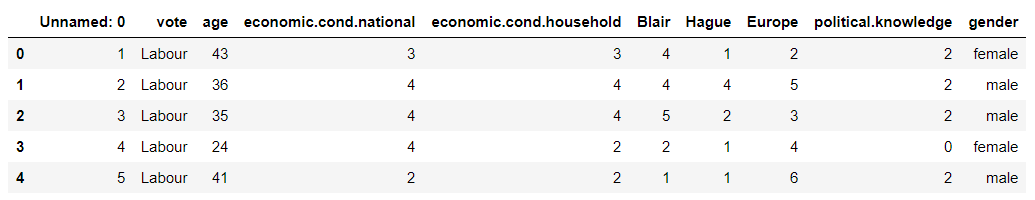
**Problem 1:**

To predict which party a voter will vote for given 1525 observations with 9 columns: vote, age, economic.cond.national, economic.cond.household, Blair, Hague, Europe, political.knowledge

The head of the data looks like this:

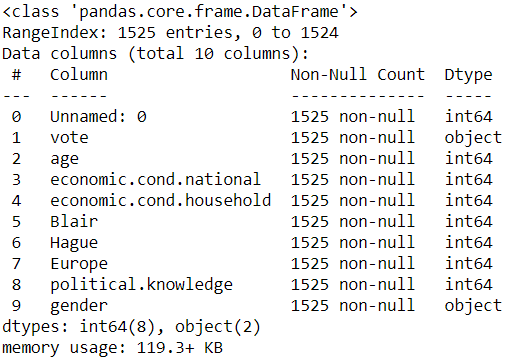


Here we see that there are 10 columns. The column: Unnamed: 0 is the voter number and is not to be considered for further analysis. This column will be removed in due course.

The shape of the data:



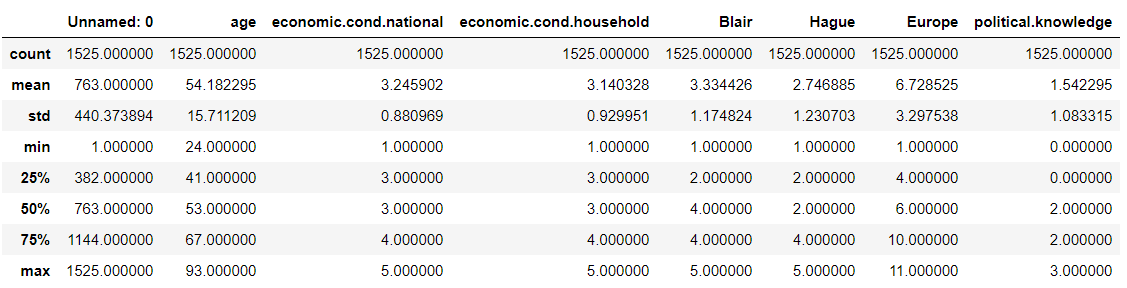
The info of the data:



Only vote and gender variables are of the object variable and the other columns are of the int64 variable.

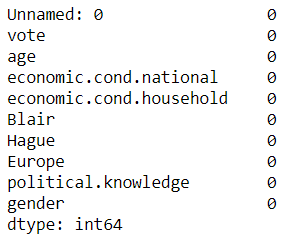
It is to be understood that the column: vote is the target column for our analysis.

The describe function of the data frame gives us the following output:

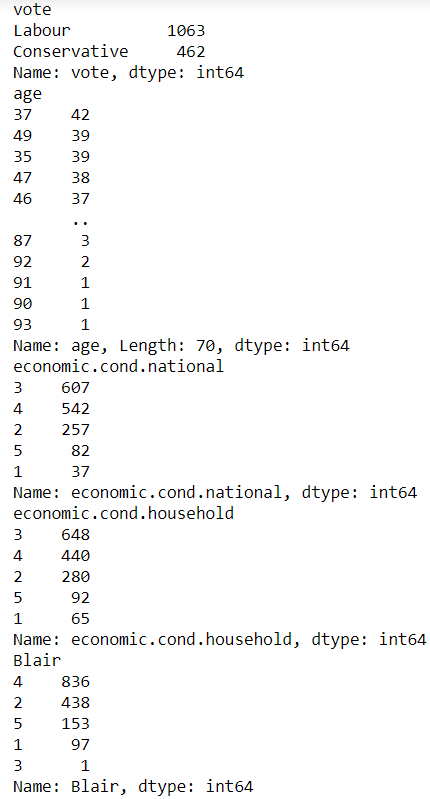


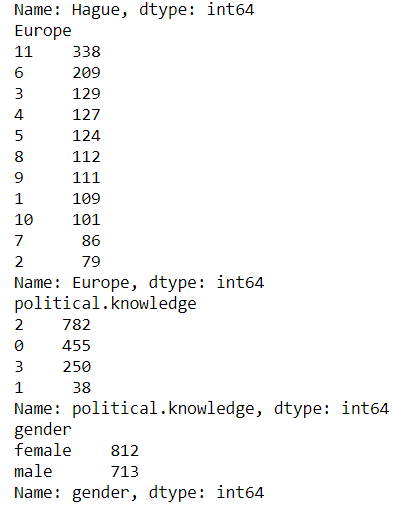
All the numerical variables except age are ordinal. The minimum of age is 24, the mean is 54 and the max is 93. The 50th percentile is almost coinciding with the mean and hence it can be concluded that the distribution is almost symmetrical.

We also see that there are no null values in the dataset:



We see that there are no duplicate records in the dataset:  
We then look at the value counts of all the columns:

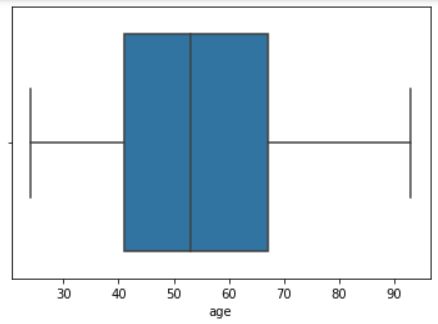


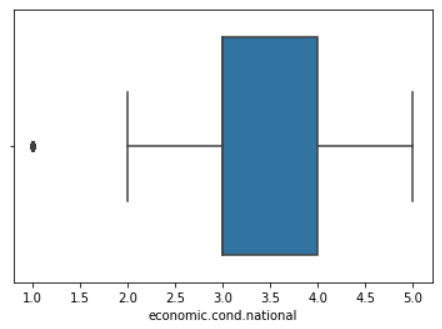


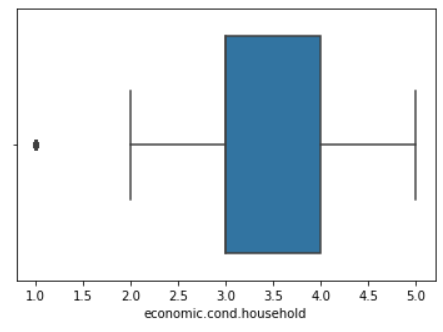
We also drop the Unnamed: 0 column from our dataset.

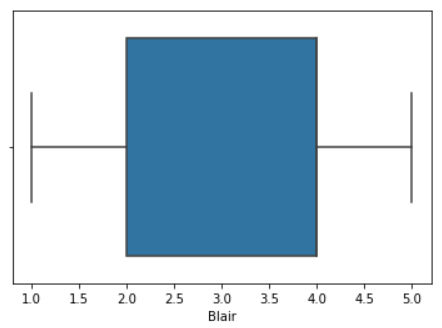
**Univariate & Multivariate Analysis:**

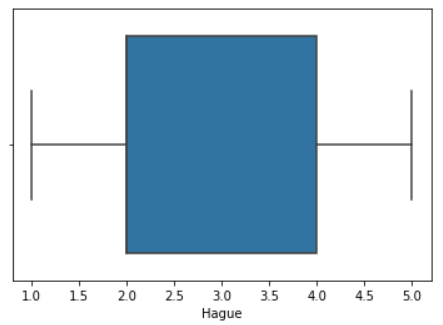
**Boxplot:**

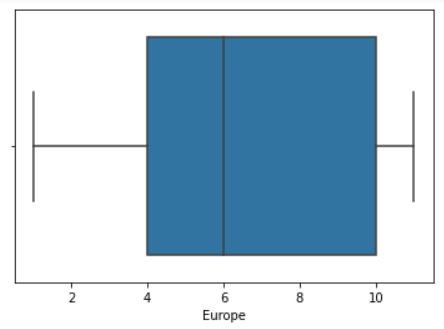


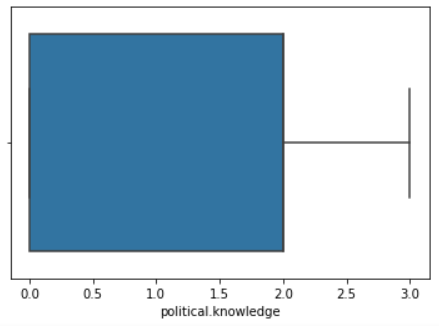






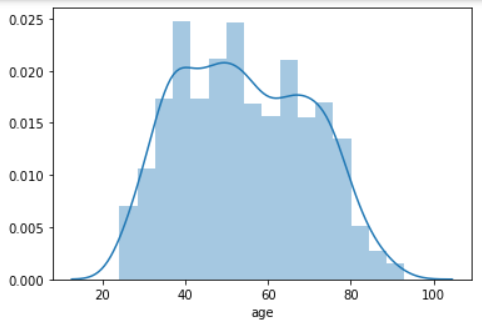


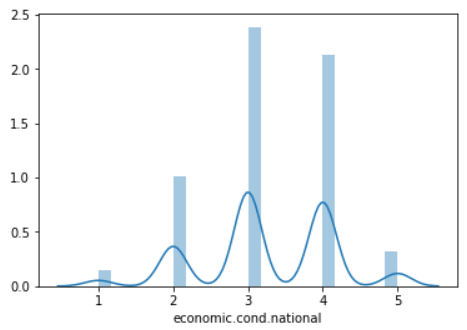


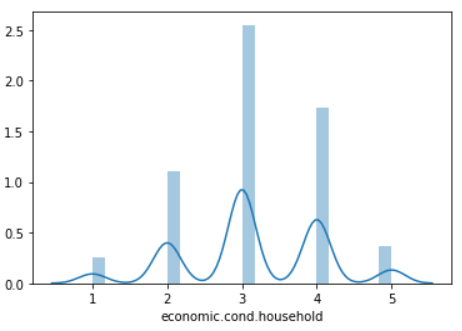


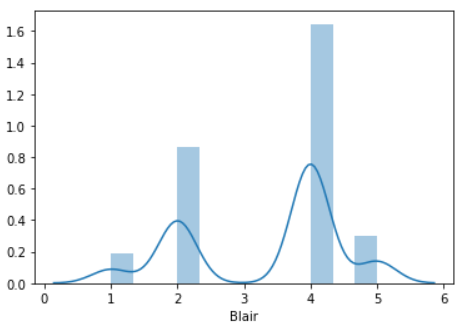
From the above plots, it can be seen that there are two outliers in the variables: economic.cond.national & economic.cond.household.

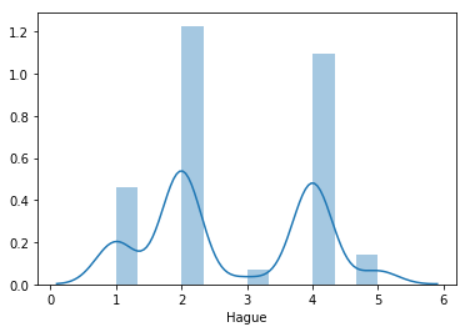
Let us also look at the distribution plot of each of the variables which gives us a histogram:

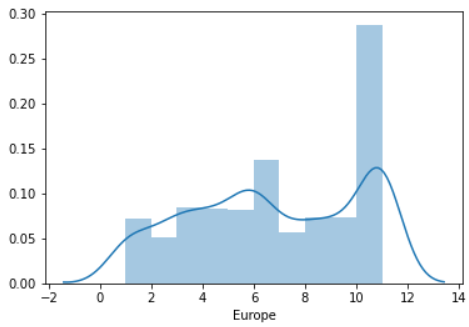


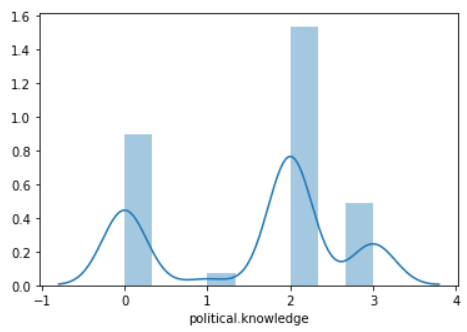




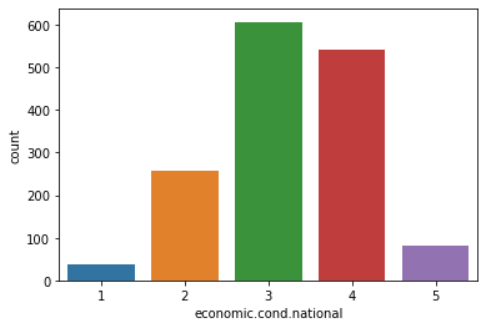




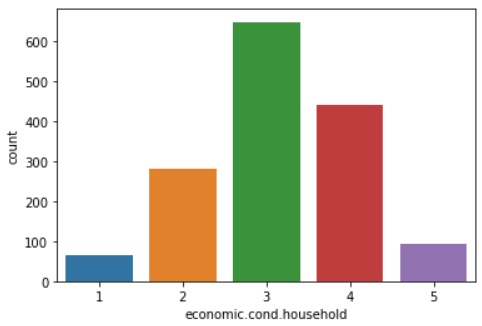




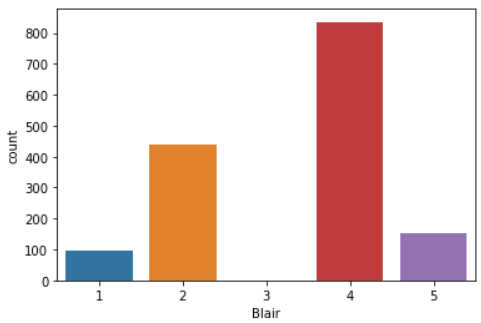
From the above plots it can be observed that the distribution plot holds good only for the age variable which is a continuous variable. As we had speculated, the age variable is a normally distributed variable. For the rest of the variables the distribution plot does not seem to be the optimal one. Hence, we first look at the individual count plots for the ordinal columns:



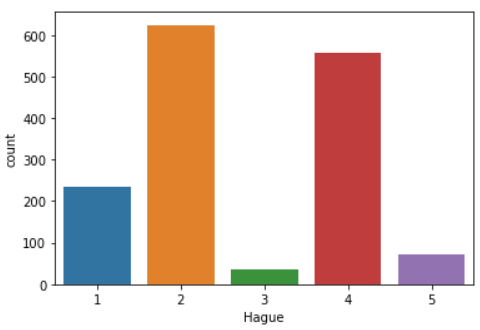
The majority of people have rated 3 on the assessment of current national economic condition. This is closely followed by the rating 4. This is to show that the people are pretty satisfied with the national economic condition.



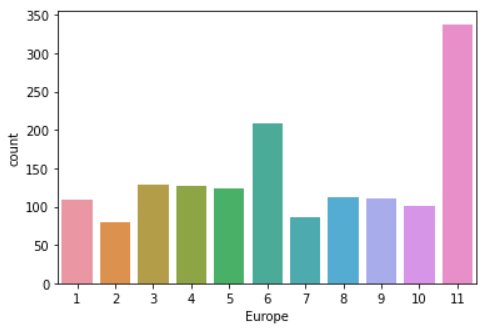
The assessments of people on the current household economic conditions is the most on the rating 3 and followed by the rating 4. It is interesting to see that the difference between the votes on the ratings 3 and 4 are slightly higher as compared to the previous plot.



The number of counts on rating Blair as 4 is the highest with 800 + counts and is followed by the rating 2 with 400 + counts. There are no counts on rating 3 for Blair. The rating 1 is the rating category with the least counts. About 100 + people have rated 5 for Blair.

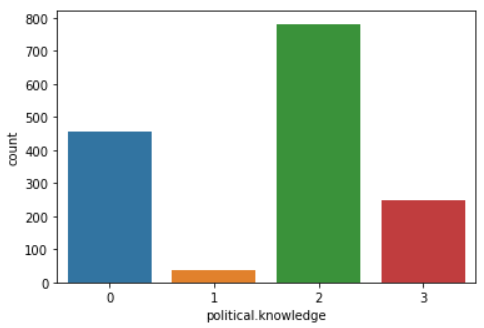


For Hague, the majority counts exist for the rating 2 followed by the rating 4. The number of people who have rated 1 for Hague is more than the number of people who have rated 5 for Hague.



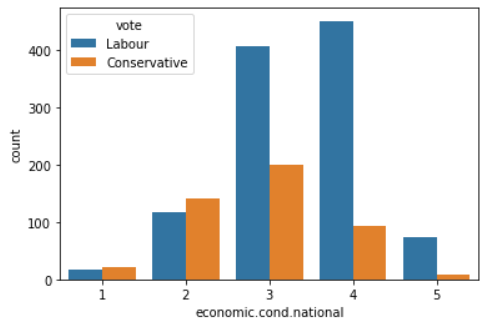
There are most counts for the rating 11 for European integration. This means that they are Eurosceptic.

There are also a high number of counts for the rating 6, goes to say that they are neutral in their opinion on European integration.



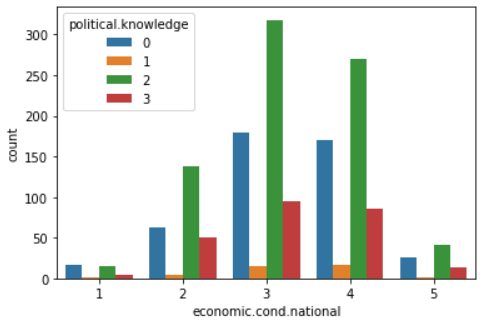
This is for the knowledge of the party’s positions on European integration. Most counts exist for the rating of 2. There are 400+ counts on rating 0. There are around 200 + counts on the rating 3 which is the highest.

**Countplots with hue to analyze two variables:**

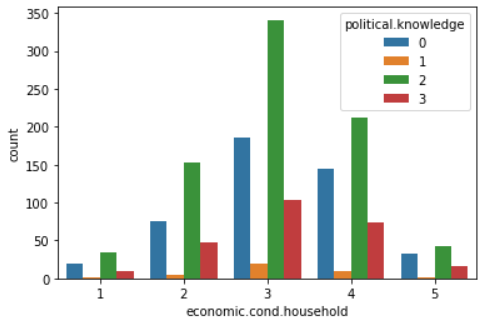


The Labour party has received the maximum votes in the ratings 3,4 and 5. The conservative party has received majority of votes by the people who have rated 2 and 1 on the national economic condition.

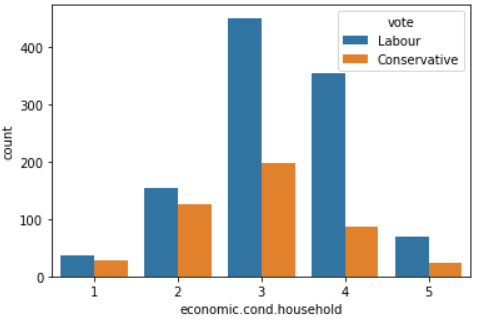
In the ratings 3,4 and 5, the count of people who have voted for the labour party is 400, 400 + and 50 + respectively. The conservative party has received the majority of votes by the people who have rated 3 and 2 on the national economic condition.



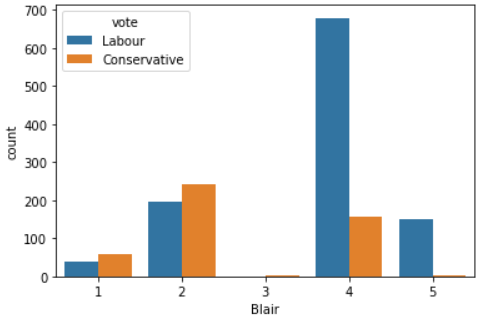
It is to observe that the people in the ratings of 2 and 3 in the political knowledge variable have rated 3 and 4 in the national economic condition. The majority of people who are rated 0 in the political knowledge have also rated 3 and 4 in the national economic condition.



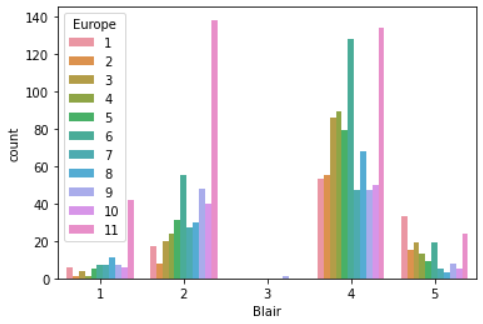
The people who are in the rating 2 of the political knowledge are the maximum number of people who have rated 3 and 4 in the household economic condition.



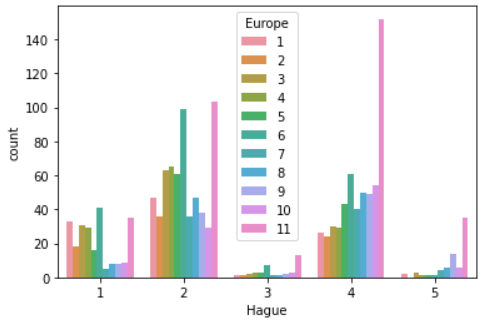
The rating 3 on the household economic condition consists of 430+ votes for the Labour party and less than 200 votes for the conservative party. The difference between the votes on the two parties is even more distinct among the people who have rated 4 on the household economic condition with the votes on the conservative party contributing to around only 50+ whereas the votes on the Labour party is 350+.

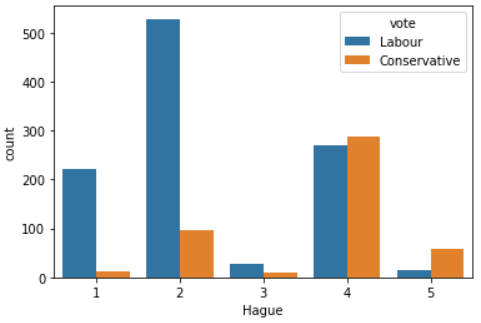


For Blair, the counts on the rating 3 are almost null. The rating 4 has majority counts from the people who have voted for Labour 680+ and has about less than 160 counts from the people who have voted for the conservative.

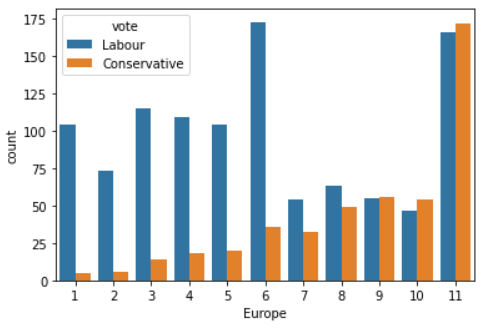


For the Blair variable, the majority counts on the rating 2 comprise of the people who have rated 11 on the Eurosceptic sentiment. For the rating 4, the counts comprise majorly of people who have rated 11 and 5 on the Eurosceptic sentiment. For the rating 5, it is interesting to note that the largest single contribution has come from people who have rated 1 on the Eurosceptic sentiment. Maybe this goes to show that people who think very strongly of Blair are more looking forward to European integration.

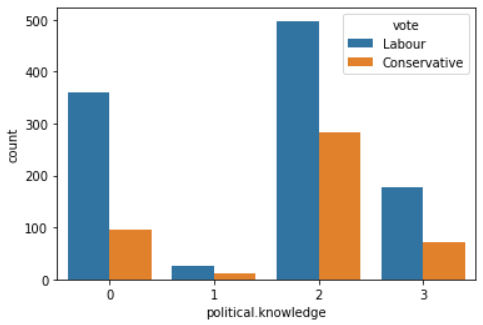




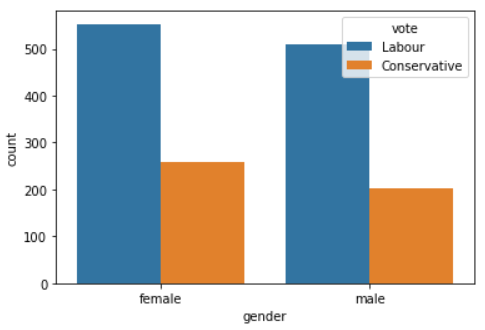
It is to be expected that the counts on the ratings 1 and 2 for Hague mostly comprise of people who have voted for the Labour party. It is however interesting to note that for people who have rated 4 for Hague,the conservative part leader, have also voted for the Labour party. The counts on this rating comprises of the votes in near to equal proportions for both the parties.



It is interesting to note that people who have voted for the Labour party are the majority of people who have rated low on the Eurosceptic sentiment. As the rating on the Europe variable goes up 6-11, the proportion of people who have voted for the Conservative party increases drastically. This does go to show that the people who have voted for the Labour party are more in favor for the European integration whereas the people who have voted for the Conservative party are more in favor of the Eurosceptic sentiment.

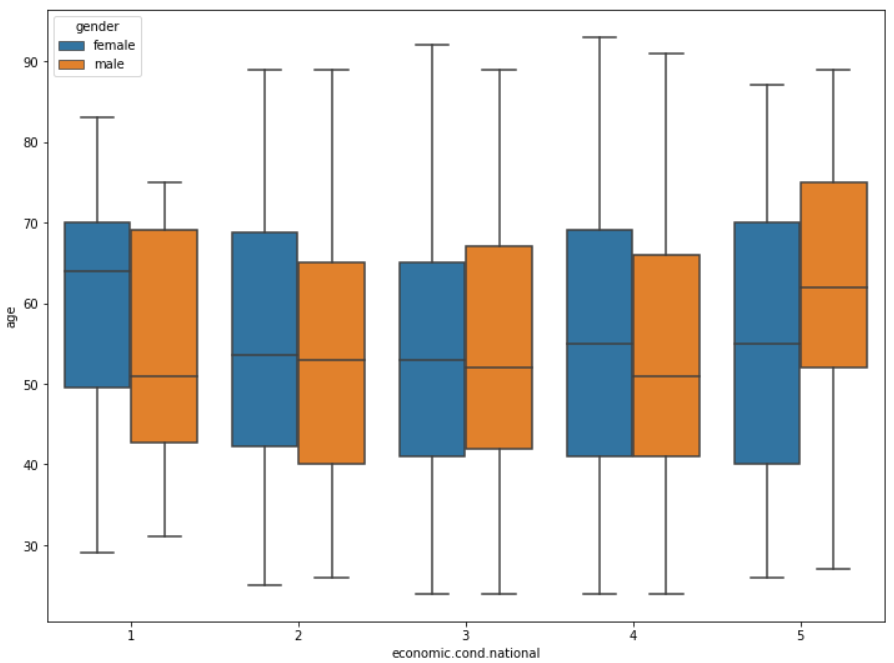


The counts of people who are rated 2 and 3 on the political knowledge have a high proportion of people who have voted for the Conservative party. The highest proportion in every rating category is of course by people who have voted for the Labour party.

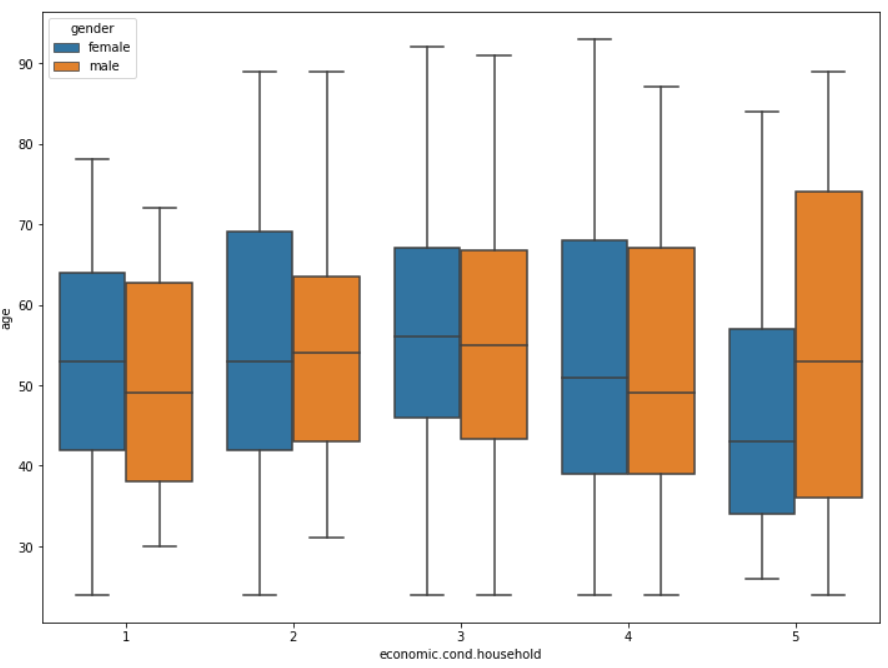


It is interesting to observe that more number of female voters over male are present among the people who have voted for the conservative party.

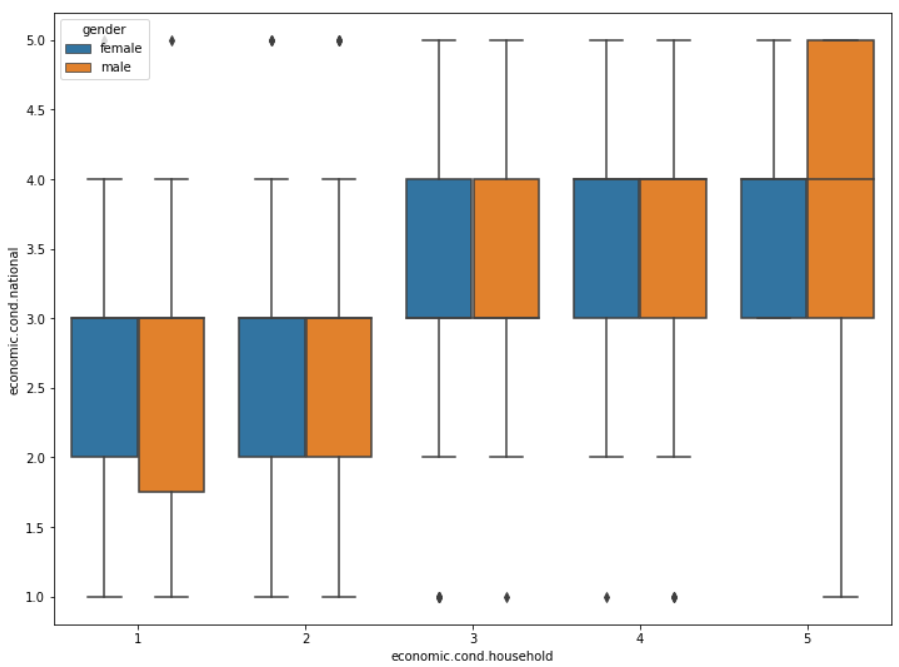
**Categorical Boxplot:**



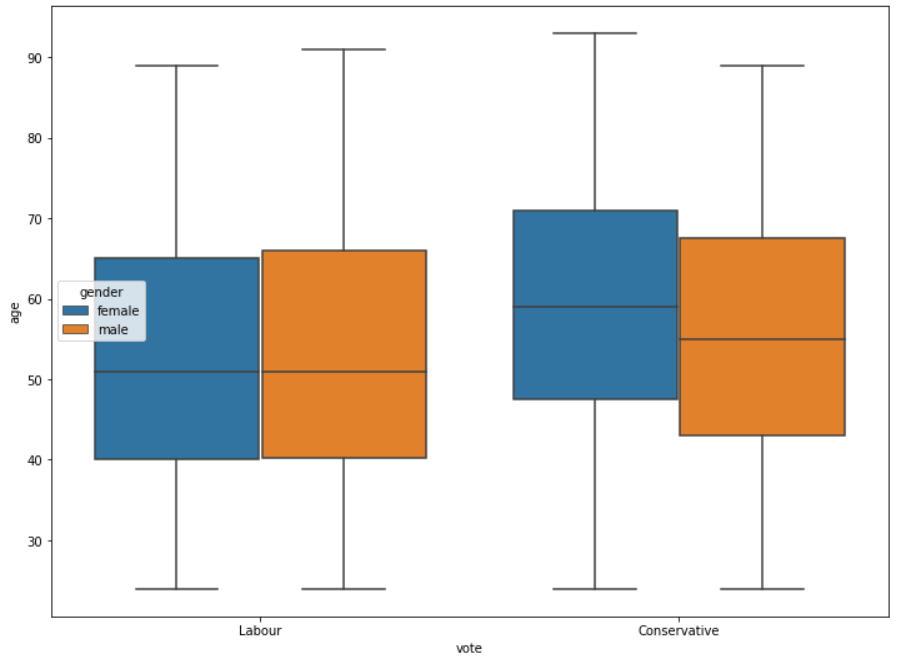
The median age of the people who have rated 3 on the national economic condition is lower than that of the people have rated 5 and 1. There is a sharp contrast between the median age of the gender classes in the rating 1 of national economic condition.



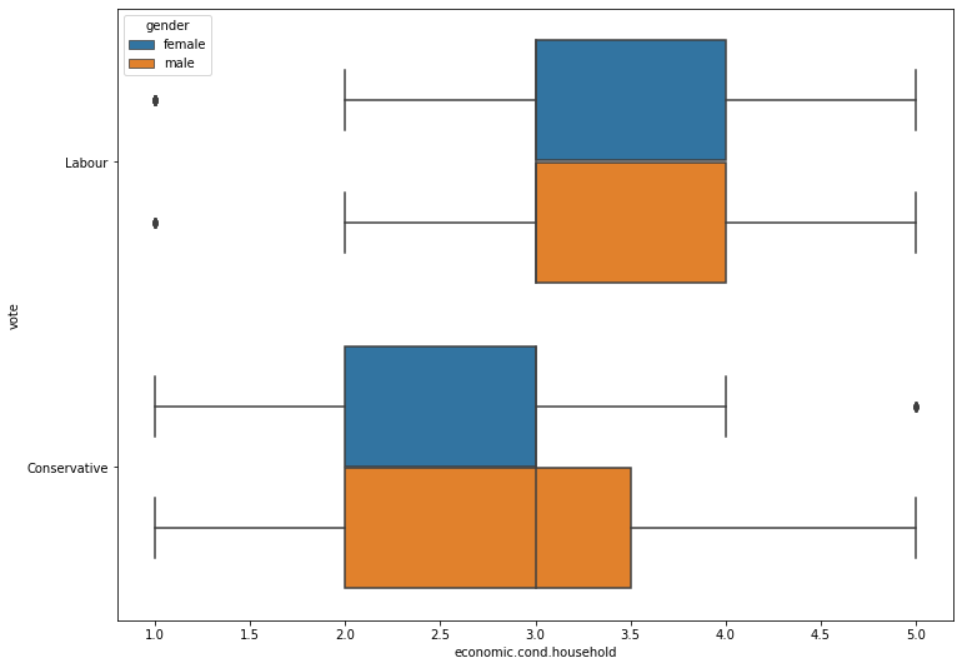
The median age of the people who have rated 3 on the household economic condition is much higher than that of the people who have rated 5 and 1. It is to also note that the majority of people have rated 4 and 5 on the household economic condition. In the rating 5, the median age of the male is much higher than that of the female. In fact the age group of women with the lowest median age have rated 5 on the household economic condition. In the rating 1, the median age of the female is much higher than that of the male.



It is interesting to note that there are outliers for both the gender classes. They have rated low on the household economic condition, rated high on the national economic condition and vice versa. In the rating category 5 of household economic condition there are majority male voters who have rated between 4 and 5 in the national economic condition.



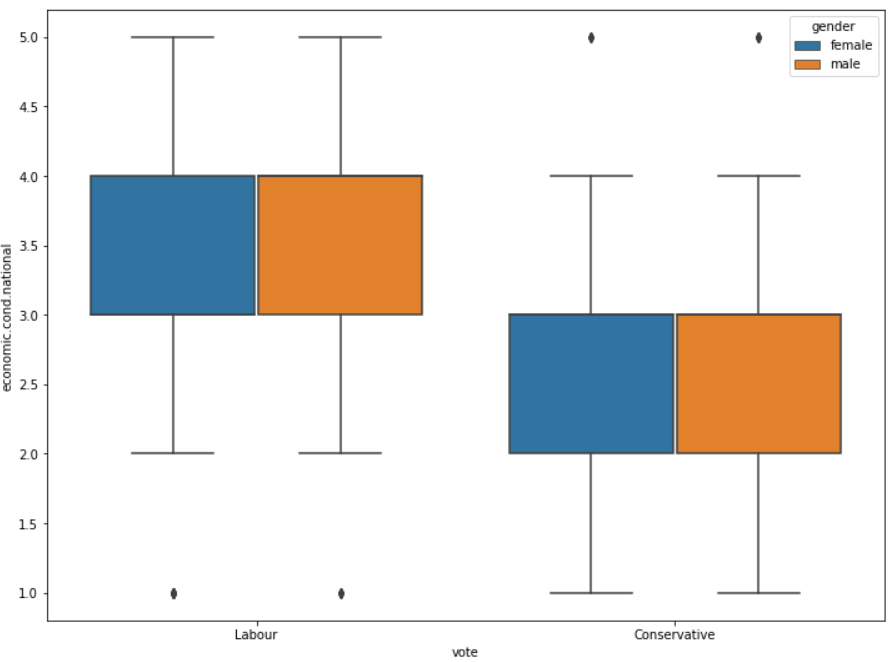
The median age of the people who have voted for the conservative party is much higher than that of the people who have voted for the Labour party. The median ages of the male and female voters are the same for the Labour party but there is a difference between the median ages in the voters for the conservative party. The elderly among the female voters have preferred the Conservative party more than the Labour party.



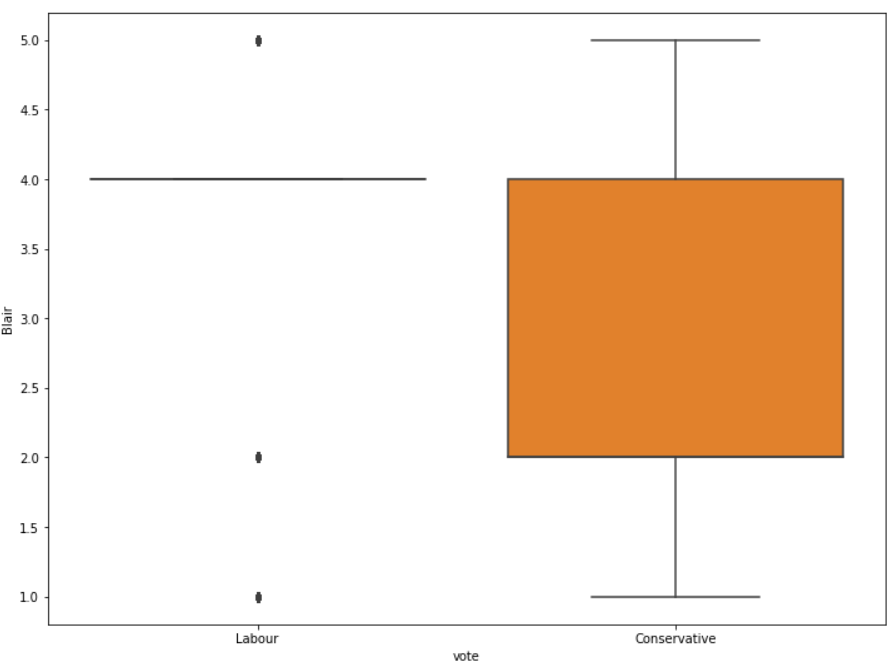
Among the people who have voted for the Labour party, the gender class distribution is almost the same.

It is also to observe that in this category, the majority distribution lies by the voters who have rated between 3 and 4 in the household economic condition.

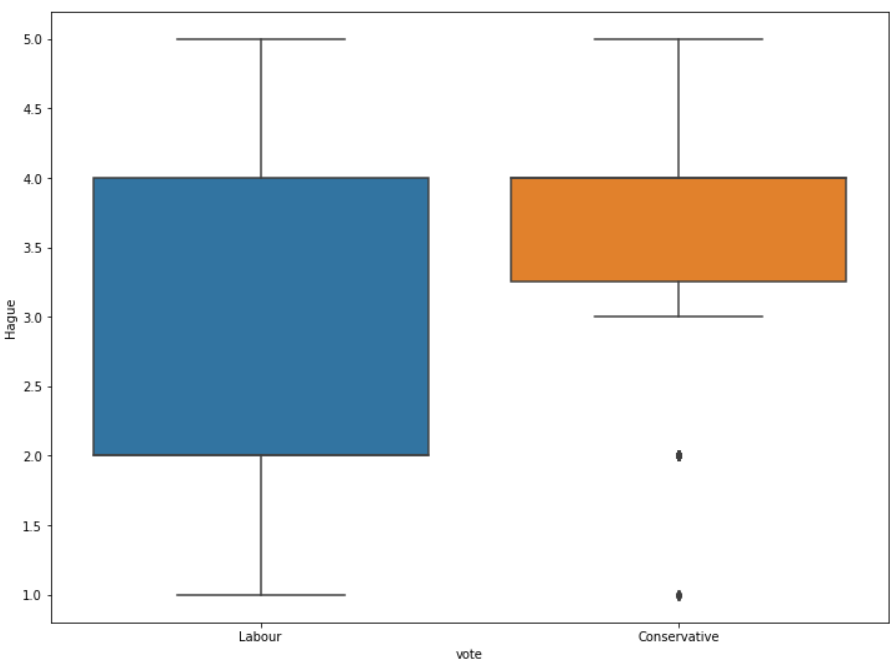
For the conservative party, the female voters have mainly chosen the rating between 2 and 3 in the household economic condition. The male voters however, have chosen the rating between 2 and 3.5 in the household economic condition. In this, majority of male voters have chosen between the rating 3 and 3.5.



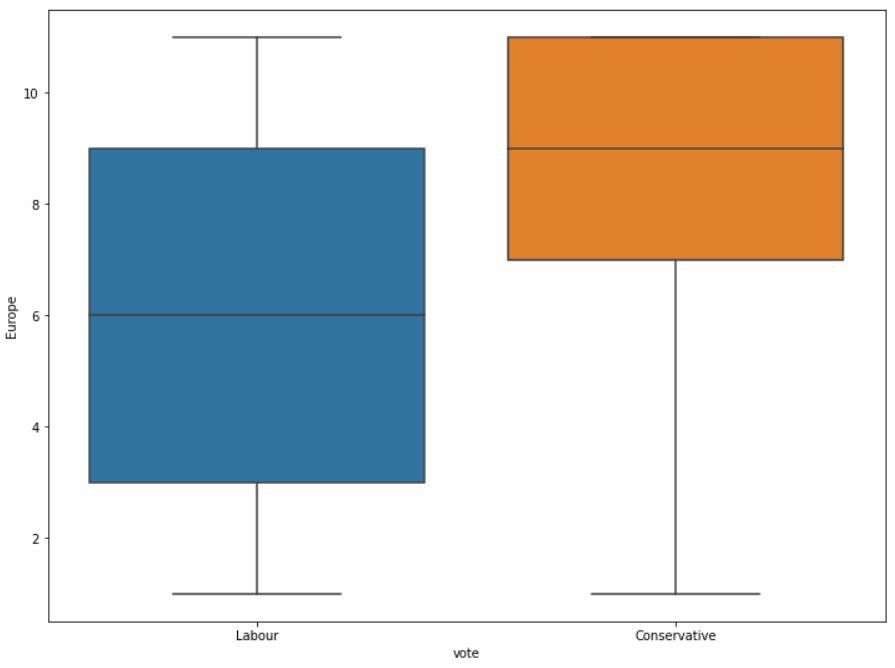
The majority of voters for the Labour party have chosen between the ratings 3 and 4 for the national economic condition national and the majority of voters for the Conservative party have chosen between the ratings 2 and 3 for the economic condition national.



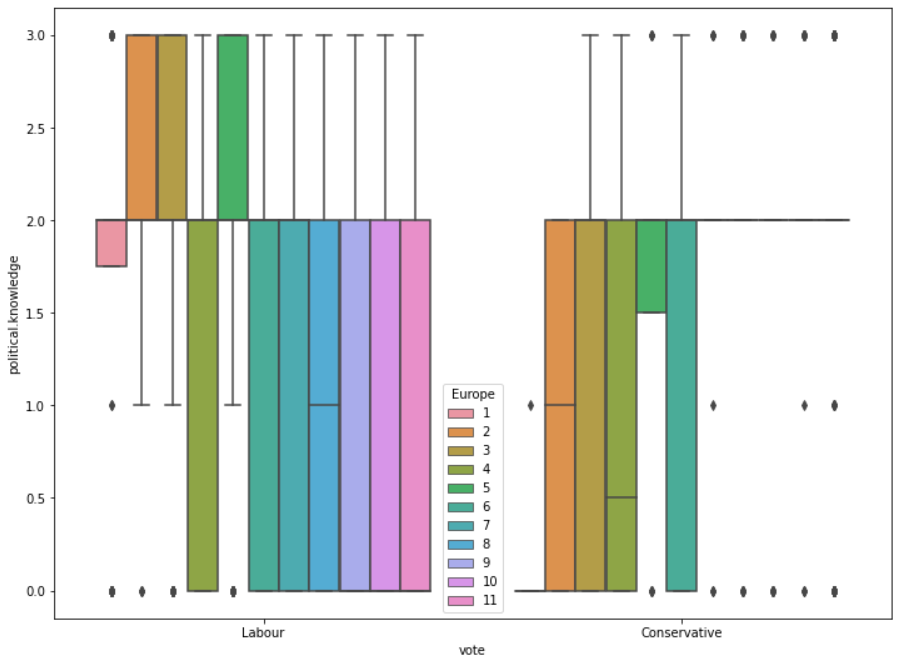
There are more data points with voters for the Conservative party and Blair than for the Labour party and Blair. This is to note that Blair is the assessment of the Labour leader.



For Hague, the majority of voter distribution is in the Labour party . Voters who have voted for the Conservative party have rated between 3.2 and 5 for Hague. There are outliers in the conservative party who have rated 1 and 2 for Hague. Hague is the assessment of the Conservative party leader.



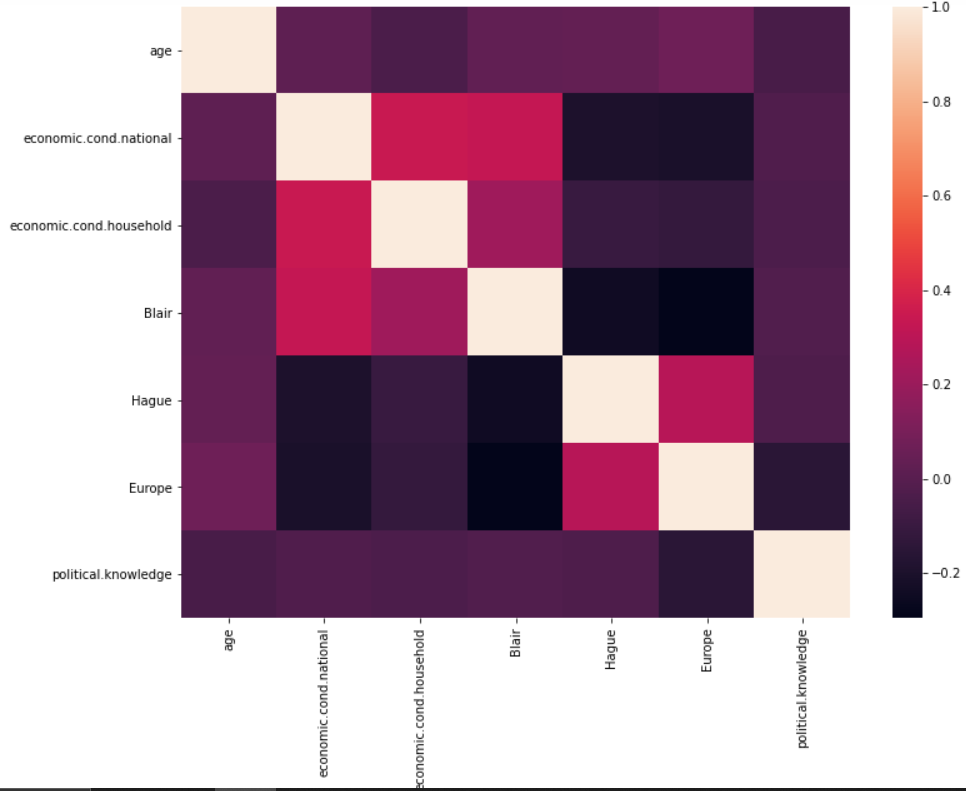
As inferred before, the conservative party is showing a high Eurosceptic sentiment. The distribution of the voters for the Labour party is more neutral towards the European integration.



The data points for the voters of the Labour party and political knowledge is a lot more than that for the Conservative party. For the votes on the Labour party, the people who have shown a high rating on the Eurosceptic sentiment have shown a low rating on the political knowledge. This means that many people who have voted for the Labour party did not have much knowledge on the party’s standpoint on the European integration and yet they have rated high on the Eurosceptic sentiment.

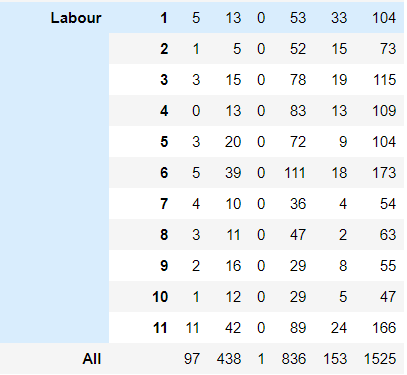
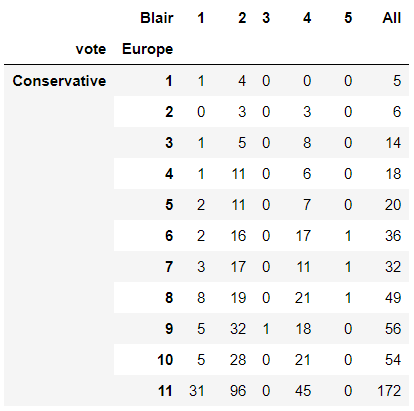
As for the votes on the conservative party, it can be seen that there are a lot of voters having a neutral standpoint on the European integration and have also rated low on the political knowledge.

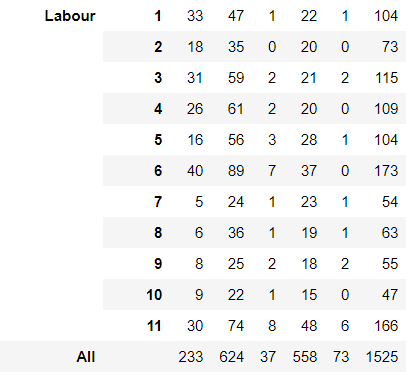
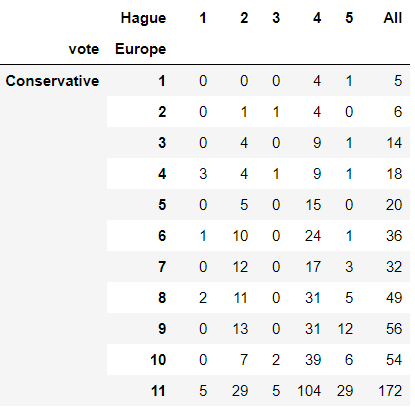
We then check for multicollinearity among the variables using a heatmap:

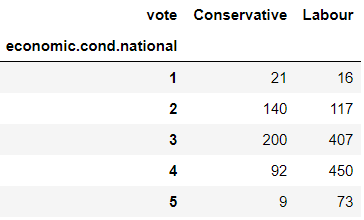


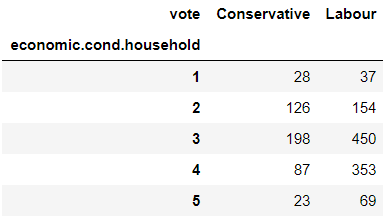
It can be seen that there is some collinearity between the three variables: economic condition household, economic conditional national and Blair.

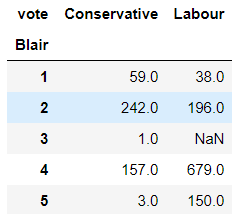
We then do a crosstab and a group-by function for the different variables and look at the numbers:

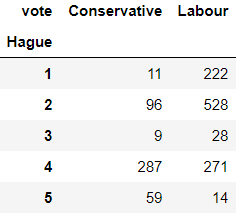


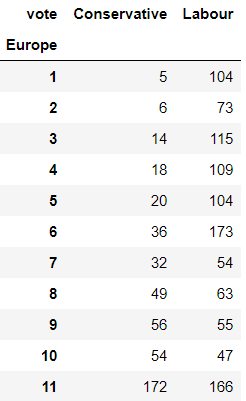


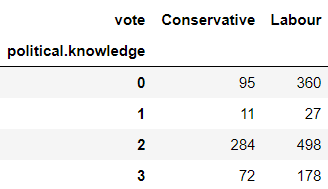




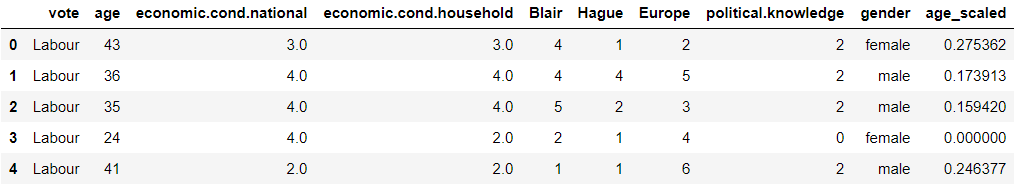








Using the MinMax Scaler, we scale the age variable and give a new name for the column as age\_scaled. The head of the data looks like this:



We then remove the original age column from our dataframe.

We then perform one hot encoding on the Gender variable and perform label encoding on the target variable: vote

Here, it is to note that Labour is 1 and Conservative is 0 in the target variable: vote.

We then split the data into 70% training and 30% testing using the train-test split.

**Logistic Regression model:**

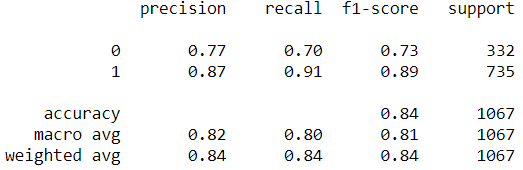
We pass a Logistic Regression model with maximum iterations of 1000.

We get the model accuracy as 84.25% on the train set and 82.31% on the test set.

Confusion matrix on the train data:



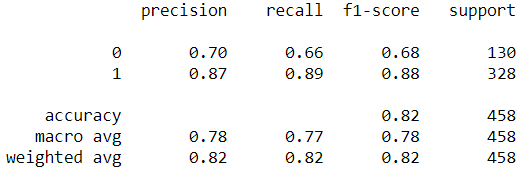
Classification report on the train data:



Confusion matrix on the test data:



Confusion matrix on the test data:



### AUC and ROC for the train and test data:

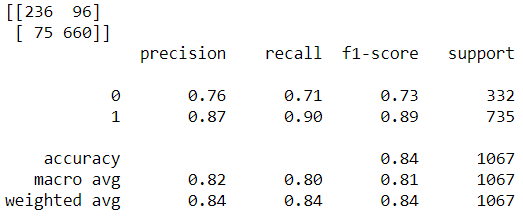
### 

**Linear Discriminant Analysis model:**

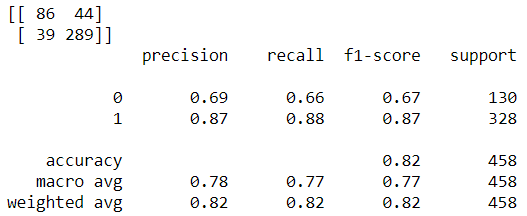
We pass a Linear Discriminant Analysis model.

We get a model accuracy score of 83.9% on the train data and 81.8% on the test data.

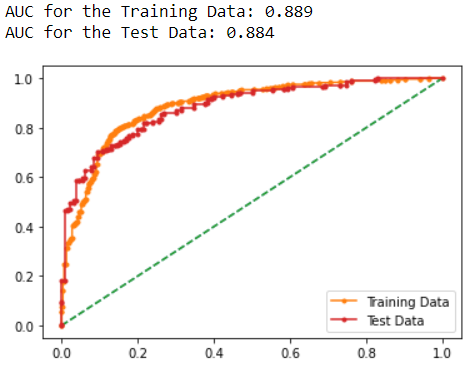
Confusion matrix and Classification report for the train data:



Confusion matrix and Classification report on the test data:



### AUC and ROC for the train and test data:

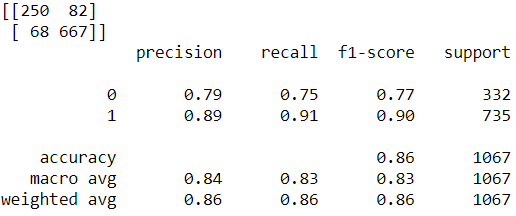


**K-Nearest Neighbors model:**

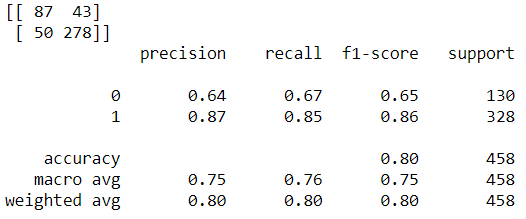
We pass a KNN model.

The model accuracy on the train data is 85.9% and on the test data the accuracy is 79.6%

The Confusion matrix and Classification report on the train data is:

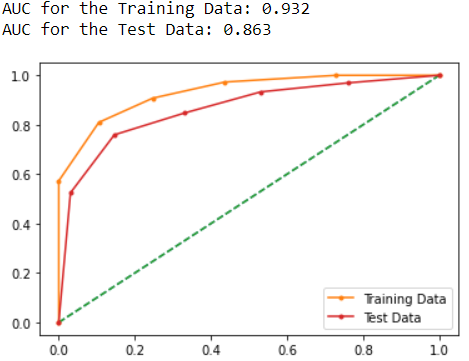


The confusion matrix and Classification report on the test data is:



The is a clear overfit model as the accuracy scores are much lower on the test data.

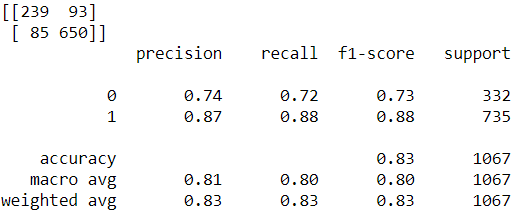
### AUC and ROC for the train and test data:



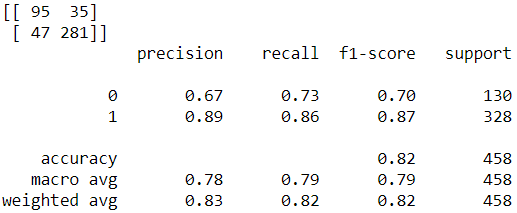
**Naive Bayes model:**

The model accuracy on the train data is 83.3% and on the test data it is 82.09%

The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:



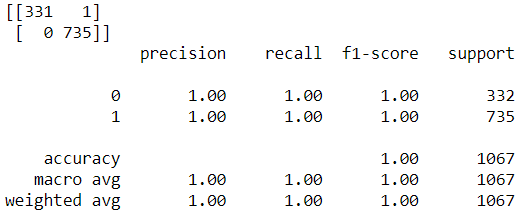
**Random Forest Classifier model:**

We run the Random forest model with number of estimators as 100 and the random state as 1.

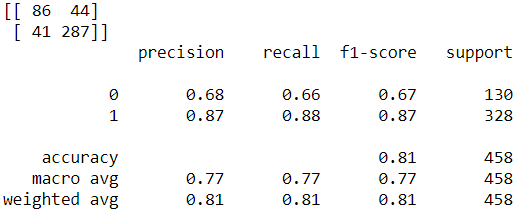
The model accuracy on the train data is 99.9% and on the test set it is 81.44%

This is a highly overfit model and is not to be considered as the difference between the train and test accuracies are more than 10% .

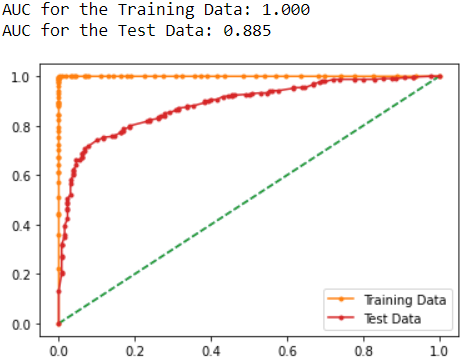
The confusion matrix and classification report on the train data are:



The confusion matrix and classification report on the test data are:



AUC and ROC for the train and test data:



We have achieved the highest AUC possible in the training data with this model. The values on the test data are much lower than the test AUC.

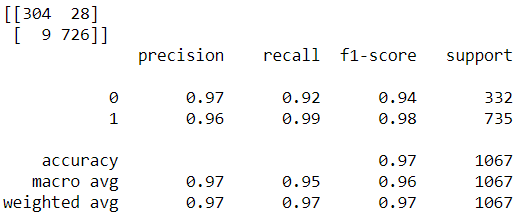
**Bagging Classifier model:**

We run the bagging classifier with the base estimator as the Random Forest model, number of estimators as 100 and the random state as 1.

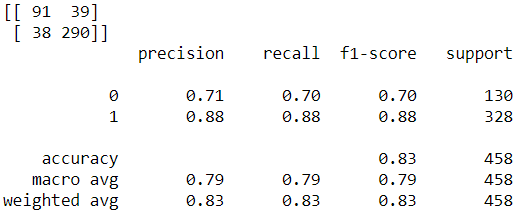
The model accuracy on the train set is 96.5% and on the test set it is 83.18%

Since we have used the random forest model as the base estimator, the bagging classifier model is also an overfit model.

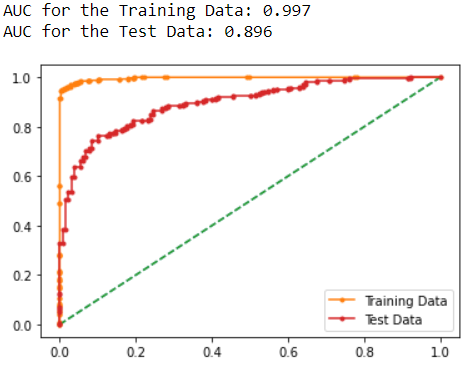
The confusion matrix and classification report on the train data are:



The confusion matrix and classification report on the test data are:



AUC and ROC for the train and test data:



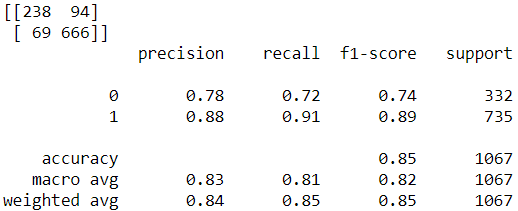
**Boosting models:**

**ADA Boost:**

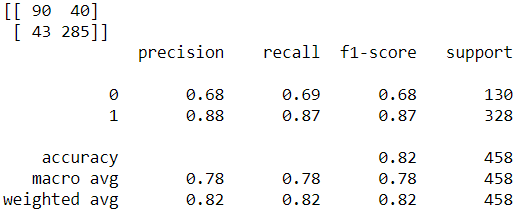
We use the AdaBoostClassifier model with number of estimators as 100 and random state as 1.

The model accuracy on the train set is 84.7% and on the test set it is 81.87%

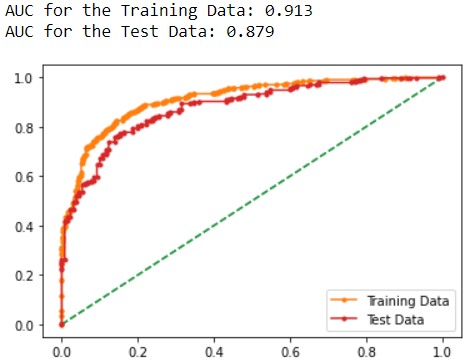
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



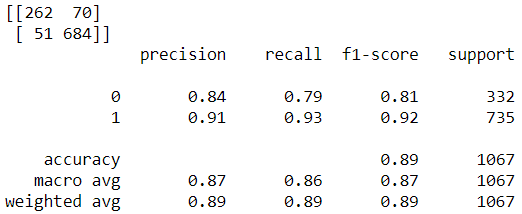
AUC and ROC for the train and test data:



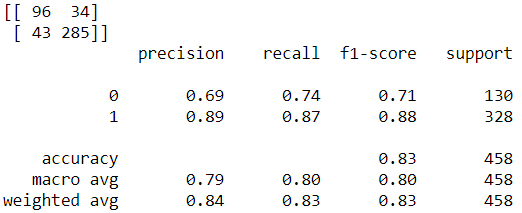
**Gradient Boosting model:**

The model accuracy score on the train data is 88.65% and on the test data it is 83.18%

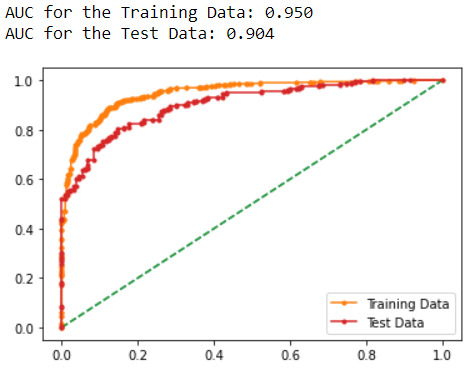
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:



We then try to tune each of the models to achieve better results:

**Model-tuning with GridSearch CV:**

**KNN model tuning:**

We pass the K Neighbors Classifier model with the Ball Tree algorithm.

Ball tree helps to overcome the inefficiencies of the KD tree for higher dimensions. Ball tree is much faster for storing multidimensional data. Ball tree partitions data in nested hyperspheres called balls. In ball tree, each node contains a hypersphere. And each hypersphere contains a subset of points which are needed to be searched.

Source: <https://debuggercafe.com/an-introduction-to-k-nearest-neighbors-in-machine-learning/>

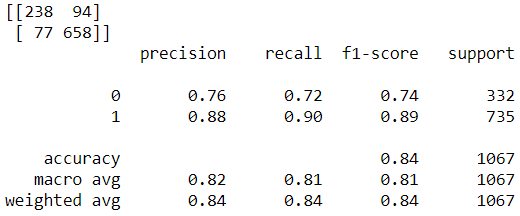
We use the following parameters in the Gridsearch:

Number of neighbors from 1 to 20, Leaf size from 20 to 40, weights as uniform and distance based and metric as Euclidean and Manhattan. We also use cross validation with 5 folds and use all the processors for instance threads with n jobs as -1. We use the scoring as accuracy.

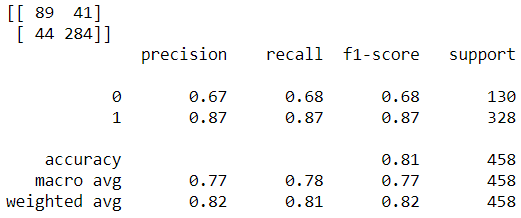
We then fit the KNN model on the train data:

The accuracy of the KNN model on the train data is 83.9% and on the test data it is 81.44 %.

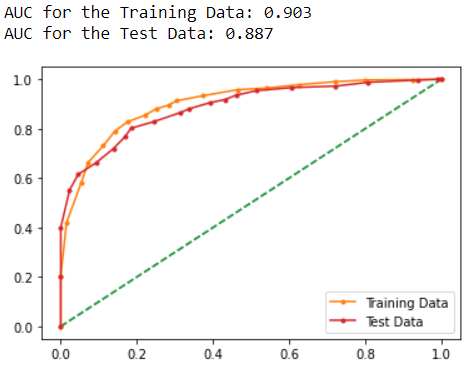
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:



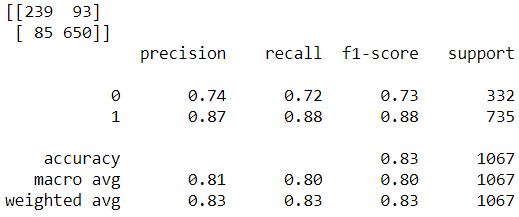
**Gaussian Naive Bayes tuning:**

We use the Naïve Bayes model with the repeated stratified K fold cross validation.

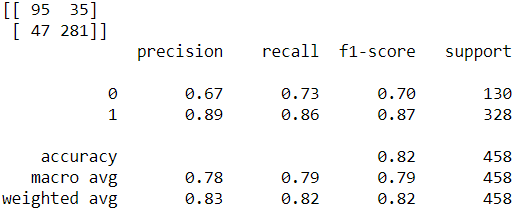
It repeats stratified K-fold n times with different randomization in each repetition. The folds are made by preserving the percentage of samples for each class. We use 5 folds and say that the cross-validator must be repeated 3 times. The random state we use for this model is 999. We use verbose as 1 and that means that the computation time for each fold and parameter candidate is displayed. The scoring is used as accuracy.

The model accuracy score on the train data is 83.31% and on the test data it is 82.09%

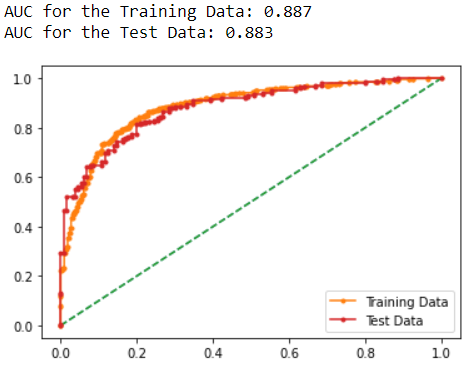
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:



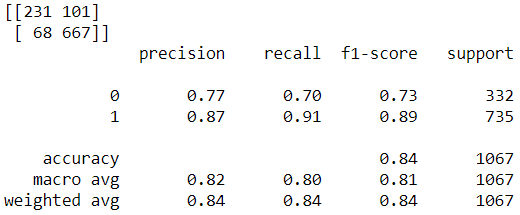
### Logistic regression tuning:

We use the logisitic regression model with the following hyperparameters:

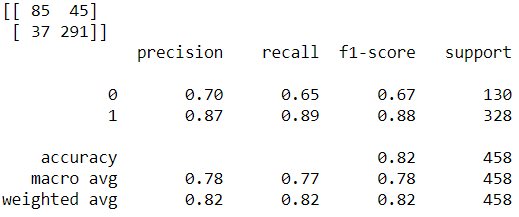
C as 0.001, 0.01, 0.1, 1, 10, 100, 1000. It is the inverse to Lambda. It controls the regularization to avoid overfitting the model. The penalty is used as l1 and l2. The penalty is added to control the properties of the regression coefficients. The maximum iterations is from 100 to 1000 with a difference of 100. This is the number of iterations for the solvers to converge. The solvers used are: newton-cg, lbfgs, liblinear, sag and saga. Refit is used as True which will refit the best estimator to the whole training set automatically. The verbose is set to 3 in this case which will display the fold and candidate parameter indexes together along with the starting time of the computation. Cross validation is used as 5.

The model accuracy on the train data is 84.16% and on the test data it is 82.09%.

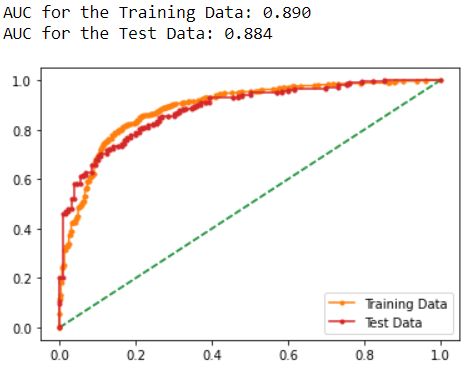
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:



### Linear Discriminant Analysis tuning:

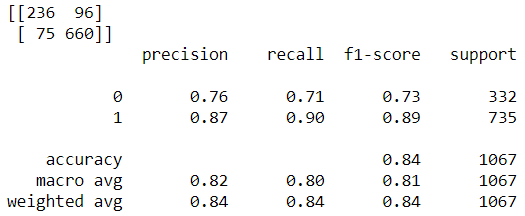
We use the lda model with the solver as lsqr that stands for least squares solution.

The Repeated Stratified K fold is used with 8 folds and the repeats as 3. The random state is used as 1.

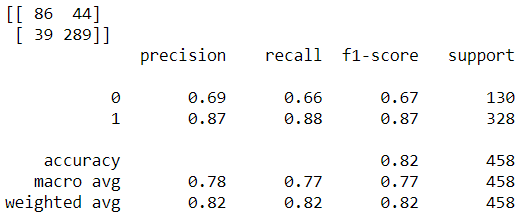
All the processors are used and the scoring is as accuracy.

The model accuracy score on the train data is 83.9% and on the test data it is 81.87%.

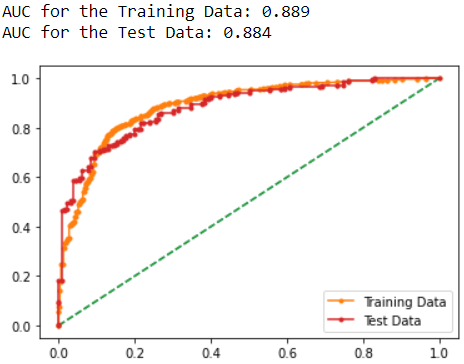
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:



### Random Forest Tuning:

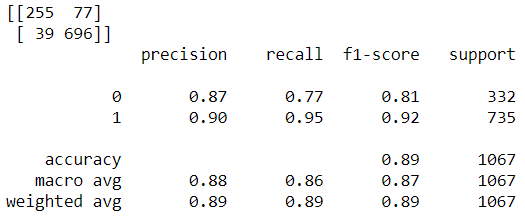
We use the Random Forest model with the following tuning parameters:

Number of estimators as 100, criterion as gini, maximum depth as none, minimum samples per split as 6, minimum samples per leaf as 4, the minimum weight fraction of leaf as 0, the maximum features as automatic, maximum leaf nodes as none, bootstrap as true, the out of bag score as false, the number of jobs as 1, verbose as 0, warm start as false and the class weight as none.

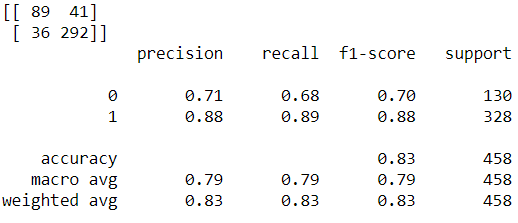
We use the random state as 1.

The model accuracy score on the train data is 89.12% and on the test data it is 83.18%

The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



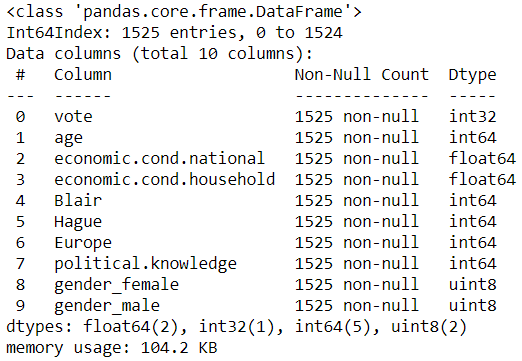
AUC and ROC for the train and test data:



**Binning the ‘age’ variable:**

We then bin the age variable into 9 bins of almost equal age intervals starting from the age 24 to 93.

We also relabel the age intervals from 1 to 9. We convert the age variable as an integer type as it now only contains the bin labels from 1 to 9. The information on the data frame with the data type looks like this:



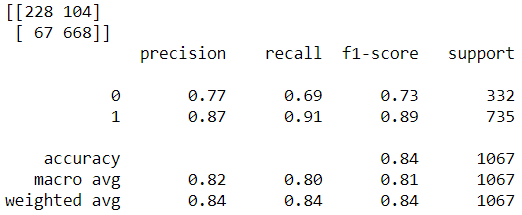
We then split the data into 70% training and 30% testing with the random state as 1.

**Logistic Regression model with tuning:**

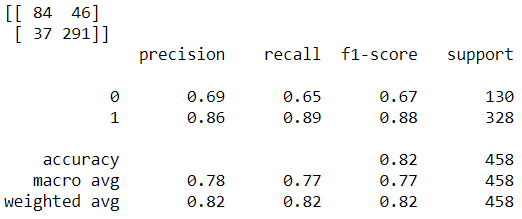
We then fit the logistic regression model with the tuning parameters discussed earlier.

The model accuracy on the train data is 83.9% and on the test data it is 81.8%

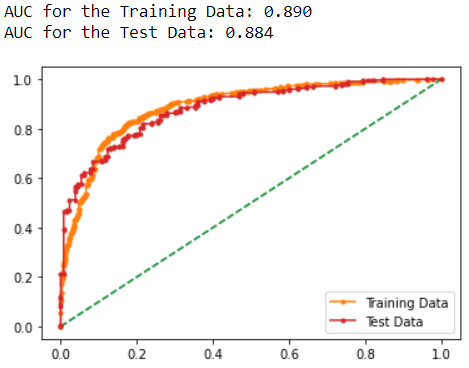
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:

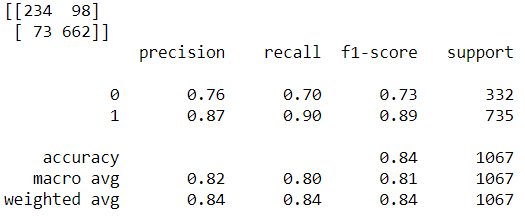


**KNN model with tuning:**

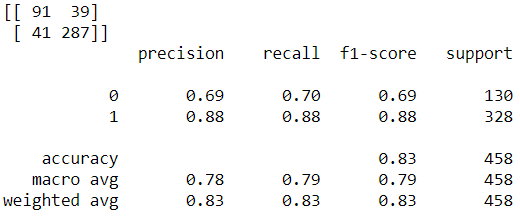
The KNN model is then fit to the train data with the tuning parameters discussed earlier.

The model accuracy on the train data is 83.97% and on the test data it is 82.53%.

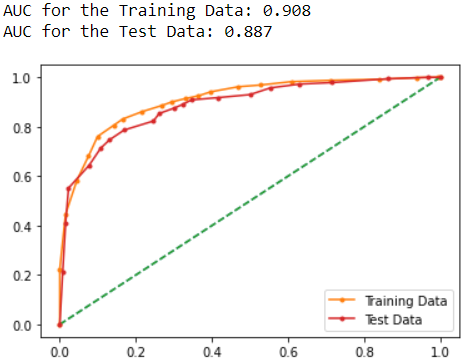
The confusion matrix and the classification report on the train data is:



The confusion matrix and the classification report on the test data is:



AUC and ROC for the train and test data:

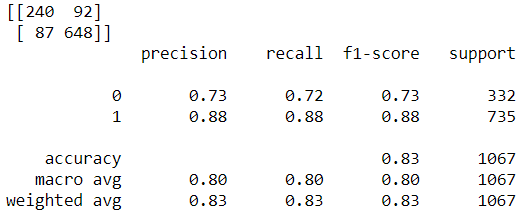


**Gaussian Naive Bayes with tuning:**

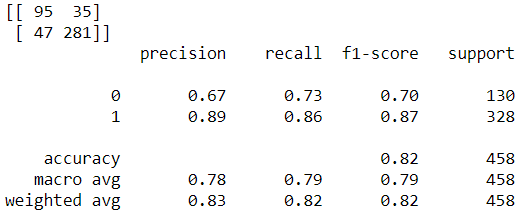
We fit the Gaussian Naïve Bayes model with the tuning parameters on the train data.

The model accuracy on the train data is 83.22% and on the test data it is 82.09%

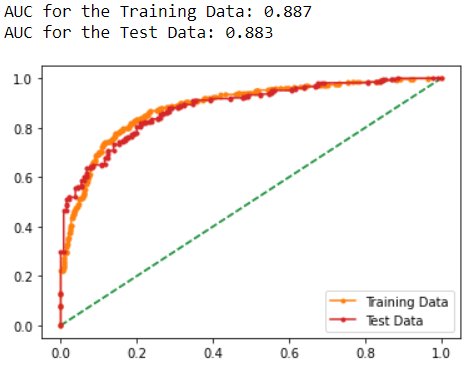
The confusion matrix and classification report on the train data look like this:



The confusion matrix and classification report on the test data look like this:



AUC and ROC for the train and test data:



### Linear Discriminant Analysis with tuning:

### We fit tuned linear discriminant analysis model on the train data with the tuning parameters discussed earlier.

### The model accuracy score on the train data is 83.7% and on the test data it is 82.5%

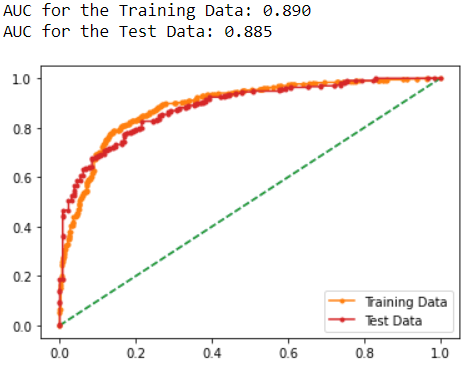
### The confusion matrix and classification report on the train data is:

### 

### The confusion matrix and classification report on the test data is:

### 

**AUC and ROC for the train and test data:**

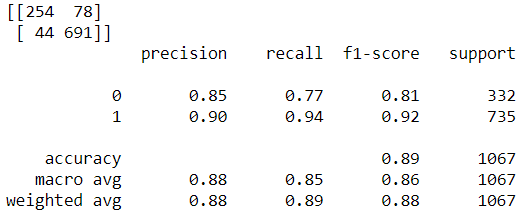


**Random Forest with Tuning:**

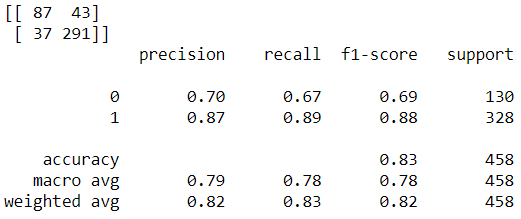
The random forest model is tuned and fit on the train data.

The model accuracy score on the train data is 88.56% and on the test data it is 82.5%

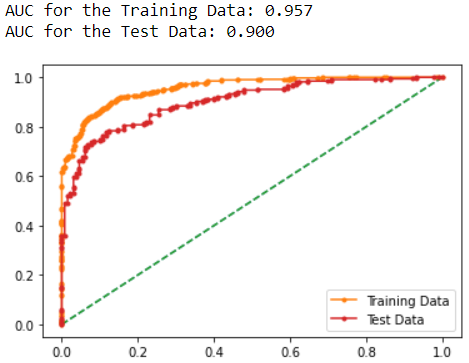
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



**AUC and ROC for the train and test data:**



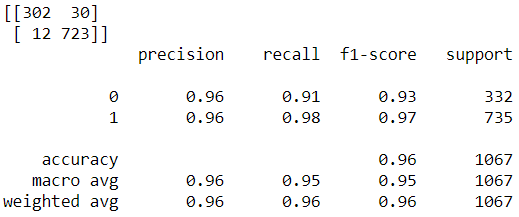
**Bagging:**

We have used the Random Forest model as the base estimator and the number of estimators are used as 100. The random state is used as 1.

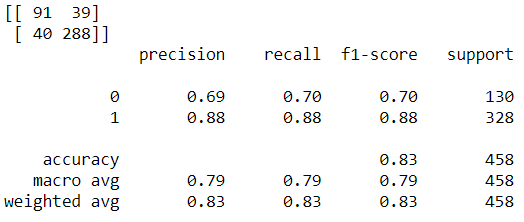
We fit the model on the train data.

The model accuracy on the train data is 96.6% and on the test data it is 82.75%

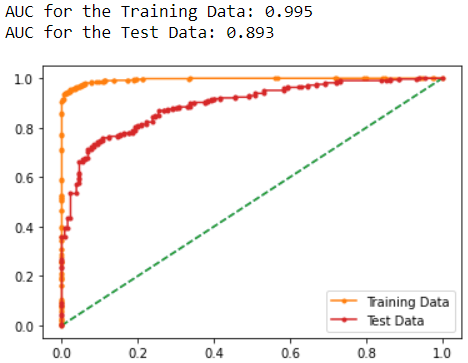
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:



**AUC and ROC for the train and test data:**



### Boosting:

### ADA Boost:

### The ADA boost model with number of estimators as 100 and learning rate as 1.0 is fit on the train data.

### The model accuracy on the train data is 83.88% and on the test data it is 81.44%

### The confusion matrix and classification report on the train data is:

### 

### The confusion matrix and classification report on the test data is:

### 

### AUC and ROC for the train and test data:

### 

### Gradient Boosting:

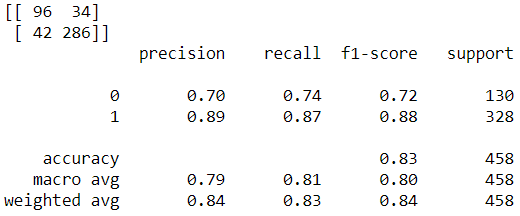
### Without Tuning:

### The model accuracy score on the train data is 88.09% and on the test data it is 83.40%.

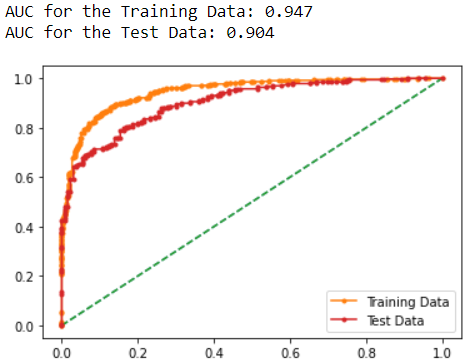
### The confusion matrix and classification report on the train data is:

### 

### The confusion matrix and classification report on the test data is:



AUC and ROC for the train and test data:

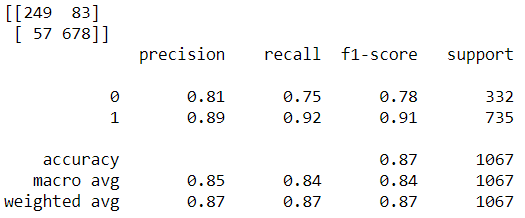


**With tuning:**

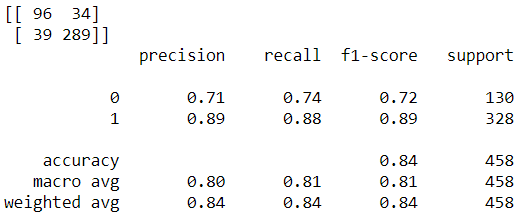
We use the tuning parameters as number of estimators: 50, 250, 500, 1000 and the max depth as 3, 5, 7, 9. We also use the learning rate as 0.01, 0.1, 0.2, 0.3. The learning rate is a technique to slow down the learning in the gradient boosting model by applying a weighting factor for the corrections by new trees when added to the model. This weighting is called as the learning rate.

The model accuracy on the train data is 86.87% and on the test data it is 84.06%

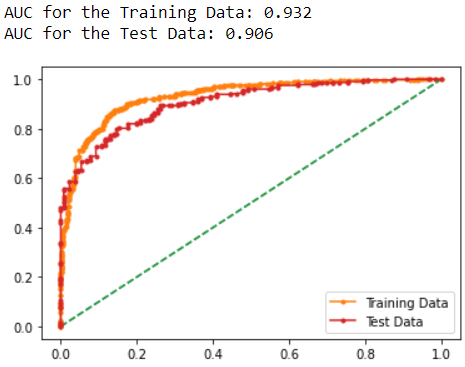
The confusion matrix and classification report on the train data is:



The confusion matrix and classification report on the test data is:

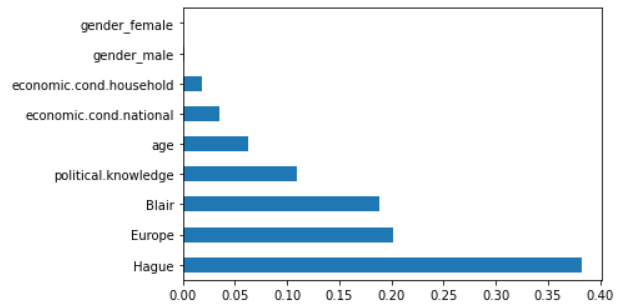


AUC and ROC for the train and test data:



Thus, it can be seen that tuning the gradient boosting model has yielded much better results.

The importance of features can be seen here:



**Final comparison of performance metrics across all the untuned models:**

**By scaling the age variable:**

**For the class 1 in the target variable: Labour party**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic regression | | Linear discriminant Analysis | | K-Nearest Neighbors | | Naïve Bayes | | Random Forest Classifier | | Bagging | | ADA Boosting | | Gradient Boosting | | |
| Accuracy | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 84.25 | 82.31 | 83.97 | 81.87 | 85.94 | 79.69 | 83.31 | 82.09 | 99.90 | 81.44 | 96.53 | 83.18 | 84.72 | 81.87 | | 88.65 | 83.18 |
| Precision | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.87 | 0.87 | 0.87 | 0.87 | 0.89 | 0.87 | 0.87 | 0.89 | 1.000 | 0.87 | 0.96 | 0.88 | 0.88 | 0.88 | | 0.91 | 0.89 |
| Recall | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.91 | 0.89 | 0.90 | 0.88 | 0.91 | 0.85 | 0.88 | 0.86 | 1.00 | 0.88 | 0.99 | 0.88 | 0.91 | 0.87 | | 0.93 | 0.87 |
| F1-score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.89 | 0.88 | 0.89 | 0.87 | 0.90 | 0.86 | 0.88 | 0.87 | 1.00 | 0.87 | 0.98 | 0.88 | 0.89 | 0.87 | | 0.92 | 0.88 |
| AUC score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.89 | 0.88 | 0.88 | 0.88 | 0.93 | 0.86 | 0.88 | 0.88 | 1.00 | 0.88 | 0.99 | 0.89 | 0.91 | 0.87 | | 0.95 | 0.90 |

**For the class 0 in the target variable: Conservative party**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic regression | | Linear discriminant Analysis | | K-Nearest Neighbors | | Naïve Bayes | | Random Forest Classifier | | Bagging | | ADA Boosting | | Gradient Boosting | | |
| Accuracy | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 84.25 | 82.31 | 83.97 | 81.87 | 85.94 | 79.69 | 83.31 | 82.09 | 99.90 | 81.44 | 96.53 | 83.18 | 84.72 | 81.87 | | 88.65 | 83.18 |
| Precision | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.77 | 0.70 | 0.76 | 0.69 | 0.79 | 0.64 | 0.74 | 0.67 | 1.00 | 0.68 | 0.97 | 0.71 | 0.78 | 0.68 | | 0.84 | 0.69 |
| Recall | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.70 | 0.66 | 0.71 | 0.66 | 0.75 | 0.67 | 0.72 | 0.73 | 1.00 | 0.66 | 0.92 | 0.70 | 0.72 | 0.69 | | 0.79 | 0.74 |
| F1-score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.73 | 0.68 | 0.73 | 0.67 | 0.77 | 0.65 | 0.73 | 0.70 | 1.00 | 0.67 | 0.94 | 0.70 | 0.74 | 0.68 | | 0.81 | 0.71 |
| AUC score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.89 | 0.88 | 0.88 | 0.88 | 0.93 | 0.86 | 0.88 | 0.88 | 1.00 | 0.88 | 0.99 | 0.89 | 0.91 | 0.87 | | 0.95 | 0.90 |

For the class 1:

* The Accuracy in the training data is the highest for the Random Forest model but this model is a highly overfit model and hence not considered. Similarly, the bagging classifier model has the training data accuracy of 96.53% but on the test data it is 83.18%

Since this is also greater than 10% we will not consider this model. The Gradient boosting model has a training accuracy of 88.65% and a testing accuracy of 83.18%. This is a good model.

* The recall of the bagging model is 0.99 on the train data and 0.88 on the test data.

This is again an overfit model and hence the next best values are in the Gradient boosting model with a train recall of 0.93 and a test recall of 0.87

* The F1 score is also high in the Gradient Boosting model with 0.92 train and 0.88 test.
* The AUC score is 0.95 for the Gradient boosting model with the train data and is 0.90 for the test data. The scores are much higher in the train data for Random Forest and Bagging models but they fail to perform equally well on the test data and hence those models are not being considered as the best model.

For the class 0:

* The Recall scores are the highest for the gradient boosting model with 0.79 – Train and 0.74 Test.
* The F1 score is much higher for the Bagging model with Train of 0.94 and test of 0.70. But the gradient boosting model is higher in the test set with 0.71 on the F1 score.
* The precision for the train data in the gradient boosting is the highest with 0.84 but is only 0.69 with the test data.

**Comparison of performance metrics across all the tuned models:**

**For the class 1 in the target variable: Labour party**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic regression | | Linear discriminant Analysis | | K-Nearest Neighbors | | Naïve Bayes | | Random Forest Classifier | |
| Accuracy in % | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 84.16 | 82.09 | 83.97 | 81.87 | 83.97 | 81.44 | 83.31 | 82.09 | 89.12 | 83.18 |
| Precision | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.87 | 0.87 | 0.87 | 0.87 | 0.88 | 0.87 | 0.87 | 0.89 | 0.90 | 0.88 |
| Recall | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.91 | 0.89 | 0.90 | 0.88 | 0.90 | 0.87 | 0.88 | 0.86 | 0.95 | 0.89 |
| F1-score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.89 | 0.88 | 0.89 | 0.87 | 0.89 | 0.87 | 0.88 | 0.87 | 0.92 | 0.88 |
| AUC score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.89 | 0.88 | 0.88 | 0.88 | 0.90 | 0.88 | 0.88 | 0.88 | 0.96 | 0.90 |

**For the class 0 in the target variable: Conservative party**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic regression | | Linear discriminant Analysis | | K-Nearest Neighbors | | Naïve Bayes | | Random Forest Classifier | |
| Accuracy in % | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 84.16 | 82.09 | 83.97 | 81.87 | 83.97 | 81.44 | 83.31 | 82.09 | 89.12 | 83.18 |
| Precision | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.77 | 0.70 | 0.76 | 0.69 | 0.76 | 0.67 | 0.74 | 0.67 | 0.87 | 0.71 |
| Recall | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.70 | 0.65 | 0.71 | 0.66 | 0.72 | 0.68 | 0.72 | 0.73 | 0.77 | 0.68 |
| F1-score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.73 | 0.67 | 0.73 | 0.67 | 0.74 | 0.68 | 0.73 | 0.70 | 0.81 | 0.70 |
| AUC score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts |
| 0.89 | 0.88 | 0.88 | 0.88 | 0.90 | 0.88 | 0.88 | 0.88 | 0.96 | 0.90 |

**Final comparison of performance metrics across all the tuned models:**

**By binning the age variable:**

**For the Class 1 in the target variable: Labour Party**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic regression | | Linear discriminant Analysis | | K-Nearest Neighbors | | Naïve Bayes | | Random Forest Classifier | | Bagging | | ADA Boosting | | Gradient Boosting | | |
| Accuracy in % | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 83.97 | 81.87 | 83.78 | 82.53 | 83.97 | 82.53 | 83.22 | 82.09 | 88.56 | 82.53 | 96.06 | 82.75 | 83.88 | 81.44 | | 86.87 | 84.06 |
| Precision | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.87 | 0.86 | 0.87 | 0.87 | 0.87 | 0.88 | 0.88 | 0.89 | 0.90 | 0.87 | 0.96 | 0.88 | 0.87 | 0.87 | | 0.89 | 0.89 |
| Recall | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.91 | 0.89 | 0.90 | 0.89 | 0.90 | 0.88 | 0.88 | 0.86 | 0.94 | 0.89 | 0.98 | 0.88 | 0.90 | 0.87 | | 0.92 | 0.88 |
| F1-score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.89 | 0.88 | 0.88 | 0.88 | 0.89 | 0.88 | 0.88 | 0.87 | 0.92 | 0.88 | 0.97 | 0.88 | 0.89 | 0.87 | | 0.91 | 0.89 |
| AUC score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.89 | 0.88 | 0.89 | 0.88 | 0.90 | 0.88 | 0.88 | 0.88 | 0.95 | 0.90 | 0.99 | 0.89 | 0.90 | 0.88 | | 0.93 | 0.90 |

* The tuned random forest model is with 88.56% in terms of the training accuracy but on the test accuracy it goes down to 82.53% The tuned Bagging model is also an overfit model with 96.06% training accuracy and 82.75% test accuracy. Again, the tuned gradient boosting model holds to be the best model with 86.87% train accuracy and 84.06% test accuracy.
* The precision is highest on the train data with 0.96 and test data with 0.88. The tuned gradient boosting model is the best with a train and test precision of 0.89
* The tuned Random Forest model has a train and test AUC of 0.95 and 0.90 whereas the gradient boosting model has a train and test AUC of 0.93 and 0.90 respectively.
* The tuned Gradient boosting model is chosen as the best model for predicting the class 1 in the target variable.

**For the Class 0 in the target variable: Conservative Party**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic regression | | Linear discriminant Analysis | | K-Nearest Neighbors | | Naïve Bayes | | Random Forest Classifier | | Bagging | | ADA Boosting | | Gradient Boosting | | |
| Accuracy in % | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 83.97 | 81.87 | 83.78 | 82.53 | 83.97 | 82.53 | 83.22 | 82.09 | 88.56 | 82.53 | 96.06 | 82.75 | 83.88 | 81.44 | | 86.87 | 84.06 |
| Precision | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.77 | 0.69 | 0.76 | 0.70 | 0.76 | 0.69 | 0.73 | 0.67 | 0.85 | 0.70 | 0.96 | 0.69 | 0.76 | 0.67 | | 0.81 | 0.71 |
| Recall | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.69 | 0.65 | 0.70 | 0.67 | 0.70 | 0.70 | 0.72 | 0.73 | 0.77 | 0.67 | 0.91 | 0.70 | 0.70 | 0.68 | | 0.75 | 0.74 |
| F1-score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.73 | 0.67 | 0.73 | 0.69 | 0.73 | 0.69 | 0.73 | 0.70 | 0.81 | 0.69 | 0.93 | 0.70 | 0.73 | 0.68 | | 0.78 | 0.72 |
| AUC score | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | Tr | Ts | | Tr | Ts |
| 0.89 | 0.88 | 0.89 | 0.88 | 0.90 | 0.88 | 0.88 | 0.88 | 0.95 | 0.90 | 0.99 | 0.89 | 0.90 | 0.88 | | 0.93 | 0.90 |

* For this class 0 as well, the scores are very good on the Gradient boosting model on all the parameters.
* Specificity = TN/(TN+FP) . For the Gradient boosting model, on the test data, TN = 96; FP = 34

Specificity = 96/(96+34) = 0.7384 i.e. The model evaluates the class 0: 73.84% = 74% correctly on unseen data. For the train data, the specificity values are = 249/332 = 75%

**Insights and recommendations:**

* The people who have voted for the Labour party are mostly in favour of European integration or neutral to it. The people who have voted for the conservative party have exhibited a strong opposition towards European integration.
* Many people who have voted for the Labour party have rated high for Hague and there are many people who have voted for the conservative party, having rated high for Blair.
* There is a class imbalance in this data (1063 for Labour and 462 for Conservative) but SMOTE cannot be applied on this as this is an election dataset. The whole purpose of the election is to find out which party has majority votes and any form of data manipulation in terms of treating the target classes can be treated as rigging.
* The tuned Gradient Boosting model has proven to bring out the best of results in predicting for the target classes ( for both Labour and Conservative).
* There are many people who have rated 0 on their knowledge on the party’s position on European Integration and have yet voted for their respective parties.
* Scaling the age variable is not preferred as much as binning the age variable.
* The top three features in predicting which party a voter will vote for are: Hague, Europe and Blair.

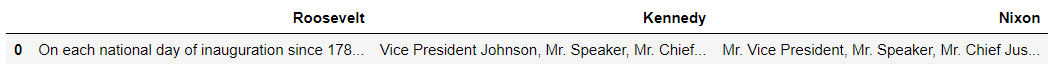
**Problem 2:**

We are required to Analyze the speeches from the following people:

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

We first load the 3 speeches as a list on the dataframe from inaugural of the nltk module in python.

The head of the data looks like this:



**To get the count of words for Roosevelt:**



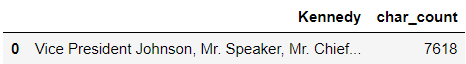
**To get the count of characters for Roosevelt:**



**To get the count of words for Kennedy:**



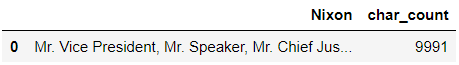
**To get the count of characters for Kennedy:**



**To get the count of words for Nixon:**



**To get the count of characters for Nixon:**



**Lower case conversion for Roosevelt:**



**Removal of Punctuation for Roosevelt:**



**Removal of StopWords for Roosevelt:**



**Common words for Roosevelt:**

We have extracted the 30 most common words in the speech of Roosevelt:



We will remove words: know, us, must from the list as they are also stop words in this context.

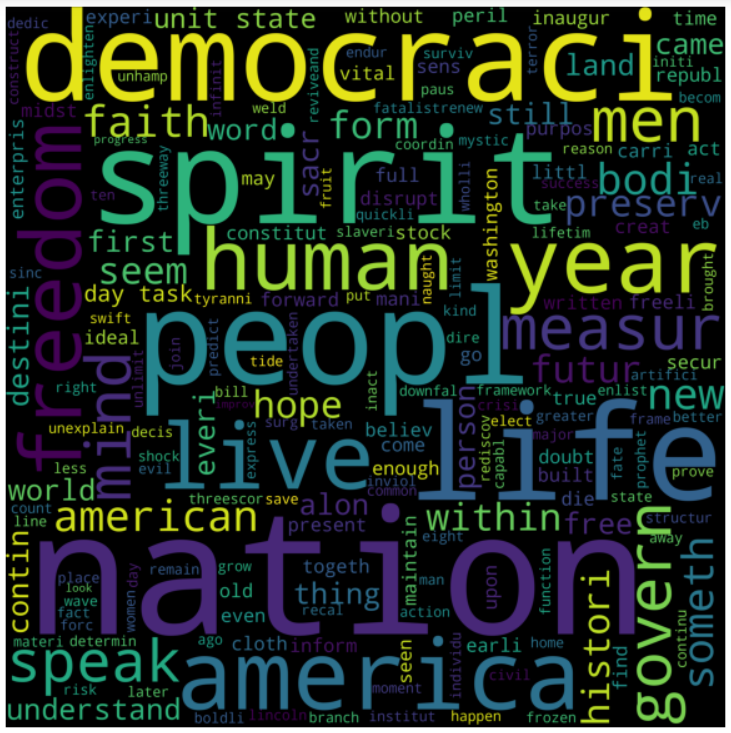
Nation, democracy and spirit are the three most common words in the inaugural speech of Roosevelt.

**Stemming for Roosevelt:**

We then perform stemming on the speech of Roosevelt:



**Wordcloud for Roosevelt’s speech after cleaning:**



**Lower case conversion for Kennedy:**



**Removal of punctuation for Kennedy:**

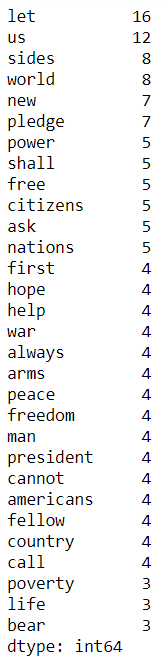


**Removal of Stopwords for Kennedy:**



**Common words removal for Kennedy:**

We then look at the 30 most common words:



We then remove the words: let and us from the speech as they are stop words.

The top 3 common words in the speech of Kennedy are: sides, world and new.

**Stemming for Kennedy:**

We then perform stemming on the speech of Kennedy.



**Wordcloud for Kennedy:**



### Lowercase conversion for Nixon:



**Removal of Punctuation for Nixon:**



### Removal of Stopwords for Nixon:



**Common words for Nixon:**



We will then remove the words: let, us, shall, make and come from the above list as they are stop words.

The top three most common words in the inaugural speech of Nixon are: peace, world and new.

**Stemming for Nixon:**

We then perform stemming on the speech for Nixon:



### Word cloud for Nixon:

### 