MACHINE LEARNING-II

KNN, NAÏVE BAYES, BAGGING, BOOSTING, TEXT ANALYTICS

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.

EDA

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

Fig 1 Sample dataset

- The dataset has 1525 rows and 10 columns.
- Unnamed:0 is not a useful column and hence we can drop the column.
- There are 7 integer variables excluding Unnamed:0 and there are 2 object variables.
- The dependent variable and the gender variables are object hence it should converted into categorical variables.
- All the independent variables in the data are categorical variables except age.

	vote	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
count	1525	1525	1525	1525	1525	1525	1525	1525
unique	2	5	5	5	5	11	4	2
top	Labour	3	3	4	2	11	2	female
freq	1063	607	648	836	624	338	782	812

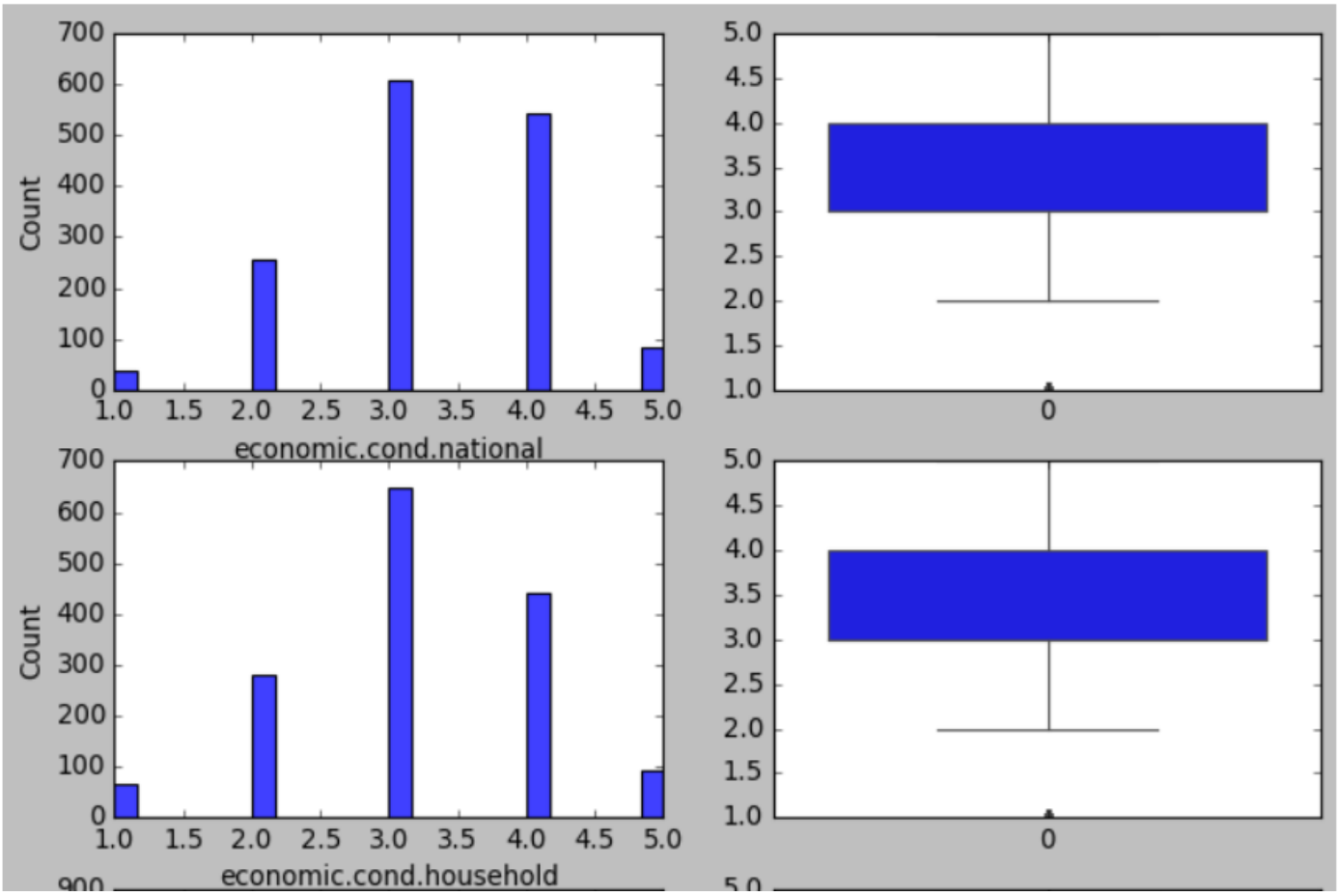
Fig 2 Descriptive statistics of categorical variables

	age
count	1525.000000
mean	54.182295
std	15.711209
min	24.000000
25%	41.000000
50%	53.000000
75%	67.000000
max	93.000000

Fig 3 Descriptive statistics of numerical variables

- Here the case of duplication does not exist as different individuals can have same outputs

Univariate and Bivariate analysis



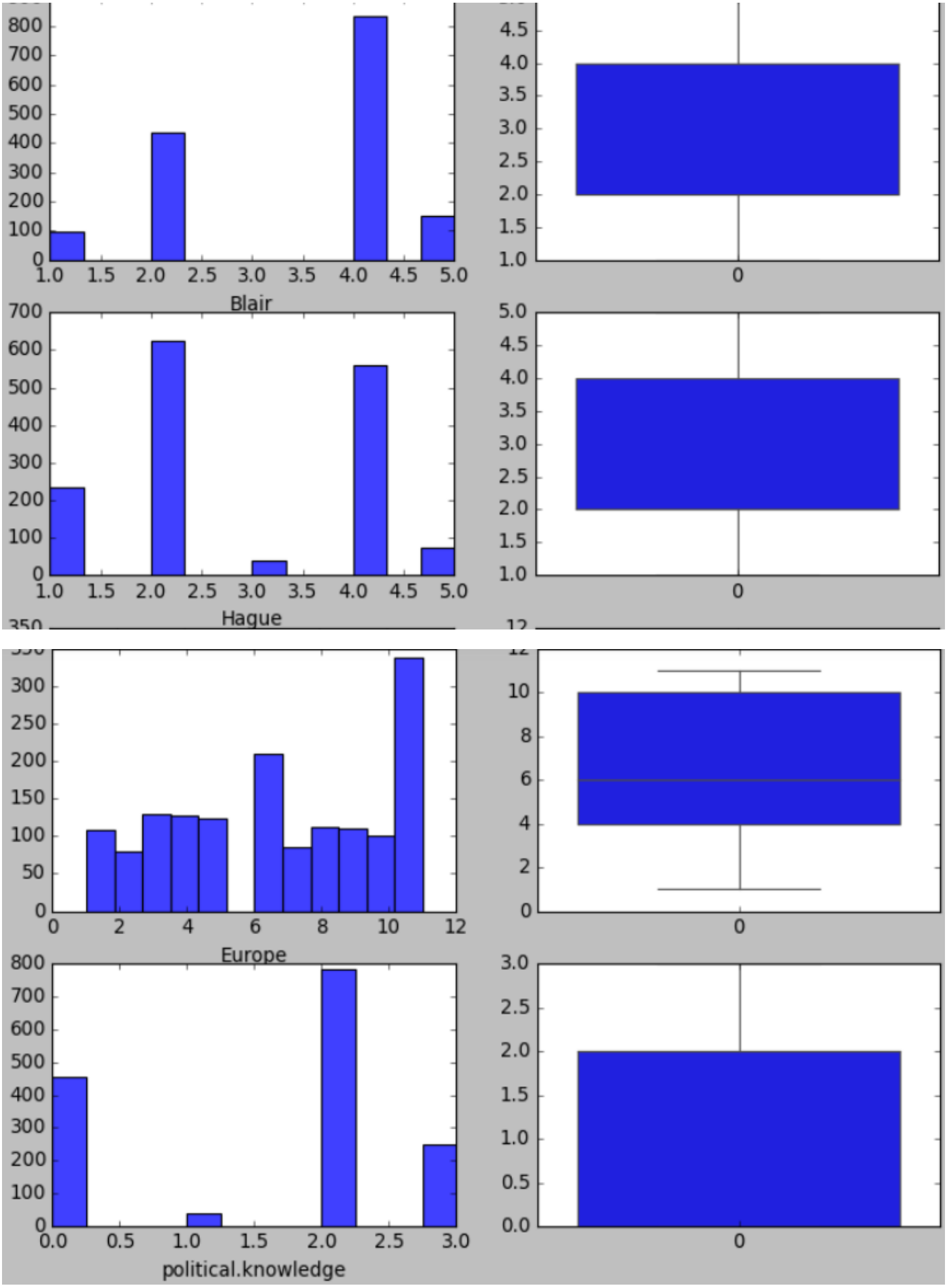


Fig 4 Univariate analysis of categorical variables

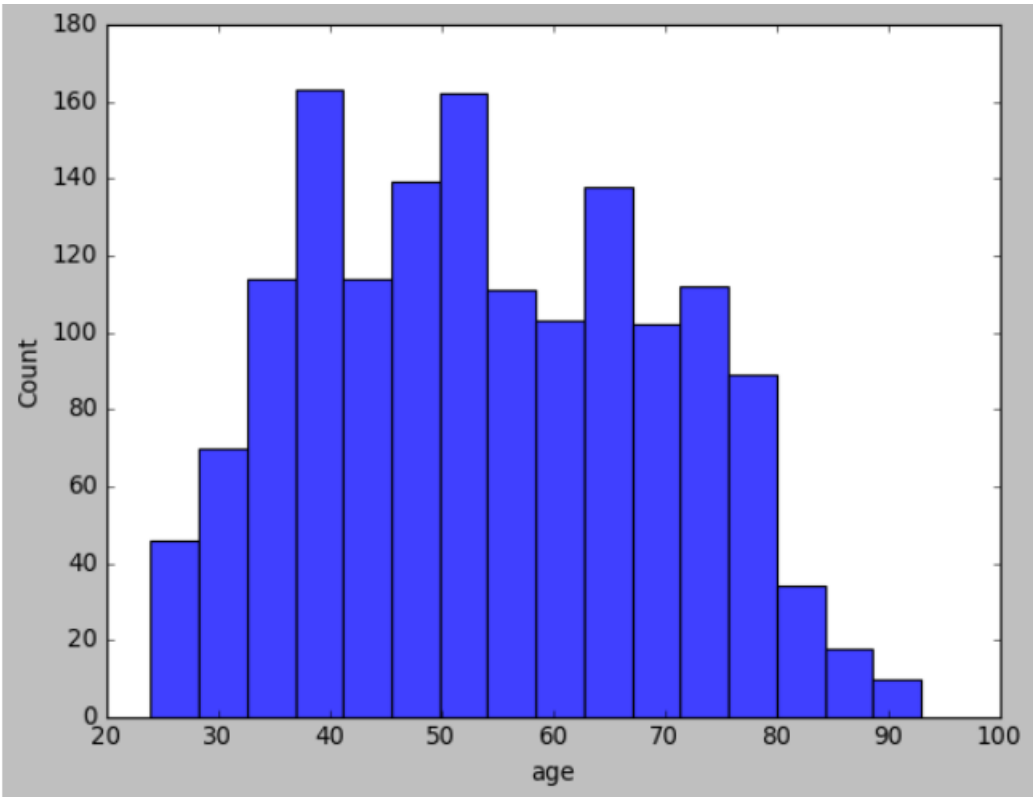


Fig 5 Univariate analysis of Numerical variables

- economic condition of the nation: 600 has rated 3 and 500+ has rated 4
- household economic condition: 600 has rated 3 and 400 has rated 4, about 250 has rated 2
- both the economic conditions of the nation and the household have a similar distribution.
- Blair: Max number of people about 800 people have rated a 4 and 400 have rated 2, no one has rated a 3 which means either high support or high opposition for the labour leaders.
- Hague: 200 has rated 1 and 600 has rated 2 and only 400 have rated 4, less support is shown here.
- Europe: about 350 has rated 11 and the majority of the people have rated more than 6 which means the majority of people do not support European integration.
- Political knowledge: max number of people have high political knowledge 2 or more than 2.

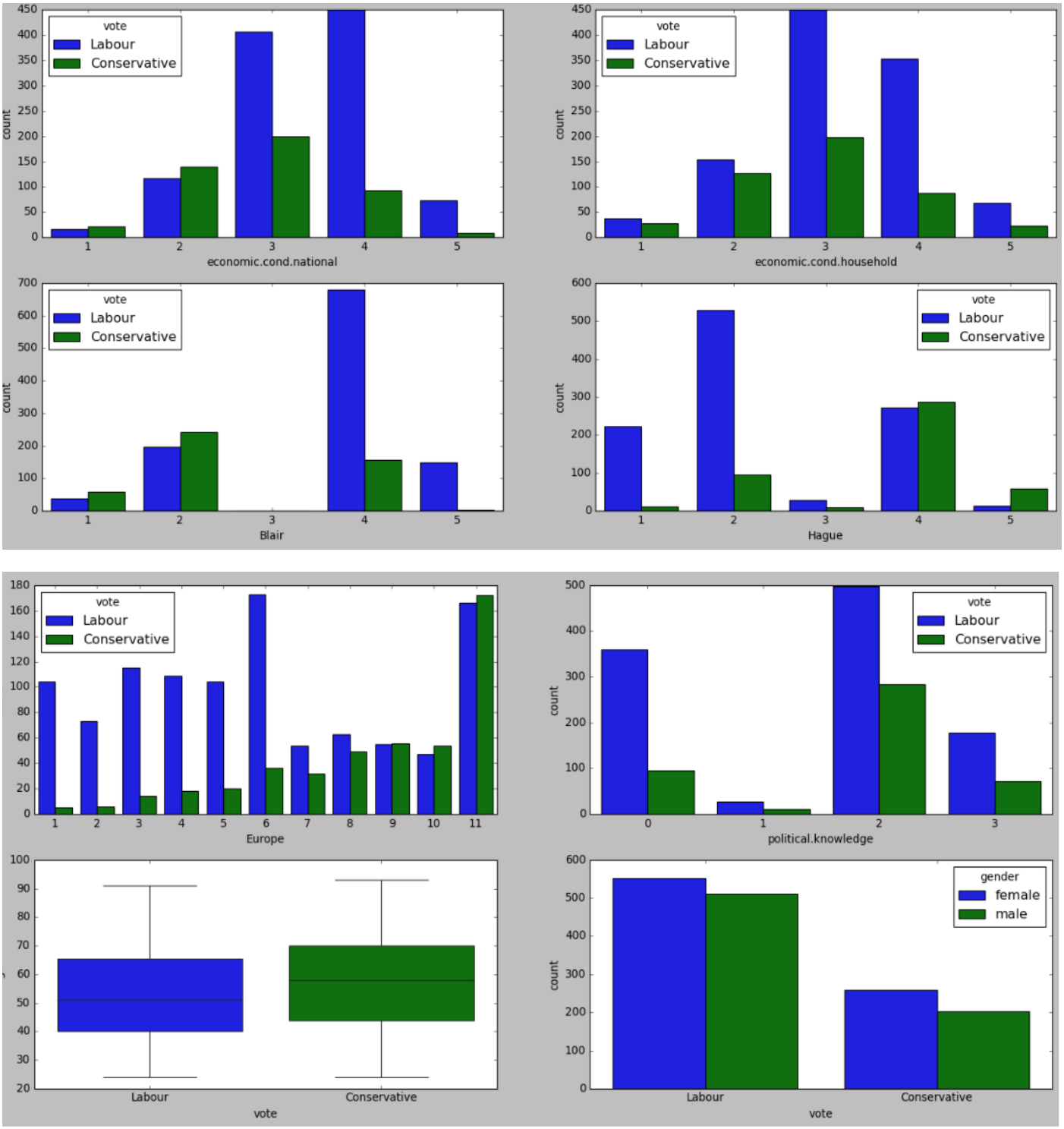


Fig 6 Count plot and boxplot of independent variables with dependent variable

- Majority of the section who says the economic condition of the nation is good supports labour party and the ones rated 2 support the conservative party.
- A similar pattern is observed in the economic condition of households too.
- People who support the labour party have a good assessment of labour leaders, and those who oppose have a bad rating of labour leaders
- Similarly people who support the conservative party have good assessments towards conservative leaders, and those who oppose have bad ratings for conservative leaders.
- People who support the Labour Party strongly oppose European integration, while supporters of the conservative party highly support European integration.
- About 400 people who support the Labour Party have 0 idea about the party's stand on European integration while the people who support the conservative party are more

aware of the party's ideas and propaganda. Even though a large number of people who support the Labour Party have given a rating of 2.

- From age it is visible that age does not affect the people's support towards the party.
- Gender also doesn't show any relevance as the party has the same proportion of both male and female voters.
- Age and Gender do not provide any meaningful predictions for a vote.

Data Preprocessing

- There are no missing values or null values in the dataset.
- outliers are not present as most of the data are categorical.

	0	1	2	3	4
vote	Labour	Labour	Labour	Labour	Labour
age	43	36	35	24	41
economic.cond.national_2	0	0	0	0	1
economic.cond.national_3	1	0	0	0	0
economic.cond.national_4	0	1	1	1	0
economic.cond.national_5	0	0	0	0	0
economic.cond.household_2	0	0	0	1	1
economic.cond.household_3	1	0	0	0	0
economic.cond.household_4	0	1	1	0	0
economic.cond.household_5	0	0	0	0	0
Blair_2	0	0	0	1	0
Blair_3	0	0	0	0	0
Blair_4	1	1	0	0	0
Blair_5	0	0	1	0	0
Hague_2	0	0	1	0	0
Hague_3	0	0	0	0	0
Hague_4	0	1	0	0	0
Hague_5	0	0	0	0	0
Europe_2	1	0	0	0	0
Europe_3	0	0	1	0	0
Europe_4	0	0	0	1	0
Europe_5	0	1	0	0	0
Europe_6	0	0	0	0	1
Europe_7	0	0	0	0	0
Europe_8	0	0	0	0	0
Europe_9	0	0	0	0	0
Europe_10	0	0	0	0	0
Europe_11	0	0	0	0	0
political.knowledge_1	0	0	0	0	0
political.knowledge_2	1	1	1	0	1
political.knowledge_3	0	0	0	0	0
gender_male	0	1	1	0	1

Fig 7 Dataset after creating dummy variables for dependent variables.

Model Building

- For naive bayes classification, it assumes all the data to be normally distributed and they are independent of each other hence scaling is not required for naive bayes classification.
- For KNN we will use Z-score method for scaling the data.
- The model was split with 70% train data and 30% test data.

Naïve Bayes classification

```
0.8031865042174321
[[206 126]
 [ 84 651]]
              precision    recall  f1-score   support

Conservative       0.71      0.62      0.66       332
      Labour       0.84      0.89      0.86       735

    accuracy                           0.80      1067
   macro avg       0.77      0.75      0.76      1067
weighted avg       0.80      0.80      0.80      1067
```

Fig 8 Naive Bayes Training data classification report

```
0.7816593886462883
[[ 86  44]
 [ 56 272]]
              precision    recall  f1-score   support

Conservative       0.61      0.66      0.63       130
      Labour       0.86      0.83      0.84       328

    accuracy                           0.78       458
   macro avg       0.73      0.75      0.74       458
weighted avg       0.79      0.78      0.78       458
```

Fig 9 Naive Bayes Testing data classification report

- The accuracy of the train and test data are quite similar.
- The precision and recall for predicting conservative party is very weak in both the models.
- Precision and recall for predicting labour party is very strong in both models.

Since the data consists of only categorical variables except 'age' performing a KNN model is unnecessary as it is a measure of distance between adjacent points and then assigning to the particular class with the shortest distance. Hence KNN can be avoided.¶

Bagging

```
0.9990627928772259
[[330   1]
 [  0 736]]
              precision    recall  f1-score   support

Conservative       1.00      1.00      1.00       331
      Labour       1.00      1.00      1.00       736

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

Fig 10 Bagging train data classification report

```
0.7947598253275109
[[ 79  52]
 [ 42 285]]
              precision    recall  f1-score   support

Conservative       0.65      0.60      0.63       131
      Labour       0.85      0.87      0.86       327

    accuracy                           0.79       458
   macro avg       0.75      0.74      0.74       458
weighted avg       0.79      0.79      0.79       458
```

Fig 11 Bagging testing data classification report

- Accuracy of the training data is 99% and for testing data is 79%.
- The recall and precision of the training data is 1.
- In testing data the Precision and recall for the conservative party is very low and for Labour Party is very good.
- The model couldn't perform well in predicting conservative party supporters.

Ada Boost classification

```
0.8406747891283973
[[230 102]
 [ 68 667]]
              precision    recall  f1-score   support

Conservative       0.77      0.69      0.73       332
      Labour       0.87      0.91      0.89       735

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

Fig 12 Ada boosting classification report

```
0.8078602620087336
[[ 85  45]
 [ 43 285]]
              precision    recall  f1-score   support

Conservative       0.66      0.65      0.66       130
      Labour       0.86      0.87      0.87       328

    accuracy                           0.81       458
   macro avg       0.76      0.76      0.76       458
weighted avg       0.81      0.81      0.81       458
```

Fig 13 Ada boosting training data classification report

- The accuracy of training data is 84% and for testing data is 82%.
- The model has performed very well in predicting labour party as the precision and recall for training and testing data shows very less variation.
- Even though the precision and recall values for conservative party in the testing data is low when it is compared to the performance of the training data the rate of variation is less. Hence the model is good.

Gradient Boost classification

```
0.8809746954076851
[[253  79]
 [ 48 687]]
              precision    recall  f1-score   support

Conservative       0.84      0.76      0.80       332
      Labour       0.90      0.93      0.92       735

    accuracy                           0.88      1067
   macro avg       0.87      0.85      0.86      1067
weighted avg       0.88      0.88      0.88      1067
```

Fig 14 Gradient boosting train data classification report

```
0.8078602620087336
[[ 84  46]
 [ 42 286]]
              precision    recall  f1-score   support

Conservative       0.67      0.65      0.66       130
      Labour       0.86      0.87      0.87       328

    accuracy                           0.81       458
   macro avg       0.76      0.76      0.76       458
weighted avg       0.81      0.81      0.81       458
```

Fig 15 Gradient boosting test data classification report

- The accuracy of the training data is 88% and for testing data is 80%.
- The model performs similar to the model of Ada boost for predicting labour party.
- In the training data for conservative party the precision and recall values are very high but for the testing data the performance has dropped quite low.

- We can see that the model for Adaboosting is better than the model for gradient boosting.

Hyperparameter Tuning

1) Ada Boosting

- AdaBoostClassifier(base_estimator=DecisionTreeClassifier(max_depth=1),
  learning_rate=0.5, n_estimators=40, random_state=1)

```
0.8369259606373008
[[221 111]
 [ 63 672]]
              precision    recall  f1-score   support

Conservative       0.78      0.67      0.72       332
      Labour       0.86      0.91      0.89       735

    accuracy                           0.84      1067
   macro avg       0.82      0.79      0.80      1067
weighted avg       0.83      0.84      0.83      1067
```

Fig 16 Ada boost tuned train data classification report

```
0.8078602620087336
[[ 81  49]
 [ 39 289]]
              precision    recall  f1-score   support

Conservative       0.68      0.62      0.65       130
      Labour       0.86      0.88      0.87       328

    accuracy                           0.81       458
   macro avg       0.77      0.75      0.76       458
weighted avg       0.80      0.81      0.81       458
```

Fig 17 Ada boost tuned test data classification report

- The accuracy of the training data is 84% and for testing data it is 80%.
- The model has not performed well in prediction for the conservative party but did a good job for labour party.
- The model is overfitting the data.

2) Gradient boost

- GradientBoostingClassifier(max_features=0.9, random_state=1, subsample=0.9)

```
0.8894095595126523
[[258  74]
 [ 44 691]]
              precision    recall  f1-score   support

Conservative       0.85      0.78      0.81       332
      Labour       0.90      0.94      0.92       735

    accuracy                           0.89      1067
   macro avg       0.88      0.86      0.87      1067
weighted avg       0.89      0.89      0.89      1067
```

Fig 18 Gradient boost tuned train data classification report

```
0.8100436681222707
[[ 86  44]
 [ 43 285]]
              precision    recall  f1-score   support

Conservative       0.67      0.66      0.66       130
      Labour       0.87      0.87      0.87       328

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

Fig 19 Gradient boost tuned test data classification report

- The accuracy of the training data is 89% and for testing data is 81%.
- The precision and recall in the training data have much variation in the testing data and hence the model is overfitting the data.

After tuning the models the performance has dropped as the training and testing data has a huge difference in the precision and recall values.

Model Comparison

	MODEL	TRAIN ACCURACY	TEST ACCURACY	Unnamed: 3	TRAIN PRECISION	Unnamed: 5	TEST PRECISION	Unnamed: 7	TRAIN RECALL	Unnamed: 9	TEST RECALL
0	NaN	NaN	NaN	Conservative	Labour	Conservative	Labour	Conservative	Labour	Conservative	Labour
1	NAiVE BAYES	80.0	78.0	0.71	0.84	0.61	0.86	0.62	0.89	0.66	0.83
2	BAGGING	100.0	79.0	1	1	0.65	0.85	1	1	0.6	0.87
3	ADA BOOSTING	85.0	82.0	0.78	0.87	0.7	0.87	0.7	0.91	0.6	0.89
4	GRADIENT BOOSTING	88.0	83.0	0.84	0.9	0.73	0.86	0.77	0.93	0.63	0.91
5	ADA BOOSTING TUNED	84.0	81.0	0.78	0.86	0.68	0.86	0.67	0.91	0.62	0.88
6	GRADIENT BOOSTING TUNED	89.0	81.0	0.85	0.9	0.67	0.87	0.789	0.94	0.66	0.87

Fig 20 All model performance details

Final Model Selection

- From the above comparison of the model models except bagging and gradient boosting tuned models all other models have similar accuracy for their training and testing data.
- Precision and recall for the labour party in both training and testing datasets, all the models have performed good.
- Precision and recall for the conservative party in both training and testing datasets, model performances are poor.
- Here recall has importance as making an error in predicting the vote can change the results of the exit poll.
- Every model has poor performance for its test data for predicting conservative party.
- From the overall evaluation of the model Ada boosting is the ideal model.
- Ada boosting has a test accuracy of 85% with comparatively good precision and recall rates.
- The variation of training score and testing scores are not so huge hence overfitting is reduced in the Ada boosting dataset.

ADA Boost model is selected as the final Model.

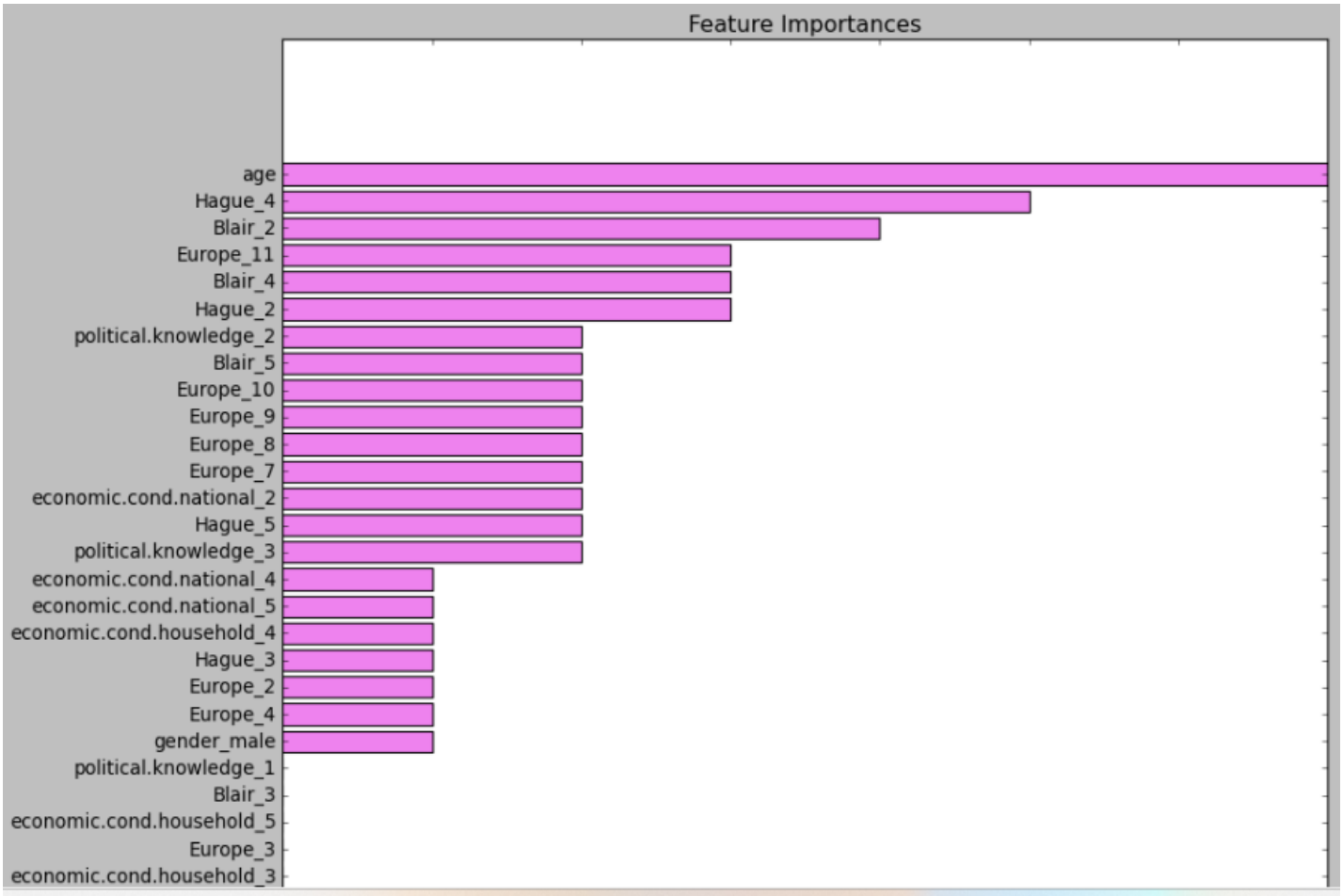


Fig 21 Final model feature importance plot

- Age is the most important parameter of the model.
- All the models show similar performances.
- Hague4, Blair 2 and europe11 has high significance too.

Insights and Recommendations

- Across all models, the performance for predicting the Conservative Party supporters is consistently poor.
- This suggests that there may be underlying factors or variables not captured by the features used in the models that influence Conservative Party support.
- Models generally perform well in predicting Labour Party supporters, with strong precision and recall rates.
- This indicates that the features used in the models are effective in capturing patterns related to Labour Party support.
- Several models exhibit signs of overfitting, where there's a significant difference between training and testing performance metrics.
- Ada Boosting consistently outperforms other models in terms of test accuracy, precision, and recall for both parties.
- Its ability to reduce overfitting and maintain a balance between training and testing performance makes it the preferred choice.
- Recall, particularly for predicting Conservative Party support, is crucial as errors in prediction can significantly impact the results of exit polls.

- Models should prioritize improving recall for Conservative Party supporters to minimize prediction errors and enhance the reliability of election forecasts.
- Fine-tuning hyperparameters or exploring different feature engineering techniques could potentially enhance the model's predictive performance further.

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

EDA

	Name	Speech
0	Roosevelt	On each national day of inauguration since 178...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

Fig 22 Sample dataset

```
Franklin D. Roosevelt's Speech Stats:
Number of Characters: 7651
Number of Words: 1453
Number of Sentences: 32

John F. Kennedy's Speech Stats:
Number of Characters: 7673
Number of Words: 1494
Number of Sentences: 27

Richard Nixon's Speech Stats:
Number of Characters: 10106
Number of Words: 1913
Number of Sentences: 20
```

Text Cleaning

Removing '\\' and '—' from the texts

```
content to stand still. As Americans, we go forward, in the service of our country, by the will of God.n
love, asking His blessing and His help, but knowing that here on earth God's work must truly be our own.n
in one another, sustained by our faith in God who created us, and striving always to serve His purpose.n
```

Fig 23 Sentences after removing '\\' and '--'

Converting to lower case

```
content to stand still. as americans, we go forward, in the service of our country, by the will of god.n
love, asking his blessing and his help, but knowing that here on earth god's work must truly be our own.n
in one another, sustained by our faith in god who created us, and striving always to serve his purpose.n
```

Fig 24 Lower case sentences of each speech

Stopwords Removing

```
Roosevelt Speech (Filtered): ['national', 'day', 'inauguration', 'since', '1789', 'people', 'renewed', 'sense', 'dedication',
'united']
Kennedy Speech (Filtered): ['vice', 'president', 'johnson', 'mr.', 'speaker', 'mr.', 'chief', 'justice', 'president', 'eisenhow
er']
Nixon Speech (Filtered): ['mr.', 'vice', 'president', 'mr.', 'speaker', 'mr.', 'chief', 'justice', 'senator', 'cook']
```

Fig 25 Tokens after removing stopwords

Stemming

```
Roosevelt Speech (Stemmed): ['nation', 'day', 'inaugur', 'sinc', '1789', 'peopl', 'renew', 'sens', 'dedic', 'unit']
Kennedy Speech (Stemmed): ['vice', 'presid', 'johnson', 'mr.', 'speaker', 'mr.', 'chief', 'justic', 'presid', 'eisenhow']
Nixon Speech (Stemmed): ['mr.', 'vice', 'presid', 'mr.', 'speaker', 'mr.', 'chief', 'justic', 'senat', 'cook']
```

Fig 26 Sentences after stemming

Most common words

The three most common words used in all three speeches are:
us : 45
nation : 37
america : 29

Word Cloud

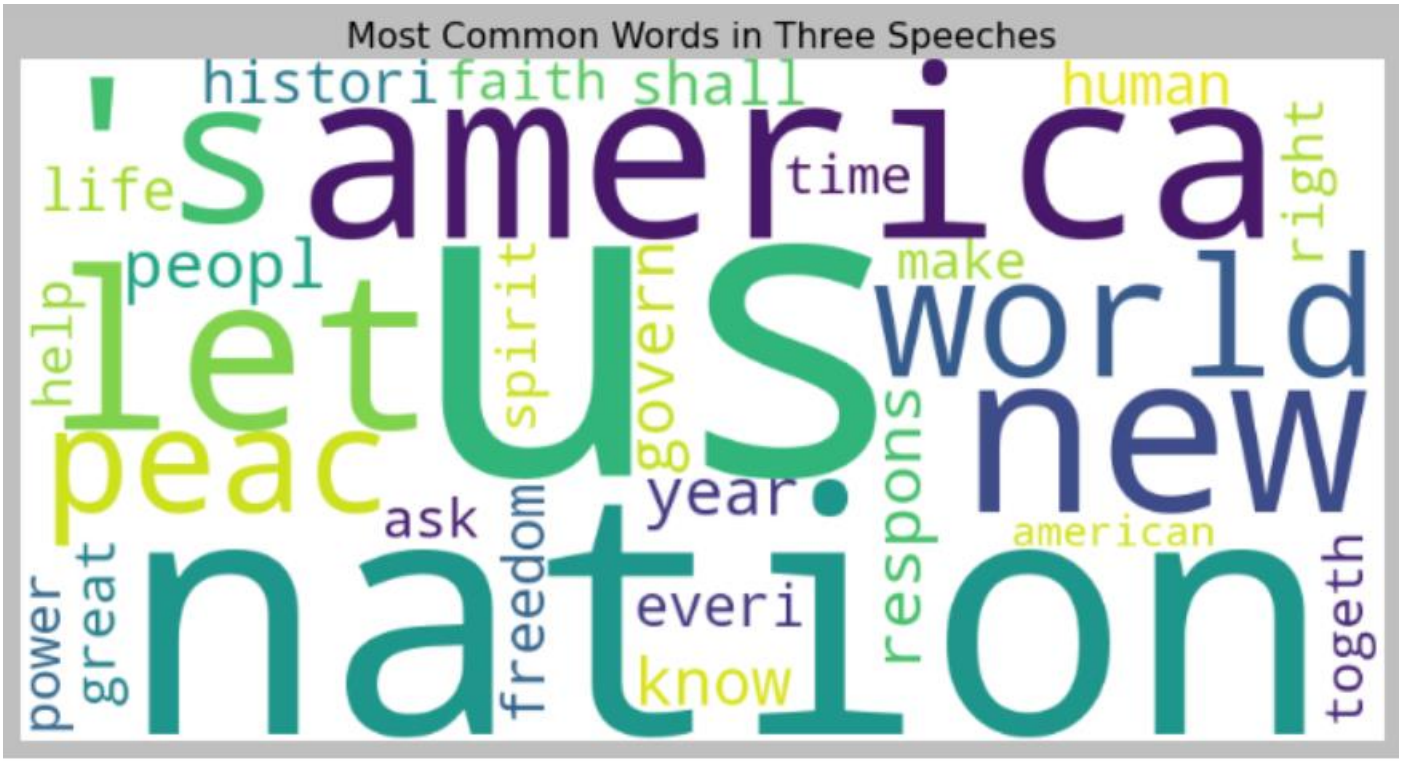


Fig 27 Wordcloud