Problem 1:

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

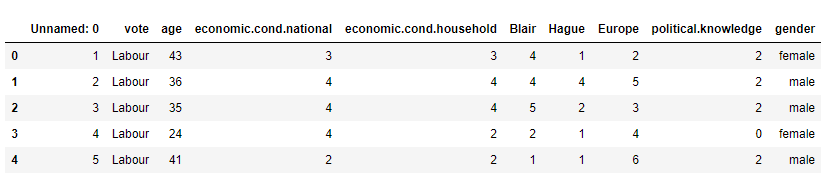
Dataset for Problem: [Election\_Data.xlsx](https://olympus.greatlearning.in/courses/13602/files/2234615/download?verifier=W5GKzi7L2HRhPARciZJ2UsMF6cIREKeFMA7xXfyV&wrap=1)

1. **Read the dataset. Do the descriptive statistics and do null value condition check. Write an inference on it. (5 Marks)**

As we start reading data, we come to know there are 9 different attributes

* **vote**: Party choice: Conservative or Labour
* **age**: in years
* **economic.cond.national**: Assessment of current national economic conditions, 1 to 5.
* **economic.cond.household**: Assessment of current household economic conditions, 1 to 5.
* **Blair**: Assessment of the Labour leader, 1 to 5.
* **Hague**: Assessment of the Conservative leader, 1 to 5.
* **Europe**: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment.
* **political.knowledge**: Knowledge of parties' positions on European integration, 0 to 3.
* **gender**: female or male.

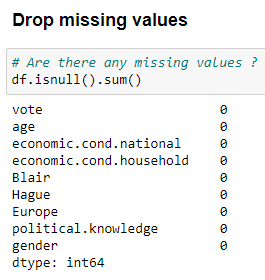
Lets see data, we need to remove first ‘Unnamed: 0’ column



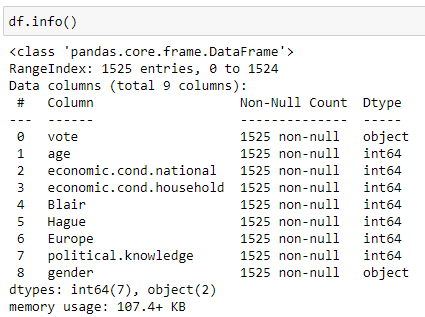
Dataset contains total 1525 records and 9 attributes.



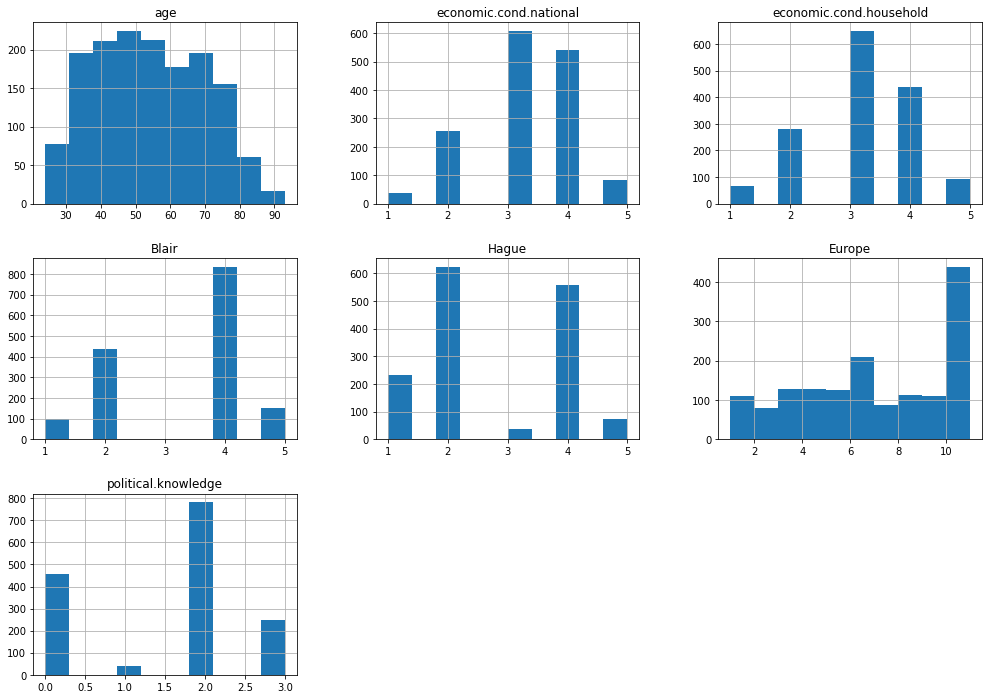
Lets drop missing values if any



Lets see datatypes and memory usage



Lets see histogram, as per above analysis most distribution seem having more than one pick and data spread is not normal

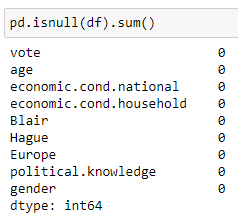


Inference

* Dataset contains 1525 records with 9 attributes
* Dataset holds two categorical and seven numerical attributes
* For attribute vote there are two different types of values
  + Conservative
  + Labour
* For attribute gender there are two different types of values
  + Male
  + Female
* Dataset contains zero missing values
* As per histogram analysis most distribution seem having more than one pick and data spread is not normal

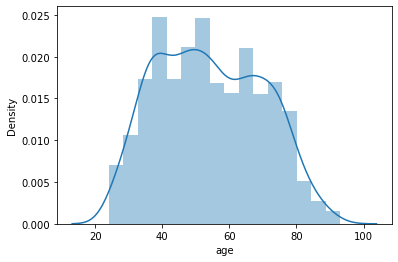
1. **Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)**

Dataset contains no null value records

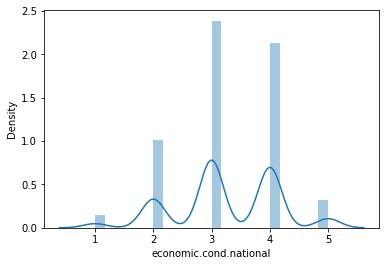


Lets start univariate analysis

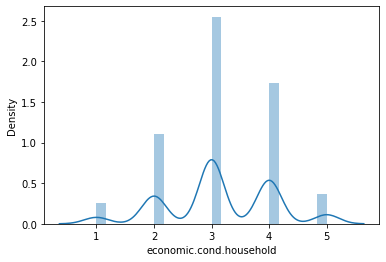
Density VS age: Distribution seem having more than one pick and trying to follow normal distribution



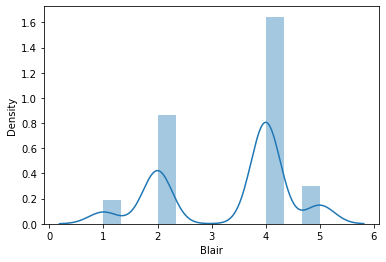
Density VS economic.cond.national: Distribution seem having more than one pick and data spread is not normal



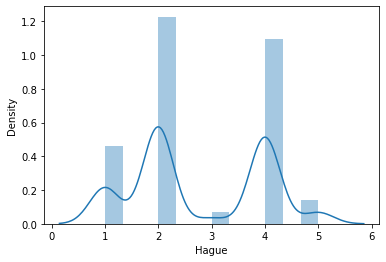
Density VS economic.cond.household: Distribution seem having more than one pick and data spread is not normal



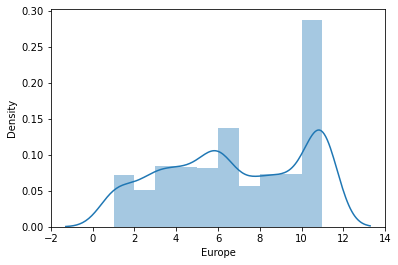
Density VS Blair: Distribution seem having more than one pick and data spread is not normal



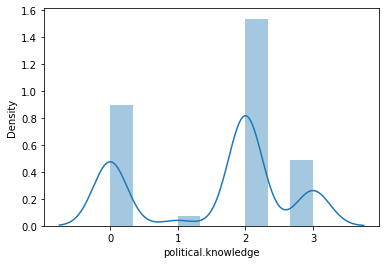
Density VS Hague: Distribution seem having more than one pick and data spread is not normal



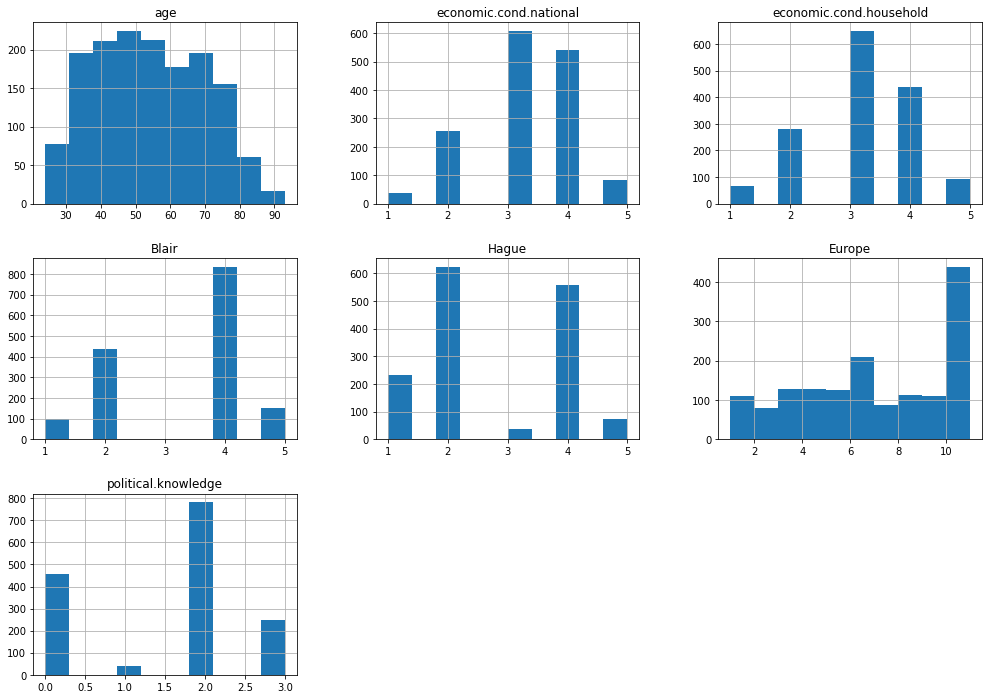
Density VS Europe: Distribution seem having more than one pick and data spread is not normal



Density VS political.knowledge: Distribution seem having more than one pick and data spread is not normal

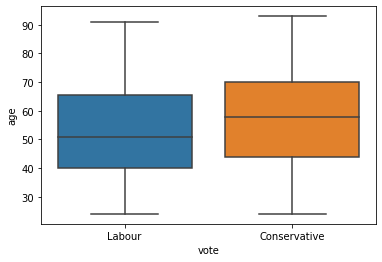


Lets see histogram, as per above analysis most distribution seem having more than one pick and data spread is not normal

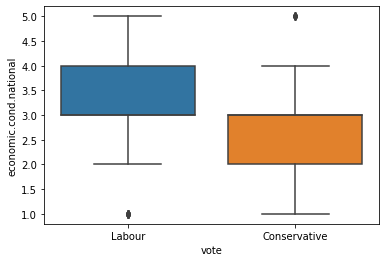


Bivariate Analysis

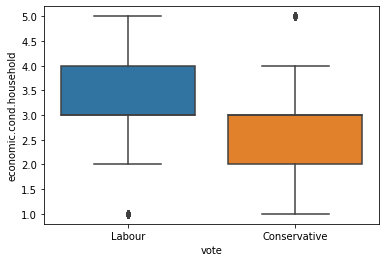
Vote VS age: new generation likes labour party more than conservative



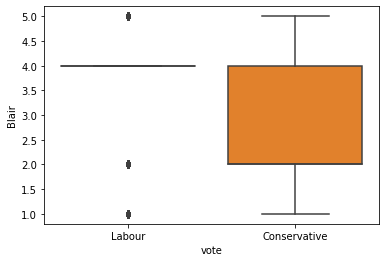
Vote VS Economic condition national: labour party hold good amount of high economic national condition followers compare to conservative party



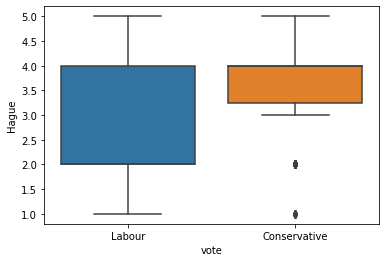
Vote VS Economic condition household: labour party hold good amount of high economic household condition followers compare to conservative party



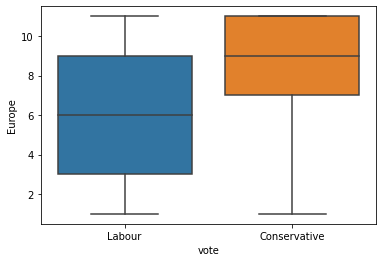
Vote VS Blair: Followers of labour party highlight blair at 4 but conservative party between 2 to 4



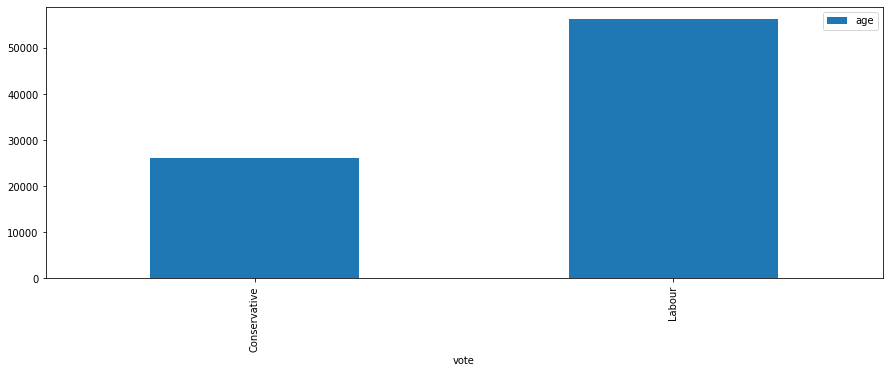
Vote VS Hague: Followers of labour party highlight hague between 2 to 4 but conservative party between 3.5 to 4



Vote VS Europe: Followers of labour party highlight high variation ‘Europe integration’ but conservative party hold high interest compare to labour



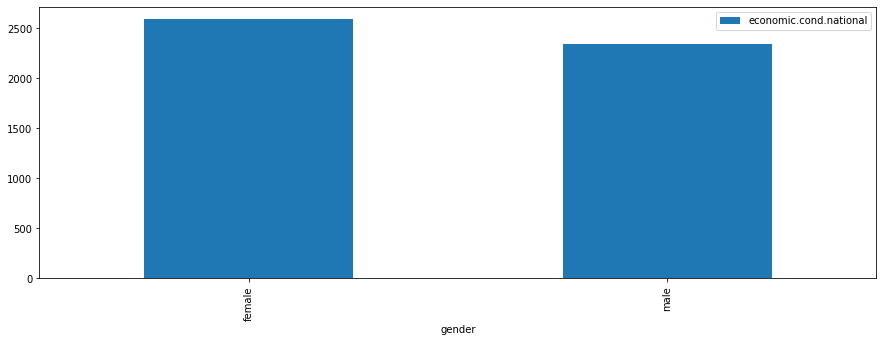
Age VS Vote: Labour party hold high age followers compare to conservative party



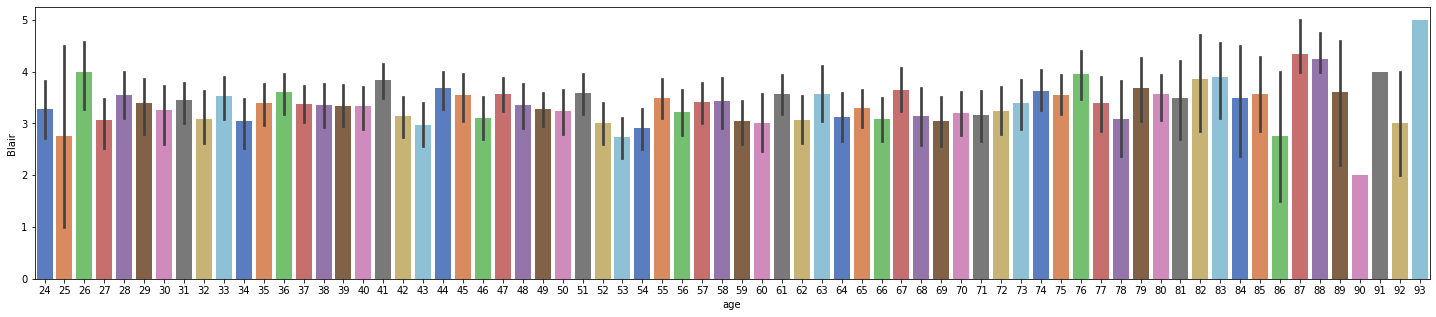
Economical condition national VS vote: Followers of labour party seem holding high economical national condition compare to conservative



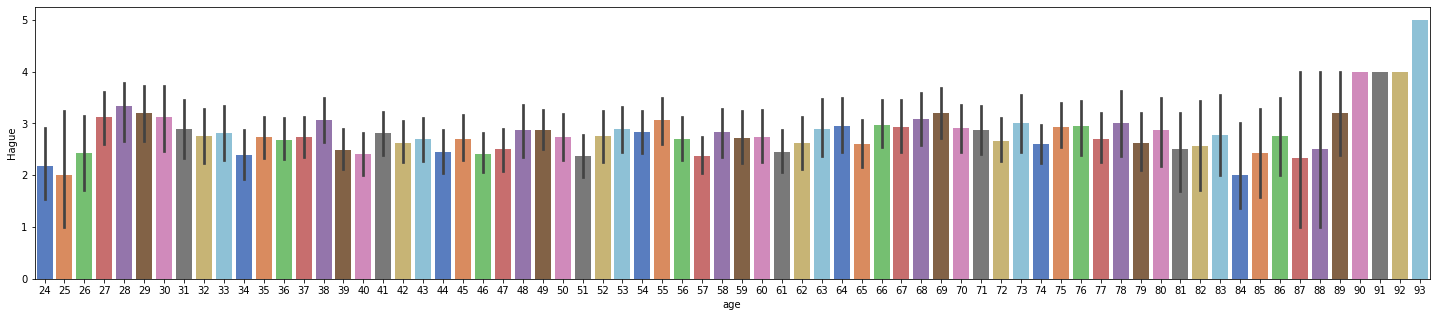
Economical condition household VS gender: Female followers seem holding high economical household condition compare to male followers



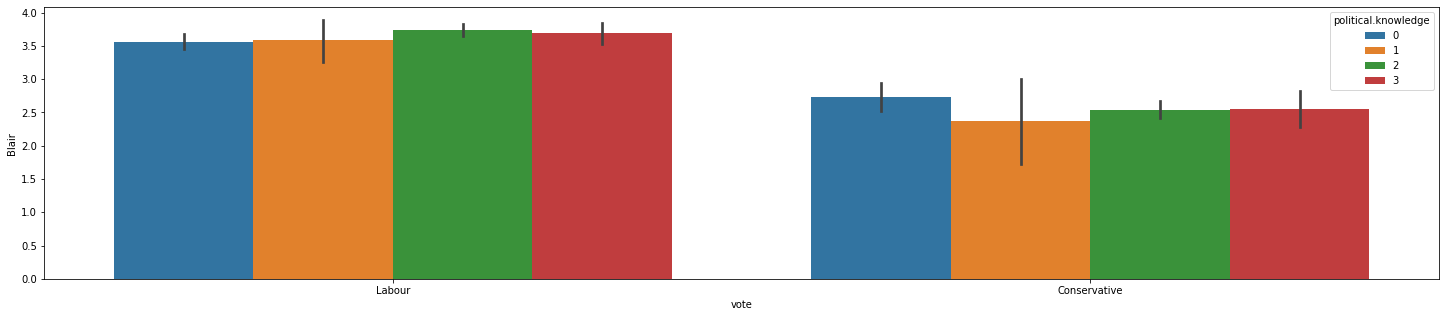
Age VS Blair: Blair seem more popular compare to Hauge in all ages



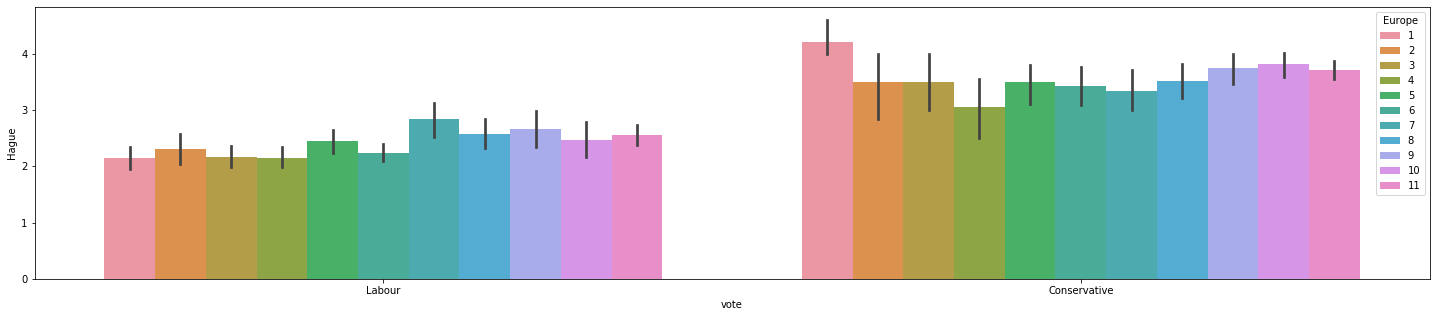
Age VS Blair: Hauge seem less popular compare to Blair in all ages



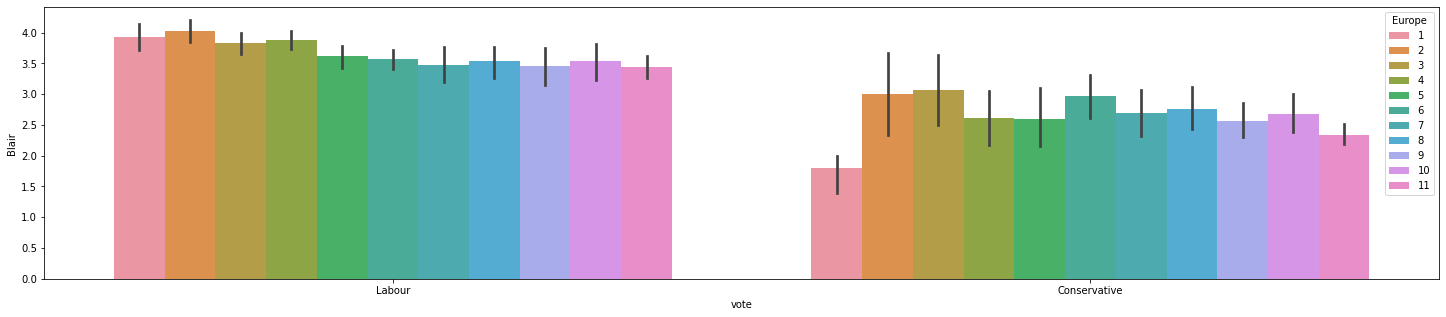
Vote VS Political knowledge: labour party voters seem having high knowledge compare to conservative party



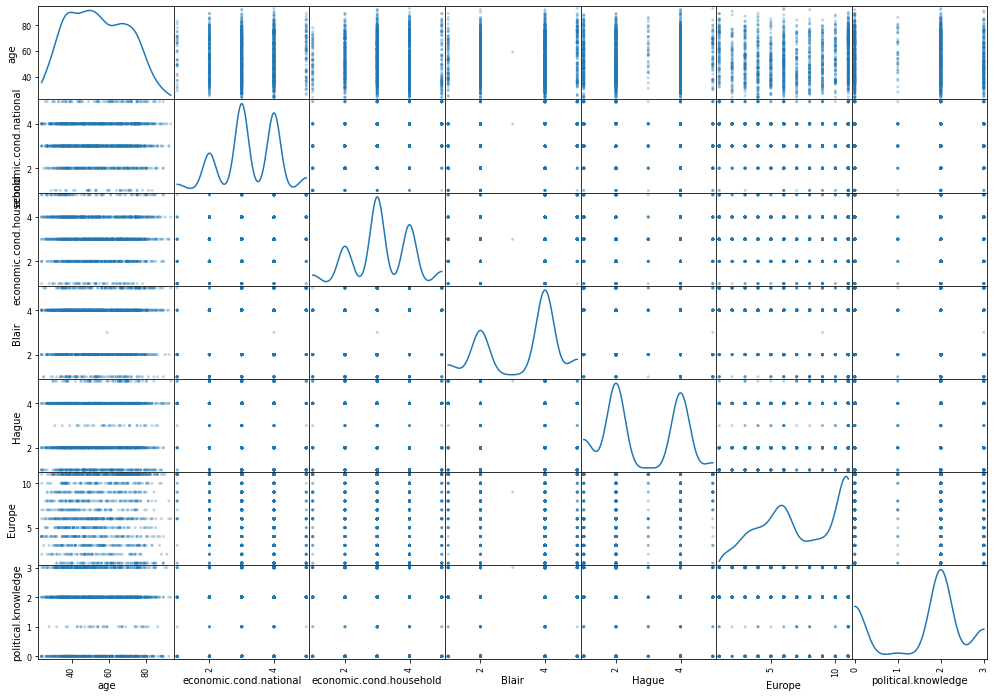
Vote VS Hauge (Hue: Europe integration) It seem conservative party followers seem to be hauge followers interested more in ‘europe integration’ compare to labour party’s hauge followers



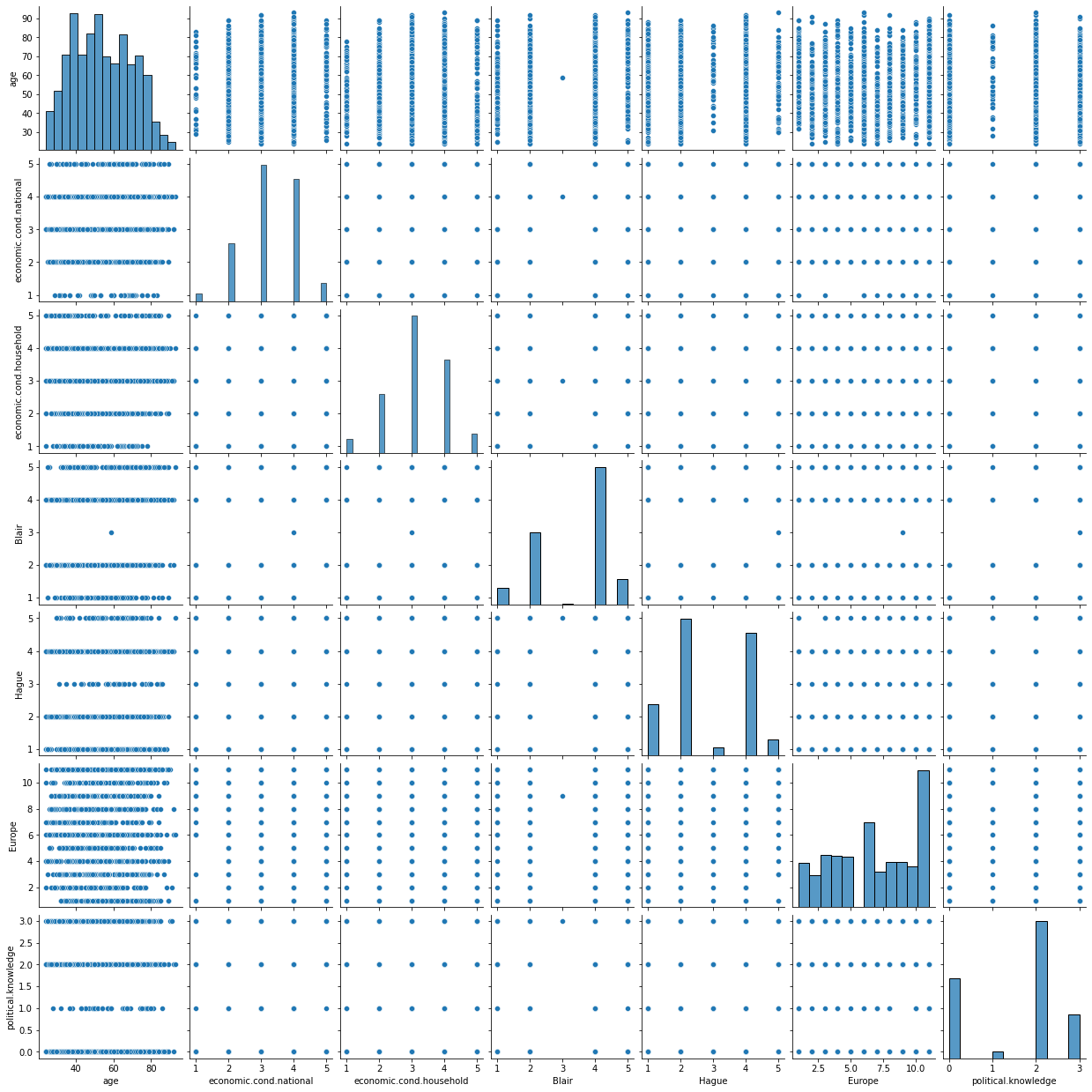
Vote VS Blair (Hue: Europe integration) It seem labour party followers seem to be blair followers interested more in ‘europe integration’ compare to conservative party’s blair followers



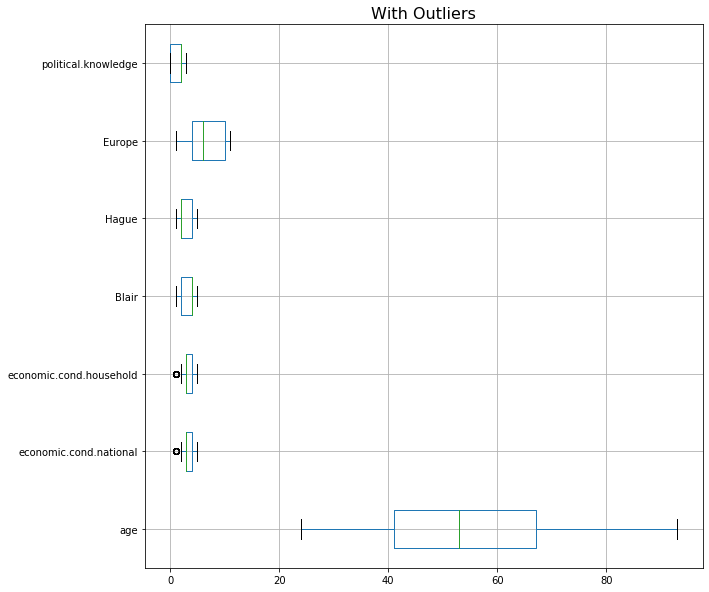
Lets see scatter plot

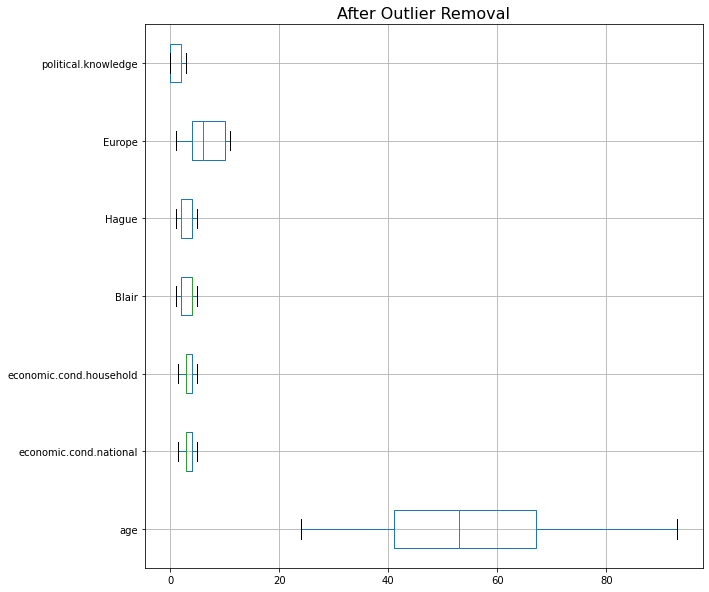


Lets see pair plot

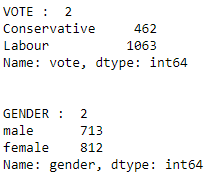


Lets see outliers, before treatment





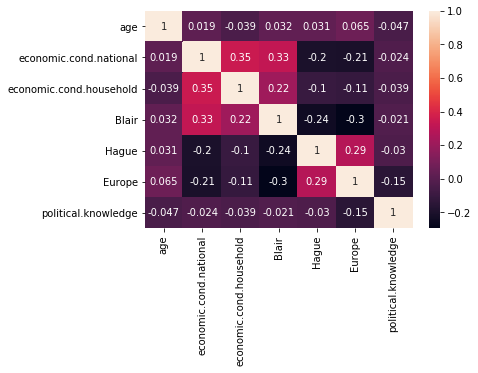
Lets see unique values for categorical variables



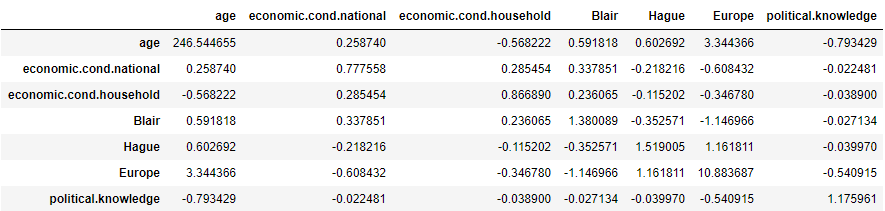
Lets check for duplicate records. We have total 8 duplicate records



Lets check correlation plot

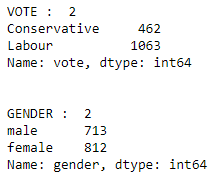


Lets see covariance plot



1. **Encode the data (having string values) for Modelling. Is Scaling necessary here or not?( 3 pts), Data Split: Split the data into train and test (70:30) (2 pts).**

We can see dataset attribute vote & gender contains variations which need to be converted from categorical to numerical

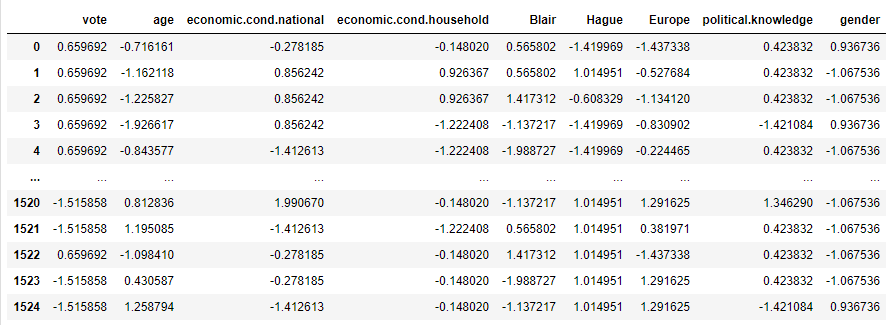


Vote contains two values ‘conservative’ and ‘labour’ only

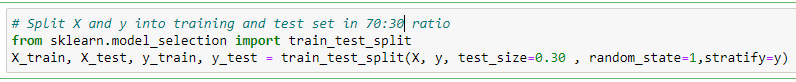
Gender contains two values ‘male’ and ‘female’ only

We use manual encoding technique.

We need to do scaling because there are different level available in dataset which need to standardize. We use standard-scaler



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

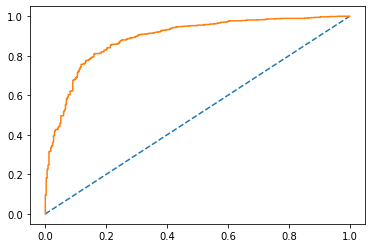


### Apply Logistic Regression and LDA (Linear Discriminant Analysis) (3 pts). Interpret the inferences of both model s (2 pts)

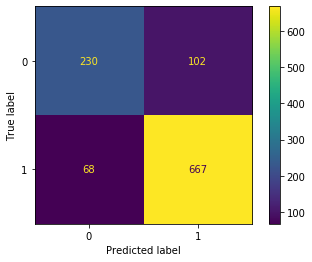
### Logistic regression model

Logistic regression model accuracy (Train data): 0.8406747891283973

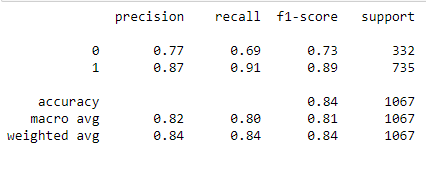
AUC/ROC Curve: AUC: 0.889

.

Confusion matrix (Train data)

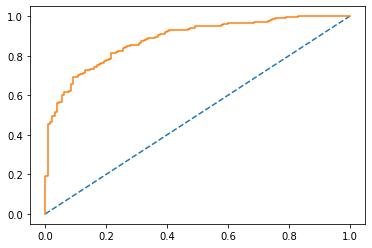


Classification report

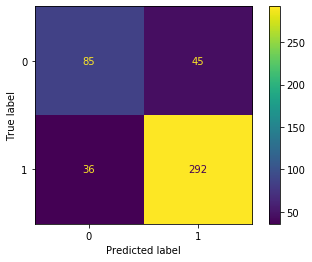


Logistic regression model accuracy (Test data): 0.8231441048034934

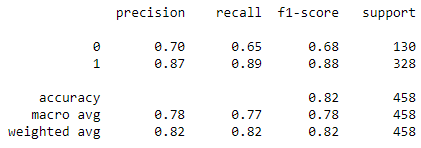
AUC/ROC Curve AUC: 0.889



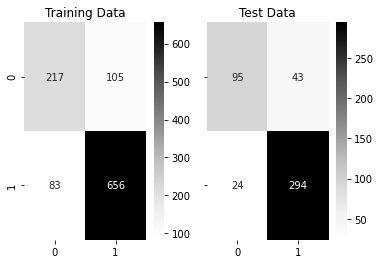
Confusion matrix (Test Data)

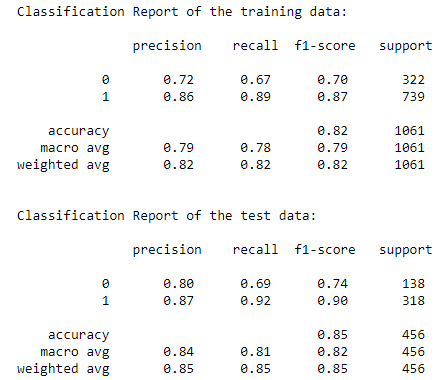


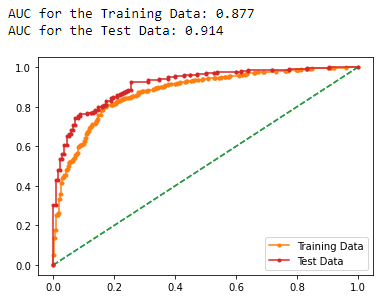
Classification report



* LDA Linear Discriminant Analysis



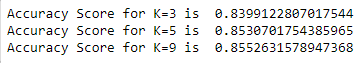




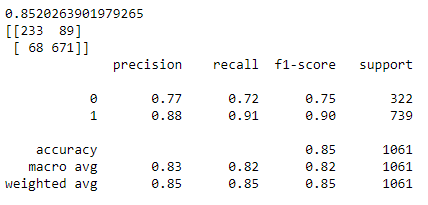
* Inferences
  + Logistic regression seem to be working good on both train data 84% and test data 82% accuracy
  + Logistic regression seem to be gaining accuracy but losing recall while moving from train data to test data
  + LDA seem to be working good on both train data 87% and test data 91% accuracy
  + LDA seem to be gaining accuracy but losing recall while moving from train data to test data

1. **Apply KNN Model and Naïve Bayes Model(5 pts). Interpret the inferences of each model (2 pts)**

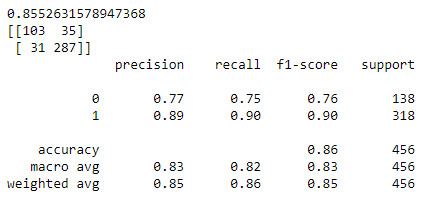
* **KNN**



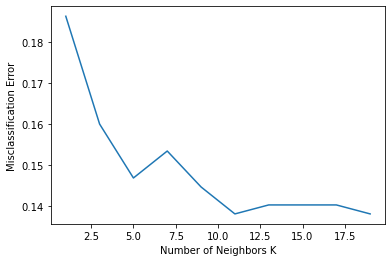
Model score train data



Model score test data

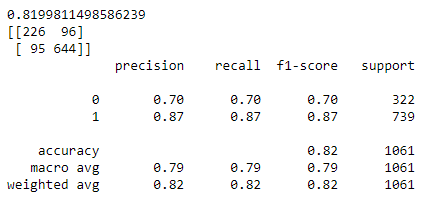


Inference: Knn seem be working good for k=5 with least misclassification error

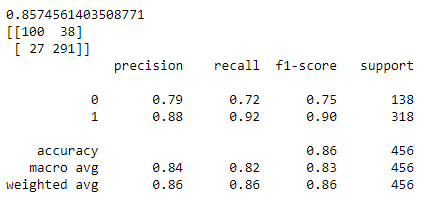


* Naïve bayes

Performance matrix on train data



Performance matrix on test data



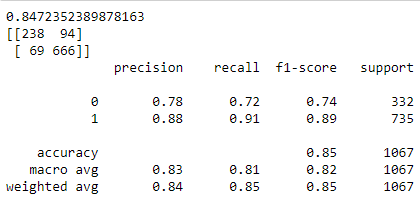
* Inferences
  + knn seem to be working good for k=5 at 85% compare to k=3 at 83%, k=9 at 85%
  + knn seem to be working good on both train data 83%(k=3) and test data 82%(k=3) accuracy
  + knn show high precision and high recall while k=3 at 83%
  + knn seem be working good for k=5 with least misclassification error
  + Naïve Bayes seem to be working good on both train data 81% and test data 82%
  + Naïve Bayes show high precision and high recall while train and test

1. **Model Tuning (2 pts) , Bagging ( 2.5 pts) and Boosting (2.5 pts).**

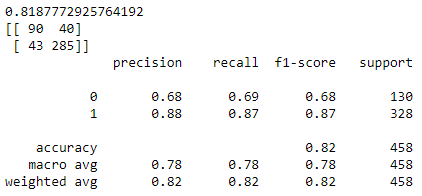
* **Adaptive boosting**



Performance matrix on train data

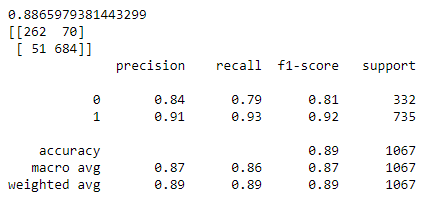


Performance matrix on test data

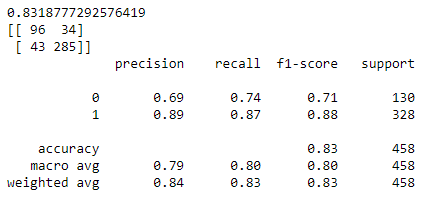


* Gradient boosting

Performance matrix on train data



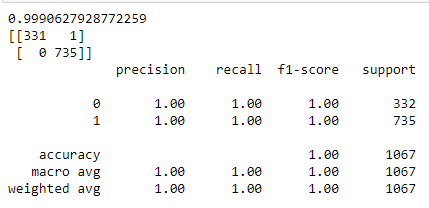
Performance matrix on test data



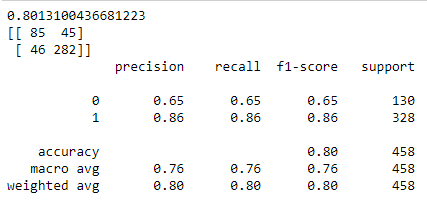
* Bagging



Performance matrix with train data



Performance matrix with test data

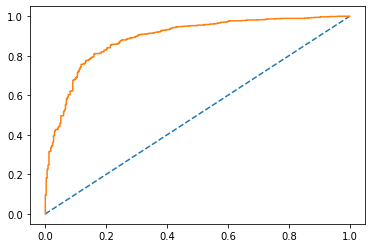


1. **Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model (4 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized (3 pts)**

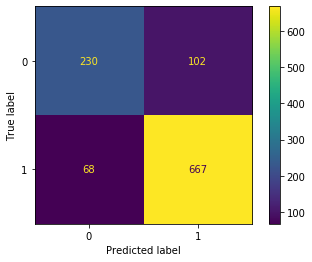
### Logistic regression model

Logistic regression model accuracy (Train data): 0.8406747891283973

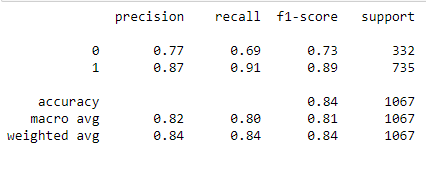
AUC/ROC Curve: AUC: 0.889

.

Confusion matrix (Train data)

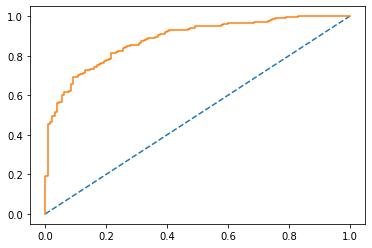


Classification report

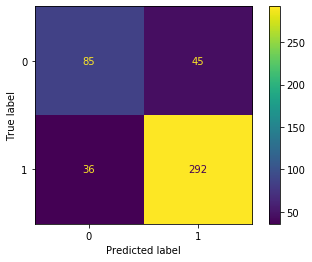


Logistic regression model accuracy (Test data): 0.8231441048034934

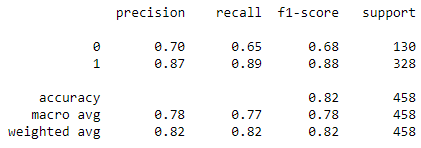
AUC/ROC Curve AUC: 0.889



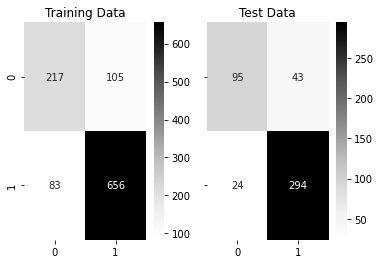
Confusion matrix (Test Data)

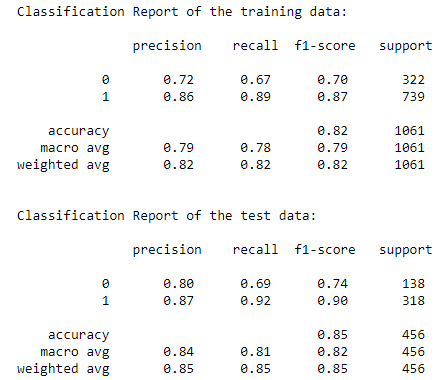


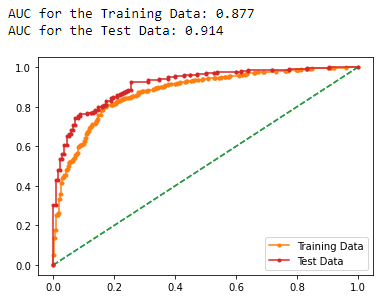
Classification report



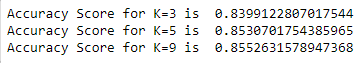
* LDA Linear Discriminant Analysis



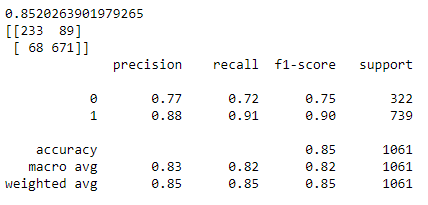




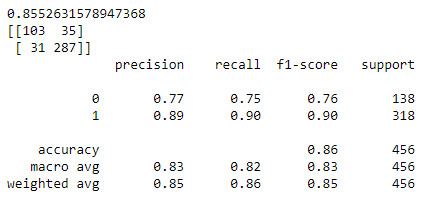
* **KNN**



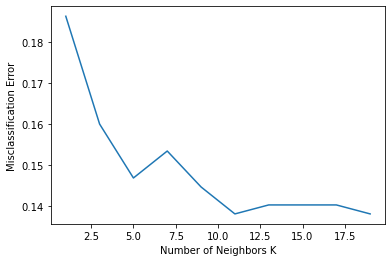
Model score train data



Model score test data

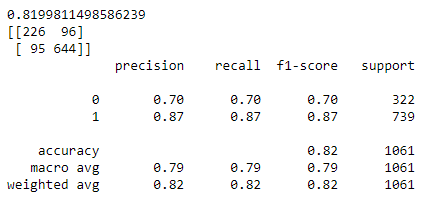


Inference: Knn seem be working good for k=5 with least misclassification error

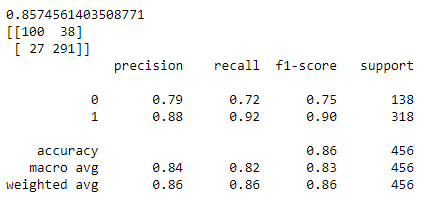


* Naïve bayes

Performance matrix on train data



Performance matrix on test data



Comparison of all four models performance matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Logit** | | | **LDA** | | | **kNN** | | | **Naïve Bayes** | | |
| **Train data score** | 84% | | | 87% | | | 83% | | | 81% | | |
| **Test data score** | 82% | | | 82% | | | 85% | | | 82% | | |
| **Performance**  **Matrix (Train Data)** |  | A | P |  | A | P |  | A | P |  | A | P |
| 0 | .77 | .69 | 0 | .77 | .69 | 0 | .77 | .72 | 0 | .70 | .70 |
| 1 | .87 | .91 | 1 | .87 | .91 | 1 | .88 | .91 | 1 | .87 | .87 |
| **Performance**  **Matrix (Test Data)** |  | A | P |  | A | P |  | A | P |  | A | P |
| 0 | .70 | .65 | 0 | .70 | .65 | 0 | .77 | .75 | 0 | .79 | .72 |
| 1 | .87 | .89 | 1 | .87 | .89 | 1 | .89 | .90 | 1 | .88 | .92 |

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

**Feedback**

**Labour party specific**

* + Having good economical national condition indicate, chance of being labour party follower is high
  + Having good economical household condition indicate, chance of being labour party follower is high
  + Labour party follower seem to be less interested in Europe integration comparative to conservative party and with high deviation among follower
  + Labour party follower seems to be holding high political knowledge compare to conservative party followers

**Conservative party specific**

* + More the age, chance of being conservative party follower is high
  + Conservative party follower seem to be highly interested in Europe integration comparatively labour party and with less deviation among follower
  + Conservative party follower seems to be holding low political knowledge compare to labour party followers

**Recommendations**

* + Dataset seem biased towards labour party's 1063 records compare to 462 records of conservative party, kindly check with business if balanced data is available for analysis.
  + Blair seems more popular compare to hauge in all age group
  + Blair's followers seem to be not in favour of Europe integration
  + Hauge's followers seem to be in favour of Europe integration
  + Economical condition household and economical condition national seem to highly positively correlated
  + Blair's followers seem to be having good economical condition household and economical condition national

Problem 2:

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

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

* Find the number of characters, words and sentences for the mentioned documents. – 3 Marks

(Hint: use .words(), .raw(), .sent() for extracting counts)

* Remove all the stopwords from all the three speeches. – 3 Marks
* Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks
* Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks [ refer to the End-to-End Case Study done in the Mentored Learning Session ]

Code Snippet to extract the three speeches:

"  
import nltk  
nltk.download('inaugural')  
from nltk.corpus import inaugural  
inaugural.fileids()  
inaugural.raw('1941-Roosevelt.txt')  
inaugural.raw('1961-Kennedy.txt')  
inaugural.raw('1973-Nixon.txt')  
"  
**Important Note: Please reflect on all that you have learned while working on this project. This step is critical in cementing all your concepts and closing the loop. Please write down your thoughts**[**here**](https://docs.google.com/forms/d/e/1FAIpQLSfqHHlmJyUkniiBiejtudIuRFk_TVCLe843wfX6Iu3QNRPmng/viewform?usp=sf_link)**.**

1. Find the number of characters, words and sentences for the mentioned documents. – 3 Marks (Hint: use .words(), .raw(), .sent() for extracting counts)

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Roosevelt Speech Results \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

Number Of Total Characters: 7262

Number Of Total Words: 1338

Number Of Total Sentence: 38

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Kennedy Speech Results \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

Number Of Total Characters: 7336

Number Of Total Words: 1365

Number Of Total Sentence: 27

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Nixon Speech Results \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

Number Of Total Characters: 9646

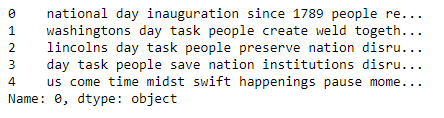
Number Of Total Words: 1802

Number Of Total Sentence: 51

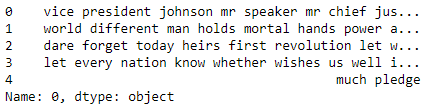
1. Remove all the stopwords from all the three speeches. – 3 Marks

Please check all three speeches after removing stopwords

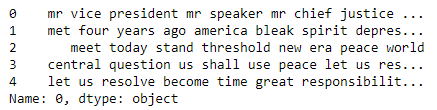
**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Roosevelt Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***



**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Kennedy Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***



**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Nixon Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***



1. Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Roosevelt Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Top three words:**

* + **Nation**
  + **Spirit**
  + **People**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Kennedy Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Top three words:**

* + **Let**
  + **World**
  + **Sides**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Nixon Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Top three words:**

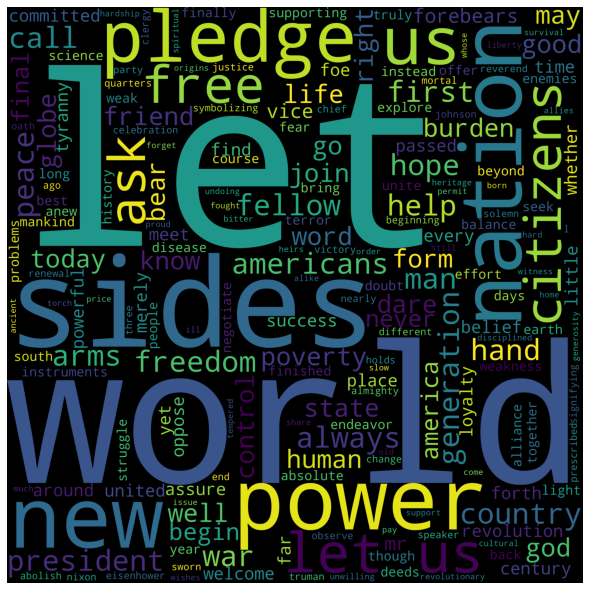
* + **World**
  + **Peace**
  + **America**

1. Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks [ refer to the End-to-End Case Study done in the Mentored Learning Session ]

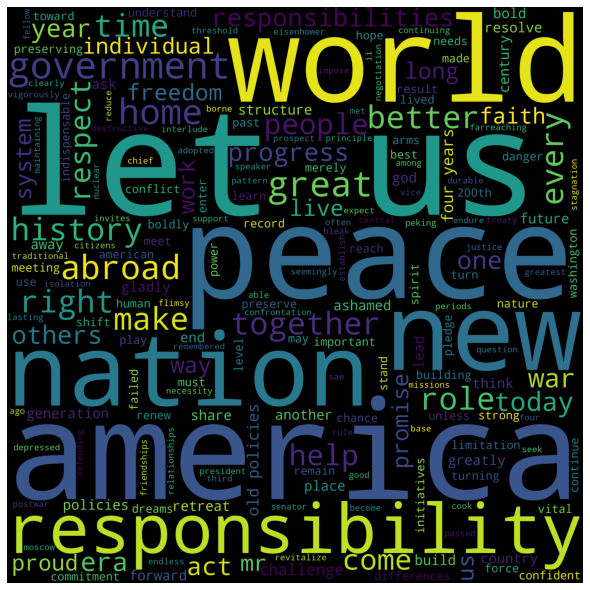
**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Roosevelt Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***



**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Kennedy Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***



**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Nixon Speech \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

****