MACHINE LEARNING PROJECT

Done By

Hariharan Manickam

Problem 1

1.1) 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.

Out[3]:

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

Out[9]:

	count	mean	std	min	25%	50%	75%	max
age	1525.0	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0

1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) 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.

Shape :

```
Out[6]: (1525, 9)
```

Null Values:

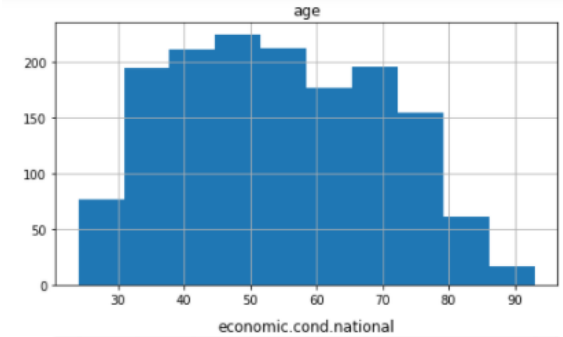
```
Out[8]: vote                    0
        age                     0
        economic.cond.national  0
        economic.cond.household 0
        Blair                   0
        Hague                   0
        Europe                  0
        political.knowledge     0
        gender                  0
        dtype: int64
```

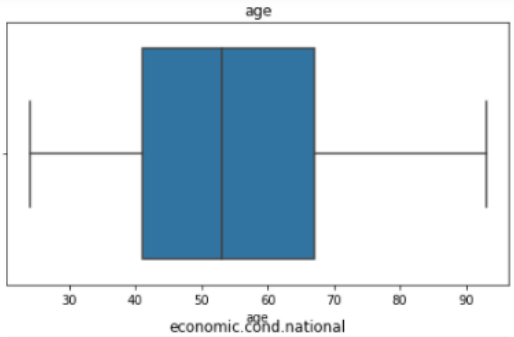
Data Type before categorising :

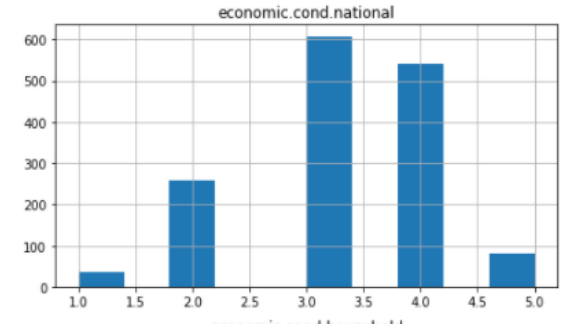
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

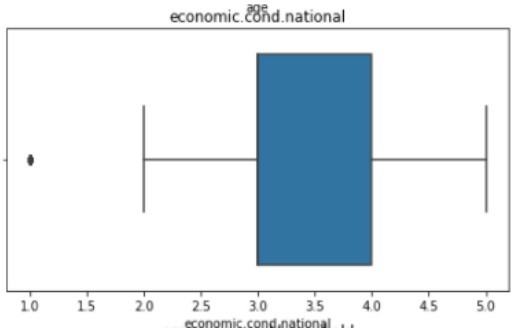
Data type after categorising :

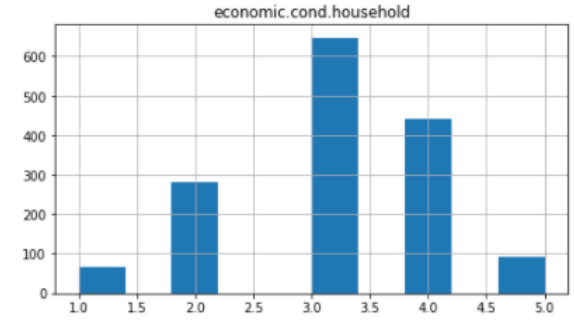
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   int8
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   int8
dtypes: int64(7), int8(2)
memory usage: 86.5 KB
```

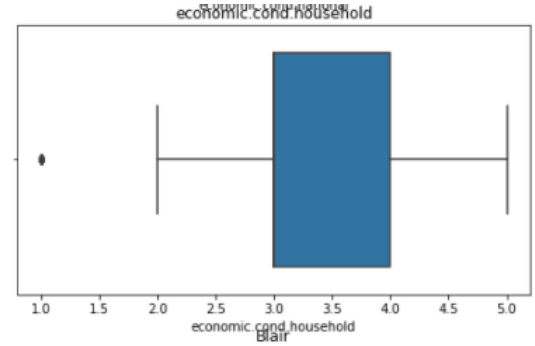

age
200
150
100
50
0
30
40
50
60
70
80
90

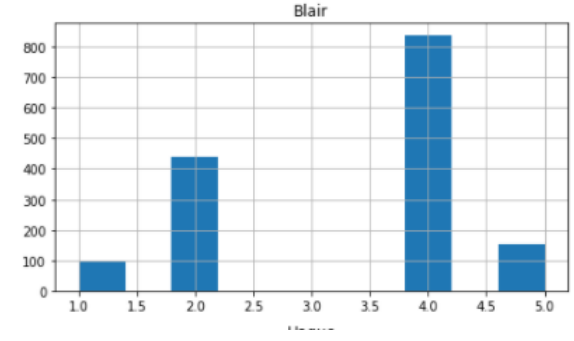

age
30
40
50
60
70
80
90
age

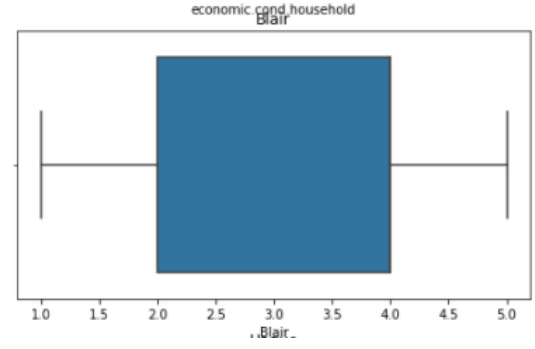

economic.cond.national
600
500
400
300
200
100
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

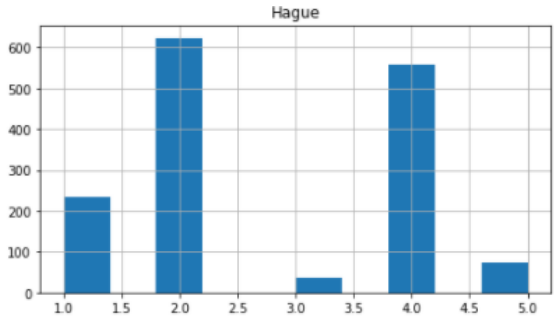

economic.cond.national
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
economic.cond.national

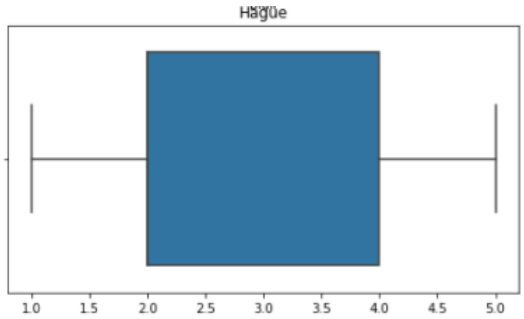

economic.cond.household
600
500
400
300
200
100
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

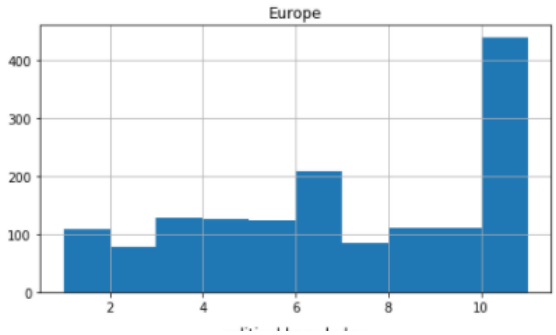

economic.cond.household
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
economic.cond.household

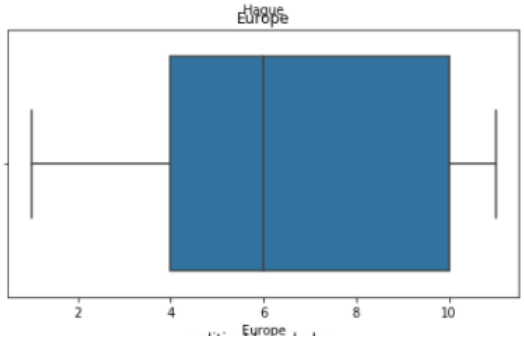

Blair
800
700
600
500
400
300
200
100
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

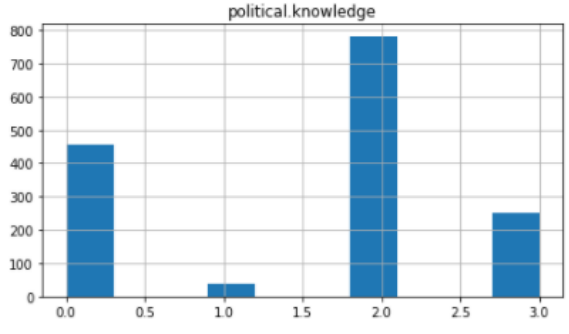

Blair
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Blair

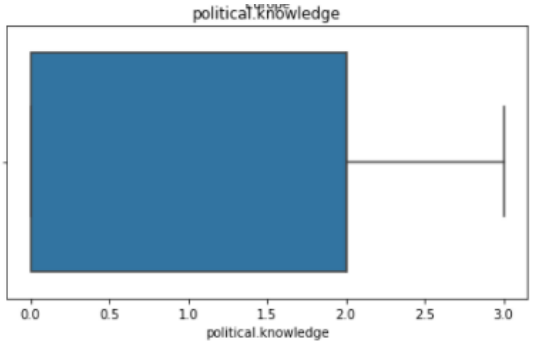












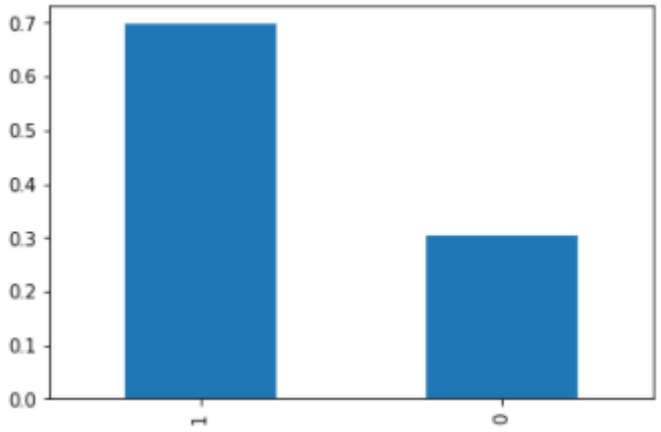
Out[16]:

vote	-0.858449
age	0.144621
economic.cond.national	-0.240453
economic.cond.household	-0.149552
Blair	-0.535419
Hague	0.152100
Europe	-0.135947
political.knowledge	-0.426838
gender	0.130239

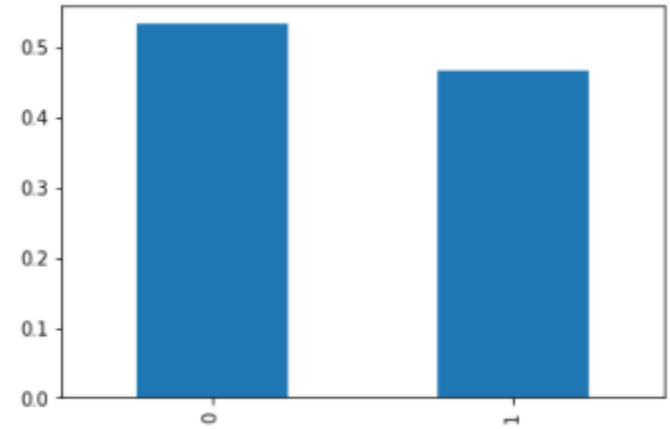
dtype: float64

Histograms

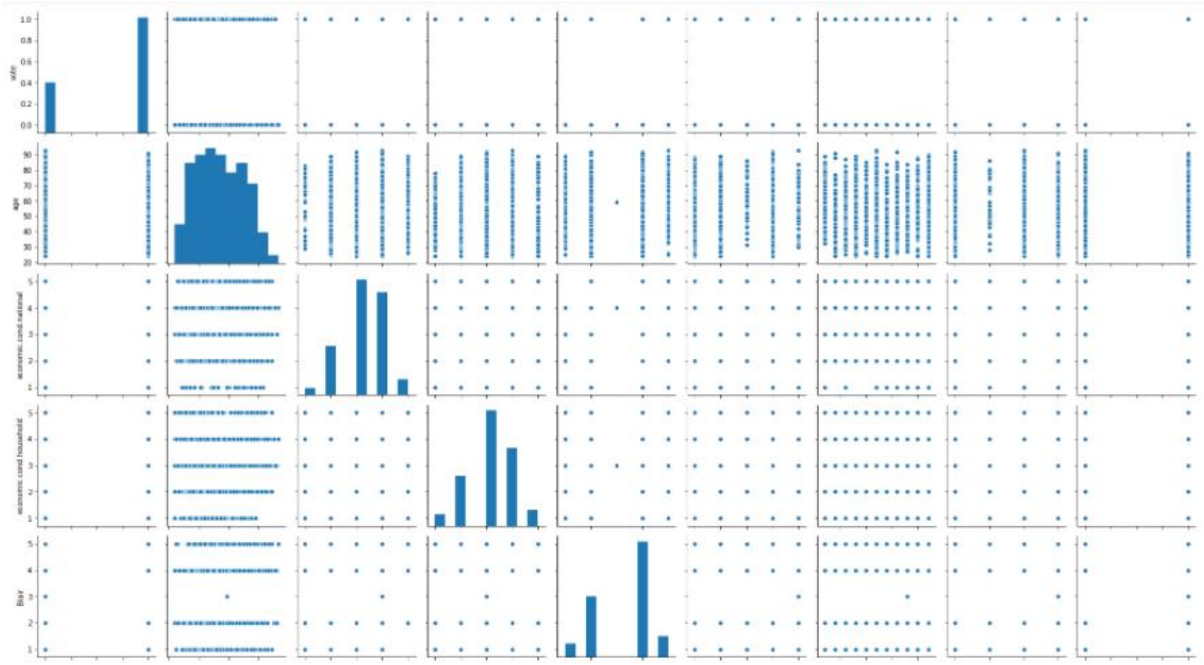
Voters (yes – 1, no – 0)

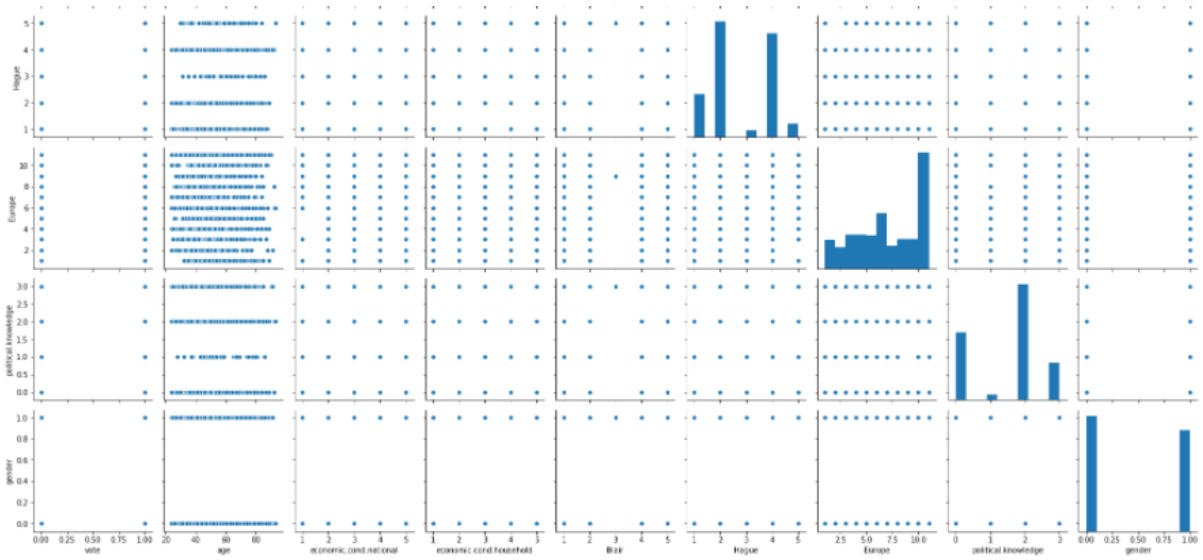


Gender (Male – 0, Female – 1)

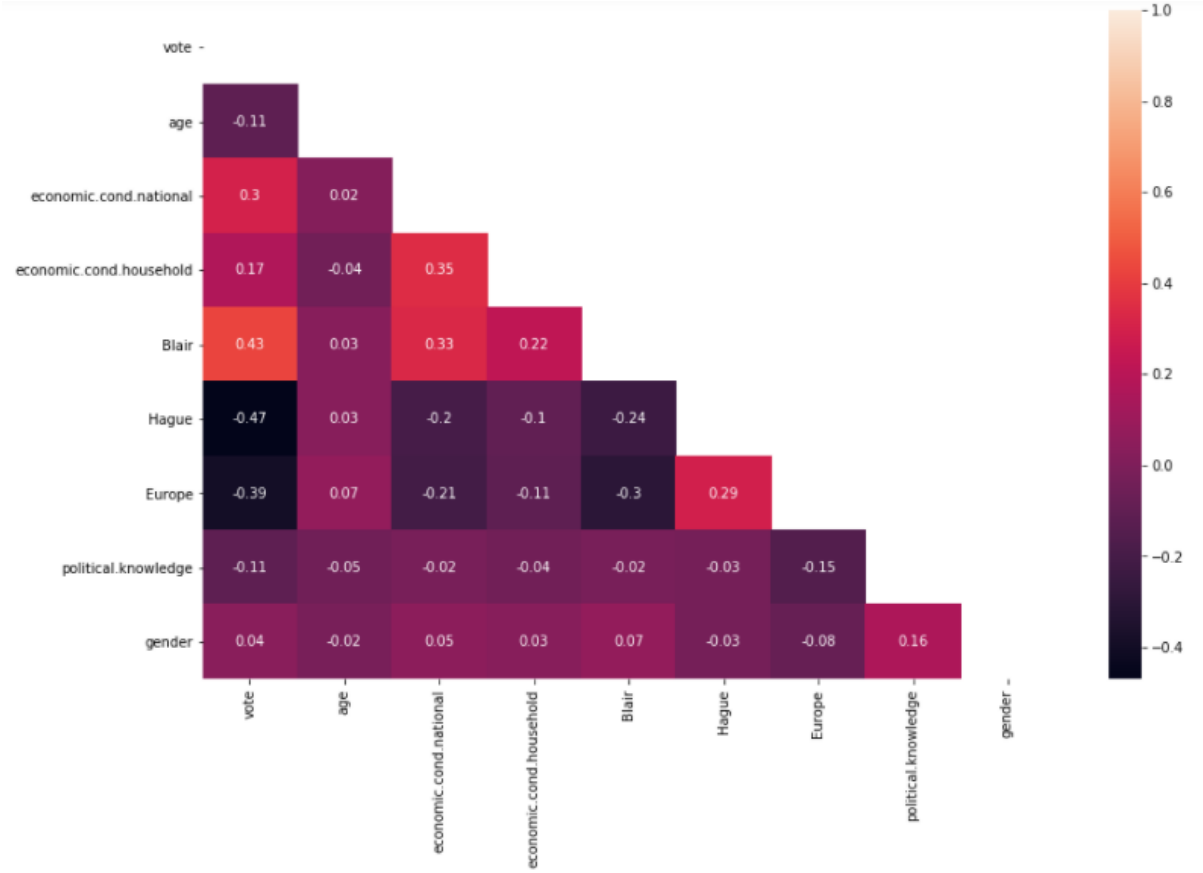


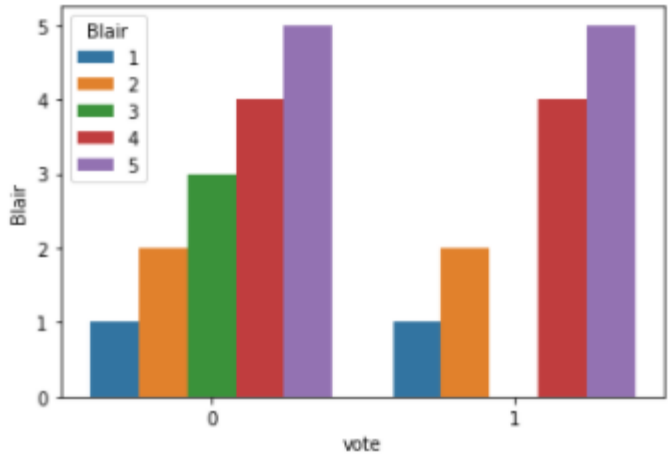
Bivariate Analysis

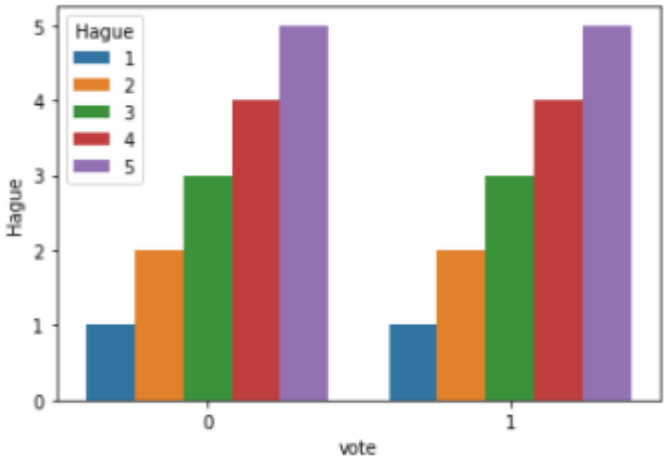




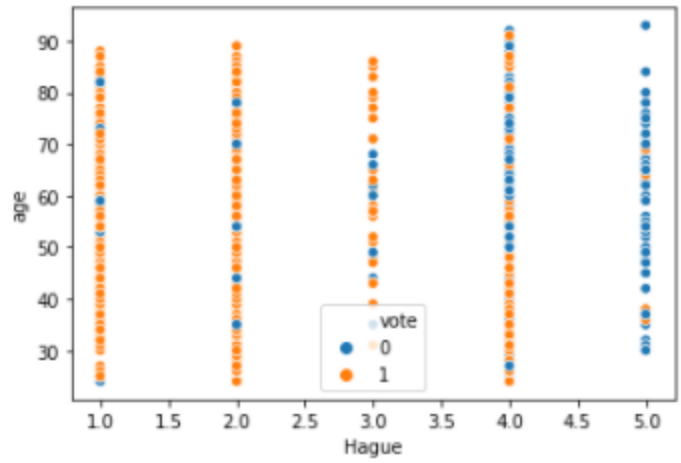
Correlation

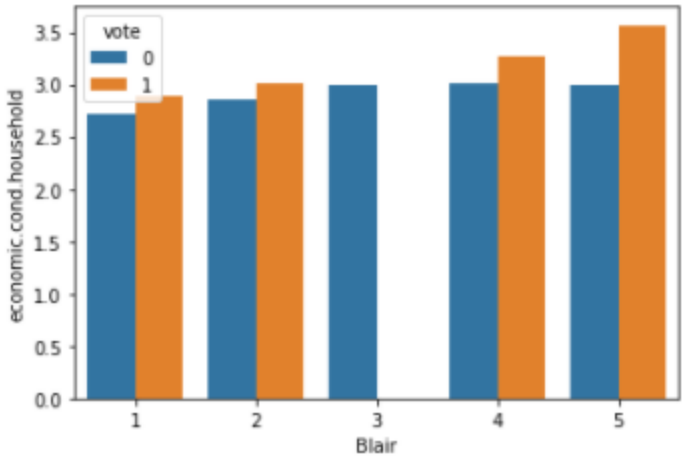


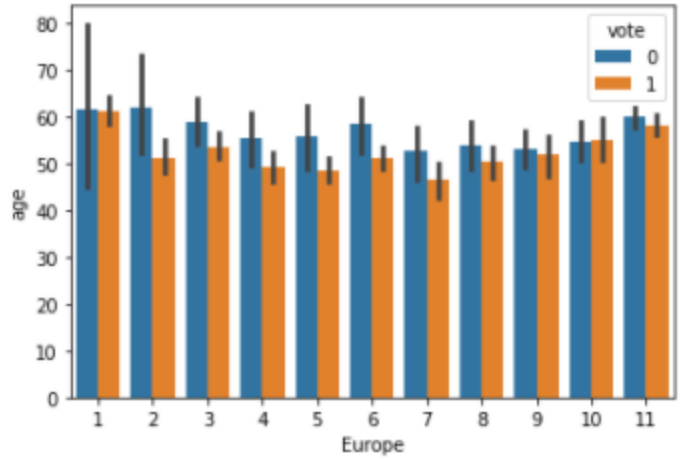


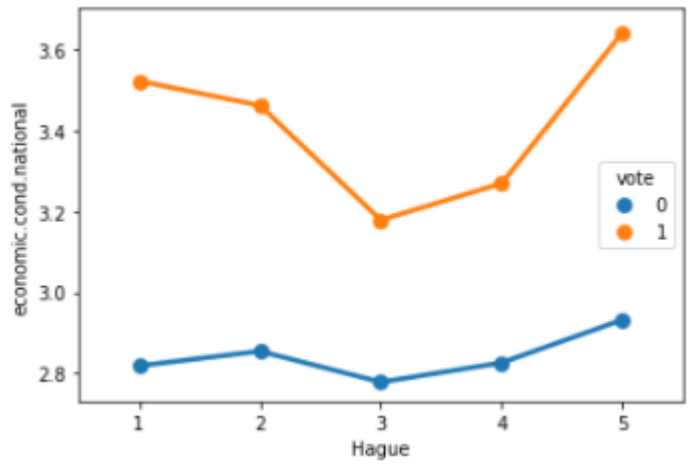


Multivariate Analysis




vote
0
1
economic.cond.household
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
Blair


vote
0
1
age
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
Europe


vote
0
1
economic.cond.national
3.6
3.4
3.2
3.0
2.8
1
2
3
4
5
Hague

1.3) 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.

- Scaling is not necessary for logistic regression, LDA and Naive Bayes models.
- It should be done for KNN model and Random forest algorithm.

1.4) 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)

Logistic Regression

```
Out[30]: LogisticRegression(C=1, intercept_scaling=1.0, penalty='l1', solver='liblinear')
```

Training Score, Confusion Matrix & Classification Report

```
0.8406747891283973
```

```
[[230 102]
 [ 68 667]]
```

	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

Testing Score, Confusion Matrix & Classification Report

```
0.8231441048034934
```

```
[[ 85  45]
 [ 36 292]]
```

	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

Hence, Linear Regression is decently fitted

LDA

```
Out[33]: LinearDiscriminantAnalysis(shrinkage=0.01, solver='lsqr')
```

Training Score, Confusion Matrix & Classification Report

```
0.8425492033739457


[[226 106]
 [ 62 673]]


              precision    recall  f1-score   support

           0       0.78      0.68      0.73       332
           1       0.86      0.92      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
```

Testing Score, Confusion Matrix & Classification Report

```
0.8275109170305677


[[ 85  45]
 [ 34 294]]


              precision    recall  f1-score   support

           0       0.71      0.65      0.68       130
           1       0.87      0.90      0.88       328

    accuracy                           0.83       458
   macro avg       0.79      0.78      0.78       458
weighted avg       0.82      0.83      0.83       458
```

Hence, LDA is decently fitted.

1.5) 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)

KNN Model

```
Out[41]: KNeighborsClassifier()
```

Training Score, Confusion Matrix & Classification Report

```
0.8631677600749765
```

```
[[248  84]
 [ 62 673]]
```

	precision	recall	f1-score	support
0	0.80	0.75	0.77	332
1	0.89	0.92	0.90	735
accuracy			0.86	1067
macro avg	0.84	0.83	0.84	1067
weighted avg	0.86	0.86	0.86	1067

Testing Score, Confusion Matrix & Classification Report

```
0.8253275109170306
```

```
[[ 92  38]
 [ 42 286]]
```

	precision	recall	f1-score	support
0	0.69	0.71	0.70	130
1	0.88	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

Hence KNN Model seems like a good fit.

Naïve Bees Model

```
Out[36]: GaussianNB()
```

Training Score, Confusion Matrix & Classification Report

```
0.8331771321462043


[[240  92]
 [ 86 649]]


              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
```

Testing Score, Confusion Matrix & Classification Report

```
0.8253275109170306


[[ 94  36]
 [ 44 284]]


              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
```

This model is a good fit.

1.6) 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.

Decision Tree Classifier

```
Out[44]: DecisionTreeClassifier()
```

Training Score, Confusion Matrix, Classification Report

```
0.9990627928772259


[[332   0]
 [  1 734]]
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	332
1	1.00	1.00	1.00	735
accuracy			1.00	1067
macro avg	1.00	1.00	1.00	1067
weighted avg	1.00	1.00	1.00	1067

Testing Score, Confusion Matrix, Classification Report

```
0.7532751091703057


[[ 81  49]
 [ 64 264]]
```

	precision	recall	f1-score	support
0	0.56	0.62	0.59	130
1	0.84	0.80	0.82	328
accuracy			0.75	458
macro avg	0.70	0.71	0.71	458
weighted avg	0.76	0.75	0.76	458

This model is overfitted.

Random Forest Model

```
Out[47]: RandomForestClassifier(random_state=1)
```

Training Score, Confusion Matrix, Classification Report

```
0.9990627928772259


[[331   1]
 [  0 735]]


              precision    recall  f1-score   support

           0       1.00      1.00      1.00       332
           1       1.00      1.00      1.00       735

    accuracy                           1.00      1067
   macro avg       1.00      1.00      1.00      1067
weighted avg       1.00      1.00      1.00      1067
```

Testing Score, Confusion Matrix, Classification Report

```
0.8187772925764192


[[ 90  40]
 [ 43 285]]


              precision    recall  f1-score   support

           0       0.68      0.69      0.68       130
           1       0.88      0.87      0.87       328

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

This model is overfitted

Bagging Classifier

```
Out[51]: BaggingClassifier(base_estimator=RandomForestClassifier(random_state=1),
                           n_estimators=100, random_state=1)
```

Training Score, Confusion Matrix, Classification Report

```
0.9653233364573571


[[304  28]
 [  9 726]]


              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
```

Testing Score, Confusion Matrix, Classification Report

```
0.8362445414847162


[[ 92  38]
 [ 37 291]]


              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
```

This model is overfitted

Adaboosting

Training Score, Confusion Matrix, Classification Report

```
0.8472352389878163
[[238  94]
 [ 69 666]]
              precision    recall  f1-score   support

           0       0.78      0.72      0.74       332
           1       0.88      0.91      0.89       735

    accuracy                           0.85      1067
   macro avg       0.83      0.81      0.82      1067
weighted avg       0.84      0.85      0.85      1067
```

Testing Score, Confusion Matrix, Classification Report

```
0.8187772925764192
[[ 90  40]
 [ 43 285]]
              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
```

This model has a good fit.

Gradient Boosting

Training Score, Confusion Matrix, Classification Report

```
0.8734770384254921


[[250  82]
 [ 53 682]]


              precision    recall  f1-score   support

           0       0.83      0.75      0.79       332
           1       0.89      0.93      0.91       735

    accuracy                           0.87      1067
   macro avg       0.86      0.84      0.85      1067
weighted avg       0.87      0.87      0.87      1067
```

Testing Score, Confusion Matrix, Classification Report

```
0.8362445414847162


[[ 97  33]
 [ 42 286]]


              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
```

This model has a good fit.

Naïve Bees with SMOTE

```
Out[64]: GaussianNB()
```

Training Score, Confusion Matrix, Classification Report

```
0.8231292517006803


[[597 138]
 [122 613]]


              precision    recall  f1-score   support

           0       0.83      0.81      0.82       735
           1       0.82      0.83      0.83       735

    accuracy                           0.82      1470
   macro avg       0.82      0.82      0.82      1470
weighted avg       0.82      0.82      0.82      1470
```

Testing Score, Confusion Matrix, Classification Report

```
0.7947598253275109


[[103  27]
 [ 67 261]]


              precision    recall  f1-score   support

           0       0.61      0.79      0.69       130
           1       0.91      0.80      0.85       328

    accuracy                           0.79       458
   macro avg       0.76      0.79      0.77       458
weighted avg       0.82      0.79      0.80       458
```

This model has a good fit, but it is less accurate.

KNN with SMOTE

```
Out[67]: KNeighborsClassifier()
```

Training Score, Confusion Matrix, Classification Report

```
0.8863945578231293


[[688  47]
 [120 615]]


              precision    recall  f1-score   support

           0       0.85      0.94      0.89       735
           1       0.93      0.84      0.88       735

    accuracy                           0.89      1470
   macro avg       0.89      0.89      0.89      1470
weighted avg       0.89      0.89      0.89      1470
```

Testing Score, Confusion Matrix, Classification Report

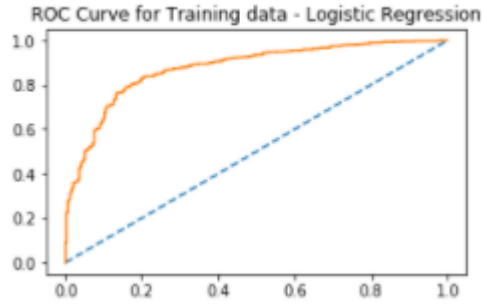
```
0.7794759825327511


[[106  24]
 [ 77 251]]


              precision    recall  f1-score   support

           0       0.58      0.82      0.68       130
           1       0.91      0.77      0.83       328

    accuracy                           0.78       458
   macro avg       0.75      0.79      0.75       458
weighted avg       0.82      0.78      0.79       458
```

This model is underfitted.

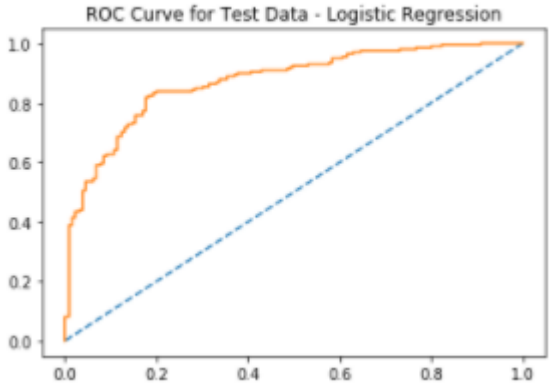
1.7) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured 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)

Logistic Regression

AUC: 0.876

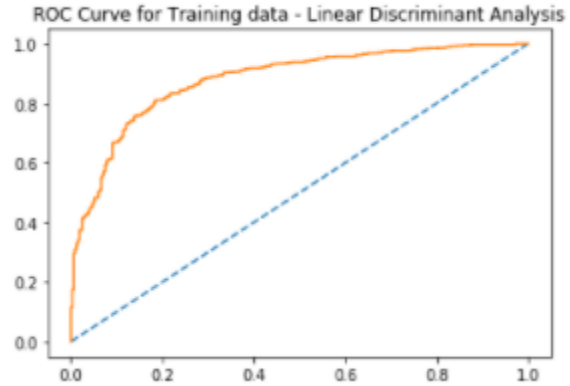


AUC: 0.874

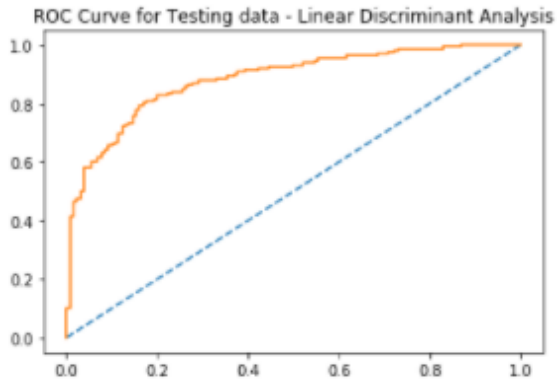


LDA

AUC: 0.879

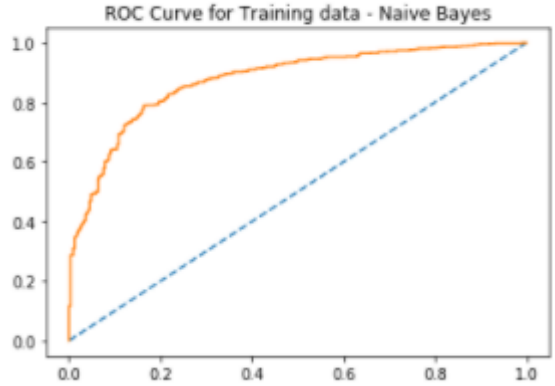


AUC: 0.883

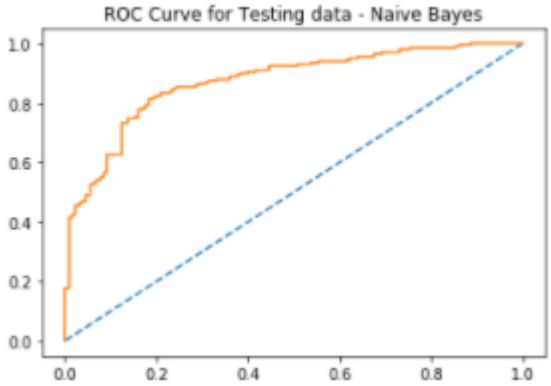


Naïve Bees

AUC: 0.875

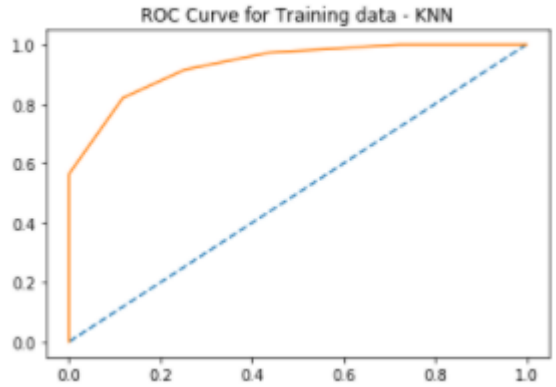


AUC: 0.873

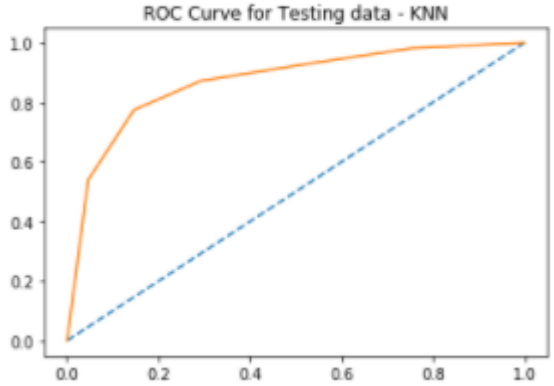


KNN

AUC: 0.932

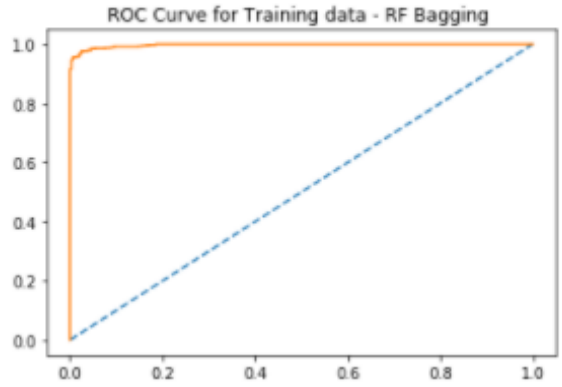


AUC: 0.871

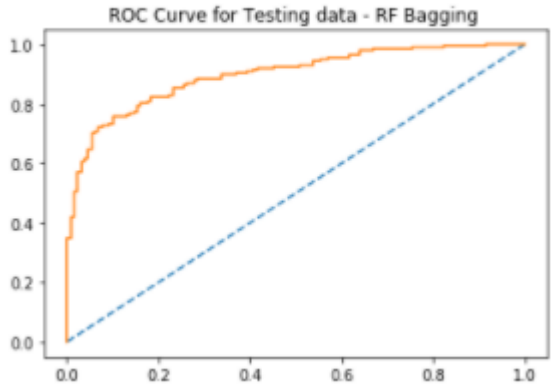


RF Bagging

AUC: 0.997



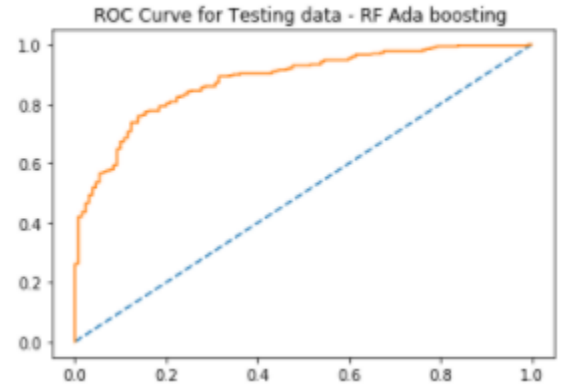
AUC: 0.898



RF Adaboosting

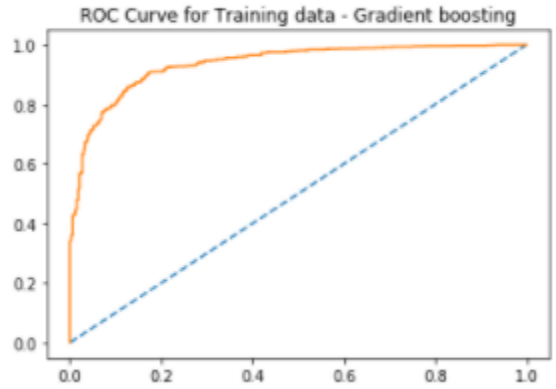
AUC: 0.913

AUC: 0.879

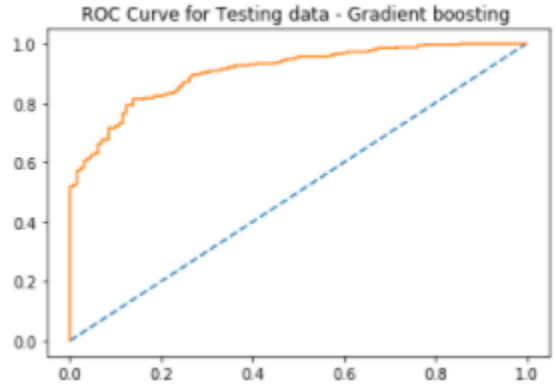


Gradient Boosting

AUC: 0.936



AUC: 0.907



1.8) 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.

- Random Forest Model has the highest accuracy
- Many models are underfitted and overfitted
- Europe & Hague seem to be negatively correlated with getting votes
- Should focus more on countries with high economic conditions

Problem 2

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

```
There were  68  sentences in President Roosevelt's speech.

There were  1536  words in President Roosevelt's speech.

There were  7571  characters in President Roosevelt's speech


There were  52  sentences in President Kennedy's speech.

There are  1546  words in President Kennedy's speech.

There are  7618  characters in President Kennedy's speech.


There were  69  sentences in President Nixon's speech.

There are  2028  words in President Nixon's speech.

There are  9991  characters in President Nixon's speech.
```

2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.

President Roosevelt's Speech

```
Before removing stopwords, the count is 1526

After removing stopwords the count is  626
Before removal:

  ['On', 'each', 'national', 'day', 'of', 'inauguration', 'since', '1789', ',', 'the', 'people', 'have', 'renewed', 'their', 's
ense', 'of', 'dedication', 'to', 'the', 'United', 'States', '.', 'In', 'Washington', "'s", 'day', 'the', 'task', 'of', 'the',
'people', 'was', 'to', 'create', 'and', 'weld', 'together', 'a', 'nation', '.', 'In', 'Lincoln', "'s", 'day', 'the', 'task', 'o
f', 'the', 'people', 'was', 'to', 'preserve', 'that', 'Nation', 'from', 'disruption', 'from', 'within', '.', 'In', 'this', 'da
y', 'the', 'task', 'of', 'the', 'people', 'is', 'to', 'save', 'that', 'Nation', 'and', 'its', 'institutions', 'from', 'disrupti
on', 'from', 'without', '.', 'To', 'us', 'there', 'has', 'come', 'a', 'time', ',', 'in', 'the', 'midst', 'of', 'swift', 'happen
ings', ',', 'to', 'pause', 'for', 'a', 'moment', 'and', 'take', 'stock', '--', 'to', 'recall', 'what', 'our', 'place', 'in', 'h


After removal:

 ['national', 'day', 'inauguration', 'since', 'people', 'renewed', 'sense', 'dedication', 'united', 'states', 'washington',
"'s", 'day', 'task', 'people', 'create', 'weld', 'together', 'nation', 'lincoln', "'s", 'day', 'task', 'people', 'preserve', 'n
ation', 'disruption', 'within', 'day', 'task', 'people', 'save', 'nation', 'institutions', 'disruption', 'without', 'us', 'com
e', 'time', 'midst', 'swift', 'happenings', 'pause', 'moment', 'take', 'stock', 'recall', 'place', 'history', 'rediscover', 'ma
y', 'risk', 'real', 'peril', 'inaction', 'lives', 'nations', 'determined', 'count', 'years', 'lifetime', 'human', 'spirit', 'li
fe', 'man', 'three-score', 'years', 'ten', 'little', 'little', 'less', 'life', 'nation', 'fullness', 'measure', 'live', 'men',
'doubt', 'men', 'believe', 'democracy', 'form', 'government', 'frame', 'life', 'limited', 'measured', 'kind', 'mystical', 'arti
```

President Kennedy's Speech

```
Before removing stopwords, the count is 1543

After removing stopwords the count is  689
Before removal:

  ['Vice', 'President', 'Johnson', ',', 'Mr.', 'Speaker', ',', 'Mr.', 'Chief', 'Justice', ',', 'President', 'Eisenhower', ',',
'Vice', 'President', 'Nixon', ',', 'President', 'Truman', ',', 'reverend', 'clergy', ',', 'fellow', 'citizens', ',', 'we', 'obs
erve', 'today', 'not', 'a', 'victory', 'of', 'party', ',', 'but', 'a', 'celebration', 'of', 'freedom', '--', 'symbolizing', 'a
n', 'end', ',', 'as', 'well', 'as', 'a', 'beginning', '--', 'signifying', 'renewal', ',', 'as', 'well', 'as', 'change', '.', 'F
or', 'I', 'have', 'sworn', 'I', 'before', 'you', 'and', 'Almighty', 'God', 'the', 'same', 'solemn', 'oath', 'our', 'forebears',
'l', 'prescribed', 'nearly', 'a', 'century', 'and', 'three', 'quarters', 'ago', '.', 'The', 'world', 'is', 'very', 'different',

After removal:

 ['vice', 'president', 'johnson', 'mr.', 'speaker', 'mr.', 'chief', 'justice', 'president', 'eisenhower', 'vice', 'president',
'nixon', 'president', 'truman', 'reverend', 'clergy', 'fellow', 'citizens', 'observe', 'today', 'victory', 'party', 'celebratio
n', 'freedom', 'symbolizing', 'end', 'well', 'beginning', 'signifying', 'renewal', 'well', 'change', 'sworn', 'almighty', 'go
d', 'solemn', 'oath', 'forebears', 'l', 'prescribed', 'nearly', 'century', 'three', 'quarters', 'ago', 'world', 'different', 'm
an', 'holds', 'mortal', 'hands', 'power', 'abolish', 'forms', 'human', 'poverty', 'forms', 'human', 'life', 'yet', 'revolutiona
ry', 'beliefs', 'forebears', 'fought', 'still', 'issue', 'around', 'globe', 'belief', 'rights', 'man', 'come', 'generosity', 's
tate', 'hand', 'god', 'dare', 'forget', 'today', 'heirs', 'first', 'revolution', 'let', 'word', 'go', 'forth', 'time', 'place',
'friend', 'foe', 'alike', 'torch', 'passed', 'new', 'generation', 'americans', 'born', 'century', 'tempered', 'war', 'disciplin
```

President Nixon's Speech

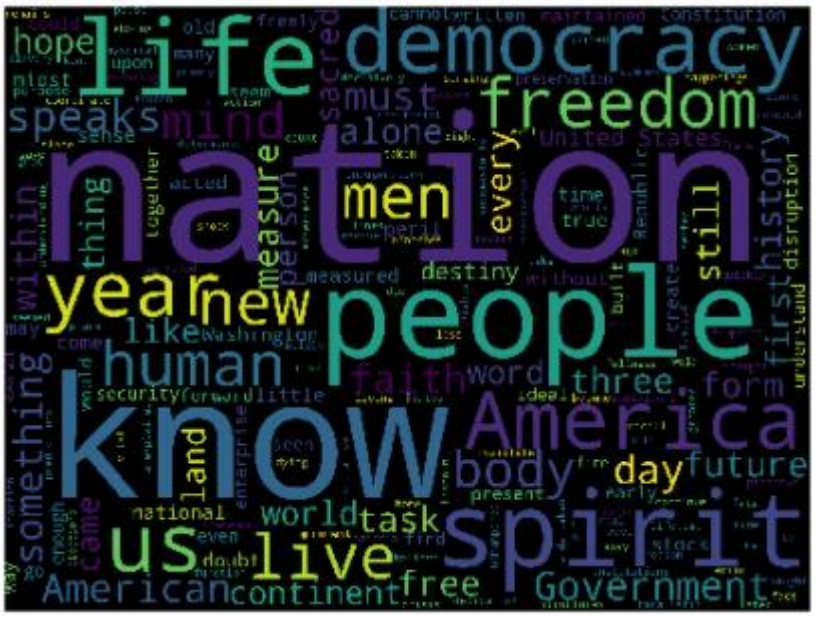
```
Before removing stopwords, the count is 2006

After removing stopwords the count is  845
Before removal:

  ['Mr.', 'Vice', 'President', ',', 'Mr.', 'Speaker', ',', 'Mr.', 'Chief', 'Justice', ',', 'Senator', 'Cook', ',', 'Mrs.', 'Eis
enhower', ',', 'and', 'my', 'fellow', 'citizens', 'of', 'this', 'great', 'and', 'good', 'country', 'we', 'share', 'together',
':', 'When', 'we', 'met', 'here', 'four', 'years', 'ago', ',', 'America', 'was', 'bleak', 'in', 'spirit', ',', 'depressed', 'b
y', 'the', 'prospect', 'of', 'seemingly', 'endless', 'war', 'abroad', 'and', 'of', 'destructive', 'conflict', 'at', 'home',
'.', 'As', 'we', 'meet', 'here', 'today', ',', 'we', 'stand', 'on', 'the', 'threshold', 'of', 'a', 'new', 'era', 'of', 'peace',
'in', 'the', 'world', '.', 'The', 'central', 'question', 'before', 'us', 'is', ':', 'How', 'shall', 'we', 'use', 'that', 'peac
e', '?', 'Let', 'us', 'resolve', 'that', 'this', 'era', 'we', 'are', 'about', 'to', 'enter', 'will', 'not', 'be', 'what', 'othe

After removal:

 ['mr.', 'vice', 'president', 'mr.', 'speaker', 'mr.', 'chief', 'justice', 'senator', 'cook', 'mrs.', 'eisenhower', 'fellow',
'citizens', 'great', 'good', 'country', 'share', 'together', 'met', 'four', 'years', 'ago', 'america', 'bleak', 'spirit', 'depr
essed', 'prospect', 'seemingly', 'endless', 'war', 'abroad', 'destructive', 'conflict', 'home', 'meet', 'today', 'stand', 'thre
shold', 'new', 'era', 'peace', 'world', 'central', 'question', 'us', 'shall', 'use', 'peace', 'let', 'us', 'resolve', 'era', 'e
nter', 'postwar', 'periods', 'often', 'time', 'retreat', 'isolation', 'leads', 'stagnation', 'home', 'invites', 'new', 'dange
r', 'abroad', 'let', 'us', 'resolve', 'become', 'time', 'great', 'responsibilities', 'greatly', 'borne', 'renew', 'spirit', 'pr
omise', 'america', 'enter', 'third', 'century', 'nation', 'past', 'year', 'saw', 'far-reaching', 'results', 'new', 'policies',
'peace', 'continuing', 'revitalize', 'traditional', 'friendships', 'missions', 'peking', 'moscow', 'able', 'establish', 'base',
```

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

```
Top 3 frequent words for President Roosevelt are below

[('nation', 12), ('know', 10), ('spirit', 9)]

Top 3 frequent words for President Kennedy are below

[('let', 16), ('us', 12), ('world', 8)]

Top 3 frequent words for President Nixon are below

[('us', 26), ('let', 22), ('america', 21)]
```

2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)

President Roosevelt's Speech



President Kennedy's Speech

President Nixon’s Speech