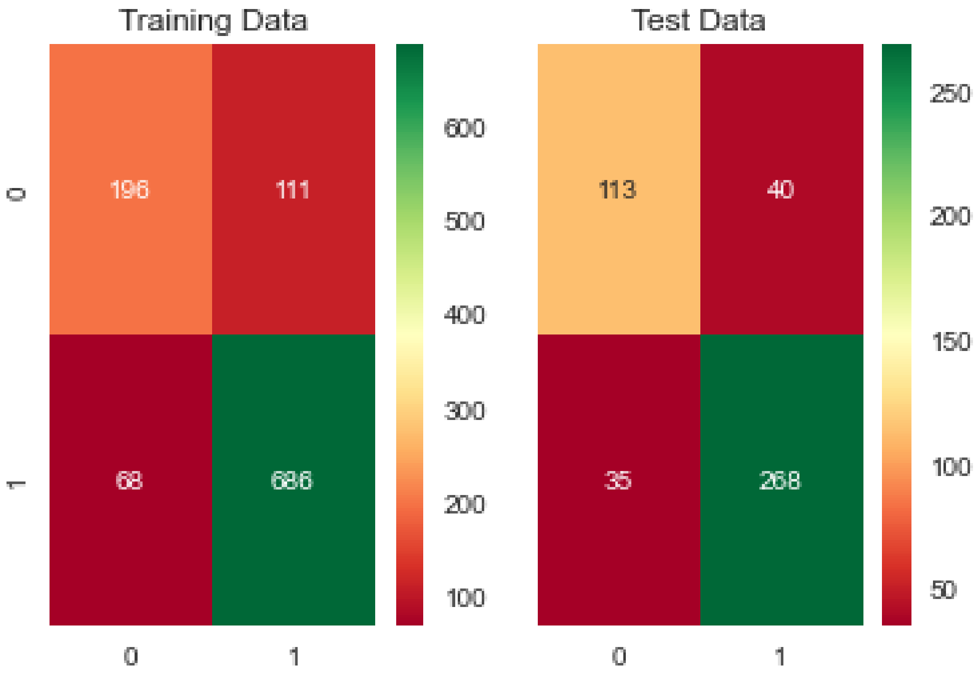
Gradient Boosting:
Accuracy score of training data:88.03%

Accuracy score of test data:83.33%

1.7) 1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)

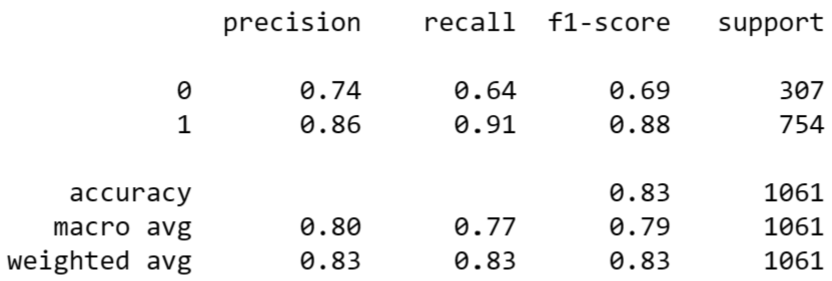
Ans:

Confusion matrix on the training and test data:

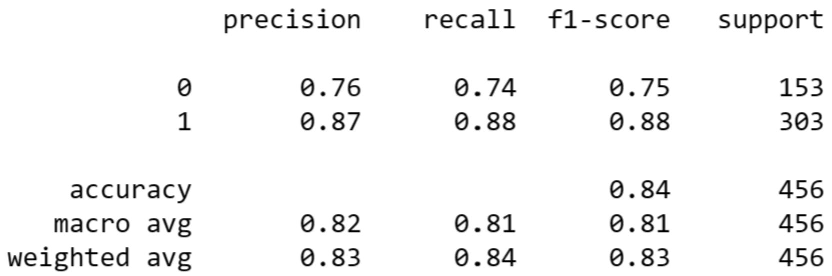


Training data:
True Negative : 196
False Positive : 111
False Negative : 68
True Positive : 686
Test data:
True Negative : 113
False Positive : 40
False Negative : 35
True Positive : 268

Classification Report:

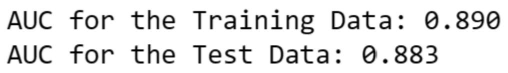


```
              precision    recall  f1-score   support

           0       0.74      0.64      0.69       307
           1       0.86      0.91      0.88       754

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

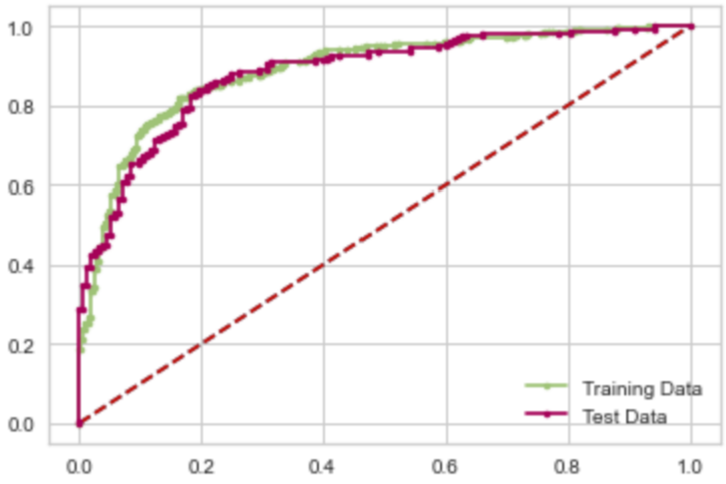


```
              precision    recall  f1-score   support

           0       0.76      0.74      0.75       153
           1       0.87      0.88      0.88       303

    accuracy                           0.84       456
   macro avg       0.82      0.81      0.81       456
weighted avg       0.83      0.84      0.83       456
```

AUC and ROC for the training and test data:



```
AUC for the Training Data: 0.890
AUC for the Test Data: 0.883
```

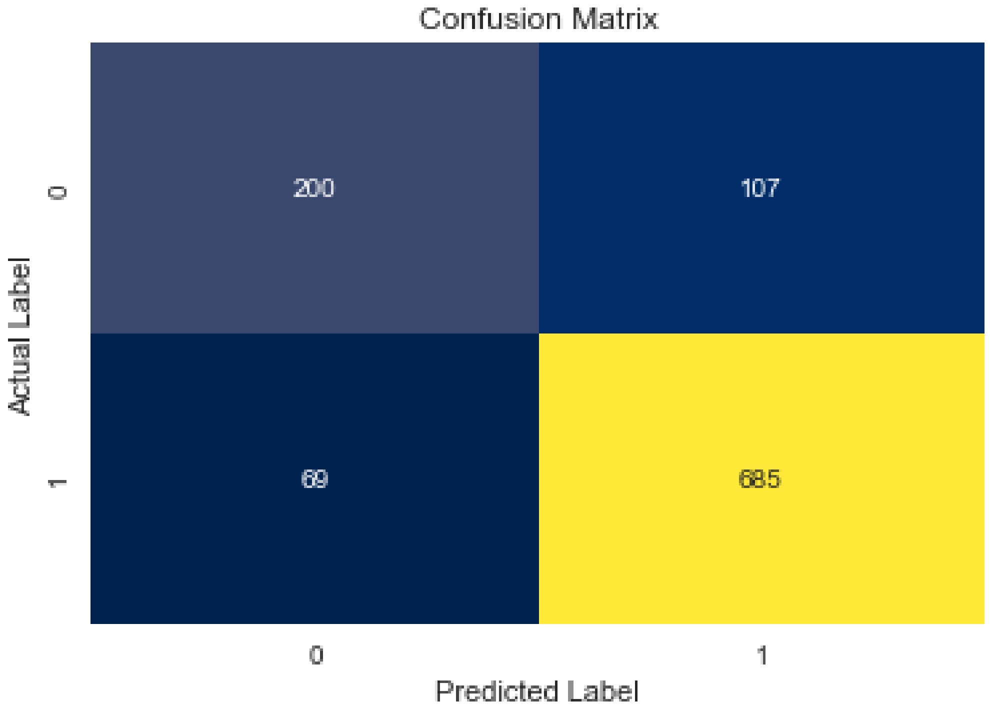


Train Data:
AUC: 89%
Accuracy: 83%
precision : 86%
recall : 91%
f1 :88%
Test Data:
AUC: 88.3%
Accuracy: 84%
precision: 87%

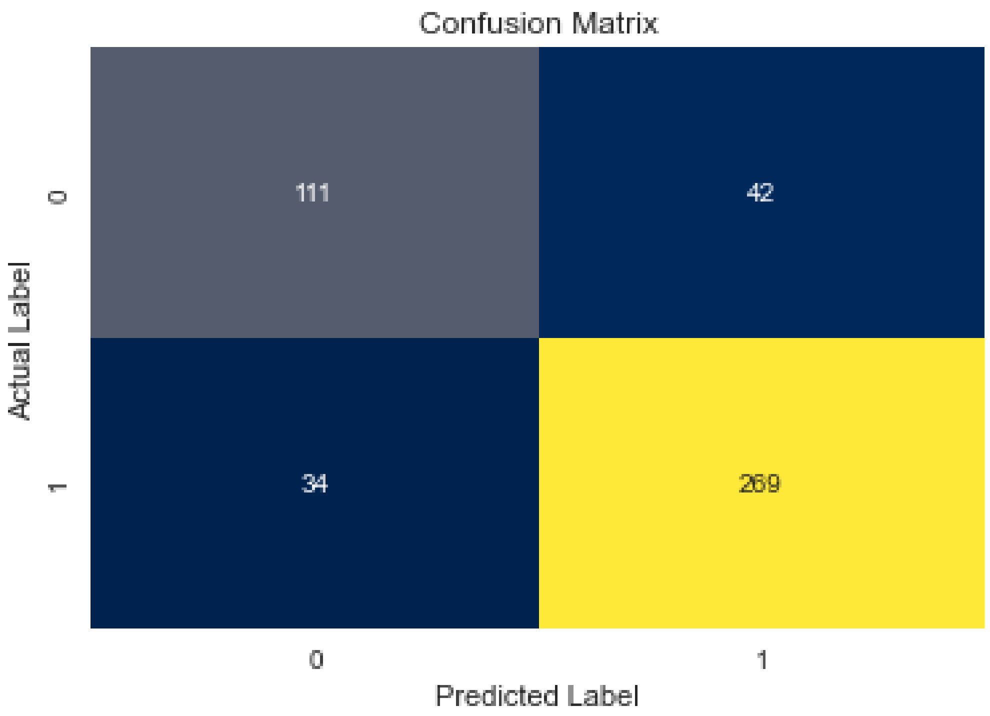
recall : 88%
f1 : 88%

Training and Test set results are almost similar, this proves no overfitting or underfitting.

LinearDiscriminantAnalysis



Training Matrix



Test Matrix

```
              precision    recall  f1-score   support

           0       0.74      0.65      0.69       307
           1       0.86      0.91      0.89       754

    accuracy                           0.83      1061
   macro avg       0.80      0.78      0.79      1061
weighted avg       0.83      0.83      0.83      1061
```

```
              precision    recall  f1-score   support

           0       0.77      0.73      0.74       153
           1       0.86      0.89      0.88       303

    accuracy                           0.83       456
   macro avg       0.82      0.81      0.81       456
weighted avg       0.83      0.83      0.83       456
```

Trainig data:
True Negative : 200
False Positive : 107
False Negative : 69
True Positive : 685
Test data:

True Negative : 111
False Positive : 42
False Negative : 34
True Positive : 269

Almost all the models performed well with accuracy between 82% to 84%. Gradient boosting improved the accuracy to 87% so it is better model for to predict which party a voter will vote

Comparing all the model ,Gradient boosting model is best model for this dataset with accuracy of 87% in both training and test set

Almost all the models performed well with accuracy between 82% to 84% with scaled data. But Gradient boosting is best and optimised model with accuracy of 87% and also best AUC,Precision,f1 score, Recall .

Problem 2:

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

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

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

The number of characters in Roosevelt speech are: 7571

The number of characters in Kennedy speech are: 7618

The number of characters in Nixon speech are: 9991

The number of Words in Roosevelt speech are: 1536

The number of Words in Kennedy speech are: 1546

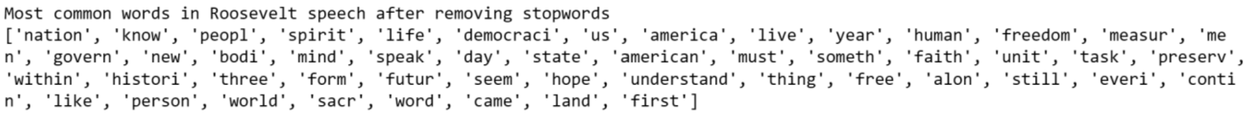
The number of Words in Nixon speech are: 2028

The number of sentences in Roosevelt's speech: 68

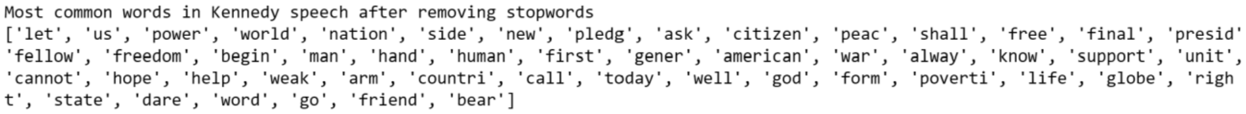
The number of sentences in Kennedy's speech: 52

The number of sentences in Nixon's speech: 69

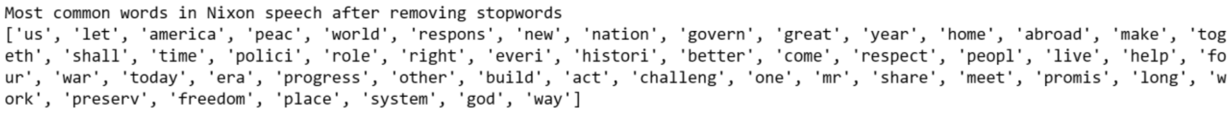
2.2 Remove all the stopwords from all three speeches.



```
Most common words in Roosevelt speech after removing stopwords
['nation', 'know', 'peopl', 'spirit', 'life', 'democraci', 'us', 'america', 'live', 'year', 'human', 'freedom', 'measur', 'me
n', 'govern', 'new', 'bodi', 'mind', 'speak', 'day', 'state', 'american', 'must', 'someth', 'faith', 'unit', 'task', 'preserv',
'within', 'histori', 'three', 'form', 'futur', 'seem', 'hope', 'understand', 'thing', 'free', 'alon', 'still', 'everi', 'conti
n', 'like', 'person', 'world', 'sacr', 'word', 'came', 'land', 'first']
```



```
Most common words in Kennedy speech after removing stopwords
['let', 'us', 'power', 'world', 'nation', 'side', 'new', 'pledg', 'ask', 'citizen', 'peac', 'shall', 'free', 'final', 'presid'
'fellow', 'freedom', 'begin', 'man', 'hand', 'human', 'first', 'gener', 'american', 'war', 'alway', 'know', 'support', 'unit',
'cannot', 'hope', 'help', 'weak', 'arm', 'countri', 'call', 'today', 'well', 'god', 'form', 'poverti', 'life', 'globe', 'righ
t', 'state', 'dare', 'word', 'go', 'friend', 'bear']
```



```
Most common words in Nixon speech after removing stopwords
['us', 'let', 'america', 'peac', 'world', 'respons', 'new', 'nation', 'govern', 'great', 'year', 'home', 'abroad', 'make', 'tog
eth', 'shall', 'time', 'polici', 'role', 'right', 'everi', 'histori', 'better', 'come', 'respect', 'peopl', 'live', 'help', 'fo
ur', 'war', 'today', 'era', 'progress', 'other', 'build', 'act', 'challeng', 'one', 'mr', 'share', 'meet', 'promis', 'long', 'w
ork', 'preserv', 'freedom', 'place', 'system', 'god', 'way']
```

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)

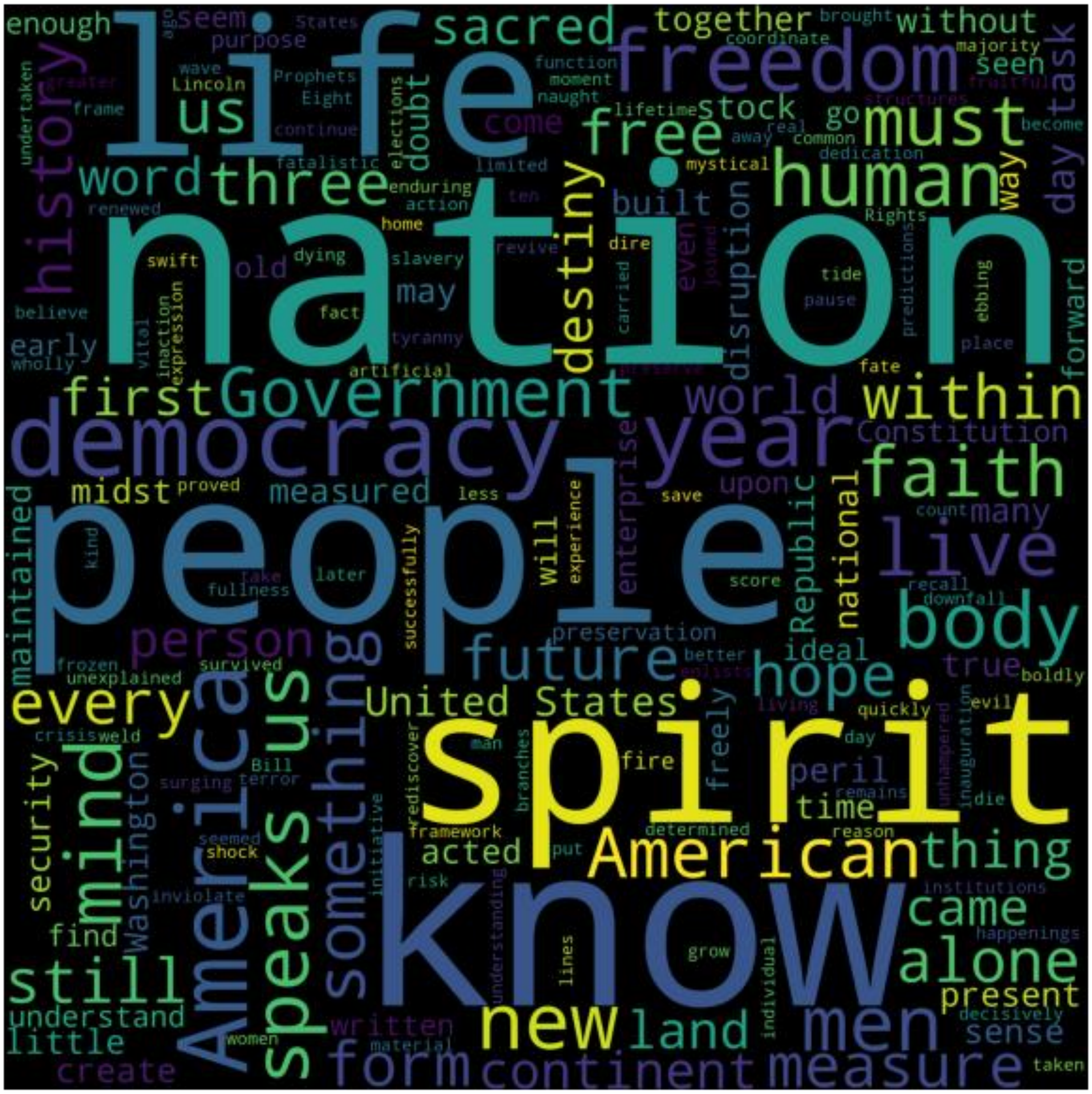
Top three words in Roosevelt's speech(after removing the stopwords): [('nation', 17), ('know', 10), ('peopl', 9)]

Top three words in Kennedy's speech(after removing the stopwords): [('let', 16), ('us', 12), ('power', 9)]

Top three words in Nixon's speech(after removing the stopwords): [('us', 26), ('let', 22), ('america', 21)]

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

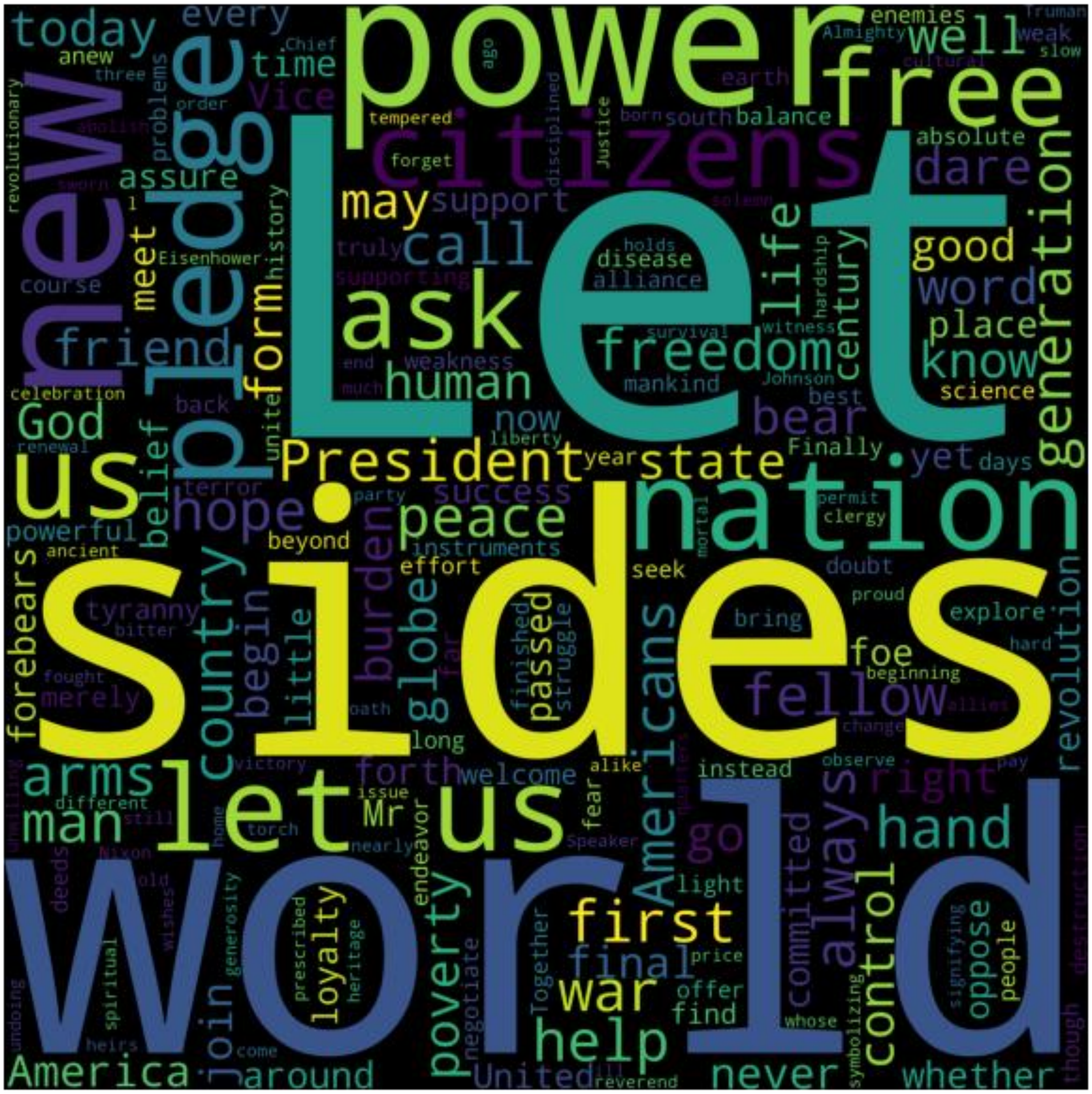
Roosevelt speech Word Cloud (after cleaning)!!



Most Frequent words are nation, people, spirit, life

Less Frequent words are weld, kind, women, moment

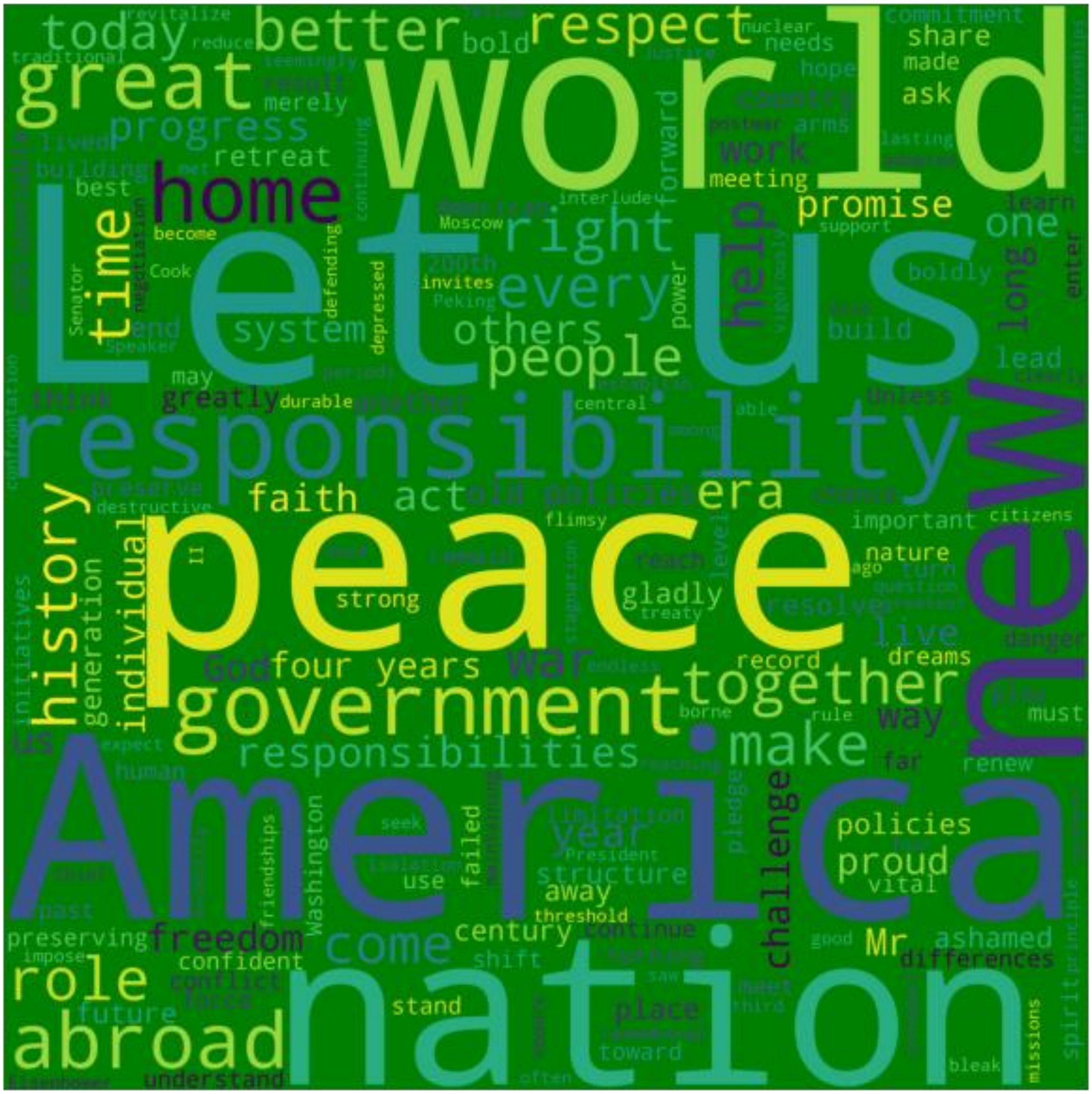
Kennedy speech



Most Frequent words are let, world, slides, power

Less Frequent words are best, wishes, slow

Nixon speech



Most Frequent words are America, let, us, nation

Less Frequent words are flimsy, adopted, saw