MACHINE LEARNING

BUSINESS REPORT

-----------------\*\*\*\*\*\*------------------

**BY:- ANIKET HIRGUDE**

***INDEX***

**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.*

*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.*

*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.*

*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)*

*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)*

*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.*

***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)***

***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.***

*PROBLEM :2*

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

***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.***

***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)***

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

***Quality of Business Report (Please refer to the Evaluation Guidelines for Business report checklist. Marks in this criteria are at the moderator's discretion)***

*Problem 1:*

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

* 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.*

*:-*

*HEAD OF THE DATA*

|  | **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 |

TABLE 1.1

THE DATA INFORMATION

<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

LET’S HAVE A LOOK AT THE DATA TYPE

vote object

age int64

economic.cond.national int64

economic.cond.household int64

Blair int64

Hague int64

Europe int64

political.knowledge int64

gender object

dtype: object

THERE ARE 2 COLUMNS IN A CATEGORICAL FORMAT (VOTE AND GENDER)

DATA SHAPE

(1525, 9)

In [10]:

DESCRIBE THE DATA

|  | **age** | **economic.cond.national** | **economic.cond.household** | **Blair** | **Hague** | **Europe** | **political.knowledge** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1525.000000 | 1525.000000 | 1525.000000 | 1525.000000 | 1525.000000 | 1525.000000 | 1525.000000 |
| **mean** | 54.182295 | 3.245902 | 3.140328 | 3.334426 | 2.746885 | 6.728525 | 1.542295 |
| **std** | 15.711209 | 0.880969 | 0.929951 | 1.174824 | 1.230703 | 3.297538 | 1.083315 |
| **min** | 24.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 0.000000 |
| **25%** | 41.000000 | 3.000000 | 3.000000 | 2.000000 | 2.000000 | 4.000000 | 0.000000 |
| **50%** | 53.000000 | 3.000000 | 3.000000 | 4.000000 | 2.000000 | 6.000000 | 2.000000 |
| **75%** | 67.000000 | 4.000000 | 4.000000 | 4.000000 | 4.000000 | 10.000000 | 2.000000 |
| **max** | 93.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 11.000000 | 3.000000 |

TABLE 1.2

CHECKING FOR THE NULL VALUES

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

WE FOUND THAT THERE ARE 8 ROWS OF duplicate NUMBERS

SO WE HAVE TO REMOVE THAT

WHILE USING DUPLICATE FUNCTION

NOW THE OUTPUT IS

Number of duplicate rows = 0

SKEWNESS OF THE DATA

age 0.134944

economic.cond.national -0.236477

economic.cond.household -0.138688

Blair -0.543655

Hague 0.140236

Europe -0.147907

political.knowledge -0.418982

dtype: float64

WERE WE KNOWN THERE ARE 2 COLUMNS THAT HAVE NEGATIVE VALUES AND THE REST OF POSITIVE

# 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.

:-

EDA

NULL VALUES

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

vote object

age int64

economic.cond.national int64

economic.cond.household int64

Blair int64

Hague int64

Europe int64

political.knowledge int64

gender object

dtype: object

SHAPE OF THE DATA AFTER REMOVING DUPLICATES VALUES

(1509, 9)

INDIVIDUAL DETAILS OF CATEGORICAL DATA

VOTE : 2

Conservative 458

Labour 1051

Name: vote, dtype: int64

GENDER : 2

male 705

female 804

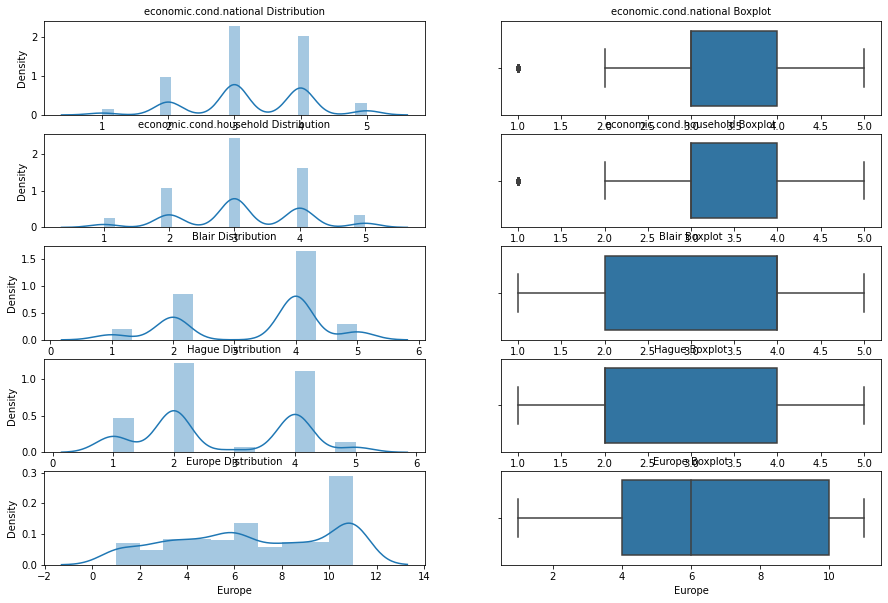
Name: gender, dtype: int64

CHECKING FOR CATEGORICAL AND NUMERICAL COLUMNS

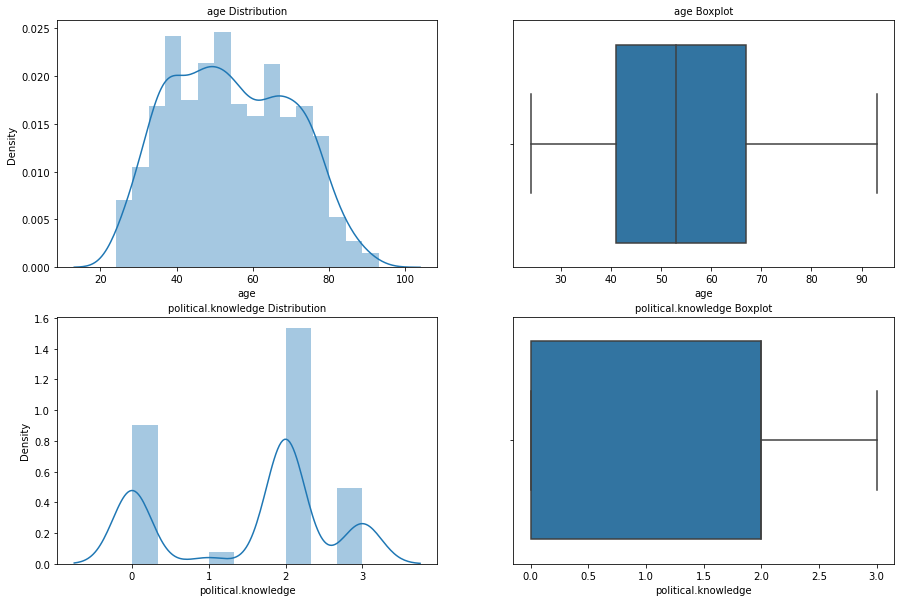
['vote', 'gender'] OBJECT TYPE

['age', 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political.knowledge'] INT TYPE

UNIVARIATE

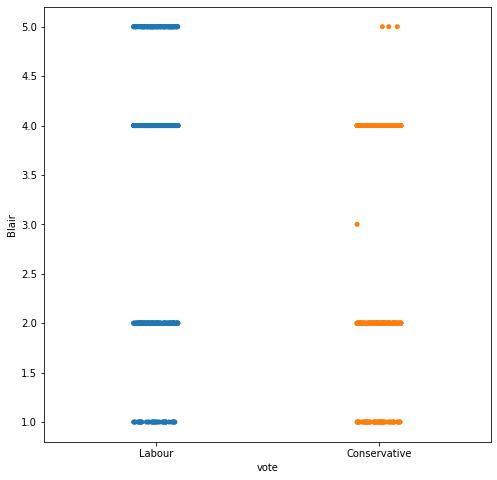


GRAPH 1.1



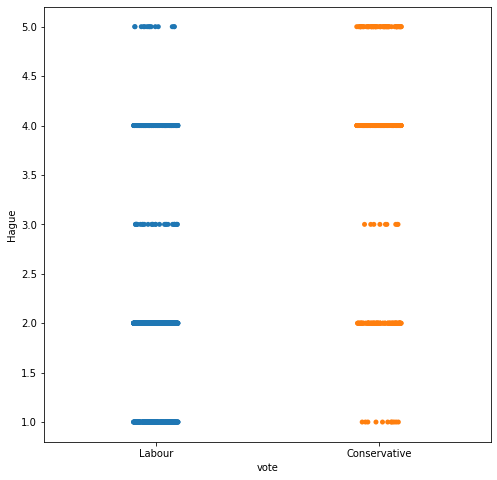
GRAPH 1.2

BIVARIATE



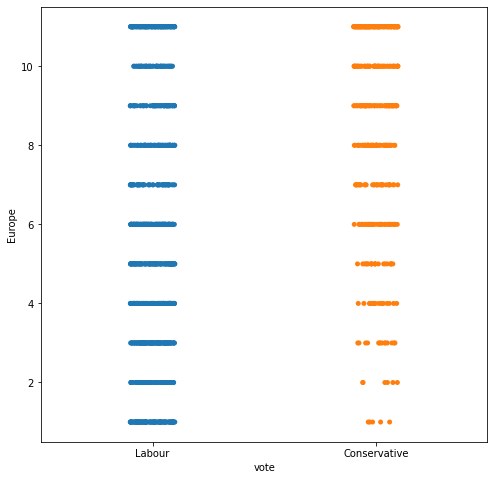
GRAPH 1.3

in Blair, labour vote will be more than conservative



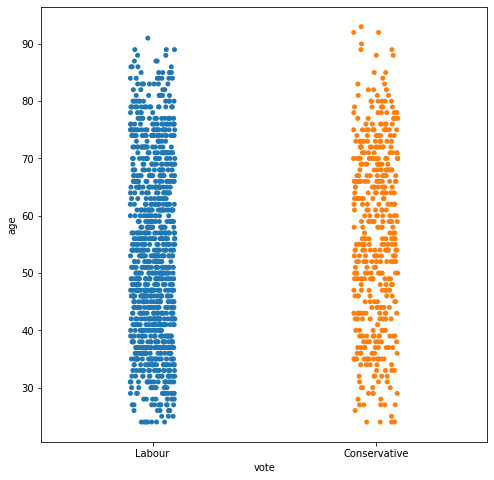
GRAPH 1.4

in Hague the labour and conservative voters will give the proper fight



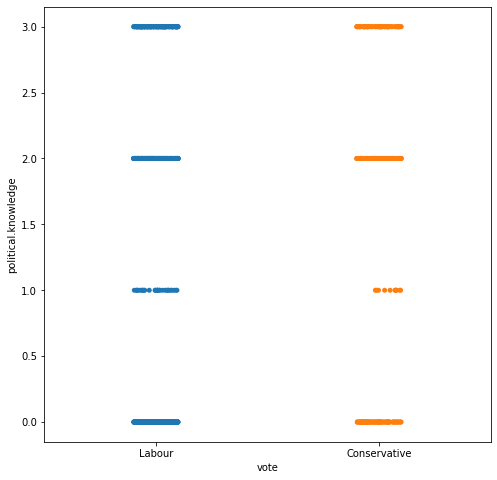
GRAPH 1.5

the labour parties have more chances to win in Europe



GRAPH 1.6

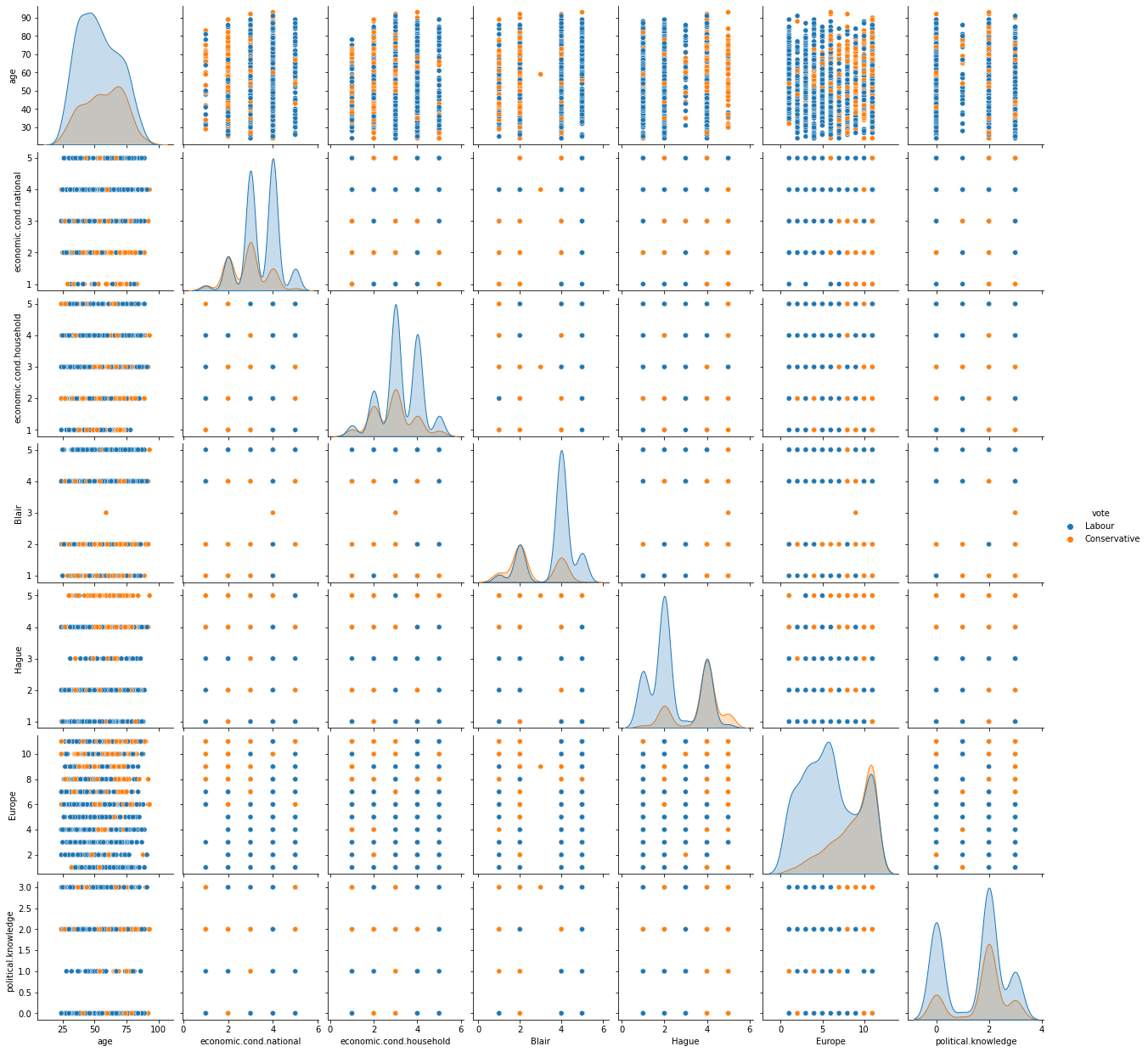
labour have more voters in the range of 30 to 75 age



GRAPH 1.7

somewhere we r looking that conservatives have low political knowledge

LET’S PLOT A PAIR PLOT FOR DEEP UNDERSTAND THE DATA



GRAPH 1.8

CHECKING FOR THE CORRELATION

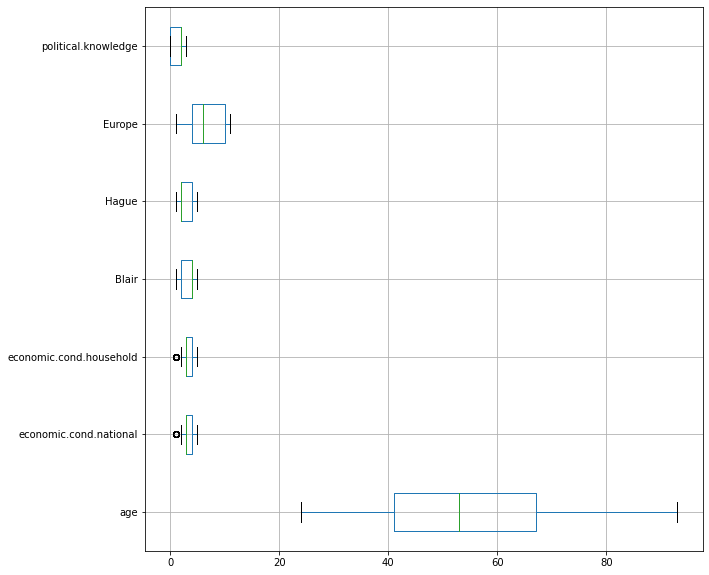


GRAPH 1.9

In the above plot pink colour represents maximum correlation and purple colour represents minimum correlation.

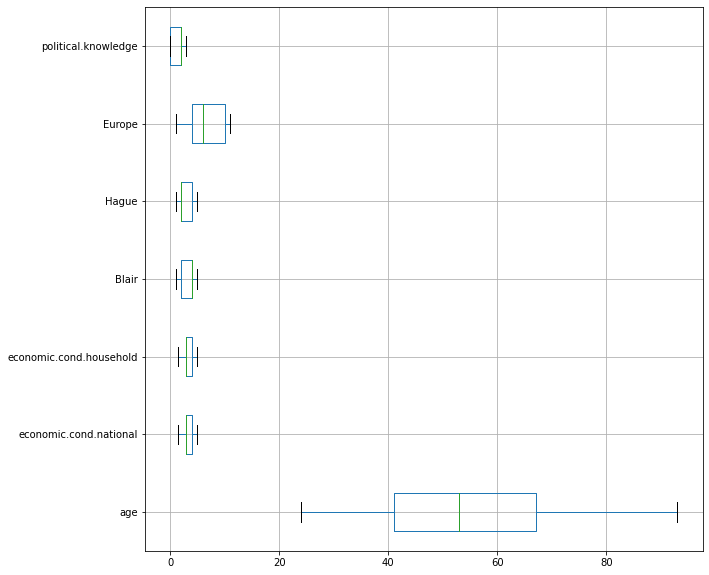
We can see none of the variables has a correlation with any other variables.

CHECKING FOR THE OUTLIERS



GRAPH 1.10

THERE IS SOME OUTLIERS IN THE ECONOMIC.COND.HOUSEHOLD AND ECONOMIC.COND.NATIONAL COLUMNS

OUTLIERS TREATMENT

GRAPH 1.11

AFTER THE OUTLIERS TREATMENT

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.

:-

VARIANCE OF THE DATA

age 246.237009

economic.cond.national 0.729936

economic.cond.household 0.787257

Blair 1.379962

Hague 1.523412

Europe 10.893439

political.knowledge 1.178370

dtype: float64

Converting categorical into dummies

|  | **age** | **economic.cond.national** | **economic.cond.household** | **Blair** | **Hague** | **Europe** | **political.knowledge** | **vote\_Labour** | **gender\_male** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 43 | 3.0 | 3.0 | 4 | 1 | 2 | 2 | 1 | 0 |
| **1** | 36 | 4.0 | 4.0 | 4 | 4 | 5 | 2 | 1 | 1 |
| **3** | 24 | 4.0 | 2.0 | 2 | 1 | 4 | 0 | 1 | 0 |
| **4** | 41 | 2.0 | 2.0 | 1 | 1 | 6 | 2 | 1 | 1 |
| **5** | 47 | 3.0 | 4.0 | 4 | 4 | 4 | 2 | 1 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **98** | 74 | 3.0 | 4.0 | 2 | 4 | 1 | 0 | 1 | 1 |
| **99** | 67 | 3.0 | 4.0 | 5 | 1 | 11 | 0 | 1 | 1 |
| **100** | 40 | 4.0 | 4.0 | 4 | 2 | 6 | 0 | 1 | 1 |
| **101** | 71 | 3.0 | 4.0 | 4 | 2 | 3 | 2 | 1 | 1 |
| **102** | 49 | 3.0 | 3.0 | 4 | 2 | 6 | 2 | 1 | 0 |

TABLE 1.3

3

# SPLITTING the data into train and test (70:30)

#Checking the number of rows and columns in Train & Test dataframe

Shape of X\_train is : (1056, 8)

Shape of y\_train is : (1056,)

Shape of X\_test is : (453, 8)

Shape of y\_test is : (453,)

CHECKING THE PERCENTAGE SPLITED DATA

69.98% data is in training set

30.02% data is in test set

X HEAD

|  | **age** | **economic.cond.national** | **economic.cond.household** | **Blair** | **Hague** | **Europe** | **political.knowledge** | **gender\_male** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 43 | 3.0 | 3.0 | 4 | 1 | 2 | 2 | 0 |
| **1** | 36 | 4.0 | 4.0 | 4 | 4 | 5 | 2 | 1 |
| **3** | 24 | 4.0 | 2.0 | 2 | 1 | 4 | 0 | 0 |
| **4** | 41 | 2.0 | 2.0 | 1 | 1 | 6 | 2 | 1 |
| **5** | 47 | 3.0 | 4.0 | 4 | 4 | 4 | 2 | 1 |

TABLE 1.4

Y HEAD

0 1

1 1

3 1

4 1

5 1

Name: vote\_Labour, dtype: uint8

Modelling

Accuracy Score for K=3 is 0.7748344370860927

Accuracy Score for K=5 is 0.7726269315673289

Accuracy Score for K=9 is 0.8101545253863135

SCALING THE DATA

['age', 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political.knowledge']

Check if the variables have been scaled or not

|  | **age** | **economic.cond.national** | **economic.cond.household** | **Blair** | **Hague** | **Europe** | **political.knowledge** | **gender\_male** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.275362 | 0.428571 | 0.428571 | 0.75 | 0.00 | 0.1 | 2 | 0 |
| **1** | 0.173913 | 0.714286 | 0.714286 | 0.75 | 0.75 | 0.4 | 2 | 1 |
| **3** | 0.000000 | 0.714286 | 0.142857 | 0.25 | 0.00 | 0.3 | 0 | 0 |
| **4** | 0.246377 | 0.142857 | 0.142857 | 0.00 | 0.00 | 0.5 | 2 | 1 |
| **5** | 0.333333 | 0.428571 | 0.714286 | 0.75 | 0.75 | 0.3 | 2 | 1 |

TABLE 1.5

**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)**

:-

APPLYING Logistic Regression

1)MODEL OF TRAIN

0 1 2 3 4 5 6 \

0 -0.009644 0.452169 0.148236 0.722248 -0.82089 -0.169225 -0.368465

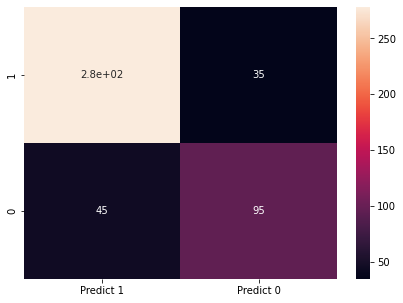
7 intercept

0 0.072629 1.419151

MODEL SCORE

0.8323863636363636

CONFUSION MATRIX



GRAPH 1.12

The confusion matrix

True Positives (TP): we correctly predicted that they do have elections 282

True Negatives (TN): we correctly predicted that they don't have elections 95

False Positives (FP): we incorrectly predicted that they do have elections (a "Type I error") 45 Falsely predict positive Type I error

False Negatives (FN): we incorrectly predicted that they don't have elections (a "Type II error") 35 Falsely predict negative Type II error

2) MODEL OF TEST

0 1 2 3 4 5 6 \

0 -0.021055 0.72097 0.16383 0.415552 -0.717147 -0.267546 -0.357888

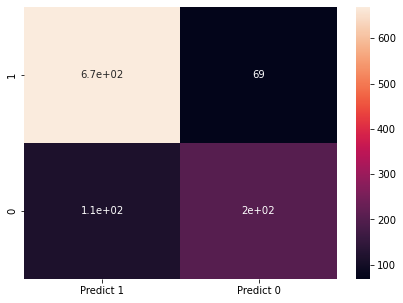
7 intercept

0 0.210146 2.574943

MODEL SCORE

0.8498896247240618

CONFUSION MATRIX



GRAPH 1.13

The confusion matrix

True Positives (TP): we correctly predicted that they do have elections 670

True Negatives (TN): we correctly predicted that they don't have elections 22

False Positives (FP): we incorrectly predicted that they do have elections (a "Type I error") 112 Falsely predict positive Type I error

False Negatives (FN): we incorrectly predicted that they don't have elections (a "Type II error") 69 Falsely predict negative Type II error

LDA Model:

# Check Correlation values

|  | **age** | **economiccondnational** | **economiccondhousehold** | **Blair** | **Hague** | **Europe** | **politicalknowledge** | **gendermale** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **age** | 1.000000 | 0.022378 | -0.041611 | 0.033965 | 0.027614 | 0.060172 | -0.044676 | -0.016097 |
| **economiccondnational** | 0.022378 | 1.000000 | 0.345991 | 0.322798 | -0.200264 | -0.206348 | -0.029153 | 0.055926 |
| **economiccondhousehold** | -0.041611 | 0.345991 | 1.000000 | 0.217194 | -0.098025 | -0.110138 | -0.041253 | 0.032208 |
| **Blair** | 0.033965 | 0.322798 | 0.217194 | 1.000000 | -0.243815 | -0.295734 | -0.021682 | 0.065222 |
| **Hague** | 0.027614 | -0.200264 | -0.098025 | -0.243815 | 1.000000 | 0.284105 | -0.029449 | -0.027201 |
| **Europe** | 0.060172 | -0.206348 | -0.110138 | -0.295734 | 0.284105 | 1.000000 | -0.150013 | -0.074583 |
| **politicalknowledge** | -0.044676 | -0.029153 | -0.041253 | -0.021682 | -0.029449 | -0.150013 | 1.000000 | 0.155655 |
| **gendermale** | -0.016097 | 0.055926 | 0.032208 | 0.065222 | -0.027201 | -0.074583 | 0.155655 | 1.000000 |

TABLE 1.6

CONFUSION MATRIX FOR Y, PRIDIC CLASS

array([[315, 143],

[105, 946]], dtype=int64)

CLASSIFICATION REPORT

precision recall f1-score support

0 0.75 0.69 0.72 458

1 0.87 0.90 0.88 1051

accuracy 0.84 1509

macro avg 0.81 0.79 0.80 1509

weighted avg 0.83 0.84 0.83 1509

**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)**

**:-**

Run the KNN with no of neighbours to be 1,3,5..19 and \*Find the optimal number of neighbours from K=1,3,5,7....19 using the Mis classification error

Hint: Misclassification error (MCE) = 1 - Test accuracy score. Calculated MCE for each model with neighbours = 1,3,5...19 and find the model with lowest MCE

# empty list that will hold accuracy scores

# perform accuracy metrics for values from 1,3,5....15

# changing to misclassification error

[0.26048565121412803,

0.22516556291390732,

0.22737306843267113,

0.20309050772626935,

0.1898454746136865,

0.19646799116997793,

0.1876379690949227,

0.1876379690949227,

0.17880794701986757,

0.17880794701986757]

KNN

Performance Matrix on train data set

0.8522727272727273

[[223 95]

[ 61 677]]

precision recall f1-score support

0 0.79 0.70 0.74 318

1 0.88 0.92 0.90 738

accuracy 0.85 1056

macro avg 0.83 0.81 0.82 1056

weighted avg 0.85 0.85 0.85 1056

Performance Matrix on test data set

0.7726269315673289

[[ 82 58]

[ 45 268]]

precision recall f1-score support

0 0.65 0.59 0.61 140

1 0.82 0.86 0.84 313

accuracy 0.77 453

macro avg 0.73 0.72 0.73 453

weighted avg 0.77 0.77 0.77 453

Naive\_bayes FOR GaussianNB

GaussianNB()

precision recall f1-score support

0 0.73 0.74 0.73 140

1 0.88 0.88 0.88 313

accuracy 0.83 453

macro avg 0.81 0.81 0.81 453

weighted avg 0.83 0.83 0.83 453

[[103 37]

[ 38 275]]

Precision: Within a given set of positively-labeled results, the fraction that were true positives = tp/(tp + fp) Recall: Given a set of positively-labeled results, the fraction of all positives that were retrieved = tp/(tp + fn) Accuracy: tp + tn / (tp + tn + fp +fn) But this measure can be dominated by larger class. Suppose 10, 90 and 80 of 90 is correctly predicted while only 2 of 0 is predicted correctly. Accuracy is 80+2 / 100 i.e. 82%

TO over come the dominance of the majority class, use weighted measure (not shown)

F is harmonic mean of precision and recal given by ((B^2 +1) PR) / (B^2P +R) When B is set to 1 we get F1 = 2PR / (P+R)

**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.**

**:-**

MODEL TUNING –

Decision Tree Classifier FIT THE DATA

DecisionTreeClassifier(random\_state=1)

TREE SCORE FOR TRAIN AND TEST

1.0

0.7505518763796909

Performance Matrix on a train data set

1.0

[[318 0]

[ 0 738]]

precision recall f1-score support

0 1.00 1.00 1.00 318

1 1.00 1.00 1.00 738

accuracy 1.00 1056

macro avg 1.00 1.00 1.00 1056

weighted avg 1.00 1.00 1.00 1056

Performance Matrix on a test data set

0.7505518763796909

[[ 98 42]

[ 71 242]]

precision recall f1-score support

0 0.58 0.70 0.63 140

1 0.85 0.77 0.81 313

accuracy 0.75 453

macro avg 0.72 0.74 0.72 453

weighted avg 0.77 0.75 0.76 453

Reducing over fitting (Regularization)

0.8153409090909091

0.7660044150110376

Importance of features in the tree building ( The importance of a feature is computed as the

(normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance )

Imp

age 0.000000

economic.cond.national 0.009259

economic.cond.household 0.000000

Blair 0.321778

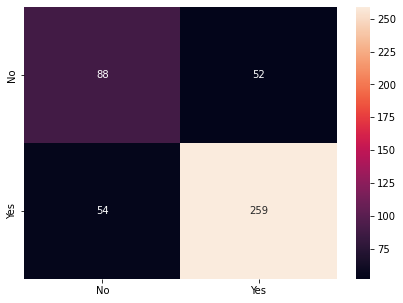
Hague 0.581238

Europe 0.046642

political.knowledge 0.041082

gender\_male 0.000000

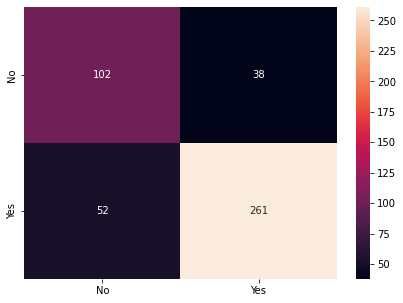
CONFUSION MATRIX GRAPH FOR ANAYLIS



GRAPH 1.14

BAGGING -

0.8013245033112583



GRAPH 1.15

BOSTING –

0.8233995584988962



GRAPH 1.16

**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)**

**:-**

|  | **age** | **economiccondnational** | **economiccondhousehold** | **Blair** | **Hague** | **Europe** | **politicalknowledge** | **gendermale** | **Prediction** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.275362 | 0.428571 | 0.428571 | 0.75 | 0.00 | 0.1 | 2 | 0 | 1 |
| **1** | 0.173913 | 0.714286 | 0.714286 | 0.75 | 0.75 | 0.4 | 2 | 1 | 1 |
| **3** | 0.000000 | 0.714286 | 0.142857 | 0.25 | 0.00 | 0.3 | 0 | 0 | 1 |
| **4** | 0.246377 | 0.142857 | 0.142857 | 0.00 | 0.00 | 0.5 | 2 | 1 | 1 |
| **5** | 0.333333 | 0.428571 | 0.714286 | 0.75 | 0.75 | 0.3 | 2 | 1 | 1 |

TABLE 1.6

Getting the probability Score

|  | **0** | **1** |
| --- | --- | --- |
| **0** | 0.195748 | 0.804252 |
| **1** | 0.044562 | 0.955438 |
| **2** | 0.081612 | 0.918388 |
| **3** | 0.976065 | 0.023935 |
| **4** | 0.028299 | 0.971701 |

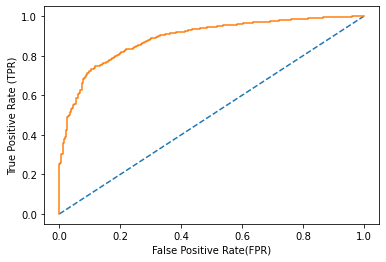
TABLE 1.7

Score OF THE MODEL

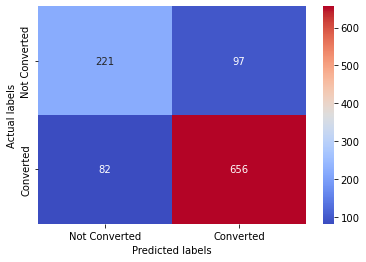
ROC\_AUC FOR TRAIN DATA-

AUC SCORE - AUC: 0.889

ROC CURVE



GRAPH 1.17



GRAPH 1.18

Accuracy - Test Data

0.8304924242424242

CLASSIFICATION REPORT

precision recall f1-score support

0 0.73 0.69 0.71 318

1 0.87 0.89 0.88 738

accuracy 0.83 1056

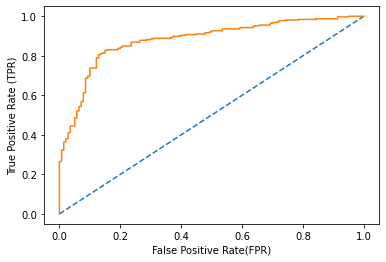
macro avg 0.80 0.79 0.80 1056

weighted avg 0.83 0.83 0.83 1056

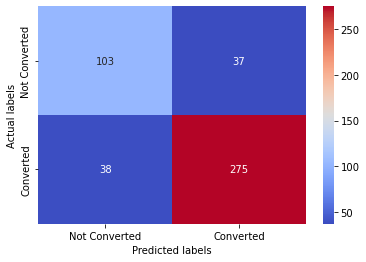
ROC\_AUC FOR TEST DATA-

SCORE - AUC: 0.889

ROC\_AUC COURVE



GRAPH 1.19



GRAPH 1.20

Accuracy - Test Data

0.8344370860927153

CLASSIFICATION REPORT

precision recall f1-score support

0 0.73 0.74 0.73 140

1 0.88 0.88 0.88 313

accuracy 0.83 453

macro avg 0.81 0.81 0.81 453

weighted avg 0.83 0.83 0.83 453

Accuracy - Test Data

\*\*\*\*\*

END

\*\*\*\*\*

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

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

:-

EDA

## Reading the first 20 words in Roosevelt’s speech

['On', 'each', 'national', 'day', 'of', 'inauguration', 'since', '1789', ',', 'the', 'people', 'have', 'renewed', 'their', 'sense', 'of', 'dedication', 'to', 'the', 'United']

In [83]:

## Reading the first 20 words in Kennedy’s speech

['Vice', 'President', 'Johnson', ',', 'Mr', '.', 'Speaker', ',', 'Mr', '.', 'Chief', 'Justice', ',', 'President', 'Eisenhower', ',', 'Vice', 'President', 'Nixon', ',']

In [84]:

## Reading the first 20 words in Nixon’s speech

['Mr', '.', 'Vice', 'President', ',', 'Mr', '.', 'Speaker', ',', 'Mr', '.', 'Chief', 'Justice', ',', 'Senator', 'Cook', ',', 'Mrs', '.', 'Eisenhower']

In [85]:

## Reading the First 5 sentences in Roosevelt’s speech

[['On', 'each', 'national', 'day', 'of', 'inauguration', 'since', '1789', ',', 'the', 'people', 'have', 'renewed', 'their', 'sense', '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', 'of', 'the', 'people', 'was', 'to', 'preserve', 'that', 'Nation', 'from', 'disruption', 'from', 'within', '.'], ['In', 'this', 'day', 'the', 'task', 'of', 'the', 'people', 'is', 'to', 'save', 'that', 'Nation', 'and', 'its', 'institutions', 'from', 'disruption', 'from', 'without', '.'], ['To', 'us', 'there', 'has', 'come', 'a', 'time', ',', 'in', 'the', 'midst', 'of', 'swift', 'happenings', ',', 'to', 'pause', 'for', 'a', 'moment', 'and', 'take', 'stock', '--', 'to', 'recall', 'what', 'our', 'place', 'in', 'history', 'has', 'been', ',', 'and', 'to', 'rediscover', 'what', 'we', 'are', 'and', 'what', 'we', 'may', 'be', '.']]

In [86]:

## Reading the First 5 sentences in speech

[['Vice', 'President', 'Johnson', ',', 'Mr', '.', 'Speaker', ',', 'Mr', '.', 'Chief', 'Justice', ',', 'President', 'Eisenhower', ',', 'Vice', 'President', 'Nixon', ',', 'President', 'Truman', ',', 'reverend', 'clergy', ',', 'fellow', 'citizens', ',', 'we', 'observe', 'today', 'not', 'a', 'victory', 'of', 'party', ',', 'but', 'a', 'celebration', 'of', 'freedom', '--', 'symbolizing', 'an', 'end', ',', 'as', 'well', 'as', 'a', 'beginning', '--', 'signifying', 'renewal', ',', 'as', 'well', 'as', 'change', '.'], ['For', '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', 'now', '.'], ['For', 'man', 'holds', 'in', 'his', 'mortal', 'hands', 'the', 'power', 'to', 'abolish', 'all', 'forms', 'of', 'human', 'poverty', 'and', 'all', 'forms', 'of', 'human', 'life', '.'], ['And', 'yet', 'the', 'same', 'revolutionary', 'beliefs', 'for', 'which', 'our', 'forebears', 'fought', 'are', 'still', 'at', 'issue', 'around', 'the', 'globe', '--', 'the', 'belief', 'that', 'the', 'rights', 'of', 'man', 'come', 'not', 'from', 'the', 'generosity', 'of', 'the', 'state', ',', 'but', 'from', 'the', 'hand', 'of', 'God', '.']]

In [87]:

## Reading the First 5 sentences in Nixon’s speech

[['Mr', '.', 'Vice', 'President', ',', 'Mr', '.', 'Speaker', ',', 'Mr', '.', 'Chief', 'Justice', ',', 'Senator', 'Cook', ',', 'Mrs', '.', 'Eisenhower', ',', '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', 'by', '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', 'peace', '?'], ['Let', 'us', 'resolve', 'that', 'this', 'era', 'we', 'are', 'about', 'to', 'enter', 'will', 'not', 'be', 'what', 'other', 'postwar', 'periods', 'have', 'so', 'often', 'been', ':', 'a', 'time', 'of', 'retreat', 'and', 'isolation', 'that', 'leads', 'to', 'stagnation', 'at', 'home', 'and', 'invites', 'new', 'danger', 'abroad', '.']]

In [88]:

## Reading the last line in Roosevelt’s speech

['As', 'Americans', ',', 'we', 'go', 'forward', ',', 'in', 'the', 'service', 'of', 'our', 'country', ',', 'by', 'the', 'will', 'of', 'God', '.']

## Reading the last line in Kennedy’s speech

['With', 'a', 'good', 'conscience', 'our', 'only', 'sure', 'reward', ',', 'with', 'history', 'the', 'final', 'judge', 'of', 'our', 'deeds', ',', 'let', 'us', 'go', 'forth', 'to', 'lead', 'the', 'land', 'we', 'love', ',', 'asking', 'His', 'blessing', 'and', 'His', 'help', ',', 'but', 'knowing', 'that', 'here', 'on', 'earth', 'God', "'", 's', 'work', 'must', 'truly', 'be', 'our', 'own', '.']

## Reading the last line in Nixon’s speech

['Let', 'us', 'go', 'forward', 'from', 'here', 'confident', 'in', 'hope', ',', 'strong', 'in', 'our', 'faith', 'in', 'one', 'another', ',', 'sustained', 'by', 'our', 'faith', 'in', 'God', 'who', 'created', 'us', ',', 'and', 'striving', 'always', 'to', 'serve', 'His', 'purpose', '.']

# Information about the speech's

Roosevelt:

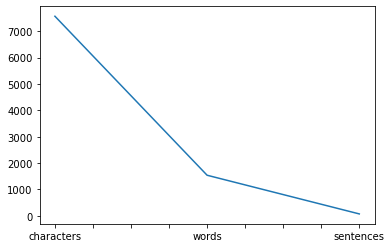
The number of characters in Roosevelt’s 1941 speech is: 7571

The number of words in Roosevelt’s 1941 speech is: 1536

The number of sentences in Roosevelt’s 1941 speech is: 68

|  | **Roosevelt\_speech** |
| --- | --- |
| **characters** | 7571 |
| **words** | 1536 |
| **sentences** | 68 |

TABLE 1



graph 1

Kennedy:

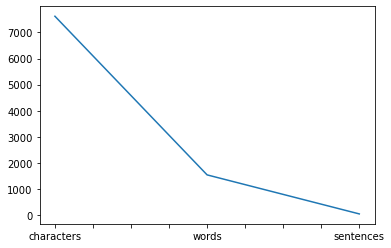
The numbers of characters in Kennedy 1961 speech is: 7618

The numbers of words in Kennedy 1961 speech is: 1546

The numbers of sentences in Kennedy 1961 speech is: 52

| **Kennedy\_speech** |
| --- |
| **characters** | 7618 |
| **words** | 1546 |
| **sentences** | 52 |

Table 2



Graph 2

Nixon:

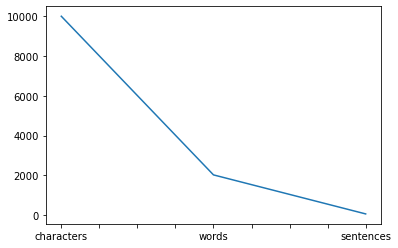
The numbers of characters in Nixon 1973 speech is: 9991

The numbers of words in Nixon 1973 speech is: 2028

The numbers of sentences in Nixon 1973 speech is: 69

|  | **Nixon\_speech** |
| --- | --- |
| **characters** | 9991 |
| **words** | 2028 |
| **sentences** | 69 |

Table 3



Graph 3

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.

:-

WE DID THIS, WITH ALL 3 DATA:

1. Converting to Lower Case
2. Cleaning the special characters (/@$: etc, from Regular Expressions package, re.sub() function for string substitution using regular expressions
3. Tokenization - Splitting the text files into words
4. Stop words - Removing words that are not useful or Removing meaningless words

Roosevelt’s speech last sentenced after removing stop-word

['As', 'Americans', ',', 'we', 'go', 'forward', ',', 'in', 'the', 'service', 'of', 'our', 'country', ',', 'by', 'the', 'will', 'of', 'God', '.']

In [104]:

Kennedy’s speech last sentenced after removing stop-word

['With', 'a', 'good', 'conscience', 'our', 'only', 'sure', 'reward', ',', 'with', 'history', 'the', 'final', 'judge', 'of', 'our', 'deeds', ',', 'let', 'us', 'go', 'forth', 'to', 'lead', 'the', 'land', 'we', 'love', ',', 'asking', 'His', 'blessing', 'and', 'His', 'help', ',', 'but', 'knowing', 'that', 'here', 'on', 'earth', 'God', "'", 's', 'work', 'must', 'truly', 'be', 'our', 'own', '.']

Nixon’s speech was last sentenced after removing stop-word

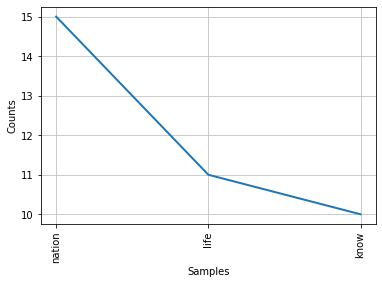
['Let', 'us', 'go', 'forward', 'from', 'here', 'confident', 'in', 'hope', ',', 'strong', 'in', 'our', 'faith', 'in', 'one', 'another', ',', 'sustained', 'by', 'our', 'faith', 'in', 'God', 'who', 'created', 'us', ',', 'and', 'striving', 'always', 'to', 'serve', 'His', 'purpose', '.']

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)

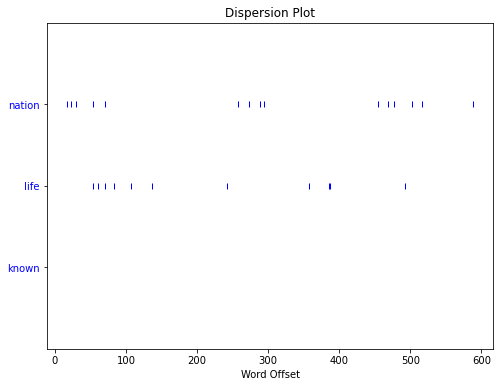
:-

Top 3 frequency occurring words of Roosevelt’s speech

[('nation', 15), ('life', 11), ('know', 10)]



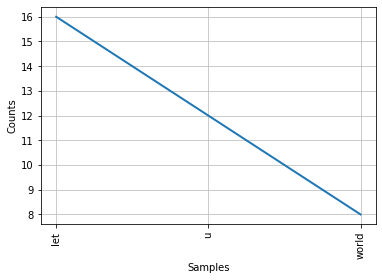
Graph 4



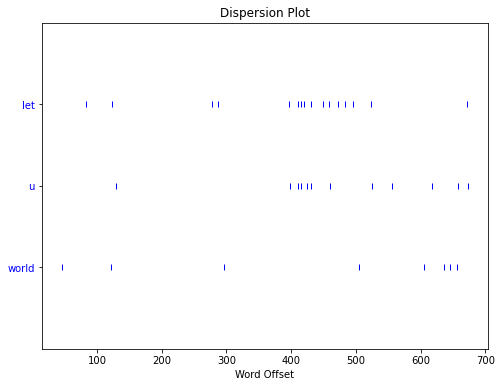
Graph 5

Top 3 frequency occurring words of Kennedy’s speech

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



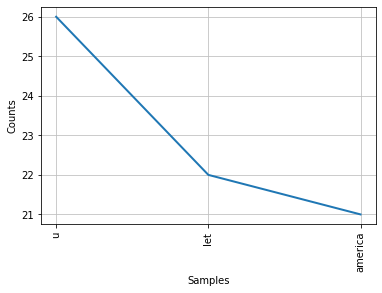
Graph 6



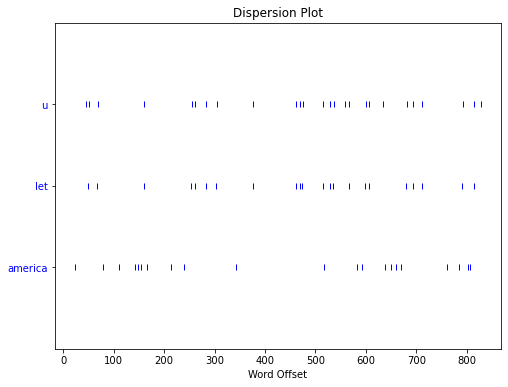
Graph 7

Top 3 frequently occurring words of Nixon’s speech

[('u', 26), ('let', 22), ('america', 21)]



graph 8

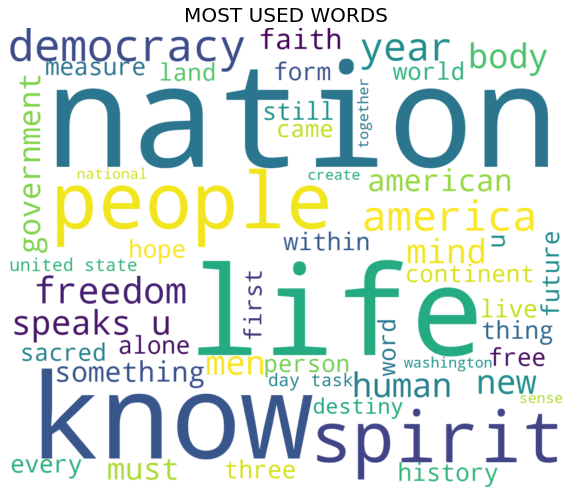


Graph 9

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

:-

Roosevelt’s speech:



Word cloud 1

Kennedy’s speech:



Word cloud 2

Nixon’s speech:



Word cloud 3

END\*\*\*