Python Assignment Report

Manan Kumar

April 13, 2024

1 Methodology

A KNN classifier model [6] with grid search [2] and decision tree regressor [3] was prepared (finally, only KNN was selected for submission) for the given task, which includes the following common major sub-tasks:

1. Columns for state and party were encoded using the label encoder function. Also, the columns for total assets and liabilities had order specifying terms like thou+, crore+, and hund+, which were removed, and the values were typecasted from string to integer. The education level field had elements of type string, which were encoded to numbers.
2. A new assets-to-liabilities ratio was introduced, keeping in mind that the liability for a candidate can be zero, so it was shifted by 1 to remove arithmetic errors.
3. All the data in the .csv file was scaled [1] for better performance. In the case of the KNN classifier, Grid Search was implemented to find the best combination of possible parameters.
4. Utilised the fact that all candidates with Adv. in their name had an education level as Graduate Professional and also the fact that candidates with Dr. in their name never had an education level below Graduate.

2 Experiment Details

lightgray **Method**	**Parameters**
KNN classifier	'metric': 'chebyshev', 'n_neighbors': 29, 'weights': 'uniform'
Decision Tree Regressor	'random_state': 1

Table 1: Methods Used

3 Data Insights

3.1 Plots

1. Percentage distribution of parties having candidates with the most criminal records. [4]
2. The percentage distribution of parties with the net worth of candidates.
3. Percentage of candidates in each education level.[4]
4. Percentage of candidates for each state who have higher education that is 12th grade or above. [5]

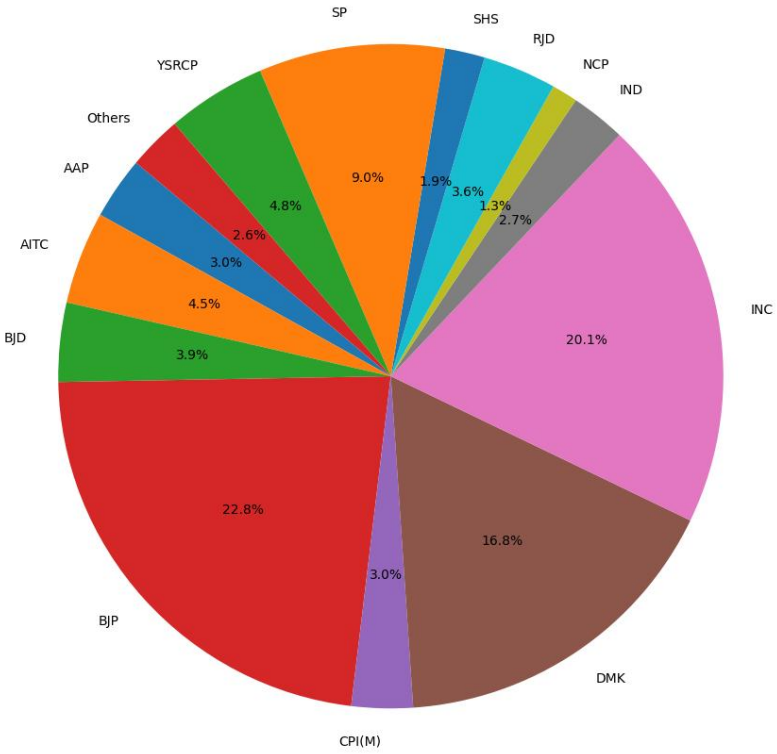

Figure 1: Distribution of Criminal Cases among Parties

3.2 Insights

1. The total number of candidates in a party with criminal records depends upon the size of the party.
   a. **INC (Indian National Congress)**: Constitutes approximately **20.15%** of the distribution.
   b. **BJP (Bharatiya Janata Party)**: Represents around **22.8%**.
   c. **Others**: This category, which is for smaller parties, accounts for **9%**.
   d. Smaller segments include **YSRCP, AAP?, NTC, BJD, CPI(M), DMK, IND, NCP, RJD, SHS, and SP**, each with percentages ranging from **3.03% to 6.48%**.
2. The total net worth of candidates in a party depends on the size of the party. The BJP (Bharatiya Janata Party) holds the highest net worth at 36.4%, followed by the INC (Indian National Congress) with 25.8%. Smaller parties like IND, TDP, YSRCP, AAP, and others collectively comprise the remaining shares. This distribution reflects the financial strength of various political entities.
3. The greatest number of candidates fall under the education level of graduate followed by post-graduates and then 12th pass.
4. Plot shows the percentage of candidates in a state that have education level 12th Pass or above.

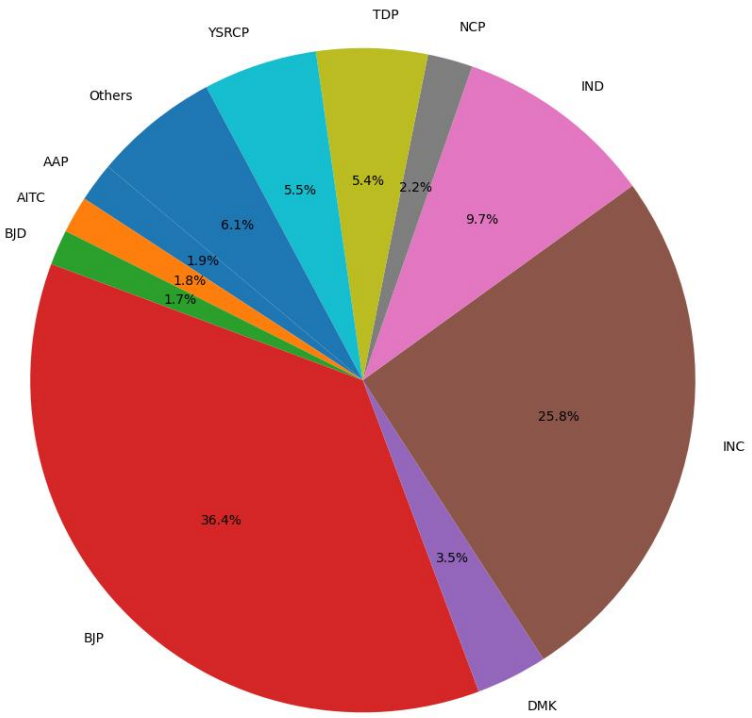

Figure 2: Distribution of Parties based on net worth

4 Results

1. **F1 score**: 0.24141 (Private) and 0.22842 (Public)
2. **Private Standings**: 77
3. **Public Standings**: 147

5 References

1. Minmax scaler geeks for geeks
2. Grid Search
3. Introduction to ML Kaggle Tutorial
4. Pie plot using matplotlib library
5. Horizontal histogram plot
6. KNN Tutorial

6 GitHub Code Repository

https://github.com/Vulcan-Fire/Kaggle-who-is-the-real-winner-/tree/main

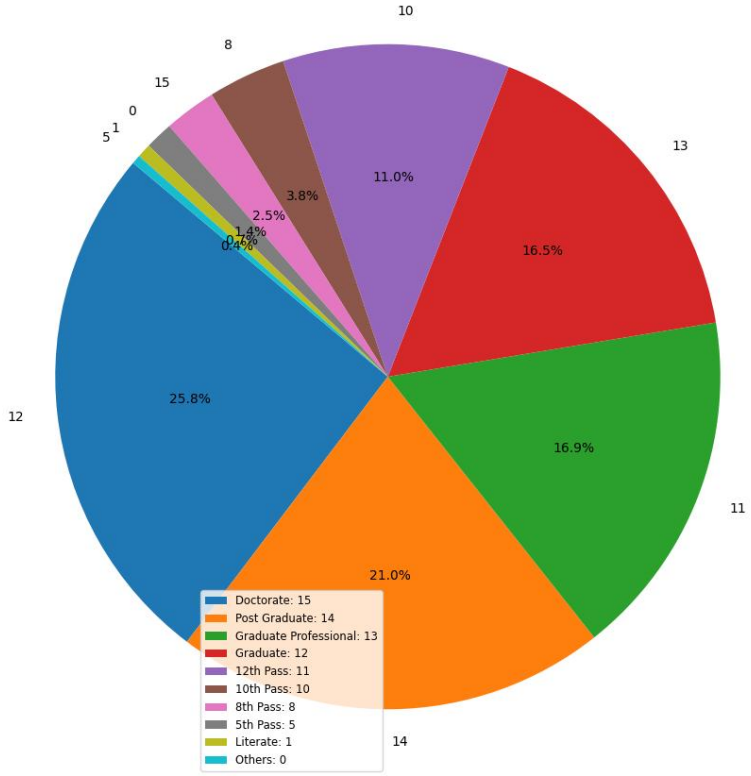

Figure 3: Distribution of candidates based on education level

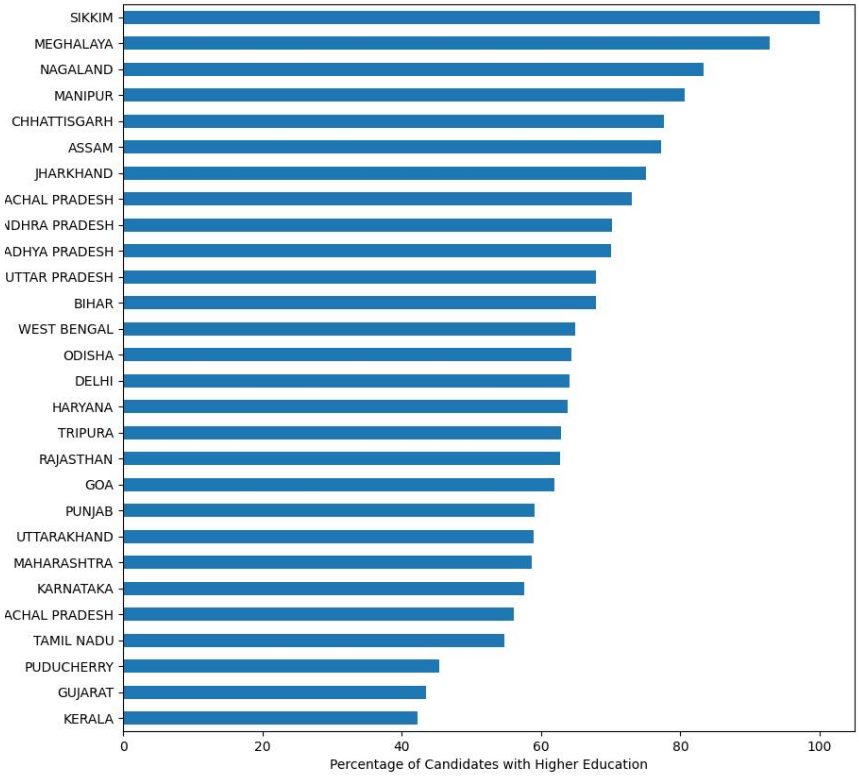

Figure 4: Percentage of higher educated candidates in each state