CS253 Python Assignment Report

Akanksha Wattamwar
221214
akankshab22@iitk.ac.in

Abstract

This report is written to give an overview of the model used for multi-class classification for the problem statement of the competition **Who is the real winner?**

1 Model Used:

This section includes the details of the various models I tried for training. The following models were used with features as **State, Party, Criminal Case**. I have not used **Liabilities** and **Total Assets** for model training as using them reduced the score of the model due to their very low correlation with our target **Education**.

1. **Decision Tree :** This model gave me the highest **f1 score** of **0.24793**. The hyperparameters used for this model are
   - max_depth=10
   - min_samples_split = 5
   - min_samples_leaf = 1

2. **K Nearest Neighbors :** This model gave the **f1 score** of **0.23006**. The hyperparameters used for this model are **n_neighbors = 1**.

3. **Random Forest Classifier:** This model gave the **f1 score** of **0.20958**. The hyperparameters used for this model are
   - max_depth=10
   - min_samples_split = 2
   - min_samples_leaf = 2
   - n_estimators = 100

4. **LinearSVC :** This model gave **f1** score of **0.157318** with the optimum value of $C = 0.01$.

I also tried training the models using Decision tree and KNN statewise but that didn't give me higher score. However, these models are included in the submitted code.

2 Data Analysis:

This section gives an idea of the distribution of candidates based on different features like State, Party, Criminal Case, Total Assets, and Liabilities.

Preprint. Under review.

1. **Distribution of Education Categories :**

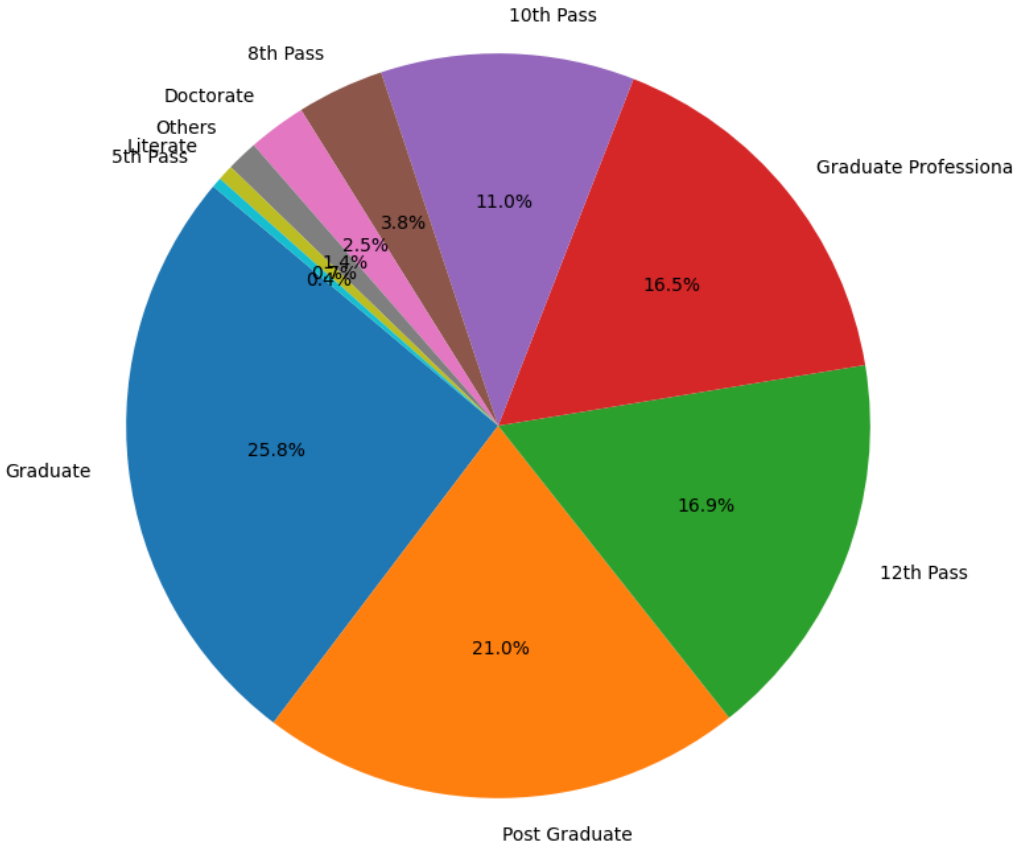

Figure 1: Percentage Distribution of Education Categories

We can see from the above figure that most of the candidates are Graduate, Post Graduate, 12th Pass or Graduate Professional.

2. **Distribution of Candidates with most criminal records among the parties:**

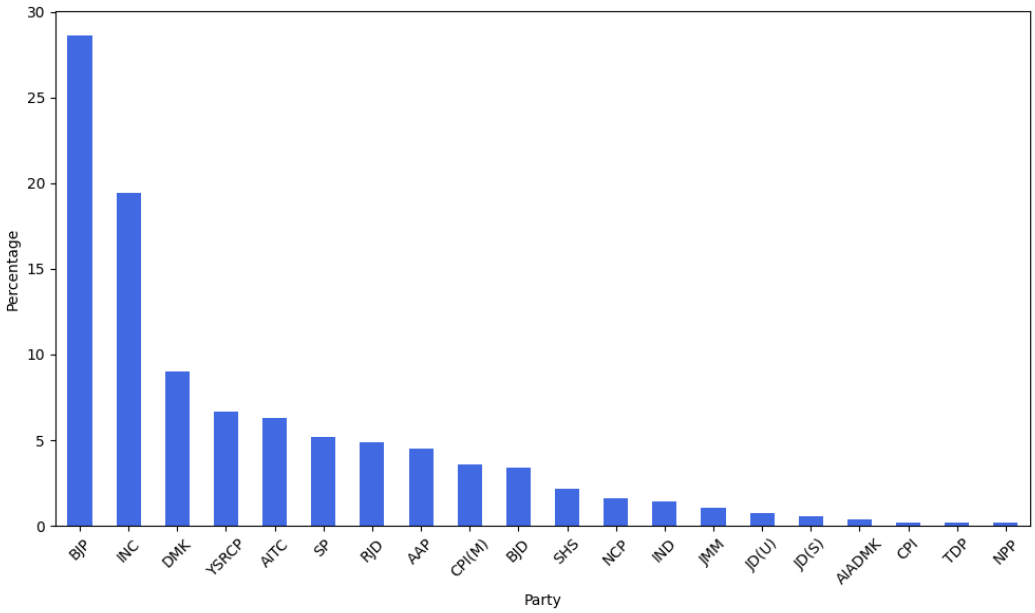

Figure 2: Percentage Distribution of Parties with Candidates having no. of Criminal Cases more than Mean of the overall Criminal Cases.

I have considered a candidate as having most criminal cases if he is above the mean of all criminal cases.
As evident from Figure 2 which is shown below, **BJP** has around **28%** of candidates with no. of criminal cases above mean. After BJP, **INC** has around **24%** of candidates with no. of criminal cases above mean.

3. **Distribution of Candidates with most criminal records in each of the party:**

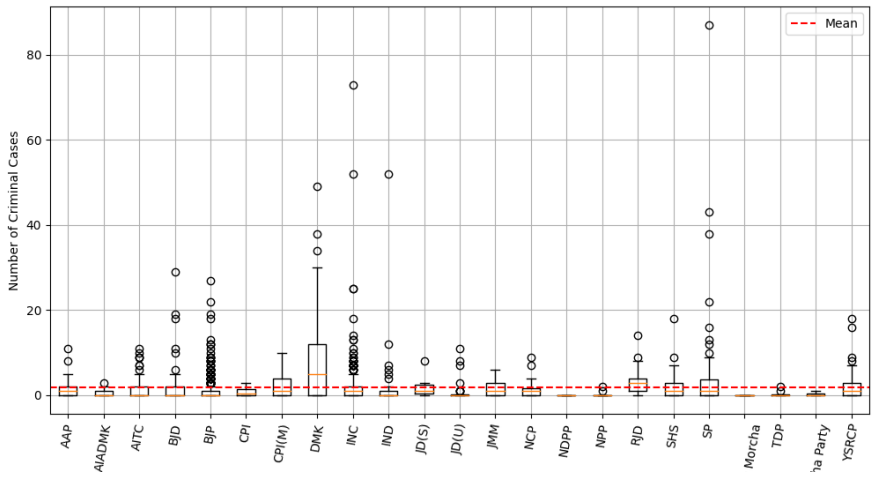

Figure 3: Distribution of Criminal Cases for Each Party

As seen from the figure above, the parties like SP and INC have candidates have more than 40 criminal cases.

4. **Distribution of most wealthy candidates among the parties:**

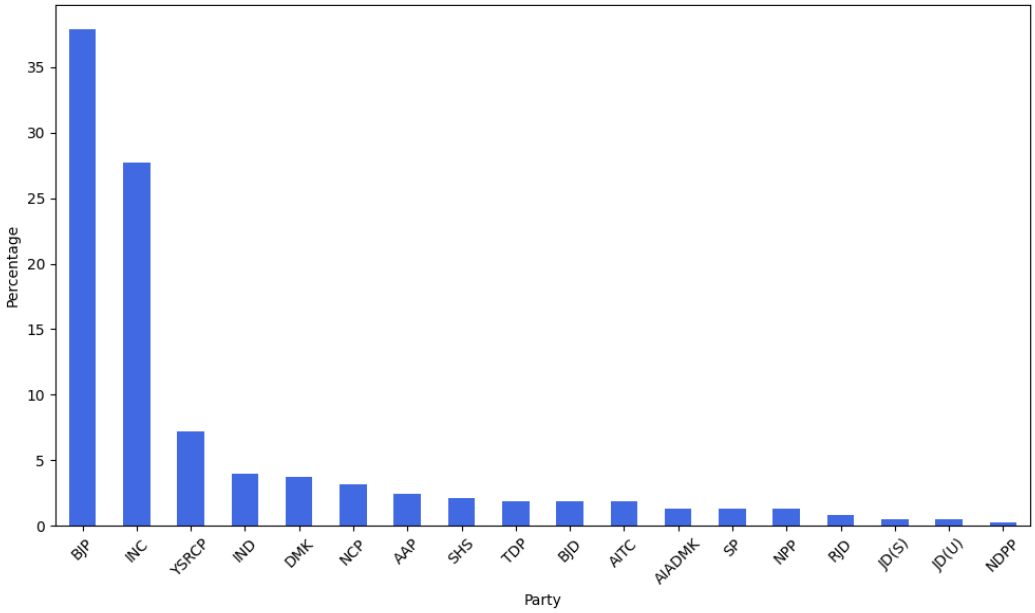

Figure 4: Percentage Distribution of Parties with Candidates having Total Assets greater than Mean

I have considered candidates having total assets value greater than the mean of total assets as most wealthy candidates.
Evidently from the below graph, BJP has around **38%** of the most wealthy candidates and INC has around **27%**.

5. **Distribution of most wealthy candidates among the parties:**
Except some outliers, **BJP** and **INC** has the most wealthy candidates.

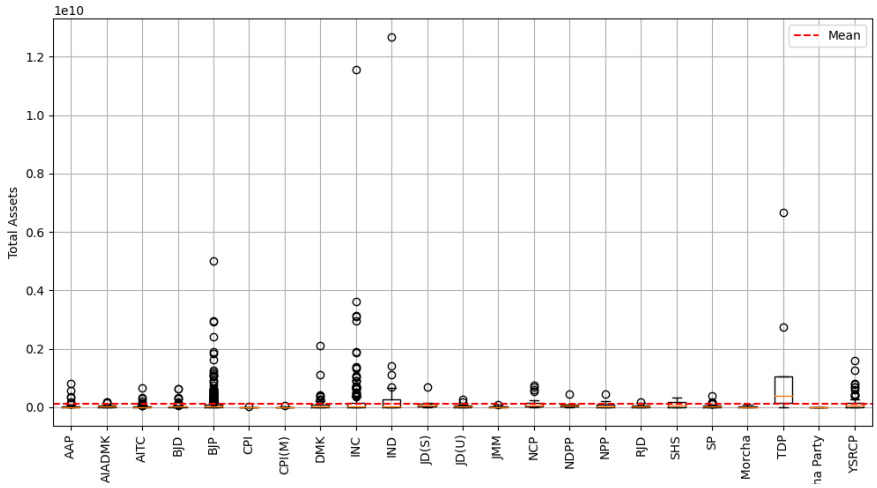

Figure 5: Distribution of Total Assets for Each Party

6. **Distribution of candidates among the parties based on liabilities:**

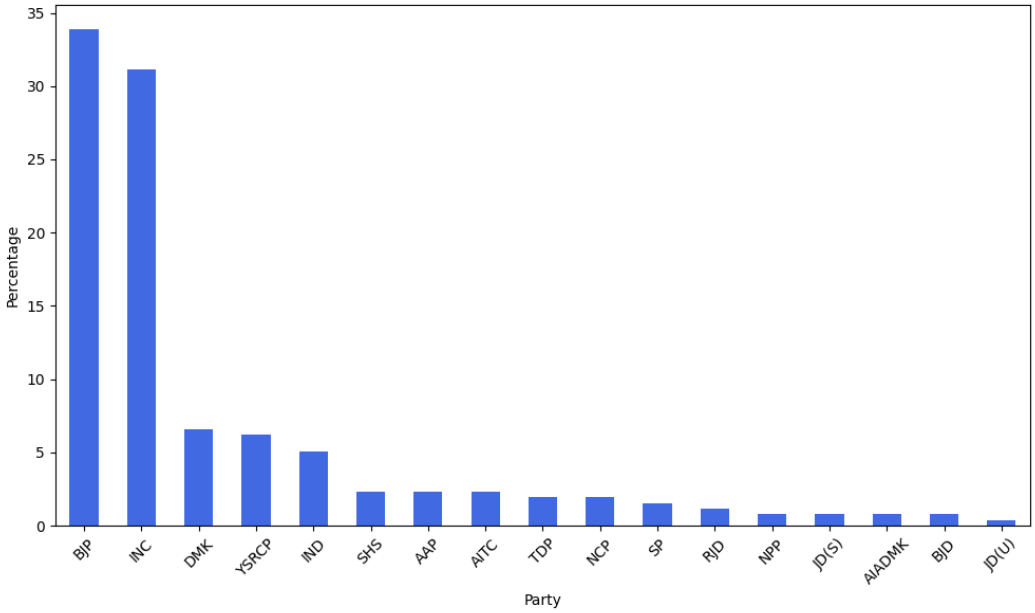

Figure 6: Percentage Distribution of Parties with Candidates having Liabilities greater than Mean

Evidently from the below graph, BJP has around **34%** of the candidates having liabilities more than mean and INC has around **32%**.

7. **Distribution of candidates in each party based on liabilities:**
All parties except **BJP** and **INC** have candidates with very less liabilities.

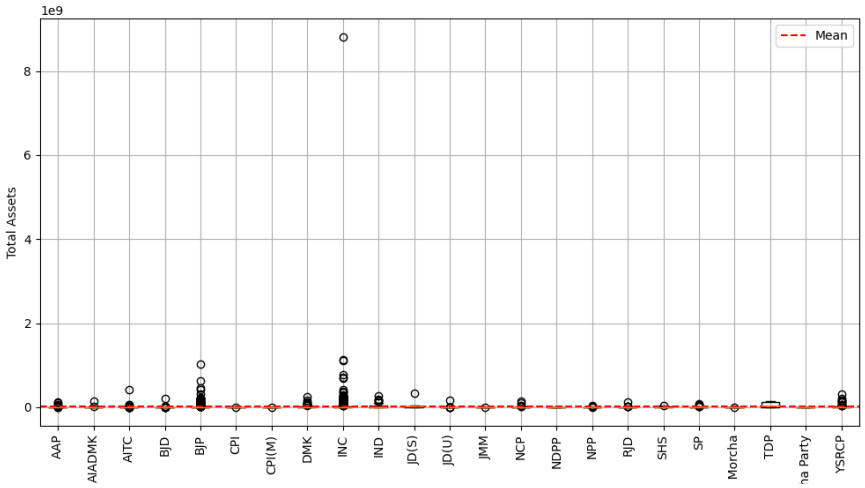

Figure 7: Distribution of Total Assets for Each Party

8. **Statewise distribution of candidates based on Education categories :**

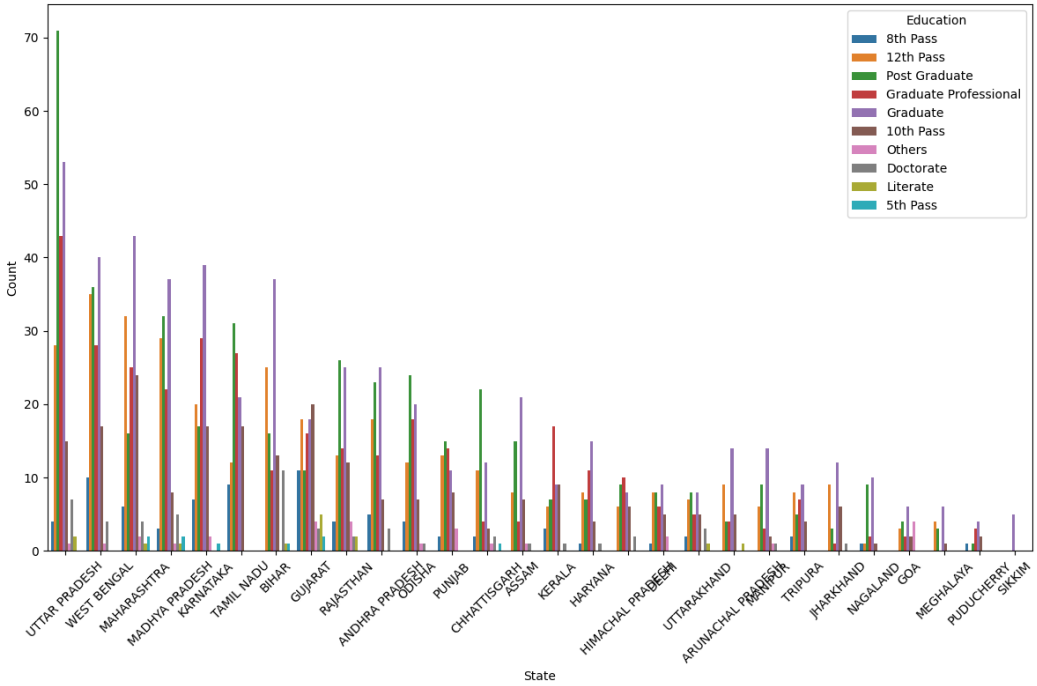

Figure 8: Most Common Education Category for Each State

Evidently, **UP** has the most of the educated candidates and **Sikkim** has the least.

3 Results:

My final **leaderboard score** is **0.24793** and **rank** is **70th**.

The link to my github repository is `https://github.com/akankshabw/Python_Assignment`