Problem 1 Machine Learning

You are hired by one of the leading news channel 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 null value condition check.

Data set :

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
Labour	47	3	4	4	4	4	2	male
Labour	57	2	2	4	4	11	2	male
Labour	77	3	4	4	1	1	0	male
Labour	39	3	3	4	4	11	0	female
Labour	70	3	2	5	1	11	2	male
Labour	39	3	3	1	2	7	0	female
Labour	66	4	3	4	4	9	2	male
Labour	59	4	4	4	1	10	2	female
Labour	66	3	3	2	5	8	0	female
Labour	77	2	3	2	1	11	2	female
Labour	51	4	4	4	4	5	0	male
Labour	43	2	4	1	4	8	0	female
Labour	41	4	4	5	4	7	2	female
Labour	79	3	3	4	2	1	0	male
Labour	37	3	1	1	1	5	2	female
Labour	38	3	3	4	4	7	0	male
Labour	53	2	1	2	4	5	2	male
Labour	59	3	3	4	2	1	2	male
Conservative	44	2	4	4	4	9	2	male
Conservative	60	3	2	4	4	2	2	female
Labour	51	3	3	4	3	6	0	female
Conservative	56	2	2	2	4	9	2	female
Labour	51	3	2	4	2	2	2	female
Labour	44	3	3	4	2	1	2	male
Labour	61	4	3	5	1	1	2	male

We are provided with the above data set of 1525 rows and 9 columns. Of the above columns, seven columns are integer data type and two columns are object data type.

```
<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
```

There are **no** Null values in the given dataset.

The given dataset has **8 duplicate rows** and have to be dropped before proceeding for further analysis.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
67	Labour	35	4	4	5	2	3	2	male
626	Labour	39	3	4	4	2	5	2	male
870	Labour	38	2	4	2	2	4	3	male
983	Conservative	74	4	3	2	4	8	2	female
1154	Conservative	53	3	4	2	2	6	0	female
1236	Labour	36	3	3	2	2	6	2	female
1244	Labour	29	4	4	4	2	2	2	female
1438	Labour	40	4	3	4	2	2	2	male

Descriptive statistics for the dataset:

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
count	1517	1517.000000	1517.0	1517.0	1517.0	1517.0	1517.0	1517.0	1517
unique	2	NaN	5.0	5.0	5.0	5.0	11.0	4.0	2
top	Labour	NaN	3.0	3.0	4.0	2.0	11.0	2.0	female
freq	1057	NaN	604.0	645.0	833.0	617.0	338.0	776.0	808
mean	NaN	54.241266	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	15.701741	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	24.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	41.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	53.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	67.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	93.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN

We have Column **'age'** as **integer type** data and remaining all columns are **object type** data.

As per the details resulted from the descriptive statistics of the dataset, we can find that:

All the variables are having 1517 values after dropping the duplicate rows.

Of the entire dataset, column **'age'** has **highest max** value of 93 and **least min** value of **24.**

```
VOTE :  2
Conservative     460
Labour          1057
Name: vote, dtype: int64


ECONOMIC.COND.NATIONAL :  5
1     37
5     82
2    256
4    538
3    604
Name: economic.cond.national, dtype: int64


ECONOMIC.COND.HOUSEHOLD :  5
1     65
5     92
2    280
4    435
3    645
Name: economic.cond.household, dtype: int64
```

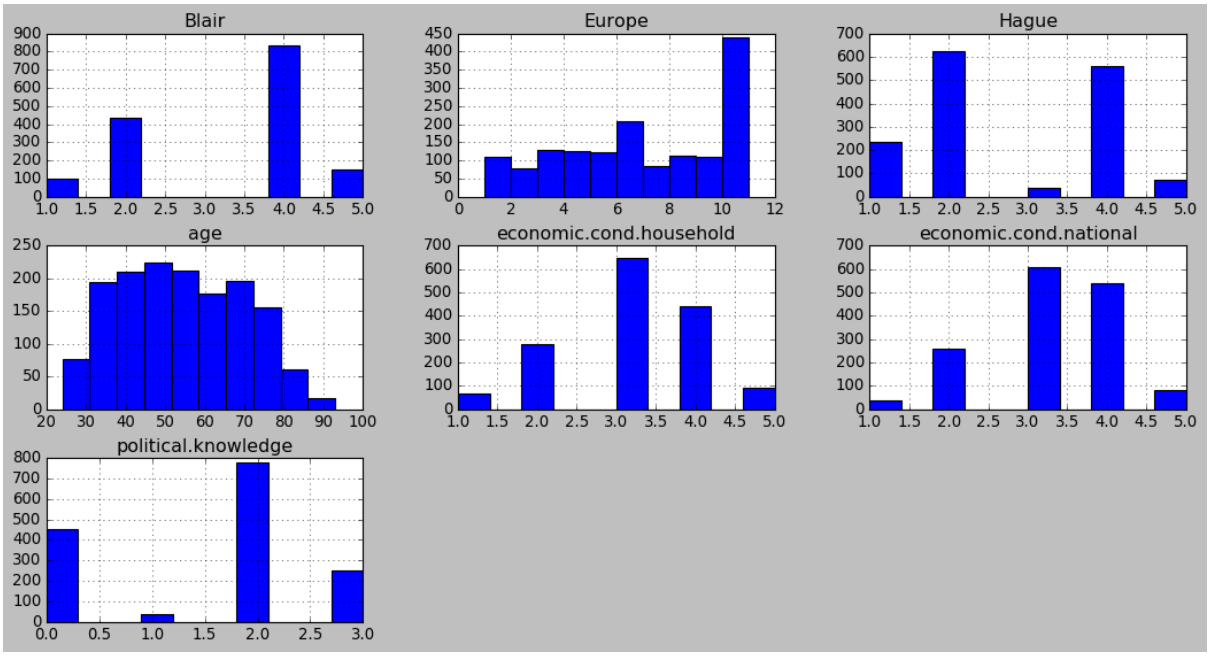
```
BLAIR :  5
3      1
1     97
5    152
2    434
4    833
Name: Blair, dtype: int64


EUROPE :  11
2      77
7      86
10    101
1     109
9     111
8     111
5     123
4     126
3     128
6     207
11    338
Name: Europe, dtype: int64
```

```
HAGUE :  5
3     37
5     73
1    233
4    557
2    617
Name: Hague, dtype: int64


POLITICAL.KNOWLEDGE :  4
1     38
3    249
0    454
2    776
Name: political.knowledge, dtype: int64


GENDER :  2
male      709
female    808
Name: gender, dtype: int64
```



From the above plot , we can see the histograms of the variables of the given dataset.

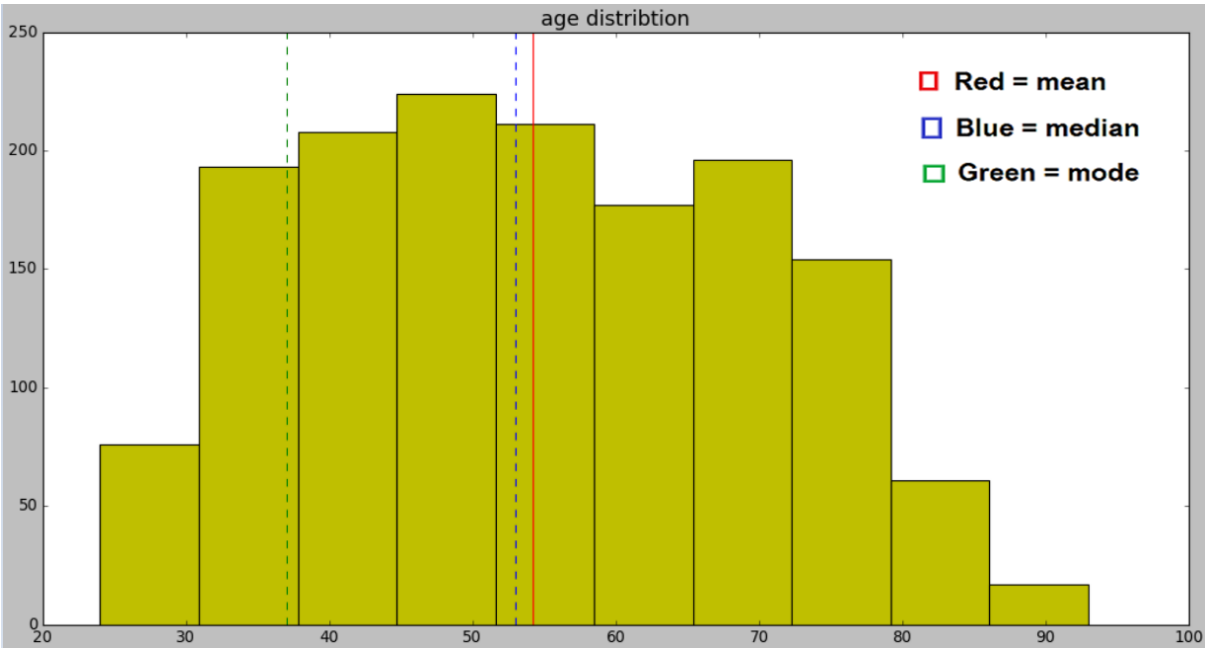
The variable **'age'** has skewness of value 0.139800 and it is almost symmetrical.

From the given dataset, we can understand that **'Labour party'** is the highest choosed party with **69.67%** value.

Minimum age of voter is **24** and **maximum** age of voter is **93** and Female voters are 6.53% more than male vot ers.

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

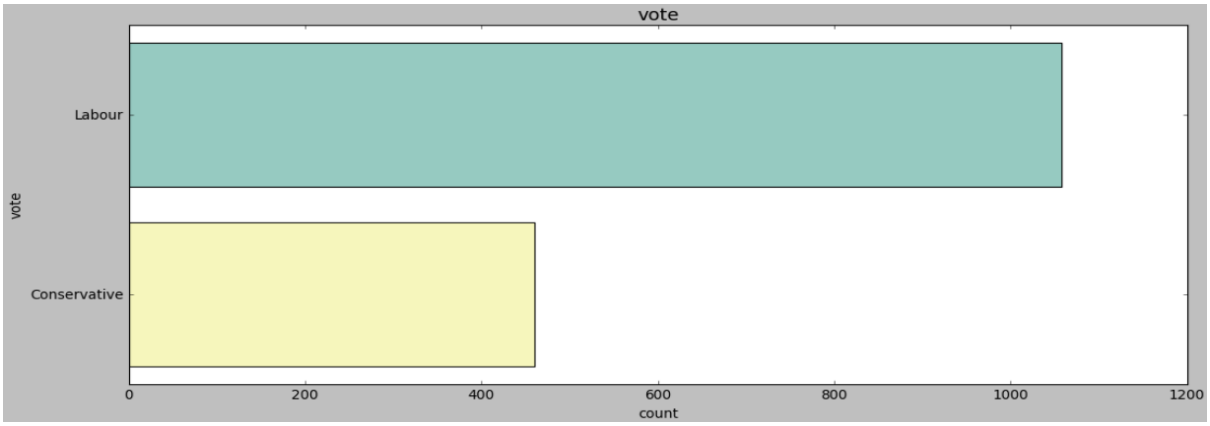
Among the variables of the given dataset, variable 'age' is integer data type and remaining all variables are obje ct data type.



This variable has **mean** of **54.241266** and **median** of **53.00**

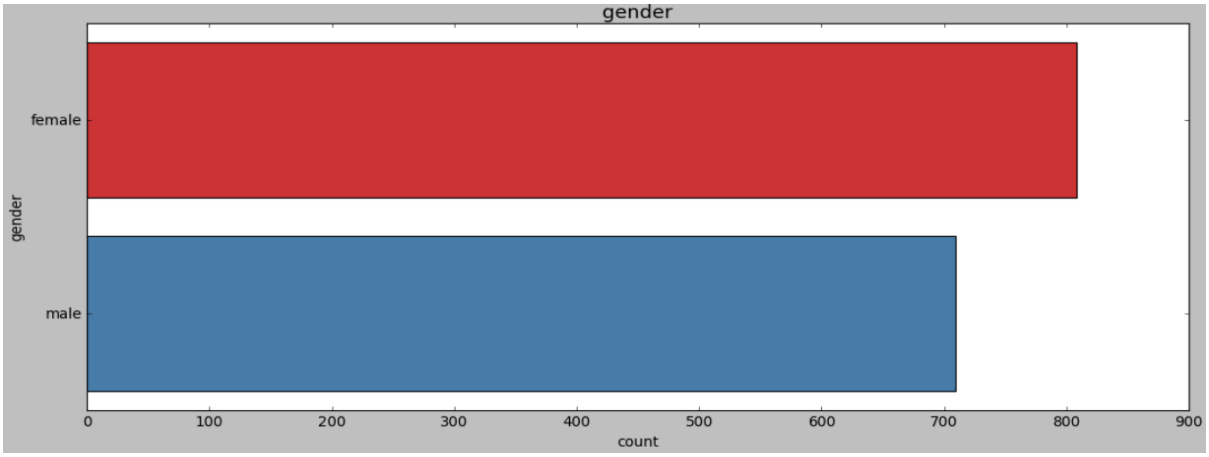
Variable:

Vote-



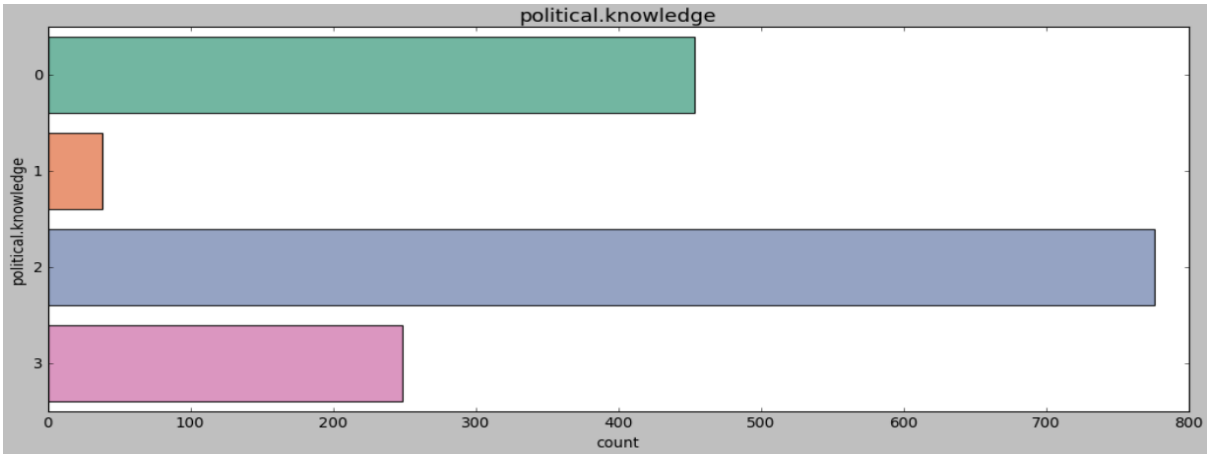
This variable has high count of **'Labour'**

Gender :



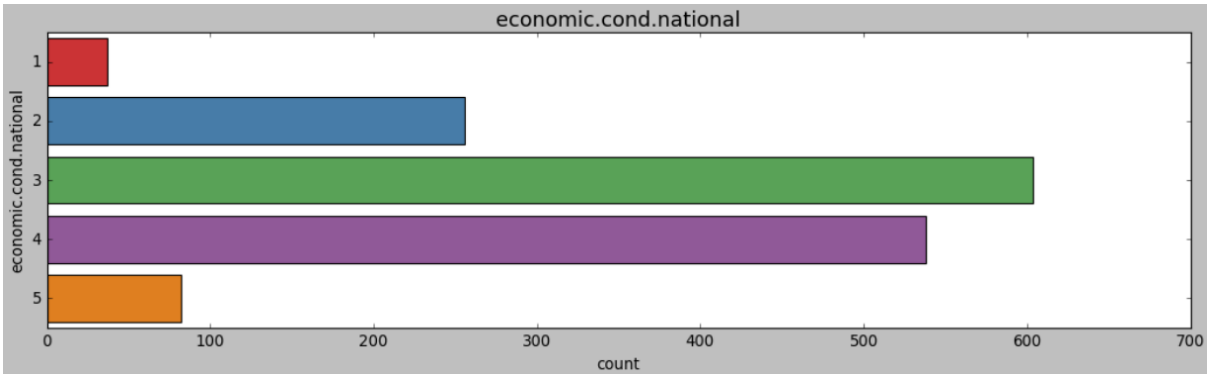
This variable has high count of **'female'**

political.knowledge:



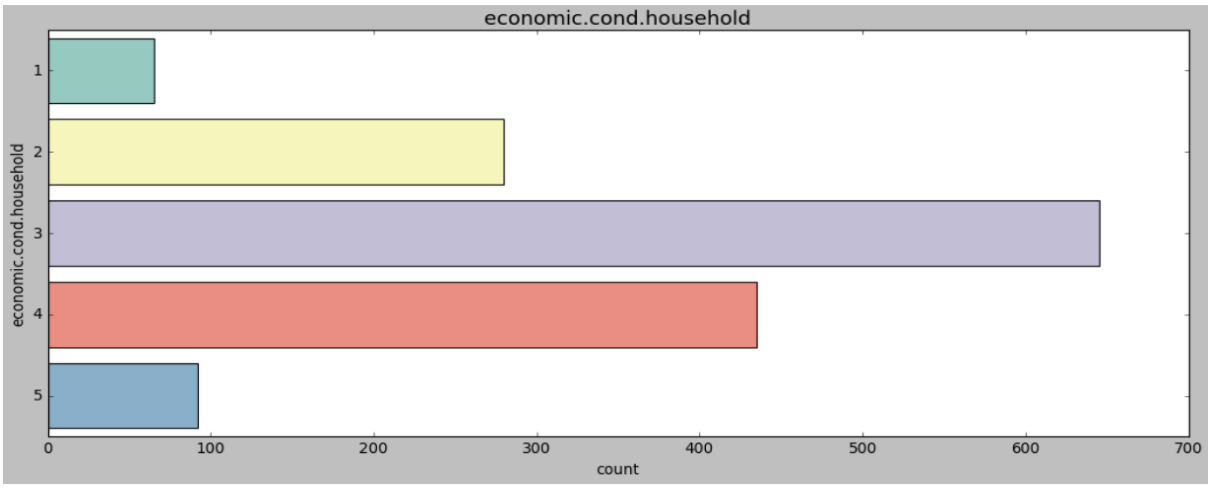
Majority of voters have political knowledge of value **'2'**

economic.cond.national:



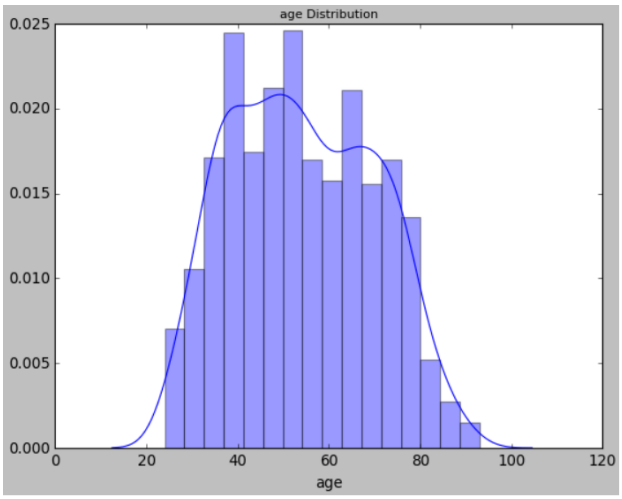
Majority of voters has current national economic condition of value **'3'**

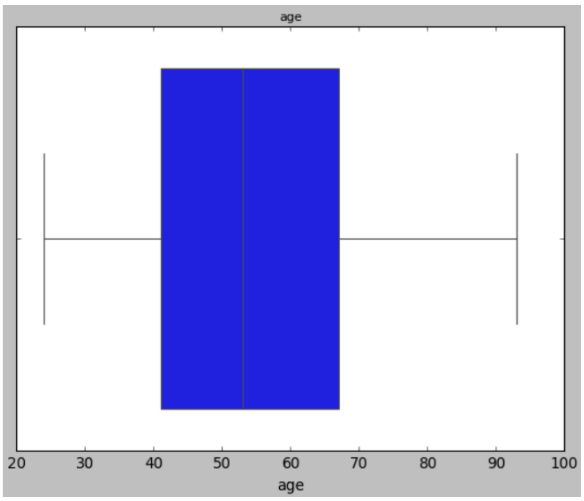
economic.cond.household:



Majority of voters has current household economic condition of value **'3'**

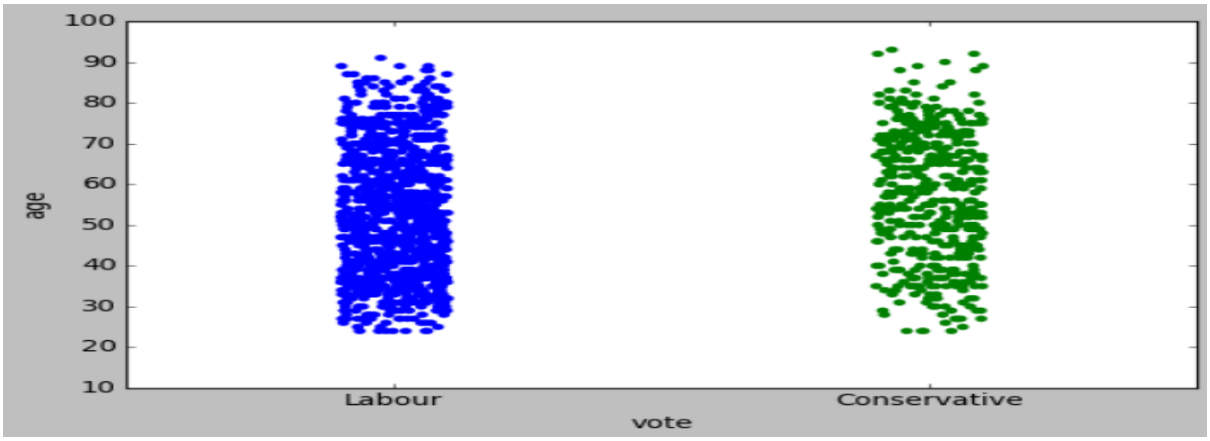
Age :



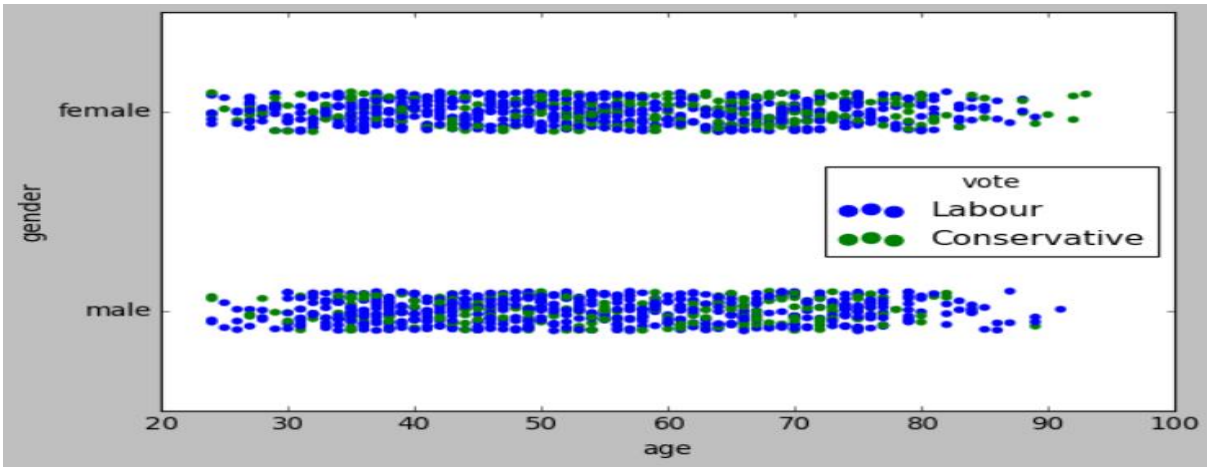


From the above plot, we can infer that the variable is almost symmetric and has **no outliers.**

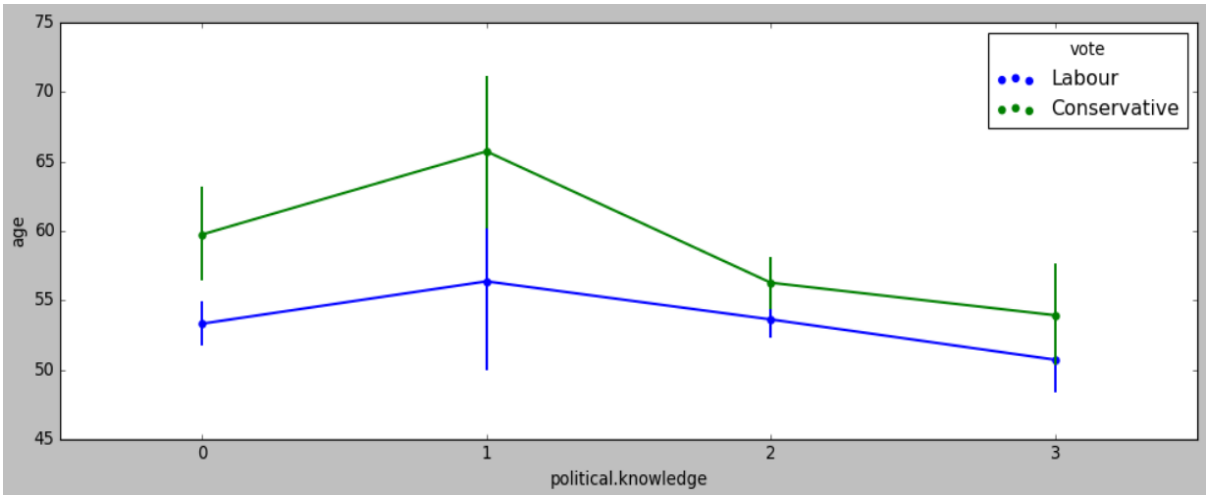
Bivariate and Multivariate Analysis :



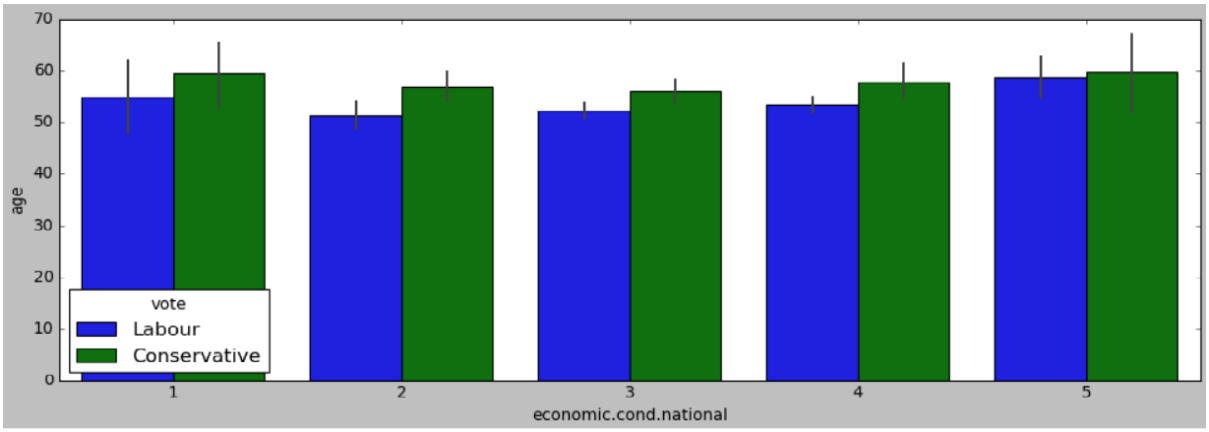
From above plot , we can infer that younger and older voters are less preferred towards conservative party.

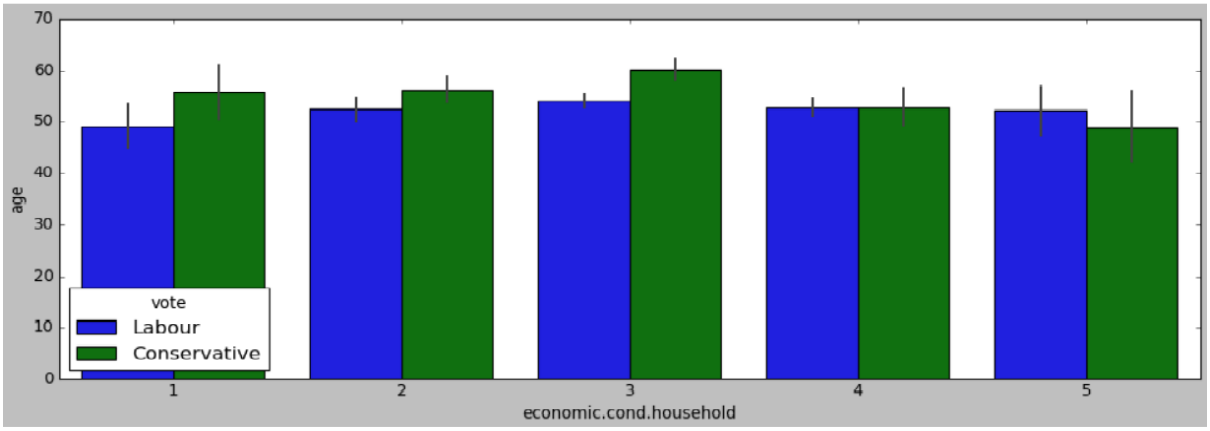


From above plot , we can infer that for an age between **30 to 85**, both **gender** have almost **similar preference** f or the two party's. But **younger and older male voters** preferred more towards **labour party** and **older female** voters preferred towards **conservative party**.

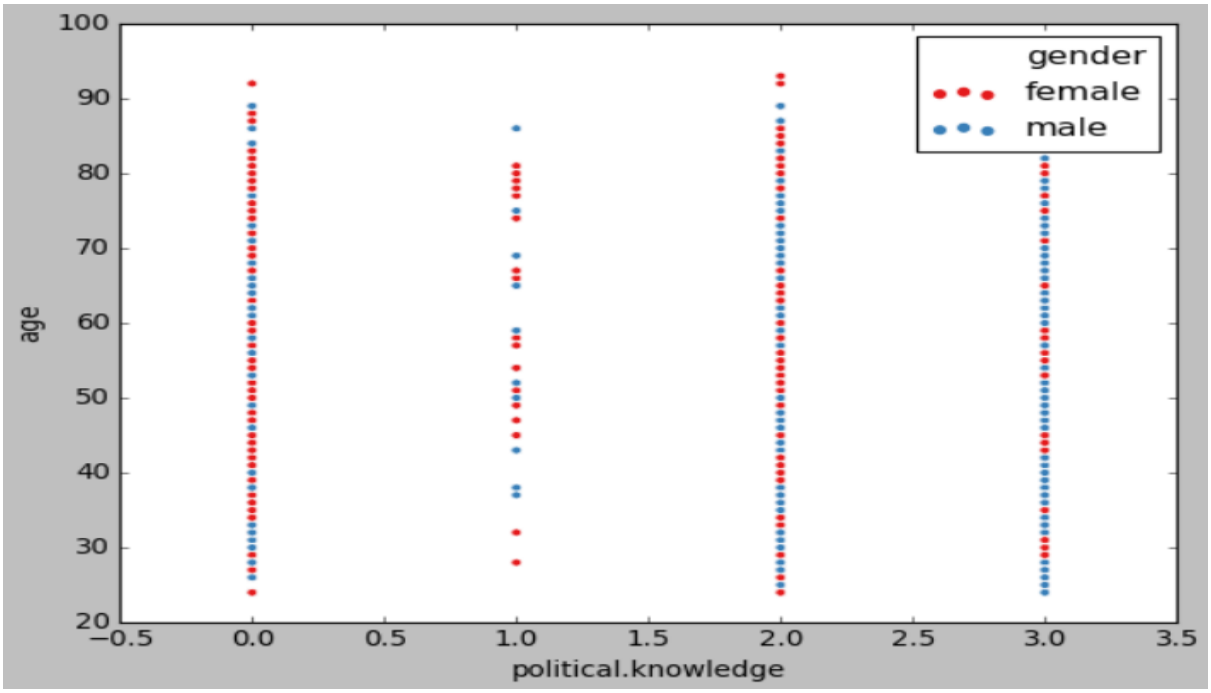


Overall, for a given political knowledge , **Older voters** preferred towards **Conservative party**.



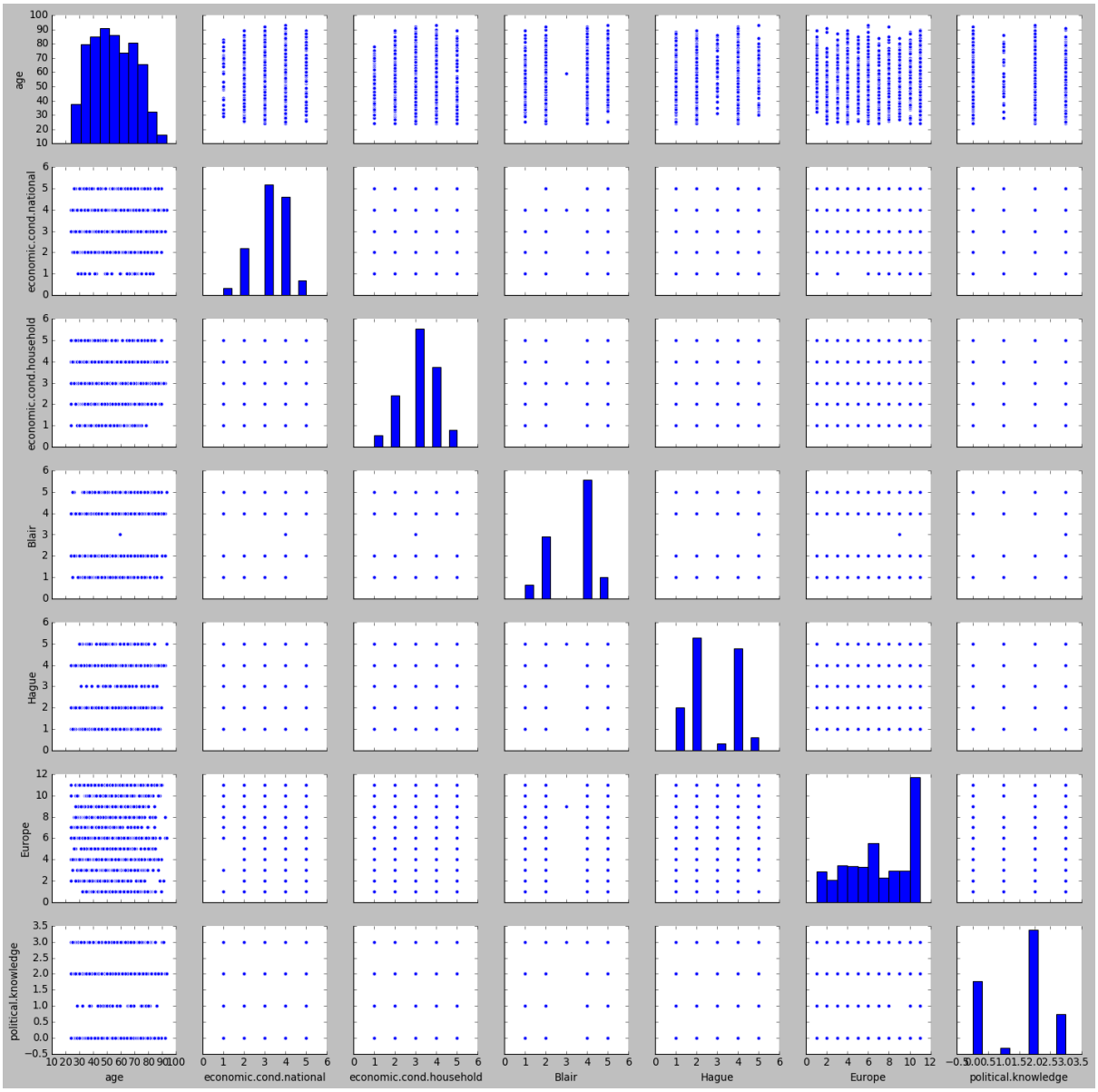


At all levels of current **national economic conditions** , voters are more preferred towards **conservative party**. In case of **current household economic conditions** , voters preferred **conservative party** at **low and mid lev el** and **labour party at high level**.



From above plot we can infer that, most of the female voters are on low side of political knowledge compared to male voters.

PairPlot:



From above plot , we can see relationship between different variables of the given dataset.

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

Among the variables of the given dataset, variable 'age' is integer data type and remaining all variables are obje ct data type. We need to encode the data for modelling.

	vote	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	Labour	3	3	4	1	2	2	female
1	Labour	4	4	4	4	5	2	male
2	Labour	4	4	5	2	3	2	male
3	Labour	4	2	2	1	4	0	female
4	Labour	2	2	1	1	6	2	male

After encoding the object type data , we have got the following dataset,

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

As there is much difference in the magnitude of value in the dataset between values among different variables. We prefer to use z score method to minimize the effect of both skewness and variability due to difference between high and low magnitude value of the data. Also scaling of variables does not affect the accuracy of the model.

Scaled Variables :

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	-0.716161	-0.278185	-0.148020	0.565802	-1.419969	-1.437338	0.423832	-0.936736
1	-1.162118	0.856242	0.926367	0.565802	1.014951	-0.527684	0.423832	1.067536
2	-1.225827	0.856242	0.926367	1.417312	-0.608329	-1.134120	0.423832	1.067536
3	-1.926617	0.856242	-1.222408	-1.137217	-1.419969	-0.830902	-1.421084	-0.936736
4	-0.843577	-1.412613	-1.222408	-1.988727	-1.419969	-0.224465	0.423832	1.067536

Splitting data into train and test (70:30)

```
x_train.head()
```

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
991	-1.289535	-1.412613	0.926367	-1.988727	1.014951	1.291625	0.423832	-0.936736
1274	-0.907286	0.856242	-0.148020	0.565802	1.014951	-0.224465	-1.421084	1.067536
649	0.430587	0.856242	-0.148020	0.565802	1.014951	0.078753	0.423832	-0.936736
677	-0.461328	-0.278185	-0.148020	0.565802	-0.608329	1.291625	-1.421084	1.067536
538	-0.652453	1.990670	-0.148020	0.565802	-0.608329	0.381971	-1.421084	1.067536

```
x_test.head()
```

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
504	1.067669	-0.278185	-0.148020	-1.137217	-0.608329	0.381971	0.423832	-0.936736
369	-0.716161	-0.278185	-1.222408	0.565802	-0.608329	0.381971	1.346290	1.067536
1075	2.214417	1.990670	2.000755	1.417312	-0.608329	-1.740556	0.423832	1.067536
1031	-0.461328	-1.412613	-0.148020	-1.137217	1.014951	0.381971	0.423832	-0.936736
1329	-1.353243	1.990670	0.926367	0.565802	1.014951	0.381971	-1.421084	1.067536

```
y_train.head()
```

```
991     0
1274    1
649     0
677     1
538     1
Name: vote, dtype: int8
```

```
y_test.head()
```

```
504     1
369     1
1075    1
1031    0
1329    1
Name: vote, dtype: int8
```

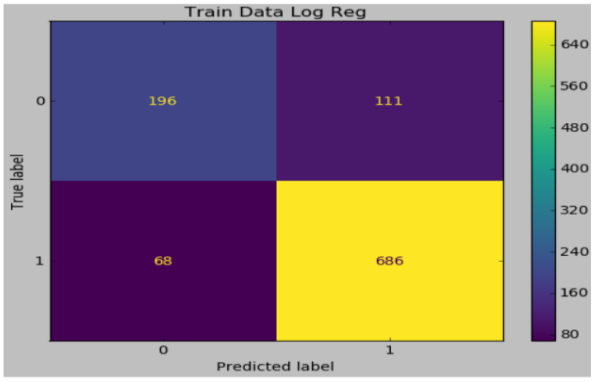
1.4. Apply Logistic Regression and LDA (Linear Discriminant Analysis).

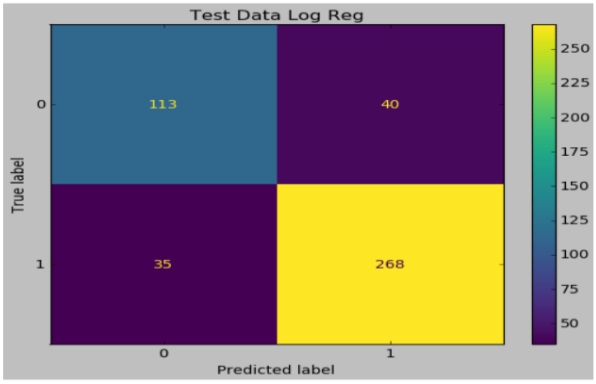
Logistic Regression:

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg',
                   verbose=True)
```

	Train Data	Test Data
Model score	0.831291235	0.835526316
AUC	0.89	0.89

Confusion Matrix:





Classification Report:

```
Classification Report of the training data:
              precision    recall  f1-score   support

           0       0.74      0.64      0.69       307
           1       0.86      0.91      0.88       754

    accuracy                           0.83      1061
   macro avg       0.80      0.77      0.79      1061
weighted avg       0.83      0.83      0.83      1061

Classification Report of the test data:
              precision    recall  f1-score   support

           0       0.76      0.74      0.75       153
           1       0.87      0.88      0.88       303

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

Feature Importance :

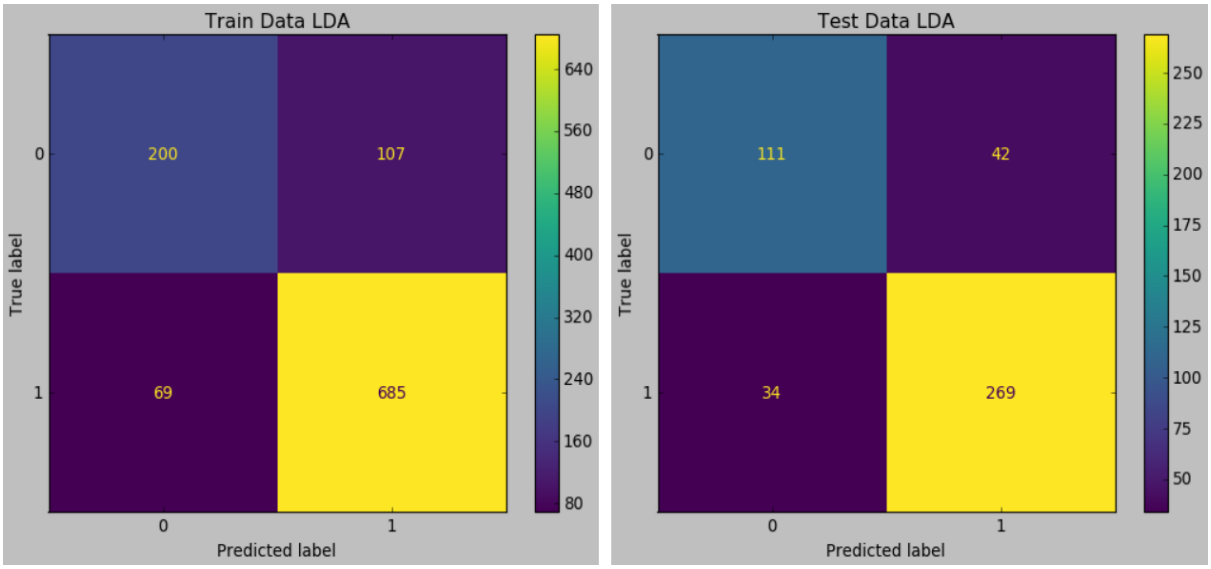
```
The coefficient for age is -0.23541759961490158
The coefficient for economic.cond.national is 0.5620327056282469
The coefficient for economic.cond.household is 0.05699068559850641
The coefficient for Blair is 0.710025879233481
The coefficient for Hague is -1.0219410257175439
The coefficient for Europe is -0.6984592770125304
The coefficient for political.knowledge is -0.3525771010165865
The coefficient for gender is 0.09934857345390612
```

From above results , The most **important feature** is **'Hague'**

LDA (Linear Discriminant Analysis):

	Train Data	Test Data
Model score	0.834118756	0.833333333
AUC	0.889	0.888

Confusion Matrix:



Classification Report :

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.65	0.69	307
1	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

Classification Report of the test data:

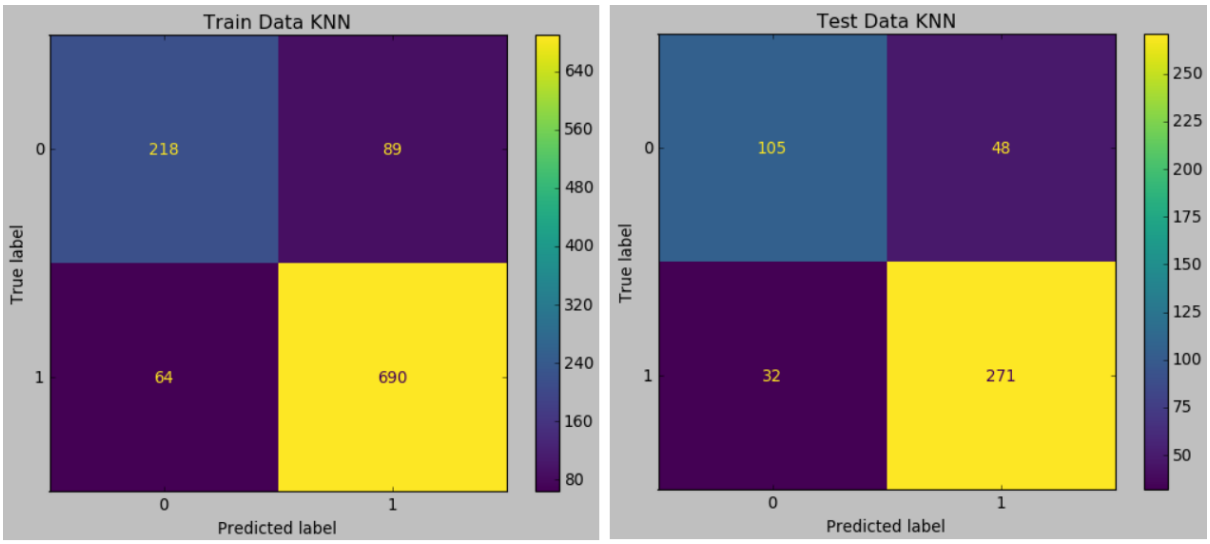
	precision	recall	f1-score	support
0	0.77	0.73	0.74	153
1	0.86	0.89	0.88	303
accuracy			0.83	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456

1.5.Apply KNN Model and Naïve Bayes Model.

KNN Model:

	Train Data	Test Data
Model score	0.855796418	0.824561404
AUC	0.927	0.87

Confusion Matrix:



Classification Report :

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.77	0.71	0.74	307
1	0.89	0.92	0.90	754
accuracy			0.86	1061
macro avg	0.83	0.81	0.82	1061
weighted avg	0.85	0.86	0.85	1061

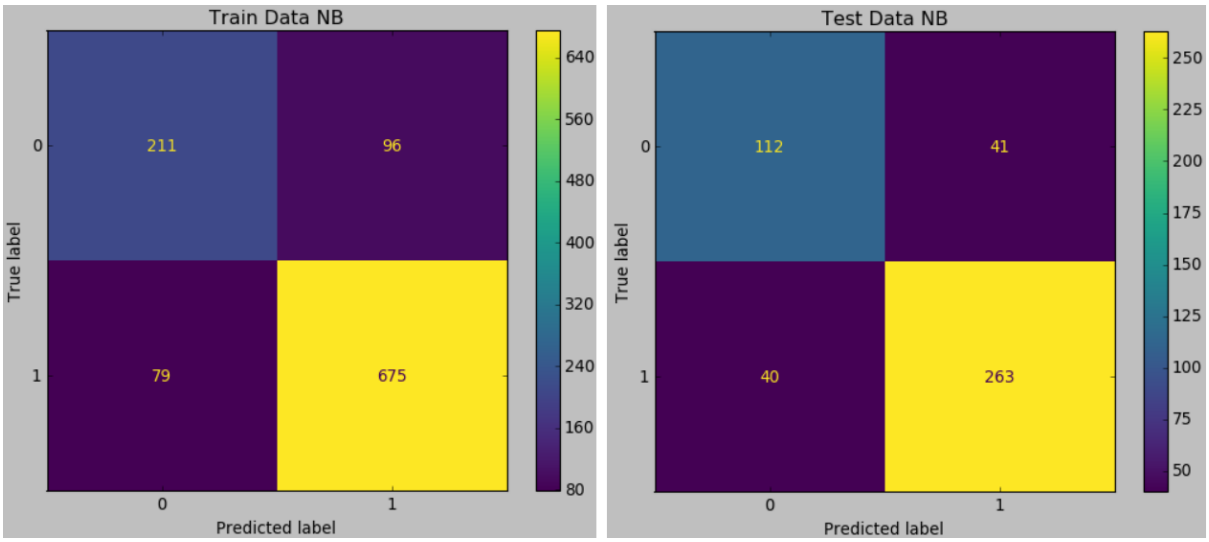
Classification Report of the test data:

	precision	recall	f1-score	support
0	0.77	0.69	0.72	153
1	0.85	0.89	0.87	303
accuracy			0.82	456
macro avg	0.81	0.79	0.80	456
weighted avg	0.82	0.82	0.82	456

Naïve Bayes Model :

	Train Data	Test Data
Model score	0.835061263	0.822368421
AUC	0.888	0.876

Confusion Matrix:



Classification Report :

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.73	0.69	0.71	307
1	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

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.74	0.73	0.73	153
1	0.87	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

1.6. Model Tuning, Bagging and Boosting.

Model Tuning helps in maximizing a model's performance without over fitting or creating too high of a variance . This can be accomplished by selecting appropriate hyper parameters. GridSearch CV is one of the hyper parame ter tuning method to select best parameters.

```
GridSearchCV(cv=5, estimator=RandomForestClassifier(),
             param_grid={'max_depth': [8, 10], 'max_features': [4, 6],
                         'min_samples_leaf': [30, 90],
                         'min_samples_split': [90, 270],
                         'n_estimators': [150, 200]})
```

We have applied GridSearch CV for random forest model and achieved the following best parameters.

```
RandomForestClassifier(max_depth=8, max_features=4, min_samples_leaf=30,
                       min_samples_split=90, n_estimators=150)
```

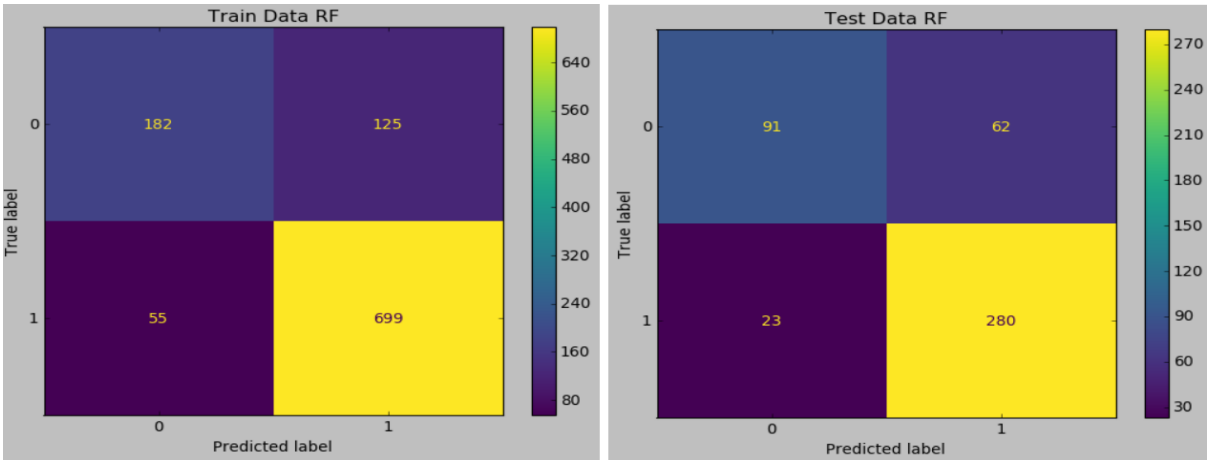
Bagging:

We are using Random Forest model for Bagging:

Random Forest Model:

	Train Data	Test Data
Model score	0.830348728	0.813596491
AUC	0.899	0.887

Confusion Matrix:



Classification Report :

```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.77      0.59      0.67       307
           1       0.85      0.93      0.89       754

    accuracy                           0.83      1061
   macro avg       0.81      0.76      0.78      1061
weighted avg       0.83      0.83      0.82      1061
```

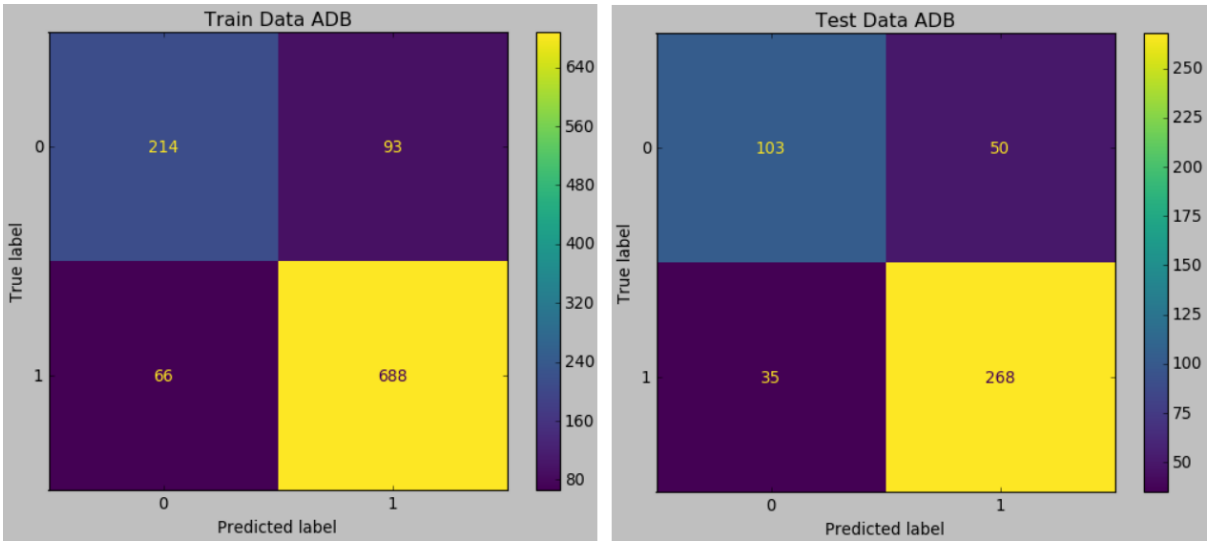
Classification Report of the test data:

	precision	recall	f1-score	support
0	0.80	0.59	0.68	153
1	0.82	0.92	0.87	303
accuracy			0.81	456
macro avg	0.81	0.76	0.77	456
weighted avg	0.81	0.81	0.81	456

Ada Boost Model:

	Train Data	Test Data
Model score	0.850141376	0.813596491
AUC	0.915	0.877

Confusion Matrix:



Classification Report :

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.76	0.70	0.73	307
1	0.88	0.91	0.90	754
accuracy			0.85	1061
macro avg	0.82	0.80	0.81	1061
weighted avg	0.85	0.85	0.85	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.75	0.67	0.71	153
1	0.84	0.88	0.86	303
accuracy			0.81	456
macro avg	0.79	0.78	0.79	456
weighted avg	0.81	0.81	0.81	456

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 mo dels and write inference which model is best/optimized.

Comparing the performance metrics from all the models, we can summarize as below,

	Log Train	Log Test	LDA Train	LDA Test	KNN Train	KNN Test	NB Train	NB Test	RF Train	RF Test	ADB Train	ADB Test
Accuracy	0.831	0.836	0.834	0.833	0.856	0.825	0.835	0.822	0.830	0.814	0.850	0.814
AUC	0.89	0.89	0.89	0.89	0.93	0.87	0.89	0.88	0.90	0.89	0.92	0.88
Recall	0.91	0.88	0.91	0.89	0.92	0.89	0.90	0.87	0.93	0.92	0.91	0.88
Precision	0.86	0.87	0.86	0.86	0.89	0.85	0.88	0.87	0.85	0.82	0.88	0.84
F1 Score	0.88	0.88	0.89	0.88	0.90	0.87	0.89	0.87	0.89	0.87	0.90	0.86

Looking at the details got from **test data** from all the models ,

Accuracy : Logistic Regression model has highest value of 0.836

AUC : Logistic Reg , LDA and Random Forest have highest value of 0.89 and KNN model has least value of 0.87

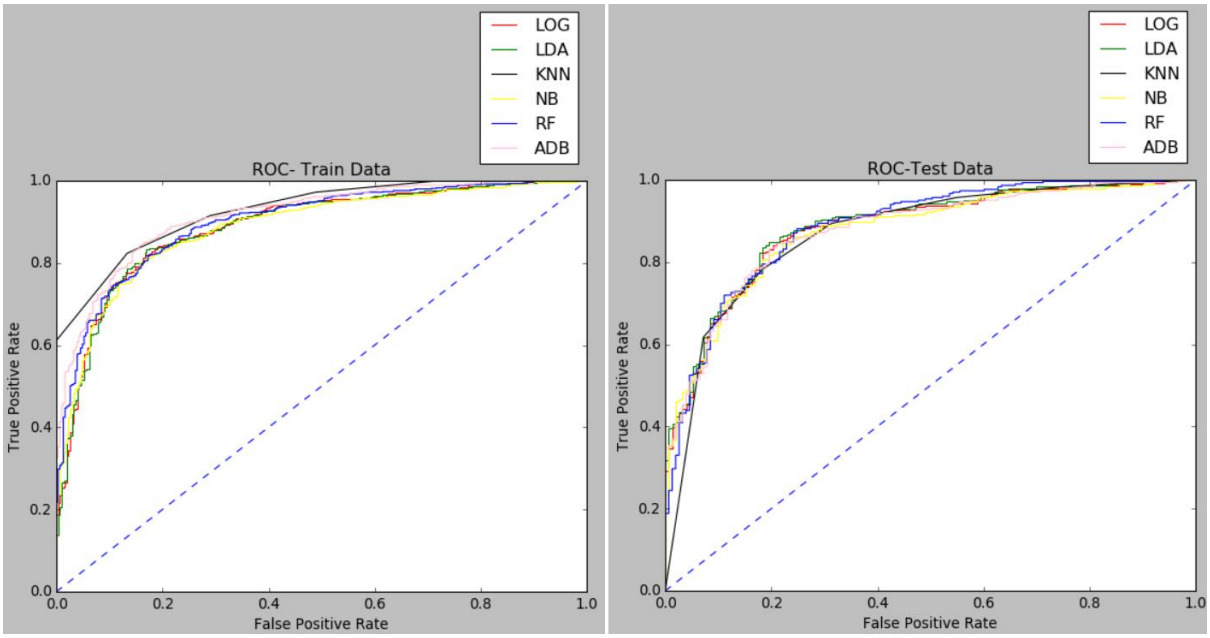
Recall : Random Forest model has highest value of 0.92 and NB model has least value 0.87

Precision : Logistic Reg and NB Model have highest value of 0.87 and RF model has least value of 0.82

F1 Score : Logistic Reg and LDA have highest value 0.88 and ADB model has least value of 0.86

Training and Test set results are almost similar in most of the models and overall measures are high in Logistic Regression model.

Therefore, **Logistic Regression model has slightly better performance than the remaining models.**



Area under curve is highest for the Logistic Reg , LDA and Random Forest models with a value of 0.89

We know that logistic regression is one of the most widely used statistical method for analyzing categorical outcome variable. Logistic regression is the more flexible and more robust method in case of violations of the assumptions also logistic regression is preferred when the dependent variable is dichotomous, while discriminant analysis is preferred when it is nominal (more than two groups). It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative). It can interpret model coefficients as indicators of feature importance.

Therefore we can use logistic regression model in predicting which party a voter will vote for and thereby we can help in creating an exit poll.

1.8. Based on these predictions, what are the insights?

Due to the importance of understanding and managing the challenges in different business domains, it is required to find an effective aid in making decisions. The results from models show that the above algorithms are a promising opportunity in predicting which party a voter will vote for through the cause and effect relationship between the independent and dependent variables of the given dataset.

The above model will be helpful in predicting the dependent variables through the independent variables by assigning the probability of employee opting for the package to the every predictor variable to give the best predictive/dependent variable.

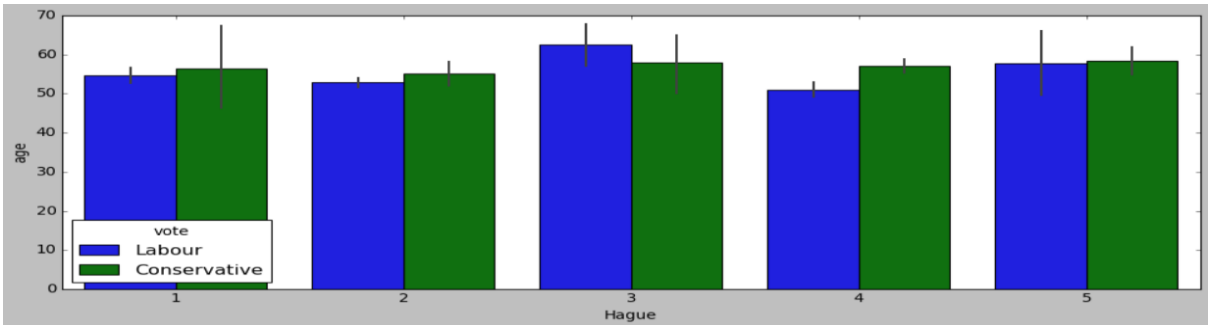
The proportion of the True positive(TP) to Predicted positive(TP+FP) is good for the models. So they will be useful in predicting the target variable.

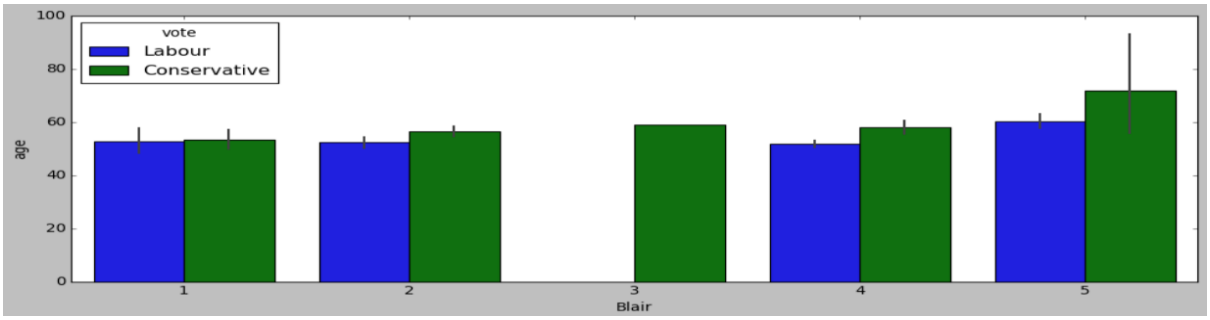
As per predictions of the model, we have got the following coefficients for the independent variables of the given dataset.

The coefficient of the different attributes of the given dataset are:

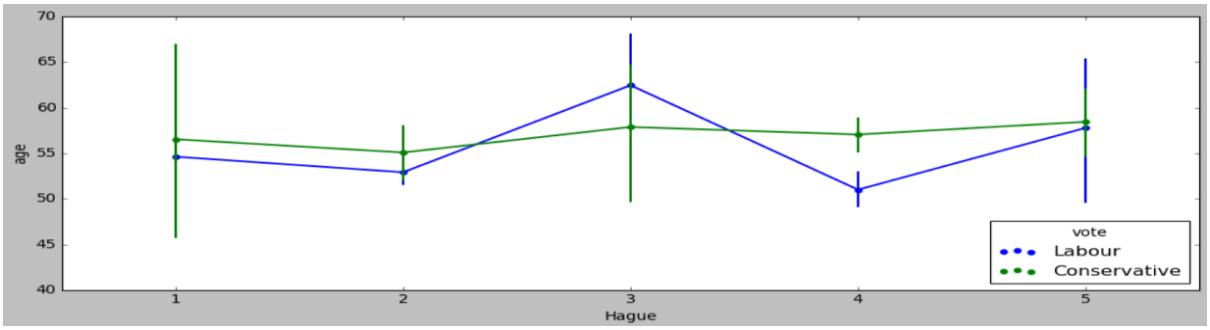
The coefficient for age is -0.23541759961490158
The coefficient for economic.cond.national is 0.5620327056282469
The coefficient for economic.cond.household is 0.05699068559850641
The coefficient for Blair is 0.710025879233481
The coefficient for Hague is -1.0219410257175439
The coefficient for Europe is -0.6984592770125304
The coefficient for political.knowledge is -0.3525771010165865
The coefficient for gender is 0.09934857345390612

'Hague' is the most important feature among all the features of the dataset and **'Blair'** is the second most important feature**. This indicates that voter is highly influenced by the assessment of the leader of the party.**

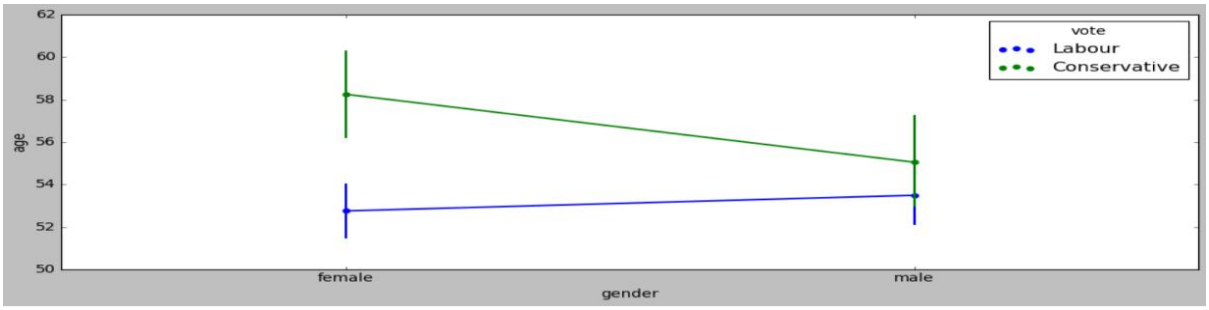




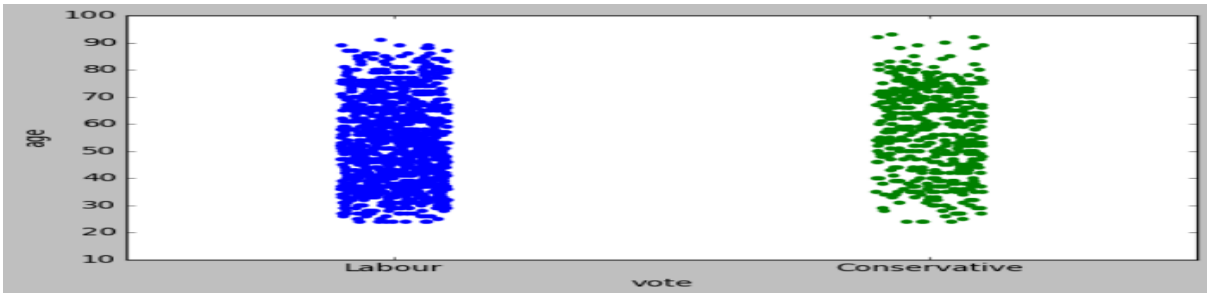
From above plots, we can infer that higher the level in variables ‘Hague’ and ‘Blair’ greater is preference towards Conservative party.



From above point plot , we can clearly understand that Labour party is more preferred at medium level of ‘Hague’



In case of both gender , greater the age of the voter ,higher is preference towards conservative party.



From above plot , we can infer that younger and older voters are less preferred towards conservative party.

Therefore we can analyze voters based on the feature importance to get the better results in predicting to predict which party a voter will vote

So, The Overall analysis of given dataset definitely helped to get insights that would help to create an exit poll that will help in predicting overall win and seats covered by a particular party.