Machine Learning Project

Sayan Mondal
Date:25/09/2022

Contents Page No.

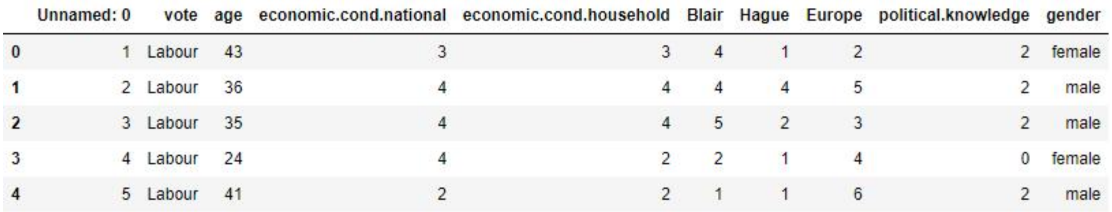
1.
List of Figures

2.
List of Figures Page No.

1.
List of Tables Page No.

1.1 Read the data set. Do the descriptive statistics and do the null value condition check. Write an inference on it.

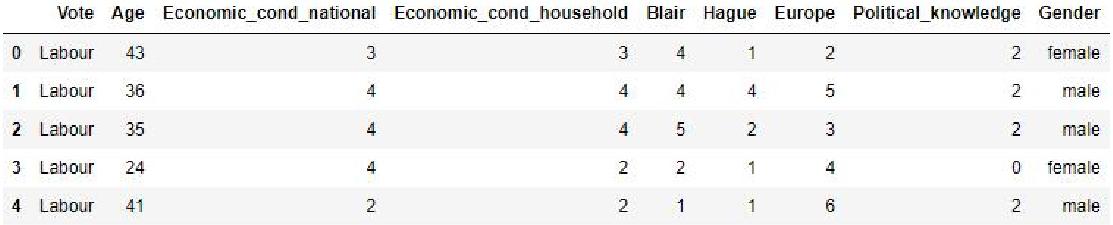
Sample of the data set-



	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

Table No.1

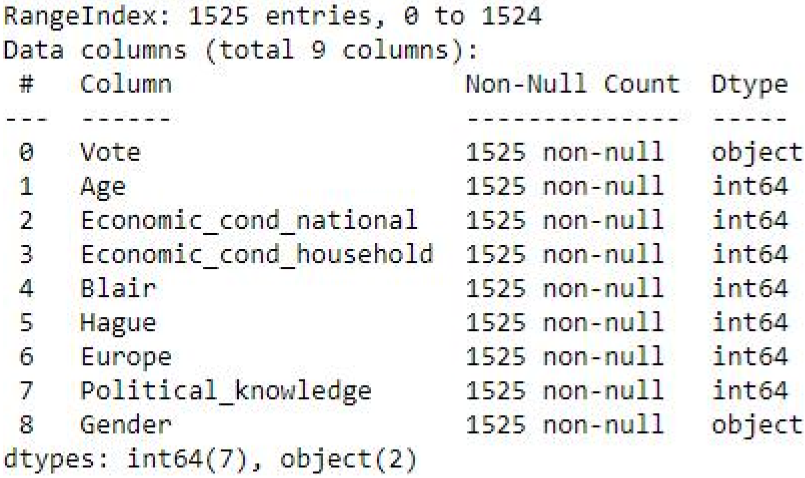
As 'Unnamed:0 ' is just an index set, we remove this column.Also as per convention we rename the columns.



	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

Table No.2

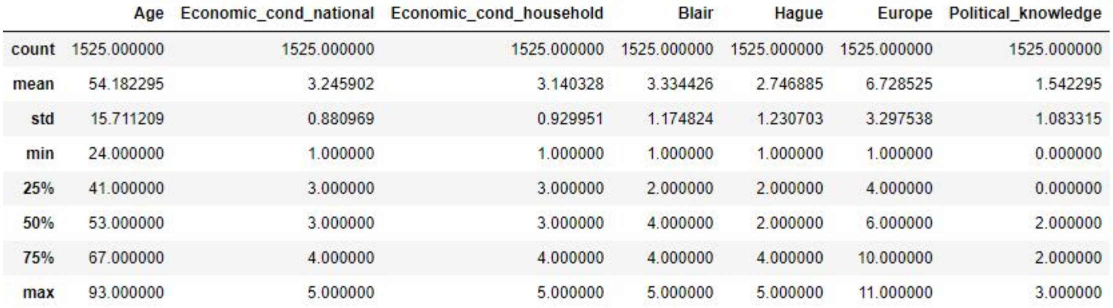
Data information-



```
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)
```

There are 1525 observations and 9 feature columns.Out of 9 columns , 'Vote' & 'Age' are categorical and rest are numerical data types.

Five point summary(Numerical data)-

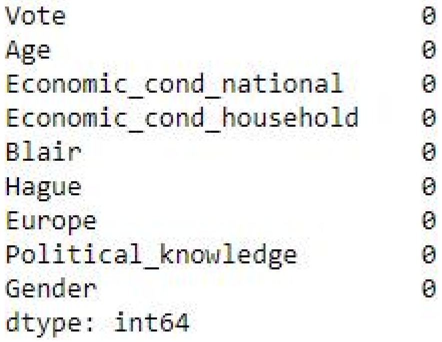


	Age	Economic_cond_national	Economic_cond_household	Blair	Hague	Europe	Political_knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
mean	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538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

Table No.3

The data distribution of ‘Age’ is almost symmetrical.It will be more evident on later discussions.

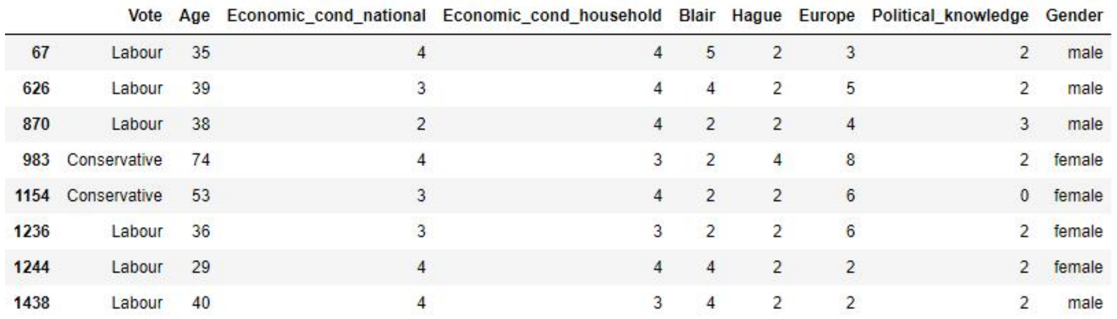
Checking missing values-



```
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
```

There is no missing value in the data set.

Duplicate rows-

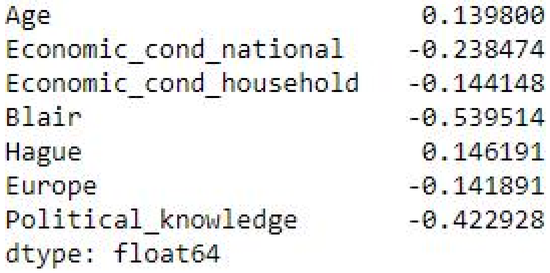


	Vote	Age	Economic_cond_national	Economic_cond_household	Blair	Hague	Europe	Political_knowledge	Gender
67	Labour	35	4	4	5	2	3	2	male
626	Labour	39	3	4	4	2	5	2	male
870	Labour	38	2	4	2	2	4	3	male
983	Conservative	74	4	3	2	4	8	2	female
1154	Conservative	53	3	4	2	2	6	0	female
1236	Labour	36	3	3	2	2	6	2	female
1244	Labour	29	4	4	4	2	2	2	female
1438	Labour	40	4	3	4	2	2	2	male

Table No.4

There are 8 duplicate rows.These duplicates need to be dropped because they do not add any value to the study, be it associated with different people.

Skew values-



```
Age                         0.139800
Economic_cond_national     -0.238474
Economic_cond_household    -0.144148
Blair                      -0.539514
Hague                       0.146191
Europe                     -0.141891
Political_knowledge        -0.422928
dtype: float64
```

Skewness is a measure of asymmetry of the probability distribution of the data.Here 2 variables are positively skewed and rest negatively skewed.Blair count has maximum skewness.

1.2 Perform Uni-variate and Bi-variate Analysis. Do exploratory data analysis. Check for Outliers.

Uni-variate analysis(Numerical variables)-

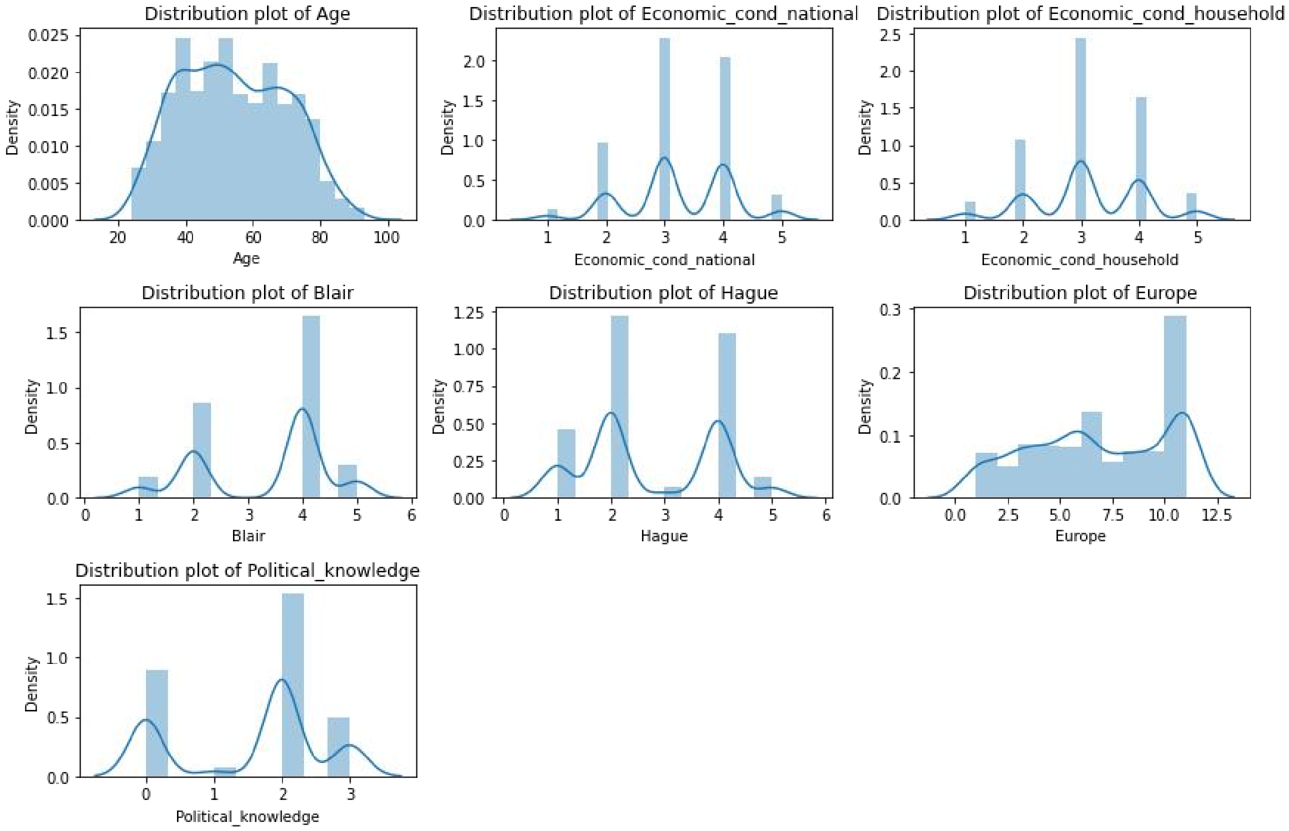


Fig. 1

'Age' is almost normally distributed.Rest of the variables are not that well spread as they are discrete in nature.

Uni-variate analysis(Categorical variables)-

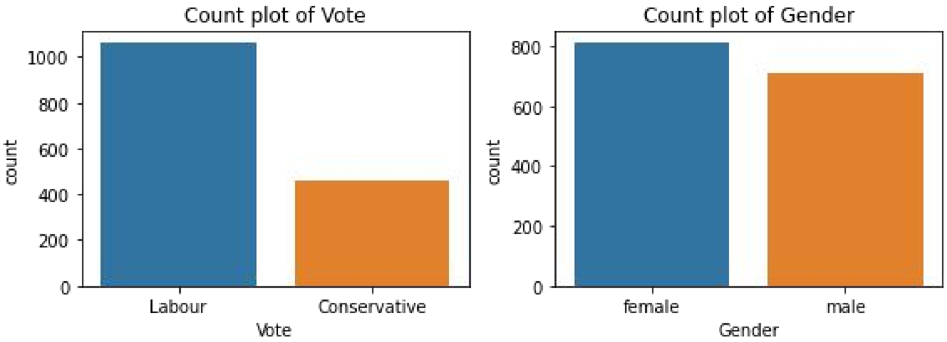


Fig. 2

The target variable consists of 70% Labour party and 30% Conservative party.The data is imbalanced here.

Bi-variate analysis-

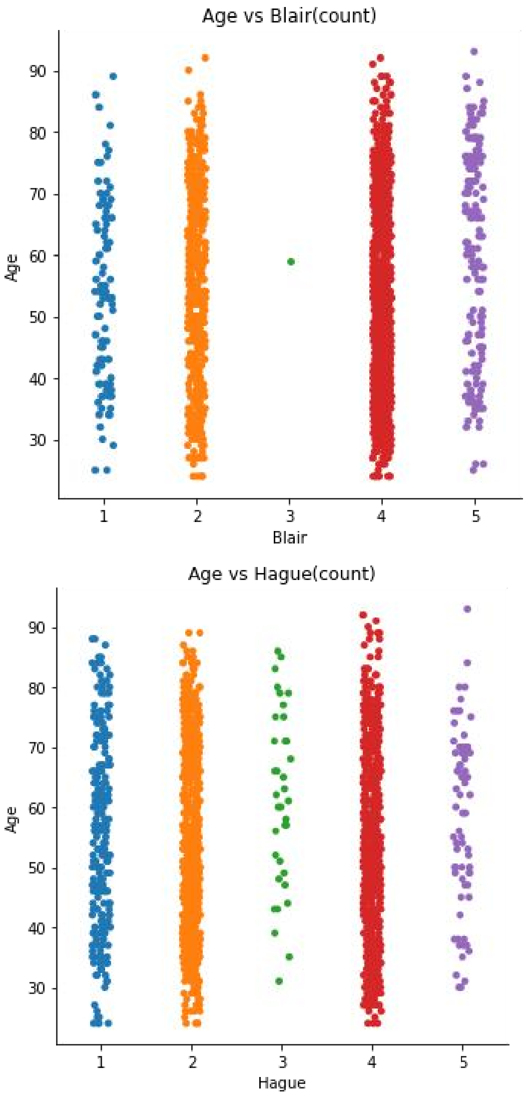
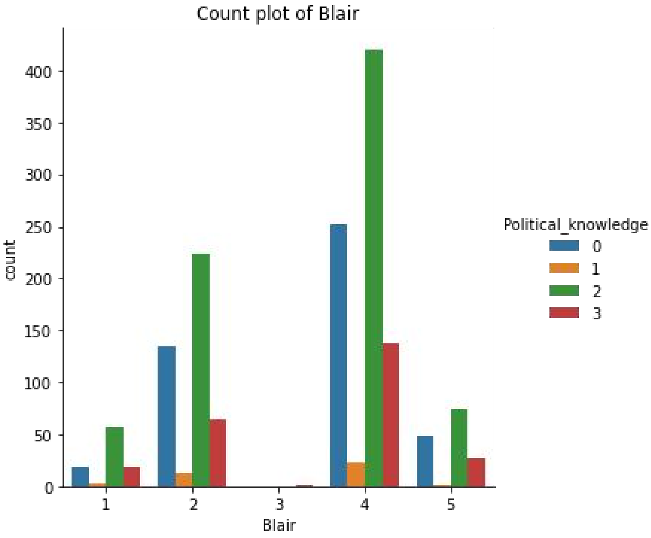


Fig. 3

People over 40 gives Blair good point than Hague.



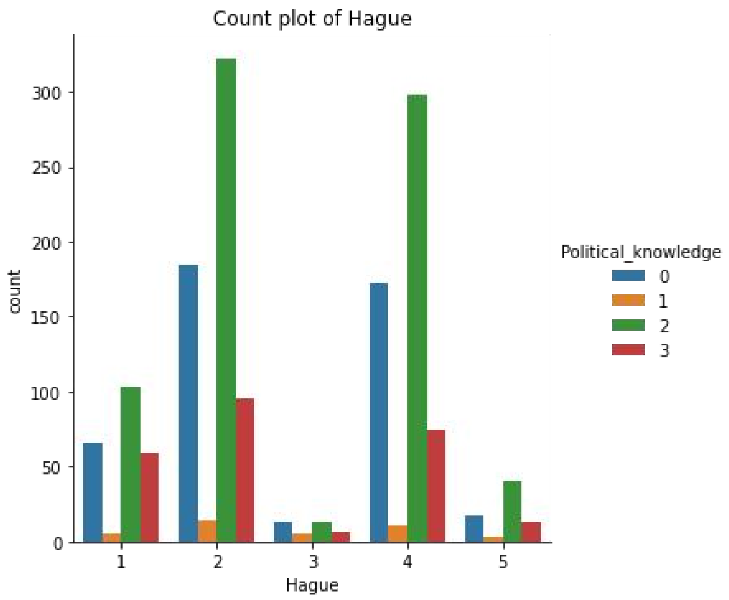


Fig. 4

People with better political knowledge gives Blair better rank than Hague.

Pair plot of Numerical variables-

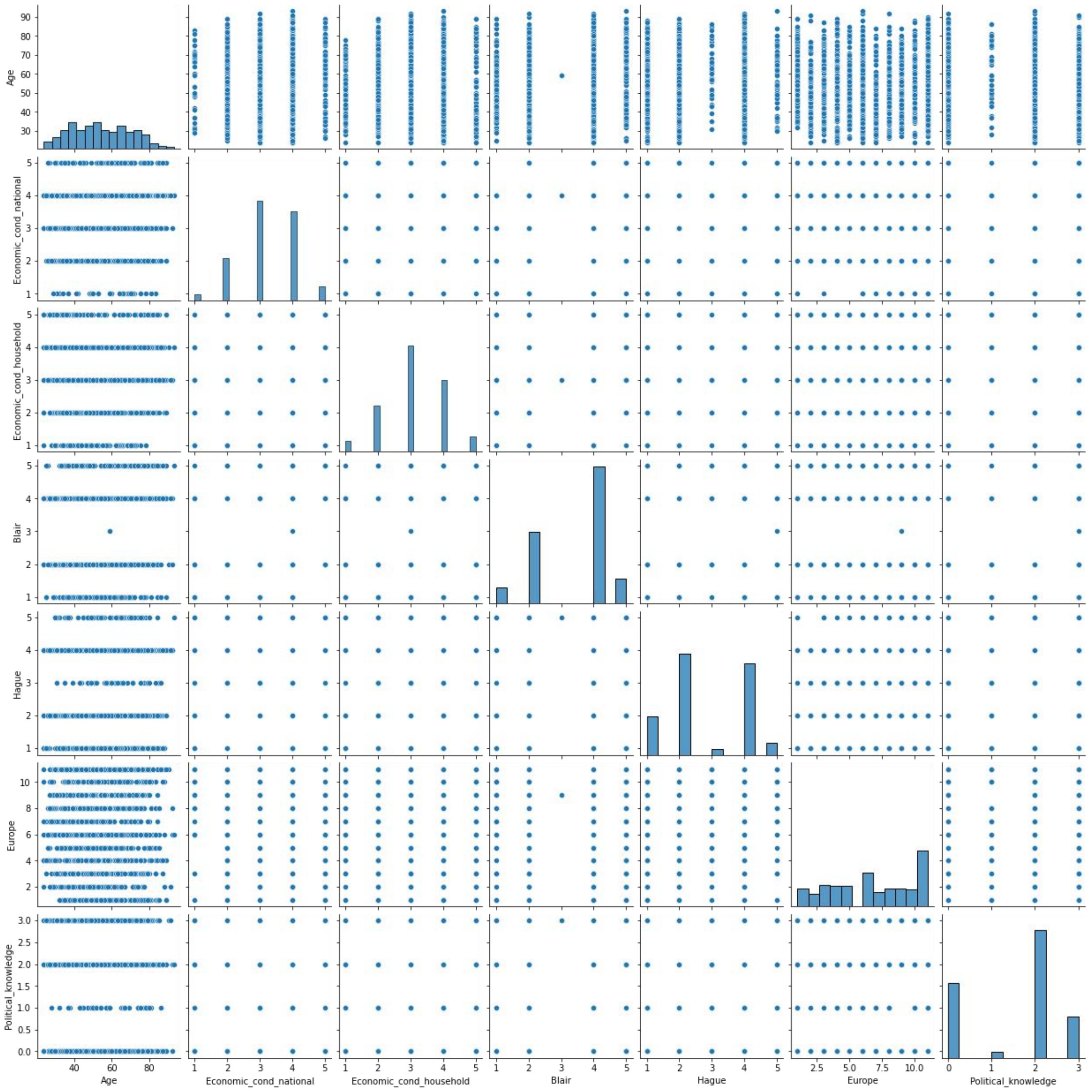


Fig. 5

We observe from the above Pair plot ,that there is not much relation between the variables.Thus we don’t have to deal with multicollinearity.Though it is a rough estimate,it will be more evident from heat map.

Heat map(Numerical variables)-

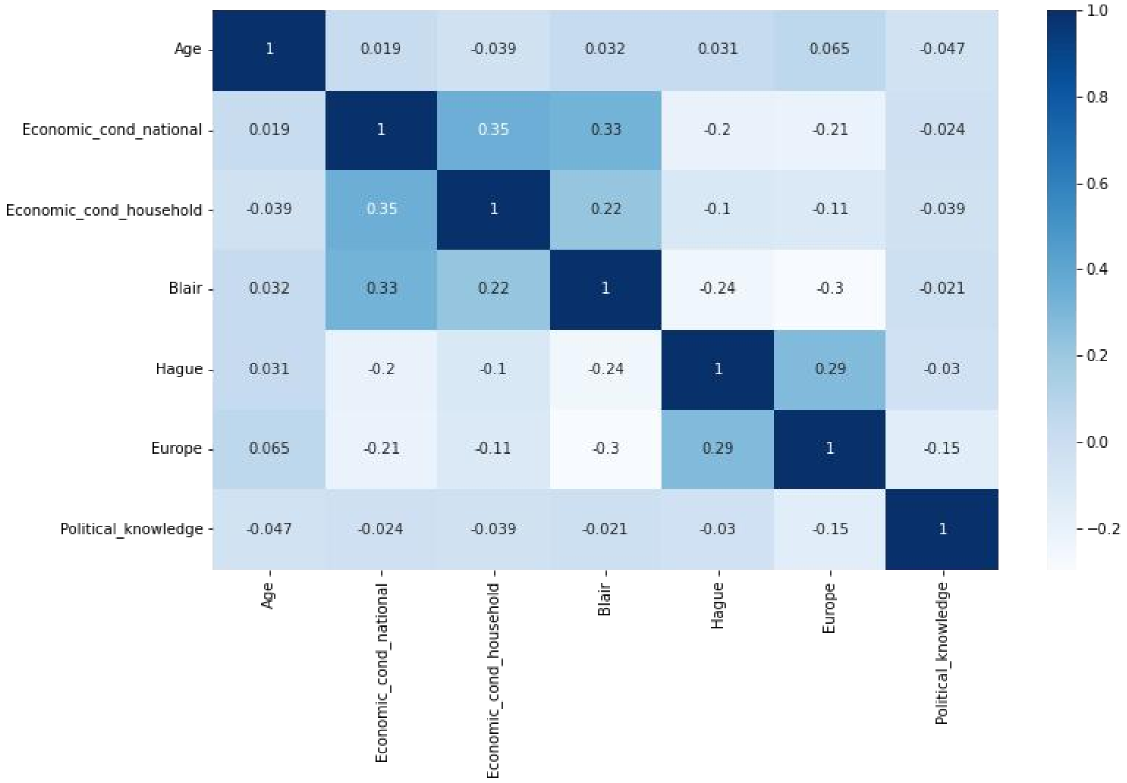


Fig. 6

It is very much evident that there is no strong relationship between the variables.The maximum positive correlation being 0.35 between Economic_cond_national and household.The maximum negative correlation being -0.3 between Blair and Europe.

Checking for Outliers-

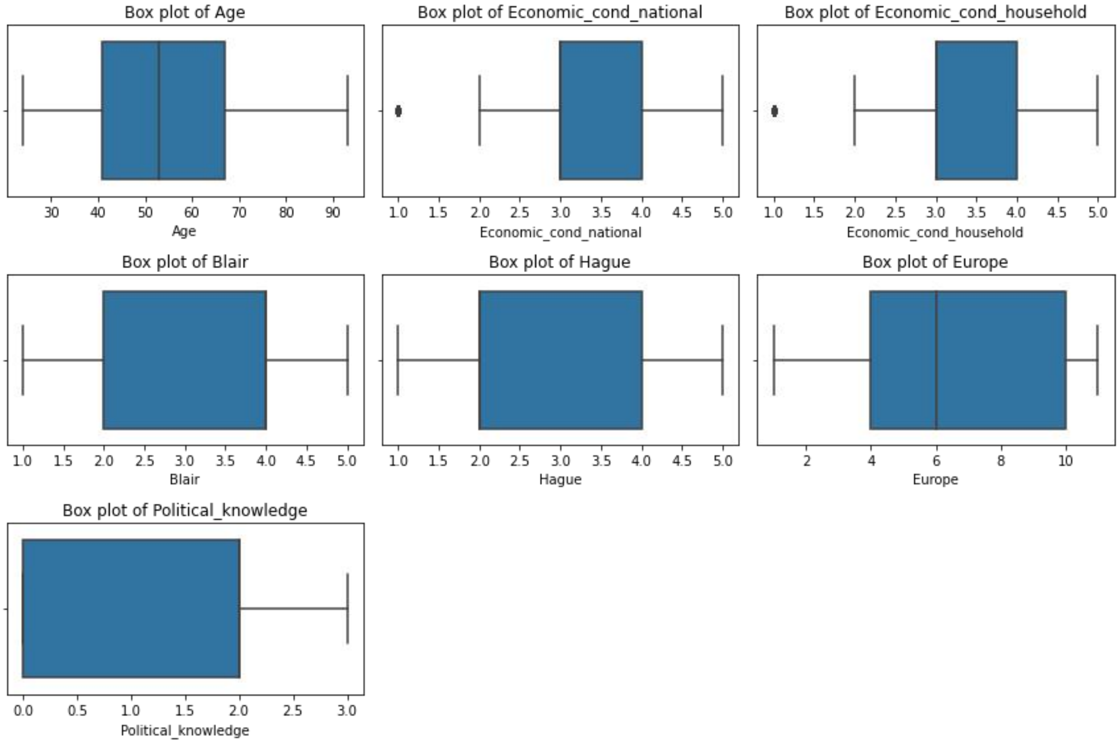
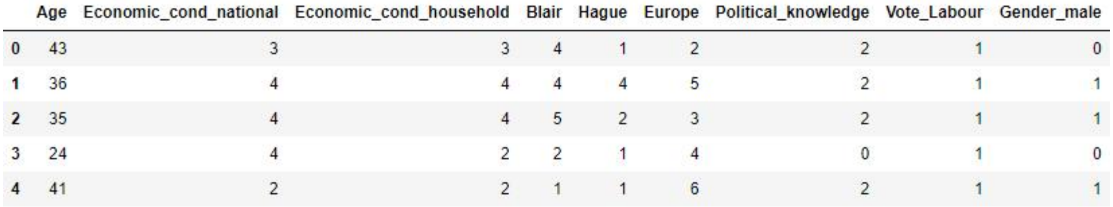


Fig. 7

There is very few outliers for Economic_con_national and household only.Outliers are to be treated for only continuous columns analyses.So we move on without treating the outliers.

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).

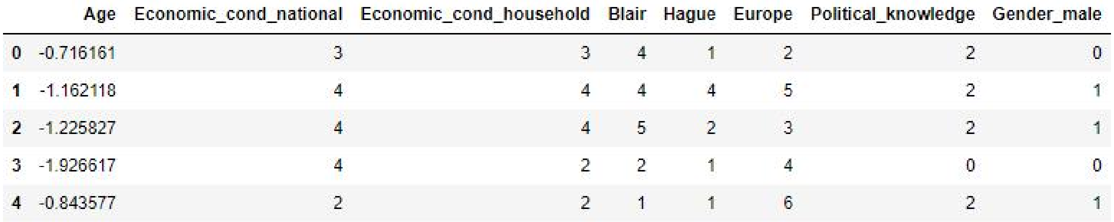
Encoding-
As Machine Learning models can not take string values,we have to encode the Categorical variables.Here are two Categorical variables 'Vote' & 'Gender'.



	Age	Economic_cond_national	Economic_cond_household	Blair	Hague	Europe	Political_knowledge	Vote_Labour	Gender_male
0	43	3	3	4	1	2	2	1	0
1	36	4	4	4	4	5	2	1	1
2	35	4	4	5	2	3	2	1	1
3	24	4	2	2	1	4	0	1	0
4	41	2	2	1	1	6	2	1	1

Here ,Vote_Labour=1 means Labour Party and Vote_labour=0 means Conservative Party.For Gender_male=1 means male Gender_male=0 means female.

Scaling-
Scaling is necessary as range of values for 'Age' lies between 24 and 93 while for other columns range is maximum 10.Also variance for 'Age' is very high in comparison to other variables.So we need to scale the data.We scale the 'Age ' variable only here.



	Age	Economic_cond_national	Economic_cond_household	Blair	Hague	Europe	Political_knowledge	Gender_male
0	-0.716161	3	3	4	1	2	2	0
1	-1.162118	4	4	4	4	5	2	1
2	-1.225827	4	4	5	2	3	2	1
3	-1.926617	4	2	2	1	4	0	0
4	-0.843577	2	2	1	1	6	2	1

Data split-
We split the data into 70:30 ratio.There are 1061 observations in the train data and 456 observations in the test data.

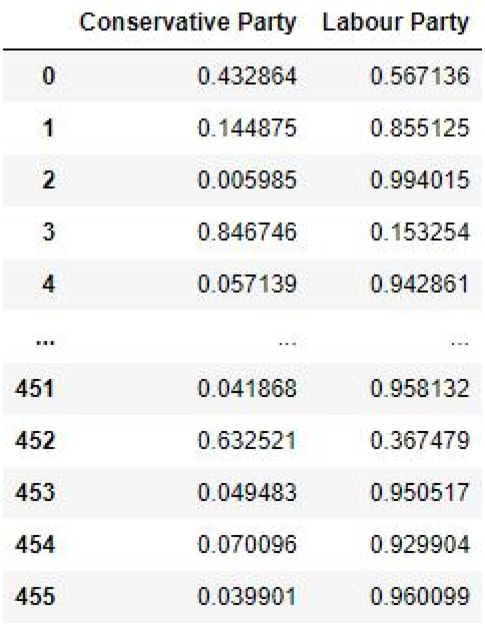
1.4 Apply Logistic Regression and LDA (linear discriminant analysis). 1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

Logistic Regression-

We are using default values for the hyper parameters to fit the model on the train data.The test accuracy and train accuracy are almost same and they are pretty good.So the model is a good fit.

Test Accuracy-0.83
Train Accuracy-0.84

Probabilities(on Test Data)-



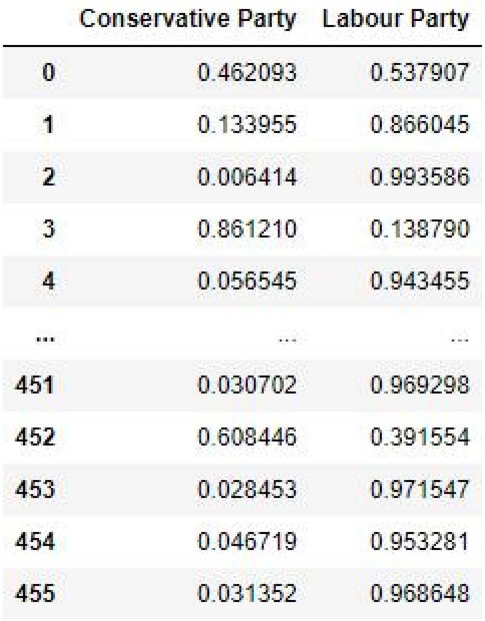
	Conservative Party	Labour Party
0	0.432864	0.567136
1	0.144875	0.855125
2	0.005985	0.994015
3	0.846746	0.153254
4	0.057139	0.942861
...	...	...
451	0.041868	0.958132
452	0.632521	0.367479
453	0.049483	0.950517
454	0.070096	0.929904
455	0.039901	0.960099

Linear Discriminant Analysis-

We are using default values for the hyper parameters to fit the model on the train data.The test accuracy and train accuracy are almost same and they are pretty good.So the model is a good fit.

Train Accuracy-0.83
Test Accuracy-0.83

Probability (on Test Data)-



	Conservative Party	Labour Party
0	0.462093	0.537907
1	0.133955	0.866045
2	0.006414	0.993586
3	0.861210	0.138790
4	0.056545	0.943455
...	...	...
451	0.030702	0.969298
452	0.608446	0.391554
453	0.028453	0.971547
454	0.046719	0.953281
455	0.031352	0.968648

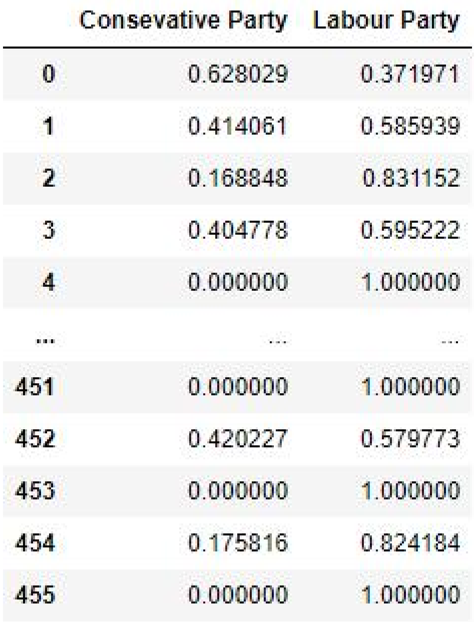
Inference-Both the model performs almost the same.Logistic Regression is slightly better.

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

KNN model-
We are using default hyper parameters except ‘weight=distance’ to fit the model on the train data.Both the test accuracy and train accuracy are pretty good.So the model is a good fit/

Train Accuracy-1.00
Test Accuracy-0.82

Probabilities(on Test Data)-

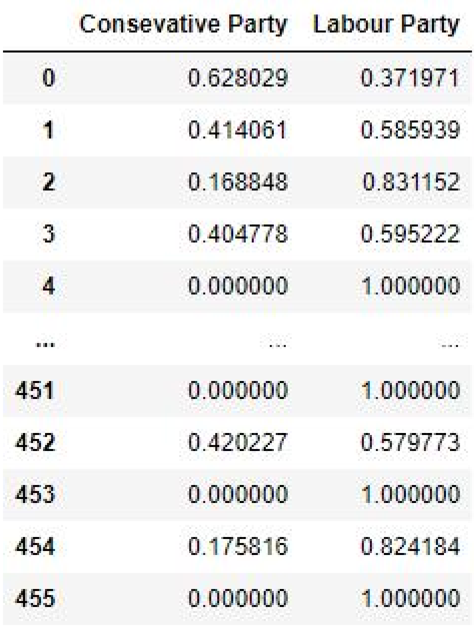


	Consevative Party	Labour Party
0	0.628029	0.371971
1	0.414061	0.585939
2	0.168848	0.831152
3	0.404778	0.595222
4	0.000000	1.000000
...	...	...
451	0.000000	1.000000
452	0.420227	0.579773
453	0.000000	1.000000
454	0.175816	0.824184
455	0.000000	1.000000

Naive Bayes Model-
We are using default values for the hyper parameters to fit the model on the train data.The test accuracy and train accuracy are almost same and they are pretty good.So the model is a good fit.

Train Accuracy-0.84
Test Accuracy-0.82

Probability (on Test Data)-



	Consevative Party	Labour Party
0	0.628029	0.371971
1	0.414061	0.585939
2	0.168848	0.831152
3	0.404778	0.595222
4	0.000000	1.000000
...	...	...
451	0.000000	1.000000
452	0.420227	0.579773
453	0.000000	1.000000
454	0.175816	0.824184
455	0.000000	1.000000

Inference-Both the model performs well.But KNN model have slightly more accuracy than Naive Bayes model.

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

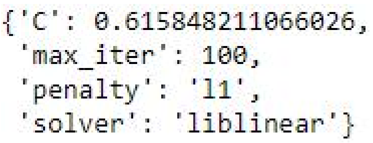
Model Tuning-
We improve the models by adjusting different hyper parameters.This procedure is called model tuning.Such a method is GridSearchCV.

1. Logistic Regression-

Parameters-

a. max_iter- maximum number of iterations taken for the solvers to converge.
b. penalty-regularization parameter.
c. solver-There are different kind of solvers:{ liblinear , ibfgs , newton-cg , saga , sag }
d. C-regularization parameter.

Best Parameters(After performing GridSearchCV)-



```
{'C': 0.615848211066026,
 'max_iter': 100,
 'penalty': 'l1',
 'solver': 'liblinear'}
```

Accuracy-
Train data=0.83
Test data=0.83

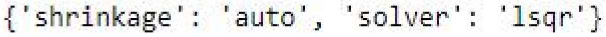
Inference-The model performs a bit better after model tuning.

2. Linear Discriminant Analysis-

Parameters-

a. shrinkage-regularization parameter
b. Solver-{ lsqr , svd , eigen }

Best Parameters(After GridSearchCV)-



```
{'shrinkage': 'auto', 'solver': 'lsqr'}
```

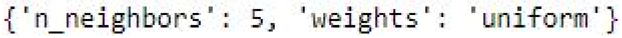
Accuracy-
Train data=0.83
Test data=0.84

Inference-The model improves slightly by tuning.

3. KNN model-

Parameters-
a. n_neighbors- number of nearest neighbors.
b. weights-
c. algorithm-

Best Parameters(After GridSearchCV)-



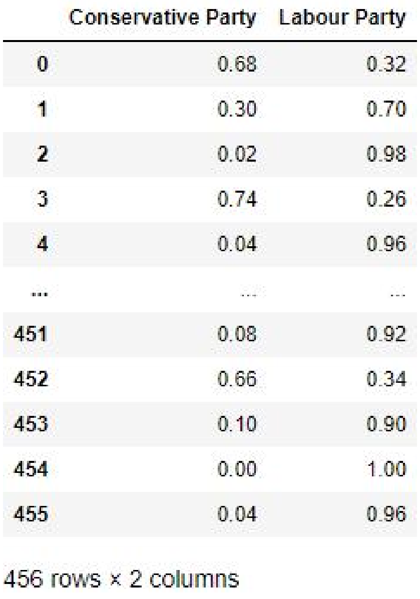
```
{'n_neighbors': 5, 'weights': 'uniform'}
```

Accuracy-
Train data=0.85
Test data=0.82

Inference-Accuracy on train data increases while on test data it remains same.

4. Naive Bayes model-
There are no specific parameters for this model.

5.Bagging-
It is an ensemble technique.Here we use Random Forest Classifier for bagging.
Parameters-
n_estimators=50(number of trees Random Forest contains)
max_features=3(square root of number of independent variables)
Accuracy-
Train Data=1.00
Test Data=0.84
Probability(on Test Data)-

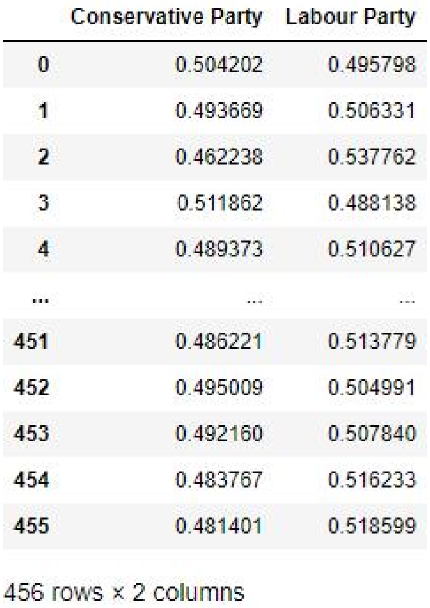


	Conservative Party	Labour Party
0	0.68	0.32
1	0.30	0.70
2	0.02	0.98
3	0.74	0.26
4	0.04	0.96
...	...	...
451	0.08	0.92
452	0.66	0.34
453	0.10	0.90
454	0.00	1.00
455	0.04	0.96

456 rows × 2 columns

Inference-The ensemble model works accurately on train data.On test data accuracy is 0.84.The model is not over fitted.

6.Boosting-
1.Ada Boosting-
It is used to boost the performance of any machine learning algorithm.It is best used with weak learners.These are models that achieve accuracy just above random chance on a classification problem.
Parameters-
n_estimators=50
Accuracy-
Train Data=0.85
Test Data=0.82
Probability(on Test Data)-

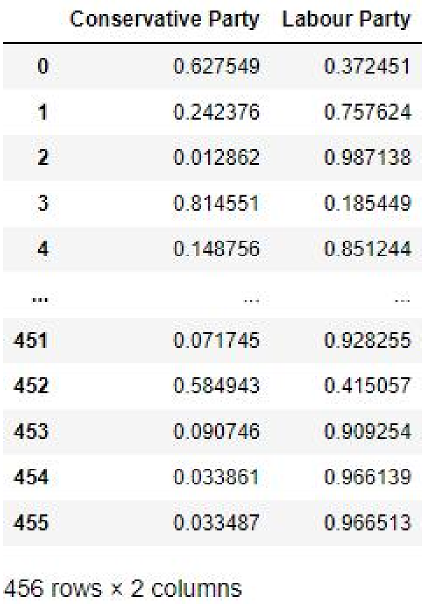


	Conservative Party	Labour Party
0	0.504202	0.495798
1	0.493669	0.506331
2	0.462238	0.537762
3	0.511862	0.488138
4	0.489373	0.510627
...	...	...
451	0.486221	0.513779
452	0.495009	0.504991
453	0.492160	0.507840
454	0.483767	0.516233
455	0.481401	0.518599

456 rows × 2 columns

Inference-
The accuracy for test and train data are significantly close.The model is not over fitted.

2.Gradient Boosting-
It is a machine learning technique used in regression and classification tasks,among others.It gives a prediction model in the form of an ensemble of weak prediction models,which are typically decision trees.
Parameters-
n_estimators=50
Accuracy-
Train Data=0.88
Test Data=0.83
Probability(on Test Data)-



	Conservative Party	Labour Party
0	0.627549	0.372451
1	0.242376	0.757624
2	0.012862	0.987138
3	0.814551	0.185449
4	0.148756	0.851244
...	...	...
451	0.071745	0.928255
452	0.584943	0.415057
453	0.090746	0.909254
454	0.033861	0.966139
455	0.033487	0.966513

456 rows × 2 columns

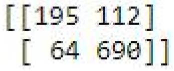
Inference-
The model is not over fitted.Accuracy score for test and train data are significantly close.

1.7 Performance Metrics:

To evaluate the performance or quality of the model,different metrics are used,and these metrics are known as performance metrics or evaluation metrics.
Performance metrics for Classification-

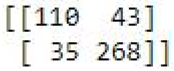
1. Accuracy
2. Confusion Matrix
3. Precision
4. Recall
5. F-Score
6. AUC(area under the curve)-ROC

1.Logistic Regression-
Accuracy-(already mentioned)
Confusion Matrix-
Train Data-



```
[[195 112]
 [ 64 690]]
```

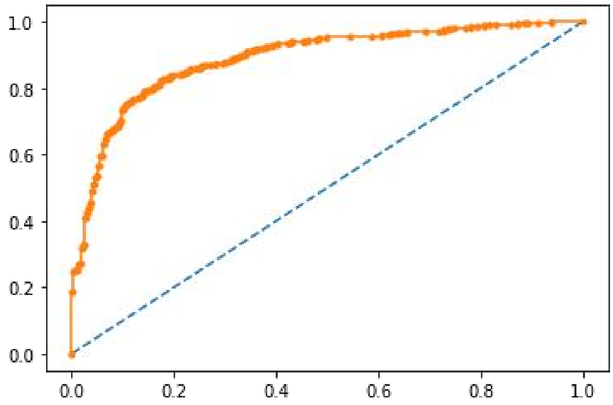
Test Data-



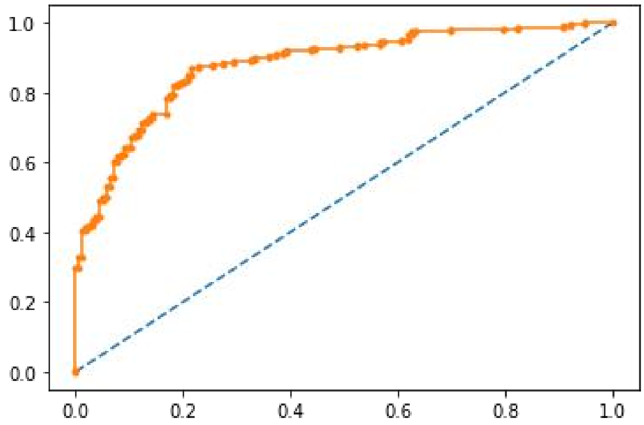
```
[[110  43]
 [ 35 268]]
```

Precision-
Train Data-
Labour Party(class-1)=0.86
Conservative Party(class-2)=0.75
Test Data-
Labour Party(class-1)=0.86
Conservative Party(class-2)=0.76
Recall-
Train Data-
Labour Party(class-1)=0.94
Conservative Party(class-2)=0.64
Test Data-
Labour Party(class-1)=0.88
Conservative Party(class-2)=0.72

F-Score-
Train Data-
Labour Party(class-1)=0.89
Conservative Party(class-2)=0.69
Test Data-
Labour Party(class-1)=0.87
Conservative Party(class-2)=0.74
AUC-ROC-
Train Data-

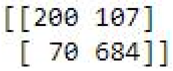


Test Data-



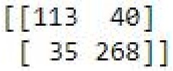
AUC-Score
Train Data=0.89
Test Data=0.87

2.LDA-
Accuracy-(already mentioned)
Confusion Matrix-
Train Data-



```
[[200 107]
 [ 70 684]]
```

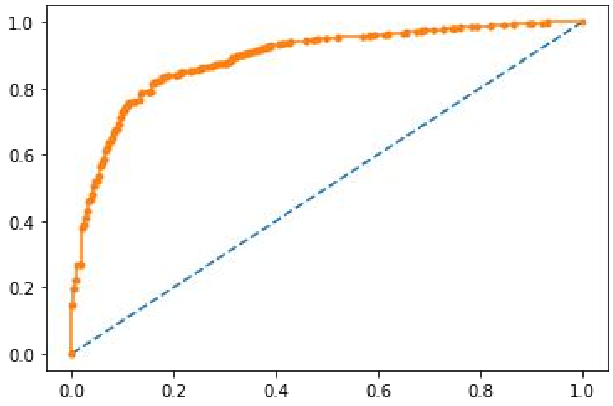
Test Data-



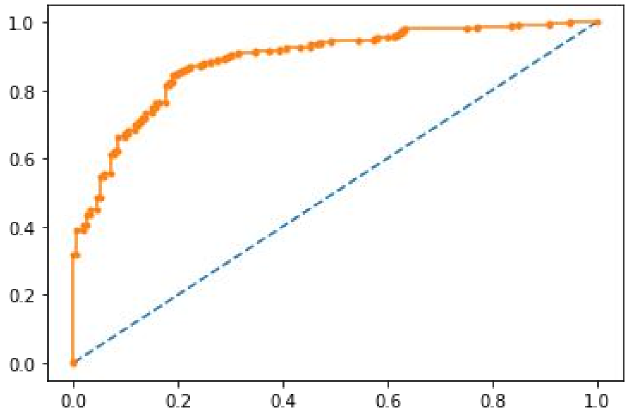
```
[[113  40]
 [ 35 268]]
```

Precision-
Train Data-
Labour Party(class-1)=0.86

Conservative Party(class-0)=0.74
Test Data-
Labour Party(class-1)=0.87
Conservative Party(class-2)=0.76
Recall-
Train Data-
Labour Party(class-1)=0.95
Conservative Party(class-2)=0.61
Test Data-
Labour Party(class-1)=0.88
Conservative Party(class-2)=0.74
F-Score-
Train Data-
Labour Party(class-1)=0.89
Conservative Party(class-2)=0.69
Test Data-
Labour Party(class-1)=0.88
Conservative Party(class-2)=0.75
AUC-ROC-
Train Data-



Test Data-



AUC-Score
Train Data=0.89
Test Data=0.8

1.8 Based on your analysis and working on the business problem,detail out appropriate insights and recommendation to help the management solve the business objective.

1. Using Logistic Regression model for predicting the outcome as it has the best performance.
2. Best features for predicting outcome in Logistic Regression are Hague,Blair,Economic_cond_national,Economic_cond_household,Political_knowledge.
3. If we can manipulate this features we can also manipulate the outcomes.
4. Gathering more data will also help our purpose.

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

Characters-
Franklin D.Roosevelt’s speech-7571
John F.Kennedy’s speech-7618
Richard Nixon’s speech-9991

Words-
Franklin D.Roosevelt’s speech-1536
John F.Kennedy’s speech-1546
Richard Nixon’s speech-2028

Sentences-
Franklin D.Roosevelt’s speech-68
John F.Kennedy’s speech-52
Richard Nixon’s speech-69

2.2 Remove all the stop words from all three speeches.

Stop words- A stop word is a commonly used word(such as ‘the’,’an’,’a’,’in’).It has no use in understanding the sentiment in Machine learning models.So we generally remove these words from the given text.
Word count-

Franklin D.Roosevelt’s speech-

Before removal of stop words-1536
After removal of stop words-657

John F.Kennedy’s speech-

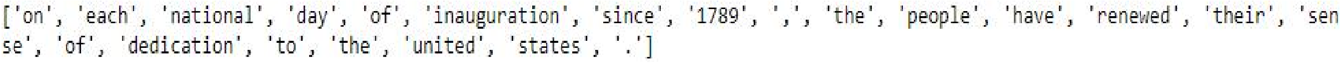
Before removal of stop words-1546
After removal of stop words-722

Richard Nixon’s speech-

Before removal of stop words-2028
After removal of stop words-853

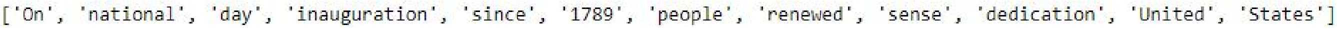
Sample sentence -

Before removal of stop words-



```
['on', 'each', 'national', 'day', 'of', 'inauguration', 'since', '1789', ',', 'the', 'people', 'have', 'renewed', 'their', 'sen
se', 'of', 'dedication', 'to', 'the', 'united', 'states', '.']
```

After removal of stop words-



```
['On', 'national', 'day', 'inauguration', 'since', '1789', 'people', 'renewed', 'sense', 'dedication', 'United', 'States']
```

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

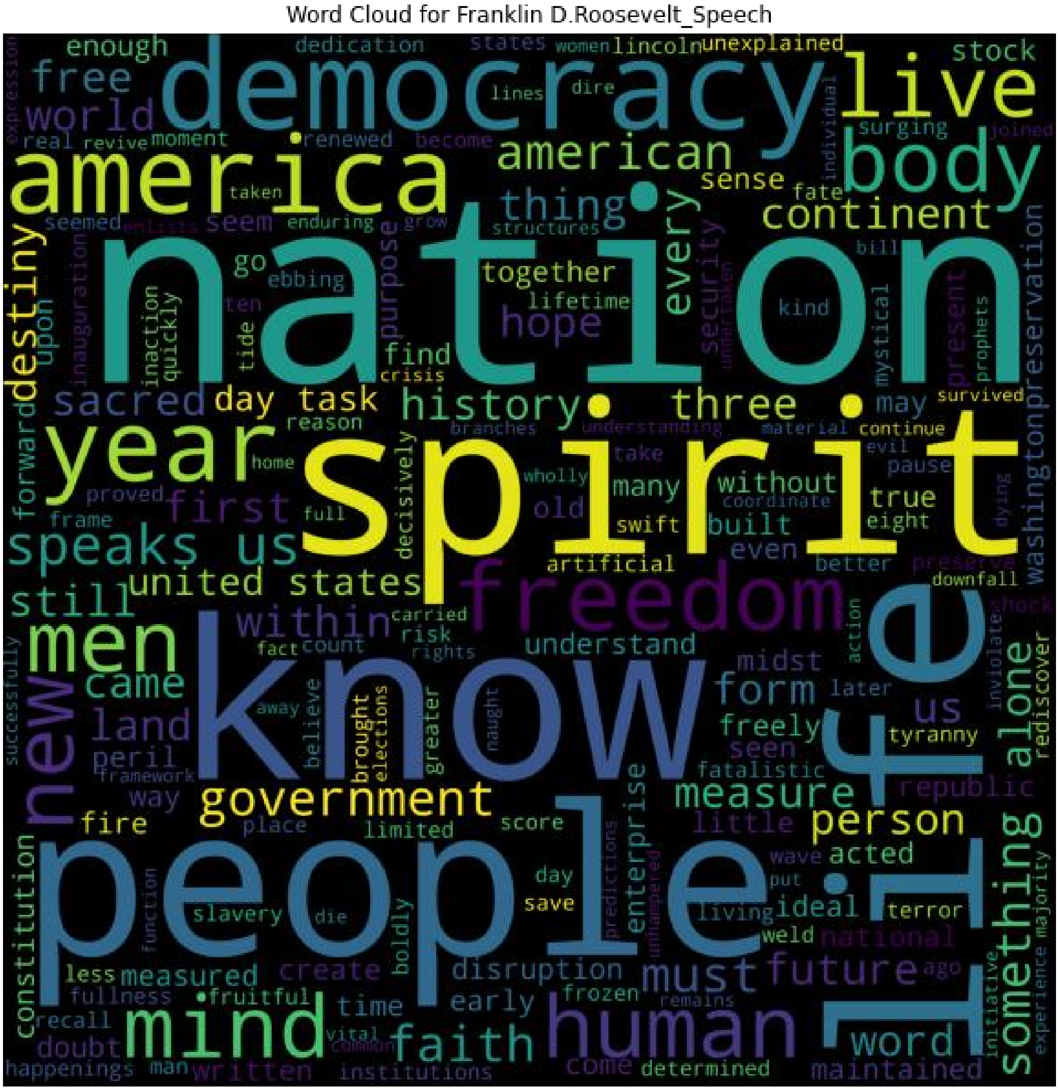
Franklin D.Roosevelt’s speech-
‘nation ’- 12 times
‘know ’- 10 times
‘spirit’,’life’ & ‘democracy ’- 9 times

John F.Kennedy’s speech-
‘let ‘-16 times
‘us ‘-12 times
‘world’, ‘sides ‘-8 times

Richard Nixon’s speech-
‘us ‘- 26 times
‘let ‘-22 times
‘america ‘-21 times

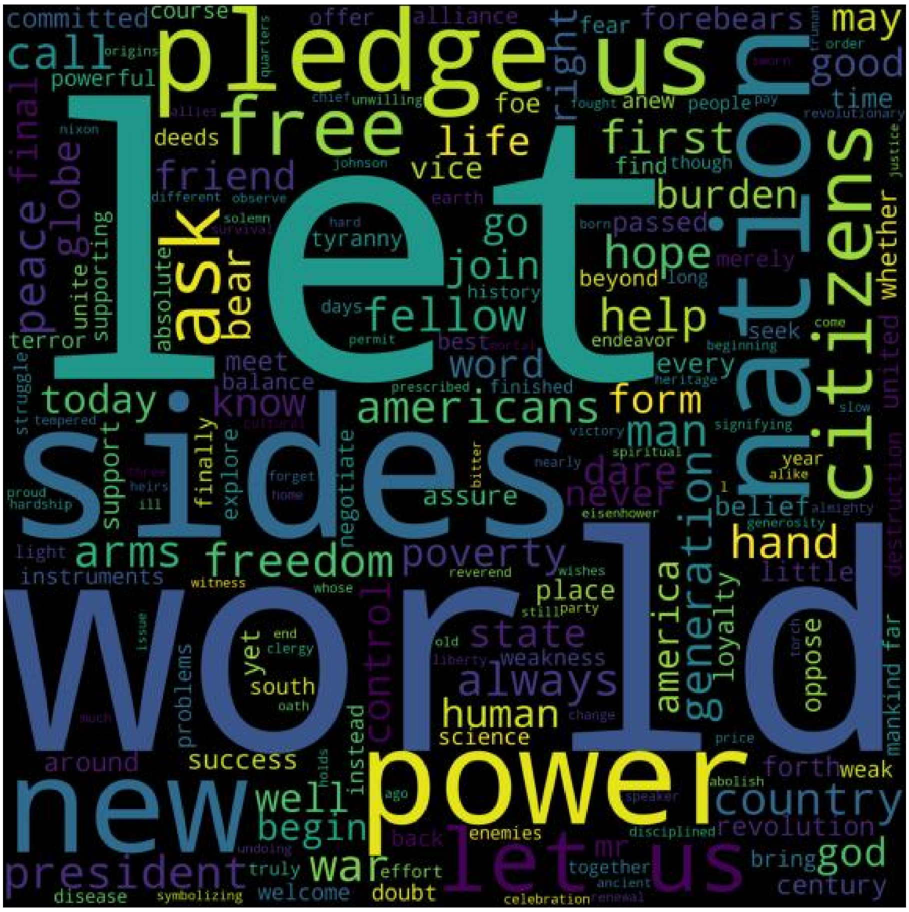
2.4 Plot the the word cloud of each of the speeches of the variable.(after removing the stop words)

A word cloud is a collection,or cluster, of words depicted in different sizes.The bigger and bolder the word appears,the more often it’s mentioned within a given text and the more important it is.


Word Cloud for Franklin D.Roosevelt_Speech

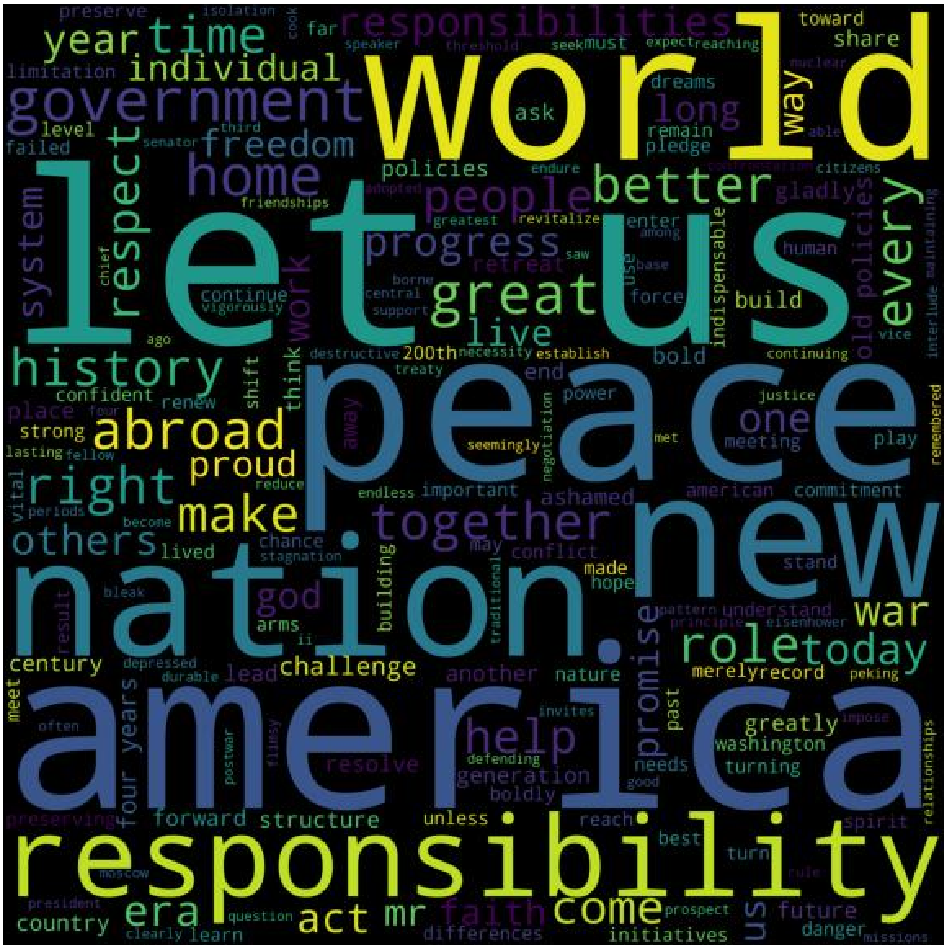
We can see the largest words are ‘nation’,’know’,’people’,’life’,’democracy’ etc which we observed as as top three words in the previous question.

Word Cloud for John F.Kennedy_Speech



We can see the largest words are ‘let’,’sides’,’world ’etc which we observed as most frequent words in the previous question.

Word Cloud for Richard Nixon_Speech



We can see the largest words are ‘us’,’let’,’america ’etc which we observed as most frequent words in the previous question.

Inference-

As we observe from word clouds of three different speeches,there are many words common among them.

the test data.