Assignment 3

Professor: Dr. Indranil Saha
Student: Kundan Kumar
Roll No. 220568
Date: 9th April, 2024

1 Methodology

1.1 Data Preprocessing Steps

1. **Dropped Useless Field like 'ID' and 'Candidate Name'**

   We dropped the 'ID' and 'Candidate' columns because they do not contribute to the predictive power of the model. The 'ID' column typically serves as a unique identifier for each data point and does not contain any meaningful information for predicting the target variable. Similarly, the 'Candidate Name' column contains the names of the candidates, which are specific to each data point and do not generalize well to new data. Therefore, removing these columns helps reduce noise and simplifies the dataset, making it more suitable for training machine learning models.

2. **Converted 'Liabilities' and 'Total Assets' from string to numeric value.**

   This preprocessing step involved converting the 'Total Assets' and 'Liabilities' fields from string representations of monetary amounts (e.g., '1.2 Crore+', '3 Lac+') to numeric values suitable for machine learning. This conversion was necessary for quantitative analysis and modeling, ensuring the data could be effectively utilized by machine learning algorithms.

3. **Dimensionality Reduction of 'Constituency' field.**

   In preprocessing the 'Constituency ' field, we categorized constituencies into three main groups: 'SC', 'ST', and 'GEN'. This step effectively reduced the dimensionality of the dataset, initially comprising over 100 unique constituency names. By simplifying the feature into broader categories based on social and demographic characteristics, i aimed to capture important information while enhancing model generalization. This categorization strategy not only facilitates easier learning for machine learning models but also ensures that important demographic factors are explicitly encoded, thereby improving the model's ability to make accurate predictions. Overall, this approach streamlines the dataset while preserving relevant information critical for predicting candidate education levels

4. **Grouping Sates into 4 categories 'EAST','WEST','NORTH' and 'SOUTH'.**

   Grouping states into four categories—'East', 'West', 'North', and 'South' and mapping 'East' to 'South' from 1-4 . This was motivated by the significant regional variations in education levels across India. For instance, southern states like Kerala and Tamil Nadu generally have higher literacy rates and better education infrastructure compared to eastern states like Bihar and Jharkhand. By categorizing states based on their geographical location, we enable the model to capture these regional disparities and incorporate them into the learning process. This approach enhances the model's ability to recognize and utilize regional educational trends, ultimately leading to more accurate predictions of candidate education levels. Additionally, it simplifies the dataset by reducing the number of distinct categories, making it easier for the model to process and learn from the data.
   UPDATE: I finally ended up not using this pre-processing,Since instead of this, using one-hot encoding on "state" field gave me better F1 Score.

5. **Mapping level of education in 'Education' field to Numeric value.**

   We assigned higher numbers to higher education levels with the belief that it would give more weight to individuals with higher levels of education in our analysis. By representing education categories with increasing numeric values, we aimed to emphasize the importance of higher education in predicting outcomes or patterns in our dataset. This approach allowed us to prioritize the educational attainment of individuals in our analysis, potentially capturing its impact more effectively in predictive modeling.

6. **Used one-hot encoding on 'Party' field and 'Constituency' field after pre-processing.**
   One-hot encoding was applied to the 'Party' field and 'Constituency' after pre-processing, since there was no discernible correlation between education and political party affiliation. This technique converts categorical variables into binary features, assigning each party a unique binary feature(1-0). It allows machine learning algorithms to interpret party affiliation effectively without imposing ordinal relationships between parties. Thus, one-hot encoding was chosen to handle the categorical nature of political party data while preserving their categorical information.

7. **Used the prefix of 'Candidate' Name like 'Dr.' or 'Adv.' to make the model better.**

   Recognizing the potential significance of prefixes like 'Dr.' and 'Adv.' in candidate names, we decided to incorporate this information into our analysis. By processing the candidate names and mapping the presence of 'Dr.' to the value 2 and 'Adv.' to 1, we aimed to capture any distinction associated with these titles. Other names were assigned the value 0. This approach was implemented to leverage any inherent association between these prefixes and educational attainment or professional status.

2 Library Used

(a) Pandas
(b) Sklearn
(c) Matplotlib

3 Experiment Details

Model	Parameters	F1 Score
BernoulliNB	$\alpha = 0.75$	0.26023
Decision tree	random_state = 45	0.20506
Random Forest	n_estimators = 100	0.21955
KNN	n_neighbors=5, weights='uniform'	0.17433

Table 1: Performance of Different Models

3.1 Data Insights

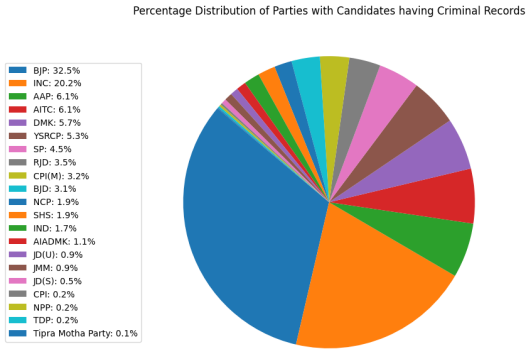

Figure 1: The percentage distribution of parties with candidates having the most criminal records.

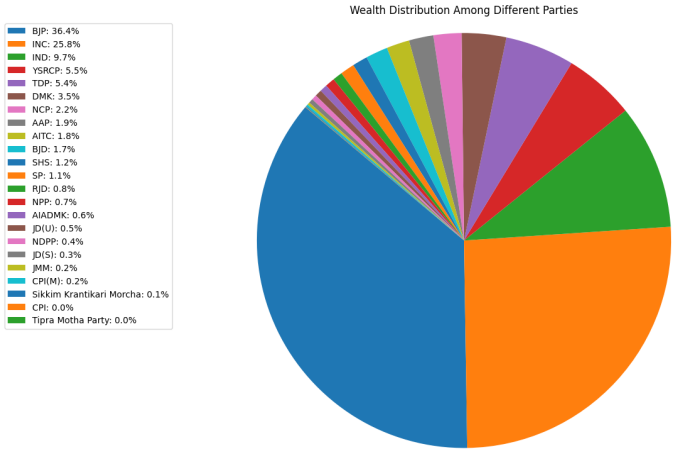

Figure 2: Wealth distribution among Politician.

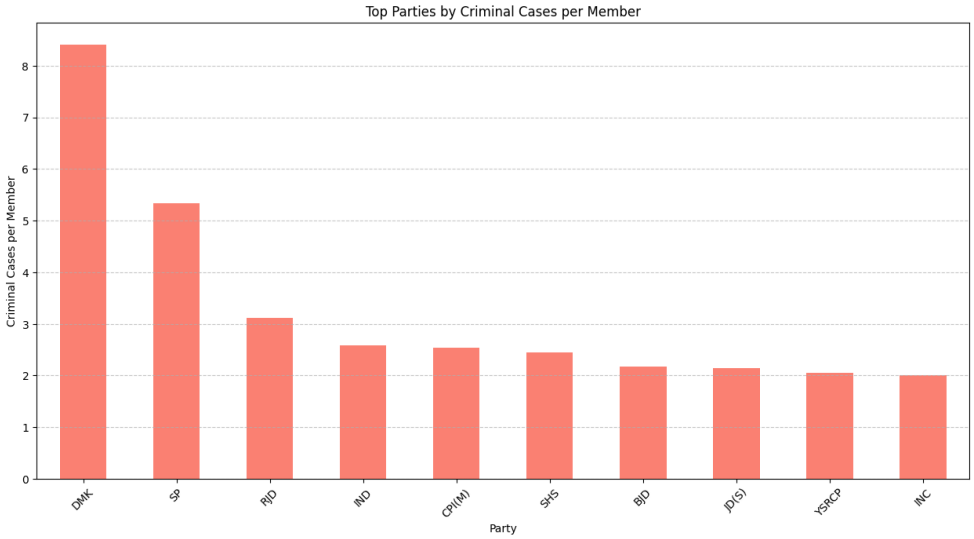

Figure 3: Average number of criminal cases per person party wise.

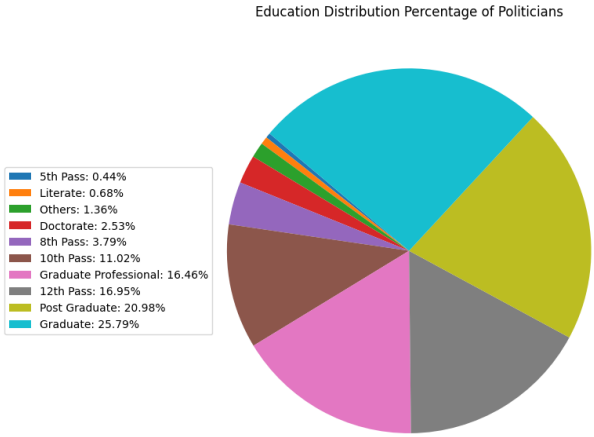

Figure 4: Education distribution of the Politician.

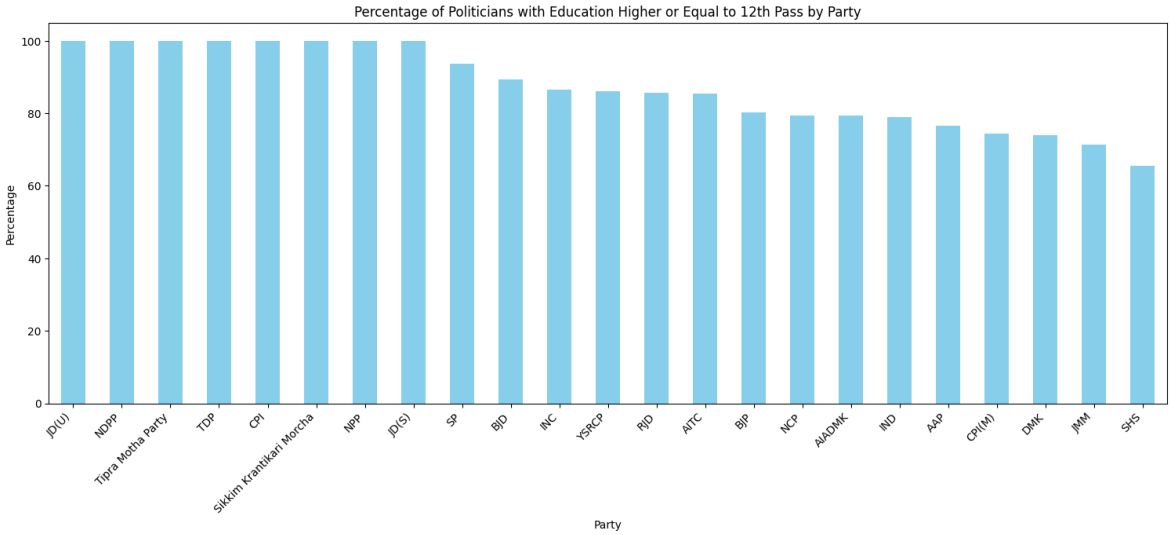

Figure 5: Percentage of politician with education higher or equal to 12th pass Party wise.

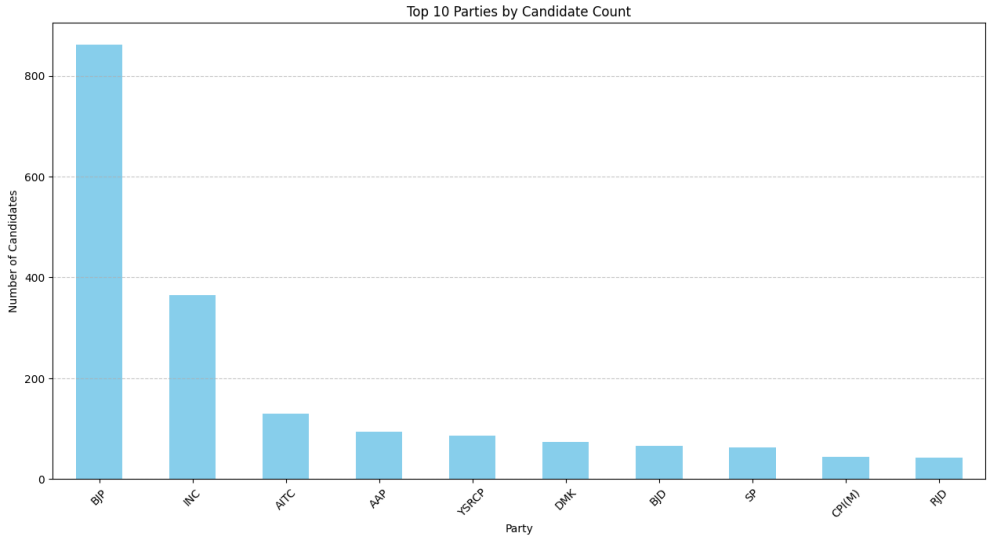

Figure 6: Candidate Distribution of the given data Set.

3.2 Some Insights from the Plots

1. Two biggest parties of India 'BJP' and 'INC' have highest percentage of member who have criminal cases against them.
2. Most of the wealth among Politician is with 'BJP' and 'INC' party. This data may be influenced by the fact that no. of politician that we have in our dataset is far more from BJP and INC Party. Figure 6 shows the distribution.
3. If you look at average number of criminal cases per person in a particular party then 'DMK' party comes at the top.
4. Education distribution of the politician shows a less percentage of politician only approx 63% is well educated , that is having a gradate degree or above.

5. Party which have most educated member is JD(U). Also INC have more educated members than BJP.

4 Results

Final Model:- Bernoulli NB

Final Public F1 Score:- 0.26023

Final Private F1 Score:- 0.25777

Public Leaderboard Rank:- 38^{th}

Private Leaderboard Rank:- 23^{th}

5 References

Tutorial on how to build ML Model:-
https://dev.to/mage_ai/buildingyourfirstmachinelearningmodel40lc

What is Corelation:-
https://medium.com/@abdallahashraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9

One Hot Encoding: https://deepchecks.com/glossary/one-hot-encoding/

How Random Forest work:-
https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Understanding Hyper-parameter of Random Forest:
https://www.analyticsvidhya.com/blog/2021/03/introduction-to-random-forest-and-its-hyper-parameters/

Understanding Bernoulli NB:-
https://medium.com/@pa3lo/naive-bayes-classifier-0f0e7c9f86c8

For Plotting using Matplotlib
https://www.w3schools.com/python/matplotlib_intro.asp