Murali b

machine learning

**CONTENTS:**

**Details: Page Number:**

**Problem 1**

**1.1 Head of the data 4  
1.2 Tail of the data 4  
1.3 Info of the data 5  
1.4 Shape of the data 5  
1.5 Summary of the data 5  
1.6 Columns names details 6  
1.7 Checking Null values 6  
1.8 Duplicates values 6  
1.9 Renaming the columns names 7  
1.10 Skewness of the data 7  
1.11 Correlation of the data 21  
1.12 Encoding the data 25  
1.13 LDA Accuracy of test data 26  
1.14 LDA Accuracy of train data 26  
1.15 Confusion matrix and classification report 26  
1.16 Confusion matrix and classification report 26  
1.17 Logistic Regression Train data accuracy 28  
1.18 Test data accuracy 28  
1.19 Confusion matrix and classification report 28  
1.20 Confusion matrix and classification report 28  
1.21 Naive Bayes test data accuracy 29  
1.22 Train data accuracy 29  
1.23 Confusion matrix and classification report 29  
1.24 Confusion matrix and classification report 30  
1.25 KNN model test data accuracy 30  
1.26 Train data accuracy 30  
1.27 Confusion matrix and classification report 30  
1.28 Confusion matrix and classification report 31  
1.32 Probability of test data 32  
1.33 Accuracy of test data 33  
1.34 Accuracy of train data 33  
1.35 Confusion matrix and classification report 33  
1.36 Confusion matrix and classification report 33  
1.38 LDA Accuracy value test data 36  
1.39 Accuracy value of train data 36  
1.40 Confusion matrix and classification report 36  
1.41 Confusion matrix and classification report 36  
1.42 KNN Model test data accuracy 39  
1.43 Train data accuracy 39  
1.44 Confusion matrix and classification report 38  
1.45 Confusion matrix and classification report 38  
1.48 Naïve Bayes Accuracy score test data 42  
1.49 Accuracy score of train data 42  
1.50 Confusion matrix and classification report 42  
1.51 Confusion matrix and classification report 42**

**Problem 2:**

**1.66 No of words – Roosevelt 57  
1.67 No of words – Kennedy 57  
1.68 No of words – Nixon 57  
1.69 No of chars – Roosevelt 57  
1.70 No of chars – Kennedy 57  
1.71 No of chars – Nixon 57  
1.72 No of sentences – Roosevelt 58  
1.73 No of sentences – Kennedy 58  
1.74 No of sentences – Nixon 58  
1.75 Removal of stopwords – Roosevelt 58  
1.76 Removal of stopwords – Kennedy 59  
1.77 Removal of stopwords – Nixon 59  
1.78 Most occurrences words – Roosevelt 60  
1.79 Most occurrences words – Kennedy 60  
1.80 Most occurrences words – Nixon 61**

**FIGURES:**

**Problem 1:**

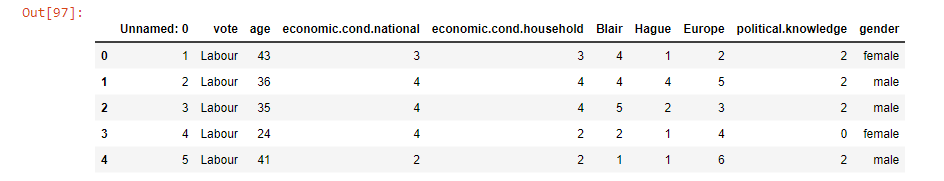
**Fig 1.1 Histograms – Univariate Analysis 8-9  
Fig 1.2 Bivariate Analysis (till Fig 1.6) 9-21  
Fig 1.11 Heatmap 22  
Fig 1.12 Multivariate Analysis 23  
Fig 1.13 Checking for outliers 24  
Fig 1.14 Removing Outliers 24  
Fig 1.15 Plot of confusion matrix train data 33  
Fig 1.16 Test data plot of confusion matrix 33  
Fig 1.17 AUC and ROC curve of train data 34  
Fig 1.18 AUC and ROC curve of test data 34  
Fig 1.19 Plotting confusion matrix test data 37  
Fig 1.20 Plotting confusion matrix test data 37  
Fig 1.21 AUC value and ROC value train data 38  
Fig 1.22 AUC value and ROC value test data 38  
Fig 1.23 Plotting confusion matrix of test data 39  
Fig 1.24 Plotting of train data 39  
Fig 1.25 AUC value and ROC curve 40  
Fig 1.26 AUC value and RO Curve 40  
Fig 1.27 K neighbors 41  
Fig 1.28 Plot of train data 41   
Fig 1.29 Plotting confusion matrix of test data 43  
Fig 1.30 AUC value and ROC curve of train data 44  
Fig 1.31 AUC value and ROC curve of test data 44  
Fig 1.32 AUC value and ROC curve of train data 46  
Fig 1.33 AUC value and ROC curve of test data 47  
Fig 1.34 AUC value and ROC curve of train data 48  
Fig 1.35 AUC value and ROC curve of test data 49  
Fig 1.36 AUC value and ROC curve of train data 51  
Fig 1.37 AUC value and ROC curve of test data 51  
Fig 1.38 AUC value and ROC curve of train data 53  
Fig 1.39 AUC value and ROC curve of test data 53  
Fig 1.40 AUC value and ROC curve of train data 55  
Fig 1.41 AUC value and ROC curve of test data 55**

**Problem 1:**

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.

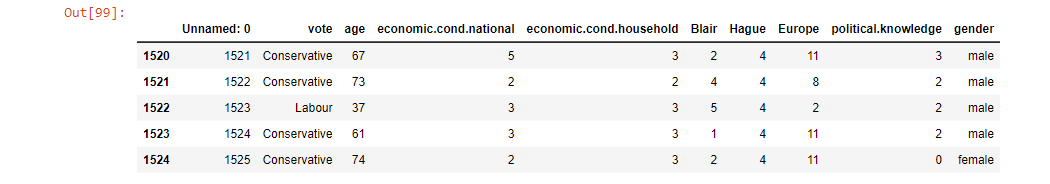
**Data Ingestion:**1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

**Displaying the top 5 data:**



Output 1.1 The above screenshot represents the top 5 head data of the dataset.

**Displaying the bottom 5 data:**



Output 1.2 The picture represents the last 5 data of the dataset.

**Info:**

Text

Description automatically generated

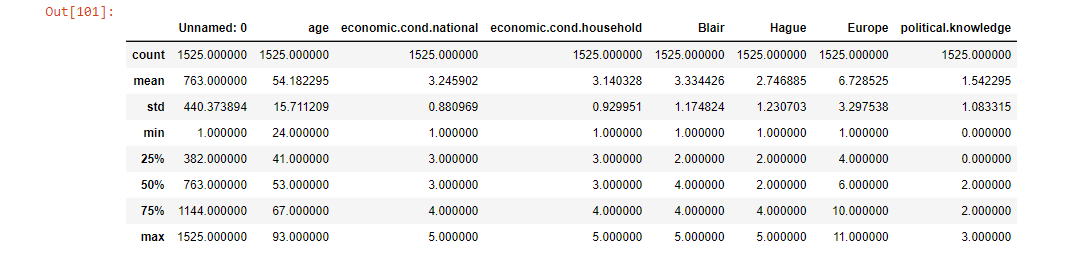
Output 1.3 The information of the dataset includes 8 integer datasets and 2 object datasets.

**Shape:**



Output 1.4 We can see the shape of the data as 1525 entries with 9 observations.

**Summary:**



Output 1.5 It displays the 5 point summary analysis such as count , mean , standard deviation , minimum and maximum value , 25th percentile , 50th percentile and 75th percentile.

**Columns:**

Text

Description automatically generated

Output 1.6 It displays the name of the column’s details.

**Checking the Null value:**

Text

Description automatically generated

Output 1.7 We can see that there are no Null values present in the dataset.

**Checking for Duplicates:**



Output 1.8 We had 8 duplicate values present so after dropping the duplicate values we can see now there is no duplicate values present in the dataset.



**Renaming the columns names and dropping columns name of Unnamed : 0**

We had renamed column names of : economic.cond.national , economic.cond.household , political.knowledge into a proper name.

Graphical user interface, text

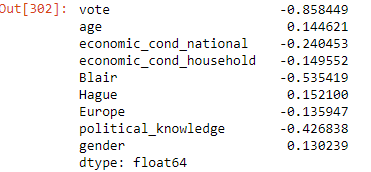
Description automatically generated

Output 1.9 So, after renaming the columns names, it displays the columns names as mentioned in the screenshot.

Chart

Description automatically generated with medium confidence

**Skewness:**



Output 1.10 We can see that from the above screenshot attached, we can see only age and Hague has positive skewness whereas all other attributes have negative skewness.

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, Correlation 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.

**Univariate Analysis: (Histograms)**

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

Fig 1.1 We can see that there is no relationship in vote attributes , and can see different range of values in age attributes , and the values keeps increasing and decreasing in economic\_cond\_national and economic\_cond\_household attributes , Blair and Hague are almost similar in range of values. Europe has different set of values present , and again same no existence of relationship or values present in gender and political\_knowledge attributes.

Bi-variate Analysis: (Countplot)

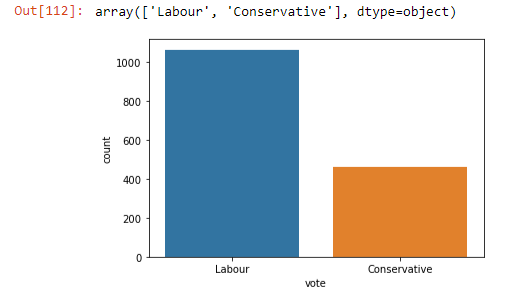


Fig 1.2 We can see more number of vote in Labour than Conservative.

Chart, bar chart, histogram

Description automatically generated

Fig 1.3 We can see random discriminant representation of values present in the Age attributes.

Chart, bar chart

Description automatically generated

Fig 1.4 From the above picture, we can see highest values ranging as 3 and lowest value as 1 in economic\_cond\_national attributes.

Chart, bar chart

Description automatically generated

Fig 1.5 Same like economic\_cond\_national attributes, we can also see here that highest values ranging as 3 and lowest value as 1 in economic\_cond\_household attributes.

Chart, bar chart

Description automatically generated

Fig 1.6 As per the attached screenshot, we can see highest number of Blair as 4 and minimum values as 1.

Chart, bar chart

Description automatically generated

Fig 1.7 We can see almost to equal distribution of Hague between values ranging between 2 and 4 but 2 shows as maximum values.

Chart, bar chart

Description automatically generated

Fig 1.8 The highest maximum value is 11 and minimum value is2 present in Europe dataset.

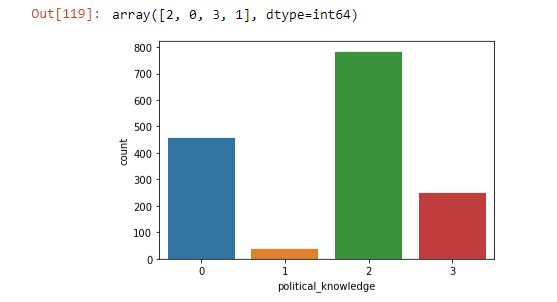


Fig 1.9 We can see that a greater number of people have political knowledge as mentioned in the above screenshot.

Chart, bar chart

Description automatically generated

Fig 1.10 Here, we can see that there are a greater number of female people than male people.

**Bi-variate Analysis:**

Chart, histogram

Description automatically generated

Fig 1.11 We can see no outliers present in age attribute but there is a steep fall and rise curve shaped representation of data.

Chart, histogram

Description automatically generated

Fig 1.12 We can see outliers present in economic\_cond\_national attributes and the representation of data shows as randomly segregated data among each other.

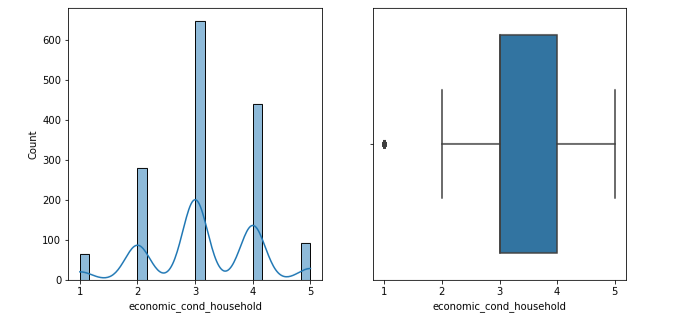


Fig 1.13 Outliers are present again in economic\_cond\_household attribute but the pattern of data representation shows as similar fashion like economic\_cond\_national

Chart, histogram

Description automatically generated

Fig 1.14 We can see no outliers present in Blair attribute and the values shows as increase and decrease pattern following a non-linear way.

Chart, histogram

Description automatically generated

Fig 1.14 No outliers present again but the data representation seems to be in similar way like Blair attribute

Chart, histogram

Description automatically generated

Fig 1.15 As we can see no outliers present but the value representation shows as linear way of data organized.

Chart, histogram

Description automatically generated

Fig 1.16 Political\_knowledge attribute doesn’t have any outliers and no proper data exists between each other in Political\_knowledge attribute.

**Scatterplot:**

Graphical user interface, application

Description automatically generated

Fig 1.17 From the above plot, we can see that there is an equal representation of data between Vote and Age.

Chart, scatter chart

Description automatically generated

Fig 1.18 The data seems to very scattered and there is no relationship between economic\_cond\_national and economic\_cond\_household

Chart, scatter chart

Description automatically generated

Fig 1.19 Blair and Hague do not have a proper relationship exist between them.

Chart, scatter chart

Description automatically generated

Fig 1.20 We can see points are linearly scattered between Europe and political\_knowledge attributes.

**Boxplot:**

Chart, box and whisker chart

Description automatically generated

Fig 1.21 From the above boxplot, we can see that Age has more values than the Vote and no outliers present. Highest value of voter with age is 70 and minimum is 40.

Chart, box and whisker chart

Description automatically generated

Fig 1.22 We can see no outliers present in 2 of the boxplot representation whereas outliers exist in remaining 3 boxplots. Highest values is 4.5 and lowest values is 1.0

Chart, box and whisker chart

Description automatically generated

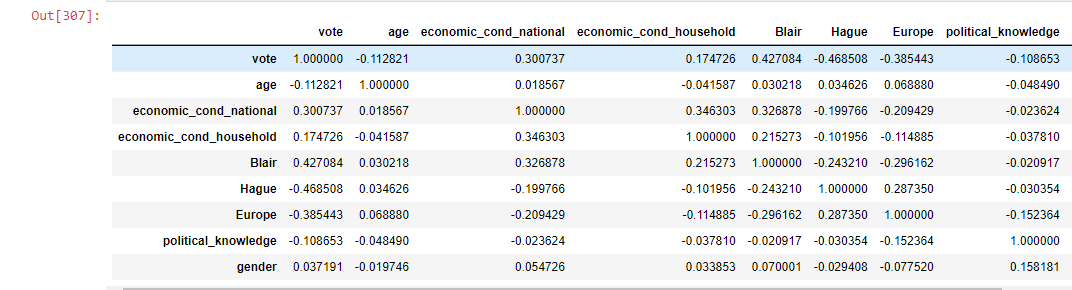
Fig 1.23 We can see equal amount of data representation and highest value is 4.0 and lowest seems to be 1.0

Chart, bar chart

Description automatically generated

Fig 1.24 We can see no outliers present in the above data representation and the data is organized in a non-linear pattern.

**Correlation:**



**Output 1.11 The above screenshot displays the correlation values among ach attributes present in the dataset.**

**Heatmap:**

Chart

Description automatically generated

**Fig 1.11 We can see highest values of positive correlation relationship found between Vote and Blair as per the heatmap whereas other attributes have lesser relationship and even cases found in negative correlation.**

**Multi-variate Analysis:**

Chart

Description automatically generated

Fig 1.12 The pairplot shows the full descriptive analysis of each and all attributes present in the dataset. It depicts the multivariate analysis between each attribute.

**Checking for Outliers:**

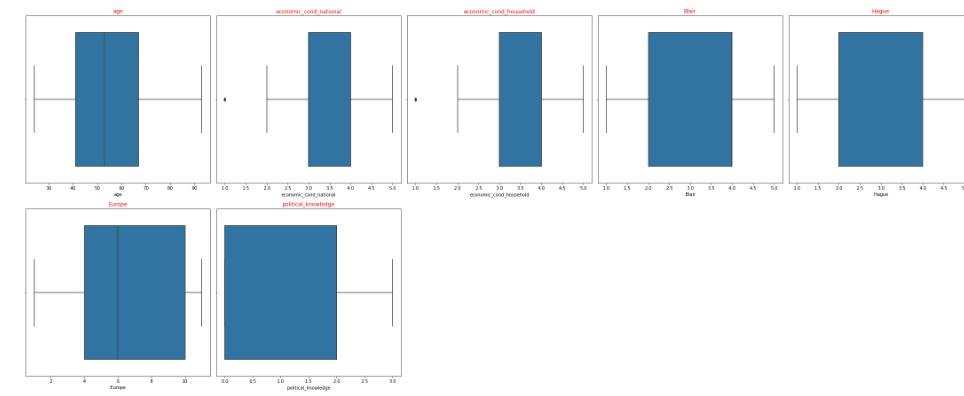


Fig 1.13 We can see outliers present only in economic\_cond\_household and economic\_cond\_national but not in other 5 attributes.

**After removing the Outliers:**

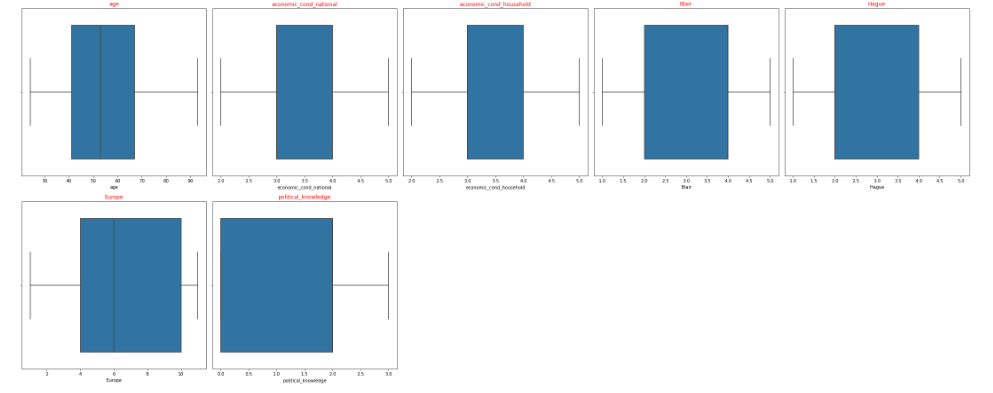


Fig 1.14 Now, we can see that there are no outliers present in the dataset now.

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.

Graphical user interface, text

Description automatically generated

Output 1.12 We have converted the object datatype into Integer type to fit in the models.

Text

Description automatically generated

Output 1.13 We have already encoded the data having string values for modelling by splitting into 70:30 as train and test data and displayed the rows and columns numbers of training and test data of both independent variables and dependent variables.

Scaling is not required for this modelling as we can see that almost all the variables are categorical independent variables and we have taken vote attribute as dependent variable for the analysis.

**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 behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)**

**LDA:**

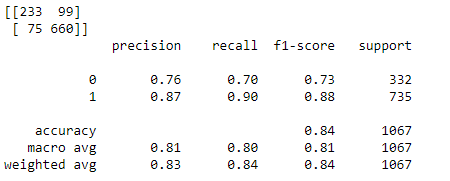
**Accuracy of Train data:**

**Output 1.13**

**Accuracy of Test data:**

**Output 1.14**

**Confusion matrix and classification report of Train data:**

**Output 1.15**

**Confusion matrix and classification report of Test data:**

Calendar

Description automatically generated**Output 1.16**

**The model gives accuracy of 84% on train data and 82% on test data.**

**Logistic Regression:**

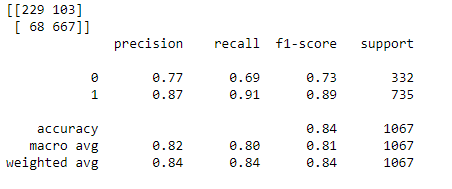
**Train data Accuracy:**

**Output 1.17**

**Test data Accuracy:**

**Output 1.18**

**Confusion matrix and classification report of Train data:**

**Output 1.19**

**Confusion matrix and classification report of Test data:**

Table

Description automatically generated**Output 1.20**

**The model gives accuracy of 84% on train data and 82% on test data. Here we can see that both the models give the same exact values when model is defined and fitted on both the test and test data so it’s not over fitted nor under fitted.**

**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 behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)**

**Naive Bayes:**

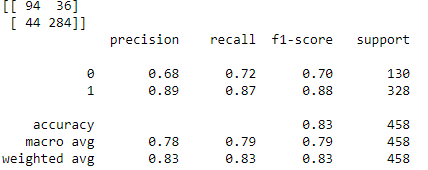
**Accuracy on Test data:**

**Output 1.21**

**Accuracy on Train data:**

**Output 1.22**

**Confusion matrix and classification report of Test data:**

**Output 1.23**

**Confusion matrix and classification report of train data:**

Table

Description automatically generated**Output 1.24**

**We can see accuracy as 82.5% on test data and 83.31% on train data.**

**KNN Model :**

**Accuracy on Test data:**

**Output 1.25**

**Accuracy on Train data:**

**Output 1.26**

**Confusion matrix and classification report of Test data:**

Table

Description automatically generated**Output 1.27**

**Confusion matrix and classification report of train data:**

Table

Description automatically generated**Output 1.28**

**We can see accuracy as 79% on test data and 86% on train data. So the accuracy is less in KNN model test data so we can choose Naïve Bayes model.**

**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. Define a logic behind choosing particular values for different hyper-parameters for grid search. 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)**

**Ans: 1.6 and 1.7 is answered in a combined format by covering all the information asked.**

**Logistic Regression:**

Text

Description automatically generated with medium confidence**Output 1.29**

**GridsearchCV output of Logistic Regression.**

Text

Description automatically generated**Output 1.30**

**The above screenshot displays the confusion matrix of train data.**

**Output 1.31**

**Probabilities of Test data:**

Graphical user interface, text, application

Description automatically generated**Output 1.32**

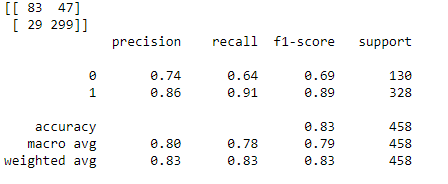
**Accuracy of test data:**

**Output 1.33**

**Accuracy of train data:**

**Output 1.34**

**Confusion matrix and classification report of test data:**

**Output 1.35**

**Confusion matrix and classification report of train data:**

Table

Description automatically generated**Output 1.36**

**Plotting of train data confusion matrix:**

Chart, treemap chart

Description automatically generated**Fig 1.15**

**Plotting of test data confusion matrix:**

Chart

Description automatically generated**Fig 1.16**

**AUC value and ROC curve of train data:**

Chart, line chart

Description automatically generated**Fig 1.17**

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generated**Fig 1.18**

**LDA:**

**Output 1.37**

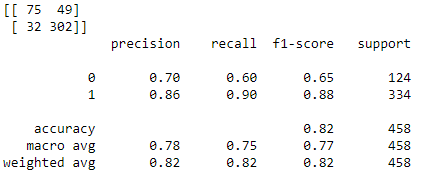
**Accuracy value of test data:**

**Output 1.38**

**Accuracy value of train data:**

**Output 1.39**

**Confusion matrix and classification report of test data:**

**Output 1.40**

**Confusion matrix and classification report of train data:**

Table

Description automatically generated**Output 1.41**

**Plotting confusion matrix of Train data:**

Chart

Description automatically generated**Fig 1.19**

**Plotting confusion matrix of test data:**

Chart

Description automatically generatedFig 1.20

**AUC value and ROC curve of train data:**

Chart, line chart

Description automatically generatedFig 1.21

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generated**Fig 1.22**

**KNN Model:**

**Test data:**

**Output 1.42**

**Train data:**

**Output 1.43**

**Confusion matrix and classification report of test data:**

Table

Description automatically generated**Output 1.44**

**Confusion matrix and classification report of train data:**

Table

Description automatically generated**Output 1.45**

**Plotting confusion matrix of train data:**

Chart, treemap chart

Description automatically generated**Fig 1.23**

**Plotting confusion matrix of test data:**

Chart

Description automatically generatedFig 1.24

**AUC value and ROC curve of train data:**

Chart, line chart

Description automatically generatedFig 1.25

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generated**Fig 1.26**

A picture containing text

Description automatically generated

**Output 1.46 As per the above-mentioned screenshot, we can see accuracy score for k=3, 5 and 9.**

**MCE values:**

Text

Description automatically generated**Output 1.47**

Chart, line chart

Description automatically generated**Fig 1.27**

**We can see the above plot when represented number of neighbors K against the values of misclassification error.**

**Naive Bayes:**

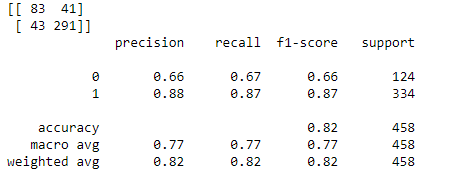
**Accuracy score of test data:**

**Output 1.48**

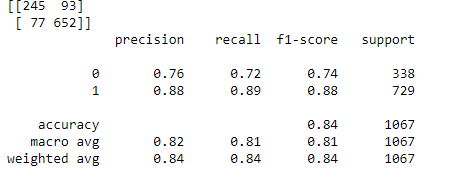
**Accuracy score of train data:**

**Output 1.49**

**Confusion matrix and classification report of test data:**

**Output 1.50**

**Confusion matrix and classification report of train data:**

**Output 1.51**

**Plotting confusion matrix of train data:**

Chart

Description automatically generated**Fig 1.28**

**Plotting confusion matrix of test data:**

Chart

Description automatically generated**Fig 1.29**

**AUC value and ROC curve of train data:**

Chart, line chart

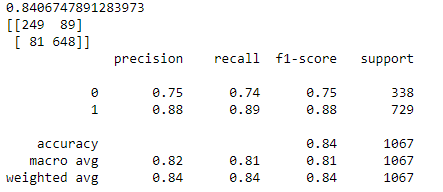
Description automatically generated**Fig 1.30**

**AUC value and ROC curve of test data:**

Chart, line chart

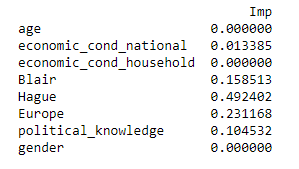
Description automatically generated**Fig 1.31**

**Navies Bayes model accuracy and confusion matrix and classification report on train data with SMOTE:**

**Output 1.52**

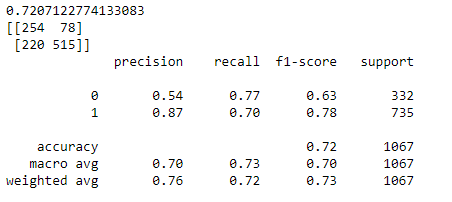
**Decision Tree:**

**Important Features:**

**Output 1.53**

**We can see Hague is the most important feature present in the dataframe followed by Europe.**

**Accuracy , confusion\_matrix and classification\_report of train data:**

**Output 1.54**

**Accuracy , confusion\_matrix and classification\_report of test data:**

Table

Description automatically generated**Output 1.55**

**AUC value and ROC curve of train data:**

Chart, line chart

Description automatically generated**Fig 1.32**

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generated**Fig 1.33**

**Bagging:**

**Accuracy , confusion\_matrix and classification\_report of train data:**

Table

Description automatically generated**Output 1.56**

**Accuracy , confusion\_matrix and classification\_report of test data:**

Table

Description automatically generatedOutput 1.57

**AUC value and ROC curve of train data:**Chart, line chart

Description automatically generated**Fig 1.34**

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generatedFig 1.35

**ADA Boosting:**

Text

Description automatically generated**Output 1.58**

**Accuracy , confusion\_matrix and classification\_report of train data:**

Table

Description automatically generated**Output 1.59**

**Accuracy , confusion\_matrix and classification\_report of test data:**

Table

Description automatically generatedOutput 1.60

**AUC value and ROC curve of train data:**

Chart, line chart

Description automatically generatedFig 1.36

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generated**Fig 1.37**

**Gradient Boosting:**

**Accuracy , confusion\_matrix and classification\_report of train data:**

Table

Description automatically generated**Output 1.61**

**Accuracy , confusion\_matrix and classification\_report of test data:**

Table

Description automatically generatedOutput 1.62

**AUC value and ROC curve of train data:**

Chart, line chart

Description automatically generatedFig 1.38

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generatedFig 1.39

**Random Forest:**

**Accuracy , confusion\_matrix and classification\_report of train data:**

Table

Description automatically generated**Output 1.63**

**Accuracy , confusion\_matrix and classification\_report of test data:**

Table

Description automatically generated**Output 1.64**

**AUC value and ROC curve of train data:**Chart, line chart

Description automatically generated**Fig 1.40**

**AUC value and ROC curve of test data:**

Chart, line chart

Description automatically generated**Fig 1.41**

**So, from all the above-mentioned models, we can see that performance is better when the models are tuned when compared to the original model. In any of the tuned models we can see no over fitting nor under fitting. But Gradient Boosting model is the best and preferred model as it has best accuracy score in terms of train and test data and even AUC score is high in both test and train data along with highest score of precision and recall values when compared to other models.**

**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.**

**We can assume that the greatest number of people have rated as 3 and 4 for household economic condition and national economic condition and the labor party has majority number of votes when compared to conservative party. Blair has a even better score when compared to Hague in terms of votes. Minimum contribution of the total population does not have any idea about political knowledge. The tuned models performed better than the original models on both the test and train data.**

**Business Recommendations:**

**As earlier mentioned, Gradient boosting model is the best and preferred model so without scaling then the model will perform optimized outcome predictions. We can still improve the models by improving the outcome and by collecting more insights about the data model. If we still tune the attributes even more then we can still get very good high outcomes.**

Problem 2:

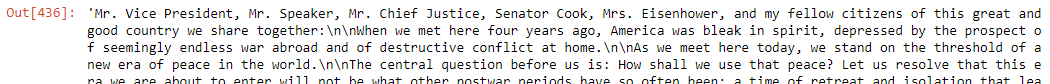
In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

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

(Hint: use .words(), .raw(), .sent() for extracting counts)

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



**Output 1.65**

**Number of words in 1941-Roosevelt:**

**Output 1.66**

**Number of words in 1961-Kennedy:**

**Output 1.67**

**Number of words in 1973-Nixon:**

**Output 1.68**

**Number of characters in 1941-Roosevelt:**

**Output 1.69**

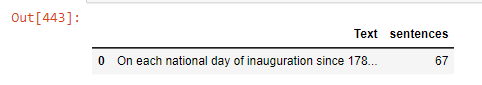
**Number of characters in 1961-Kennedy:**

**Output 1.70**

**Number of characters in 1973-Nixon:**

**Output 1.71**

**Number of sentences in 1941-Roosevelt:**

**Output 1.72**

**Number of sentences in 1961-Kennedy:**

Graphical user interface, application

Description automatically generated**Output 1.73**

**Number of sentences in 1973-Nixon:**

A picture containing graphical user interface

Description automatically generated**Output 1.74**

**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.**

Text

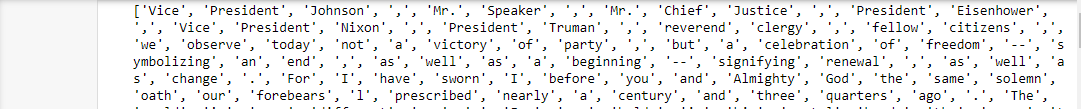
Description automatically generated

**After removing stopwords of Roosevelt:**

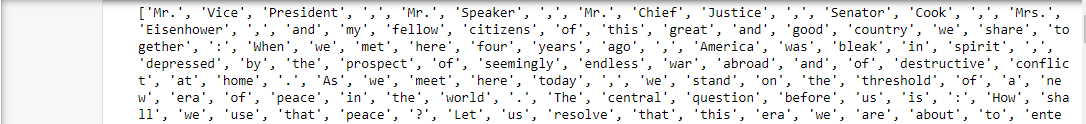
Chart, scatter chart

Description automatically generated**Output 1.75**

**After removing stopwords of Kennedy:**

**Output 1.76**

**After removing stopwords of Nixon:**

**Output 1.77**

**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)**

Graphical user interface

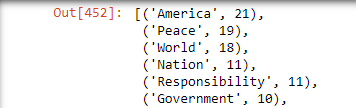
Description automatically generated with medium confidence

**Output 1.78 As per the above screenshot, we can see top three words in Roosevelt as Nation which occurs 12 times, Spirit occurs 9 times, Life occurs 9 times.**

Text

Description automatically generated

**Output 1.79 As per the above screenshot, we can see top three words in Kennedy as World occurs 8 times , Sides occurs 8 times and Pledge occurs 7 times.**



**Output 1.80 As mentioned in the screenshot, we can see top three words in Nixon as America occurs 21 times, Peace occurs 19 times and World occurs 18 times.**