MACHINE LEARNING

|  |  |  |
| --- | --- | --- |
| S.no | Content | Pg.no |
| 1 | Problem 1: Election Data | 5 |
| 1.1 | Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. | 5 |
| 1.2 | Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. | 7 |
| 1.3 | Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30) | 13 |
| 1.4 | Apply Logistic Regression and LDA (linear discriminant analysis). | 15 |
| 1.5 | Apply KNN Model and Naïve Bayes Model. Interpret the results. | 25 |
| 1.6 | Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. | 31 |
| 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. Final Model: Compare the models and write inference which model is best/optimized. | 33 |
| 1.8 | Based on these predictions, what are the insights? | 34 |
| 2 | Problem 2 :  Inaugural corpora | 35 |
| 2.1 | Find the number of characters, words, and sentences for the mentioned documents. | 35 |
| 2.2 | Remove all the stopwords from all three speeches. | 36 |
| 2.3 | Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. | 38 |
| 2.4 | Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) | 38 |

Table of contents

|  |  |  |
| --- | --- | --- |
| S.no | Contents | Pg.no |
|  | Head of the Election data | 5 |
|  | Basic information of Election data | 6 |
|  | Statistical information of Election data | 6 |
|  | Null values in Election data | 6 |
|  | Head of the Election data after encoding | 11 |
|  | Head of the Election data after scaling | 14 |
|  | Statistical information after scaling | 14 |
|  | Statistical information after converting categorical data | 15 |
|  | Classification report for training data without smote | 17 |
|  | Classification report for testing data without smote | 19 |
|  | Classification report for training data with smote | 19 |
|  | Classification report for testing data with smote | 20 |
|  | Coefficients of LDA model before smote | 21 |
|  | Classification report for testing data without smote | 22 |
|  | Confusion Matrix for training after SMOTE | 22 |
|  | Confusion Matrix for testing after SMOTE | 24 |
|  | Confusion Matrix for testing after SMOTE | 26 |
|  | Classification report with smote | 27 |
|  | Classification report with smote – Navie Bayes | 27 |
|  | Confusion Matrix after smote – Navie bayes | 28 |
|  | Confusion Matrix after smote – KNN | 29 |
|  | Classification Report after Smote - Random forest | 30 |
|  | Confusion matrix for bagging | 31 |
|  | Classification Report bagging | 32 |
|  | Confusion matrix for boosting | 32 |
|  | Classification Report Boosting | 33 |
|  | Comparison table between models | 33 |

List of Tables

List of Figures

|  |  |  |
| --- | --- | --- |
| S.no | Content | Pg.no |
|  | Voters | 7 |
|  | Age | 7 |
|  | Gender | 8 |
|  | Economic.cond.national | 8 |
|  | Economic.cond.Household | 9 |
|  | Blair | 9 |
|  | Hague | 9 |
|  | Europe | 10 |
|  | Political Knowledge | 10 |
|  | Correlation plot | 11 |
|  | Bivariate Analysis | 16 |
|  | Vote variable on train data | 16 |
|  | Vote after SMOTE | 18 |
|  | Confusion Matrix for training data without SMOTE | 18 |
|  | Confusion Matrix for testing data with SMOTE | 19 |
|  | ROC after smote | 21 |
|  | Feature Importance before smote | 20 |
|  | ROC before smote | 22 |
|  | Feature Importance after smote | 22 |
|  | ROC after smote | 24 |
|  | Score vs K value | 25 |
|  | Score vs K values – zoomed | 26 |
|  | ROC-KNN | 27 |
|  | ROC – Navie Bayes | 29 |
|  | ROC – Random forest | 31 |
|  | Word cloud | 38 |
|  |  |  |

**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.**

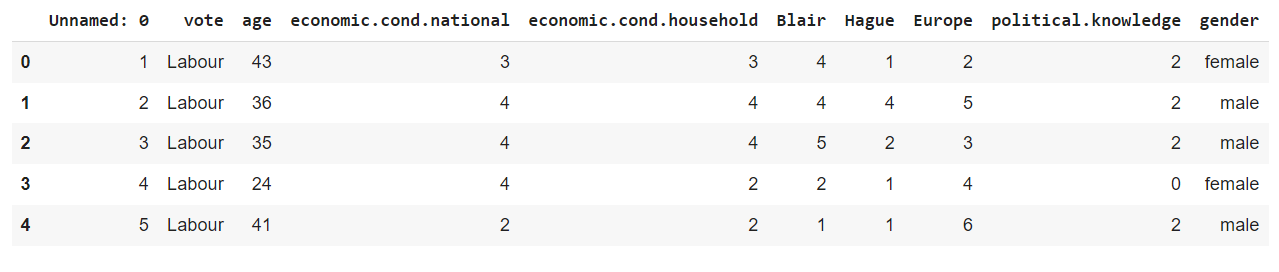
|  |
| --- |
| Data Dictionary |
|  |
| 1. vote: Party choice: Conservative or Labour |
| 2. age: in years |
| 3. economic.cond.national: Assessment of current national economic conditions, 1 to 5. |
| 4. economic.cond.household: Assessment of current household economic conditions, 1 to 5. |
| 5. Blair: Assessment of the Labour leader, 1 to 5. |
| 6. Hague: Assessment of the Conservative leader, 1 to 5. |
| 7. Europe: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment. |
| 8. political.knowledge: Knowledge of parties' positions on European integration, 0 to 3. |
| 9. gender: female or male. |

1. Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

To check the data, we must import all the libraries and load the dataset.

First, see the first five rows of the data. This helps to know about the data and columns.

Table 1: Head of the Election data



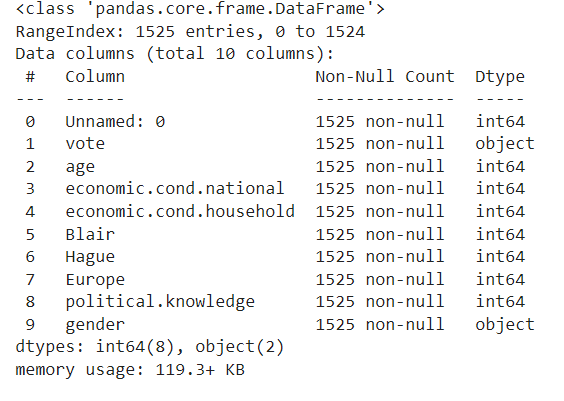
Next, check the basic information of the dataset, this can be done by using info() function.

From the information of the Election data, there are 1525 rows and 10 columns.

There are two object type which are to be treated.

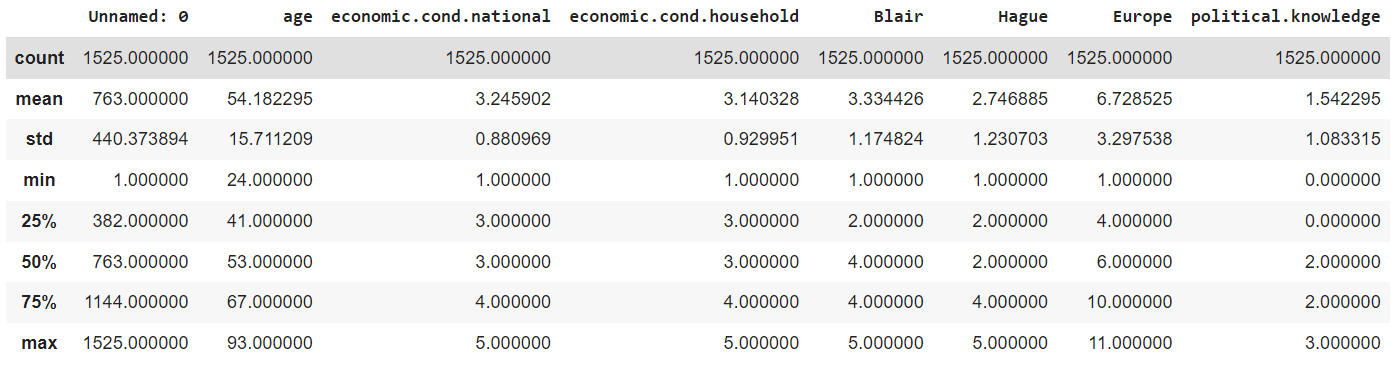
No null values can be seen here.

Table 2: Basic information of Election data



Check the statistical description of the data

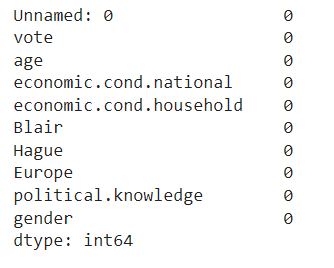
Table 3: statistical information of Election data



We have categorical data but this is considered as numeric data type.

These should be changed to categorical data type.

Table 4: Null values after the treatment



There are no duplicates in the data.

The first column unnamed:0, is actually an index for the data.

This should be changed.

1. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

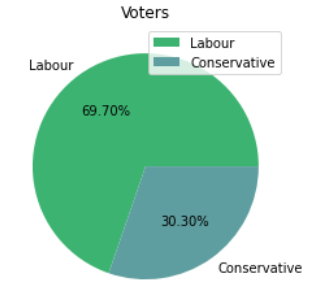
Univariate Analysis:

Voting’s:

There are two types of parties: Labour, Conservative.

70% of the people are voting for Labour and 30% are voting for Conservative party.

Figure 1: voters



Age:

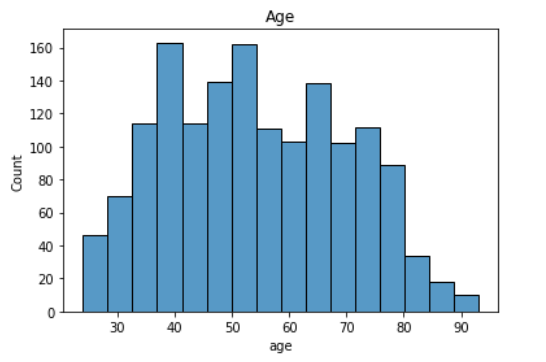
The minimum age for voting in Europe is 16.

In the data the minimum age found is 24 and maximum is 93.

Most of the votes were received by the age group 35-40, 50-55.

Low voting rate is from the age above 80.

Figure 2: Age

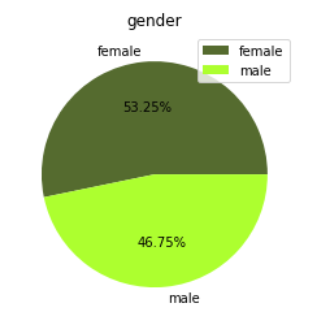


Gender:

There are more females than men who cast their votes.

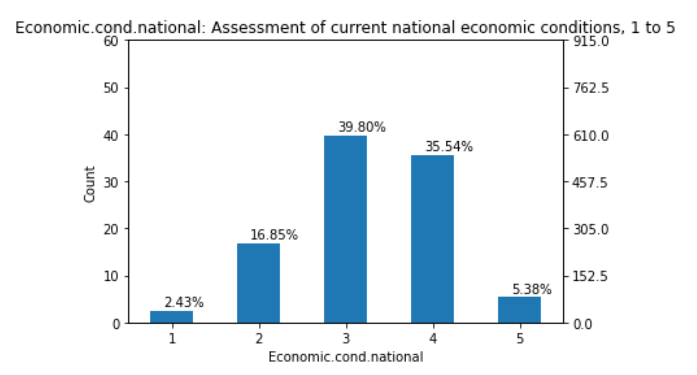
53% of the females cast their vote and 48% of the male castheir votes.

Figure 3: Gender



Economic. cond. national: Assessment of current national economic conditions:

Figure 4: Economic cond national



The economic national condition of the people is defined on a scale of 1 to 5.

The 40% people are on the scale 3. And 35% on scale of 4.

The lowest scale 1 are not casting their vote

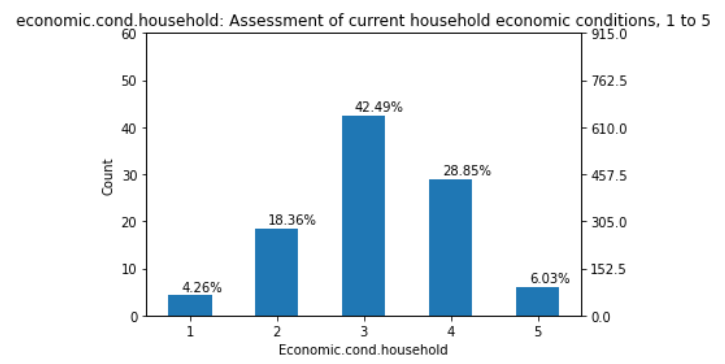
Economic. cond. Household: Assessment of current household economic conditions:

The economic hosehold condition of the people is defined on a scale of 1 to 5.

The 42% people are on the scale 3. And 29% on scale of 4.

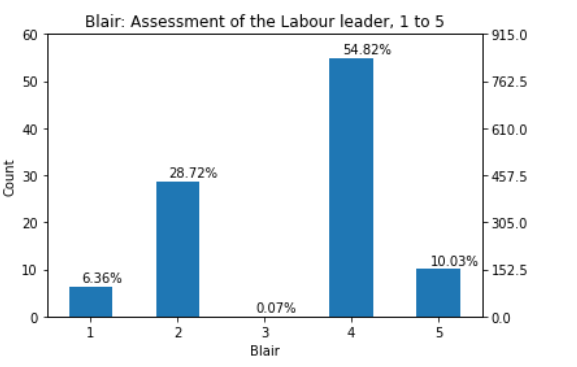
The people on scale 3,4 are more likely to caste vote.

Figure 5: Economic cond Household



Blair: Assessment of the Labour leader, 1 to 5

Figure 6: Blair



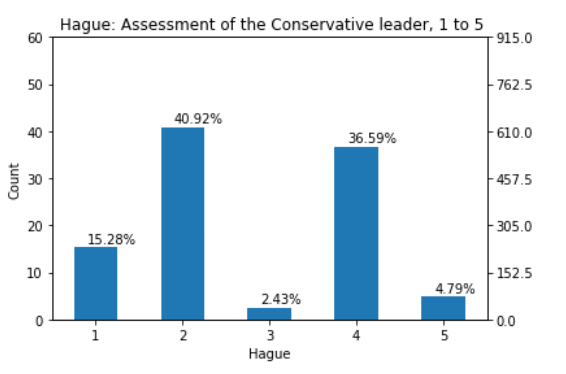
The people will be scaled by their knowledge for labour leader.

The people on a scale three are not casting their votes.

People on a scale 4 are casting their votes.

Hague: Assessment of the Conservative leader, 1 to 5

Figure 7: Hague



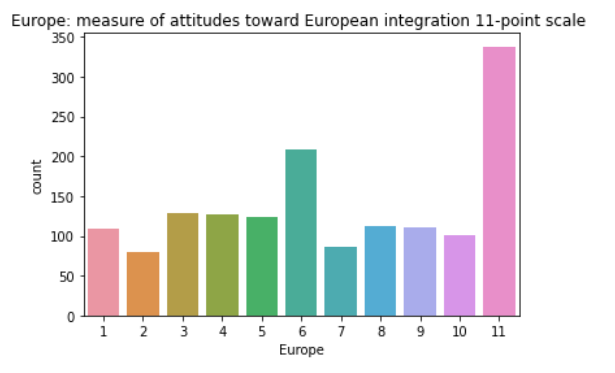
The people will be scaled by their knowledge for labour leader.

The people on a scale three are not casting their votes.

People on a scale 2 and 4 are casting their votes.

Europe: Measure of attitudes towards European integration 11 point scale:

Figure 8: Europe

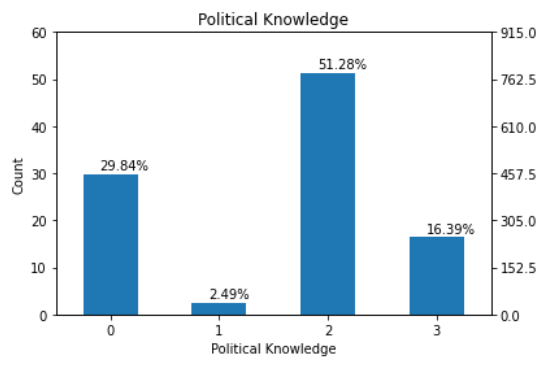


Most of the people are with high attitude towards European integration.

People with high attitude towards European integration are casting their votes.

Political Knowledge:

Figure 9: Political Knowledge

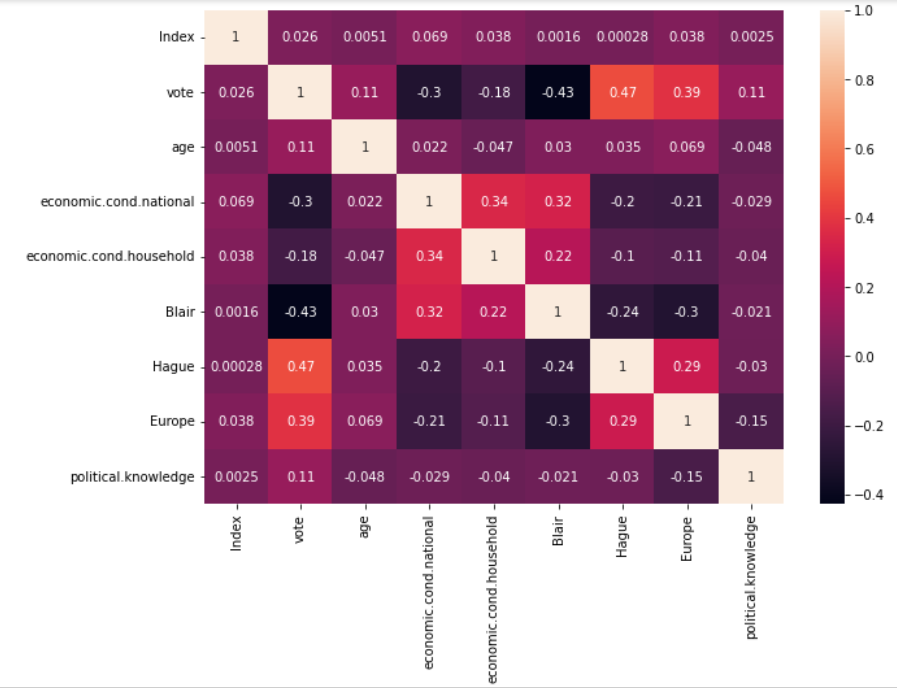


People with knowledge of 2 scale are casting more votes.

There are 30% of the people with 0 knowledge on politics.

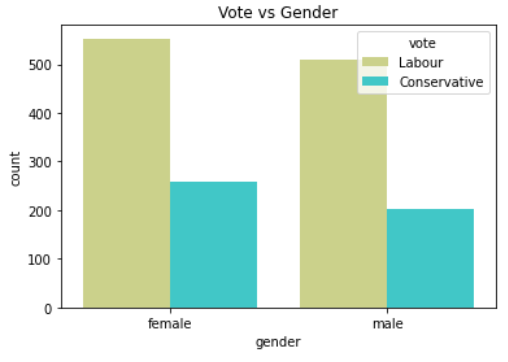
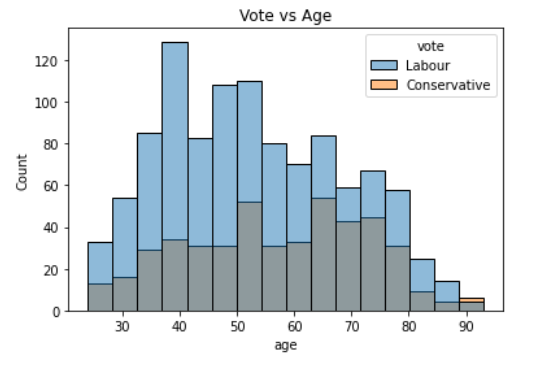
Correlation Plot:

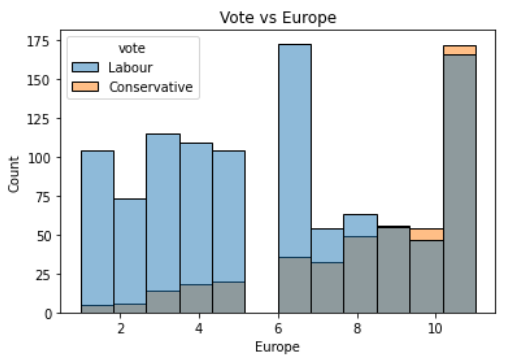
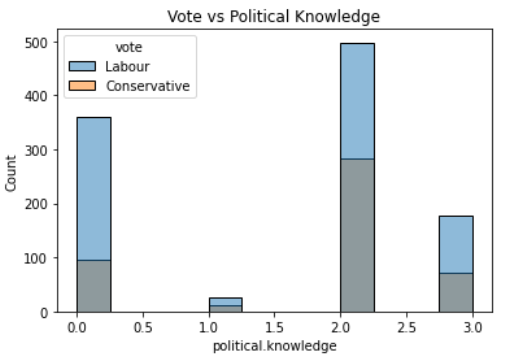
Figure 10:Correlation PLot

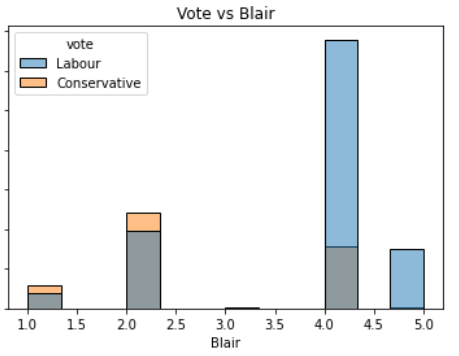
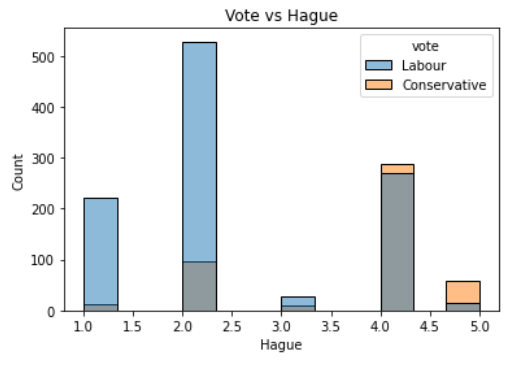


Bivariate Analysis:

Figure 11: Bivariate Analysis



Insights:

* All the age groups people are choosing Labour party to rule them.
* Both male and females are selecting Labour party.
* People with high attitude towards Europe are choosing Conservative party.
* As attitude increases, number of people choosing Conservative Party increasing.
* People with high and low knowledge are not casting their votes. But who casted chosen Labour Party.
* Blair is the knowledge on the Labour party. People with high knowledge on the Labour party vote for them. People with low knowledge vote for Conservative party.
* Hague is the knowledge on the Conservative party. People with high knowledge on the Conservative party vote for them. People with low knowledge vote for Labour party.
* There is no high correlation between the variables.

Outlier Treatment:

There are outliers in Economic. cond. Household and Economic. cond. National. These are to treated.

The data points which are greater than the maximum value are compressed to the maximum value.

The data points which are lesser than the minimum value are compressed to the minimum value.

By this, the outliers are treated.

1. **Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30)**

Encoding:

The vote column is object data type. The values are to be encoded.

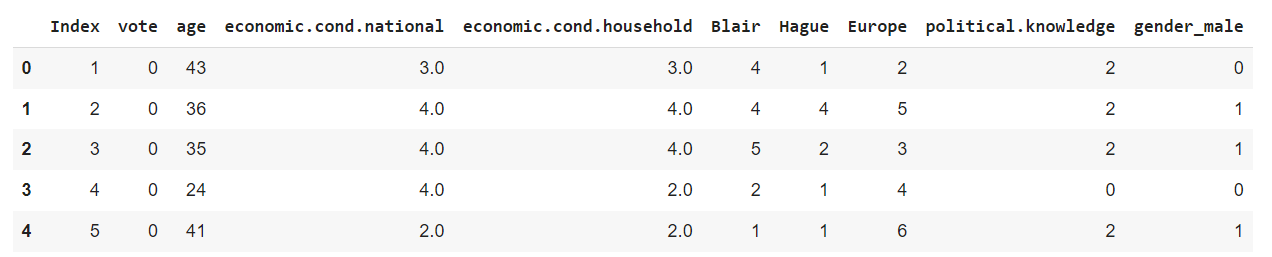
There are two values : Labour, Conservative.

These are encoded by 0,1 respectively.

Other columns are also to be encoded by One Hot Encoding.

After the encoding, the data is as following.

Table 5: Head of the data after encoding



Scaling:

In this case Scaling is required.

As we could see the data. The scale of few columns are different from one another.

Scale of age is different. The range of age variable is [24-93]

Range of economic.cond.national, economic.cond.household, Blair and Hague is [1-5]

Range of Europe is [1-11]

Range of Political knowledge is [0-3]

Hence to bring all of this into a same scale, scaling is required.

Therefore, the data is standardized by zscore.

Table 6: Head of the data after scaling

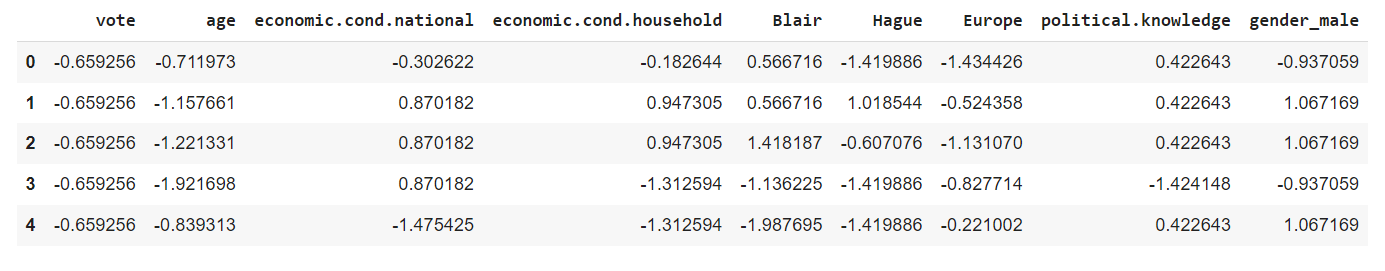
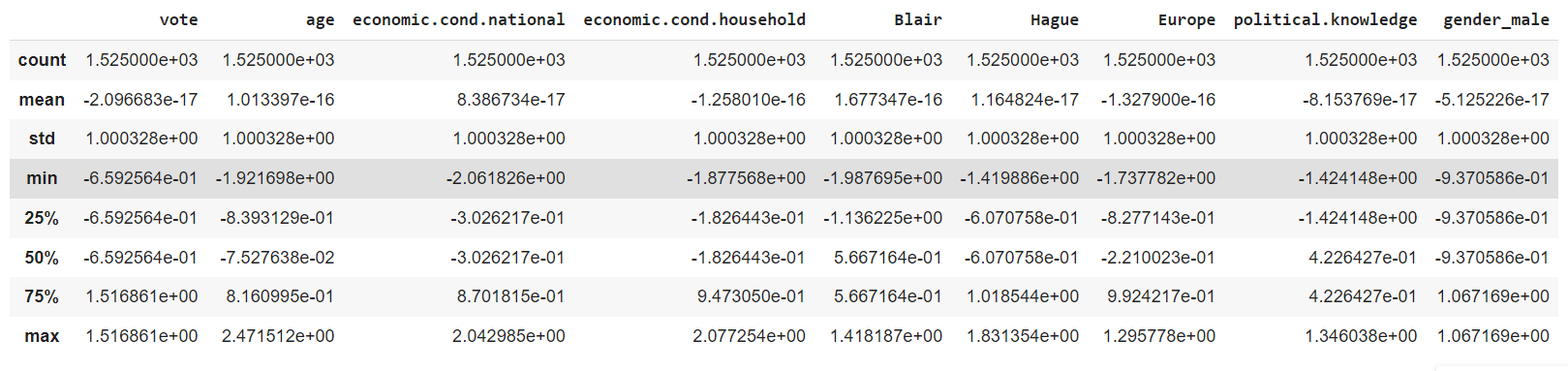


Table 7: statistical information after scaling



**Inferences:**

Scaling does not change the values much. Scaling is just about brings all variables on the same scale.

When data is rescaled the median, mean(μ), and standard deviation(σ) are all rescaled by the same constant. You will multiply by the scaling constant k to determine the new mean, median, or standard deviation. The variance(σ2) is rescaled by multiplying by the scaling constant squared.

Let’s check for the variable age.

Mean of age before scaling is 54 and variance is 245

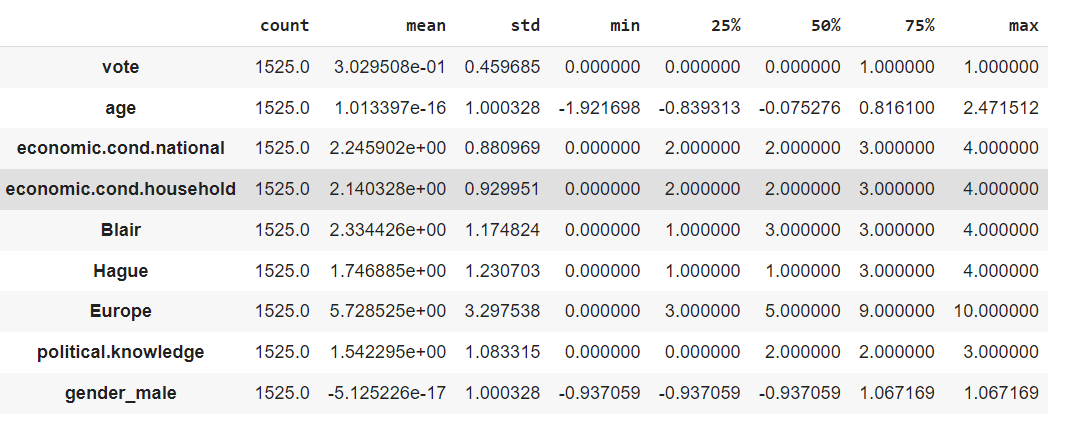
Mean of age after scaling is 1.1 x 10\*\*-6 and variance is 1

We should convert the few numeric data type to categorical data type.

Data such as rating is considered to be categorical.

Here we could see the difference for the variables when converted into categorical data.

Table 8: statistical information after converting categorical data



Train Test Spilt:

Number transactions X\_train dataset: (1067, 8)

Number transactions y\_train dataset: (1067, 1)

Number transactions X\_test dataset: (458, 8)

Number transactions y\_test dataset: (458, 1)

When checked the proportion of the dependent variable that is vote here have imbalnced data.

There are 0: 1063 and 1: 462.

This is to be treated as it the model is prepared with this data there might be skewness.

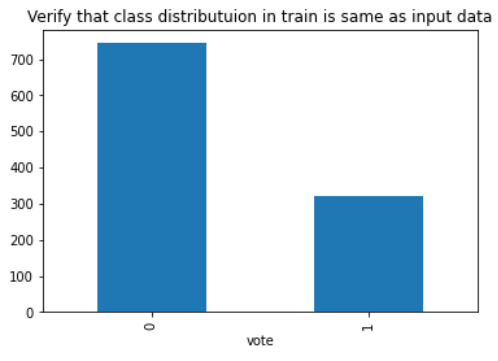
The model predicts 0’s well as there is much data to be trained when divided.

Therefore the data must be balanced first to train models to make predictions.

We are using SMOTE.

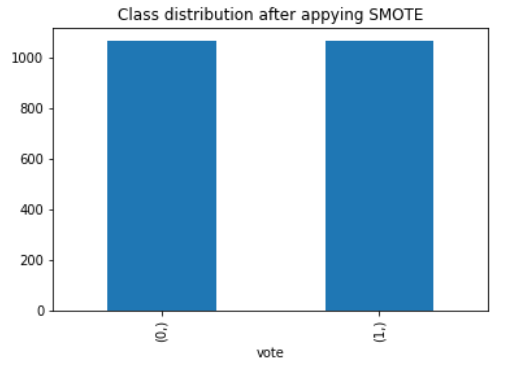
Synthetic Minority Oversampling Technique (SMOTE) is a machine learning technique that solves problems that occur when using an imbalanced data set. Imbalanced data sets often occur in practice, and it is crucial to master the tools needed to work with this type of data.

Figure 12: Vote variable distribution on train data



After applying SMOTE,

Figure 13: vote variable after SMOTE



Here, the data is balanced now. There are 1063 0’s and 1063 1’s.

The shape of the data has changed to 2126 rows and 8 columns.

Index column is dropped before scaling as it is not necessary.

Logistic Regression:

Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable.

In this data set, according to the business need the dependent variable is “Vote” and all other remaining variables are dependent variables.

Let’s split the data into dependent and independent data. Here I used ‘x’ for the independent variables and ‘y’ for the dependent variable.

To train and test the data, we divide the data into two parts in the 70:30 ratio.70% of the data is to train the model and 30% of the data is to test the model.

Here the data must be divided as array, to perform logistic regression.After fitting the model, we calculate the precision, accuracy, recall for training and test data.

Table 9: Classification report for training data without Smote

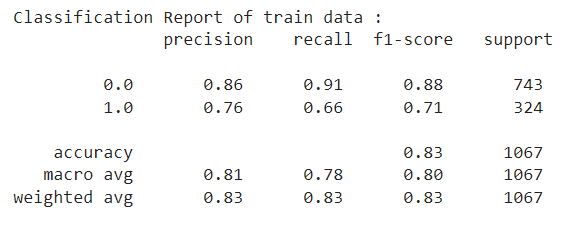
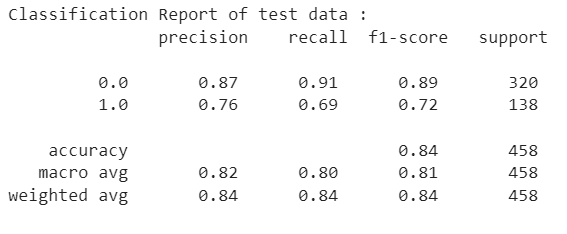


Table 10: Classification report for testing data without Smote



Confusion Matrix:

A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm.

a table of confusion (sometimes also called a confusion matrix) is a table with two rows and two columns that reports the number of true positives, false negatives, false positives, and true negatives. This allows more detailed analysis than simply observing the proportion of correct classifications (accuracy). Accuracy will yield misleading results if the data set is unbalanced; that is, when the numbers of observations in different classes vary greatly.

Figure 14: Confusion matrix for training data without Smote

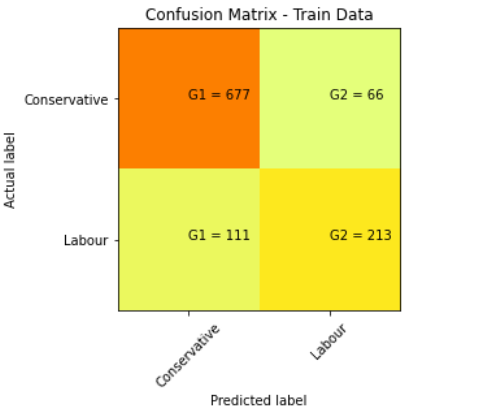
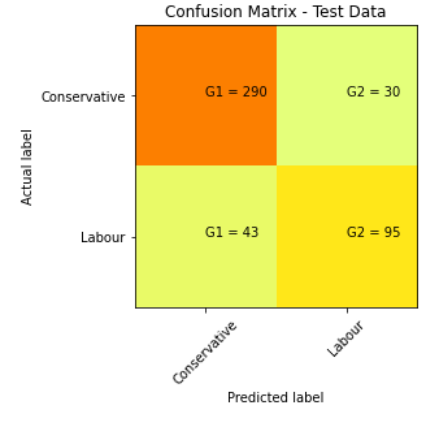


Figure 14: Confusion matrix for testing data without Smote



For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (290 + 95) / (290 + 95+30+43) = 0.84

Precision = TP/(TP + FP)

Precision = 290 / (290 + 43) = 0.87

Recall = TP / (TP + FN)

Recall = 290 / (290 + 30) = 0.90

Table 11: Classification report for testing data with Smote

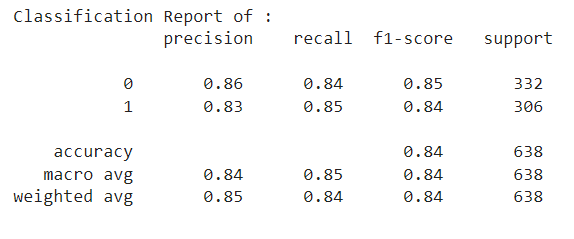


Table 12: Classification report for training data with Smote

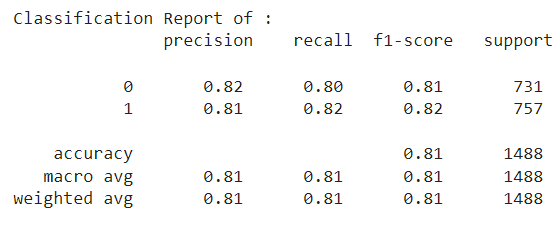
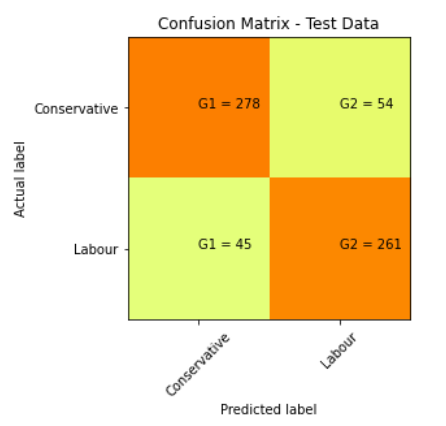


Figure 15: Confusion matrix for testing data with Smote



The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 261
* True Positive: Positive value which is correctly mapped. 278
* False Negative: Positive value, predicted as negative. 54
* False Positive: Negative value, predicted as positive.45

For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (278 + 261) / (278 + 261+54+45) = 0.84

Precision = TP/(TP + FP)

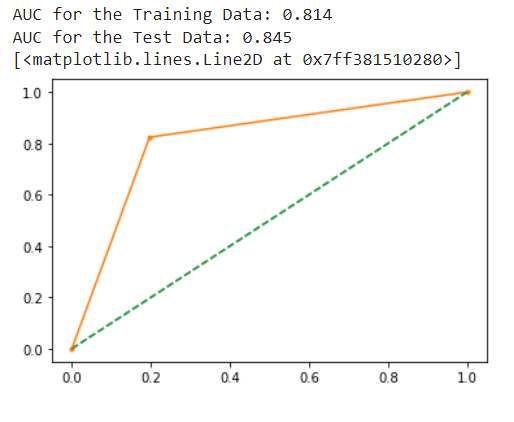
Precision = 278 / (278 + 45) = 0.86

Recall = TP / (TP + FN)

Recall = 278 / (278 + 54) = 0.83

AUC and ROC:

Figure 16: ROC after smote – logistic Regression



**Linear Discriminate Analysis:**

LDA is a supervised classification technique that is considered a part of crafting competitive machine learning models. This category of dimensionality reduction is used in areas like image recognition and predictive analysis in marketing

The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable.

In this data set, according to the business need the dependent variable is ‘Contraceptive method used’ and all other remaining variables are dependent variables.

Let’s split the data into dependent and independent data. Here I used ‘x’ for the independent variables and ‘y’ for the dependent variable.

To train and test the data, we divide the data into two parts in the 70:30 ratio.

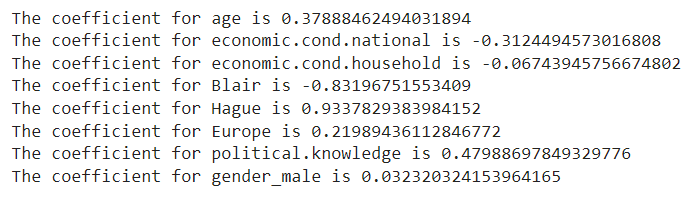
70% of the data is to train the model and 30% of the data is to test the model.

To perform LDA we need import few required libraries.

Now fit the data to the model.

The coefficients are as follows:

Table 13: Coefficients of LDA model before smote

****

Hague is playing the major role in the selection of the party.

The intercept of the data is -2.27

Figure 17: Feature Importance before smote

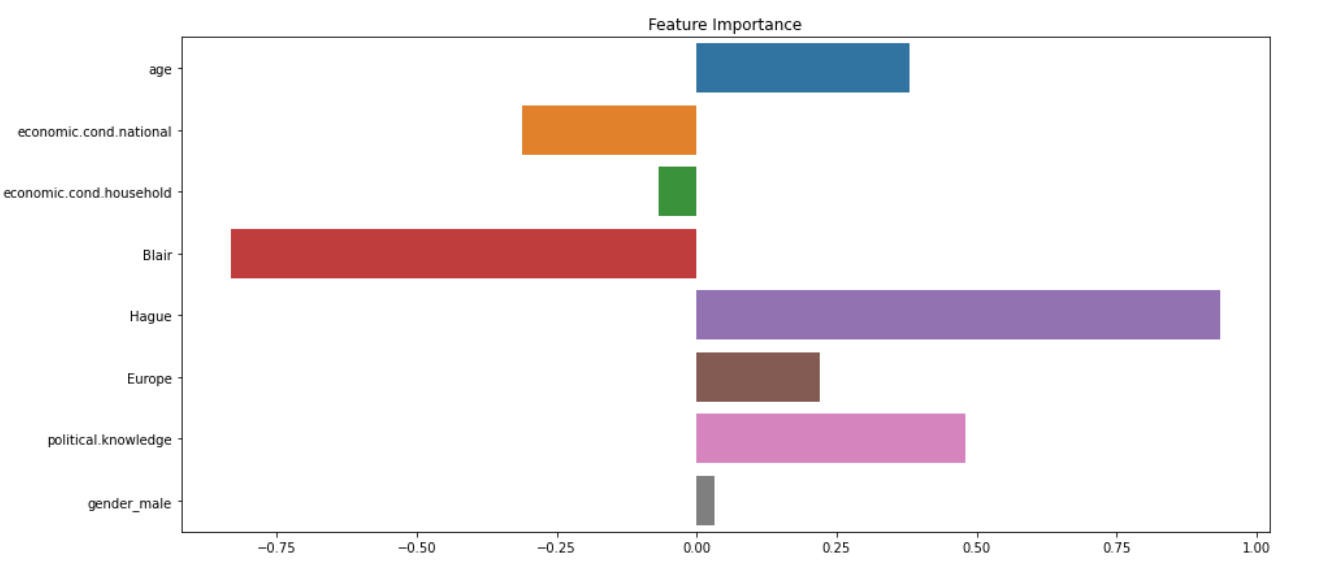


Table 14: Classification report for testing data without Smote

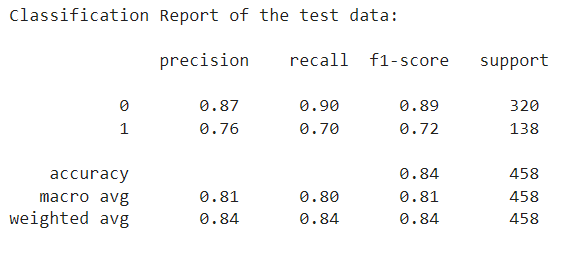
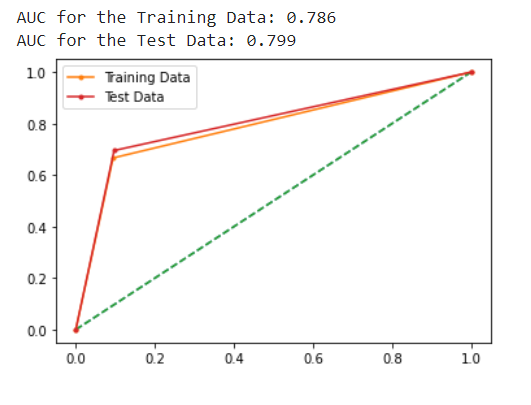
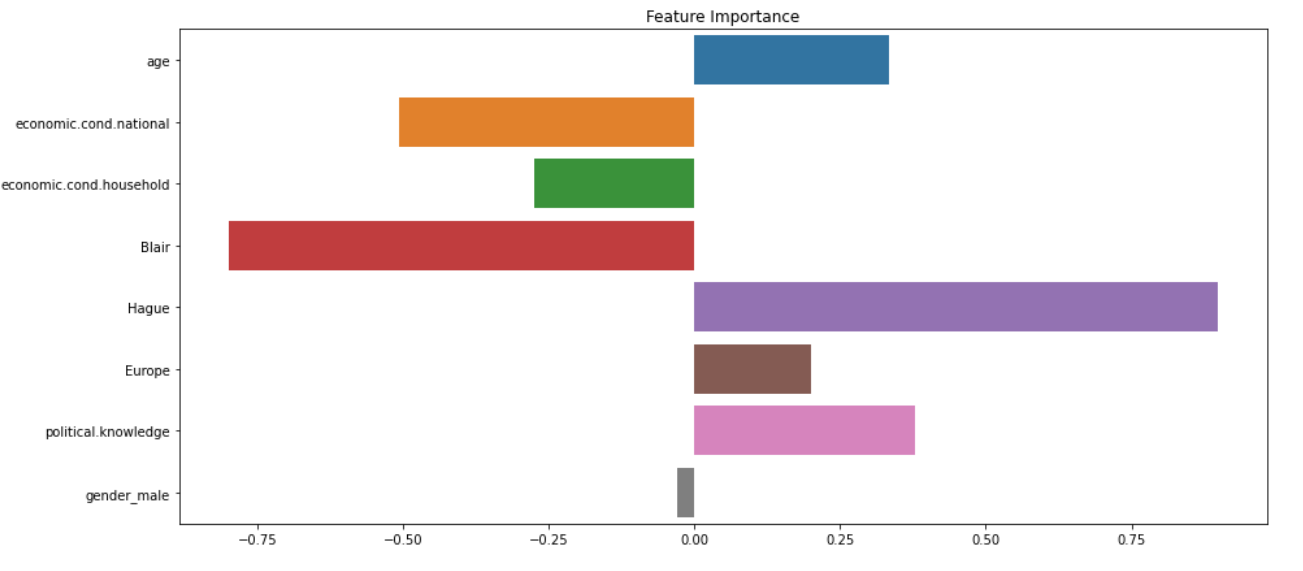


Figure 18: ROC before smote – LDA



After SMOTE – LDA:

Figure 19: Feature Importance after smote



There is no much difference in the coefficients, just difference in points or decimals.

Table 15: Confusion matrix after SMOTE- LDA

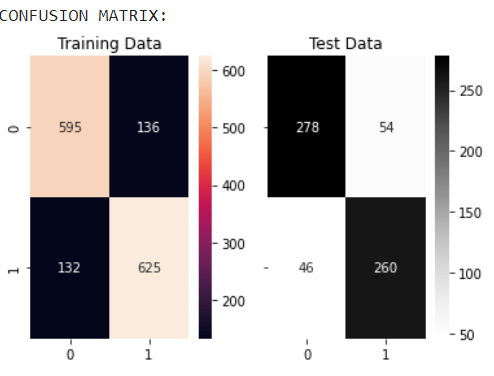
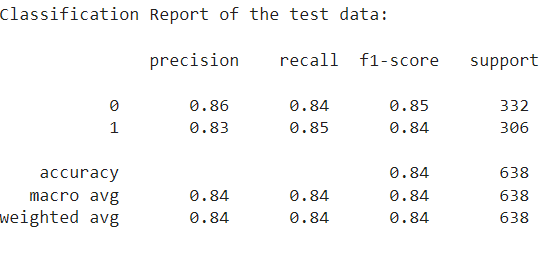


Table 16: Classification report for testing data with Smote



The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 260
* True Positive: Positive value which is correctly mapped. 278
* False Negative: Positive value, predicted as negative. 54
* False Positive: Negative value, predicted as positive.46

For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (278 + 260) / (278 + 261+54+46) = 0.84

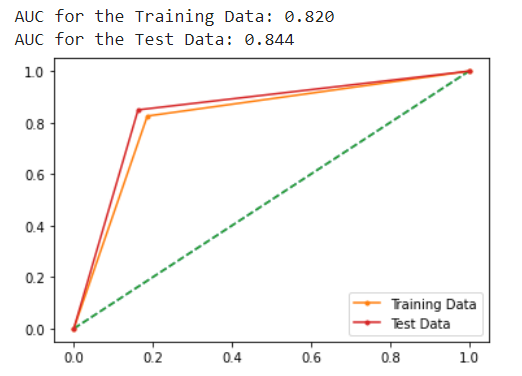
Precision = TP/(TP + FP)

Precision = 278 / (278 + 46) = 0.86

Recall = TP / (TP + FN)

Recall = 278 / (278 + 54) = 0.83

Figure 20: ROC after smote – LDA



KNN – KneighborsClassifier

I have made models on imbalanced data and balanced data here also.

The main importance is given to balanced data, so we focus on balanced data. The model after applying SMOTE.

There is improvement in before models like Logistic Regression and LDA for before and after balancing the data.

Balanced data have better accuracy.

For imbalanced data the k value is 33.

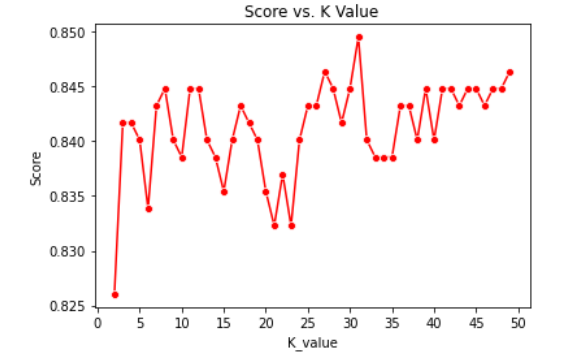
For balanced data the k value is 31.

This can be found by K=sqrt(n)

Where is n is number of rows. The k values would be around 35 – 40.

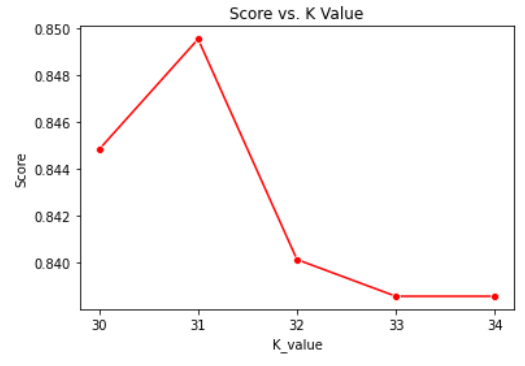
When plotted the line plot with score, we can get better score at a point which is considered as K value.

Figure 21: Score vs K value



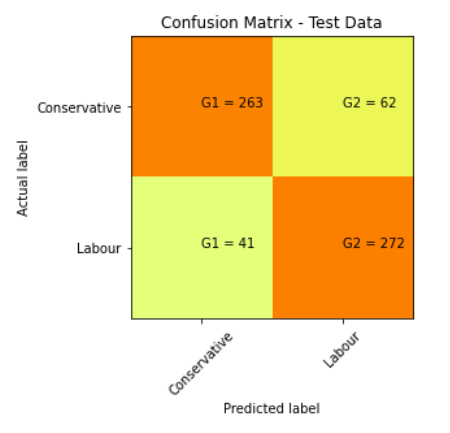
When zoomed to check the exact point:

Figure 22: Score vs K value - zoomed



The K-value is 31. AT K=31, the score is high.

Table 17: Confusion matrix after SMOTE- KNN



The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 263
* True Positive: Positive value which is correctly mapped. 272
* False Negative: Positive value, predicted as negative. 62
* False Positive: Negative value, predicted as positive.41

For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (272 + 263) / (272 + 263+62+41) = 0.833

Precision = TP/(TP + FP)

Precision = 278 / (278 + 46) = 0.858

Recall = TP / (TP + FN)

Recall = 278 / (278 + 54) = 0.837

Table 18: Classification Report after SMOTE - KNN

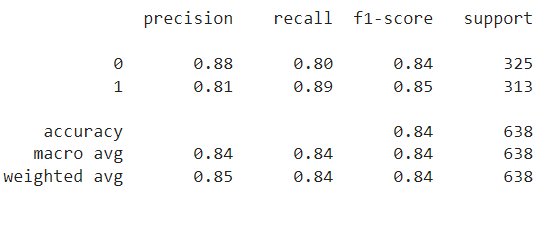
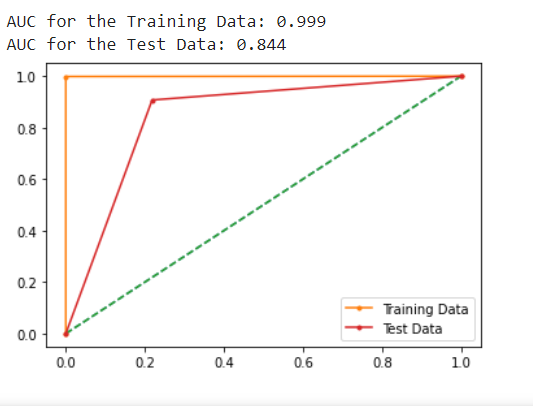


Figure 23: ROC -KNN



Naïve Bayes Model

Table 19: Classification Report after SMOTE-Navie Bayes

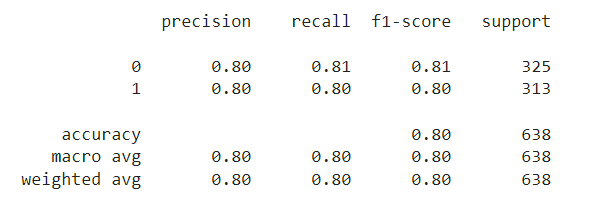
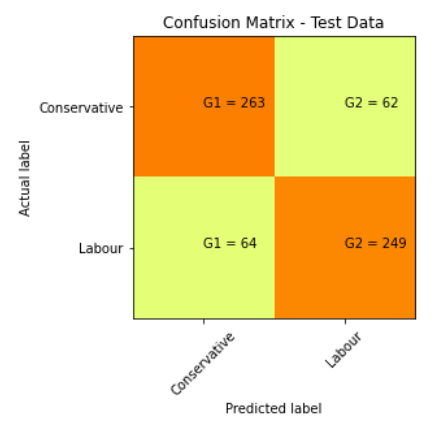


Table 20: Confusion matrix after SMOTE- Navie Bayes



The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 263
* True Positive: Positive value which is correctly mapped. 249
* False Negative: Positive value, predicted as negative. 62
* False Positive: Negative value, predicted as positive.64

For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (249 + 263) / (249 + 263+62+64) = 0.80

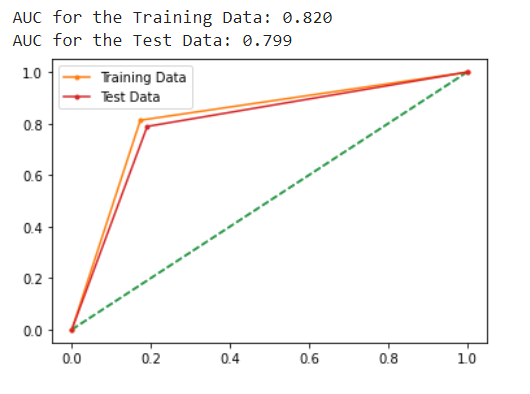
Precision = TP/(TP + FP)

Precision = 249 / (249 + 64) = 0.79

Recall = TP / (TP + FN)

Recall = 249 / (249 + 62) = 0.80

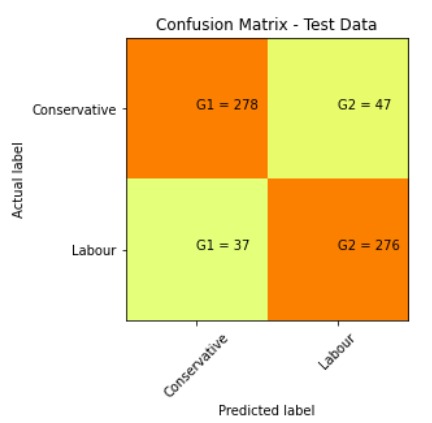
Figure 24: ROC – Naïve Bayes



Model Tunning:

Model tuning is the experimental process of finding the optimal values of hyperparameters to maximize model performance. Hyperparameters are the set of variables whose values cannot be estimated by the model from the training data. These values control the training process.

Table 21: Confusion matrix after SMOTE-



The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 278
* True Positive: Positive value which is correctly mapped. 276
* False Negative: Positive value, predicted as negative. 47
* False Positive: Negative value, predicted as positive.37

For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (278 + 276) / (278 + 276+37+47) = 0.87

Precision = TP/(TP + FP)

Precision = 276/ (276 + 37) = 0.88

Recall = TP / (TP + FN)

Recall = 276 / (276 + 47) = 0.85

Table 22: Classification Report after SMOTE- Random forest

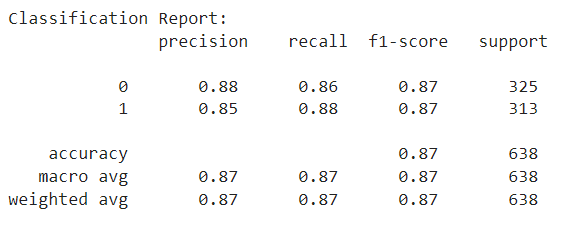
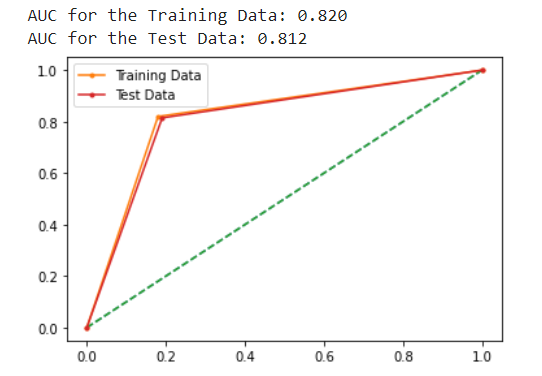


Figure 25 : ROC - Random forest

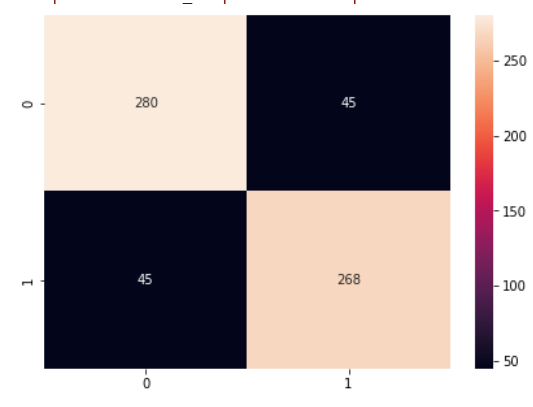


Bagging:

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

It also improves the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

Table 23: Confusion matrix for bagging



The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 268
* True Positive: Positive value which is correctly mapped. 280
* False Negative: Positive value, predicted as negative. 45
* False Positive: Negative value, predicted as positive.45

For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (280 + 268) / (280 + 268+45+45) = 0.86

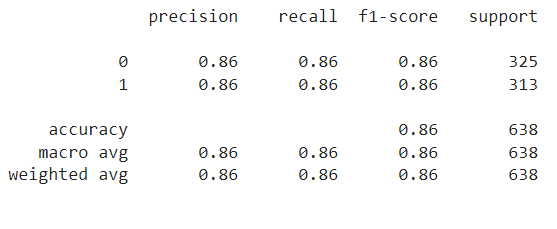
Precision = TP/(TP + FP)

Precision = 276/ (276 + 37) = 0.86

Recall = TP / (TP + FN)

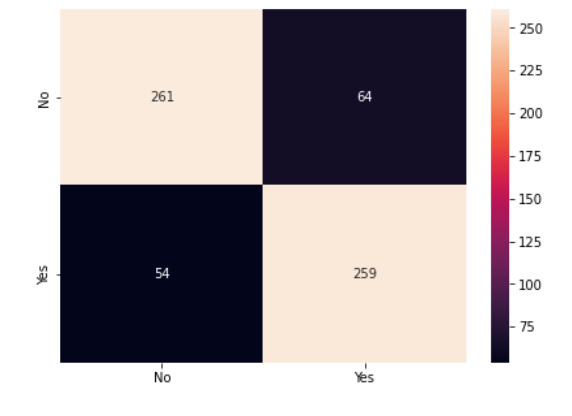
Recall = 276 / (276 + 47) = 0.85

Table 24: Classification Report Bagging



Boosting:

Table 25: Confusion matrix for boosting



The confusion matrix analysis on testing data:

* True Negative: Negative value which is correctly mapped. 259
* True Positive: Positive value which is correctly mapped. 261
* False Negative: Positive value, predicted as negative. 64
* False Positive: Negative value, predicted as positive.54

For testing data without SMOTE:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy = (261 + 259) / (261 + 259+54+64) = 0.815

Precision = TP/(TP + FP)

Precision = 261/ (261 + 54) = 0.83

Recall = TP / (TP + FN)

Recall = 261 / (261 + 64) = 0.80

Table 26: Classification Report Boosting

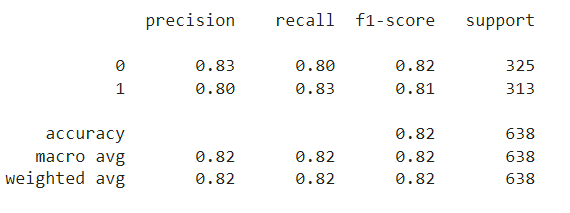


Table 27: Comparison Table between Models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | LDA | Random Forest | K-Nearest Neighbour | Naïve Bayes | Bagging | Boosting |
| Accuracy | 84 | 84 | 87 | 85 | 83 | 87 | 82 |
| Precision | 86 | 86 | 88 | 86 | 79 | 86 | 83 |
| Recall | 83 | 83 | 85 | 84 | 80 | 88 | 80 |
| AUC Train | 81 | 82 | 82 | 99 | 82 | 99 | 83 |
| Auc Test | 84 | 84 | 81 | 85 | 83 | 87 | 82 |
| Overfit/  Underfit | Overfit | Overfit | Overfit | Good | Overfit | Good | Overfit |

From the above table, I could conclude that KNN and Bagging models is the best fit for this data.

The training data for the KNN is 99% of which 85% of the data is explained.

Basically, bagging is the technique to avoid overfitting.

This technique can also be used in this model.

**Insights on the data:**

There should be some measures taken regarding:

* People with the age below 30 and above 80 voting rate is very low.
* People with high and low national economic standards are not casting their votes.
* People should have enough political knowledge to choose their leader.
* There should be more events conducted to change the attitude of the people towards Europe.
* Knowledge on Conservative leaders is less when compared to Labour leaders.
* Most of the people chose the Labour leader, hence the data is unbalanced.

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

**2.1 Find the number of characters, words, and sentences for the mentioned documents.**

**The data is loaded and worked on the count of the files mention are as following:**

The 1973-Nixon.txt file has :

The number od characters : 9991

The number of words : 2028

The number of sentences : 69

The number od vocabulary : 516

The 1961-Kennedy.txt file has :

The number od characters : 7618

The number of words : 1546

The number of sentences : 52

The number od vocabulary : 546

The 1941-Roosevelt.txt file has :

The number od characters : 7571

The number of words : 1536

The number of sentences : 68

The number od vocabulary : 502

**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.**

Stop words are words that are so common they are basically ignored by typical tokenizers. By default, NLTK (Natural Language Toolkit) includes a list of 40 stop words, including: “a”, “an”, “the”, “of”, “in”, etc.

The stopwords in nltk are the most common words in data. The stopwords and puntuations are listed below:

< ,( ,, ,we ,yourself ,because ,wouldn't ,` ,they ,or ,wasn ," ,below ,+ ,of ,then ,i ,further ,- ,himself ,? ,) ,from ,itself ,each ,she ,yours ,her ,to ,own ,has ,when ,any ,didn't ,out ,down ,that'll ,| ,having ,ma ,me ,his ,don't ,shan ,s ,about ,while ,other ,once ,needn ,this ,d ,by ,not ,mustn ,after ,again ,aren ,very ,have ,she's ,[ ,did ,some ,too ,you ,on ,ll ,shan't ,shouldn ,shouldn't ,will ,in ,no ,under ,weren ,herself ,~ ,my ,} ,just ,over ,aren't ,couldn't ,hadn't ,hasn ,the ,don ,hasn't ,$ ,same ,\_ ,as ,-- ,\* ,won ,^ ,if ,were ,m ,should've ,. ,am ,against ,# ,you'll ,ours ,between ,ve ,that ,mightn't ,] ,isn't ,he ,such ,how ,you'd ,until ,it's ,all ,wasn't ,both ,' ,\ ,was ,these ,a ,few ,mightn ,there ,weren't ,doing ,for ,hers ,> ,won't ,/ ,mustn't ,what ,and ,ain ,into ,is ,an ,{ ,which ,@ ,had ,only ,up ,them ,couldn ,themselves ,does ,= ,re ,your ,nor ,their ,who ,haven't ,wouldn ,been ,off ,theirs ,being ,doesn't ,now ,& ,those ,% ,needn't ,during ,but ,are ,y ,be ,isn ,hadn ,why ,before ,its ,where ,it ,here ,; ,should ,ourselves ,yourselves ,didn ,doesn ,myself ,you've ,t ,so ,our ,whom ,do ,you're ,most ,more ,haven ,: ,o ,through ,with ,than ,him ,above ,! ,can ,at.

**The stop words are removed in the file 1961-Kennedy.txt:**

The word count before removing the stopwords is 1546

The word cound of the Kennedy file after removing stopwords : 696

The sample Kennedy file after the removal of stopwords and puntuations:

vice president johnson mr. speaker mr. chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change sworn almighty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands power abolish forms human poverty forms human life yet revolutionary beliefs forebears fought still issue around globe belief rights man come generosity.

**The stop words are removed in the file 1973-Nixon.txt:**

The word count before removing the stopwords is 2028

The word cound of the Nixon file after removing stopwords : 632

The file after the removal of stopwords and puntuations:

national day inauguration since 1789 people renewed sense dedication united states washington 's day task people create weld together nation lincoln 's day task people preserve nation disruption within day task people save nation institutions disruption without us come time midst swift happenings pause moment take stock recall place history rediscover may risk real peril inaction lives nations determined count years lifetime human spirit life man three-score years ten little little less life nation fullness measure live men doubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplained reason tyranny slavery become surging wave future freedom ebbing tide americans know true eight years ago life republic seemed frozen fatalistic terror proved true midst shock acted acted quickly boldly decisively later years

**The stop words are removed in the file 1941-Roosevelt.txt:**

The word count before removing the stopwords is 1536

The word cound of the Roosevelet file after removing stopwords : 848

The file after the removal of stopwords and puntuations:

mr. vice president mr. speaker mr. chief justice senator cook mrs. eisenhower fellow citizens great good country share together met four years ago america bleak spirit depressed prospect seemingly endless war abroad destructive conflict home meet today stand threshold new era peace world central question us shall use peace let us resolve era enter postwar periods often time retreat isolation leads stagnation home invites new danger abroad let us resolve become time great responsibilities greatly borne renew spirit promise america enter third century nation past year saw far-reaching results new policies peace continuing revitalize traditional friendships missions peking moscow able establish base new durable pattern relationships among nations world america 's bold initiatives 1972 long remembered year greatest progress since end world war ii toward lasting

**2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words.**

'people': 594,

'government': 570, '

us': 502,

'must': 374,

'upon': 371,

'great': 346,

'may': 343,

'world': 343,

'states': 335,

'nation': 328.

The ‘people’ is the most common word occurring among all the files.

The top three words is '

people': 594,

'government': 570, '

us': 502.

**2.4) Plot the word cloud of each of the three speeches**

Figure 26: Word cloud

****