MACHINE LEARNING

BUSINESS REPORT

THAKUR ARUN SINGH

AUGUST 2021

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

CONTENTS

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.

PROBLEM 1.1

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

Resolution:

First, we import all the necessary libraries seaborn,numpy,pandas,sklearn etc to perform our analysis

Next, we import the data set “Election_Dataset_Two Classes”

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

Tail:

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1520	Conservative	67	5	3	2	4	11	3	male
1521	Conservative	73	2	2	4	4	8	2	male
1522	Labour	37	3	3	5	4	2	2	male
1523	Conservative	61	3	3	1	4	11	2	male
1524	Conservative	74	2	3	2	4	11	0	female

- By looking at the data, the first column can be removed since it is not of much significance.
- In Total, we have 9 columns which are valid to perform our analysis.
- All 8 features are Independent variable whereas 9th Feature vote is dependent variable.it is basically Target variable model wants to predict. All features except vote and gender are Numerical. Vote and gender are Nominal Categorical.

Data description:

	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

- In above description we can see mean values and 50% are almost same, mean and median are almost Coherent.
- Gender and vote seems to be Categorical Nominal variables, where order is not important aspect
- All other variables are Categorical Ordinal Variables, Ratings
- All features seems to be somewhat equally distributed around mean.

Information about the data:

```
<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
```

We can see gender and vote are of Object Datatype, we will try to convert it into integer Data type further.

Total 1525 data points are there, 1525 different people, No null value can be detected here.

Checking null 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
```

No null values present in the data set.

Checking for any duplicate records:

```
Total no of duplicate values = 8
```

We will check manually whether these are exact duplicates or partial duplicates.

	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

Shape of the data:

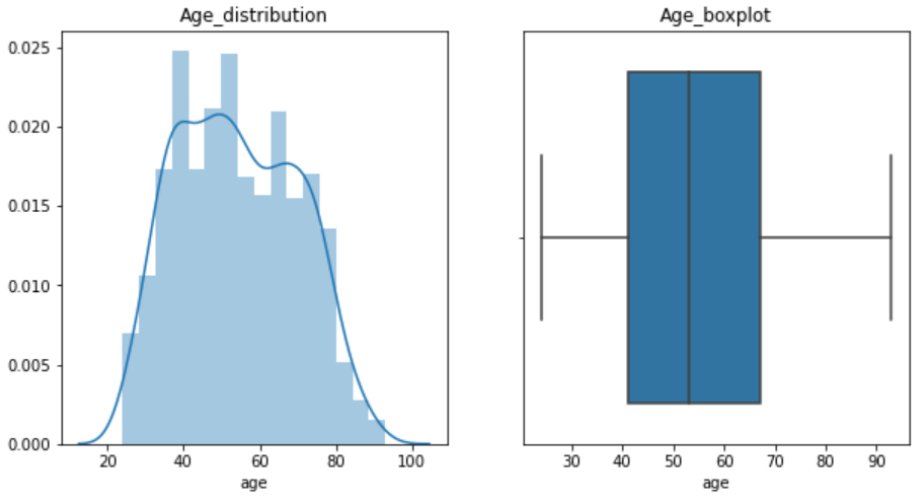
Shape of the dataset is 1525 rows and 9 Columns.

PROBLEM 1.2

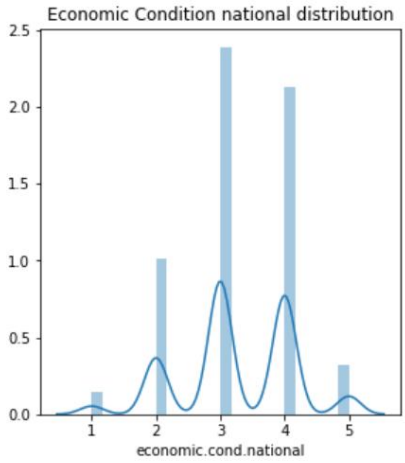
Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

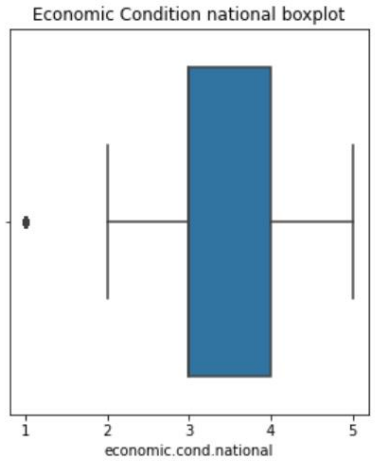
Resolution:

Univariate Analysis:

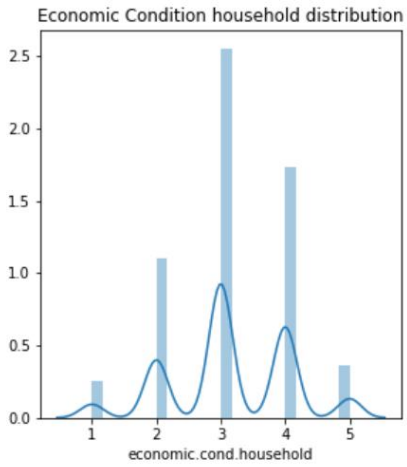


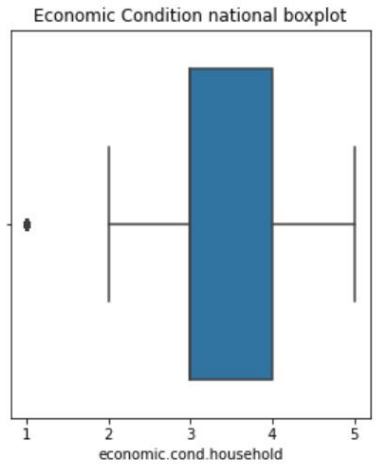
It seems that Age is normally distributed and not much skewed, all age groups are covered, it has no outliers as well.



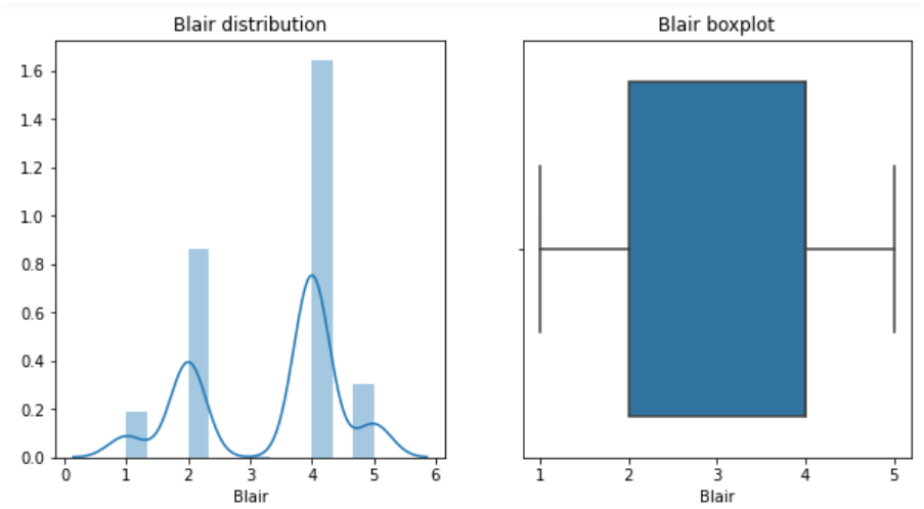


In above plot we can see spikes at almost every rating, but even this is showing most people have given 3 ratings. Very less people have given 1 and 5 rating, from this we can say that this particular nation neither have great economic condition nor poor condition.

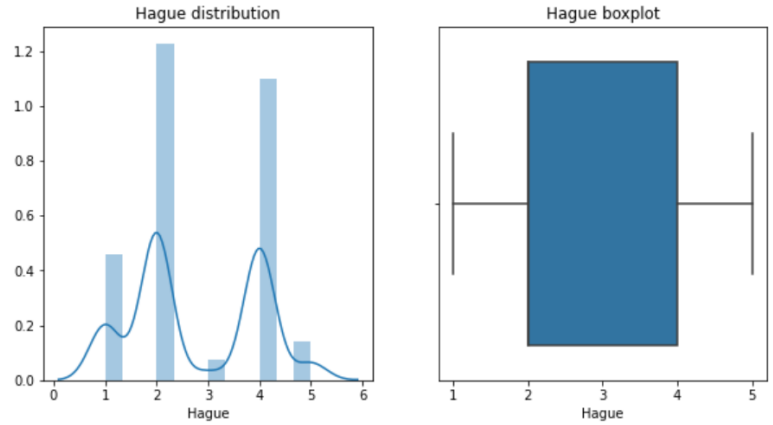




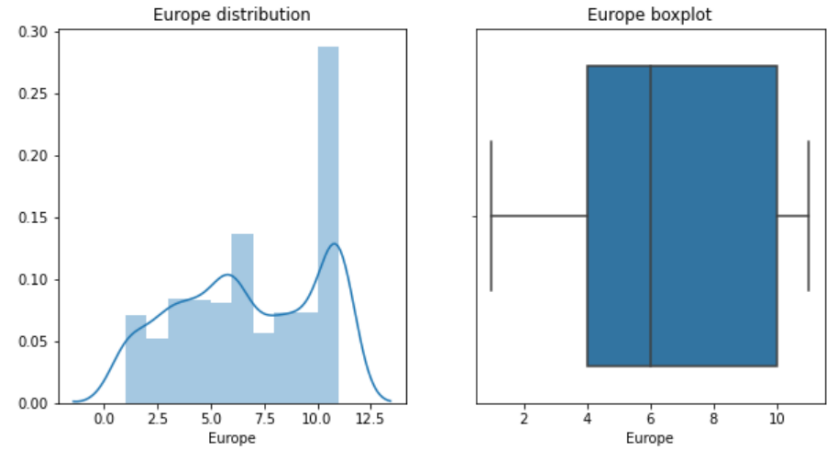
In above plot we can see spikes at almost every rating, but even this is showing most people have given 3 ratings. Very less people have given 1 and 5 rating; from this we can say that this particular nation's people neither have great economic condition nor poor condition.



Labour leader Blair has received 2 and 4 score mostly.4 is the highest frequency.

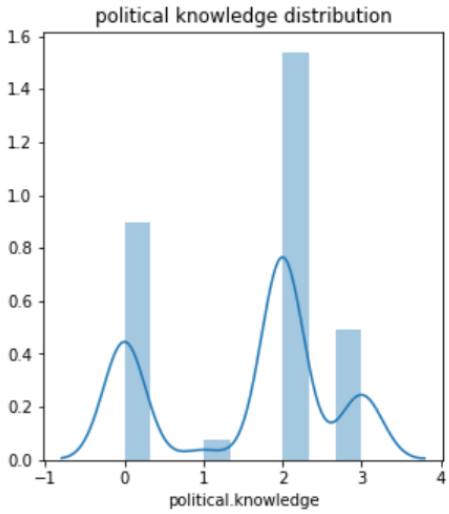


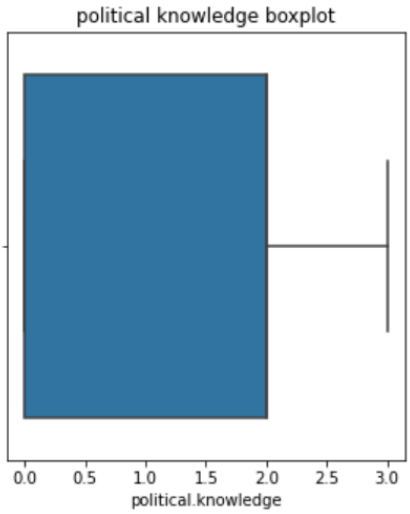
Conservative Party leader Hague has received 2 and 4 score mostly.2 is the highest frequency.



From the Europe plot, the 50% people have rated 4-10 it means they have Eurosceptic sentiment in increasing order.

10 score seems to be have Highest Frequency here; it says that most of the people are against Europe Integration.



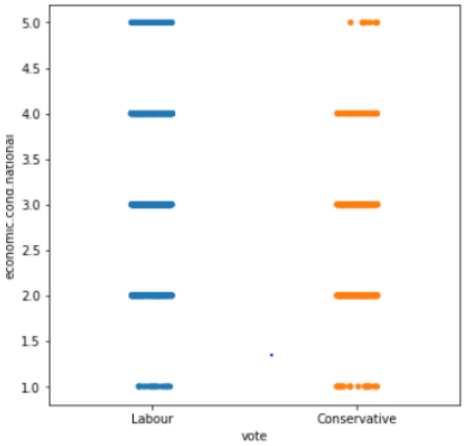


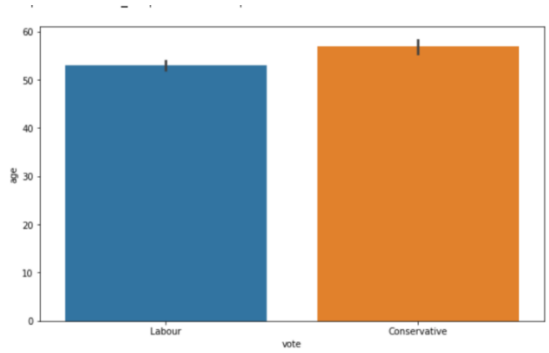
Only around 25% People have average to High political knowledge, 75% people have less Political Knowledge 0,1,2.

Outliers: Economic Household conditions and National Household conditions, Rating 1 is outlier as very less number of people has given it. But we will not treat this value as these are valid outliers, we will keep it.

Bivariate Analysis

Scatterplot of Europe vs. Age

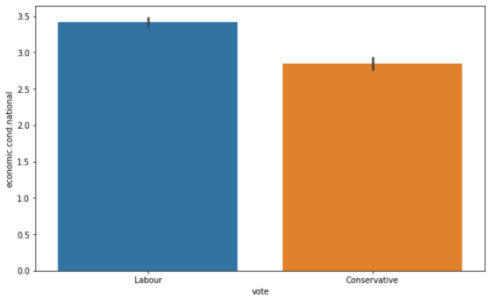




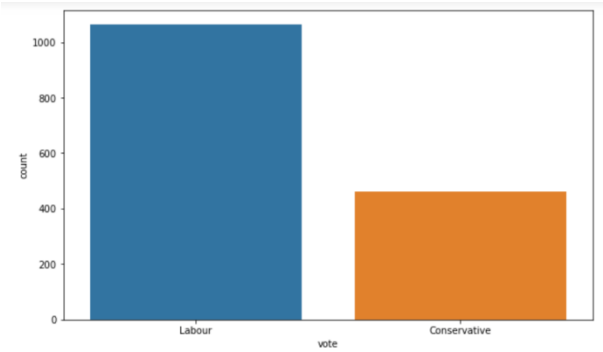
Higher Europe (Eurosceptic) score, it has converted into vote to the Conservative Party.

All age group people are voting to both the parties, so it is no more age specific, But Euro Skepticism is playing important role here

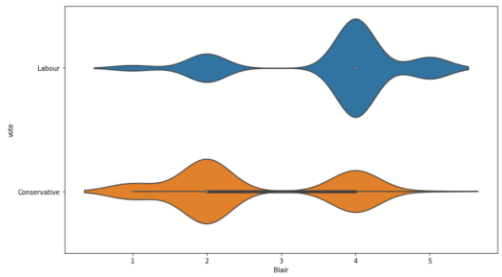
Almost all age groups are voting to Labour and Conservative Party.



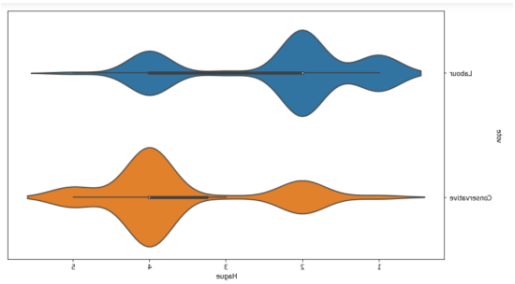
People who are giving good assessment score on National Economic condition are tend to give vote to Labour party, while lesser score turn into vote to Conservative party, this is not the case always happens.



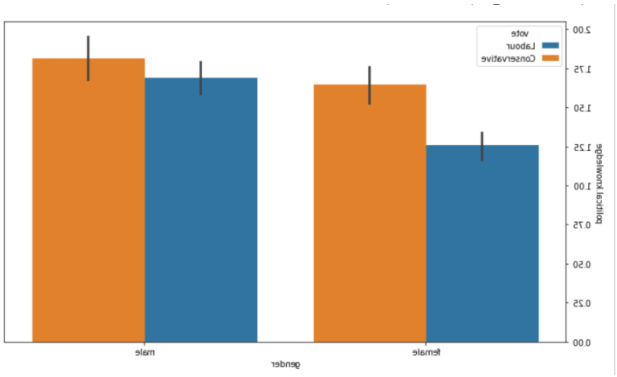
This data has more than 1000 readings who are voting to Labour party while less than 500 who are voting to conservative party it is unbalanced data, might affect to prediction of model.



If person’s rating is high to Blair as a party leader, vote will be mostly go into Labour Party's basket. while low Blair score convert the vote into Conservative Party Favor.

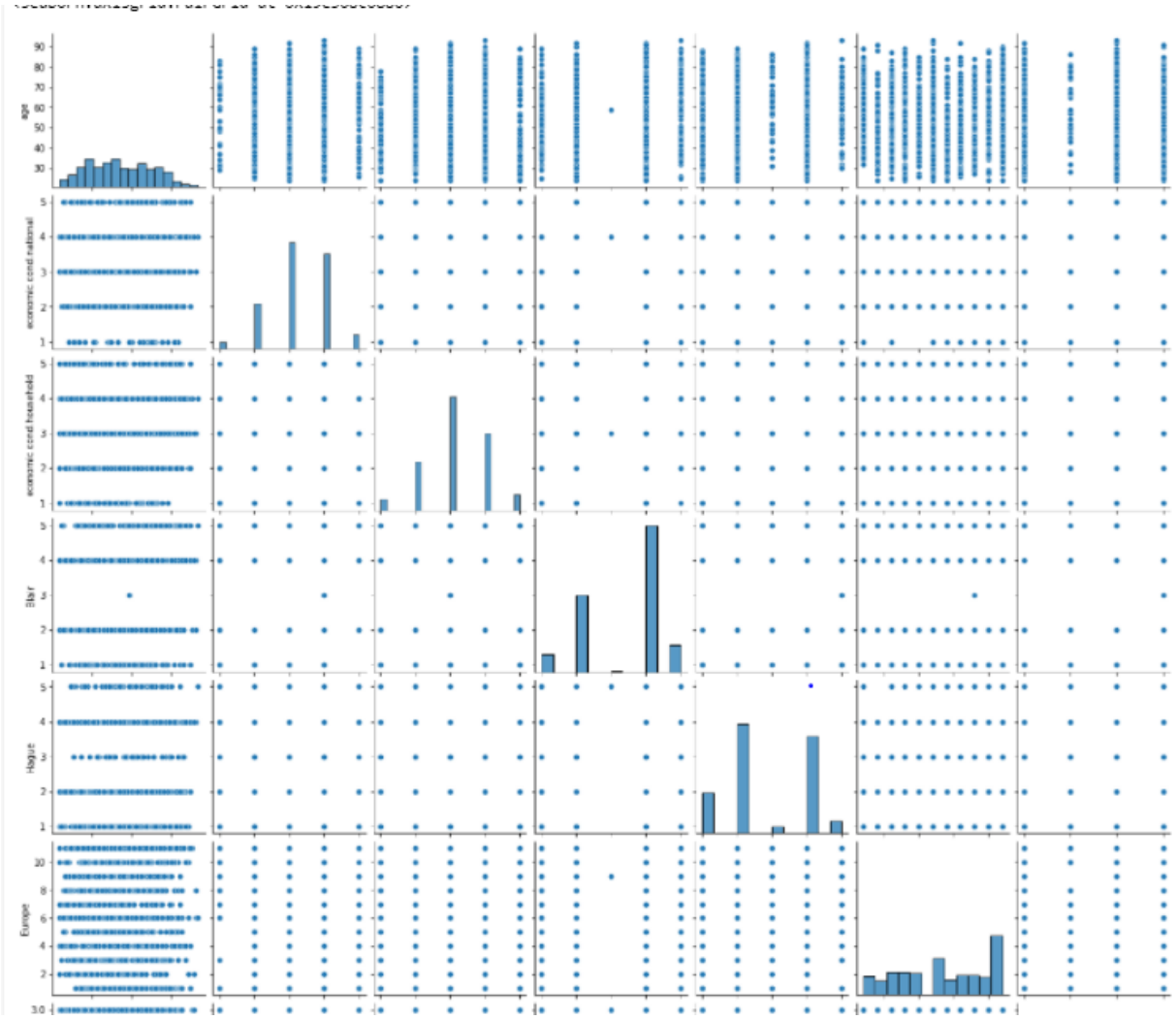


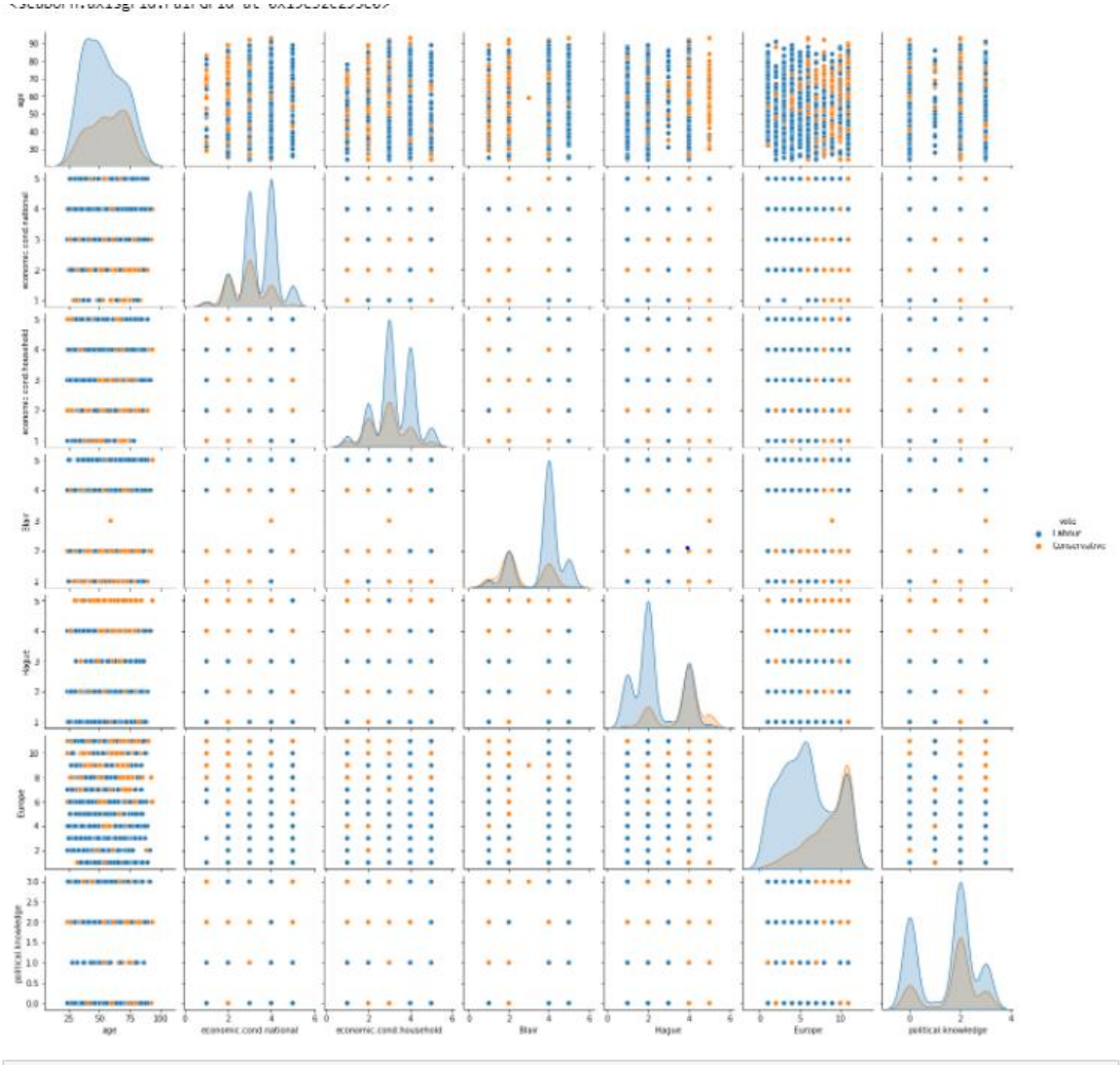
If person’s rating is high to Hague as a party leader, vote will be mostly go into Conservative Party's basket. while low Hague score convert the vote into Labour party Favor.



Political Knowledge of Males seems to be higher than average knowledge of females. In both cases higher political Knowledge people are voting to Conservative Party.

Pair plot :

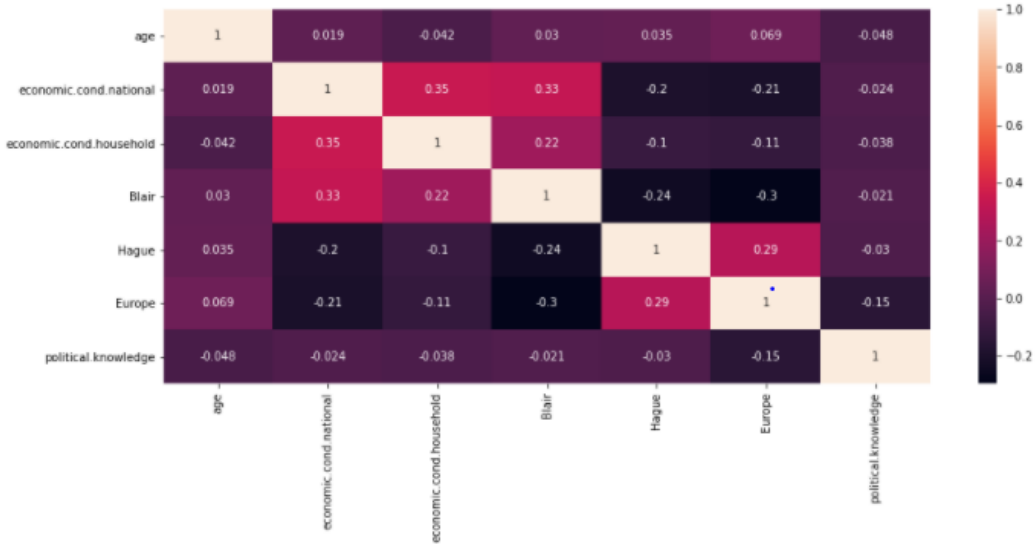

age
economic cond.national
economic cond.household
Blair
Hague
Europe



Heat map for correlation:

```
sns.heatmap(data_df.corr(),annot=True)
```

Out[31]: <AxesSubplot:>



Blair and Economic Household Condition Rating, Economic national Condition rating shows good correlation.

Europe and Blair are inversely related, means if Person is more Eurosceptic there will be less chances he will vote to Blair as a party leader of Labour party

If person is giving good assessment score to Hague he must be with Conservative party and he or she is Highly Eurosceptic.

It is not always true, but people having good political Knowledge prefer Europe Integration.

PROBLEM 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).

Resolution:

We have to encode the features vote and Gender, to feed the data to model.

Here vote and gender are Nominal Categorical variables, we can use one hot encoding to feed data to model.

We will use get dummies function to convert data into 0's and 1's.

vote_labour=1 ; Vote has been given to Labour Party.

vote_labour=0 ; Vote has been given to Conservative Party.

gender_male=0; Female

gender_male=1; male

After Scaling, the first 5 rows,

age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
43	3	3	4	1	2	2	1	0
36	4	4	4	4	5	2	1	1
35	4	4	5	2	3	2	1	1
24	4	2	2	1	4	0	1	0
41	2	2	1	1	6	2	1	1

All features are having different ranges and different scales like somewhere it is 1-11, 0-4 etc.

We will do scaling for KNN only.

Algorithm	Problem Type	Features might need scaling?
KNN	Either	Yes
Linear regression	Regression	No (unless regularized)
Logistic regression	Classification	No (unless regularized)
Naive Bayes	Classification	No
Decision trees	Either	No
Random Forests	Either	No
AdaBoost	Either	No
Neural networks	Either	Yes

Data Split

Split the data into train and test (70:30).

We will calculate VIF to see if there any issue of Multicollinearity or not.

```
age VIF =  1.03
economic.cond.national VIF =  1.28
economic.cond.household VIF =  1.16
Blair VIF =  1.34
Hague VIF =  1.32
Europe VIF =  1.28
political.knowledge VIF =  1.09
vote_Labour VIF =  1.67
gender_male VIF =  1.03
```

All VIF score is less than 4 hence no issue of Multi-collinearity.

Let’s copy all the predictor variables into X data frame. And copy target into the y data frame.

Split X and y into training and test set in 70:30 ratios.

We have created 4 variables here, X_train, X_test, y_train, y_test.

We will check the distribution percentage for target variable in train and test data.

Train target variable:

```
1 y_train.value_counts(1)
```

```
1    0.697282
0    0.302718
Name: vote_Labour, dtype: float64
```

Test Target variable:

```
1 y_test.value_counts(1)
```

```
1    0.696507
0    0.303493
Name: vote_Labour, dtype: float64
```

Almost 70-30 is the distribution of vote is achieved, as original dataset is also contains 70:30

Labour: Conservative distribution.

PROBLEM 1.4

Apply Logistic Regression and LDA (linear discriminant analysis)

Resolution:

Apply Logistic Regression:

First we will create Logistic Regression and then we will fit it on train data **X_train, y_train.**

We will use solver as 'newton-cg' with 10,000 iterations.

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg',
                   verbose=True)
```

Predicting on Training and Test dataset

```
ytrain_predict = model.predict(X_train)
ytest_predict = model.predict(X_test)
```

Getting the probabilities:

```
ytest_predict_prob=model.predict_proba(X_test)
pd.DataFrame(ytest_predict_prob).head()
```

First 5 rows of Predicted probabilities.

	0	1
0	0.616214	0.383786
1	0.186461	0.813539
2	0.187993	0.812007
3	0.163937	0.836063
4	0.052483	0.947517

Apply Linear Discriminant Analysis:

Build LDA model

We will create LDA classifier and then we will fit it on training data *X_train, y_train.*

```
clf = LinearDiscriminantAnalysis()
model=clf.fit(X_train,y_train)
```

Then we will predict the values on train and test data *X_train , X_test.*

With a cut off value 0.5

PROBLEM 1.5

Apply KNN Model and Naïve Bayes Model. Interpret the results

Resolution:

Here we have to note one thing is data is not scaled, and KNN algorithm is sensitive to this kind of data, so we will scale the data using *Z-score* before feeding to the model.

We will import zscore from scipy.stats

After applying z score to the data, and after scaling all the dependent variables,

Data is converted from -1 to +1.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_male
0	-0.711973	-0.279218	-0.150948	0.566716	-1.419886	-1.434426	0.422643	-0.937059
1	-1.157661	0.856268	0.924730	0.566716	1.018544	-0.524358	0.422643	1.067169
2	-1.221331	0.856268	0.924730	1.418187	-0.607076	-1.131070	0.422643	1.067169
3	-1.921698	0.856268	-1.226625	-1.136225	-1.419886	-0.827714	-1.424148	-0.937059
4	-0.839313	-1.414704	-1.226625	-1.987695	-1.419886	-0.221002	0.422643	1.067169

From *sklearn.neighbors* we will import *KNeighborsClassifier.*

then we will create KNeighborsClassifier()

then we will fit it on train data x_train and y_train.

Apply Naïve Bayes Model:

From sklearn.naive_bayes we will import GaussianNB

Then we will create Naïve Bayes model and fit it on (X_train, y_train).

```
NB_model = GaussianNB()
NB_model.fit(X_train, y_train)
```

```
GaussianNB()
```

PROBLEM 1.6

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

Resolution:

Apply Random Forest:

For Bagging we will use Random Forest algorithm as Suggested.

First we will import RandomForestClassifier from sklearn.ensemble.

Then we will create Random forest Classifier with n_estimators as 100.

n_estimators : This is the number of trees we want to build before taking the maximum voting of predictions. Higher number of trees gives better performance.

Then we will fit the model on (X_train, y_train) dataset.

```
from sklearn.ensemble import AdaBoostClassifier

ADB_model = AdaBoostClassifier(n_estimators=100,random_state=1)
ADB_model.fit(X_train,y_train)
```

Apply Bagging

Here we will create Bagging Classifier, Where we will select trees as 100, and base

Estimator as Classification and Regression tree.

Then we will fit the model on train data

```
from sklearn.ensemble import BaggingClassifier
Bagging_model=BaggingClassifier(base_estimator=cart,n_estimators=100,random_state=1)
Bagging_model.fit(X_train, y_train)
```

```
BaggingClassifier(base_estimator=DecisionTreeClassifier(), n_estimators=100,
                  random_state=1)
```

Ada boost
First we will import AdaBoostClassifier from sklearn.ensemble library.

Then we will create Ada boost model.

Then we will fit the model on (X_train, y_train) dataset.

```
from sklearn.ensemble import AdaBoostClassifier

ADB_model = AdaBoostClassifier(n_estimators=100,random_state=1)
ADB_model.fit(X_train,y_train)
```

```
AdaBoostClassifier(n_estimators=100, random_state=1)
```

PROBLEM 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.

Resolution:

So far we have created and trained the models Logistic Regression, LDA, Naive Bayes, KNN, Random Forest, Ada boost on Training data set. Now we will check Performance of all of them.

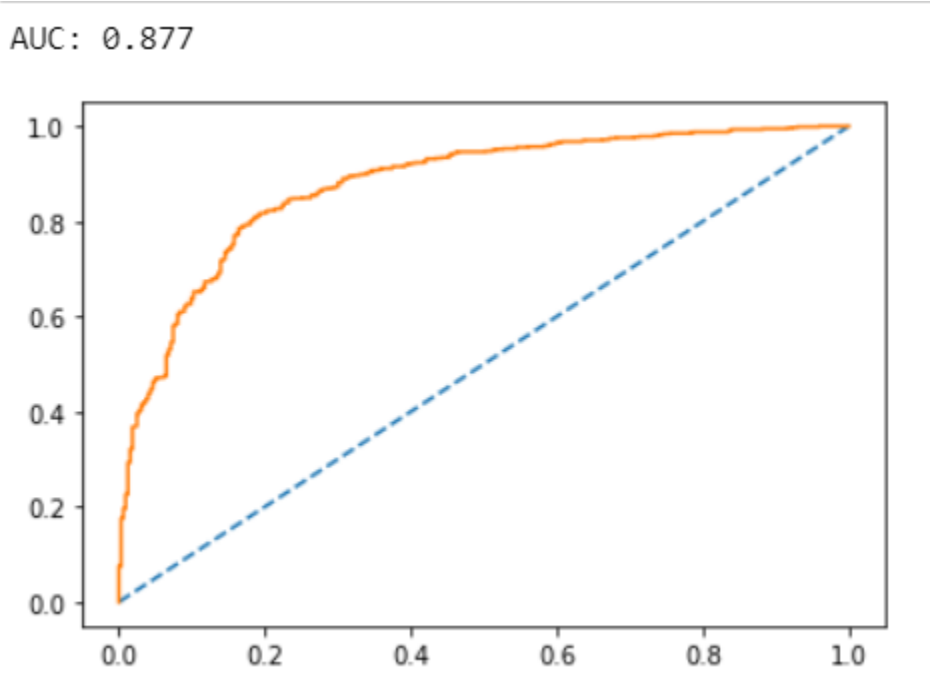
1. **Logistic Regression**

   Accuracy and AUC-ROC:

   On Train data:

   83% is the Accuracy on train data.

   AUC and ROC curve for training data:

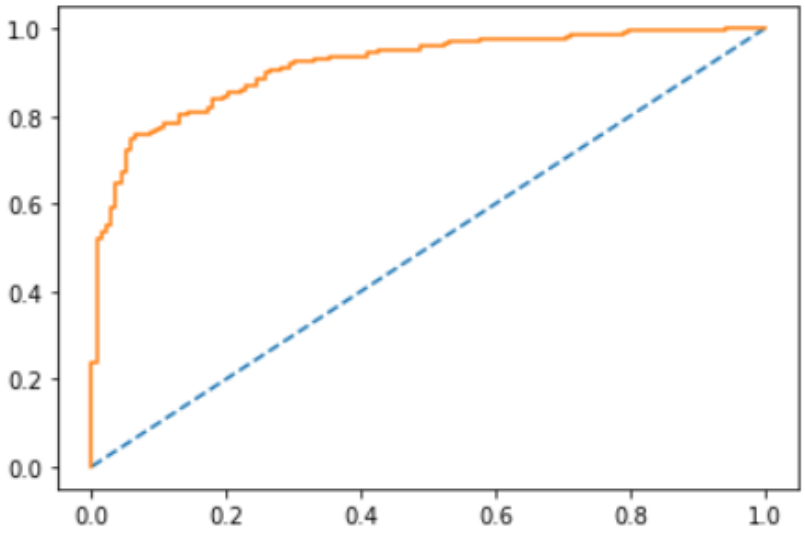


On Test data:

84.93% is the Accuracy on train data.

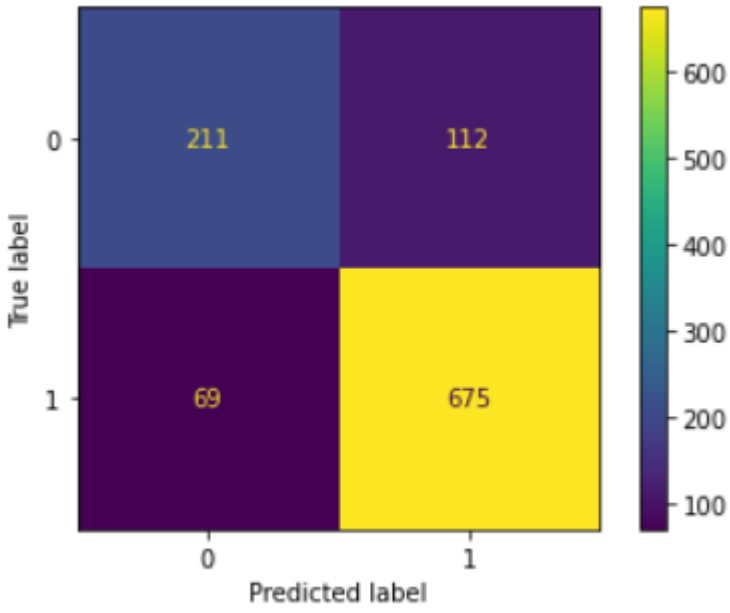
AUC and ROC curve for training data:

```
AUC: 0.914
```

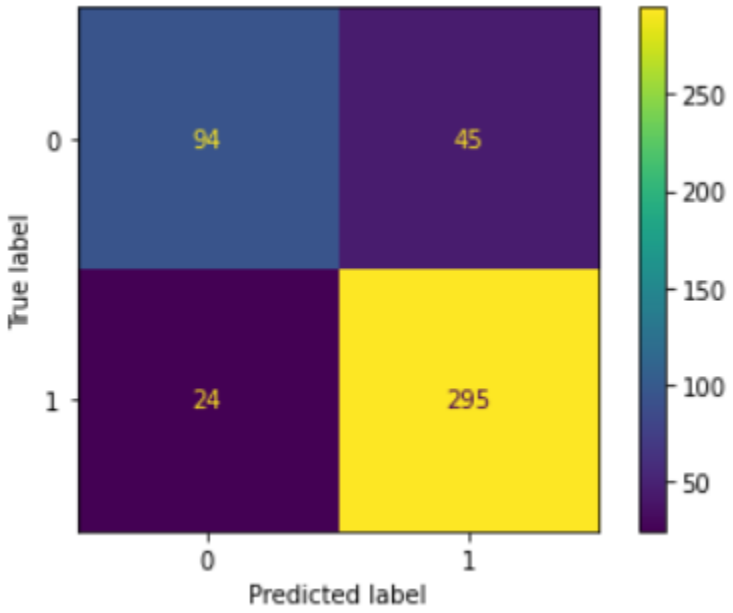


Confusion Matrix

On Train data



On Test Data : Confusion Matrix,



Classification Report

On Train data:

	precision	recall	f1-score	support
0	0.75	0.65	0.70	323
1	0.86	0.91	0.88	744
accuracy			0.83	1067
macro avg	0.81	0.78	0.79	1067
weighted avg	0.83	0.83	0.83	1067

```
lr_train_precision  0.86
lr_train_recall  0.91
lr_train_f1  0.88
```

On Test Data ,

	precision	recall	f1-score	support
0	0.80	0.68	0.73	139
1	0.87	0.92	0.90	319
accuracy			0.85	458
macro avg	0.83	0.80	0.81	458
weighted avg	0.85	0.85	0.85	458

```
lr_test_precision  0.87
lr_test_recall  0.92
lr_test_f1  0.9
```

We can see both 0's and 1's are equally Important here, because vote to both leaders matter to us. Logistic Regression has performed really well no over fitting issue can be Observed here.

Precision, recall and f1 score to predict 0, is 75%, 65% , 70% on Train data.

It has 83%-84% Accuracy on test data, while f1 score is also 88-90%.

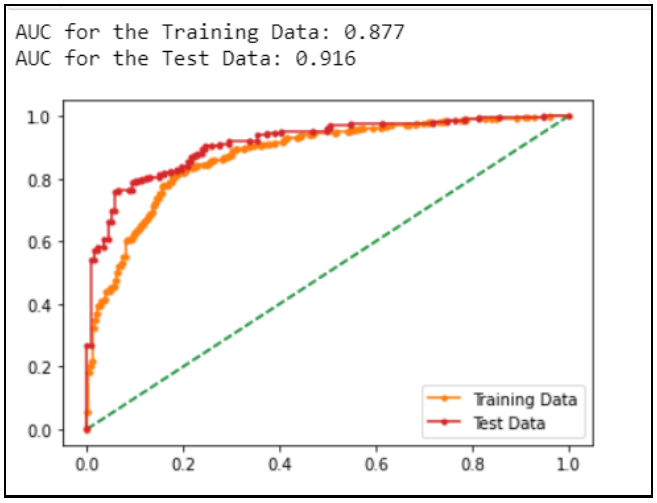
2. Linear Discriminant Analysis

Accuracy and AUC-ROC:

On Train and Test data:

83% and 85% is the Accuracy on train data and test data respectively.

AUC and ROC curve for training and test data:



Confusion Matrix for LDA:

Classification Report for LDA:

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.67	0.70	323
1	0.86	0.90	0.88	744
accuracy			0.83	1067
macro avg	0.80	0.78	0.79	1067
weighted avg	0.83	0.83	0.83	1067

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.79	0.71	0.75	139
1	0.88	0.92	0.90	319
accuracy			0.85	458
macro avg	0.83	0.81	0.82	458
weighted avg	0.85	0.85	0.85	458

LDA has also performed really well no over fitting issue can be Observed here. On Test data and train data only + - 2% is the difference.

To Predict 0, the precision, recall and f1 score has fallen as compare to Predict 1.

It has 83%-85% Accuracy, while f1 score is also 88-90%.

3. **K Nearest Neighbour**

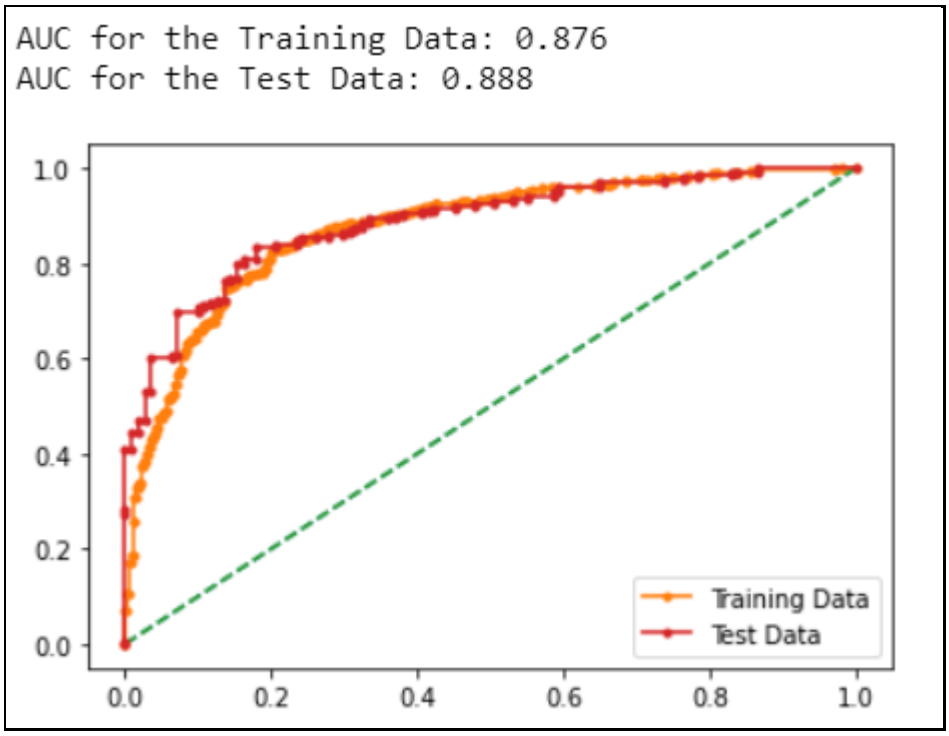
   Accuracy and AUC-ROC:

   On Train data:

   87% is the Accuracy on train data.

   82% is the Accuracy on test data.

   AUC and ROC curve for and test data:



Confusion Matrix for KNN:

Classification Report for KNN :

Classification report for training data:

	precision	recall	f1-score	support
0	0.81	0.75	0.78	351
1	0.89	0.92	0.91	792
accuracy			0.87	1143
macro avg	0.85	0.83	0.84	1143
weighted avg	0.87	0.87	0.87	1143

Classification report for test data:

	precision	recall	f1-score	support
0	0.69	0.73	0.71	111
1	0.89	0.86	0.87	271
accuracy			0.82	382
macro avg	0.79	0.80	0.79	382
weighted avg	0.83	0.82	0.83	382

Here in KNN we can see the value is little lower in case of test data

Accuracy and f1 score, but again even this model will perform well,

We need to hyperparameter tuning to improve performance.

It is predicting 1 with higher f1 score but it is showing poor performance to predict 0.

4. **Naive Bayes Algorithm**

   Accuracy and AUC-ROC:

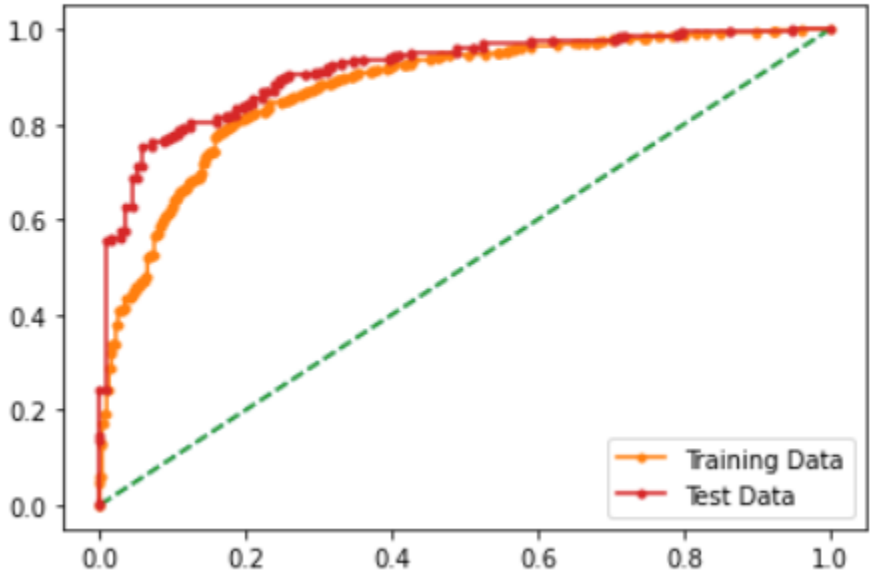
   On Train data:

   82% is the Accuracy on train data.

   84% is the Accuracy on test data.

   AUC and ROC curve for train and test data:

```
AUC for the Training Data: 0.876
AUC for the Test Data: 0.915
```



Confusion Matrix:

Classification Report for Naïve Bayes Algorithm:

On Train data:

```
0.8219306466729147
[[223 100]
 [ 90 654]]
              precision    recall  f1-score   support

           0       0.71      0.69      0.70       323
           1       0.87      0.88      0.87       744

    accuracy                           0.82      1067
   macro avg       0.79      0.78      0.79      1067
weighted avg       0.82      0.82      0.82      1067
```

On Test Data:

```
0.8471615720524017
[[101  38]
 [ 32 287]]
              precision    recall  f1-score   support

           0       0.76      0.73      0.74       139
           1       0.88      0.90      0.89       319

    accuracy                           0.85       458
   macro avg       0.82      0.81      0.82       458
weighted avg       0.85      0.85      0.85       458
```

We need hyper parameter tuning to improve performance.

It is predicting 1 with higher f1 score but it is showing poor performance to predict 0.

Random Forest

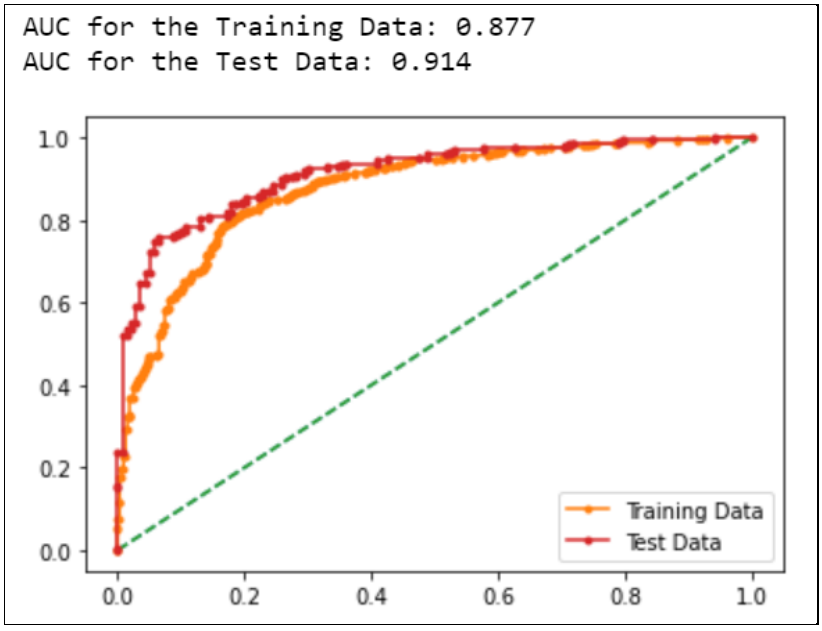
Accuracy and AUC-ROC:

On Train-test data:

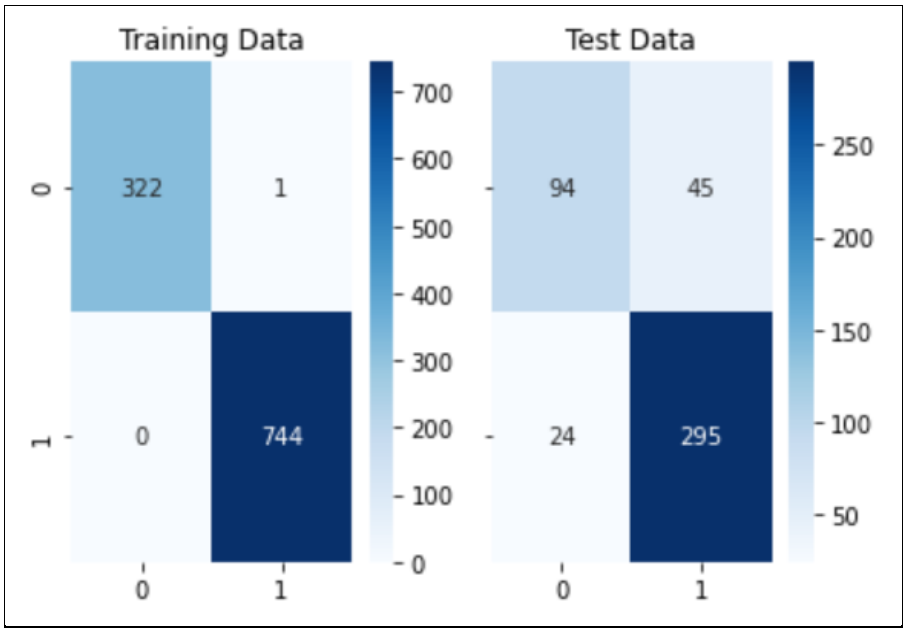
100% is the Accuracy on train data.

85% is the Accuracy on test data.

AUC and ROC curve for train and test data:



Confusion Matrix for Random Forest:



Classification Report for Random Forest:

On Train Data:

```
0.9990627928772259
[[322   1]
 [  0 744]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       323
           1       1.00      1.00      1.00       744

    accuracy                           1.00      1067
   macro avg       1.00      1.00      1.00      1067
weighted avg       1.00      1.00      1.00      1067
```

On Test Data:

```
0.8493449781659389
[[ 94  45]
 [ 24 295]]
              precision    recall  f1-score   support

           0       0.80      0.68      0.73       139
           1       0.87      0.92      0.90       319

    accuracy                           0.85       458
   macro avg       0.83      0.80      0.81       458
weighted avg       0.85      0.85      0.85       458
```

Random Forest is clearly an over fitting one, as almost all metrics has reduced on test data.

5. **Bagging**

Accuracy and AUC-ROC:

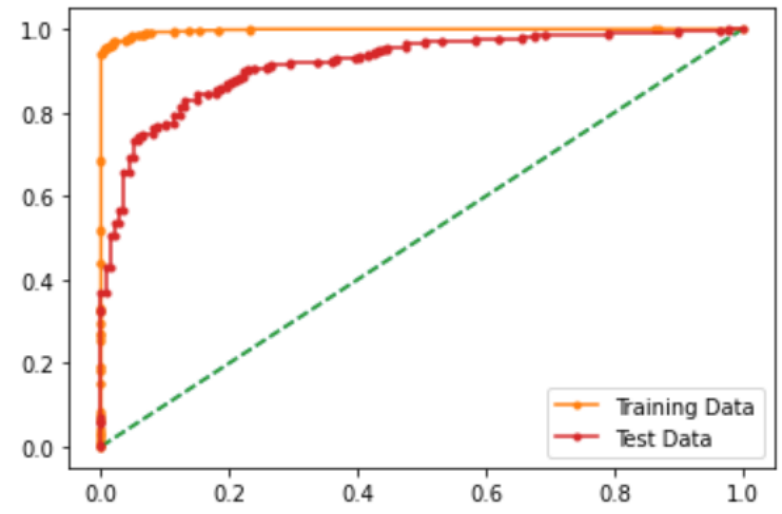
On Train-test data:

99.7% is the Accuracy on train data.

91.8% is the Accuracy on test data.

AUC and ROC curve for train and test data:

```
AUC for the Training Data: 0.997
AUC for the Test Data: 0.918
```



Confusion Matrix for Bagging :

Classification Report for Bagging:

On Train Data,

```
0.971883786316776
[[298  25]
 [  5 739]]
              precision    recall  f1-score   support

           0       0.98      0.92      0.95       323
           1       0.97      0.99      0.98       744

    accuracy                           0.97      1067
   macro avg       0.98      0.96      0.97      1067
weighted avg       0.97      0.97      0.97      1067
```

On Test data,

```
0.8427947598253275
[[ 93  46]
 [ 26 293]]
              precision    recall  f1-score   support

           0       0.78      0.67      0.72       139
           1       0.86      0.92      0.89       319

    accuracy                           0.84       458
   macro avg       0.82      0.79      0.81       458
weighted avg       0.84      0.84      0.84       458
```

Model has stability in predicting 1 while Predicting 0, Bagging Model is Over fitting.

6. **Ada Boost**

Accuracy and AUC-ROC:

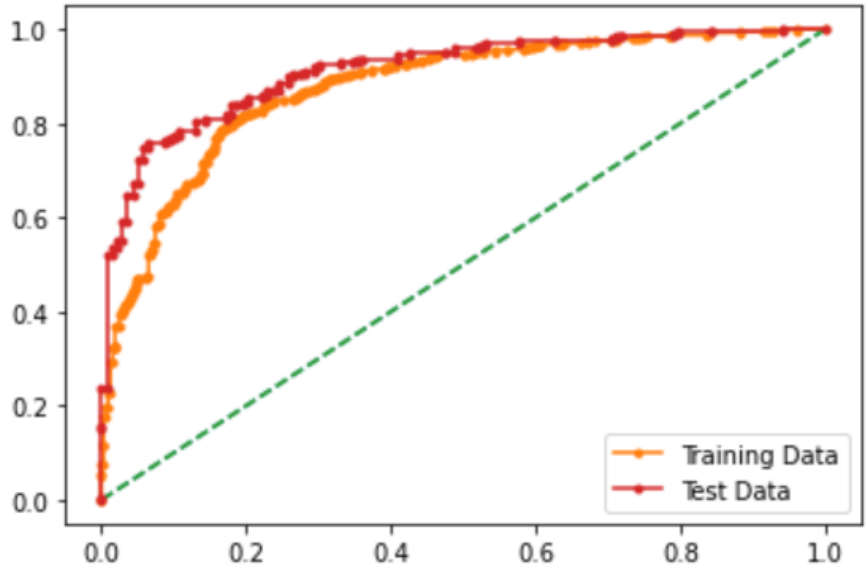
On Train-test data:

84.5% is the Accuracy on train data.

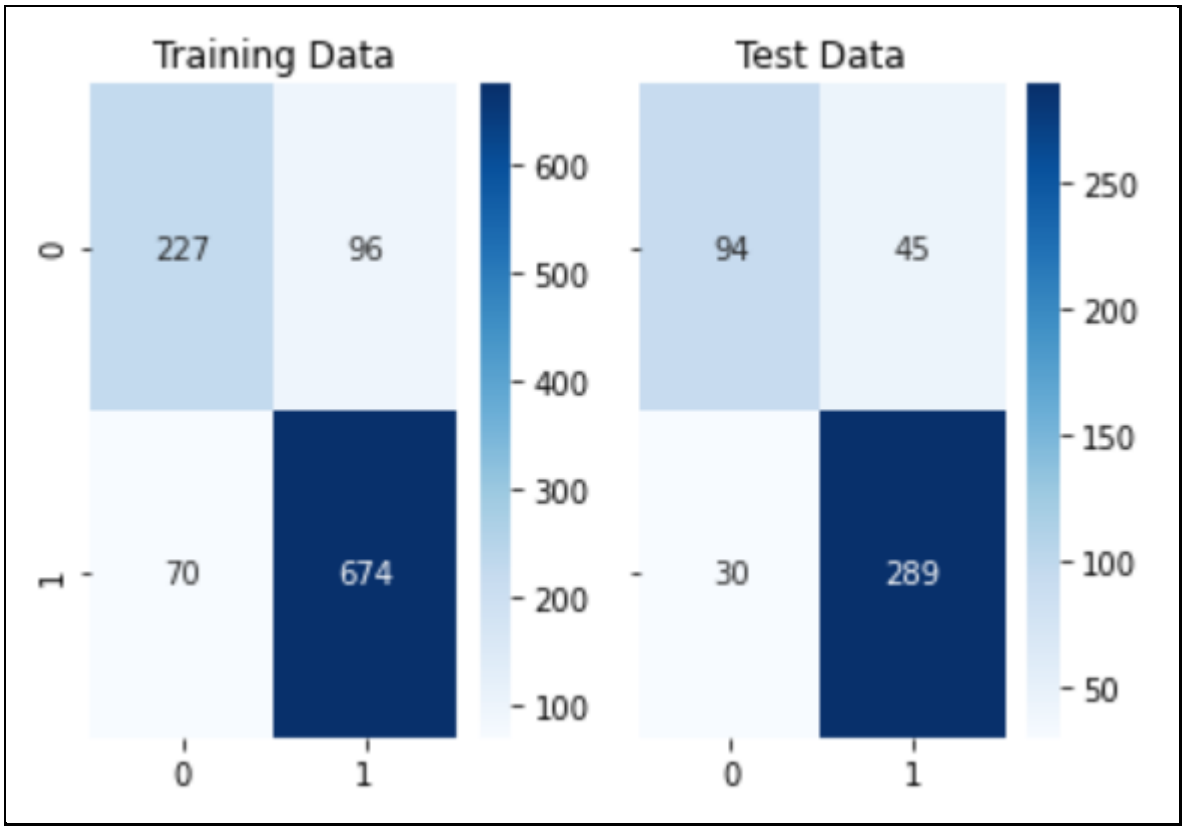
83.62% is the Accuracy on test data.

AUC and ROC curve for train and test data:

```
AUC for the Training Data: 0.877
AUC for the Test Data: 0.914
```



Confusion Matrix for Ada Boost:



Classification Report for Ada Boost:

On Train Data,

```
0.8444236176194939
[[227  96]
 [ 70 674]]
              precision    recall  f1-score   support

           0       0.76      0.70      0.73       323
           1       0.88      0.91      0.89       744

    accuracy                           0.84      1067
   macro avg       0.82      0.80      0.81      1067
weighted avg       0.84      0.84      0.84      1067
```

On Test Data,

```
0.8362445414847162
[[ 94  45]
 [ 30 289]]
              precision    recall  f1-score   support

           0       0.76      0.68      0.71       139
           1       0.87      0.91      0.89       319

    accuracy                           0.84       458
   macro avg       0.81      0.79      0.80       458
weighted avg       0.83      0.84      0.83       458
```

Model is Performing well as compare to other models.

All Model Comparison by its performance metrics:

	Logit Train	Logit Test	LDA Train	LDA Test	KNN Train	KNN Test	Naive Bayes Train	Naive Bayes Test	RF Train	RF Test	Ada Boost Train	Ada Boost Test	Ada Boost Train_0	Ada Boost Test_0
Accuracy	0.83	0.85	0.83	0.84	0.87	0.71	0.82	0.85	1.00	0.85	0.84	0.84	0.84	0.84
AUC	0.88	0.91	0.88	0.91	0.88	0.89	0.88	0.91	0.88	0.91	0.90	0.91	0.90	0.91
Recall	0.91	0.92	0.90	0.91	0.92	0.86	0.88	0.90	1.00	0.92	0.70	0.91	0.70	0.68
Precision	0.86	0.87	0.86	0.87	0.89	0.89	0.87	0.88	1.00	0.87	0.76	0.87	0.76	0.76
F1 Score	0.88	0.90	0.88	0.89	0.91	0.87	0.87	0.89	1.00	0.90	0.73	0.89	0.73	0.71

Now the Train data seems to be unbalanced, like 30-70% Conservative Party and Labour party votes, so we will use SMOTE oversampling method on train data here to balance the Data.

Naive Bayes with SMOTE

Performance metrics for Naive Bayes with SMOTE:

On Train Data:

```
0.8131720430107527
[[595 149]
 [129 615]]
              precision    recall  f1-score   support

           0       0.82      0.80      0.81       744
           1       0.80      0.83      0.82       744

    accuracy                           0.81      1488
   macro avg       0.81      0.81      0.81      1488
weighted avg       0.81      0.81      0.81      1488
```

On Test Data:

```
0.8427947598253275
[[109  30]
 [ 42 277]]
              precision    recall  f1-score   support

           0       0.72      0.78      0.75       139
           1       0.90      0.87      0.88       319

    accuracy                           0.84       458
   macro avg       0.81      0.83      0.82       458
weighted avg       0.85      0.84      0.84       458
```

Model is overfitting here, so we will see what happens with SMOTE and Ada Boost, Linear Regression.

SMOTE with Ada Boost:

Classification Report:

On Train Data,

```
0.8145161290322581
[[601 143]
 [133 611]]
              precision    recall  f1-score   support

           0       0.82      0.81      0.81       744
           1       0.81      0.82      0.82       744

    accuracy                           0.81      1488
   macro avg       0.81      0.81      0.81      1488
weighted avg       0.81      0.81      0.81      1488
```

On Test Data,

```
0.8362445414847162
[[110  29]
 [ 46 273]]
              precision    recall  f1-score   support

           0       0.71      0.79      0.75       139
           1       0.90      0.86      0.88       319

    accuracy                           0.84       458
   macro avg       0.80      0.82      0.81       458
weighted avg       0.84      0.84      0.84       458
```

This is again over fitting / under fitting issue while predicting 0's and 1.

SMOTE with Logistic Regression:

Classification Report:

On Train Data,

```
0.821236559139785
[[611 133]
 [133 611]]
              precision    recall  f1-score   support

           0       0.82      0.82      0.82       744
           1       0.82      0.82      0.82       744

    accuracy                           0.82      1488
   macro avg       0.82      0.82      0.82      1488
weighted avg       0.82      0.82      0.82      1488
```

On Test Data,

```
0.8362445414847162
[[114  25]
 [ 50 269]]
              precision    recall  f1-score   support

           0       0.70      0.82      0.75       139
           1       0.91      0.84      0.88       319

    accuracy                           0.84       458
   macro avg       0.81      0.83      0.82       458
weighted avg       0.85      0.84      0.84       458
```

Not exactly but Logistic Regression seems much stable as Compare to other models, after applying SMOTE-balanced data.

Model Tuning:

Now we will use Model tuning for various Algorithms to check whether the performance is Improving or not by changing its hyper parameters.

For Naive Bayes Model, will use K-fold validation how model performs on Limited dataset.

```
Cross Validation Score: [0.82242991 0.8411215  0.81308411 0.81308411 0.81308411 0.82242991
 0.79439252 0.87735849 0.81132075 0.81132075] [0.82242991 0.8411215  0.81308411 0.81308411 0.81308411 0.82242991
 0.79439252 0.87735849 0.81132075 0.81132075]
Average Score: 0.8219626168224299
```

Average Score is 0.82, which is good score, it is performing well.

Let's check using GRID search, for KNN

```
params = {'n_neighbors':[2,4,6,8,10,12,14,16,18],
          'algorithm': ['auto','ball_tree','kd_tree','brute'],
          'leaf_size':list(range(1,30)),
          'p':[1,2],
          'metric':['minkowski','euclidean','manhattan','chebyshev', 'mahanalobis']}
```

We are here using different distance metrics, P value is

When p = 1, this is equivalent to using manhattan_distance

P=2,Power parameter for the Minkowski metric.

We are selecting leaf size from 1 to 30, to get the optimal value.

Before fitting the data, we are scaling train data by using z score

Scaled data:

age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_male
0.497751	-0.279218	-0.150948	-1.136225	1.831354	0.385710	-1.424148	1.067169
-0.902983	-0.279218	0.924730	0.566716	-1.419886	-1.131070	1.346038	1.067169
-0.138946	-1.414704	-0.150948	-1.136225	1.831354	1.295778	0.422643	-0.937059
-0.457295	-0.279218	-0.150948	0.566716	-0.607076	-0.827714	-1.424148	-0.937059
-0.202616	-0.279218	-1.226625	0.566716	-0.607076	-1.434426	0.422643	-0.937059
...	...	...	...	...	...	...	...
0.816100	0.856268	-0.150948	1.418187	-0.607076	-1.434426	0.422643	1.067169
-1.730689	1.991754	2.000408	-1.136225	-1.419886	-0.827714	1.346038	1.067169
-1.285001	0.856268	2.000408	0.566716	1.018544	0.082354	0.422643	-0.937059
-1.157661	0.856268	0.924730	0.566716	-0.607076	0.082354	0.422643	-0.937059
-1.412340	-1.414704	-1.226625	0.566716	1.018544	-0.524358	1.346038	1.067169

The values set for algorithm,

```
GridSearchCV(estimator=KNeighborsClassifier(),
             param_grid={'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
                         'leaf_size': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,
                                       13, 14, 15, 16, 17, 18, 19, 20, 21, 22,
                                       23, 24, 25, 26, 27, 28, 29],
                         'metric': ['minkowski', 'euclidean', 'manhattan',
                                    'chebyshev', 'mahanalobis'],
                         'n_neighbors': [2, 4, 6, 8, 10, 12, 14, 16, 18],
                         'p': [1, 2]},
             verbose=1)
```

These are the best Parameters for KNN:

```
{'algorithm': 'auto',
 'leaf_size': 1,
 'metric': 'minkowski',
 'n_neighbors': 16,
 'p': 2}
```

Train and Test Accuracy for KNN after Grid Search CV.

```
Train Accuracy is :0.8433945756780402

Test Accuracy is :0.8272251308900523
```

It is almost same as earlier.

Lets check using GRID search, for Logistic Regression and Random Forest

We will import Pipeline from sklearn.pipeline library.

In this we will create two Classifier – Logistic Regression and Random Forest will see the best performing model with its best Parameters.

```
{'classifier': LogisticRegression(C=11.288378916846883, solver='liblinear'),
 'classifier__C': 11.288378916846883,
 'classifier__penalty': 'l2',
 'classifier__solver': 'liblinear'}
```

Logistic Regression Classifier is giving best results, with Solver Liblinear

With Classifier Penalty as Ridge.

```
Train Accuracy is :0.8294283036551078

Test Accuracy is :0.8493449781659389
```

Classification Report:

```
Classification report for Train set:
              precision    recall  f1-score   support

           0       0.75      0.65      0.70       323
           1       0.86      0.91      0.88       744

    accuracy                           0.83      1067
   macro avg       0.80      0.78      0.79      1067
weighted avg       0.83      0.83      0.83      1067
```

```
Classification report for Test set:
              precision    recall  f1-score   support

           0       0.80      0.68      0.73       139
           1       0.87      0.92      0.90       319

    accuracy                           0.85       458
   macro avg       0.83      0.80      0.81       458
weighted avg       0.85      0.85      0.85       458
```

The Logistic Regression is Performing well on Test data and Train data, it seems to be stable. It is predicting 0 and 1 clearly.

It has low scores while predicting 0 but for 1 Precision, Recall and F1 is good on both train and test data.

Prediction of 0 i.e vote to conservative Party, prediction is little tougher as there is very less data points conveying vote to Conservative Party. This can be minimized if we ask for more data.

If we check the first 10 votes, It is clearly, mentioning the vote to Labour party.

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=uint8)
```

First 400 votes,

```
{0: 96, 1: 304}
```

Here out of 400, 96 votes are to the Conservative party and 304 votes are to the Labour party so clearly it is indicating the Labour party will win.

Again there are some assumptions and data limits we have, so as of now this is the scenario.

PROBLEM 1.8

Based on these predictions, what are the insights?

Steps Performed and Insights:

To predict results, we have started with Exploratory data Analysis where we got some hidden insights by looking at Dataset, by using univariate analysis, Multivariate analysis. Null value, Outlier detection. From EDA we have drawn below mentioned Insights:

Assessment to leader Blair and Economic Household Condition Rating, Economic national Condition rating shows good correlation.

If person is giving good assessment score to Hague he must be with Conservative party and he or she is Highly Eurosceptic.

It is not always true, but people having good political Knowledge are preferring Europe Integration.

Modelling

As **the** target variable here is Categorcal i.e Conservative and Labour, we can use Logistic Regression, LDA, KNN, Naive Bayes, RF, Bagging, Boosting.

After all this we have created a Final Classification report where all performance metrics are mentioned.

We have used all mentioned Algorithms. If we check the Target Variable we have almost 70% data of Labour Party and 30 % Conservative Party. We have built all algorithms on this data first, after this we have used SMOTE technique of Oversampling to balance Imbalanced data. We can see in the Classification report that Accuracy, Precision has been dropped.

After all this to do Model tunning we have used Gridsearch CV, K fold cross validation to Improve performance of Existing Base models.

Results:

To get the crystal clear picture, we need more data points to avoid Imbalance data issue. but from this data, we can say that to win the election one must be with vision Europe integration, as this vision has helped labour party.

The Model Logistic Regression has good score and seems to be very stable model in this Scenario. If we predict on Test Data, The labour Party seems to be get Higher number of Votes as compare to Conservative Party. Again the question arises here does this sample represent the whole nation, any bias in the data, like from Particular region data is collected where Labour Party has higher influence. There are less votes to Conservative party, but again it needs more data to predict correct results. But as of now this data tells us that Labour party will win these elections.

^^^************************************

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.

PROBLEM 2.1

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

Resolution:

The number of characters for Roosevelt is: 7571

The number of words for Roosevelt speech is: 1323

The number of sentences for Roosevelt speech is: 68

The number of characters for John F. Kennedy is: 7618

The number of words for John F. Kennedy speech is: 1364

The number of sentences for John F. Kennedy speech is: 52

The number of characters for John F. Kennedy is: 9991

The number of words for John F. Kennedy speech is: 1769

The number of sentences for John F. Kennedy speech is: 69

	president	text	char_count	word_count	sents_count
1941-Roosevelt	Roosevelt - 1941	On each national day of inauguration since 178...	7571	1323	68
1961-Kennedy	Kennedy - 1961	Vice President Johnson, Mr. Speaker, Mr. Chief...	7618	1364	52
1973-Nixon	Nixon - 1973	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991	1769	68

PROBLEM 2.2

Remove all the stop words from all three speeches.

Resolution:

Download stopwords from nltk

After removing stop words from Franklin D. Roosevelt

```
['national day inauguration since 1789 people renewed sense dedication united states washingtons day task people create weld to
gether nation lincolns day task people preserve nation disruption within day task people save nation institutions disruption wi
thout come time midst swift happenings pause moment take stock recall place history rediscover may risk real peril inaction liv
es nations determined count years lifetime human spirit life man threescore years ten little little less life nation fullness m
easure live men doubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplain
ed reason tyranny slavery become surging wave future freedom ebbing tide americans know true eight years ago life republic seem
ed frozen fatalistic terror proved true midst shock acted acted quickly boldly decisively later years living years fruitful yea
rs people democracy brought greater security hope better understanding lifes ideals measured material things vital present futu
re experience democracy successfully survived crisis home put away many evil things built new structures enduring lines maintai
ned fact democracy action taken within threeway framework constitution united states coordinate branches government continue fr
eely function bill rights remains inviolate freedom elections wholly maintained prophets downfall american democracy seen dire
predictions come naught democracy dying know seen reviveand grow know cannot die built unhampered initiative individual men wom
en joined together common enterprise enterprise undertaken carried free expression free majority know democracy alone forms gov
ernment enlists full force mens enlightened know democracy alone constructed unlimited civilization capable infinite progress i
mprovement human life know look surface sense still spreading every continent humane advanced end unconquerable forms human soc
iety nation like person bodya body must fed clothed housed invigorated rested manner measures objectives time nation like perso
n mind mind must kept informed alert must know understands hopes needs neighbors nations live within narrowing circle world nat
ion like person something deeper something permanent something larger sum parts something matters future calls forth sacred gua
rding present thing find difficult even impossible hit upon single simple word yet understand spirit faith america product cent
uries born multitudes came many lands high degree mostly plain people sought early late find freedom freely democratic aspirati
on mere recent phase human history human history permeated ancient life early peoples blazed anew middle ages written magna cha
rta americas impact irresistible america new world tongues peoples continent newfound land came believed could create upon cont
inent new life life new freedom vitality written mayflower compact declaration independence constitution united states gettysbu
rg address first came carry longings spirit millions followed stock sprang moved forward constantly consistently toward ideal g
ained stature clarity generation hopes republic cannot forever tolerate either undeserved poverty selfserving wealth know still
far go must greatly build security opportunity knowledge every citizen measure justified resources capacity land enough achieve
purposes alone enough clothe feed body nation instruct inform mind also spirit three greatest spirit without body mind men know
nation could live spirit america killed even though nations body mind constricted alien world lived america know would perished
spirit faith speaks daily lives ways often unnoticed seem obvious speaks capital nation speaks processes governing sovereigntie
s 48 states speaks counties cities towns villages speaks nations hemisphere across seas enslaved well free sometimes fail hear
heed voices freedom privilege freedom old old story destiny america proclaimed words prophecy spoken first president first inau
gural 1789 words almost directed would seem year 1941 preservation sacred fire liberty destiny republican model government just
ly considered deeply finally staked experiment intrusted hands american people lose sacred fireif smothered doubt fear reject d
estiny washington strove valiantly triumphantly establish preservation spirit faith nation furnish highest justification every
sacrifice may make cause national defense face great perils never encountered strong purpose protect perpetuate integrity democ
racy muster spirit america faith america retreat content stand still americans go forward service country god']
```

After removing stop words from John F. Kennedy

```
['vice president johnson speaker chief justice president eisenhower vice president nixon president truman reverend clergy fello
w citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change sworn
almighty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands power
abolish forms human poverty forms human life yet revolutionary beliefs forebears fought still issue around globe belief rights
man come generosity state hand god dare forget today heirs first revolution word go forth time place friend foe alike torch pas
sed new generation americans born century tempered war disciplined hard bitter peace proud ancient heritage unwilling witness p
ermit slow undoing human rights nation always committed committed today home around world every nation know whether wishes well
ill pay price bear burden meet hardship support friend oppose foe order assure survival success liberty much pledge old allies
whose cultural spiritual origins share pledge loyalty faithful friends united little cannot host cooperative ventures divided l
ittle dare meet powerful challenge odds split asunder new states welcome ranks free pledge word one form colonial control passe
d away merely replaced far iron tyranny always expect find supporting view always hope find strongly supporting freedom remembe
r past foolishly sought power riding back tiger ended inside peoples huts villages across globe struggling break bonds mass mis
ery pledge best efforts help help whatever period required communists may seek votes right free society cannot help many poor c
annot save rich sister republics south border offer special pledge convert good words good deeds new alliance progress assist f
ree men free governments casting chains poverty peaceful revolution hope cannot become prey hostile powers neighbors know join
oppose aggression subversion anywhere americas every power know hemisphere intends remain master house world assembly sovereign
states united nations last best hope age instruments war far outpaced instruments peace renew pledge supportto prevent becoming
merely forum invective strengthen shield new weak enlarge area writ may run finally nations would make adversary offer pledge r
equest sides begin anew quest peace dark powers destruction unleashed science engulf humanity planned accidental selfdestructio
n dare tempt weakness arms sufficient beyond doubt certain beyond doubt never employed neither two great powerful groups nation
s take comfort present course sides overburdened cost modern weapons rightly alarmed steady spread deadly atom yet racing alter
uncertain balance terror stays hand mankinds final war begin anew remembering sides civility sign weakness sincerity always sub
ject proof never negotiate fear never fear negotiate sides explore problems unite instead belaboring problems divide sides firs
t time formulate serious precise proposals inspection control arms bring absolute power destroy nations absolute control nation
s sides seek invoke wonders science instead terrors together explore stars conquer deserts eradicate disease tap ocean depths e
ncourage arts commerce sides unite heed corners earth command isaiah undo heavy burdens oppressed go free beachhead cooperation
may push back jungle suspicion sides join creating new endeavor new balance power new world law strong weak secure peace preser
ved finished first 100 days finished first 1000 days life administration even perhaps lifetime planet begin hands fellow citize
ns mine rest final success failure course since country founded generation americans summoned give testimony national loyalty g
raves young americans answered call service surround globe trumpet summons call bear arms though arms need call battle though e
mbattled call bear burden long twilight struggle year year rejoicing hope patient tribulation struggle common enemies man tyran
ny poverty disease war forge enemies grand global alliance north south east west assure fruitful life mankind join historic eff
ort long history world generations granted role defending freedom hour maximum danger shrink responsibility welcome believe wou
ld exchange places people generation energy faith devotion bring endeavor light country serve glow fire truly light world fello
w americans ask country ask country fellow citizens world ask america together freedom man finally whether citizens america cit
izens world ask high standards strength sacrifice ask good conscience sure reward history final judge deeds go forth lead land
love asking blessing help knowing earth gods work must truly']
```

After removing stop words from Richard Nixon.

```
['vice president speaker chief justice senator cook mrs eisenhower fellow citizens great good country share together met four y
ears ago america bleak spirit depressed prospect seemingly endless war abroad destructive conflict home meet today stand thresh
old new era peace world central question use peace resolve era enter postwar periods often time retreat isolation leads stagnat
ion home invites new danger abroad resolve become time great responsibilities greatly borne renew spirit promise america enter
third century nation past year saw farreaching results new policies peace continuing revitalize traditional friendships mission
s peking moscow able establish base new durable pattern relationships among nations world americas bold initiatives 1972 long r
emembered year greatest progress since end world war ii toward lasting peace world peace seek world flimsy peace merely interlu
de wars peace endure generations come important understand necessity limitations americas role maintaining peace unless america
work preserve peace peace unless america work preserve freedom freedom clearly understand new nature americas role result new p
olicies adopted past four years respect treaty commitments support vigorously principle country right impose rule another force
continue era negotiation work limitation nuclear arms reduce danger confrontation great powers share defending peace freedom wo
rld expect others share time passed america make every nations conflict make every nations future responsibility presume tell p
eople nations manage affairs respect right nation determine future also recognize responsibility nation secure future americas
role indispensable preserving worlds peace nations role indispensable preserving peace together rest world resolve move forward
beginnings made continue bring walls hostility divided world long build place bridges understanding despite profound difference
s systems government people world friends build structure peace world weak safe strong respects right live different system wou
ld influence others strength ideas force arms accept high responsibility burden gladly gladly chance build peace noblest endeav
or nation engage gladly also act greatly meeting responsibilities abroad remain great nation remain great nation act greatly me
eting challenges home chance today ever history make life better america ensure better education better health better housing b
etter transportation cleaner environment restore respect law make communities livable insure godgiven right every american full
equal opportunity range needs great reach opportunities great bold determination meet needs new ways building structure peace a
broad required turning away old policies failed building new era progress home requires turning away old policies failed abroad
shift old policies new retreat responsibilities better way peace home shift old policies new retreat responsibilities better wa
y progress abroad home key new responsibilities lies placing division responsibility lived long consequences attempting gather
power responsibility washington abroad home time come turn away condescending policies paternalism washington knows best person
expected act responsibly responsibility human nature encourage individuals home nations abroad decide locate responsibility pla
ces measure others today offer promise purely governmental solution every problem lived long false promise trusting much govern
ment asked deliver leads inflated expectations reduced individual effort disappointment frustration erode confidence government
people government must learn take less people people remember america built government people welfare work shirking responsibil
ity seeking responsibility lives ask government challenges face together ask government help help national government great vit
al role play pledge government act act boldly lead boldly important role every one must play individual member community day fo
rward make solemn commitment heart bear responsibility part live ideals together see dawn new age progress america together cel
ebrate 200th anniversary nation proud fulfillment promise world americas longest difficult war comes end learn debate differenc
es civility decency reach one precious quality government cannot provide new level respect rights feelings one another new leve
l respect individual human dignity cherished birthright every american else time come renew faith america recent years faith ch
allenged children taught ashamed country ashamed parents ashamed americas record home role world every turn beset find everythi
ng wrong america little right confident judgment history remarkable times privileged live americas record century unparalleled
worlds history responsibility generosity creativity progress proud system produced provided freedom abundance widely shared sys
tem history world proud four wars engaged century including one bringing end fought selfish advantage help others resist aggres
sion proud bold new initiatives steadfastness peace honor made breakthrough toward creating world world known structure peace l
ast merely time generations come embarking today era presents challenges great nation generation ever faced answer god history
conscience way use years stand place hallowed history think others stood think dreams america think recognized needed help far
beyond order make dreams come true today ask prayers years ahead may gods help making decisions right america pray help togethe
r may worthy challenge pledge together make next four years best four years americas history 200th birthday america young vital
```

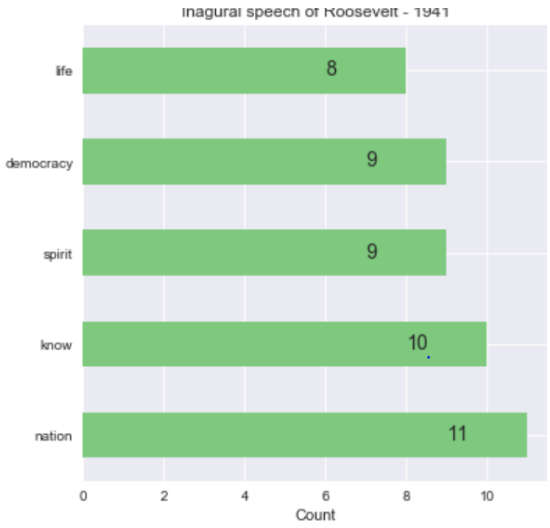
PROBLEM 2.3

Which word occurs the most number of times in his inaugural address for each president? Mention the top three words.

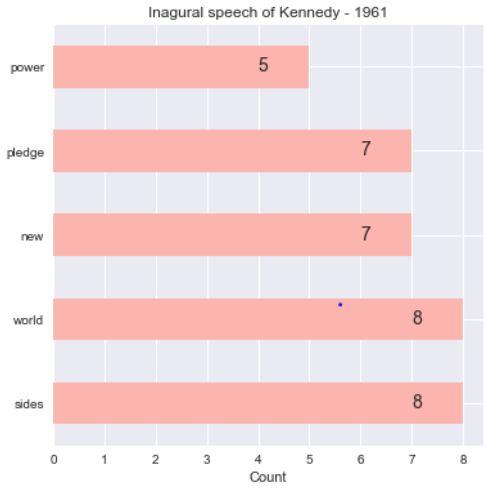
(After removing the stopwords)

Resolution:

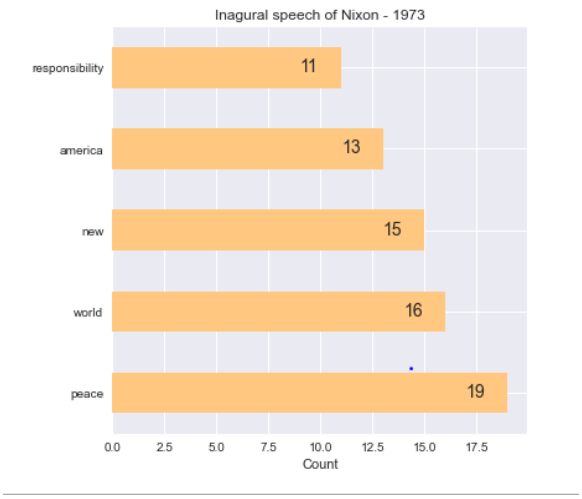
Words that occurs the most in speech Franklin D. Roosevelt is Nation



Words that occurs the most in speech Kennedys Speech is Sides and world



Words that occurs the most in speech Richard Nixon is peace:

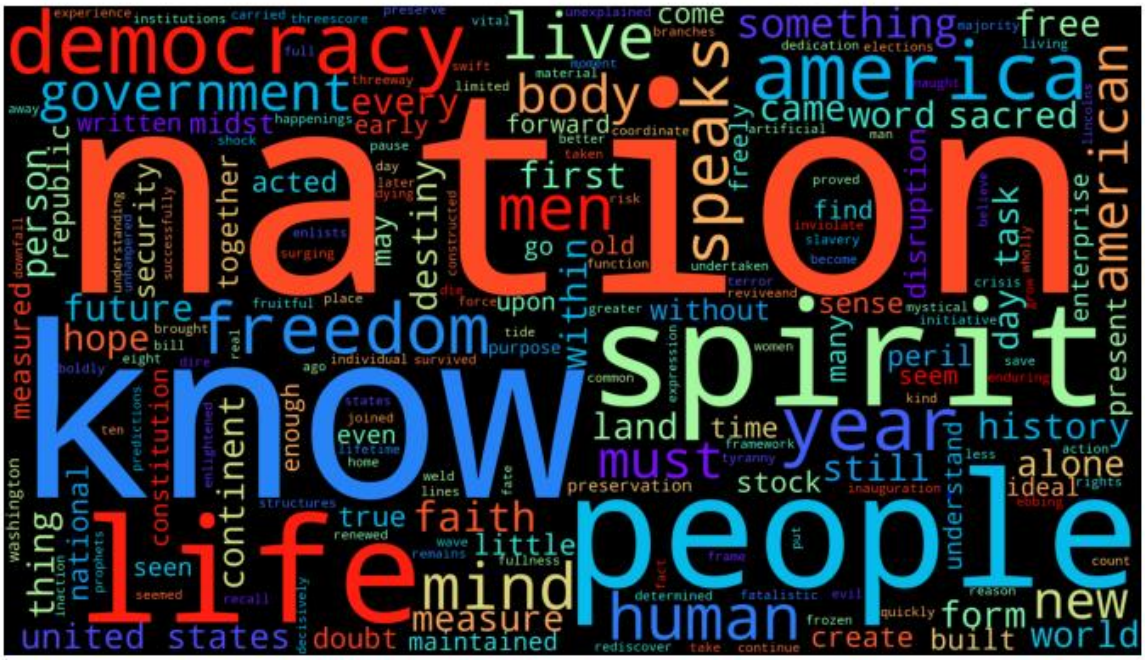


PROBLEM 2.4

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

Resolution:

WordCloud for Franklin D. Roosevelt



WordCloud for John F. Kennedy

WordCloud for Richard Nixon

The End

Thakur Arun Singh

*****************************^^*****************************