

Machine Learning - Project Report

Problem 1: Analyze Recent Elections	2
1.1	2
1.2	4
1.3	8
1.4	8
1.5	9
1.6	9
1.7	10
1.8	15
Problem 2: Speeches of the Presidents	16
2.1	16
2.2	17
2.3	17
2.4	18

Problem 1: Analyze Recent Elections

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.

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

Ans 1.1

The data is read using the read excel function in python.

Viewing the first few rows:

	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

Descriptive Statistics:

	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

Checking for Null Values:

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

Viewing the number of duplicate rows:

```
Total no. of duplicates values = 8
```

Viewing the Shape:

```
(1525, 9)
```

Seeing the distribution of the vote and gender columns:

Vote:

```
Labour      1063
Conservative  462
Name: vote, dtype: int64
```

Gender:

```
female      812
male        713
Name: gender, dtype: int64
```

Seeing the data types:

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

1.1 Inference:

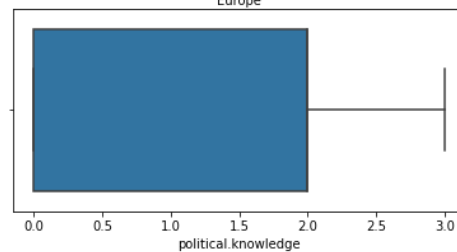
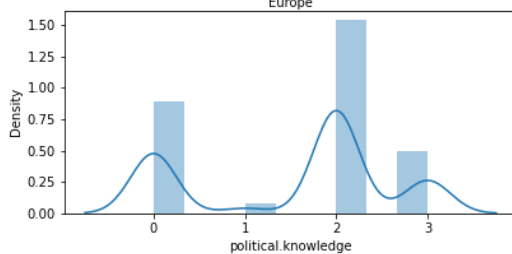
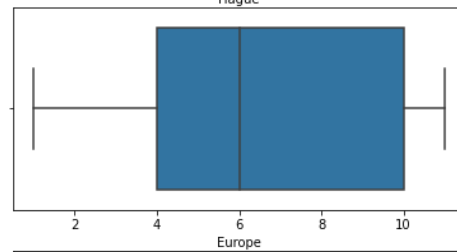
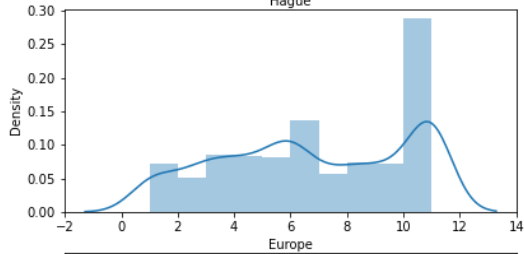
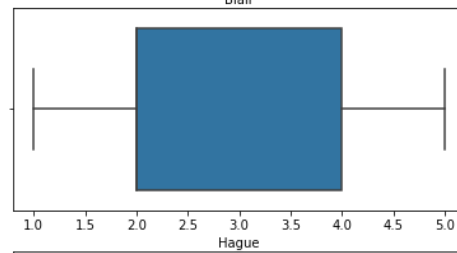
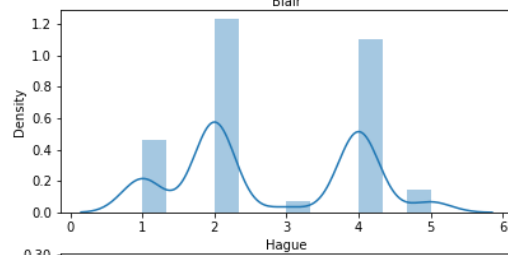
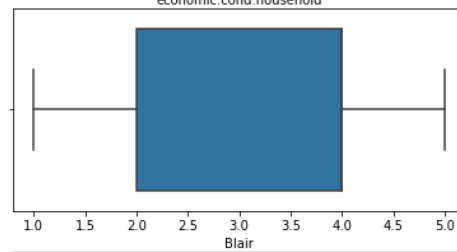
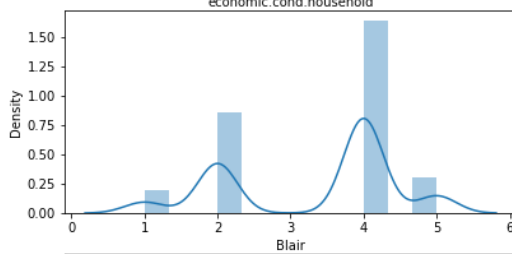
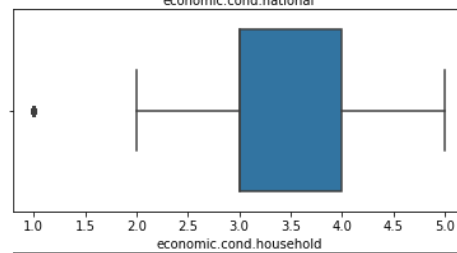
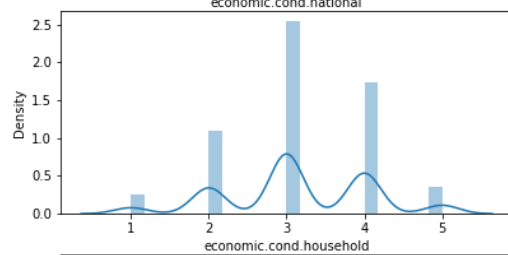
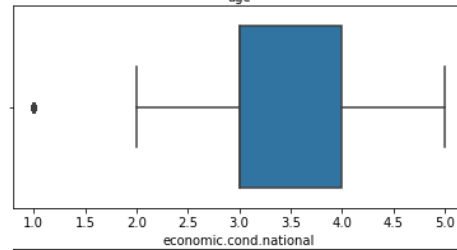
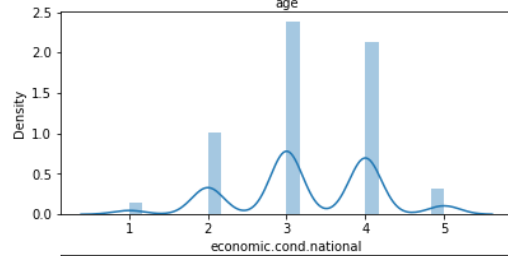
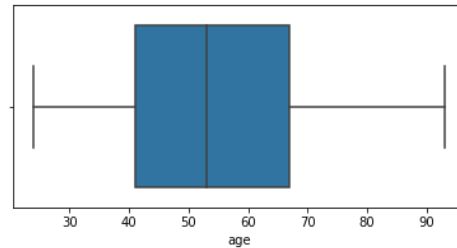
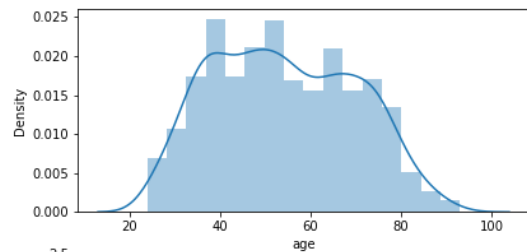
- There are no Null Values.
- Two columns - gender and vote have string data, and hence are categorical.
- Most of the votes were given to the Labour party.
- There are 8 duplicate rows.
- There are 1525 rows and 9 columns.

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

Ans 1.2

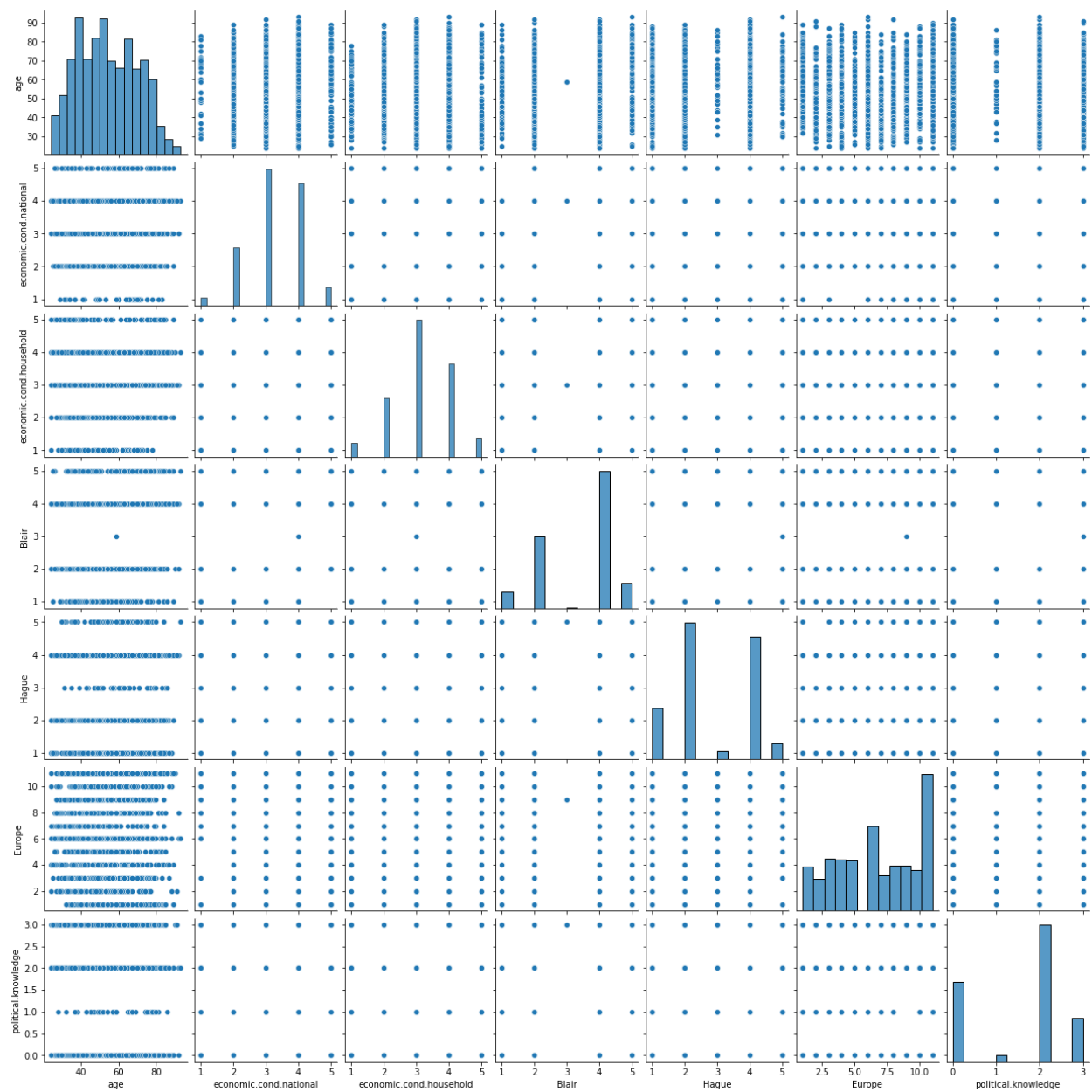
Univariate Analysis:

Dist Plot & Boxplot:

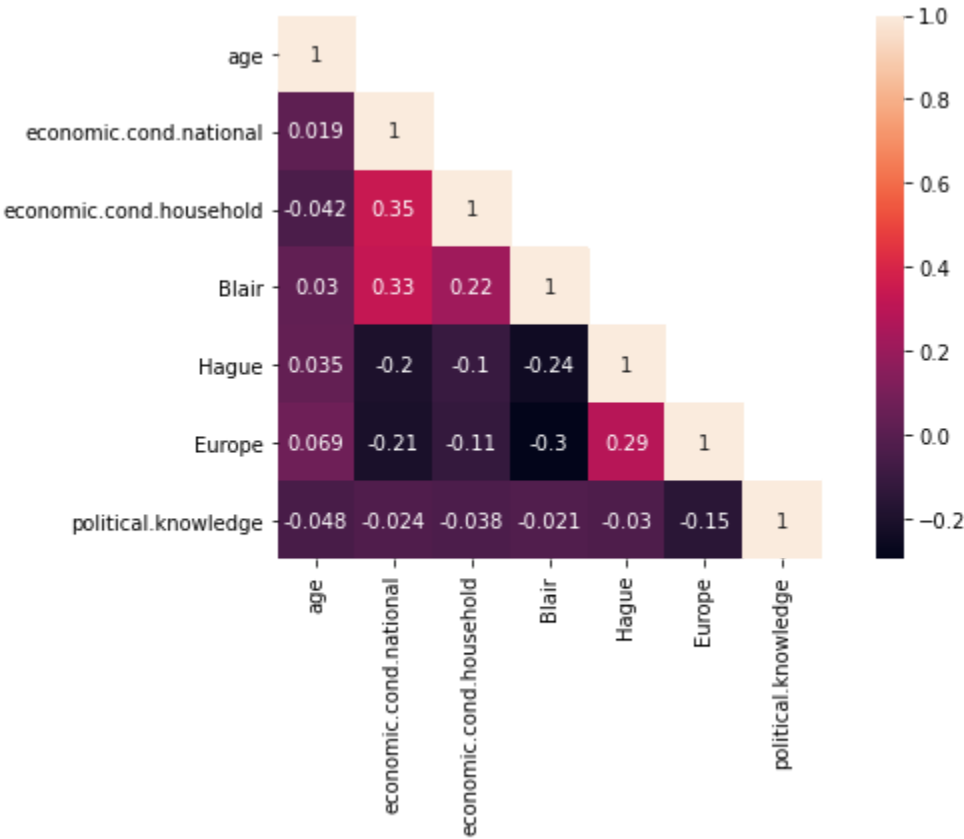


Multi-Variate Analysis:

Pairplot:



Correlation Plot:



Outlier Analysis:

	Outlier %
age	0.00
economic.cond.national	2.43
economic.cond.household	4.26
Blair	0.00
Hague	0.00
Europe	0.00
political.knowledge	0.00

1.2 Inference:

Density is the only variable that is not evenly distributed.

- The column `economic.cond.national` has a medium correlation with Europe, Blair and `economic.cond.household`.
- The column `economic.cond.household` has a medium correlation with Blair.
- The column Blair has a medium correlation with Europe and Hague.
- The column Hague has a medium correlation with Europe.
- The column Europe has a medium correlation with `political.knowledge`.
- `Economic.cond.household` and `economic.cond.national` have outliers while the other columns don't have outliers.

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).

Ans 1.3

The columns Vote & Gender have been encoded using `pd.get_dummies`:

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
0	43	3	3	4	1	2	2	1	0
1	36	4	4	4	4	5	2	1	1
2	35	4	4	5	2	3	2	1	1
3	24	4	2	2	1	4	0	1	0
4	41	2	2	1	1	6	2	1	1

No Scaling is not required in this case as the numbers are more or less in the same range.

The data has been split into train and test using the `train_test_split` function, `vote_labor` is the target column.

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

Ans 1.4

The Logistic Regression Model has been applied using the `LogisticRegression` package from the SK Learn library.

The Linear Discriminant Analysis Model has been applied using the LinearDiscriminantAnalysis package from the SK Learn library.

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

Ans 1.5

The KNN Model has been applied using the KNeighborsClassifier package from the SK Learn library.

The Naïve Bayes Model has been applied using the GaussianNB package from the SK Learn library.

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

Ans 1.6

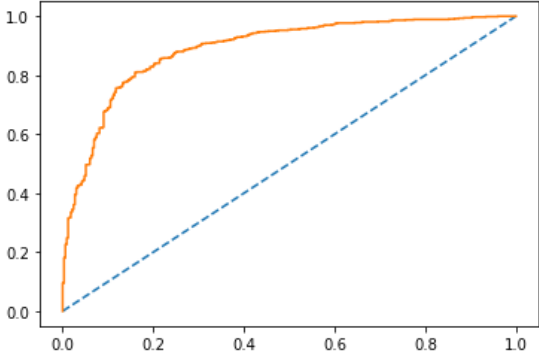
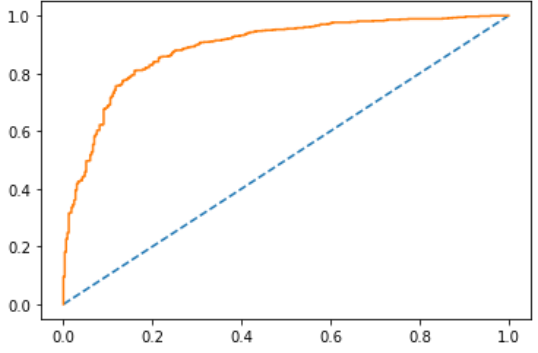
For Bagging - BaggingClassifier & DecisionTreeClassifier packages from the SK Learn Library have been used.

For Boosting - AdaBoostClassifier package from the SK Learn Library have been used.

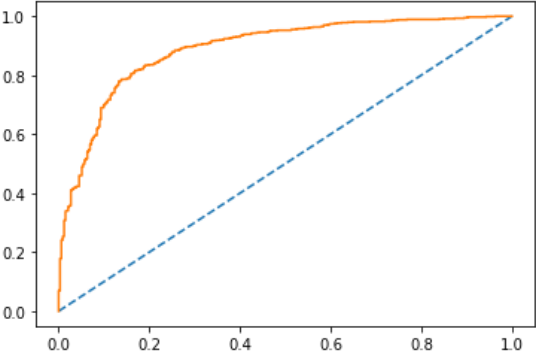
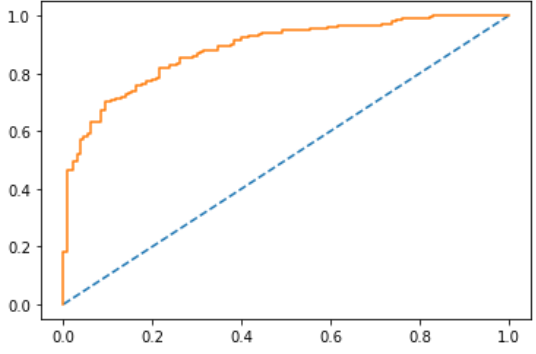
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

Ans 1.7

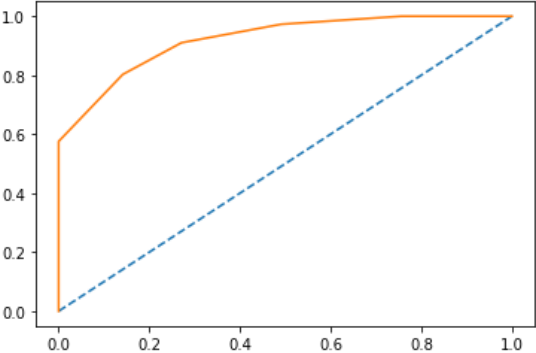
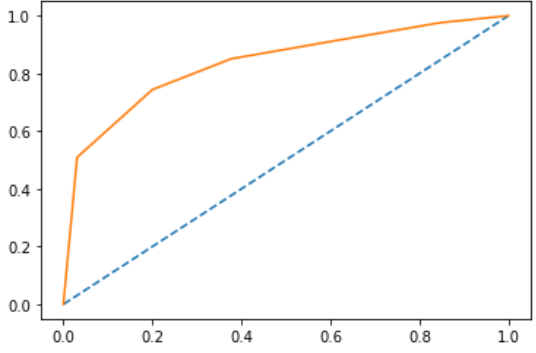
Model Comparison:

	Logistic Regression	
	Train Dataset	Test Dataset
Accuracy	0.84	0.82
Confusion Matrix	$\begin{bmatrix} 230 & 102 \\ 68 & 667 \end{bmatrix}$	$\begin{bmatrix} 85 & 45 \\ 36 & 292 \end{bmatrix}$
ROC Curve:		
ROC_AUC score:	0.889	0.882

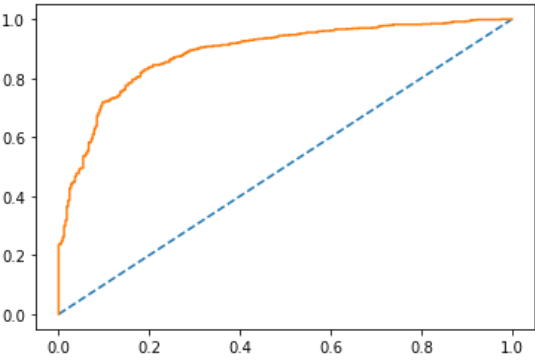
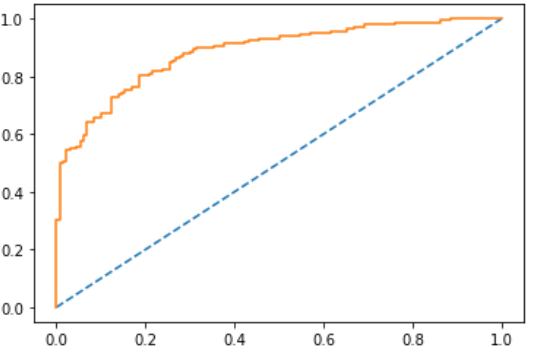


	Linear Discriminant Analysis	
	Train Dataset	Test Dataset
Accuracy	0.84	0.82
Confusion Matrix	$\begin{bmatrix} 233 & 99 \\ 75 & 660 \end{bmatrix}$	$\begin{bmatrix} 86 & 44 \\ 39 & 289 \end{bmatrix}$
ROC Curve:		
ROC_AUC score:	0.889	0.884

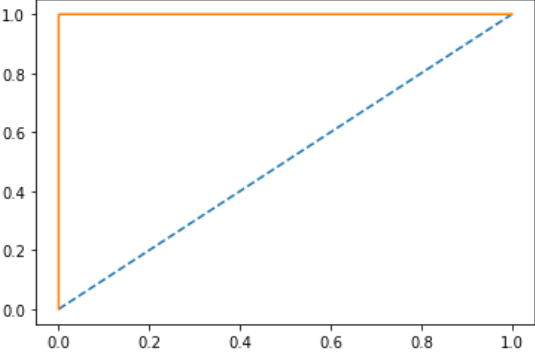
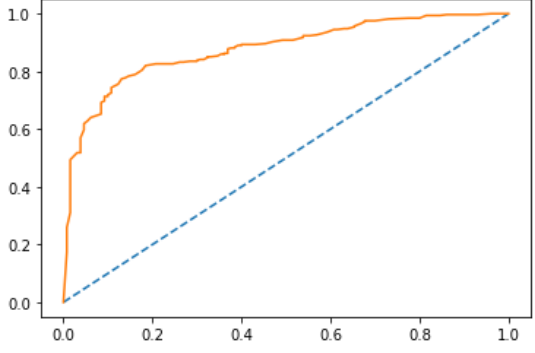


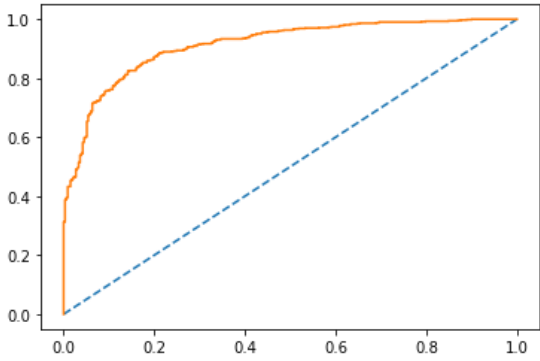
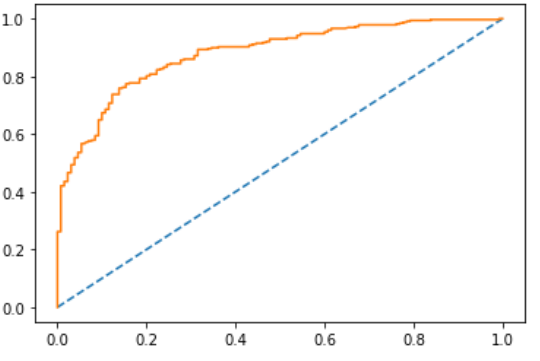
	KNN	
	Train Dataset	Test Dataset
Accuracy	0.85	0.79
Confusion Matrix	$\begin{bmatrix} 242 & 90 \\ 66 & 669 \end{bmatrix}$	$\begin{bmatrix} 81 & 49 \\ 49 & 279 \end{bmatrix}$
ROC Curve:		
ROC_AUC score:	0.921	0.835



	Naïve Bayes Model	
	Train Dataset	Test Dataset
Accuracy	0.83	0.83
Confusion Matrix	$\begin{bmatrix} 240 & 92 \\ 86 & 649 \end{bmatrix}$	$\begin{bmatrix} 94 & 36 \\ 44 & 284 \end{bmatrix}$
ROC Curve:		
ROC_AUC score:	0.886	0.885



	Bagging Model	
	Train Dataset	Test Dataset
Accuracy	0.99906	0.80
Confusion Matrix	$\begin{bmatrix} 331 & 1 \\ 0 & 735 \end{bmatrix}$	$\begin{bmatrix} 83 & 47 \\ 46 & 282 \end{bmatrix}$
ROC Curve:		
ROC_AUC score:	1.000	0.877

	ADABOOST Model	
	Train Dataset	Test Dataset
Accuracy	0.85	0.82
Confusion Matrix	$\begin{bmatrix} 238 & 94 \\ 69 & 666 \end{bmatrix}$	$\begin{bmatrix} 90 & 40 \\ 43 & 285 \end{bmatrix}$
ROC Curve:		
ROC_AUC score:	0.913	0.879

1.7 Inference:

On the basis of the Test Accuracy and AUC score the best model is the ADABOOST Model.

1.8 Based on these predictions, what are the insights?

Ans 1.8

- There are rows with duplicate values and columns which have outliers - hence the quality of the data can be improved by the data team.
- Most of the votes were given to the Labour party.
- A lot of columns are correlated hence we can in the future test performance excluding a few columns.
- ADABOOST Model is the best model in terms of performance.

Problem 2: Problem 2: Speeches of the Presidents:

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:

- President Franklin D. Roosevelt in 1941
- President John F. Kennedy in 1961
- President Richard Nixon in 1973

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

Ans 2.1

Roosevelt:

```
Num sentences: 31
Num chars: 6249
Num words: 1360
```

Kennedy:

```
Num sentences: 27
Num chars: 6255
Num words: 1390
```

Nixon:

```
Num sentences: 21
Num chars: 8223
Num words: 1819
```

2.2 Remove all the stopwords from all three speeches.

Ans 2.2

Stop Words are removed using - Stop Words Package from the NLTK Library and the Punctuation package from the String Library.

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)

Roosevelt:

- know
- spirit
- life

Kennedy:

- us
- world
- let

Nixon:

- us
- america
- peace

