**Machine Learning**

Student Name: Suneel Kumar Pentapalli

Date: 05/Aug/2023



Contents

[***Problem 1*** 3](#_Toc142843461)

[1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks) 3](#_Toc142843462)

[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. 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. 7](#_Toc142843463)

[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). 12](#_Toc142843464)

[1.4 Apply Logistic Regression and LDA (linear discriminant analysis). 12](#_Toc142843465)

[1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results. 14](#_Toc142843466)

[1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. 17](#_Toc142843467)

[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. 18](#_Toc142843468)

[1.8 Based on these predictions, what are the insights? (5 marks) 20](#_Toc142843469)

[***Problem 2*** 22](#_Toc142843470)

[2.1 Find the number of characters, words, and sentences for the mentioned documents. 22](#_Toc142843471)

[2.2 Remove all the stopwords from all three speeches. 23](#_Toc142843472)

[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) – 3 Marks 25](#_Toc142843473)

[2.4 Plot the word cloud of each of the speeches of the variable. 26](#_Toc142843474)

# 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.

Dataset for Problem: [Election\_Data.xlsx](https://olympus.mygreatlearning.com/courses/90007/files/8584387/download?verifier=pJ8zNX2FHgzFxE31sx93pd9ipC6D0Jy7qJs7kRMy&wrap=1)

Data Ingestion: 11 marks

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

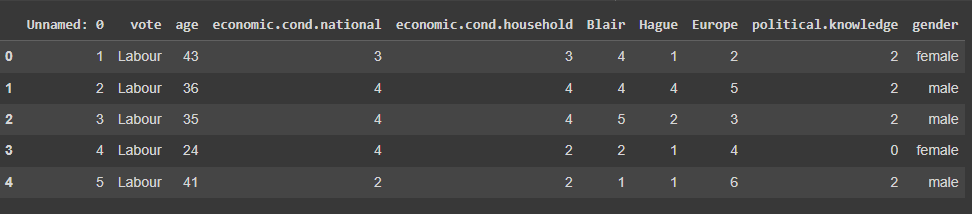
Given dataset has 10 features and 1525 records.

Below are the feature names along with the respective datatypes,

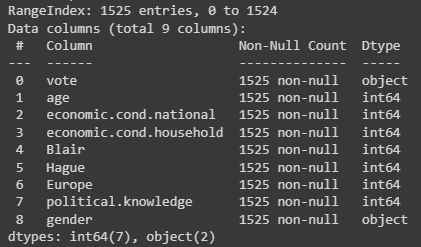
A screenshot of a computer program

Description automatically generated

Given data has 8 features with integer datatype and 2 features with object data types, sample data for the same is as follows,

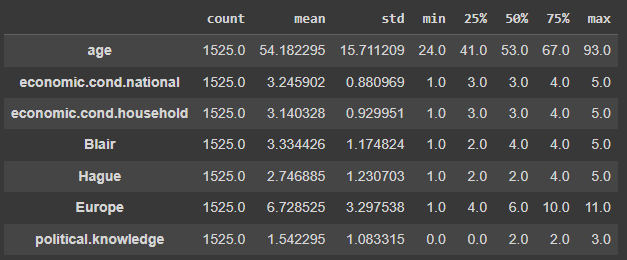


As we can see there is additional index column “**Unnamed: 0**” in the dataset which doesn’t add any value to our analysis. Hence removing it.



**Five Point summary:**

Five point summary after removing unnecessary columns is as follows,



**age:** The data we got for analysis has voters with mean age of 54 years.

Minimum age of the voter in the data is 24 years and max is 93 years and median lies near 53 years.

Current **national economic conditions(economic.cond.national)** and household economic conditions(**(economic.cond.household)**  seems similar with mean of 3.24 and 3.14 respectively.

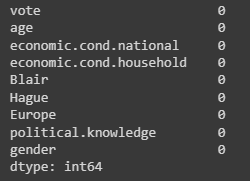
Mean rating for labour leader(**Blair**) is 3.3 which is pretty high compared to mean rating of conservative leader(**Hague**) which is 2.7

**Europe:** on a scale of 1 to 11, mean rating of euroscepticism is 6.7 which is a bit higher than median rating 6.

**political.knowledge:** mean rating of knowledge of parties and positions on a scale of 1 to 3 is 1.54

**Null Values:**

As shown below, none of the features have null entries,



**Duplicate Check:**

There are 8 duplicate entries found in the dataset, as they add no value to our analysis, we have removed them

A screenshot of a computer

Description automatically generated

Record count after removing duplicates is 1517.

Lets explore the skewness part of all the features,

Below graph shows the distribution of each value/range from the dataset, its evident that the data is none of the columns are evenly distributed. Out of all columns, age column seems a bit normally distributed but it has a little skewness towards left.

In remaining columns, peak in graph represents more records with that particular x-value.

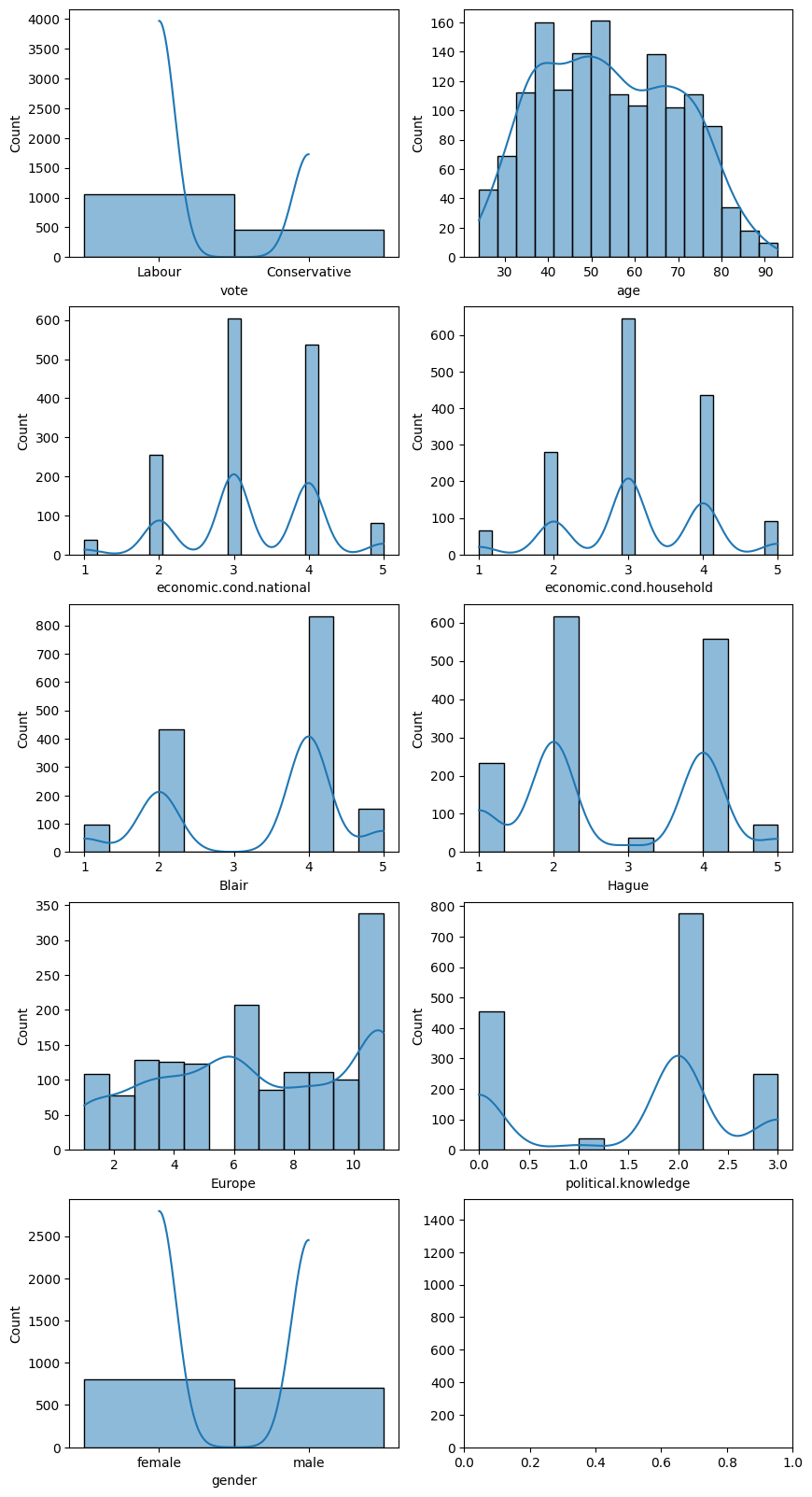


Fig 1-Hist Plots

### 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. 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.

Data Preparation: 4 marks

Null checks, data type checks has been already discussed in above section 1.1, let continue with further analysis here,

Correlation Chart:

Below chart shows how variation in each attribute change/impact the values in other attribute.

Correlation values in below graph range from -1 to +1, where -1 represents the attributes are highly negatively corelated and +1 represents high positive correlation. Based on below information we can conclude there is very less correlation between any two variables.

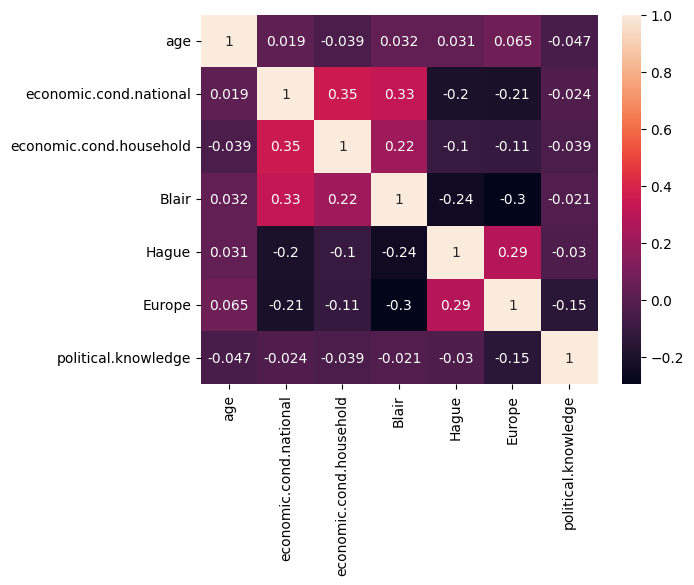
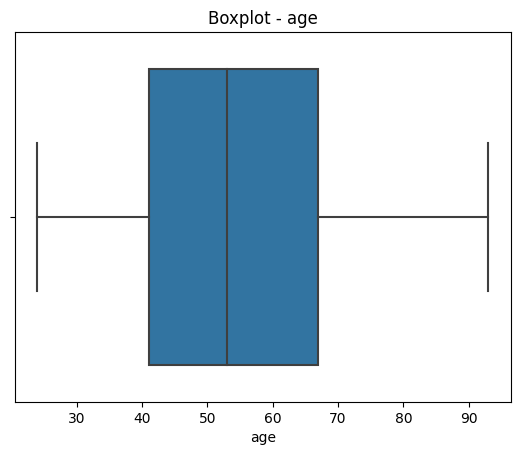


Fig 2- Heat Map (Co-relation Plot)

**Univariate Analysis:**

**Univariate analysis** is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

As part of univariate analysis, lets create box plots to check for outliers in our data. **Outliers** can be identified as points that lie outside plots whiskers. In below graphs, we can see **outliers** in attributes “**economic.cond.national**” and “**economic.cond.household**”.

A graph with a blue rectangle

Description automatically generated

A graph with a blue rectangle

Description automatically generated A diagram of a box plot

Description automatically generated A blue box plot with black lines

Description automatically generated A blue square with black lines

Description automatically generatedA diagram of a box plot

Description automatically generated

Fig 3 - Box Plots

**Outlier Proportion:**

Above mentioned attributes “**economic.cond.national**” and “**economic.cond.household**” have outlier count of 37 and 65 records respectively that constitutes to about 2.4% and 4.2% of total records respectively.

Note: We are not removing outliers in this case as, though values were out of range but all the values are valid.

**Bivariate Analysis:**

In bivariate analysis we study distribution and patterns between two variables as shown below,

Below count plot on vote choices among age groups shows majority of age groups having “Labour” party as their choice with highest “Labour” party votes in lower age groups.

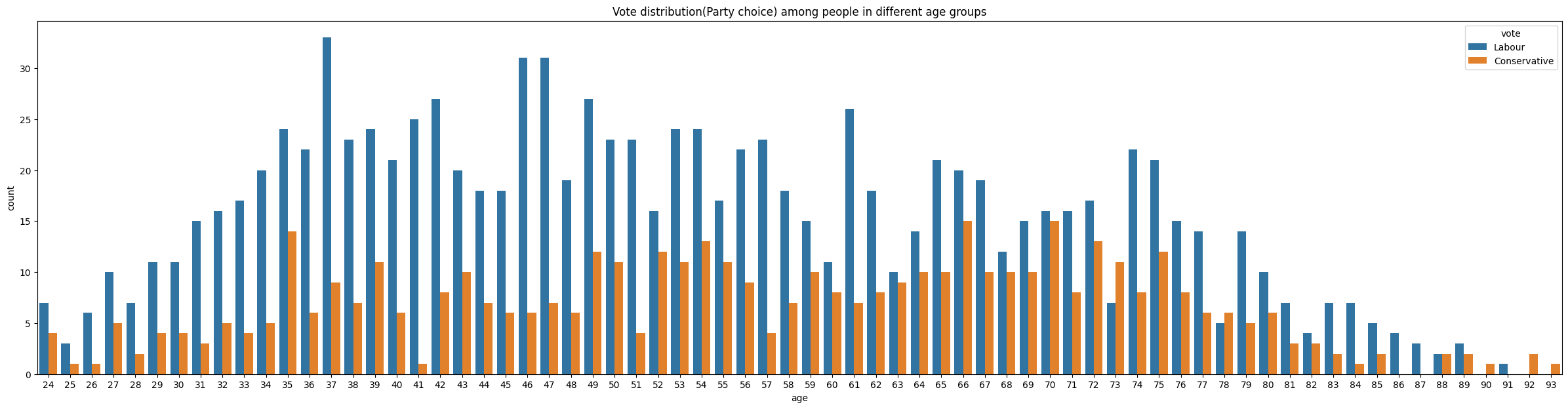
****

Fig 4- Vote Distribution choice among people in different age groups

Similarly vote choices of people among all other features are as follows,

From the below plots, “Distribution of votes among – Europe” is interesting as it shows increase in Eurosceptic sentiment drawing more “Conservative” votes.

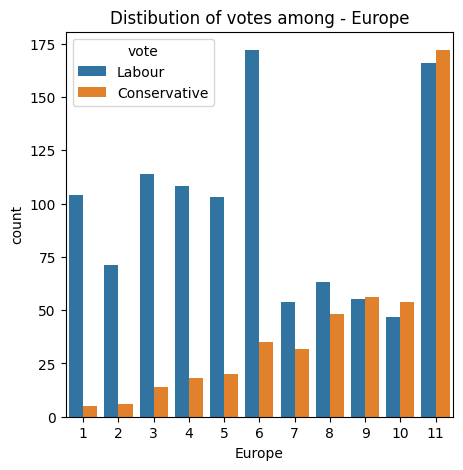
A graph of a number of blue and orange bars

Description automatically generated A graph of blue and orange bars

Description automatically generated A graph of blue and orange bars

Description automatically generated ­A graph of blue and orange bars

Description automatically generated

­­A graph of a number of people

Description automatically generated with medium confidence

Fig 5 – Distribution of Votes (Hist Plots)

### 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).

Now lets go with splitting our data in below steps,

1. **Split the data into predictor(X) and target(Y) set.** X will have all these columns (age,economic\_cond\_national,economic\_cond\_household,Blair,Hague,Europe,political\_knowledge,gender\_male) and Y will have our variable of interest(vote\_Labour*<this has been created as part of encoding, though equal importance will be given to both Conservative and Labour vote categories in analysis>*)

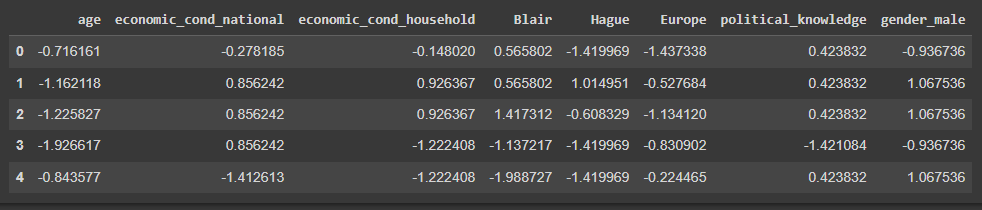
As the models we are going to apply use distance based algorithms except bayes model, hence its mandatory to scale our data.

Using Zscore technique, I have scaled data before proceeding with step 2,

**Z-score normalization** refers to the process of normalizing every value in a dataset such that the mean of all of the values is 0 and the standard deviation is 1.

We use the following formula to perform a z-score normalization on every value in a dataset:

**New value = (x – μ) / σ**



1. **Split above predictor and target sets into train(X\_train,Y\_train) and test data(X\_test,Y\_test) sets.**  
   After splitting data into train and test splits in 70:30 ratio, below are the counts of train and test datasets,

Record count of train data set : 1061

Record count of test data set : 456

Note: while splitting the data into train and test sets, we have used hyper parameter **stratify=Y,** to make the vote classes in same proportion in both train and test sets.

### 1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

(4 marks)

***Note: 0 in our target feature represents a Conservative Vote and 1 represents Labour vote. All below inferences are based on the same.***

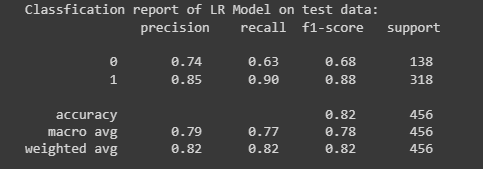
**Logistic Regression Model:**

Logistic regression is a machine learning method for binary classification. It predicts the probability of an event occurrence based on one or more features.

Logistic Regression Model score on train data: 0.83

Logistic Regression Model score on test data: 0.85

As model scores is pretty much same for both test and train data, hence we can assume there is no overfitting or underfitting of our model.

Classification report on test data:  


When using classification models in machine learning, there are three common metrics that we use to assess the quality of the model,

1. **Precision:** Percentage of correct positive predictions relative to total positive predictions.

Of all the votes that are predicted/classified as **Conservative** votes, 74% of the votes are actually voted for **Conservative** party.

Similarly, of all the votes that are predicted as **Labour** votes, 85% of the votes as actually voted for **Labour** party.

1. **Recall:** Percentage of correct positive predictions relative to total actual positives.

Of all the Votes that actually did vote for **Conservative** Party, our LR model was able to predict this outcome correctly for 64% of the votes.

Similarly, of all the votes that actually did vote for **Labour** Party, our LR model predicted this outcome correctly for 90% of the votes.

3. **F1 Score:** A weighted harmonic mean of precision and recall. The closer to 1, the better the model.

* F1 Score: 2 \* (Precision \* Recall) / (Precision + Recall)

Since f1 score for Labour votes classification is more near to 1 than that of Conservative votes classification, hence this model is more accurate an precise for classifying Labour votes. The reason for this variation in accuracies between Labour and Conservative classification is, The data set has more records with Labour votes, hence our model is able to predict labour votes more precisely that conservative votes. The distribution of Labour and Conservative Votes in our dataset is as follows,



**Linear Discriminant Analysis:**

when a response variable has more than two possible classes then we typically prefer touse a method known as linear discriminant analysis, often referred to as LDA.

LDA and logistic regression models are both used for classification, but LDA is far more stable than logistic regression when it comes to making predictions for multiple classes and is therefore the preferred algorithm to use when the response variable can take on more than two classes.

Linear Discriminant Analysis Model score on train data: 0.83

Linear Discriminant Analysis Model score on test data: 0.85

As model scores is pretty much same for both test and train data, hence we can assume there is no overfitting or underfitting of our model.

Classification report of LDA Model on test data:

A screenshot of a test

Description automatically generated

1. **Precision:**

Of all the votes that are predicted/classified as **Conservative** votes, 74% of the votes are actually voted for **Conservative** party.

Similarly, of all the votes that are predicted as **Labour** votes, 85% of the votes as actually voted for **Labour** party.

1. **Recall:**

Of all the Votes that actually did vote for **Conservative** Party, our LDA model was able to predict this outcome correctly for 64% of the votes.

Similarly, of all the votes that actually did vote for **Labour** Party, our LDA model predicted this outcome correctly for 90% of the votes.

To conclude, both LR and LDA models performance is same when classifying our votes.

### 1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.

(4 marks)

**KNN Model:**

KNN algorithm stores the entire training dataset as a reference. When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.

Next, the algorithm identifies the K nearest neighbors to the input data point based on their distances. In the case of classification, the algorithm assigns the most common class label among the K neighbors as the predicted label for the input data point. For regression, it calculates the average or weighted average of the target values of the K neighbors to predict the value for the input data point.

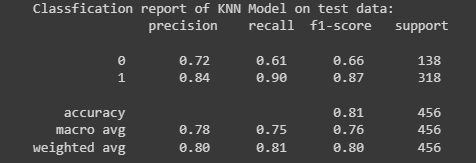
KNN model score on training and testing election data is as follows,

KNN Model score on train data: 0.87

KNN Model score on test data: 0.81

Here, model score on training data is 6 points more than that of test data, this implies overfitting of our model.

Classification Report of KNN Model on test data:



1. **Precision:**

Of all the votes that are predicted/classified as **Conservative** votes, 72% of the votes are actually voted for **Conservative** party.

Similarly, of all the votes that are predicted as **Labour** votes, 84% of the votes as actually voted for **Labour** party.

1. **Recall:**

Of all the Votes that actually did vote for **Conservative** Party, our KNN model was able to predict this outcome correctly for 61% of the votes.

Similarly, of all the votes that actually did vote for **Labour** Party, our KNN model predicted this outcome correctly for 90% of the votes.

**Naïve Bayes Model:**

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

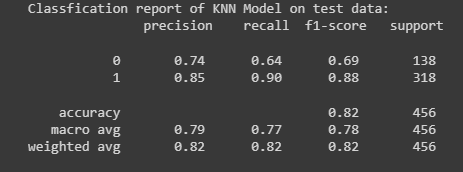
Naïve Bayes model scores on training and testing data is as follows,

Naive Bayes Model score on train data: 0.84

Naive Bayes Model score on test data: 0.82

As model scores is pretty much same for both test and train data, hence we can assume there is no overfitting or underfitting of our model.

Classification Report of Naïve Bayes model test data:



1. **Precision:**

Of all the votes that are predicted/classified as **Conservative** votes, 74% of the votes are actually voted for **Conservative** party.

Similarly, of all the votes that are predicted as **Labour** votes, 85% of the votes as actually voted for **Labour** party.

1. **Recall:**

Of all the Votes that actually did vote for **Conservative** Party, our NB model was able to predict this outcome correctly for 64% of the votes.

Similarly, of all the votes that actually did vote for **Labour** Party, our NB model predicted this outcome correctly for 90% of the votes.

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

(7 marks)

After applying tuning on all the above models, these are the model scores when model is passed with the best parameters suggested by GridSearchCV.

Logistic Regression after Model Tuning:

Model score on train data: 0.835

Model score on test data: 0.842

Classification Report on test data:

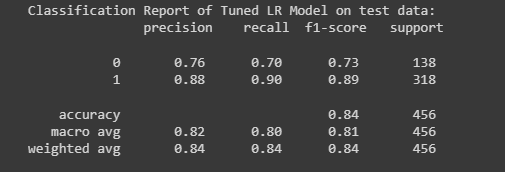


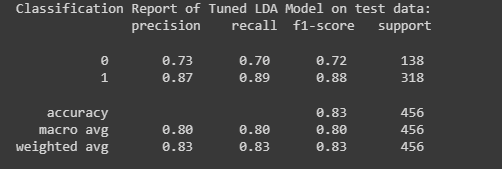
Fig1 -

Linear Discriminant Analysis after Model Tuning:

Model score on train data: 0.835

Model score on test data: 0.833

Classification Report on test data:

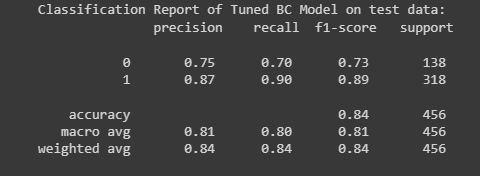


Bagging Classifier after Model Tuning:

Model score on train data: 0.92

Model score on test data: 0.83

Classification Report on test data:



ADABoosting Classifier after Model Tuning:

Model score on train data: 0.836

Model score on test data: 0.837

Classification Report on test data:

A screenshot of a computer

Description automatically generated

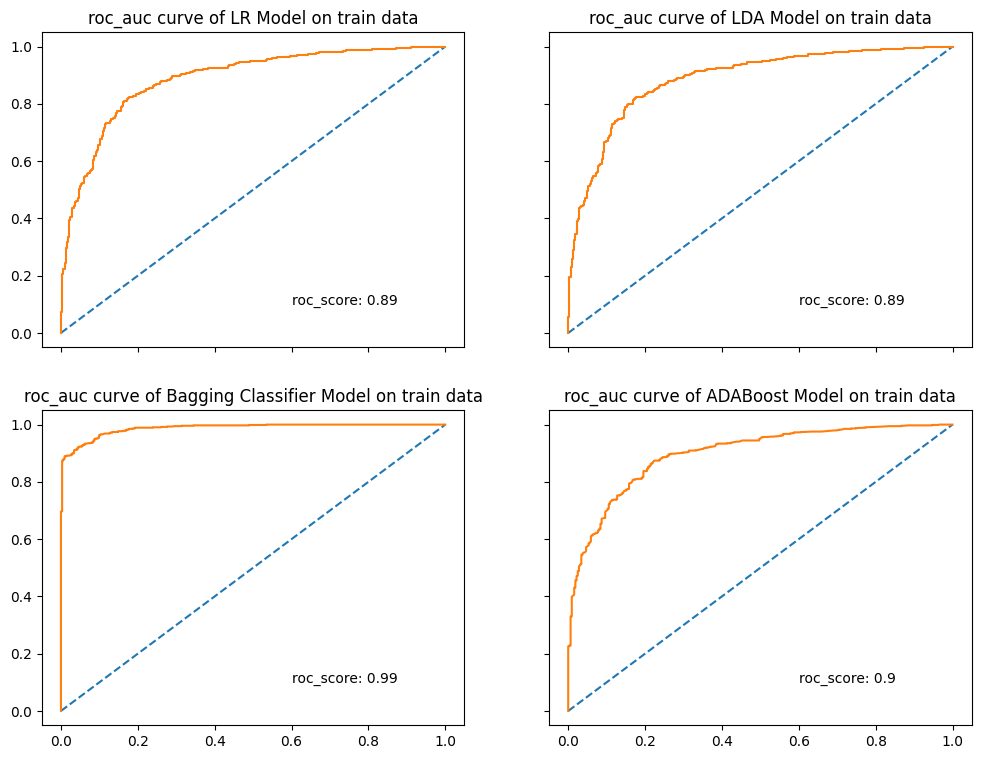
### 1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

(7 marks)

Model accuracy that we discussed in above section is threshold variant, i.e., the change of threshold values change the accuracy scores accordingly.

Hence we use ROC\_AUC\_score as it is invariant to threshold and invariant to scale.

ROC\_Curve along with ROC\_AUC\_Score for all the models on train data:



ROC\_Curve along with ROC\_AUC\_Score for all the models on test data:

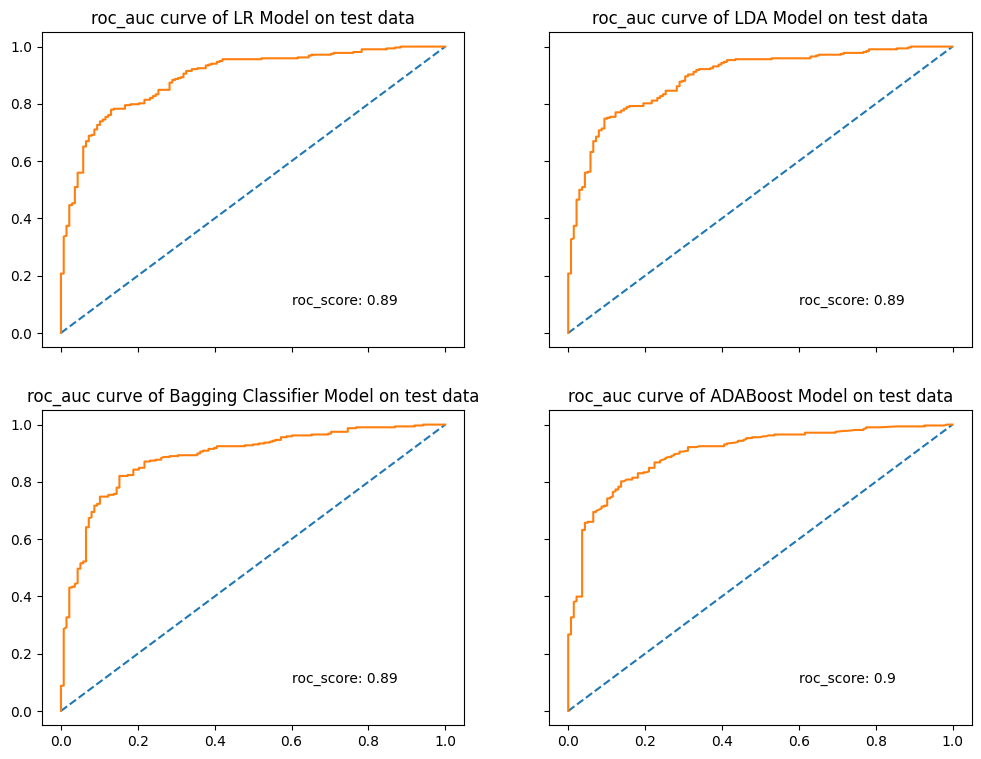


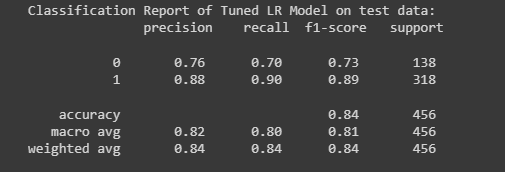
Fig 6 – Roc Curves on LR, LDR, Bagging Classifier and ADA Boost Models

### 1.8 Based on these predictions, what are the insights? (5 marks)

All the performance metrics have been discussed in above, where its observed that Logistic regression model performed better overall when compared to all other models.

Though roc\_score for ADA Boost is one point higher, metrics in classification seems better for Logistic Regression model, Hence we will go ahead with LR Model.

Insights from LR Model:



1. **Precision:**

Of all the votes that are predicted/classified as **Conservative** votes, 76% of the votes are actually voted for **Conservative** party.

Similarly, of all the votes that are predicted as **Labour** votes, 88% of the votes as actually voted for **Labour** party.

1. **Recall:**

Of all the Votes that actually did vote for **Conservative** Party, our NB model was able to predict this outcome correctly for 70% of the votes.

Similarly, of all the votes that actually did vote for **Labour** Party, our NB model predicted this outcome correctly for 90% of the votes.

1. **F1-score:**

A weighted harmonic mean of precision and recall. The closer to 1, the better the model.

For the model picked, f1 scores for the classes Conservative and Labour is 0.73 and 0.89 repectively.

# Problem 2

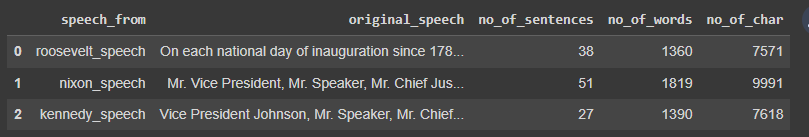
In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

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

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

### 2.1 Find the number of characters, words, and sentences for the mentioned documents.

Sentence, word and character counts of each president’s speech is as follows before removing any stop words.



As per above data, its evident that the speech given by Nixon is longer compared to other president’s speeches.

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

All the stop words along with punctuation have been removed from all the three speeches, the final text is as follows,

**Roosevelt’s Speech:**

*national day inauguration since 1789 people renewed sense dedication united states washingtons day task people create weld together nation lincolns day task people preserve nation disruption within day task people save nation institutions disruption without come time midst swift happenings pause moment take stock recall place history rediscover may risk real peril inaction lives nations determined count years lifetime human spirit life man threescore years ten little little less life nation fullness measure live men doubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplained reason tyranny slavery become surging wave future freedom ebbing tide americans know true eight years ago life republic seemed frozen fatalistic terror proved true midst shock acted acted quickly boldly decisively later years living years fruitful years people democracy brought greater security hope better understanding lifes ideals measured material things vital present future experience democracy successfully survived crisis home put away many evil things built new structures enduring lines maintained fact democracy action taken within threeway framework constitution united states coordinate branches government continue freely function bill rights remains inviolate freedom elections wholly maintained prophets downfall american democracy seen dire predictions come naught democracy dying know seen reviveand grow know cannot die built unhampered initiative individual men women joined together common enterprise enterprise undertaken carried free expression free majority know democracy alone forms government enlists full force mens enlightened know democracy alone constructed unlimited civilization capable infinite progress improvement human life know look surface sense still spreading every continent humane advanced end unconquerable forms human society nation like person bodya body must fed clothed housed invigorated rested manner measures objectives time nation like person mind mind must kept informed alert must know understands hopes needs neighbors nations live within narrowing circle world nation like person something deeper something permanent something larger sum parts something matters future calls forth sacred guarding present thing find difficult even impossible hit upon single simple word yet understand spirit faith america product centuries born multitudes came many lands high degree mostly plain people sought early late find freedom freely democratic aspiration mere recent phase human history human history permeated ancient life early peoples blazed anew middle ages written magna charta americas impact irresistible america new world tongues peoples continent newfound land came believed could create upon continent new life life new freedom vitality written mayflower compact declaration independence constitution united states gettysburg address first came carry longings spirit millions followed stock sprang moved forward constantly consistently toward ideal gained stature clarity generation hopes republic cannot forever tolerate either undeserved poverty selfserving wealth know still far go must greatly build security opportunity knowledge every citizen measure justified resources capacity land enough achieve purposes alone enough clothe feed body nation instruct inform mind also spirit three greatest spirit without body mind men know nation could live spirit america killed even though nations body mind constricted alien world lived america know would perished spirit faith speaks daily lives ways often unnoticed seem obvious speaks capital nation speaks processes governing sovereignties 48 states speaks counties cities towns villages speaks nations hemisphere across seas enslaved well free sometimes fail hear heed voices freedom privilege freedom old old story destiny america proclaimed words prophecy spoken first president first inaugural 1789 words almost directed would seem year 1941 preservation sacred fire liberty destiny republican model government justly considered deeply finally staked experiment intrusted hands american people lose sacred fireif smothered doubt fear shall reject destiny washington strove valiantly triumphantly establish preservation spirit faith nation furnish highest justification every sacrifice may make cause national defense face great perils never encountered strong purpose protect perpetuate integrity democracy muster spirit america faith america retreat content stand still americans go forward service country god*

**Nixon’s Speech:**

*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 depressed prospect seemingly endless war abroad destructive conflict home meet today stand threshold new era peace world central question shall use peace resolve era enter postwar periods often time retreat isolation leads stagnation home invites new danger abroad resolve become time great responsibilities greatly borne renew spirit promise america enter third century nation past year saw farreaching results new policies peace continuing revitalize traditional friendships missions peking moscow able establish base new durable pattern relationships among nations world americas bold initiatives 1972 long remembered year greatest progress since end world war ii toward lasting peace world peace seek world flimsy peace merely interlude wars peace endure generations come important understand necessity limitations americas role maintaining peace unless america work preserve peace peace unless america work preserve freedom freedom clearly understand new nature americas role result new policies adopted past four years shall respect treaty commitments shall support vigorously principle country right impose rule another force shall continue era negotiation work limitation nuclear arms reduce danger confrontation great powers shall share defending peace freedom world shall expect others share time passed america make every nations conflict make every nations future responsibility presume tell people nations manage affairs respect right nation determine future also recognize responsibility nation secure future americas role indispensable preserving worlds peace nations role indispensable preserving peace together rest world resolve move forward beginnings made continue bring walls hostility divided world long build place bridges understanding despite profound differences systems government people world friends build structure peace world weak safe strong respects right live different system would influence others strength ideas force arms accept high responsibility burden gladly gladly chance build peace noblest endeavor nation engage gladly also act greatly meeting responsibilities abroad remain great nation remain great nation act greatly meeting challenges home chance today ever history make life better america ensure better education better health better housing better transportation cleaner environment restore respect law make communities livable insure godgiven right every american full equal opportunity range needs great reach opportunities great bold determination meet needs new ways building structure peace abroad required turning away old policies failed building new era progress home requires turning away old policies failed abroad shift old policies new retreat responsibilities better way peace home shift old policies new retreat responsibilities better way progress abroad home key new responsibilities lies placing division responsibility lived long consequences attempting gather power responsibility washington abroad home time come turn away condescending policies paternalism washington knows best person expected act responsibly responsibility human nature encourage individuals home nations abroad decide locate responsibility places measure others today offer promise purely governmental solution every problem lived long false promise trusting much government asked deliver leads inflated expectations reduced individual effort disappointment frustration erode confidence government people government must learn take less people people remember america built government people welfare work shirking responsibility seeking responsibility lives ask government challenges face together ask government help help national government great vital role play pledge government act act boldly lead boldly important role every one must play individual member community day forward make solemn commitment heart bear responsibility part live ideals together see dawn new age progress america together celebrate 200th anniversary nation proud fulfillment promise world americas longest difficult war comes end learn debate differences civility decency reach one precious quality government cannot provide new level respect rights feelings one another new level respect individual human dignity cherished birthright every american else time come renew faith america recent years faith challenged children taught ashamed country ashamed parents ashamed americas record home role world every turn beset find everything wrong america little right confident judgment history remarkable times privileged live americas record century unparalleled worlds history responsibility generosity creativity progress proud system produced provided freedom abundance widely shared system history world proud four wars engaged century including one bringing end fought selfish advantage help others resist aggression proud bold new initiatives steadfastness peace honor made breakthrough toward creating world world known structure peace last merely time generations come embarking today era presents challenges great nation generation ever faced shall answer god history conscience way use years stand place hallowed history think others stood think dreams america think recognized needed help far beyond order make dreams come true today ask prayers years ahead may gods help making decisions right america pray help together may worthy challenge pledge together make next four years best four years americas history 200th birthday america young vital began bright beacon hope world go forward confident hope strong faith one another sustained faith god created striving always serve purpose*

**Kennedy’s Speech:**

*vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change sworn almighty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands power abolish forms human poverty forms human life yet revolutionary beliefs forebears fought still issue around globe belief rights man come generosity state hand god dare forget today heirs first revolution word go forth time place friend foe alike torch passed new generation americans born century tempered war disciplined hard bitter peace proud ancient heritage unwilling witness permit slow undoing human rights nation always committed committed today home around world every nation know whether wishes well ill shall pay price bear burden meet hardship support friend oppose foe order assure survival success liberty much pledge old allies whose cultural spiritual origins share pledge loyalty faithful friends united little cannot host cooperative ventures divided little dare meet powerful challenge odds split asunder new states welcome ranks free pledge word one form colonial control shall passed away merely replaced far iron tyranny shall always expect find supporting view shall always hope find strongly supporting freedom remember past foolishly sought power riding back tiger ended inside peoples huts villages across globe struggling break bonds mass misery pledge best efforts help help whatever period required communists may seek votes right free society cannot help many poor cannot save rich sister republics south border offer special pledge convert good words good deeds new alliance progress assist free men free governments casting chains poverty peaceful revolution hope cannot become prey hostile powers neighbors know shall join oppose aggression subversion anywhere americas every power know hemisphere intends remain master house world assembly sovereign states united nations last best hope age instruments war far outpaced instruments peace renew pledge supportto prevent becoming merely forum invective strengthen shield new weak enlarge area writ may run finally nations would make adversary offer pledge request sides begin anew quest peace dark powers destruction unleashed science engulf humanity planned accidental selfdestruction dare tempt weakness arms sufficient beyond doubt certain beyond doubt never employed neither two great powerful groups nations take comfort present course sides overburdened cost modern weapons rightly alarmed steady spread deadly atom yet racing alter uncertain balance terror stays hand mankinds final war begin anew remembering sides civility sign weakness sincerity always subject proof never negotiate fear never fear negotiate sides explore problems unite instead belaboring problems divide sides first time formulate serious precise proposals inspection control arms bring absolute power destroy nations absolute control nations sides seek invoke wonders science instead terrors together explore stars conquer deserts eradicate disease tap ocean depths encourage arts commerce sides unite heed corners earth command isaiah undo heavy burdens oppressed go free beachhead cooperation may push back jungle suspicion sides join creating new endeavor new balance power new world law strong weak secure peace preserved finished first 100 days finished first 1000 days life administration even perhaps lifetime planet begin hands fellow citizens mine rest final success failure course since country founded generation americans summoned give testimony national loyalty graves young americans answered call service surround globe trumpet summons call bear arms though arms need call battle though embattled call bear burden long twilight struggle year year rejoicing hope patient tribulation struggle common enemies man tyranny poverty disease war forge enemies grand global alliance north south east west assure fruitful life mankind join historic effort long history world generations granted role defending freedom hour maximum danger shrink responsibility welcome believe would exchange places people generation energy faith devotion bring endeavor light country serve glow fire truly light world fellow americans ask country ask country fellow citizens world ask america together freedom man finally whether citizens america citizens world ask high standards strength sacrifice ask good conscience sure reward history final judge deeds go forth lead land love asking blessing help knowing earth gods work must truly own*

### 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) – 3 Marks

Most common 10 words used by each president are as follows,

**Words occurring most number of times in Roosevelt’s speech:**

('nation', 11),

('know', 10),

('spirit', 9),

('democracy', 9),

('life', 8),

('people', 7),

('america', 7),

('years', 6),

('freedom', 6),

('human', 5)

**Words occurring most number of times in Nixon’s speech:**

('peace', 19),

('world', 16),

('new', 15),

('america', 13),

('responsibility', 11),

('government', 10),

('great', 9),

('home', 9),

('abroad', 8),

('nation', 8)

**Words occurring most number of times in Kennedy’s speech:**

('world', 8),

('sides', 8),

('new', 7),

('pledge', 7),

('citizens', 5),

('power', 5),

('shall', 5),

('free', 5),

('nations', 5),

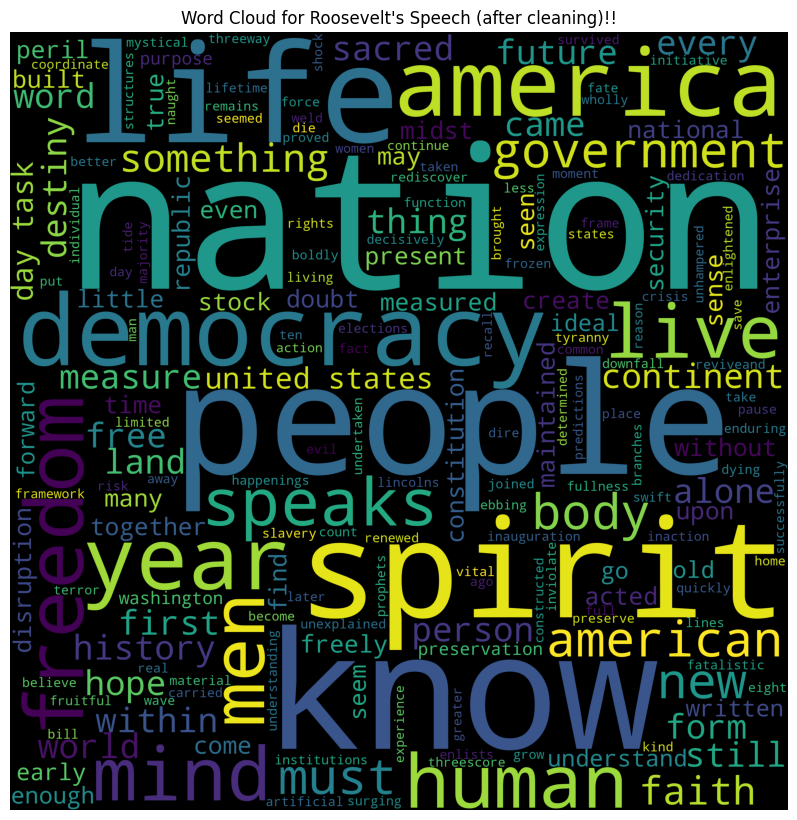
('ask', 5)

### 2.4 Plot the word cloud of each of the speeches of the variable.

Below is the word cloud to visualize all the words used in speech by all each president.

Note: This word cloud is after stopword removal

Word cloud for Roosevelt’s Speech:



Word cloud for Nixon’s Speech:

A black background with words

Description automatically generated

Word cloud for Kennedy’s Speech:

A black background with words

Description automatically generated

**Glossary**

Fig 1-Hist Plots ……………………………………………………………………………………Page 6

Fig 2- Heat Map (Co-relation Plot) …………………………………………………………. Page 7

Fig 3 - Box Plots …………………………………………………………………………………. Page 9

Fig 4- Vote distribution choice among people in different age groups……………. Page 10

Fig 5 – Distribution of Votes (Hist Plots) ……………………………………………………...Page 11

Fig 6 – Roc Curve on LR, LDR, Bagging Classifier and ADA Boost Models …………. Page 20