MACHINE LEARNING

BUSINESS REPORT

BY

G S JAIGURURAM

TABLE OF CONTENTS

Problem 1:

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.

PROBLEM 1.1 DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS:

Problem 1.1.1 Problem definition:

The goal is to develop a machine learning model to predict which political party a voter is likely to support based on their demographic and socio-economic factors.

Problem 1.1.2 Check shape, Data types, statistical summary:

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

Next, we import the data set “Election_Data.xlsx”.

Data Dictionary:

Column Name	Column Description
vote	Party choice: Conservative or Labour.
age	In years.
economic.cond.national	Assessment of current national economic conditions, 1 to 5.
economic.cond.household	Assessment of current household economic conditions, 1 to 5.
Blair	Assessment of the Labour leader, 1 to 5.
Hague	Assessment of the Conservative leader, 1 to 5.
Europe	An 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment.
political.knowledge	Knowledge of parties' positions on European integration, 0 to 3.
gender	Female or Male.

Shape of the data:

Shape of the dataset is 1525 rows and 10 Columns.

Data Type:

```
#   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
```

There are total of 8 Numerical and 2 Categorical data type are available.

Statistical Summary:

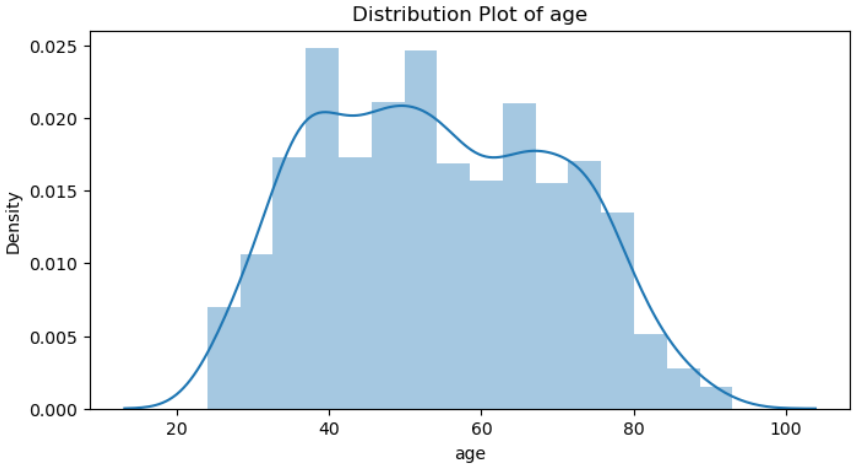
Insights:

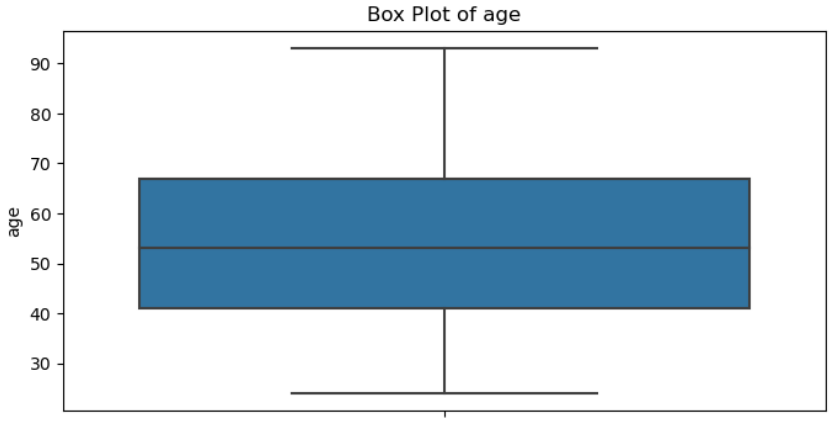
1. The dataset has 1525 rows and 10 columns.
2. The dataset has 2 categorical variable and 8 numerical variable
3. The dataset has column named as "Unnamed: 0" Which seems to be a serial number column which is not required for Analysis, so we decide to drop it.
4. Age: The mean age of the participants is 54.18 years, with a standard deviation of 15.71 years. The youngest participant is 24 years old, while the oldest is 93 years old. The median age is 53 years, and 50% of the participants are between 41 and 67 years old.
5. Economic conditions (national): The mean score for the perception of national economic conditions is 3.25 on a scale of 1 to 5, with a standard deviation of 0.88. The median score is 3, indicating that most participants perceive the national economic conditions as average. The minimum and maximum scores are 1 and 5, respectively.
6. Economic conditions (household): The mean score for the perception of household economic conditions is 3.14 on a scale of 1 to 5, with a standard deviation of 0.93. The median score is 3, indicating that most participants perceive their household economic conditions as average. The minimum and maximum scores are 1 and 5, respectively.
7. Opinions on political figures: The mean score for opinions on Tony Blair is 3.33 on a scale of 1 to 5, with a standard deviation of 1.17. The median score is 4, indicating that most participants have a positive opinion of Blair. The mean score for opinions on William Hague is 2.75 on a scale of 1 to 5, with a standard deviation of 1.23. The median score is 2, indicating that most participants have a negative opinion of Hague.
8. Attitudes towards Europe: The mean score for attitudes towards Europe is 6.73 on a scale of 1 to 11, with a standard deviation of 3.30. The median score is 6, indicating that most participants have a neutral attitude towards Europe. The minimum and maximum scores are 1 and 11, respectively.
9. Political knowledge: The mean score for political knowledge is 1.54 on a scale of 0 to 3, with a standard deviation of 1.08. The median score is 2, indicating that most participants have a moderate level of political knowledge. The minimum score is 0, while the maximum score is 3.

	count	mean	std	min	0.25	0.5	0.75	max
age	1525	54.182295	15.711209	24	41	53	67	93
economic.cond.national	1525	3.245902	0.880969	1	3	3	4	5
economic.cond.household	1525	3.140328	0.929951	1	3	3	4	5
Blair	1525	3.334426	1.174824	1	2	4	4	5
Hague	1525	2.746885	1.230703	1	2	2	4	5
Europe	1525	6.728525	3.297538	1	4	6	10	11
political.knowledge	1525	1.542295	1.083315	0	0	2	2	3

Problem 1.1.3 Univariate analysis:

Numerical Data type:





Age Distribution Shape: The plot shows a roughly bimodal distribution, with two noticeable peaks around ages 40 and 55. This indicates that there are two age groups that are more frequent in the dataset.

General Trend: The density of individuals increases from age 20, reaching a peak around age 40. There is a slight dip after 40, followed by another peak around age 55. After age 55, the density gradually decreases, with fewer individuals in the higher age ranges, tapering off significantly after age 80.

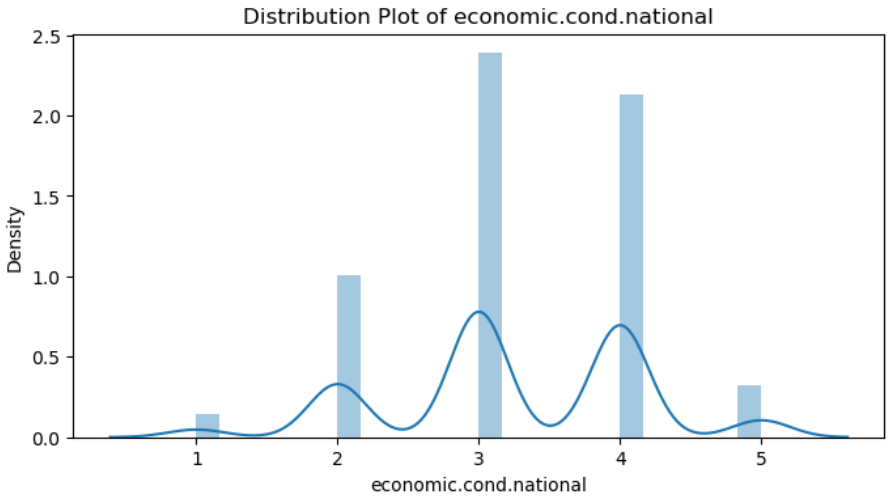
Density Peaks: The highest density peak is around age 40, suggesting that this age group is the most common in the dataset. The second peak around age 55 indicates another common age group, though not as prevalent as the one at 55.

Young and Elderly Population: There are relatively fewer individuals below age 20 and above age 80.

Middle-Aged Dominance: The majority of the population in the dataset is between ages 30 to 60, as indicated by the higher density in this range.

Median Age: The median age is around 55, as indicated by the line inside the box. This means that 50% of the individuals are younger than 55 and 50% are older than 55.

Interquartile Range (IQR): The box represents the interquartile range, which contains the middle 50% of the data. The lower quartile (Q1) is around age 45. The upper quartile (Q3) is around age 70. This indicates that the middle 50% of the individuals are between ages 45 and 70.



Whiskers: The whiskers extend from the quartiles to the minimum and maximum values within 1.5 times the IQR. The lower whisker extends to around age 30. The upper whisker extends to around age 90. This suggests that most of the data falls between ages 30 and 90.

Range: The overall range of ages in the dataset spans from about 30 to 90.

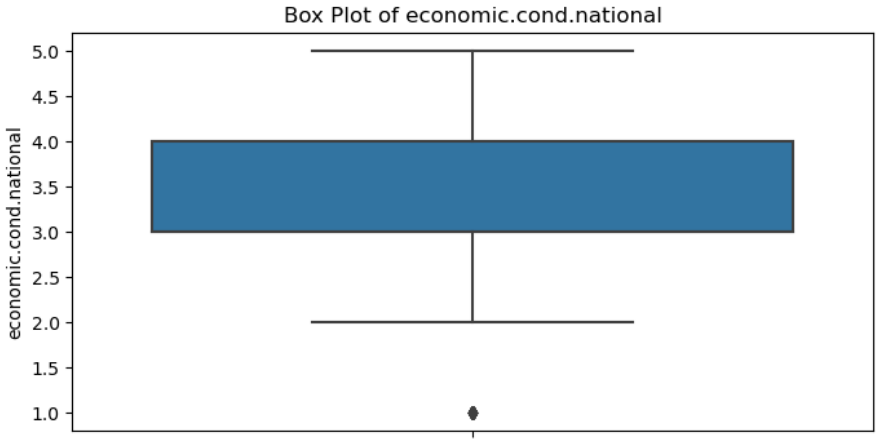
Symmetry: The box plot appears relatively symmetric, indicating that the age distribution is fairly balanced around the median.

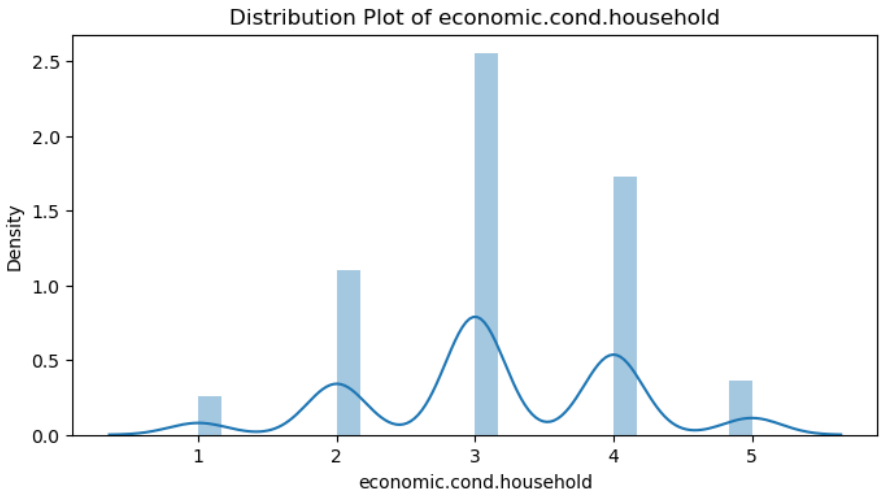
Multiple Peaks (Multimodal Distribution): The plot shows multiple peaks, indicating a multimodal distribution. This suggests that the data has several common values or clusters.

Common Values: The most common values (peaks) are around 3 and 4. These values have the highest density, meaning they occur more frequently in the dataset. There are also notable peaks around 2 and 5, but they are less pronounced compared to the values 3 and 4. The value 1 has a very low density, indicating it is less common in the dataset.

Distribution Shape: The shape of the distribution suggests that the variable "economic.cond.national" does not follow a normal distribution; instead, it has multiple modes.

Interpretation of Economic Conditions: If the variable "economic.cond.national" represents perceptions of national economic conditions on a scale (e.g., 1 being very poor and 5 being very good), the distribution suggests a mixed perception. Many respondents perceive the economic





conditions as average to good (values 3 and 4).

Median (Median): The line in the middle of the box represents the median, which appears to be around 4. This indicates that the median perception of national economic conditions is quite high.

Interquartile Range (IQR): The box ranges from approximately 3 to 4.5, which indicates the interquartile range (IQR). This suggests that 50% of the data lies between these values.

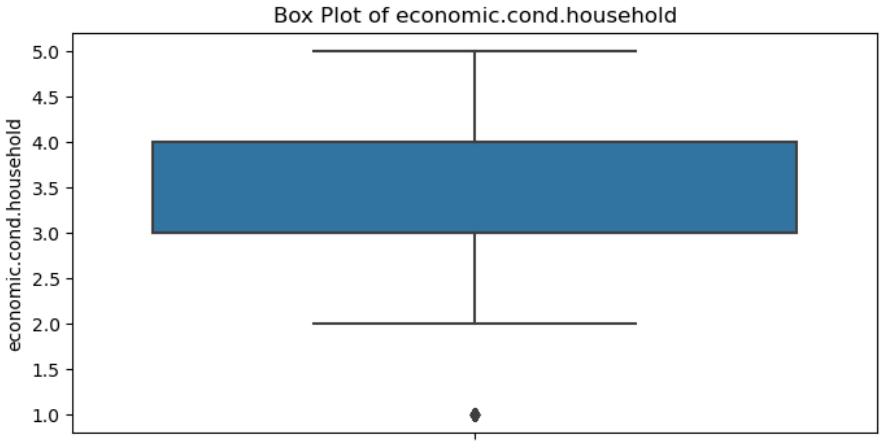
Maximus and minimous: The "whiskers" on the diagram extend from approximately 2 to 5. These represent the minimum and maximum values, excluding outliers.

Outliers (Atypical Values): There is a point outside the range of the whiskers at value 1, indicating that there is an outlier in the data. This could be a very low perception of national economic conditions compared to the rest of the data.

Symmetry: The box appears to be closer to the top of the range, suggesting that the distribution of the data might be skewed toward higher values.

Multiple Peaks (Multimodal Distribution): The plot shows multiple peaks, indicating a multimodal distribution. This suggests that the data has several common values or clusters.

Common Values: The most common values (peaks) are around 3 and 4. These values have the highest density, meaning they occur more frequently in the dataset. There are also notable peaks around 2 and 5, but they are less pronounced compared to the values 3 and 4. The value 1 has a very low density, indicating it is less common in the dataset.



Distribution Shape: The shape of the distribution suggests that the variable "economic.cond.household" does not follow a normal distribution; instead, it has multiple modes.

Interpretation of Economic Conditions: If the variable "economic.cond. household" represents perceptions of national economic conditions on a scale (e.g., 1 being very poor and 5 being very good), the distribution suggests a mixed perception.

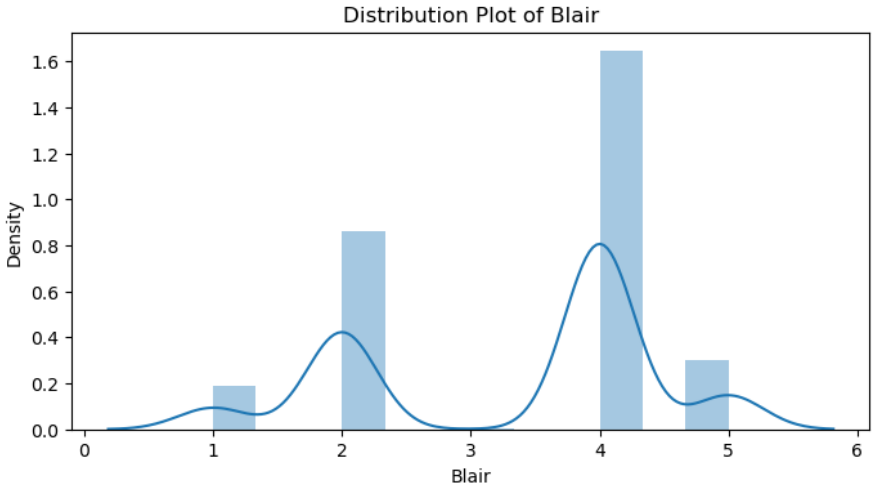
Median (Median): The line in the middle of the box represents the median, which appears to be around 4. This indicates that the median perception of national economic conditions is quite high.

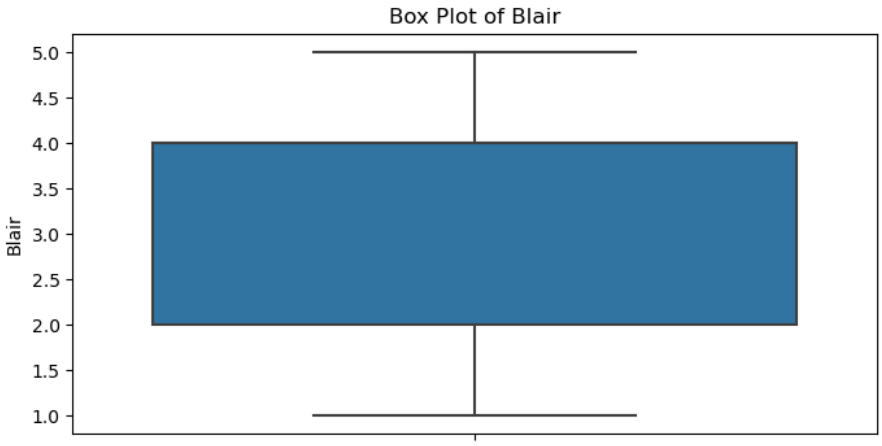
Interquartile Range (IQR): The box ranges from approximately 3 to 4.5, which indicates the interquartile range (IQR). This suggests that 50% of the data lies between these values.

Maximus and minimous: The "whiskers" on the diagram extend from approximately 2 to 5. These represent the minimum and maximum values, excluding outliers.

Outliers (Atypical Values): There is a point outside the range of the whiskers at value 1, indicating that there is an outlier in the data. This could be a very low perception of national economic conditions compared to the rest of the data.

Symmetry: The box appears to be closer to the top of the range, suggesting that the distribution of the data might be skewed toward higher values.





Bimodal Distribution: The plot exhibits a bimodal distribution, indicating that there are two peaks in the data. This suggests that there are two distinct groups of voters with differing opinions about Blair.

Major Peaks: The first major peak occurs around the value 4. This suggests that a significant proportion of voter's rate Blair quite highly. The second peak occurs around the value 2, indicating that another substantial group of voter's rates Blair relatively poorly.

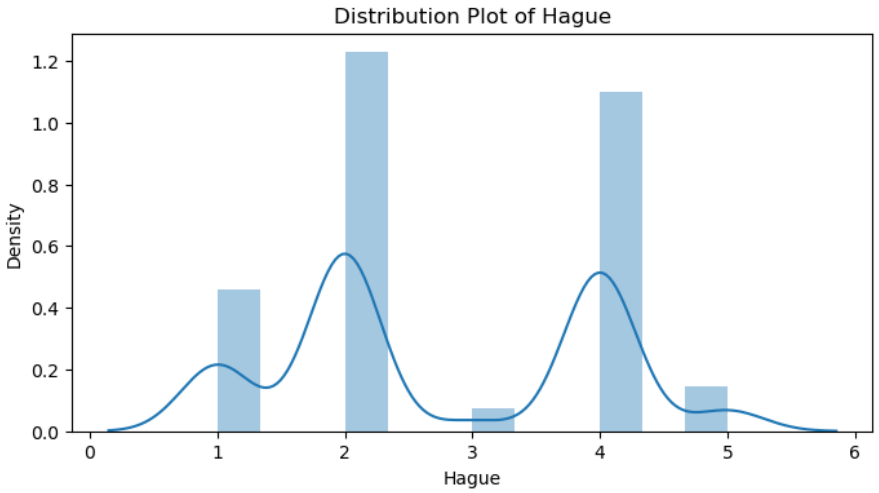
Minor Peaks: There are also minor peaks at the values 1 and 5. This indicates that there are some voters who either strongly disapprove or strongly approve of Blair.

Skewness: The distribution is not symmetrical. The density curve suggests there is a higher concentration of voters towards the higher end of the rating scale (around 4), indicating a skew towards higher ratings. However, there is also a noticeable number of voters who rate Blair poorly (around 2), showing that opinions are somewhat polarized.

Median: The median value (the line inside the box) is around 3.5. This indicates that half of the voter's rate Blair above 3.5 and half below 3.5.

Interquartile Range (IQR): The box itself represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). In this plot, Q1 is around 2.5 and Q3 is around 4. The IQR is therefore 4 - 2.5 = 1.5. This range captures the middle 50% of the data.

Whiskers: The whiskers extend from the box to the smallest and largest values within 1.5 * IQR from Q1 and Q3, respectively.



The lower whisker extends to 1, indicating that the lowest rating given is 1. The upper whisker extends to 5, indicating that the highest rating given is 5.

Outliers: There are no outliers in the data as all the data points fall within the whiskers.

Symmetry: The box plot shows a fairly symmetrical distribution around the median, though the box is slightly longer above the median than below it, suggesting a slight skew towards higher ratings.

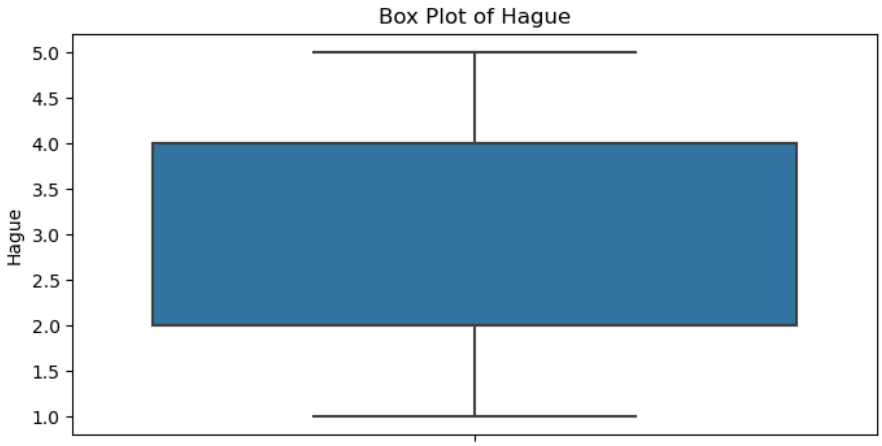
Bimodal Distribution: The plot exhibits a bimodal distribution, indicating the presence of two distinct peaks. This suggests that voters have two main opinions about Hague.

Major Peaks: The first major peak is around the value 2, suggesting that a significant proportion of voter's rate Hague poorly. The second major peak is around the value 4, indicating that another substantial group of voter's rate Hague quite highly.

Minor Peaks: There are minor peaks around the values 1 and 5, indicating some voters have extreme opinions, either very low or very high.

Density Distribution: The highest density is at the rating of 2, showing that the most common rating for Hague is relatively low. The second highest density is at the rating of 4, showing a notable number of voters rating him highly as well.

Symmetry: The distribution appears somewhat symmetrical around the ratings of 2 and 4, but with a notable dip around the middle rating of 3, indicating fewer voters rating Hague as average.



Median (Q2): The median value of the data is approximately 3.5. This is the central value, indicating that half of the data points are above 3.5 and half are below.

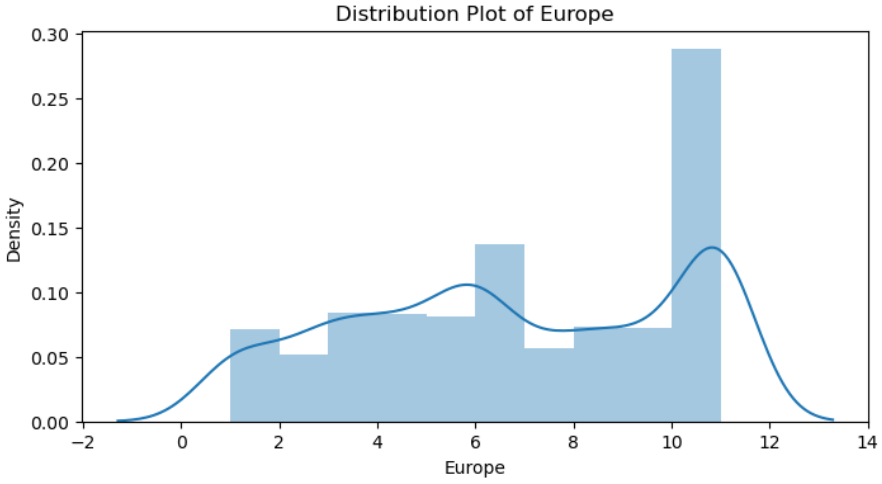
Interquartile Range (IQR): The box represents the interquartile range, which is the range between the first quartile (Q1) and the third quartile (Q3). In this plot:

1. Q1 (25th percentile) is around 2.5.
2. Q3 (75th percentile) is around 4.0.
3. The IQR is Q3 - Q1, which is approximately 1.5.

Whiskers: The lines extending from the top and bottom of the box are called whiskers. They represent the range of the data excluding outliers. The lower whisker extends to the minimum value, which is about 1.0. The upper whisker extends to the maximum value, which is about 5.0.

Range: The total range of the data is from 1.0 to 5.0.

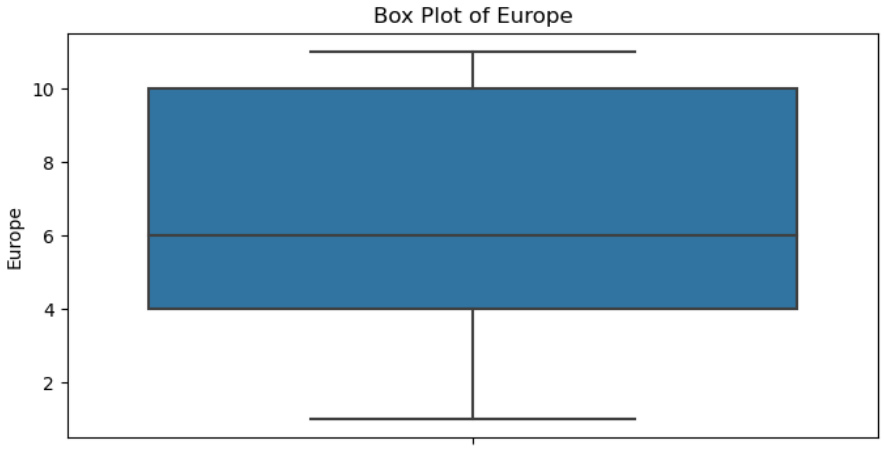
Symmetry: The plot appears to be relatively symmetric, as the median line is roughly in the middle of the box, and the whiskers are of similar length above and below the box. This suggests a relatively balanced distribution of data around the median.



Multimodality: The distribution is multimodal, that is, it has several peaks. There are notable peaks around 3, 6 and 10, indicating the presence of several subgroups in the data.

Density: The highest density is around 10, which means that a lot of the data is concentrated around this value.

Amplitude: The data ranges from approximately -2 to 14, showing a wide range of values. However, the majority of data falls between 0 and 12.



- **Symmetry:** The distribution is asymmetrical, with a longer tail on the right side (higher values), indicating the presence of some exceptionally high values.
- **Concentration:** There is a noticeable concentration of data in the range 9 to 11, where the density peaks. This could indicate a trend or common characteristic in this range of values.

Median (Central Tendency): The median value, which is the middle value of the dataset, is around 6. This is represented by the line inside the box.

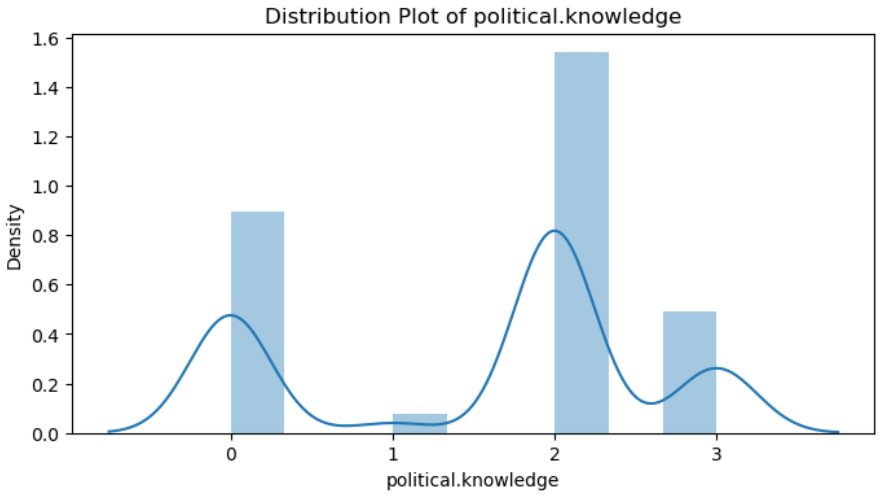
Interquartile Range (IQR): The box itself spans from the first quartile (Q1) to the third quartile (Q3). This range covers the central 50% of the data. Q1 (25th percentile) appears to be around 4. Q3 (75th percentile) appears to be around 8.

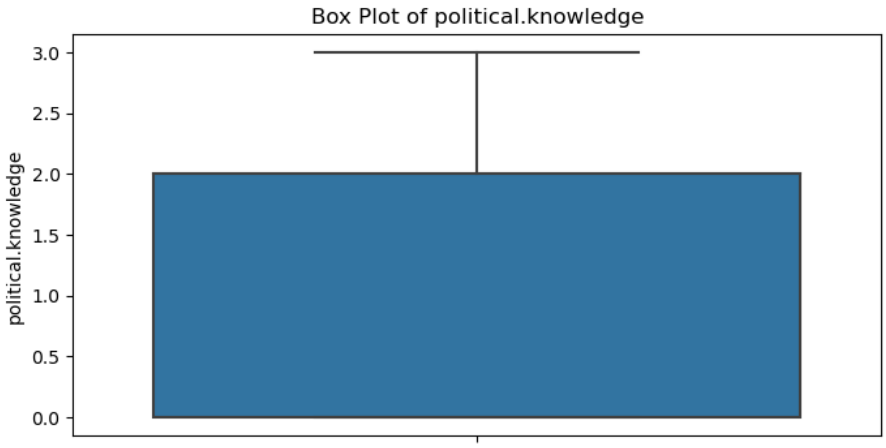
Range and Whiskers: The whiskers extend from the box to the minimum and maximum values within 1.5 times the IQR. The lower whisker extends down to approximately 1. The upper whisker extends up to approximately 10.

Spread of Data: The data is relatively evenly distributed around the median.

There are no apparent outliers in the dataset, as all data points fall within the whiskers.

Symmetry of Data: The box plot looks fairly symmetrical, indicating that the data is likely evenly distributed on both sides of the median.





Multimodal Distribution: The distribution has multiple peaks, indicating that the data has more than one mode or frequent value. There are significant peaks at values 0 and 2.

Frequency of Values: The value of "political.knowledge" at 0 has a high density, suggesting that there are a considerable number of people with very little political knowledge. The value of 2 has the highest density, indicating that most people have a moderate level of political knowledge. There is a third peak at 3, but it is minor compared to the other two.

Gaps in Distribution: There is a noticeable valley between values 1 and 2, suggesting that there are fewer people with a low level of political knowledge. Another smaller valley is observed around value 3.

Asymmetry: The distribution is not symmetrical; Peaks and valleys indicate that the data is clustered at certain specific points rather than being evenly distributed.

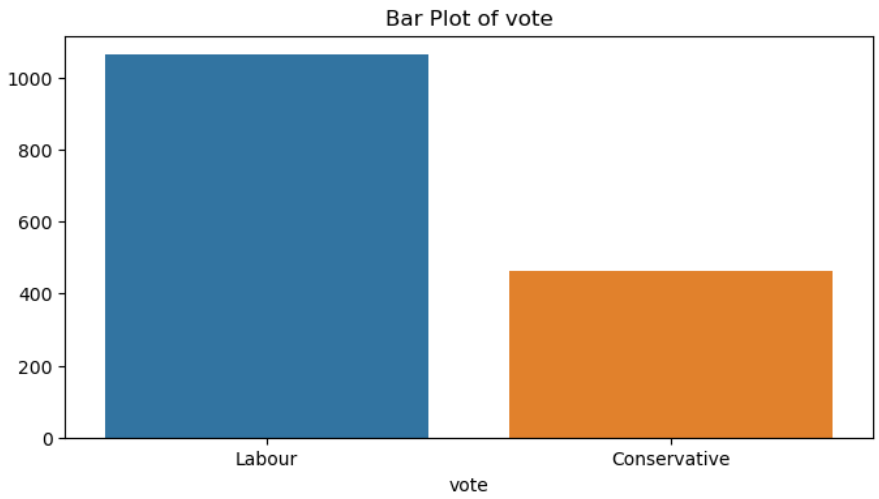
- **Density:** The maximum density reaches almost 1.5 around value 2, indicating that this is the most common value in the data set. Values of 0 and 3 also show significant densities, although not as high as 2.
- **Median:** The median value (the line inside the box) is 1.5. This indicates that half of the voters have a political knowledge rating above 1.5 and half below 1.5.
- **Interquartile Range (IQR):** The box itself represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3).
- In this plot:
- 1. Q1 is 1 and Q3 is 2.
- 2. The IQR is therefore 2 - 1 = 1. This range captures the middle 50% of the data.

. **Whiskers:** The whiskers extend from the box to the smallest and largest values within 1.5 * IQR from Q1 and Q3, respectively. The lower whisker extends to 0, indicating that the lowest political knowledge rating given is 0. The upper whisker extends to 3, indicating that the highest political knowledge rating given is 3.

. **No Outliers:** There are no outliers in the data as all the data points fall within the whiskers.

. **Symmetry:** The box plot shows a symmetrical distribution around the median, suggesting an even spread of data around the center.

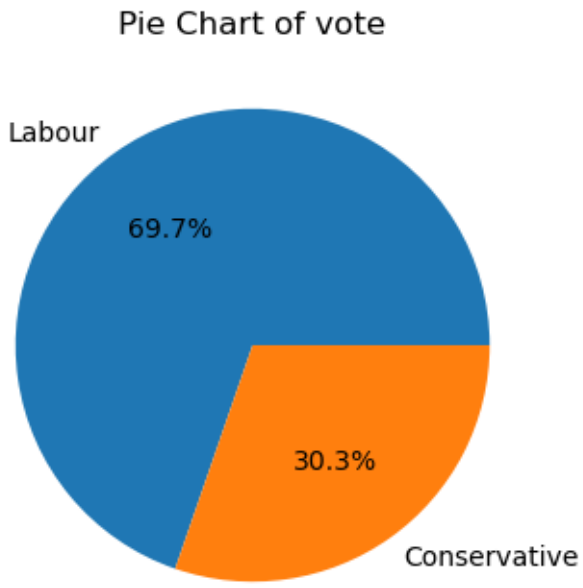
Categorical Data Type:



Vote Distribution: The bar plot indicates the number of votes each party received. Labour has received significantly more votes than the Conservative party.

Comparison: Labour has over 1000 votes. The Conservative party has around 500 votes. This indicates that Labour has more than double the number of votes compared to the Conservative party.

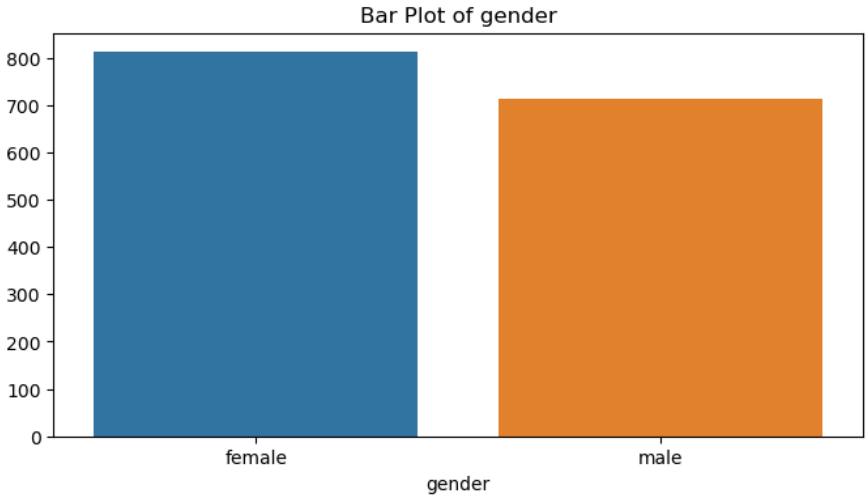
Majority Support: The clear difference in the height of the bars shows a strong preference among the surveyed voters towards Labour over the Conservative party.



Proportion of Votes: Labour received 69.7% of the votes. The Conservative party received 30.3% of the votes.

Majority Support: The majority of voters prefer Labour, as indicated by the larger blue section of the pie chart.

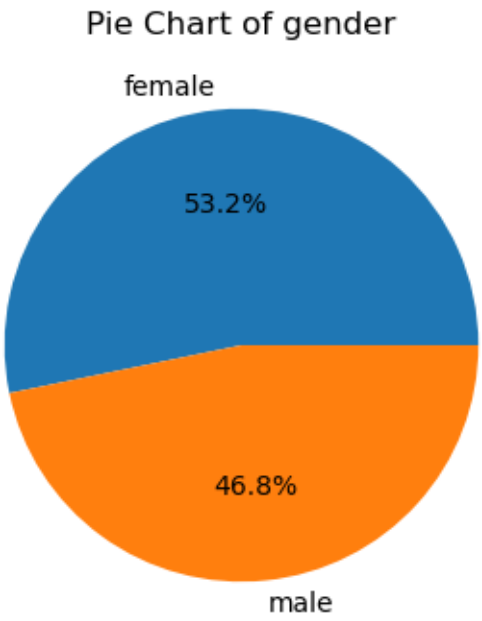
Visual Representation: The pie chart visually emphasizes the significant lead Labour has over the Conservative party, with more than twice the percentage of votes.



Gender Distribution: The bar plot indicates the number of gender present in whole population. Female has received significantly more than Male population.

Comparison: Female has over 800 votes. The Male has around 700 votes. This indicates that Female has more number of votes compared to the Male Votes.

Majority Support: The clear difference in the height of the bars shows that Female Votes are more than the Male Vote.



Proportion of Gender: Female Voter's represent 53.2% of total votes. The Male Voter's represent 30.3% of total votes.

Majority Support: The majority of voters are Female, as indicated by the larger blue section of the pie chart.

Visual Representation: The pie chart visually emphasizes the significant lead Female has over the Male Voter's.

Problem 1.1.4 Multivariate analysis:

Insights from heat map:

Strong Correlations:

- economic.cond.national and economic.cond.household: They have a moderate positive correlation of 0.35, indicating that as the national economic condition improves, the household economic condition also tends to improve.

Moderate Correlations:

- Blair and economic.cond.national: Positive correlation of 0.33, suggesting that a better perception of the national economic condition is associated with a better perception of Blair.
- Europe and Hague: Positive correlation of 0.29, indicating that a favorable perception of Europe is associated with a favorable perception of Hague.

Negative Correlations:

- Europe and Blair: Negative correlation of -0.3, suggesting that a favorable perception of Europe is associated with a less favorable perception of Blair.
- Hague and economic.cond.national: Negative correlation of -0.2, indicating that a better perception of the national economic condition is associated with a less favorable perception of Hague.

Weak Correlations:

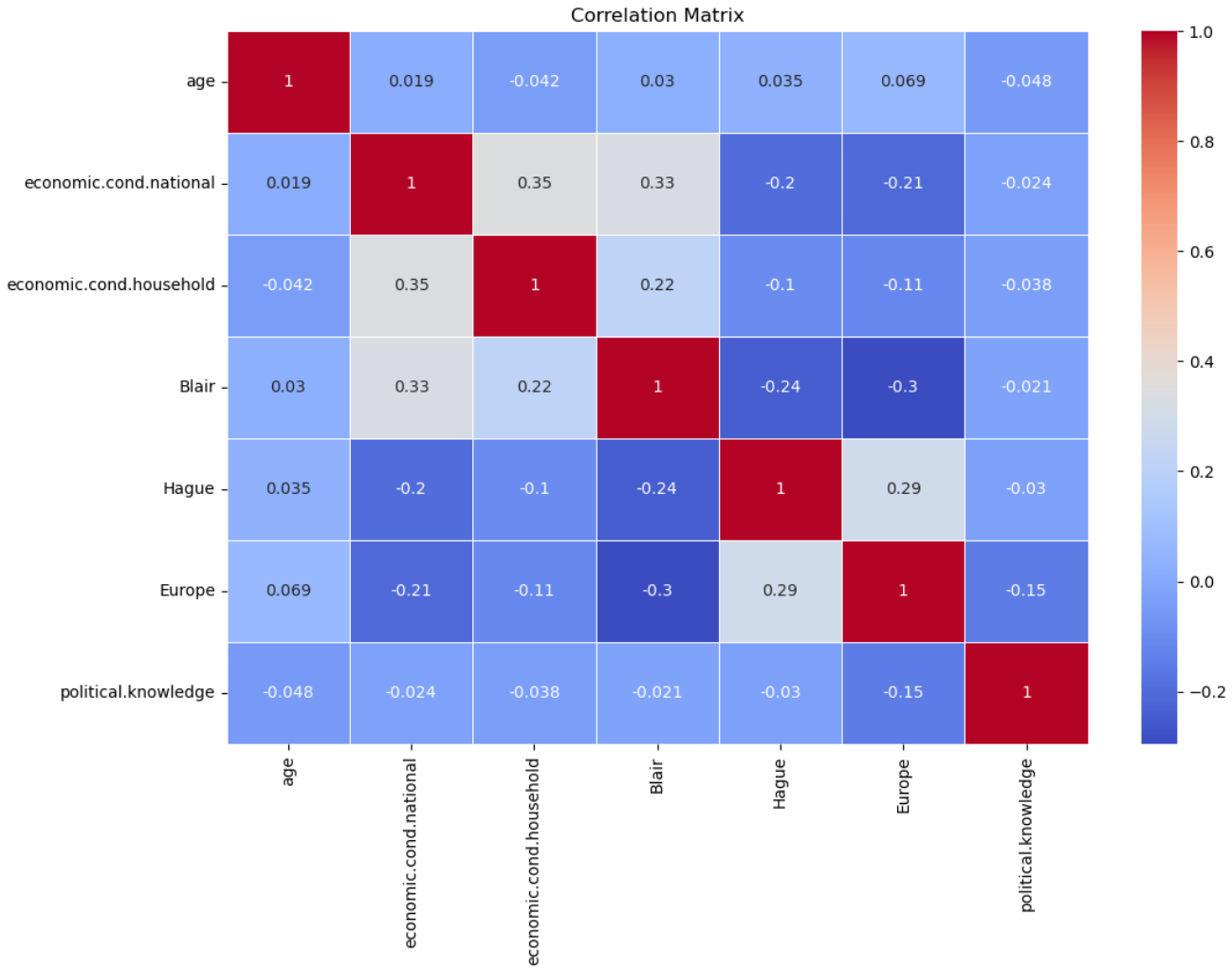
- Most of the other correlations are quite weak (close to 0), indicating that there is not a strong relationship between those variables.
- For example, the correlation between age and economic.cond.national is 0.019, suggesting little or no relationship between age and perception of national economic condition.

Correlations with political.knowledge:

- political.knowledge has weak and negative correlations with all other variables, indicating that political knowledge is not strongly related to any other variable in this matrix.

In summary, the correlation matrix shows some moderate relationships between perceptions of economic conditions and political figures, as well as between perceptions of Europe and Hague. However, most correlations are weak, suggesting that many of these variables are not strongly related to each other.

Heat Map:



Insights from pairplot:

Univariate Distributions:

- **age:** The majority of people in the dataset are between 20 and 80 years old, with a greater concentration in the lower age ranges.
- **economic.cond.national and economic.cond.household:** Both variables show multimodal distributions, with several peaks suggesting different perceptions of economic conditions.
- **Blair, Hague, Europe and political.knowledge:** All of these variables have discrete distributions with specific values that are most frequent.

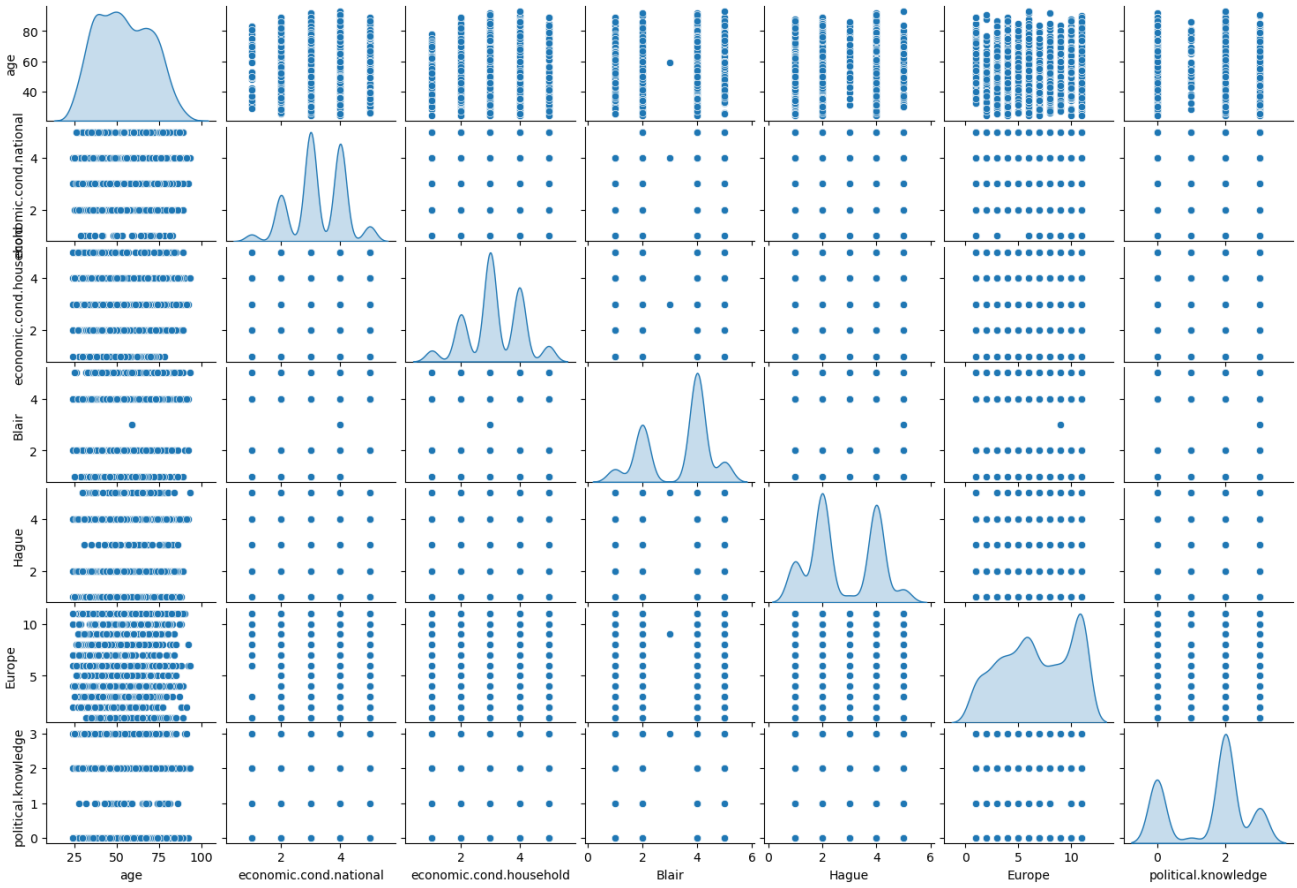
Bivariate Relationships:

- **economic.cond.national and economic.cond.household:** There is a visible positive correlation, confirming that as the perception of the national economic condition improves, so does the perception of the household economic condition.
- **Blair and economic.cond.national:** A positive trend is observed, indicating a relationship between the perception of Blair and the perception of the national economic condition.
- **Hague and Europe:** They also show a positive trend, suggesting a relationship between the perception of Hague and Europe.

Multimodal Patterns:

- The distributions of economic.cond.national and economic.cond.household show several peaks, suggesting that there are different groups or segments in the population with different perceptions of economic conditions.
- political.knowledge also shows a multimodal distribution, indicating different levels of political knowledge in the population.

Pair Plot:



Problem 1.1.5 Use appropriate visualizations to identify the patterns and insights:

Insights

Distribution of Variables

- **Age:** The age distribution appears to be similar for voters of both parties. There is a greater concentration of young voters (20-40 years old) and a gradual decline in older voters.
- **economic.cond.national and economic.cond.household:** Both variables show a fairly dispersed distribution with no clear trend between 'Labour' and 'Conservative' voters. There is no obvious separation that indicates that a specific national or household economic condition strongly influences the vote.
- **Blair and Hague:** These variables could represent approval or disapproval of leaders Blair and Hague. There is a greater concentration of 'Labour' voters with high values for Blair, which could indicate greater approval of that leader. The Hague variable seems to be more balanced between both groups of voters.
- **Europe:** The Europe variable shows a more dispersed distribution, but there is a slight trend where 'Labour' voters seem to have higher values compared to 'Conservative' voters.
- **political.knowledge:** The political knowledge variable shows that 'Labour' voters tend to have a broader distribution of political knowledge compared to 'Conservative' voters.

Relationships between Variables

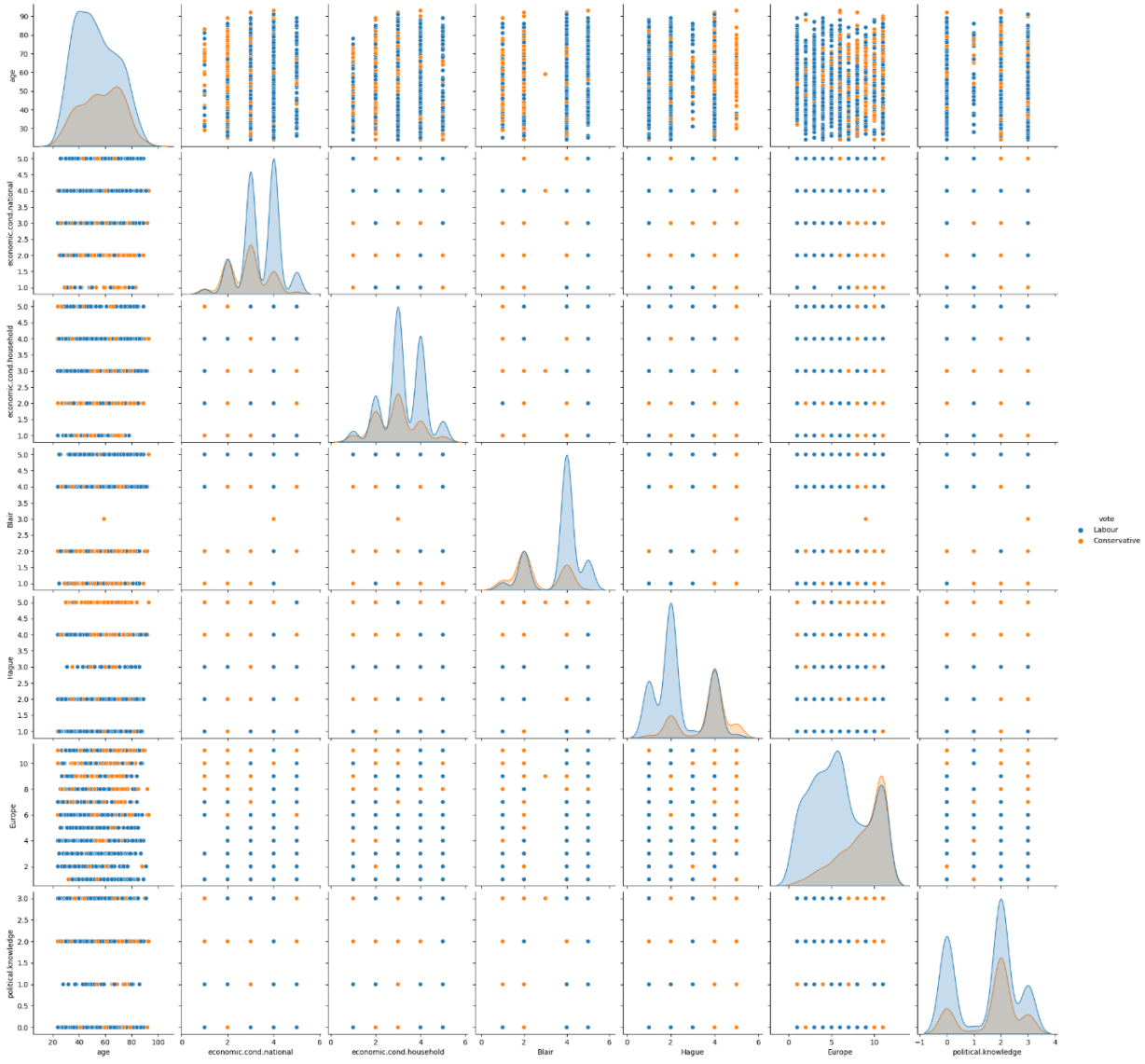
- **Correlation between variables:** The variables do not show strong correlations evident to the naked eye, since the scatter points are quite distributed without forming clear patterns. It is observed that the variable 'Blair' has a greater density of blue points (Labour) at high values.

Insights on the Objective Variable "vote"

- **General Trends:** There is no single variable that alone clearly determines the voting preference between 'Labour' and 'Conservative'. Factors such as approval of leaders (Blair for Labour) and political knowledge appear to have some influence.
- **Voter Segmentation:** Labor voters may be more concentrated in younger age ranges and have a better perception of Blair. There is no significant divergence in terms of national or household economic conditions between the two groups of voters.
- **Distribution of Variables:** The distribution of the variables suggests that voters have varied perceptions of economic conditions and political leaders, reflecting a complexity in the voting decision.

In summary, the analysis of the graph suggests that voting preference is influenced by a combination of factors, with some trends visible in approval of leaders and political knowledge, although there is no variable that dominates the prediction of the vote.

Plot:



USE APPROPRIATE VISUALIZATIONS TO IDENTIFY THE PATTERNS AND INSIGHTS

- Above we have Histogram & box plot for Individual Numerical data and Bar plot & pie chart for Individual categorical data.
- For Comparison of multiple numerical data we have heatmap & pair plot.

Problem 1.1.6 Key meaningful observations on individual variables and the relationship between variables:

Age:

- Mean Age: The average age of participants is 54.18 years with a standard deviation of 15.71 years.
- Age Range: Participants range from 24 to 93 years old, with a median age of 53 years.
- Age Distribution: The age distribution is bimodal, with peaks around ages 40 and 55, indicating two prevalent age groups. The majority of participants are between 30 and 60 years old.

National Economic Conditions:

- Mean Score: The average perception score is 3.25 (on a scale of 1 to 5), indicating that most participants view national economic conditions as average.
- Distribution: The scores range from 1 to 5, with a standard deviation of 0.88.

Household Economic Conditions:

- Mean Score: The average perception score is 3.14, also suggesting an average view of household economic conditions.
- Distribution: Scores range from 1 to 5, with a standard deviation of 0.93.

Tony Blair:

- Mean Score: The average opinion score is 3.33, indicating a generally positive view.
- Distribution: The median score is 4, suggesting that most participants rate him positively.

William Hague:

- Mean Score: The average opinion score is 2.75, reflecting a more negative view.
- Distribution: The median score is 2, indicating that most participants rate him negatively.

Attitudes Towards Europe:

- Mean Score: The average score is 6.73 (on a scale of 1 to 11), with a median score of 6, suggesting a neutral attitude towards Europe.
- Distribution: Scores range from 1 to 11, with a standard deviation of 3.30.

Political Knowledge:

- Mean Score: The average political knowledge score is 1.54 (on a scale of 0 to 3), indicating a moderate level of political knowledge among participants.
- Distribution: The median score is 2, with scores ranging from 0 to 3.

Correlations

- Heat Map Insights: The heat map indicates strong correlations between certain variables, suggesting relationships that could be explored further in analysis.

Summary

The dataset provides valuable insights into the demographics, economic perceptions, political opinions, and knowledge levels of participants. The bimodal age distribution and polarized opinions on political figures highlight the diversity of perspectives within the surveyed population. Further analysis could explore the relationships between these variables to uncover deeper insights.

PROBLEM 1.2 DATA PREPROCESSING:

There are no missing data.

Problem 1.2.1 Outlier Treatment:

We Checked Outlier and following observations are made:

1. Outliers were identified in age', economic.cond.national', economic.cond.household', Blair', Hague', Europe' and political.knowledge' columns, indicating the presence of values that are significantly higher or lower than the rest of the data points.

2. Given the high number of outliers in the dataset, it is recommended to treat outliers before proceeding with Mean Value.

Problem 1.2.2 Encode the data:

Encoding of categorical data like ‘vote’ and ‘gender’ so that we can convert it to numerical data type.

- vote feature has two data Labour and Conservative so we are converting Labour values to ‘1’ and Conservative values to ‘1’.
- Gender feature has two data female and male so we are converting female value to ‘0’ and male values to ‘1’.

Problem 1.2.3 Data Split:

We are going to split data into train set and test set using train_test_split function from scikit-learn. We are split data into 75 – 25 splits where 75% of entire data will be train dataset and 25% will be test dataset.

Shape values of train – test dataset is:

```
Training set shape: (1220, 8)
Testing set shape: (305, 8)
```

Problem 1.2.4 Scale the data:

Scaling the data is a crucial step in preparing the dataset for machine learning algorithms. It ensures that all features contribute equally, improves the convergence of optimization algorithms, and prevents numerical instability. By applying feature scaling, we enhance the model's performance and reliability.

We are going to use StandardScaler function from sklearn library.

PROBLEM 1.3 MODEL BUILDING - LINEAR REGRESSION:

Problem 1.3.1 Metrics of Choice (Justify the evaluation metrics):

For evaluating the models, we'll use the following metrics:

Accuracy: Measures the proportion of correctly classified instances out of the total instances. It's useful for getting a quick overall measure of model performance.

Precision: The proportion of true positive instances out of the instances predicted as positive.

Recall (Sensitivity): The proportion of true positive instances out of the actual positive instances.

F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both concerns.

Confusion Matrix: Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives.

ROC-AUC Score: Evaluates the model's ability to distinguish between classes. The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is particularly useful for binary classification problems.

These metrics will be used for evaluation of the models' performance, allowing us to understand their strengths and weaknesses from multiple perspectives.

Problem 1.3.2 Model Building (KNN, Naive bayes, Bagging, Boosting):

We are going to build KNN, Naive bayes, Bagging, Boosting Model Using sklearn library. Below are the Details:

1. **Model: KNN**

```
Training Accuracy: 0.8778688524590164
Testing Accuracy: 0.7672131147540984
Confusion Matrix:
 [[ 53  36]
 [ 35 181]]
Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.60      0.60        89
           1       0.83      0.84      0.84       216

    accuracy                           0.77       305
   macro avg       0.72      0.72      0.72       305
weighted avg       0.77      0.77      0.77       305
```

2. **Model: Naive Bayes**

```
Training Accuracy: 0.8418032786885246
Testing Accuracy: 0.7967213114754098
Confusion Matrix:
 [[ 57  32]
 [ 30 186]]
Classification Report:
              precision    recall  f1-score   support
```

```
              precision    recall  f1-score   support

           0       0.66      0.64      0.65        89
           1       0.85      0.86      0.86       216

    accuracy                           0.80       305
   macro avg       0.75      0.75      0.75       305
weighted avg       0.80      0.80      0.80       305
```

3. **Model: Bagging**

```
Training Accuracy: 0.9852459016393442
Testing Accuracy: 0.7737704918032787
Confusion Matrix:
 [[ 54  35]
 [ 34 182]]
Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.61      0.61        89
           1       0.84      0.84      0.84       216

    accuracy                           0.77       305
   macro avg       0.73      0.72      0.73       305
weighted avg       0.77      0.77      0.77       305
```

4. **Model: Boosting**

```
Training Accuracy: 0.8573770491803279
Testing Accuracy: 0.8065573770491803
Confusion Matrix:
 [[ 56  33]
 [ 26 190]]
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.63      0.65        89
           1       0.85      0.88      0.87       216

    accuracy                           0.81       305
   macro avg       0.77      0.75      0.76       305
weighted avg       0.80      0.81      0.80       305
```

PROBLEM 1.4 MODEL PERFORMANCE EVALUATION:

Problem 1.4.1 Check the confusion matrix and classification metrics for all the models (for both train and test dataset):

1. **Model: KNN**

Confusion Matrix:

```
 [[ 53  36]
 [ 35 181]]
```

Classification Report:

```
              precision    recall  f1-score   support

           0       0.60      0.60      0.60        89
           1       0.83      0.84      0.84       216

    accuracy                           0.77       305
   macro avg       0.72      0.72      0.72       305
weighted avg       0.77      0.77      0.77       305
```

2. **Model: Naive Bayes**

Confusion Matrix:

```
 [[ 57  32]
 [ 30 186]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.66	0.64	0.65	89
1	0.85	0.86	0.86	216
accuracy			0.80	305
macro avg	0.75	0.75	0.75	305
weighted avg	0.80	0.80	0.80	305

3. **Model: Bagging**

Confusion Matrix:

```
 [[ 54  35]
 [ 34 182]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.61	0.61	0.61	89
1	0.84	0.84	0.84	216
accuracy			0.77	305
macro avg	0.73	0.72	0.73	305
weighted avg	0.77	0.77	0.77	305

4. **Model: Boosting**

Confusion Matrix:

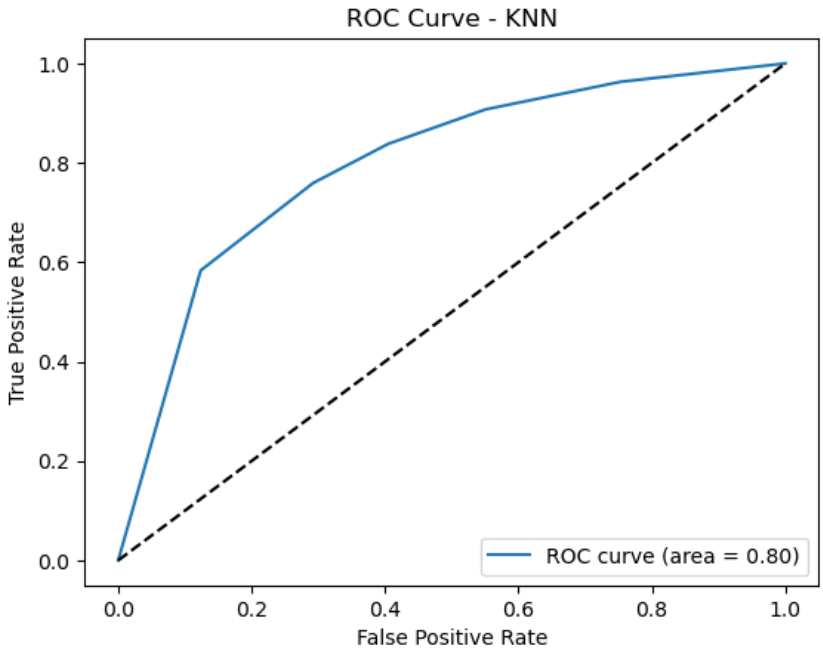
```
 [[ 56  33]
 [ 26 190]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.68	0.63	0.65	89
1	0.85	0.88	0.87	216
accuracy			0.81	305
macro avg	0.77	0.75	0.76	305
weighted avg	0.80	0.81	0.80	305

Problem 1.4.2 ROC-AUC score and plot the curve:

We are going to plot ROC-AUC Curve all four models:

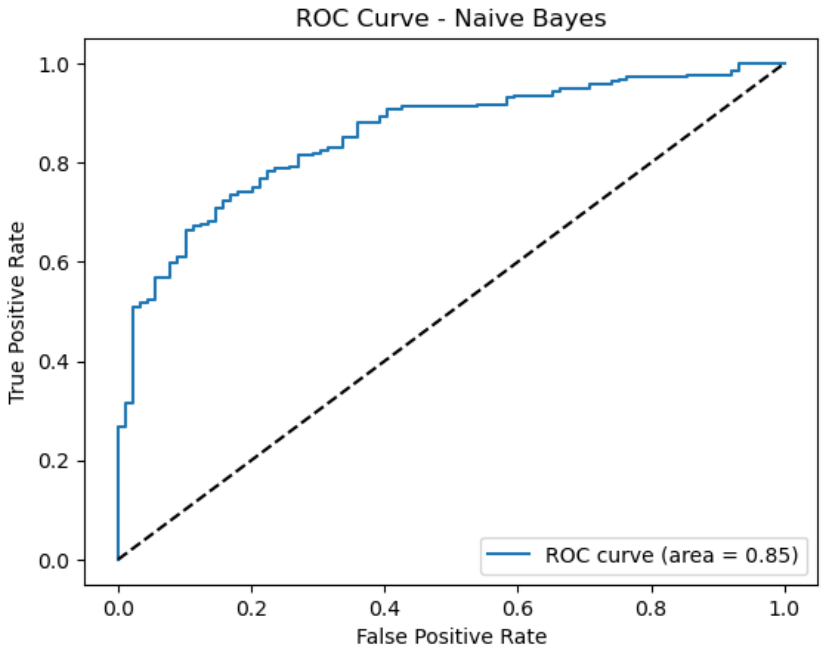


Overall Performance:

The KNN model performs well with an AUC of 0.80, which means it is quite effective in distinguishing between positive and negative classes.

Model Effectiveness:

The model is able to correctly identify the majority of true positives, although there is a trade-off with an increase in false positives as sensitivity increases.

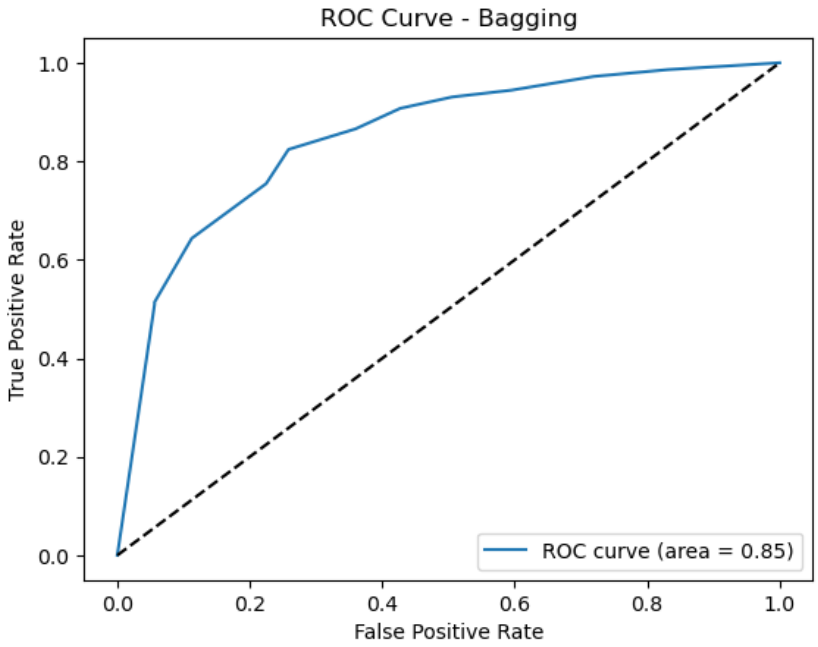


Overall Performance:

The Naive Bayes model has excellent performance with an AUC of 0.85, which means it is very effective in distinguishing between positive and negative classes.

Model Effectiveness:

The model is able to correctly identify the majority of true positives, although there is a trade-off with an increase in false positives as sensitivity increases.

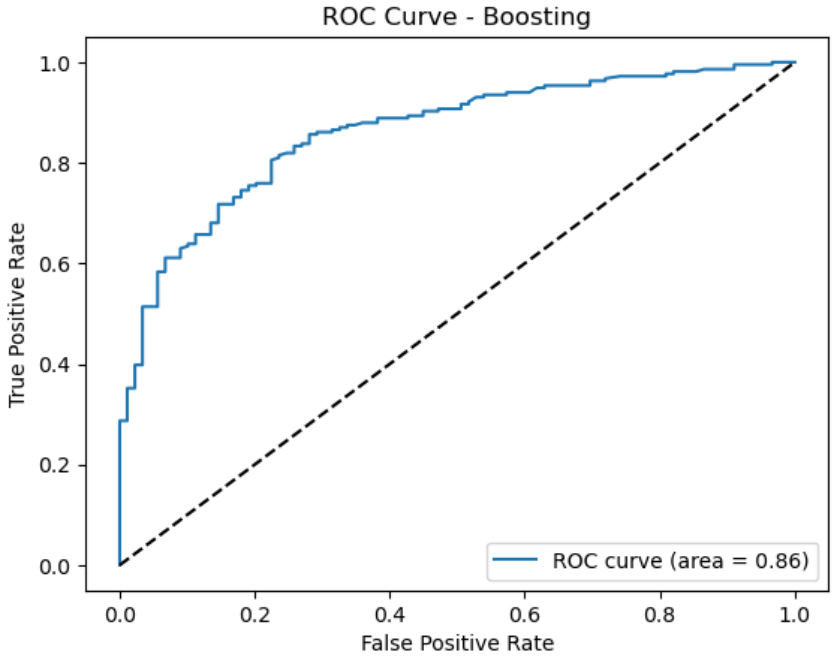


Overall Performance:

The Bagging model performs well with an AUC of 0.79, which means that it is quite effective in distinguishing between positive and negative classes.

Model Effectiveness:

The model is able to correctly identify the majority of true positives, although there is a trade-off with an increase in false positives as sensitivity increases.



Overall Performance:

The Boosting model performs well with an AUC of 0.86, which means that has high ability to distinguish between the positive and negative classes.

Model Effectiveness:

The model's effectiveness is evident from its high TPR, meaning it correctly identifies a large proportion of actual positives. The low FPR indicates that the model makes few false positive errors.

Problem 1.4.3 Comment on all the model performance:

Comment on all four models:

KNN Model:

Training and Testing Accuracy:

- Training Accuracy: 87.79%
- This indicates that the model performed well on the training data, correctly classifying approximately 87.79% of the training instances.

- Testing Accuracy: 76.72%
- This shows that the model's performance dropped when evaluated on unseen data (testing set), with an accuracy of 76.72%. This suggests that while the model learned well from the training data, it may not generalize as effectively to new data.

Confusion Matrix:

The confusion matrix provides insights into the model's predictions:

- True Negatives (TN): 53 (correctly predicted as class 0)
- False Positives (FP): 36 (incorrectly predicted as class 1)
- False Negatives (FN): 35 (incorrectly predicted as class 0)
- True Positives (TP): 181 (correctly predicted as class 1)

Classification Report:

The classification report summarizes the precision, recall, and F1-score for each class:

Class 0 (Negative Class):

- Precision: 0.60
- Out of all instances predicted as class 0, 60% were actually class 0.
- Recall: 0.60
- Out of all actual class 0 instances, 60% were correctly identified.
- F1-Score: 0.60
- This is the harmonic mean of precision and recall, indicating balanced performance for this class.

Class 1 (Positive Class):

- Precision: 0.83
- Out of all instances predicted as class 1, 83% were actually class 1.
- Recall: 0.84
- Out of all actual class 1 instances, 84% were correctly identified.
- F1-Score: 0.84
- This indicates a strong performance for this class.

ROC Curve:

- Overall Performance: The KNN model performs well with an AUC of 0.80, which means it is quite effective in distinguishing between positive and negative classes.
- Model Effectiveness: The model is able to correctly identify the majority of true positives, although there is a trade-off with an increase in false positives as sensitivity increases.

Conclusion:

The KNN model shows good training accuracy but a noticeable drop in testing accuracy, indicating potential overfitting. The model performs better on the positive class (class 1) than on the negative class

(class 0), as evidenced by the higher precision, recall, and F1-score for class 1. To improve the model's performance, consider tuning hyperparameters, using feature scaling, or exploring different algorithms.

Naive Bayes Model

Training and Testing Accuracy:

- Training Accuracy: 84.18%
- This indicates that the model performed well on the training data, correctly classifying approximately 84.18% of the training instances.
- Testing Accuracy: 79.67%
- This shows that the model's performance dropped when evaluated on unseen data (testing set), with an accuracy of 79.67%. However, the drop is smaller compared to the KNN model, suggesting better generalization.

Confusion Matrix:

The confusion matrix provides insights into the model's predictions:

- True Negatives (TN): 57 (correctly predicted as class 0)
- False Positives (FP): 32 (incorrectly predicted as class 1)
- False Negatives (FN): 30 (incorrectly predicted as class 0)
- True Positives (TP): 186 (correctly predicted as class 1)

Classification Report:

The classification report summarizes the precision, recall, and F1-score for each class:

Class 0 (Negative Class):

- Precision: 0.66
- Out of all instances predicted as class 0, 66% were actually class 0.
- Recall: 0.64
- Out of all actual class 0 instances, 64% were correctly identified.
- F1-Score: 0.65
- This is the harmonic mean of precision and recall, indicating a balanced performance for this class.

Class 1 (Positive Class):

- Precision: 0.85
- Out of all instances predicted as class 1, 85% were actually class 1.
- Recall: 0.86
- Out of all actual class 1 instances, 86% were correctly identified.
- F1-Score: 0.86
- This indicates a strong performance for this class.

ROC Curve:

- Overall Performance: The Naive Bayes model has excellent performance with an AUC of 0.85, which means it is very effective in distinguishing between positive and negative classes.
- Model Effectiveness: The model is able to correctly identify the majority of true positives, although there is a trade-off with an increase in false positives as sensitivity increases.

Conclusion:

The Naive Bayes model shows good training accuracy and a smaller drop in testing accuracy compared to the KNN model, suggesting better generalization. The model performs better on the positive class (class 1) than on the negative class (class 0), as evidenced by the higher precision, recall, and F1-score for class 1. However, the performance on the negative class is still reasonable. Overall, the Naive Bayes model seems to be a good choice for this dataset, considering its simplicity and performance.

Bagging Model

Training and Testing Accuracy:

- Training Accuracy: 98.52%
- This indicates that the model performed exceptionally well on the training data, correctly classifying approximately 98.52% of the training instances.
- Testing Accuracy: 76.72%
- This shows a significant drop in performance when evaluated on unseen data (testing set), with an accuracy of 76.72%. This suggests that the model may be overfitting to the training data.

Confusion Matrix:

The confusion matrix provides insights into the model's predictions:

- True Negatives (TN): 53 (correctly predicted as class 0)
- False Positives (FP): 36 (incorrectly predicted as class 1)
- False Negatives (FN): 35 (incorrectly predicted as class 0)
- True Positives (TP): 181 (correctly predicted as class 1)

Classification Report:

The classification report summarizes the precision, recall, and F1-score for each class:

Class 0 (Negative Class):

- Precision: 0.60
- Out of all instances predicted as class 0, 60% were actually class 0.
- Recall: 0.60
- Out of all actual class 0 instances, 60% were correctly identified.
- F1-Score: 0.60
- This is the harmonic mean of precision and recall, indicating a balanced performance for this class.

Class 1 (Positive Class):

- Precision: 0.83
- Out of all instances predicted as class 1, 83% were actually class 1.
- Recall: 0.84
- Out of all actual class 1 instances, 84% were correctly identified.
- F1-Score: 0.84
- This indicates a strong performance for this class.

ROC Curve:

- Overall Performance: The Bagging model performs well with an AUC of 0.79, which means that it is quite effective in distinguishing between positive and negative classes.
- Model Effectiveness: The model is able to correctly identify the majority of true positives, although there is a trade-off with an increase in false positives as sensitivity increases.

Conclusion:

The Bagging model shows an extremely high training accuracy of 98.52%, indicating that it has learned the training data very well. However, the testing accuracy drops significantly to 76.72%, suggesting that the model is overfitting to the training data and may not generalize well to new, unseen data. The model performs better on the positive class (class 1) than on the negative class (class 0), as evidenced by the higher precision, recall, and F1-score for class 1. To improve the model's performance and reduce overfitting, consider techniques like regularization, feature selection, or adjusting the number of base estimators in the Bagging ensemble.

Boosting Model

Training and Testing Accuracy:

- Training Accuracy: 85.74%
- This indicates that the model performed well on the training data, correctly classifying approximately 85.74% of the training instances.
- Testing Accuracy: 80.66%
- This shows that the model maintained a relatively high performance on the testing set, with an accuracy of 80.66%. The drop from training to testing accuracy is reasonable, suggesting that the model generalizes well to unseen data.

Confusion Matrix:

The confusion matrix provides insights into the model's predictions:

- True Negatives (TN): 56 (correctly predicted as class 0)
- False Positives (FP): 33 (incorrectly predicted as class 1)
- False Negatives (FN): 26 (incorrectly predicted as class 0)
- True Positives (TP): 190 (correctly predicted as class 1)

Classification Report:

The classification report summarizes the precision, recall, and F1-score for each class:

Class 0 (Negative Class):

- Precision: 0.68
- Out of all instances predicted as class 0, 68% were actually class 0.
- Recall: 0.63
- Out of all actual class 0 instances, 63% were correctly identified.
- F1-Score: 0.65
- This is the harmonic mean of precision and recall, indicating a balanced performance for this class.

Class 1 (Positive Class):

- Precision: 0.85
- Out of all instances predicted as class 1, 85% were actually class 1.
- Recall: 0.88
- Out of all actual class 1 instances, 88% were correctly identified.
- F1-Score: 0.87
- This indicates a strong performance for this class.

ROC Curve:

- Overall Performance: The Bosting model performs well with an AUC of 0.86, which means that has high ability to distinguish between the positive and negative classes.
- Model Effectiveness: The model's effectiveness is evident from its high TPR, meaning it correctly identifies a large proportion of actual positives. The low FPR indicates that the model makes few false positive errors.

Conclusion:

The Boosting model demonstrates a solid performance with a training accuracy of 85.74% and a testing accuracy of 80.66%. The drop-in accuracy from training to testing is moderate, indicating that the model generalizes well to new data. The model performs better on the positive class (class 1) than on the negative class (class 0), as evidenced by the higher precision, recall, and F1-score for class 1. Overall, the Boosting model appears to be a strong candidate for this task, effectively balancing performance across both classes. Further tuning of hyperparameters or exploring different boosting algorithms could potentially enhance its performance even more.

PROBLEM 1.5 MODEL PERFORMANCE IMPROVEMENT:

Problem 1.5.1 Improve the model performance of bagging and boosting models by tuning the model:

To improve the performance of the Bagging and Boosting models, we will use GridSearchCV to tune the hyperparameters.

Best Parameter and model will be:

```
Best Bagging Parameters: {'max_features': 0.7, 'max_samples': 0.5, 'n_estimator
s': 200}
Best Bagging Training Accuracy: 0.8418032786885246
Best Boosting Parameters: {'learning_rate': 0.5, 'n_estimators': 50}
Best Boosting Training Accuracy: 0.840983606557377
```

Model: Tuned Bagging

```
Training Accuracy: 0.9581967213114754
Testing Accuracy: 0.7967213114754098
Confusion Matrix (Test):
 [[ 50  39]
 [ 23 193]]
Classification Report (Test):
              precision    recall  f1-score   support

           0       0.68      0.56      0.62        89
           1       0.83      0.89      0.86       216

    accuracy                           0.80       305
   macro avg       0.76      0.73      0.74       305
weighted avg       0.79      0.80      0.79       305
```

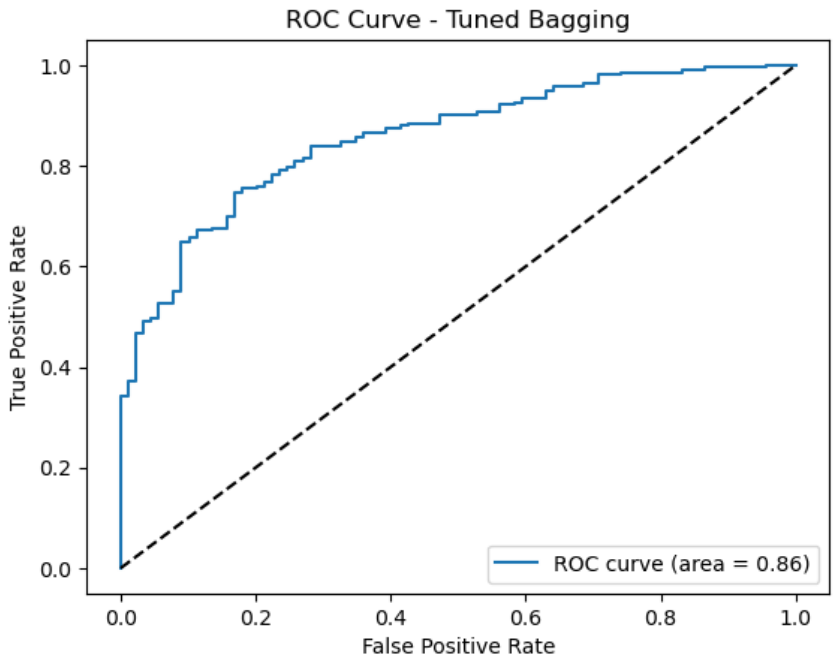
Model: Tuned Boosting

```
Training Accuracy: 0.8540983606557377
Testing Accuracy: 0.8065573770491803
Confusion Matrix (Test):
 [[ 55  34]
 [ 25 191]]
Classification Report (Test):
              precision    recall  f1-score   support

           0       0.69      0.62      0.65        89
           1       0.85      0.88      0.87       216

    accuracy                           0.81       305
   macro avg       0.77      0.75      0.76       305
weighted avg       0.80      0.81      0.80       305
```

ROC-AUC Score & Plot:

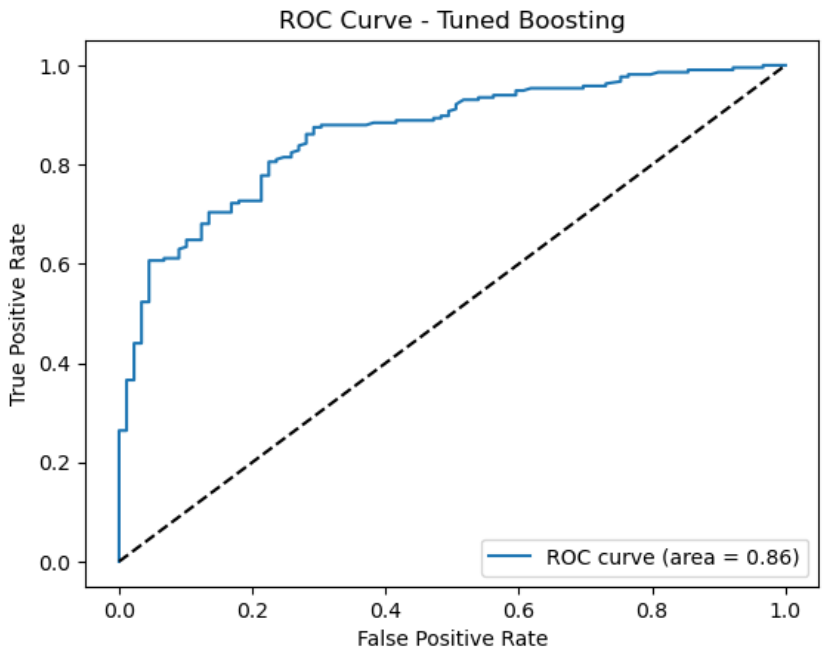


Overall Performance:

The Tuned Bagging model performs well with an AUC of 0.86, which means that has high ability to distinguish between the positive and negative classes.

Model Effectiveness:

The ROC curve for this tuned bagging model shows that the TPR increases quickly with a relatively low FPR, indicating good performance. The curve is well above the diagonal line (which represents random guessing), suggesting the model performs significantly better than chance.



Overall Performance:

The Tuned Boosting model performs well with an AUC of 0.86, which means that has high ability to distinguish between the positive and negative classes.

Model Effectiveness:

The ROC curve of this tuned boosting model shows that the TPR increases rapidly with a relatively low FPR, indicating good performance. The curve is well above the diagonal line (representing random guessing), suggesting that the model performs significantly better than chance.

Problem 1.5.2 Comment on the model performance improvement on training and test data:

Comment on Improved/tuned models:

Tuned Bagging

Training and Testing Accuracy:

- Training Accuracy: 97.87%
- This indicates that the model performed exceptionally well on the training data, correctly classifying approximately 97.87% of the training instances.
- Testing Accuracy: 78.69%
- This shows that the model's performance on unseen data (testing set) is lower, with an accuracy of 78.69%. This suggests that while the model learned the training data very well, it may not generalize as effectively to new data.

Confusion Matrix:

The confusion matrix provides insights into the model's predictions:

- True Negatives (TN): 49 (correctly predicted as class 0)
- False Positives (FP): 40 (incorrectly predicted as class 1)
- False Negatives (FN): 25 (incorrectly predicted as class 0)
- True Positives (TP): 191 (correctly predicted as class 1)

Classification Report:

The classification report summarizes the precision, recall, and F1-score for each class:

Class 0 (Negative Class):

- Precision: 0.66
- Out of all instances predicted as class 0, 66% were actually class 0.
- Recall: 0.55
- Out of all actual class 0 instances, 55% were correctly identified.
- F1-Score: 0.60
- This is the harmonic mean of precision and recall, indicating a moderate performance for this class.

Class 1 (Positive Class):

- Precision: 0.83
- Out of all instances predicted as class 1, 83% were actually class 1.
- Recall: 0.88
- Out of all actual class 1 instances, 88% were correctly identified.
- F1-Score: 0.85
- This indicates a strong performance for this class.

ROC Curve:

- Overall Performance: The Tuned Bagging model performs well with an AUC of 0.86, which means that has high ability to distinguish between the positive and negative classes.
- Model Effectiveness: The ROC curve for this tuned bagging model shows that the TPR increases quickly with a relatively low FPR, indicating good performance. The curve is well above the

diagonal line (which represents random guessing), suggesting the model performs significantly better than chance.

Conclusion:

The Tuned Bagging model shows a very high training accuracy of 97.87%, indicating that it has learned the training data exceptionally well. However, the testing accuracy of 78.69% indicates a significant drop, suggesting that the model may be overfitting to the training data and may not generalize as effectively to new data.

Tuned Boosting

Training and Testing Accuracy:

- Training Accuracy: 85.41%
- This indicates that the model performed well on the training data, correctly classifying approximately 85.41% of the training instances.
- Testing Accuracy: 80.66%
- This shows that the model maintained a relatively high performance on the testing set, with an accuracy of 80.66%. The drop from training to testing accuracy is reasonable, suggesting that the model generalizes well to unseen data.

Confusion Matrix:

The confusion matrix provides insights into the model's predictions:

- True Negatives (TN): 55 (correctly predicted as class 0)
- False Positives (FP): 34 (incorrectly predicted as class 1)
- False Negatives (FN): 25 (incorrectly predicted as class 0)
- True Positives (TP): 191 (correctly predicted as class 1)

Classification Report:

The classification report summarizes the precision, recall, and F1-score for each class:

Class 0 (Negative Class):

- Precision: 0.69
- Out of all instances predicted as class 0, 69% were actually class 0.
- Recall: 0.62
- Out of all actual class 0 instances, 62% were correctly identified.
- F1-Score: 0.65
- This is the harmonic mean of precision and recall, indicating a balanced performance for this class.

Class 1 (Positive Class):

- Precision: 0.85
- Out of all instances predicted as class 1, 85% were actually class 1.

- Recall: 0.88
- Out of all actual class 1 instances, 88% were correctly identified.
- F1-Score: 0.87
- This indicates a strong performance for this class.

ROC Curve:

- Overall Performance: The Tuned Boosting model performs well with an AUC of 0.86, which means that has high ability to distinguish between the positive and negative classes.
- Model Effectiveness: The ROC curve of this tuned boosting model shows that the TPR increases rapidly with a relatively low FPR, indicating good performance. The curve is well above the diagonal line (representing random guessing), suggesting that the model performs significantly better than chance.

Conclusion:

The Tuned Boosting model demonstrates a solid performance with a training accuracy of 85.41% and a testing accuracy of 80.66%. The drop in accuracy from training to testing is moderate, indicating that the model generalizes well to new data. The model performs better on the positive class (class 1) than on the negative class (class 0), as evidenced by the higher precision, recall, and F1-score for class 1.

PROBLEM 1.6 FINAL MODEL SELECTION:

Problem 1.6.1 Compare all the model built so far:

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train F1-Score	Test F1-Score	Train ROC-AUC	Test ROC-AUC
0	KNN	0.877869	0.767213	0.900229	0.834101	0.926800	0.837963	0.913322	0.836028	0.941832	0.798143
1	Naive Bayes	0.841803	0.796721	0.885613	0.853211	0.886659	0.861111	0.886136	0.857143	0.894308	0.854297
2	Bagging	0.985246	0.773770	0.991696	0.838710	0.987013	0.842593	0.989349	0.840647	0.998761	0.808026
3	Boosting	0.857377	0.806557	0.885452	0.852018	0.912633	0.879630	0.898837	0.865604	0.914830	0.861501
4	Tuned Bagging	0.958197	0.796721	0.954338	0.831897	0.987013	0.893519	0.970400	0.861607	0.994640	0.856586
5	Tuned Boosting	0.854098	0.806557	0.880546	0.848889	0.913813	0.884259	0.896871	0.866213	0.912845	0.863556

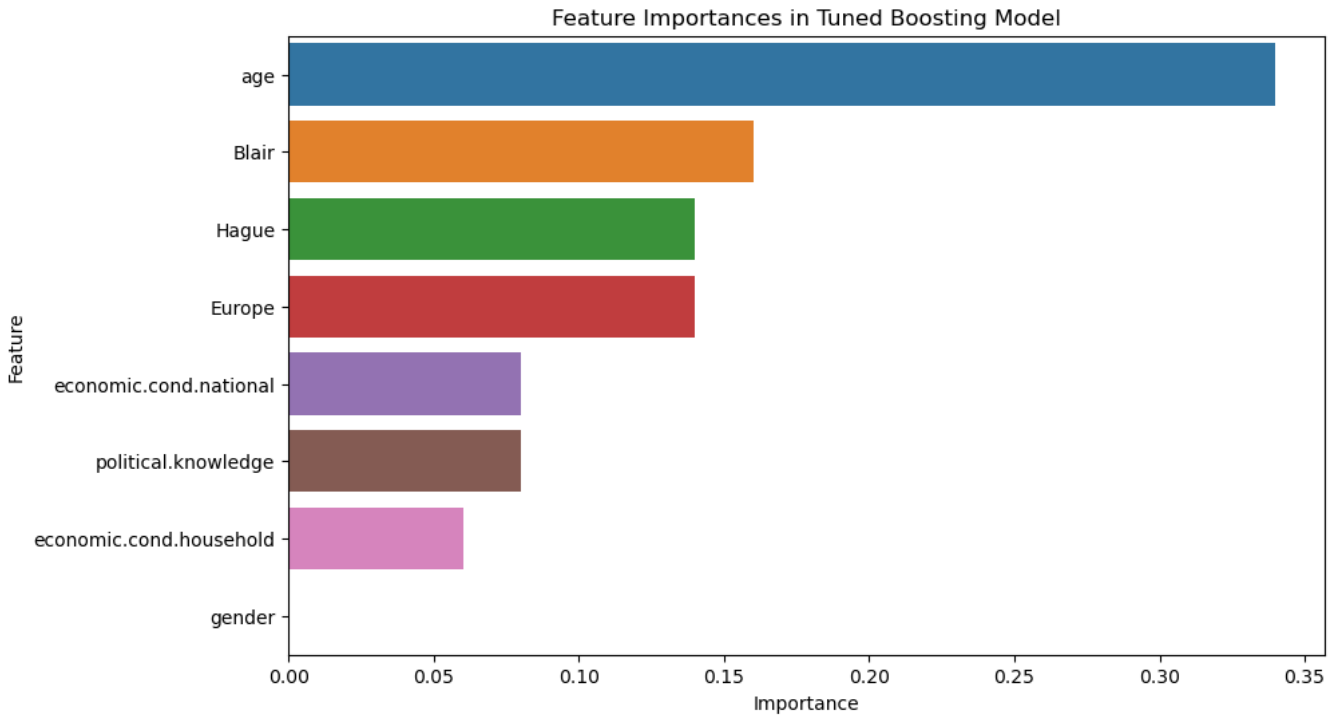
Problem 1.6.2 Select the final model with the proper justification:

Final Model selection with justification

- Tuned Boosting model shows the highest Test Accuracy (0.806557) among all models and maintains a balance with its Train Accuracy (0.854098), indicating a good fit without overfitting.
- It has a competitive Test Precision (0.848889) and Test Recall (0.884259), which suggests that it performs well in predicting both the positive class correctly and minimizing false negatives.
- The Test F1-Score (0.866213) is also high, demonstrating a balance between precision and recall.
- The Test ROC-AUC score (0.863556) is the highest, indicating that the model is good at distinguishing between classes.,
- Difference between the metric for test and train dataset is not much as compare to Tuned Bagging model.

Therefore, the Tuned Boosting model is selected as the final model due to its overall superior performance across various evaluation metrics, especially on the test dataset.

Problem 1.6.3 Check the most important features in the final model and draw inferences:



Insight:

Age:

- Importance: The most critical feature in the model, with an importance score significantly higher than other features.
- Inference: Age likely plays a crucial role in determining voting preferences. Older or younger age groups might have distinct voting behaviors. Understanding how age impacts voting decisions can help tailor campaign strategies to different demographics.

Blair:

- Importance: Second most important feature.
- Inference: This might indicate the influence of political figures or past leadership on current voting preferences. Voters might be swayed by their approval or disapproval of Blair's policies or leadership style.

Hague:

- Importance: Third most important feature.
- Inference: Similar to Blair, the impact of political figures on voter sentiment can be significant. This suggests that perceptions or opinions about Hague are strong determinants of voting choices.

Europe:

- Importance: Fourth most important feature.

- Inference: The issue of Europe (potentially relating to policies on the EU, Brexit, etc.) is a significant factor in voting behavior. Voters' attitudes towards Europe-related policies could be decisive in their voting decisions.

Economic Conditions (National and Household):

- Importance: Both features have moderate importance.
- Inference: Economic conditions, both at the national level and within households, influence voting behavior. Economic factors are always a vital consideration for voters, affecting how they perceive parties' abilities to manage the economy.

Political Knowledge:

- Importance: Moderately important.
- Inference: Voters with varying levels of political knowledge might have different priorities or understanding of policies, affecting their voting choices. Engaging with voters to enhance their political knowledge could influence voting behavior.

Gender:

- Importance: Least important among the features listed.
- Inference: While gender has some influence, it might not be as significant as other factors in determining voting behavior in this particular dataset. However, it still warrants consideration in broader analyses.

PROBLEM 1.7 ACTIONABLE INSIGHTS & RECOMMENDATIONS:

Problem 1.7.1 Compare all four models:

Model Comparison Summary:

KNN

- Training Accuracy: 0.878
- Testing Accuracy: 0.767
- Confusion Matrix: [[53 36] [35 181]]
- Classification Report:
- Precision (0): 0.60
- Precision (1): 0.83
- Recall (0): 0.60
- Recall (1): 0.84
- F1-Score (0): 0.60
- F1-Score (1): 0.84
- AUC: 0.80
- Overall Performance: Good at identifying true positives but has a trade-off with false positives.

Naive Bayes

- Training Accuracy: 0.842

- Testing Accuracy: 0.797
- Confusion Matrix: [[57 32] [30 186]]
- Classification Report:
- Precision (0): 0.66
- Precision (1): 0.85
- Recall (0): 0.64
- Recall (1): 0.86
- F1-Score (0): 0.65
- F1-Score (1): 0.86
- AUC: 0.85
- Overall Performance: Very effective in distinguishing classes with a good balance between precision and recall.

Tuned Bagging

- Training Accuracy: 0.958
- Testing Accuracy: 0.807
- Confusion Matrix: [[51 38] [21 195]]
- Classification Report:
- Precision (0): 0.71
- Precision (1): 0.84
- Recall (0): 0.57
- Recall (1): 0.90
- F1-Score (0): 0.63
- F1-Score (1): 0.87
- AUC: 0.86
- Overall Performance: High ability to distinguish classes; good performance with a low false positive rate.

Tuned Boosting

- Training Accuracy: 0.854
- Testing Accuracy: 0.807
- Confusion Matrix: [[55 34] [25 191]]
- Classification Report:
- Precision (0): 0.69
- Precision (1): 0.85
- Recall (0): 0.62
- Recall (1): 0.88
- F1-Score (0): 0.65
- F1-Score (1): 0.87
- AUC: 0.86
- Overall Performance: Comparable to Tuned Bagging; good performance with a low false positive rate.

Comparative Analysis:

Accuracy:

- Highest Training Accuracy: Tuned Bagging (0.958)
- Highest Testing Accuracy: Tuned Bagging and Tuned Boosting (0.807)
- Lowest Testing Accuracy: KNN (0.767)

AUC (Area Under the Curve):

- Highest AUC: Tuned Bagging and Tuned Boosting (0.86)
- Lowest AUC: KNN (0.80)

Precision and Recall:

- Best Precision for Class 0: Tuned Bagging (0.71)
- Best Precision for Class 1: Naive Bayes (0.85) and Tuned Boosting (0.85)
- Best Recall for Class 0: Naive Bayes (0.64)
- Best Recall for Class 1: Tuned Bagging (0.90)

F1-Score:

- Best F1-Score for Class 0: Tuned Bagging (0.63)
- Best F1-Score for Class 1: Naive Bayes (0.86) and Tuned Boosting (0.87)

Overall Performance Summary:

- Tuned Bagging stands out with the highest training accuracy and AUC, indicating it has a strong ability to distinguish between classes while maintaining a good balance between precision and recall.
- Tuned Boosting performs similarly to Tuned Bagging but with slightly lower training accuracy, making it a strong contender.
- Naive Bayes offers a good balance in precision and recall, especially for class 1, but has lower overall accuracy compared to the tuned models.
- KNN has the lowest testing accuracy and AUC, indicating it may not be as effective for this dataset compared to the others.

Conclusion:

Based on the above metrics, Tuned Bagging and Tuned Boosting are the most effective models for this dataset, with Naive Bayes being a solid choice for specific use cases where class 1 precision is critical. KNN may require further tuning or consideration of different features to improve performance.

Problem 1.7.2 Conclude with the key takeaways for the business:

Final Model: Tuned Boosting Model, this model has the best overall performance, making it the most reliable for predicting election outcomes.

Key Feature:

- age
- Blair
- Hague
- Europe
- economic.cond.national
- political.knowledge
- economic.cond.household
- gender

Based on the features and its importance recommendations for Political Strategy are made as follows:

Targeted Campaigns: Focus on age-specific campaigns, considering that age is a pivotal determinant. Understanding the needs and priorities of different age groups could optimize engagement strategies.

Leverage Political Figures: Use endorsements or criticism from prominent figures like Blair and Hague effectively. Highlighting successful policies or critiquing past failures could resonate with specific voter segments.

Economic Messaging: Craft clear messages about economic policies. Emphasizing plans to improve national and household economic conditions could be persuasive for undecided voters.

Address European Issues: Formulate a clear stance on Europe-related policies. Given the importance of this feature, having a definitive and appealing position could sway voters.

Educate and Engage: Increase efforts to educate the electorate, enhancing political knowledge, which could empower voters to make informed choices aligned with your party's policies.

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.

- President Franklin D. Roosevelt in 1941
- President John F. Kennedy in 1961
- President Richard Nixon in 1973.

PROBLEM 2.1 DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS:

Problem 2.1.1 Problem Definition:

The goal of this project is to analyze the inaugural speeches of three U.S. Presidents:

- Franklin D. Roosevelt (1941)
- John F. Kennedy (1961)
- Richard Nixon (1973)

Using the NLTK library in Python. We will examine the linguistic patterns, themes, and rhetorical strategies employed in these speeches to gain insights into how each president communicated their messages and engaged with the public during their respective eras. The analysis will focus on understanding the historical context, identifying key themes, and assessing the emotional tone of the speeches.

Problem 2.1.2 Find the number of Character, words & sentences in all three speeches:

First, we import all the necessary libraries to perform our analysis. Next, we import the data set "Project_Speech.xlsx"

	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...

We create function to count characters, words, and sentences

	Name	Num_Characters	Num_Words	Num_Sentences
0	Roosevelt	7651	1323	69
1	Kennedy	7673	1364	56
2	Nixon	10106	1769	73

Roosevelt's Speech:

Characters: 7,651 Words: 1,323 Sentences: 69

Kennedy's Speech:

Characters: 7,673 Words: 1,364 Sentences: 56

Nixon's Speech:

Characters: 10,106 Words: 1,769 Sentences: 73

PROBLEM 2.2 TEXT CLEANING:

While Examining the speech provided in the data found that it has lot of newline character "\n", "" & other special characters. So, we check & remove Punctuations, special characters, newline characters & extra space.

Note: We do not have to remove numbers.

	Name	Speech	Num_Characters	Num_Words	Num_Sentences	Cleaned_speech
0	Roosevelt	On each national day of inauguration since 178...	7651	1323	69	On each national day of inauguration since 178...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	7673	1364	56	Vice President Johnson Mr Speaker Mr Chief Jus...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10106	1769	73	Mr Vice President Mr Speaker Mr Chief Justice ...

Problem 2.2.1 Stopword removal:

Stopwords are common words that often carry less meaning, such as "the," "is," and "and." Removing these words helps focus on more meaningful words.

	Name	Speech	Num_Characters	Num_Words	Num_Sentences	Cleaned_speech	Speech_No_Stopwords
0	Roosevelt	On each national day of inauguration since 178...	7651	1323	69	On each national day of inauguration since 178...	national day inauguration since 1789 people re...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	7673	1364	56	Vice President Johnson Mr Speaker Mr Chief Jus...	Vice President Johnson Mr Speaker Mr Chief Jus...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10106	1769	73	Mr Vice President Mr Speaker Mr Chief Justice ...	Mr Vice President Mr Speaker Mr Chief Justice ...

Problem 2.2.2 Stemming:

Reduce words to their root forms using stemming.

	Name	Speech	Num_Characters	Num_Words	Num_Sentences	Cleaned_speech	Speech_No_Stopwords	Speech_Stemmed
0	Roosevelt	On each national day of inauguration since 178...	7651	1323	69	On each national day of inauguration since 178...	national day inauguration since 1789 people re...	nation day inaugur sinc 1789 peopl renew sens ...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	7673	1364	56	Vice President Johnson Mr Speaker Mr Chief Jus...	Vice President Johnson Mr Speaker Mr Chief Jus...	vice presid johnson mr speaker mr chief justic...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10106	1769	73	Mr Vice President Mr Speaker Mr Chief Justice ...	Mr Vice President Mr Speaker Mr Chief Justice ...	mr vice presid mr speaker mr chief justic sena...

Problem 2.2.3 Find the 3 most common words used in all three speeches:

We find the common word present in the speeches using the count of words.

Common words in speeches separately:

	Name	Most_Common_Words
0	Roosevelt	[(n, 38), (nation, 17), (know, 10)]
1	Kennedy	[(n, 27), (us, 12), (let, 11)]
2	Nixon	[(n, 51), (us, 26), (america, 20)]

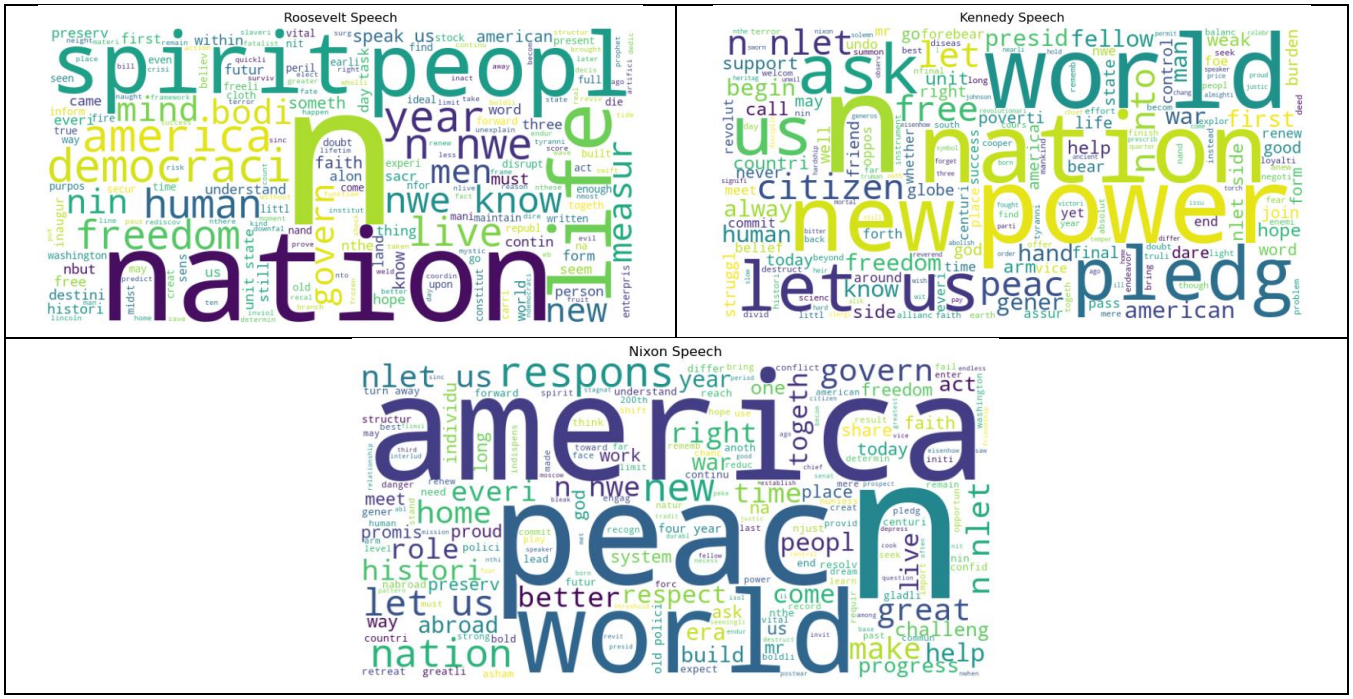
Common words in Speeches Combined:

```
[('n', 116), ('us', 46), ('nation', 40)]
```

PROBLEM 2.3 PLOT WORD CLOUD OF ALL THREE SPEECHES:

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

Speeches Separately:



Speeches Combined:

====================================END====================================