CS253 Python Assignment

Apoorv Tandon

GitHub

1 Methodology

1.1 Data Exploration and Feature Engineering

Data was listed and various functions of matplotlib used to gain insights into how the data was spread out and possible correlations/patterns in the data.
I found that Names starting with Dr. and Adv. indicated a higher level of education, whereas no other details of the Candidate column contributed to improving the models. So binary(true/false) value columns 'Dr' and 'Adv' were created and 'Candidate' dropped from training data.
Similar process was carried out for the Constituency names: the presence of SC or ST in the name seemed to have a correlation with Education Levels, and so these were included as features in the training data.

As a result, three features(ID, Candidate, Constituency) were dropped and four(Dr, Adv, SC, ST) were included.

1.2 PreProcessing

To make the Assets and Liabilities useful, string data(e.g. 'Crore+') was converted into numerical format and the pre existing columns were replaced by purely numerical data.
Due to large variation of Criminal Case, Assets and Liabilities, the columns were grouped into 10 bins(labelling) uniformly using KBinsDiscretizer from sklearn.preprocessing library. This minimized effect of outliers and removed need for standardization/normalization

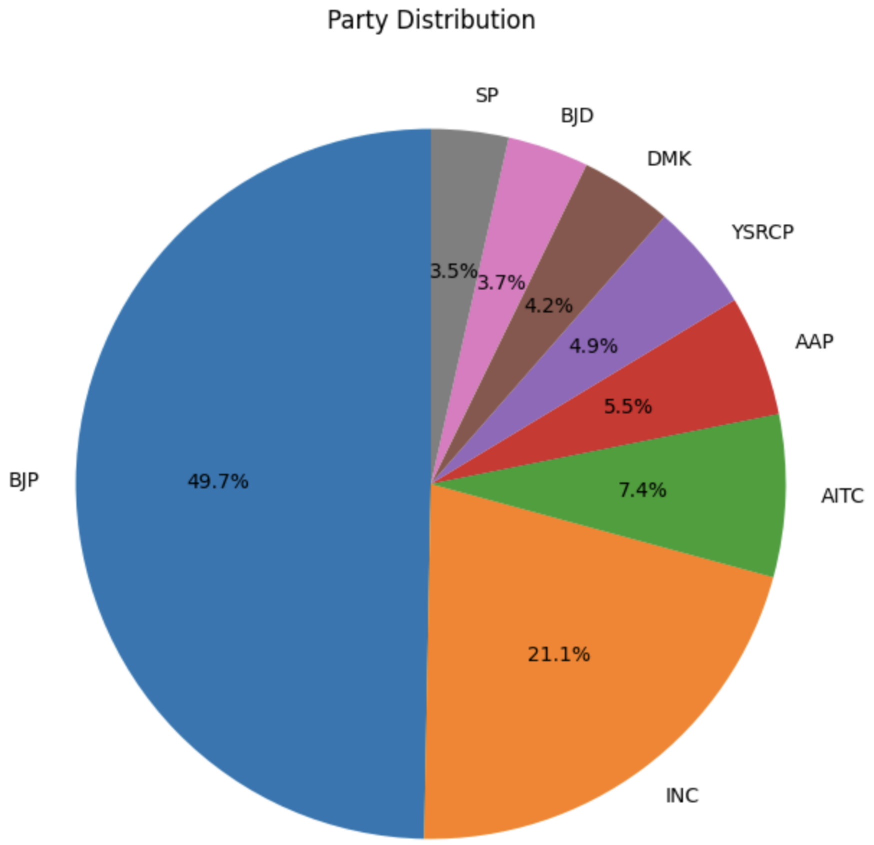
One-hot encoding was used for 'Party' and 'state' columns which collaterally increased the dimensionality of the feature vector.

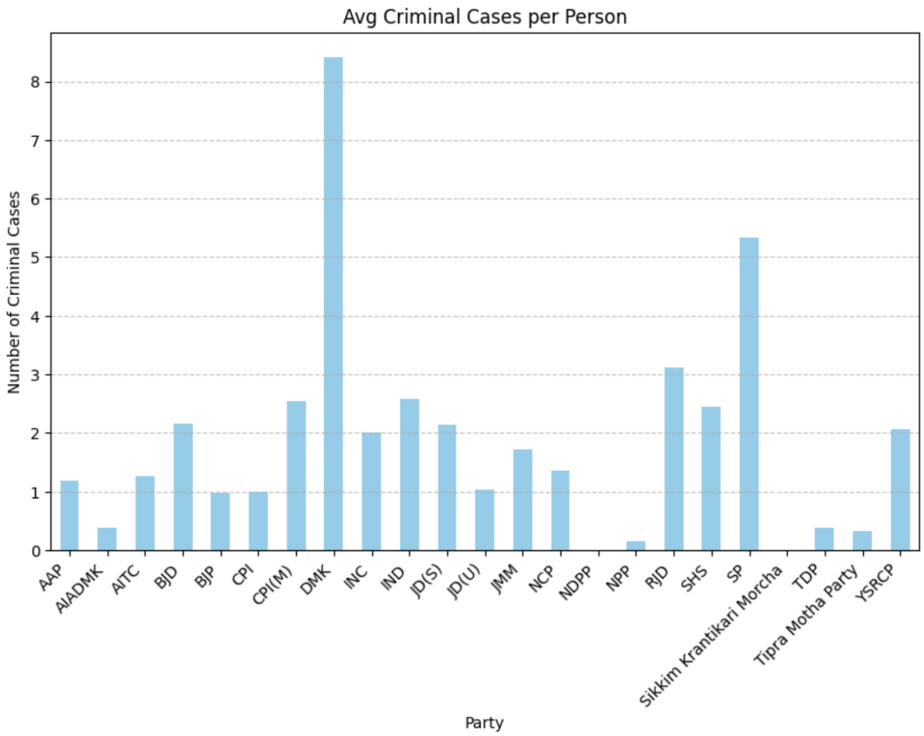
2 Models Used

S No	Model	Parameter Optimum	$F1_{score}$
1	DecisionTreeClassifier	NA	0.21182131322307654
2	RandomForestClassifier	n_estimators=43	0.23611794836727118
3	KNeighborsClassifier	n_neighbors=31	0.2526379012340768
4	MultinomialNB	alpha=0.06	0.2423617508799282(5Fold)

2.1 Data Insights

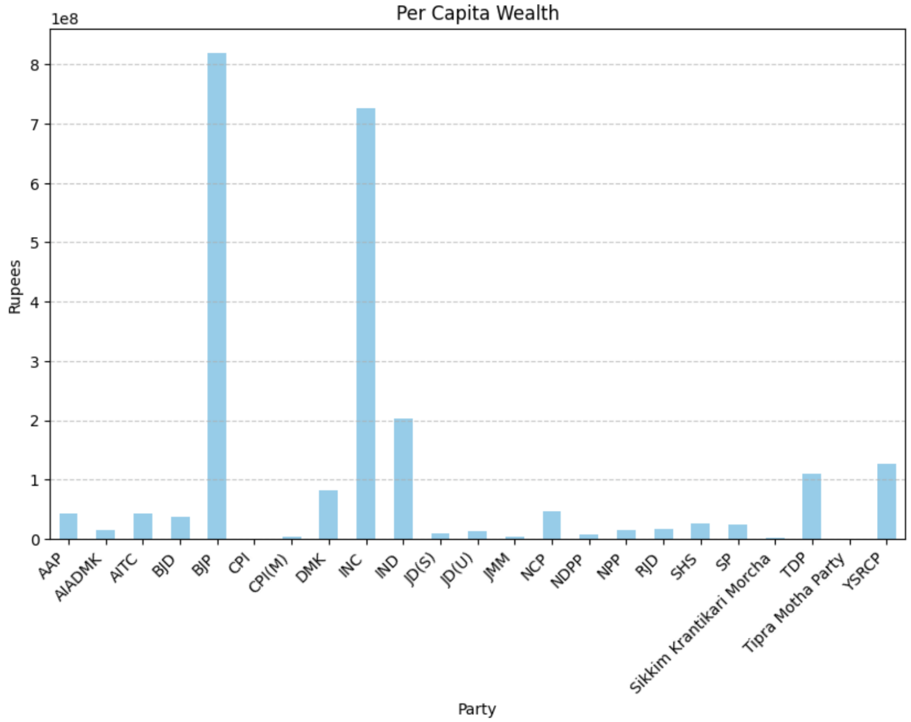
2.1.1 Graphs

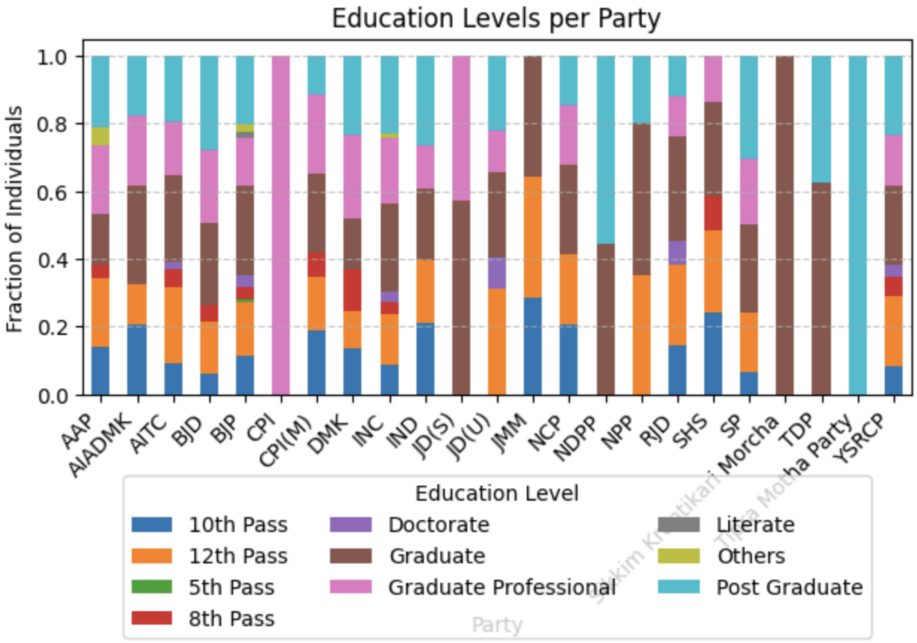




2.1.2 Inferences

The plots lead to interesting inferences.
Some of the most popular parties have the smallest proportions of highly educated members.
These popular parties also happen to have the highest per capita wealth.
Amongst the most criminally involved parties, most have a nearly uniform composition amongst all levels of education.


Per Capita Wealth
1e8
Rupees
0
1
2
3
4
5
6
7
8
AAP
AIADMK
AITC
BJD
BJP
CPI
CPI(M)
DMK
INC
IND
JD(S)
JD(U)
JMM
NCP
NDPP
NPP
RJD
SHS
SP
Sikkim Krantikari Morcha
TDP
Tipra Motha Party
YSRCP
Party


Education Levels per Party
Fraction of Individuals
0.0
0.2
0.4
0.6
0.8
1.0
AAP
AIADMK
AITC
BJD
BJP
CPI
CPI(M)
DMK
INC
IND
JD(S)
JD(U)
JMM
NCP
NDPP
NPP
RJD
SHS
SP
Sikkim Krantikari Morcha
TDP
Tipra Motha Party
YSRCP
Party
Education Level
10th Pass
12th Pass
5th Pass
8th Pass
Doctorate
Graduate
Graduate Professional
Literate
Others
Post Graduate

3 Results

Since cross validation was used to determine efficacy of the MultinomialNB, we obtain mean score of 0.2423617508799282.

Public Leaderboard Rank: 21
Private Leaderboard Rank:

4 References

For KNN training
RandomForestClassifier
DecisionTreeClassifier
KFold
cross$_val_score$
KBinsDiscretizer
MultinomialNB
Basics of ML
Preventing Overfitting by using NB Classifiers
Using KFold Validation, Cross Validation to better detect overfitting
Discretization to overcome very varied data
hyperparameter tuning(Stack Exchange)