MACHINE LEARNING -2
PROJECT

BUSINESS REPORT

Prepared By : Sahid
COURSE : DSBA
DATE : 01-09-2024

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Project Overview & Objectives

Problem 1: Predicting Voter Preferences

Overview:

1. **Context**: CNBE, a prominent news channel, aims to provide insightful election coverage using data-driven analysis. A comprehensive survey has been conducted involving 1525 voters, capturing various demographic and socio-economic factors.
2. **Dataset**: The dataset comprises 9 variables including voters' age, gender, assessments of national and household economic conditions, evaluations of political leaders, attitudes toward European integration, and political knowledge.
3. **Approach**:

    - Conduct Exploratory Data Analysis (EDA) to understand the dataset and identify patterns.
    - Pre-process the data by handling missing values, encoding categorical variables, and scaling features.
    - Build multiple machine learning models (KNN, Naive Bayes, Bagging, Boosting) for predicting voter party support.
    - Evaluate and improve the models using appropriate metrics and techniques.
    - Select the final model and provide actionable insights based on model performance.

Objectives:

1. Primary Objective: Develop a predictive model to forecast which political party a voter is likely to support.
2. Secondary Objectives:

    - Gain insights into the factors that influence voters' preferences.
    - Create an accurate and reliable exit poll mechanism.
    - Assist in predicting the overall election outcomes, including which party is likely to secure the majority of seats.
    - Provide strategic recommendations to political parties based on the model's findings.

Problem 2: Analyzing Presidential Speeches

Overview:

1. **Context**: The project focuses on analyzing speeches from three US Presidents (Franklin D. Roosevelt in 1941, John F. Kennedy in 1961, and Richard Nixon in 1973). Presidential speeches play a crucial role in shaping public opinion and policy.
2. **Dataset**: The speeches are sourced from the inaugural corpora in the nltk library. The dataset comprises text files of the speeches which will be analyzed for common words and themes.
3. **Approach**:

    - Perform Exploratory Data Analysis (EDA) to count the characters, words, and sentences in each speech.
    - Clean the text by removing stopwords and applying stemming.
    - Identify the most frequent words used in each speech.
    - Visualize the common words using word clouds to highlight key themes.

Objectives:

1. Primary Objective: Analyze the speeches to identify the most common words used in each President's speech.
2. Secondary Objectives:

    - Understand the recurring themes and focus areas of each President.
    - Provide visual representations (word clouds) to illustrate the predominant words and themes.
    - Offer insights into the rhetoric and priorities during the respective time periods of each President's term.
    - Contribute to historical and linguistic analysis of presidential speeches.

2. Problem 1: Predictive Modeling for Voter Party Support

2.1 Problem Definition & Exploratory Data Analysis (EDA)

Problem Definition

Context: Elections are a fundamental aspect of democratic societies, where accurately predicting voter behavior can immensely influence the strategies adopted by political parties and provide a more nuanced understanding of public opinion. CNBE, a renowned news channel, is dedicated to delivering insightful and data-driven election coverage. To this end, CNBE has conducted a comprehensive survey involving 1525 voters, capturing a wide range of demographic and socio-economic factors such as age, gender, evaluations of national and household economic conditions, assessments of political leaders, attitudes towards European integration, and political knowledge. The rich dataset gathered from this survey serves as the foundation for developing a predictive model aimed at forecasting which political party a voter is likely to support. Such a model will not only enable CNBE to produce accurate exit polls but also offer a reliable prediction of which party is poised to secure the majority of seats, thus providing valuable real-time insights during election coverage.

Exploratory Data Analysis (EDA)

Data Description:

The dataset consists of survey responses from 1525 voters, capturing a range of demographic and socio-economic factors that are potential predictors of political party support. The dataset includes the following variables:

1. **vote**: This is the target variable representing the political party choice of the voter, either Conservative or Labour.
2. **age**: This variable represents the age of the voter in years.
3. **economic.cond.national**: This variable captures the voter's assessment of the current national economic conditions on a scale from 1 to 5, where 1 indicates very poor conditions and 5 indicates excellent conditions.
4. **economic.cond.household**: This variable captures the voter's assessment of the current household economic conditions on a scale from 1 to 5, where 1 indicates very poor conditions and 5 indicates excellent conditions.
5. **Blair**: This variable represents the voter's assessment of the Labour leader (Tony Blair) on a scale from 1 to 5, where 1 indicates a very poor opinion and 5 indicates a very good opinion.
6. **Hague**: This variable represents the voter's assessment of the Conservative leader (William Hague) on a scale from 1 to 5, where 1 indicates a very poor opinion and 5 indicates a very good opinion.
7. **Europe**: This variable measures the respondent's attitudes toward European integration on an 11-point scale, with higher scores representing a more Eurosceptic sentiment.
8. **political.knowledge**: This variable indicates the level of the voter's knowledge about the positions of the political parties on European integration, ranging from 0 to 3.
9. **gender**: This variable represents the gender of the voter, either female or male.

Data Overview :

The dataset has 1525 rows and 10 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

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

(Table 1: Top five rows of the dataset for Problem1)

Descriptive Statistics :

	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
mean	763.000000	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	440.373894	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315
min	1.000000	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	382.000000	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000
50%	763.000000	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000
75%	1144.000000	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000
max	1525.000000	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000

(Table 2: Descriptive statistics of the dataset for Problem1)

Based on the descriptive statistics for the dataset metrics, the following observations can be made:

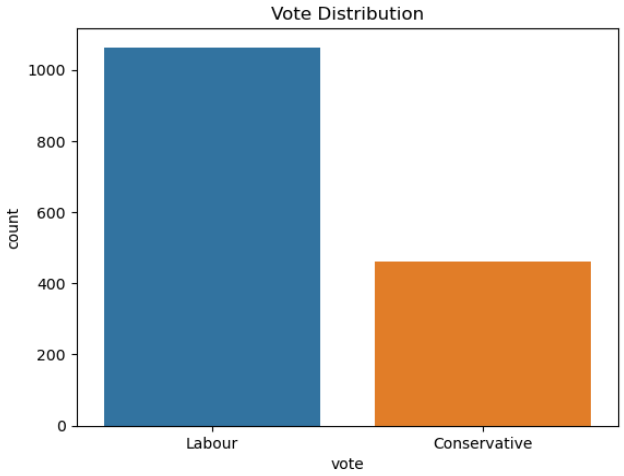
1. **Voter Age Distribution:**
    - The average age of voters is approximately 54 years, with a wide age range from 24 to 93 years. This indicates a diverse voter base with a concentration in middle-aged to older groups, suggesting that campaign strategies might need to address issues pertinent to this age group.
2. **Economic Conditions:**
    - Both national and household economic conditions are rated around the middle of the scale, with moderate variability. Most voters perceive economic conditions as neither extremely positive nor negative, reflecting a general sense of economic stability or uncertainty.
3. **Perceptions of Political Figures:**
    - Blair: The average rating for Blair is relatively high (3.33 out of 5), with most voters rating him positively. This suggests that Blair is generally well-regarded, which could be leveraged in campaigns or policy endorsements.
    - Hague: Hague receives a lower average rating (2.75 out of 5), with greater variability in opinions. This indicates a more polarized view, which may impact his political standing and the effectiveness of strategies aimed at improving his public image.

4. **Views on Europe**:
    - Voters' views on Europe are moderately positive with significant variability. This suggests that European-related policies or stances could be a key issue, but opinions vary widely, necessitating nuanced messaging that resonates with diverse viewpoints.
5. **Political Knowledge:**
    - The average level of political knowledge is low to moderate, with a significant proportion of voters having limited political knowledge. This indicates that educational campaigns or clear, straightforward communication may be necessary to engage and inform voters effectively.

Univariate Analysis

For performing Univariate analysis we will take a look at the Countplots , Boxplots and Histograms of variables such as **'age' 'economic.cond.national' 'economic.cond.household' 'Blair' 'Hague' 'Europe' 'political.knowledge'** which are highly relevant for developing a predictive model to forecast which political party a voter is likely to support.

Observations on vote

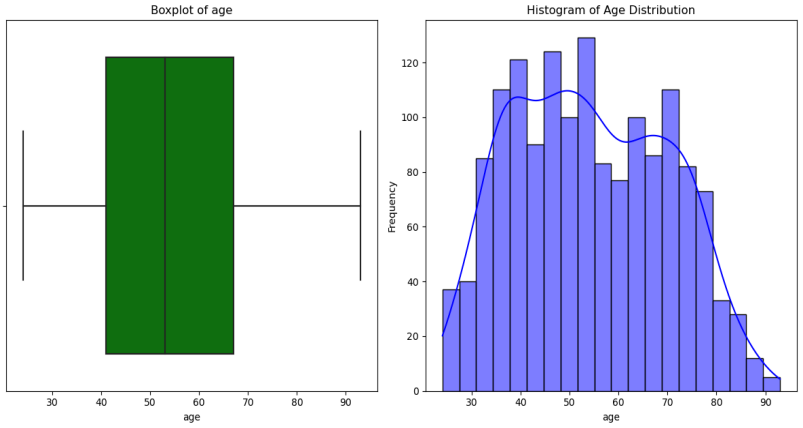


(Fig1 : Plot showing the countplot for age)

Observations:

The vote distribution looks like it's skewed towards 'labour' , almost more than twice as much as 'conservative'

Observations on age

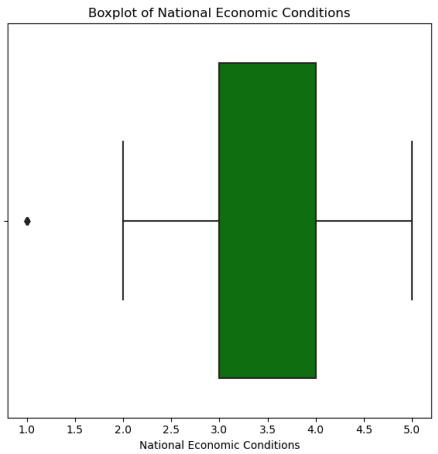


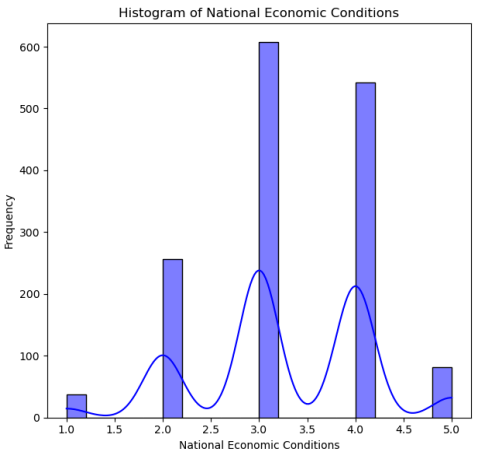
(Fig2 : Plot showing the boxplot and distribution for age)

Observations:

● The Distribution of age is slightly right skewed, almost normally distributed. ● There are no outliers in this variable

Observations on National Economic Conditions



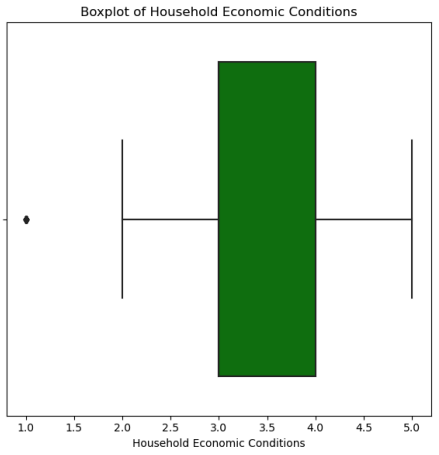


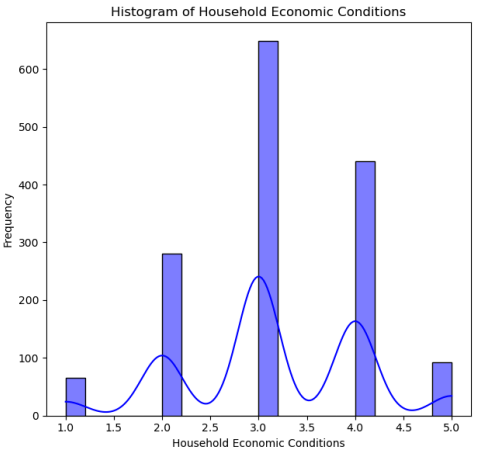
(Fig3 : Plot showing the boxplot and distribution for National Economic Conditions)

Observations:

- The Distribution of National Economic Conditions is slightly left skewed.
- There is one outlier in this variable

Observations on Household Economic Conditions



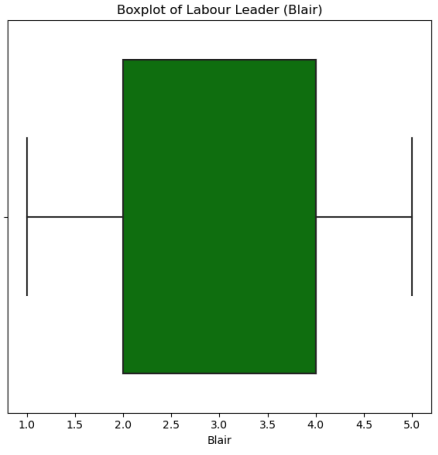


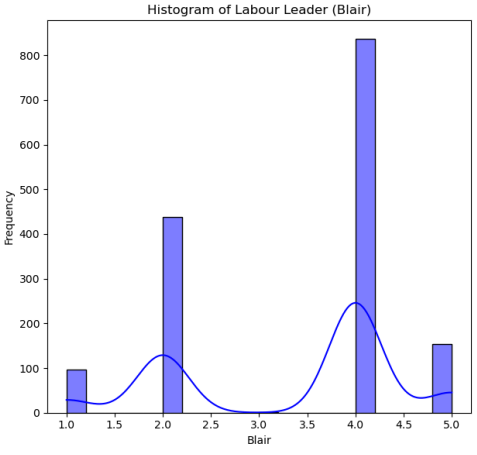
(Fig4 : Plot showing the boxplot and distribution for Household Economic Conditions)

Observations:

- The Distribution of Household Economic Conditions is slightly left skewed.
- There is one outlier in this variable

Observations on Labour Leader (Blair)



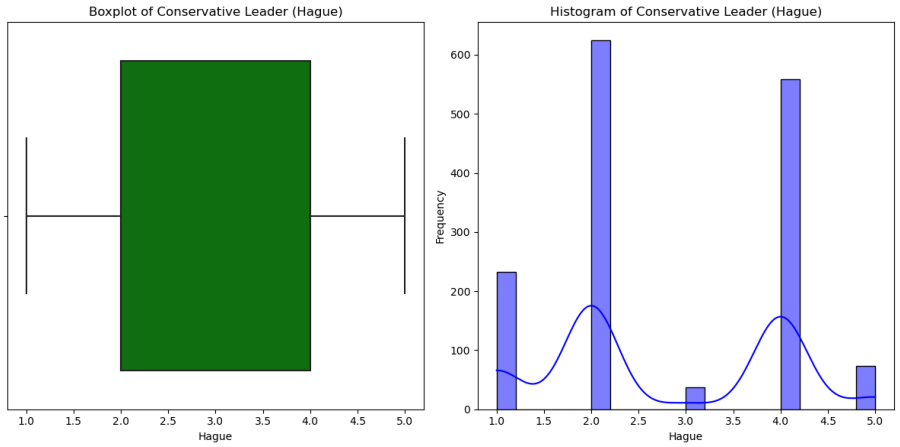


(Fig5 : Plot showing the boxplot and distribution for Labour Leader (Blair))

Observations:

- The Distribution of Labour Leader (Blair) is normal.
- There are no outlier in this variable

Observations on Conservative Leader (Hague)

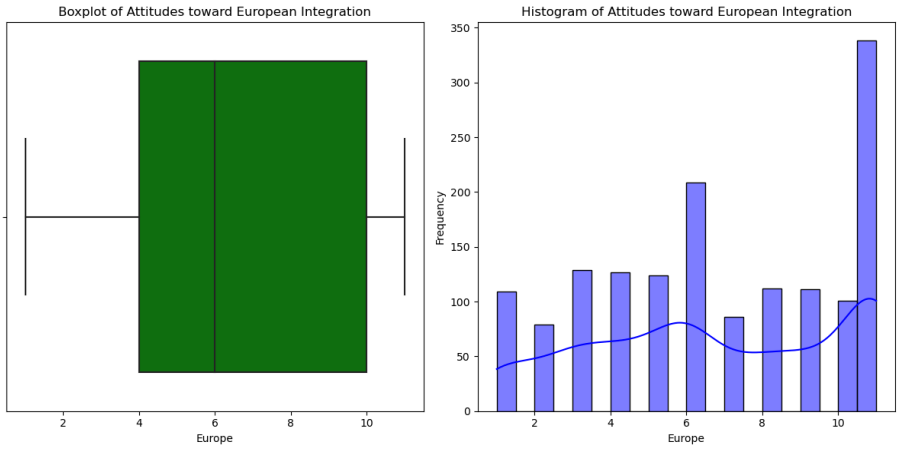


(Fig6 : Plot showing the boxplot and distribution for Conservative Leader (Hague))

Observations:

- The Distribution of Conservative Leader (Hague) is normal.
- There are no outlier in this variable

Observations on Attitudes toward European Integration

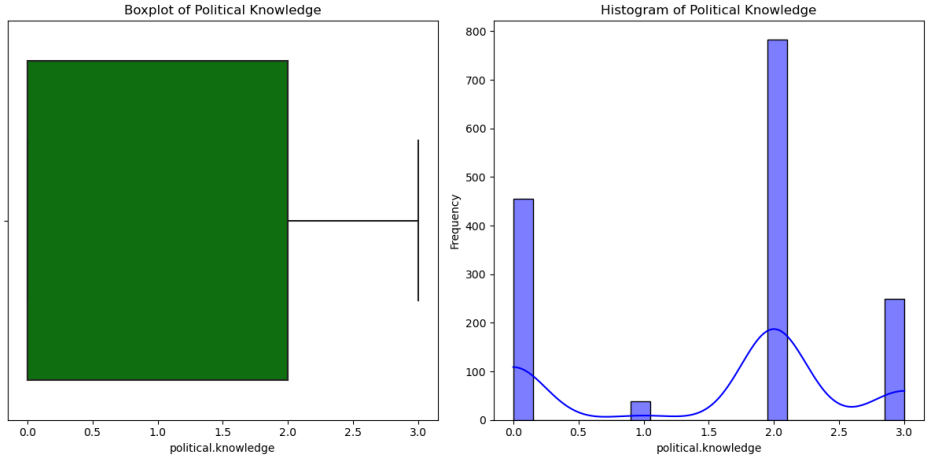


(Fig7 : Plot showing the boxplot and distribution of Attitudes toward European Integration)

Observations:

- The Distribution of Attitudes toward European Integration is skewed to the left
- There are no outlier in this variable

Observations on Political Knowledge

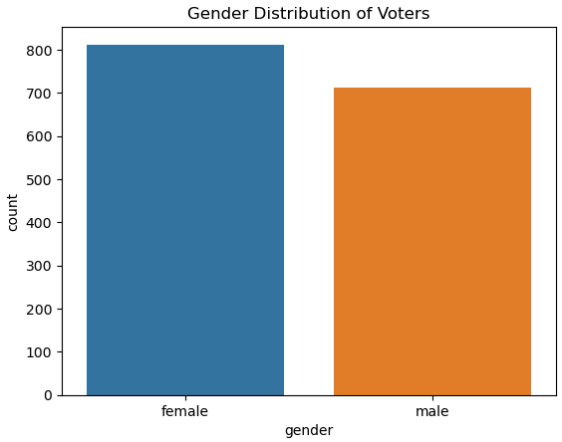


(Fig8 : Plot showing the boxplot and distribution of Political Knowledge)

Observations:

- The Distribution of Political Knowledge is skewed to the right
- There are no outlier in this variable

Observations on Gender Distribution of Voters

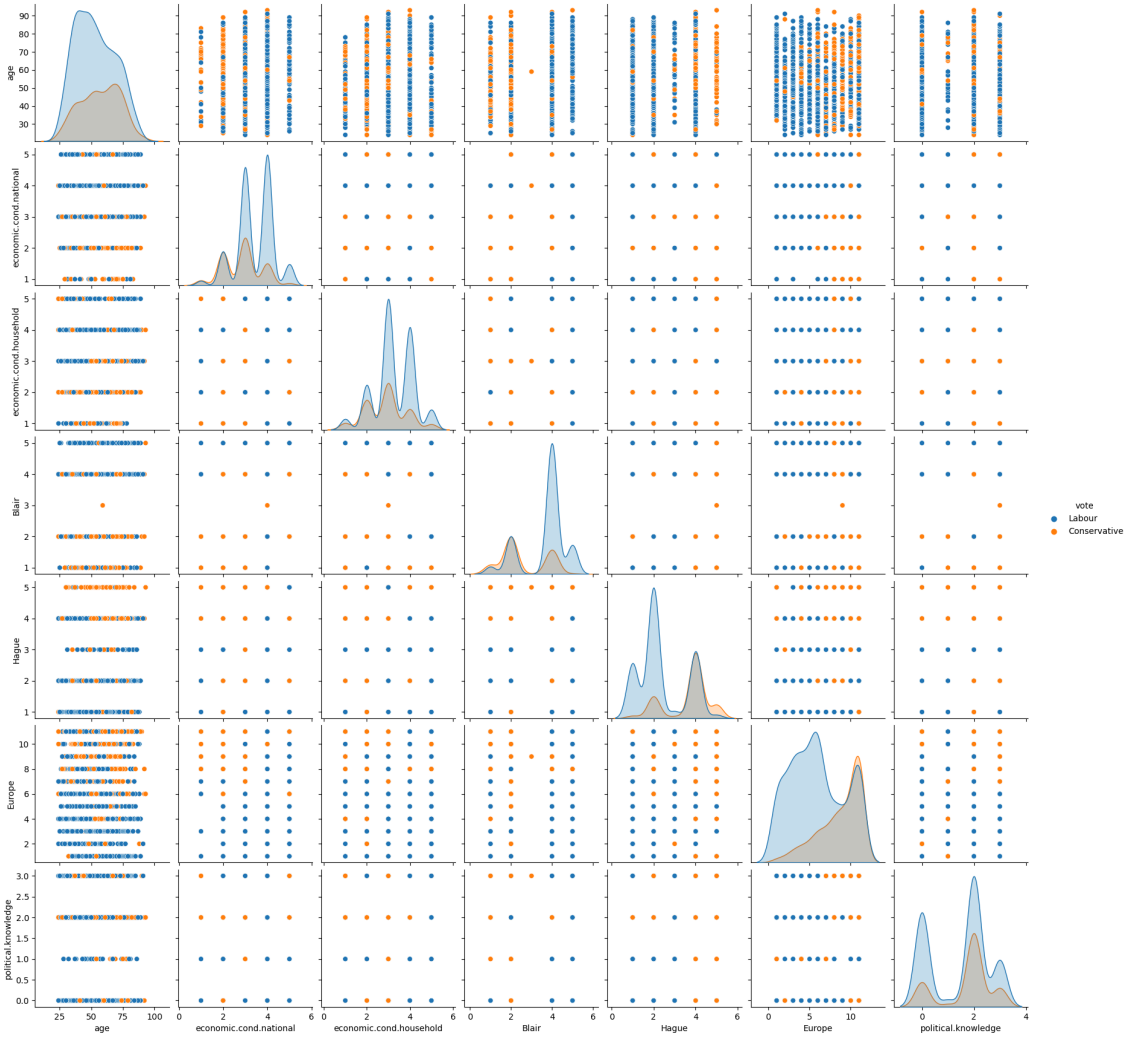


(Fig9 : Plot showing the countplot for Gender)

Observations:

The Gender distribution voters looks like it's skewed towards 'female' , little more than 'male'

Multivariate analysis



(Fig10 : Pair plot showing separation between Labour and Conservative voters.)

Brief analysis of the pairplot:

Age Distribution:

The age distribution shows that both Labour and Conservative voters are spread across a wide age range, but Labour voters seem to have a higher concentration in the younger age groups, while Conservative voters are more evenly spread across ages.

Economic Conditions (National and Household):

The economic.cond.national and economic.cond.household features appear to have a similar distribution across both Labour and Conservative voters, with no strong visible separation between the two groups. However, there are some clusters in these features that could indicate a certain pattern, which might be worth further exploration.

Leader Perceptions (Blair and Hague):

The Blair and Hague features, representing perceptions of the leaders, show more noticeable differences between Labour and Conservative voters. Labour voters generally rate Blair more positively, while Conservative voters tend to rate Hague higher, which aligns with expectations given their party affiliations.

Europe:

The Europe feature also shows some separation between Labour and Conservative voters, though it is less pronounced. This might suggest differing views on Europe between the two groups, but it's not as distinct as the leader perception features.

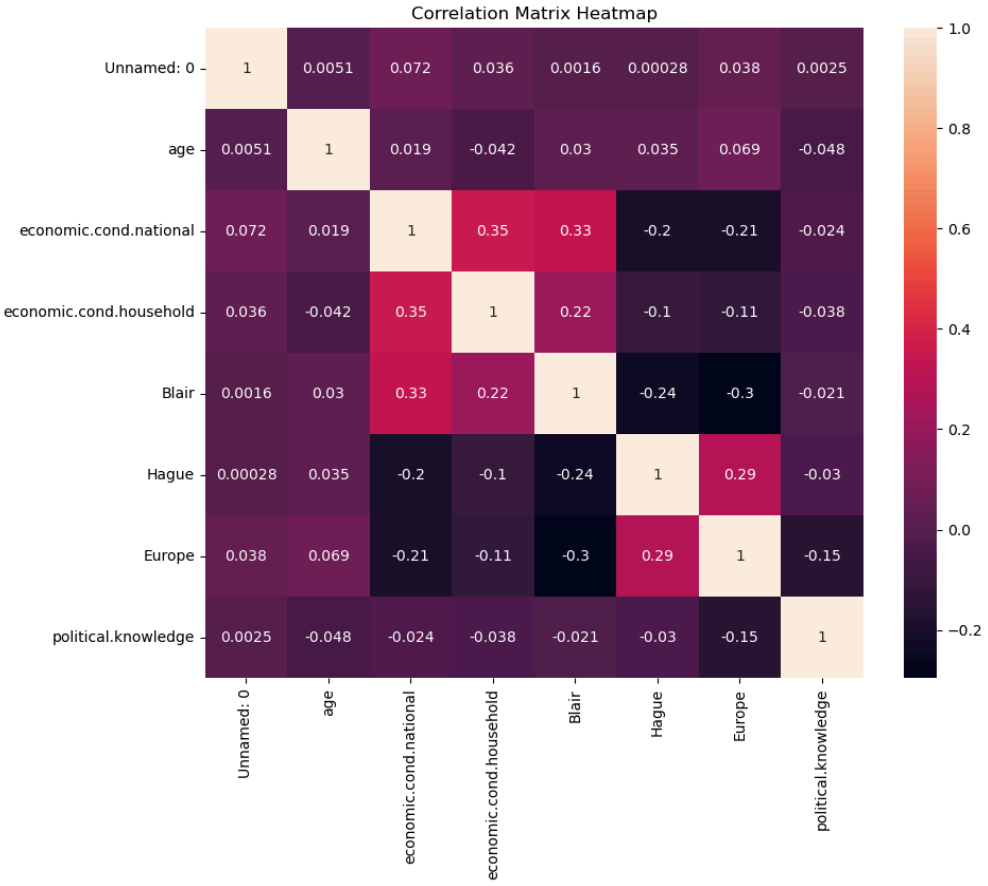
Political Knowledge:

The political.knowledge feature doesn't show a strong separation between Labour and Conservative voters. Both groups seem to be evenly distributed across different levels of political knowledge.

Overall Insights:

The most distinct separation between Labour and Conservative voters is seen in their perceptions of Blair and Hague. Age and economic conditions show some patterns, but they are not as strongly differentiated. Europe and political knowledge features are more evenly distributed across the two voter groups.

Correlation of Numerical Variables



(fig 11 showing heatmap of Correlation of relevant numerical variables)

Observations

Economic Conditions & Leadership:

- Moderate Positive Correlations: economic.cond.national with economic.cond.household (0.346), and Blair with both economic.cond.national (0.327) and economic.cond.household (0.215). This indicates that positive economic perceptions are linked with favorable ratings of Blair.

Leadership Polarization:

- Negative Correlations: Blair with Hague (-0.243) and Europe (-0.296). Voters who favor Blair tend to rate Hague and Europe negatively.

Political Knowledge:

- Weak Correlations: Political knowledge has minimal impact on other variables, with the strongest being a weak negative correlation with Europe (-0.152).

Conclusion:

- Economic Sentiment Influences Leader Ratings: Voter opinions on the economy are tied to their ratings of political leaders.
- Polarization: There is a clear divide between voters supporting Blair versus Hague and Europe.
- Political Knowledge: This factor shows little correlation with other variables, indicating a limited role in shaping voter preferences.

2.2 Data Preprocessing

Missing Values & Outliers: Detection and treatment.

Check for missing values： There are no missing values.

Outlier Detection and Treatment

To ensure the accuracy and reliability of our analysis, we performed outlier detection and treatment on the numerical columns in our dataset. This process helps in mitigating the potential skewing effect of extreme values, ensuring a more robust model.

Methodology :

Detection:

- Outliers were detected using the Interquartile Range (IQR) method.
- First, the 1st quartile (Q1) and the 3rd quartile (Q3) of each numerical column were calculated.
- The IQR was then computed as the difference between Q3 and Q1.
- Outliers were identified as any data points below the lower bound (Q1 - 1.5 * IQR) or above the upper bound (Q3 + 1.5 * IQR).

Treatment:

- Instead of removing outliers, values outside the lower and upper bounds were capped at these bounds.
- This method helps in retaining the data's structure while minimizing the impact of extreme values.

Reporting:

- The count of detected outliers per column was recorded before applying the treatment, showing how many data points were adjusted.

- The treatment ensures that the dataset is free from extreme values, which could otherwise skew the analysis.

Data Encoding Methods

- The target variable vote, originally categorical with labels 'Labour' and 'Conservative', was encoded into binary format for modeling purposes. 'Labour' was mapped to 1, and 'Conservative' was mapped to 0.

- The gender variable, which had categories such as 'Male' and 'Female', was one-hot encoded to convert it into numerical format. This involved creating a binary indicator column, where 'Female' was represented as 1 and 'Male' as 0 (using drop_first=True to avoid multicollinearity).

This transformation ensures that the categorical data is in a format compatible with our analytical and modeling techniques, enhancing the overall effectiveness of the analysis.

Train-Test Split and Feature Scaling

Splitting the Data:

- The dataset was divided into features (predictor variables) and the target variable ('vote'). The features were all columns except 'vote', and the target variable indicated the party preference (Labour or Conservative).
- The data was then split into training and testing sets, with 70% of the data allocated to training and 30% to testing. A fixed random seed was used to ensure that the results are reproducible.

Feature Scaling:

- To normalize the features and ensure that they are on a similar scale, the data was standardized using a scaling method. The training data was scaled first, and then the same scaling parameters were applied to the test data. This ensures consistency and that the test data is processed in the same way as the training data.
- KNN and Naive Bayes benefit from scaling because they rely on distance metrics or probability calculations that assume normally distributed data.Bagging and Boosting models (like Random Forest or XGBoost) do not require scaling, but scaling won't hurt their performance either.If we skip scaling,it might negatively impact the performance of models like KNN.

Summary Statistics:

- After scaling, summary statistics were generated for both the training and testing sets. These statistics were then compared to ensure that the distribution of features in both sets was consistent and representative of the overall dataset. This step helps to verify that the data split has not introduced any significant bias.

(Table 3: Summary statistics for the test & training set for Problem1)

	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set
Unnamed: 0	1067	769.587629	445.810814	1	379.5	768	1157.5	1525	458	747.652838	427.524726	4	388	758	1109.75	1520
age	1067	54.561387	15.804764	24	41	54	68	93	458	53.299127	15.472148	24	41	52	65	92
economic.cond.na	1067	3.264292	0.894778	1	3	3	4	5	458	3.203057	0.847336	1	3	3	4	5
economic.cond.ho	1067	3.145267	0.92621	1	3	3	4	5	458	3.128821	0.939528	1	3	3	4	5
Blair	1067	3.353327	1.174221	1	2	4	4	5	458	3.290393	1.176333	1	2	4	4	5
Hague	1067	2.763824	1.226078	1	2	2	4	5	458	2.707424	1.241862	1	2	2	4	5
Europe	1067	6.680412	3.304443	1	4	6	10	11	458	6.840611	3.282259	1	4	7	10	11
political.knowledge	1067	1.561387	1.065724	0	0	2	2	3	458	1.497817	1.123158	0	0	2	2	3

Observations

- Consistency: The summary statistics between the training and testing sets are quite consistent, indicating that the data split was well-balanced. This suggests that the model trained on the training set should perform similarly on the testing set.
- Key Variables: The slight variations in means and distributions across some variables, such as age and perception of Europe, are minor and should not significantly impact the model's performance. Both sets adequately represent the overall dataset.

2.3 Model Building

Choice of Metrics

In the model-building phase, selecting appropriate metrics is crucial for evaluating model performance and ensuring that the model meets the business objectives. For this project, the following metrics were chosen:

1. **Accuracy:** Measures the proportion of correctly classified instances among all instances. It provides a general sense of how well the model performs but can be misleading in imbalanced datasets.
2. **Precision and Recall:**
    - **Precision** evaluates the proportion of true positive predictions out of all positive predictions made by the model, which is useful when the cost of false positives is high.

    - **Recall** assesses the proportion of true positives out of all actual positives, which is critical when the cost of false negatives is high.
3. **F1 Score:** The harmonic mean of precision and recall, providing a single metric that balances both aspects. It is especially useful for imbalanced datasets where both false positives and false negatives are of concern.
4. **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** Measures the model's ability to discriminate between classes, with a value of 1 indicating perfect separation and 0.5 indicating no discrimination.
5. **Confusion Matrix:** Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives, helping to understand the model's classification performance.

These metrics were selected to provide a comprehensive evaluation of model performance, ensuring that the chosen model aligns with the objectives of accurate prediction and effective classification of voter support.

Model Construction:

1. **K-Nearest Neighbors (KNN)**
    - **Description:** KNN is a non-parametric method used for classification and regression. It classifies a data point based on how its neighbors are classified.
    - **Configuration:** The number of neighbors (k) and the distance metric are key parameters. For this project, the default parameters were used initially.
2. **Naive Bayes**
    - **Description:** Naive Bayes is a probabilistic classifier based on Bayes' Theorem with the assumption of independence between features.
    - **Configuration:** Typically involves choosing the variant of Naive Bayes (e.g., Gaussian, Multinomial) based on the data characteristics. The Gaussian variant was used for continuous features.
3. **Bagging (Bootstrap Aggregating)**
    - **Description:** Bagging improves the stability and accuracy of machine learning algorithms by combining predictions from multiple models trained on different subsets of the data.
    - **Configuration:** The key parameters include the number of estimators, maximum samples, and maximum features. Various combinations were tested to find the optimal configuration.
4. **Boosting**
    - **Description:** Boosting combines multiple weak learners to create a strong learner. It sequentially trains models to correct errors of the previous models.
    - **Configuration:** Important parameters include the number of estimators, learning rate, and maximum depth of trees. Different values were explored to optimize model performance.

Summary

- **Selection Criteria:** Models were selected based on their suitability for classification tasks and their ability to handle the dataset's characteristics. Hyperparameters were tuned to optimize performance.
- **Implementation:** Models were implemented using standard libraries and configured based on best practices and parameter tuning results.

2.4 Model Performance Evaluation

K-Nearest Neighbors (KNN) Model Evaluation

Training Data Evaluation:

- **Precision, Recall, F1-Score:**
  - **Class 0:** Precision: 0.80, Recall: 0.75, F1-Score: 0.77
  - **Class 1:** Precision: 0.89, Recall: 0.92, F1-Score: 0.90
  - **Overall Accuracy:** 0.86
  - **Macro Average:** Precision: 0.84, Recall: 0.83, F1-Score: 0.84
  - **Weighted Average:** Precision: 0.86, Recall: 0.86, F1-Score: 0.86
- **Confusion Matrix:**
  - True Positives (Class 1): 673
  - True Negatives (Class 0): 248
  - False Positives: 84
  - False Negatives: 62

Test Data Evaluation:

- **Precision, Recall, F1-Score:**
  - **Class 0:** Precision: 0.69, Recall: 0.71, F1-Score: 0.70
  - **Class 1:** Precision: 0.88, Recall: 0.87, F1-Score: 0.88
  - **Overall Accuracy:** 0.83
  - **Macro Average:** Precision: 0.78, Recall: 0.79, F1-Score: 0.79
  - **Weighted Average:** Precision: 0.83, Recall: 0.83, F1-Score: 0.83
- **Confusion Matrix:**
  - True Positives (Class 1): 286
  - True Negatives (Class 0): 92
  - False Positives: 38
  - False Negatives: 42

Key Observations:

Training Data:

- The KNN model performs well on the training data with an overall accuracy of 86%.
- The recall is slightly lower for Class 0 (Labour) compared to Class 1 (Conservative), indicating that the model is slightly less effective at identifying all Labour voters.

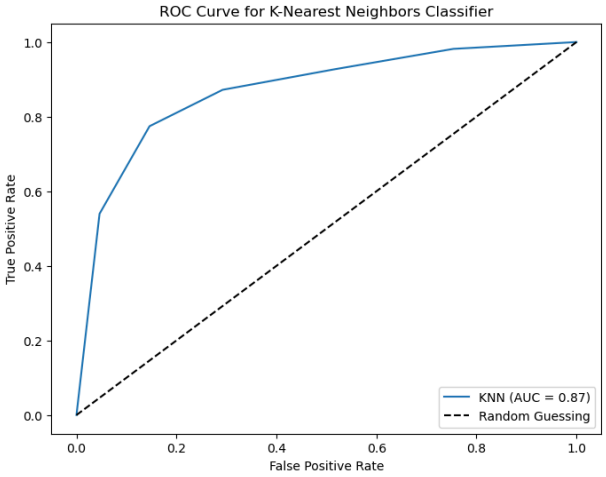
Test Data:

- The accuracy on the test data is 83%, which is slightly lower than on the training data, suggesting some generalization but not significant overfitting.
- The precision and recall for Class 1 (Conservative) are notably higher than for Class 0 (Labour), indicating that the model is better at predicting Conservative voters.

- The model has a tendency to misclassify some Labour voters as Conservative (as seen in the confusion matrix with 38 false positives in the test data).

ROC-AUC Score:

- The ROC-AUC score of 0.87 confirms that the KNN model has a strong ability to differentiate between the positive and negative classes. This high score suggests that the model is effective in distinguishing between the two classes and performs well in binary classification tasks.



ROC Curve:

The ROC curve illustrates the model's performance across different classification thresholds, showing that the KNN model achieves a good balance between the True Positive Rate (TPR) and the False Positive Rate (FPR). The curve's position close to the top-left corner indicates that the model is successful in minimizing false positives while maximizing true positives.

(fig 12 showing ROC curve of the model 'KNN's performance)

Naive Bayes Classifier Model Evaluation

Training Data Evaluation

- **Precision, Recall, and F1-Score:**
  - **Class 0 (Negative):** Precision = 0.74, Recall = 0.72, F1-Score = 0.73
  - **Class 1 (Positive):** Precision = 0.88, Recall = 0.88, F1-Score = 0.88
  - **Overall Accuracy:** 83%
- **Confusion Matrix:**
  - True Negatives (TN): 240
  - False Positives (FP): 92
  - False Negatives (FN): 86
  - True Positives (TP): 649

Evaluation on Test Data

- **Precision, Recall, and F1-Score:**
  - **Class 0 (Negative):** Precision = 0.68, Recall = 0.72, F1-Score = 0.70
  - **Class 1 (Positive):** Precision = 0.89, Recall = 0.87, F1-Score = 0.88

    - **Overall Accuracy:** 83%

- **Confusion Matrix:**
    - True Negatives (TN): 94
    - False Positives (FP): 36
    - False Negatives (FN): 44
    - True Positives (TP): 284

ROC-AUC Score

- **ROC-AUC Score:** 0.88

  The ROC-AUC score indicates that the Naive Bayes model performs very well in distinguishing between the two classes. A score of 0.88 reflects strong classification performance, showing the model's effectiveness at distinguishing between positive and negative classes.

ROC Curve

- The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various threshold settings. The curve's position indicates that the Naive Bayes model maintains a good balance between sensitivity and specificity, corroborated by its high ROC-AUC score.

Key Observations:

Training Data:

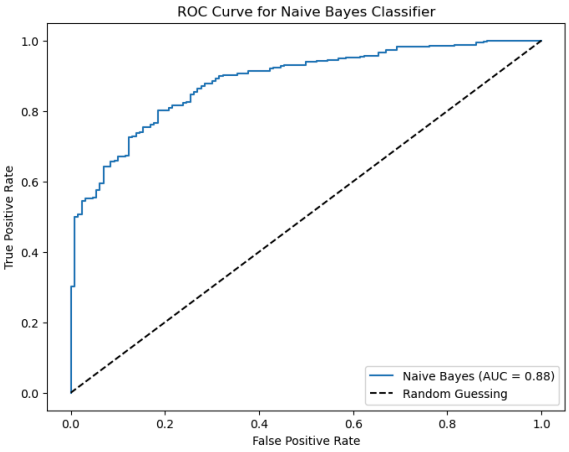
- The Naive Bayes model achieves an accuracy of 83% on the training data.
- The recall for both classes is equal at 88%, indicating the model's consistency in identifying true positives for both Labour and Conservative voters.
- Precision for Class 0 (Labour) is slightly lower at 0.74, reflecting some misclassification as Conservative voters.

Test Data:

- The model maintains an accuracy of 83% on the test data, showing good generalization.
- Similar to the training data, the model is better at predicting Conservative voters, with higher precision and recall for Class 1.
- The model has a comparable performance for Labour voters on the test data, with slightly better recall (0.72) than precision (0.68).

ROC-AUC Score:

- The high ROC-AUC score of 0.88 underscores the model's strong ability to differentiate between positive and negative classes, indicating reliable classification performance.



ROC Curve:

The ROC curve demonstrates that the Naive Bayes model effectively balances between true positives and false positives across different thresholds, confirming its strong performance.

(fig 13 showing ROC curve of the model ' Naive Bayes's performance)

Comparison with KNN:

Both models achieve the same overall accuracy on the test data (83%). KNN shows slightly higher precision and recall for Class 1 (Conservative) on both training and test data, making it marginally better at predicting Conservative voters. Naive Bayes performs similarly but with a more balanced recall for Labour voters compared to KNN, although its precision for Labour voters is slightly lower.

Bagging Classifier Model Evaluation

Training Data:

- **Accuracy:** The Bagging Classifier achieved an accuracy of **98%** on the training data, demonstrating excellent performance in identifying both classes.
- **Precision, Recall, and F1-Score:**
  - **Class 0 (Labour):** Precision = 0.96, Recall = 0.99, F1-Score = 0.97.
  - **Class 1 (Conservative):** Precision = 0.99, Recall = 0.98, F1-Score = 0.99.
- **Confusion Matrix:**
  - True Negatives (TN): 328
  - False Positives (FP): 4
  - False Negatives (FN): 14
  - True Positives (TP): 721

These results indicate that the model is highly effective in correctly classifying both classes with minimal errors on the training data.

Test Data:

- **Accuracy:** The accuracy on the test data dropped to **81%**, which suggests some degree of overfitting compared to the training performance.
- **Precision, Recall, and F1-Score:**
    - **Class 0 (Labour):** Precision = 0.66, Recall = 0.70, F1-Score = 0.68.
    - **Class 1 (Conservative):** Precision = 0.88, Recall = 0.86, F1-Score = 0.87.
- **Confusion Matrix:**
    - True Negatives (TN): 91
    - False Positives (FP): 39
    - False Negatives (FN): 46
    - True Positives (TP): 282

While the model performs well overall, the drop in precision and recall for Class 0 on the test data indicates some challenges in correctly identifying Labour voters, leading to more misclassifications compared to the training data.

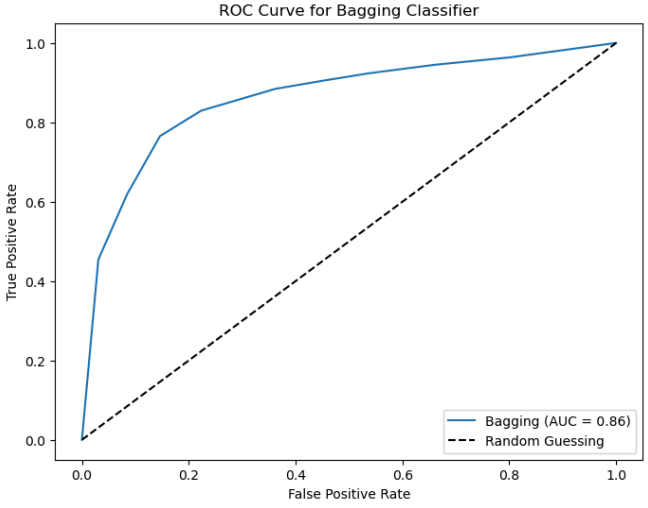
ROC-AUC Score:

- The ROC-AUC score of **0.86** demonstrates the model's good ability to distinguish between the two classes. However, it indicates that there is still room for improvement, particularly when compared to models with higher ROC-AUC scores.

Key Observations:

- The Bagging Classifier exhibits near-perfect performance on the training data, which may suggest overfitting since the test data performance drops significantly.
- The model's precision and recall for predicting Conservative voters (Class 1) remains relatively high, but its ability to correctly identify Labour voters (Class 0) is less robust on the test data.
- The higher number of false negatives and false positives on the test data further supports the observation of potential overfitting.

The overall evaluation of the Bagging Classifier shows that while it performs well, especially in classifying Conservative voters, it does not generalize as effectively to the test data, leading to a decrease in overall accuracy and some challenges in predicting Labour voters.



ROC Curve:

The ROC curve for the Bagging model shows a reasonable trade-off between true positive rate and false positive rate, demonstrating that the model has a good overall performance, but there is room for improvement compared to models with higher ROC-AUC scores.

(fig 14 showing ROC curve of the ‘ Bagging ’ model’s performance)

Comparison with KNN and Naive Bayes:

Bagging achieves superior performance on the training data, but this comes at the cost of overfitting, as evidenced by the drop in test accuracy. KNN and Naive Bayes exhibit more consistent performance across both training and test data, with KNN slightly outperforming in precision and recall for Conservative voters. Bagging does not provide a significant advantage over KNN and Naive Bayes in test data performance, making the latter two more reliable for generalization.

Boosting Classifier Model Evaluation

Train Data

- **Precision, Recall, and F1-Score:**
  - **Class 0 (Negative):** Precision = 0.84, Recall = 0.79, F1-Score = 0.81
  - **Class 1 (Positive):** Precision = 0.91, Recall = 0.93, F1-Score = 0.92
  - **Overall Accuracy:** 89%
- **Confusion Matrix:**
  - **True Negatives (TN):** 262
  - **False Positives (FP):** 70
  - **False Negatives (FN):** 51
  - **True Positives (TP):** 684

Test Data

- **Precision, Recall, and F1-Score:**
  - **Class 0 (Negative):** Precision = 0.69, Recall = 0.74, F1-Score = 0.71
  - **Class 1 (Positive):** Precision = 0.89, Recall = 0.87, F1-Score = 0.88
  - **Overall Accuracy:** 83%
- **Confusion Matrix:**
  - **True Negatives (TN):** 96
  - **False Positives (FP):** 34
  - **False Negatives (FN):** 43
  - **True Positives (TP):** 285

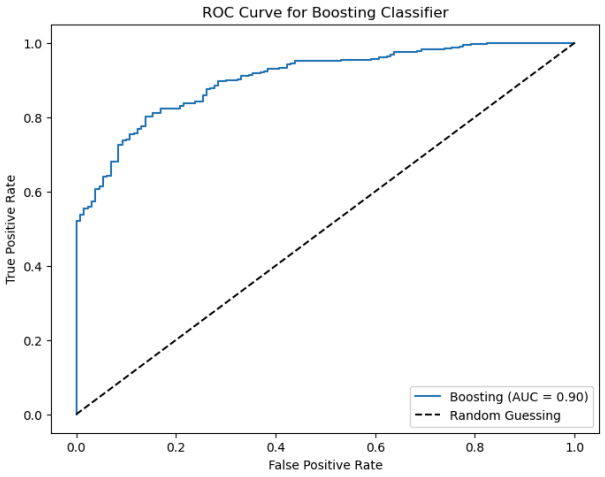
ROC-AUC Score

- **ROC-AUC Score:** 0.904
- The ROC-AUC score of 0.904 indicates that the Boosting model performs well in distinguishing between the two classes, showing strong predictive ability.

Key Observations:

- **Training Data:**
  - The Boosting model demonstrates strong performance on the training data with high precision, recall, and F1-scores for both classes, resulting in an overall accuracy of 89%.

    - The confusion matrix indicates a few more false positives and false negatives than desired, suggesting that while the model is effective, there is still room for improvement in reducing misclassifications.
- **Test Data:**
    - The model maintains good performance on the test data with an overall accuracy of 83%, slightly lower than the training data, indicating a decent balance between learning and generalization.
    - Precision and recall for Class 1 (Positive) remain high, but there is a slight decrease in the model's ability to predict Class 0 (Negative) accurately, with lower precision (0.69) and recall (0.74).
- **ROC-AUC Score:**
    - The high ROC-AUC score of 0.904 reflects strong performance in distinguishing between the classes, making the Boosting model a reliable option for predictive tasks.



ROC Curve:

The ROC curve for the Boosting model demonstrates a favorable trade-off between the true positive rate and false positive rate, indicating that the model is highly effective at distinguishing between positive and negative classes.

(fig 15 showing ROC curve of the ‘ Boosting’ model’s performance)

Comparison with Other Models:

- Boosting provides a balanced and strong performance, with good generalization from the training data to the test data, similar to KNN and Naive Bayes but with slightly higher overall accuracy.
- KNN and Naive Bayes show slightly lower accuracy but more consistent recall for Class 0 (Labour).
- Bagging demonstrated strong training performance but suffered from overfitting, making Boosting a better choice for this dataset.

2.5 Model Performance Improvement

Tuning Bagging and Boosting Models

To enhance the performance of our predictive models, we conducted hyperparameter tuning using Grid Search for both Bagging and Boosting models. The following sections detail the tuning process and the resulting improvements in model performance.

Bagging Model

Before Tuning:

- **Train Data:**
  - **Precision:** 0.96 (Class 0), 0.99 (Class 1)
  - **Recall:** 0.99 (Class 0), 0.98 (Class 1)
  - **F1-Score:** 0.97 (Class 0), 0.99 (Class 1)
  - **Accuracy:** 0.98
  - **ROC-AUC Score:** 0.86
- **Test Data:**
  - **Precision:** 0.66 (Class 0), 0.88 (Class 1)
  - **Recall:** 0.70 (Class 0), 0.86 (Class 1)
  - **F1-Score:** 0.68 (Class 0), 0.87 (Class 1)
  - **Accuracy:** 0.81
  - **ROC-AUC Score:** 0.90

After Tuning:

- **Best Parameters:** {'max_features': 0.5, 'max_samples': 0.5, 'n_estimators': 200}
- **Train Data:**
  - **Precision:** 0.96 (Class 0), 0.92 (Class 1)
  - **Recall:** 0.82 (Class 0), 0.99 (Class 1)
  - **F1-Score:** 0.88 (Class 0), 0.95 (Class 1)
  - **Accuracy:** 0.93
  - **ROC-AUC Score:** 0.89
- **Test Data:**
  - **Precision:** 0.70 (Class 0), 0.84 (Class 1)
  - **Recall:** 0.57 (Class 0), 0.91 (Class 1)
  - **F1-Score:** 0.63 (Class 0), 0.87 (Class 1)
  - **Accuracy:** 0.81
  - **ROC-AUC Score:** 0.89

Analysis: Tuning improved the Bagging model's performance on the training data by increasing the precision for Class 1 and achieving higher accuracy. However, the ROC-AUC score on the test data slightly decreased compared to the pre-tuning model. The recall for Class 0 on the test data decreased, indicating a trade-off between precision and recall.

Boosting Model

Before Tuning:

- **Train Data:**
  - **Precision:** 0.84 (Class 0), 0.91 (Class 1)
  - **Recall:** 0.79 (Class 0), 0.93 (Class 1)
  - **F1-Score:** 0.81 (Class 0), 0.92 (Class 1)
  - **Accuracy:** 0.89
  - **ROC-AUC Score:** 0.90
- **Test Data:**
  - **Precision:** 0.69 (Class 0), 0.89 (Class 1)
  - **Recall:** 0.74 (Class 0), 0.87 (Class 1)
  - **F1-Score:** 0.71 (Class 0), 0.88 (Class 1)
  - **Accuracy:** 0.83
  - **ROC-AUC Score:** 0.90

After Tuning:

- **Best Parameters:** {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50}
- **Train Data:**
    - **Precision:** 0.83 (Class 0), 0.89 (Class 1)
    - **Recall:** 0.75 (Class 0), 0.93 (Class 1)
    - **F1-Score:** 0.79 (Class 0), 0.91 (Class 1)
    - **Accuracy:** 0.87
    - **ROC-AUC Score:** 0.91
- **Test Data:**
    - **Precision:** 0.70 (Class 0), 0.90 (Class 1)
    - **Recall:** 0.75 (Class 0), 0.87 (Class 1)
    - **F1-Score:** 0.72 (Class 0), 0.88 (Class 1)
    - **Accuracy:** 0.84
    - **ROC-AUC Score:** 0.91

Analysis: Tuning the Boosting model led to improvements in the ROC-AUC score for both training and test data, indicating a better overall model performance in distinguishing between classes. Precision and recall for Class 1 increased, and the model maintained a balanced performance with higher accuracy.

Conclusion: After tuning, the Boosting model demonstrated superior performance over the Bagging model, as indicated by the higher ROC-AUC scores. While both models showed improvements, Boosting proved to be more effective in the context of this dataset.

2.6 Final Model Selection

Comparative Analysis of All Models

To determine the best model for predicting voter support, we evaluated multiple models, focusing on their ROC-AUC scores, precision, recall, and overall accuracy. The models considered were Bagging and Boosting, with the following results:

- Bagging Model:
    - ROC-AUC Score: 0.89
    - Train Accuracy: 93%
    - Test Accuracy: 81%
    - Train F1-Score: 0.93
    - Test F1-Score: 0.80
- Boosting Model:
    - ROC-AUC Score: 0.91
    - Train Accuracy: 87%
    - Test Accuracy: 84%
    - Train F1-Score: 0.87
    - Test F1-Score: 0.84

Final Model Selection with Justification

After comparing the models, the **Boosting model** was selected as the final model for the following reasons:

1. Higher ROC-AUC Score: The Boosting model achieved a ROC-AUC score of 0.91, which is higher than the Bagging model's score of 0.89. This indicates that the Boosting model has a better ability to distinguish between the two classes (Labour and Conservative voters).
2. Balanced Performance: The Boosting model showed consistent performance across both the training and test datasets, with an accuracy of 84% on the test data compared to the Bagging model's 81%. This suggests that the Boosting model generalizes better to unseen data.
3. Precision and Recall: The Boosting model also exhibited better precision and recall on the test data, particularly for the minority class (Labour voters), ensuring more reliable predictions.

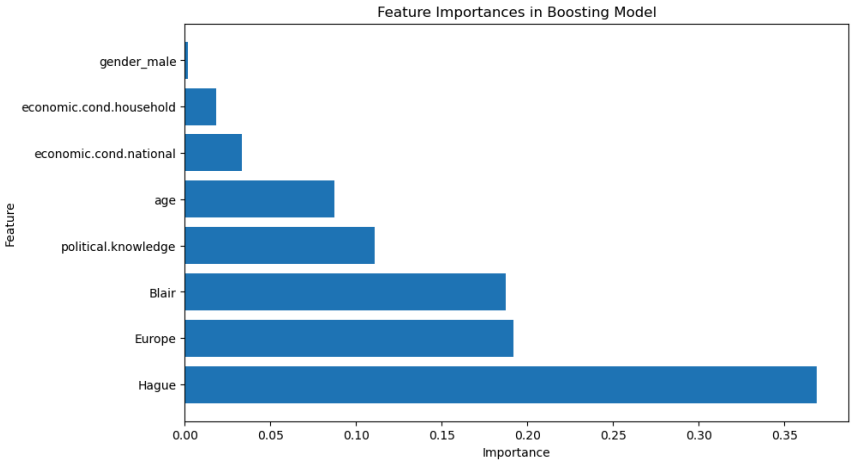
Given these points, the Boosting model is preferred for its overall robustness and ability to capture the complexities in the data.

Feature Importance and Inferences

Using the Boosting model, we assessed feature importance to understand which factors most influenced voter predictions. Below are the key features and their importance:

Feature	Importance
Hague	0.369
Europe	0.192
Blair	0.187
Political Knowledge	0.111
Age	0.087
Economic Cond. National	0.033
Economic Cond. Household	0.019
Gender (Male)	0.002

(Table 4: Table showing feature importance scores derived from the Boosting model)



(fig 16 showing feature importance scores of the ' Boosting' model's performance)

Inferences:

1. Leadership Perception: The perceptions of leaders (Hague and Blair) were the most significant factors in determining voter preference, emphasizing the importance of political leadership in elections.
2. Europe: Opinions on Europe also played a substantial role, reflecting the relevance of international issues in voter decision-making.
3. Political Knowledge: Voters with higher political knowledge were more likely to make informed decisions, highlighting the impact of voter education on election outcomes.

2.7 Business Insights & Recommendations

Conclusions from Model Comparisons

- Boosting vs. Bagging: After tuning, the Boosting model outperformed the Bagging model with a higher ROC-AUC score (0.91 vs. 0.89). This indicates that the Boosting model is better at distinguishing between voters who support Labour and those who support Conservative.
- Model Performance: The Boosting model showed stronger generalization capabilities, reflected in its consistent performance across both training and test datasets. It achieved an accuracy of 84% on the test data, with an ROC-AUC score that suggests a robust ability to classify voters accurately.
- Feature Importance Analysis: The most influential features in predicting voter preference were related to perceptions of political leaders (Hague, Blair), stance on Europe, and political knowledge. Age and economic conditions were moderately influential, while gender had a minimal impact.

Strategic Recommendations

1. Focus on Leader Perceptions:
    - Engage with Public Perceptions: Since voters' perceptions of leaders like Hague and Blair are the most critical factors influencing their political choices, political campaigns should prioritize shaping these perceptions. This could involve targeted media campaigns, public engagements, and messaging that highlights leadership qualities and achievements.
2. Address European Issues:
    - Strategic Messaging on Europe: The stance on Europe is a significant factor for voters. Campaigns should clearly articulate their position on European matters, addressing voter concerns and aligning messaging with the electorate's values and interests regarding Europe.
3. Enhance Political Knowledge:
    - Educational Campaigns: With political knowledge being an important predictor, increasing voter awareness through educational campaigns could be beneficial. Providing clear and accessible information about political issues and party policies may help in persuading undecided voters or reinforcing the support of existing voters.
4. Target Younger Voters:
    - Youth Engagement: Given that age plays a role in voter preference, campaigns should consider tailored strategies to engage younger voters, such as through social media, campus events, and youth-focused policy discussions.
5. Economic Messaging:
    - Economic Conditions: Although economic conditions are less influential than leader perceptions and Europe, they still play a role. Campaigns should ensure that their

economic policies are communicated effectively, emphasizing how they will positively impact both national and household economies.

6. Minimize Gender Bias:
    - Gender-Neutral Messaging: Given the minimal influence of gender in the model, campaigns should aim to craft messages that resonate across gender lines, ensuring inclusivity and broad appeal.

3. Problem 2: Speech Analysis Using NLTK

3.1 Problem Definition & Exploratory Data Analysis (EDA)

Problem Definition

Context : Presidential speeches have long been a crucial tool for shaping public opinion and setting the agenda for a nation's policies and initiatives. Each speech is a reflection of the political, social, and economic priorities of its time. This project focuses on analyzing the inaugural speeches of three influential US Presidents: Franklin D. Roosevelt in 1941, John F. Kennedy in 1961, and Richard Nixon in 1973. These speeches, delivered during critical junctures in American history, are expected to highlight different themes and priorities based on the unique challenges and circumstances of their respective eras. With the advent of Natural Language Processing (NLP) techniques, particularly utilizing the nltk (Natural Language Toolkit) library in Python, it is now possible to systematically analyze these historical texts to identify recurring themes and keywords. By examining the most common words used in these speeches, this project seeks to uncover the underlying rhetoric and focal points of each President's address, thus providing valuable insights into their leadership and the socio-political context of their terms.

Exploratory Data Analysis (EDA)

Data Description for Problem 2: Analyzing Presidential Speeches

The dataset for this problem comprises the inaugural speeches of three US Presidents, extracted from the nltk inaugural corpora in Python. The speeches are presented as text files, each corresponding to one of the following Presidents and years:

1. President Franklin D. Roosevelt (1941): This speech, delivered during a critical period of World War II, is expected to emphasize themes such as unity, strength, and resilience.
2. President John F. Kennedy (1961): Known for its inspirational tone, this speech is anticipated to focus on themes like change, progress, and civic responsibility.

3. President Richard Nixon (1973): Delivered amidst political turmoil, this speech likely addresses themes of stability, governance, and national progress.

For each speech, the dataset includes:

The full text of the speech.
The number of characters, words, and sentences contained within the speech, which can be derived during exploratory data analysis.

The primary focus of the analysis is on the textual content of these speeches, aiming to identify the most common words and the recurring themes through text cleaning, word frequency analysis, and visualization techniques such as word clouds.

We analyzed the speeches by counting the total number of characters, words, and sentences in each speech. This process involved calculating the length of the text for characters, tokenizing the text to count words, and splitting the text into sentences. The results provided a clear understanding of the length and complexity of each speech, with detailed counts for each metric

Character Count:

- Franklin D. Roosevelt : 7651 characters
- Richard Nixon: 10106 characters
- John F. Kennedy : 7673 characters

Word Count:

- Franklin D. Roosevelt : 1453 words
- John F. Kennedy : 1494 words
- Richard Nixon : 1913 words

Sentence Count:

- Franklin D. Roosevelt : 32 sentences
- John F. Kennedy : 27 sentences
- Richard Nixon : 20 sentences

3.2 Text Cleaning

- **Stopword Removal and Stemming**
    - To enhance the clarity of the speech content and focus on meaningful words, stopwords (common words like "and," "the," etc.) were removed. This step was essential to eliminate words that do not contribute significantly to the content analysis.
    - Next, stemming was applied using the Porter Stemmer to reduce words to their root forms. This process helps in consolidating different variations of the same word, improving the accuracy of subsequent analysis.

- **Analysis of Common Words**
    - After cleaning the text by removing stopwords and applying stemming, the most common words used across all three speeches were identified. These words include "nation," "day," "inaugur," "sinc," and "peopl." The frequent use of these terms reflects the central themes of the speeches, which focused on national unity, renewal, and dedication.

3.3 Visualization

Word Clouds for Each Speech

Word clouds were generated to visually represent the most frequent words in each speech. This visualization technique highlights the prominence of specific terms, providing an intuitive understanding of the key themes and topics addressed by each speaker.

(fig 17 showing word cloud for Kennedy)

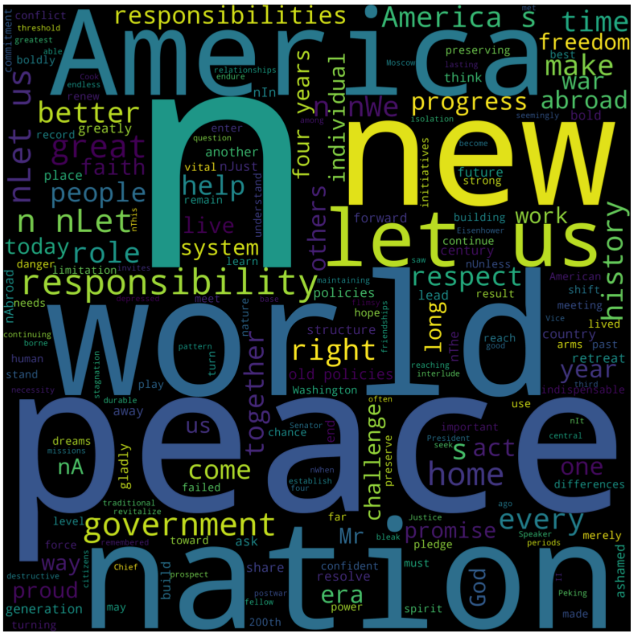
John F. Kennedy: Kennedy's word cloud reveals a strong emphasis on themes of renewal and progress, with frequently used words highlighting his vision for the future

Word Cloud for Roosevelt

(fig 18 showing word cloud for Roosevelt)

<u>Franklin D. Roosevelt</u>: The word cloud for Roosevelt's speech emphasizes terms related to national unity and dedication, reflecting his focus on collective effort and commitment.

Word Cloud for Nixon



(fig 19 showing word cloud for Nixon)

<u>Richard Nixon:</u> Nixon's word cloud showcases terms related to leadership and policy, underlining his focus on governance and national issues.

3.4 Insights & Conclusions

Summary & Key Takeaways of Findings

1. **Speech Length and Structure:**
    - Franklin D. Roosevelt and John F. Kennedy delivered speeches of similar lengths, both in terms of characters and words. Roosevelt's speech was slightly shorter in word count but comparable in character count to Kennedy's.
    - Richard Nixon delivered a notably longer speech, with the highest character and word count, reflecting a more extensive discussion or detailed presentation.
2. **Text Analysis:**
    - **Stopword Removal:** The analysis identified a significant number of stopwords in each speech, which were removed to focus on meaningful content. This step reduced the volume of non-informative words and highlighted the core vocabulary used by each speaker.
    - **Stemming:** Applying stemming to the remaining words revealed the root forms, which helped in aggregating similar terms and provided a clearer view of the key topics.
3. **Word Cloud Visualization:**
    - **Roosevelt's Speech:** The word cloud highlighted themes of unity and dedication, showing a focus on national solidarity and collective effort.
    - **Kennedy's Speech:** The visualization emphasized terms related to progress and renewal, indicating a forward-looking vision and emphasis on innovation.
    - **Nixon's Speech:** Key terms related to leadership and governance were prominent, reflecting a focus on policy and national issues.

Key Takeaways:

- Each speech had unique thematic focuses, with Roosevelt and Kennedy concentrating on unity and progress, while Nixon addressed leadership and policy.
- The use of word clouds and text analysis tools provided valuable insights into the themes and priorities communicated by each speaker.

Conclusion

Voter preferences

In the analysis of voter preferences between Labour and Conservative parties, we evaluated and compared the performance of Bagging and Boosting models. The key findings and recommendations from this analysis are summarized below:

- **Model Performance:** The Boosting model emerged as the superior model with an ROC-AUC score of 0.91, outperforming the Bagging model (0.89). This demonstrates Boosting's greater effectiveness in accurately distinguishing between Labour and Conservative voters.
- **Feature Importance:** The analysis revealed that perceptions of political leaders (Hague and Blair), stance on Europe, and political knowledge are the most influential factors in determining voter preference. Age and economic conditions also play roles, though less significant, while gender had minimal impact.
- **Strategic Recommendations:** Based on the insights gained:

- Enhance Engagement on Leader Perceptions: Political campaigns should focus on shaping and highlighting positive perceptions of key leaders.
- Clarify Stance on European Issues: Clear communication regarding positions on Europe is crucial.
- Increase Political Knowledge: Educational initiatives should be implemented to boost voter awareness and understanding.
- Target Younger Voters: Develop strategies specifically aimed at engaging younger demographics.
- Communicate Economic Policies Effectively: Ensure that economic policies are clearly articulated to address voter concerns.
- Maintain Inclusivity: Craft messages that appeal broadly, given the minimal impact of gender on voter preferences.

By implementing these strategies, political campaigns can better align their efforts with the factors that most influence voter behavior, leading to more effective and targeted outreach efforts.

Analysis of the inaugural speeches

The analysis of the inaugural speeches provided valuable insights into the rhetorical styles and thematic focuses of the speakers:

1. **Length and Structure:** The comparative analysis of character, word, and sentence counts revealed that Richard Nixon's speech was the longest, reflecting a more detailed presentation compared to Franklin D. Roosevelt and John F. Kennedy, whose speeches were of similar length.
2. **Text Cleaning:**
   - Stopword Removal: By removing common but non-informative words, the analysis underscored the core vocabulary used by each speaker, emphasizing their main themes.
   - Stemming: Stemming reduced words to their root forms, which facilitated a clearer understanding of the central topics and helped in identifying recurring themes across the speeches.
3. **Word Clouds:**
   - The word clouds visually represented the most frequently used terms in each speech, highlighting Roosevelt's focus on unity and dedication, Kennedy's emphasis on progress and renewal, and Nixon's attention to leadership and governance.

Overall: The combination of quantitative and qualitative analysis methods provided a comprehensive view of the content and stylistic elements of the speeches. This approach not only highlighted the unique aspects of each speaker's message but also offered insights into their rhetorical strategies and thematic priorities. The use of word clouds effectively illustrated the most significant terms, enhancing our understanding of the key messages conveyed during these pivotal addresses.

5. References

- **Previous jupyter notebooks**
- **Google**
- **Course Resources**