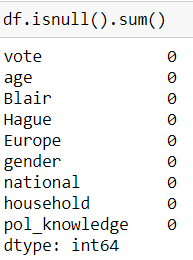
|  |  |  |
| --- | --- | --- |
| Image | |  |
| Machine Learning | | |
| Ritesh Singh |  | |

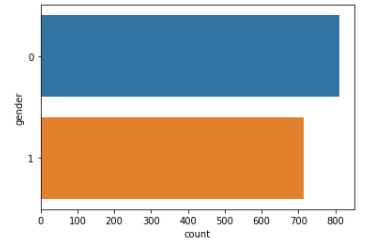
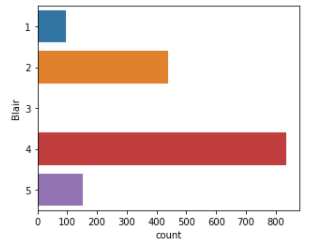
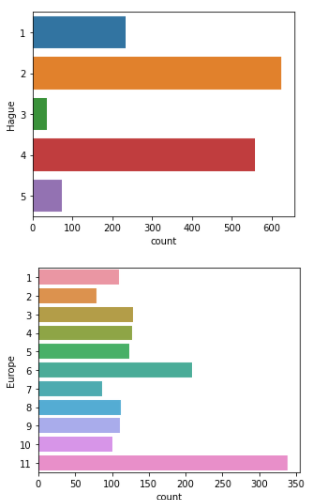
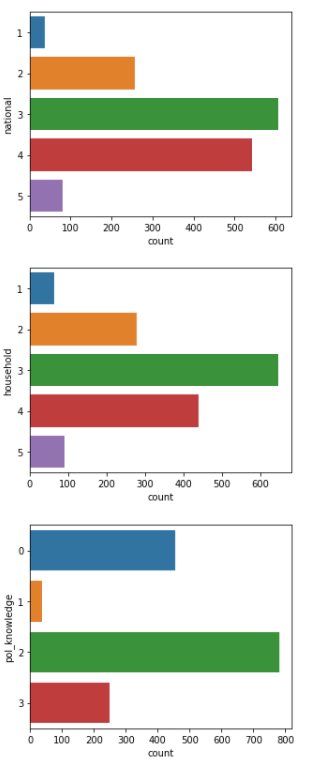
|  |
| --- |
| PROBLEM 1 **You are hired by one of the leading news channels CNBE who wants to analyse recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.**   * 1. **Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.** |

Data consists of 9 columns and 1525 rows. Column names are changed for simplicity. Data is ordinal as observations are in scale (e.g., out of 5).

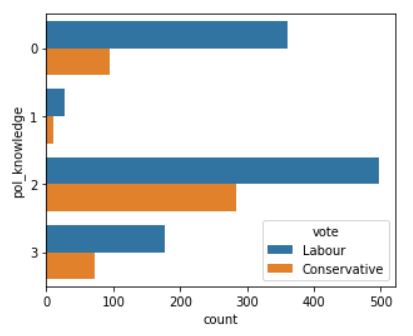
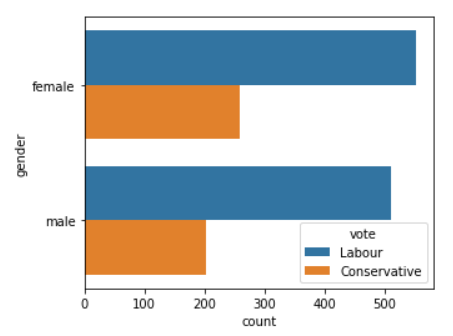
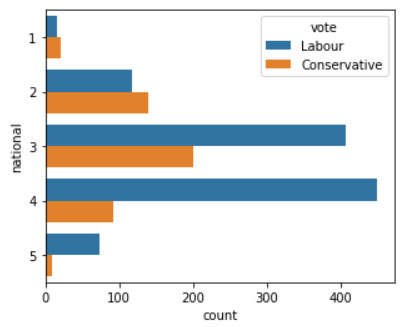
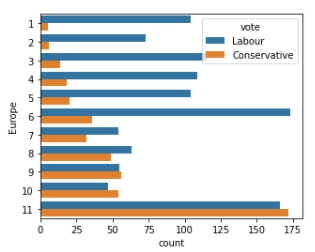
 

There are 1525 rows and no null values in the sample. Every participant in the survey has marked their ratings. People older than 23 years are considered for the survey. Most of the participant’s ages are between 38 years to 68 years. Minimum in Political knowledge is 0, that means there are some participants who believe that political party do not have any political Knowledge. Duplicates are not dropped because here any two participants can give similar ratings.

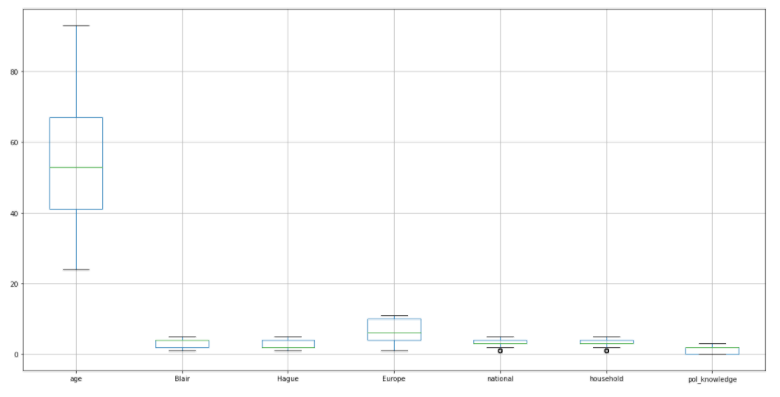
* 1. **Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.**

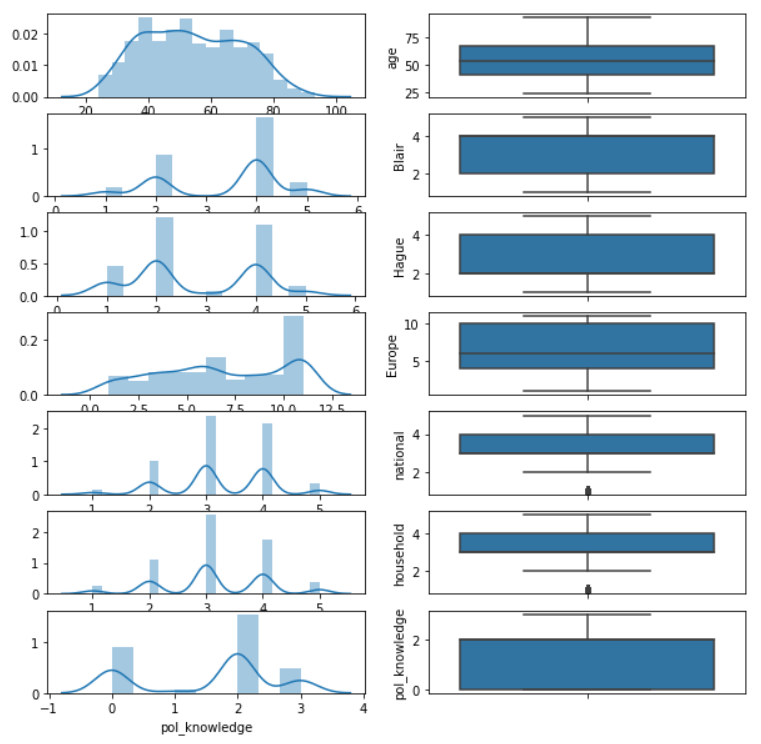
* Female participants are more than Male participants.
* Many participants gave a rating of 4 to Labor and 2 to Conservative leader.
* Many participants believe that parties are Eurosceptic as lot of them gave 11.
* More number of people gave 2 in political Knowledge and close to 450 believe they do not have any knowledge of parties' positions on European integration.

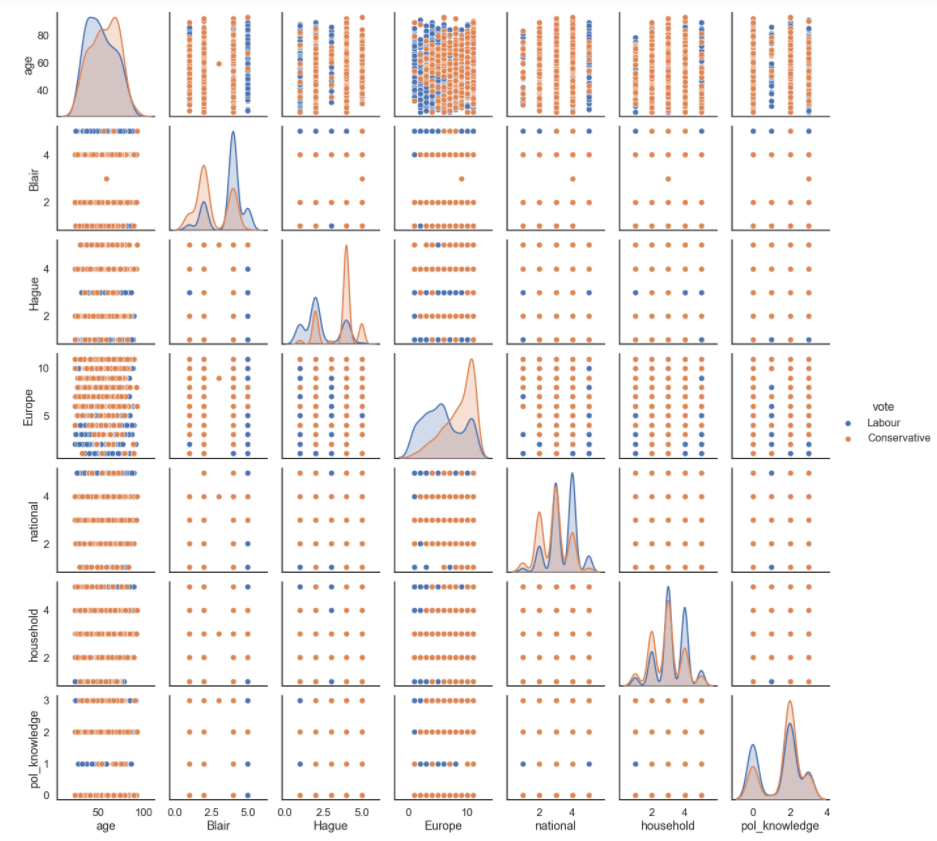
* Most of the female and male participants are inclined towards Labour party.
* More people are voting to Labour party for Current national economic conditions.
* Eurosceptic sentiment is more in Conservative party at scale 11.

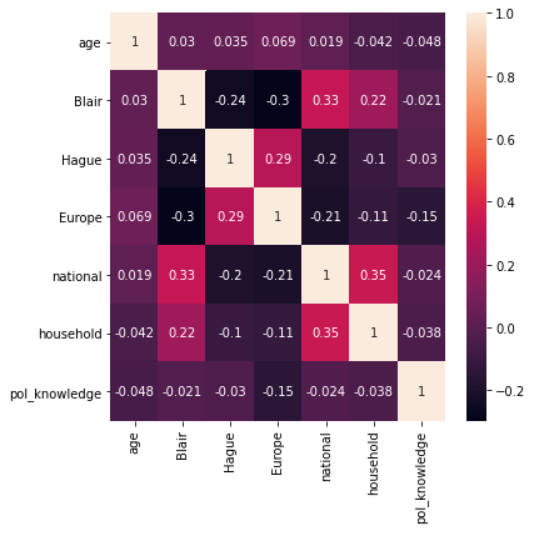


* There are some outliers in National and household.



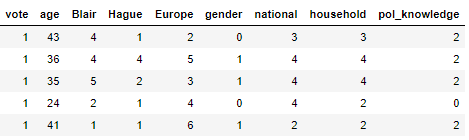
* These outliers in National and household are the lowest ratings (probably 1)





* After looking at the correlation matrix we can say that no two features are related to each other.

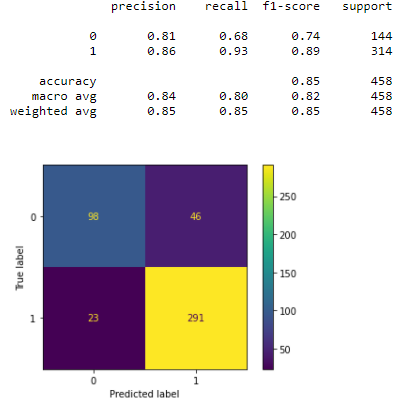
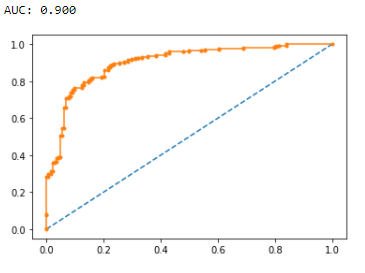
* 1. **Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).**



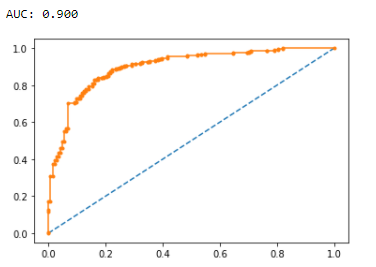
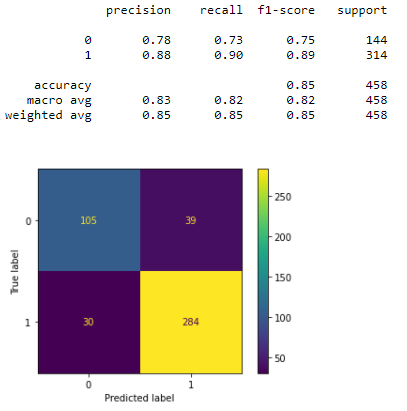
* Two features gender and vote are not numeric and are encoded into numeric values.
* Scaling here is not necessary because most of the features are ordinal and after scaling though values will lie below 1 but these values can again be considered as ordinal. For example, all the rows having 3 in household will have same value. Also the results will not be affected by scaling.
* Splitting is done in 70:30 ration (refer ipynb file)
  1. **Apply Logistic Regression and LDA (linear discriminant analysis).**

**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)**

Logistic Regression

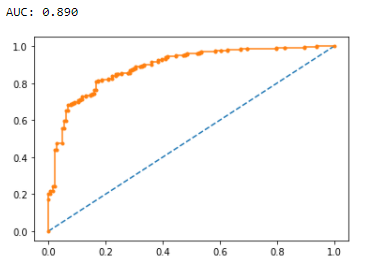
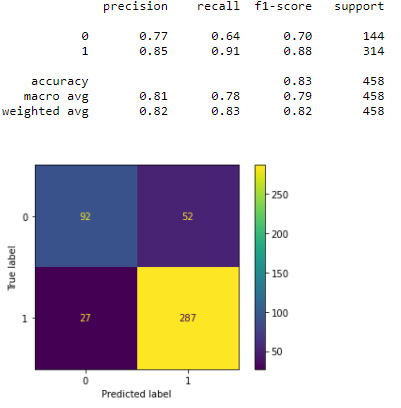
 

LDA

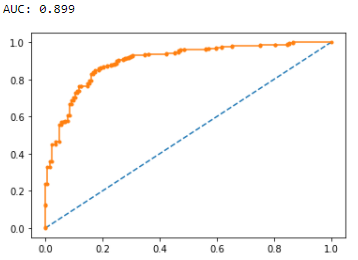
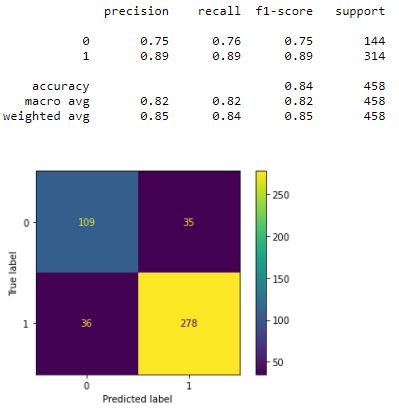


**1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results. (4 marks)**

KNN

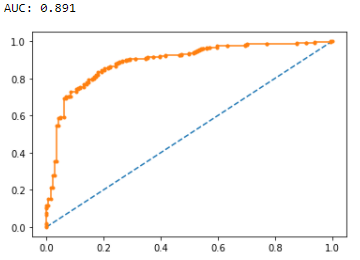
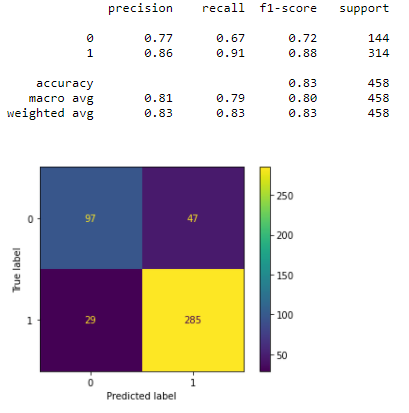


NB

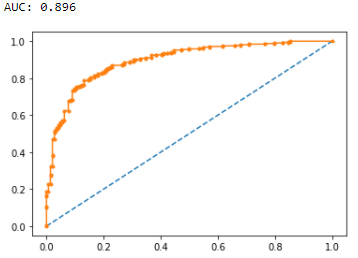
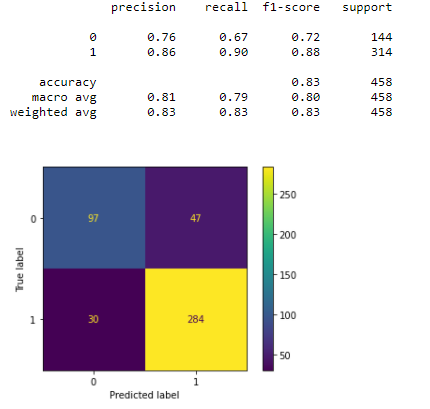


**1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.**

Bagging



Gradient Boosting





* After model tuning and using grid search, we have this comparison table, we can refer to Q 1.6 for graphical representation.
* Every model is overfitting except for KNN

and Random forest, we can see the noticeable % difference between the test and train accuracy of each model.

* 1. **Based on these predictions, what are the insights? (5 marks)**
* If we see the number of observations for conservative and labour, the data is not balanced. We have more observations for labour and close to 500 observations for conservative.
* If sampling technique is correct, then we can say that 70% of the people will vote Labour.
* Female participants are more than Male participants.
* Many participants gave a rating of 4 to Labor and 2 to Conservative leader.
* Many participants believe that parties are Eurosceptic as lot of them gave 11.
* More number of people gave 2 in political Knowledge and close to 450 believe they do not have any knowledge of parties' positions on European integration.
* Most of the female and male participants are inclined towards Labour party.
* More people are voting to Labour party for Current national economic conditions.
* Eurosceptic sentiment is more in Conservative party at scale 11.
* Including demographic details would have helped the analysis.
* Factors like education, healthcare and environment in survey can give better results.

#### 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:**

**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.**

Characters in Roosevelt: 7571

Words in Roosevelt: 1536

Sentences in Roosevelt: 68

Characters in Kennedy: 7618

Words in Kennedy: 1546

Sentences in Kennedy: 52

Characters in Nixon: 9991

Words in Nixon: 2028

Sentences in Nixon: 69

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

After removing stop words: 2126

Most common words: 1019

**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 most common words in 1941-Roosevelt.txt: ['nation', 'know', 'spirit']

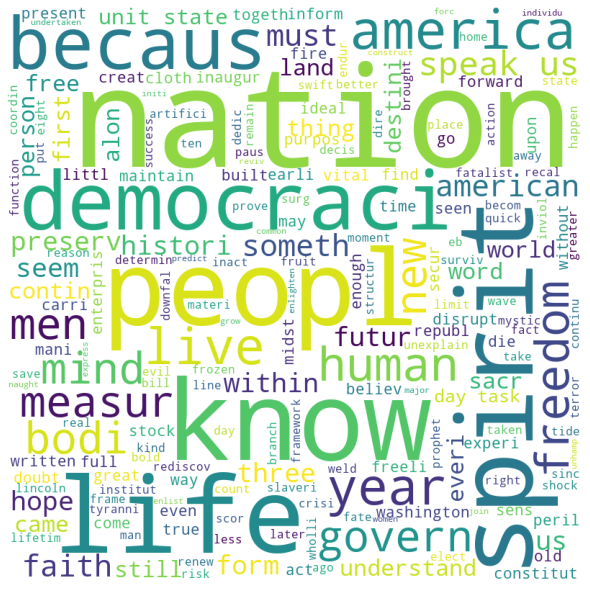
3 most common words in 1961-Kennedy.txt: ['let', 'us', 'world']

3 most common words in 1973-Nixon.txt: ['us', 'let', 'america']

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

3 most common words in 1941-Roosevelt.txt: ['nation', 'know', 'spirit']

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*WORD CLOUD FOR 1941-Roosevelt.txt \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

****

3 most common words in 1961-Kennedy.txt : ['let', 'us', 'power']

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*WORD CLOUD FOR 1961-Kennedy.txt \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

****

3 most common words in 1973-Nixon.txt: ['us', 'let', 'america']

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*WORD CLOUD FOR 1973-Nixon.txt \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

****