MACHINE LEARNING
PROJECT BUSINESS REPORT

Sanjana M
PGP-DSBA Online

Table of Contents

Table of Figures

Table of Tables

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

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

Election data has originally has 1525 entries with 10 columns.

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

Table 1: Original data head

Out of these 10, 1 is a serial number columns which is dropped, ‘vote’ and ‘gender’ columns are of categorical variables. The rest are of int variables namely 'age', 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political.knowledge'

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               1525 non-null   int64 
 1   vote                     1525 non-null   object
 2   age                      1525 non-null   int64 
 3   economic.cond.national   1525 non-null   int64 
 4   economic.cond.household  1525 non-null   int64 
 5   Blair                    1525 non-null   int64 
 6   Hague                    1525 non-null   int64 
 7   Europe                   1525 non-null   int64 
 8   political.knowledge      1525 non-null   int64 
 9   gender                   1525 non-null   object
dtypes: int64(8), object(2)
memory usage: 119.3+ KB
```

Table 2: Data info

“age” is a discrete data with mean of 54.18 years, minimum being 24v years and maximum being 93 years.
The rest of the entries are ordinal in nature with multiple levels to show intensity.

	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

Table 3: Describing of numeric variables of data

On original data,
‘Vote’ has 2 group
‘Labour’ with 1063 count and ‘Conservative’ with count of 462. An imbalance can be observed here.
‘Gender’ has 2 groups
‘female’ with count of 812, ‘male’ with count of 713. Balance is seen here

```
vote    No of Levels: 2
Labour          1063
Conservative     462
Name: vote, dtype: int64


gender   No of Levels: 2
female    812
male      713
Name: gender, dtype: int64
```

Table 4: Categorical variables

On null check, we can see that no entries are left unfilled

```
vote                       0
age                        0
economic.cond.national     0
economic.cond.household    0
Blair                      0
Hague                      0
Europe                     0
political.knowledge        0
gender                     0
dtype: int64
```

Table 5: Null check

On checking, we found that there were 8 duplicate entries, which were dropped.
After all the necessary column, rows dropping, the data shape is (1517, 9)
Meaning, 1517 rows and 9 columns

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

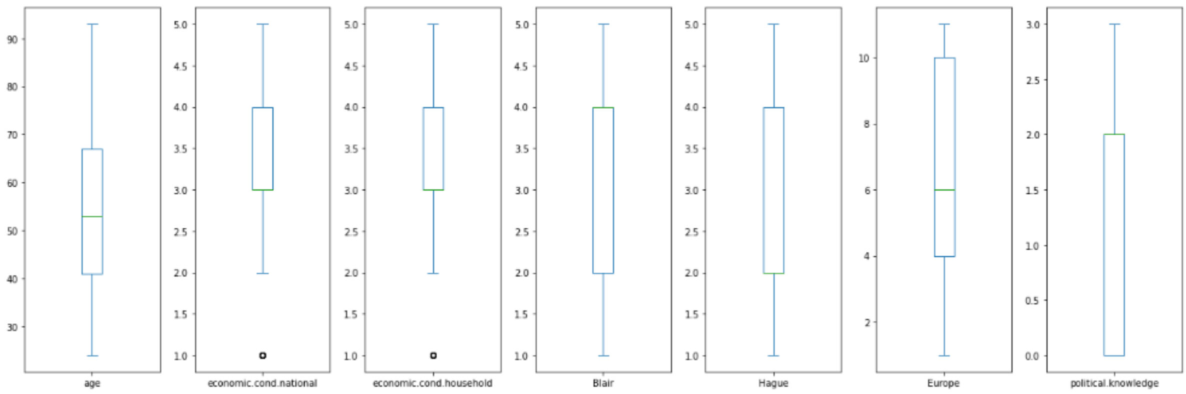


Figure 1: Box plots

'Age' has no outliers. 'Economic.cond.national' and . 'Economic.cond.household' are ordinal variables hence cant have outliers.
As we can see that 'Age' variable is having discrete values so to convert this to ordinal values we will use binning.

8 age_ bins are created : '20s', '30s', '40s', '50s', '60s', '70s', '80s', '90s'
['20s' < '30s' < '40s' < '50s' < '60s' < '70s' < '80s' < '90s']

Each entry is marked with the appropriate age bin as shown below:

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender	age_bins
0	Labour	43	3	3	4	1	2	2	female	40s
1	Labour	36	4	4	4	4	5	2	male	30s

Table 6: Data with age_bins

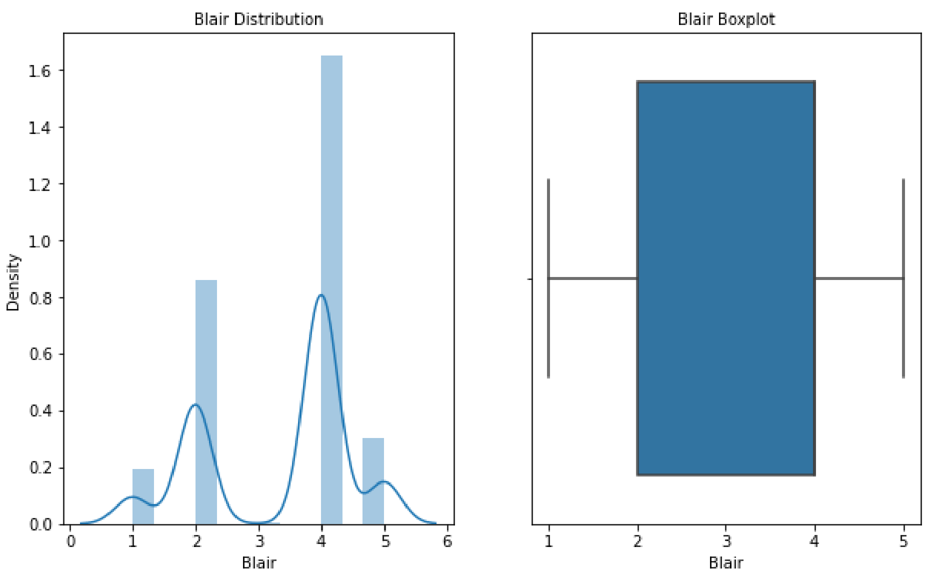


Figure 2: Blair-graphical representation

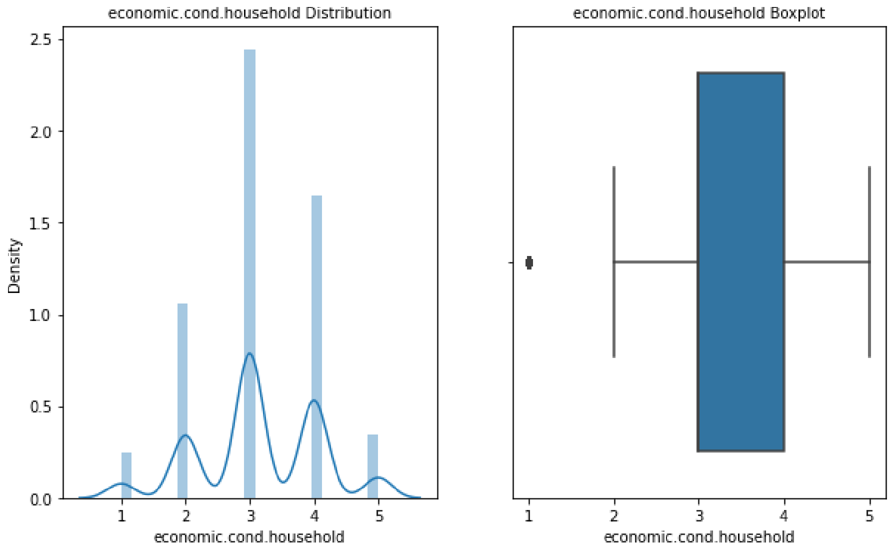


Figure 3: economic.cond.household-graphical representation

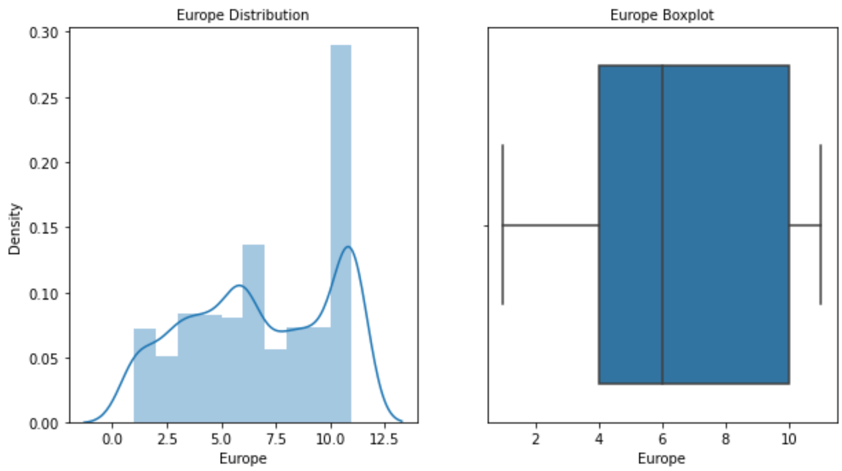


Figure 4: Europe-graphical representation

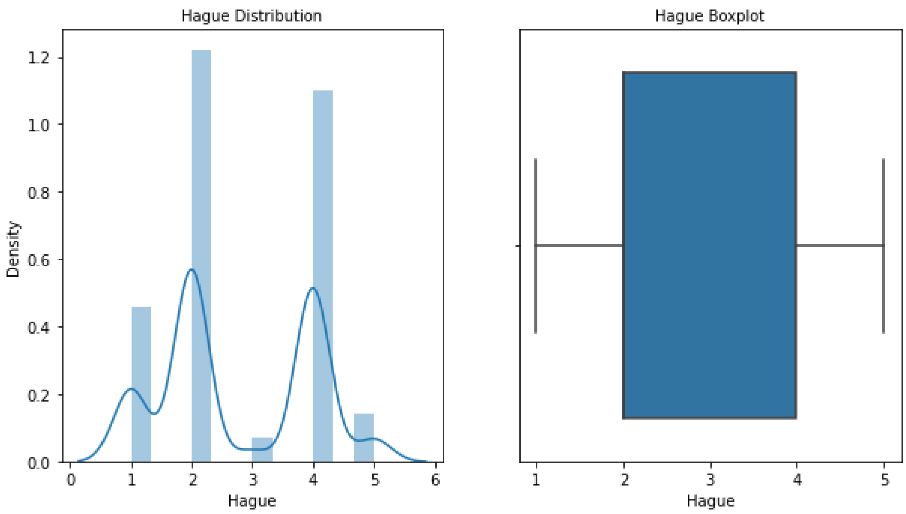


Figure 5: Hague-graphical representation

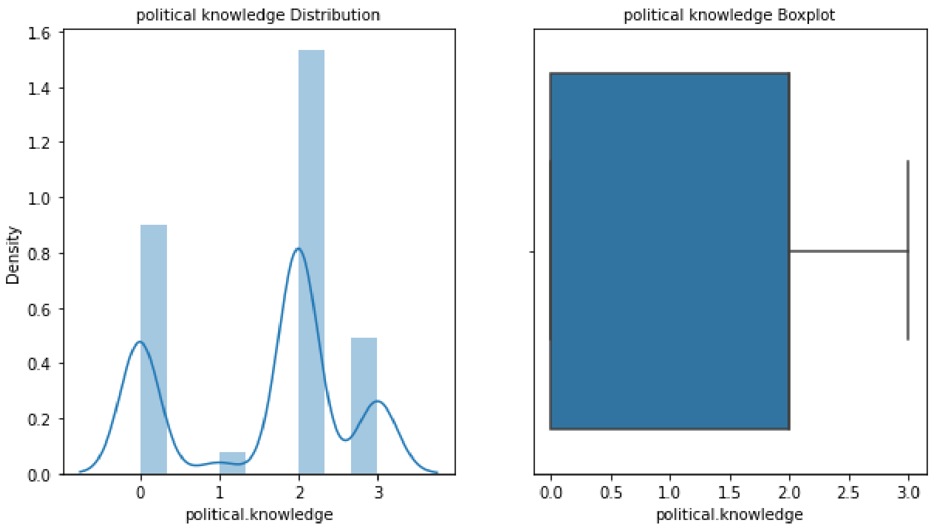


Figure 6:Political knowledge-graphical representation

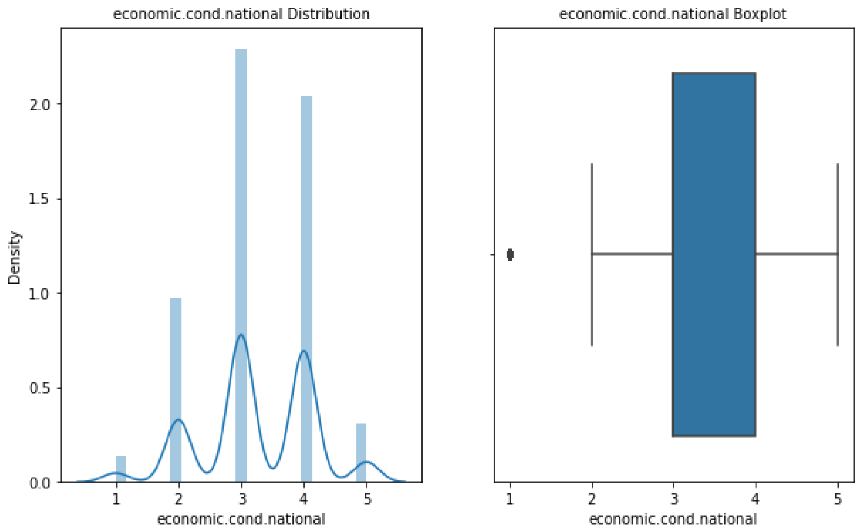


Figure 7: economic.cond.national-graphical representation

After necessary data processing, ‘vote’ has 1057 ‘Labour’ entries and 460 ‘Conservative’ entries

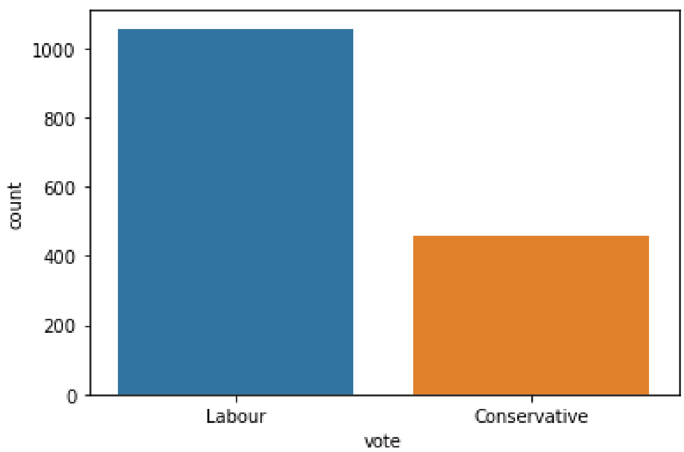


Figure 8: Vote-countplot

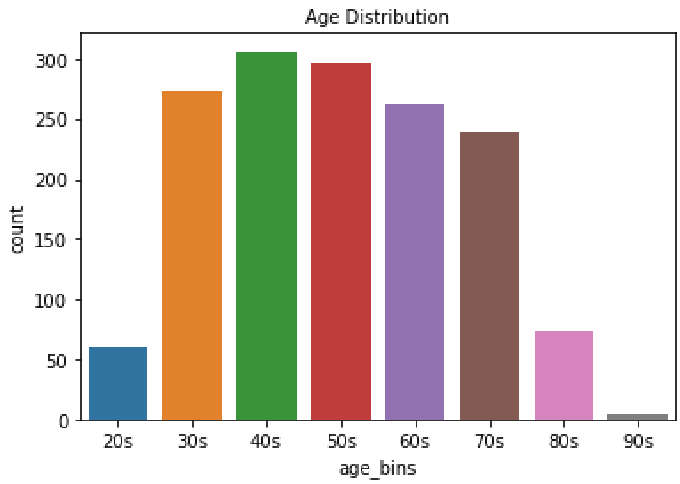


Figure 9: Age bins-countplot

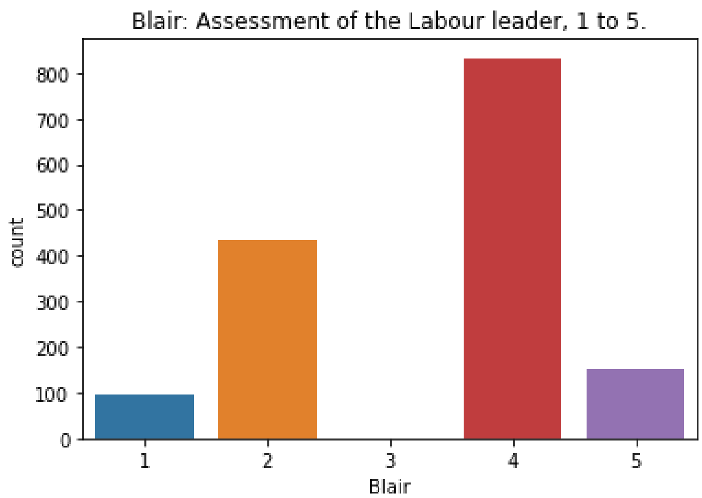


Figure 10: Balir-countplot

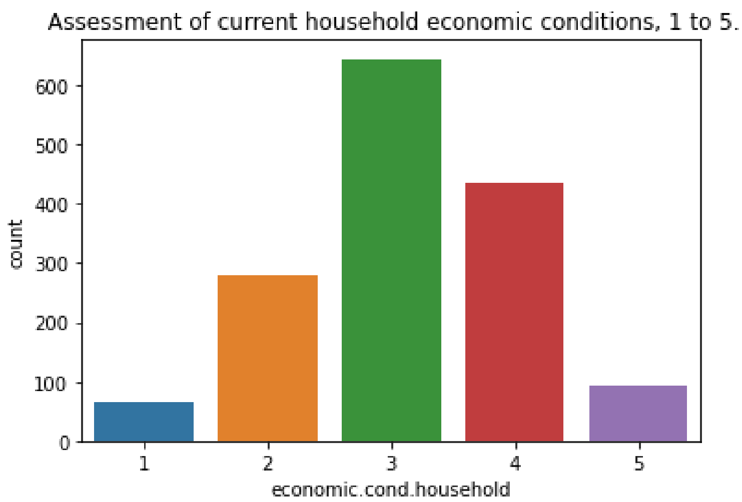


Figure 11: economic.cond.household-countplot

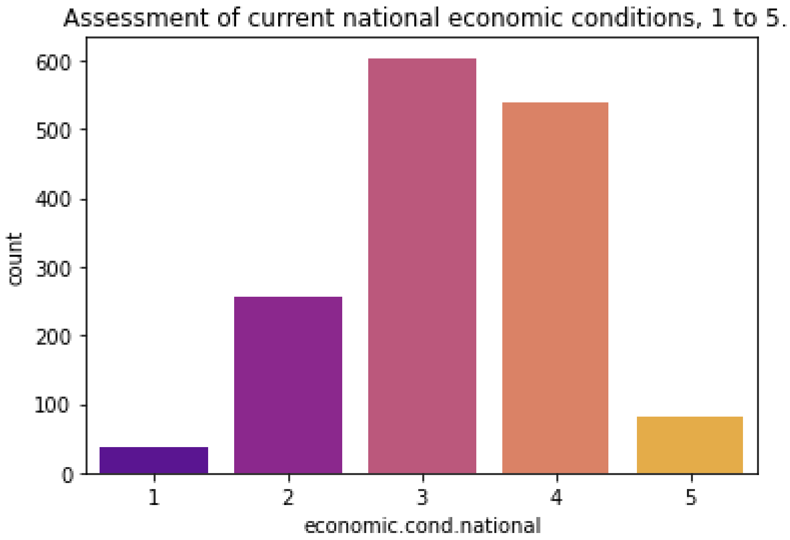


Figure 12: economic.cond.national-countplot

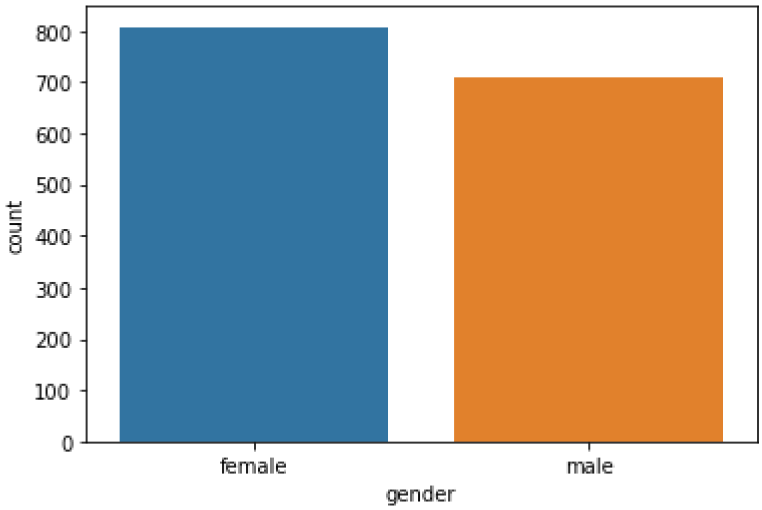


Figure 13: Gender-countplot

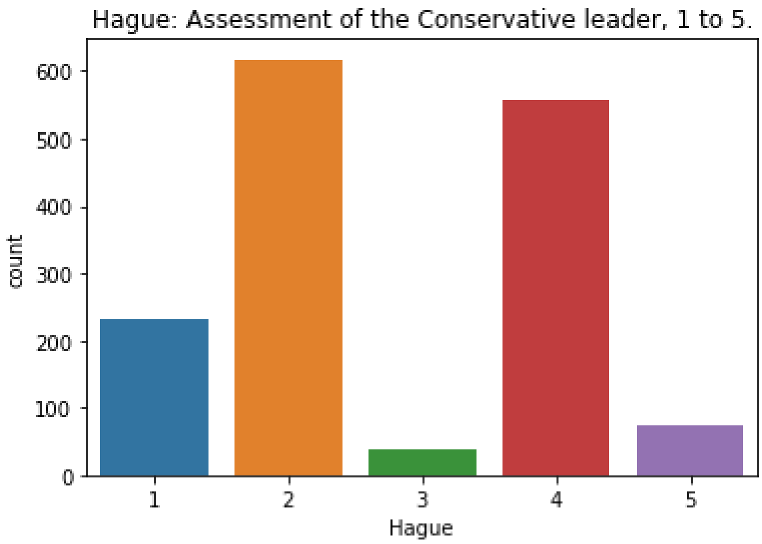


Figure 14: Hague-countplot

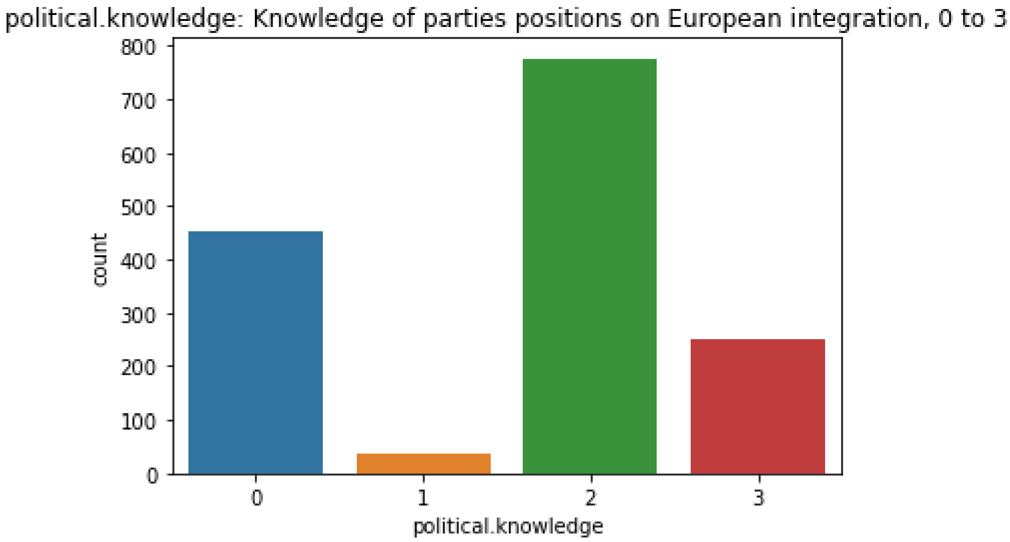


Figure 15: Political knowledge-countplot

Below plot represents that the Labour Party is getting More Votes in each Age Group

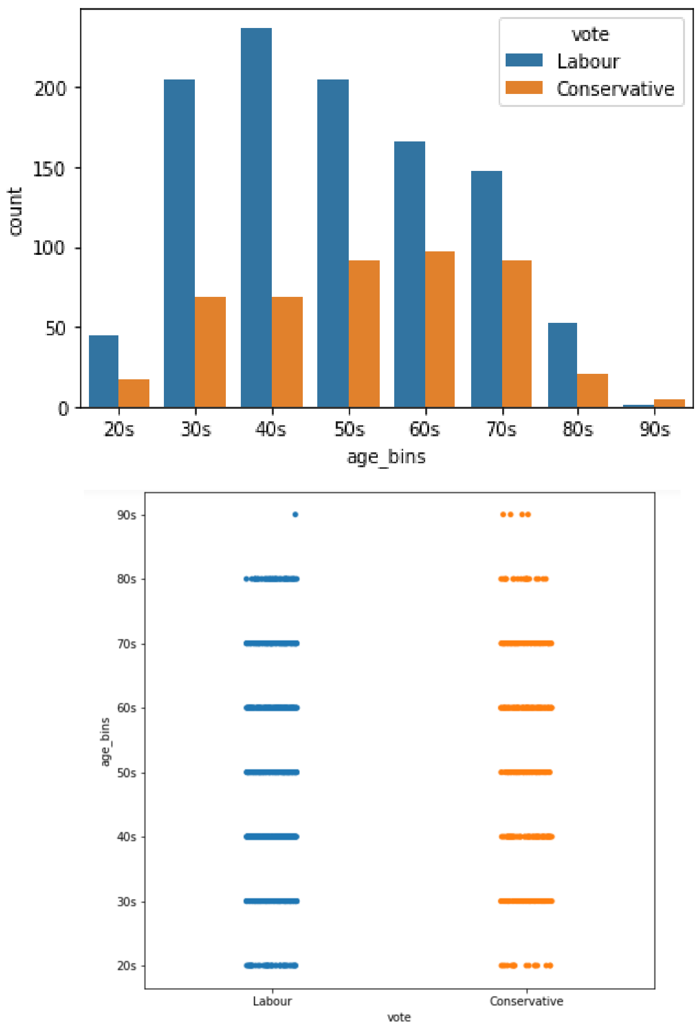
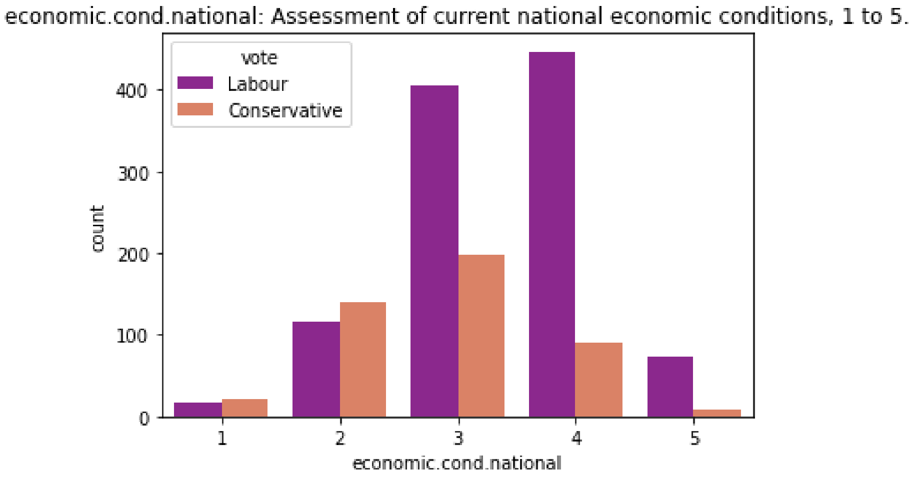


Figure 16: Age_bins and Vote

Economin.condi in household and national gets more votes for either of the parties when their score is 3 (or 4)



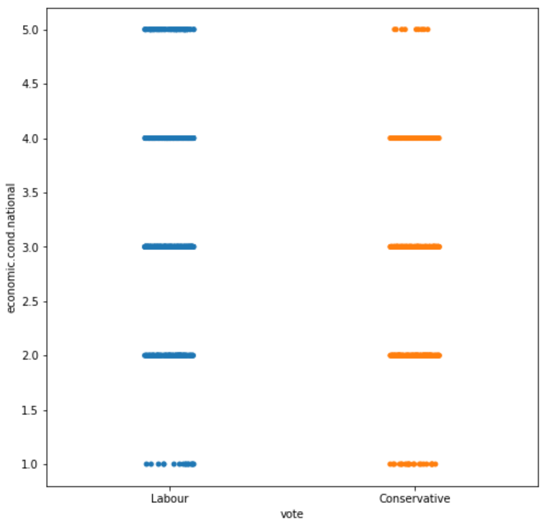


Figure 17: economic.cond.national and Vote

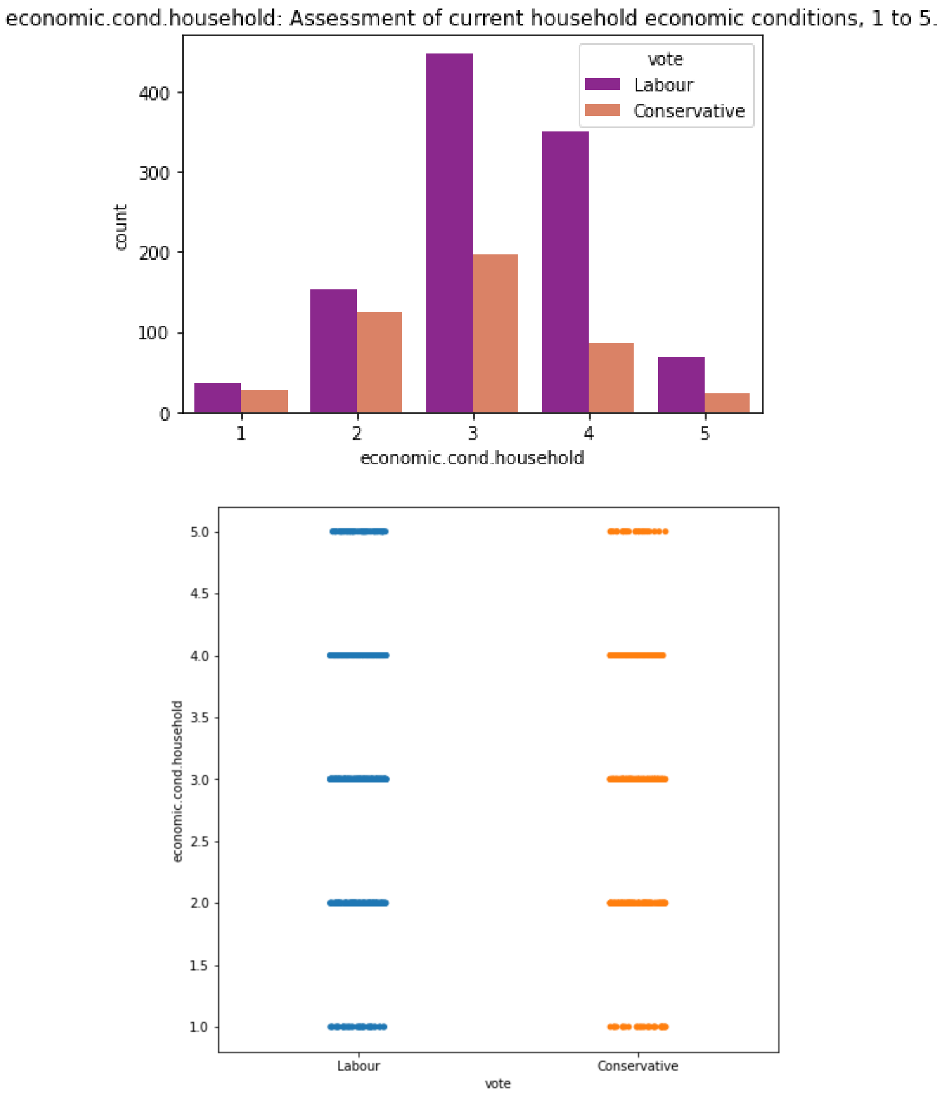


Figure 18: economic.cond.household and Vote

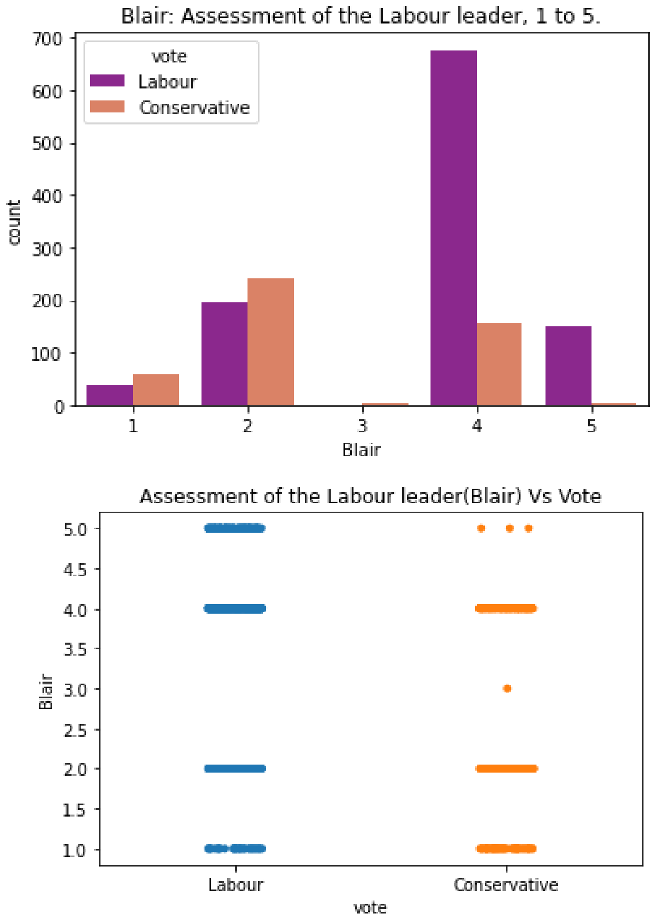


Figure 19: Blair and Vote

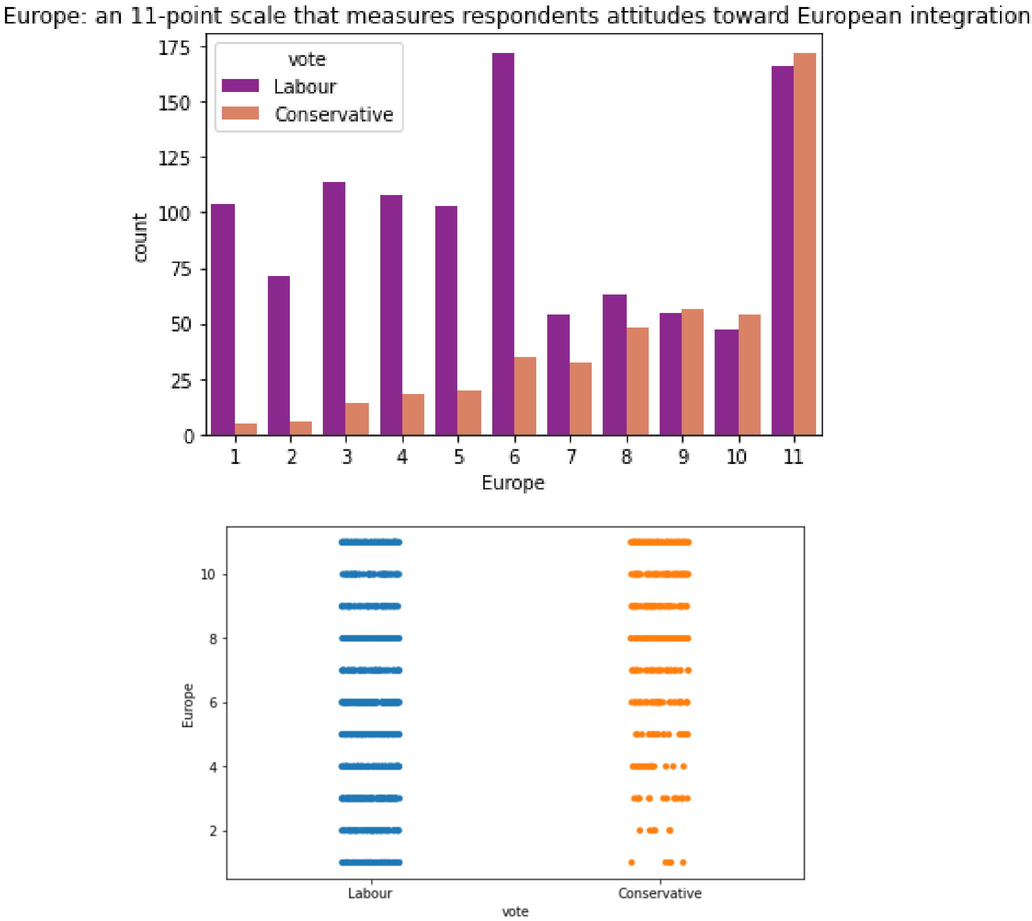


Figure 20: Europe and Vote

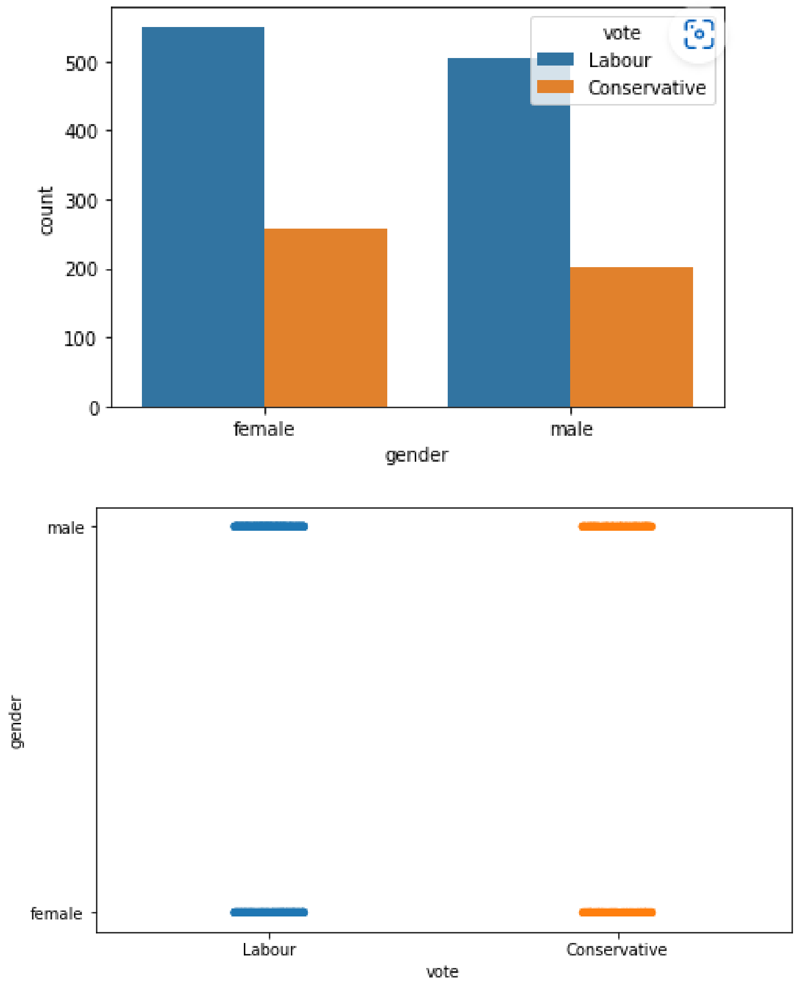
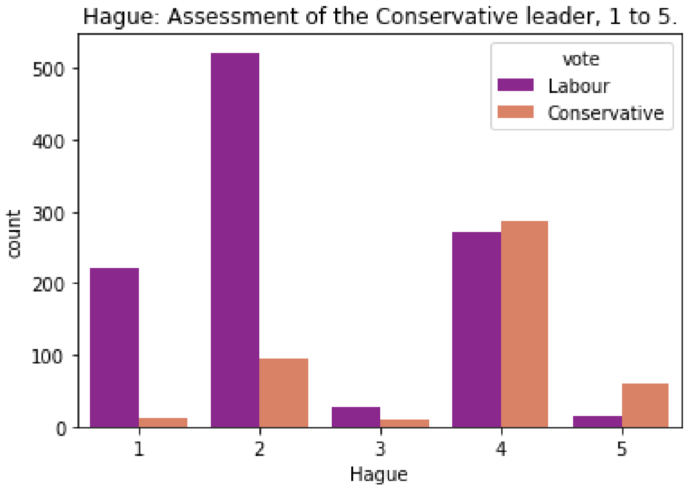


Figure 21: Gender and Vote



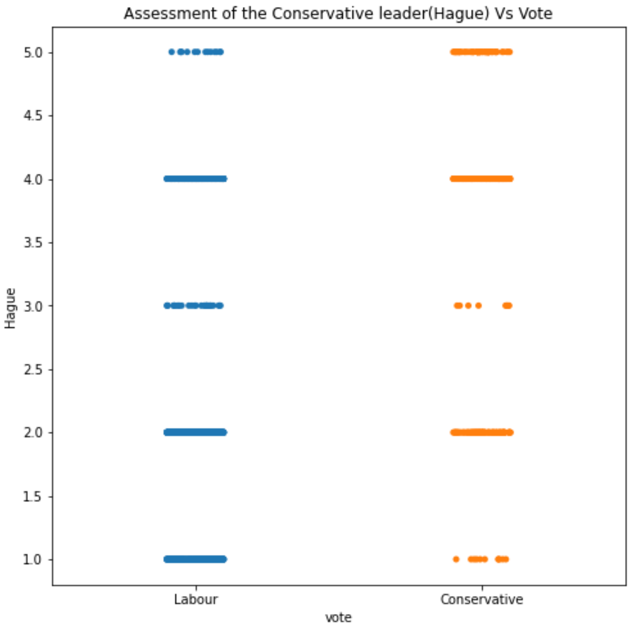


Figure 22: Hague and Vote

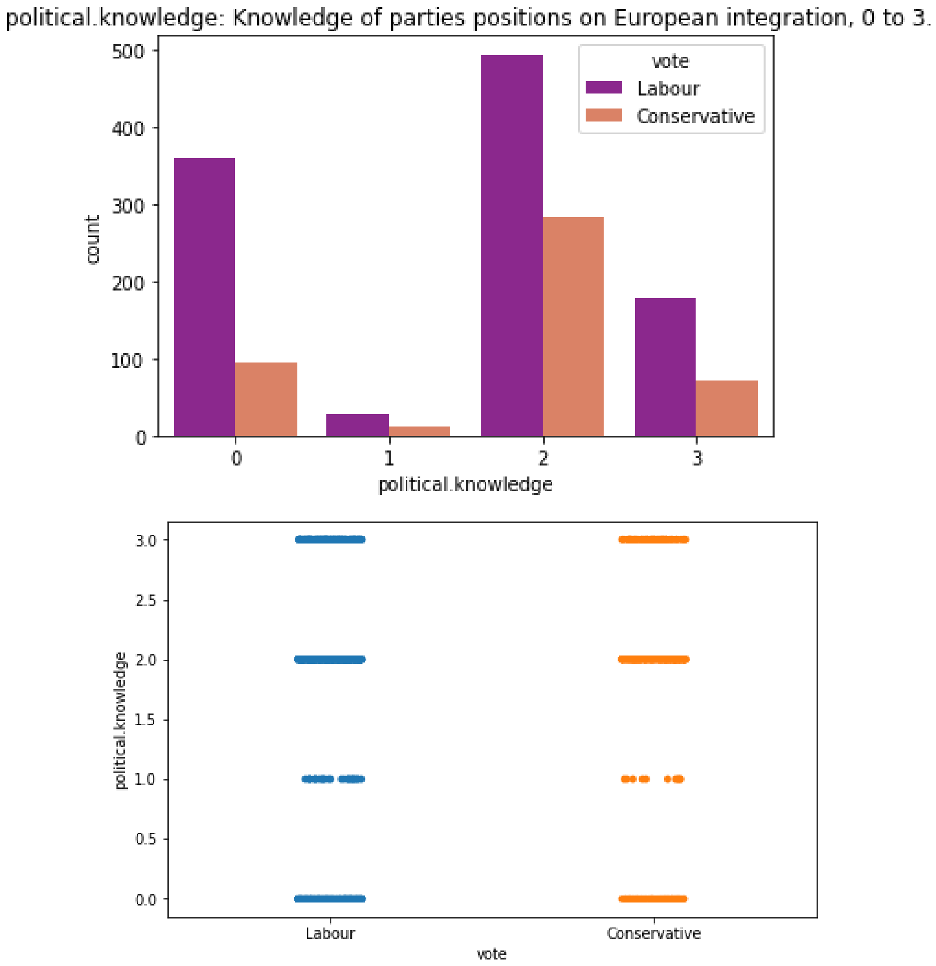


Figure 23: Political knowledge and Vote

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
age	1.000000	0.018687	-0.038868	0.032084	0.031144	0.064562	-0.046598
economic.cond.national	0.018687	1.000000	0.347687	0.326141	-0.200790	-0.209150	-0.023510
economic.cond.household	-0.038868	0.347687	1.000000	0.215822	-0.100392	-0.112897	-0.038528
Blair	0.032084	0.326141	0.215822	1.000000	-0.243508	-0.295944	-0.021299
Hague	0.031144	-0.200790	-0.100392	-0.243508	1.000000	0.285738	-0.029906
Europe	0.064562	-0.209150	-0.112897	-0.295944	0.285738	1.000000	-0.151197
political.knowledge	-0.046598	-0.023510	-0.038528	-0.021299	-0.029906	-0.151197	1.000000

Table 7: Correlation

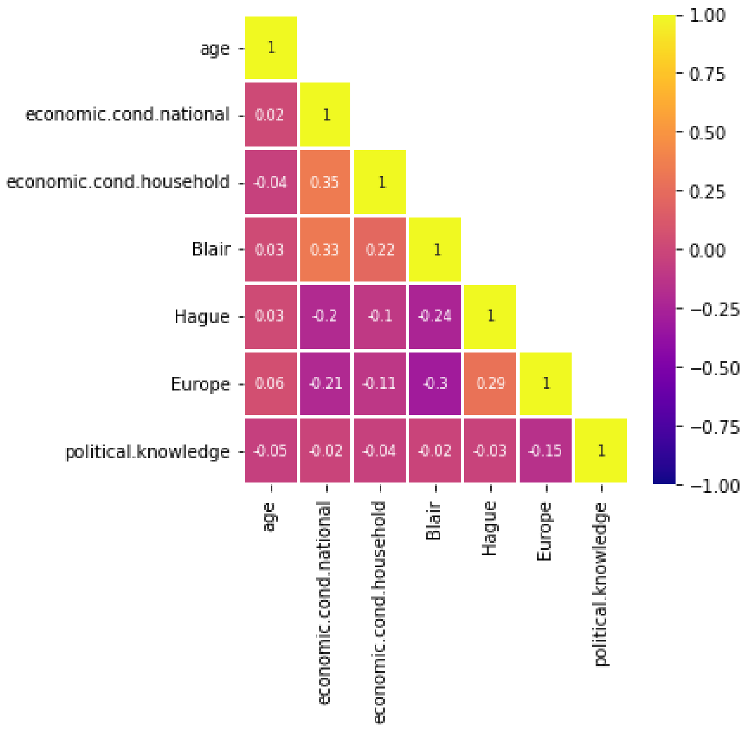


Figure 24: Heatmap to show correlation

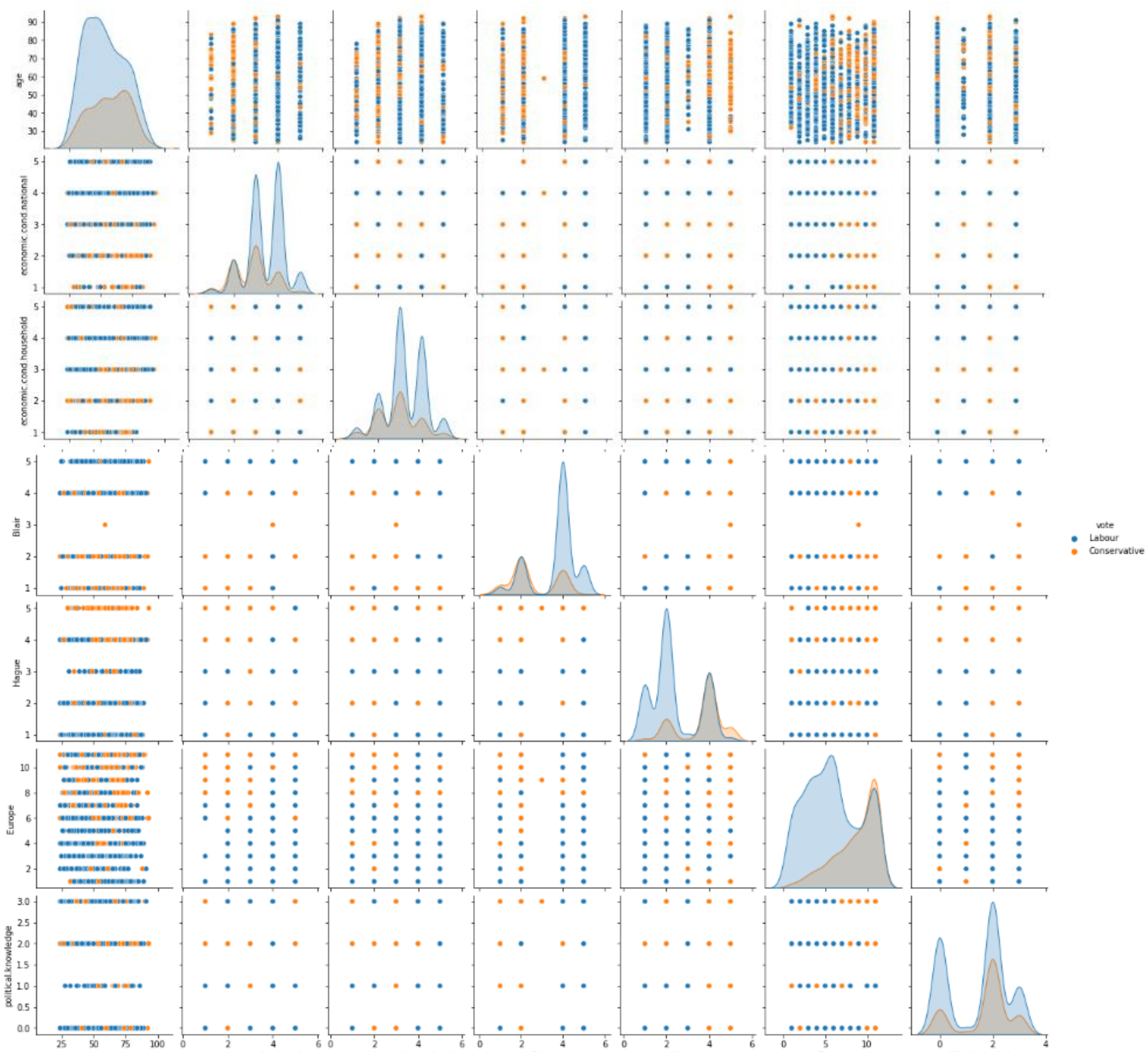


Figure 25: Pairplot

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

The data is encoded

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	age_bins	gender_male
0	Labour	43	3	3	4	1	2	2	40s	0
1	Labour	36	4	4	4	4	5	2	30s	1
2	Labour	35	4	4	5	2	3	2	30s	1
3	Labour	24	4	2	2	1	4	0	20s	0
4	Labour	41	2	2	1	1	6	2	40s	1

Table 8: Gender encoded data

	vote	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	age_bins	gender_male
0	Labour	3	3	4	1	2	2	2	0
1	Labour	4	4	4	4	5	2	1	1
2	Labour	4	4	5	2	3	2	1	1
3	Labour	4	2	2	1	4	0	0	0
4	Labour	2	2	1	1	6	2	2	1

Table 9: Age_bins encoded

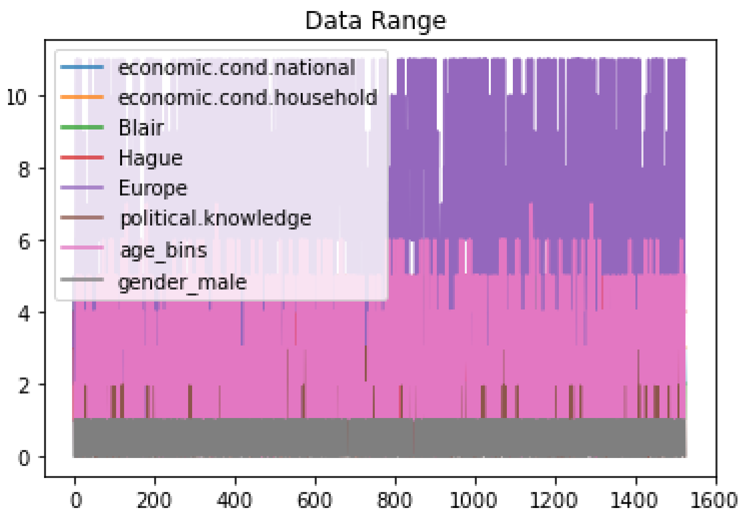


Figure 26: Data's variables' range

Since Above figure shows that points ranges are 0 -11 and most of the variables are ordinal variables so there is no need of scaling.

Independent (X) and dependent (Y) variables are segregated.
'Vote' is dependent variable.
Rest are independent variables.

	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	age_bins	gender_male
0	3	3	4	1	2	2	2	0
1	4	4	4	4	5	2	1	1
2	4	4	5	2	3	2	1	1
3	4	2	2	1	4	0	0	0
4	2	2	1	1	6	2	2	1

Table 10 :Head of Independent variables(X)

```
0    Labour
1    Labour
2    Labour
3    Labour
4    Labour
Name: vote, dtype: object
```

Table 11: Head of Dependent variable (Y)

Data is split into train and test in 70:30 ration using the train_test_split

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30 , random_state=1)
```

Figure 27: Data split:

Training data for Independent variables have 1061 entries with 8 columns, Dependent variable has 1061 columns.
Testing data for Independent variables have 456 entries with 8 columns, Dependent variable has 456 columns.

```
X_train:  (1061, 8)
X_test:  (456, 8)
y_train:  (1061,)
y_test:  (456,)
```

Figure 28: Train and test data shape

1.4 Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks)

LogisticRegression with solver 'newton-cg' is applied with maximum iteration of 10000

Logistic Regression - Training Data
Model score: 0.83 , Recall score for Labour: 0.64 , Recall score for Conservative: 0.91

Logistic Regression - Testing Data
Model score: 0.83 , Recall score for Labour: 0.73 , Recall score for Conservative: 0.88

AUC for testing data is 0.8899
AUC for testing data is 0.8833

Classification report for Logistic Regression-Training data:

	precision	recall	f1-score	support
Conservative	0.74	0.64	0.68	307
Labour	0.86	0.91	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.77	0.78	1061
weighted avg	0.83	0.83	0.83	1061

Table 12: Logistic regression-Classification report-Training data

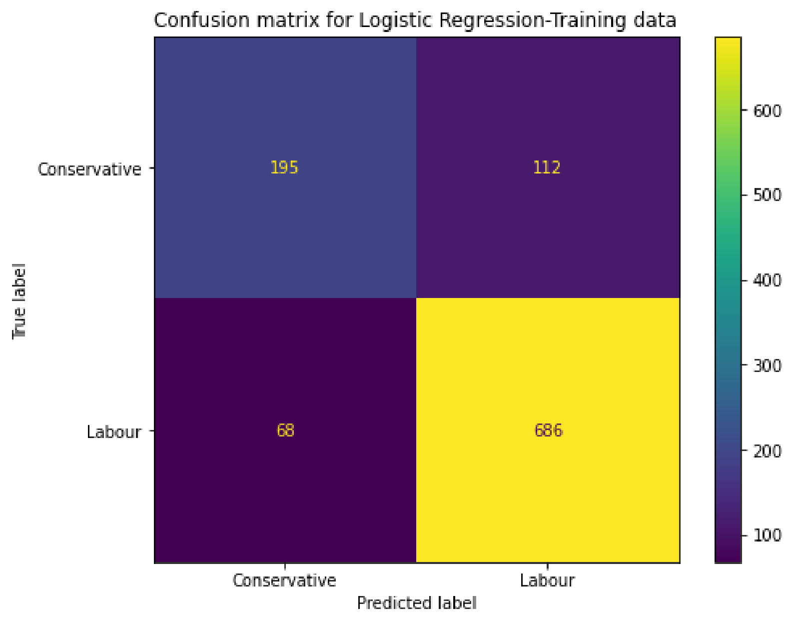


Figure 29: Logistic regression-Confusion matrix-Training data

AUC for Logistic Regression-Training data: 0.88993

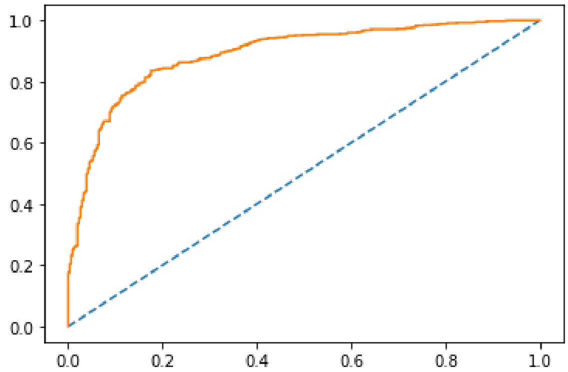


Figure 30: Logistic regression-AUC&ROC-Training data

Classification report for Logistic Regression-Testing data:

	precision	recall	f1-score	support
Conservative	0.76	0.73	0.74	153
Labour	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 13: Logistic regression-Classification report-Testing data

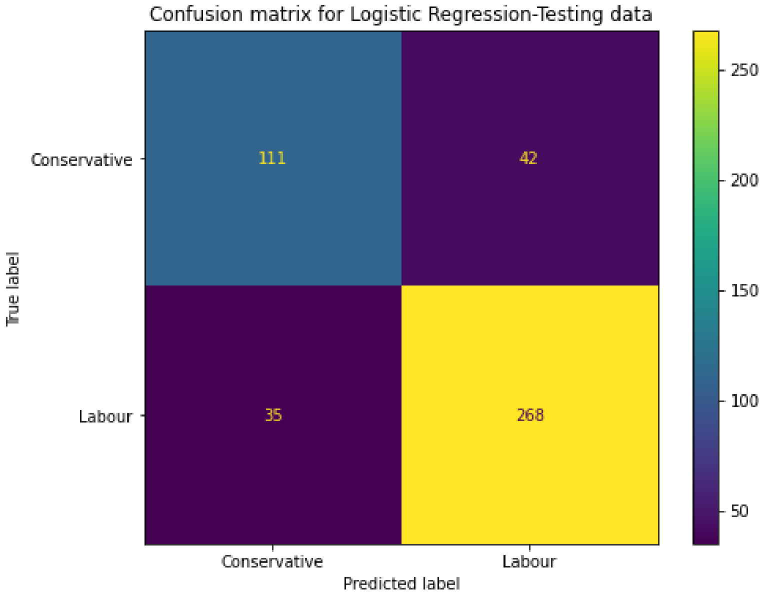


Figure 31: Logistic regression-Confusion matrix -Testing data

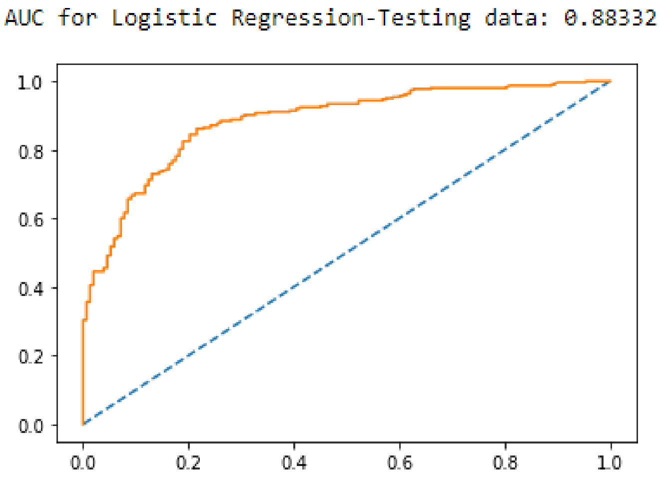


Figure 32: Logistic regression-AUC&ROC -Testing data

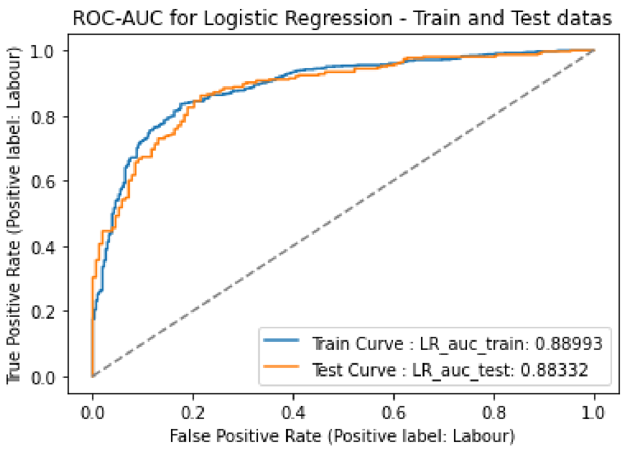


Figure 33: Logistic regression-AUC and ROC- Train and Test data

Upon tuning the LR model, AUC for train is 0.8899, AUC for test is 0.8832
Shows not much difference is made.

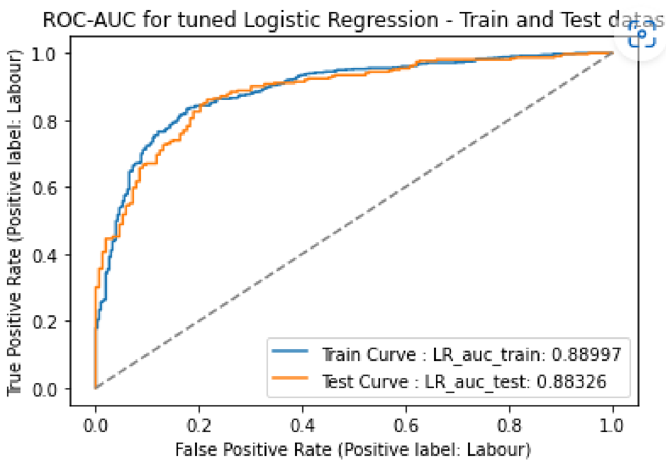


Figure 34: Tuned Logistic regression

Linear discriminant analysis is applied for the data

Linear Discriminant Analysis - Training Data
Model score: 0.83 , Recall score for Labour: 0.65 , Recall score for Conservative: 0.91

Linear Discriminant Analysis - Testing Data
Model score: 0.84 , Recall score for Labour: 0.73 , Recall score for Conservative: 0.89

AUC for training data is 0.889
AUC for testing data is 0.888

Classification report for Linear Discriminant Analysis model on Training data:

	precision	recall	f1-score	support
Conservative	0.74	0.65	0.69	307
Labour	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Table 14: LDA-Classification report -Training data

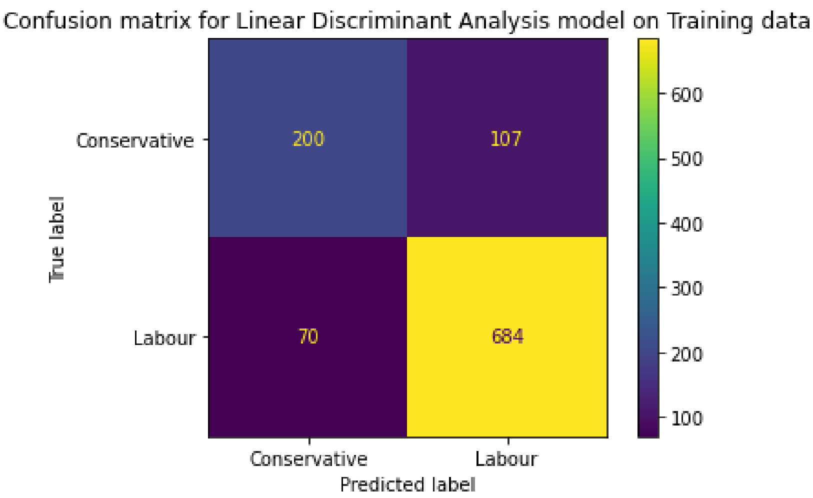


Figure 35:LDA-Confusion matrix-Training data

Classification report for Linear Discriminant Analysis model on Testing data:

	precision	recall	f1-score	support
Conservative	0.78	0.73	0.75	153
Labour	0.87	0.89	0.88	303
accuracy			0.84	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.84	0.84	0.84	456

Table 15: LDA-Classification report-Testing data

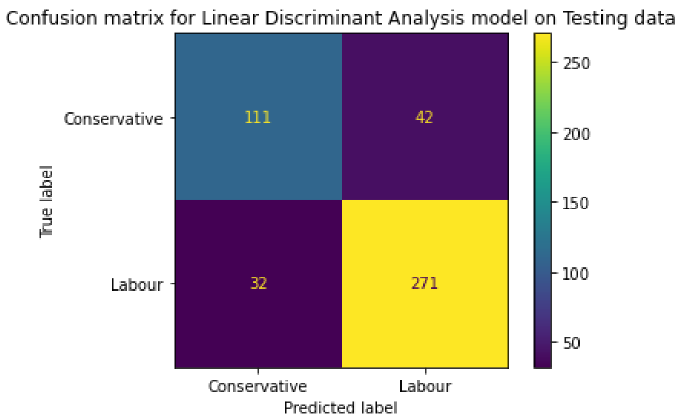


Figure 36: LDA-Confusion matrix-Testing data

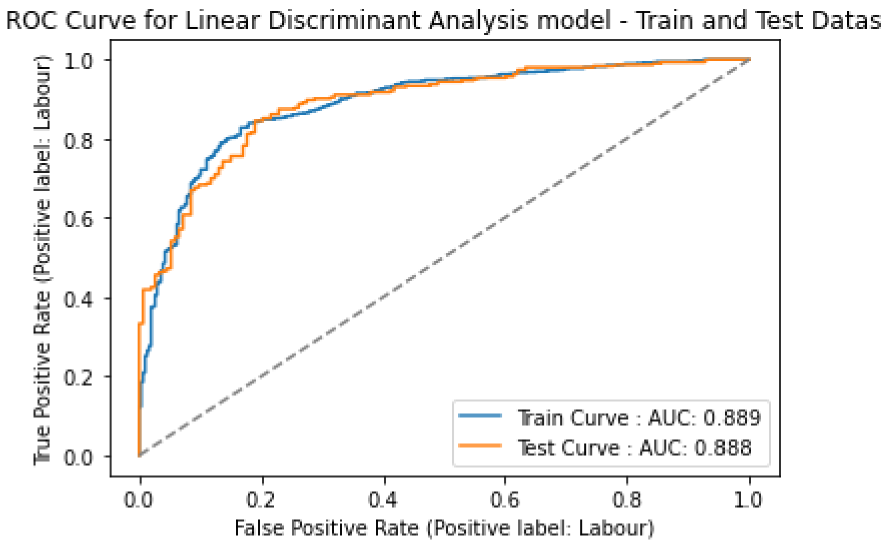


Figure 37: LDA- ROC & AUC - Train and test data

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results. (4 marks)

KNN is applied on the data

KNN on Training data:
Model score: 0.86 Recall for Conservative: 0.71 Recall for Labour: 0.92

KNN on Training data:
Model score: 0.81 Recall for Conservative: 0.65 Recall for Labour: 0.89

AUC for training data is 0.921
AUC for testing data is 0.892

Classification Report - KNN Model - Training data

	precision	recall	f1-score	support
Conservative	0.79	0.71	0.75	307
Labour	0.89	0.92	0.90	754
accuracy			0.86	1061
macro avg	0.84	0.82	0.83	1061
weighted avg	0.86	0.86	0.86	1061

Table 16: KNN-Classification report-Training data

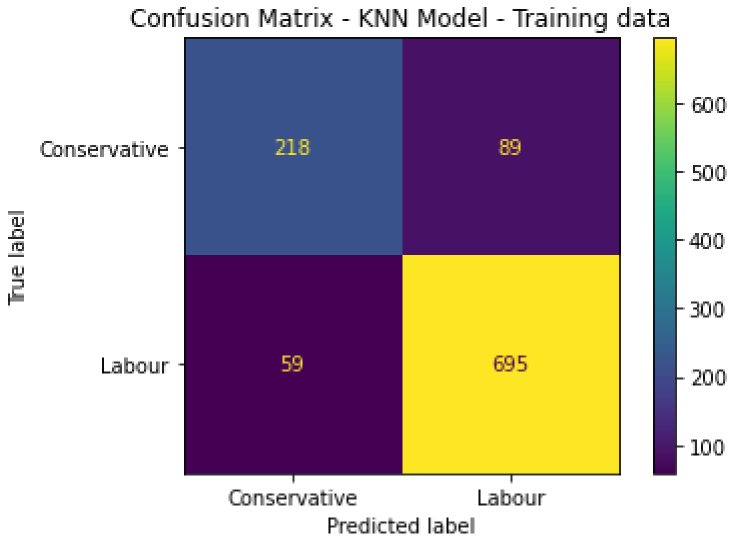


Figure 38: KNN-Confusion matrix-Training data

Classification report - KNN Model - Testing data

	precision	recall	f1-score	support
Conservative	0.75	0.65	0.70	153
Labour	0.84	0.89	0.86	303
accuracy			0.81	456
macro avg	0.79	0.77	0.78	456
weighted avg	0.81	0.81	0.81	456

Table 17: KNN-Classification report-Testing data

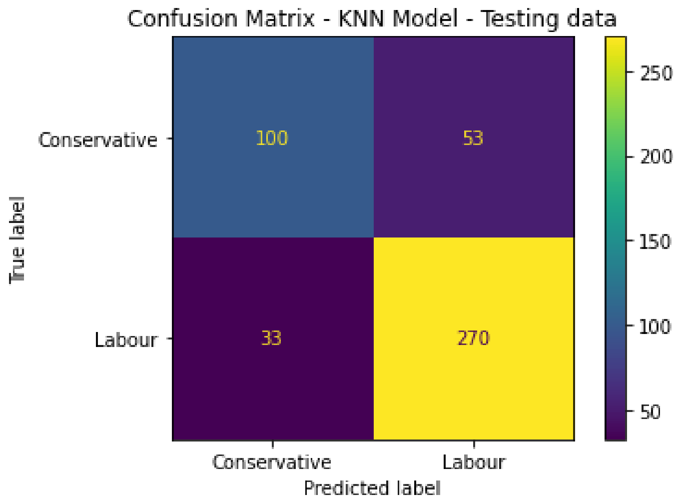


Figure 39: KNN-Confusion matrix-Testing data

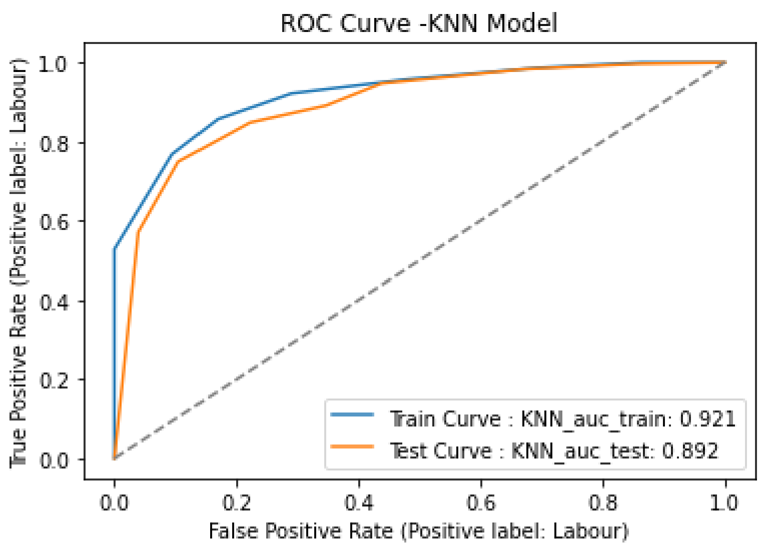


Figure 40: KNN- ROC & AUC- Train and test data

Gaussian Naïve Baye's is applied on the data

GaussianNB on Training data:
Score: 0.84 Recall for Conservative: 0.69 Recall for Labour: 0.89

GaussianNB on Testing data
Score: 0.82 Recall on Conservative: 0.73 Recall on Labour: 0.87

AUC for training data 0.888
AUC for testing data 0.877

Classification matrix for GaussianNB Model on Training data

	precision	recall	f1-score	support
Conservative	0.73	0.69	0.71	307
Labour	0.88	0.89	0.89	754
accuracy			0.84	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.84	0.83	1061

Table 18: GaussianNB-Classification report-Training data

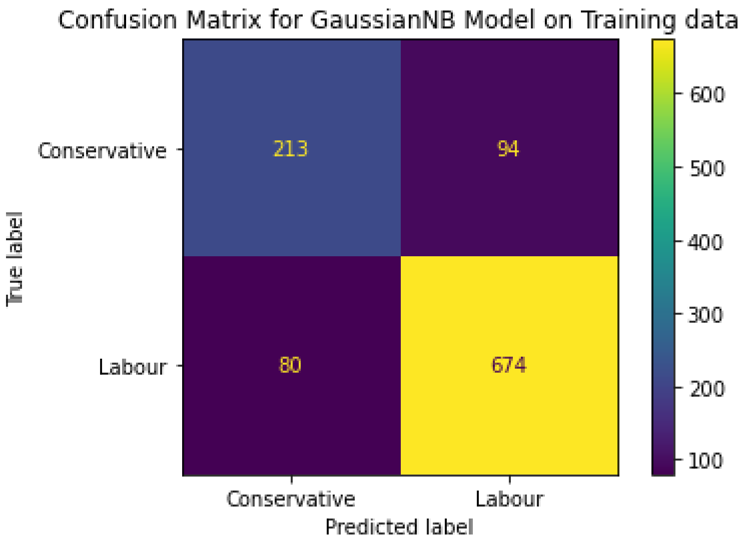


Figure 41: GaussianNB-Confusion matrix-Training data

Classification matrix for GaussianNB Model on Testing data

	precision	recall	f1-score	support
Conservative	0.74	0.73	0.73	153
Labour	0.86	0.87	0.87	303
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

Table 19: GaussianNB-Classification report-Testing data

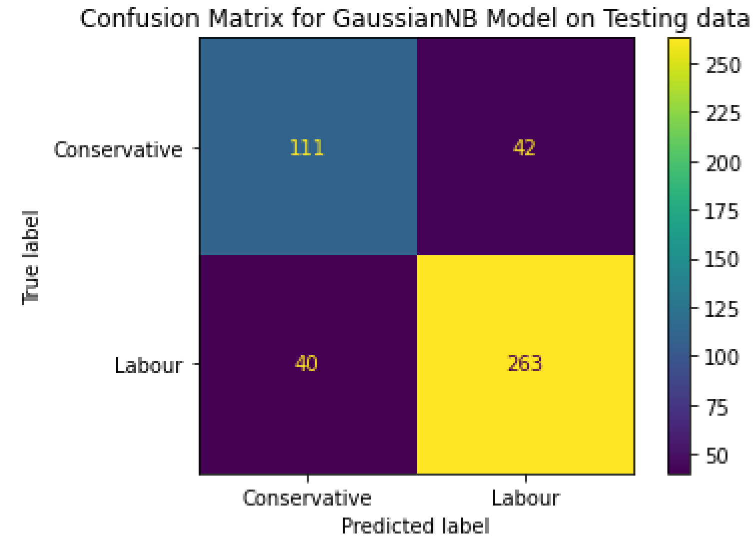


Figure 42: GaussianNB-Confusion matrix -Testing data

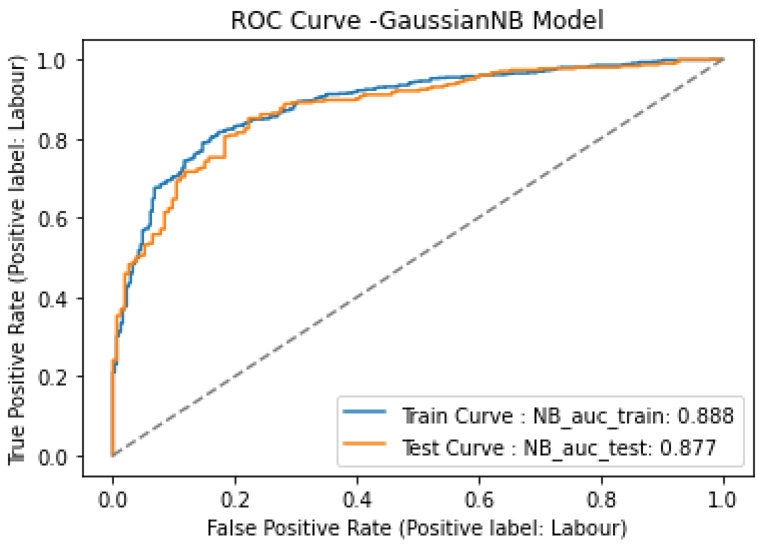


Figure 43: GaussianNB-AUC & ROC- Train and test data

The overall recall for the model looks better for GaussianNB
Model score of testig data is closer to training data for the same model.
KNN and Gaussian NB are very good models with GaussianNB being more preferred

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

Model tuning is a process of maximising a model's performance without overfitting or creating too high of a variance. This can be done by using hyperparameters. Methods like Grid Search, Random search and Bayesian optimization can be used for selecting appropriate hyperparameters.

Bagging or Bootstrap aggregating, is an ensemble learning technique that helps tp improve the performance and acuuracy of an ML algorithm. Bagging avoids overfitting of data specifically for decision tree algorithms. It is used for creating multiple models parallel.
It randomly samples the data with replacement and uses it for training

Boosting is an ensemble modelling technique that attempts to build a string classifier from a number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built using training data. The second model is build which tries to correct the errors in the first model. This continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.
The computational expense is very high
There are 2 kinds of boosting: Ada boosting and Gradient boosting

Bagging Classifier on Naïve Baye's is applied on the data-

Bagging NB Classifier on Training data:
Model score: 0.83 Recall for Conservative: 0.69 Recall for Labour: 0.89

Bagging NB Classifier on Testing data:
Model score: 0.82 Recall for Conservative: 0.72 Recall for Labour: 0.87

AUC for training data 0.888
AUC for testing data 0.877

	precision	recall	f1-score	support
Conservative	0.73	0.69	0.71	307
Labour	0.88	0.90	0.89	754
accuracy			0.84	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.84	0.83	1061

Figure 44:: BaggingNBClassifier-Classification report-Training data

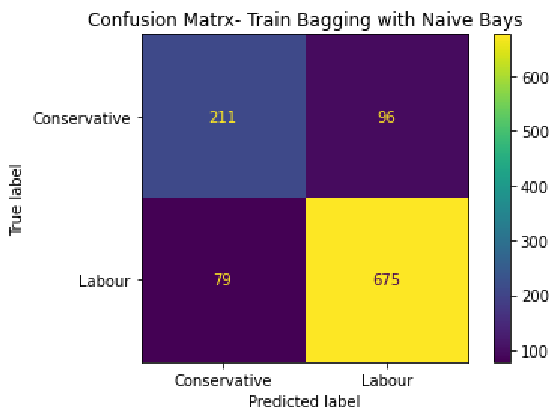


Figure 45: BaggingNBClassifier-Confusion matrixt-Training data

	precision	recall	f1-score	support
Conservative	0.74	0.73	0.73	153
Labour	0.86	0.87	0.87	303
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

Figure 46: BaggingNBClassifier-Classification report-Testing data

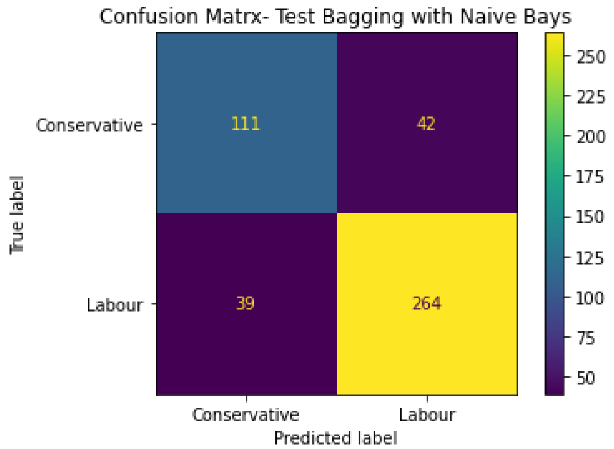


Figure 47: BaggingNBClassifier-Confusion matrix-Testing data

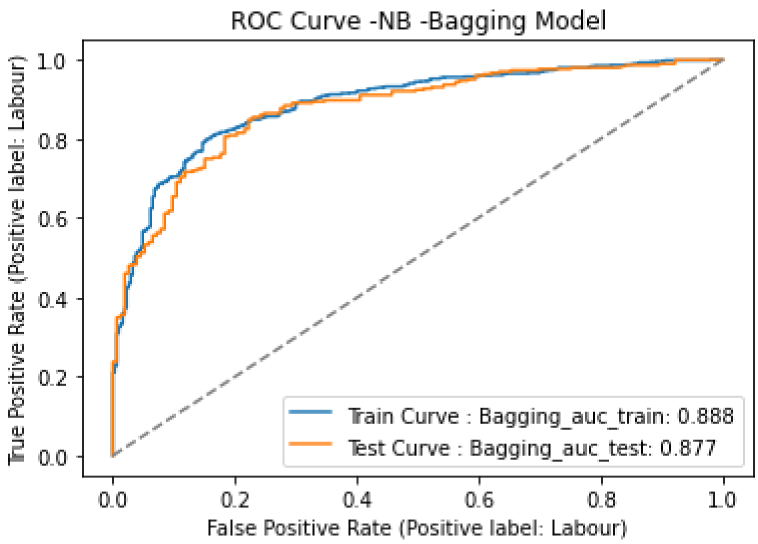


Figure 48: BaggingNBClassifier- ROC and AUC- Train and Test data

Random Forest Classifier is applied on the data-

Random Forest Classifier on Training data:
Model score: 0.99 Recall for Conservative: 0.98 Recall for Labour: 1.0

Random Forest Classifier on Testing data:
Model score: 0.81 Recall for Conservative: 0.64 Recall for Labour: 0.9

AUC for training data 0.888
AUC for testing data 0.877

	precision	recall	f1-score	support
Conservative	0.99	0.98	0.99	307
Labour	0.99	1.00	0.99	754
accuracy			0.99	1061
macro avg	0.99	0.99	0.99	1061
weighted avg	0.99	0.99	0.99	1061

Table 20: RandomForest-Classification report- Train data

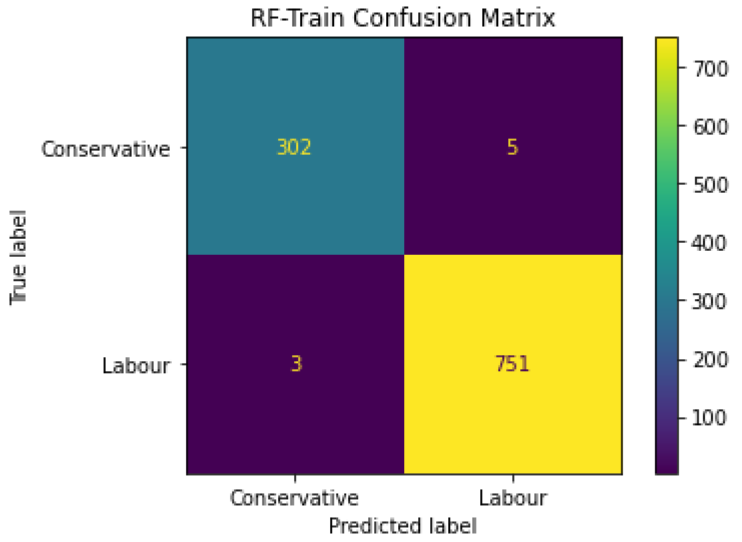


Figure 49: RandomForest-Confusion matrix - Train data

	precision	recall	f1-score	support
Conservative	0.76	0.64	0.70	153
Labour	0.83	0.90	0.86	303
accuracy			0.81	456
macro avg	0.80	0.77	0.78	456
weighted avg	0.81	0.81	0.81	456

Table 21 RandomForest-Classification report- Test data

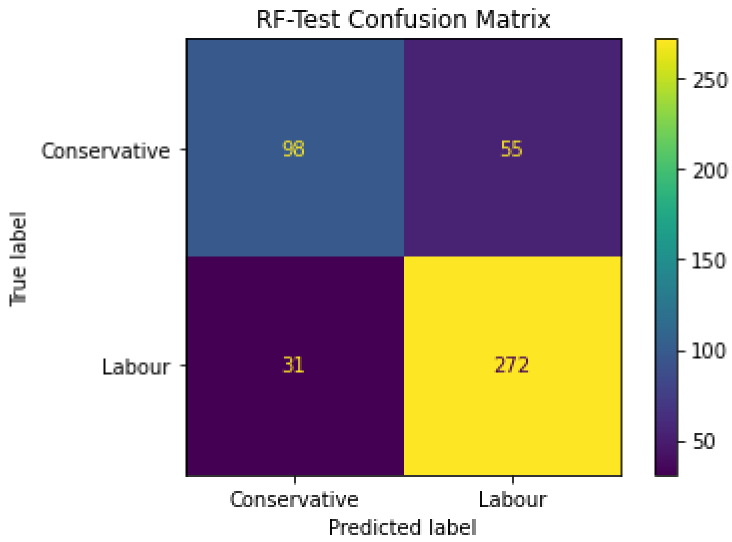


Figure 50: RandomForest-Confusion matrix - Train data

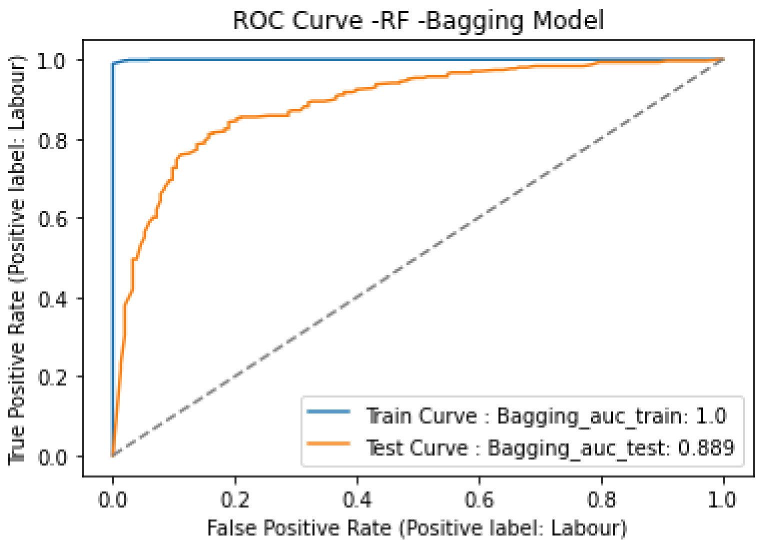


Figure 51: RandomForest-ROC & AUC- Train and test data

Bagging Classifier RandomForest is applied on the data-

Bagging RF Classifier on Training data:
Model score: 0.96 Recall for Conservative: 0.89 Recall for Labour: 0.99

Bagging RF Classifier on Testing data:
Model score: 0.83 Recall for Conservative: 0.67 Recall for Labour: 0.91

AUC for training data 0.996
AUC for testing data 0.894

	precision	recall	f1-score	support
Conservative	0.98	0.89	0.93	307
Labour	0.96	0.99	0.97	754
accuracy			0.96	1061
macro avg	0.97	0.94	0.95	1061
weighted avg	0.96	0.96	0.96	1061

Figure 52: BaggingRF-Classification report-Train data

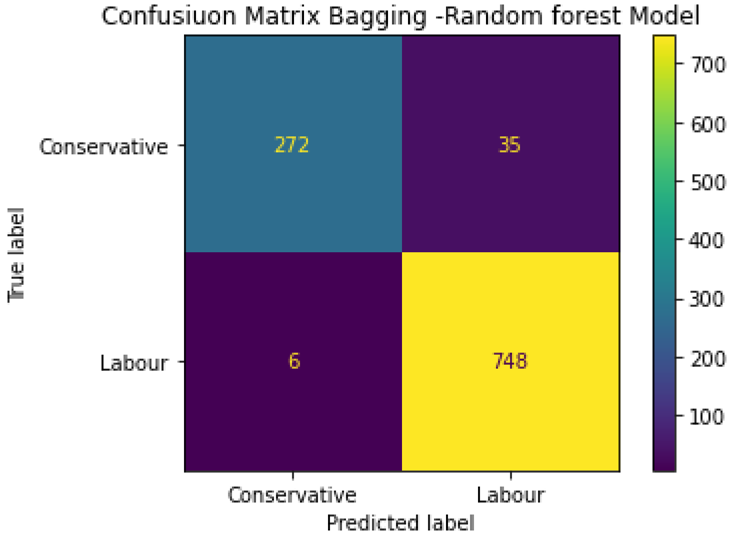


Figure 53: BaggingRF-Confusion matrix -Train data

	precision	recall	f1-score	support
Conservative	0.79	0.67	0.73	153
Labour	0.85	0.91	0.88	303
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.83	456

Figure 54: BaggingRF-Classification report-Test data

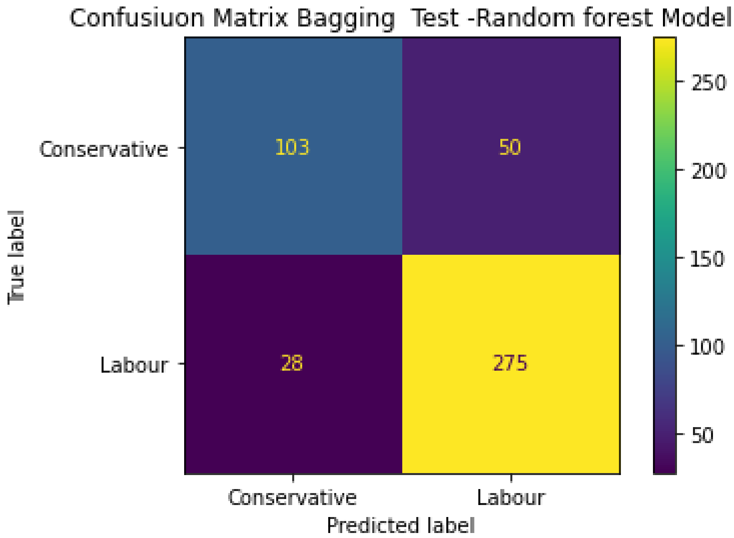


Figure 55: BaggingRF-Confusion matrix -Test data

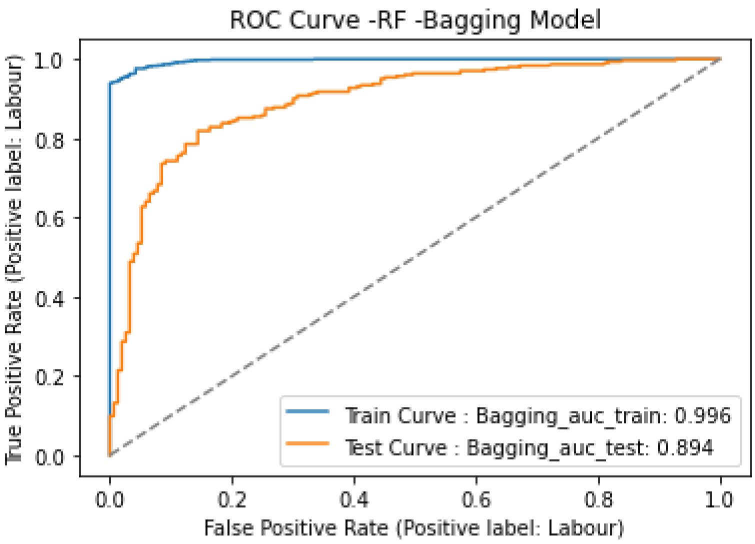


Figure 56: BaggingRF-ROC & AUC -Train and test data

Ada boosting Classifier is applied on the data-

Ada boosting Classifier on Training data:
Model score: 0.94 Recall for Conservative: 0.68 Recall for Labour: 0.91

Ada boosting Classifier on Testing data:
Model score: 0.82 Recall for Conservative: 0.69 Recall for Labour: 0.89

AUC for training data 0.906
AUC for testing data 0.88

	precision	recall	f1-score	support
Conservative	0.75	0.68	0.71	307
Labour	0.87	0.91	0.89	754
accuracy			0.84	1061
macro avg	0.81	0.79	0.80	1061
weighted avg	0.84	0.84	0.84	1061

Figure 57:AdaBoosting-Classification report-Train data

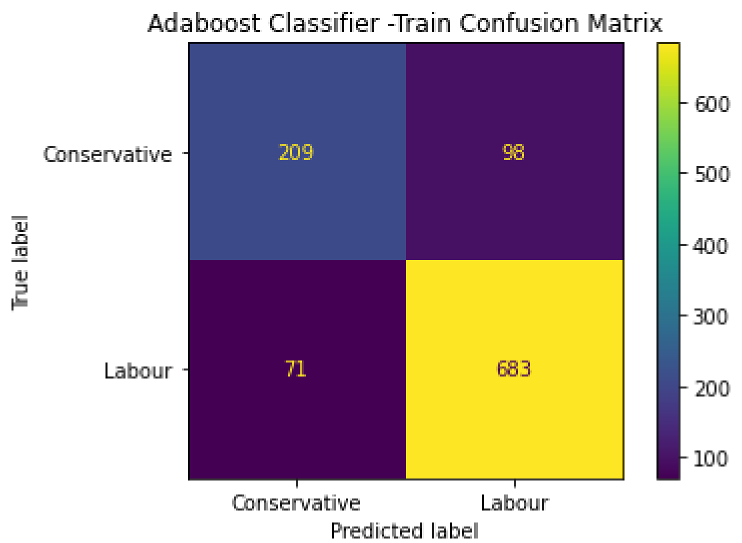


Figure 58: AdaBoosting-Confusion matrix-Train data

	precision	recall	f1-score	support
Conservative	0.76	0.69	0.72	153
Labour	0.85	0.89	0.87	303
accuracy			0.82	456
macro avg	0.80	0.79	0.80	456
weighted avg	0.82	0.82	0.82	456

Figure 59: AdaBoosting-Classification matrix-Test data

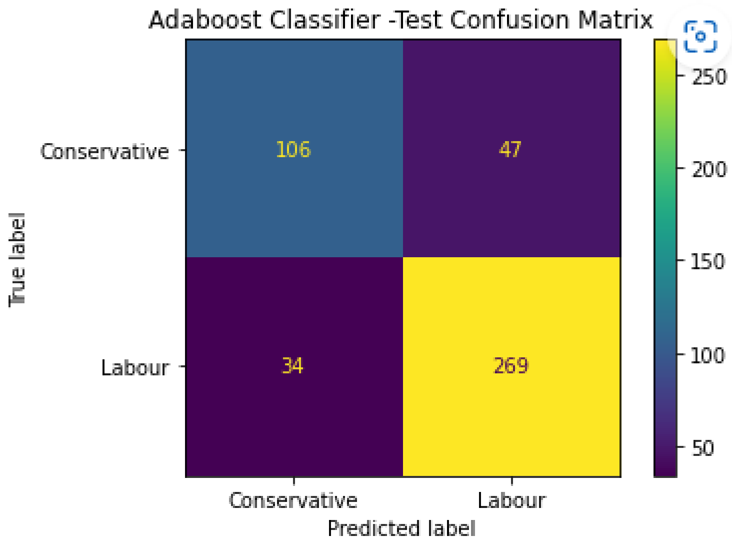


Figure 60: AdaBoosting-Confusion matrix-Test data

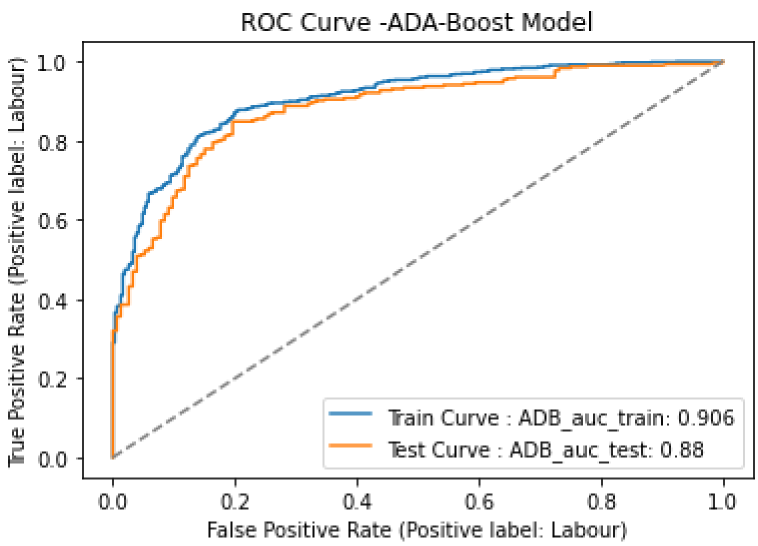


Figure 61: AdaBoosting-ROC & AUC-Train and test data

Gradient boosting Classifier is applied on the data-

Gradient boosting Classifier on Training data:
Model score: 0.89 Recall for Conservative: 0.77 Recall for Labour: 0.93

Gradient boosting Classifier on Testing data:
Model score: 0.83 Recall for Conservative: 0.69 Recall for Labour: 0.90

AUC for training data 0.956
AUC for testing data 0.895

	precision	recall	f1-score	support
Conservative	0.83	0.77	0.80	307
Labour	0.91	0.93	0.92	754
accuracy			0.89	1061
macro avg	0.87	0.85	0.86	1061
weighted avg	0.89	0.89	0.89	1061

Figure 62: GradientBoosting-Classification report-Train data

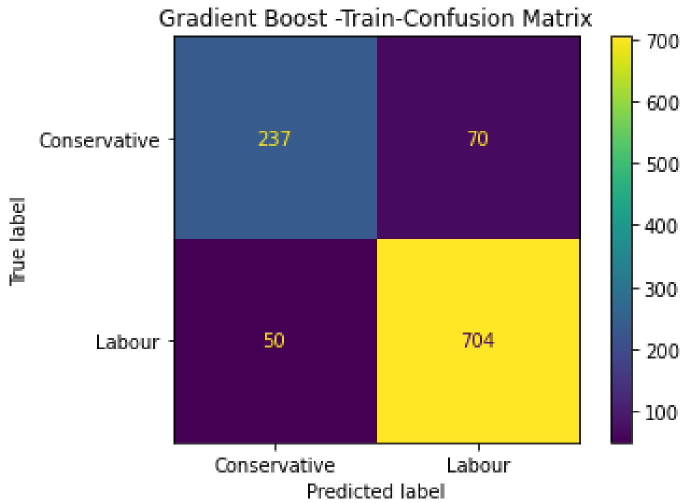


Figure 63: GradientBoosting-Confusion matrix -Train data

	precision	recall	f1-score	support
Conservative	0.78	0.69	0.73	153
Labour	0.85	0.90	0.88	303
accuracy			0.83	456
macro avg	0.82	0.80	0.80	456
weighted avg	0.83	0.83	0.83	456

Figure 64: GradientBoosting-Classification report-Test data

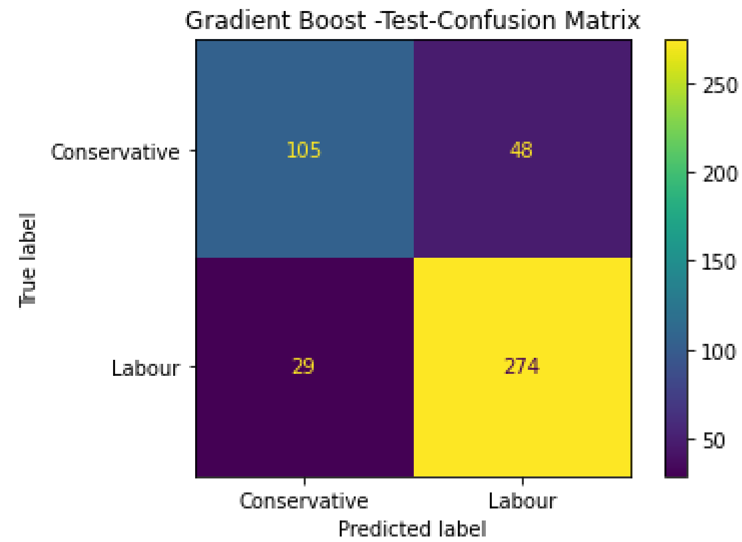


Figure 65: GradientBoosting-Confusion matrix-Test data

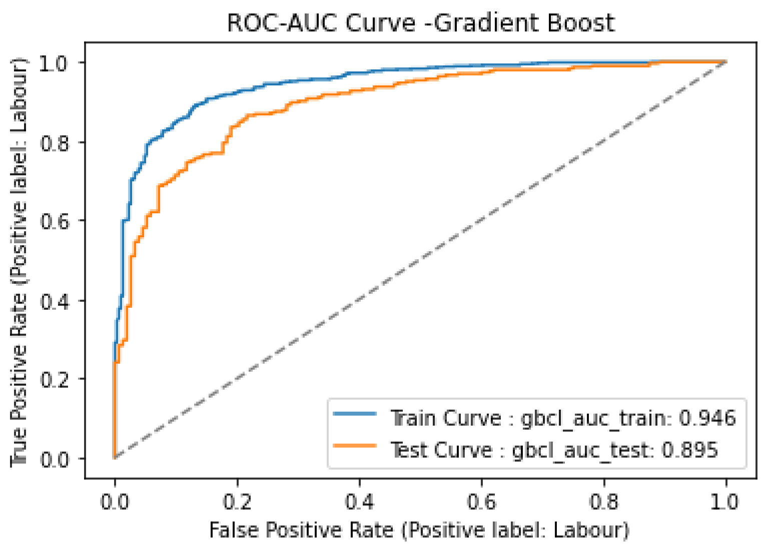


Figure 66: GradientBoosting-ROC & AUC -Train and test data

DecisionTree Classifier is applied on the data-

DecisionTree Classifier on Training data:
Model score: 0.99 Recall for Conservative: 1.0 Recall for Labour: 0.99

DecisionTree Classifier on Testing data:
Model score: 0.76 Recall for Conservative: 0.66 Recall for Labour: 0.82

AUC for training data 1.0
AUC for testing data 0.739

	precision	recall	f1-score	support
Conservative	0.97	1.00	0.99	307
Labour	1.00	0.99	0.99	754
accuracy			0.99	1061
macro avg	0.99	0.99	0.99	1061
weighted avg	0.99	0.99	0.99	1061

Figure 67: DecisionTree-Classification report-Train data

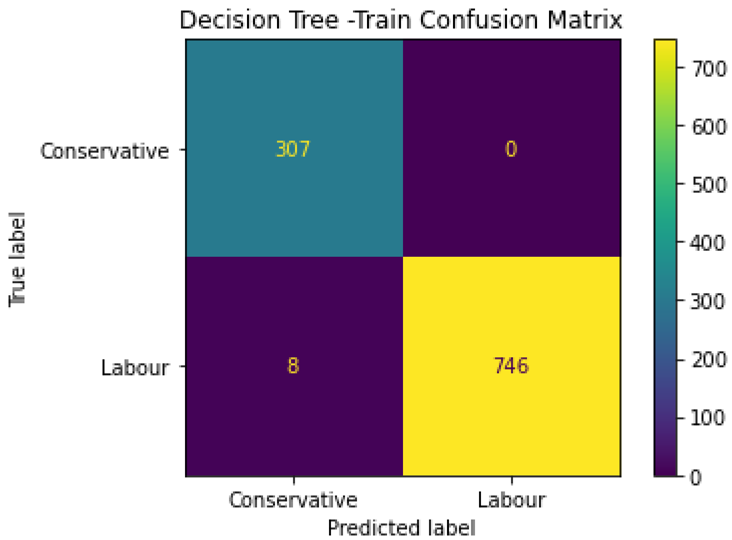


Figure 68: DecisionTree-Confusion matrix -Train data

	precision	recall	f1-score	support
Conservative	0.64	0.66	0.65	153
Labour	0.83	0.82	0.82	303
accuracy			0.76	456
macro avg	0.73	0.74	0.74	456
weighted avg	0.76	0.76	0.76	456

Figure 69: DecisionTree-Classification report-Test data

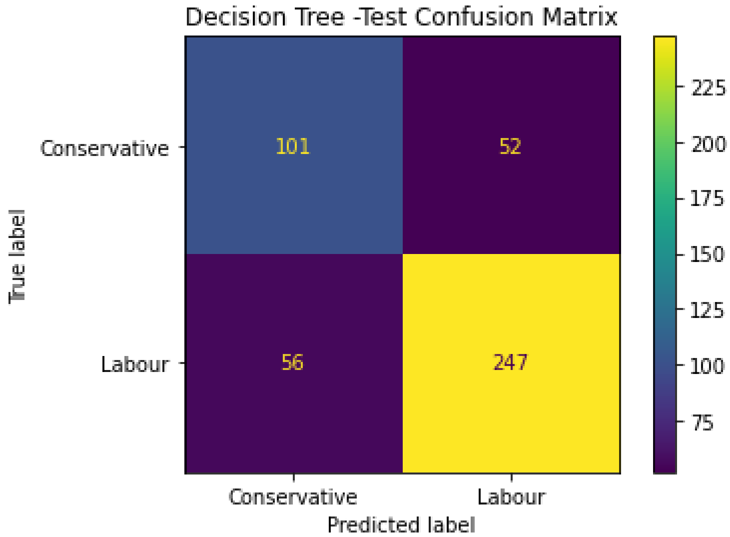


Figure 70: DecisionTree-Confusion matrix -Test data

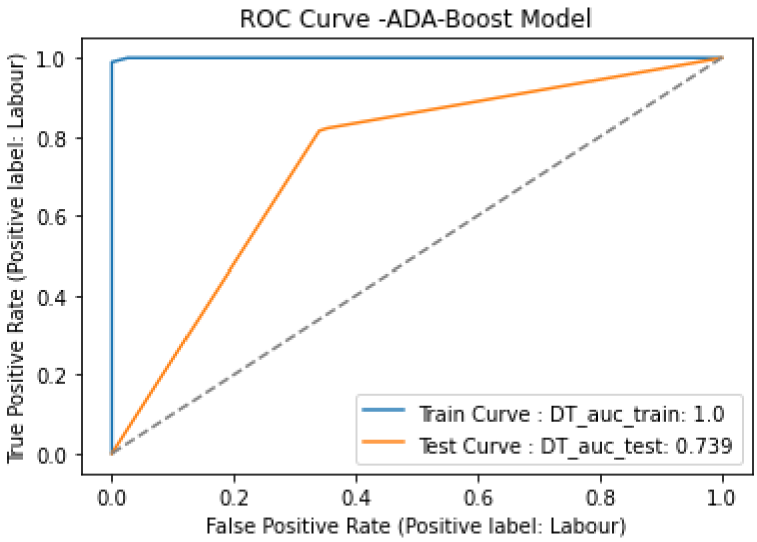


Figure 71: DecisionTree-ROC & AUC -Train and test data

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)

	Train Recall	Test Recall	Accuracy Train	Accuracy Test
Naive-Bayes	0.89	0.87	0.84	0.82
LDA	0.91	0.89	0.83	0.84
ADABoost	0.91	0.89	0.84	0.82
GradientBoost	0.93	0.90	0.89	0.83
KNN	0.92	0.89	0.86	0.81
DecisionTree	0.99	0.82	0.99	0.76
RF	1.00	0.90	0.99	0.81
Bagging	0.99	0.91	0.96	0.83

Table 22: Culmination of models

So as per the test data, best performing model is - Linear Discriminant Analysis

Best Performing models are - Decision Tree , Random Forest and Bagging

However are these best performing models overfitted??

Let's look at the performance on the test data set

So we will select models which have performed approximately similar on the train and test data set and apply SMOTE on the same to check if the performance improves or not eg. Naive Bayes and KNN

SMOTE (Synthetic Minority Oversampling Technique) is a statistical technique for increasing the number of cases in minority of a dataset. It is used to treat data imbalance.

Naïve Baye's with SMOTE-
Model score for train data: 0.83
Model score for test data: 0.80

	precision	recall	f1-score	support
Conservative	0.83	0.83	0.83	754
Labour	0.83	0.83	0.83	754
accuracy			0.83	1508
macro avg	0.83	0.83	0.83	1508
weighted avg	0.83	0.83	0.83	1508

Table 23: NB with SMOTE-Classification test-Train data

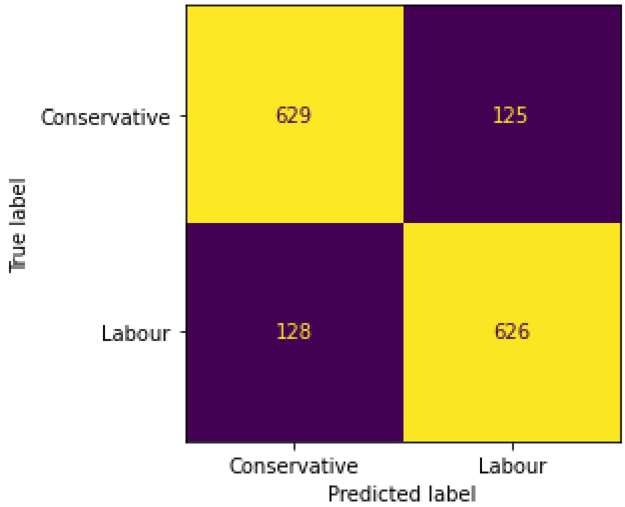


Figure 72: NB wit SMOTE-Confusion matrix-Train data

	precision	recall	f1-score	support
Conservative	0.67	0.78	0.72	153
Labour	0.88	0.80	0.84	303
accuracy			0.80	456
macro avg	0.77	0.79	0.78	456
weighted avg	0.81	0.80	0.80	456

Table 24: NB with SMOTE-Classification report-Test data

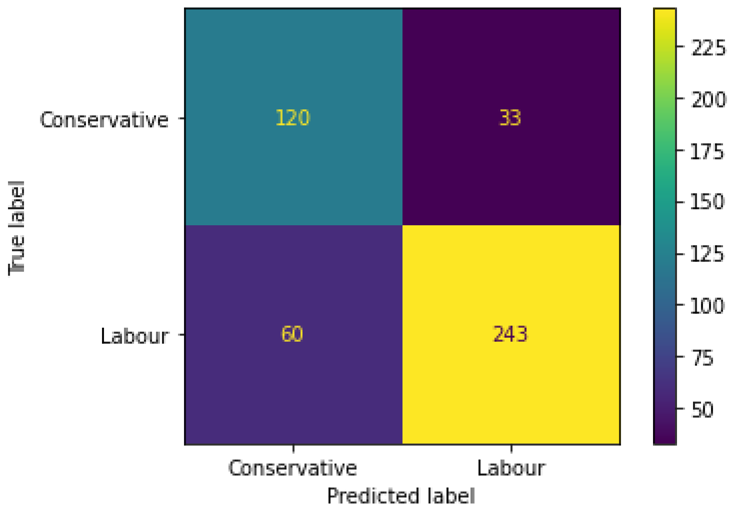


Figure 73: NB with SMOTE-Confusion matrix-Test data

KNN with SMOTE-

Model score with train data: 0.89

Model score with test data: 0.81

	precision	recall	f1-score	support
Conservative	0.86	0.95	0.90	754
Labour	0.94	0.84	0.89	754
accuracy			0.90	1508
macro avg	0.90	0.90	0.90	1508
weighted avg	0.90	0.90	0.90	1508

Table 25: KNN with SMOTE-Classification report-Train data

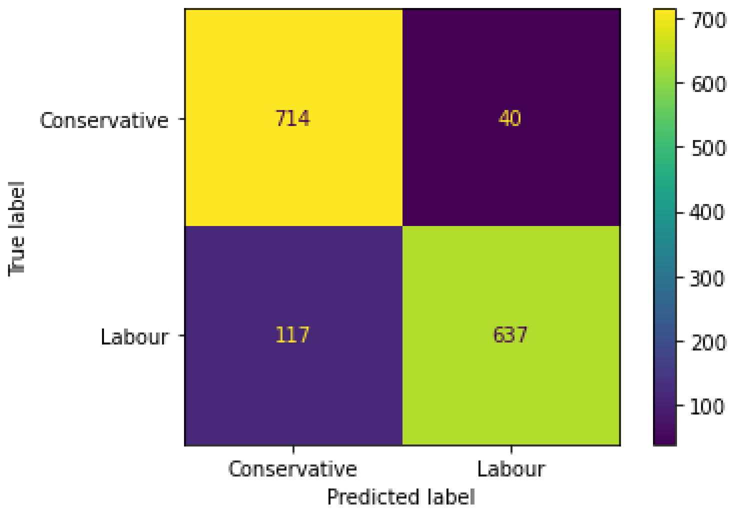


Figure 74: KNN with SMOTE-Confusion matrix -Train data

	precision	recall	f1-score	support
Conservative	0.69	0.80	0.74	153
Labour	0.89	0.82	0.85	303
accuracy			0.81	456
macro avg	0.79	0.81	0.80	456
weighted avg	0.82	0.81	0.81	456

Table 26: KNN with SMOTE-Classification report-Test data

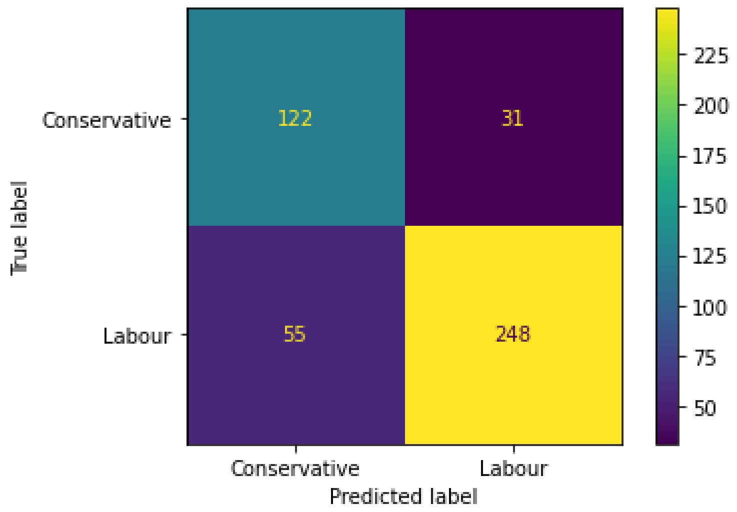


Table 27: KNN with SMOTE-Confusion matrix -Test data

Conclusion after SMOTE- Recall for Naive Bayes decresed significantly. Huge Difference between the train and test dataset Recall value,Accuracy for KNN

	Accuracy Train	Accuracy Test
Naive-Bayes SMOTE	0.832228	0.796053
KNN SMOTE	0.895889	0.811404

Table 28: SMOTE models

Cross Validation on Naive Bayes Model:

Accuracy score - 0.76821192, 0.82119205, 0.82781457, 0.78807947, 0.87417219, 0.85430464, 0.81456954, 0.87417219, 0.79333333, 0.86

Recall scores across all iterations of 10 folds - - 0.76821192, 0.82119205, 0.82781457, 0.78807947, 0.87417219, 0.85430464, 0.81456954, 0.87417219, 0.79333333, 0.86

Average recall score across all iterations of 10 fold cv - 0.83

After 10 fold cross validation, scores both on data set a for all 10 folds are almost same.
Hence our model is valid

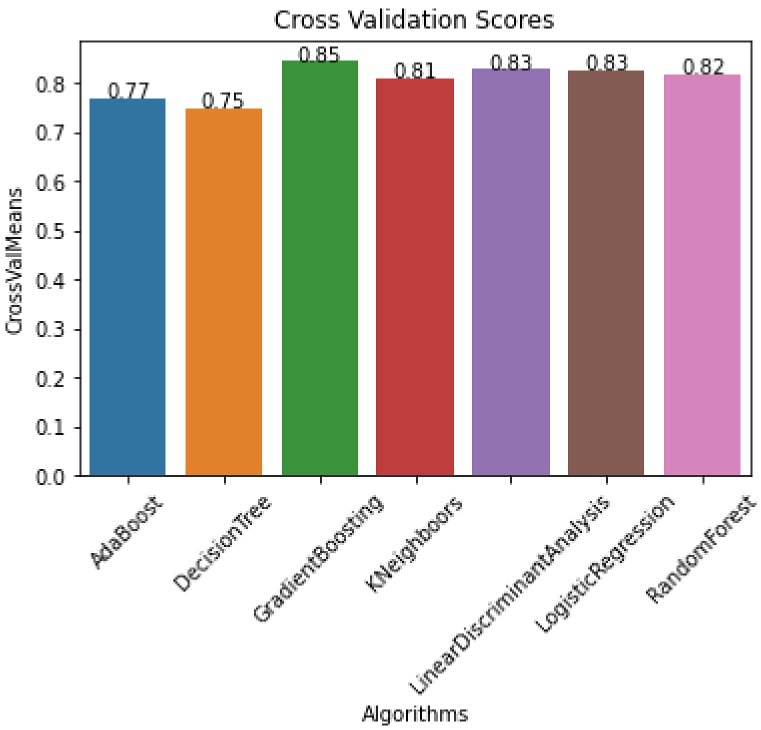


Figure 75: Cross validation scores across models

	Algorithms	CrossValMeans	CrossValErrors
0	AdaBoost	0.770984	0.040393
1	DecisionTree	0.748360	0.026929
2	GradientBoosting	0.845415	0.037358
3	KNeighboors	0.811524	0.024736
4	LinearDiscriminantAnalysis	0.828487	0.030502
5	LogisticRegression	0.827526	0.025642
6	RandomForest	0.819053	0.034136

Table 29: CrossValidation means and errors across models

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

Overall, 'Labour' has the highest chance of winning

Their vote bank is the population in their 30s and 40s

The party needs to focus on senior citizens to affect their vote bank positively

When economic.cond.national score is 3 or 4, they are most likely to vote for Labour

When Blair score is 4, they are most likely to go for Labour

Focusing on population in Hague category 1 and 2 will have a positive impact on the vote bank

'Conservative' has very weak chances of winning

An improvement can be seen by targeting voters in their 60s and 70s

Economic.cond.national of level has promising results towards this party with room for improvement

Voters with Blair score 4 and 5 are least likely to go with this party, this can be tackled as a long term goal

Hague score 4 and 5 show promising results and can be focused upon

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. (3 Marks)

1. President Franklin D. Roosevelt in 1941 -

Number of characters: 7571
Number of words: 1360
Number of sentences: 69

2. President Richard Nixon in 1973 -

Number of characters: 9991
Number of words: 1819
Number of sentences: 70

3. President John F. Kennedy in 1961 -

Number of characters: 7618
Number of words: 1390
Number of sentences: 56

2.2 Remove all the stopwords from all three speeches. (3 Marks)

Stopwords from nltk.corpus is imported and used as stopwords. The txt document is split at stopwords to remove them.
Later on, punctuations are removed too. The text is converted into all lower case.
Stemming is done, where the words are converted to their base word
Eg; Running, Runner -> Run

President Franklin D. Roosevelt in 1941-

	Text	totalwords	char_count	avg_word	No_of_stopwords
0	on each national day of inauguration since 178...	1360	7571	4.539706	632

Table 30: President Franklin D. Roosevelt in 1941- analysis

Original speech:

"On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn Washington\'s day the task of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people was to preserve that Nation from disruption from within.\n\nIn this day the task of the people is to save that Nation and its institutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a moment and take stock -- to recall what our place in history has been, and to rediscover what we are and what we may be. If we do not, we risk the real peril of inaction.\n\nLives of nations 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 more, a little less. The life of a nation is the fullness of the measure of its will to live.\n\nThere are men who doubt this. There are men who believe that democracy, as a form of Government and a frame of life, is limited or measured by a kind of mystical and artificial fate that, for some unexplained reason, tyranny and slavery have become the surging wave of the future -- and that freedom is an ebbing tide.\n\nBut we Americans know that this is not true.\n\nEight years ago, when the life of this Republic.."

Speech after processing looks something like this:

"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 us 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 us 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… "

President Richard Nixon in 1973-

	Text	totalwords	char_count	avg_word	No_of_stopwords
0	mr vice president mr speaker mr chief justice ...	1819	9991	4.465091	899

Table 31: President Richard Nixon in 1973-analysis

Original speech:

“Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.\n\nLet 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.\n\nThis past year saw far-reaching results from our 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 the 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 remembered as the year of the greatest progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flimsy peace which is merely an interlude between wars, but a peace…”

Speech after processing looks something like this :

“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 us shall use peace let us resolve era enter postwar periods often time retreat isolation leads stagnation home invites new danger abroad let us 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...”

President John F. Kennedy in 1961-

	Text	totalwords	char_count	avg_word	No_of_stopwords
0	vice president johnson mr speaker mr chief jus...	1390	7618	4.461871	618

Table 32: President John F. Kennedy in 1961-analysis

Original speech:

"Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well as a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our forebears l prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for which our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity of the state, but from the hand of God.\n\nWe dare not forget today that we are the 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 by a hard and bitter peace, proud of our ancient heritage -- and unwilling to witness.."

Speech after preprocessing looks something like this:

"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 let 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 let every nation know whether wishes us well ill shall pay price bear burden…"

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)

Top 3 most occurring words in the speech of President Franklin D. Roosevelt in 1941 are : 'nation', 'know', 'demoracy' with frequency of 11,10,9 respectively

```
nation       11
know         10
democracy     9
spirit        9
life          8
us            8
people        7
america       7
years         6
freedom       6
dtype: int64
```

Table 33: President Franklin D. Roosevelt in 1941-Most occurring words

Top 3 most occurring words in the speech of President Richard Nixon in 1973 are : 'us', 'let', 'peace' with frequency of 26,22,19 respectively

```
us                26
let               22
peace             19
world             16
new               15
america           13
responsibility    11
government        10
great              9
home               9
dtype: int64
```

Table 34: President Richard Nixon in 1973-Most occuring words

Top 3 most occurring words in the speech of President John F. Kennedy in 1961 are : 'let', 'us', 'sides' with frequency of 16,12,8 respectively

```
let         16
us          12
sides        8
world        8
pledge       7
new          7
ask          5
citizens     5
nations      5
free         5
dtype: int64
```

Table 35: President John F. Kennedy in 1961-Most occurring words

2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) [refer to the End-to-End Case Study done in the Mentored Learning Session] (3 Marks)

Word cloud of President Franklin D. Roosevelt in 1941

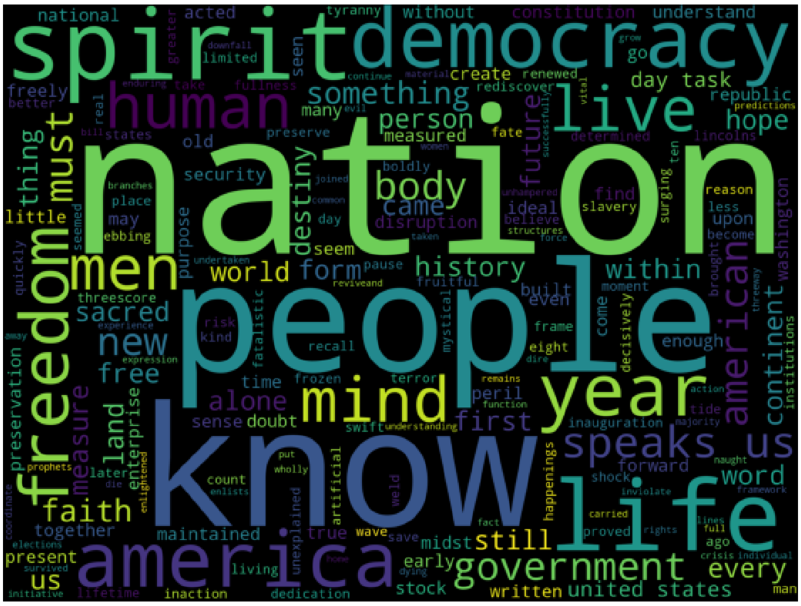


Figure 76: Word cloud of President Franklin D. Roosevelt in 1941

Word cloud of President Richard Nixon in 1973

Figure 77: Word cloud of President Richard Nixon in 1973

Word cloud of President John F. Kennedy in 1961:

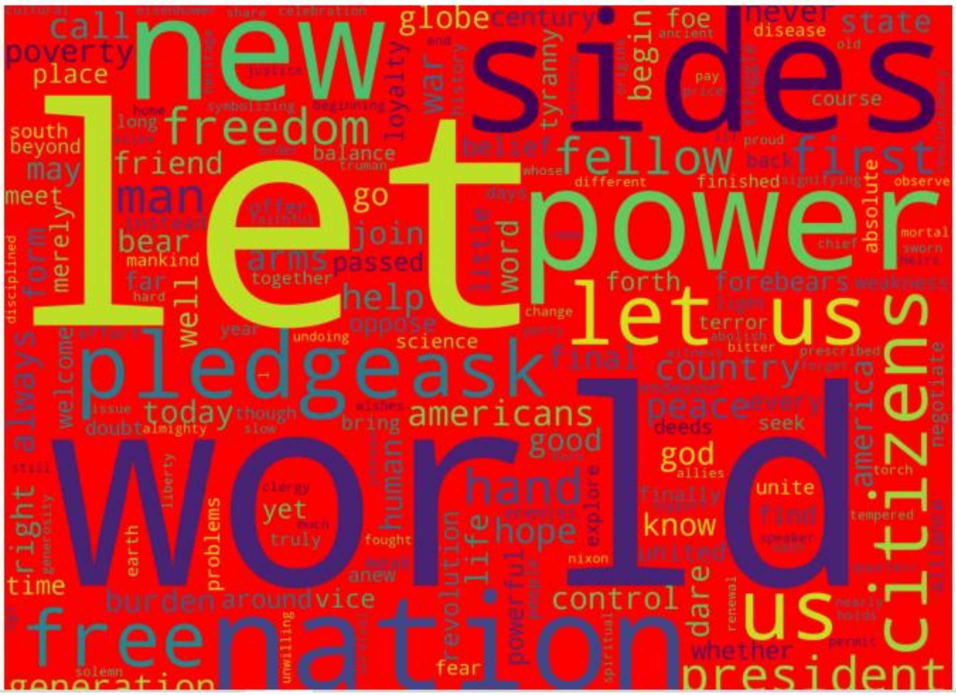


Figure 78: Word cloud of President John F. Kennedy in 1961