**Machine Learning PROJECT REPORT - Clustering**

Arusa Khalfay

**Table of Contents:**

[**1 Part 1: Clustering 4**](#_ksb7dbvkoy4y)

[1.1 Clustering: Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc. 4](#_lye3rca2ph1w)

[1.2 Clustering: Treat missing values in CPC, CTR and CPM using the formula given. 6](#_xmvuwoenhpq0)

[1.3 Clustering: Check if there are any outliers. Do you think treating outliers is necessary for K-Means clustering? Based on your judgement decide whether to treat outliers and if yes, which method to employ. 8](#_2l4darolbkz4)

[1.4 Clustering: Perform z-score scaling and discuss how it affects the speed of the algorithm. 11](#_8d5g2wvincyd)

[1.5 Clustering: Perform Hierarchical by constructing a Dendrogram using WARD and Euclidean distance. 12](#_n7bznwpw95y)

[1.6 Clustering: Make an Elbow plot (up to n=10) and identify optimum number of clusters for k-means algorithm. 13](#_gld99b2zf65r)

[1.7 Clustering: Print silhouette scores for up to 10 clusters and identify optimum number of clusters. 14](#_q6d0zjsto1yf)

[1.8 Clustering: Profile the ads based on optimum number of clusters using silhouette score and your domain understanding [Hint: Group the data by clusters and take sum or mean to identify trends in Clicks, spend, revenue, CPM, CTR, & CPC based on Device Type. Make bar plots]. 14](#_hcwmpoy38s26)

[1.9 Clustering: Conclude the project by providing a summary of your learnings. 15](#_hyc44l3mbp6y)

**List of Figures:**

| Figure 1 | Checking head of given dataset | 4 |
| --- | --- | --- |
| Figure 2 | Info of dataset | 5 |
| Figure 3 | Summary of dataset | 5 |
| Figure 4 | Univariate Analysis | 7 |
| Figure 5 | Train dataset | 13 |
| Figure 6 | Test dataset | 13 |
| Figure 7 | Logistic Regression Results | 15 |
| Figure 8 | Logistic Regression Coeff | 16 |
| Figure 9 | LDA results | 17 |
| Figure 10 | KNN results | 18 |
| Figure 11 | NaiveNaive Bayes Results | 18 |

# 

# 

# 

# 

# 

# 

# 

# 1. Problem 1

Problem Statement:

You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

## 1.1 Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.

As per the instructions given above we are reading the following data:

1. **Checking Head** - Shows information about the first five rows of the dataset.



Figure 1: Checking head for the given dataset

1. **Checking info** of the dataset

On checking the info of the dataset, we observe that there are 23065 rows and 19 columns.

We also notice that there are 6 columns of the data type float, 7 of integer and 6 of object.

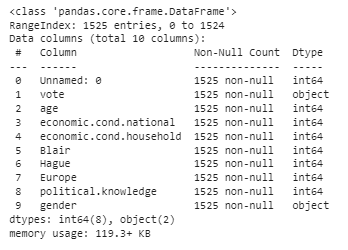


Figure 2: Info of the dataset

1. **Checking Summary** of the dataset

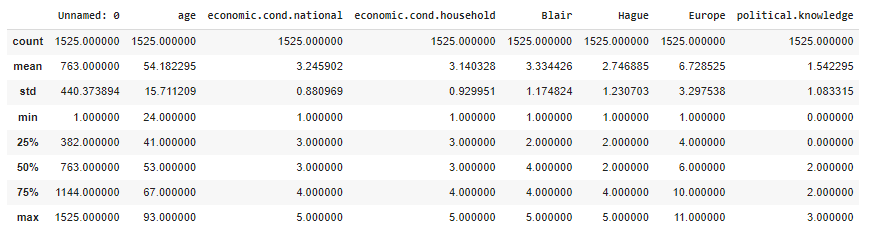
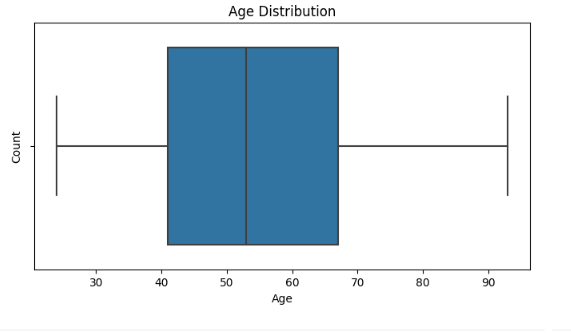
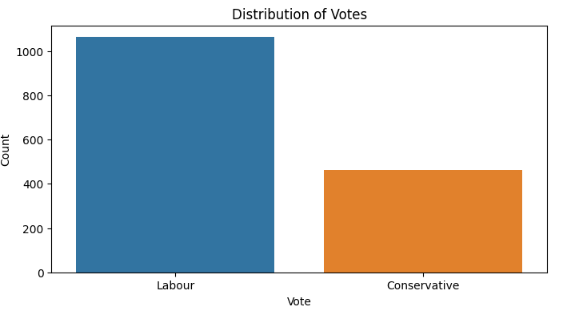


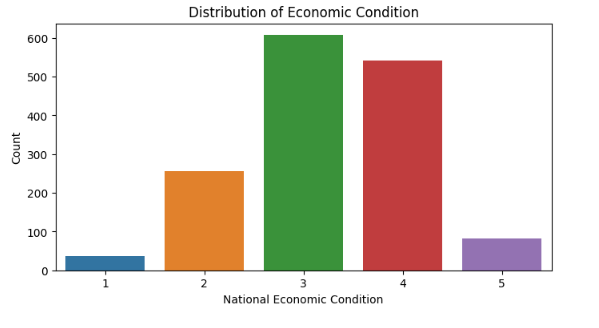
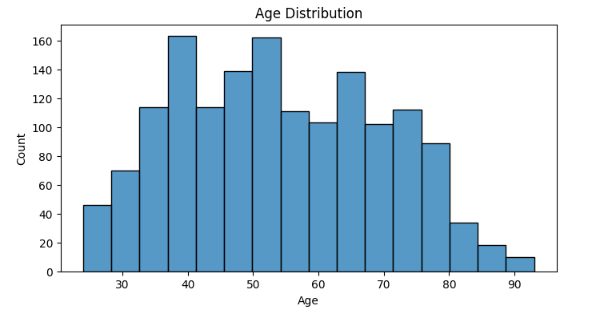
Figure 3: Summary of the dataset

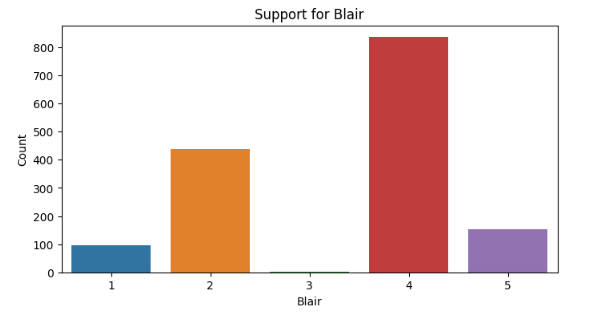
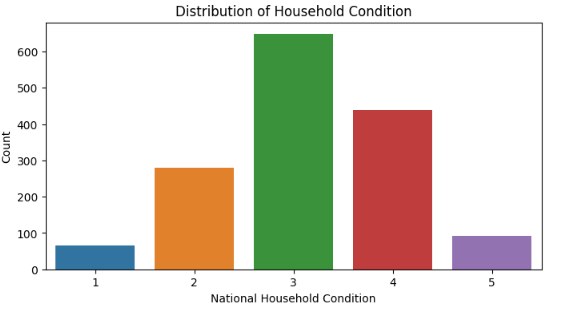
1. **Checking null values** - No null values
2. **Checking Duplicate values** in the dataset

There are no duplicate values in the dataset.

## 1.2 Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.







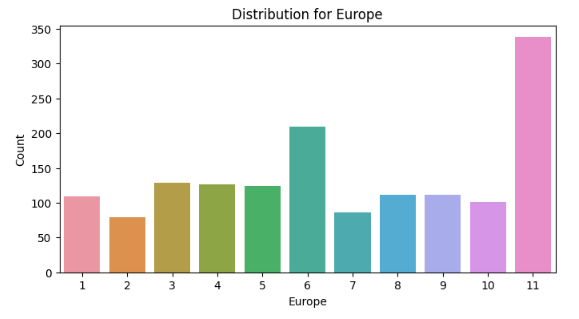
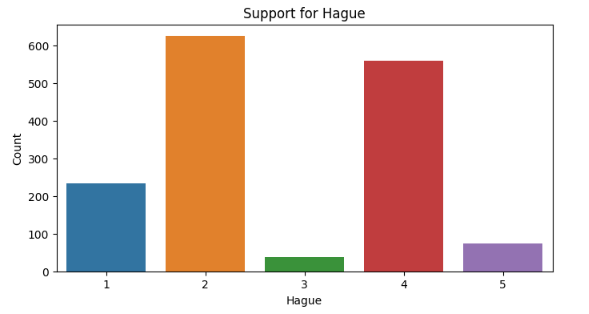
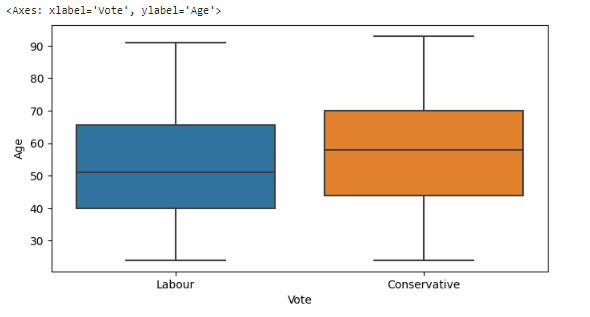


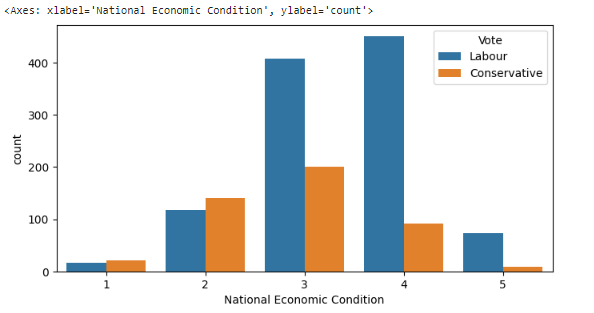
Fig4: Univariate Analysis

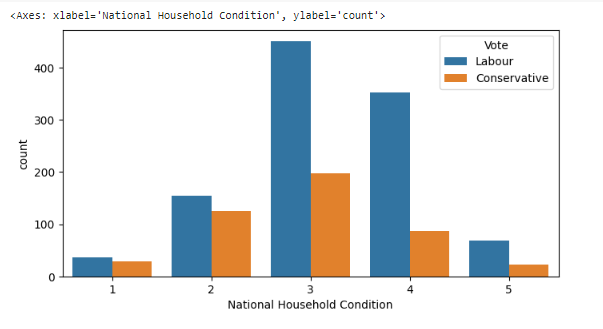
Inferences:

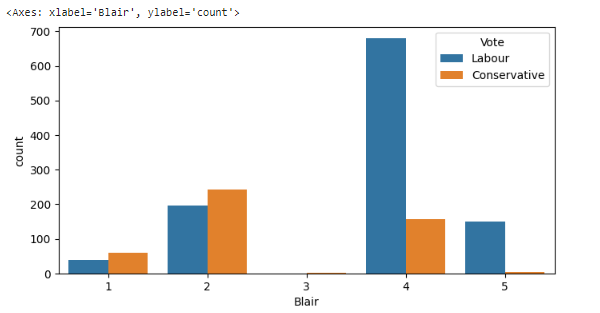
1. The variable 'gender' demonstrates a near-equal distribution of male and female observations.
2. The economic conditions at both household and national levels demonstrate distributions that align with commonly accepted knowledge.
3. The graphical representations of 'Hague' and 'Blair' reveal that only a small number of individuals hold neutral opinions. 'Blair' receives more positive opinions, while 'Hague' receives a greater number of negative opinions.
4. The 'Europe' variable displays proportional representation across its categories, with the exception of '11' category having a comparatively lower count.
5. The distribution of 'age' does not adhere to a Normal Distribution pattern and exhibits a kurtosis value of -0.94, suggesting a flatter distribution than expected in a normal distribution (where kurtosis is 0).
6. According to the Box plot, there are no exceptional data points in the 'age' variable, indicating the absence of outliers.

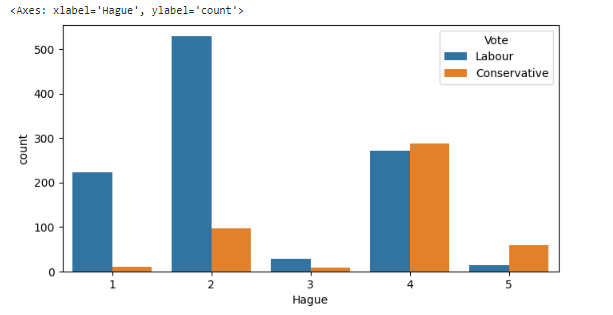
Bivariate Analysis:

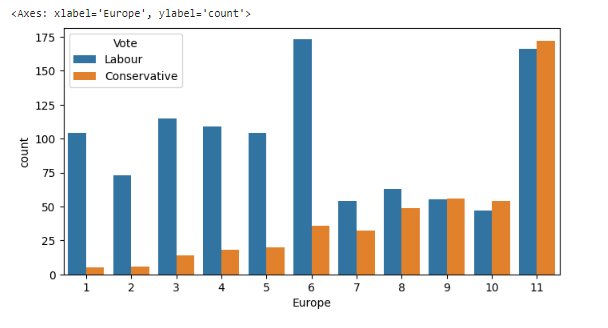


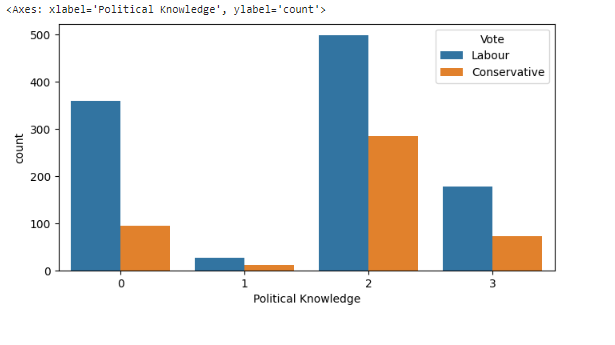


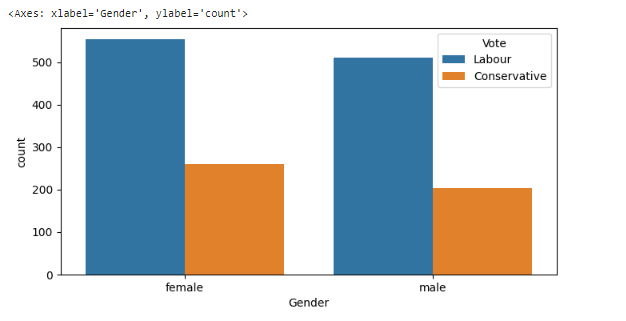


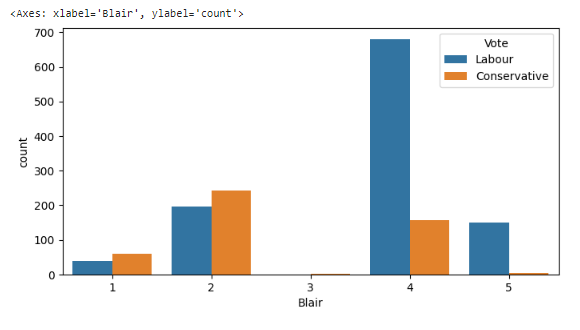


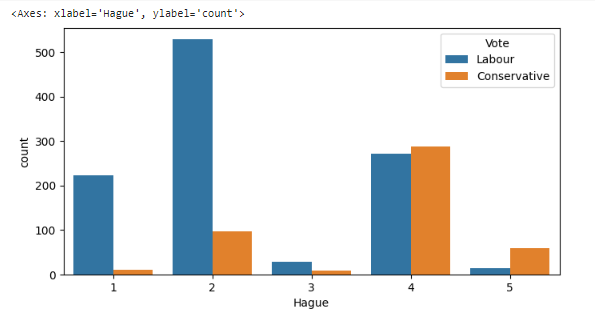


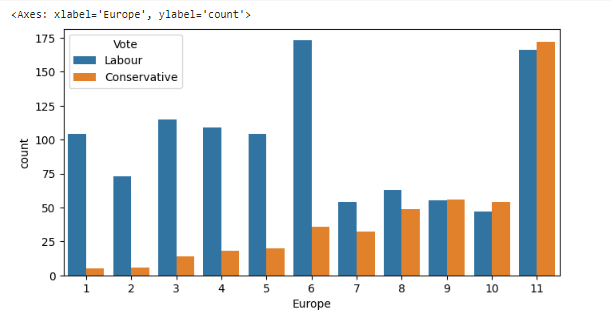


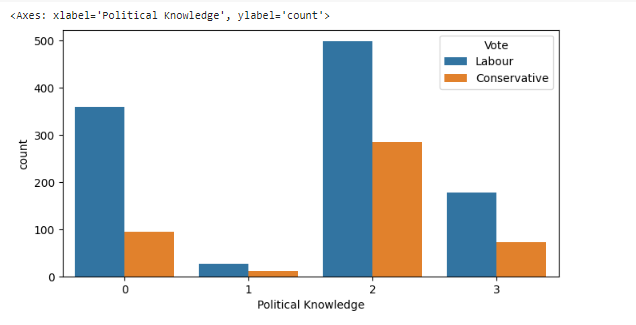


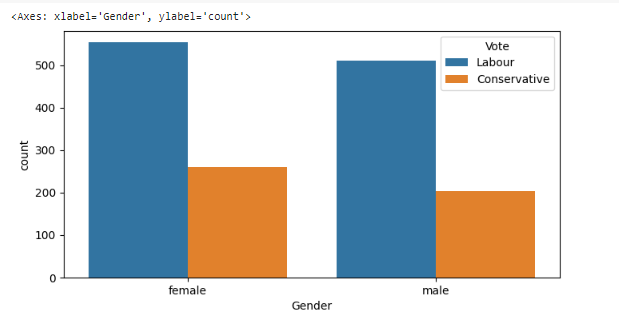












Inferences:

1. The box plot demonstrates a noticeable trend where individuals aged 60 and above exhibit a higher inclination to vote for the 'Conservative' party.
2. Hague is affiliated with the 'Conservative' party, whereas Blair is a member of the 'Labour' party.
3. Individuals expressing stronger Eurosceptic sentiments are more likely to cast their votes in favor of the 'Conservative' party.  
   Individuals who receive lower assessments for both national and household economic conditions show a propensity to vote for the 'Conservative' party.
4. The number of voters is evenly distributed across genders, indicating an almost equal representation of male and female voters.

## 

## 1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not?( 2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get\_dummies(drop\_first=True)) Data split, ratio defined for the split, train-test split should be discussed.

We utilize the pd.get\_dummies() function to encode the 'gender' variable, ensuring that we avoid colinearity issues in our dataset by dropping the first column. Additionally, we manually encode the 'vote' variable.

Scaling variables is a matter of choice for linear models like Linear Regression, Logistic Regression, LDA, and tree-based models such as Random Forest. However, it becomes necessary to scale variables when employing distance-based models like KNN and SVM. This is to prevent variables with larger numerical values from exerting excessive influence on the model's performance.

Importantly, scaling should be performed after the train-test split. This means that the scaling parameters (such as mean or min-max values) learned from the training set should be applied to scale the test set. This ensures that the test data remains unseen during the training phase.

Since none of the features exhibit a normal distribution, we can employ the Min-Max Scaler for our particular problem. With that settled, we can now proceed to divide the dataset into training and testing subsets. We employ the train\_test\_split function, specifying a test size of 0.3 to achieve a 70:30 split, and set the random\_state to 100 to obtain reproducible results.

To ensure equal class proportions in both the training and test sets, we stratify the target variable (denoted as 'y'). Consequently, the training set comprises 1067 rows and 8 features, while the test set consists of 458 rows and 8 features.

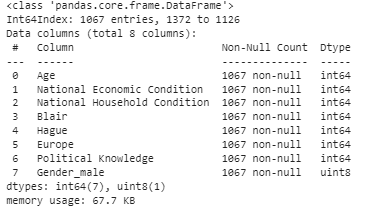


Fig5: Train data info

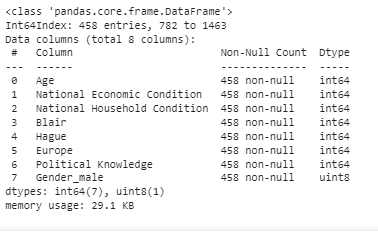
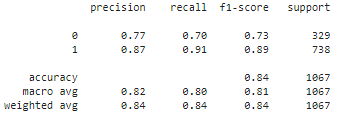


Fig6: Test data info

## 1.4 Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Logistic Regression:

The results after applying Loqistic Reqressionon the train and test sets are as follows —



Confusion Matrix, Accuracy Score:



Fig: Logistic Regression results

Logistic Regression provides us with the ability to examine the coefficient values associated with each feature, allowing us to understand the extent to which each feature contributes to the final outcome or decision. By plotting these coefficient values in a bar plot, we can gain a clearer and more intuitive understanding of their respective influences. This visualization enables us to easily compare and interpret the impact of each feature on the logistic regression model's output.

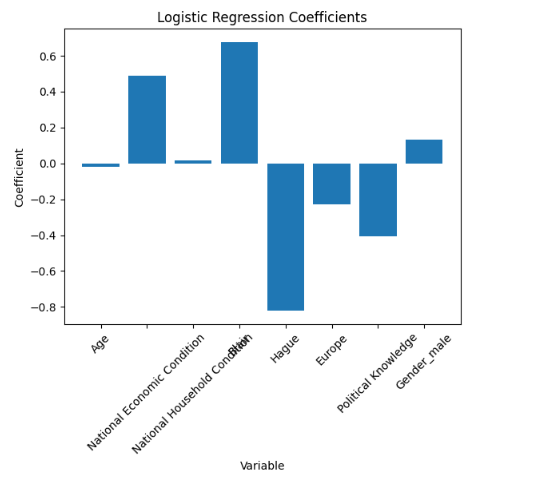


Fig: Logistic Regression Coeff

Inferences-

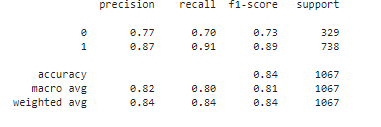
1.Independent variables with larger absolute values have a greater impact on the target variable

2. Train accuracy and test accuracy & ROC-AUC scores are very similar, hence this model does not suffer from overfitting.

3.Variables ’age’, ’economic.cond.household’ & ‘gender\_maIe’ are least important.

LDA:

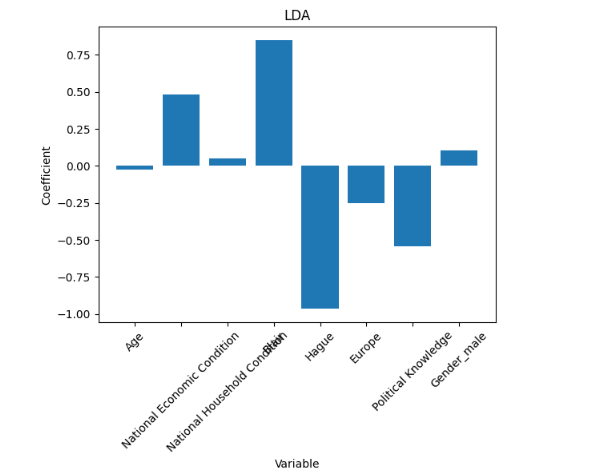
The results after applying LDA on train and test sets are as follows -



Accuracy Score, Confusion Matrix:



Fig: LDA results



Inferences:

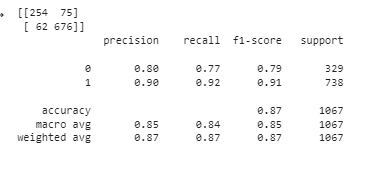
* Based on the coefficients obtained from the logistic regression model, we can infer that the predictors 'Hague' have the highest importance in influencing the target variable, followed by 'Blair', 'political.knowledge', 'economic.cond.national', and 'Europe'. The order of importance suggests that 'Hague' has the strongest impact, followed by 'Blair', and so on.
* Among the predictors, 'age', 'gender\_male', and 'economic.cond.household' are found to be the least important. These predictors have minimal influence on the target variable and do not significantly contribute to the logistic regression model's decision-making process.
* The similarity between the train accuracy, test accuracy, and ROC-AUC scores indicates that this logistic regression model does not suffer from overfitting. The model performs equally well on both the training and test datasets, suggesting that it generalizes effectively to unseen data. This consistency is a positive indication, as it implies that the model's performance is reliable and not overly influenced by the training data.

## 1.5 Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

KNN Model:

Accuracy and Confusion Matrix:

Train Data:



Test Data:

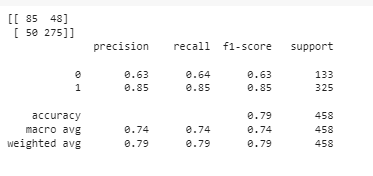


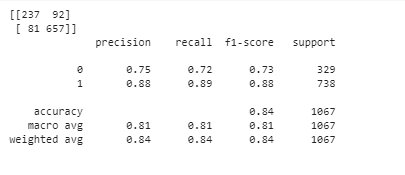
Fig: KNN results

A significant difference in ROC-AUC score of train and test set implies the KNN model suffers from overfitting.

Naive Bayes:

Accuracy and Confusion Matrix:

Train Data:



Test Data:

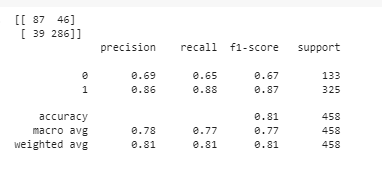
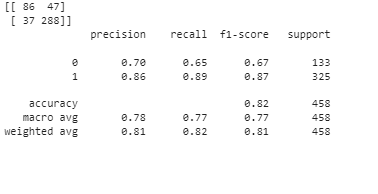


Fig: Naives Bayes Results

Similar accuracy for training and test dataset.

Hence the model is generalized.

## 1.6 Model Tuning (4 pts) , Bagging ( 1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best\_params. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.



## 1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts

**Different classification methods used to model, accuracy, recall value, AUC are as follows:**

| **Model** | **Accuracy Score** | **Confusion Matrix** |
| --- | --- | --- |
| Logistic Regression | 0.77 | [[ 230 99] [ 69 669]] |
| Linear Discriminant Analysis | 0.77 | [[ 299 100] [ 68 670]] |
| K-Nearest Neighbors | 0.80 | [[ 254 75] [ 62 676]] |
| Naive Bayes | 0.63 | [[ 85 48] [ 50 275]] |

* Accuracy on test set: All models are approximately equal.
* ROC- AUC Score: on test set: All models are approximately equal.

## 1.8 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

In terms of the confusion matrices, all models show some degree of misclassification, but the logistic regression and LDA models have the highest number of true positives and true negatives, and the lowest number of false positives and false negatives.

Overall, the logistic regression and LDA models appear to be the best optimized models for this dataset, based on their high accuracy and ROC scores, and low number of misclassifications. However, it's important to note that further tuning of hyperparameters or testing of other models may still lead to even better performance.

**Based on the analysis, the following insights and recommendations can be made to help the management solve the business objective:**

1. The logistic regression and linear discriminant analysis models have the highest accuracy and ROC score compared to the other models. This suggests that these models are the best fit for predicting the voting pattern of the respondents. The management can use these models to identify the key factors that influence the voting decision of the respondents.

2. The features that have the highest impact on the voting decision are the assessments of the Labour and Conservative leaders, as evident from the feature importance analysis conducted on the best-performing models. This suggests that the voters are heavily influenced by the perception of the party leaders. The management can use this insight to develop effective communication strategies that highlight the strengths of their party leader.

3. The age of the respondent has been converted into age groups, which is a useful feature for analysis. However, the analysis can be further improved by including additional demographic factors such as education, income, and occupation. This can provide a more comprehensive understanding of the factors that influence the voting decision of the respondents.

4. The Naive Bayes model has the lowest accuracy and ROC score among all the models. This suggests that this model may not be the best fit for predicting the voting pattern of the respondents. The management should avoid using this model for making any important business decisions.

5. The K-nearest neighbors model has an accuracy score of 0.7817, which is slightly lower than the other models. However, this model can be useful for identifying similar voters based on their demographic and attitudinal factors. The management can use this model to segment their voter base and tailor their communication strategies accordingly.

6. Finally, the business objective of the management is to identify the key factors that influence the voting decision of the respondents. The insights and recommendations outlined above can help the management achieve this objective. By understanding the factors that influence the voting decision, the management can develop effective communication strategies that resonate with their target audience and improve their chances of winning the election.

**Problem 2:**

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

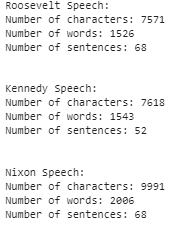
1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

## 2.1 Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)

To determine the number of characters, we can utilize the raw() function from the inaugural corpora. This function returns the entire text as a single string, and by applying the len() function to it, we can obtain the count of characters.

To count the number of words, we can employ the words() function from the inaugural corpora. However, it is important to note that this function considers punctuation symbols as words as well. To exclude these punctuation symbols from the word count, we need to apply a condition or filter to exclude them.

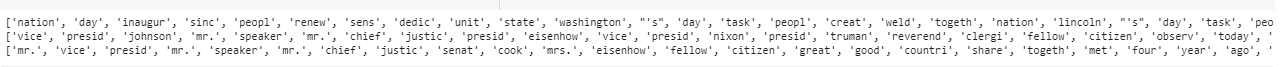
To count the number of sentences, we can use the sents() function from the inaugural corpora. This function splits the text into sentences and returns them as a list. By applying the len() function to this list, we can obtain the count of sentences.



## 2.2 Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.

Stopwords refer to commonly occurring words that do not contribute significantly to the analysis or understanding of the text. It is recommended to remove these stopwords during the preprocessing step. The NLTK (Natural Language Toolkit) package provides a built-in list of 179 stopwords that can be utilized for this purpose.

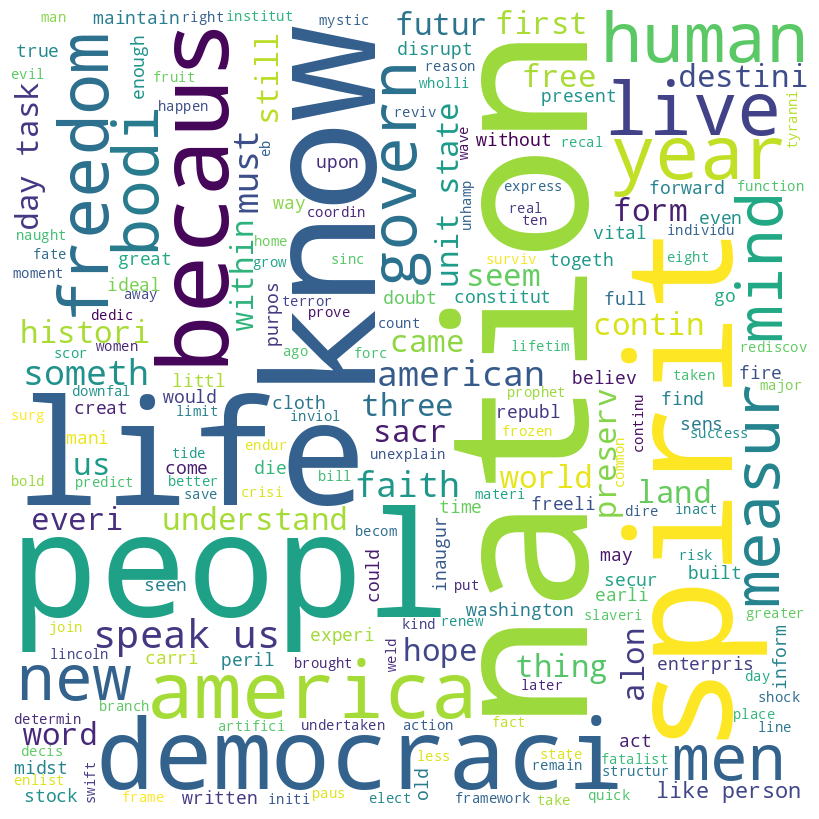
In addition to removing stopwords, it is also advised to perform other preprocessing steps. These steps may include converting all words to lowercase to ensure case insensitivity, removing numerical characters from the text, and eliminating punctuation marks. These actions help to clean the text and ensure that the subsequent analysis focuses on the most meaningful and relevant information present in the documents.



## 2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)



## 2.4 Plot the word cloud of each of the three speeches. (after removing the stopwords)



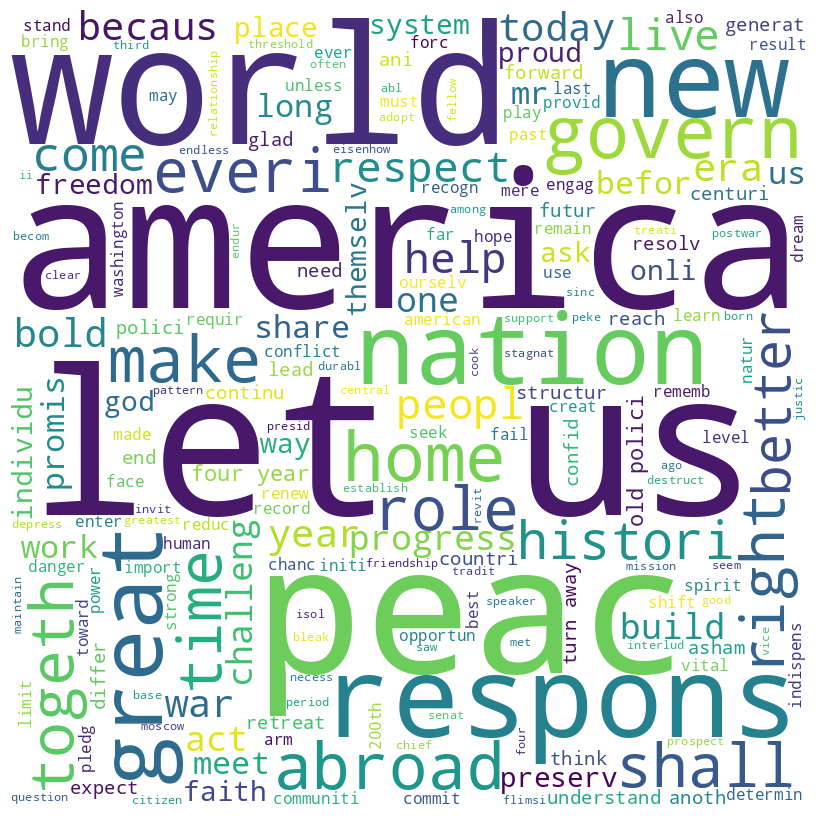
****

Fig: Word Cloud