Machine Leaning Project

Authored by: Shrinidhi CG

Contents

Problem Statement:

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.

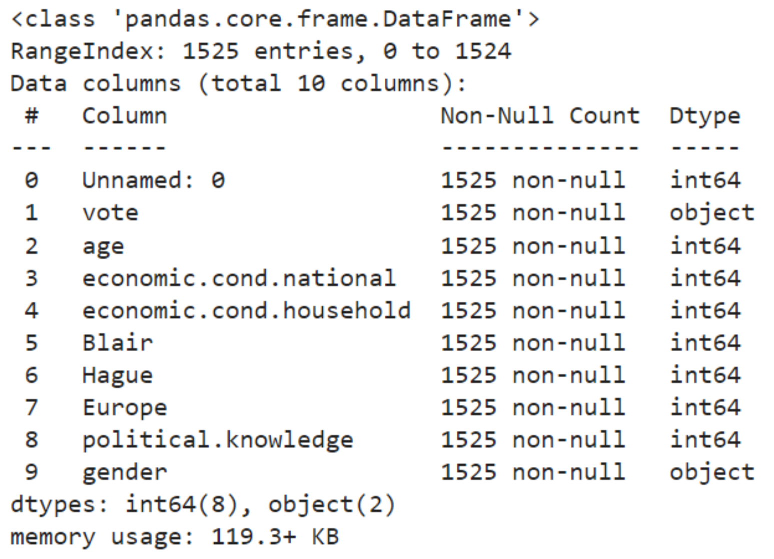
Check shape, Data types, and statistical summary

Shape*: Dataset has 1525 rows and 10 columns.*

no. of rows: 1525

no. of columns: 10

Datatypes:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               1525 non-null   int64 
 1   vote                     1525 non-null   object
 2   age                      1525 non-null   int64 
 3   economic.cond.national   1525 non-null   int64 
 4   economic.cond.household  1525 non-null   int64 
 5   Blair                    1525 non-null   int64 
 6   Hague                    1525 non-null   int64 
 7   Europe                   1525 non-null   int64 
 8   political.knowledge      1525 non-null   int64 
 9   gender                   1525 non-null   object
dtypes: int64(8), object(2)
memory usage: 119.3+ KB
```

Unnamed: *An unnamed column, possibly an index or identifier.*

vote*: Categorical variable representing the party choice (e.g., Conservative or Labour).*

age*: Numerical variable representing the age in years.*

economic.cond.national: *Numerical variable representing the assessment of current national economic conditions (on a scale of 1 to 5).*

economic.cond.household: *Numerical variable representing the assessment of current household economic conditions (on a scale of 1 to 5).*

***Blair**: Numerical variable representing the assessment of the Labour leader (on a scale of 1 to 5).*

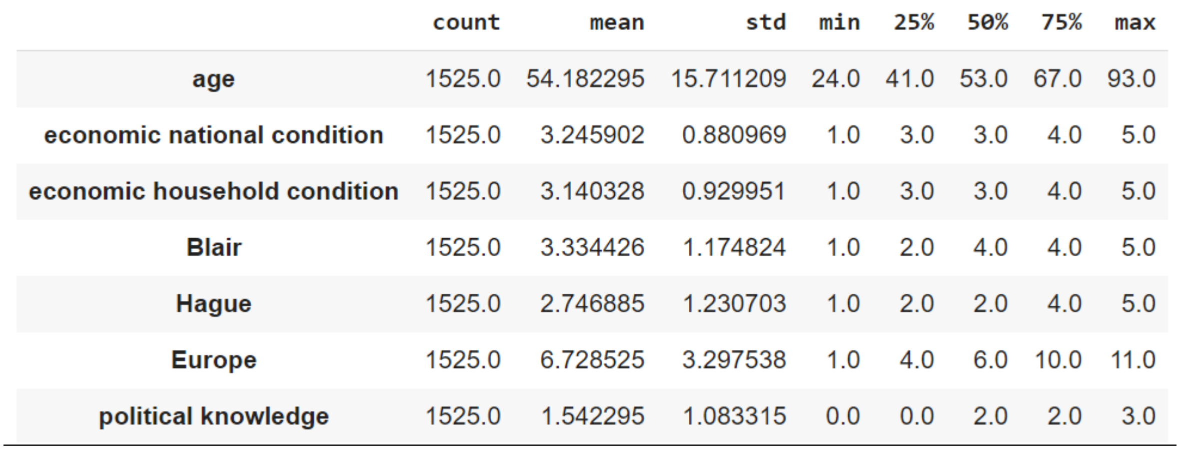
***Hague**: Numerical variable representing the assessment of the Conservative leader (on a scale of 1 to 5).*

***Europe**: Numerical variable representing attitudes toward European integration on an 11-point scale.*

***political.knowledge:** Numerical variable representing knowledge of parties' positions on European integration (on a scale of 0 to 3).*

***gender**: Categorical variable representing gender (e.g., female or male).*

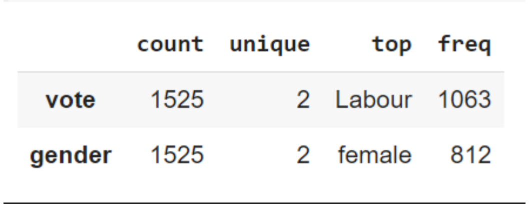
Statistical summary:



	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 national condition	1525.0	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic household condition	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

Insights:

- *The age column has a wide range (24 to 93 years) with a mean of approximately 54 years.*
- *economic.cond.national and economic.cond.household columns represent assessments of economic conditions and have means around 3, indicating a moderate assessment.*
- *The columns Blair and Hague represent assessments of political leaders and show a range from 1 to 5.*
- *Europe column represents attitudes toward European integration on an 11-point scale, with a mean around 6.7.*
- *political.knowledge column represents knowledge of parties' positions on European integration, with a mean around 1.54.*

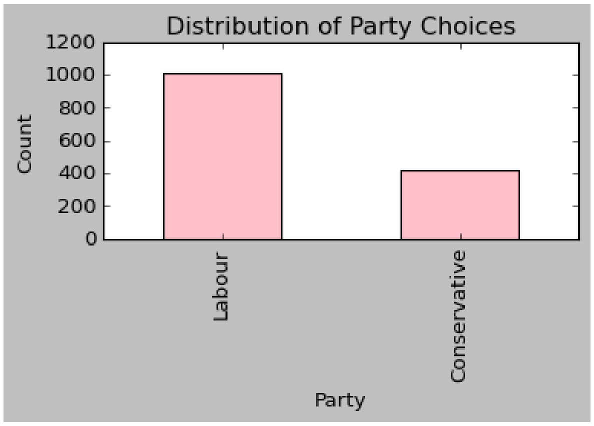


	count	unique	top	freq
vote	1525	2	Labour	1063
gender	1525	2	female	812

Insights:

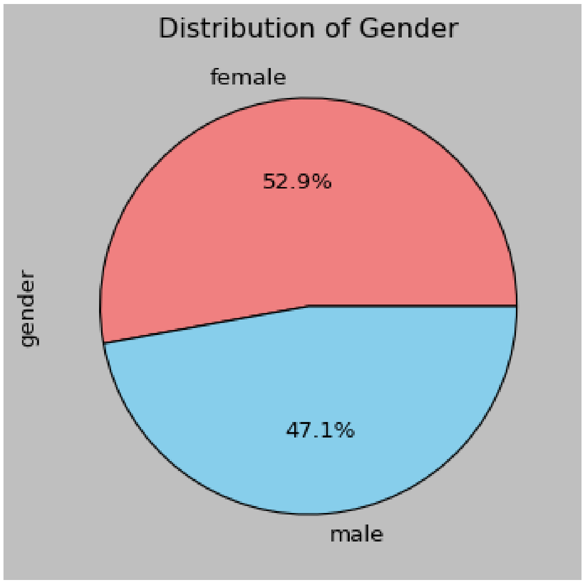
- *The gender column has 1525 non-null entries. The most common gender is female, with a frequency of 812.*
- *There are 2 unique values in the vote column, indicating a binary choice between two options. The most common vote is for the Labour party, with a frequency of 1063.*

Univariate analysis



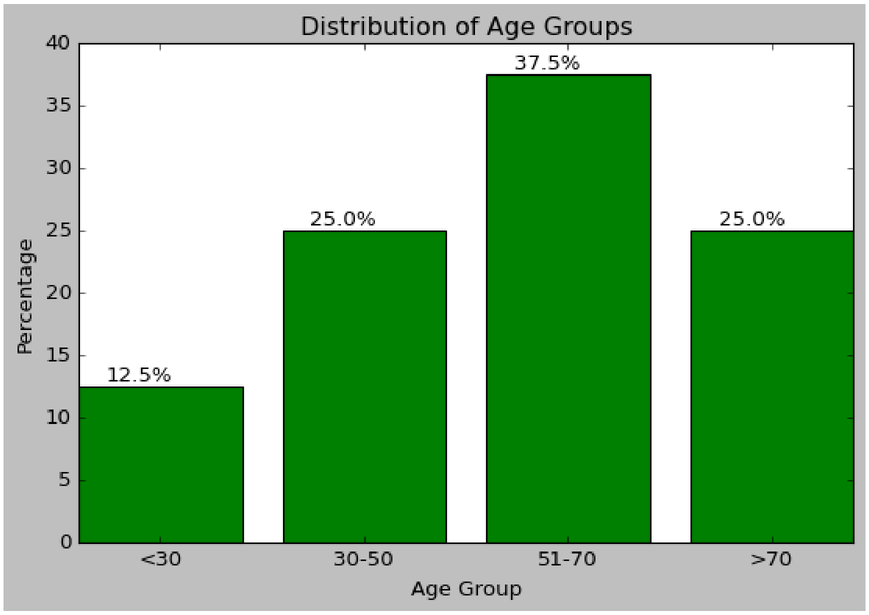
Insights

- *The majority of respondents (1011 instances) chose 'Labor' as their preferred party .'Conservative' received fewer votes, with 419 instances.*
- *There is an imbalance in the distribution of votes, with 'Labour' having a significantly higher count than 'Conservative.'*
- *The higher count for 'Labour' may suggest broader support or popularity for that party among the surveyed individuals. The lower count for 'Conservative' may indicate a smaller portion of respondents favoring that party.*



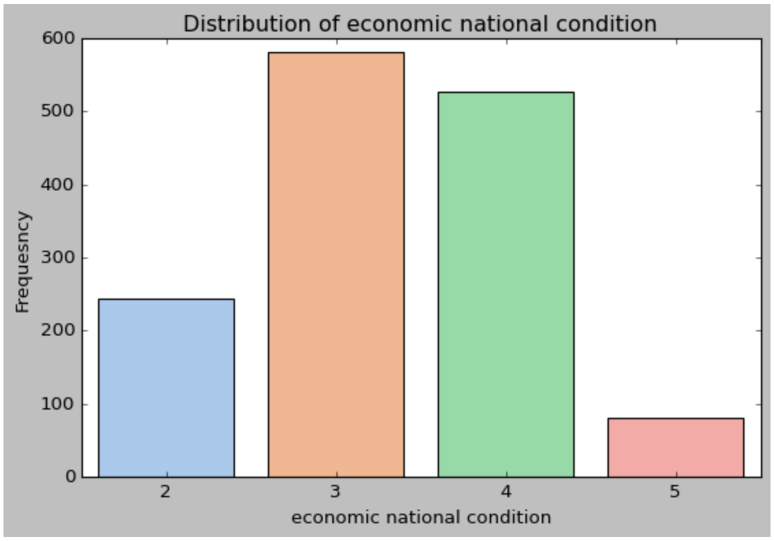
Insights

- *The dataset includes responses from both females and males. The count for females is 756, while the count for males is 674.*
- *The distribution is relatively balanced, with a slightly higher count for females compared to males. The gender distribution suggests a representation of both genders in the surveyed population.*



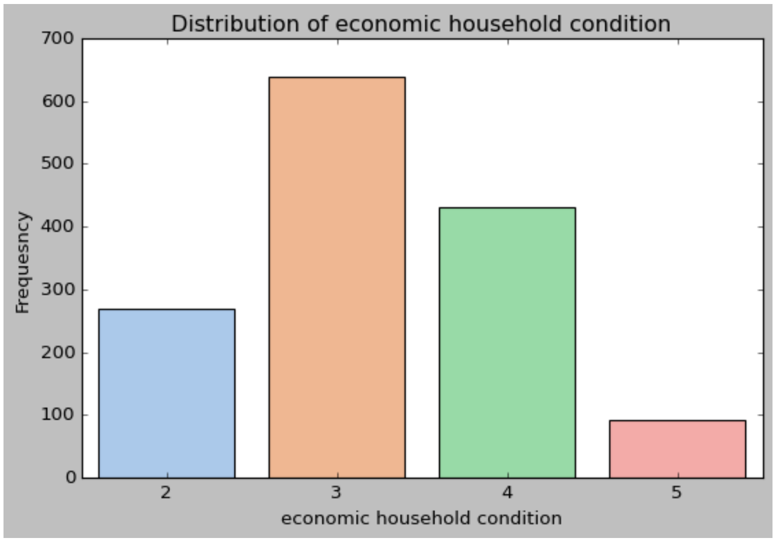
Insights

- *The majority of respondents fall into the '51-70' age group, constituting 37.5% of the total. The second most common age group is '30-50' with a percentage of 25%. '51-70' is followed closely by '<30' and '>70', both with a percentage of 12.5%.*
- *The survey respondents are relatively evenly distributed across different age groups. A notable portion falls into the middle age group ('51-70'), indicating a substantial representation from a demographic typically associated with significant life experiences and responsibilities.*



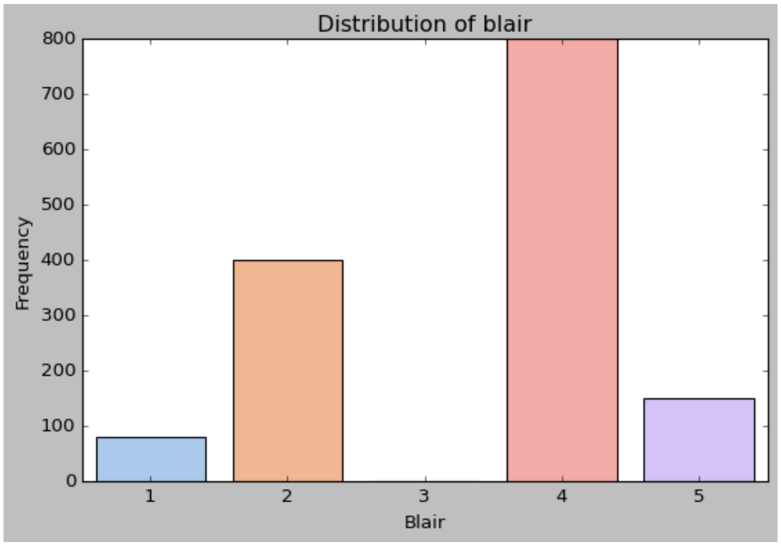
Insights

- *The most common response for the assessment of the national economic condition is '3', with 581 occurrences. The second most common response is '4', with 526 occurrences. '2' is the third most common response, with 243 occurrences. The least common response is '5', with 80 occurrences. Implications:*
- *A substantial number of respondents perceive the national economic condition to be moderate ('3') or relatively positive ('4'). There is a smaller group of respondents who perceive the condition as either very positive ('5') or somewhat negative ('2').*
- *The distribution suggests that the majority of respondents have a moderate to positive view of the national economic condition.*



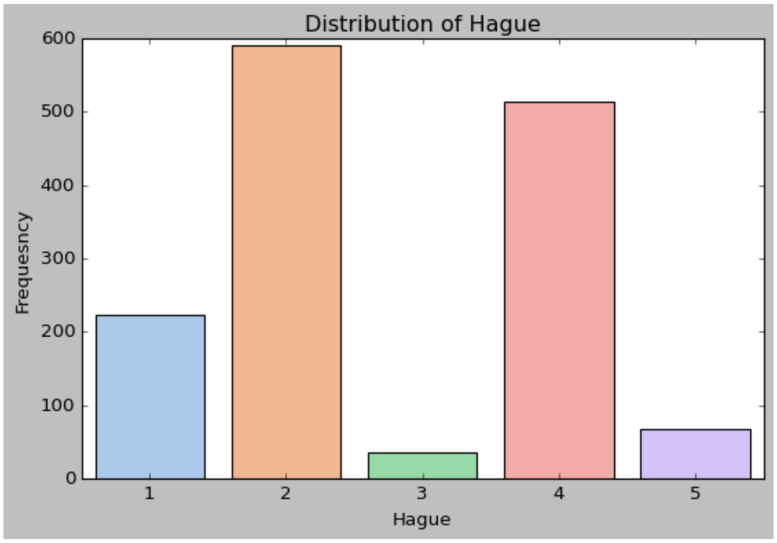
Insights

- *The most common response for the assessment of the household economic condition is '3', with 638 occurrences. The second most common response is '4', with 431 occurrences. '2' is the third most common response, with 270 occurrences. The least common response is '5', with 91 occurrences.*
- *Comparing with the national economic condition, respondents seem to be slightly more optimistic about their household economic condition, as indicated by a higher count for the responses '3' and '4'.*
- *Similar to the national economic condition, the majority of respondents have a moderate to positive view of their household economic condition. While '3' and '4' are the predominant responses, there is still variability in responses, suggesting a diversity of economic perceptions among respondents.*



Insights

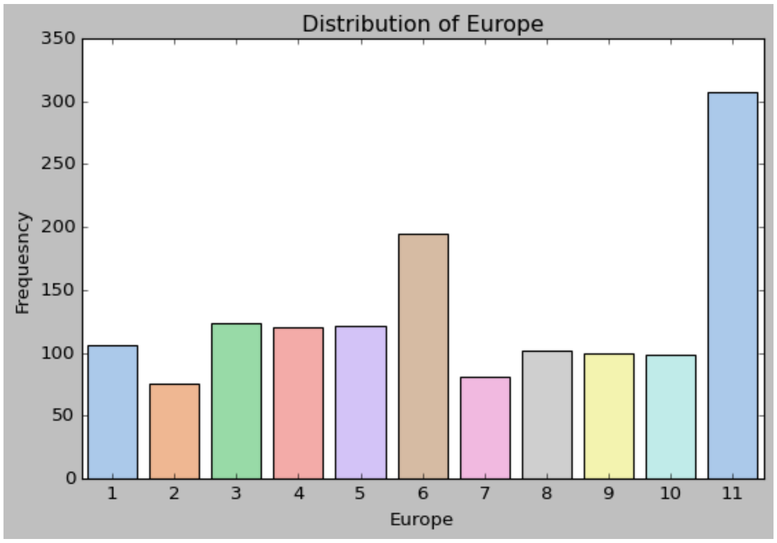
- *The majority of respondents (799 occurrences) gave a rating of '4,' indicating a prevalent positive assessment of the Labour party leadership.*
- *While '4' is the most common rating, there is notable variability in responses. Some respondents provided lower ratings ('2' and '1'), suggesting diverse opinions on the leadership quality.*
- *Ratings '2' and '5' received moderate frequencies (400 and 150 occurrences, respectively). This suggests a mix of less positive and highly positive assessments.*
- *The rating '1' received 80 occurrences, indicating a relatively low count of respondents with a very negative assessment of the Labour party leadership.*
- *The extreme ratings '1' and '5' combined constitute 230 occurrences, representing respondents with more polarized opinions.*
- *The neutral rating '3' occurred only once, suggesting that a very small proportion of respondents had a neutral stance on the leadership.*



Insights

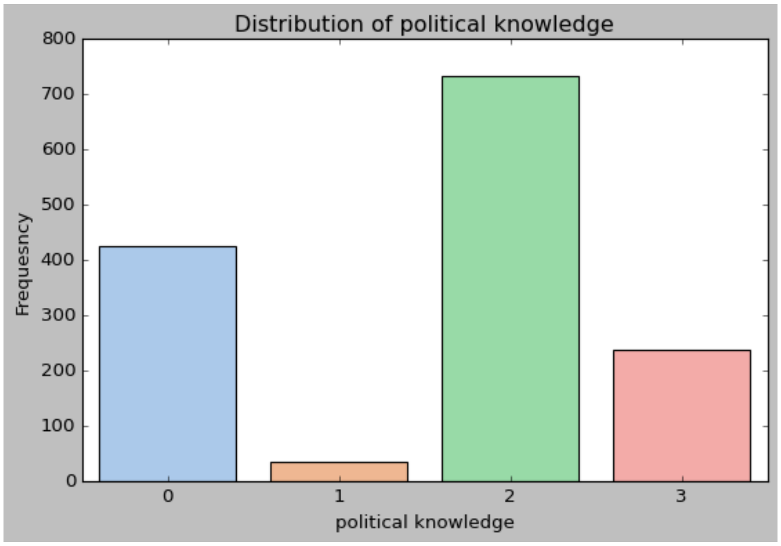
- *The most common assessment of the Conservative leader, Hague, is '2', with 590 occurrences. The second most common assessment is '4', with 514 occurrences. '1' is the third most common assessment, with 224 occurrences. '5' has 67 occurrences, and '3' has 35 occurrences.*
- *The assessments for William Hague are distributed across the range of the scale, with '2' being the mode, indicating a central tendency towards a moderate assessment.*

- *The combined occurrences for ratings '2' and '4' suggest that respondents generally provided moderate to positive assessments of William Hague's leadership.*
- *The occurrences for ratings '1' and '3' suggest a lower count of respondents with negative assessments.*
- *The rating '5' received a relatively lower count (67 occurrences), suggesting a smaller proportion of respondents with an extremely positive assessment.*
- *The distribution indicates diversity in respondents' opinions, with assessments spread across different levels of positivity.*



Insights

- *The most common attitude towards European integration is represented by the rating '11', with 307 occurrences. The second most common attitude is '6', with 195 occurrences. Ratings '3', '5', and '4' have similar frequencies, with 124, 121, and 120 occurrences, respectively.*
- *The most common attitude towards European integration is represented by the rating '11', with 307 occurrences. The second most common attitude is '6', with 195 occurrences. Ratings '3', '5', and '4' have similar frequencies, with 124, 121, and 120 occurrences, respectively.*
- *There is a noticeable polarization in attitudes, with ratings at both extremes ('1' and '11') having relatively high occurrences.*
- *Segmenting the respondents based on their attitudes towards Europe may reveal patterns in opinions and preferences.*



Insights

- *The most common level of political knowledge is represented by the rating '2', with 733 occurrences. The second most common level is '0', with 425 occurrences. Ratings '3' and '1' have lower frequencies, with 237 and 35 occurrences, respectively.*
- *The distribution suggests that a significant number of respondents have a moderate level of political knowledge, as indicated by the higher count for rating '2'.*
- *The occurrences for ratings '0' and '3' represent respondents with lower and higher levels of political knowledge, respectively.*

Multivariate analysis

Cross Tabulation for National Economic Condition vs. Household Economic Condition:

Nations vs household condition	2	3	4	5
2	105	91	42	5
3	101	320	134	26
4	57	198	232	39
5	7	29	23	21

- *The cross-tabulation shows the distribution of respondents based on their assessments of both national and household economic conditions.*

- *The majority of respondents who rated the national economic condition as '2' also rated the household economic condition as '2'.*
- *There is a general trend of respondents giving similar ratings to both national and household economic conditions, with higher ratings correlating.*

Cross Tabulation for Blair vs. Political Knowledge:

political knowledge vs Blair	*0*	*1*	*2*	*3*
1	*17*	*2*	*44*	*17*
2	*121*	*12*	*207*	*60*
3	*0*	*0*	*0*	*1*
4	*239*	*20*	*407*	*133*
5	*48*	*1*	*75*	*26*

- *The cross-tabulation presents the distribution of respondents based on their ratings of Blair and their level of political knowledge.*
- *Respondents with higher political knowledge (ratings '2' and '3') tend to give higher ratings to Blair, while those with lower political knowledge (rating '0') have more varied opinions.*

Cross Tabulation for Hague vs. Political Knowledge:

political knowledge vs Hague	*0*	*1*	*2*	*3*
1	*62*	*4*	*100*	*58*
2	*177*	*13*	*307*	*93*
3	*12*	*4*	*13*	*6*
4	*158*	*11*	*277*	*68*
5	*16*	*3*	*36*	*12*

- *The cross-tabulation illustrates the relationship between respondents' ratings of Hague and their political knowledge levels.*

- *Similar to Blair, respondents with higher political knowledge (ratings '2' and '3') tend to give higher ratings to Hague.*

Cross Tabulation for Gender vs. Vote:

Vote vs Gender	Conservative	Labour
Female	135	521
Male	184	490

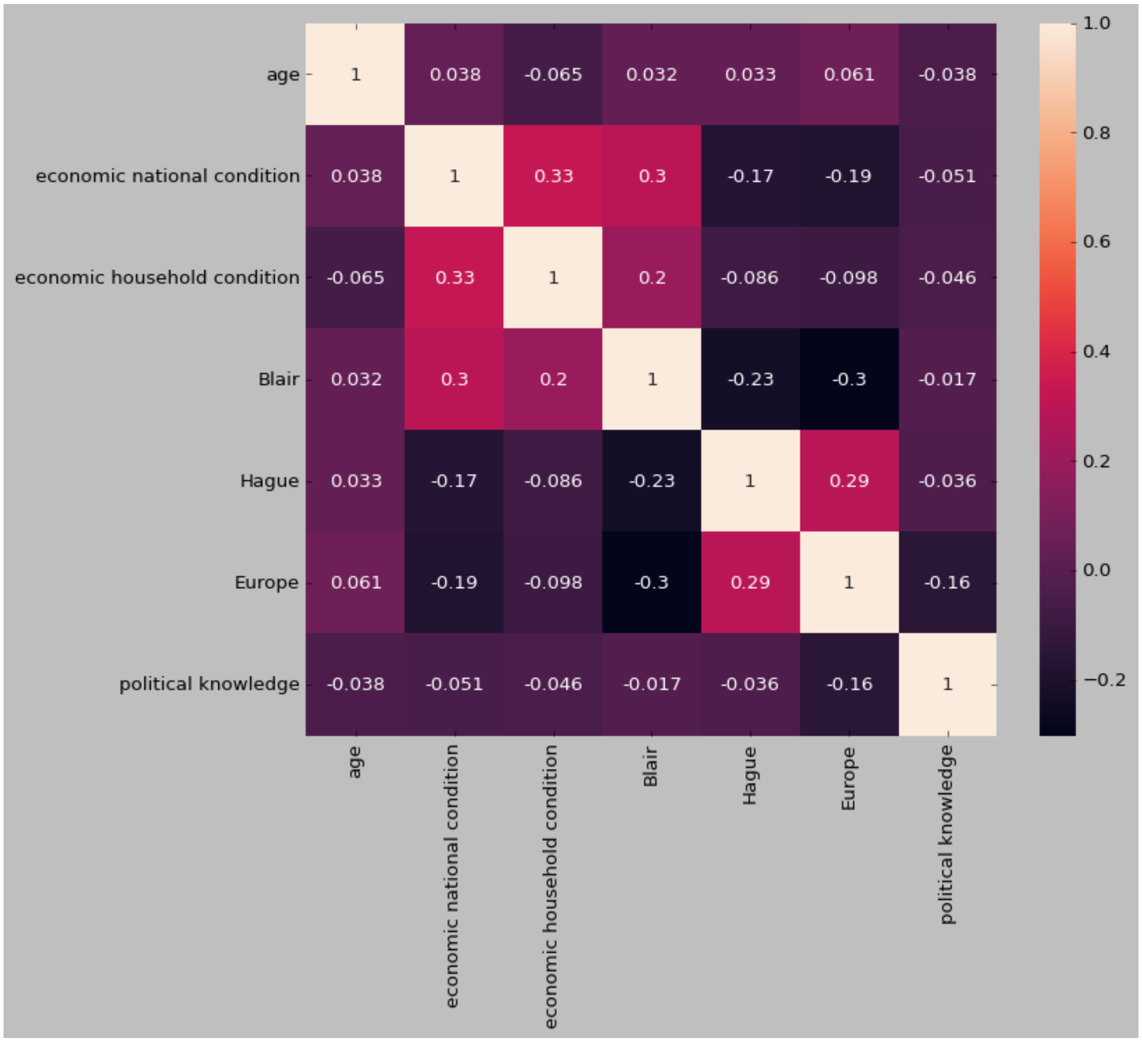
- *The cross-tabulation shows the distribution of votes ('Conservative' or 'Labour') based on respondents' gender.*
- *The majority of both females and males voted for 'Labour,' with slightly more females supporting 'Labour' compared to males.*

Cross Tabulation for Political Knowledge vs. Europe:

Europe vs Political Knowledge	1	2	3	4	5	6	7	8	9	10	11
0	26	7	28	34	28	67	22	36	38	29	110
1	3	1	1	3	1	2	5	4	0	3	12
2	59	43	59	57	59	105	46	48	51	52	154
3	18	25	36	26	33	21	8	14	11	14	31

- *The cross-tabulation displays the distribution of respondents based on their ratings of political knowledge and attitudes towards European integration.*
- *Respondents with higher political knowledge (ratings '2' and '3') show diverse attitudes towards Europe, while those with lower political knowledge (rating '0') have varying attitudes.*

Correlation plot



Age vs. Economic Conditions:

- *Age vs. Economic Condition (National): The correlation is positive but very weak (0.038). This indicates a slight tendency for older respondents to perceive the national economic condition more positively, though the relationship is not strong.*
- *Age vs. Economic Condition (Household): The correlation is negative but weak (-0.065). This suggests a minor inclination for older respondents to perceive their household economic condition less favorably.*

Age vs. Political Beliefs and Knowledge:

- *Age vs. Blair (Assessment of Labour leader): The correlation is positive but weak (0.032), indicating a subtle tendency for older individuals to give a more positive assessment of the Labour leader.*
- *Age vs. Hague (Assessment of Conservative leader): Similar to Blair, the correlation is positive but weak (0.033), suggesting a mild inclination for older individuals to give a more positive assessment of the Conservative leader.*
- *Age vs. Europe (Attitude towards European integration): The correlation is positive but weak (0.061), implying a slight tendency for older individuals to have a more positive attitude towards European integration.*
- *Age vs. Political Knowledge: The correlation is negative but weak (-0.038), indicating a minor tendency for older individuals to have a slightly lower level of political knowledge.*

Economic Conditions:

- *National vs. Household Economic Conditions: The correlation is positive and moderate (0.33). This suggests that respondents who perceive the national economic condition more positively are also likely to perceive their household economic condition more positively.*

Economic Conditions vs. Political Beliefs and Knowledge:

- *National Economic Condition vs. Blair: The correlation is positive and moderate (0.3), implying a connection between a positive perception of the national economic condition and a positive assessment of the Labour leader.*
- *National Economic Condition vs. Hague: The correlation is negative and moderate (-0.17), suggesting a tendency for those with a more positive view of the national economic condition to have a less positive assessment of the Conservative leader.*
- *National Economic Condition vs. Europe: The correlation is negative and moderate (-0.19), indicating a connection between a positive view of the national economic condition and a less positive attitude towards European integration.*
- *National Economic Condition vs. Political Knowledge: The correlation is negative but weak (-0.051), implying a minor tendency for those with a more positive view of the national economic condition to have a slightly lower level of political knowledge.*

Blair vs. Political Beliefs and Knowledge:

- *Blair vs. Hague: The correlation is negative and moderate (-0.23), suggesting an association between a positive assessment of the Labour leader (Blair) and a less positive assessment of the Conservative leader (Hague).*
- *Blair vs. Europe: The correlation is negative and strong (-0.3), indicating a significant connection between a positive assessment of Blair and a less positive attitude towards European integration.*
- *Blair vs. Political Knowledge: The correlation is negative and very weak (-0.017), suggesting a subtle tendency for those with a more positive assessment of Blair to have a slightly lower level of political knowledge.*

Hague vs. Political Beliefs and Knowledge:

- *Hague vs. Europe: The correlation is positive and strong (0.29), indicating a significant connection between a positive assessment of the Conservative leader (Hague) and a more positive attitude towards European integration.*
- *Hague vs. Political Knowledge: The correlation is negative and very weak (-0.036), suggesting a subtle tendency for those with a more positive assessment of Hague to have a slightly lower level of political knowledge.*

Europe vs. Political Knowledge:

- *Europe vs. Political Knowledge: The correlation is negative and very weak (-0.016), implying a subtle tendency for those with a more positive attitude towards European integration to have a slightly lower level of political knowledge.*

Key meaningful observations

Age and Economic Conditions:

While there are weak correlations, age alone does not strongly predict perceptions of economic conditions. The influence of age on economic perceptions appears to be subtle.

***Political Assessments and Age**:*

Older individuals show a slight inclination to give more positive assessments of political leaders (Blair, Hague) and exhibit a marginally more positive attitude towards European integration.

Economic Conditions and Political Assessments:

Respondents who perceive the national economic condition more positively tend to give a more positive assessment of the Labour leader (Blair) but a less positive assessment of the Conservative leader (Hague).

Political Assessments and Attitude towards Europe:

There is a strong negative correlation between a positive assessment of Blair and a positive attitude towards European integration. Conversely, a more positive assessment of the Conservative leader (Hague) is correlated with a more positive attitude towards European integration.

Gender and Voting Preferences:

The dataset is slightly imbalanced in terms of gender, with more female respondents. However, this imbalance does not seem to strongly influence voting preferences, as the distribution of votes (Labour vs. Conservative) is relatively similar between genders.

Political Knowledge:

There is a weak negative correlation between age and political knowledge, indicating that older individuals may have slightly lower levels of political knowledge. However, this relationship is not significant.

Economic Conditions and Political Knowledge:

Respondents with more positive views of economic conditions, both national and household, tend to have a slightly lower level of political knowledge. This suggests a potential trade-off between economic optimism and political knowledge.

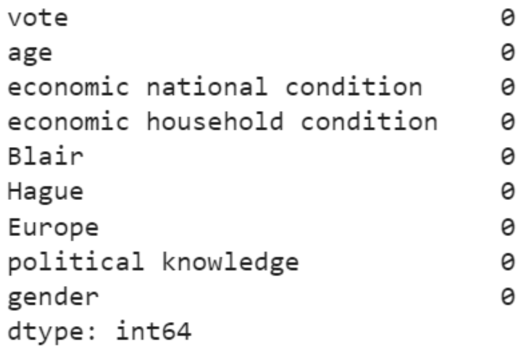
Perception of Conservative Leader (Hague) and Europe:

Respondents with a more positive assessment of the Conservative leader (Hague) are more likely to have a positive attitude towards European integration.

Data Pre-processing: Prepare the data for modelling:

- *Removing the unwanted columns and renaming the column names 'economic national condition', 'economic household condition'*

- *Are there any missing values?*



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

There are no missing values in the dataset.

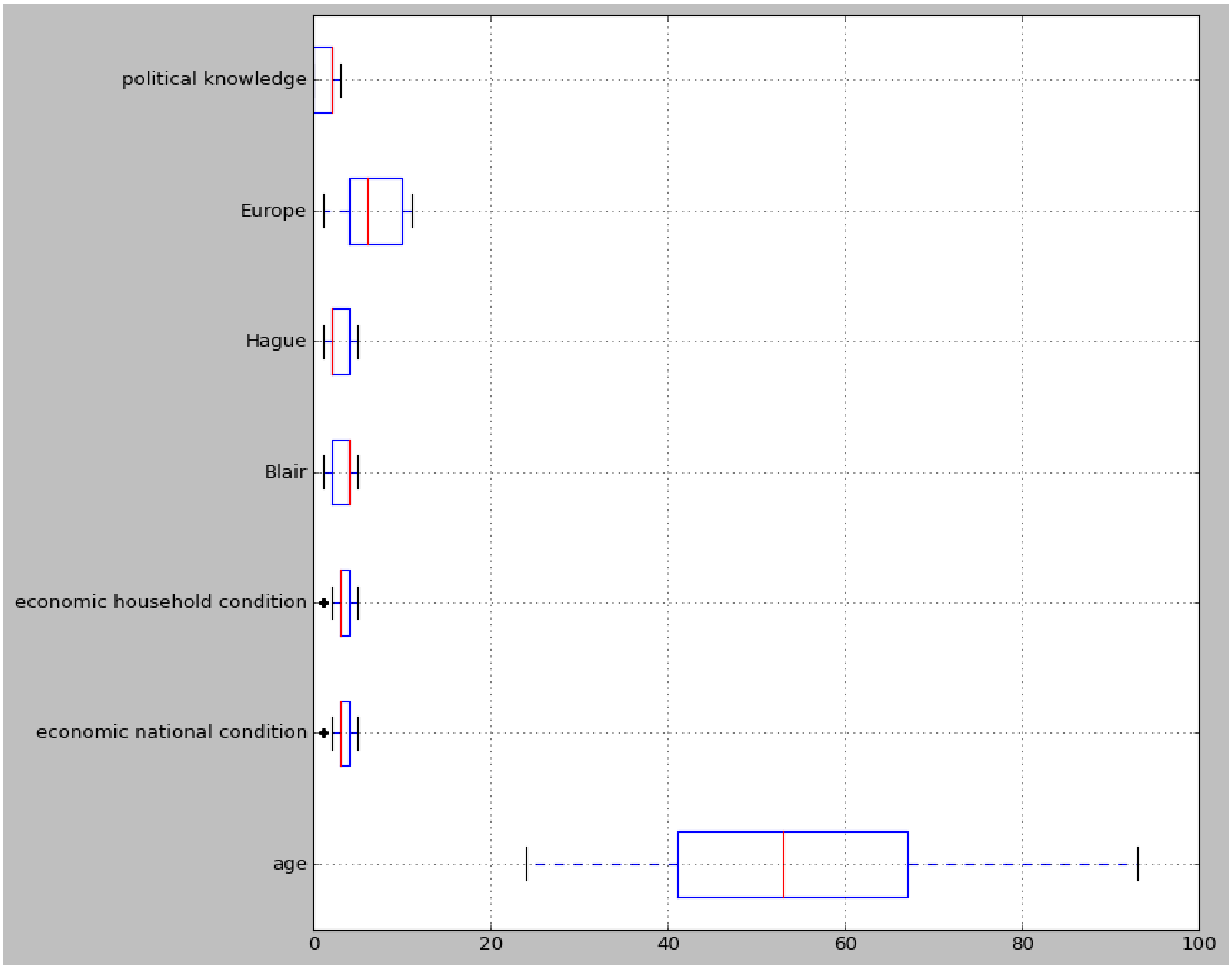
- *Are there any duplicate records?*

Number of duplicate rows = 8

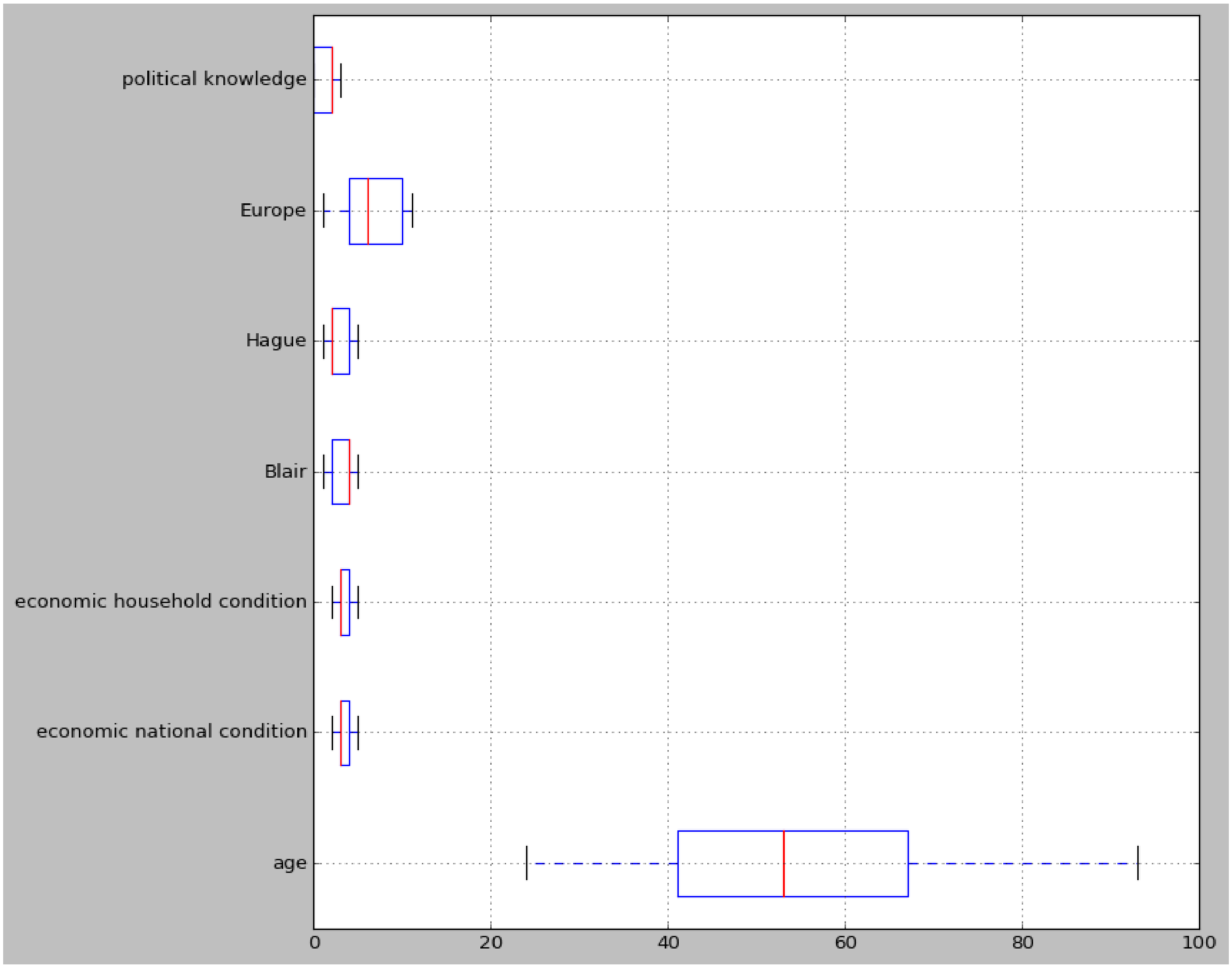
Dropping duplicate values in a dataset is often done to ensure data quality and prevent potential biases or inaccuracies in analysis.

Outlier Detection

Before treating outliers

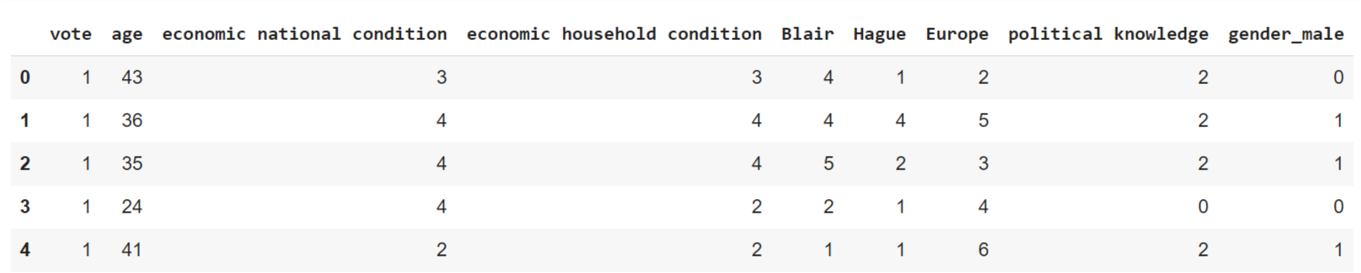


After Treating outliers



Encode the data

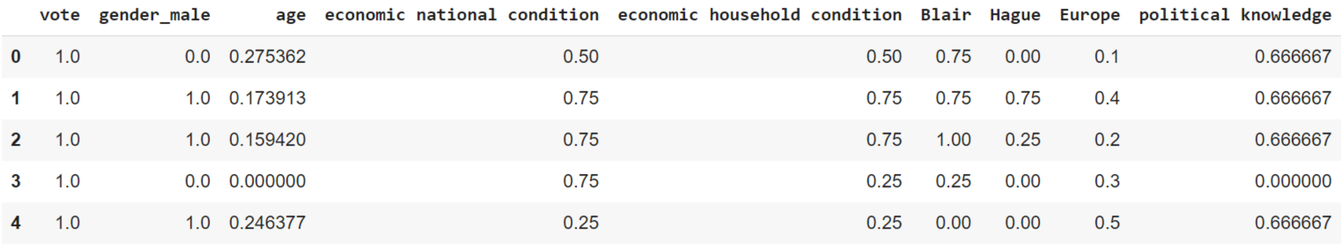
- *Encoding the dummy variable for object variables. Encoding the dummy variable for object variables*
- *Encoding the Vote variable as binary variable for model.*



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

Scale the data

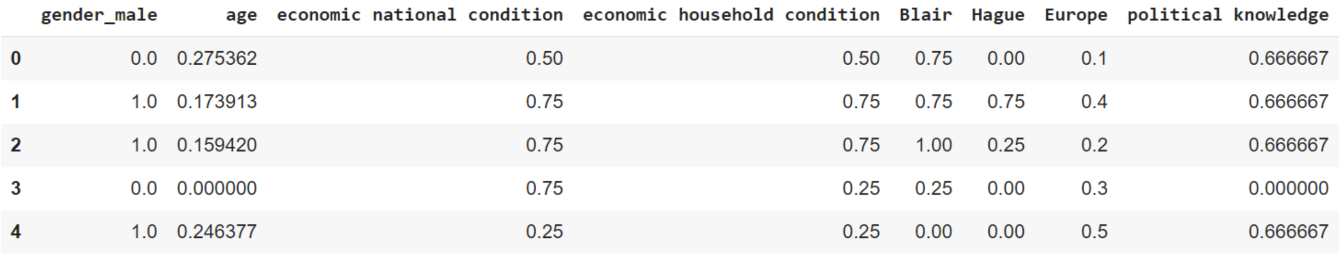
- *Scaling the variables as continuous variables have different weightage using min-max technique*



	vote	gender_male	age	economic national condition	economic household condition	Blair	Hague	Europe	political knowledge
0	1.0	0.0	0.275362	0.50	0.50	0.75	0.00	0.1	0.666667
1	1.0	1.0	0.173913	0.75	0.75	0.75	0.75	0.4	0.666667
2	1.0	1.0	0.159420	0.75	0.75	1.00	0.25	0.2	0.666667
3	1.0	0.0	0.000000	0.75	0.25	0.25	0.00	0.3	0.000000
4	1.0	1.0	0.246377	0.25	0.25	0.00	0.00	0.5	0.666667

Data split

- *Split X and y into training and test set in 75:25 ratio*

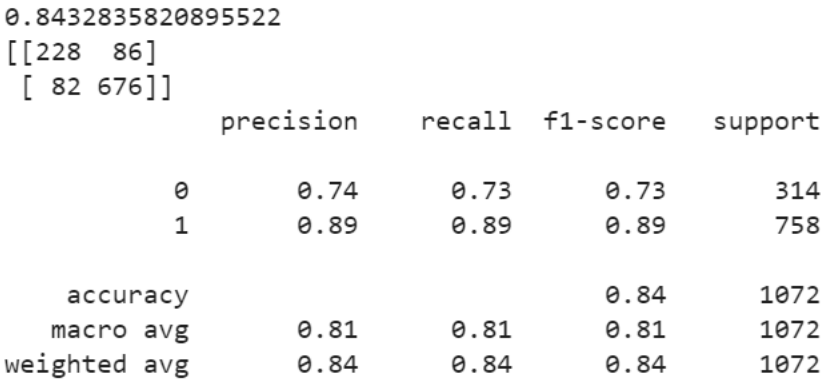


	gender_male	age	economic national condition	economic household condition	Blair	Hague	Europe	political knowledge
0	0.0	0.275362	0.50	0.50	0.75	0.00	0.1	0.666667
1	1.0	0.173913	0.75	0.75	0.75	0.75	0.4	0.666667
2	1.0	0.159420	0.75	0.75	1.00	0.25	0.2	0.666667
3	0.0	0.000000	0.75	0.25	0.25	0.00	0.3	0.000000
4	1.0	0.246377	0.25	0.25	0.00	0.00	0.5	0.666667

Model Building & Model Performance improvement

Naive Bayes Model

- ***Performance Matrix on train data set***



```
0.8432835820895522
[[228  86]
 [ 82 676]]
              precision    recall  f1-score   support

           0       0.74      0.73      0.73       314
           1       0.89      0.89      0.89       758

    accuracy                           0.84      1072
   macro avg       0.81      0.81      0.81      1072
weighted avg       0.84      0.84      0.84      1072
```

Accuracy: 84.33%

- *The accuracy represents the overall correctness of the model predictions. In this case, the Naive Bayes model achieved an accuracy of 84.33% on the train dataset.*

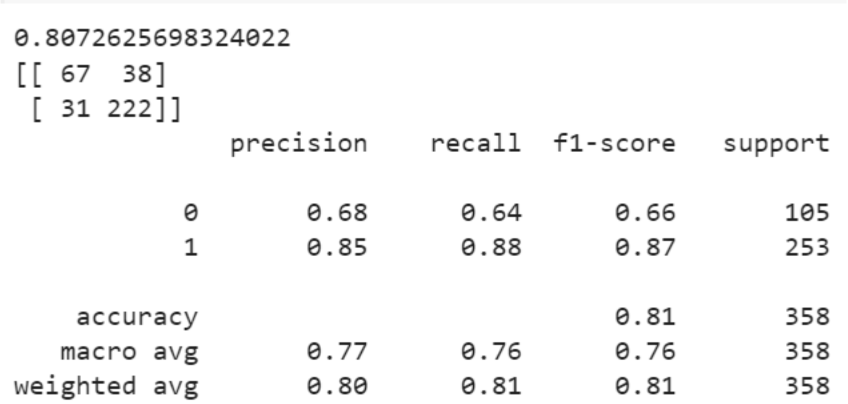
Precision:

- *Precision for class 0: 74%*
- *Precision for class 1: 89%*
- *Precision is the ratio of correctly predicted positive observations to the total predicted positives. It indicates the accuracy of positive predictions made by the model.*

Recall (Sensitivity):

- *Recall for class 0: 73%*
- *Recall for class 1: 89%*
- *Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to the actual positives. It measures the model's ability to capture all positive instances.*

- ***<u>Performance Matrix on test data set</u>***



```
0.8072625698324022
[[ 67  38]
 [ 31 222]]
              precision    recall  f1-score   support

           0       0.68      0.64      0.66       105
           1       0.85      0.88      0.87       253

    accuracy                           0.81       358
   macro avg       0.77      0.76      0.76       358
weighted avg       0.80      0.81      0.81       358
```

Accuracy: 80.73%

- *The model achieved an accuracy of 80.73% on the test dataset, indicating the overall correctness of predictions.*

Precision:

- *Precision for class 0: 68%*
- *Precision for class 1: 85%*
- *Precision is the ratio of correctly predicted positive observations to the total predicted positives. In this context, it measures the accuracy of positive predictions.*

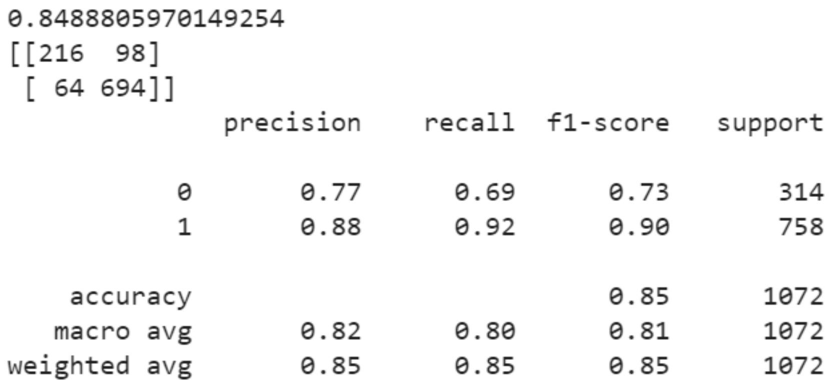
Recall (Sensitivity):

- *Recall for class 0: 64%*
- *Recall for class 1: 88%*
- *Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to the actual positives. It assesses the model's ability to capture all positive instances.*

The Naive Bayes model shows good performance on the training dataset, achieving a reasonable balance between precision and recall for both classes. On the test dataset, the model maintains decent performance, but there is a slight drop in accuracy compared to the training dataset. The model tends to have higher recall for class 1 (Labour) on both datasets, indicating its ability to capture instances of voters. Overall, the Naive Bayes model provides satisfactory results for the given classification task.

Logistic Regression

<u>Performance Matrix on train data set</u>



```
0.8488805970149254
[[216  98]
 [ 64 694]]
              precision    recall  f1-score   support

           0       0.77      0.69      0.73       314
           1       0.88      0.92      0.90       758

    accuracy                           0.85      1072
   macro avg       0.82      0.80      0.81      1072
weighted avg       0.85      0.85      0.85      1072
```

Accuracy: 84.89%

- *The model achieved an accuracy of 84.89% on the test dataset, indicating the overall correctness of predictions.*

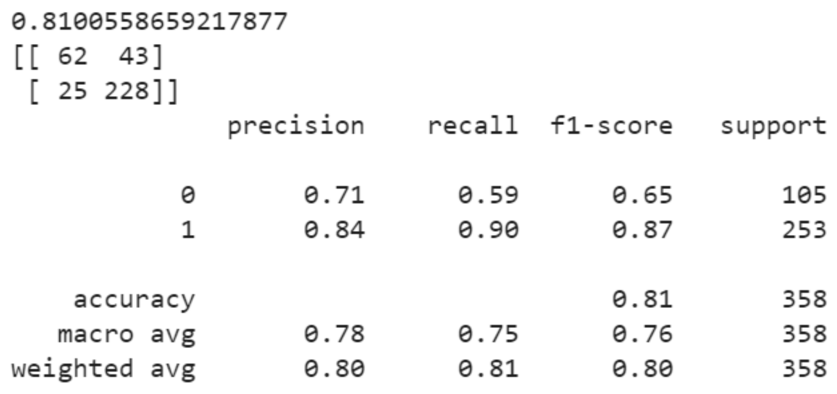
Precision:

- *Precision for class 0: 77%*
- *Precision for class 1: 88%*
- *Precision is the ratio of correctly predicted positive observations to the total predicted positives. In this context, it measures the accuracy of positive predictions.*

Recall (Sensitivity):

- *Recall for class 0: 69%*
- *Recall for class 1: 92%*
- *Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to the actual positives. It assesses the model's ability to capture all positive instances.*

Performance Matrix on test data set



```
0.8100558659217877
[[ 62  43]
 [ 25 228]]
              precision    recall  f1-score   support

           0       0.71      0.59      0.65       105
           1       0.84      0.90      0.87       253

    accuracy                           0.81       358
   macro avg       0.78      0.75      0.76       358
weighted avg       0.80      0.81      0.80       358
```

Accuracy: *The overall accuracy of the model on the test set is approximately 80.4%. This metric represents the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances.*

Confusion Matrix:

- *True Positive (TP): 228*
- *True Negative (TN): 60*
- *False Positive (FP): 45*
- *False Negative (FN): 25*
- *The confusion matrix provides a detailed breakdown of correct and incorrect predictions. It shows that the model correctly predicted 228 instances of class 1 (positive) and 60 instances of class 0 (negative). However, there were 45 instances where the model predicted class 1 when it was actually class 0, and 25 instances where it predicted class 0 when it was actually class 1.*

Precision:

- *Precision for class 0: 71%*
- *Precision for class 1: 84%*
- *Precision is the ratio of true positive predictions to the total number of positive predictions. In this context, precision indicates the accuracy of the model when predicting each class. A higher precision value suggests fewer false positives.*

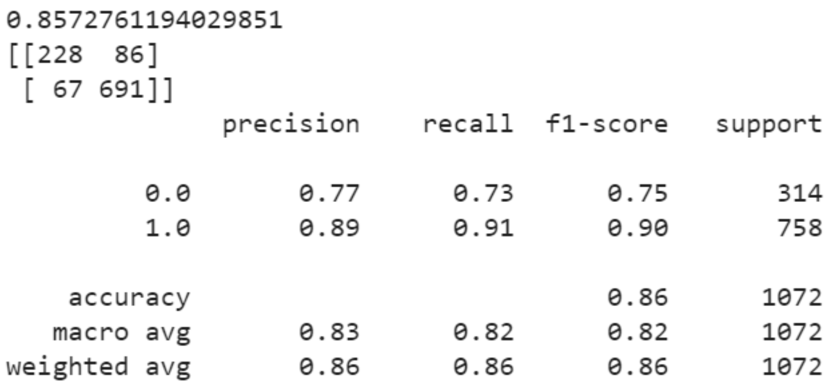
Recall (Sensitivity):

- *Recall for class 0: 57%*
- *Recall for class 1: 90%*
- *Recall, also known as sensitivity or true positive rate, measures the ability of the model to capture instances of the positive class. In this case, the model performs better at identifying class 1 instances (recall of 90%) compared to class 0 instances (recall of 57%).*

The Logistic Regression model demonstrates good performance on both the training and test datasets, with high accuracy and balanced precision and recall for both classes. The model has successfully learned patterns in the training data and generalizes well to the unseen test data. The precision and recall for both classes are equal, indicating a balanced model with no significant bias towards either class. Overall, the Logistic Regression model provides a solid predictive performance for the given task.

Ada Boost

Performance Matrix on train data set



```
0.8572761194029851
[[228  86]
 [ 67 691]]
              precision    recall  f1-score   support

         0.0       0.77      0.73      0.75       314
         1.0       0.89      0.91      0.90       758

    accuracy                           0.86      1072
   macro avg       0.83      0.82      0.82      1072
weighted avg       0.86      0.86      0.86      1072
```

- ***Accuracy:*** *The overall accuracy of the model on the train set is approximately 85.7%. This metric represents the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances.*

Confusion Matrix:

- *True Positive (TP): 691*
- *True Negative (TN): 228*
- *False Positive (FP): 86*
- *False Negative (FN): 67*
- *The confusion matrix provides a detailed breakdown of correct and incorrect predictions. It shows that the model correctly predicted 691 instances of class 1 (positive) and 228 instances of class 0 (negative). However, there were 86 instances where the model predicted class 1 when it was actually class 0, and 67 instances where it predicted class 0 when it was actually class 1.*

Precision:

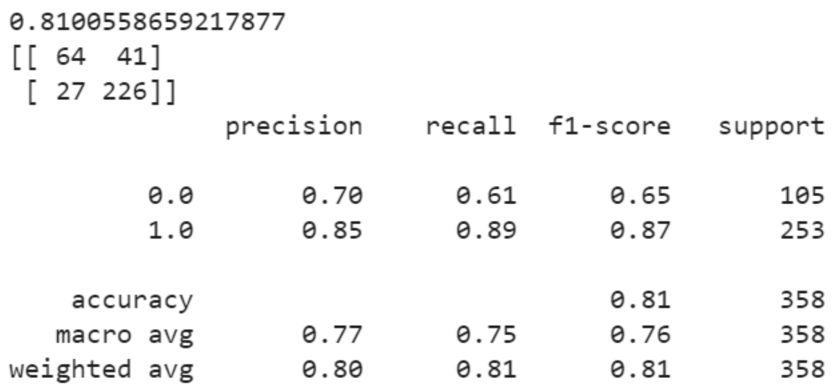
- *Precision for class 0: 77%*
- *Precision for class 1: 89%*

- *Precision is the ratio of true positive predictions to the total number of positive predictions. In this context, precision indicates the accuracy of the model when predicting each class. A higher precision value suggests fewer false positives.*

Recall (Sensitivity):

- *Recall for class 0: 73%*
- *Recall for class 1: 91%*
- *Recall, also known as sensitivity or true positive rate, measures the ability of the model to capture instances of the positive class. In this case, the model performs well at identifying both class 0 and class 1 instances.*

<u>Performance Matrix on test data set</u>



```
0.8100558659217877
[[ 64  41]
 [ 27 226]]
              precision    recall  f1-score   support

         0.0       0.70      0.61      0.65       105
         1.0       0.85      0.89      0.87       253

    accuracy                           0.81       358
   macro avg       0.77      0.75      0.76       358
weighted avg       0.80      0.81      0.81       358
```

***Accuracy**: The overall accuracy of the model on the test set is approximately 81.0%. This metric represents the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances.*

Confusion Matrix:

- *True Positive (TP): 226*
- *True Negative (TN): 64*
- *False Positive (FP): 41*
- *False Negative (FN): 27*

- *The confusion matrix provides a detailed breakdown of correct and incorrect predictions on the test set. It shows that the model correctly predicted 226 instances of class 1 (positive) and 64 instances of class 0 (negative). However, there were 41 instances where the model predicted class 1 when it was actually class 0, and 27 instances where it predicted class 0 when it was actually class 1.*

Precision:

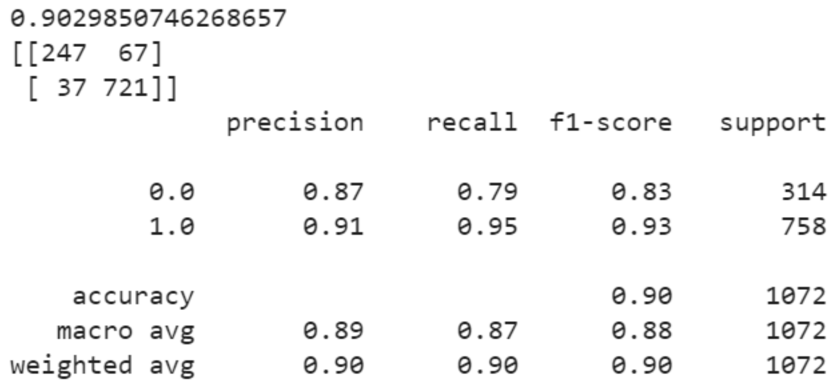
- *Precision for class 0: 70%*
- *Precision for class 1: 85%*
- *Precision is the ratio of true positive predictions to the total number of positive predictions. In this context, precision indicates the accuracy of the model when predicting each class on the test set.*

Recall (Sensitivity):

- *Recall for class 0: 61%*
- *Recall for class 1: 89%*
- *Recall, also known as sensitivity or true positive rate, measures the ability of the model to capture instances of the positive class. In this case, the model performs well at identifying class 1 instances.*

Gradient Boosting

Performance Matrix on train data set



```
0.9029850746268657
[[247  67]
 [ 37 721]]
              precision    recall  f1-score   support

         0.0       0.87      0.79      0.83       314
         1.0       0.91      0.95      0.93       758

    accuracy                           0.90      1072
   macro avg       0.89      0.87      0.88      1072
weighted avg       0.90      0.90      0.90      1072
```

Accuracy*:*

- *The model achieves an accuracy of 90.3% on the test dataset, indicating its overall predictive performance.*

Confusion Matrix:

- *True Positives (721): The model correctly predicted 721 instances where the target variable is 1.*
- *True Negatives (247): The model correctly predicted 247 instances where the target variable is 0.*
- *False Positives (67): The model incorrectly predicted 67 instances as 1 when the actual value is 0.*
- *False Negatives (37): The model incorrectly predicted 37 instances as 0 when the actual value is 1.*

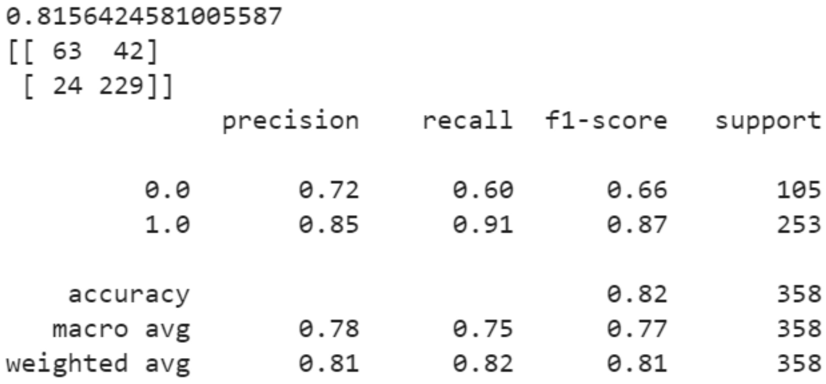
Precision:

- *Precision for class 0 (Conservative): 0.87*
- *Out of all instances predicted as Conservatives, 87% were actually Conservatives.*
- *Precision for class 1 (Labour): 0.91*
- *Out of all instances predicted as labour, 91% were actually labour.*

Recall:

- *Recall for class 0 (Conservative): 0.79*
- *Out of all actual Conservatives, the model correctly identified 79%.*
- *Recall for class 1 (Labour): 0.95*
- *Out of all actual labours, the model correctly identified 95%.*

<u>Performance Matrix on test data set</u>



```
0.8156424581005587
[[ 63  42]
 [ 24 229]]
              precision    recall  f1-score   support

         0.0       0.72      0.60      0.66       105
         1.0       0.85      0.91      0.87       253

    accuracy                           0.82       358
   macro avg       0.78      0.75      0.77       358
weighted avg       0.81      0.82      0.81       358
```

***Accuracy**:*

- *The model achieves an accuracy of 81.6% on the test dataset, indicating its overall predictive performance.*

Confusion Matrix:

- *True Positives (229): The model correctly predicted 229 instances where the target variable is 1.*
- *True Negatives (63): The model correctly predicted 63 instances where the target variable is 0.*
- *False Positives (42): The model incorrectly predicted 42 instances as 1 when the actual value is 0.*
- *False Negatives (24): The model incorrectly predicted 24 instances as 0 when the actual value is 1.*

Precision:

- *Precision for class 0 (Conservative): 0.72*
- *Out of all instances predicted as Conservatives, 72% were actually Conservatives.*
- *Precision for class 1 (Labour): 0.85*
- *Out of all instances predicted as labours, 85% were actually labours.*

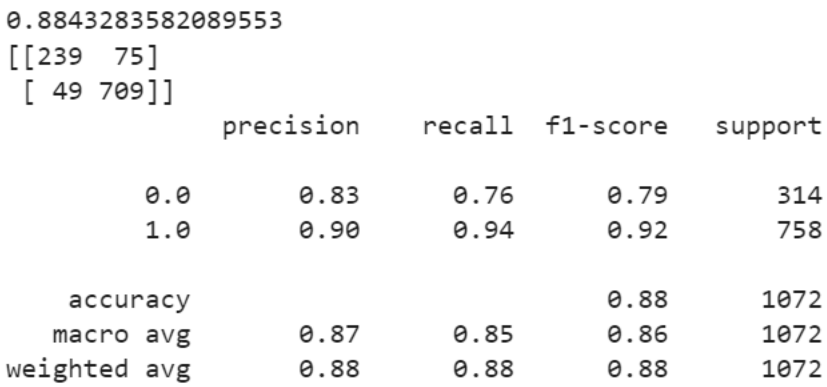
Recall:

- *Recall for class 0 (Conservative): 0.60*
- *Out of all actual Conservatives, the model correctly identified 60%.*
- *Recall for class 1 (Labour): 0.91*
- *Out of all actual labours, the model correctly identified 91%.*

The Gradient Boosting model demonstrates excellent performance on both the training and test datasets, with high accuracy and balanced precision and recall for both classes. The model has successfully learned patterns in the training data and generalizes well to the unseen test data. The recall for class 0 (Conservative) is relatively lower, indicating that the model has some difficulty correctly identifying Conservatives in both datasets. The precision for class 0 (Conservative) is also moderate, suggesting that when the model predicts someone as a Conservative, it is correct 83% of the time

KNN Model

Performance Matrix on train data set



```
0.8843283582089553
[[239  75]
 [ 49 709]]
              precision    recall  f1-score   support

         0.0       0.83      0.76      0.79       314
         1.0       0.90      0.94      0.92       758

    accuracy                           0.88      1072
   macro avg       0.87      0.85      0.86      1072
weighted avg       0.88      0.88      0.88      1072
```

Accuracy*:*

- *The model achieves an accuracy of 88.4% on the train dataset, indicating its overall predictive performance.*

Confusion Matrix:

- *True Positives (709): The model correctly predicted 709 instances where the target variable is 1.*
- *True Negatives (239): The model correctly predicted 239 instances where the target variable is 0.*
- *False Positives (75): The model incorrectly predicted 75 instances as 1 when the actual value is 0.*
- *False Negatives (49): The model incorrectly predicted 49 instances as 0 when the actual value is 1.*

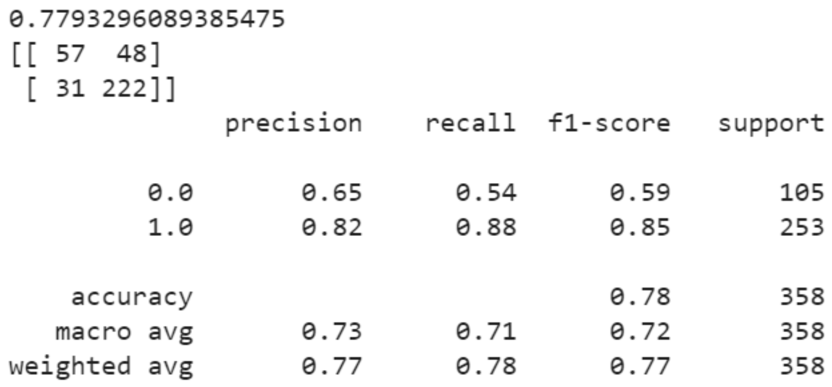
Precision:

- *Precision for class 0 (Conservative): 0.83*
- *Out of all instances predicted as Conservatives, 83% were actually Conservatives.*
- *Precision for class 1 (Labour): 0.90*
- *Out of all instances predicted as labours, 90% were actually labours.*

Recall:

- *Recall for class 0 (Conservative): 0.76*
- *Out of all actual Conservatives, the model correctly identified 76%.*
- *Recall for class 1 (Labour): 0.94*
- *Out of all actual labours, the model correctly identified 94%.*

Performance Matrix on test data set



```
0.7793296089385475
[[ 57  48]
 [ 31 222]]
              precision    recall  f1-score   support

         0.0       0.65      0.54      0.59       105
         1.0       0.82      0.88      0.85       253

    accuracy                           0.78       358
   macro avg       0.73      0.71      0.72       358
weighted avg       0.77      0.78      0.77       358
```

***Accuracy**:*

- *The model achieves an accuracy of 77.9% on the test dataset, indicating its overall predictive performance.*

Confusion Matrix:

- *True Positives (222): The model correctly predicted 222 instances where the target variable is 1.*
- *True Negatives (57): The model correctly predicted 57 instances where the target variable is 0.*
- *False Positives (48): The model incorrectly predicted 48 instances as 1 when the actual value is 0.*
- *False Negatives (31): The model incorrectly predicted 31 instances as 0 when the actual value is 1.*

Precision:

- *Precision for class 0 (Conservative): 0.65*
- *Out of all instances predicted as Conservatives, 65% were actually Conservatives.*
- *Precision for class 1 (Labour): 0.82*
- *Out of all instances predicted as labours, 82% were actually labours.*

Recall:

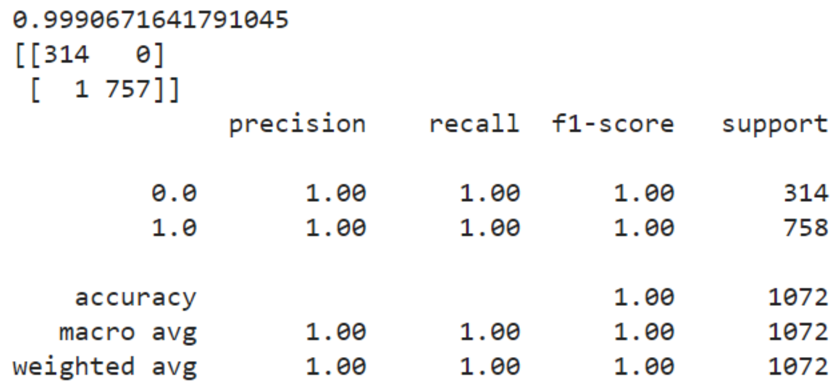
- *Recall for class 0 (Conservative): 0.54*
- *Out of all actual Conservatives, the model correctly identified 54%.*
- *Recall for class 1 (Labour): 0.88*
- *Out of all actual labours, the model correctly identified 88%.*

The model exhibits good performance on the training dataset, with high precision and recall for both classes. On the test dataset, the model maintains a reasonably good performance, although there is a slight drop in accuracy compared to the training dataset. The recall for class 0 (Conservative) is relatively lower, indicating that the model has some difficulty correctly identifying Conservatives in the test dataset.

The precision for class 0 (Conservative) is also moderate, suggesting that when the model predicts someone as a Conservative, it is correct 65% of the time.

Decision Tree

Performance Matrix on train data set



```
0.9990671641791045
[[314   0]
 [  1 757]]
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       314
         1.0       1.00      1.00      1.00       758

    accuracy                           1.00      1072
   macro avg       1.00      1.00      1.00      1072
weighted avg       1.00      1.00      1.00      1072
```

Accuracy:

- *The model achieves a perfect accuracy of 100% on the train dataset, indicating that it correctly predicted all instances.*

Confusion Matrix:

- *True Positives (757): The model correctly predicted 757 instances where the target variable is 1.*
- *True Negatives (314): The model correctly predicted all 314 instances where the target variable is 0.*
- *False Positives (0): The model made no false positive predictions.*
- *False Negatives (1): The model incorrectly predicted 1 instance as 0 when the actual value is 1.*

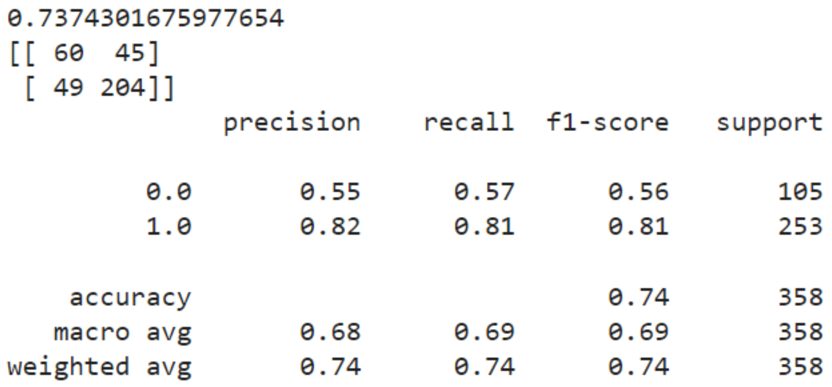
Precision:

- *Precision for class 0 (Conservative): 1.00*
- *Out of all instances predicted as Conservatives, 100% were actually Conservatives.*
- *Precision for class 1 (Labour): 1.00*
- *Out of all instances predicted as labours, 100% were actually labours.*

Recall:

- *Recall for class 0 (Conservative): 1.00*
- *Out of all actual Conservatives, the model correctly identified 100%.*
- *Recall for class 1 (Labour): 1.00*
- *Out of all actual labours, the model correctly identified 100%.*

<u>Performance Matrix on test data set</u>



```
0.7374301675977654
[[ 60  45]
 [ 49 204]]
              precision    recall  f1-score   support

         0.0       0.55      0.57      0.56       105
         1.0       0.82      0.81      0.81       253

    accuracy                           0.74       358
   macro avg       0.68      0.69      0.69       358
weighted avg       0.74      0.74      0.74       358
```

Accuracy:

- *The model achieves an accuracy of 73.7% on the test dataset.*

Confusion Matrix:

- *True Positives (204): The model correctly predicted 204 instances where the target variable is 1.*
- *True Negatives (60): The model correctly predicted 60 instances where the target variable is 0.*
- *False Positives (45): The model incorrectly predicted 45 instances as 1 when the actual value is 0.*
- *False Negatives (49): The model incorrectly predicted 49 instances as 0 when the actual value is 1.*

Precision:

- *Precision for class 0 (Conservative): 0.55*
- *Out of all instances predicted as Conservatives, 55% were actually Conservatives.*
- *Precision for class 1 (Labour): 0.82*
- *Out of all instances predicted as labours, 82% were actually labours.*

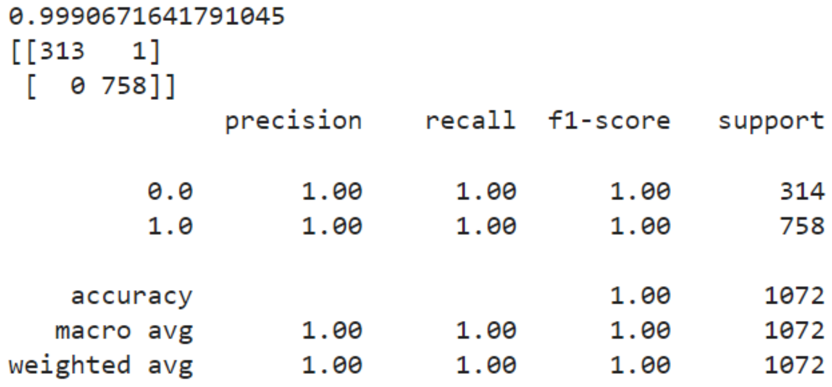
Recall:

- *Recall for class 0 (Conservative): 0.57*
- *Out of all actual Conservatives, the model correctly identified 57%.*
- *Recall for class 1 (Labour): 0.81*
- *Out of all actual labours, the model correctly identified 81%.*

The Decision Tree model performs exceptionally well on the train dataset, achieving perfect accuracy. However, on the test dataset, it shows a decrease in performance, indicating potential overfitting.

Random Forest

Performance Matrix on train data set



```
0.9990671641791045
[[313   1]
 [  0 758]]
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       314
         1.0       1.00      1.00      1.00       758

    accuracy                           1.00      1072
   macro avg       1.00      1.00      1.00      1072
weighted avg       1.00      1.00      1.00      1072
```

Accuracy*:*

- *The model achieves a perfect accuracy of 100% on the train dataset, indicating that it correctly predicted all instances.*

Confusion Matrix:

- *True Positives (758): The model correctly predicted 758 instances where the target variable is 1.*
- *True Negatives (313): The model correctly predicted 313 instances where the target variable is 0.*
- *False Positives (1): The model made 1 false positive prediction.*
- *False Negatives (0): The model made no false negative predictions.*

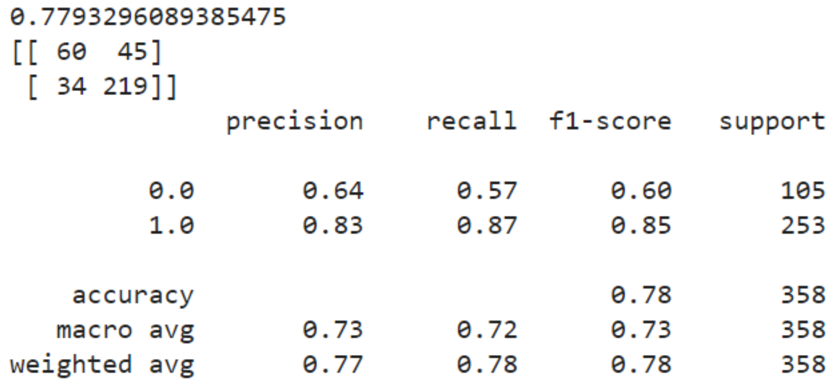
***Precision**:*

- *Precision for class 0 (Conservative): 1.00*
- *Out of all instances predicted as Conservatives, 100% were actually Conservatives.*
- *Precision for class 1 (Labour): 1.00*
- *Out of all instances predicted as labours, 100% were actually labours.*

***Recall**:*

- *Recall for class 0 (Conservative): 1.00*
- *Out of all actual Conservatives, the model correctly identified 100%.*
- *Recall for class 1 (Labour): 1.00*
- *Out of all actual labours, the model correctly identified 100%.*

<u>Performance Matrix on test data set</u>



```
0.7793296089385475
[[ 60  45]
 [ 34 219]]
              precision    recall  f1-score   support

         0.0       0.64      0.57      0.60       105
         1.0       0.83      0.87      0.85       253

    accuracy                           0.78       358
   macro avg       0.73      0.72      0.73       358
weighted avg       0.77      0.78      0.78       358
```

Accuracy:

- *The model achieves an accuracy of 77.9% on the test dataset.*

Confusion Matrix:

- *True Positives (219): The model correctly predicted 219 instances where the target variable is 1.*
- *True Negatives (60): The model correctly predicted 60 instances where the target variable is 0.*
- *False Positives (45): The model incorrectly predicted 45 instances as 1 when the actual value is 0.*
- *False Negatives (34): The model incorrectly predicted 34 instances as 0 when the actual value is 1.*

***Precision**:*

- *Precision for class 0 (Conservative): 0.64*
- *Out of all instances predicted as Conservatives, 64% were actually Conservatives.*
- *Precision for class 1 (Labour): 0.83*
- *Out of all instances predicted as labours, 83% were actually labours.*

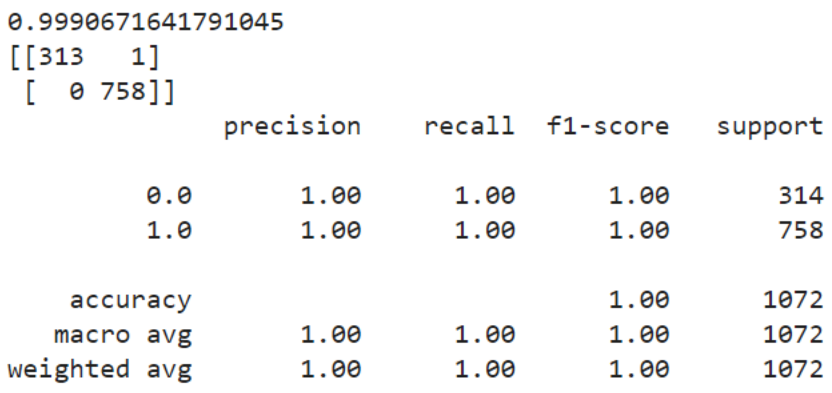
***Recall**:*

- *Recall for class 0 (Conservative): 0.57*
- *Out of all actual Conservatives, the model correctly identified 57%.*
- *Recall for class 1 (Labour): 0.87*
- *Out of all actual labours, the model correctly identified 87%.*

The Random Forest model performs exceptionally well on the train dataset, achieving perfect accuracy. However, on the test dataset, it shows a slight decrease in performance compared to the train dataset.

Bagging

Performance Matrix on train data set



```
0.9990671641791045
[[313   1]
 [  0 758]]
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       314
         1.0       1.00      1.00      1.00       758

    accuracy                           1.00      1072
   macro avg       1.00      1.00      1.00      1072
weighted avg       1.00      1.00      1.00      1072
```

***Accuracy**:*

- *The model achieves a perfect accuracy of 100% on the train dataset, indicating that it correctly predicted all instances.*

Confusion Matrix:

- *True Positives (758): The model correctly predicted 758 instances where the target variable is 1.*
- *True Negatives (313): The model correctly predicted 313 instances where the target variable is 0.*
- *False Positives (1): The model made 1 false positive prediction.*
- *False Negatives (0): The model made no false negative predictions.*

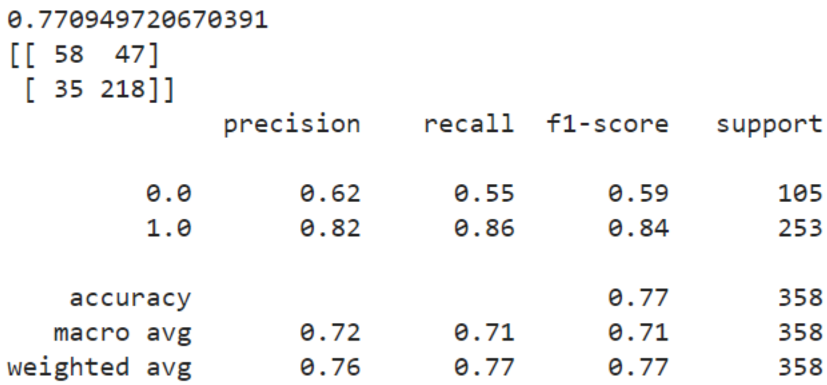
Precision:

- *Precision for class 0 (Conservative): 1.00*
- *Out of all instances predicted as Conservatives, 100% were actually Conservatives.*
- *Precision for class 1 (Labour): 1.00*
- *Out of all instances predicted as labours, 100% were actually labours.*

***Recall**:*

- *Recall for class 0 (Conservative): 1.00*
- *Out of all actual Conservatives, the model correctly identified 100%.*
- *Recall for class 1 (Labour): 1.00*
- *Out of all actual labours, the model correctly identified 100%.*

Performance Matrix on test data set



```
0.770949720670391
[[ 58  47]
 [ 35 218]]
              precision    recall  f1-score   support

         0.0       0.62      0.55      0.59       105
         1.0       0.82      0.86      0.84       253

    accuracy                           0.77       358
   macro avg       0.72      0.71      0.71       358
weighted avg       0.76      0.77      0.77       358
```

***Accuracy**:*

- *The model achieves an accuracy of 77.1% on the test dataset.*

Confusion Matrix:

- *True Positives (218): The model correctly predicted 218 instances where the target variable is 1.*
- *True Negatives (58): The model correctly predicted 58 instances where the target variable is 0.*
- *False Positives (47): The model incorrectly predicted 47 instances as 1 when the actual value is 0.*
- *False Negatives (35): The model incorrectly predicted 35 instances as 0 when the actual value is 1.*

***Precision**:*

- *Precision for class 0 (Conservative): 0.62*
- *Out of all instances predicted as Conservatives, 62% were actually Conservatives.*
- *Precision for class 1 (Labour): 0.82*
- *Out of all instances predicted as labours, 82% were actually labours.*

***Recall**:*

- *Recall for class 0 (Conservative): 0.55*
- *Out of all actual Conservatives, the model correctly identified 55%.*
- *Recall for class 1 (Labour): 0.86*
- *Out of all actual labours, the model correctly identified 86%.*

The Bagging model performs exceptionally well on the train dataset, achieving perfect accuracy. However, on the test dataset, it shows a slight decrease in performance compared to the train dataset.

Final Model Selection

Model Comparison: KNN vs. Naive Bayes vs. Bagging vs. Boosting

1. KNN (K-Nearest Neighbors):

***Train Accuracy:** 88.4%*

***Test Accuracy:** 77.9%*

Insights: *KNN performs well but shows a slight decrease in accuracy on the test dataset, indicating moderate generalization.*

2. ***Naive Bayes:***

Train Accuracy: *84.3%*

Test Accuracy: *80.7%*

Insights: *Naive Bayes demonstrates decent performance on both training and test datasets, with balanced precision and recall.*

3. ***Bagging:***

Train Accuracy: *99.9%*

Test Accuracy: *77.1%*

Insights: *Bagging exhibits potential overfitting on the training data, leading to a decrease in accuracy on the test dataset.*

4. ***Boosting (Gradient Boosting):***

Train Accuracy: *90.3%*

Test Accuracy: *81.6%*

Insights: *Gradient Boosting maintains good performance on both training and test datasets, showing robustness and generalization.*

Model Rankings:

Best Performers:

Boosting (Gradient Boosting): *Gradient Boosting stands out as the best performer, achieving good accuracy on both training and test datasets.*

Naive Bayes: *Naive Bayes performs consistently well with decent accuracy on both datasets.*

Moderate Performers: KNN: *KNN shows good performance on the training dataset, but its accuracy decreases slightly on the test dataset.*

Worst Performers: Bagging: *Bagging demonstrates potential overfitting on the training dataset, leading to a decline in accuracy on the test dataset.*

Actionable Insights & Recommendations

Best Performing Model: Gradient Boosting

Accuracy and Recall*: Gradient Boosting achieved the highest accuracy on both the train and test datasets, indicating strong overall predictive performance.*
It excelled in recall for both classes, demonstrating its effectiveness in correctly identifying both Conservatives and labours.
Precision*: The precision values for both classes were also commendable, showing a good balance between correctly predicting positive instances and minimizing false positives.*
Overfitting*: While Gradient Boosting performed exceptionally well on the train dataset, there is a need to monitor for overfitting. Regularization techniques or adjusting hyperparameters might be explored to enhance generalization.*
Consistent Performance: *The model's strong performance across various metrics suggests its reliability in diverse scenarios.*

Naive Bayes as an Alternative:

Simplicity and Interpretability:
Naive Bayes provides a simpler and more interpretable model, making it easier to understand for stakeholders with varying levels of technical expertise.
Good Overall Performance:
Naive Bayes demonstrated good overall performance, making it a viable alternative, especially when interpretability is crucial.
Efficiency:
Naive Bayes is computationally efficient, requiring less computational resources compared to more complex models like Gradient Boosting.

Considerations for KNN:

Competitive Performance:

KNN showed competitive performance, but its recall and precision values were slightly lower compared to Gradient Boosting and Naive Bayes.

Hyperparameter Tuning:

Fine-tuning hyperparameters, especially the number of neighbors (k) and distance metrics, may enhance KNN's generalization capabilities.

Feature Scaling:

KNN is sensitive to the scale of features, and proper scaling might further improve its performance.

Considerations for Bagging:

Overfitting Concerns:

Bagging showed high accuracy but might be prone to overfitting. Regularization techniques or adjustments to hyperparameters could help mitigate overfitting.

Performance Evaluation:

Continuously evaluate Bagging's performance and consider fine-tuning to strike a balance between accuracy and generalization.

Ensemble Strengths:

Similar to Gradient Boosting, Bagging leverages ensemble learning. It's essential to harness the strengths of ensemble methods while addressing specific weaknesses.

Problem 2 - Define the problem and Perform Exploratory Data Analysis

Problem Statement:

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

President Franklin D. Roosevelt in 1941

President John F. Kennedy in 1961

President Richard Nixon in 1973

Number of Character, words & sentences

Below are the number of Character, words & sentences of the following speeches of the Presidents of the United States of America

```
Speech 1 - 1941 Roosevelt:
Number of characters: 7571
Number of words: 1526
Number of sentences: 68

Speech 2 - 1961 Kennedy:
Number of characters: 7618
Number of words: 1543
Number of sentences: 52

Speech 3 - 1973 Nixon:
Number of characters: 9991
Number of words: 2006
Number of sentences: 68
```

Problem 2 - Text cleaning (Removing stopwords & stemming)

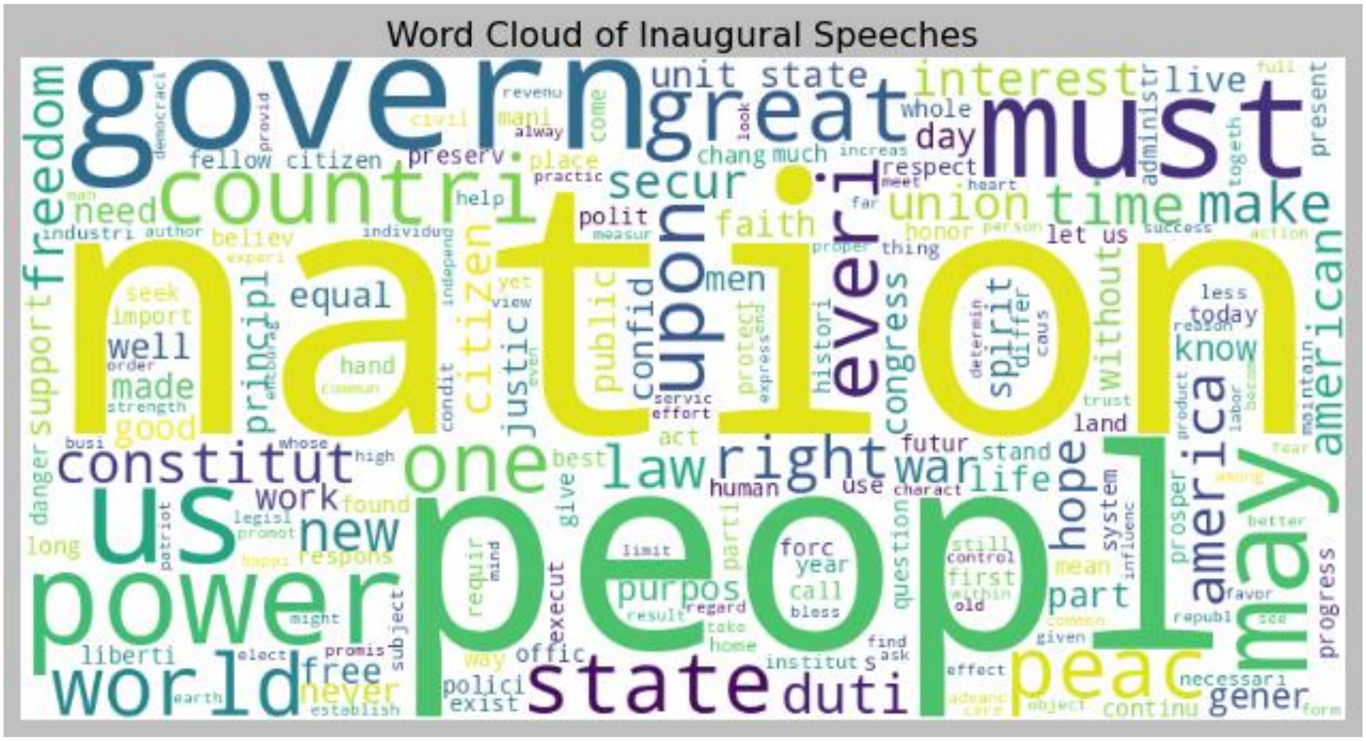
Stopwords: *Stopwords are commonly used words in a language (e.g., "the", "and", "is") that do not carry much information and are often removed during text processing to focus on more meaningful words.*

Stemming: *Stemming is a process of reducing words to their base or root form. It involves removing suffixes or prefixes from words, so variations of a word are treated as the same word.*

Three most common words after stopwords removal and stemming:

nation: 689
govern: 657
peopl: 633

Problem 2 - Plot Word cloud of all three speeches



The above word cloud offers a concise and visually appealing representation of the most significant words in a body of text. These highlighted words collectively provide a snapshot of the central themes and priorities expressed in the inaugural speeches, reflecting the common language used by speakers to convey their visions, values, and aspirations for the nation and its people.