Siddhi Padekar

Predictive Modelling

CNBE

Election data

Contents Executive

Summary………………................................................................................................................................3

Introduction............................................................................................................................................3

Data Description ....................................................................................................................................3

Sample of the dataset: ...........................................................................................................................3

Exploratory Data Analysis.......................................................................................................................4

Q1: Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed................................................................................................................................................4

Q2: Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers . Interpret the inferences for each . Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct……………………………………………………………………………………………………………………………………………..4

Q3: Encode the data (having string values) for Modelling. Is Scaling necessary here or not?( 2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get\_dummies(drop\_first=True)) Data split, ratio defined for the split, train-test split should be discussed…………………………………………………………………………………………………………………………………….….12

Q4: Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting) .......................................................................................................................................................……14

Q5: Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting) ………………………………………………17

Q6: Model Tuning (4 pts) , Bagging ( 1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best\_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.....................................................................................................................................21

Q7 : Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)………………………………………………………………………………………………27

Q7: Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific………………………………………………29

Executive Summary

CNBE news channel is the channel that is devoted exclusively to delivering news continuously, without pause throughout the entire day. The data set consists of survey which was conducted on 1525 voters with 9 variables. In this problem statement we must create an exit poll that will help in predicting overall win and seats covered by a particular party.

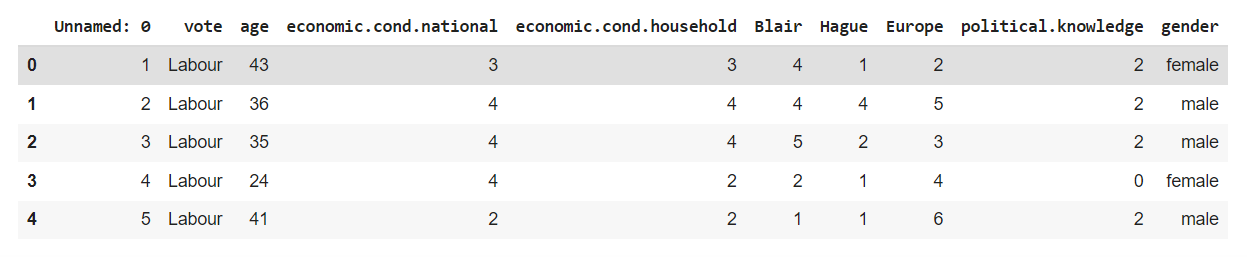
Introduction

The purpose of this whole exercise is to explore the dataset. Build a model, to predict which party a voter will vote for based on the given information to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Data Description

1. Unnamed : Sr Nos 1 to 1525
2. Vote: Labour & Conservative
3. Age: above 20 yrs. age
4. economic. cond. national: 1 to 5
5. economic. cond. household: 1 to 5
6. Blair: 1 to 5
7. Hague: 1 to 5
8. Europe: 1 to 11
9. political. Knowledge: 0 to 3 (score)
10. Gender: Male & Female

Sample of the dataset:



Dataset has 10 variables with 8 integer types and 2 object Type.

Vote category counts are as follows:

|  |  |
| --- | --- |
| Vote category | Counts |
| Conservative | 462 |
| Labour | 1063 |

Counts of males & Female are as follows:

|  |  |
| --- | --- |
| Gender | Counts |
| Male | 713 |
| Female | 812 |

p.s : One column is only Sr No hence we have dropped this column from the data set.

Shape of dataset after dropping Column is 1525 Rows and 9 Columns

7 Columns integer type 2 columns object type

Check for missing values in the dataset:

vote 0

age 0

economic.cond.national 0

economic.cond.household 0

Blair 0

Hague 0

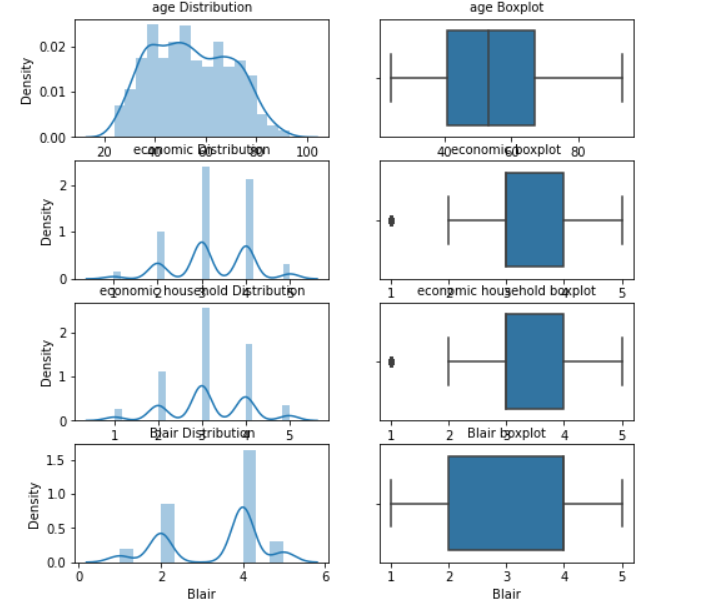
Europe 0

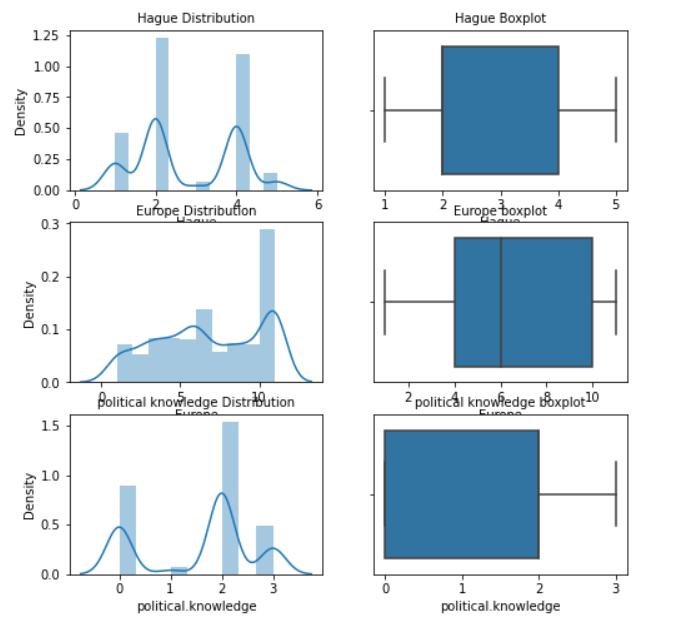
political.knowledge 0

gender 0

dtype: int64

Univariate Analysis :





Univariate analysis means analysis of one variable or one feature. Univariate basically tells us how data in each feature is distributed and also tells us about central tendencies like mean, median, and mode.From the above figure we can see

For age distribution : The distribution is slightly positively-skewed with no outliers.

For economic.cond.national distribution & economic.cond.household : The distribution is slightly negative-skewed with some outliers.

For Blair distribution : The distribution is slightly negative-skewed with no outliers.

For Hague distribution : The distribution is slightly positively-skewed with no outliers.

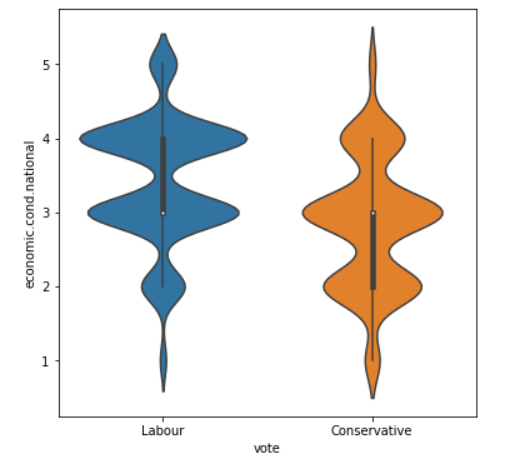
For Europe distribution : The distribution is slightly negative-skewed with no outliers.

For Hague distribution : The distribution is slightly positively-skewed with no outliers.

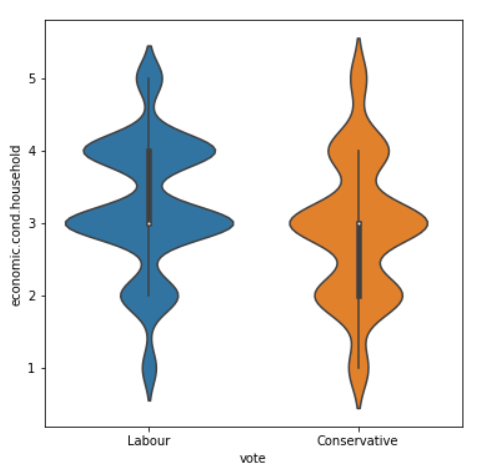
For Political. Knowledge distribution : The distribution is negative-skewed with no outliers.

Bivariate Analysis :

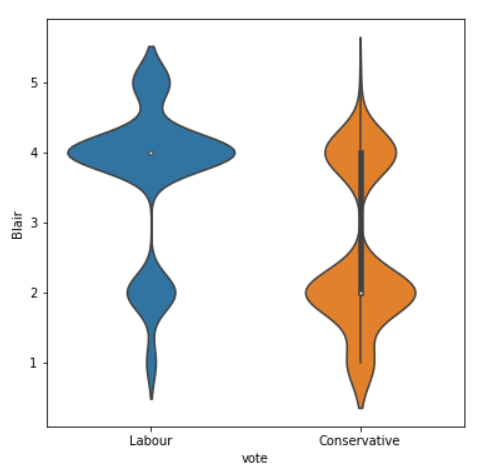
A violin plot depicts distributions of numeric data for one or more groups using density curves. The width of each curve corresponds with the approximate frequency of data points in each region.



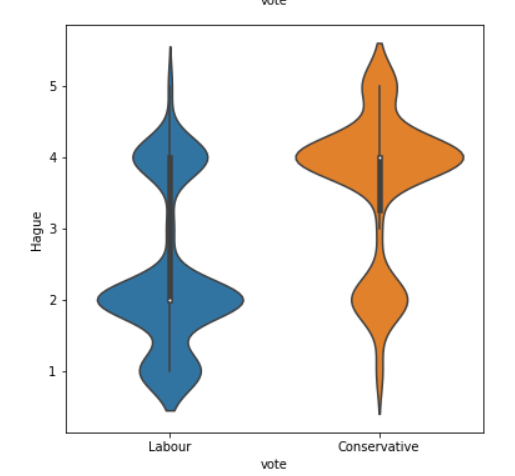
From above we can see that interquartile range for Labour vote range from 3 to 4 for economic.cond.national. The mean for both the vote class the distribution is same at 3. However the density shape is inverse in both class. Labour is negatively skewed and Conservative is positively skewed.



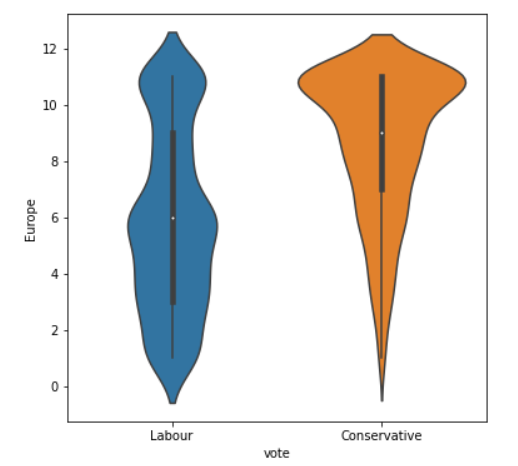
From above we can see that interquartile range for Labour vote range from 3 to 4 for economic.cond.national. The mean for both the vote class the distribution is same at 3. However the density shape is inverse in both class. Labour is negatively skewed and Conservative is positively skewed.



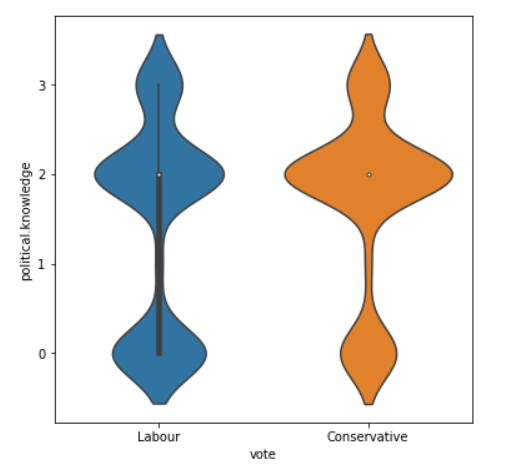
From above we can see that there is no interquartile range for Labour vote for under Blair distribution. The mean for both the vote class distribution is different – Labour has mean of 4 and conservative at 2. However, the density shape is different in both classes. Labour is negatively skewed, and Conservative is slightly positively skewed.



From above we can see that there is interquartile range for Labour vote from 2to 4 and for conservative vote 3.5 to 4 under Hague distribution. The mean for both the vote class distribution is different – Labour has mean of 2 and conservative at 4. However, the density shape is different in both classes. Labour is positively skewed, and Conservative is negatively skewed.



From above we can see that there is interquartile range for Labour vote from 3 to 9 and for conservative vote 7 to 11 under Europe distribution. The mean for both the vote class distribution is different – Labour has mean of 6 and conservative at 7. However, the density shape is different in both classes. Labour is normally skewed, and Conservative is negatively skewed.



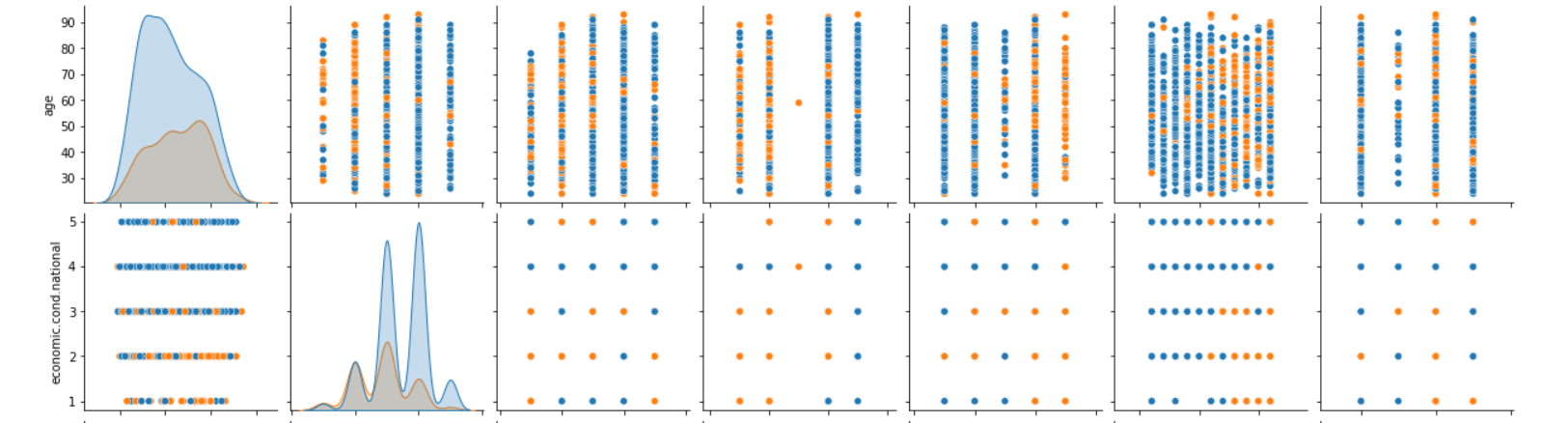
From above we can see that there is interquartile range for Labour vote from 0 to 2 and for conservative vote noting under political. Knowledge distribution. The mean for both the vote class distribution is same at 2. However, the density shape is also slightly same with negatively skewed.

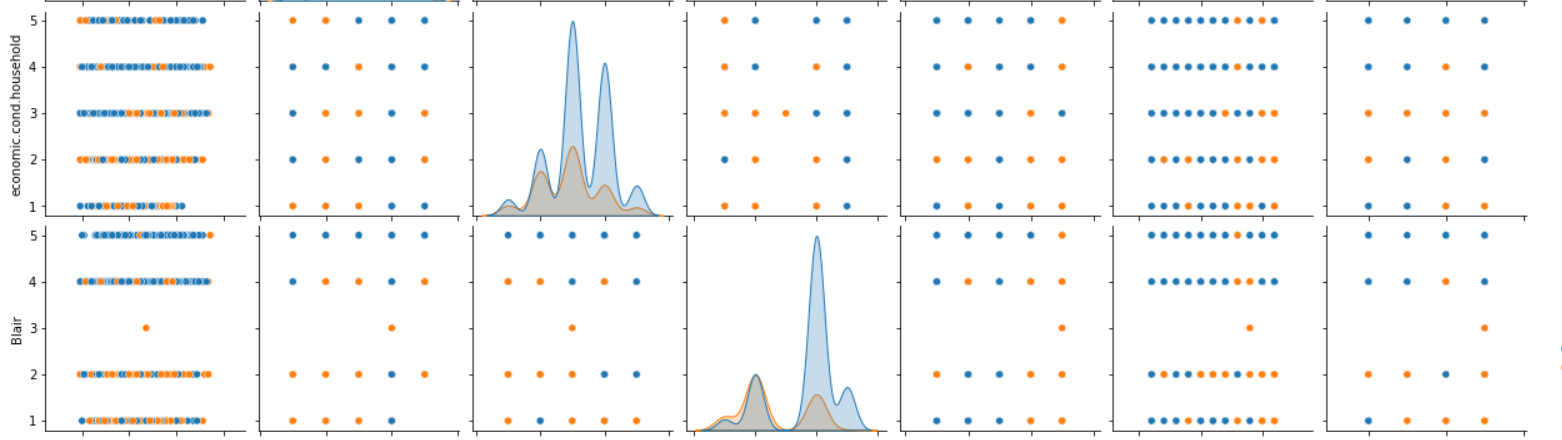
Multivariate Analysis :

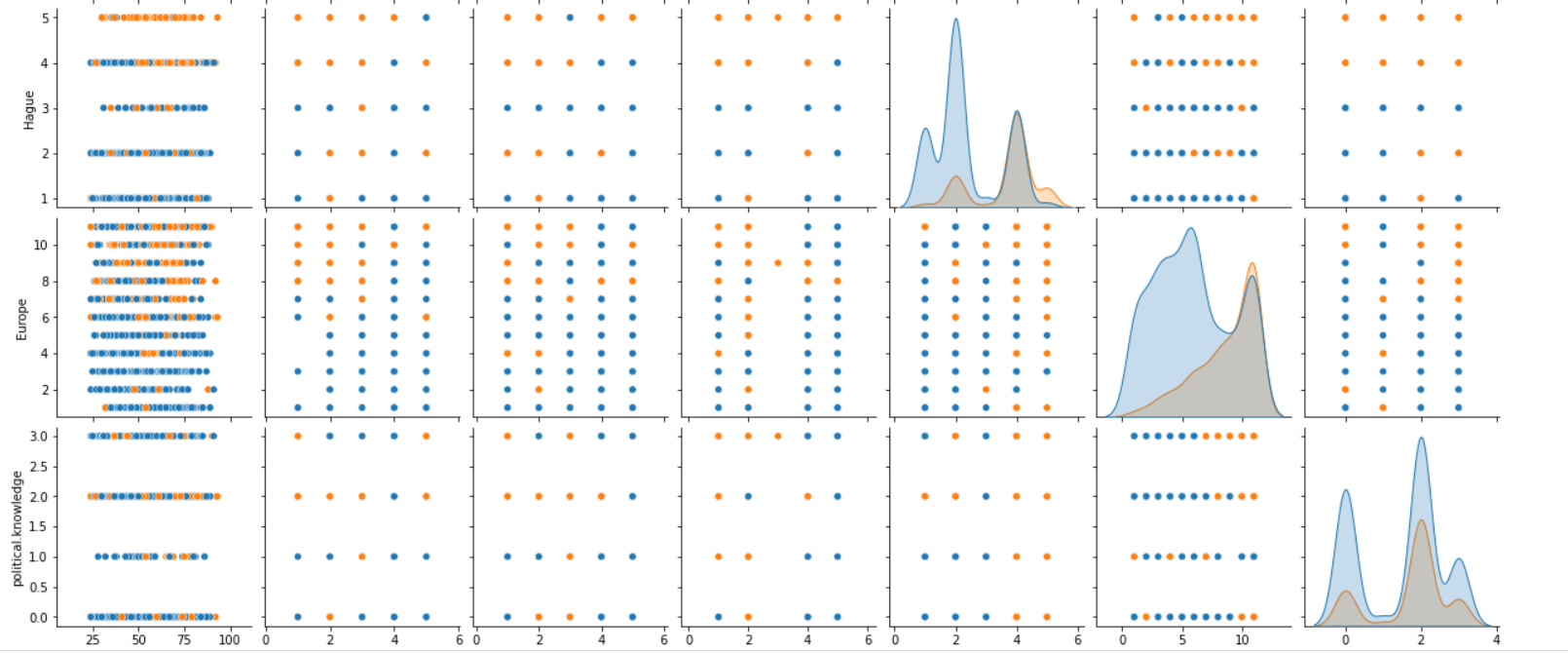
Multivariate analysis (MVA) is a Statistical procedure for analysis of data involving more than one type of measurement or observation.

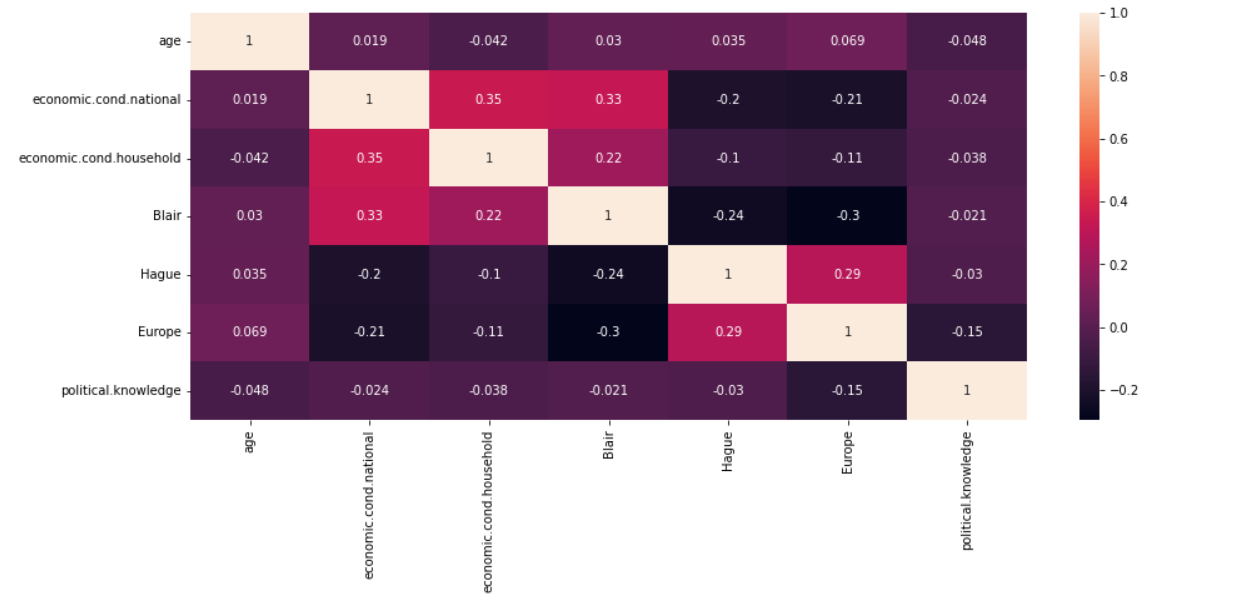
Below is the distribution of each variable with respect to Vote.







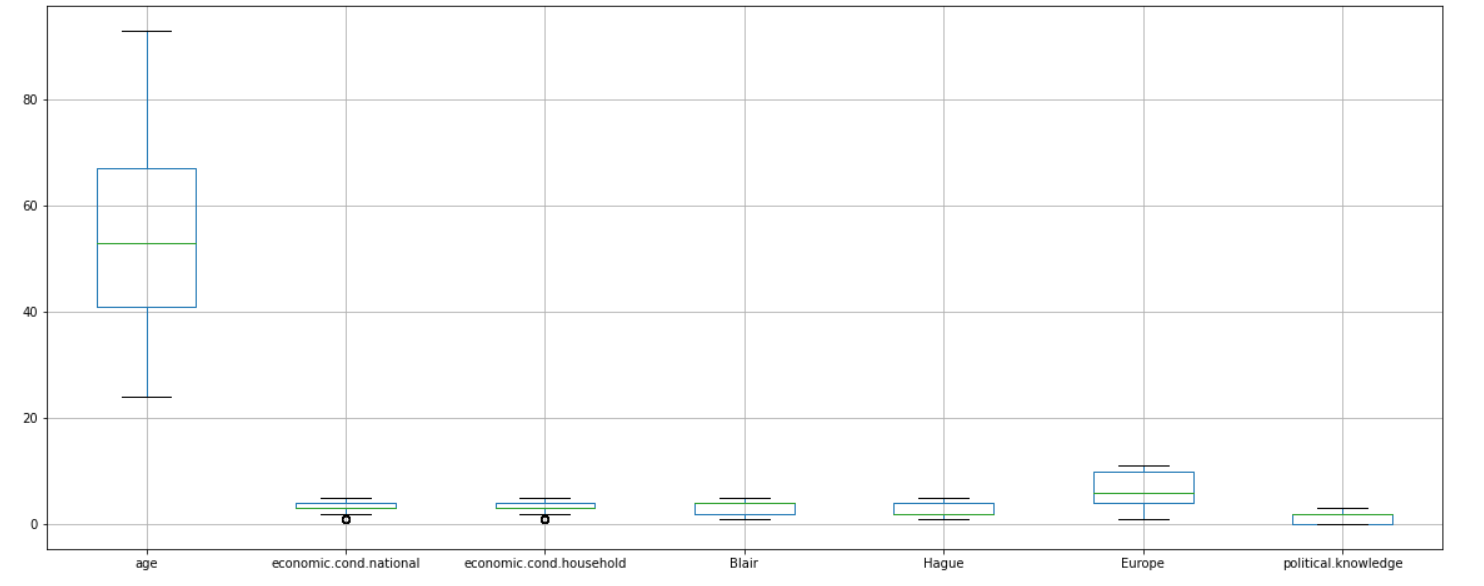




Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are; that is as one increases so does the other and the closer to 1 the stronger this relationship is.

The darker areas in the above diagram shows the variables close to -1 which has negative correlation. Eg. economic.cond.national with relates to Hauge has negative correlation close to -1. Like wise economic.cond.national with relates to Europe also has negative correlation.which means A high score on the economic.cond.national variable would predict a low score on the Hauge variable and A high score on the economic.cond.national variable would predict a low score on the Europe variable.

Outliers :

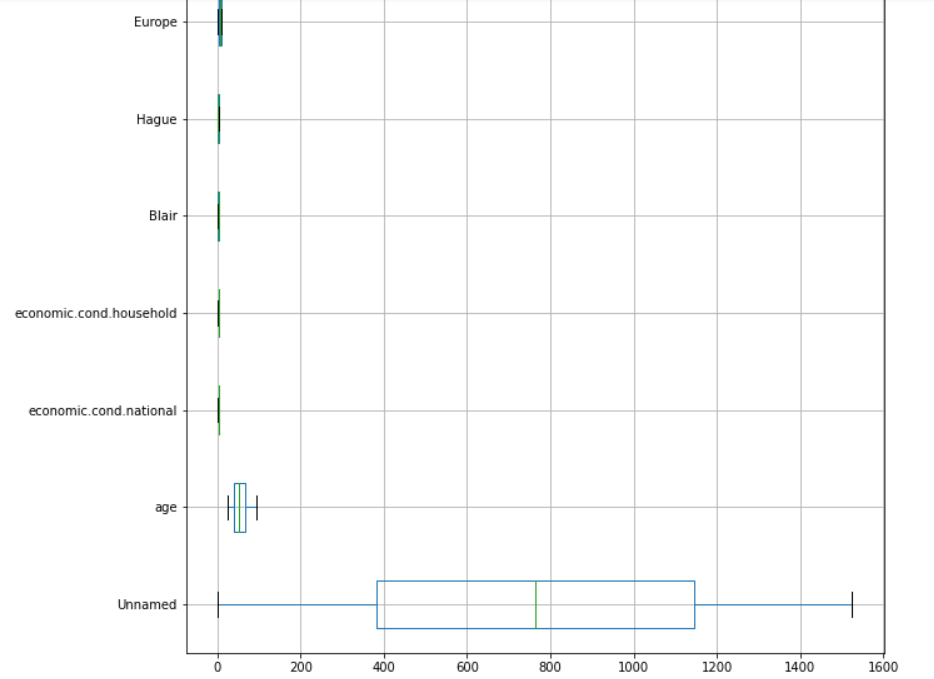


An outlier is an observation that lies an abnormal distance from other values in a random sample from a population

There are some outliers under economic.cond.national & economic.cond.household

Hence need to treat the outliers.

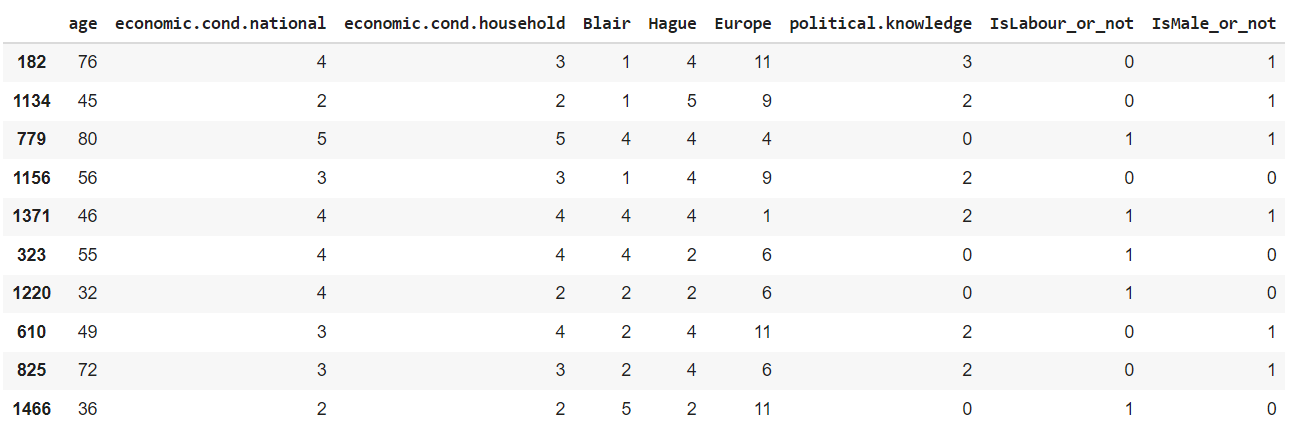
Dataset sample after removing outliers



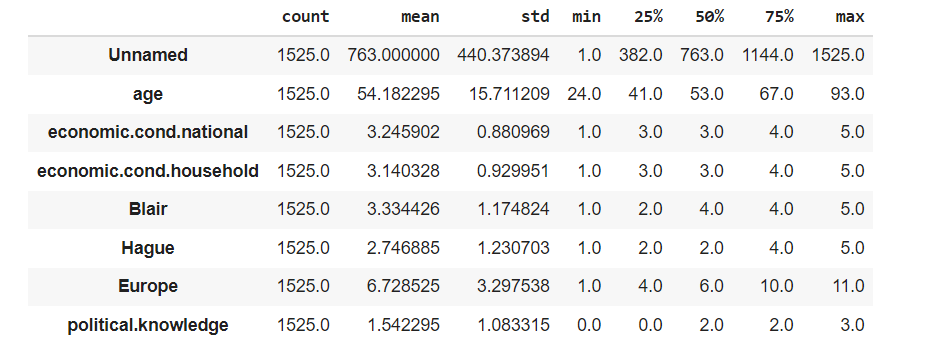
Categorising the object data :

Vote and Gender are object variable and can be categorised .

Hence Vote and Gender is categorised as IsLabour\_or\_not and IsMale\_or\_not respectively.



Is Scaling necessary here or not?

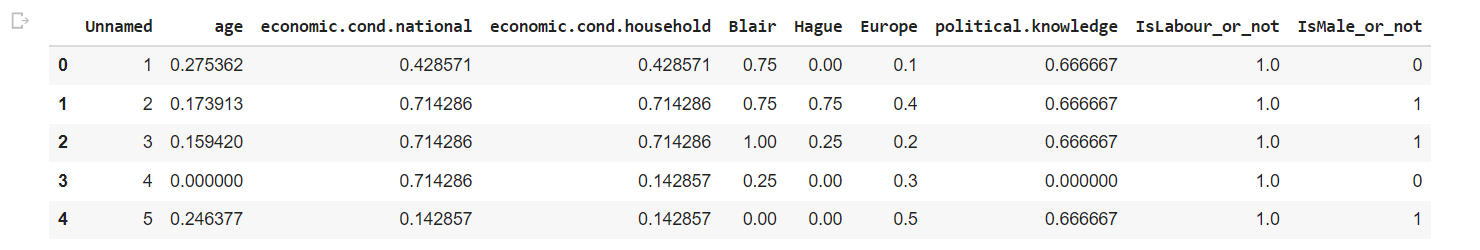


Scaling is required because most of the variables vary in magnitude.

From the above we can see age has big difference in stand deviation and mean as compared to other variables. Hence need to scale the data.

Scaling technique used in this problem is min-max method.

Dataset after scaling the data.



Split the data into train and test (70:30) :

Train-Test Split Evaluation. The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm.

**Dependent variable : Vote (“IsLabour\_or\_not”)**

Data is further split into 70% Training and 30% testing dataset for **vote** variable which is now renamed as **“IsLabour\_or\_not”**

Dependent variable has 2 categories : Labour & Conservative

Counts of people as follows:

Conservative 462

Labour 1063

Logistic Regression and LDA (Linear Discriminant Analysis)

Logistic regression and linear discriminant analyses are **multivariate statistical methods** which can be used for the evaluation of the associations between various covariates and a categorical outcome. Both methodologies have been extensively applied in research, especially in medical and sociological sciences.

While both are appropriate for the development of linear classification models, linear discriminant analysis makes more assumptions about the underlying data. Hence, it is assumed that logistic regression is the **more flexible and more robust method** in case of violations of these assumptions.

Accuracy Score

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

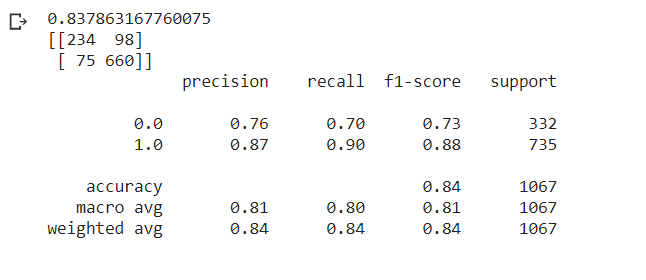
AUC ROC Curve

The Area Under the Curve (AUC) is **the measure of the ability of a classifier to distinguish between classes** and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

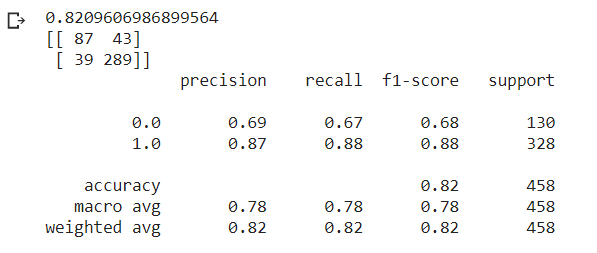
In general, an AUC of 0.5 suggests no discrimination , 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

**LDA (Linear Discriminant Analysis)**

Train Data

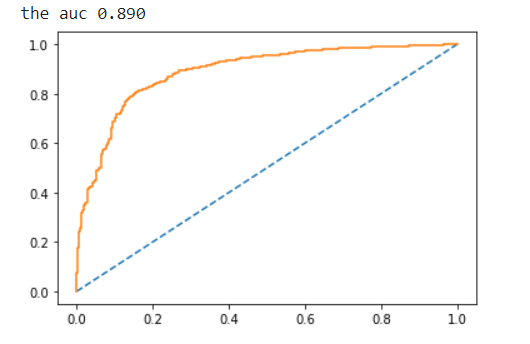


Test Data

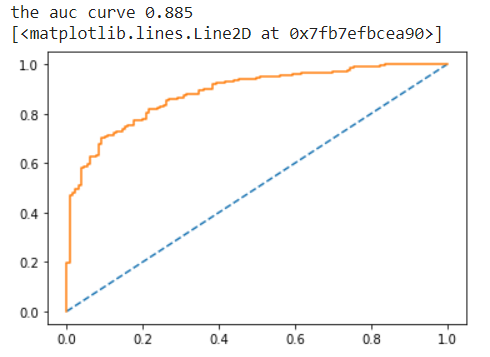


AUC ROC curve for

Train data



Test Data

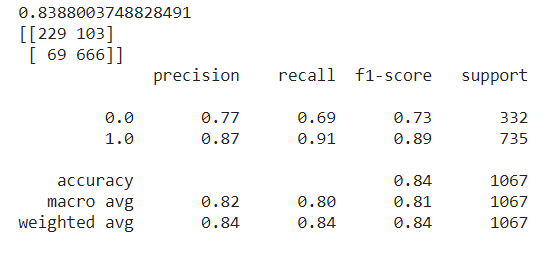


Accuracy score and AUR ROC curve under LDA for train and test data are almost same. Accuracy score for both the dataset are excellent however we cannot conclude anything form this model hence this model cannot be used.

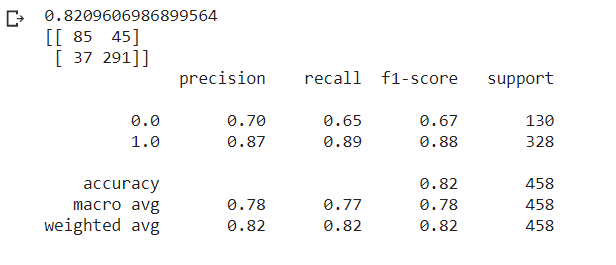
|  |  |
| --- | --- |
| **Train Data** | **0.84 (Accuracy score)** |
| **Test Data** | **0.82 (Accuracy score)** |

**Logistic Regression**

Train Data

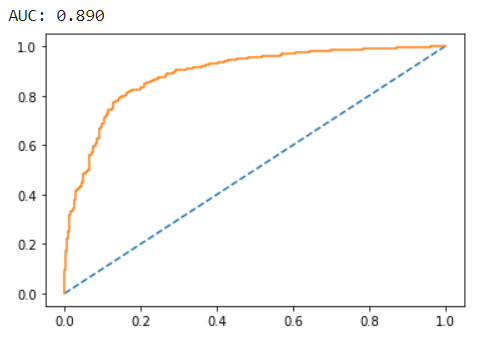


Test Data

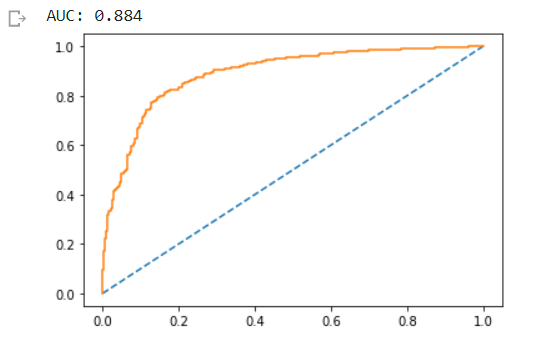


AUC ROC curve for

Train Data



Test Data



Accuracy score and AUR ROC curve under Logistic Regression for train and test data are almost same. Accuracy score for both the dataset are excellent however we cannot conclude anything form this model hence this model cannot be used.

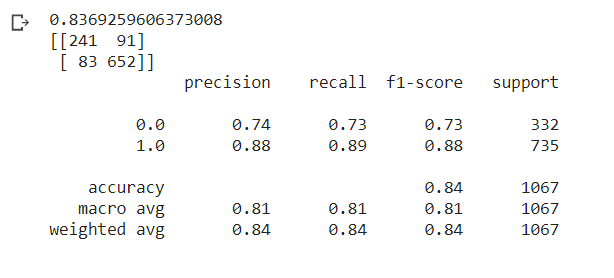
|  |  |
| --- | --- |
| **Train Data** | **0.84 (Accuracy score)** |
| **Test Data** | **0.82 (Accuracy score)** |

**Naïve Bayes Model**

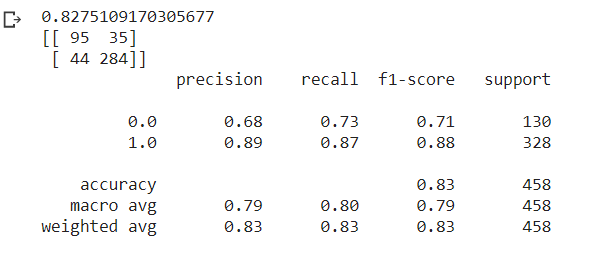
It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

The accuracy matches the expected value calculated by the probability framework of 75% and the composition of the training dataset. This majority class naive classifier is the method that should be used to calculate a baseline performance on your classification predictive modeling problems.

Train Data

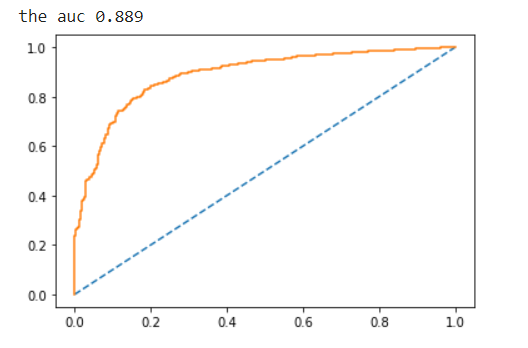


Test Data

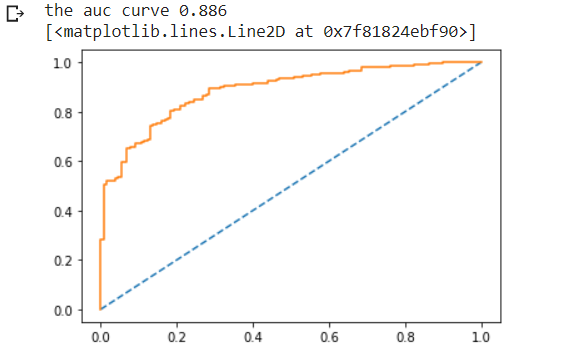


AUC ROC curve

Train Data



Test Data



Accuracy score and AUR ROC curve under Naïve Bayes for train and test data are almost same. Accuracy score for both the dataset are excellent however we cannot conclude anything form this model hence this model cannot be used.

|  |  |
| --- | --- |
| **Train Data** | **0.84 (Accuracy score)** |
| **Test Data** | **0.83 (Accuracy score)** |

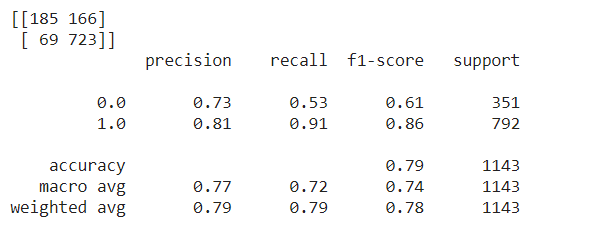
**KNN Model**

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

KNN Score

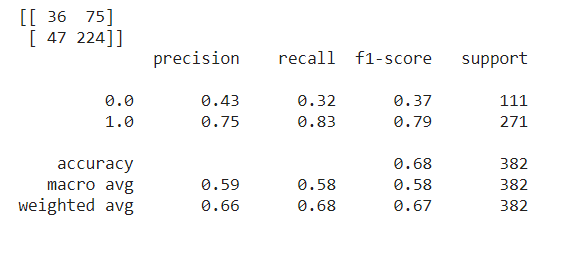
Train Data





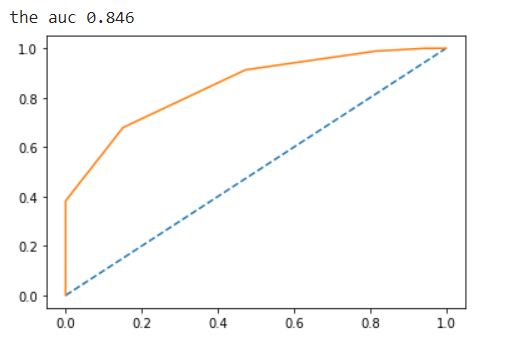
Test Data



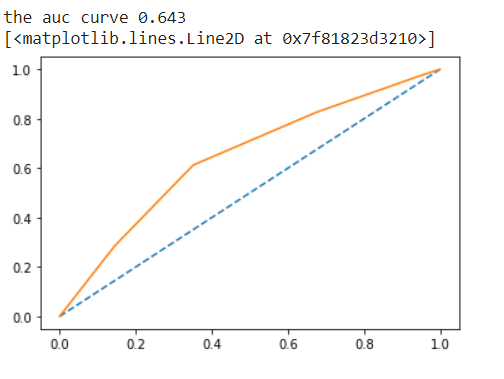


AUC\_ROC cureve

Train Data



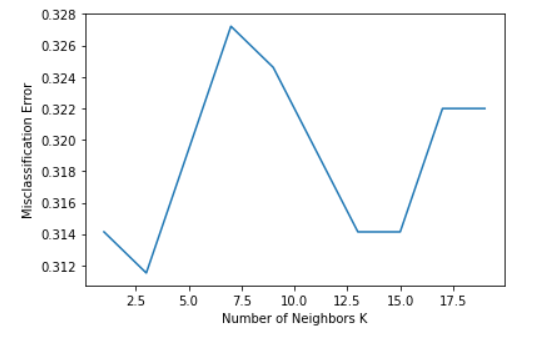
Test Data



Accuracy score and AUR ROC curve under KNN for train and test data are different. Accuracy score for train dataset is more than test dataset.hence we can say this model is perfect for our predictions.

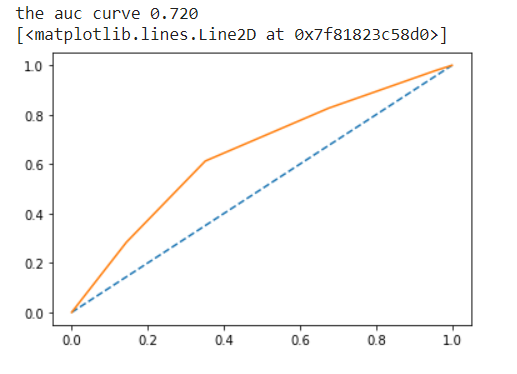
Number of K :

**K=7**

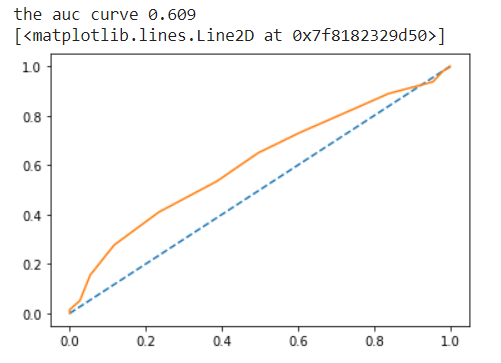


**AUC ROC Cureve after n classifier**

**Train Data**



**Test Data**



**Model Tuning**

Tuning is the process of maximizing a model's performance without overfitting or creating too high of a variance. In machine learning, this is accomplished by selecting appropriate “**hyperparameters**.” Hyperparameters can be thought of as the “dials” or “knobs” of a machine learning model.

Tuning is usually a trial-and-error process by which you change some hyperparameters (for example, the number of trees in a tree-based algorithm or the value of alpha in a linear algorithm), run the algorithm on the data again, then compare its performance on your validation set in order to determine which set of hyperparameters results in the most accurate model.

All [machine learning](https://www.datarobot.com/wiki/machine-learning/) [algorithms](https://www.datarobot.com/wiki/algorithm/) have a “default” set of hyperparameters, which [Machine Learning Mastery](https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/) defines as “a configuration that is external to the [model](https://www.datarobot.com/wiki/model/) and whose value cannot be estimated from data.” Different algorithms consist of different hyperparameters. For example, regularized [regression](https://www.datarobot.com/wiki/regression/) models have coefficients penalties, decision trees have a set number of branches, and [neural networks](https://www.datarobot.com/wiki/neural-network/) have a set number of layers. When building models, analysts and data scientists choose the default configuration of these hyperparameters after running the model on several datasets.

While the generic set of hyperparameters for each algorithm provides a starting point for analysis and will generally result in a well-performing model, it may not have the optimal configurations for your particular dataset and business problem. In order to find the best hyperparameters for your data, you need to tune them.

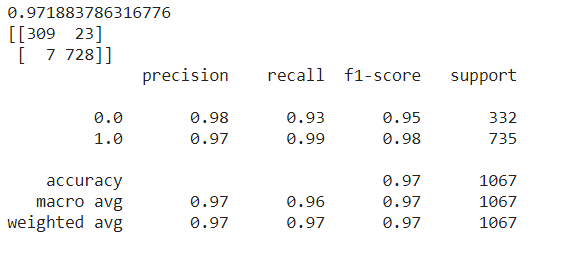
Model tuning allows you to customize your models so they generate the most accurate outcomes and give you highly valuable [insights](https://www.datarobot.com/wiki/insights/) into your data, enabling you to make the most effective business decisions.

Bagging and Boosting

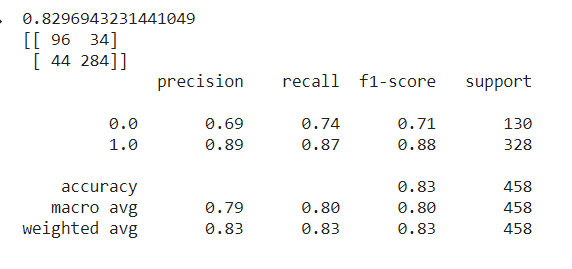
Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

Bagging – Random Forest classifier

Train Data

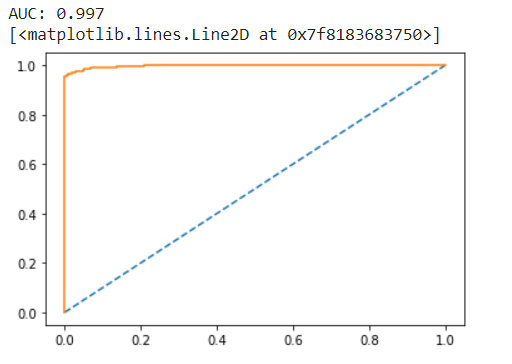


Test Data

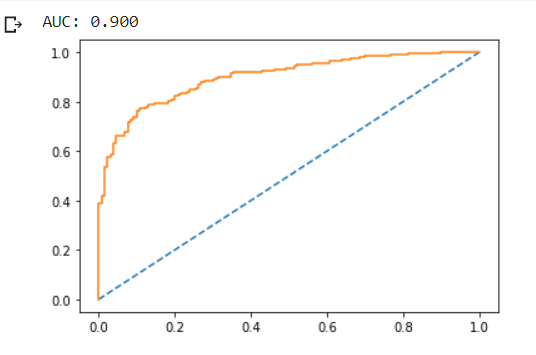


AUC ROC curve

Train Data



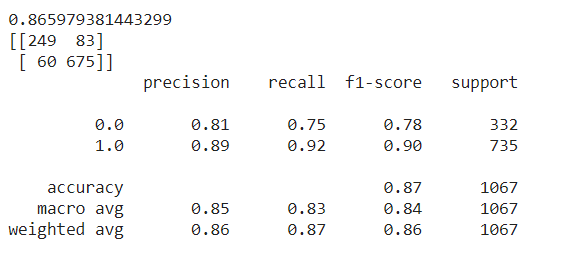
Test Data



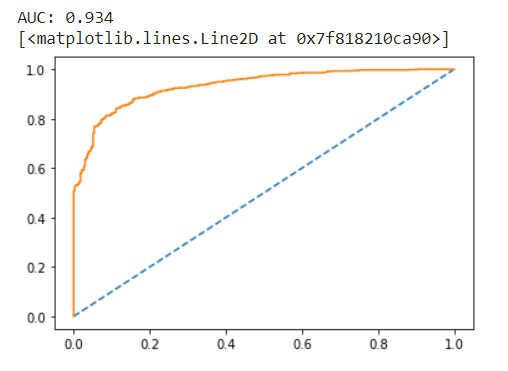
Boosting

ADA Boosting

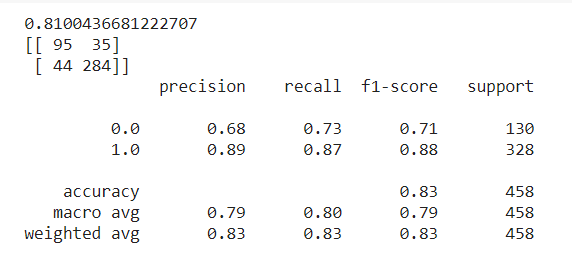
Train Data

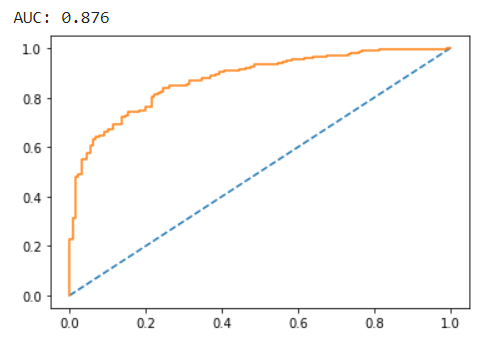


AUC\_ROC CURVE



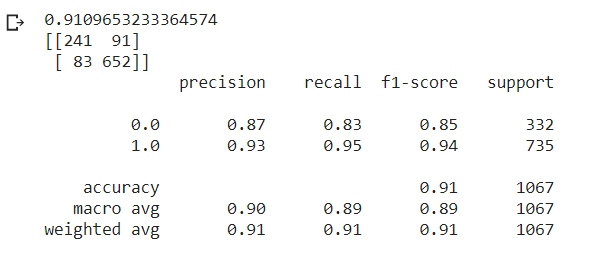
Test data



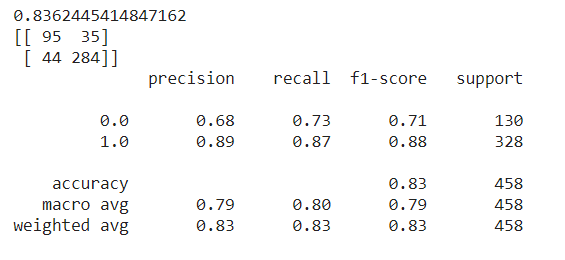


Gradient Boosting

Train



Test



Performance metrics

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Linear Discriminant Analysis | | |
|  | Accuracy | Confusion matrics (F1 Score) | ACU\_ROC score |
| Train | 0.84 | 0.73 | 0.89 |
| Test | 0.82 | 0.68 | 0.885 |

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Logistic Regression | | |
|  | Accuracy | Confusion matrics (F1 Score) | ACU\_ROC score |
| Train | 0.84 | 0.73 | 0.89 |
| Test | 0.82 | 0.67 | 0.88 |

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Naïve Bayes Model | | |
|  | Accuracy | Confusion matrics (F1 Score) | ACU\_ROC score |
| Train | 0.84 | 0.73 | 0.89 |
| Test | 0.82 | 0.71 | 0.89 |

|  |  |  |  |
| --- | --- | --- | --- |
| Data | KNN | | |
|  | Accuracy | Confusion matrics (F1 Score) | ACU\_ROC score |
| Train | 0.79 | 0.61 | 0.85 |
| Test | 0.68 | 0.37 | 0.64 |

From the above models we can see KNN model is best for the problem in hand on the basis of different measures because all the models except KNN gives almost similar scores for train and test data however in KNN it shows how train data is more perfect than test data.

The random forest algorithm is actually a bagging algorithm: also here, we draw random bootstrap samples from your training set. However, in addition to the bootstrap samples, we also draw random subsets of features for training the individual trees; in bagging, we provide each tree with the full set of features.

Bagging under random Forest

The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Bagging | | |
|  | Accuracy | Confusion matrics (F1 Score) | ACU\_ROC score |
| Train | 0.971 | 0.95 | 0.997 |
| Test | 0.83 | 0.71 | 0.9 |

Bagging is Best fitted model.

Boosting

The term 'Boosting' refers to a family of algorithms which converts weak learner to strong learners. Boosting is an ensemble method for improving the model predictions of any given learning algorithm. The idea of boosting is to train weak learners sequentially, each trying to correct its predecessor.

Gradient Boosting

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Boosting | | |
|  | Accuracy | Confusion matrics (F1 Score) | ACU\_ROC score |
| Train | 0.971 | 0.85 | 0.96 |
| Test | 0.83 | 0.71 | 0.9 |