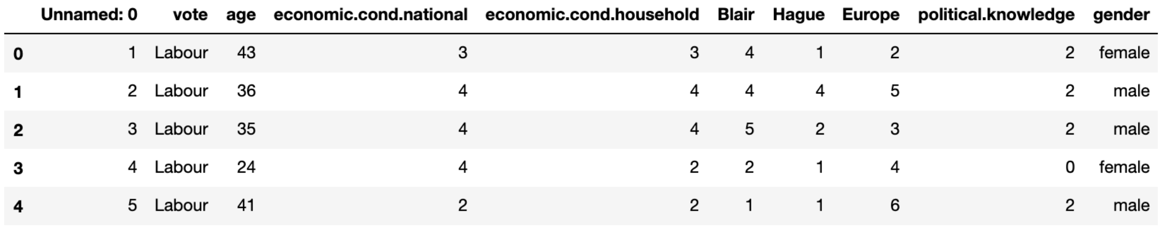
Project- Machine Learning

Problem 1:

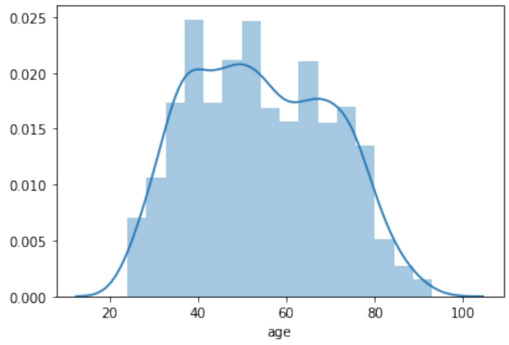
You are hired by one of the leading news channel CNBE who wants to analyse recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help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

- The data of the election survey looks like this

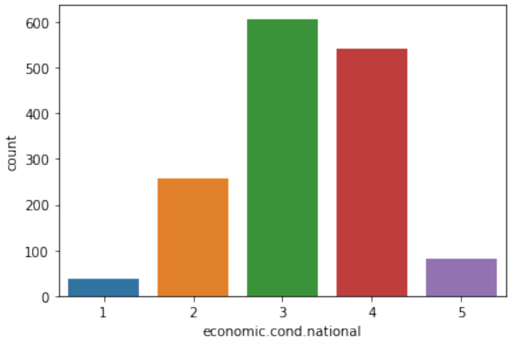


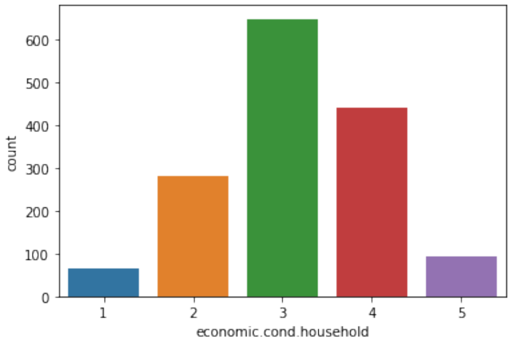
	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	1	Labour	43	3	3	4	1	2	2	female
1	2	Labour	36	4	4	4	4	5	2	male
2	3	Labour	35	4	4	5	2	3	2	male
3	4	Labour	24	4	2	2	1	4	0	female
4	5	Labour	41	2	2	1	1	6	2	male

- The data has 10 columns and 1525 entries, in which the column named Unnamed: 0 is of no use for us we will remove it
- There is no null values in the dataset
- There are no duplicate records in the dataset as well
- Except vote and gender rest of the columns are of integer datatype. Vote and gender being object data type
- Except age everything else seems to be categorical variable.
- On Analysing the data individually
    - Most of the people in the survey are between the age 25 and 80

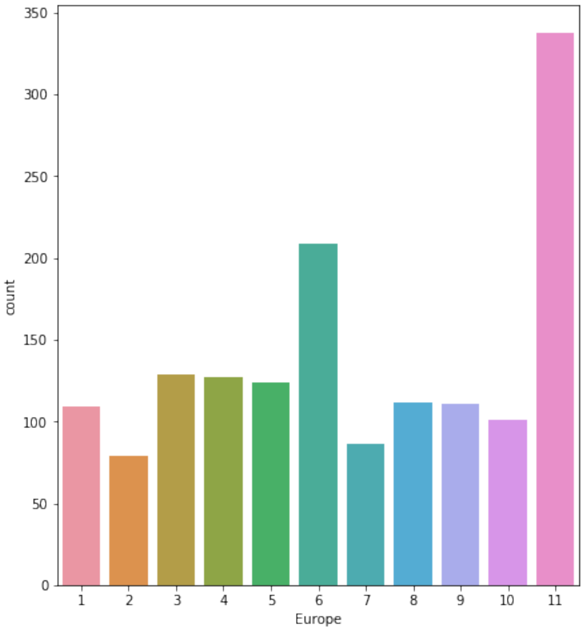


    - The national economic condition is between 3 and 4 but the household economic condition being predominantly 3

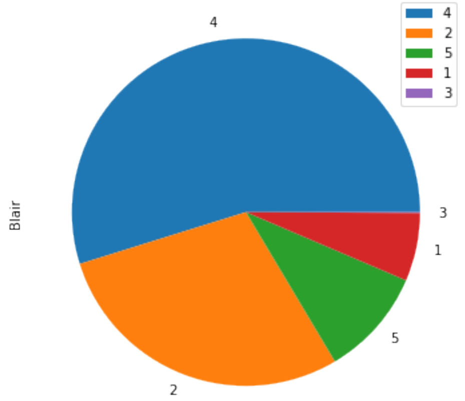




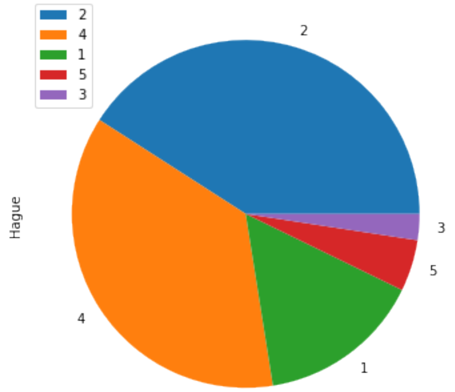
- ○ People sentiment over Europe’s integration is high



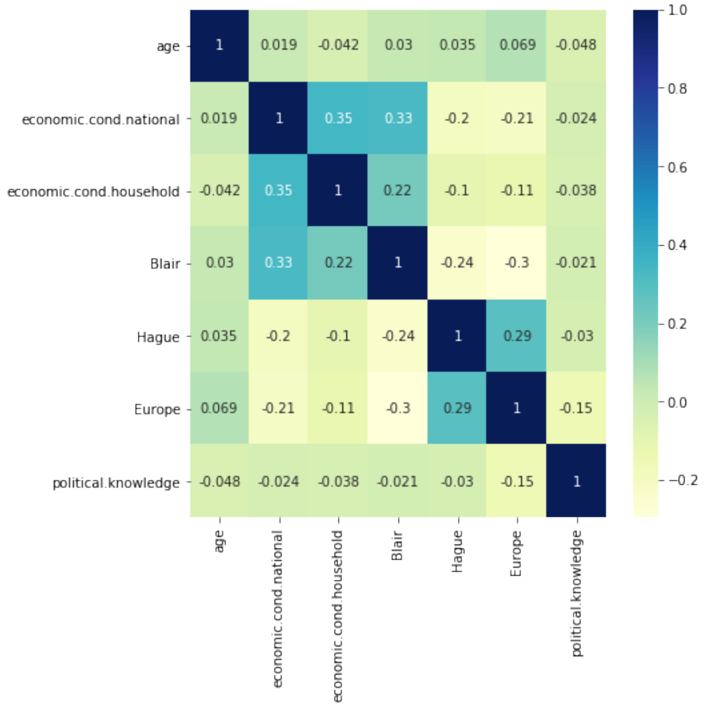
- The labour’s leaders assessment is more at 4



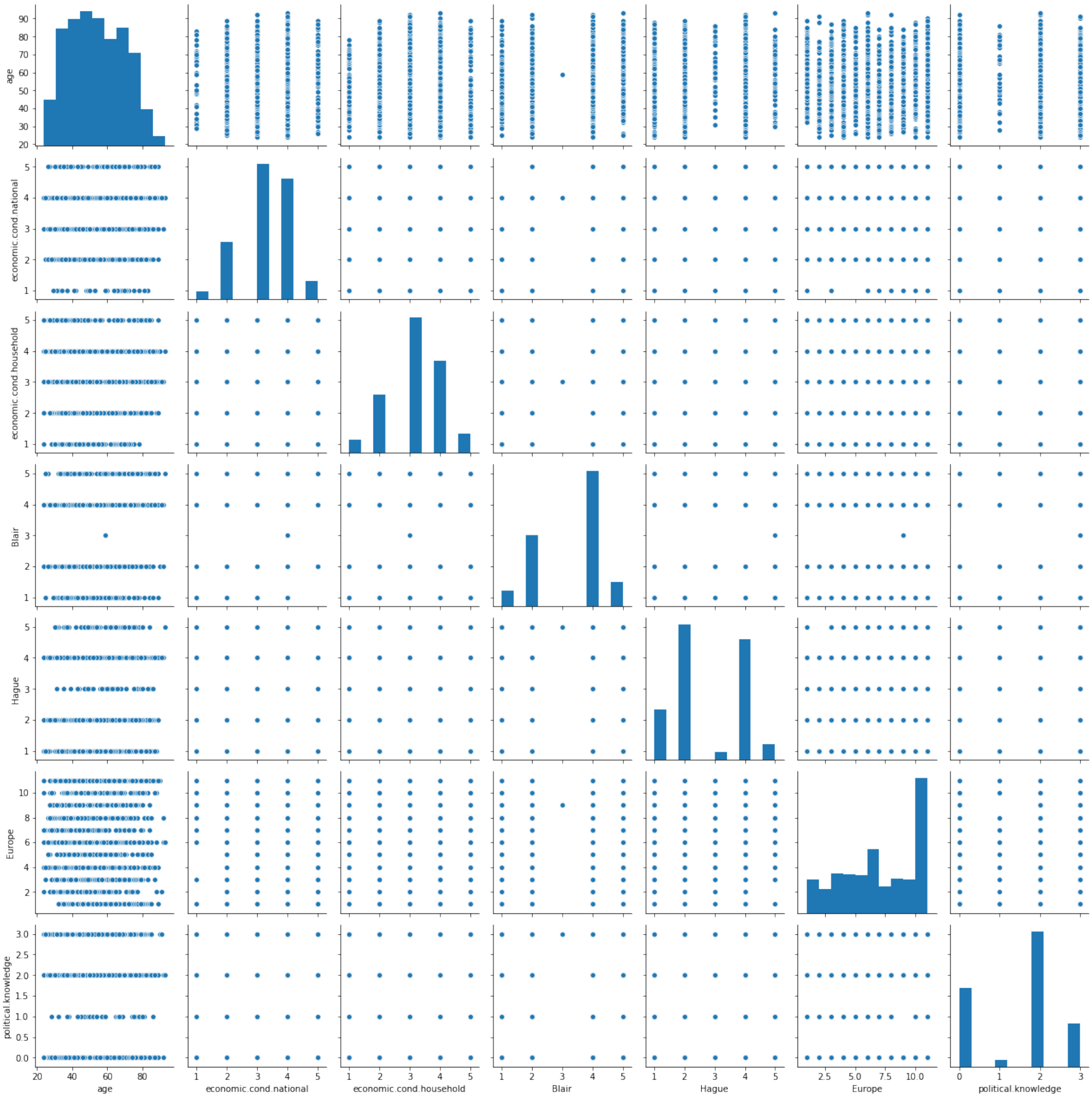
- The assessment on conservative’s leaders is at 2 though



- There is no much correlation between the data’s

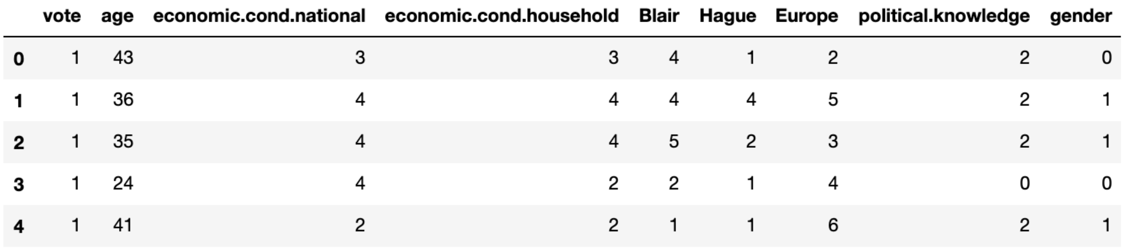


- When we draw a pair plot for our dataset it looks like this



- From above pair plot we can clearly all the variables are scattered and there is no correlation between them

- Before splitting our dataset lets encode the string variables gender and votes. Post encoding the data looks like this

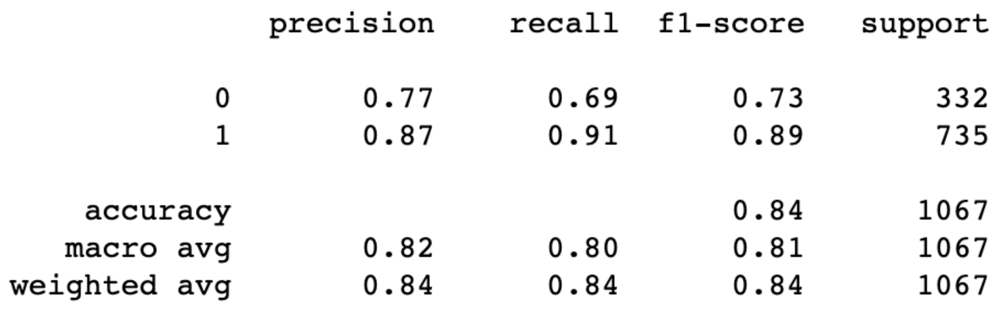


	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	1	43	3	3	4	1	2	2	0
1	1	36	4	4	4	4	5	2	1
2	1	35	4	4	5	2	3	2	1
3	1	24	4	2	2	1	4	0	0
4	1	41	2	2	1	1	6	2	1

- Let's split the data into 70% training and 30% test data and apply different models on the data

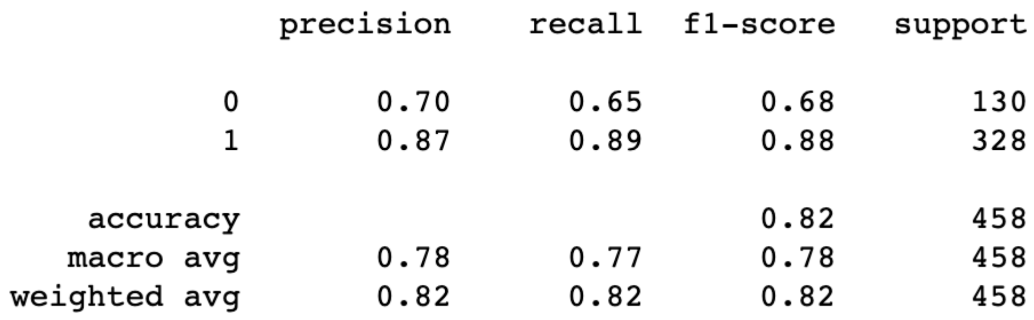
1. Logistic Regression and LDA:

    a. The classification report for Logistic Regression training dataset is



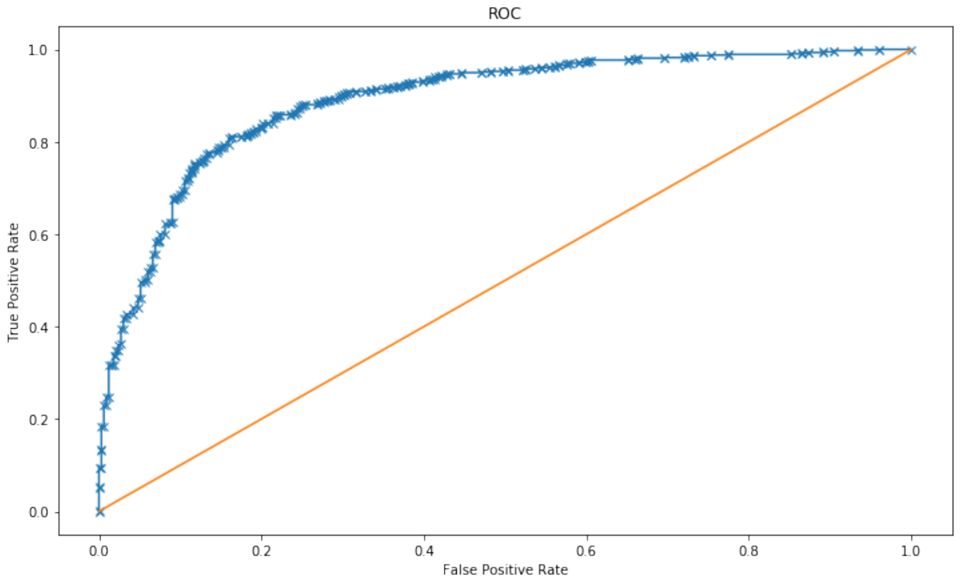
```
              precision    recall  f1-score   support

           0       0.77      0.69      0.73       332
           1       0.87      0.91      0.89       735

    accuracy                           0.84      1067
   macro avg       0.82      0.80      0.81      1067
weighted avg       0.84      0.84      0.84      1067
```

    b. The classification report for Logistic Regression test dataset is

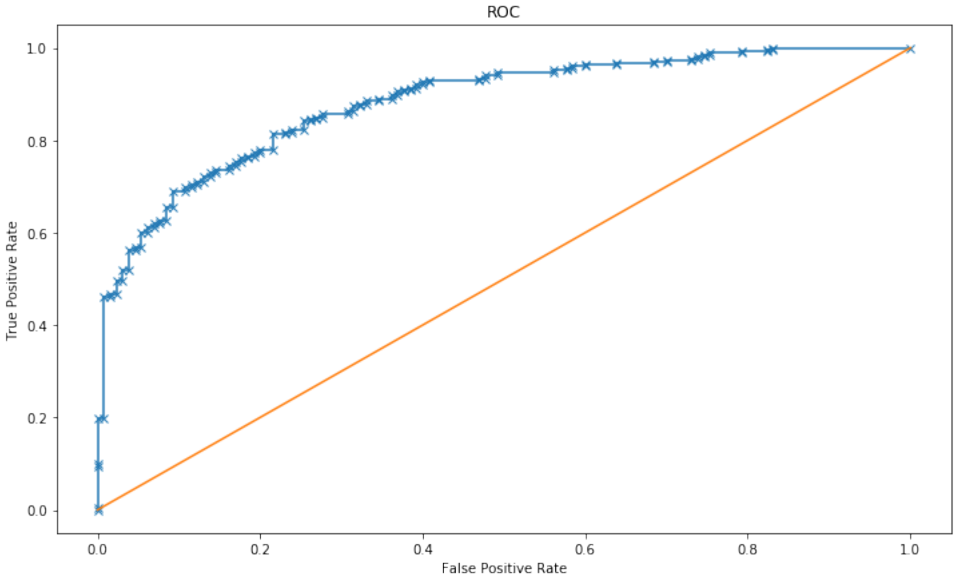


```
              precision    recall  f1-score   support

           0       0.70      0.65      0.68       130
           1       0.87      0.89      0.88       328

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.78       458
weighted avg       0.82      0.82      0.82       458
```

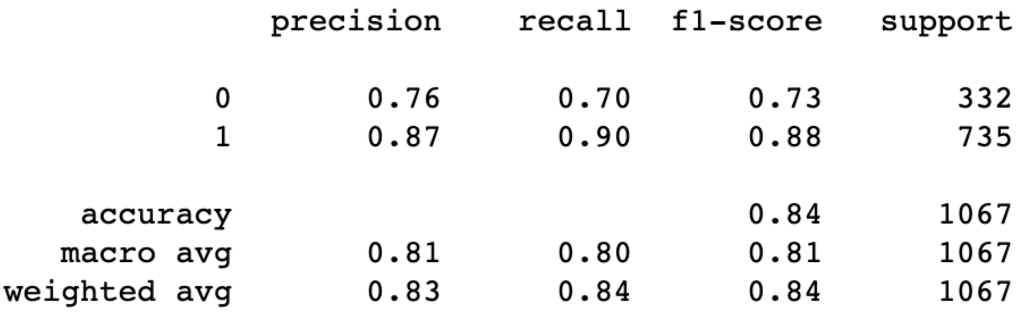
    c. The ROC-AUC score for training data is 0.89, ROC Curve for training dataset using Logistic Regression is



d. The ROC-AUC score for testing dataset is 0.88 ROC Curve for testing dataset using Logistic Regression is

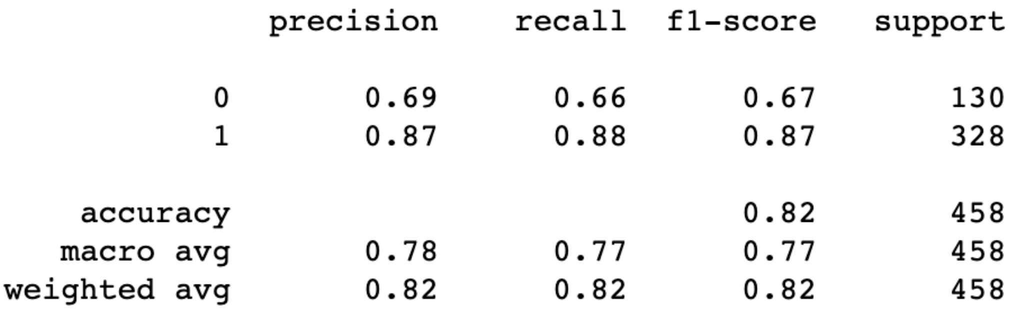


e. The classification report for LDA training dataset is



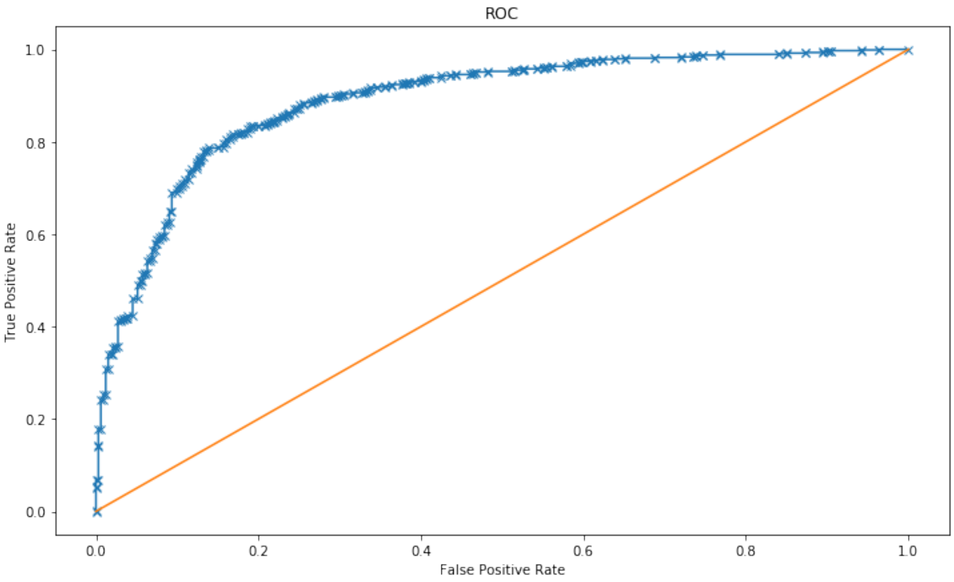
```
              precision    recall  f1-score   support

           0       0.76      0.70      0.73       332
           1       0.87      0.90      0.88       735

    accuracy                           0.84      1067
   macro avg       0.81      0.80      0.81      1067
weighted avg       0.83      0.84      0.84      1067
```

f. The classification report of LDA testing dataset is

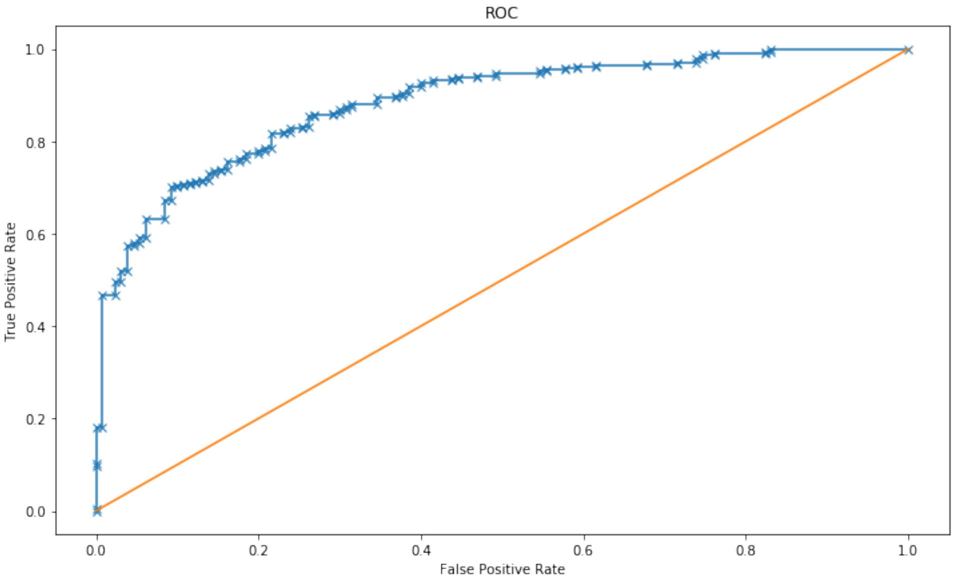


```
              precision    recall  f1-score   support

           0       0.69      0.66      0.67       130
           1       0.87      0.88      0.87       328

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.77       458
weighted avg       0.82      0.82      0.82       458
```

g. The ROC-AUC score for testing dataset is 0.88 ROC Curve for training dataset using LDA is



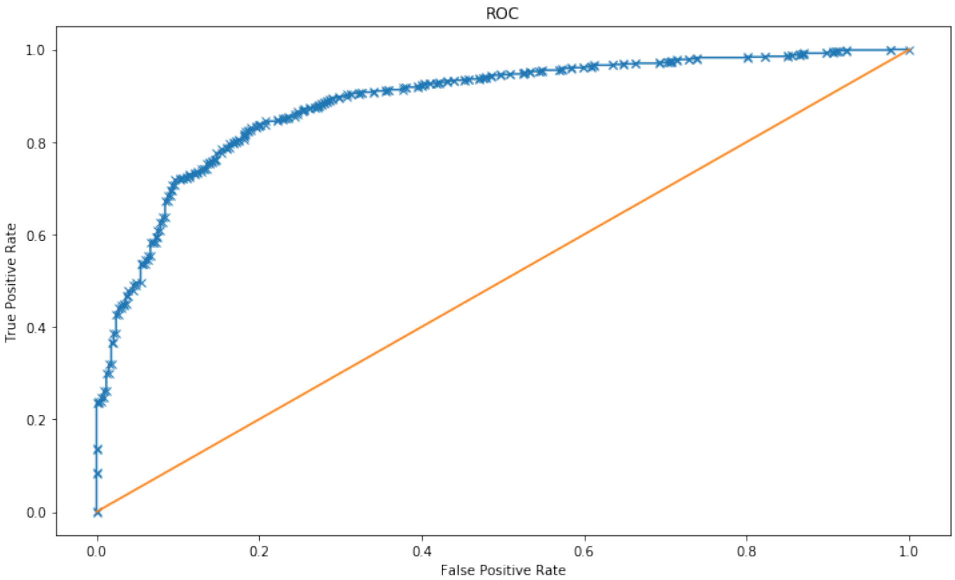
h. The ROC-AUC score for testing dataset is 0.88 ROC Curve for testing dataset using LDA is



2. NN Model, Naïve Bayes Model and support vector machine (SVM) model

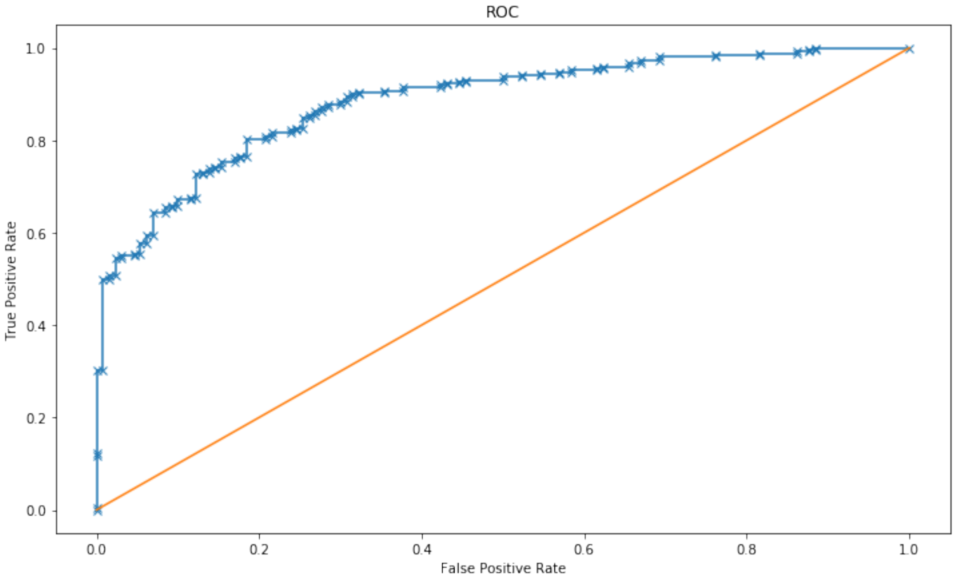
a. The Classification report and ROC_AUC curve for training data in Naïve Bayes model is And the AUC is **0.886**

```
              precision    recall  f1-score   support

           0       0.74      0.72      0.73       332
           1       0.88      0.88      0.88       735

    accuracy                           0.83      1067
   macro avg       0.81      0.80      0.80      1067
weighted avg       0.83      0.83      0.83      1067
```

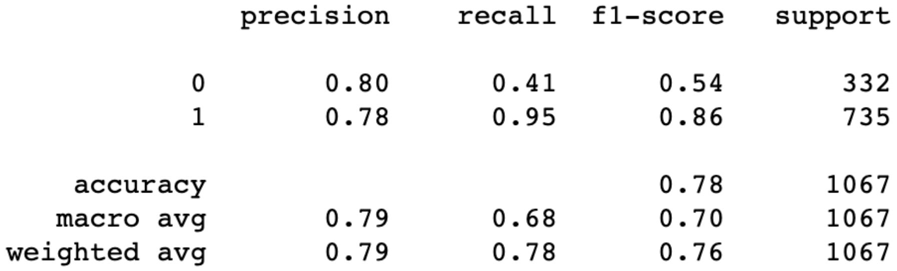


b. The Classification report and ROC_AUC curve for testing data in Naïve Bayes model is And the AUC is **0.884**

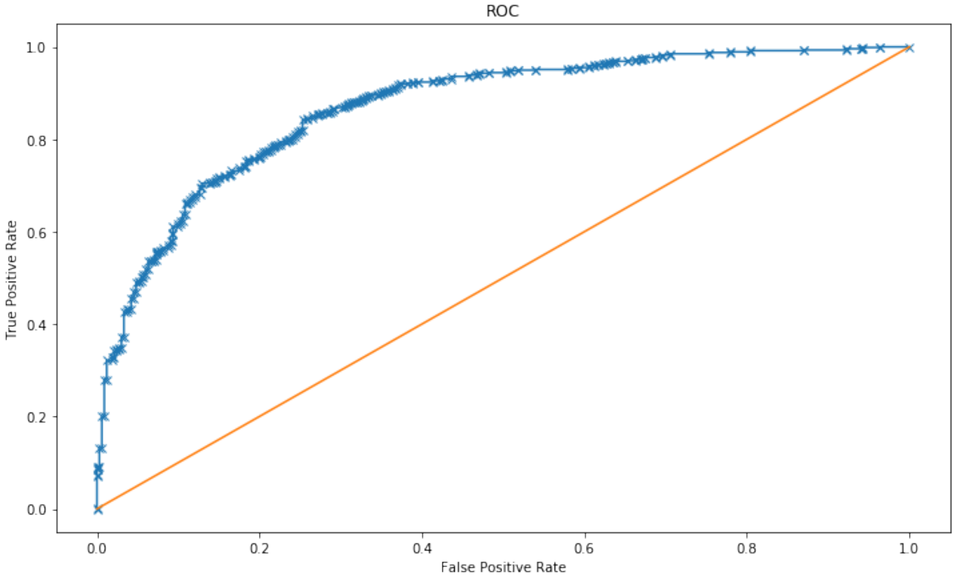
```
              precision    recall  f1-score   support

           0       0.68      0.72      0.70       130
           1       0.89      0.87      0.88       328

    accuracy                           0.83       458
   macro avg       0.78      0.79      0.79       458
weighted avg       0.83      0.83      0.83       458
```



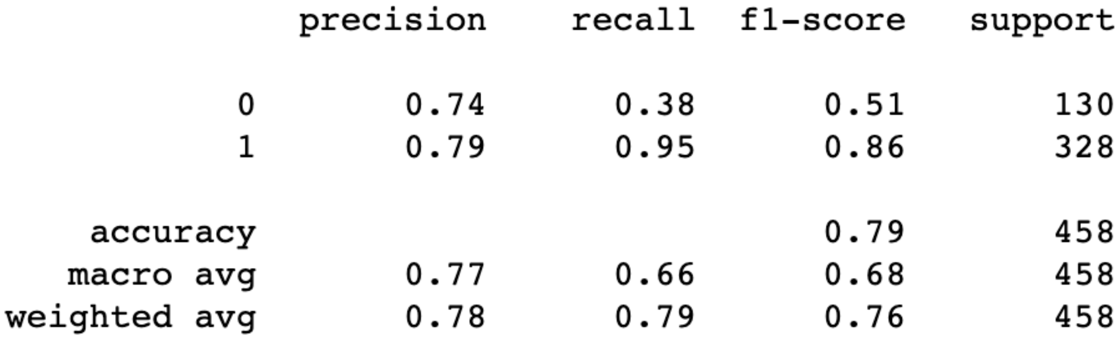
c. The Classification report and ROC_AUC curve for training data in SVM model is And the AUC is **0.87**



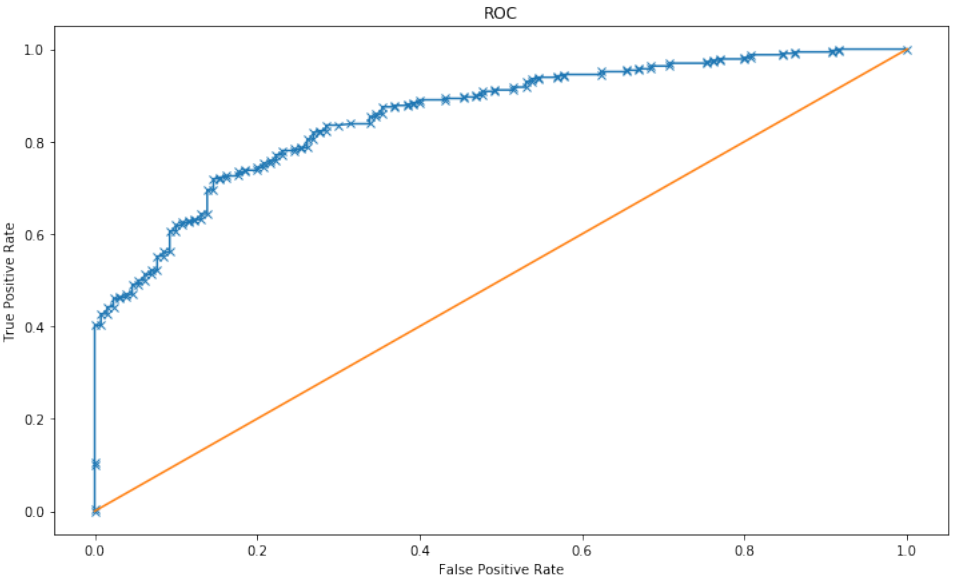
```
              precision    recall  f1-score   support

           0       0.80      0.41      0.54       332
           1       0.78      0.95      0.86       735

    accuracy                           0.78      1067
   macro avg       0.79      0.68      0.70      1067
weighted avg       0.79      0.78      0.76      1067
```



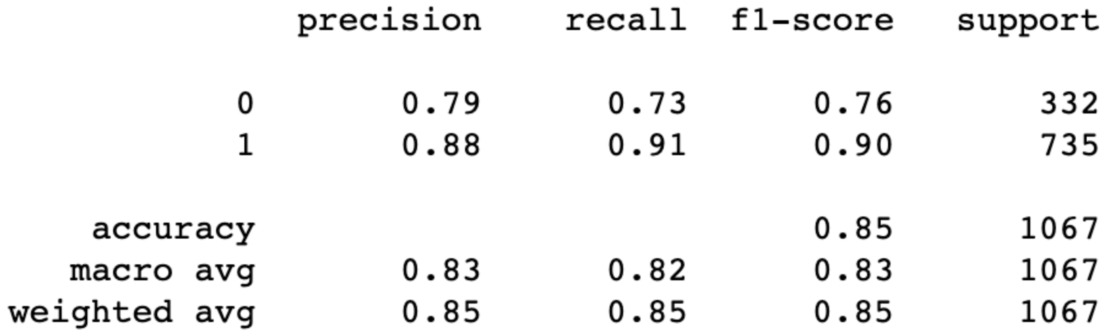
d. The Classification report and ROC_AUC curve for testing data in SVM model is And the AUC is **0.857**



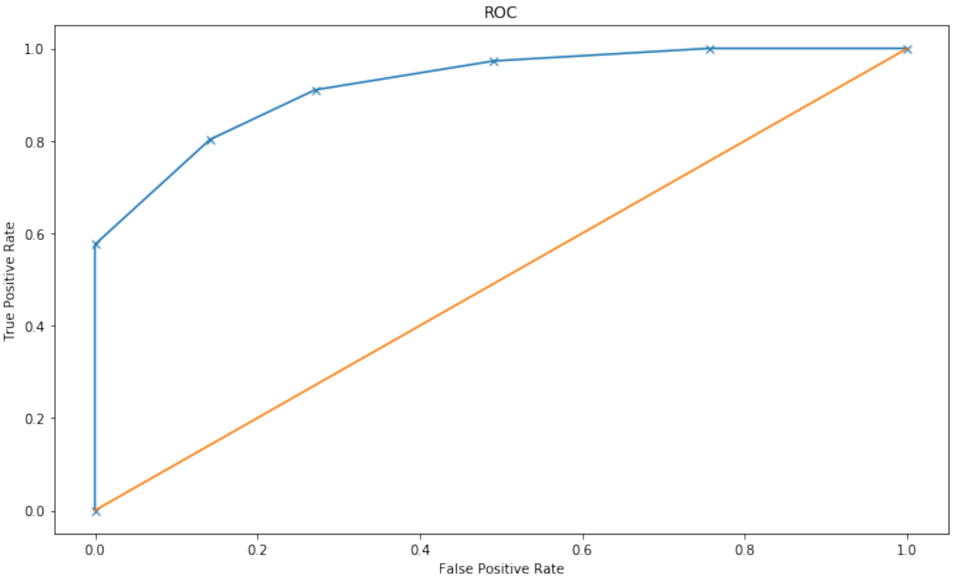
```
              precision    recall  f1-score   support

           0       0.74      0.38      0.51       130
           1       0.79      0.95      0.86       328

    accuracy                           0.79       458
   macro avg       0.77      0.66      0.68       458
weighted avg       0.78      0.79      0.76       458
```



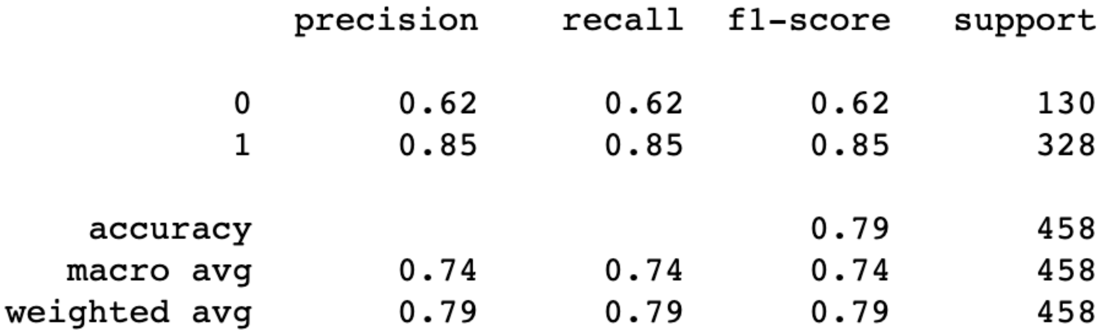
e. The Classification report and ROC_AUC curve for training data in KNN model is and the AUC is **0.92**



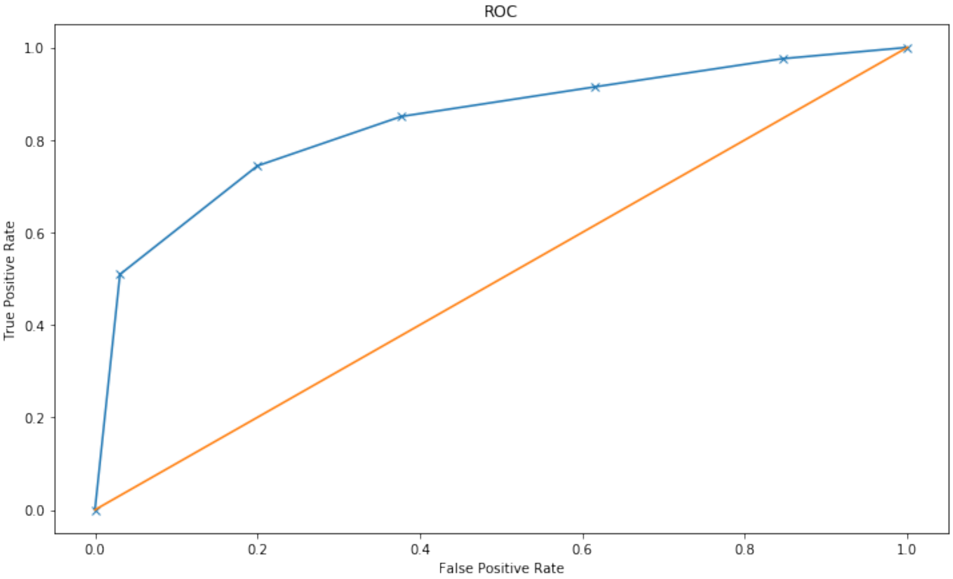
```
              precision    recall  f1-score   support

           0       0.79      0.73      0.76       332
           1       0.88      0.91      0.90       735

    accuracy                           0.85      1067
   macro avg       0.83      0.82      0.83      1067
weighted avg       0.85      0.85      0.85      1067
```



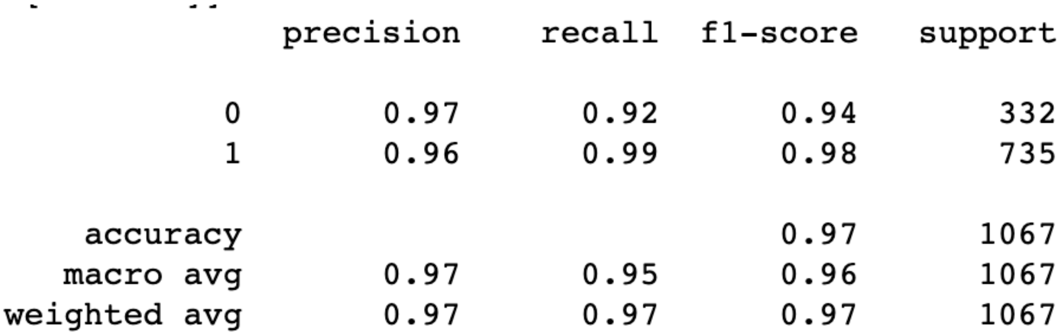
f. The Classification report and ROC_AUC curve for testing data in KNN model is and the AUC is **0.835**



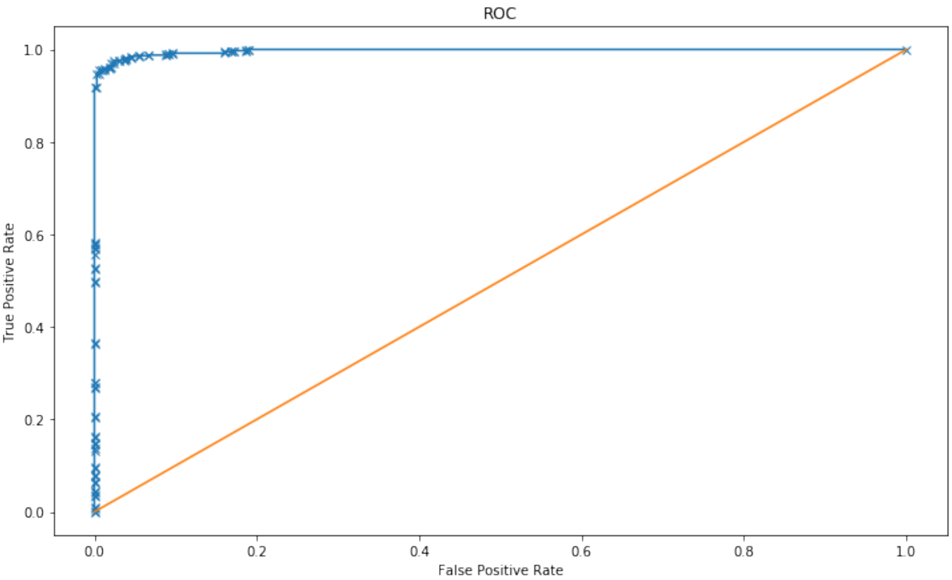
```
              precision    recall  f1-score   support

           0       0.62      0.62      0.62       130
           1       0.85      0.85      0.85       328

    accuracy                           0.79       458
   macro avg       0.74      0.74      0.74       458
weighted avg       0.79      0.79      0.79       458
```



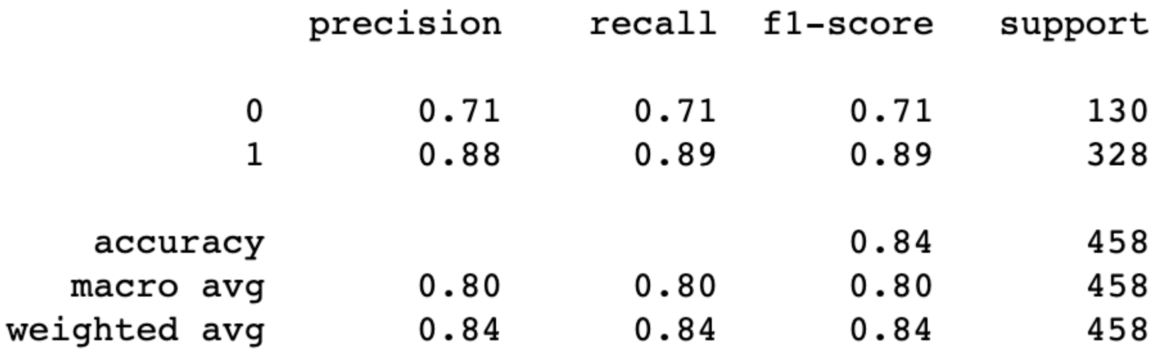
3. Model Tuning, Bagging (Random Forest should be applied for Bagging) and Boosting
   a. The classification Report and AUC_ROC curve for training data using bagging is and AUC is **0.99**



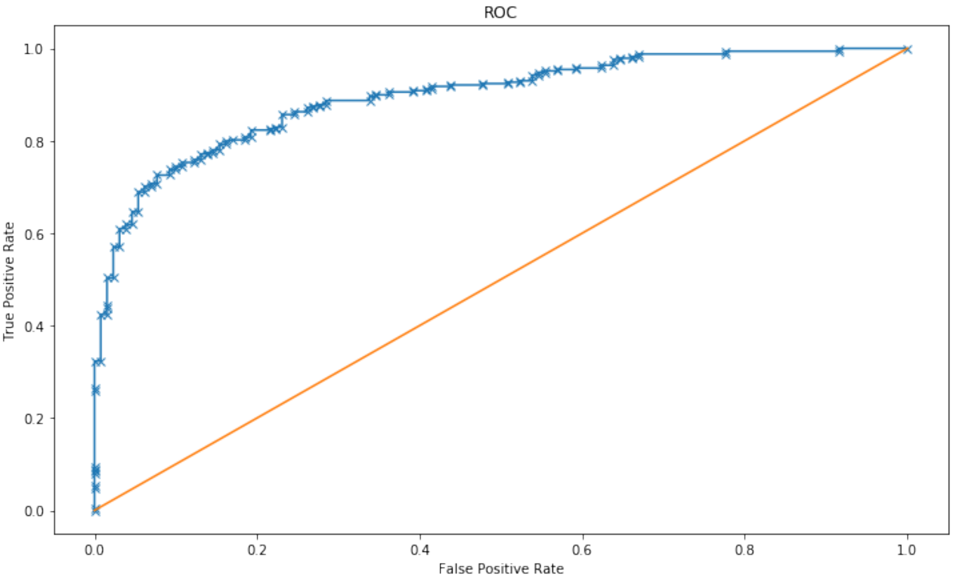
```
              precision    recall  f1-score   support

           0       0.97      0.92      0.94       332
           1       0.96      0.99      0.98       735

    accuracy                           0.97      1067
   macro avg       0.97      0.95      0.96      1067
weighted avg       0.97      0.97      0.97      1067
```



b. The classification Report and AUC_ROC curve for testing data using bagging is and AUC is **0.89**



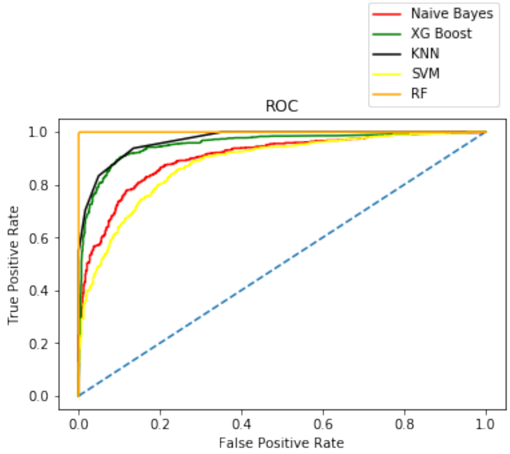
```
              precision    recall  f1-score   support

           0       0.71      0.71      0.71       130
           1       0.88      0.89      0.89       328

    accuracy                           0.84       458
   macro avg       0.80      0.80      0.80       458
weighted avg       0.84      0.84      0.84       458
```



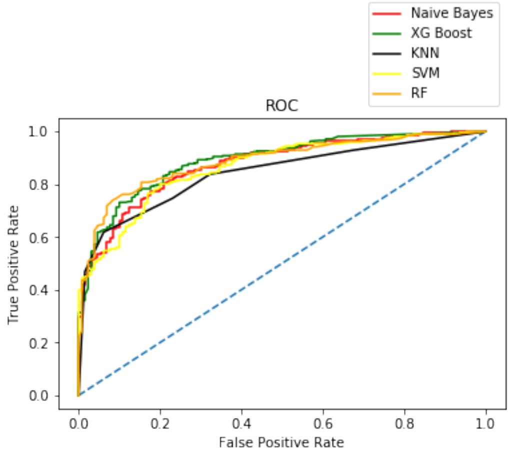
c. For model tuning we will use SMOTE technique as the data between the votes is not balanced we will use SMOTE technique to balance it

d. For Boosting we are using XGBoosting

e. After applying SMOTE and XGBoosting to all the models below is the comparison ROC curve for training data



f. After applying SMOTE and XGBoosting to all the models below is the comparison ROC curve for testing data



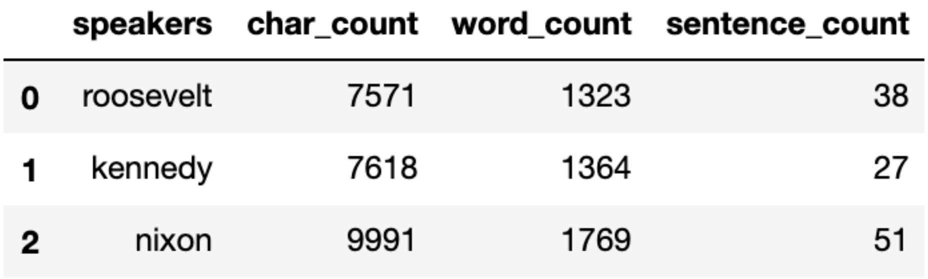
4. By Analysing all the models **XGBoosting** model seems to be better performing as there is not much difference between Training and test data. The AUC score for training data is 0.93 and Testing data is 0.89
5. We also cross validated the model and seems to have very less difference between training and test dataset

6. Inference: The two important parameters for predicting the voters who will vote for Labour are Blair and Europe if a voter is highly sentimental towards Europe and has high assessment on labour leaders he is tend to vote for LABOUR

Problem 2:

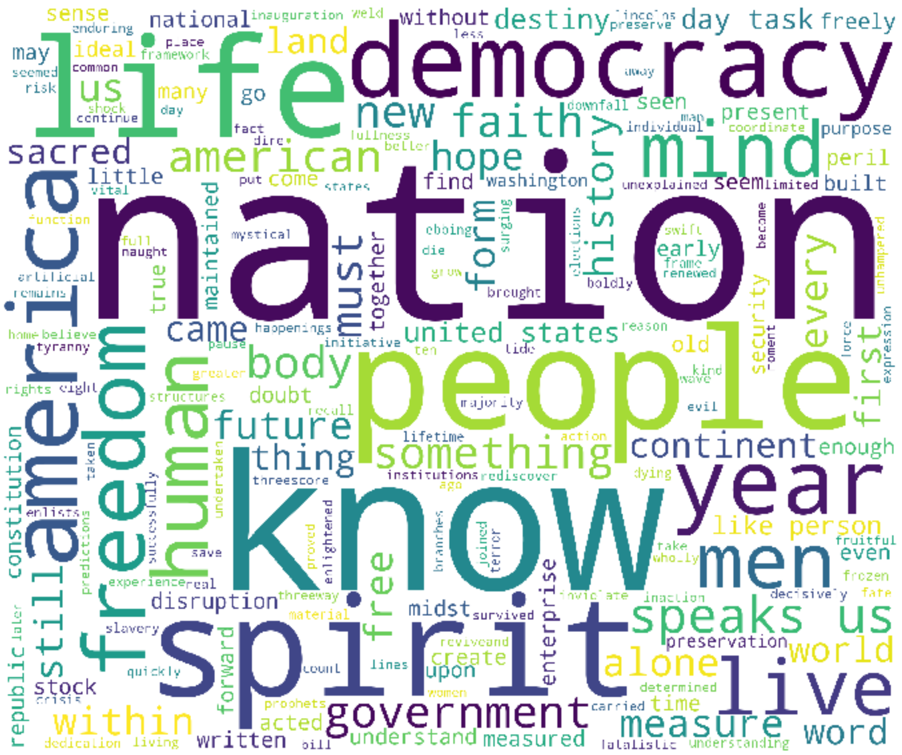
In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America

1. The number of characters, words and sentence count in each of the president is

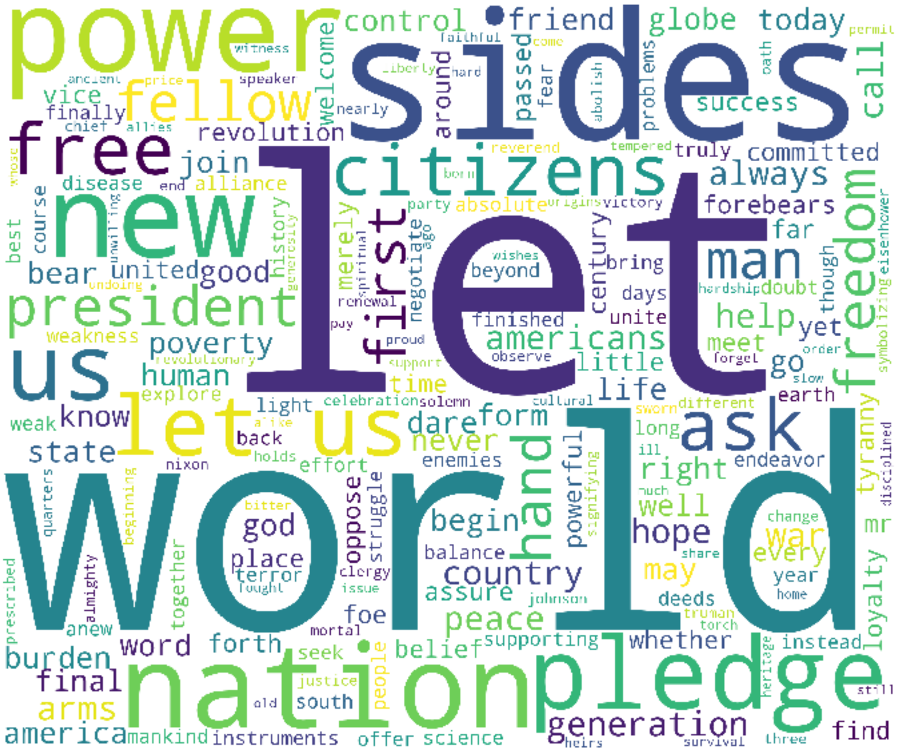


	speakers	char_count	word_count	sentence_count
0	roosevelt	7571	1323	38
1	kennedy	7618	1364	27
2	nixon	9991	1769	51

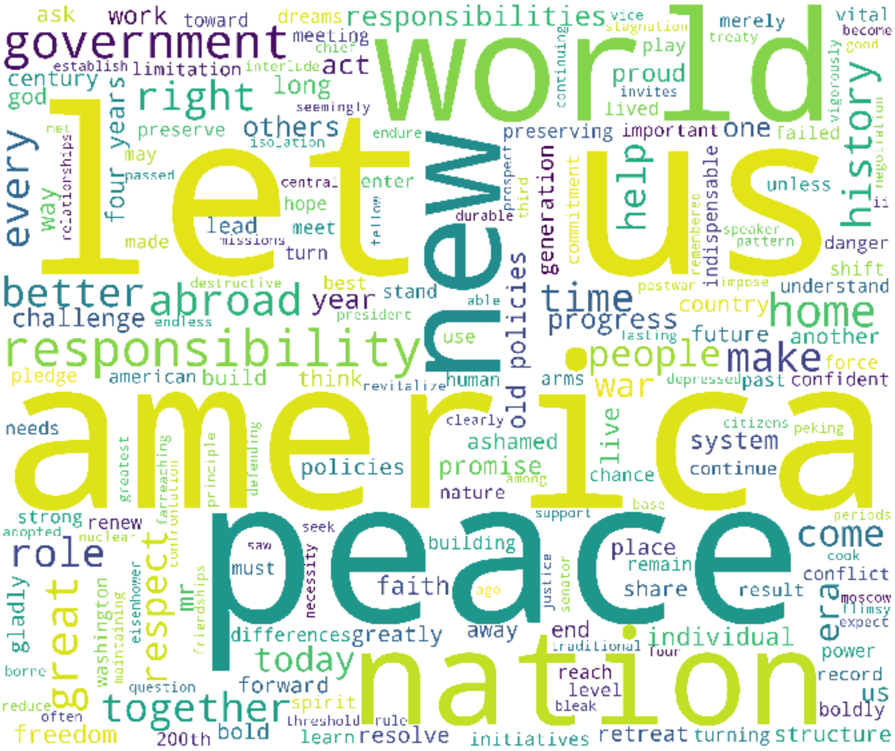
2. We will use nltk.corpus to get the stop words of English and we will remove them in each of the presidents speech
3. The words that occurs most of the time in Roosevelt's speech are **nation**, **know**, **democracy**
4. The words that occurs most of the time in Kennedy's speech are **let**, **us**, **sides**
5. The words that occurs most of the time in Nixon's speech are **us**, **let**, **peace.**
6. The Word cloud for Roosevelt's speech is



7. The word cloud for Kennedy’s speech is
8.



9. The word cloud for Nixon’s speech is


ask
work
toward
dreams
responsibilities
merely
vital
government
world
right
long
act
century
god
let us
new
every
years
others
preserve
preserving
important
one
failed
history
help
proud
play
lived
lead
hope
meet
turn
enter
unless
danger
better
abroad
year
stand
challenge
responsibility
time
progress
home
people
make
future
country
understand
another
war
arms
confident
past
human
build
american
think
pledge
america
peace
needs
policies
promise
ashamed
system
continue
chance
nature
role
great
respect
strong
renew
must
faith
building
place
remain
share
result
come
conflict
era
gladly
washington
mr
differences
greatly
away
end
individual
today
forward
nation
reach
level
power
record
us
boldly
together
spirit
freedom
200th
bold
learn
resolve
initiatives
retreat
turning
structure