**Business Report of Machine Learning Project**

**BUSINESS REPORT PROB:1**

You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

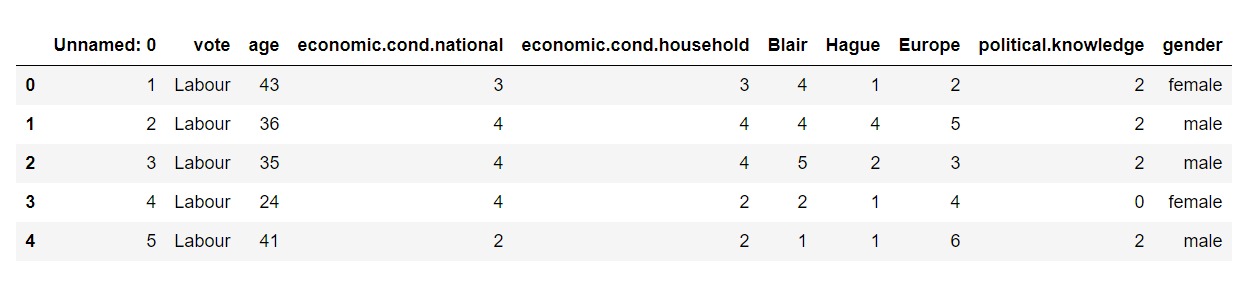
**EXECUTIVE SUMMARY**

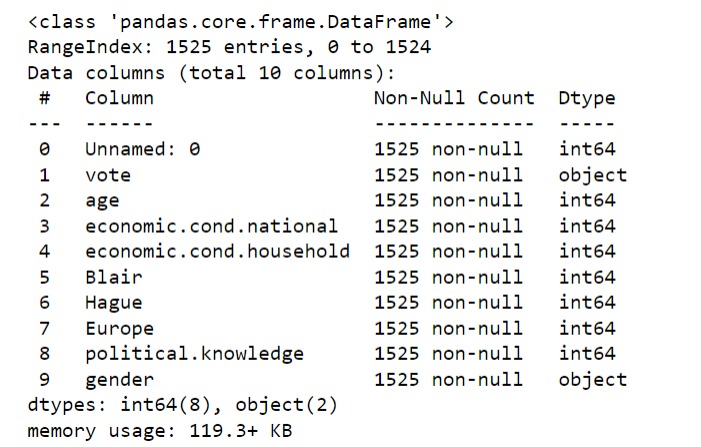
A Leading News Channel CNBE. wants to analyze the recent elections. The dataset consists of 1525 voters with 9 variables. Based on the different attributes of the dataset we have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis , visualization & apply various supervised learning algorithms i.e. Logistics Regression , LDA ,KNN & Navie Bayes to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party. Explore the dataset using central tendency and other parameters. The data consists of 1525 different voters with 9 unique activities . Analyse the different attributes of the voters which can help in predicting which party a voter will vote for on the basis of the given information.This assignment should help the news channel to create an exit poll that will help in predicting overall win and seats covered by a particular party.

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

****



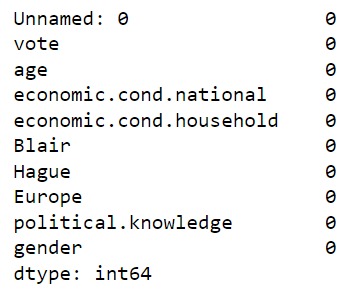
**Observation**

* The election dataset has 1525 rows and 10 columns
* All the variables are integer(int64) except 'vote' and 'gender' is object
* We can observe in the dataset looks like all the variables are categorical except 'age' column
* Removing the unwanted variable "Unnamed: 0" ,which is not giving a meaningful information.

**Data Dimensions**

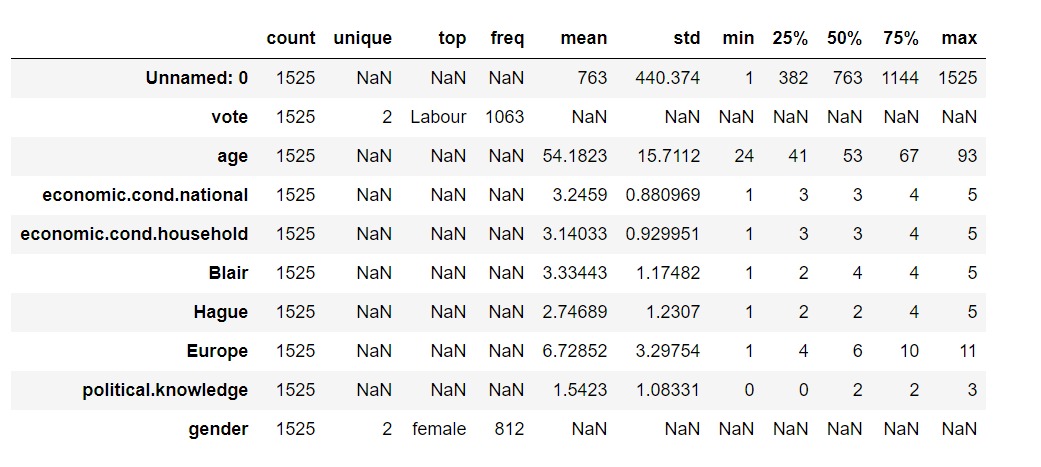
* The number of rows in the dataframe is 1525
* The number of columns in the dataframe is 10

**Checking the missing values**



* There are no missing values in the dataframe

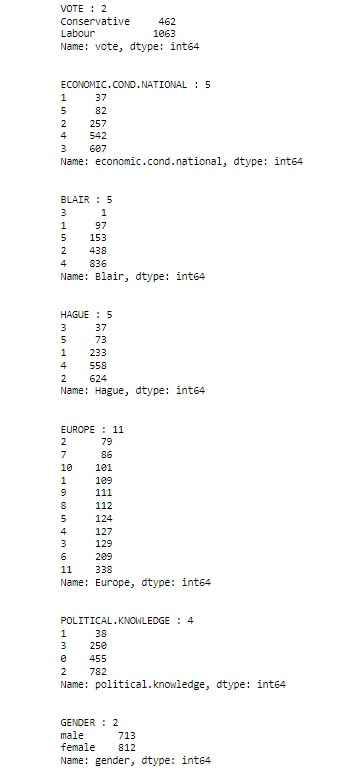
**Descriptive Analysis for the dataset**

****

**Observation**

* As we can observe that dataset looks good
* The mean and median for the only integer column 'age' is almost same indicating the column is normally distributed

**Getting the unqiue count for categorical variables in the dataset**

****

**Observation:**

* 'Vote' has two unique values which is " Labour" and " Conservative", which is also a dependent variable.
* 'gender' has two unique values 'male' and female'.
* 'Europe' being the highest have 11 unique values

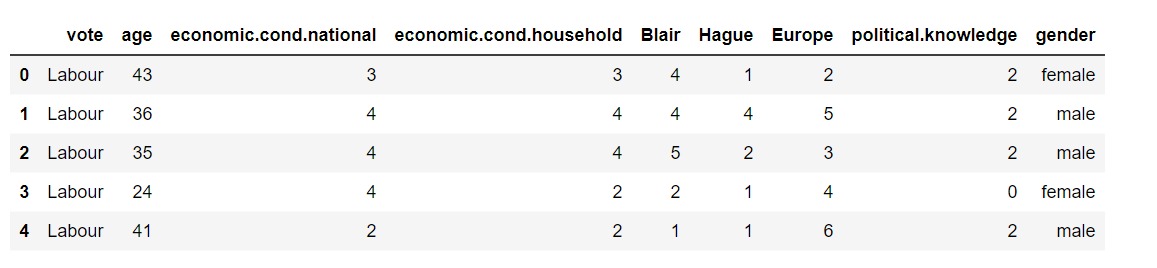
**Checking duplicates**

****

**Observation:**

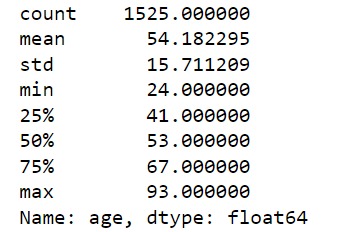
No duplicates in the dataset

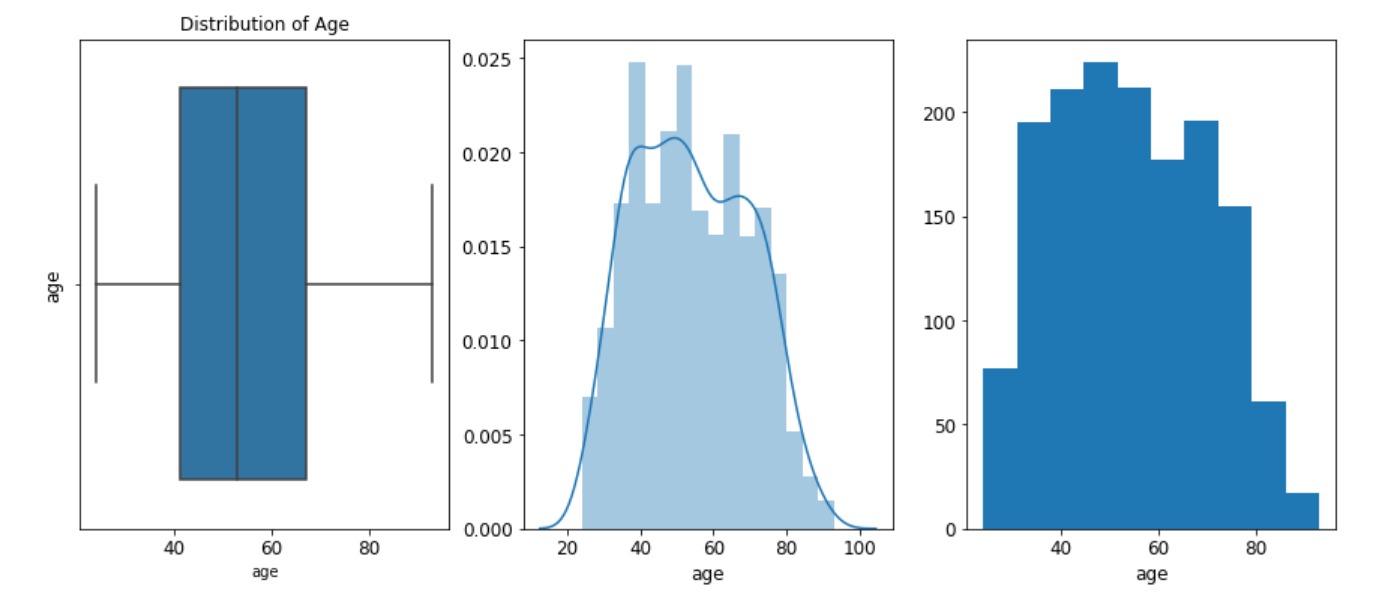
**Removing or dropping the "Unnamed: 0"**

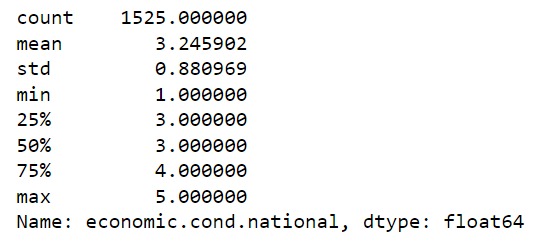
****

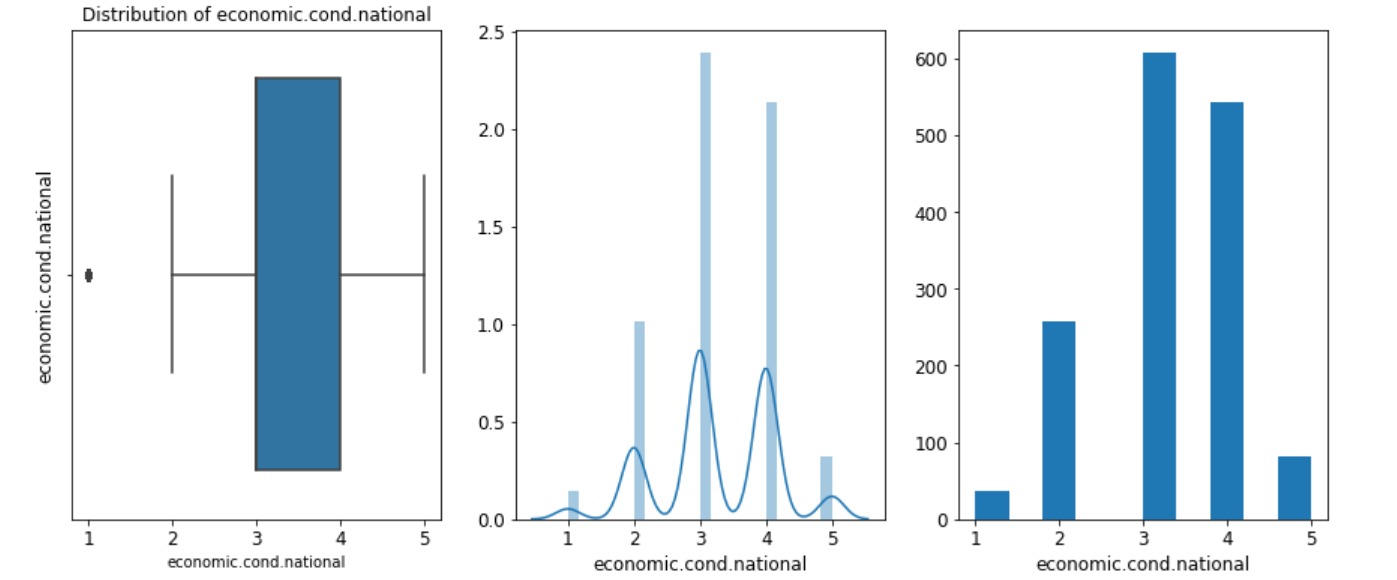
**1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.**

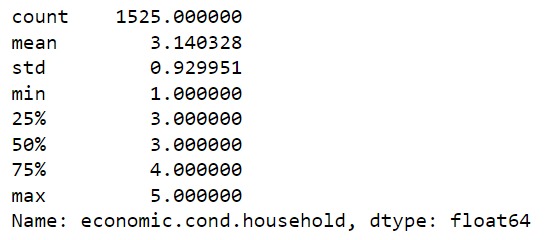
**Univariate Analysis**

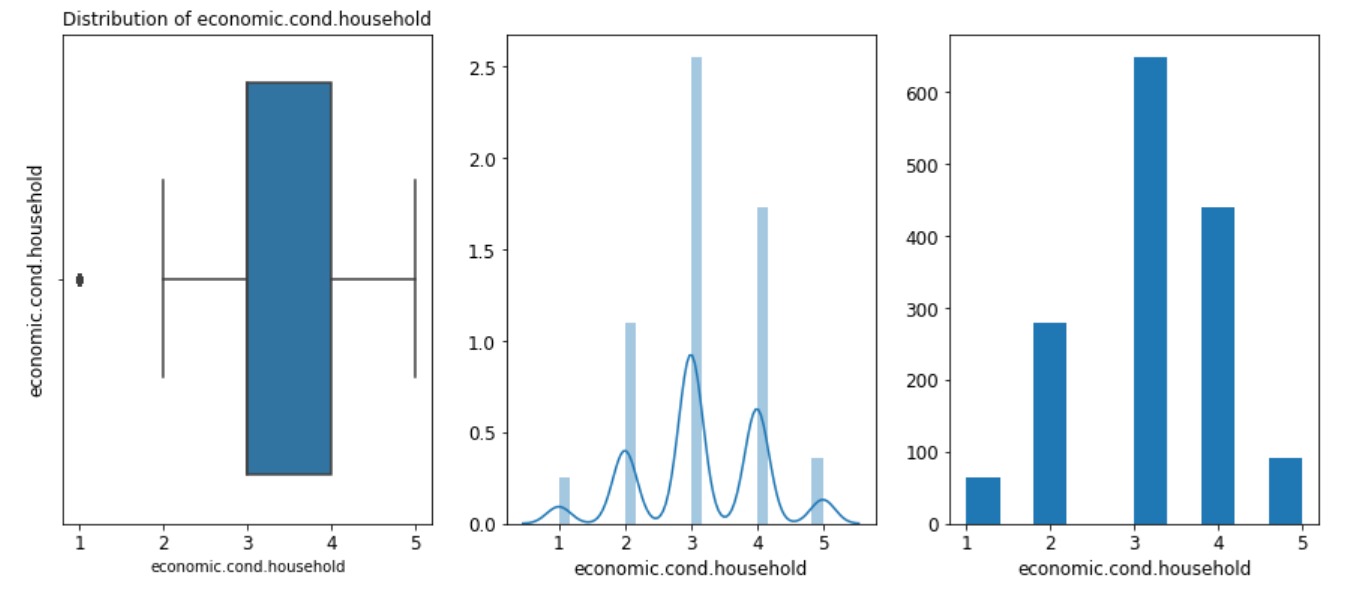
****

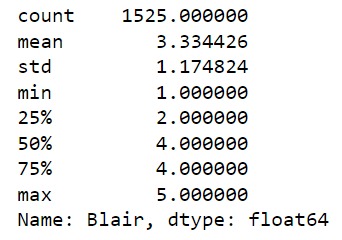
****

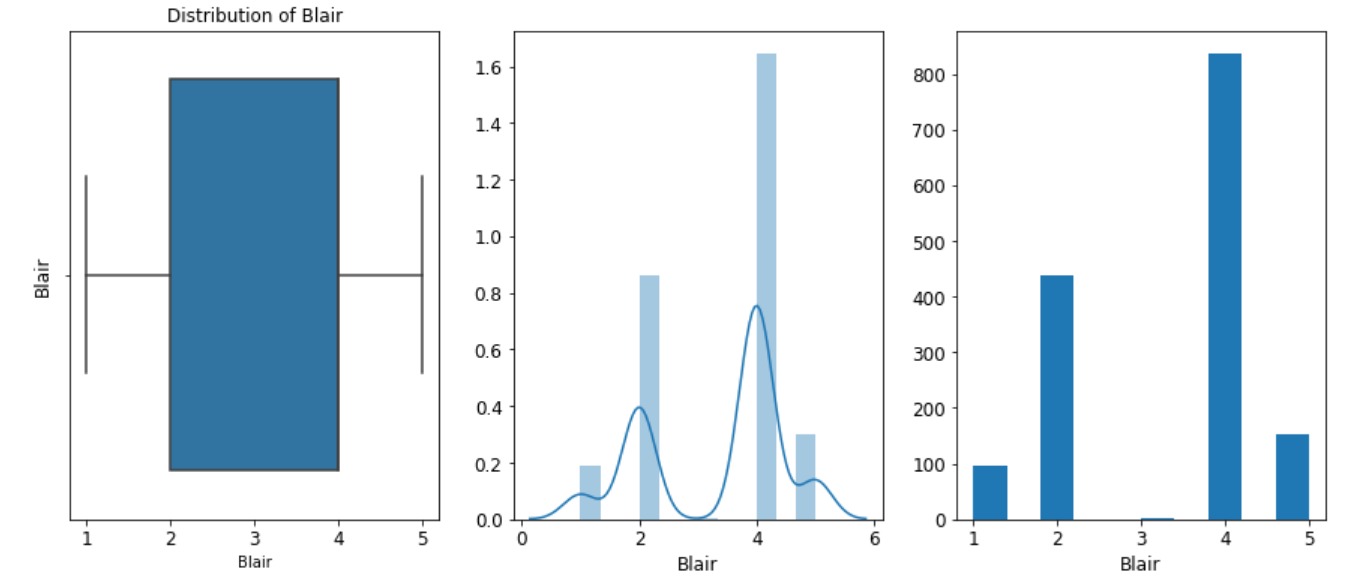


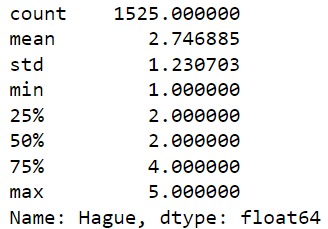


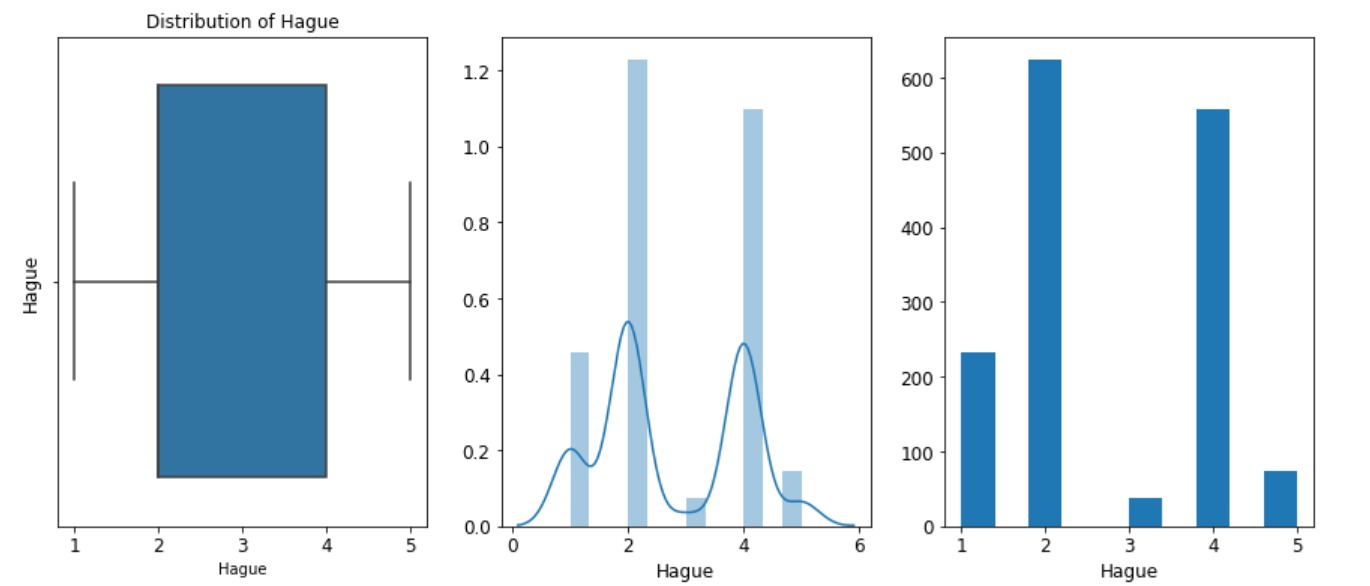


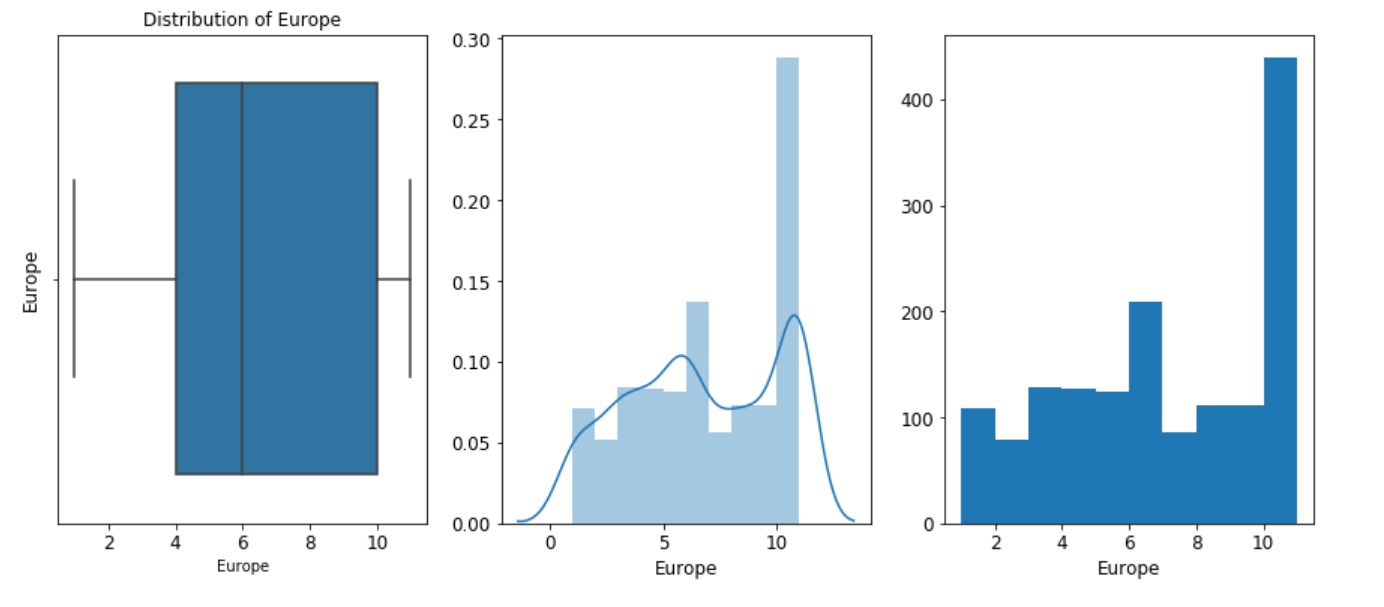
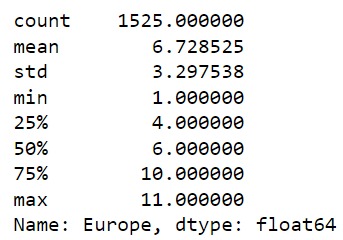


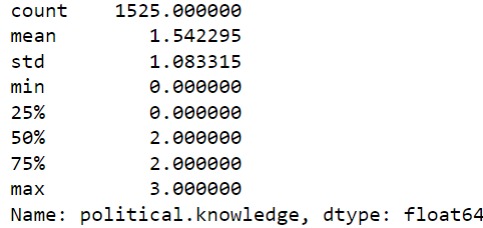


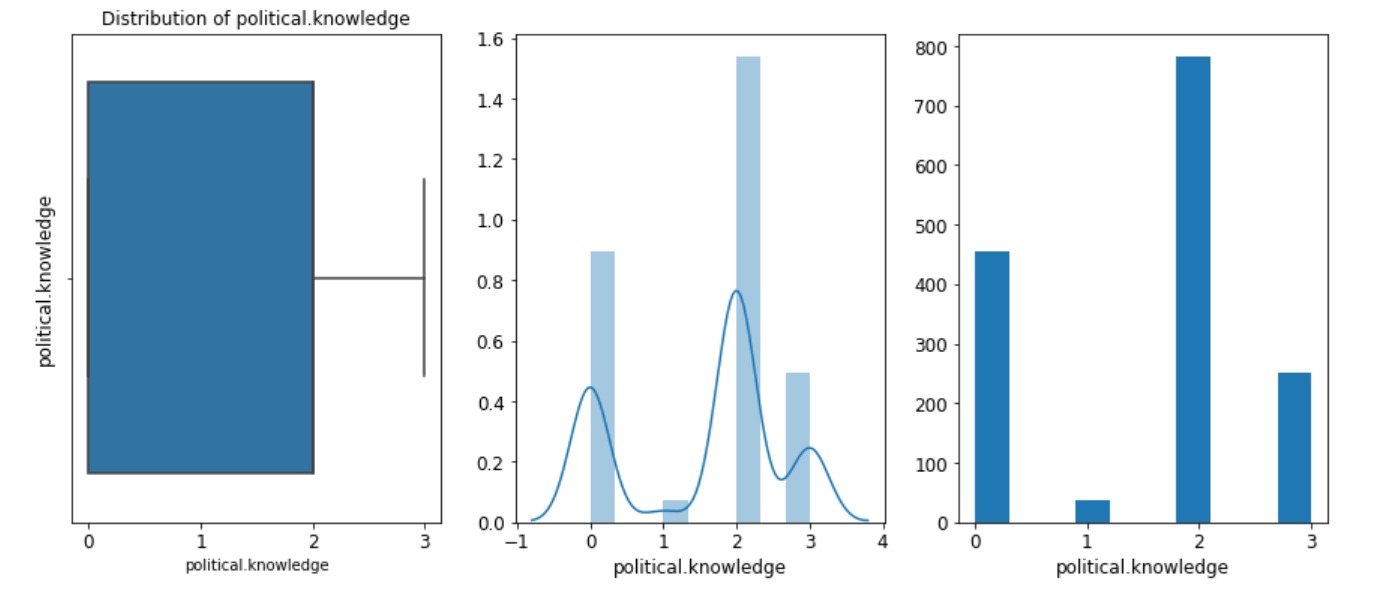




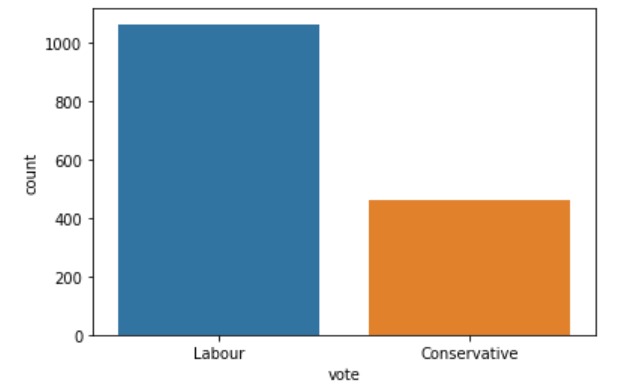


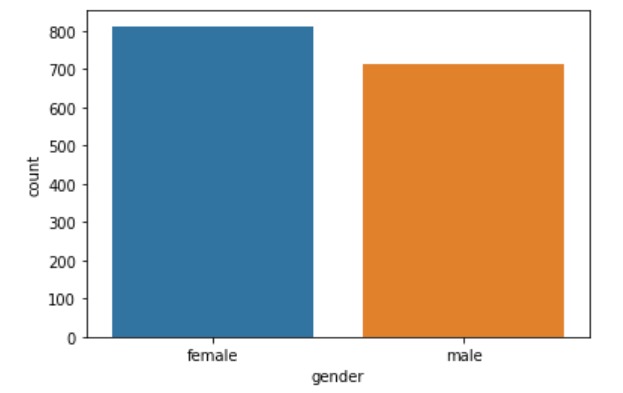






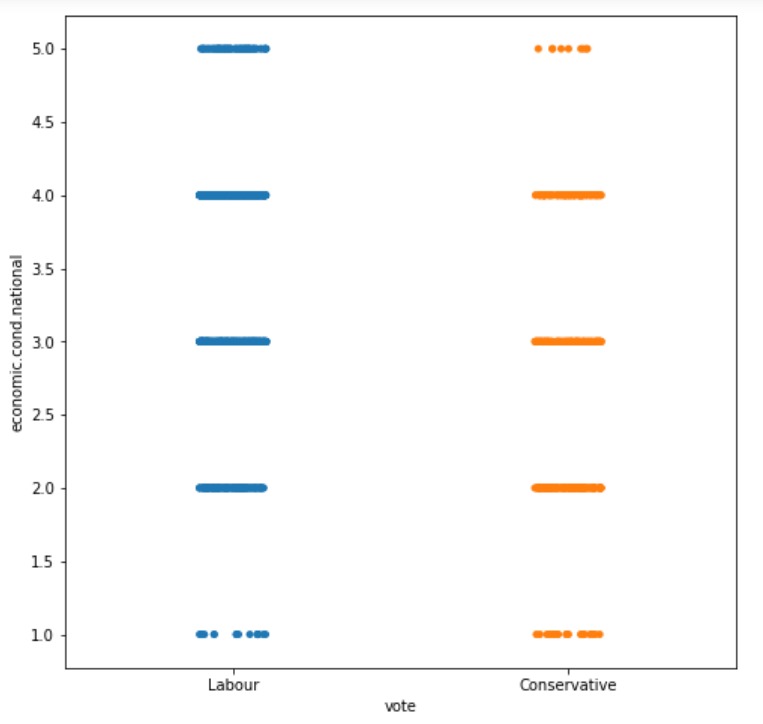
**Categorical Variable**

****

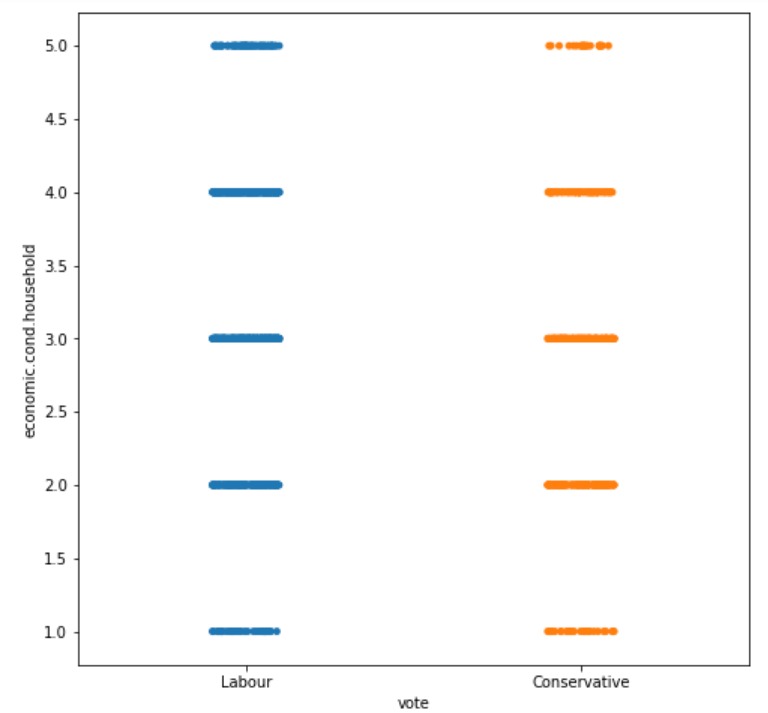


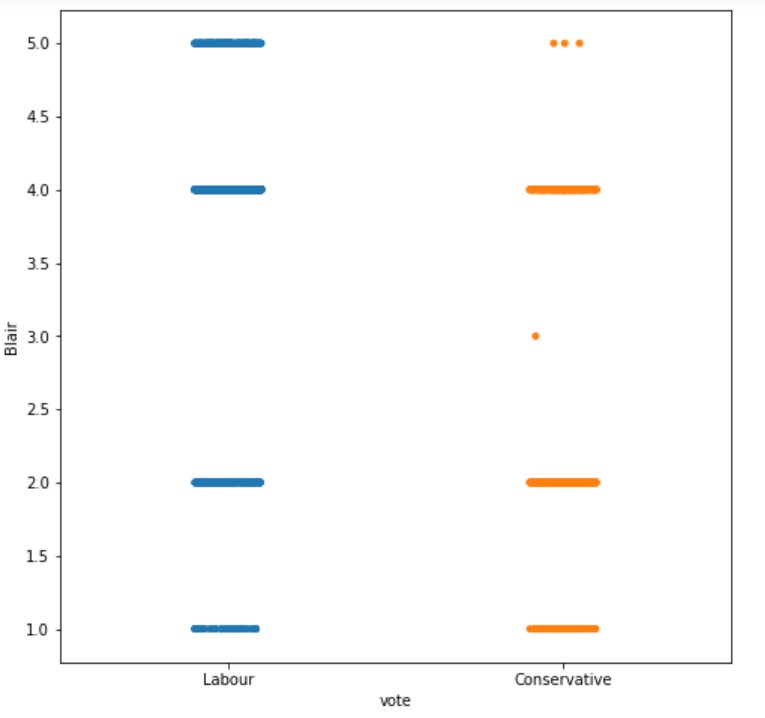
**Bivariate Analysis**

'vote' and 'economic.cond.national'

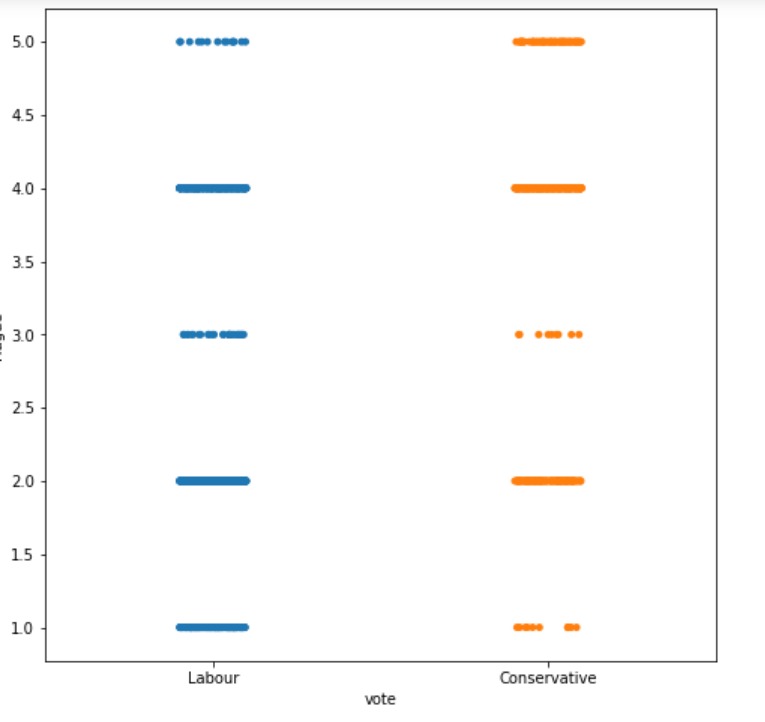
****

'vote' and 'economic.cond.household'

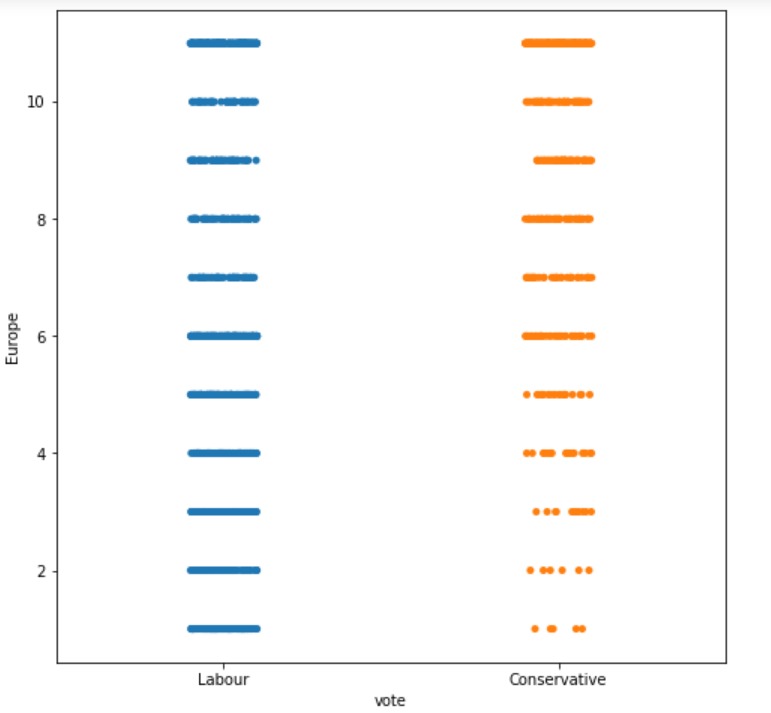


'Vote' and 'Blair

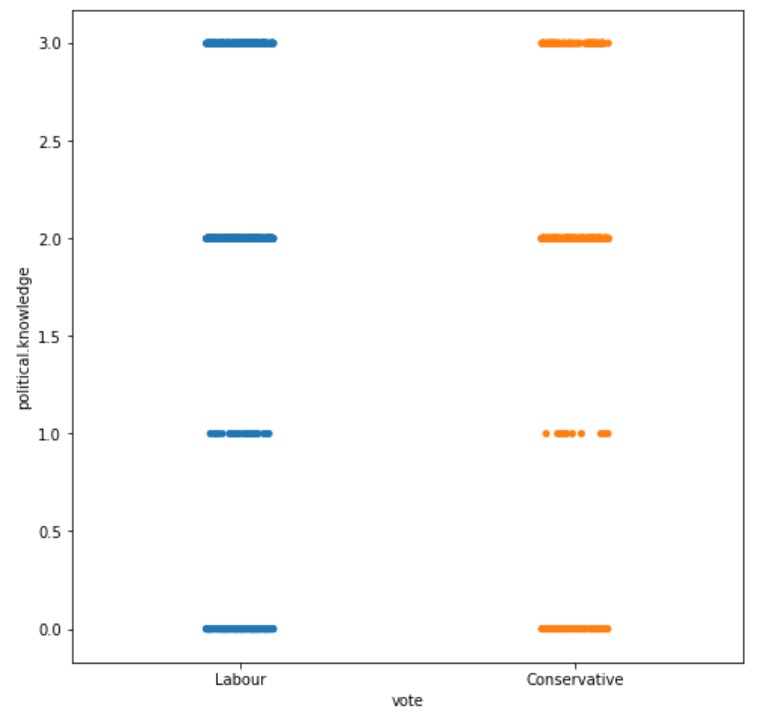
'Vote' and 'Hague



'Vote' and 'Europe'



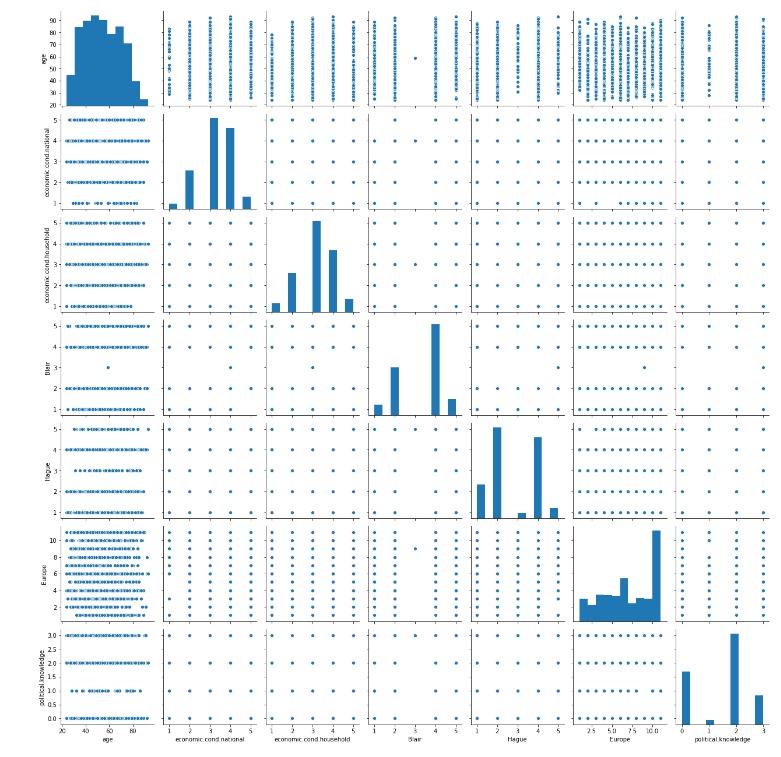
Vote' and 'political,knowledge'



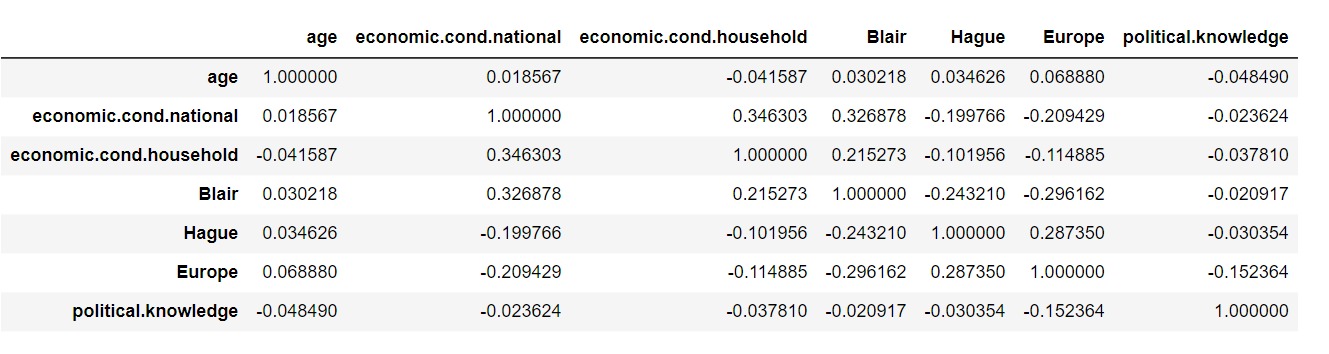
**Observation**

* The average 'age' people have voted in the election is 54.18 and the distribution is normal
* The average Assessment of current national economies is 3.29 and is categorical type and with dependent variable 'vote' of 5 have voted less for 'Conservative' who has current economic rating of 5 & 'Labour' who has current economic rating of 1
* The average Assessment of current national economies is 3.14 and is categorical type and with dependent variable 'vote' of 5 have voted less for 'Labour' who has less voted in 1.
* The 'Blair' assessment of the leader labour which has 3 rating have got only one vote tp by 'Conservative' and np votes for 'Labour' as per scatter plot
* The'Hague' assessment of 'Conservative' leader which has 1 rating 'Conservative' got less votes compared to 'labour'.
* In Europe on the scale of 1 the 'Conservative' have got less votes comared to 'labour' who have got higher votes who have the higher attitude towards european Inegration
* People who have 1 rating of 'political.knowledge' on european integration have less voted for both 'Conservative' and ' labour'
* The most number have voted in 'gender' is female
* The labour has highest votes as per bar plot

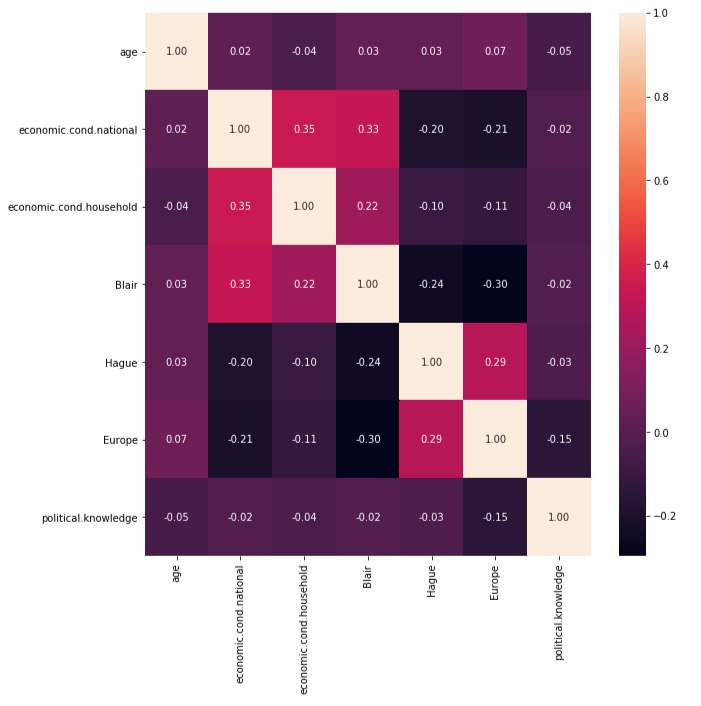
**Pair Plot for all the variables**



**Correlation Matrix**

****

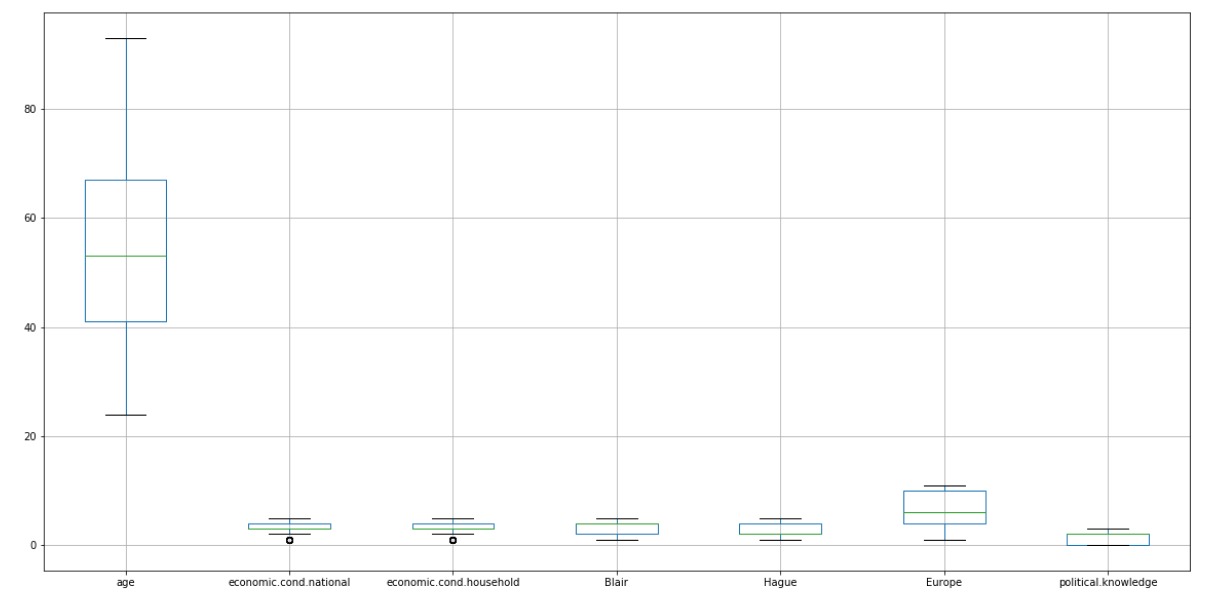
**Heat Map for visualization**

****

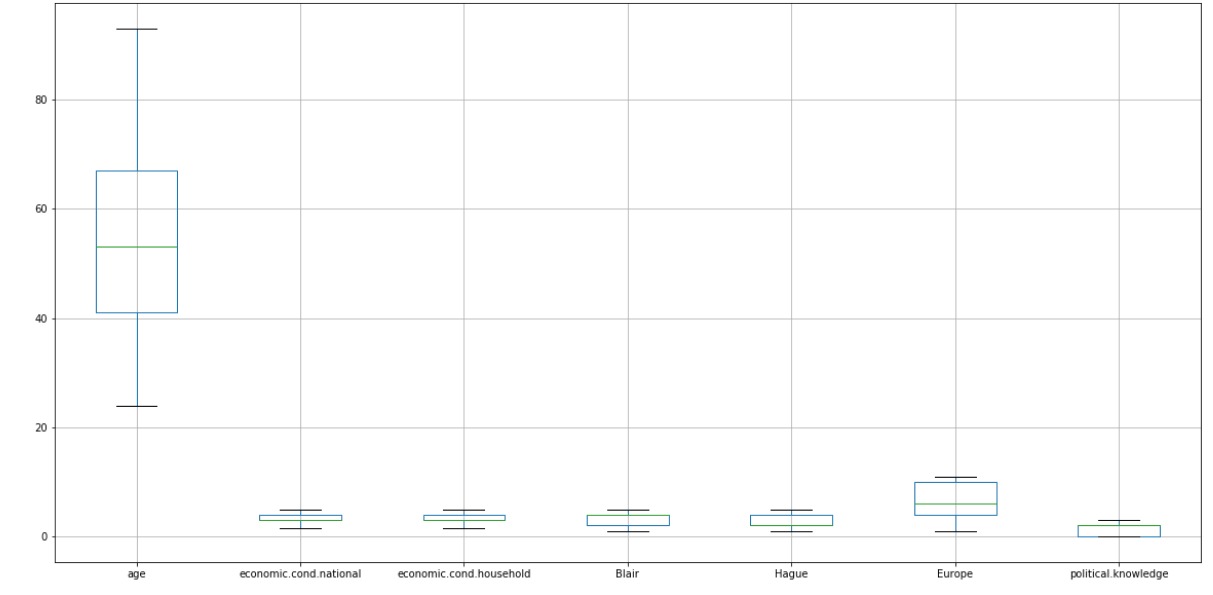
**Observation**

* As referring to pairplot the distribution are normal in 'age' and all the other variables are categorical type
* As referring to heatmap, find below which are strongly correlated -'economic.cond.national' and 'economic.cond.household'
* 'economic.cond.national' and 'Blair'

**Outliers checking**

****

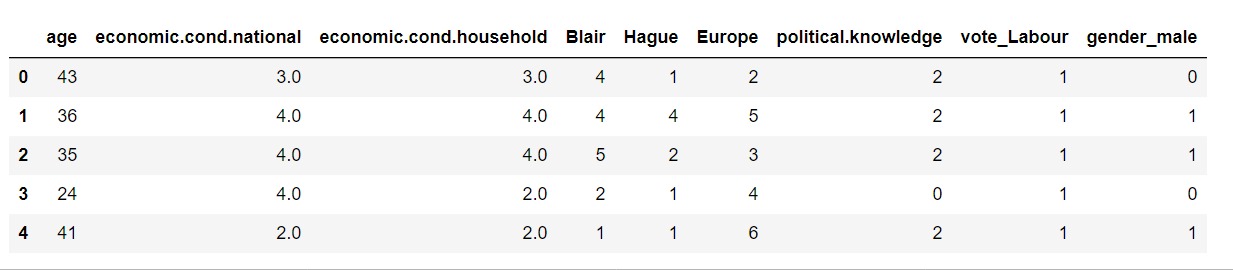
**After Treating Outliers**

****

**Observation** T

* he data have some outliers presence in the 'economic.cond.national' and 'economic.cond.household' . we are going treating because in Gaussian Naive Bayes, outliers will affect the shape of the Gaussian distribution and have the usual effects on the mean etc.
* After treating the outliers have been removed

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

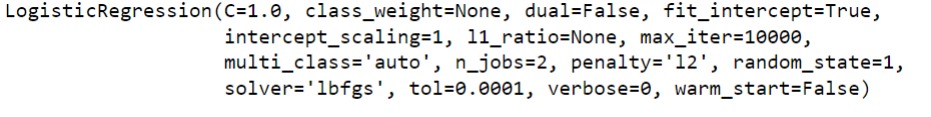


Is scaling necessary here or not?

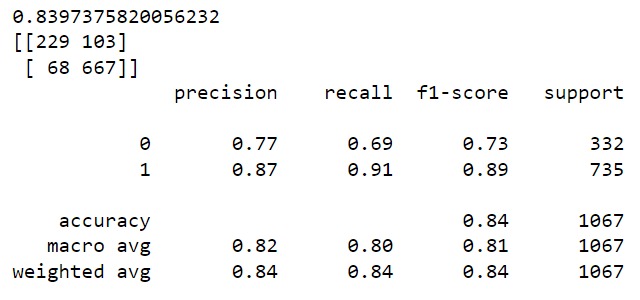
We are not going to scale the data for Logistic Regression & LDA - it's coefficients using the variation between the classes (check this), so the scaling doesn't matter either. Naive Baye's models as it is not necessary, But in case of KNN it is necessary to scale the data as it is distance based algorithm (typically based on Euclidean distance).scaling the data gives similar weightage to allt he variables

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

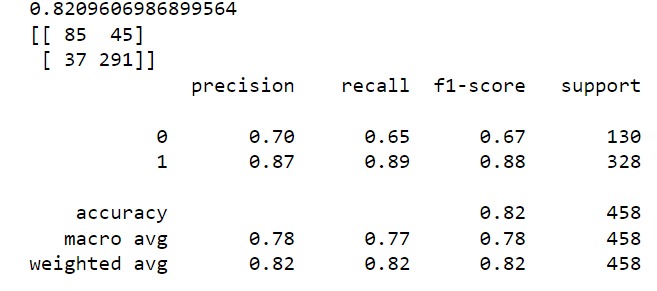
Logistic regression

****

**Performance matrix on train data**



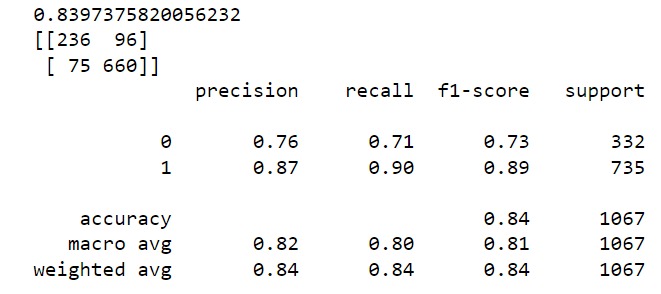
**Performance matrix on test data**



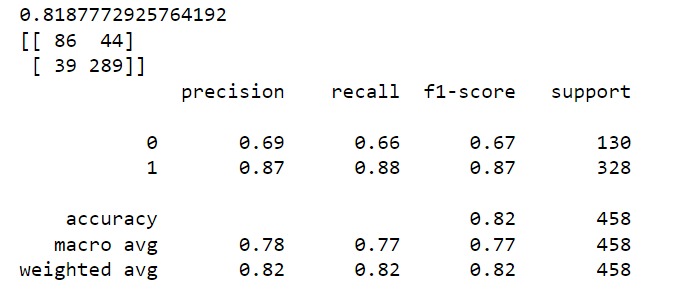
**Linear Discriminant Analysis**

****

**Performance Matrix on train data**

****

**Performance Matrix on test data**

****

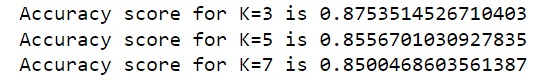
**Observation:**

* The model is not overfitting or underfitting.
* Training and Testing results shows that the model is excellent with good precision and recall values.
* Training and Testing results shows that the model is excellent with good precision and recall values.
* The LDA Model is better than Logistic Regression model as for the precision both are the same but the recall score is more LDA model

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

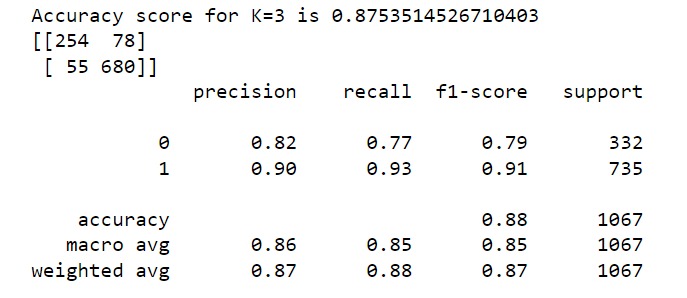
**KNN Model**

**Performance matrix for train data**

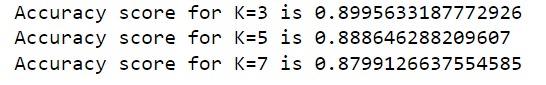
****

**Observation:**

The accuracy score for K=3 has better accuracy compared other value in training data. so we will make build matrix on the value k=3

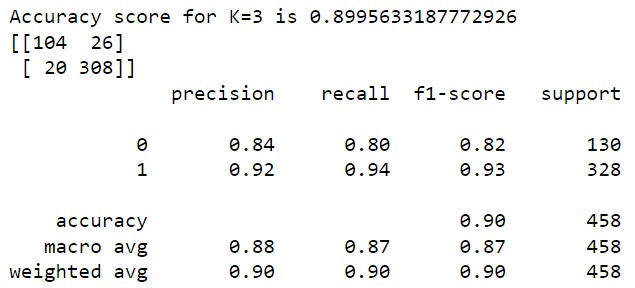
****

**Performance matrix on Test data**

****

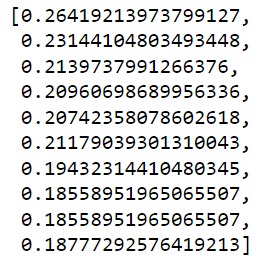
**Observation:**

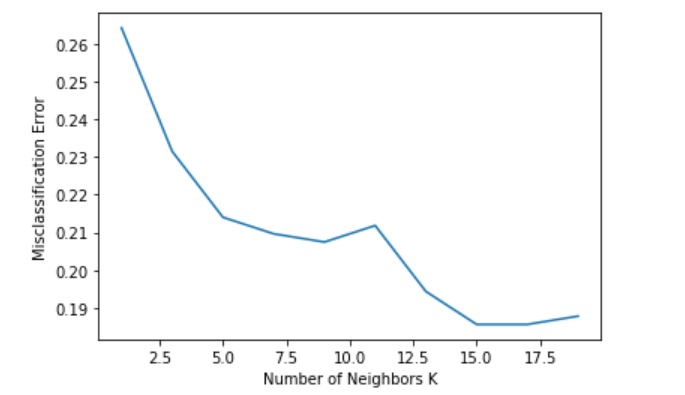
The accuracy score for K=3 has better accuracy compared other value in test data. so we will make build matrix on the value k=3

****

**Add on:**

**As we will check the Mis Classification Error(MCE) in Which model has lowest MCE**

****

****

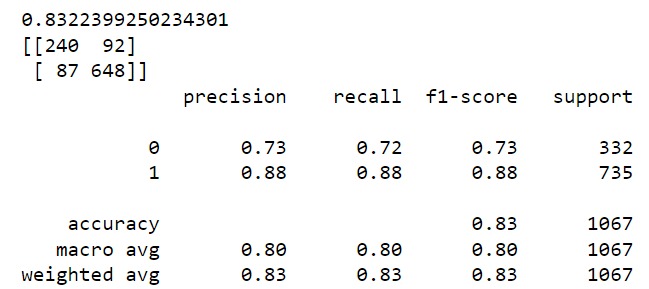
**Observation:**

The Lowest k value having the least MCE score is k=15

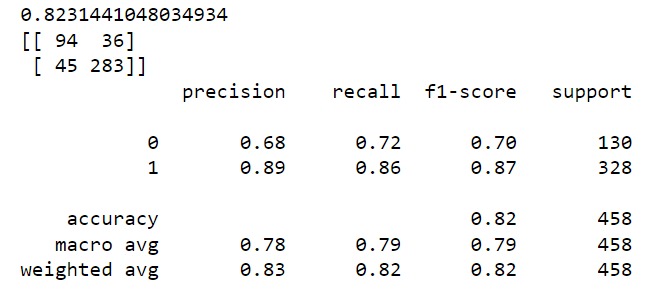
**Naive Bayes**

****

**Performance matrix for train data**

****

**Performance Matrix for test data**

****

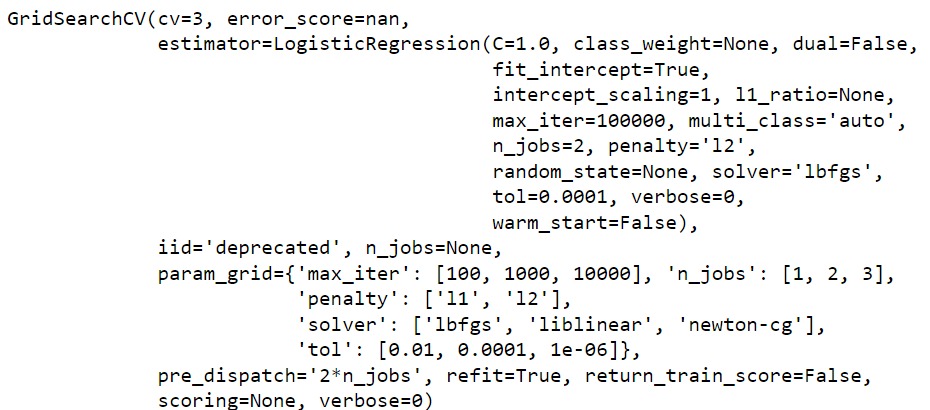
**Observation:**

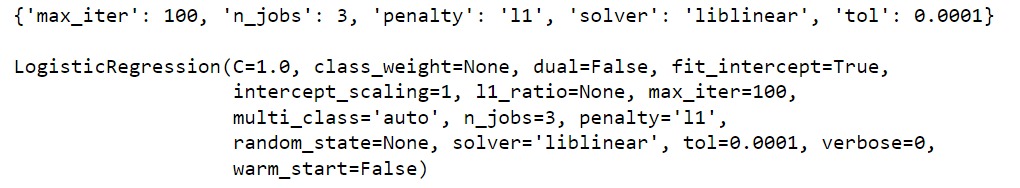
* The model is not overfitting or underfitting.
* Training and Testing results shows that the model is excellent with good precision and recall values. Training and Testing results shows that the model is excellent with good precision and recall values.
* KNN performs better compared to Naive Bayies as per the above accuracy results

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

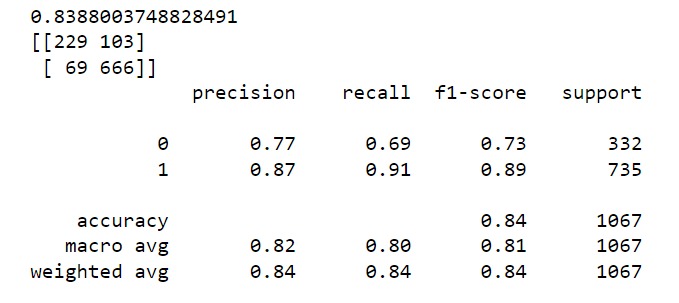
Model Tuning

Invoking the Logistic regression by applying GridSearchCV

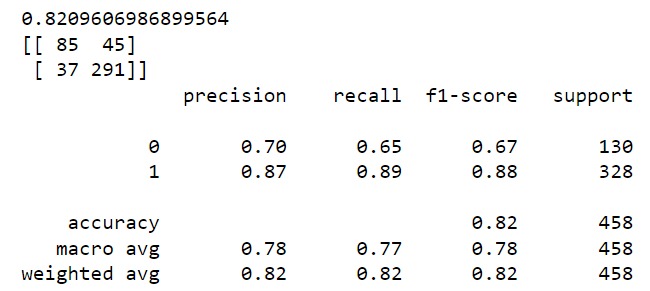


****

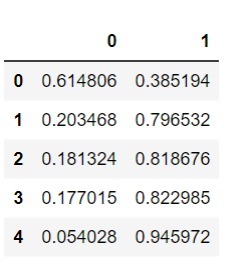
**Performance Matrix of training data after tuning**

****

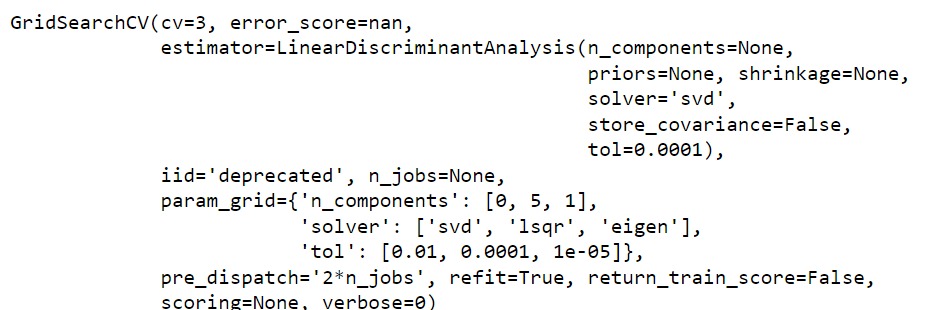
**Performance Matrix test data after tuning**

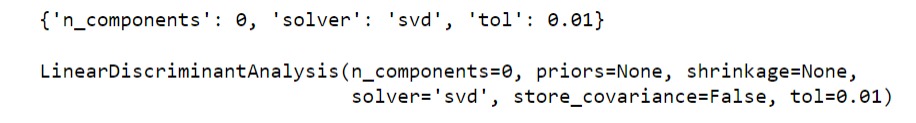
****

**Getting probabilities for Train data Set**

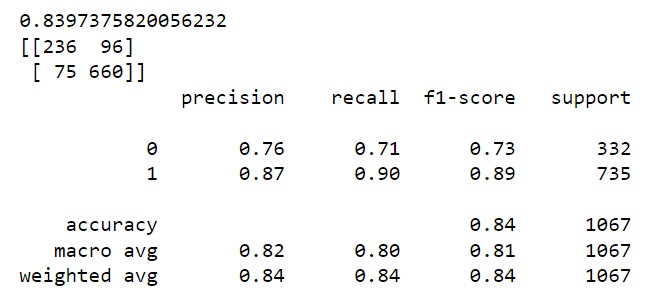
****

**Invoking Linear Discriminant Analysis by applying GridSearchCV**

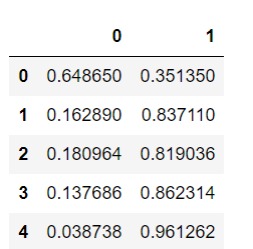
****



**Performance Matrix on training data after tunning**

****

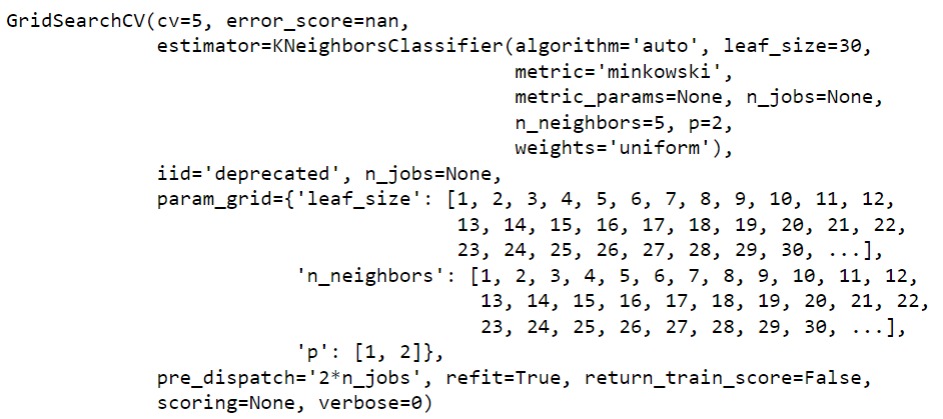
**Performance Matrix on test data after tuning**

****

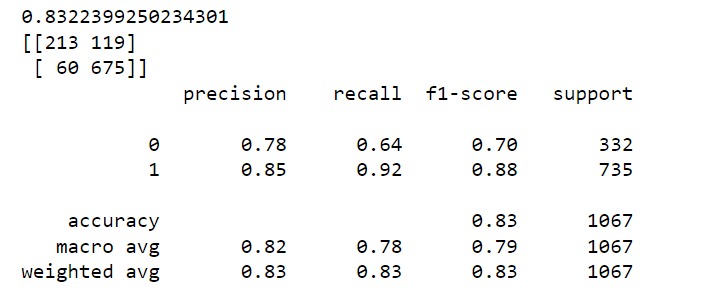
**Observation:**

* After tuning the data, their is no overfitting or underfitting in the model, all the performance matrix result is good
* Their is no much changes in the performance the accuracy and the recall still the same after tuning still the logistic regression performed better compared to LDA model

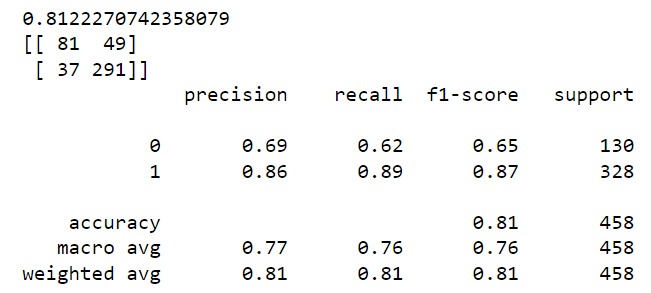
**KNN model**

****

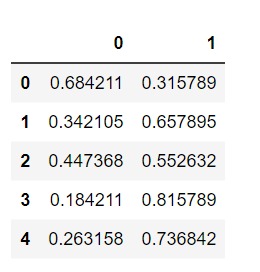
**Performance Matrix on training data after tuning**

****

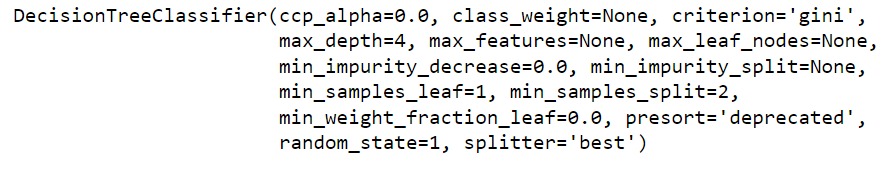
**Performance Matrix on test data after tuning**

****

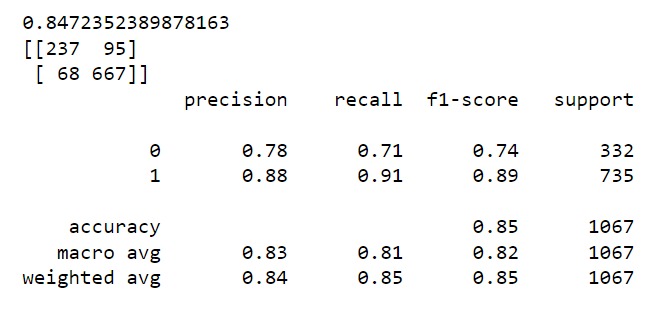
**Getting probabilities for train data**

****

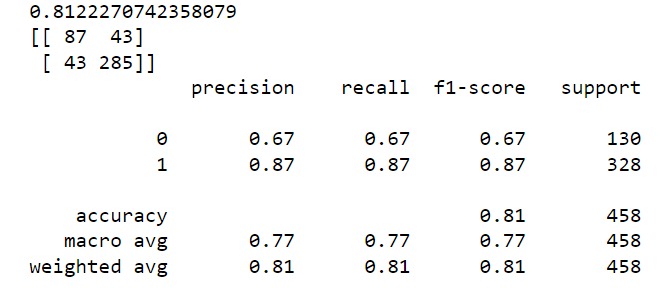
**Decision Tree Classifier**

****

**Performance Matrix on training data**

****

**Performance Matrix on test data after tuning**

****

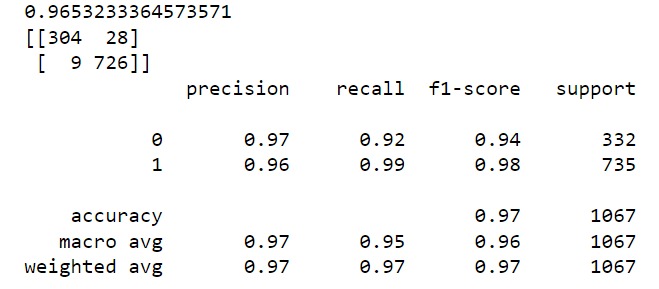
**Observation**

* The Decision Tree classifier looks good , there is no over fitting or underfitting in the model
* The training data performed well on accuracy and recall compared test data

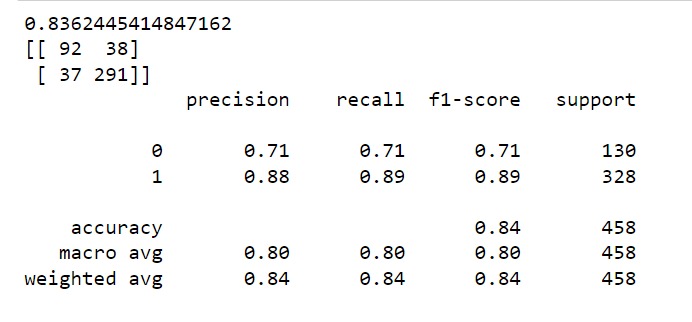
**Bagging**

****

**Performance Matrix on training data**

****

**Performance Matrix on test data**

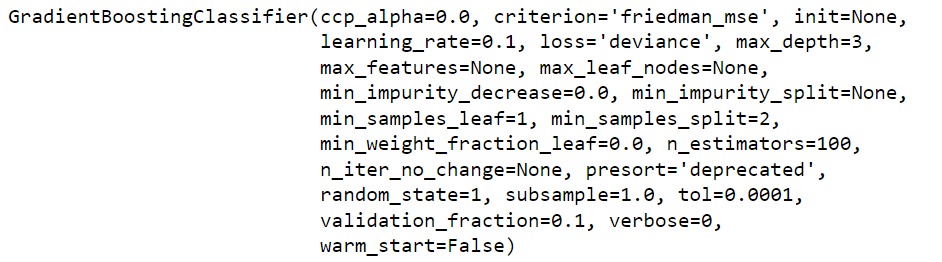
****

**Observation**

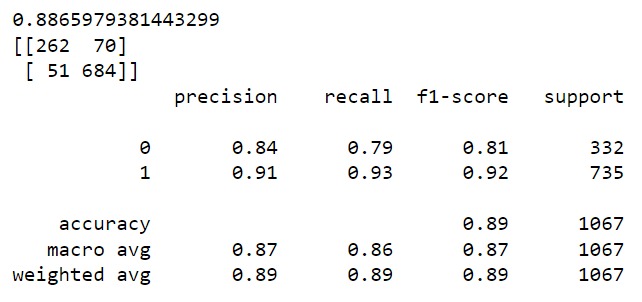
* We carried out the Bagging Classifier by using Random Classifier The accuracy for training set holds good compare to test data
* The data holds good.their is no overfitting and underfitting in model

**Boosting**

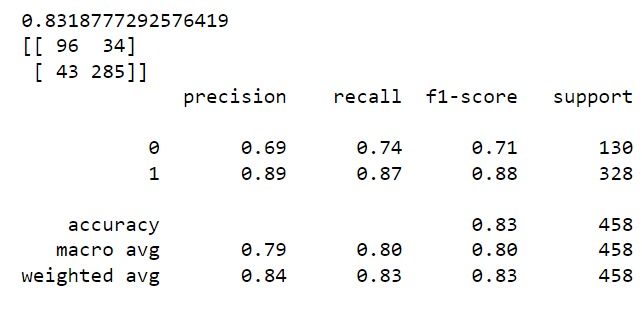
**Gradient Boosting**

****

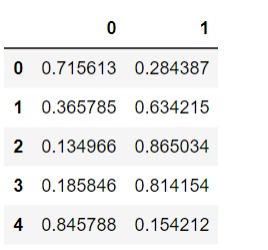
**Performance Matrix for training set**

****

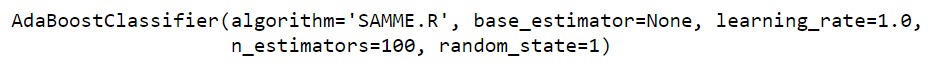
**Performance Matrix for test set**

****

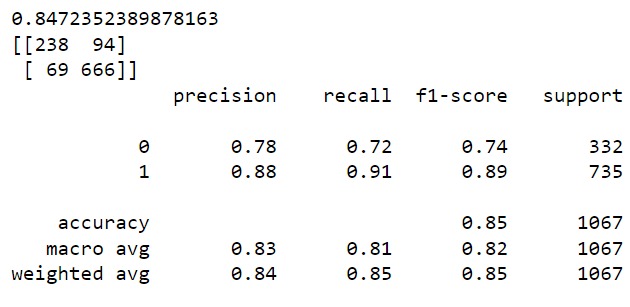
**Getting Probabilities for Training set**

****

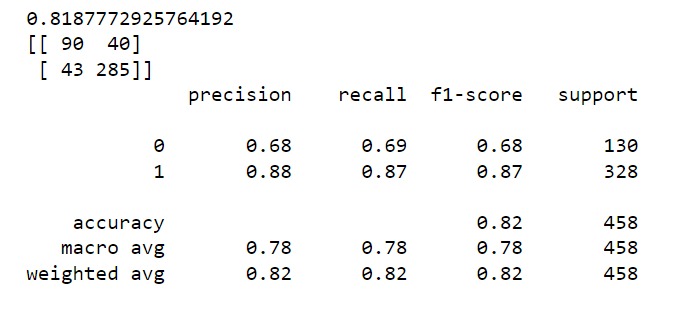
**Ada Boosting**

****

**Performance Matrix for training set**

****

**Performance Matrix for test set**

****

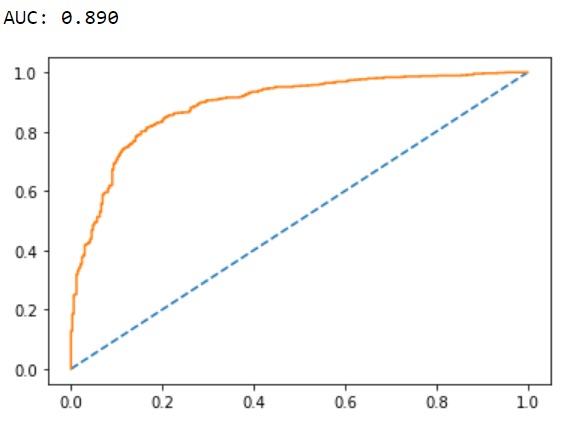
**Observation:**

* As for the Boosting technique we have used Gradient Boosting and Ada Boosting
* Gradient Boosting performs well compared to
* Ada boosting The data is nither overtfitting or underfitting

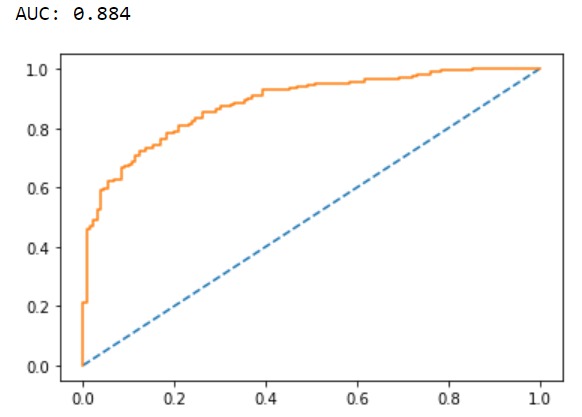
## 1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

**Logistic Regression**

**AUC ROC curve for Logistic Regression Train**

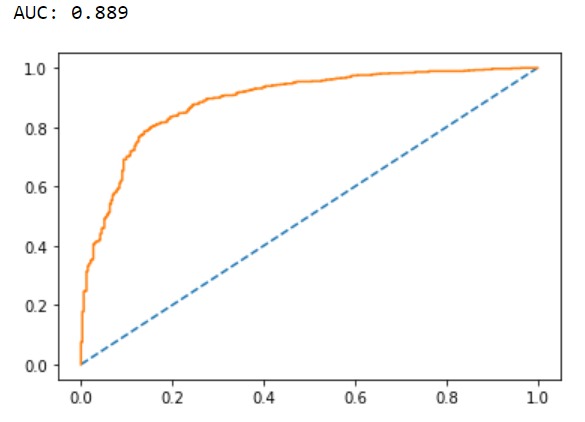
****

**AUC ROC curve for Logistic Regression test**

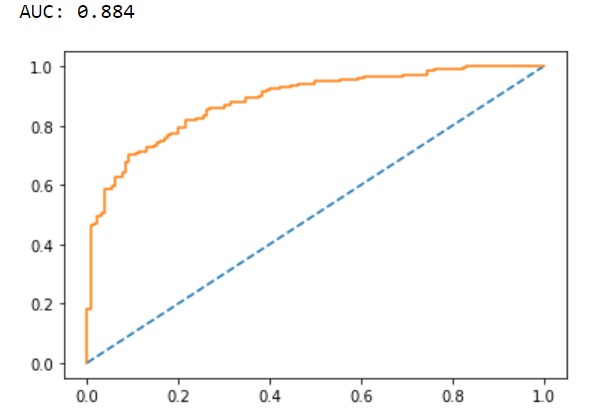
****

**Linear Discriminant Analysis**

**AUC ROC curve for LDA train**

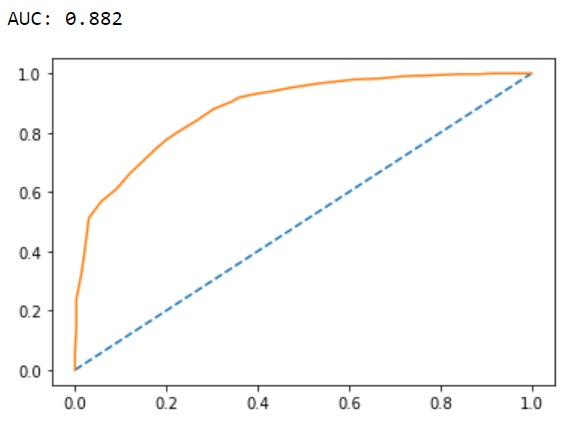
****

**AUC ROC curve for LDA test**

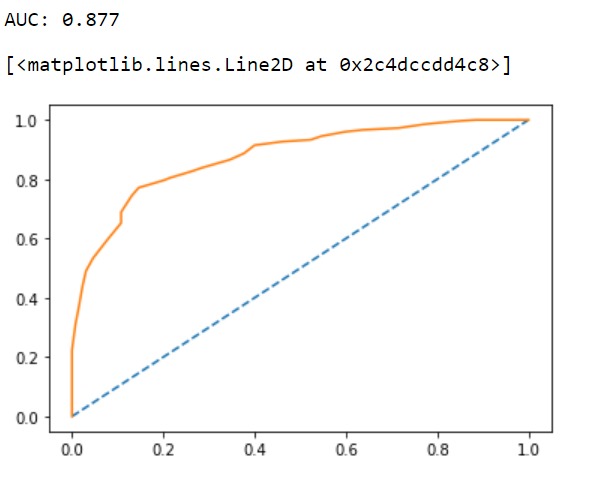
****

**KNN Model**

**AUC ROC curve for KNN train**

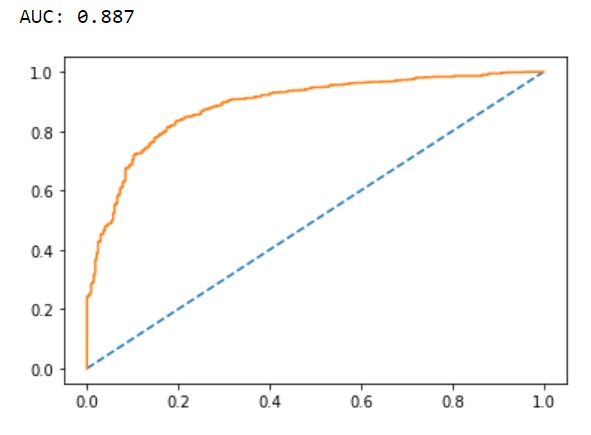
****

**AUC ROC Curve for KNN test**

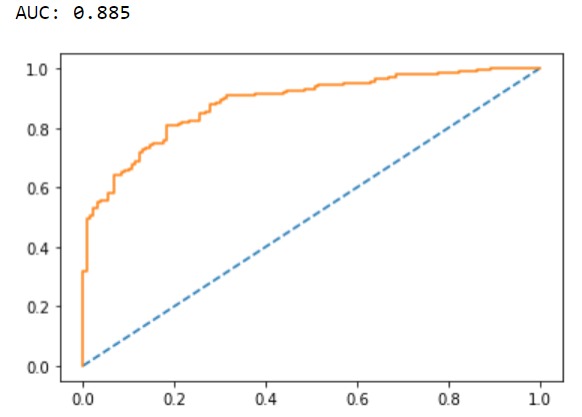
****

**Naive Bayes**

**AUC ROC curve for Naive Bayes Train**

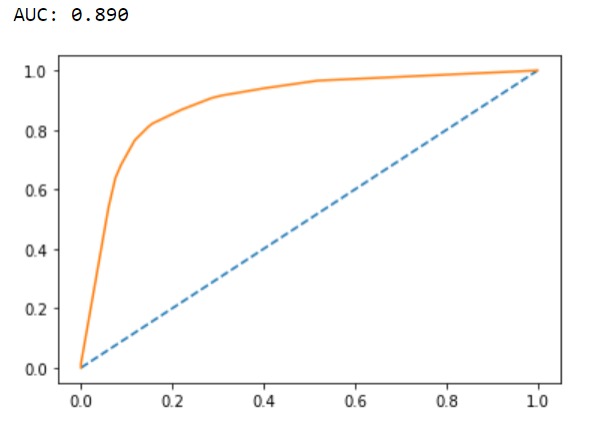
****

**AUC ROC curve for Naive Bayes Test**

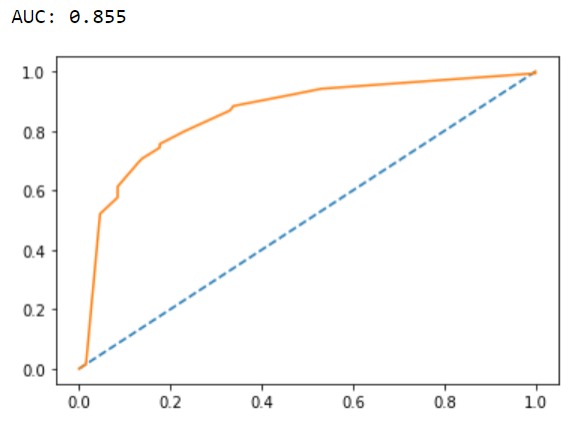
****

**Decision Tree model**

**AUC ROC Curve for Decision Tree Train**

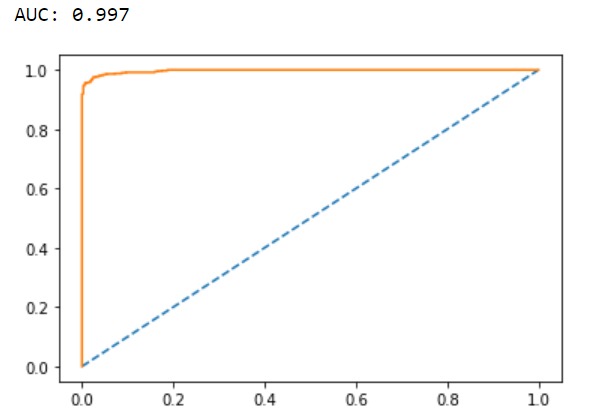
****

**AUC ROC Curve for Decision Tree Test**

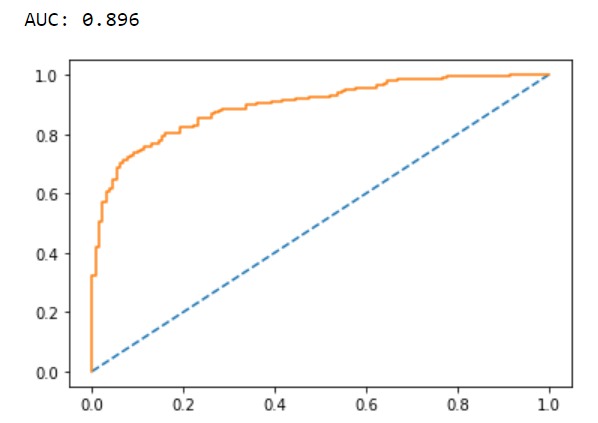
****

**Bagging Model**

**AUC ROC Curve for Bagging model Train**

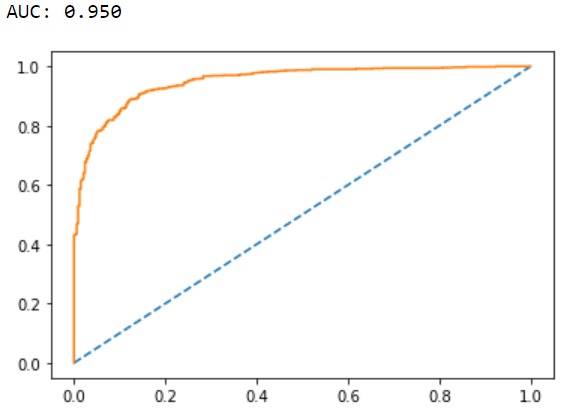


**AUC ROC Curve for Bagging model test**

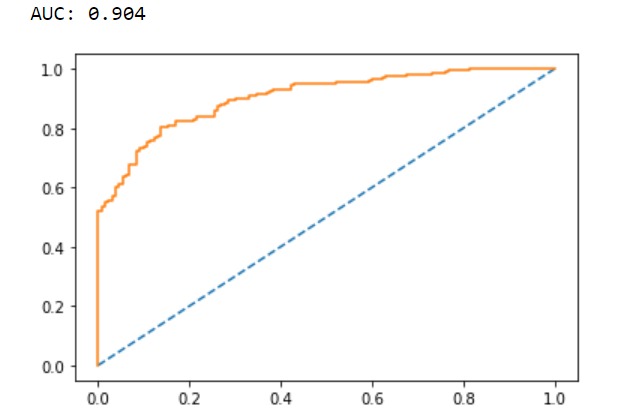
****

**Boosting**

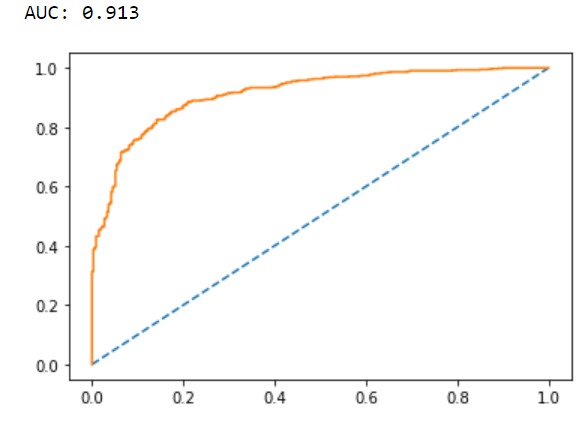
**AUC ROC curve for Gradient Boosting for Train**

****

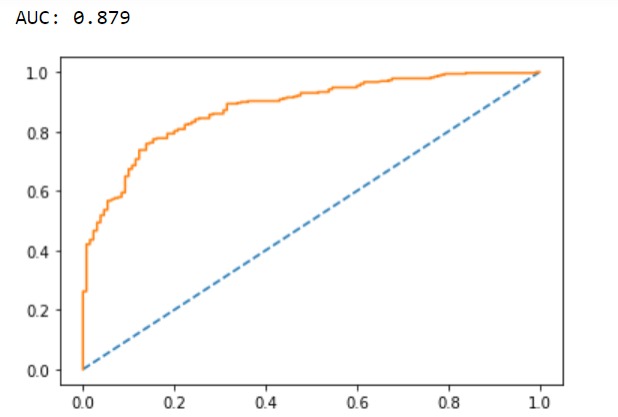
**AUC ROC curve for Gradient Boosting in Test**

****

**AUC ROC Curve for Ada Boosting Train**

****

**AUC ROC Curve for Ada Boosting Test**

****

**Observation:**

* The AUC score for Logistic Regression for train is 0.890 and for test is 0.884
* The AUC score for LDA for train is 0.889 and for test is 0.884
* The AUC score for KNN model for train is 0.781 and test is 0.958
* The AUC score for NB model for train is 0.887 and 0.885
* The AUC score for DT model for train is 0.890 and 0.855
* The AUC score for Bagging Model for train is 0.997 and 0.897
* The AUC score for Gradient Boosting for train is 0.950 and 0.904
* The AUC score for Ada Boosting for train is 0.913 and 0.879.
* As referring to the accuracy score Gradient Boosting has better performance and the accuracy is more compared to all the other models

## 1.8 Based on these predictions, what are the insights?[¶](http://localhost:8888/notebooks/ML%20_%20Project%20_%20Pratheek%20_.ipynb#1.8-Based-on-these-predictions,-what-are-the-insights?)

* Accuracy on all the models appears to be similar to each other on train and test sets.
* AUC is exactly same for both the train and the test data for all the models. While the KNN 78% and 95% on train and test data respectively.
* From the summary of the confusion matrix, we can see that the actual and the predicted data are very close to each other. This is the reflection of the right fit model.
* F1 score for both the models are almost same for both the models on train and test data.
* Model tuning on decision tree and random forest models gives the better results than other models. But however, bagging on Random forest performs well on both the train and test sets with god accuracy and overall other performance measures as well. - Boosting techniques has also shown a good performance results compare to all model.
* Therefore there is no overfitting or under fitting issues in this case study.