

Machine Learning Group Assignment

**Gautam Kumar**

**Manish Rawat**

**Sharmila Sengupta Chowdhry**

**Tanay Tilak**

**Tushar Bansal**

**Table of Contents**

Contents

[Q1. Read the dataset. Do the descriptive statistics and do null value condition check.](#_heading=h.30j0zll) 3

[Q2. Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers. Interpret the inferences for each](#_heading=h.1fob9te) 4

[Q3 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)](#_heading=h.3znysh7) 12

[Q4. Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models](#_heading=h.2et92p0) 13

[Q6. Model Tuning, Bagging and Boosting.](#_heading=h.tbhhjl69g6zt) 23

[Q7. 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 all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized](#_heading=h.9oxfccrrprtb) 29

[Q8. Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.](#_heading=h.4d34og8) 30

*PROBLEM 1: MACHINE LEARNING – ELECTION DATA*

## Q1. Read the dataset. Do the descriptive statistics and do null value condition check.

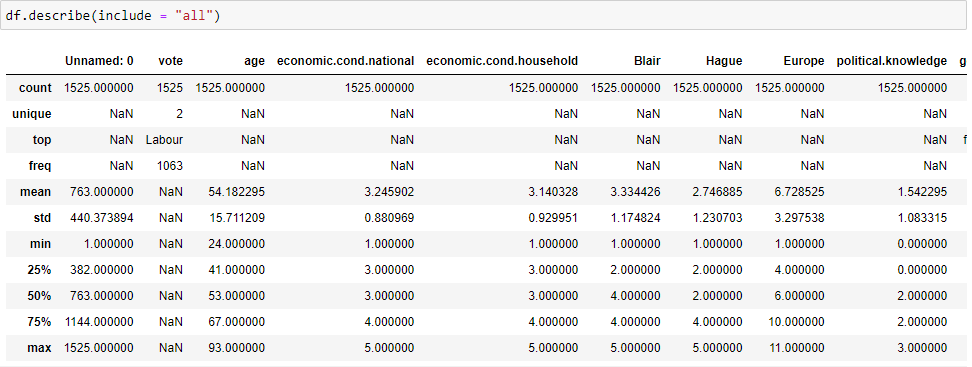
Read the data, checking the header and tail data to validate that the data is uploaded correctly.

Please refer to below the primary observations on the Election data:

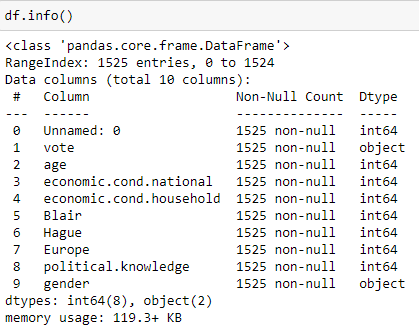
* 1. **Read the data file of Election data**



* 1. **Descriptive statistics**



* 1. **Null value condition check**



**Observation**

Data is successfully read and there are no missing values in the data. We will drop the ’Unnamed’ column, as it is not useful for description or modelling. ‘Vote’ is the target variable, divided into Conservative and Labour party votes, and it is an object data type. Apart from Age, all other integer data types are ratings of one form or another - 'economic.cond.national', and 'economic.cond.household' (both ratings of voter opinions), political knowledge, Europe, Hague and Blair are integer data types. From the above data description we can see that, the mean age of voters is 54 years, 75% voters are below the age of 67 and the maximum age of voter is 93 years. The distribution is very close to normal. Blair's average ratings (at 3.3) are higher than Hague's (2.7). In terms of percentiles, 50% of Blair's ratings are 4 or under, whereas 50% of Hague's ratings are 2 or under. Most voters are female, although the distribution between genders is close to equal.

## Q2. Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers. Interpret the inferences for each

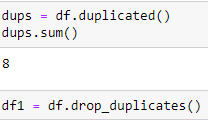
* 1. **Rename columns:**
  + economic.cond.national : eco\_nn
  + economic.cond.household: eco\_hh
  + political.knowledge: pol



* 1. **Drop the unnamed column**



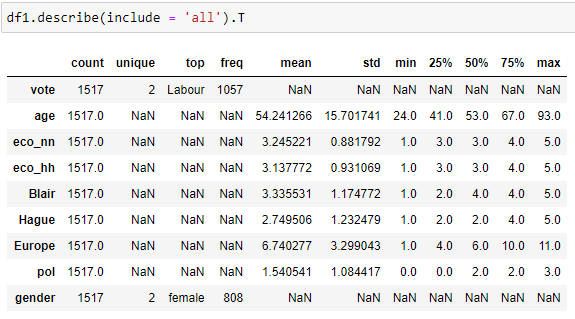
* 1. **Checking for the duplicates and dropping them**



**Observation**

We removed 8 duplicate records. We shall look at the descriptive analysis again.

* 1. **Descriptive statistics after dropping the duplicated records**



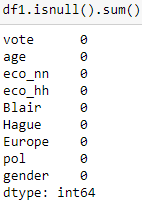
**Observation**

From the above descriptive statistics we see that:

* The mean age of voters is 53 years and 25% is between 24 and 41 years of age. 75% voters are below the age of 67 and the maximum age of voter is 93 years. The distribution is slightly right skewed.
* Blair's average ratings (at 3.3) are higher than Hague's (2.7). In terms of percentiles, 50% of Blair's ratings are 4 or under, whereas 50% of Hague's ratings are 2 or under.
* Most voters are female (812), although the distribution between genders is close to equal.
* Two political parties fighting to get power. Labour Class is the majority class with 1057 records
  1. **Checking for shape indicate that there are 1517 indexes/rows and 9 column**



* 1. **Checking for null values again**

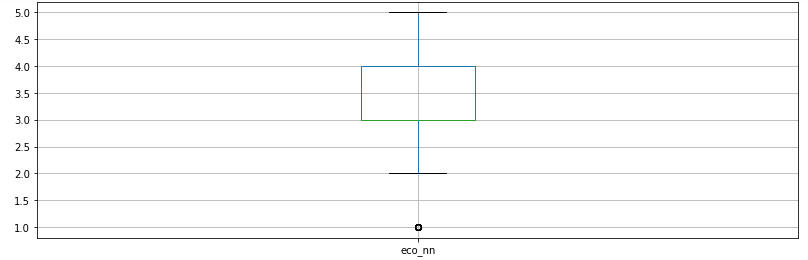
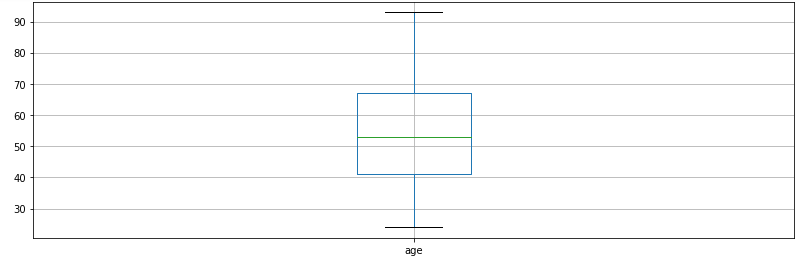


**Observation**

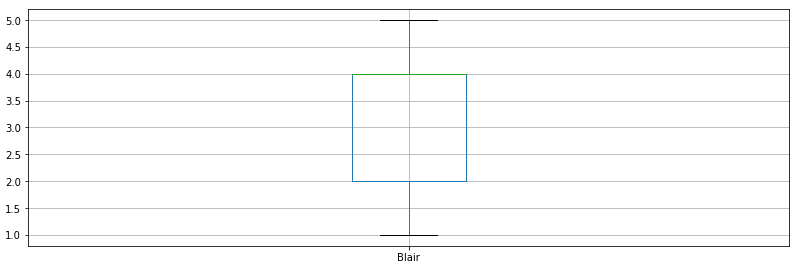
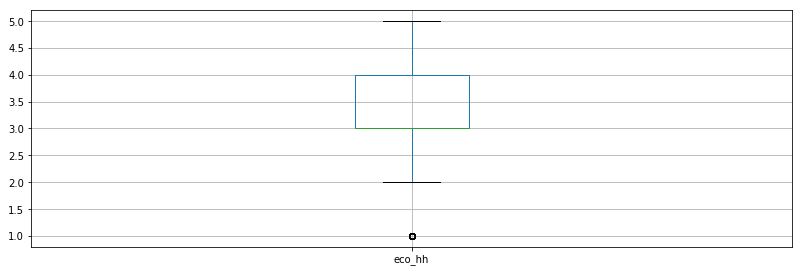
No missing values in the data set.

* 1. **Boxplot creation to check for the outliers**

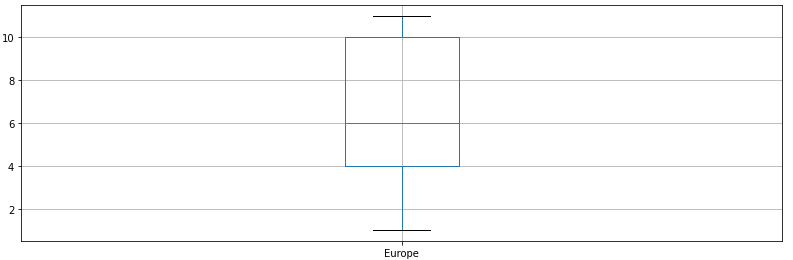
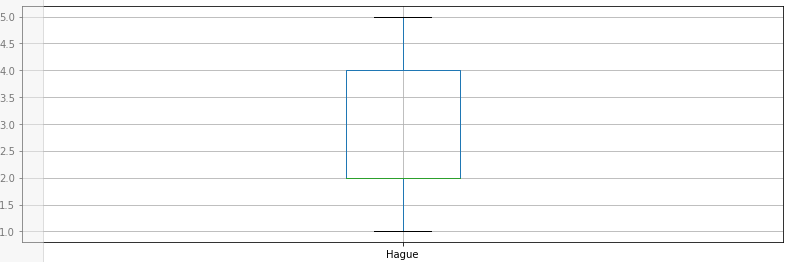
Age Eco\_nn

****

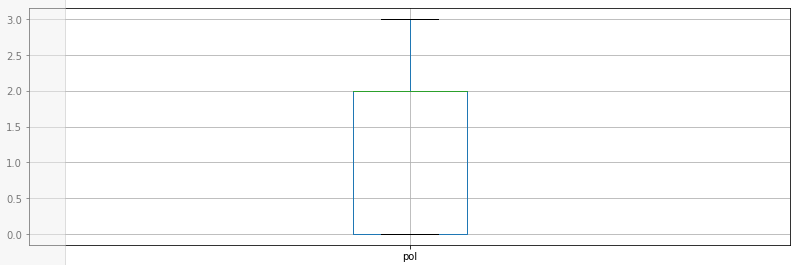
Eco\_hh Blair

****

Hague Europe



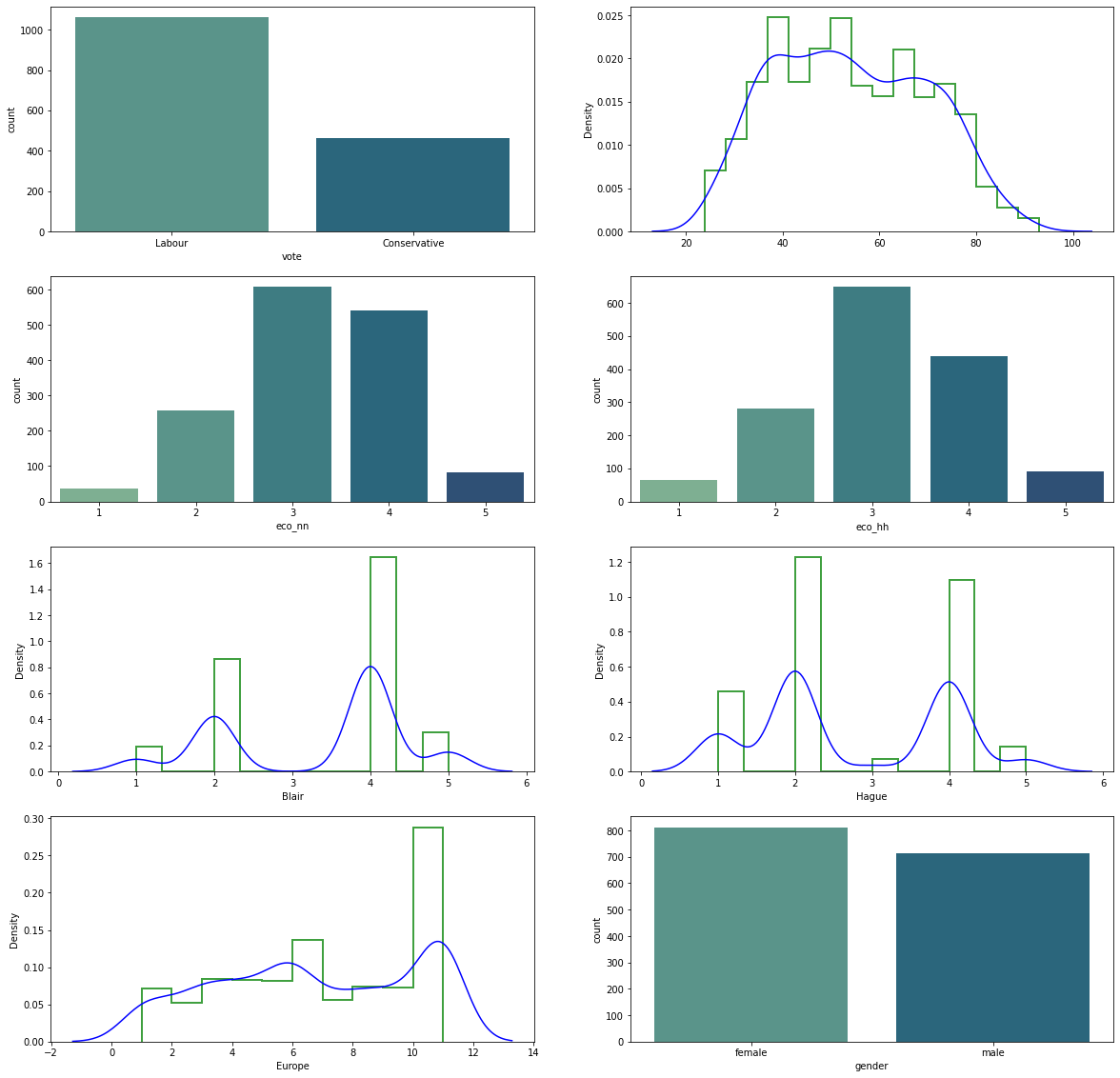
**Pol**

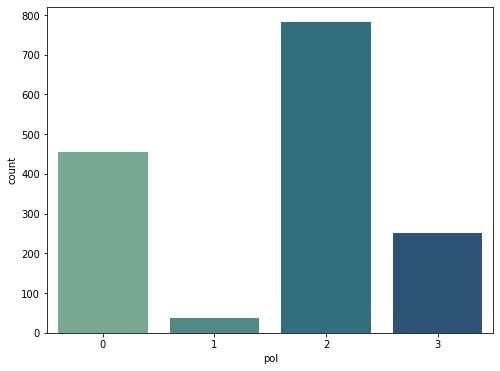


**Observation**

From the above box plots we can infer that there are outliers in the eco\_nn and eco\_hh ratings. However, since they are within the rating scope, we should not treat them as outliers.

* 1. **Univariate, Bi-variate and Multi-variate Analysis**
     1. **Univariate Analysis**

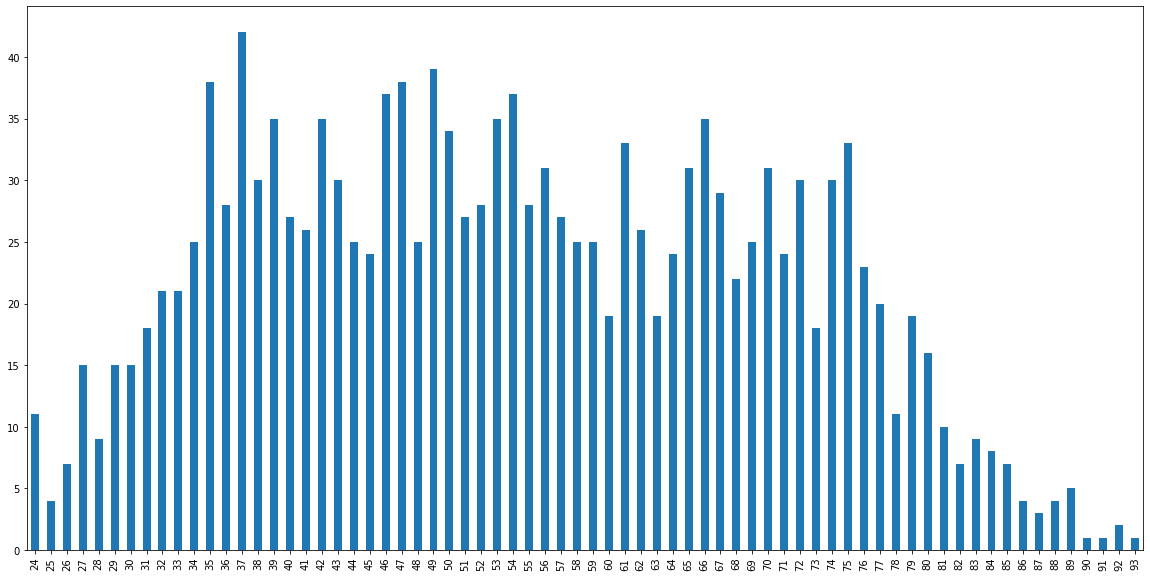




**Observations**

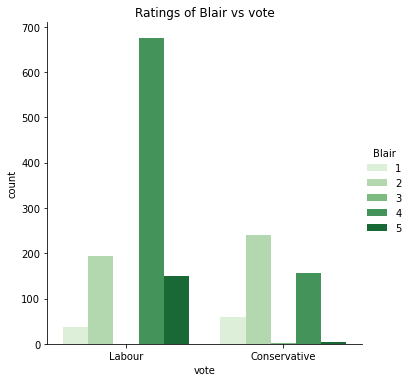
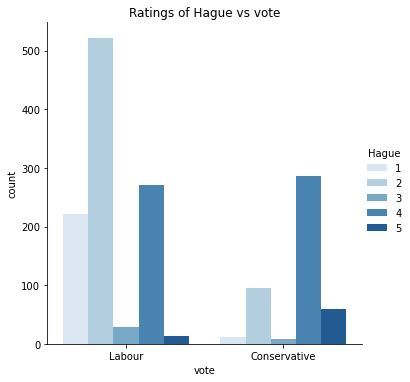
* As seen earlier, Labour is the Majority Class
* Minor Skewness in Age Distribution can be seen
* Eco\_nn and Eco\_hh, the parameter for national and household economy ratings, have similar distributions. Most voters rate these two as 3, followed by 4.
* Blair and Hague follow the similar pattern. First the graph increases from 1 to 2 but decreases from point 2 to point 3. And then again increases from 3 to 4. Voters are fairly decisive in whether they like or dislike the political leaders.
* Europe (Euroscepticism) ratings have highest no. of records on the range of 10-11, depicting that most voters do not favour the idea of the EU.
* A significant no. of people (more than half) do not know the political parties’ position on EU integration.

**Age Distribution Plot**

****

* **Observations:**
* Majority of the Data is between 35-75yr
* Low no. of data points after age of 85
  + 1. **Bivariate Analysis**

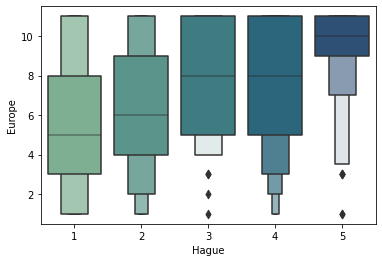
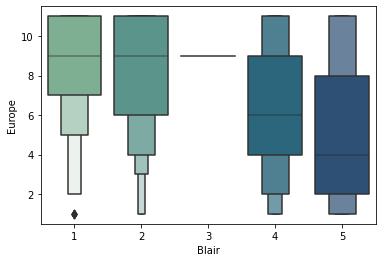
**Rating of Blair vs vote Rating of Hague vs vote**

**Observations**

* **Rating of Blair vs vote:** This is a bifurcation of Blair's ratings by the classification of votes. As is expected, Voters who vote for the Labour party, rate Blair much higher than voters who vote for the Conservative party, with the dominant rating being 4 for Blair.
* **Rating of Hague vs vote:** This is a bifurcation of Hague's ratings by the classification of votes. As is expected, Voters who vote for the Conservative party, rate Hague mostly as a 4. However, Hague got more rating of 5 from Labour voters, than Blair (from Conservative voters).

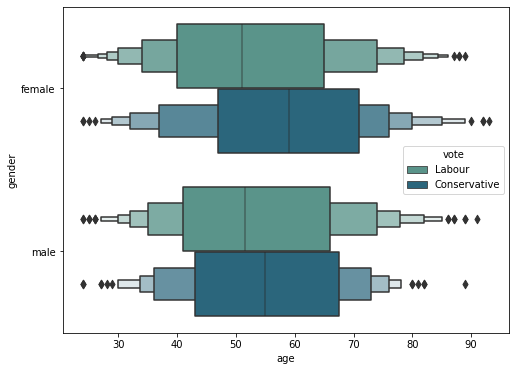
**Rating of Blair vs Europe Rating of Hague vs Europe**

**Observations**

* The boxenplot throws a lot of light on ratings of each of the candidates and how their voters feel about being a part of EU (Europe scepticism, denoted here as Europe). Most voters who rate Hague highly are Eurosceptic.
* Most voters who rate Blair poorly are Eurosceptic. Voters who rate Blair above 4, are less critical of EU.

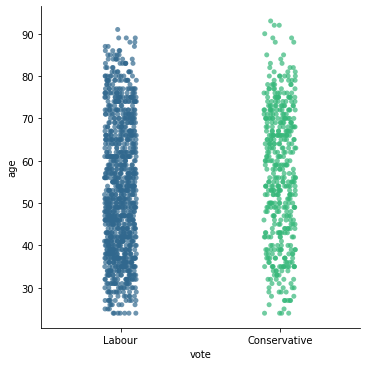
**Age vs Gender**

****

**Observations**

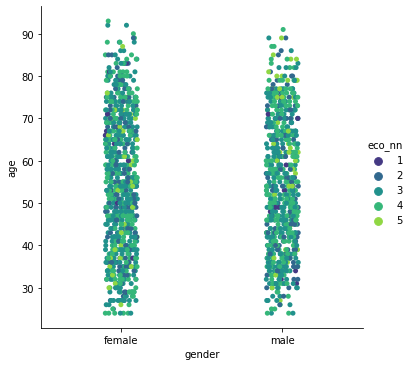
* It can be seen from the graph that outliers can be seen on both sides of the Boxenplot for both genders and both parties. However, these are within the scope of acceptable range, so no action is required.
* The median age of Conservatives Party is higher than the Labour Party in both Genders, the difference of median ages is more pronounced for female voters.

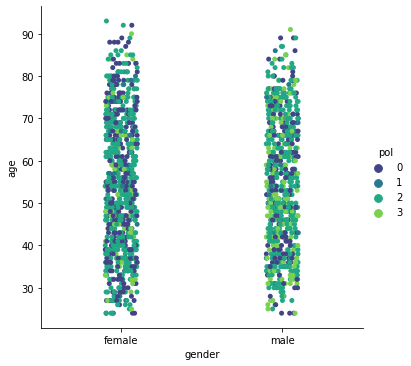
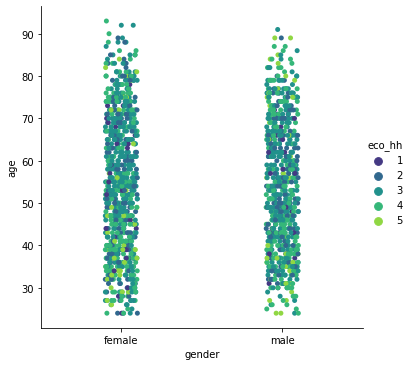
**Age vs vote**



**Observations**

* In case of Conservative Party, it can be seen that the data is more scattered on the lower age (below 45 Years). While data is more compact for the age between 50-75yr.
* In case of Labour the data is compact throughout and it is getting scattered on the higher age.
  + 1. **Multivariate Analysis**

**Gender vs Age vs pol Gender vs Age vs eco\_nn Gender vs Age vs eco\_hh**

**** 

**Observations**

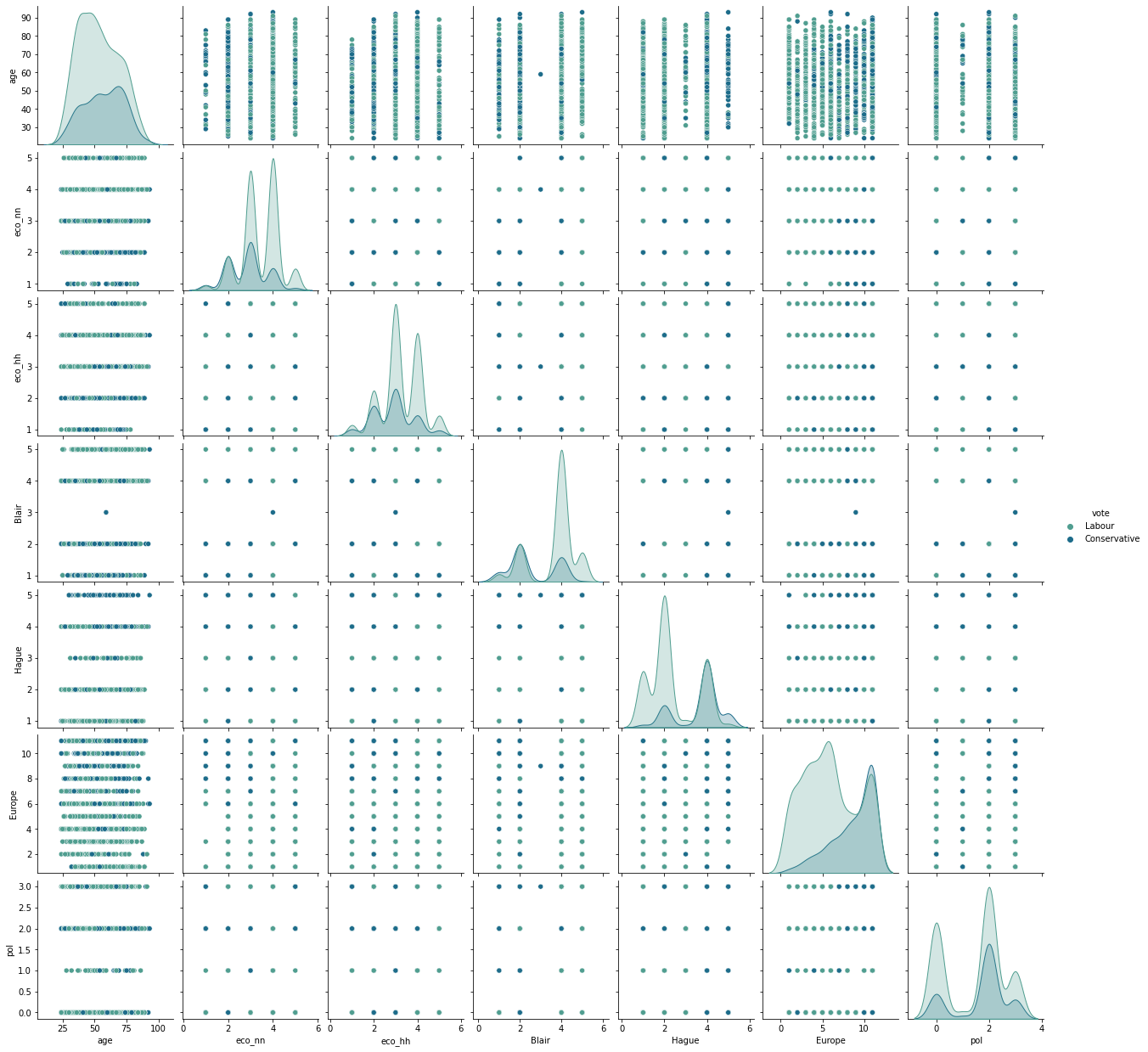
* Male voters have more information about their political party’s EU agenda. Female voters across all age groups have little or no information about their political party’s position on EU.
* Middle aged voters are rating eco\_hh lower than older and younger voters.
  1. **Correlation Matrix**



**Observation**

From the above heatmap we can infer that the Independent variables are Independent of each other. The highest correlation is 35% between eco\_nn and eco\_hh. A higher rating for Blair accompanies a higher eco\_nn rating, by a weak 33%.

* 1. **Pairplot**



**Observation**

The most telling observation from the pairplot, which is not as clearly covered by other plots, is that Europe (Euroscepticism) divides the voters between Labour and Conservative parties more clearly than any other variable.

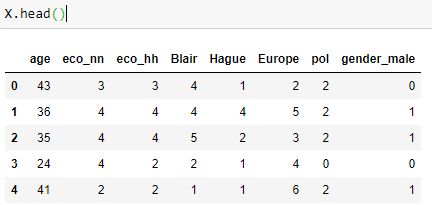
## Q3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30)

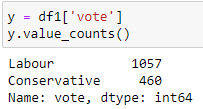
* 1. **Encoding the data**

We are encoding the data keeping the ‘vote’ variable out of the dataset and saving in a new dataframe X and saving ‘vote’ variable in y.







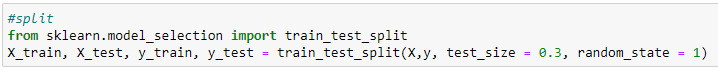


* 1. **Is scaling necessary here or not?**

**Observation**

We scaled the data for a version of the KNN model (KNN2) but the impact on the model performance was not much. Also, the scale of Independent variables are similar in range. Scaling will also change the values of the ordinal variables, and given the marginal value scaling is offering us in this case, scaling is not necessary.

* 1. **Splitting Training & Test data (70:30)**

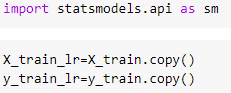


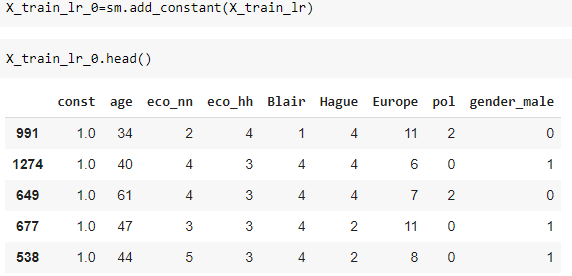
**Observation**

While splitting the data into (70:30) ratio train data comprise of 1061 rows of records and test data comprise of 456 rows of records.

## Q4. Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models

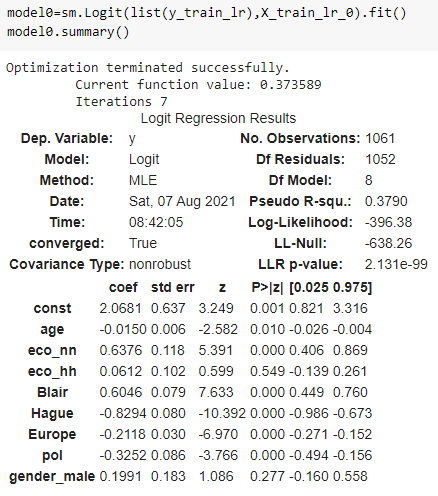
1. **Logistic Regression Model**
   1. **Model preparation:**





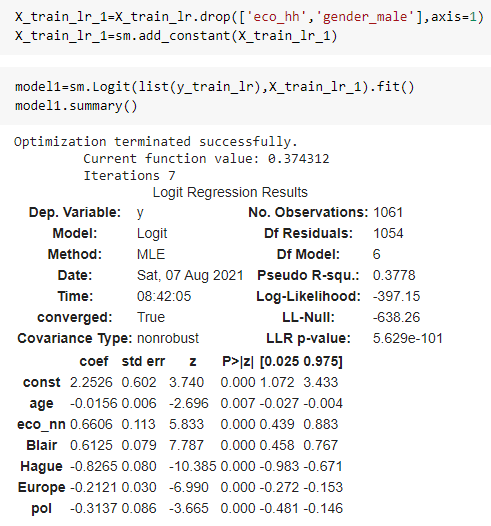
* 1. **Model Building:**

**4.2.1 Model 0:**



**Observation:** variable “eco\_hh” and “gender\_male” seem to be insignificant

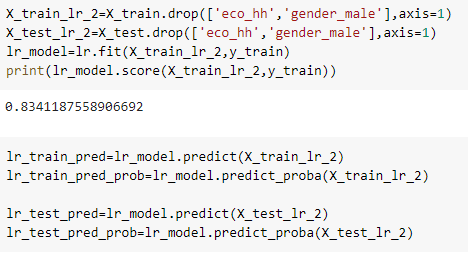
**4.2.2 Model 1: After dropping insignificant variables: eco\_hh and gender\_male**



**Observation**

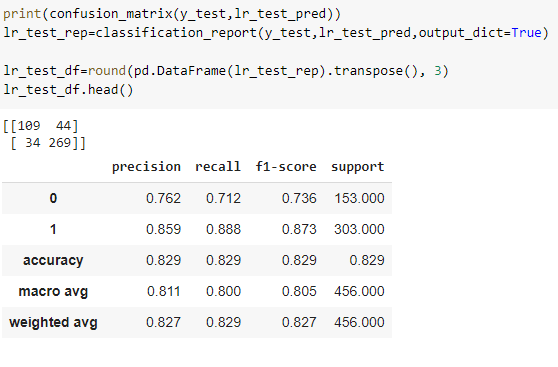
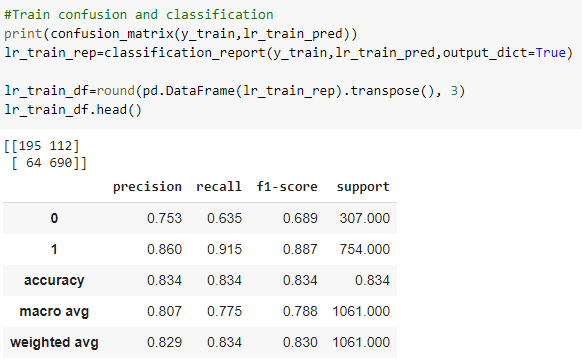
There is no insignificant value in the Model 1, hence we will be using Model 1 for prediction.

* 1. **Fit the model and do the prediction**



* 1. **Building confusion matrix**





**Observation**

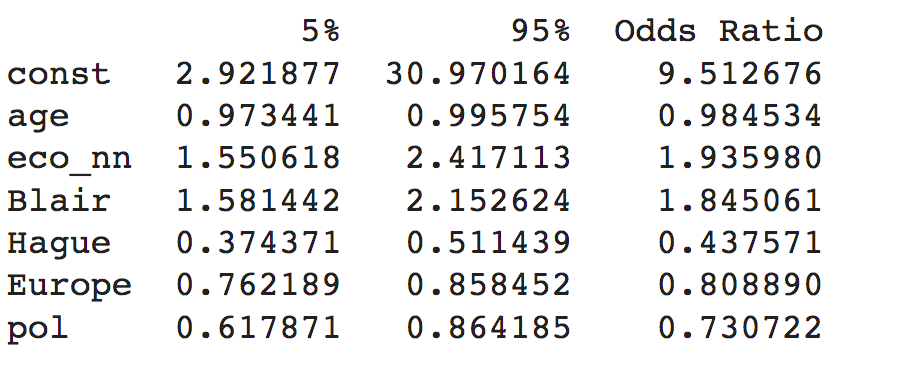
**Inference on Train Data:**

* + While building the Logit model on the full dataset, we found out that Eco\_hh and Gender\_male seem to be an insignificant variable (Based on p-value). Thus we dropped both the variables while developing the Logit Final Model.
  + The Logit Train Model is able to predict with 83% accuracy.
  + Precision for Labour Party is 86% implying that of all the predictions of Labour voters, 86% of the time our model predicted it correctly.
  + The Recall for Labour Party is 91.5% implying that out of all who actually voted for the Labour Party, our model identifies them 91.5% of the time as Labour voters.
  + For Conservatives, the recall is 63.5%, while the precision is 75.3%, implying that the model is leaving out 36.5% Conservatives votes and misclassifying 24.7% Labour votes as conservatives.

**Inference on Test Data:**

* + The Logit Test Model is able to predict with 82.9% accuracy, which is fairly good.
  + Precision and recall for Labour are 85.9% and 88.8%, respectively.
  + For Conservatives votes, Precision and recall are 76.2% and 71.2%, respectively, better than train performance.

The Logit model has low variance. Logistic Regression has the great benefit of explanation of how features can influence the target or dependent variable. Below is the odds ratio and its interpretation:

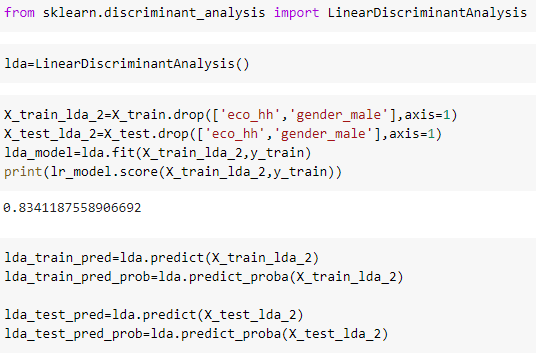


**Interpretation**

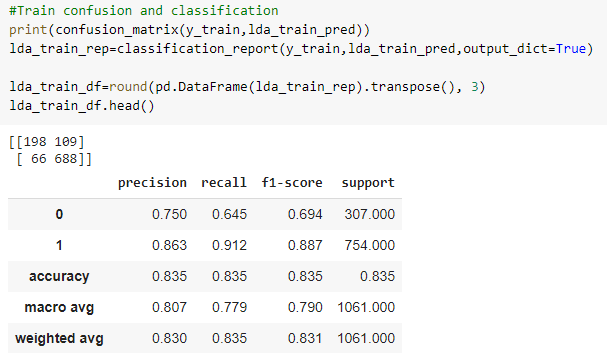
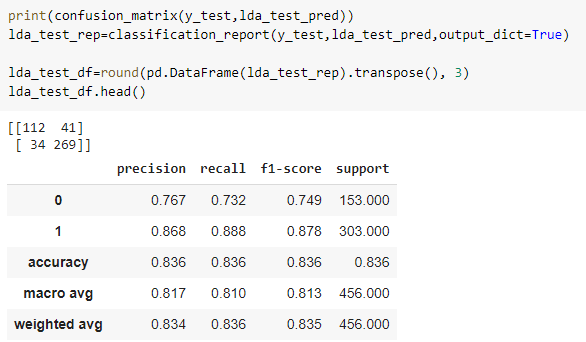
* + eco\_nn: With every one unit increase in eco\_nn, a voter has a 93.5% more likelihood of voting for the Labour party.
  + Blair: With every one unit increase in rating for Mr. Blair, the voter is 84.5% more likely to vote for the Labour party.
  + Hague: With every one unit increase in rating for Mr. Hague, the voter is 56.3% less likely to vote for the Labour party.
  + Europe: With every one unit increase in Eurosceptic rating, the voter is 19.2% less likely to vote for the Labour party.
  + pol: With every one unit increase in rating for political party's EU agenda, the voter is 27% less likely to vote for the Labour party. A higher rating means more Eurosceptic. More Eurosceptic voters are likely to vote for Conservative party.

Let us now see how the LDA model performs.

1. **LDA Model**
   1. **Model building and predicting train and test**



* 1. **Confusion Matrix**

**Observation:**

**Inference on Train Data:**

* + On Train data, the model is 83.5% accurate, comparable to the Logit accuracy.
  + The model is 86.3% precise (precision) on the Labour votes it predicts (86.3% of the Labour votes identified by the model are correct), while a high 91.2% of all actual Labour votes are correctly identified by the model during training (recall).
  + 75% of Conservative votes identified by the model are correct (precision), while 64.5% of all Conservative votes are correctly identified by the model (recall).

The Train LDA performance is very close to the Train Logit performance. Let us evaluate the performance of the test data on the LDA model.

**Inference on Test Data:**

* + The accuracy of the test data is 83.6%, proving that the model does not have variance.
  + Precision and Recall of Labour votes are 86.8% and 88.8%, respectively, while for Conservatives, the same parameters are 76.7% and 73.2%, respectively.

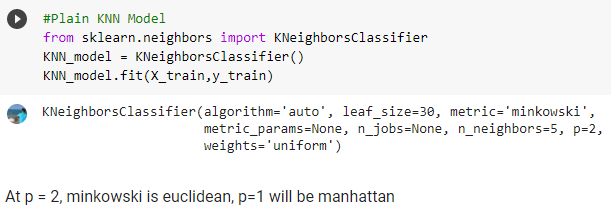
The LDA model performance is slightly better than the Logistic model. Given a choice we should deploy the Logistic Regression model due to its explanatory power.

Q5. Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model.

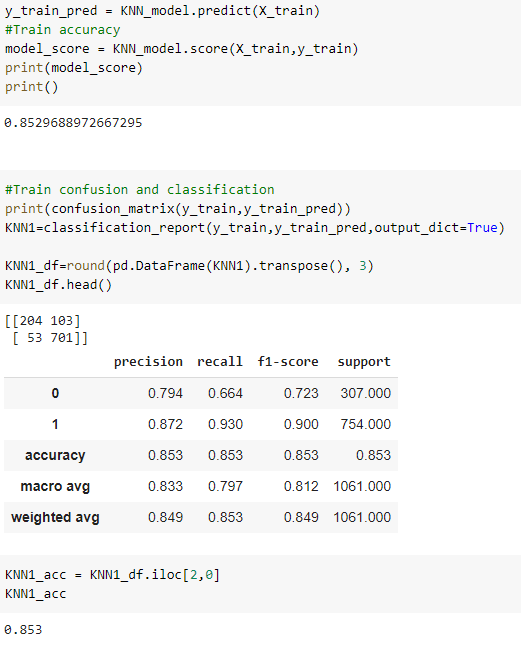
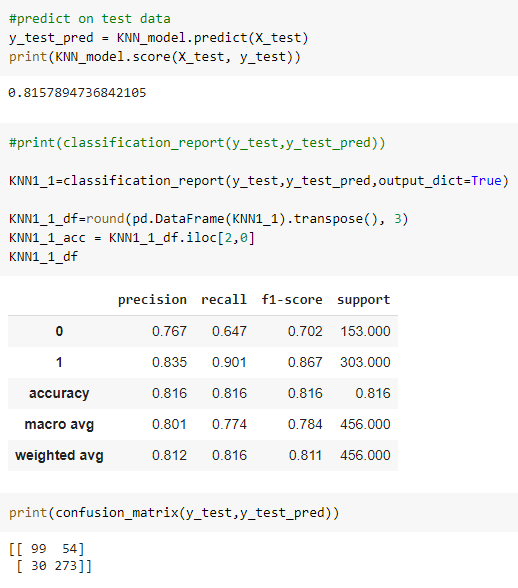
KNN has a lot of advantages: It is quick for a small dataset like ours. It is a simple algorithm with high levels of accuracy possible. No assumptions about the data are necessary.

We start with a base KNN Model. The subsequent variations of KNN has scaled data, one with an optimum K value and a KNN with parameters selected with a Grid search algorithm.

1. **KNN Model**
   1. **KNN Model 1**



KNN Model 1 is the base KNN Model. We see the train and test performance of this model below.

**Observation**

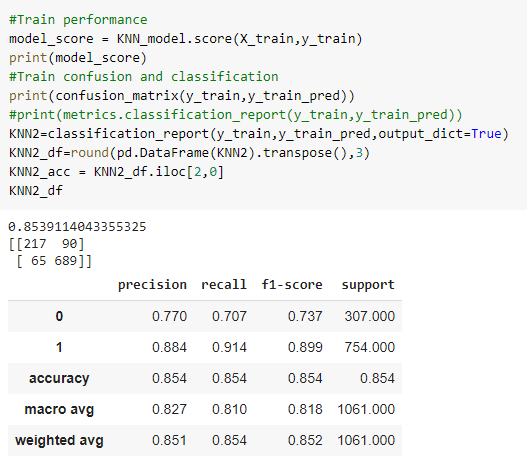
**Inference on Train Data:**

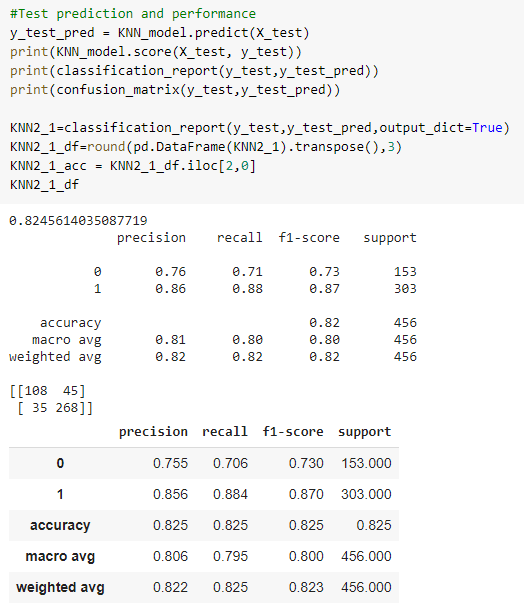
* + On Train data, the model is 83.5% accurate.
  + The model is 87.2% precise (precision) on the Labour votes it predicts and it predicts an impressive 93% Labour votes correctly (recall).
  + 79.4% of Conservative votes identified by the model are correct (precision), while 66.4% of all Conservative votes are correctly identified by the model (recall).

**Inference on Test Data:**

* + On Test data, the model is 81.6% accurate.
  + The model is 83.5% precise (precision) on the Labour votes it predicts and it predicts an impressive 90.1% Labour votes correctly (recall).
  + 76.7% of Conservative votes identified by the model are correct (precision), while 66.7% of all Conservative votes are correctly identified by the model (recall).
  1. **KNN Model 2 : Standard Scaler**

Although our data does not have the kind of range that will impact the model, since KNN is known to get affected by distance of the data spread, we try another iteration of the model by scaling the data and check its performance.

**Train prediction and performance Test prediction and performance**



**Observation**

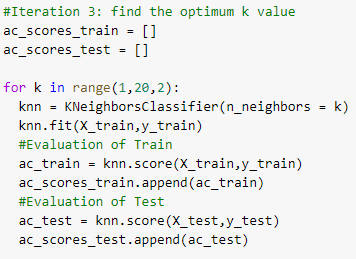
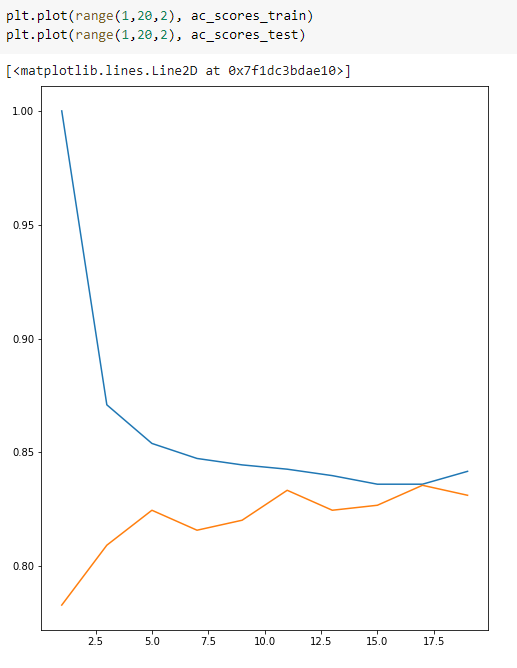
**Inference**

* + Scaling improved the test accuracy from 81.6 % to 82.5%
  + More Conservative votes are being correctly classified correctly by the model, with recall stepping up from 64.7% to 70.6%
  + Precision of Labour votes have increased marginally to 85.6% from 83.5% in the previous version of the model (KNN1).

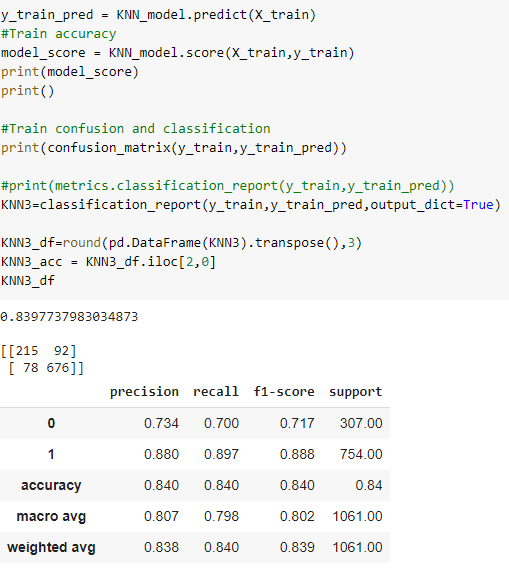
The default K-value being used is 5, which may not be appropriate. We iterate it once more by trying to find an optimum K value.

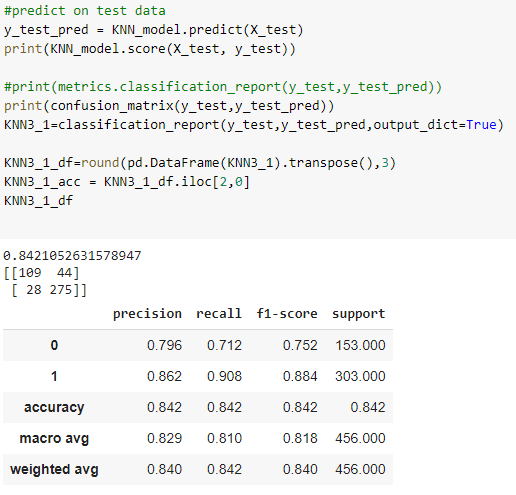
* 1. **Model 3 : Finding the optimum K values and creating the model**

Using multiple K values and iterating the same in KNN3 model and plotting the KNN3 scores of test and train to find the optimum K value.



Building the KNN3 model with optimum K value =16:

**Train prediction and performance Test prediction and performance**



**Observation**

On all performance parameters KNN3 is better than any other KNN model so far.

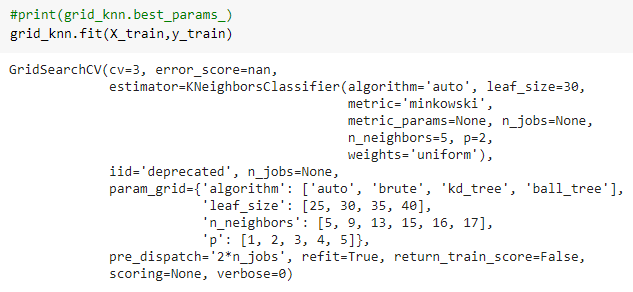
Accuracy is at 82.3% (as as against 82.3% (KNN2), 81.6% (KNN1)).

Precision for Labour party votes is 86.2% (as against 85.6% (KNN2), 83.5% (KNN1)), Recall for Labour party at 90.8% (as against 88.4 (KNN1), 90.1%(KNN2))

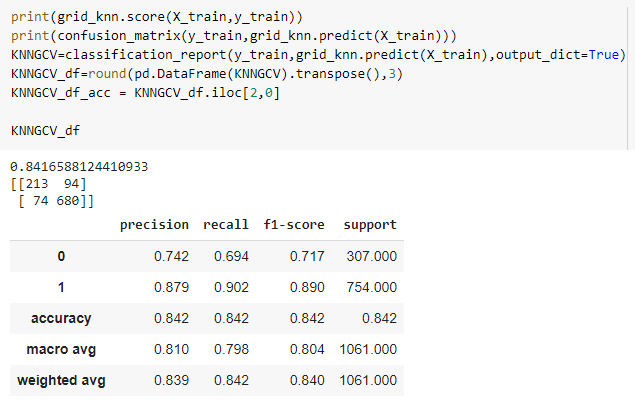
For Conservative party votes, the model predicts with 79.6% precision (as against 75.5% (KNN2), 76.7% (KNN1)) and 71.2% recall (as against 64.7% (KNN1), 70.6% (KNN2))

Based on the above 3 KNN models, it seems all the model are performing fairly good. However, if the business has to choose anyone one model we would recommend to use KNN3 model which is performing fairly well for both the parties i.e.; 'Labour' & 'Conservative'

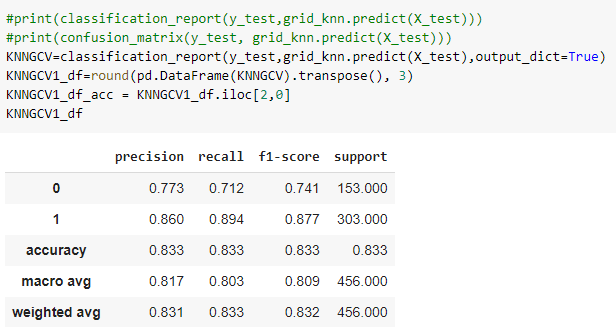
* 1. **Model 4 : Model having best parameters using Grid Search**



**Train prediction and performance**



**Test prediction and performance**



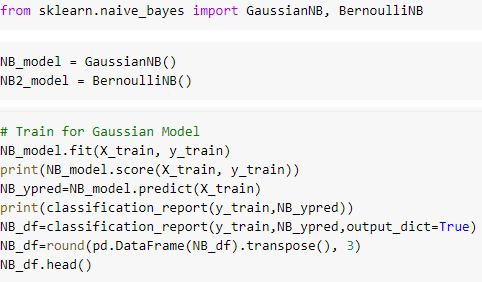
**Observation**

Gridsearch CV didn’t show any improvement, because the best params came out to be the same as the previous iteration.

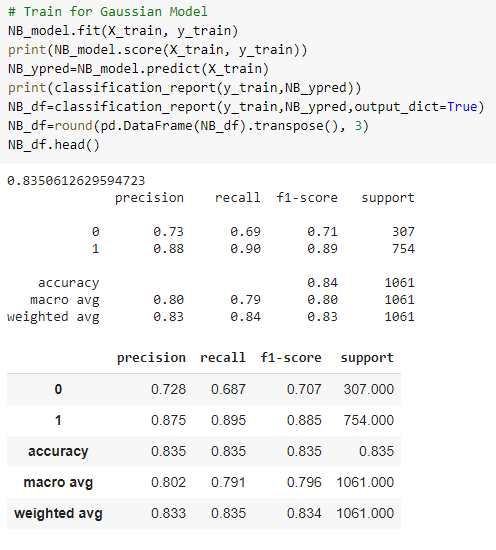
1. **Naïve Bayes**

Gaussian Naive Bayes is supposed to work better with continuous variables and Bernoulli Naive Bayes, which is more suited for binary variables. We have a mix of discrete and multinomial values in our dataset.

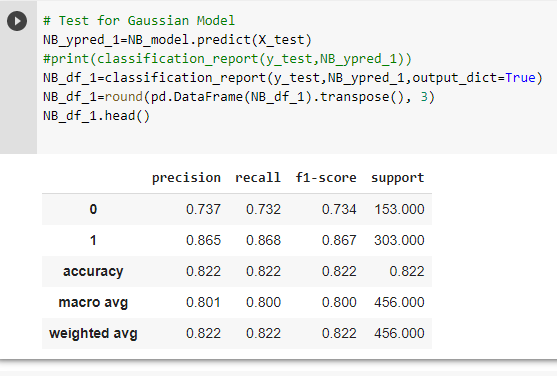
* 1. **Model 1 – Gaussian Model**



**Train for Gaussian Model**



**Test for Gaussian Model**



**Observation**

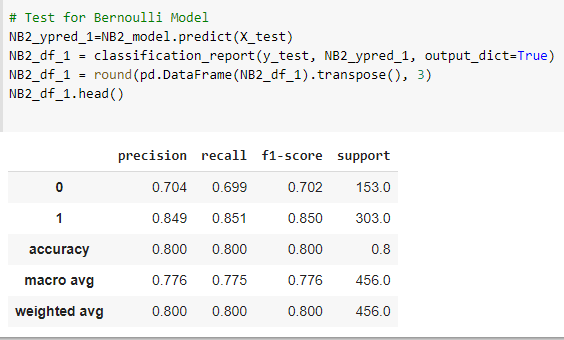
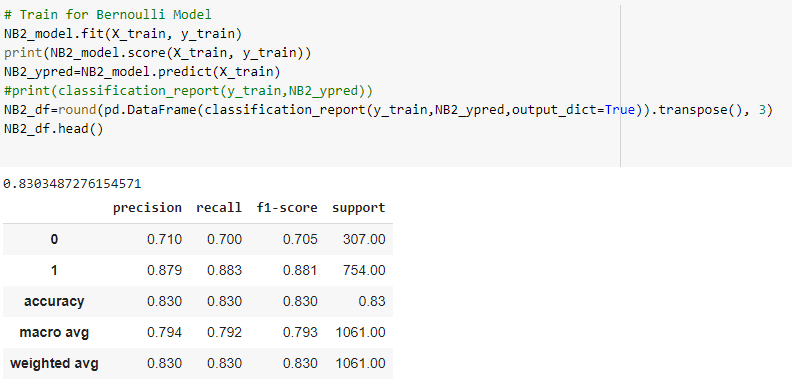
**Inference on Train Data:**

* + Accuracy is 83.5%
  + Labour vote precision is 87.5% and recall is 89.5%, as against Conservative vote precision at 72.8% and recall at 68.7%.

**Inference on Test Data:**

* + Accuracy is 82.2%
  + Labour vote precision is 86.5% and recall is 86.8%, as against Conservative vote precision at 73.7% and recall at 73.2%.
  1. **Model 2 – Bernoulli Model**

**Train for Bernoulli Model Test for Bernoulli Model**



**Observation**

**Inference on Train Data:**

* + Accuracy is 83%
  + Labour vote precision is 87.9% and recall is 88.3%, as against Conservative vote precision at 71% and recall at 70%.

**Inference on Test Data:**

* + Accuracy is 80%
  + Labour vote precision is 84.9% and recall is 85.1%, as against Conservative vote precision at 70.4% and recall at 69.9%.

We can see from the above two explanations that the Gaussian Naive Bayes Model is working better than the Bernoulli Model.

The KNN3 Model is the best model in this section.

## Q6. Model Tuning, Bagging and Boosting.

* 1. **Boosting**

Boosting algorithms combine low accurate classifiers to create a highly accurate classifier. Combining a set of weak classifiers, gives boosting algorithms robustness to variance, thereby they offer more stability in deployment..

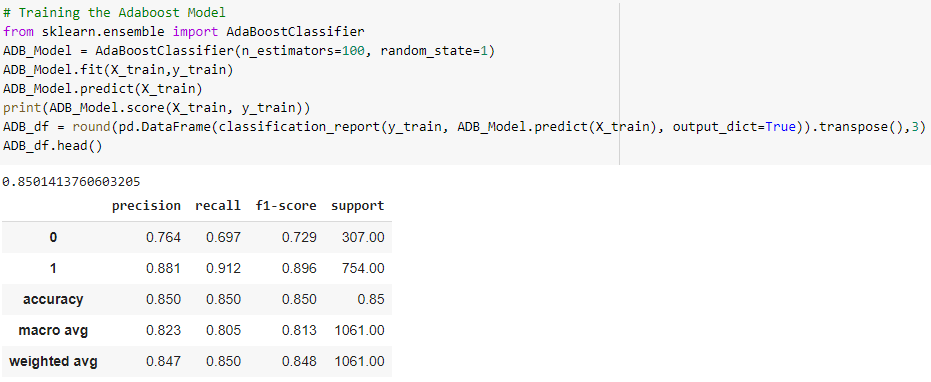
**6.1.1 Adaboost Classifier**

AdaBoost is an iterative ensemble method that uses multiple classifiers to increase the accuracy of classifiers. Adaboost creates each iteration in sequence, after setting the weights of classifiers and training the data sample in each iteration to ensure the accurate predictions of unusual observations. Each iteration prefers the misclassified data from the previous classification.

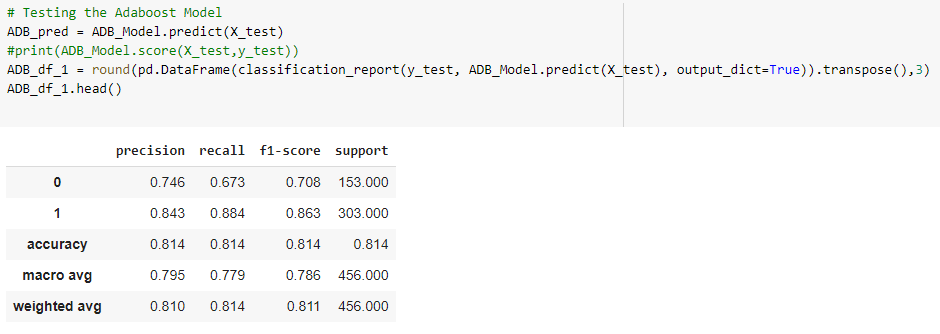
So, new weak learners are added sequentially to focus on training the model on the more difficult patterns, until the algorithm identifies a model that correctly classifies these samples.

It also assigns weights to classifiers, the higher the classification power, the higher the weight they get.

**Train prediction and performance**



**Test prediction and performance**



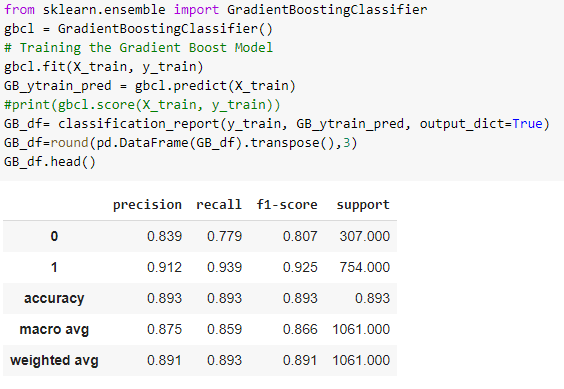
**Observation**

* + The Train Accuracy is 85%, compared to Test Accuracy of 81.4%.
  + Out of all who actually voted for Labour, 91.2% of them our model predicted correctly in train and 88.4% in test data.
  + Recall for Conservative class is 69.7% for train and 67.3% for test.
  + Of all the voters the model identifies as Labour, 88% are actually Labour voters in the training set. For test data, this drops to 84.3%.

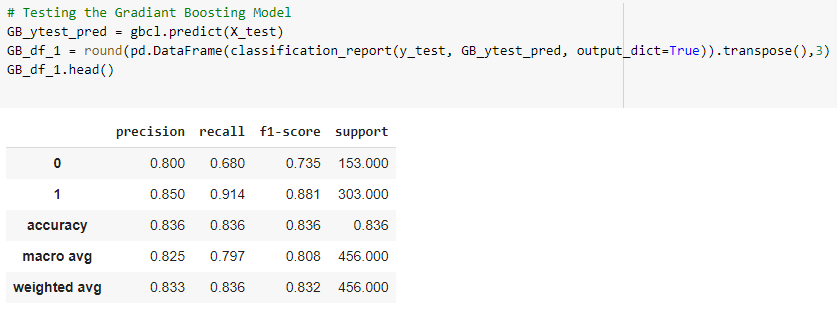
**6.1.2 Gradient boost Classifier**

While gradient boosting also adjusts based on incorrect predictions, it fits each new tree entirely based on the errors of the previous tree's predictions. Building a machine learning algorithm that solely focuses on the errors requires a comprehensive method that sums errors to make accurate final predictions. This method leverages residuals, the difference between the model's predictions and actual values.

**Train prediction and performance**



**Test prediction and performance**

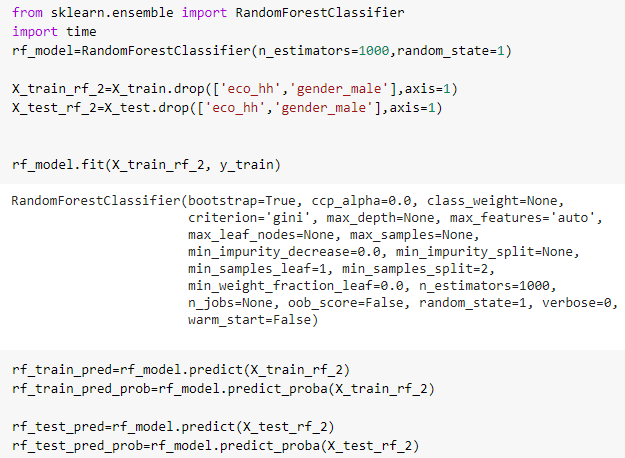


**Observation**

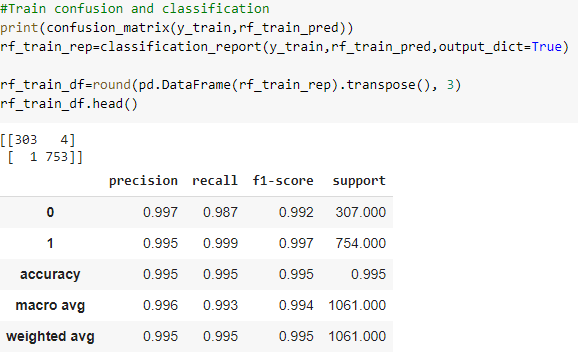
* + The Train Accuracy 89%, for Gradient Boost model is more than the test Accuracy (84%).
  + Out of all who actually voted for Labour (Recall), 85% of the time our model predicted correctly.
  + The test model identifies Conservative voters with a precision of 80%, and Labour voters with a precision of 85%.
  + Recall value of Labour vote prediction on test data is at 91.4%, the same for Conservative votes is 68% (depicting only 68% of Conservative votes classified as Conservative).
  + Precision for Gradient boost is higher than the Naive Bayes Models.
  1. **Bagging**

**6.2.1 Random Forest**

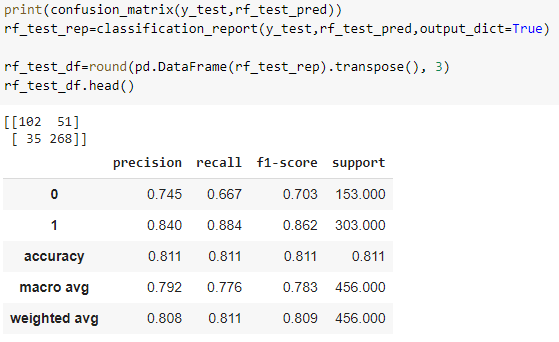
**Model building and predicting on train and test data**



**Train confusion matrix and classification report**



**Test confusion matrix and classification report**



**Observation**

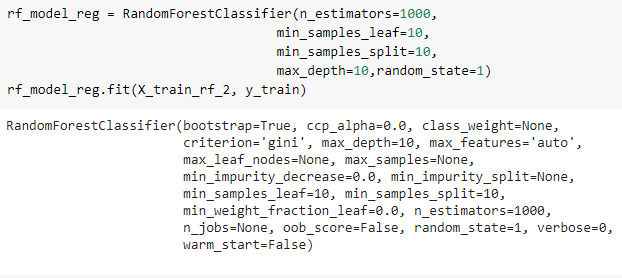
* + From the confusion matrix, we can see False Negatives and False Positives are very Low for train Data and because of that Accuracy and Precision is nearly 100%.
  + For Train as Model score (Accuracy), Recall and Precision is more than 99%.
  + As all the Metrics are very high for RF model, approx. 99% which can result in overfitting.
  + If we see the Metrics for Test Data and see all the metrics has dropped sharply (more than 20%) which proves that the model is overfitted, and has high variance.

**6.2.2 Regularize the RF Model through Grid Search**

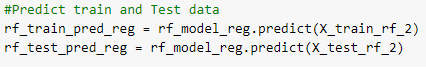
**Finding the best parameters**



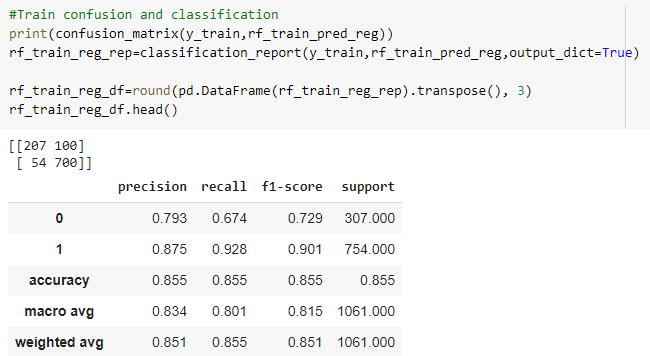
**Model creation using best parameters**



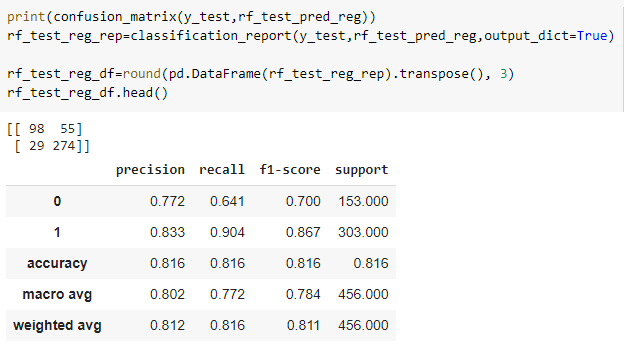
**Predict train and test data**



**Train confusion matrix and classification report**



**Test confusion matrix and classification report**

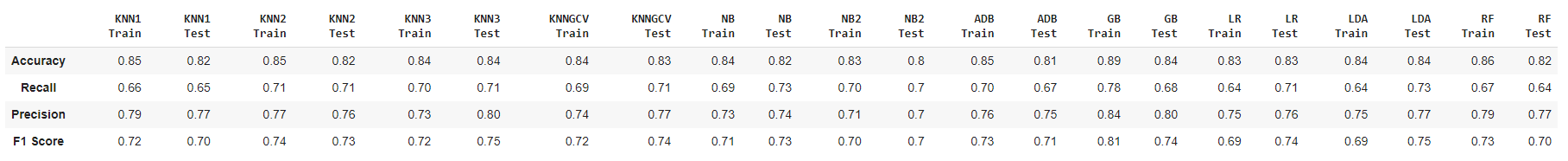


**Observation**

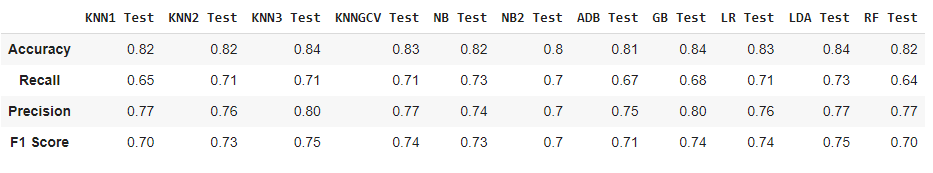
* + From the confusion Matrix, we can see False Negatives and False Positives have improved in the train Data and hence Accuracy and Precision has dropped down.
  + As all the Metrics have improved after passing the best parameters in the RF model, which to an extent addressed the issue of overfitting, hence the model is regularized.

## Q7. 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 all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized

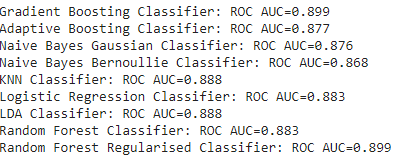
* 1. **Comparison of Train and Test of each models**

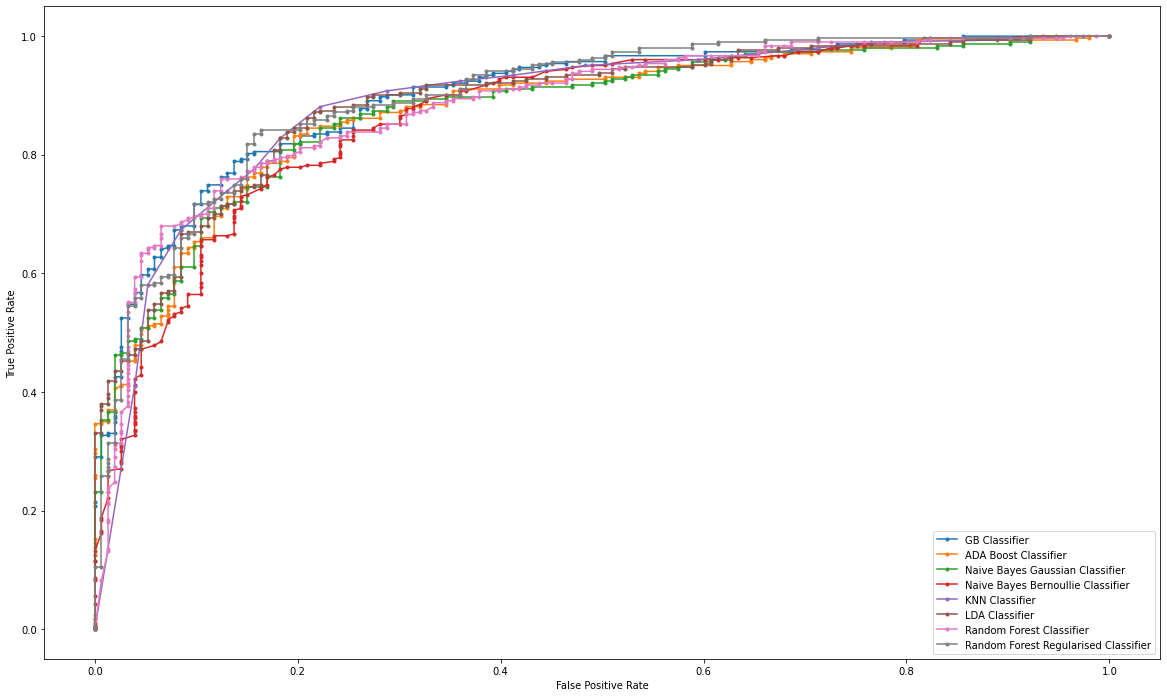


* 1. **Comparison of Final Models**



* 1. **ROC AUC**





**Observation**

Looking at all the model's test score KNN3 and LDA stands out in terms of Accuracy.

* + Both models are 84% accurate, for test and train, which is quite good
  + At 88.8% AUC Score is same for both the models
  + Test Recall for KNN3 is a bit lower than LDA (71% compared to 73%), but KNN3 has more comparable train and test figures
  + Test Precision for KNN3 is 80%, as compared to 77% for LDA, although the train results are a bit further away for KNN3 than LDA (73% and 75%, respectively)
  + F1 Score of the minority class (Conservatives) for Test is 75.2% for KNN3, as against 74.9% for LDA

Although, Gradient boosting classifier and Random Forest Regularized Classifier has the highest AUC score of 89.9%, given all the aspects, we recommend the KNN3 Model because of it proven performance as well as its simplicity and ease of implementation.

## Q8. Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

In all, we used the data provided to us to build 13 Machine Learning models, including regularized ones. To create an exit poll, we recommend using the KNN3 Model which has one of the best accuracy, precision and recall performance across train and test to arrive at a prediction of results.

Additionally, the logistic regression model helped us gain the following insights which might be helpful for the exit poll:

* + age: With each unit increase in age, there is a 2% decrease in likelihood of voting for the Labour party.
  + eco\_nn: With every one unit increase in eco\_nn, a voter has a 93.5% more likelihood of voting for the Labour party.
  + Blair: With every one unit increase in rating for Mr. Blair, the voter is 84.5% more likely to vote for the Labour party.
  + Hague: With every one unit increase in rating for Mr. Hague, the voter is 56.3% less likely to vote for the Labour party.
  + Europe: With every one unit increase in Eurosceptic rating, the voter is 19.2% less likely to vote for the Labour party.
  + pol: With every one unit increase in rating for awareness of political party's EU agenda, the voter is 27% less likely to vote for the Labour party.

\* \* \*