**Machine Learning**

**Project**

**-R.PRAKASH**

|  |  |  |
| --- | --- | --- |
| **Problem Number** | **Problem Statement** | **Page No** |
| 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. | 6 |
| 1.1 | Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it | 6 |
| 1.2 | Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers | 7 |
| 1.3 | Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). | 18 |
| 1.4 | Apply Logistic Regression and LDA (linear discriminant analysis). | 22 |
| 1.5 | Apply KNN Model and Naïve Bayes Model. Interpret the results. | 24 |
| 1.6 | Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. | 27 |
| 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. Final Model: Compare the models and write inference which model is best/optimized. | 30 |
| 1.8 | Based on these predictions, what are the insights? (5 marks) | 30 |
| 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 | 34 |
| 2.1 | Find the number of characters, words, and sentences for the mentioned documents. – 3 Marks | 34 |
| 2.2 | Remove all the stopwords from all three speeches. | 36 |
| 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) | 36 |
| 2.4 | Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks | 37 |

**List of Figures:**

|  |  |
| --- | --- |
| **Name of figure** | **Pg.no** |
| Fig 1.1-Datatype | 6 |
| Fig-1.2- Null values | 7 |
| Fig 1.3- Frequency distribution | 8 |
| Fig 1.4-Count plot | 8 |
| Fig 1.5- Boxplot-Age | 8 |
| Fig 1.6- Frequency distribution-national condition | 9 |
| Fig 1.7-Count plot- national condition | 9 |
| Fig 1.8-Count plot-national condition vs vote | 10 |
| Fig 1.9- Frequency distribution-Economic condition | 10 |
| Fig 1.10-Count plot-economic condition | 11 |
| Fig 1.11-Count plot-economic condition vs vote | 11 |
| Fig 1.12- Frequency distribution-Blair | 11 |
| Fig 1.13-Count plot-Blair | 12 |
| Fig 1.14-Count plot-Blair vs Vote | 12 |
| Fig 1.15- Frequency distribution-Hague | 13 |
| Fig 1.16-Count plot-Hague | 13 |
| Fig 1.17-Count plot-Hague vs Vote | 13 |
| Fig 1.18- Frequency distribution-Europe | 14 |
| Fig 1.19-Count plot-Hague | 14 |
| Fig 1.20-Count plot-Hague vs Vote | 14 |
| Fig 1.21- Frequency distribution-Political | 15 |
| Fig 1.22-Count plot-Political | 16 |
| Fig 1.23-Count plot-Political vs Vote | 16 |
| Fig 1.24- Frequency distribution-Gender | 16 |
| Fig 1.25-Count plot-Gender | 17 |
| Fig 1.26-Count plot- Gender vs Vote | 17 |
| Fig 1.27-Correlation Matrix | 17 |
| Fig 1.28-Encoding | 17 |
| Fig 2.1-Word cloud roosevolt | 37 |
| Fig 2.2-Word cloud kennedy | 38 |
| Fig 2.3-Word cloud Nixon | 39 |

|  |  |
| --- | --- |
| **List of Tables** | **Page no** |
| Table-1.1-First five rows of the datase | 6 |
| Table-1.2-Description | 7 |
| Table-1.3-Before Scaling | 19 |
| Table-1.4-After Scaling | 19 |
| Table-1.5-Independent Variables | 20 |
| Table-1.6-Logistic regression | 22 |
| Table-1.7-LDA | 23 |
| Table-1.8-Naïve bayes | 24 |
| Table-1.9-KNN | 24 |
| Table-1.10-Adaboost | 27 |
| Table-1.11-Gradient boost | 28 |
| Table-1.12-Random forest(Bagging)30 | 30 |
| Table-1.13-Final model | 30 |
| Table-2.1-Data frame | 34 |
| Table-2.2-Characters | 34 |
| Table-2.4-Words | 35 |
| Table-2.5-Count of Words-Before stop words removal | 36 |
| Table-2.6-Count of Words-After stop words removal | 36 |

**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.**

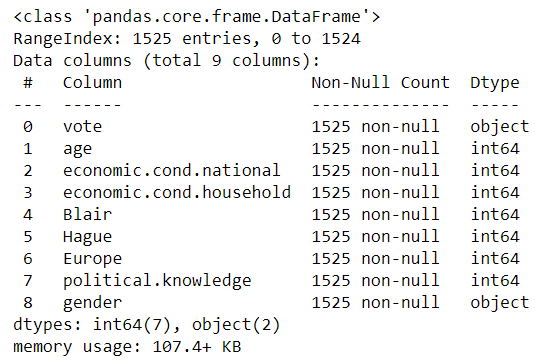
**Data Dictionary:**

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

**1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks).**

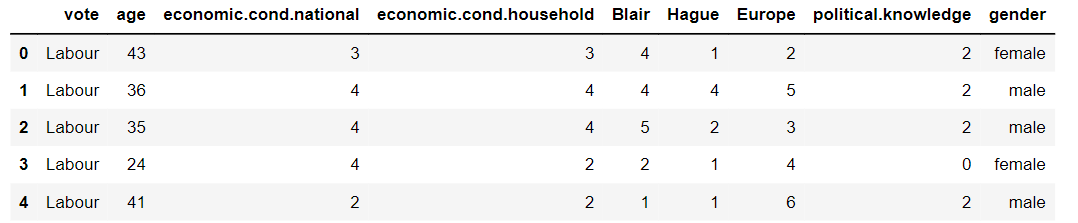
**Data inference:**

* There are 1525 rows and 9 columns in the data set.
* In the dataset, two variables are of object datatype and six variables are of integer datatype.



**Fig 1.1-Datatype**

* The first five rows of the dataset are as follows

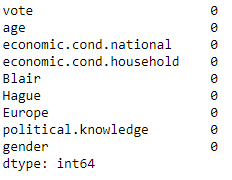


**Table-1.1-First five rows of the dataset**

**Duplicates:**

* There are ***eight*** duplicates in the dataset. These are removed as our dataset is large enough to handle it.

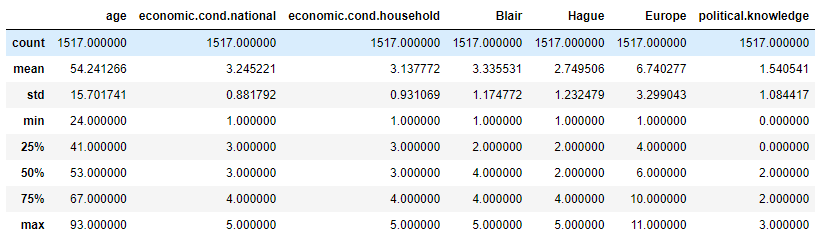
**Null values:**

****

**Fig-1.2- Null values**

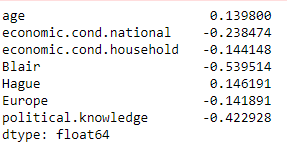
* There are no null values in the dataset.

**Data description:**

****

**Table-1.2-Description**

**Skewness:**



* skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data.
* Between -0.5 and 0.5 = Nearly symmetrical=
* Between -1 and -0.5 = Slightly Negatively skewed
* Between 0.5 and 1= Slightly Positively skewed
* Lower than -1 and greater than 1= Extremely skewed+
* Except Blair, all other variables are nearly symmetric.

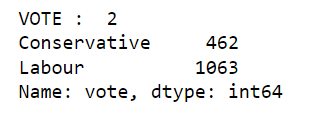
**1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical 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**

**Exploratory Data Analysis:**

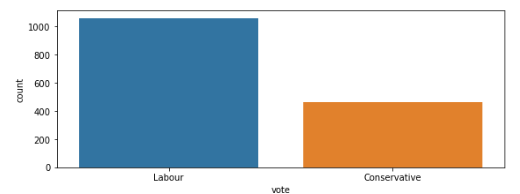
**Univariate and Bivaraiate Analysis:**

**a) Vote variable:**

* This variable indicates the party choice of the voters.
* This is the target variable of our dataset.
* There are two categories in this variable. They are the conservative party and Liberal party.
* Votes to conservative party is 462 and to the labour party is 1063.
* Nearly 70 percent of the voters voted for the Labour party and 30 percent of the voters voted for the conservative party.
* This model is slightly unbalanced.

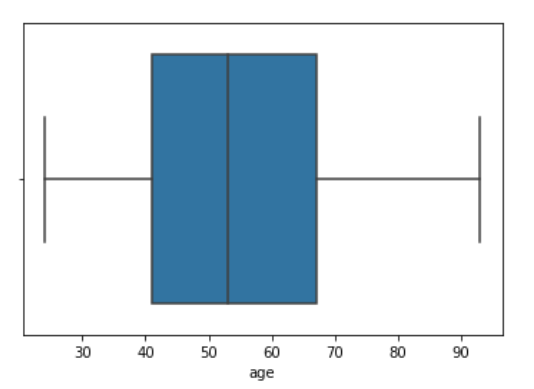
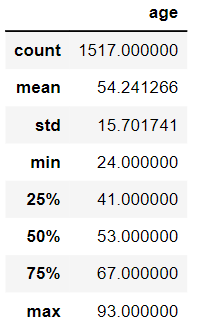
 

**Fig 1.3- Frequency distribution-Vote**



**Fig 1.4-Count plot-Vote**

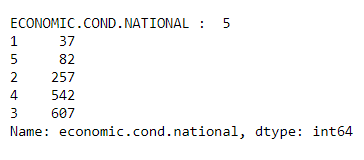
**b) Age:**



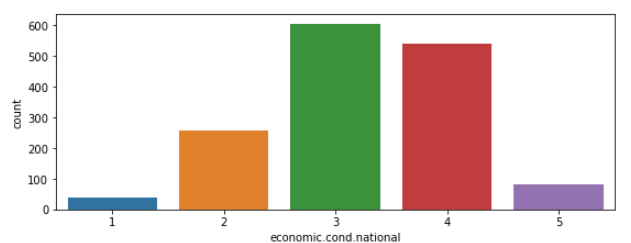
**Fig 1.5- Boxplot-Age**

* The minimum age of voters considered in the dataset is 24 years.
* Almost 75 percent of the voters are below the age of 67 years.
* Median age is 53.
* The maximum age of the voter is 93 years.
* There are **no outliers** in the dataset.

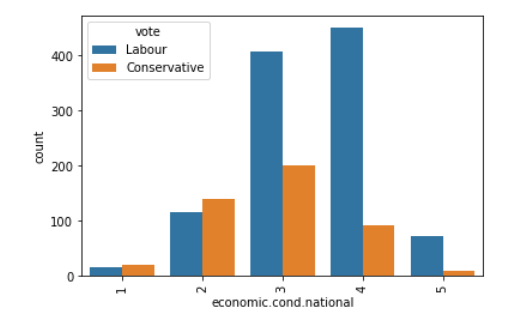
**C) economic.cond.national:**



**Fig 1.6- Frequency distribution-national condition**



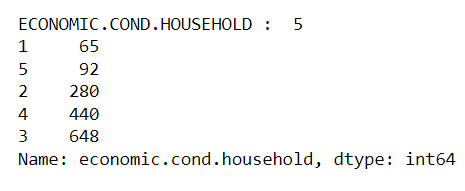
**Fig 1.7-Count plot- national condition**



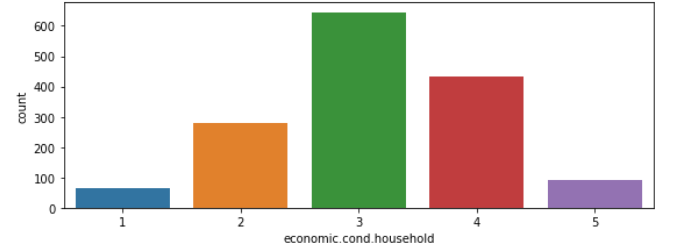
**Fig 1.8-Count plot-national condition vs vote**

* This variable indicates the satisfactory level of the voters on current national economic conditions.
* It ranges from 1 to 5 where 1 indicates lowest satisfaction and 5 indicates highest satisfaction.
* Here most of the voters chose the moderate level. i.e. 3
* It seems that currently the labour party is in power. So, people who have a satisfication level of 3 and above voted in favor of the labour party compared to the conservative party.
* The level of 1 and 2 voted more in favour of the conservative party.

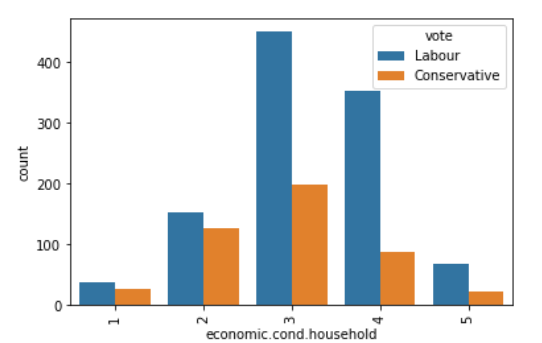
**d) economic.cond.household:**



**Fig 1.9- Frequency distribution-Economic condition**



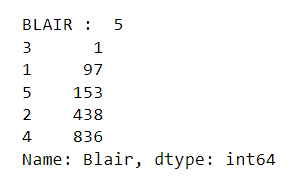
**Fig 1.10-Count plot-economic condition**



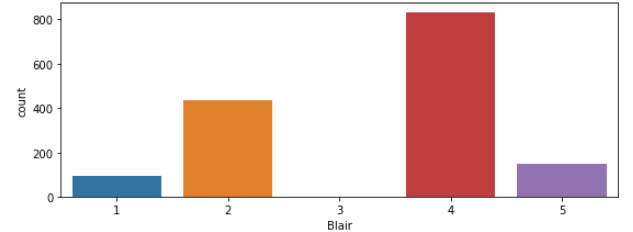
**Fig 1.11-Count plot-economic condition vs vote**

* This variable indicates the satisfactory level of the voters on current household economic conditions in the range of 1 to 5.
* Similarly, voters with satisfaction level of 3 and above, seems to have chosen labour party in higher proportion.
* Though the voters with lower satisfactory level of of 1 and 2 voted more in favour of labour party but the difference to the conservative party is less.

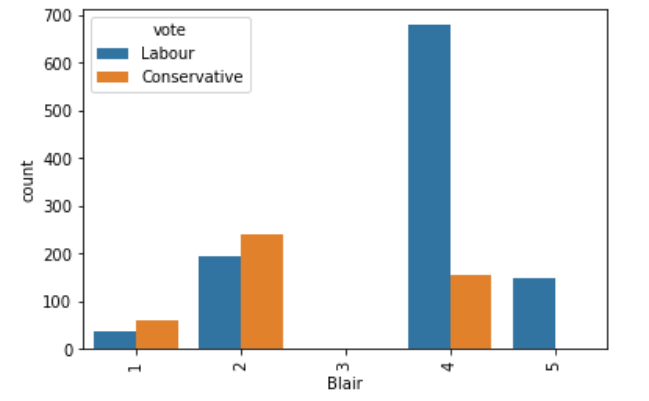
**e) Blair**



**Fig 1.12- Frequency distribution-Blair**

****

**Fig 1.13-Count plot-Blair**

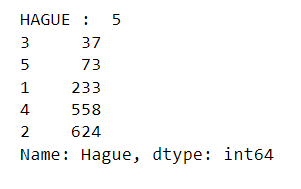
****

**Fig 1.14-Count plot-Blair vs Vote**

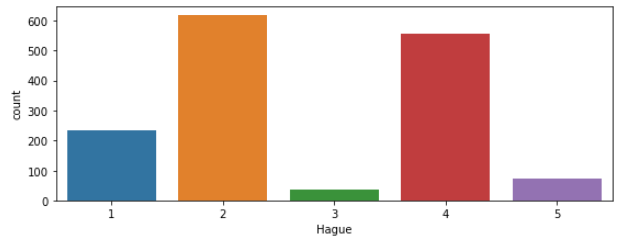
* Blair is Tony Blair who is contesting the election from Labour Party seat.
* Range 1 indicates low satisfaction and 5 indicates high satisfaction.
* Voters who gave a Satisfactory level of 4 and 5 voted more in favour of the labour party.
* Only one person gave a satisfactory level of 3 for blair.
* Low satisfication levels of 1 and 2 for blair preferred for the conservative party.

**f)Hague**

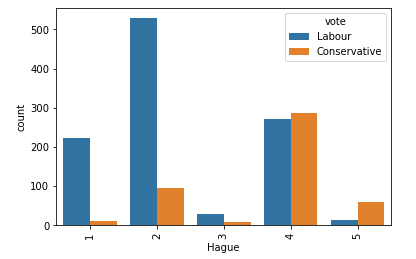
* Hague is the candidate selected from the conservative party.
* It is clear that most of the voters seem to have lower satisfaction towards hague as many voters gave scores of 1 and 2.
* Voters with a satisfactory level of 4 and above voted more in number for the conservative party.
* Low score towards Hague resulted in more votes in favour of the labour party.



**Fig 1.15- Frequency distribution-Hague**

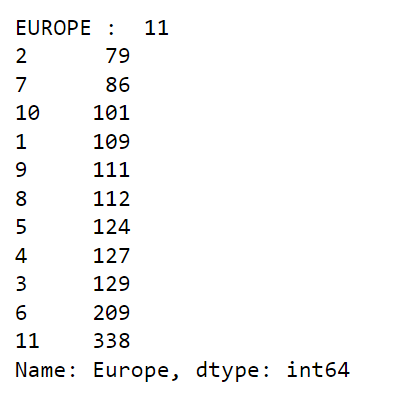


**Fig 1.16-Count plot-Hague**

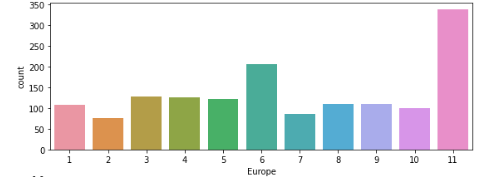


**Fig 1.17-Count plot-Hague vs Vote**

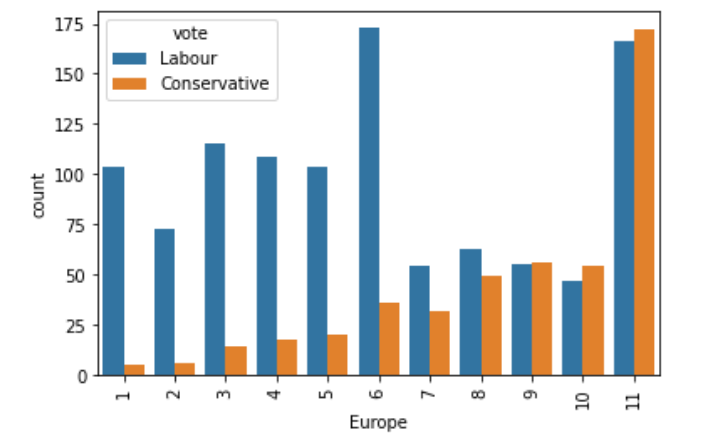
**g) Europe:**



**Fig 1.18- Frequency distribution-Europe**



**Fig 1.19-Count plot-Hague**

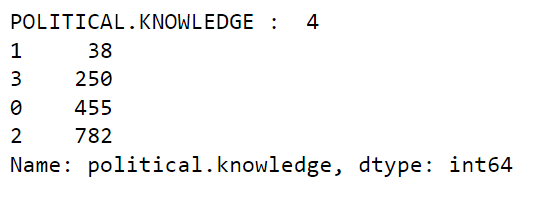


**Fig 1.20-Count plot-Hague vs Vote**

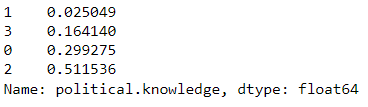
* It is an 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment.
* More the score, more the preference for the conservative party among the voters and vice versa.
* Voters who gave a score of 1 to 6 voted more in proportion to the labour party.
* It indicates that the conservative party may be skeptical towards EU integration.

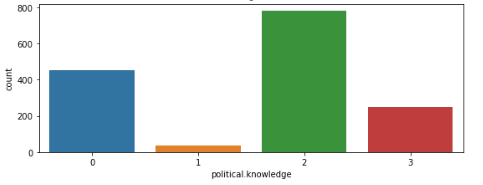
**h) Political Knowledge**:

* It denotes the voter’s knowledge of parties positions on European integration, 0 to 3.
* Nearly 30 percent of the voters do not have political knowledge.
* Irrespective of their political knowledge, voting for the labour party is high in number.

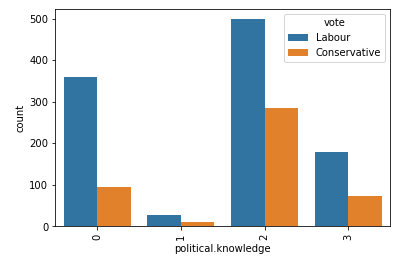


**Fig 1.21- Frequency distribution-Political**



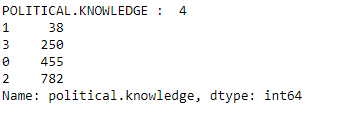


**Fig 1.22-Count plot-Political**

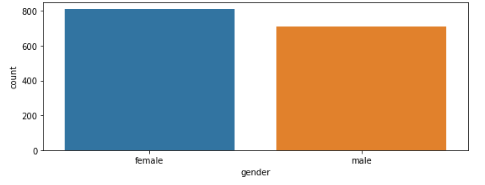


**Fig 1.23-Count plot-Political vs Vote**

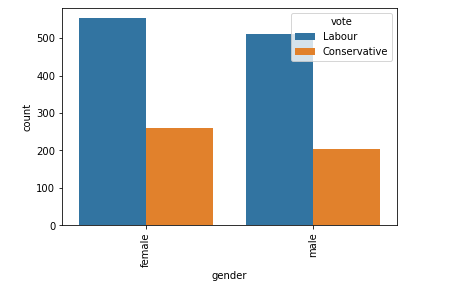
**g) Gender:**



**Fig 1.24- Frequency distribution-Gender**



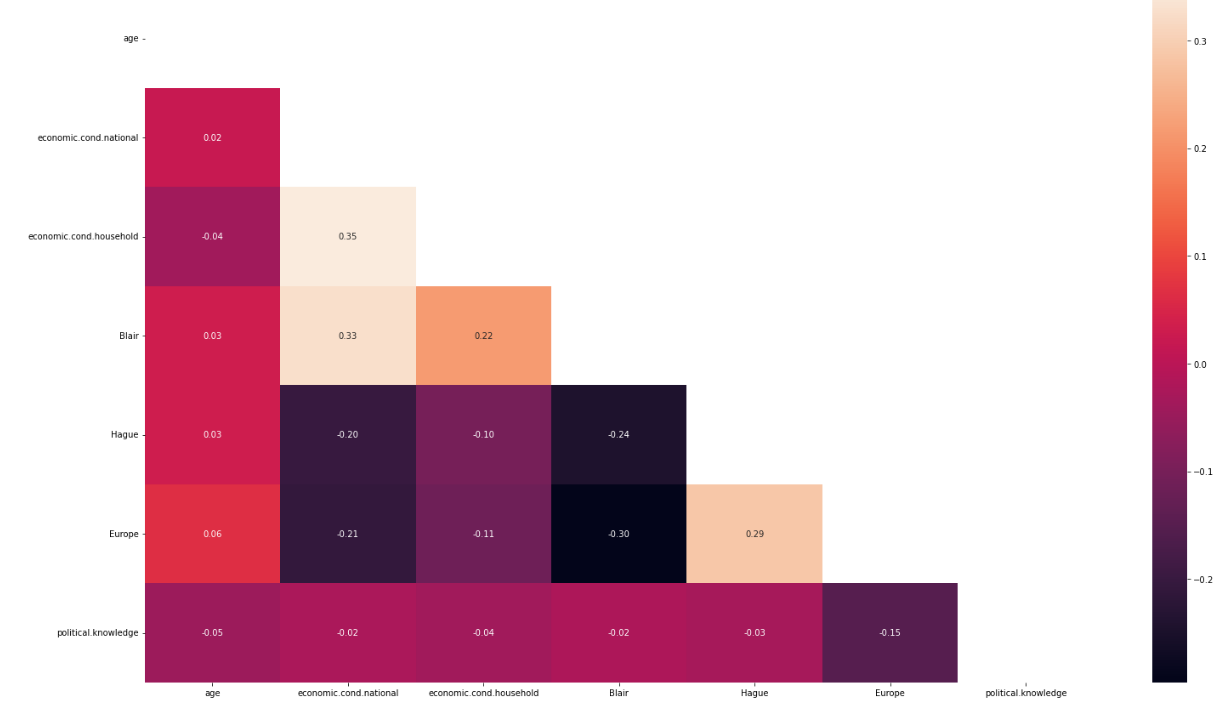
**Fig 1.25-Count plot-Gender**



**Fig 1.26-Count plot- Gender vs Vote**

* There are more female voters in the data set.
* Irrespective of gender, voting for the parties is almost equally distributed.

**Multivariate Analysis:**



**Fig 1.27-Correlation Matrix**

There is no major correlation in the dataset.

**Moderate Positive Correlation:**

* economic.cond.national and economic.cond.household
* Blair and economic.cond.national
* Blair and economic.cond.household
* Europe and Haugue:

**Moderate negative correlation:**

* Hague and economic.cond.national
* Hague and Blair
* Europe and Blair
* Europe and economic.cond.national

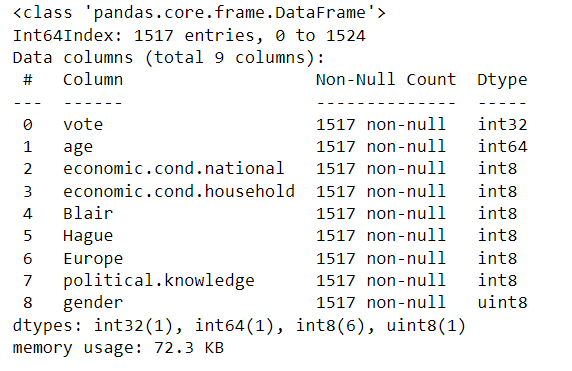
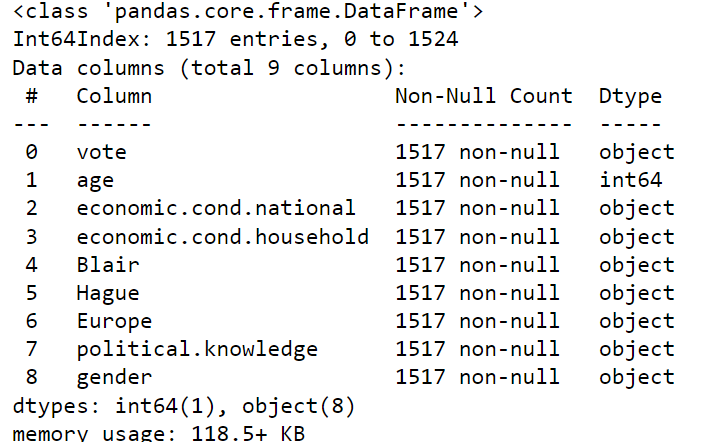
**Data Preparation**

**1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (4 Marks)**

**Encoding:**

Encoding is done to convert the categorical variable into numeric in order to feed into the model.

* Variable “age” is a continuous variable.
* Variables economic.cond.national, economic.cond.household, Blair, Hague, Europe and political.knowledge are ordinal data. So, they are encoded with label encoding.
* Variable “Gender” is a nominal data thus there is no order present. So, it is encoded by one-hot encoding.
* Our target variable “Vote” is encoded with label encoding as 1’s and 0’s. Here, 0’s represent conservative party and 1’s represent labour party.



**Fig 1.28-Encoding**

**Scaling:**

Scaling is the technique to bring the data points closer to each other. In this data set, scaling is done for three reasons:

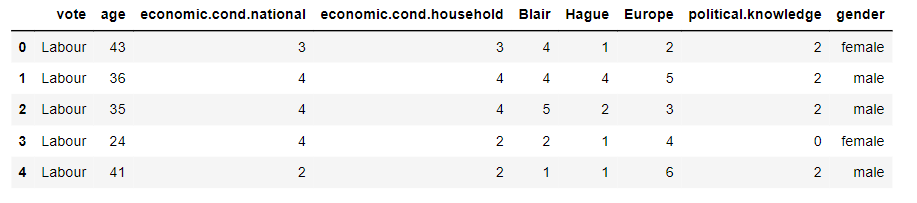
1)For easier comparison among different measurement units. For example, in this dataset variable ‘age’ is given as a continuous variable whereas the variable “gender” is a categorical variable.

2)High variance between variables. Variables with high variance may get higher weightage than the other variables particularly in ***distant based algorithms*** such as KNN, thereby influencing the model building.

3) It is also time-consuming to train and run the algorithm with this wide variation. So, it is necessary to scale down the data points around the mean as zero.

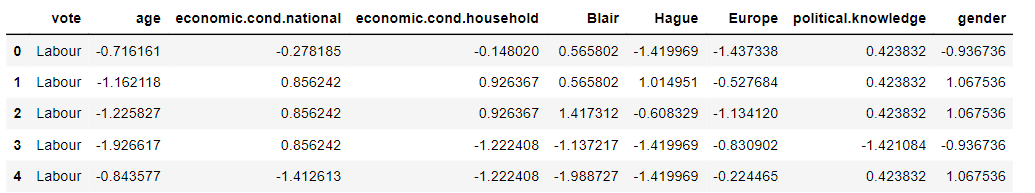
4) Here, we use standard scaler to scale the variables.

**Before scaling:**



**Table-1.3-Before Scaling**

**After Scaling**

****

**Table-1.4-After Scaling**

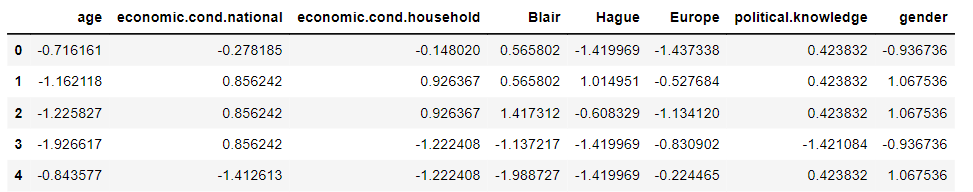
**Splitting the dataset:**

•Dataset is split into independent variables and dependent variable (Target variable).

•Independent variables-Includes all variables except ‘vote’ variable

•Dependent variable- It includes only the target variable ‘vote’.

Below is the first five rows of independent variables.



**Table-1.5-Independent Variables**

Then, it is split into training data and testing data.

* Training set-To train the model.
* Testing set-To evaluate whether the model can generalise well to new/unseen data. For instance, if the training accuracy is extremely high while the testing accuracy is poor then this is a good indicator that the model is probably overfitted.
* Here, 30 percent of the dataset is taken as testing set.

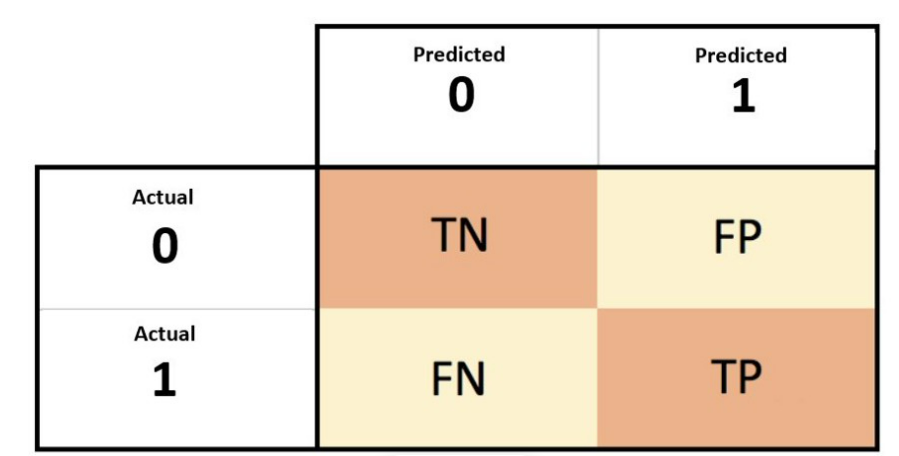
**Modeling**

Logistic Regression:

* It is a supervised classification algorithm.
* **Hyperpararameter tuning:** By using grid search, we find the best parameters for the algorithm.
* Solver- There are various solvers such as 'sag','lbfgs','newton-cg','liblinear'.
* Penalty- It is done to regularize the model to avoid overfitting and to get the best predictions. Different penalties are 'l2','l1', 'elasticnet','none'.
* Tol- It tells algorithm when to stop.
* After grid search, we found the best parameter for the model as 
* Then, the model is built by using the above parameters and used for prediction.

**Performance Metrices:**

**Confusion matrix:**



True Negative (TN)-Actually voted for conservative party and also predicted the same.

False Positive (FP)- Actually voted for the conservative party but predicted as labour party.

False negative (FN)- Actually voted for the labour party and but predicted as conservative party.

True positive (TP)- Actually voted for labour party and also predicted the same.

**ROC(Receiver operating characteristics)-AUC(Area under the curve) curve:**

* ROC is used to visualize the model performance.
* X-axis denotes False positivity rate and y- axis denotes the True positive rate.
* It helps us visualize how well our machine learning classifier is performing
* AUC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. More area under the curve, better the model.
* When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives.
* When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values.

**Logistic Regression:**

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training data | Testing data |
| classification report |  |  |
| Confusion matrix |  |  |
|  |  | |

**Table-1.6-Logistic regression**

There is neither overfitting nor underfitting in this dataset. **As the difference between train and test accuracies is less than 10%, it is a valid model.**

**Linear Discriminant Analysis:**

* LDA is also a supervised classification algorithm. Here, the model is built and predicted.
* Probability cut-off was varied to improve the predictability. But, the default cutoff of 0.5 is much better than other cutoffs.
* **Model Tuning:**
* Best parameters:

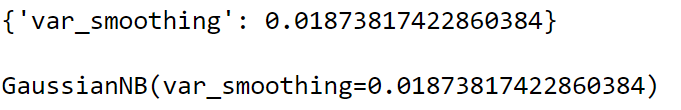


* ‘svd’: Singular value decomposition (default). This solver Does not compute the covariance matrix.
* After tuning, there is no significant difference found in the model.

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training data | Testing data |
| classification report |  |  |
| Confusion matrix |  |  |
|  |  | |

**Table-1.7-LDA**

**GaussianNB:**

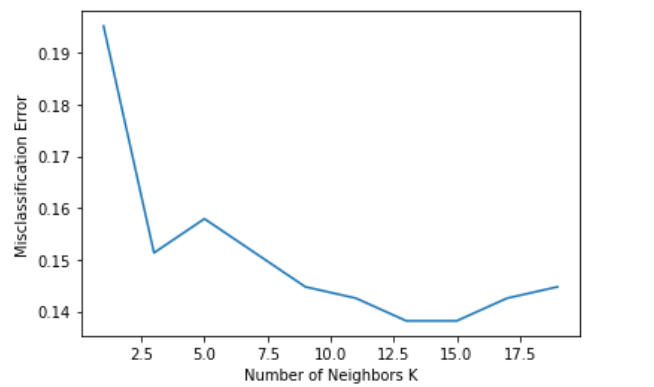
* Now GaussianNB classifier is built. The classifier is trained using training data.
* We can use fit() method for training it.
* After building a classifier, our model is ready to make predictions.
* We can use predict() method with test set features as its parameters
* **Model tuning:**
* Best parameter
* 
* Var\_smoothing- Portion of the largest variance of all features that are added to variances for calculation stability.

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training data | Testing data |
| classification report |  |  |
| Confusion matrix |  |  |
|  |  | |

**Table-1.8-Naïve bayes**

**KNN Model:**

* It is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) method
* For default k=5, we are getting accuracy in test data as 0.85.
* **Model tuning**- In order to improve the accuracy of the model, we adjust the value of K.



* From the graph, it is clear the k=15 is giving the best accuracy. So, we use this to build the model.

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training data | Testing data |
| classification report |  |  |
| Confusion matrix |  |  |
|  |  | |

**Table-1.9-KNN**

1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks)

**Boosting**

Boosting is the sequential model-building process. **Boosting** is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers.

**Techniques in boosting:**

**Adaboost:**

An additive model where shortcomings of previous models are identified by high-weight data points.

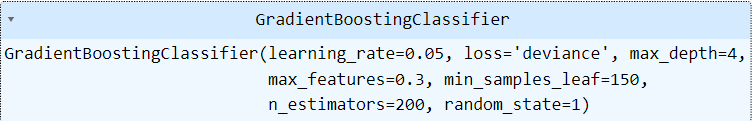
|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training data | Testing data |
| classification report |  |  |
| Confusion matrix |  |  |
| ROC Curve |  | |

**Table-1.10-Adaboost**

**Gradient Boosting:**

An additive model where shortcomings of previous models are identified by the gradient.

Model tuning:



After model tuning, the model has improved.

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training data | Testing data |
| classification report |  |  |
| Confusion matrix |  |  |
|  |  | |

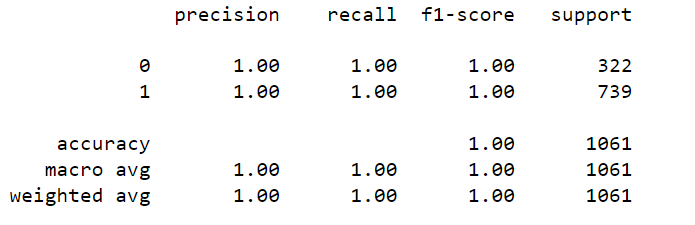
**Table-1.11-Gradient boost**

**Bagging:**

Here the models learns in parallel and combines them for determining the model average.

**Random forest:**

* When the model is trained, it is found to be overfitted.



* Here bagging decreases the [variance](https://www.geeksforgeeks.org/mathematics-mean-variance-and-standard-deviation/)and helps to avoid [overfitting](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/) in the random forest model.

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training data | Testing data |
| classification report |  |  |
| Confusion matrix |  |  |
|  |  | |

**Table-1.12-Random forest(Bagging)**

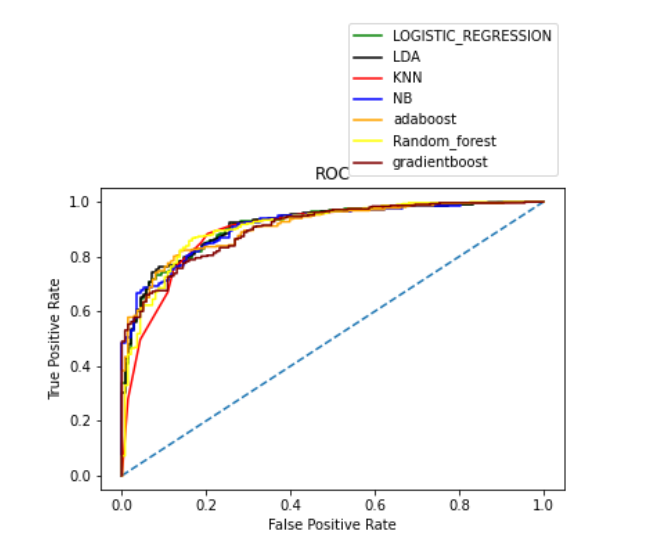
In this model accuracy between the train data and test data is more than 10 percent. So, this indicates the presence of overfitting even after bagging.

**1.7.Final Model: Compare the models and write inference which model is best/optimized. (7 marks)**

**Test data for all models:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic regression (Test) | LDA (Test) | Naïve Bayes | KNN (Test) | Adaboost | Gradient boost | Random forest  (Bagging) |
| Accuracy | 0.86 | 0.85 | 0.86 | 0.86 | 0.84 | 0.86 | 0.85 |
| Recall | 0.93 | 0.92 | 0.92 | 0.92 | 0.90 | 0.92 | 0.92 |
| AUC | 0.913 | 0.914 | 0.912 | 0.89 | 0.908 | 0.91 | 0.91 |
| F1 score | 0.90 | 0.90 | 0.90 | 0.90 | 0.88 | 0.90 | 0.90 |
| Precision | 0.87 | 0.87 | 0.88 | 0.89 | 0.87 | 0.88 | 0.87 |

**Table-1.13-Final model**

****

True Negative (TN)-Actually voted for conservative party and also predicted the same.

False Positive (FP)- Actually voted for the conservative party but predicted as labour party.

False negative (FN)- Actually voted for the labour party but predicted as conservative party.

True positive (TP)- Actually voted for labour party and also predicted the same.

In this dataset, True positive and True Negative are equally important.

**F1 Score:**

* It is best when we need to take both recall and precision into account.
* F1 score is best for model KNN as it has a F1 score of 90 percent as the difference between train and test data is just 1%.

**Sensitivity/Recall:**

* False negative (FN) is it fails to identify the voters of the labour party.
* Sensitivity/recall is being influenced by false negativity.
* Increase in sensitivity, decreases the false negativity.
* Recall is better for all the models than precision.

**Precision:**

* It tells how sure one is of determining the true positive.
* That is, how precisely it finds the labour party.
* Here, the precision is better for all the models.
* But in KNN, it performs very good.

**Accuracy:**

* It tells how accurately the model classifies the voter as conservative and labour party.
* This metric is important as both the classes are important in our dataset.
* All model have accuracy of at least 85 percent.

**AUC**:

* The AUC score ranges well within the range of 0.5-1. It tells that the model classifies the really well.

**Overall insights:**

* So, when taking into account all the performance metrics KNN outperforms the other models, It also shows less variability in performance between train and test data

**1.8 Based on these predictions, what are the insights?**

**Insights:**

* Overall labour party has more vote shares than the conservative party.
* The satisfactory levels of ECONOMIC.COND.NATIONAL and ECONOMIC.COND.HOUSEHOLD influenced the voter’s choice.
* Blair is the most preferred candidate than hague.
* There is significant impact of gender in voting pattern between the parties.
* The influence of the conservative party is more between the age group of 65-75 compared to other age groups.
* Eurosceptic voters preferred the conservative party than the labour party and vice versa.
* Expect random forest, there is no presence of overfitting in the models.

**Business recommendations:**

* Nearly 30 percent of the voters are unaware of party’s position on European integration. So, by enough campaigning, the conservative party can pull more vote share from the above group.
* By taking a moderate stand on Europe integration, conservative party can improve its vote share.
* Conservative party can change the candidate as hague has less influence compared to blair.
* By enough welfare measures, both parties can improve the satisfactory level of voters on both household and national economic conditions.

**Problem 2:**

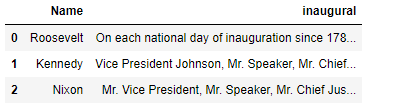
**In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:**

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**

**a) Number of characters:**

For better understanding, we have taken the Inaugural speeches in the form of a data frame.

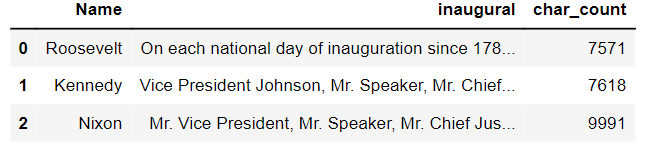


**Table-2.1-Data frame**

Here we used len() function to determine the number of characters.

**Number of characters before removing punctuations:**

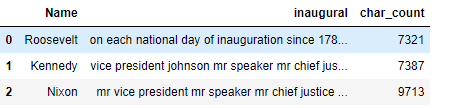
* Roosevelt’s inaugural speech is 7571
* Kennedy’s inaugural speech is 7618
* Nixon’s inaugural speech is 9991



**Table-2.2-Characters**

The total number of characters by combining all the texts is 25180

**Number of characters after removing punctuations:**



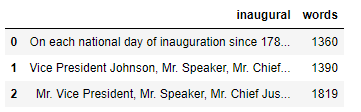
The total number of characters by combining all the texts is 24421

**b) Number of words:**

By using split() function, we find the number of words.

Number of words without removing punctuations:

* Roosevelt- 1360 words.
* Kennedy- 1390 words
* Nixon – 1819 words
* Total words- 4569



**Table-2.4-Words**

Number of words after removing punctuations:

* Roosevelt- 1338 words.
* Kennedy- 1365 words
* Nixon – 1802 words
* Total words- 4505.

**Number of sentences:**

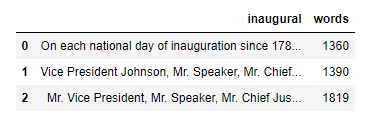
By using sents() function, we find the number of sentences

* Roosevelt- 68 sentences
* Kennedy- 69 sentences
* Nixon- 52 sentences

**2.2 Remove all the stopwords from all three speeches.**

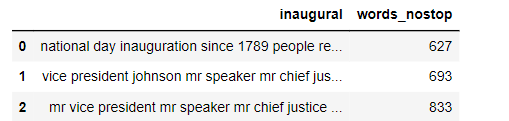
Stop words are words which do not help us in the data analysis part. For example- a, an, etc

Number of words before removal of stop words:



**Table-2.5-Count of Words-Before stop words removal**

Number of words after removal of stop words:



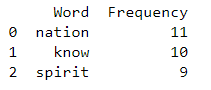
**Table-2.6-Count of Words-After stop words removal**

* From the above we can say that almost half of the unnecessary words is removed.
* Sample sentence- *‘national day inauguration since 1789 people renewed sense* dedication’
* When we look at the inaugural column we can see the sentence without stop words. Removing the stop words will help us in improving the computation speed and accurate analysis.

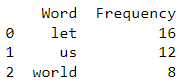
**2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopword)**

* Now we need to split the sentences into words. This process is called as Tokenization.
* The top 3 words which has maximum occurrence in the speeches are

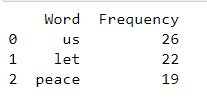
**Roosevolt:**



**Kennedy:**



**Nixon:**



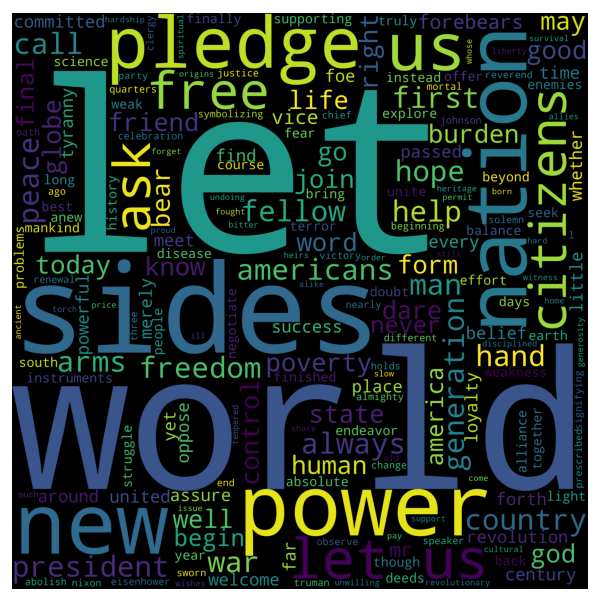
**2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)**

**Roosevolt:**

****

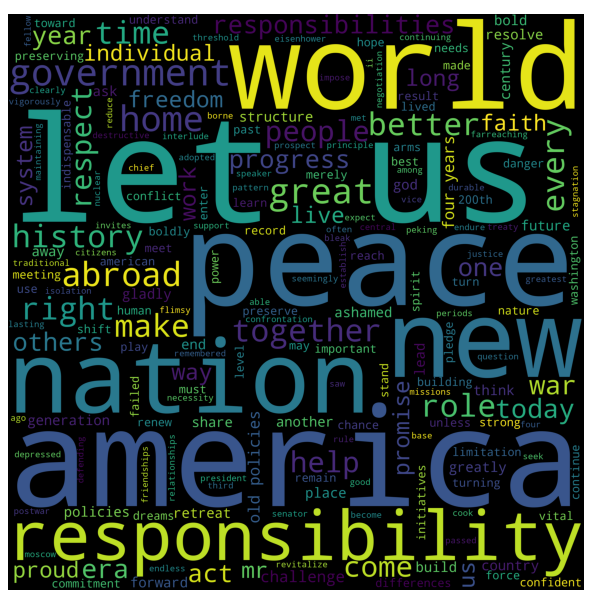
**Fig 2.1-Word cloud roosevolt**

**Kennedy:**

****

**Fig 2.2-Word cloud kennedy**

**Nixon:**



**Fig 2.3-Word cloud Nixon**