Python Assignment Report

Mahaarajan J

13th April 2024

1 Methodology

1.1 Data Preprocessing Steps

The data is pre-processed in the following way after loading them into dataframes:

1. A dictionary is prepared to map the 10 education categories to integers from 0 to 9. And it is applied on the **Education** column of the data (both test and train) to map it.
2. Similar dictionaries are created to map the fields **Party** and **state**. These maps are then applied to the respective columns of the data.
3. A inverse of the education map is also created to map the integers back to the original strings they represent.

1.2 Feature engineering

It was found that in the **Candidates** column some names had the designation **Dr.** or **Adv.** which prompted me to create 2 new features (columns) in the data **IsDoctor** and **IsAdv** which have entry 0 when false and 1 when true by applying **startswith** method on the column **Candidates**.
These features clearly do indicate the education (target variable) of the individual and on inclusion were found to increase the **F1_score**

1.3 Dimensionality reduction techniques

I have reduced the dimensionality by removing certain irrelevant columns such as **ID**,**Candidate** and **Constituency**.

I have also removed the **Total Assets** and **Liabilities** because since this is the data of politicians their education will not be much dependent on their wealth as most of this wealth would have been accumulated after they became politicians.

This is justified as I was able to see a drop in **F_1 scores** when these columns were included

1.4 Normalization, standardization, or transformation used

I standardised the data using the **StandardScaler** method. This method centers the data (mean=0) and scales it so that variance becomes 1. It is a common transformation step which greatly improves the performance particularly for algorithms that are sensitive to the scale of input features, such as support vector machines (SVMs), k-nearest neighbors (KNN), and neural networks. It also helps to compare different features of different units

2 Experiment Details

In this experiment I have used different models to learn the data and I shall list all them with the appropriate hyperparameters used and F1_score.

I passed a list of hyper-parameters and applied **GridSearchCV** method on them while optimising the **F1_score_weighted** to obtain the best parameters. The table containing the models used. Best hyperparameters obtained and the F_1 scores obtianed by submissions are:

Table 1: Model Evaluation

Model Name	**Best Hyperparameters**	**F1 Score**
KNN	'algorithm': 'ball_tree', 'leaf_size': 10, 'n_neighbors': 20, 'p': 1, 'weights': 'uniform'	0.265
Random Forest Classifier	'n_estimators':100, 'max_depth': None, 'min_samples_split':5, 'min_samples_leaf':2	0.237
Extratrees Classifier	'n_estimators':200, 'max_depth': 10, 'min_samples_split':5, 'min_samples_leaf':2	0.214
Decisiontrees Classifier	'max_depth': 15, 'min_samples_split':5, 'min_samples_leaf':2	0.242
Support Vector Classifier	'C': 100, 'gamma':0.01, 'kernel':'linear'	0.196

Hence the best performing model is **KNN** which has resulted in my best F_1 score submission of **0.26498**

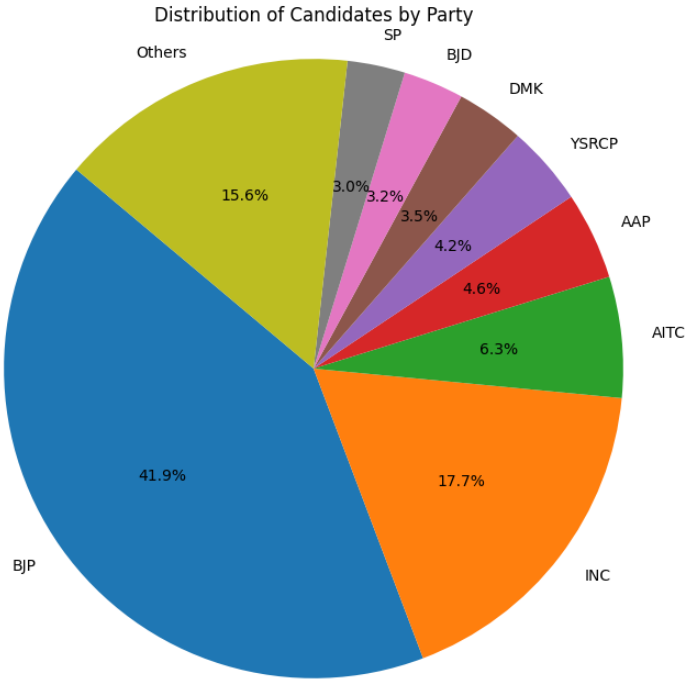
3 Data Insights

The following data insights have been made on the **train.csv** file provided to us.

For the First 2 parts let me define what **most criminal records** and **most wealth** mean
I shall define most as anything above and including the 75th percentile of the data.
Which on calculation for **Criminal Cases** and **Wealth** (assets-liabilities) using the **quantile** function in **pandas** came out as **2** and **7Cr** respectively.

3.1 Distribution of candidates among parties

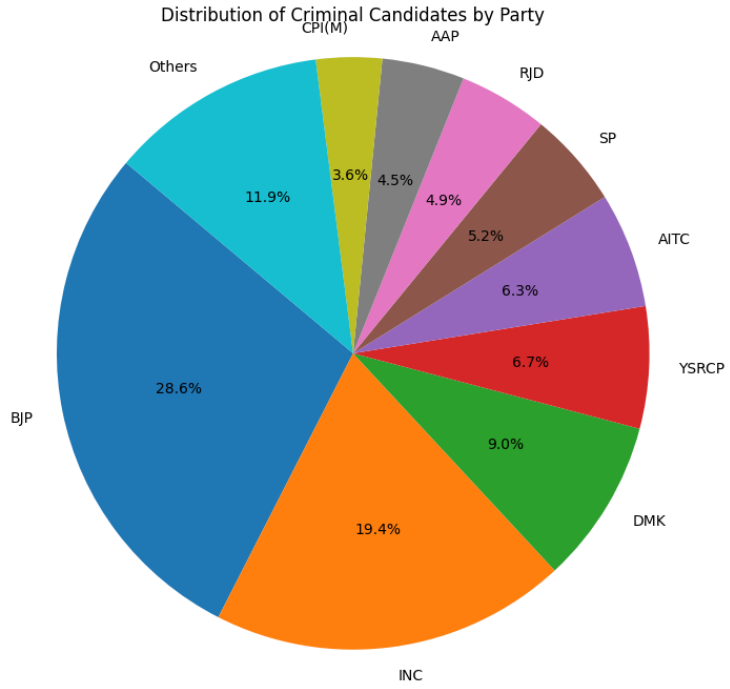
Pie chart



In the above pie chart we have only included parties which have more than 3% candidates.
It is visible that majority of the candidates are from BJP and about 60% of the candidates are just from the 2 largest parties

3.2 Distribution of candidates with most criminal cases among parties

Pie chart



Analysis

The data shows a significant shift in party distribution when considering individuals with high criminal cases.

BJP and INC, major parties overall, still have a notable presence in this group.

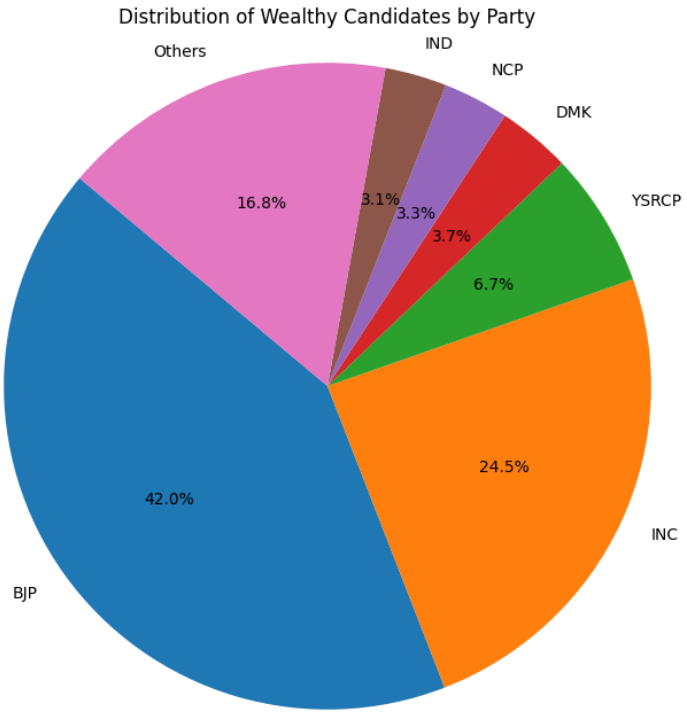
The variation in percentages suggests parties may have a higher proportion of candidates with legal issues relative to their overall representation.

This highlights challenges in public perception, credibility, and governance, emphasizing the need for transparency and accountability in candidate selection processes.

Voters may consider such data during elections, reflecting on candidates' integrity and ethical standards, while media and civil society play a role in scrutinizing candidates and parties.

3.3 Distribution of candidates with most wealth among parties

Pie chart



Analysis

The data on wealthy candidates shows that major parties like BJP and INC maintain a significant presence among them, similar to their overall distribution.

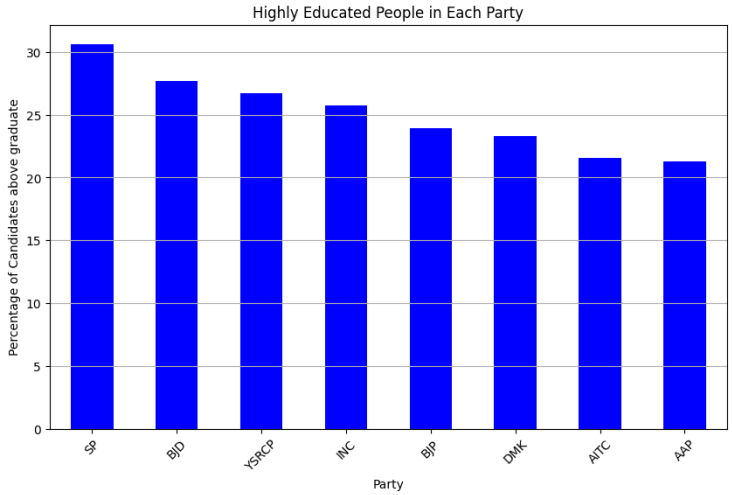
This indicates that wealthier individuals are active in politics across party lines.

Additionally, the presence of wealthy candidates in regional parties and among independents suggests a diverse landscape where financial resources play a role in electoral dynamics.

Voters may consider wealth as a factor in candidate evaluation, and parties may strategically field wealthy candidates for various reasons, including campaign funding and influence.

3.4 Percentage of highly educated candidates (more than or PG level) in the major parties

Bar graph



Analysis

The data on post-graduate percentages in different parties reveals a diverse educational background among candidates.

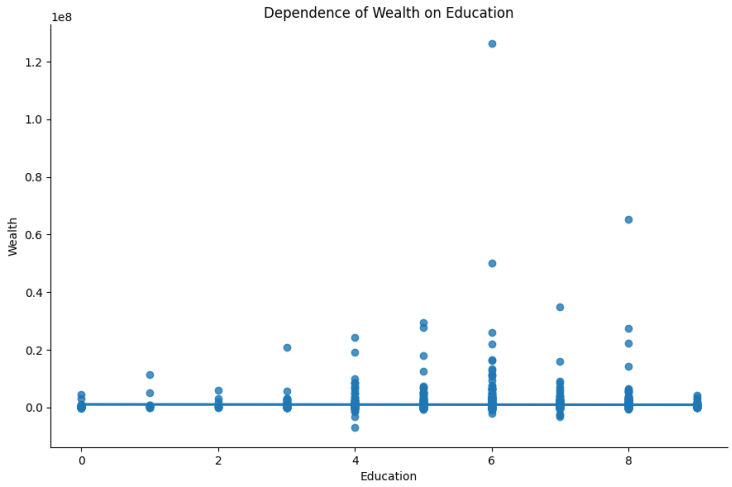
Parties like SP, BJD, and YSRCP have higher post-graduate percentages, while national parties like INC and BJP also have significant numbers.

This diversity indicates a mix of expertise and educational qualifications within political leadership, potentially influencing voter perceptions and policy formulation.

Overall, the data underscores the importance of educational diversity in political representation and governance.

3.5 Relation between Wealth and Education

Linear Regression Plot



Analysis

This linear regression model was made using the **Seaborn** library and illustrates the relation between wealth and education level.

The education level here has been encoded from 0-9 with 9 being Doctorate.

The line is parallel to the X axis which indicates that there is no linear relation between the variables which implies that Education and wealth may be independent from each other.

This is also plausible as in the political field one would think that the wealth of the individual does not depend on the education they had but would depend on the party they are from and the political clout they have.

This was the reason in the ML model used finally we did not include assets and liabilities as features.

4 Results

Final F1 score on public: **0.26498**
Final F1 score on private: **0.26740**
Public Leaderboard Rank: **24**
Private Leaderboard Rank: **10**

5 References

The following references were used during the assignment:

1. Scikit-learn models (All my models were imported from scikit library) and also GridCV library for finding the optimum values of hyperparameters, and other Python libraries like Pandas used for data preprocessing and model training. [Link]

2. Also used matplotlib and numpy for creating the pie charts used in this report and overleaf for LaTeX. [Link]

3. The video shared by Saqib Sarwar on F1 Score to understand the concept. [Link]

4. Kaggle forums and discussions for insights on Random Forest Classifier Model.

5. Scikit website for learning about various models and examples like k-nearest neighbours, linear regression.[Link]

6. The web for learning about KNN, Random Forest, Naive Bayes, GridCV, and Linear Regression.

7. Python lectures from CS253 course.

6 Code

Github Link: [Code]