PGP - Data Science and Business Analytics

MACHINE LEARNING PROJECT REPORT

Akshaya Nallathambi

3rd October, 2021

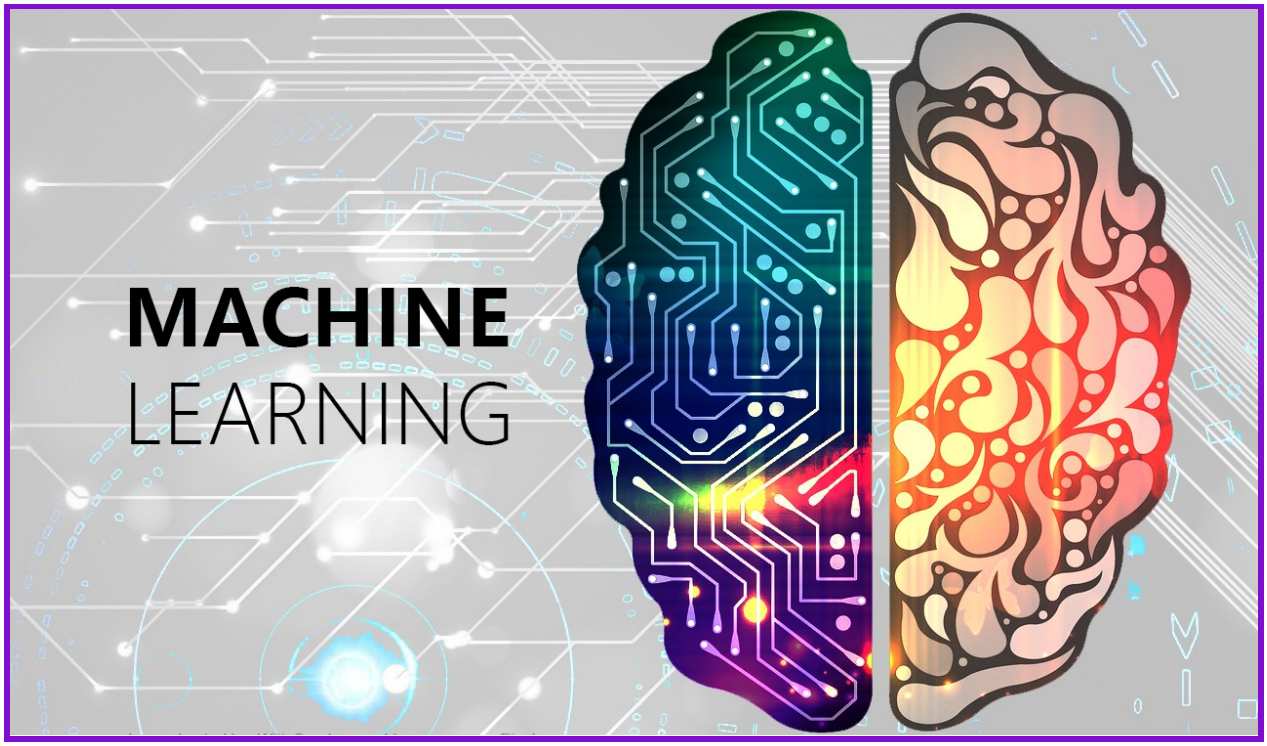


Table Of Contents

Problem 1

Problem 2

List of Figures

List of Tables

List of Graphs

Problem 1

Problem statement-

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 Description-

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.

Sample of the dataset-

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

FIGURE 1

There are 9 variables out of which 7 are int values and 2 are object values . The data given is for 1525 individuals. The dataset does not contain any null values.

Types of variables in the data frame-

vote	object	Categorical
age	int64	Continuous
economic.cond.national	int64	Continuous
economic.cond.household	int64	Continuous
Blair	int64	Continuous
Hague	int64	Continuous
Europe	int64	Continuous
political.knowledge	int64	Continuous
gender	object	Categorical

TABLE 1

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

The below table gives the first 5 rows of sample data.

vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
Labour	43	3	3	4	1	2	2	female
Labour	36	4	4	4	4	5	2	male
Labour	35	4	4	5	2	3	2	male
Labour	24	4	2	2	1	4	0	female
Labour	41	2	2	1	1	6	2	male

FIGURE 2

The image below gives the basic information of the data set. It is clear that the variables have int and object data types with 9 columns and 1525 rows. The dataset has no null values but consists of 8 duplicates which have been removed. The memory usage is 107.4+ KB.

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

FIGURE 3

The below image gives the five point summary of the continuous variables in the data set. It is clear that the given data needs scaling as the numbers are of different magnitude. There is skewness in the data when we take into account the standard deviation and the maximum value across all columns.

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

FIGURE 4

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

The given data has variables that have int and object data types with 9 columns and 1525 rows. The dataset has no null values. There are both categorical variables and continuous variables in the given data set. So univariate and bivariate analysis can be done in both.

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 (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. [1]

The histograms are used for numerical variables and count plots are used for categorical variables to perform univariate analysis.

It is clear from the graph (*GRAPH 1*) that all the numerical variables are skewed.

Skewness:

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. [2]

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

TABLE 2

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

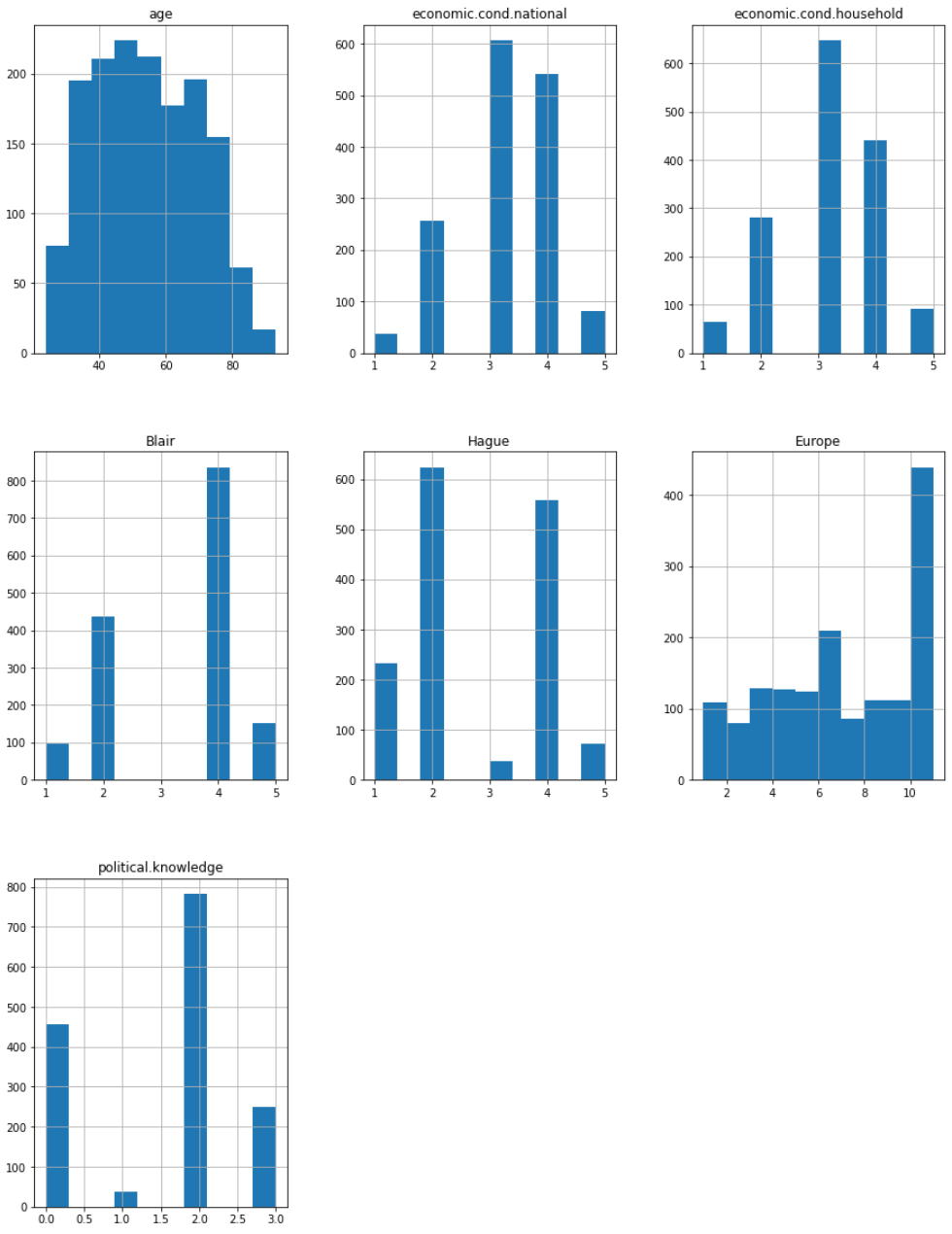
The table below shows the skewness value corresponding to each variable in the given data set.

	Skewness
age	0.144478
economic.cond.national	-0.240216
economic.cond.household	-0.149405
Blair	-0.534892
Hague	0.151950
Europe	-0.135813
political.knowledge	-0.426418

FIGURE 5

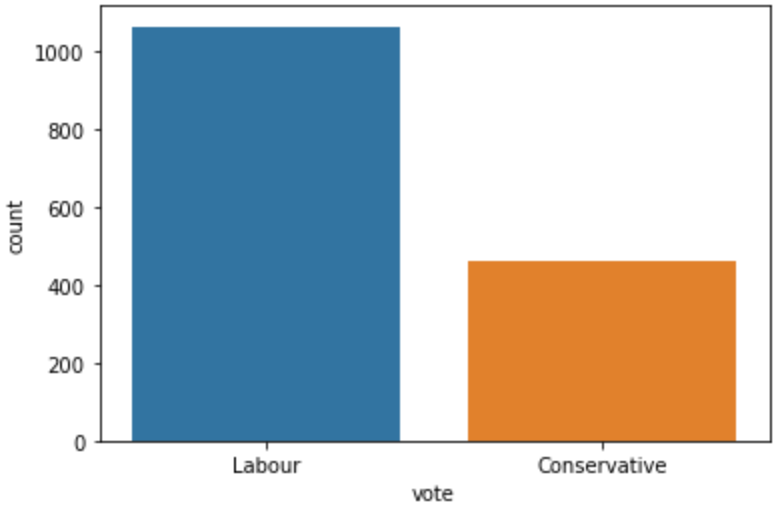
All the variables have a negative skew value except for “age” and “Hague”. Therefore, all the other variables other than “age” and “Hague” are left skewed variables. “Hague” and “age” are right skewed variables.

Univariate Analysis for Continuous variables-

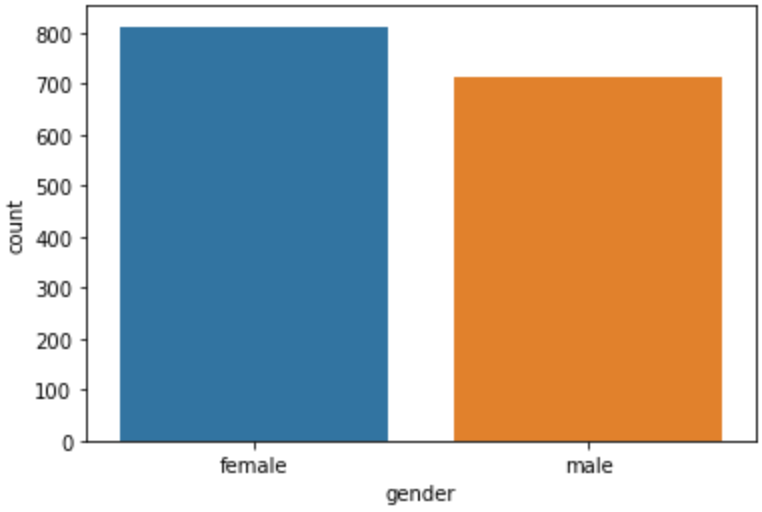


GRAPH 1

Univariate Analysis for Categorical variables-



GRAPH 2



GRAPH 3

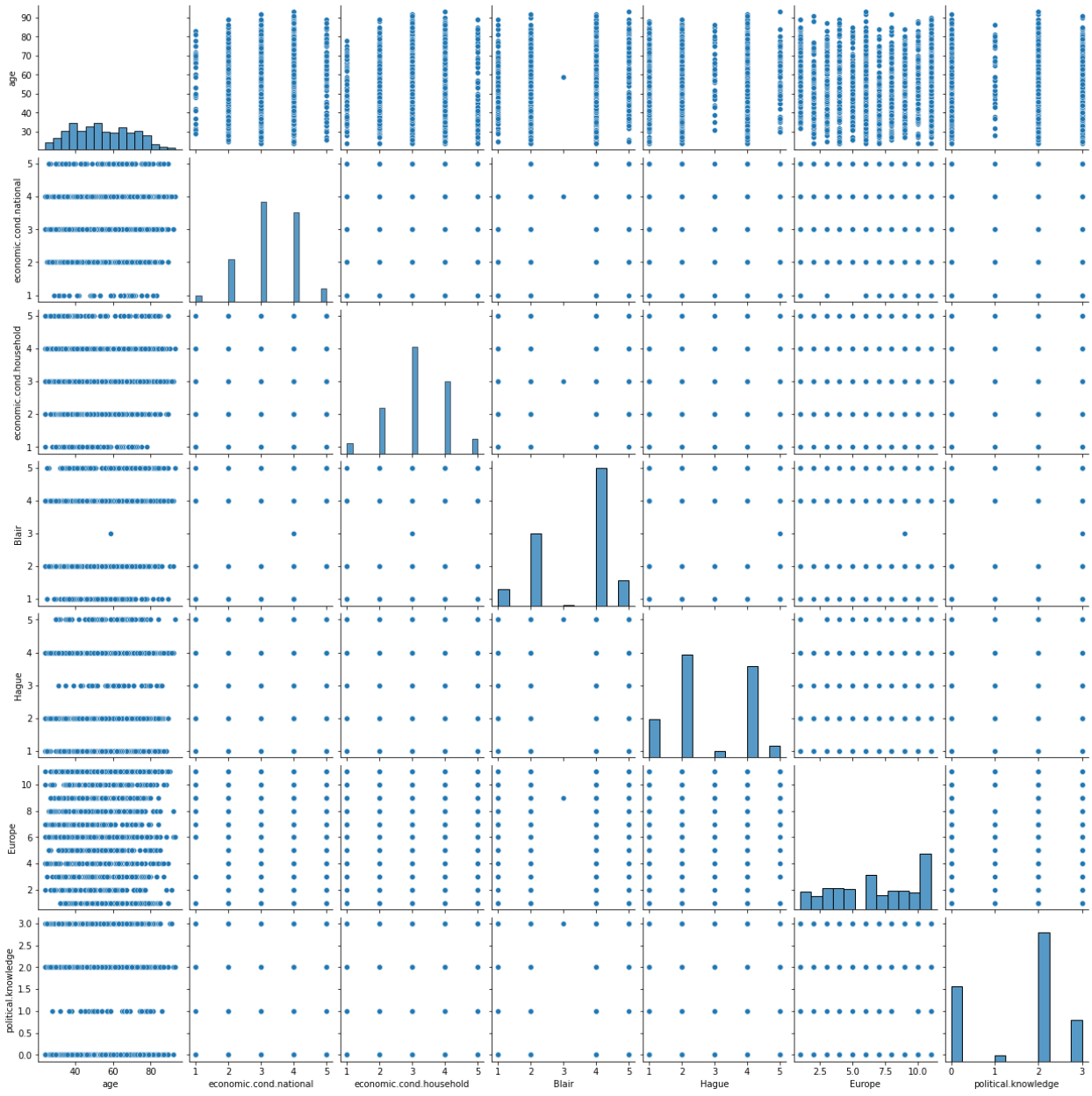
Bivariate analysis:

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. [3]

The pairplot is generally used for numerical variables and box plots are used for categorical with numerical variables to perform bivariate analysis.

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. [4]

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed. [5]

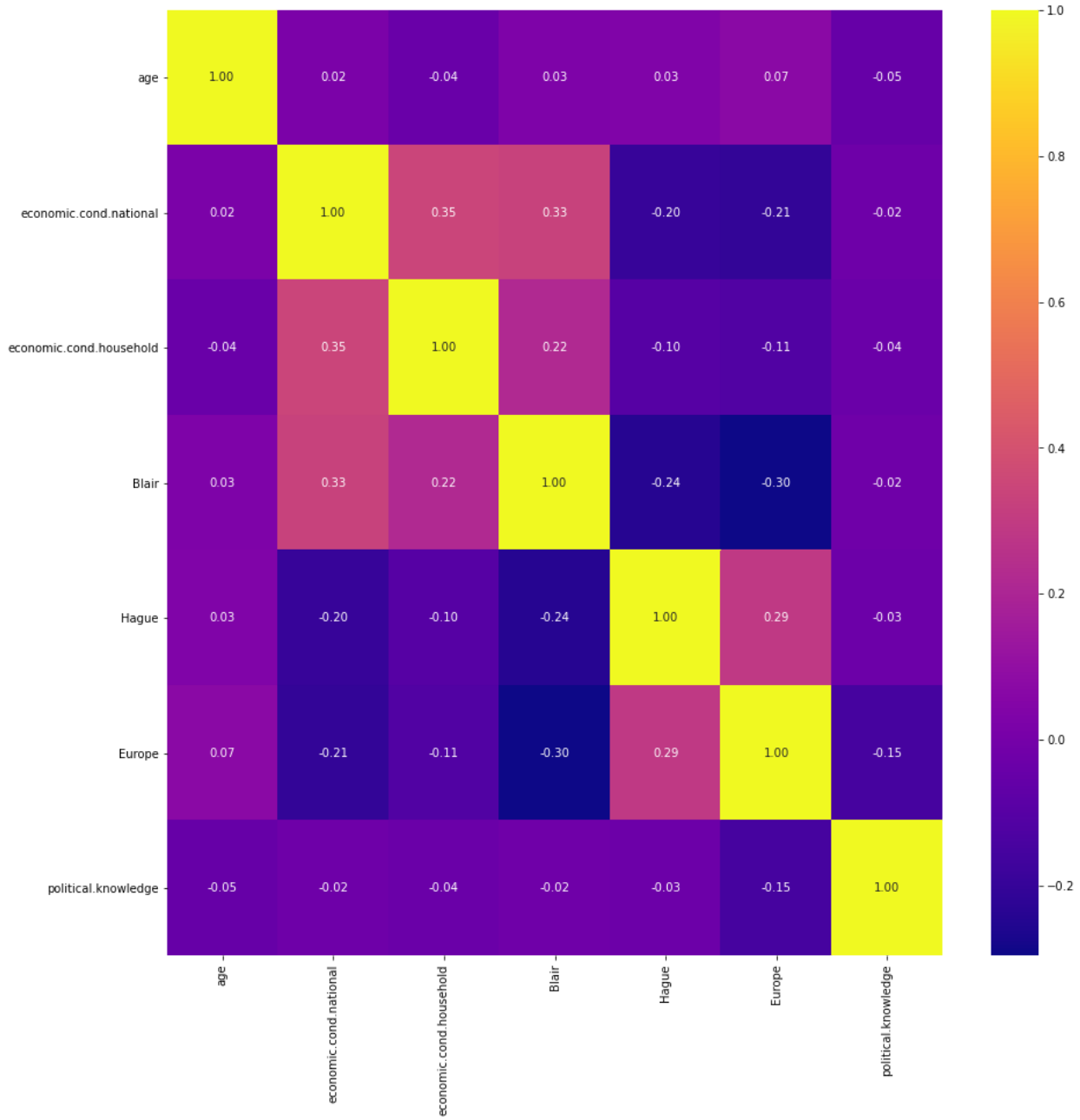


GRAPH 4

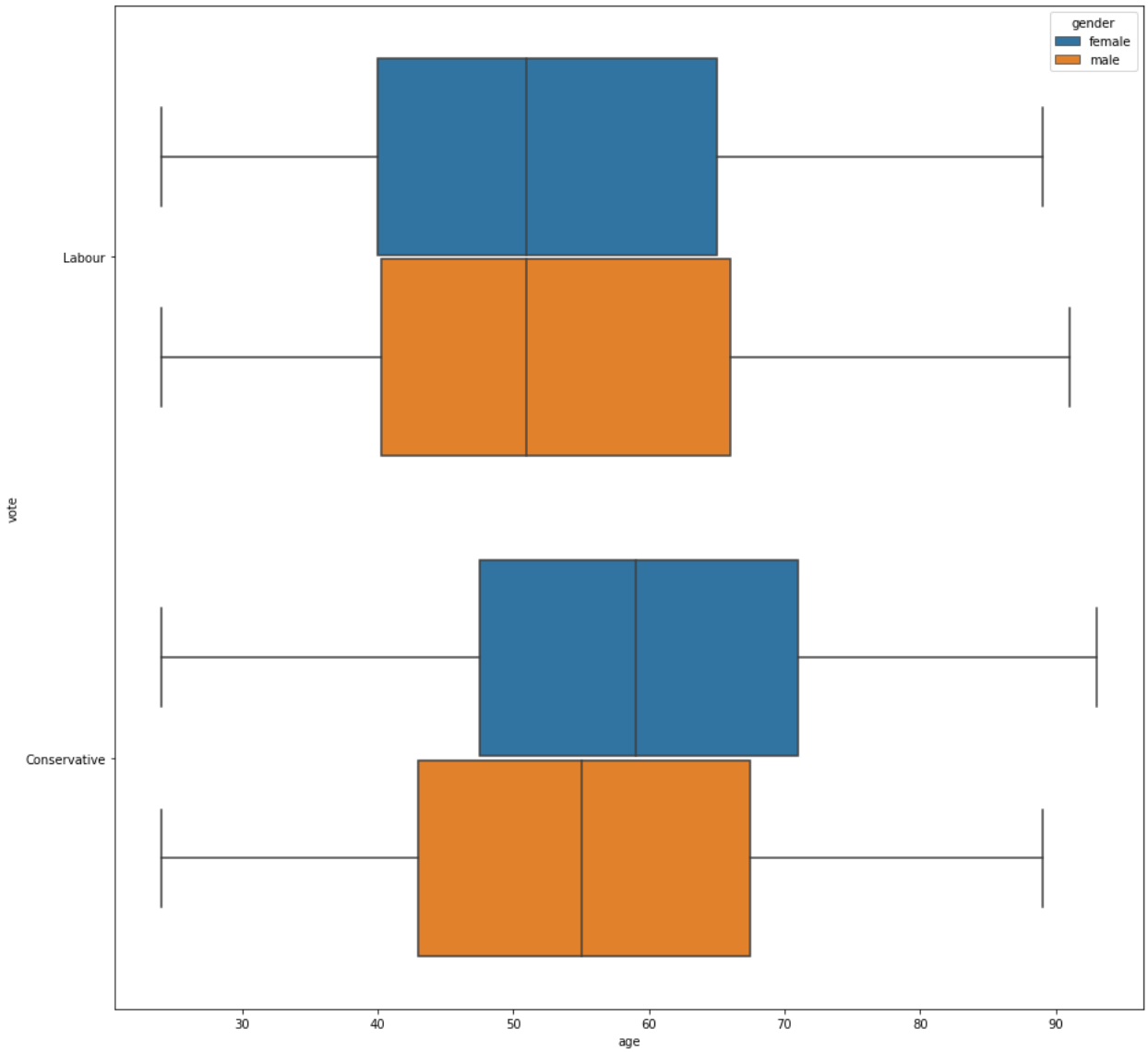
The below are the findings from the pairplot generated -

- None of the variables are correlated with each other

The heat map can also be used to check the association between two variables. All the boxes with a value higher than 0.8 are highly correlated. It is clear that none of the variables have a value more than 0.8. The heat map for all the numerical variable is below,



GRAPH 5

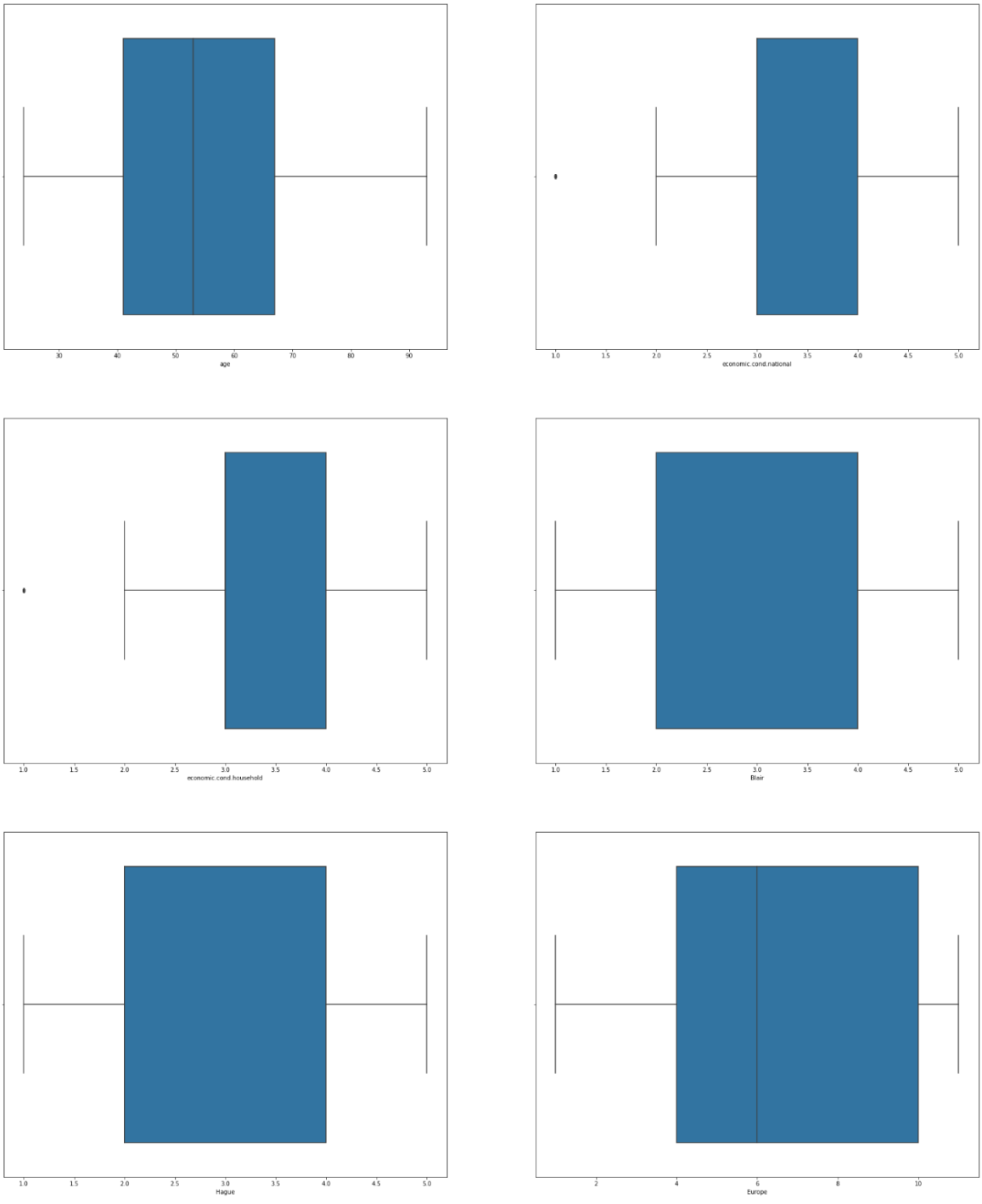


GRAPH 6

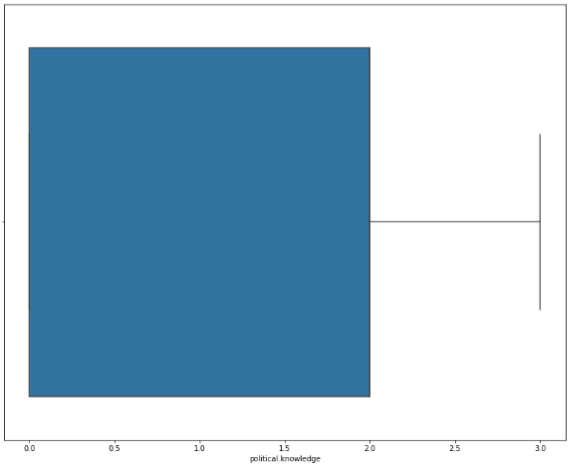
The below are the findings from the boxplot generated -

- In the labour vote category, both female and male have almost equal distribution in age. The oldest age is around 90 and it is male.
- In the conservative vote category, the age of the female is the largest that is above 90. The median of female is 60 and that of male is around 55.

There are outliers in all the continuous variables except "*economic.cond.national*" and "*economic.cond.household*". This is evident from the box plots below,



GRAPH 7



GRAPH 8

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).

Scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. [6]

Scaling converts variables with different scales of measurements into a single scale.

This is done only for the numerical variables.

The data is scaled using the formula $\frac{X-\mu}{\sigma}$.

μ: Mean

σ: Standard deviation

The process of scaling is necessary in the given data set as the variables of the data set are of different scales i.e. one variable has one digit number and other has two digit numbers. For e.g. in our data set “age” has values in two digits and “Blair” has only one

digit number. Since the data in these variables are of different scales, it is tough to compare these variables. Therefore scaling is done in the given data set.

We encode the object data types into categorical variables by encoding them as numerical variables to perform the modeling. There are two object type variables in the given data set that need encoding.

The below table gives the encoded value for each unique data in the respective variable.

VARIABLE: vote

Labour	1
Conservative	0

TABLE 3

VARIABLE: gender

female	0
male	1

TABLE 4

From the given data, 30 percent is taken for the test size and 70 percent is taken for the training. The target variable here is “vote”. The data is fitted into the Logistic Regression Model and LDA.

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

LOGISTIC REGRESSION:

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". [7]

The below is the classification report of the training data-

	precision	recall	f1-score	support
0	0.75	0.65	0.70	323
1	0.86	0.91	0.88	744
accuracy			0.83	1067
macro avg	0.81	0.78	0.79	1067
weighted avg	0.83	0.83	0.83	1067

FIGURE 6

The below is the confusion matrix for training data-

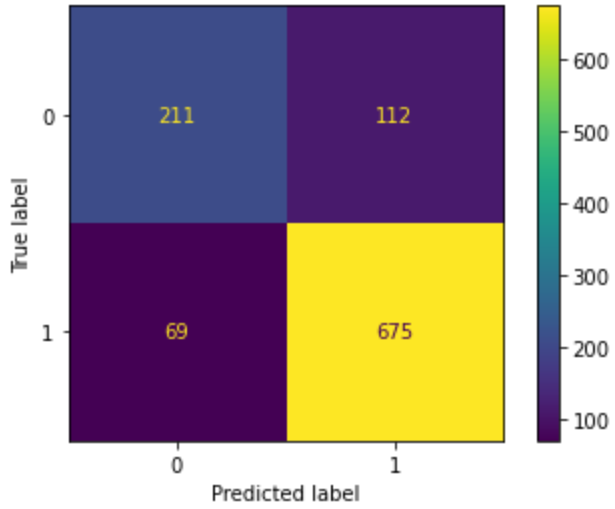
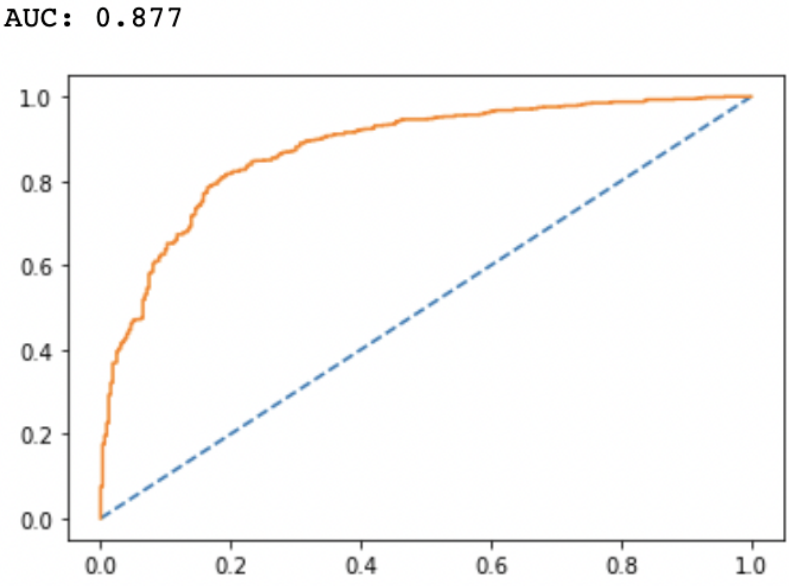


FIGURE 7

The training data accuracy is 83.03%.

ROC curve and AUC score for training data-



GRAPH 9

The below is the classification report of the testing data-

	precision	recall	f1-score	support
0	0.80	0.68	0.73	139
1	0.87	0.92	0.90	319
accuracy			0.85	458
macro avg	0.83	0.80	0.81	458
weighted avg	0.85	0.85	0.85	458

FIGURE 8

The below is the confusion matrix for testing data-

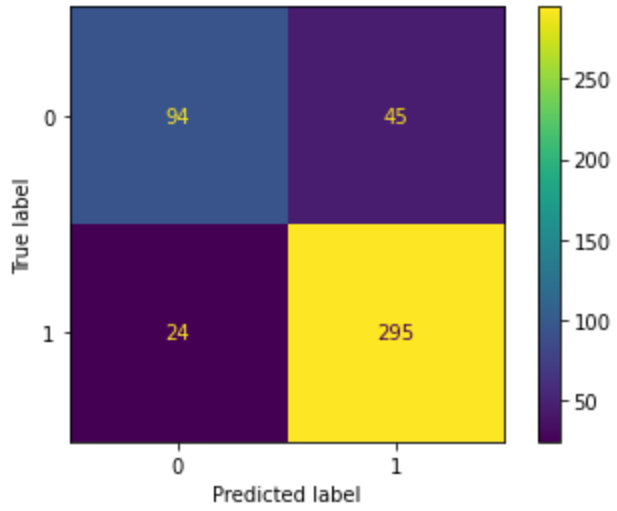
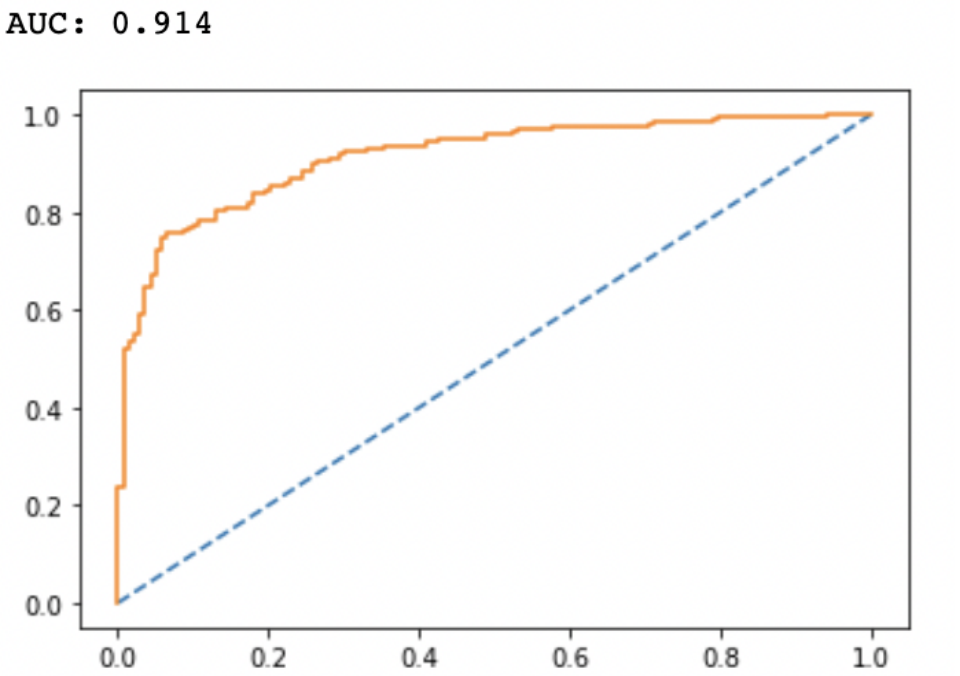


FIGURE 9

The testing data accuracy is 84.93%.

ROC curve and AUC score for testing data-



GRAPH 10

LINEAR DISCRIMINANT ANALYSIS:

Linear Discriminant Analysis (LDA) is a dimensionality reduction technique. As the name implies dimensionality reduction techniques reduce the number of dimensions (i.e. variables) in a dataset while retaining as much information as possible. For instance, suppose that we plotted the relationship between two variables where each color represents a different class. [8]

The below is the classification report of the training data-

	precision	recall	f1-score	support
0	0.73	0.67	0.70	323
1	0.86	0.90	0.88	744
accuracy			0.83	1067
macro avg	0.80	0.78	0.79	1067
weighted avg	0.82	0.83	0.82	1067

FIGURE 10

The below is the confusion matrix for training data-

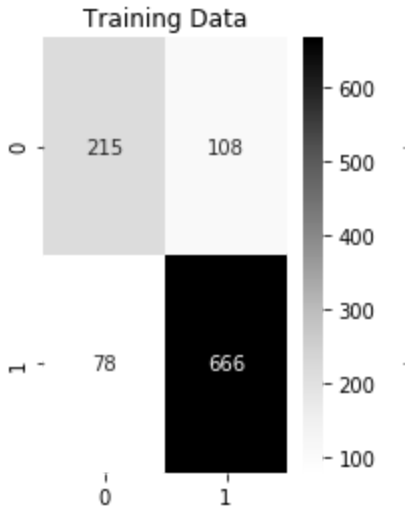
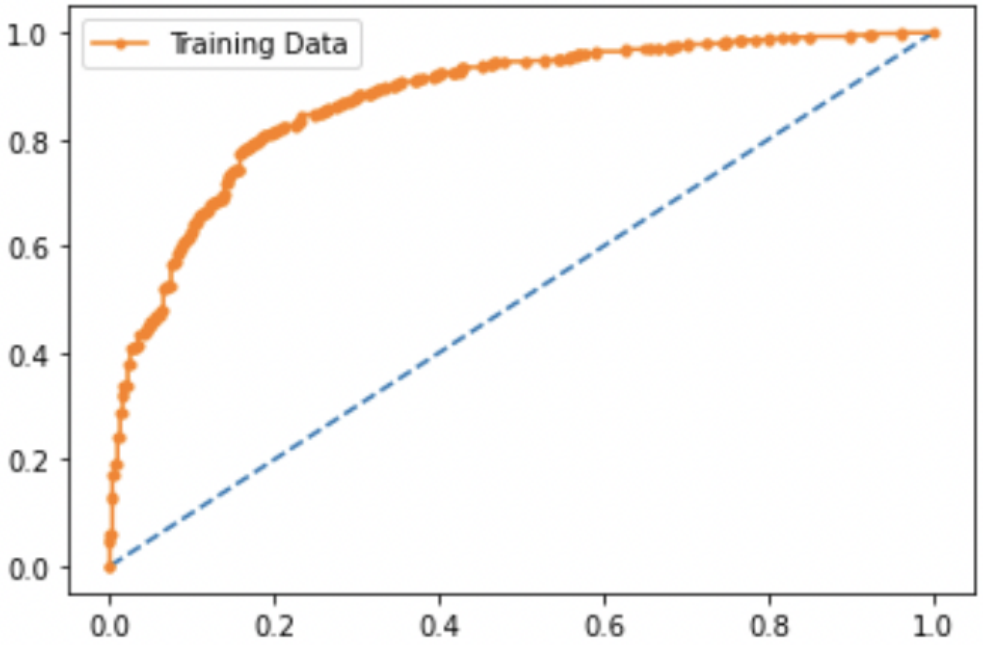


FIGURE 11

ROC curve and AUC score for training data-

AUC for the Training Data: 0.876



GRAPH 11

The below is the classification report of the testing data-

	precision	recall	f1-score	support
0	0.77	0.70	0.73	139
1	0.87	0.91	0.89	319
accuracy			0.84	458
macro avg	0.82	0.80	0.81	458
weighted avg	0.84	0.84	0.84	458

FIGURE 12

The below is the confusion matrix for testing data-

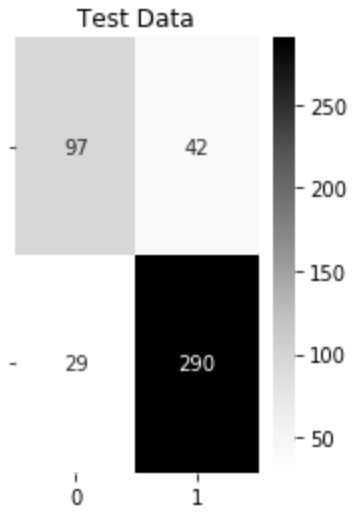
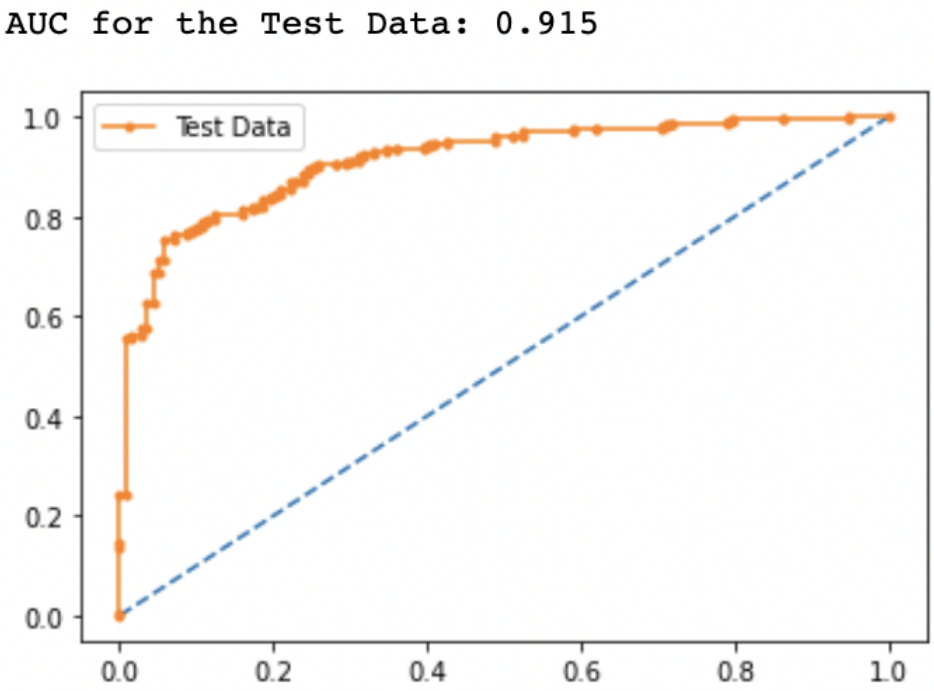


FIGURE 13

ROC curve and AUC score for testing data-



GRAPH 12

The train and test data doesn't show a large difference. Hence there is no overfitting or underfitting issue.

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

K-NEAREST NEIGHBORS:

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification of predictive problems in industry. The following two properties would define KNN well −

- Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data. [9]

NAÏVE BAYES MODEL:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. [10]

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

MODEL TUNING:

Tuning is usually a trial-and-error process by which you change some hyperparameters (for example, the number of trees in a tree-based algorithm or the value of alpha in a linear algorithm), run the algorithm on the data again, then compare its performance on your validation set in order to determine which set of hyperparameters results in the most accurate model.

All machine learning algorithms have a "default" set of hyperparameters, which Machine Learning Mastery defines as "a configuration that is external to the model and whose value cannot be estimated from data." Different algorithms consist of different hyperparameters. For example, regularized regression models have coefficients penalties, decision trees have a set number of branches, and neural networks have a set number of layers. When building models, analysts and data scientists choose the default configuration of these hyperparameters after running the model on several datasets. [11]

BAGGING:

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once. After several data samples are generated, these weak models are then trained independently, and depending on the type of

task—regression or classification, for example—the average or majority of those predictions yield a more accurate estimate.

As a note, the random forest algorithm is considered an extension of the bagging method, using both bagging and feature randomness to create an uncorrelated forest of decision trees. [12]

BOOSTING:

In machine learning, boosting is an ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones. Boosting is based on the question posed by Kearns and Valiant (1988, 1989): "Can a set of weak learners create a single strong learner?" A weak learner is defined to be a classifier that is only slightly correlated with the true classification (it can label examples better than random guessing). In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification. [13]

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.

The below is the classification report of the training data-

	precision	recall	f1-score	support
0	0.76	0.59	0.67	323
1	0.84	0.92	0.88	744
accuracy			0.82	1067
macro avg	0.80	0.76	0.77	1067
weighted avg	0.82	0.82	0.81	1067

FIGURE 14

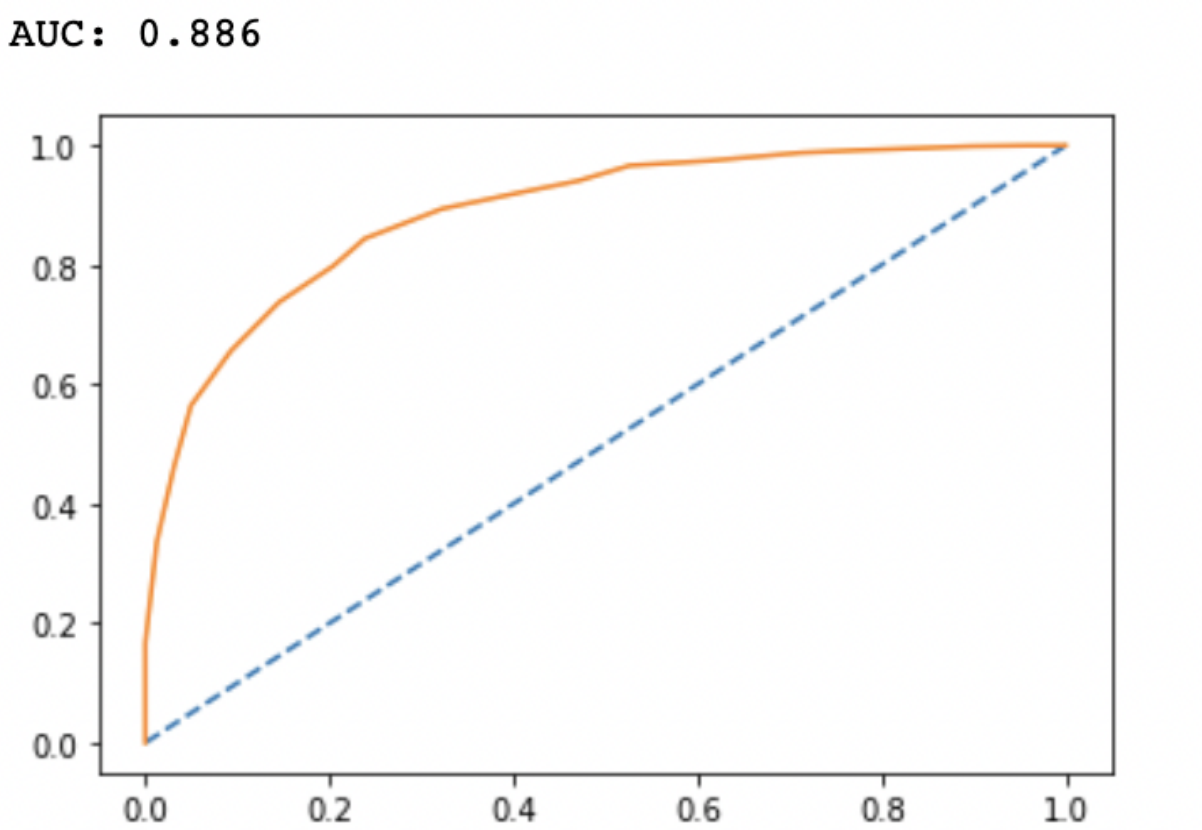
The below is the confusion matrix for training data-

```
[[192 131]
 [ 59 685]]
```

FIGURE 15

The training data accuracy is 82.19%.

ROC curve and AUC score for training data-



GRAPH 13

The below is the classification report of the testing data-

	precision	recall	f1-score	support
0	0.73	0.61	0.67	139
1	0.84	0.90	0.87	319
accuracy			0.81	458
macro avg	0.79	0.76	0.77	458
weighted avg	0.81	0.81	0.81	458

FIGURE 16

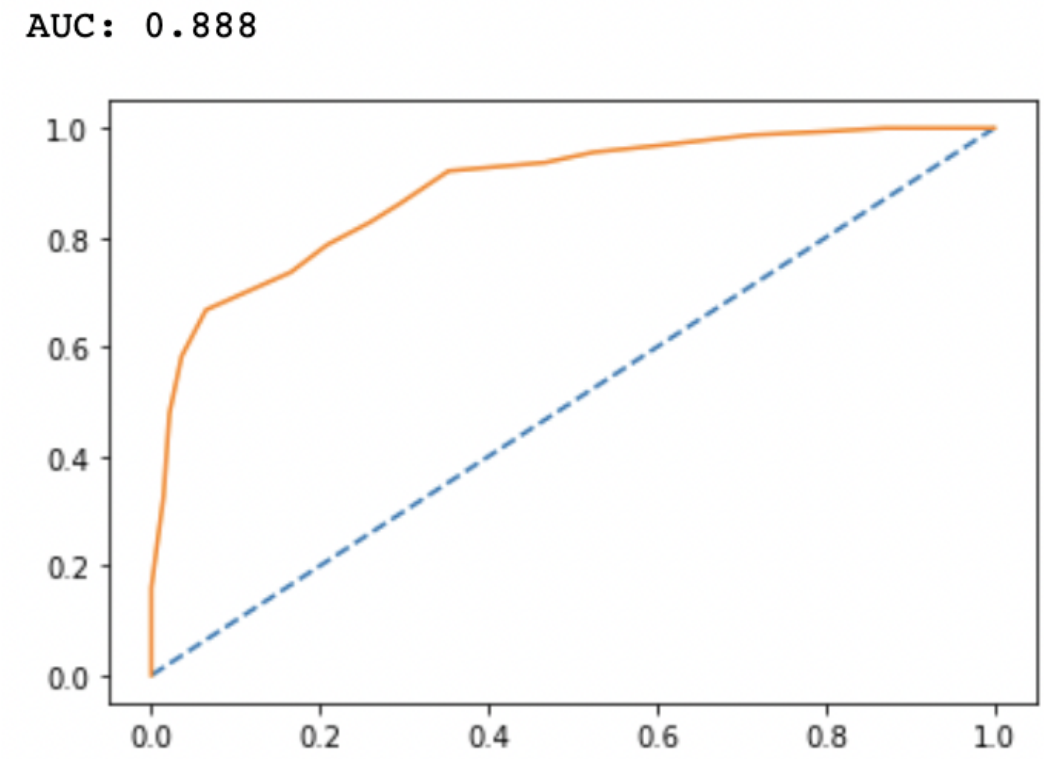
The below is the confusion matrix for testing data-

```
[[ 85  54]
 [ 31 288]]
```

FIGURE 17

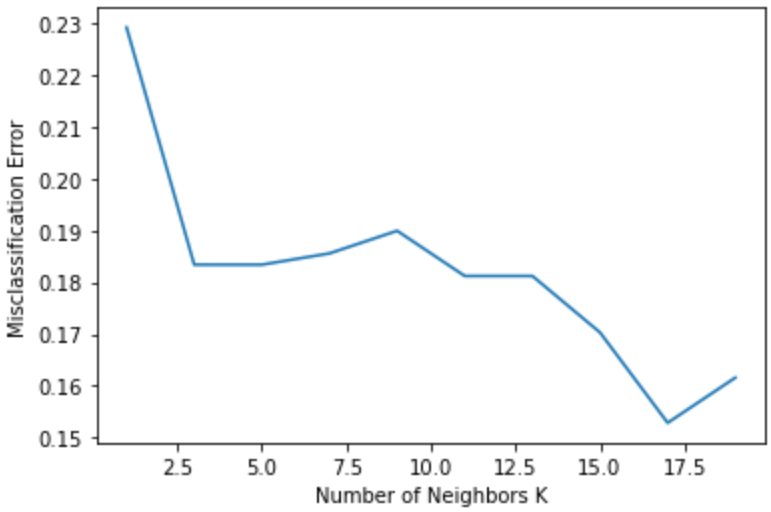
The testing data accuracy is 81.44%.

ROC curve and AUC score for testing data-



GRAPH 14

Determining the number of neighbours –



GRAPH 15

On applying models on Naïve Bayes, Adaptive Boosting, GBCL, and Bagging, the final results are –

	Logit Train	Logit Test	LDA Train	LDA Test	KNN Train	KNN Test	NB Train	NB Test	ADB Train	ADB Test	gbcl Train	gbcl Test	DT Train	DT Test	Bagging Train	Bagging Test
Accuracy	0.83	0.85	0.83	0.84	0.82	0.84	0.82	0.85	0.84	0.84	0.89	0.84	1.0	0.77	1.0	0.82
AUC	0.88	0.91	0.88	0.91	0.89	0.89	0.87	0.91	0.90	0.91	0.95	0.91	1.0	0.72	1.0	0.89
Recall	0.91	0.92	0.91	0.92	0.92	0.92	0.88	0.90	0.92	0.91	0.94	0.92	1.0	0.84	1.0	0.90
Precision	0.86	0.87	0.86	0.87	0.84	0.86	0.87	0.88	0.87	0.87	0.90	0.86	1.0	0.83	1.0	0.85
F1 Score	0.88	0.90	0.88	0.90	0.88	0.89	0.87	0.89	0.90	0.89	0.92	0.89	1.0	0.83	1.0	0.87

FIGURE 18

1.8 Based on these predictions, what are the insights?

Insights and recommendations –

• The Labour Party constituted of around 70% of the votes when compared to the 30% Conservative Party

• A gender-wise breakup reveals that around 53% were women among the 70% of Labour Party voters and 47% of 70% Labour Party voters were men

• Similarly 52% of the 30% were female voters who voted for Conservative Party and the rest 48% were male voters

• Also there is an evidence that older people preferred the Conservative Party while the nextgen / younger audience preferred to vote for the Labour Party

• People who voted for Labour had a strong opinion and support towards Blair and a strong opposition towards Hague – whereas people voting for Conservative did not show much partiality between the two. Though the Conservative voters did not oppress Blair, they supported Hague to a greater extent

Problem 2

Problem statement-

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.

SPEECH: President Franklin D. Roosevelt in 1941

Number of characters	7571
Number of words	1526
Number of sentences	68

TABLE 5

SPEECH: President John F. Kennedy in 1961

Number of characters	7618
Number of words	1543
Number of sentences	52

TABLE 6

SPEECH: President Richard Nixon in 1973

Number of characters	9991
Number of words	2006
Number of sentences	68

TABLE 7

2.2 Remove all the stopwords from all three speeches.

	Number of words before removal of stop words	Number of words after removal of stop words
Roosevelt	1526	657
Kennedy	1543	669
Nixon	2006	788

TABLE 8

After removing the stop words the number of real words in Roosevelt's speech is 657.

After removing the stop words the number of real words in Kennedy's speech is 669.

After removing the stop words the number of real words in Nixon's speech is 788.

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)

SPEECH: President Franklin D. Roosevelt in 1941

nation	12
know	10
spirit	9

TABLE 9

SPEECH: President John F. Kennedy in 1961

sides	8
world	8
new	7

TABLE 10

SPEECH: President Richard Nixon in 1973

america	21
peace	19
world	18

TABLE 11

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

SPEECH: President Franklin D. Roosevelt in 1941

FIGURE 19

SPEECH: President John F. Kennedy in 1961

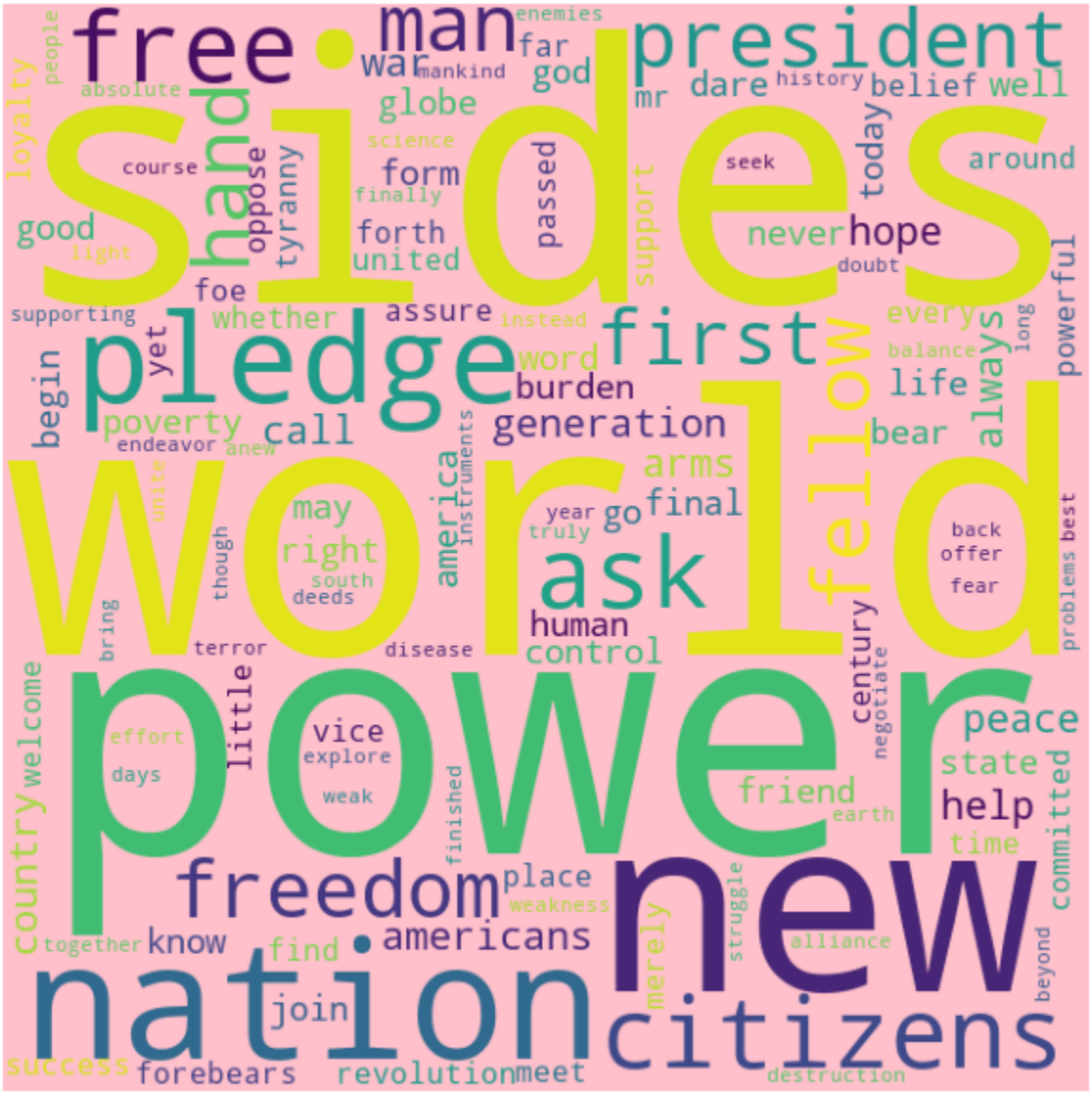


FIGURE 20

SPEECH: President Richard Nixon in 1973

FIGURE 21

References

Websites-

[1] https://www.statisticshowto.com/univariate/

[2] https://www.spss-tutorials.com/skewness/

[3] https://en.wikipedia.org/wiki/Bivariate_analysis

[4] https://pythonbasics.org/seaborn-pairplot/

[5] https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd5

[6] https://en.wikipedia.org/wiki/Feature_scaling

[7] https://en.wikipedia.org/wiki/Logistic_regression

[8] https://towardsdatascience.com/linear-discriminant-analysis-in-python-76b8b17817c2

[9]https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm

[10] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[11] https://www.datarobot.com/wiki/tuning/

[12] https://www.ibm.com/cloud/learn/bagging

[13] https://en.wikipedia.org/wiki/Boosting_(machine_learning)

End of Project