ML-2 Project

(Coded)

DSBA

By:

E. AuroRajashri

List of Content

Problem 1

Context

CNBE, a prominent news channel, is gearing up to provide insightful coverage of recent elections, recognizing the importance of data-driven analysis. A comprehensive survey has been conducted, capturing the perspectives of 1525 voters across various demographic and socio-economic factors. This dataset encompasses 9 variables, offering a rich source of information regarding voters' characteristics and preferences.

Objective

The primary objective is to leverage machine learning to build a predictive model capable of forecasting which political party a voter is likely to support. This predictive model, developed based on the provided information, will serve as the foundation for creating an exit poll. The exit poll aims to contribute to the accurate prediction of the overall election outcomes, including determining which party is likely to secure the majority of seats.

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

1.1   Define the problem and perform Exploratory Data Analysis

1.1.1 Problem Definition

- Imported necessary libraries like NumPy, Pandas,matplotlib,seaborn.
- Loaded the given dataset to dataframe election

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1	Labour	43	3	3	4	1	2	2	female
2	Labour	36	4	4	4	4	5	2	male
3	Labour	35	4	4	5	2	3	2	male
4	Labour	24	4	2	2	1	4	0	female
5	Labour	41	2	2	1	1	6	2	male
6	Labour	47	3	4	4	4	4	2	male
7	Labour	57	2	2	4	4	11	2	male
8	Labour	77	3	4	4	1	1	0	male
9	Labour	39	3	3	4	4	11	0	female
10	Labour	70	3	2	5	1	11	2	male

Fig 1: Dataset Head rows

1.1.2 Check shape, Data types, statistical summary

- Dataset has shape of 1525 rows and 9 columns. And it has 7 integer datatypes and 2 object datatypes.

```
(1525, 9)
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1525 entries, 1 to 1525
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: 119.1+ KB
```

Fig 2: Dataset Info

- Below is the dataset statistical Summary

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

Fig 3: Dataset Statistical Summary

- There are 9 duplicates in the dataset as shown below and all got dropped as it has no meaning to the analysis. So finally, dataset has 1517 rows and 9 columns

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
68	35	4	4	5	2	3	2	1
627	39	3	4	4	2	5	2	1
871	38	2	4	2	2	4	3	1
984	74	4	3	2	4	8	2	0
1155	53	3	4	2	2	6	0	0
1185	61	3	3	4	2	6	0	0
1237	36	3	3	2	2	6	2	0
1245	29	4	4	4	2	2	2	0
1439	40	4	3	4	2	2	2	1

Fig 4: Dataset duplicates

1.1.3 Univariate analysis and Bivariate analysis

- **Univariate analysis**
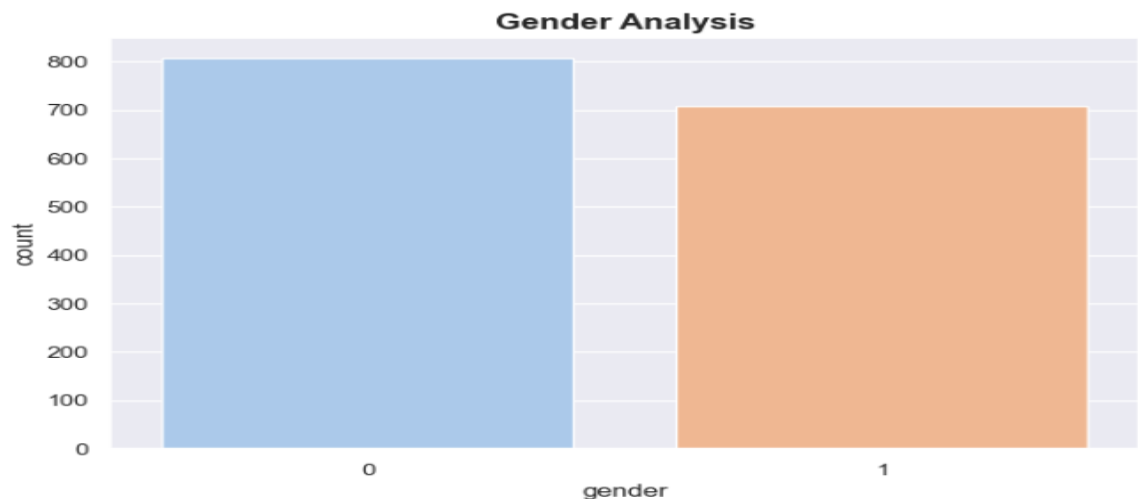
Fig 4: Gender Analysis

Fig 5: Age

5

Blair Analysis   Hague Analysis   political knowledges Analysis
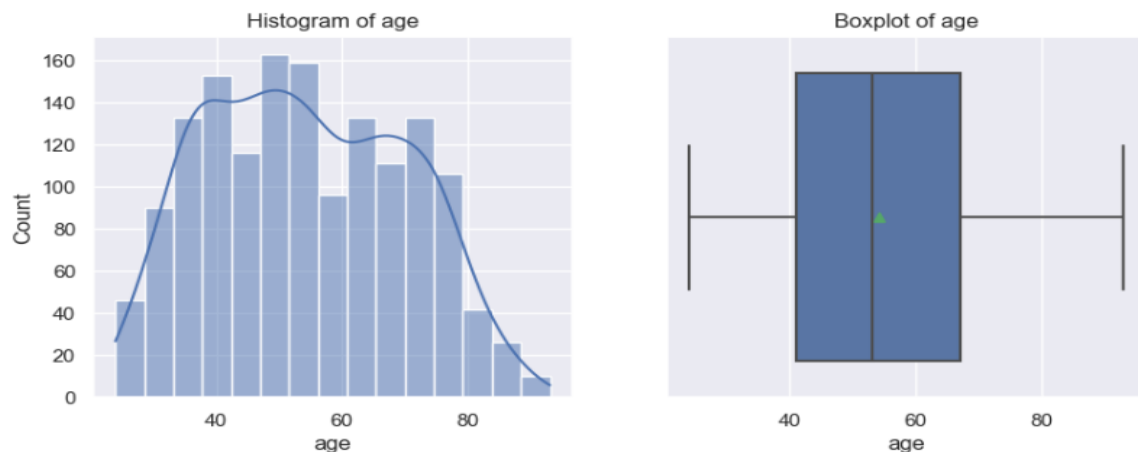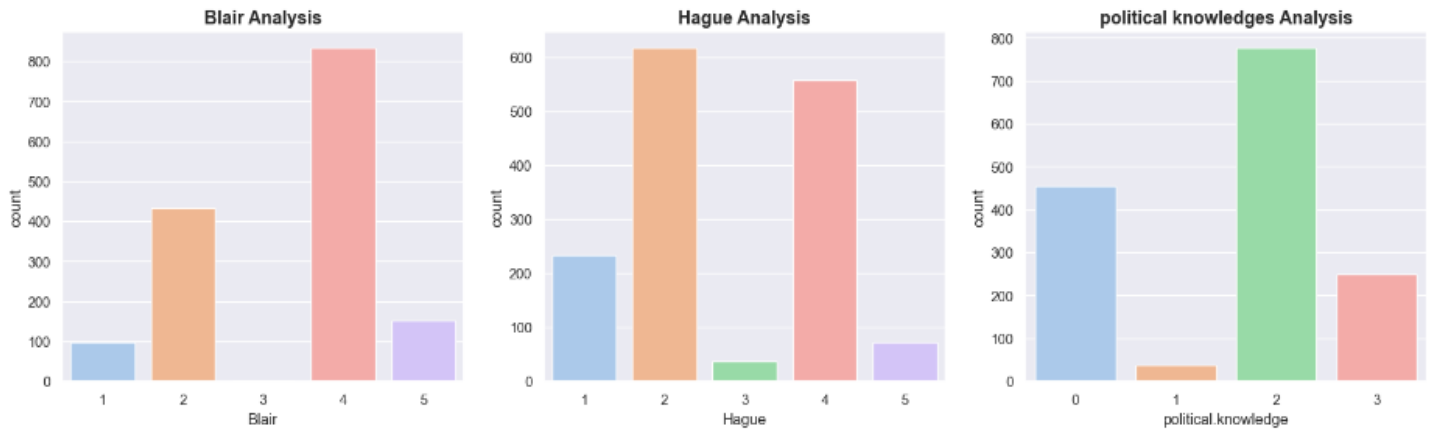
1. 53% percent of voters are female whereas, 47% are men,
2. Age variable is very slightly right skewed with the presence of no outliers. The variable ranges from 24 to 93.
3. Most of the people have rated Blair good implying that people are happy with labor party.
4. Nearly half of the people have rated Hague poorly.

- **Bivariate analysis**

Fig 8: Pair plot of numeric variables

1. As per the visulisation we can see here is a slight positive correlation between the ratings of economic conditions of the nation & the households. This slight positive correlation exists with the ratings of labour party leader, Tony Blair as well and there is a slight negative correlation with the conservative party leader, William Hague. Implying people are generally happy with the current economic conditions and would like Labour party to continue.

2. Blair & Hague have a weak negative correlation between them as is obvious as they are standing against each other in the general election.
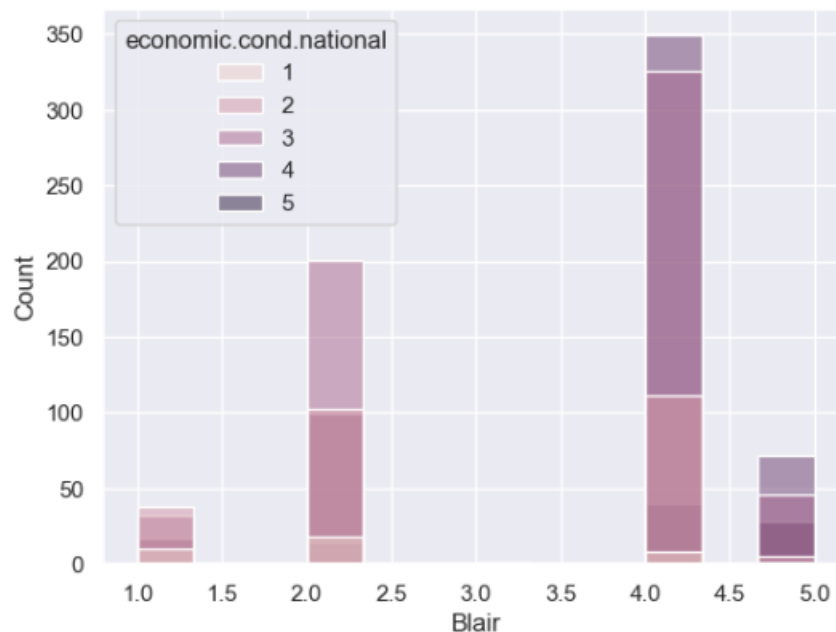
1.2   Data Preprocessing
1.2.1 Outlier treatment

- We were only interested in the outliers for the age variable since other numerical features are ordinal in nature. There are no outliers in the age variable.
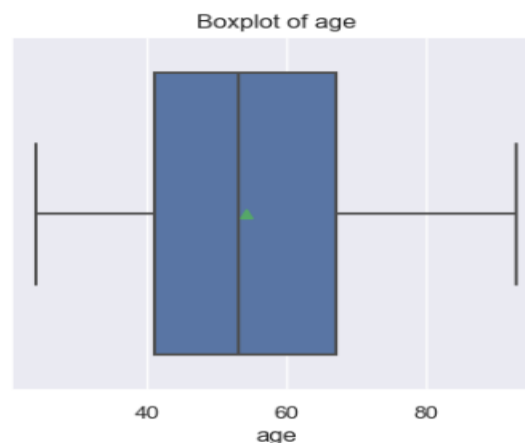

Fig 11: Boxplot of Age

1.2.2 **Encode the data**

Here, gender variable and vote need to be encoded since all the other independent features are numerical in nature, so we opt for One hot encoding with dropping the dummy variable.

```
gender                          vote
0    812                        0    1063
1    713                        1     462
Name: count, dtype: int64      Name: count, dtype: int64
```

Fig 13: One hot encoding

1.2.3 **scaling**

- We have to scale the numerical variables as distance based algorithms like KNN will give highly inaccurate results.

After scaling:

```
array([[0.        , 0.27536232, 0.5       , ..., 0.1       , 0.66666667,
        0.        ],
       [0.        , 0.17391304, 0.75      , ..., 0.4       , 0.66666667,
        1.        ],
       [0.        , 0.15942029, 0.75      , ..., 0.2       , 0.66666667,
        1.        ],
       ...,
       [0.        , 0.1884058 , 0.5       , ..., 0.1       , 0.66666667,
        1.        ],
       [1.        , 0.53623188, 0.5       , ..., 1.        , 0.66666667,
        1.        ],
       [1.        , 0.72463768, 0.25      , ..., 1.        , 0.        ,
        0.        ]])
```

Fig 16: after scaling

1.2.4 **Data Split**

- Data splitted into train and test data in .30 size
- From Sklearn model selection library imported train test split

1.3 Model Building

1.3.1 Metrics of choice

1) Logistic regression

```
▾          LogisticRegression
LogisticRegression(max_iter=2000, random_state=0)
```

```
              precision    recall  f1-score   support

           0       0.86      0.91      0.88       754
           1       0.74      0.64      0.69       307

    accuracy                           0.83      1061
   macro avg       0.80      0.77      0.79      1061
weighted avg       0.83      0.83      0.83      1061
```
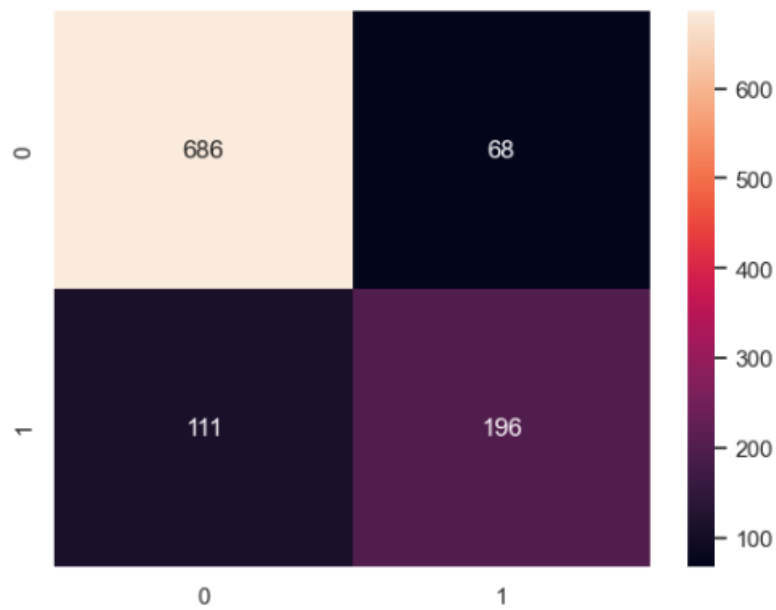
```
[[268  35]
 [ 40 113]]
                precision    recall  f1-score   support

           0        0.87      0.88      0.88       303
           1        0.76      0.74      0.75       153

    accuracy                            0.84       456
   macro avg        0.82      0.81      0.81       456
weighted avg        0.83      0.84      0.83       456
```

ROC AND AUC for training data
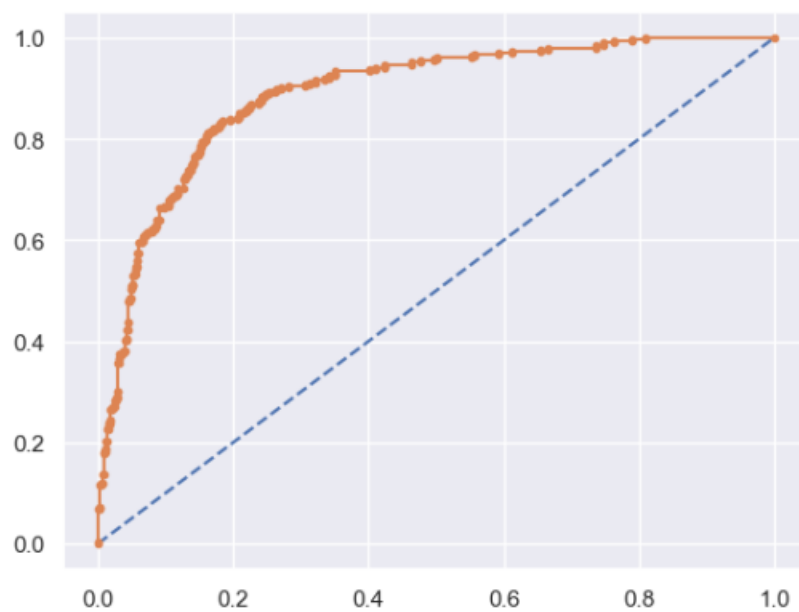
ROC AND AUC for test data

AUC: 0.883

2) LDA

Train data:

```
[[685 107]
 [ 69 200]]
              precision    recall  f1-score   support

           0       0.91      0.86      0.89       792
           1       0.65      0.74      0.69       269

    accuracy                           0.83      1061
   macro avg       0.78      0.80      0.79      1061
weighted avg       0.84      0.83      0.84      1061
```

Test data:

```
[[269  42]
 [ 34 111]]
              precision    recall  f1-score   support

           0       0.89      0.86      0.88       311
           1       0.73      0.77      0.74       145

    accuracy                           0.83       456
   macro avg       0.81      0.82      0.81       456
weighted avg       0.84      0.83      0.83       456
```
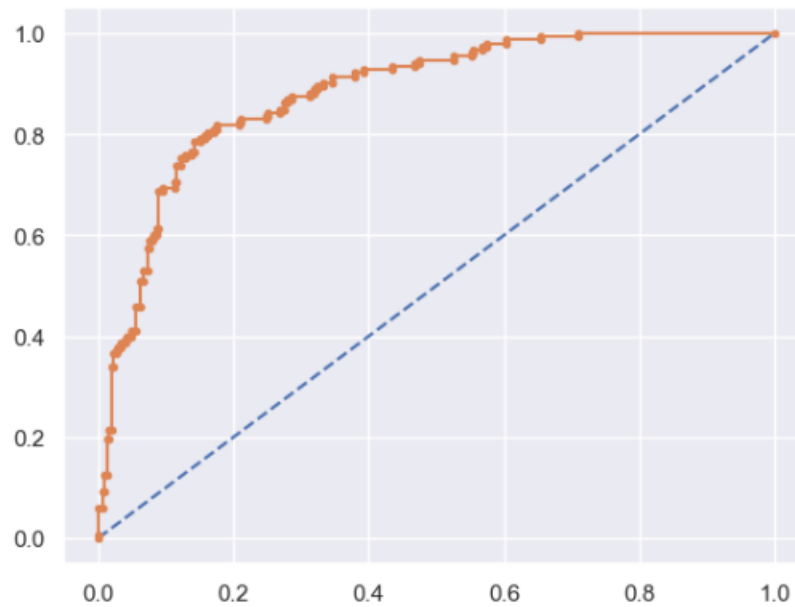

KNN model:

```
              precision    recall  f1-score   support

           0       0.91      0.84      0.87       327
           1       0.66      0.78      0.72       129

    accuracy                           0.82       456
   macro avg       0.78      0.81      0.79       456
weighted avg       0.84      0.82      0.83       456
```
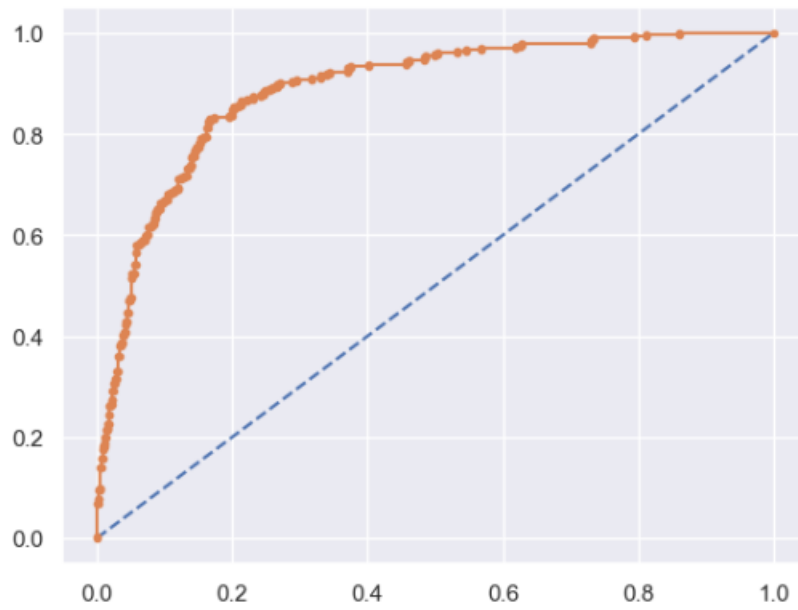
Training data:

Test data:

AUC: 0.861

Naive bayes
Train data

```
[[675  96]
 [ 79 211]]
              precision    recall  f1-score   support

           0       0.90      0.88      0.89       771
           1       0.69      0.73      0.71       290

    accuracy                           0.84      1061
   macro avg       0.79      0.80      0.80      1061
weighted avg       0.84      0.84      0.84      1061
```
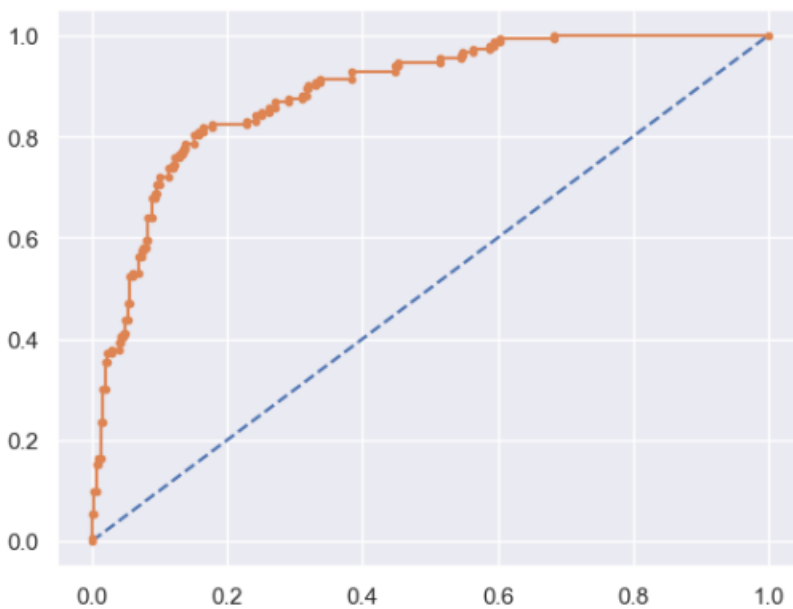

Test data

```
[[263  41]
 [ 40 112]]
              precision    recall  f1-score   support

           0       0.87      0.87      0.87       304
           1       0.73      0.74      0.73       152

    accuracy                           0.82       456
   macro avg       0.80      0.80      0.80       456
weighted avg       0.82      0.82      0.82       456
```
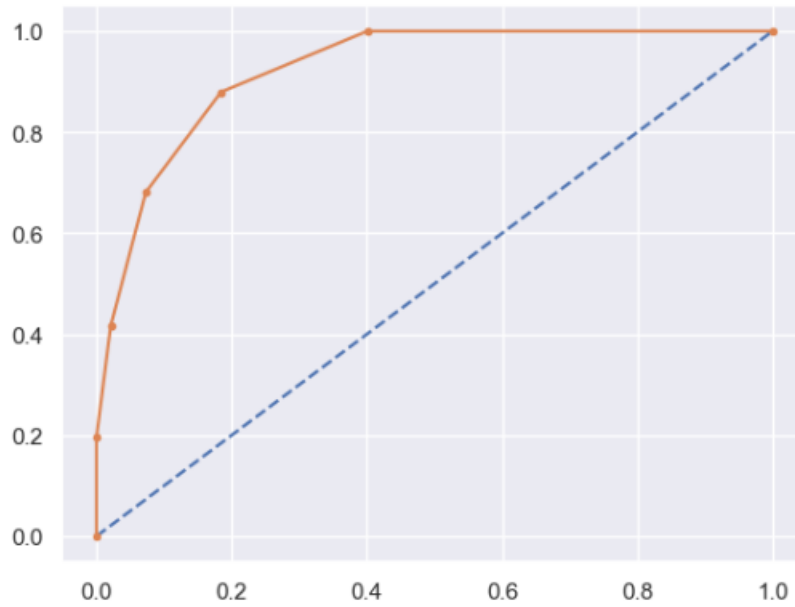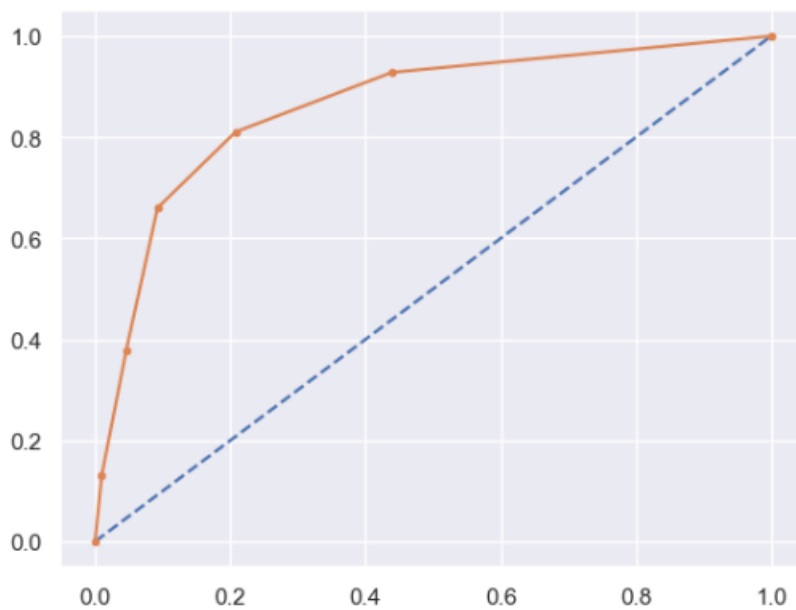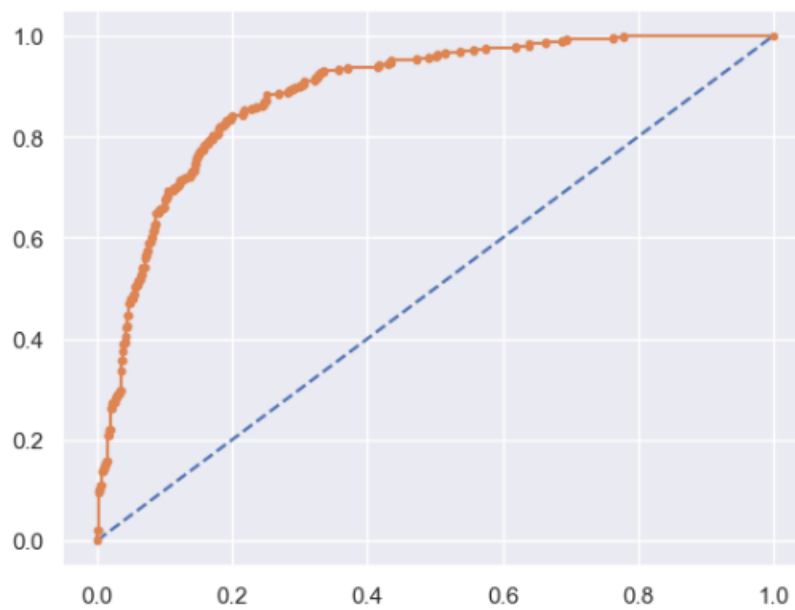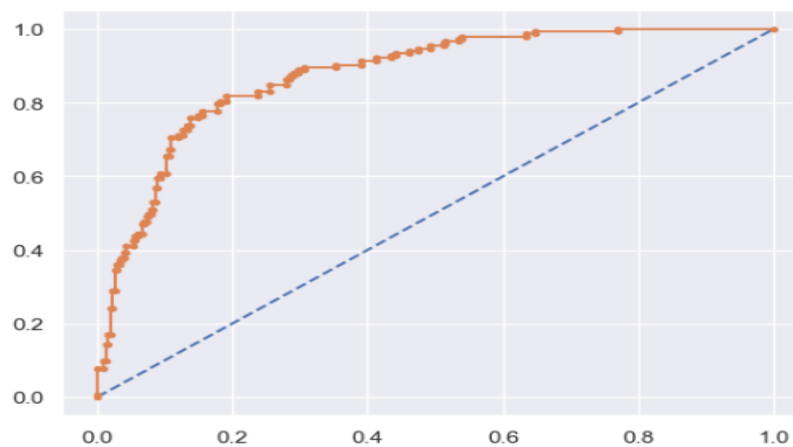
Model Tuning, Bagging (Random Forest applied for Bagging) and Boosting

```
▸            GridSearchCV
▸ estimator: AdaBoostClassifier
    ▸ AdaBoostClassifier
```

```
0.8369462770970783
[[702  52]
 [121 186]]
              precision    recall  f1-score   support

           0       0.85      0.93      0.89       754
           1       0.78      0.61      0.68       307

    accuracy                           0.84      1061
   macro avg       0.82      0.77      0.79      1061
weighted avg       0.83      0.84      0.83      1061
```

```
0.8092105263157895
[[271  32]
 [ 55  98]]
              precision    recall  f1-score   support

           0       0.83      0.89      0.86       303
           1       0.75      0.64      0.69       153

    accuracy                           0.81       456
   macro avg       0.79      0.77      0.78       456
weighted avg       0.81      0.81      0.80       456
```
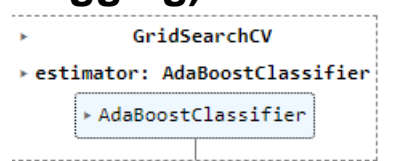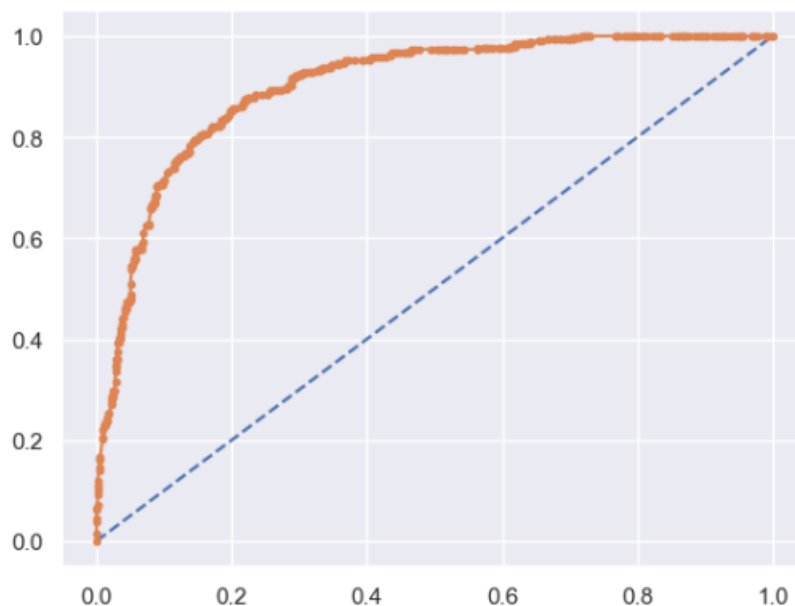
AUC: 0.884

Random forest:

```
0.8576814326107446
[[708  46]
 [105 202]]
              precision    recall  f1-score   support

           0       0.87      0.94      0.90       754
           1       0.81      0.66      0.73       307

    accuracy                           0.86      1061
   macro avg       0.84      0.80      0.82      1061
weighted avg       0.85      0.86      0.85      1061
```
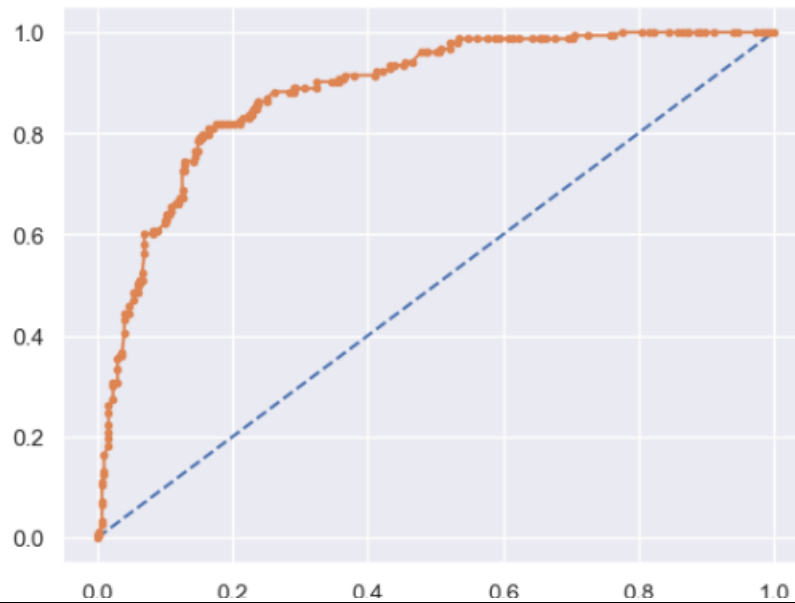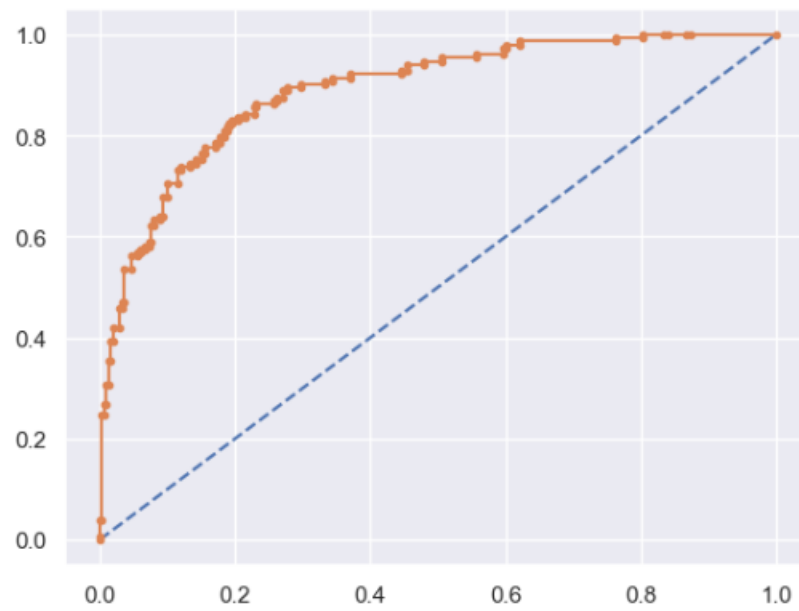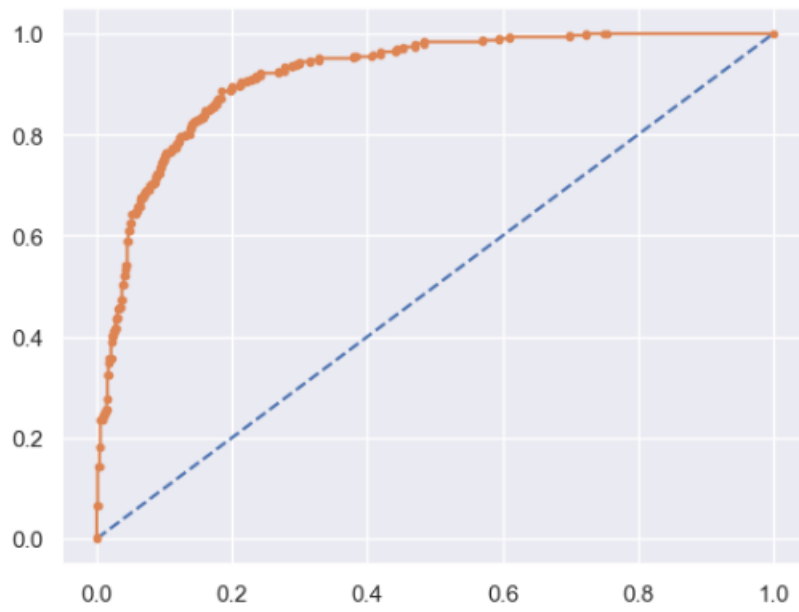
Problem 2

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

Code Snippet to extract the three speeches:

```
"
import nltk
nltk.download('inaugural')
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.raw('1941-Roosevelt.txt')
inaugural.raw('1961-Kennedy.txt')
inaugural.raw('1973-Nixon.txt')
"
```

2.1 Define the problem

2.1.1 Problem Definition

- Imported the NLTK library and downloads the 'inaugural' corpus using nltk.download('inaugural').
- Imported the 'inaugural' corpus from NLTK using from nltk.corpus import inaugural
- Retrieved the raw text content of the speeches given by three different presidents

2.1.2 Find the number of Character, words & sentences in all three speeches

1. Number of Characters in each file
- The number of characters in each file is found by using len function

```
Number of characters in Roosevelt file : 7571
Number of characters in Kennedy file : 7618
Number of characters in Nixon file : 9991
```

2. Number of words in each text file:
- Using the split() to split up the words based on space between each word and counting the total number of words by using the len() function.

```
Number of words in Kennedy file : 1390
Number of words in Nixon file : 1819
Number of words in Roosevelt file : 1360
```

3. Number of Sentences.
- Below we are counting the total number of sentence in each text file, by using lambda function.
- Using pd.Dataframe to move the data as dictionary and then with lambda function we are checking each sentece which ends with ".".

	Text	sentences
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	68

	Text	sentences
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	52

	Text	sentences
0	On each national day of inauguration since 178...	67

2.2 Text Cleaning

2.2.1 Stop word removal - Stemming

- Imported stopwords from nltk.corpus and imported word_tokenize from nltk.tokenize.
- Imported PorterStemmer from nltk.stem
- We need these to remove all the English predefined words from each text file separately and with the help of tokenize we would separate each word and remove all the words from the text file.
- Output of stop word removal and stemming of Roosevelt 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', 'th
e', 'people', 'was', 'to', 'create', 'and', 'weld', 'together', 'a', 'nation', '.', 'In', 'Lincoln', "'s", 'day', 'the', 'ta
sk', 'of', 'the', 'people', 'was', 'to', 'preserve', 'that', 'Nation', 'from', 'disruption', 'from', 'within', '.', 'In', 't
his', 'day', 'the', 'task', 'of', 'the', 'people', 'is', 'to', 'save', 'that', 'Nation', 'and', 'its', 'institutions', 'fro
m', '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', 'ou
r', 'place', 'in', 'history', 'has', 'been', ',', 'and', 'to', 'rediscover', 'what', 'we', 'are', 'and', 'what', 'we', 'ma
y', 'be', '.', 'If', 'we', 'do', 'not', ',', 'we', 'risk', 'the', 'real', 'peril', 'of', 'inaction', '.', 'Lives', 'of', 'na
tions', 'are', 'determined', 'not', 'by', 'the', 'count', 'of', 'years', ',', 'but', 'by', 'the', 'lifetime', 'of', 'the',
'human', 'spirit', '.', 'The', 'life', 'of', 'a', 'man', 'is', 'three-score', 'years', 'and', 'ten', ':', 'a', 'little', 'mo
re', ',', 'a', 'little', 'less', '.', 'The', 'life', 'of', 'a', 'nation', 'is', 'the', 'fullness', 'of', 'the', 'measure',
'of', 'its', 'will', 'to', 'live', '.', 'There', 'are', 'men', 'who', 'doubt', 'this', '.', 'There', 'are', 'men', 'who', 'b
elieve', 'that', 'democracy', ',', 'as', 'a', 'form', 'of', 'Government', 'and', 'a', 'frame', 'of', 'life', ',', 'is', 'lim
ited', 'or', 'measured', 'by', 'a', 'kind', 'of', 'mystical', 'and', 'artificial', 'fate', 'that', ',', 'for', 'some', 'unex
plained', 'reason', ',', 'tyranny', 'and', 'slavery', 'have', 'become', 'the', 'surging', 'wave', 'of', 'the', 'future', '--
', 'and', 'that', 'freedom', 'is', 'an', 'ebbing', 'tide', '.', 'But', 'we', 'Americans', 'know', 'that', 'this', 'is', 'no
t', 'true', '.', 'Eight', 'years', 'ago', ',', 'when', 'the', 'life', 'of', 'this', 'Republic', 'seemed', 'frozen', 'by',
'a', 'fatalistic', 'terror', ',', 'we', 'proved', 'that', 'this', 'is', 'not', 'true', '.', 'We', 'were', 'in', 'the', 'mids
```

- Output of stop word removal and stemming of Kennedy 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', '--', 'symbolizin
g', 'an', 'end', ',', 'as', 'well', 'as', 'a', 'beginning', '--', 'signifying', 'renewal', ',', 'as', 'well', 'as', 'chang
e', '.', 'For', 'I', 'have', 'sworn', 'I', 'before', 'you', 'and', 'Almighty', 'God', 'the', 'same', 'solemn', 'oath', 'ou
r', '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', 'al
l', 'forms', 'of', 'human', 'poverty', 'and', 'all', 'forms', 'of', 'human', 'life', '.', 'And', 'yet', 'the', 'same', 'revo
lutionary', 'beliefs', 'for', 'which', 'our', 'forebears', 'fought', 'are', 'still', 'at', 'issue', 'around', 'the', 'glob
e', '--', 'the', 'belief', 'that', 'the', 'rights', 'of', 'man', 'come', 'not', 'from', 'the', 'generosity', 'of', 'the', 's
tate', ',', 'but', 'from', 'the', 'hand', 'of', 'God', '.', 'We', 'dare', 'not', 'forget', 'today', 'that', 'we', 'are', 'th
e', 'heirs', 'of', 'that', 'first', 'revolution', '.', 'Let', 'the', 'word', 'go', 'forth', 'from', 'this', 'time', 'and',
'place', ',', 'to', 'friend', 'and', 'foe', 'alike', ',', 'that', 'the', 'torch', 'has', 'been', 'passed', 'to', 'a', 'new',
'generation', 'of', 'Americans', '--', 'born', 'in', 'this', 'century', ',', 'tempered', 'by', 'war', ',', 'disciplined', 'b
y', 'a', 'hard', 'and', 'bitter', 'peace', ',', 'proud', 'of', 'our', 'ancient', 'heritage', '--', 'and', 'unwilling', 'to',
'witness', 'or', 'permit', 'the', 'slow', 'undoing', 'of', 'those', 'human', 'rights', 'to', 'which', 'this', 'Nation', 'ha
s', 'always', 'been', 'committed', ',', 'and', 'to', 'which', 'we', 'are', 'committed', 'today', 'at', 'home', 'and', 'aroun
d', 'the', 'world', '.', 'Let', 'every', 'nation', 'know', ',', 'whether', 'it', 'wishes', 'us', 'well', 'or', 'ill', ',',
'that', 'we', 'shall', 'pay', 'any', 'price', ',', 'bear', 'any', 'burden', ',', 'meet', 'any', 'hardship', ',', 'support',
```

- Output of stop word removal and stemming of Nixon speech:

```
['Mr.', 'Vice', 'President', ',', 'Mr.', 'Speaker', ',', 'Mr.', 'Chief', 'Justice', ',', 'Senator', 'Cook', ',', 'Mrs.', 'Ei
senhower', ',', 'and', 'my', 'fellow', 'citizens', 'of', 'this', 'great', 'and', 'good', 'country', 'we', 'share', 'togethe
r', ':', 'When', 'we', 'met', 'here', 'four', 'years', 'ago', ',', 'America', 'was', 'bleak', 'in', 'spirit', ',', 'depresse
d', 'by', 'the', 'prospect', 'of', 'seemingly', 'endless', 'war', 'abroad', 'and', 'of', 'destructive', 'conflict', 'at', 'h
ome', '.', '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', 't
hat', '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', '.', 'Let', 'us', 'resolve',
'that', 'this', 'will', 'be', 'what', 'it', 'can', 'become', ':', 'a', 'time', 'of', 'great', 'responsibilities', 'greatly',
'borne', ',', 'in', 'which', 'we', 'renew', 'the', 'spirit', 'and', 'the', 'promise', 'of', 'America', 'as', 'we', 'enter',
'our', 'third', 'century', 'as', 'a', 'nation', '.', 'This', 'past', 'year', 'saw', 'far-reaching', 'results', 'from', 'ou
r', 'new', 'policies', 'for', 'peace', '.', 'By', 'continuing', 'to', 'revitalize', 'our', 'traditional', 'friendships',
',', 'and', 'by', 'our', 'missions', 'to', 'Peking', 'and', 'to', 'Moscow', ',', 'we', 'were', 'able', 'to', 'establish', 't
he', 'base', 'for', 'a', 'new', 'and', 'more', 'durable', 'pattern', 'of', 'relationships', 'among', 'the', 'nations', 'of',
'the', 'world', '.', 'Because', 'of', 'America', "'s", 'bold', 'initiatives', ',', '1972', 'will', 'be', 'long', 'remembere
d', 'as', 'the', 'year', 'of', 'the', 'greatest', 'progress', 'since', 'the', 'end', 'of', 'World', 'War', 'II', 'toward',
'a', 'lasting', 'peace', 'in', 'the', 'world', '.', 'The', 'peace', 'we', 'seek', 'in', 'the', 'world', 'is', 'not', 'the',
'flimsy', 'peace', 'which', 'is', 'merely', 'an', 'interlude', 'between', 'wars', ',', 'but', 'a', 'peace', 'which', 'can',
```

2.2.2 Three most common words used in all three speeches

- Using loop, finding out the most common words used in the speeches
- Roosevelt's speech: Nation, Spirit, Life

```
[('Nation', 12),
 ('Spirit', 9),
 ('Life', 9),
 ('Democracy', 9),
 ('America', 7),
 ('Years', 6),
 ('Freedom', 6),
 ('Human', 5),
 ('Body', 5),
 ('Mind', 5),
 ('Speaks', 5),
 ('Day', 4),
 ('States', 4),
 ('Government', 4),
 ('Faith', 4),
 ('United', 3),
 ('Task', 3),
 ('History', 3),
 ('Nations', 3),
```

- Kennedy's speech: World, Sides, Pledge

```
[('World', 8),
 ('Sides', 8),
 ('Pledge', 7),
 ('Citizens', 5),
 ('Power', 5),
 ('Free', 5),
 ('Nations', 5),
 ('President', 4),
 ('Fellow', 4),
 ('Freedom', 4),
 ('Americans', 4),
 ('Peace', 4),
 ('Hope', 4),
 ('Arms', 4),
 ('Country', 4),
 ('Call', 4),
 ('Today', 3),
 ('God', 3),
 ('Human', 3),
```

- Nixon's speech: America, Peace, World

```
[('America', 21),
 ('Peace', 19),
 ('World', 18),
 ('Nation', 11),
 ('Responsibility', 11),
 ('Government', 10),
 ('Great', 9),
 ('Home', 9),
 ('Abroad', 8),
 ('Years', 7),
 ('Policies', 7),
 ('Role', 7),
 ('Make', 7),
 ('History', 7),
 ('Time', 6),
 ('Today', 5),
 ('Responsibilities', 5),
 ('Progress', 5),
 ('Respect', 5),
```

2.3 Plot Word cloud of all three speeches

2.3.1 most common words used in all three speeches in the form of word clouds

- Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud.
- making sure to creating stopword list and cleaning the text.
- Imported Image froom PIL library
- Below is the word cloud of Roosevelt's speech: cloud mask
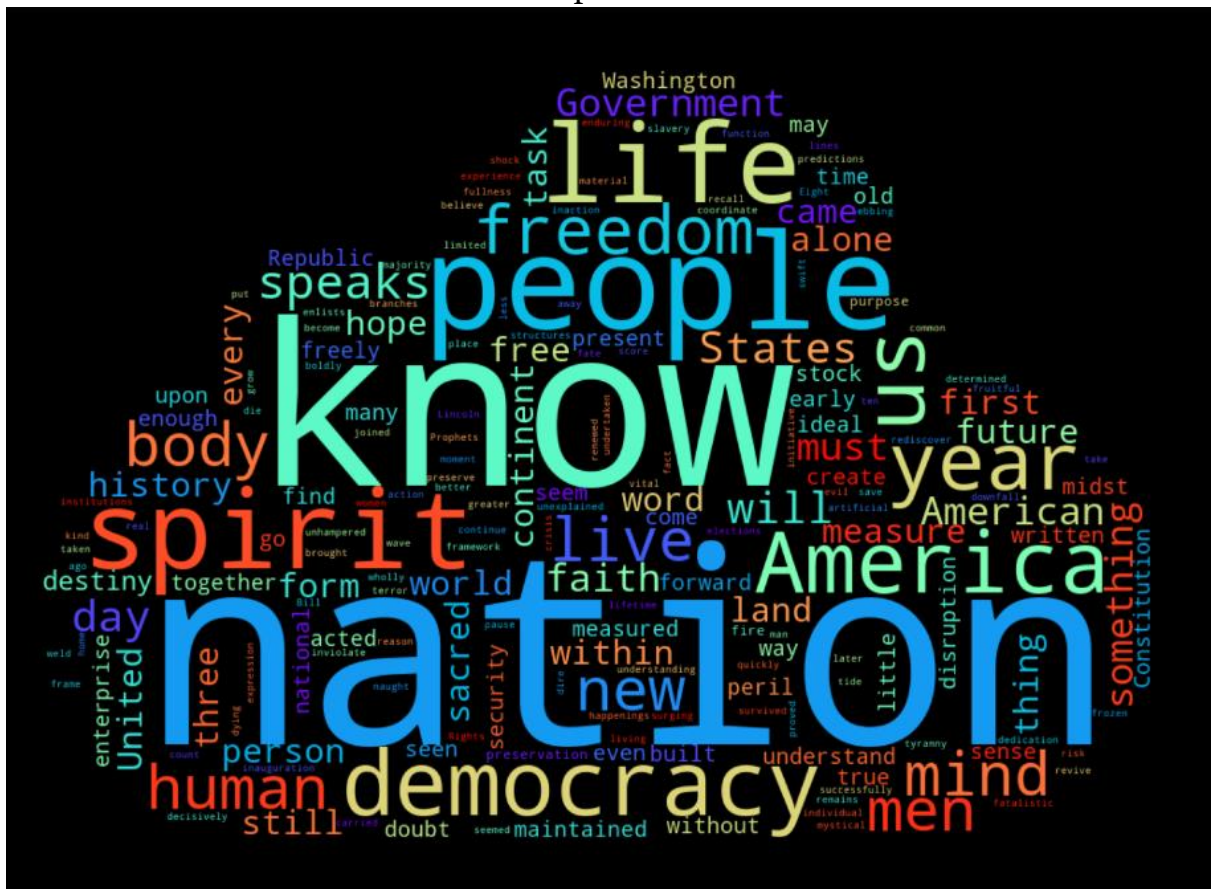
- Below is the word cloud of Kennedy's speech:

- Below is the word cloud of Nixon's speech: