MACHINE LEARNING

Project Report

By Ruchita Parulekar

Table of Contents

[**Problem 1** 3](#_Toc152323582)

[**1.1** **Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.** 3](#_Toc152323583)

[**1.2** **Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.** 5](#_Toc152323584)

[**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).** 10](#_Toc152323585)

[**1.4 Apply Logistic Regression and LDA (linear discriminant analysis).** 12](#_Toc152323586)

[**1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.** 13](#_Toc152323587)

[**1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.** 14](#_Toc152323588)

[**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.** 15](#_Toc152323589)

[**1.8 Based on these predictions, what are the insights?** 25](#_Toc152323590)

[Best model selection 25](#_Toc152323591)

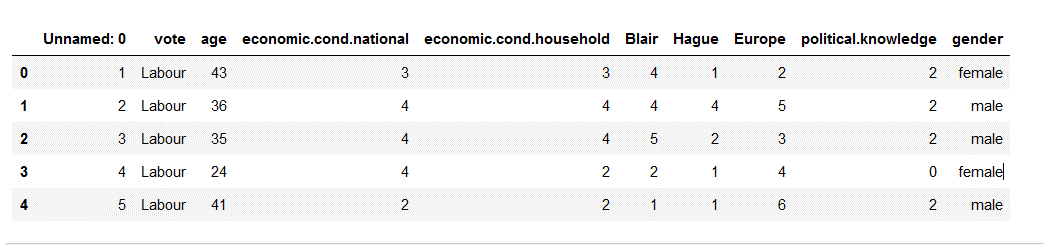
[Recommendations: 26](#_Toc152323592)

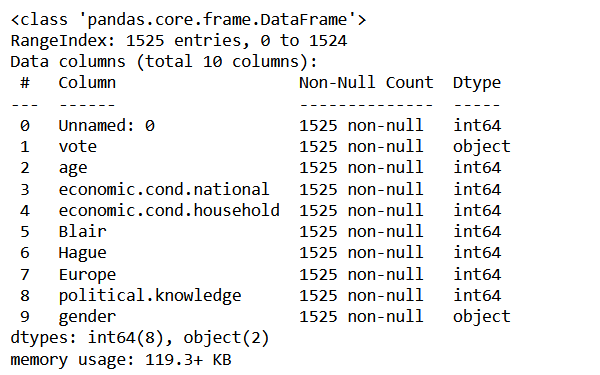
# **Problem 1**

*You are hired by one of the leading news channels CNBE who wants to analyse 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.*

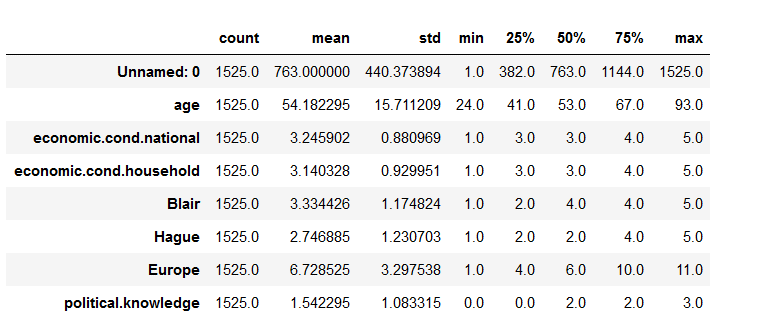
## **Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.**

**Solution:**





* The data frame has 1525 rows and 10 columns, and there are no missing values in any of the columns.
* It has 2 (Two) Object Datatypes and 8 (eight) Integer Datatypes.



* The data frame has 1525 rows and 10 columns, and there are no missing values in any of the columns.
* It has 2 (Two) Object Datatypes and 8 (eight) Integer Datatypes.

A screen shot of a computer code

Description automatically generated

* There are no missing values in the data.
* Some Column names contain ‘.’ in name that can affect the model, so we will replace the ‘.’ With ‘\_’ operator.
* ‘Unnamed: 0’ is a variable that simply represents the index in the data. Hence, it should be dropped as it is of no use in the model.

A screen shot of a computer code

Description automatically generated A close-up of a computer screen

Description automatically generated

* There are total 8 rows that are duplicated, we cannot clearly say that this is the same person or different. So, we will not remove the duplicates.

A white background with black text

Description automatically generated

* Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.
* Only Two variables are positively skewed, rest are negatively skewed.

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

**Solution:**

**A graph of a number of lines

Description automatically generated with medium confidence**

* We can see that all the numerical Variables are normally distributed.
* There are outliers present in “economic\_cond\_natianal” and “economic\_cond\_household” variables that can be seen from the boxplots.
* It is not advisable to treat the outliers. We will move forward without treating the outliers .

A graph of a number of blue lines

Description automatically generated with medium confidence

* Pairplot tells us about the interaction of each variable with every other variable present.
* As such there is no strong relationship present between the variables. There is a mixture of positive and negative relationships though which is expected.
* Overall, it’s a rough estimate of the interactions, clearer picture can be obtained by heatmap values and different kinds of plot.

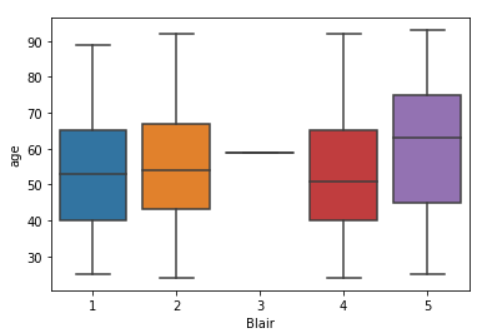
A graph of a bar chart

Description automatically generated with medium confidence A graph of blue and orange bars

Description automatically generated

* People who have voted for Labour Party are a lot more than those who voted for Conservative Party.
* People from the age of 35-65 have voted more for the Labour party.

A graph of different colored dots

Description automatically generated with medium confidence 

* People above the age of 45 years generally thinks that Blair is doing a good job.

A diagram of different colored dots

Description automatically generated A chart with different colored squares

Description automatically generated

* Hague has slightly more concentration of neutral points than that of Blair for people above 50 years of age.

A graph of different colored bars

Description automatically generated A graph of a graph with numbers and a chart

Description automatically generated with medium confidence

* Blair has more points in terms of economic households than Hague.

A graph of different colored bars

Description automatically generated A graph of different colored bars

Description automatically generated

* Blair has more points in terms of economic national than Hague.

A graph of different colored lines

Description automatically generated A graph of different colored lines

Description automatically generated

* In the whole Europe if we look at the data then Blair is leading.

Covariance Matrix: -

A screenshot of a computer

Description automatically generated

Correlation Matrix: -

A screenshot of a computer

Description automatically generated

Heatmap: -

A screenshot of a computer screen

Description automatically generated

* Multicollinearity is an important issue which can harm the model.
* Heatmap is a good way of identifying this issue.
* It gives us a basic ides of relationship the variables have with each other.

Observations:

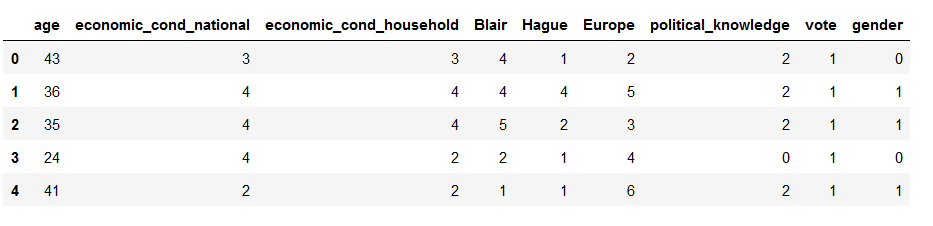
* Highest positive correlation is between “economic\_cond\_national” and “economic\_cond\_household” (35%). But the good thing is that it’s not huge.
* Highest negative correlation is between “Blair” and “Europe” (30%) but this is also not huge.

Thus, Multicollinearity wont’ be an issue is this dataset.

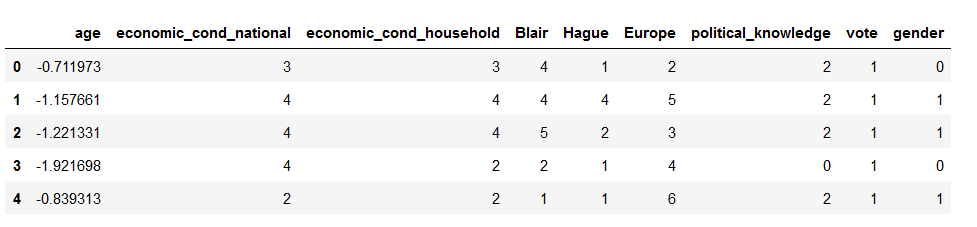
## **1.3. Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).**

**Solution:**

* As many machine learning models cannot work with string values, we will encode categorical variables and convert their datatypes to integer type.
* From the info of the dataset, we know there are 2 categorical type variables, so we need to encode these 2 variables with the suitable technique.

****

* In order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.
* After looking at the data we only need to scale the “age” variable as rest of the variables are in the range 0-10 at max.
* We will use z-score scaling here to scale the age variable.

****

* Now, we will split the data in 70:30 ratio, where train data is 70% and the test data is 30%.

Train data:

**A screenshot of a computer

Description automatically generated**

Test Data:

**A screenshot of a computer

Description automatically generated**

## **1.4 Apply Logistic Regression and LDA (linear discriminant analysis).**

**Solution:**

**The hyper parameters that are involved in the model building of Logistic Regression:**

* Penalty
* Solver
* Max\_iter
* Tol

A screenshot of a calculator

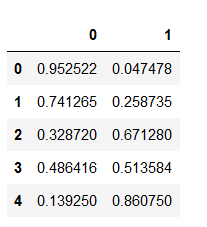
Description automatically generated

Accuracy of Train data: 0. 8397375820056232

Accuracy of Test data: 0.8187772925764192

* There is no underfitting or overfitting present as accuracy for both test and train data are not very different.

**The hyper parameters are not involved in model building of Linear Discriminant Analysis:**



Accuracy of Train data: 0.8369259606373008

Accuracy of Test data: 0.8187772925764192

* There is no underfitting or overfitting present as accuracy for both test and train data are not very different.

## **1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.**

**Solution:**

**The hyper parameters that are involved in the model building of KNN Model:**

* weights
* n\_neighbors

A close up of numbers

Description automatically generated

Accuracy of Train data: 0.9990627928772259

Accuracy of Test data: 0.7969432314410481

* There is a huge Overfitting as accuracy is more for train set than for the test set.

**The hyper parameters are not involved in the model building of Bayes Model:**

A screenshot of a calculator

Description automatically generated

Accuracy of Train data: 0.8331771321462043

Accuracy of Test data: 0.8253275109170306

* There is no underfitting or overfitting present as accuracy for both test and train data are not very different.

## **1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.**

**Solution:**

**The hyper parameters are not involved in the model building of Bagging Model:**

Accuracy of Train data: 0.9962511715089035

Accuracy of Test data: 0.8056768558951966

* There is a huge Overfitting as accuracy is more for train set than for the test set.

**The hyper parameters are not involved in the model building of AdaBoosting Model:**

Accuracy of Train data: 0.8397375820056232

Accuracy of Test data: 0.8231441048034934

* There is no underfitting or overfitting present as accuracy for both test and train data are not very similar.

**The hyper parameters are not involved in the model building of Gradient Model:**

Accuracy of Train data: 0.8734770384254921

Accuracy of Test data: 0.8362445414847162

* There is no underfitting or overfitting present as accuracy for both test and train data are not very similar.

## **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.**

**Solution:**

Usually there are many performance metrics that are used in assessing the strength of the model to understand how the model has performed as well as to take an informed decision on whether to go forward with the model in the real time scenario or not.

**Logistic Regression**

Accuracy of Train data: 0. 8397375820056232

Accuracy of Test data: 0.8187772925764192

***AUC and ROC for training and testing data respectively***

**A graph with a line

Description automatically generatedA graph with a blue line

Description automatically generated**

***Confusion Matrix for training and testing data respectively***

**A chart with numbers and a few colored squares

Description automatically generated with medium confidence A chart with numbers and a few colored squares

Description automatically generated with medium confidence**

***Classification Report for training and testing data respectively***

**A screenshot of a graph

Description automatically generated**

**A screenshot of a computer

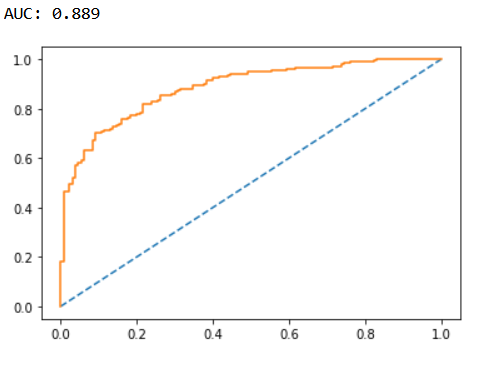
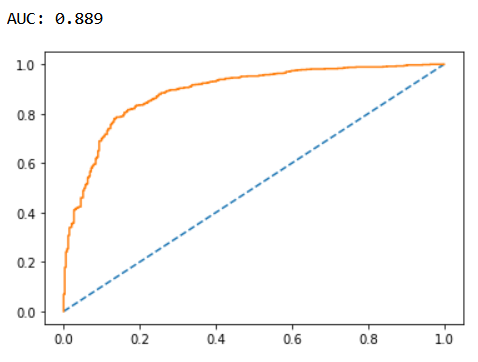
Description automatically generated**

**Linear Discriminant Analysis**

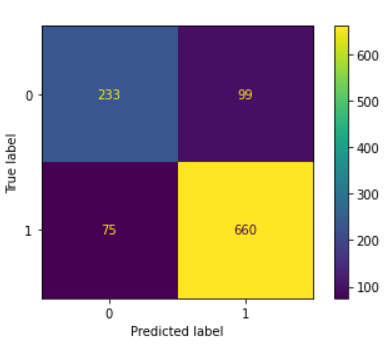
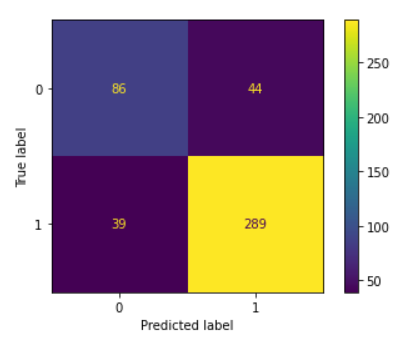
Accuracy of Train data: 0.8369259606373008

Accuracy of Test data: 0.8187772925764192

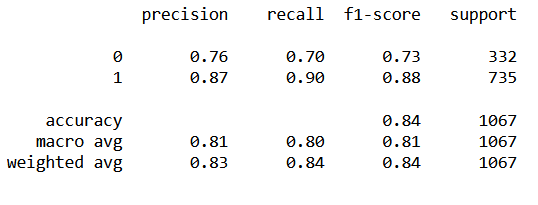
***AUC and ROC for training and testing data respectively***

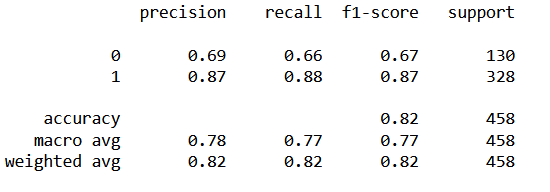
****

***Confusion Matrix for training and testing data respectively***

** **

***Classification Report for training and testing data respectively***

****

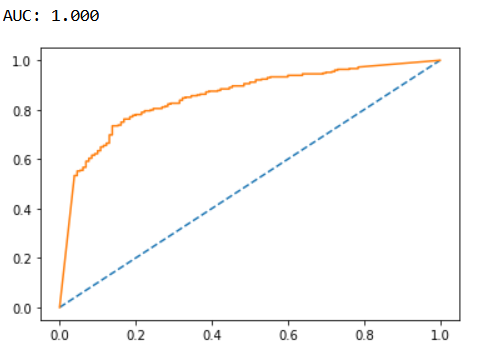
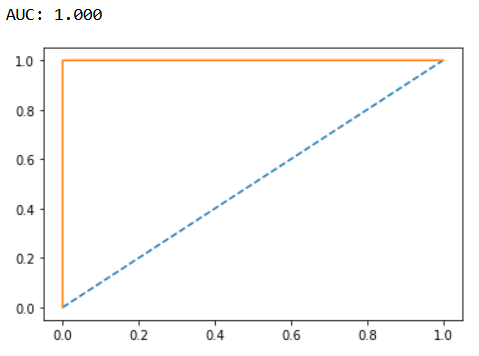
****

**KNN Model**

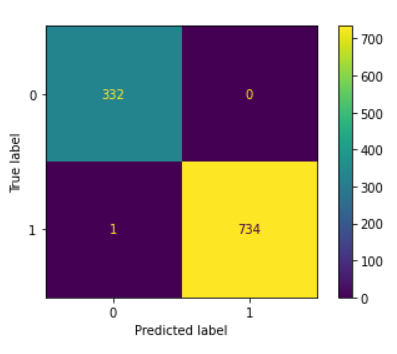
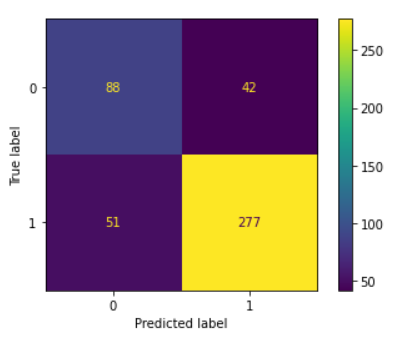
Accuracy of Train data: 0.9990627928772259

Accuracy of Test data: 0.7969432314410481

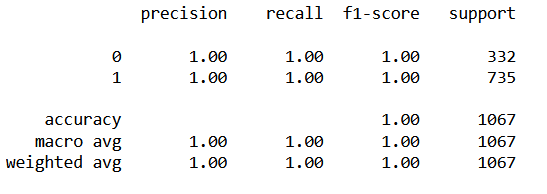
***AUC and ROC for training and testing data respectively***

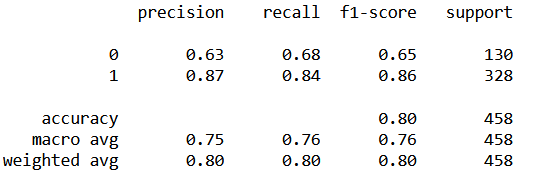
****

***Confusion Matrix for training and testing data respectively***

** **

***Classification Report for training and testing data respectively***

****

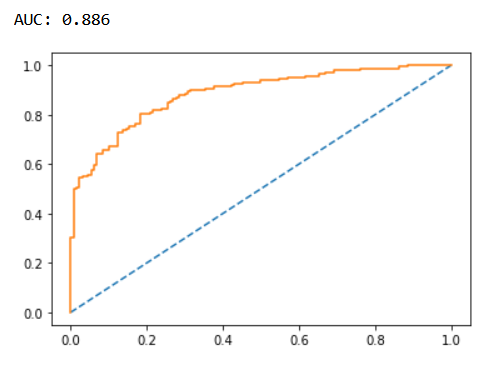
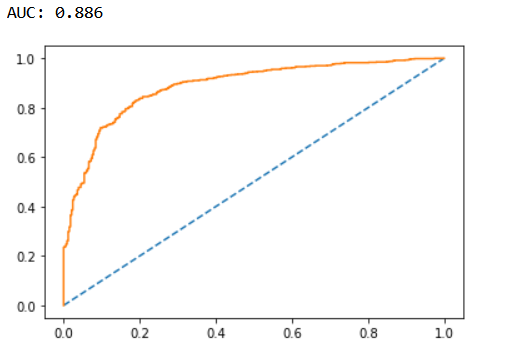
****

**Bayes Model**

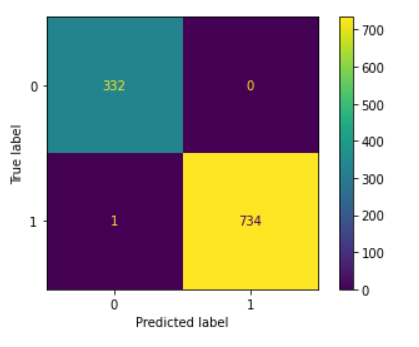
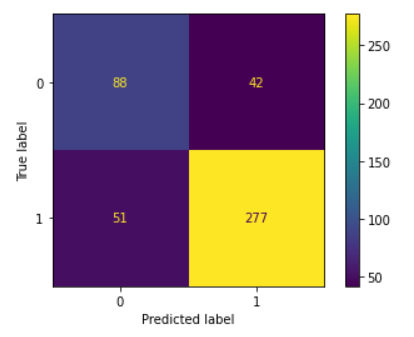
Accuracy of Train data: 0.8331771321462043

Accuracy of Test data: 0.8253275109170306

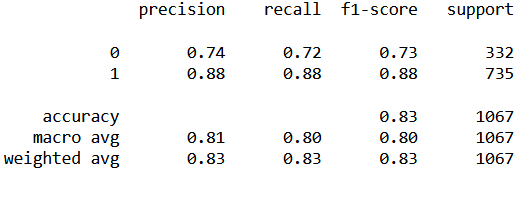
***AUC and ROC for training and testing data respectively***

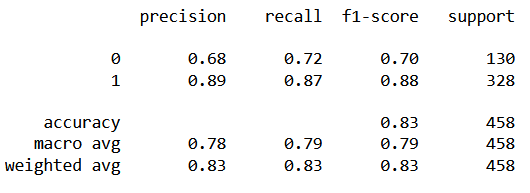
****

***Confusion Matrix for training and testing data respectively***

** **

***Classification Report for training and testing data respectively***

****

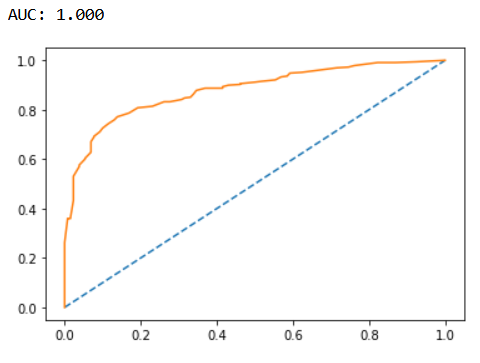
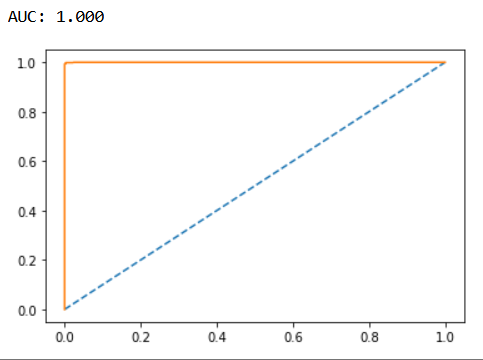
****

**Bagging Model**

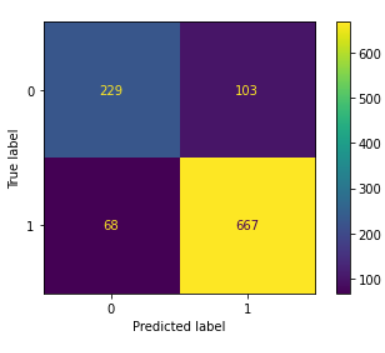
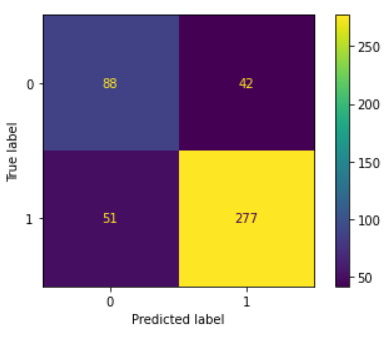
Accuracy of Train data: 0.9962511715089035

Accuracy of Test data: 0.8056768558951966

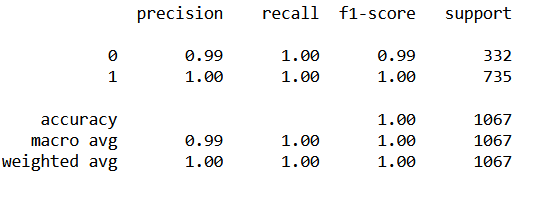
***AUC and ROC for training and testing data respectively***

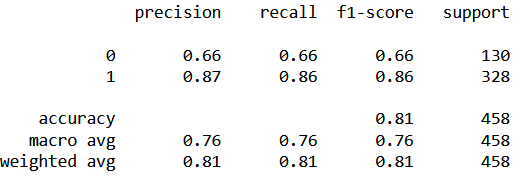
****

***Confusion Matrix for training and testing data respectively***

** **

***Classification Report for training and testing data respectively***

****

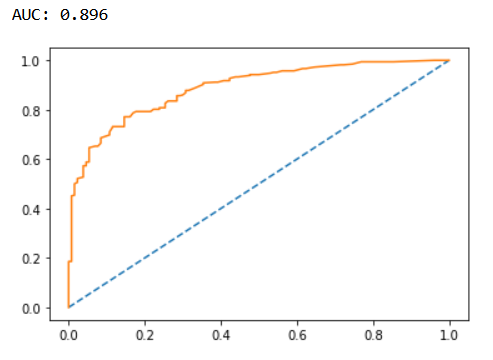
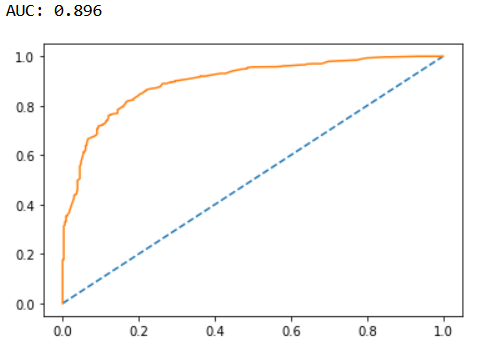
****

**Ada Boosting Model**

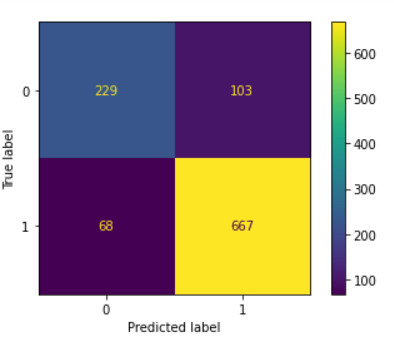
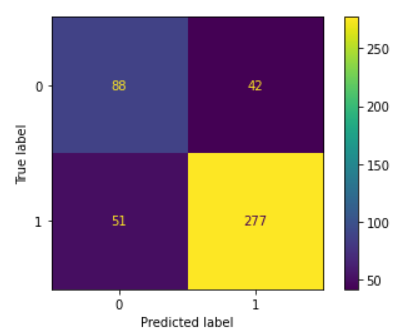
Accuracy of Train data: 0.8397375820056232

Accuracy of Test data: 0.8231441048034934

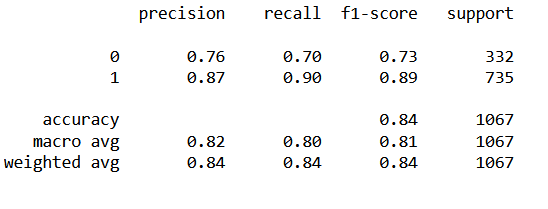
***AUC and ROC for training and testing data respectively***

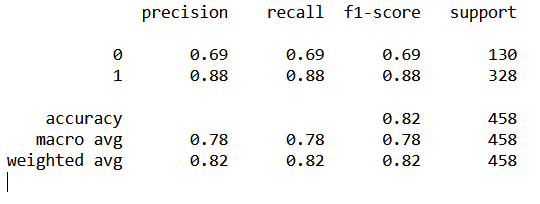
****

***Confusion Matrix for training and testing data respectively***

** **

***Classification Report for training and testing data respectively***

****

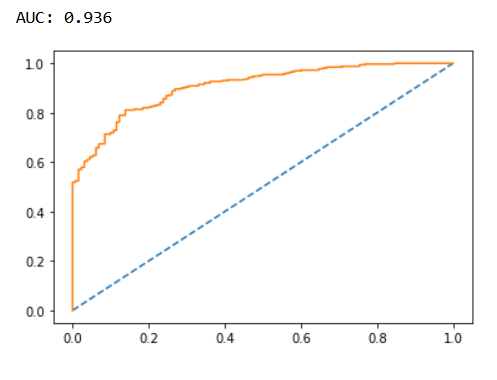
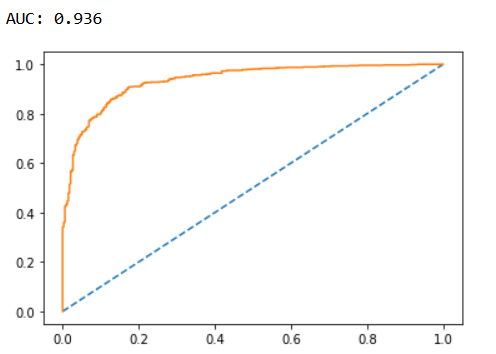
****

**Gradient Boosting Model**

Accuracy of Train data: 0.8734770384254921

Accuracy of Test data: 0.8362445414847162

***AUC and ROC for training and testing data respectively***

****

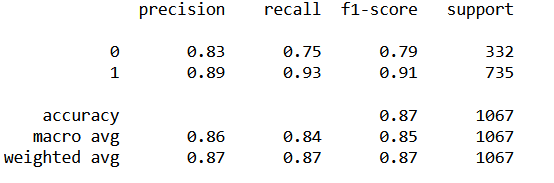
***Confusion Matrix for training and testing data respectively***

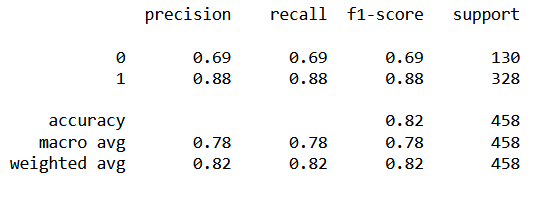
**A chart with numbers and labels

Description automatically generated with medium confidence A chart with numbers and a few colored squares

Description automatically generated with medium confidence**

***Classification Report for training and testing data respectively***

****

****

## **1.8 Based on these predictions, what are the insights?**

**Solution:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | | | | AUC | | | | Recall | | | | F1 score | | | Precision | | |
|  | **Train** | | **Test** | | **Train** | | **Test** | | **Train** | | **Test** | | **Train** | | **Test** | **Train** | | **Test** |
| Logistic Regression | **0.84** | **0.82** | | **0.89** | | **0.89** | | **0.93** | | **0.89** | | **0.89** | | **0.88** | | **0.87** | **0.86** | |
| LDA | **0.84** | **0.82** | | **0.88** | | **0.88** | | **0.90** | | **0.88** | | **0.88** | | **0.87** | | **0.87** | **0.87** | |
| KNN | **1.00** | **0.80** | | **1.00** | | **1.00** | | **1.00** | | **0.84** | | **1.00** | | **0.86** | | **1.00** | **0.87** | |
| Bayes | **0.83** | **0.83** | | **0.88** | | **0.88** | | **0.88** | | **0.87** | | **0.88** | | **0.88** | | **0.88** | **0.89** | |
| Bagging | **1.00** | **0.81** | | **1.00** | | **1.00** | | **1.00** | | **0.86** | | **1.00** | | **0.86** | | **1.00** | **0.87** | |
| AdaBoosting | **0.84** | **0.82** | | **0.89** | | **0.89** | | **0.90** | | **0.88** | | **0.89** | | **0.88** | | **0.87** | **0.88** | |
| Gradient Boosting | **0.87** | **0.82** | | **0.936** | | **0.936** | | **0.93** | | **0.88** | | **0.91** | | **0.88** | | **0.89** | **0.88** | |

* From all the inferences above, we see that mostly all the models have similar performance.
* The Accuracy score for all the models are above 85% for both test and train data.

Best model selection:

* With this, it is also very clear that the Bayes model has performed above all the rest of the models.
* With an Accuracy value of 83%, it is predicting the highest percentage of both our Train and Test set.
* If we still look at the Recall value, the Bayes model is able to identify 87% of the true positives correctly.
* The LDA, Logistic Regression and a few other models gives a higher Recall value, however the Accuracy of the Bayes model is slightly higher therefore it would be better to consider the Bayes model for doing the prediction.
* Similarly, we see that the Area Under the Curve (AUC) captured is 88% for train data and 88% for the test data. It is quite good. Therefore, it is safe to say that this model can be used for making predictions on any unseen data that is fed to the model.

## Recommendations:

* People from the age of 35-65 have voted more for the Labour party, which is Blair’s Party.
* Hague has slightly more concentration of neutral points than that of Blair for people above 50 years of age.
* Hague’s party should focus more on the voter group of 30-50.
* Blair has a more chance of winning than Hague.

# 