Machine Learning Report
- Coded Project

Table of Contents

Table of Figures

Problem 1

Context

CNBE, a prominent news channel, is gearing up to provide insightful coverage of recent elections, recognizing the importance of data-driven analysis. A comprehensive survey has been conducted, capturing the perspectives of 1525 voters across various demographic and socio-economic factors. This dataset encompasses 9 variables, offering a rich source of information regarding voters' characteristics and preferences.

Objective

The primary objective is to leverage machine learning to build a predictive model capable of forecasting which political party a voter is likely to support. This predictive model, developed based on the provided information, will serve as the foundation for creating an exit poll. The exit poll aims to contribute to the accurate prediction of the overall election outcomes, including determining which party is likely to secure the majority of seats.

Data Description

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

Define the problem and perform Exploratory Data Analysis

- Problem definition - Check shape, Data types, and statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables

Importing Data

Out[66]:

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

EDA

There are 1525 Rows and 9 columns

```
no. of rows:  1525
no. of columns:  9
```

There are 2 object datatypes and 7 integer datatypes.

```
<class 'pandas.core.frame.DataFrame'>
Index: 1525 entries, 1 to 1525
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: 119.1+ KB
```

```
Shape of the dataset: (1525, 9)
```

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

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

                          max
age                      93.0
economic.cond.national    5.0
economic.cond.household   5.0
Blair                     5.0
Hague                     5.0
Europe                   11.0
political.knowledge       3.0
```

1. Statistical Summary

There are 0 missing values in all the columns.

```
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
```

```
Variable: vote
Unique values: ['Labour' 'Conservative']
Value counts:
vote
Labour          1063
Conservative     462
Name: count, dtype: int64

Variable: gender
Unique values: ['female' 'male']
Value counts:
gender
female    812
male      713
Name: count, dtype: int64

Variable: age
Minimum value: 24
Maximum value: 93
Mean: 54.18
Standard deviation: 15.71

Variable: economic.cond.national
Minimum value: 1
Maximum value: 5
Mean: 3.25
Standard deviation: 0.88

Variable: economic.cond.household
Minimum value: 1
Maximum value: 5
Mean: 3.14
Standard deviation: 0.93

Variable: Blair
Minimum value: 1
Maximum value: 5
Mean: 3.33
Standard deviation: 1.17

Variable: Hague
Minimum value: 1
Maximum value: 5
Mean: 2.75
Standard deviation: 1.23

Variable: Europe
Minimum value: 1
Maximum value: 11
Mean: 6.73
Standard deviation: 3.30

Variable: political.knowledge
Minimum value: 0
Maximum value: 3
Mean: 1.54
Standard deviation: 1.08
```

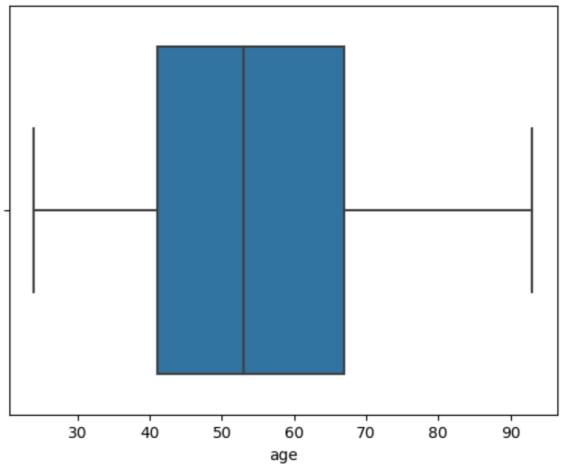
The labour vote is 1063 and the vote for conservative is 462, the labour vote is double the times of conservative votes.

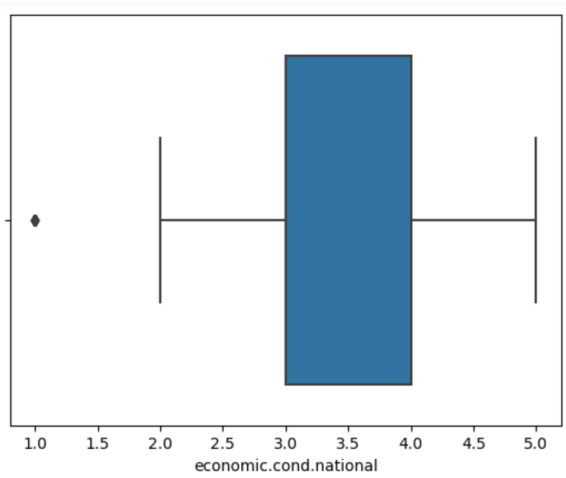
The vote cast by the gender female is more than the gender male.

The mean age cast is 54.

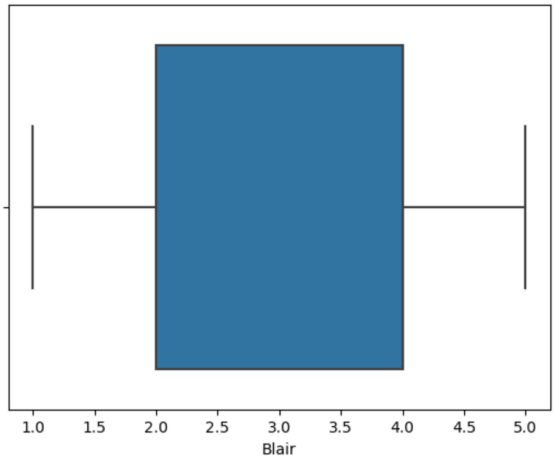
Boxplot of numerical variables

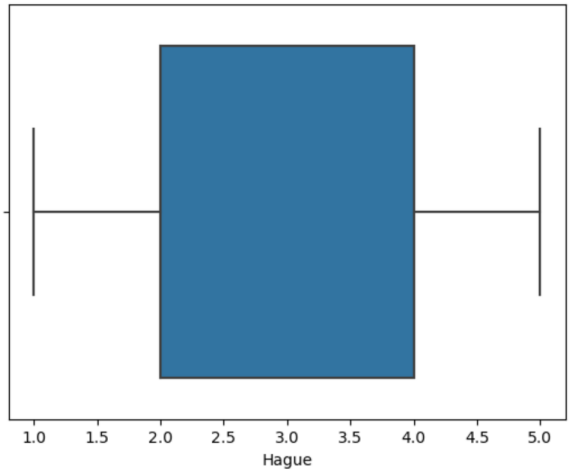
2. Boxplot of Age & Economic cond national



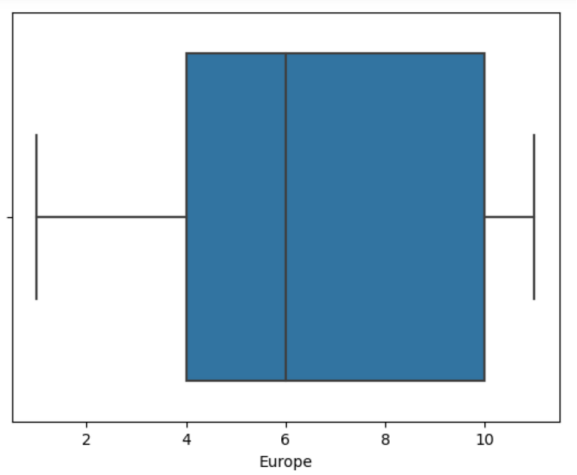


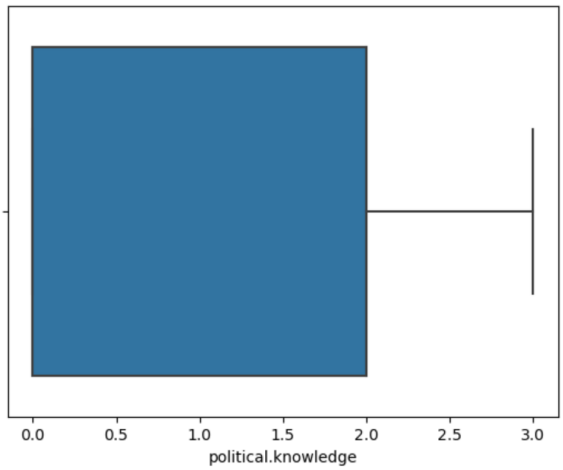
3. Boxplot of Blari & Hague



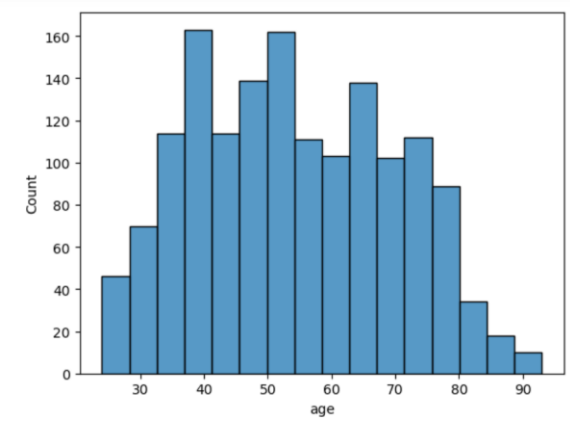


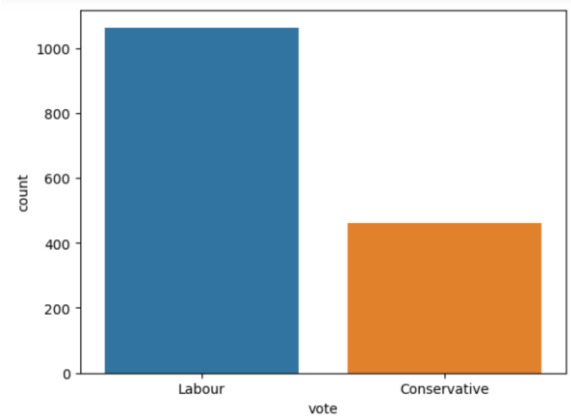
4. Boxplot of Europe & political knowledge



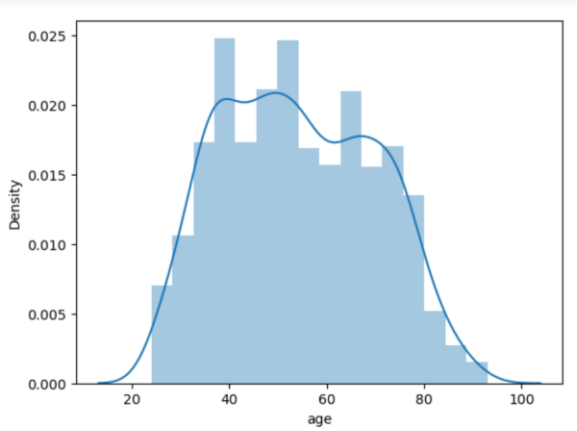


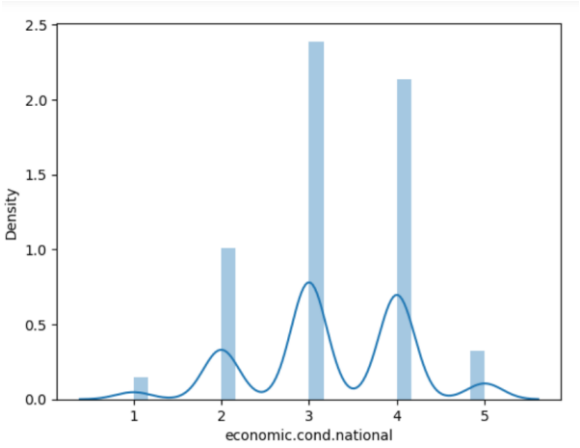
5. Displot of Age & countplot of vote



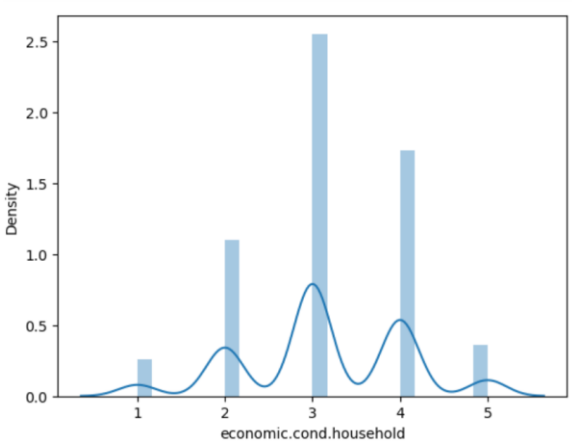


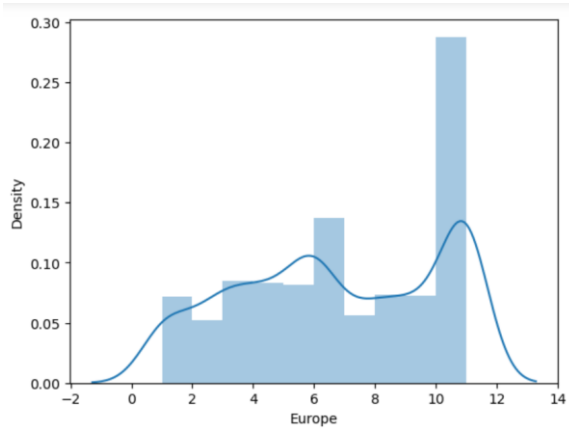
6. Displot of Age& Econoic cond national



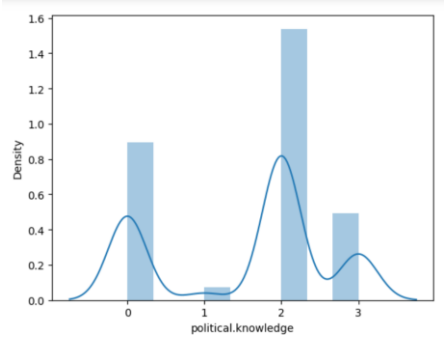


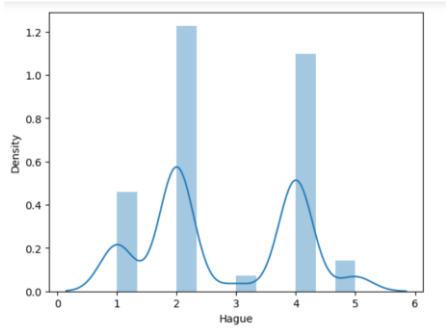
7. Displot of economic cond household & Europe





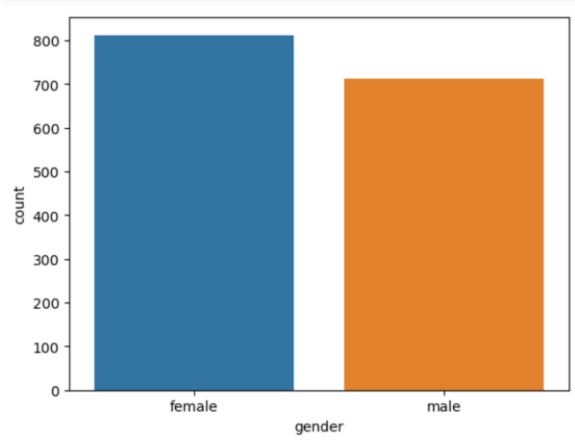
8. Distplot of ploitical knowledge & Hague

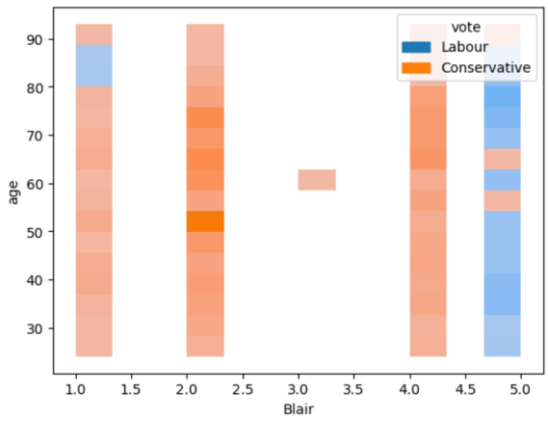


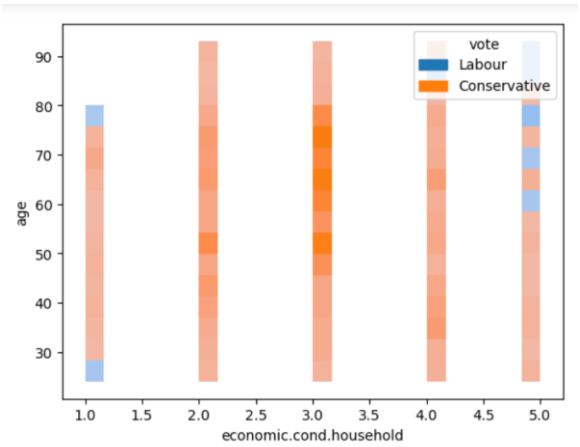


Bivariate Analysis

9. Countplot of gender & Blair

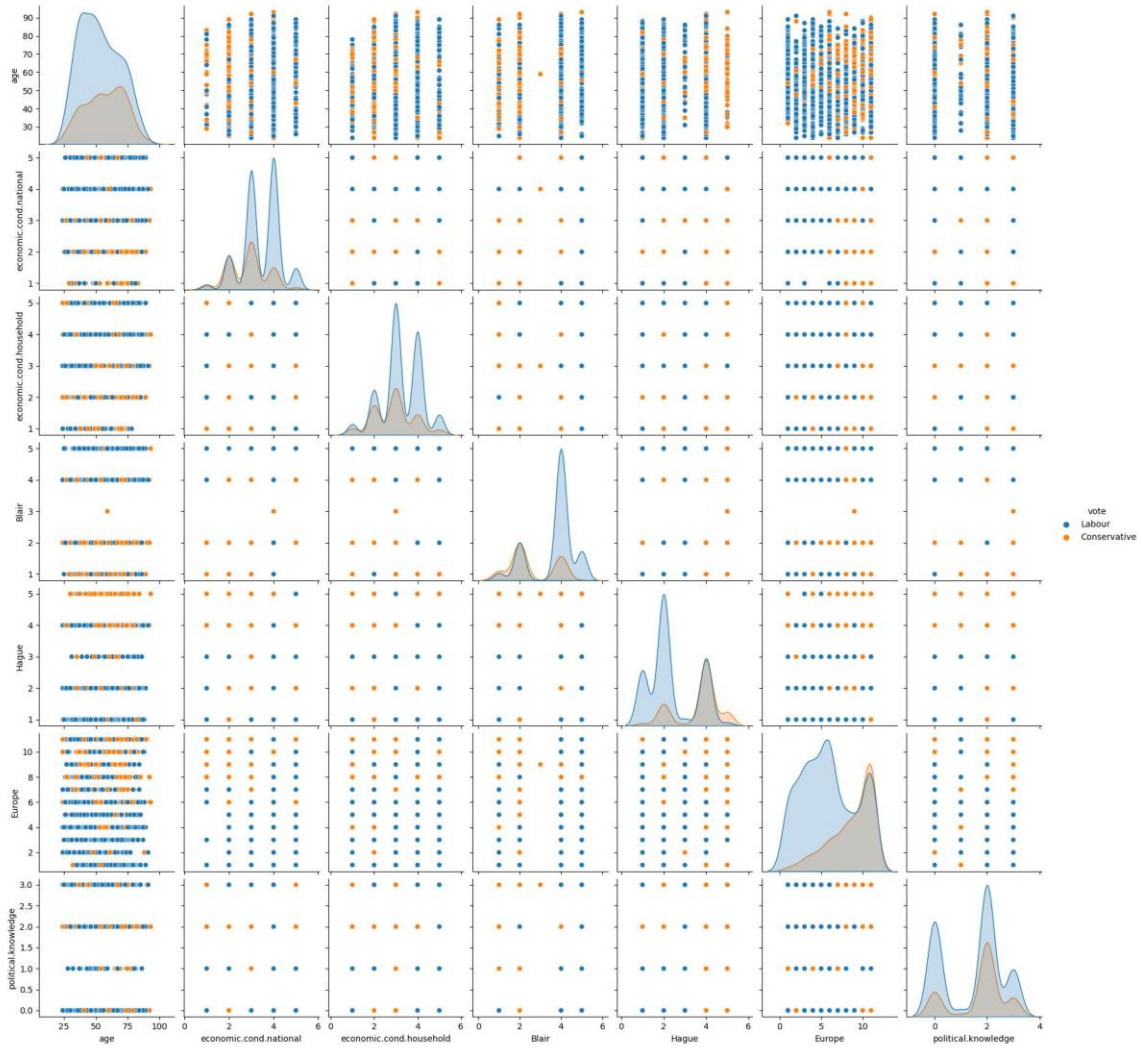






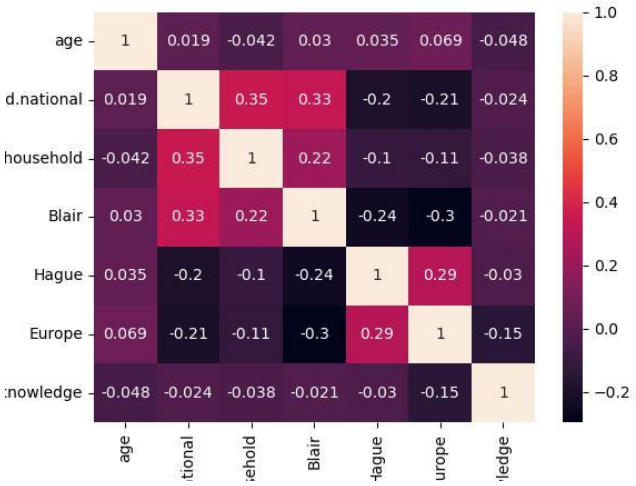
Pair plot

10. Pairplot



Heat map plot

11. Heat map plot



Key Observations:

The dataset contains 1525 rows and 9 columns. There are no missing values in the dataset.

There are few outliers in the column - economic.cond.national, economic.cond.household

The majority of the voters are between the ages of 30 and 60. The mean age of the voter is 54.

There are slightly more female voters than male voters.

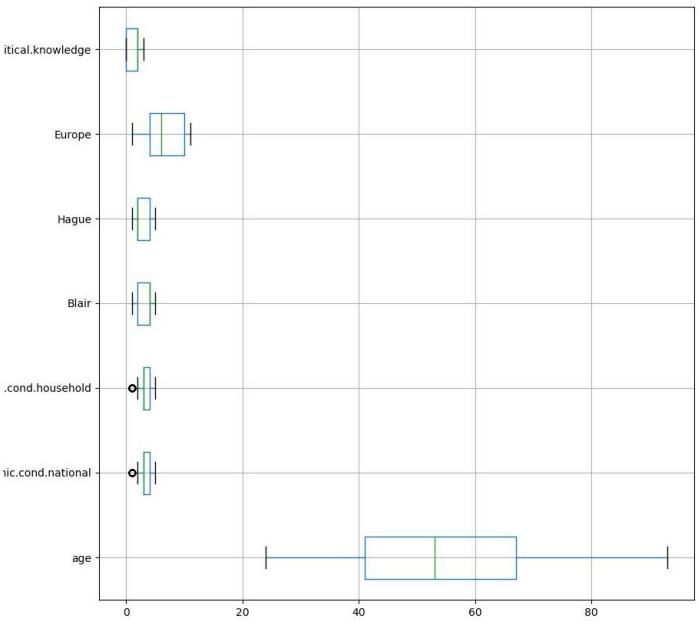
There is a strong correlation between the variables economic.cond.national and economic.cond.household with 0.35 based on the heatmap Correlation matrix. There is a strong correlation between the variables economic.cond.national and Blair with 0.33 based on the heatmap Correlation matrix.

Most of the votes are casted to the labour party than the conservative party.

Outlier Treatment

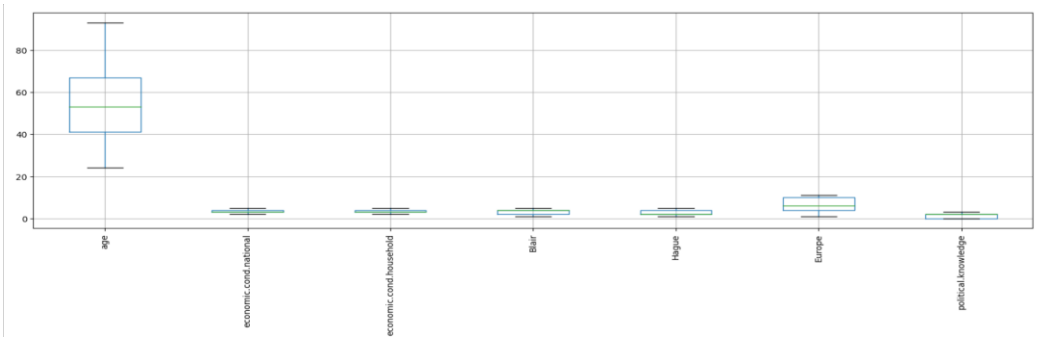
Before Outlier treatment

12. Before outlier treatment



After Outlier treatment

13. After outlier treatment



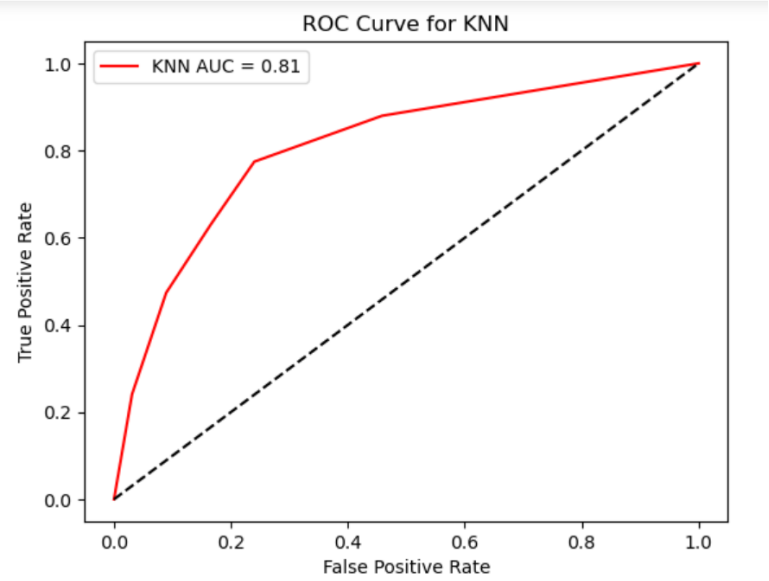
Model Building

KNN & Naïve Bayes

```
KNN Accuracy: 0.7751091703056768
Naive Bayes Accuracy: 0.8144104803493449
KNN Precision: 0.6086956521739131
Naive Bayes Precision: 0.6904761904761905
KNN Recall: 0.631578947368421
Naive Bayes Recall: 0.6541353383458647
KNN F1 Score: 0.6199261992619925
Naive Bayes F1 Score: 0.6718146718146719
```

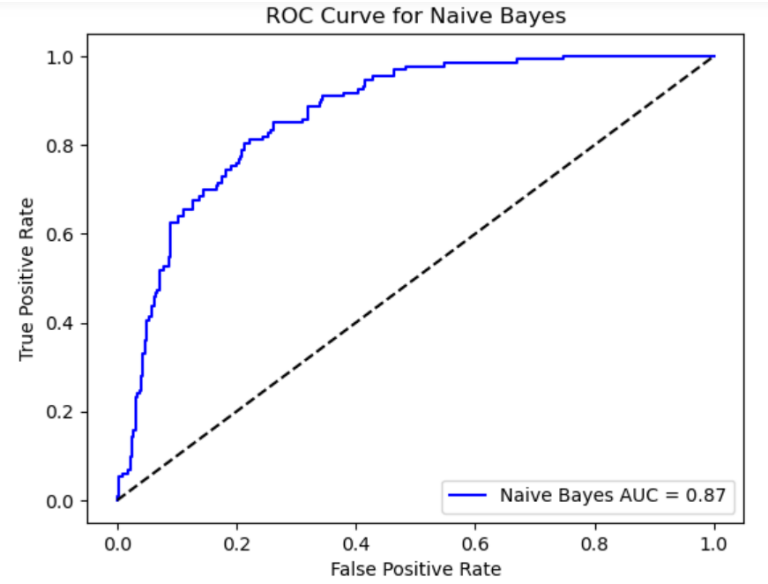
```
Confusion matrix for KNN train dataset:
[[679  59]
 [ 72 257]]
Classification report for KNN train data set:
              precision    recall  f1-score   support

           0       0.90      0.92      0.91       738
           1       0.81      0.78      0.80       329

    accuracy                           0.88      1067
   macro avg       0.86      0.85      0.85      1067
weighted avg       0.88      0.88      0.88      1067

Confusion matrix for KNN test dataset:
[[271  54]
 [ 49  84]]
Classification report for KNN test data set:
              precision    recall  f1-score   support

           0       0.85      0.83      0.84       325
           1       0.61      0.63      0.62       133

    accuracy                           0.78       458
   macro avg       0.73      0.73      0.73       458
weighted avg       0.78      0.78      0.78       458
```

14. Roc curve for KNN



```
Confusion matrix for Naive Bayes train dataset:
[[657  81]
 [ 92 237]]
Classification report for Naive Bayes train dataset:
              precision    recall  f1-score   support

           0       0.88      0.89      0.88       738
           1       0.75      0.72      0.73       329

    accuracy                           0.84      1067
   macro avg       0.81      0.81      0.81      1067
weighted avg       0.84      0.84      0.84      1067

Confusion matrix for Naive Bayes test dataset:
[[286  39]
 [ 46  87]]
Classification report for Naive Bayes test dataset:
              precision    recall  f1-score   support

           0       0.86      0.88      0.87       325
           1       0.69      0.65      0.67       133

    accuracy                           0.81       458
   macro avg       0.78      0.77      0.77       458
weighted avg       0.81      0.81      0.81       458
```

15. ROC curve for Naive Bayes



Bagging & Boosting

```
Bagging Accuracy: 0.8144104803493449
Boosting Accuracy: 0.7729257641921398
Bagging Precision: 0.6818181818181818
Boosting Precision: 0.6013986013986014
Bagging Recall: 0.6766917293233082
Boosting Recall: 0.6466165413533834
Bagging F1 Score: 0.6792452830188679
Boosting F1 Score: 0.6231884057971013
Confusion matrix for Bagging train dataset:
[[738   0]
 [  1 328]]
```

Classification report for Bagging train dataset:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	738
1	1.00	1.00	1.00	329
accuracy			1.00	1067
macro avg	1.00	1.00	1.00	1067
weighted avg	1.00	1.00	1.00	1067

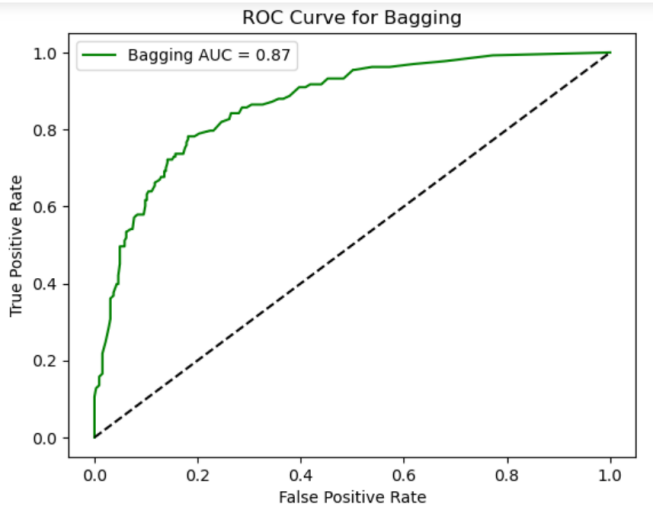
Confusion matrix for Bagging test dataset:

```
[[283  42]
 [ 43  90]]
```

Classification report for Bagging test dataset:

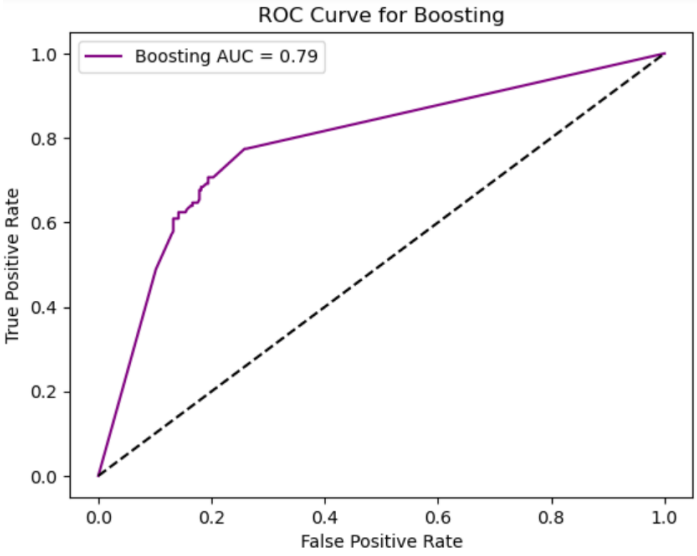
	precision	recall	f1-score	support
0	0.87	0.87	0.87	325
1	0.68	0.68	0.68	133
accuracy			0.81	458
macro avg	0.77	0.77	0.77	458
weighted avg	0.81	0.81	0.81	458

16. ROC curve for Bagging



```
Confusion matrix for Boosting train dataset:
[[738   0]
 [  1 328]]
Classification report for Boosting train dataset:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       738
           1       1.00      1.00      1.00       329

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

Confusion matrix for Boosting test dataset:
[[268  57]
 [ 47  86]]
Classification report for Boosting test dataset:
              precision    recall  f1-score   support

           0       0.85      0.82      0.84       325
           1       0.60      0.65      0.62       133

    accuracy                           0.77       458
   macro avg       0.73      0.74      0.73       458
weighted avg       0.78      0.77      0.78       458
```

17. ROC curve for Boosting



After tuning the bagging & Boosting classifier, we got the results below.

```
Bagging Train Accuracy: 0.9990627928772259
Bagging Train Precision: 1.0
Bagging Train Recall: 0.9969604863221885
Bagging Train F1 Score: 0.9984779299847794
Boosting Train Accuracy: 0.9990627928772259
Boosting Train Precision: 1.0
Boosting Train Recall: 0.9969604863221885
Boosting Train F1 Score: 0.9984779299847794
Bagging Test Accuracy: 0.8209606986899564
Bagging Test Precision: 0.6976744186046512
Bagging Test Recall: 0.6766917293233082
Bagging Test F1 Score: 0.6870229007633587
Boosting Test Accuracy: 0.7838427947598253
Boosting Test Precision: 0.6180555555555556
Boosting Test Recall: 0.6691729323308271
Boosting Test F1 Score: 0.6425992779783394
```

```
Top 3 Models:
- Model: BaggingClassifier(estimator=DecisionTreeClassifier(), n_estimators=100,
                  random_state=42)
  Accuracy: 0.8144104803493449
  Precision: 0.6818181818181818
  Recall: 0.6766917293233082
  F1 Score: 0.6792452830188679
  ROC AUC: 0.8669404279930595
- Model: GaussianNB()
  Accuracy: 0.8144104803493449
  Precision: 0.6904761904761905
  Recall: 0.6541353383458647
  F1 Score: 0.6718146718146719
  ROC AUC: 0.8667669172932331
- Model: AdaBoostClassifier(estimator=DecisionTreeClassifier(), n_estimators=100,
                   random_state=42)
  Accuracy: 0.7729257641921398
  Precision: 0.6013986013986014
  Recall: 0.6466165413533834
  F1 Score: 0.6231884057971013
  ROC AUC: 0.7859919028340082
```

From the above code, we can see that the top 3 models are: Bagging classifier, Gaussian Naive Bayes and adaboosting classifier. We can choose either Bagging classifier or Gaussian Naive Bayes, due to their accuracy, precision, Recall, F1 score and ROC AUC.
The most important features in the boosting model are:

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

Suggestions:

1. The Conservative Party should focus on improving the national and household economic conditions to increase their chances of winning the vote.
2. The Labour Party should focus on maintaining their strong support among voters who are satisfied with the current economic conditions and who have a positive view of their leader, Blair.

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

Code Snippet to extract the three speeches:

Problem 2- Define the problem and Perform Exploratory Data Analysis

-Problem Definition- Find the number of Characters, words & sentences in all three speeches

```
Speech 1:
Total characters: 7571
Total words: 1360
Total sentences: 68

Speech 2:
Total characters: 7618
Total words: 1390
Total sentences: 57

Speech 3:
Total characters: 9991
Total words: 1819
Total sentences: 72
```

Problem 2- Text cleaning

- Stopword removal- Stemming- find the 3 most common words used in all three speeches

18. Three Common words in speeches

```
Three most common words in speech 1: [('nation', 17), ('it', 14), ('the', 10)]
Three most common words in speech 2: [('let', 16), ('us', 12), ('power', 9)]
Three most common words in speech 3: [('us', 26), ('let', 22), ('america', 21)]
```

Problem 2- Plot Word cloud of all three speeches

- Show the most common words used in all three speeches in the form of word clouds

19. Word cloud