Assignment 3 : Python Report

Utpal Dwivedi (211132)
Dept. of Biological Engineering
utpaldwi21@iitk.ac.in

1 Model Used : Decision Tree

The model used by me that gave the best performance on the unknown test data is **Decision Tree** , even though I tried Random Forest and kNN classifier with hyperparameter tuning.

In **Data Preprocessing**, I modified the categorical data into numerical one by using Label Encoder. 5 features were extracted from train.csv file and stored in encoded format.

The **hyperparameters** used in Decision Tree can be seen below :
'criterion': ['gini', 'entropy'],
'max_depth': [None, 10, 20, 30],
'min_samples_split': [2, 5, 10],
'min_samples_leaf': [1, 2, 4]

I have also used **Grid Search Cross Validation Technique** to get the optimum parameters with cv=5. Out of the top 2 performing codes written by me in Decision Tree with **f1 score** nearly **0.22** and **0.2** respectively, I trained over all the training data and got score 0.22 while in case of 0.2 score I divided the training data into two sets , **80%** for training and **20%** for validation.

Other Models Experimented :

1. **Random Forest** n_estimators=100, max_depth=7, random_state=1
2. **kNN** k=10

In conclusion, the thorough experimentation and evaluation process with various machine learning models led to the selection of Decision Tree as the primary model due to its superior performance on the test data. The combination of thoughtful feature engineering, hyperparameter tuning, and model validation techniques contributed to the success of this project.

2 Data Analysis

2.1 Parties with Candidates having the Most Criminal Records

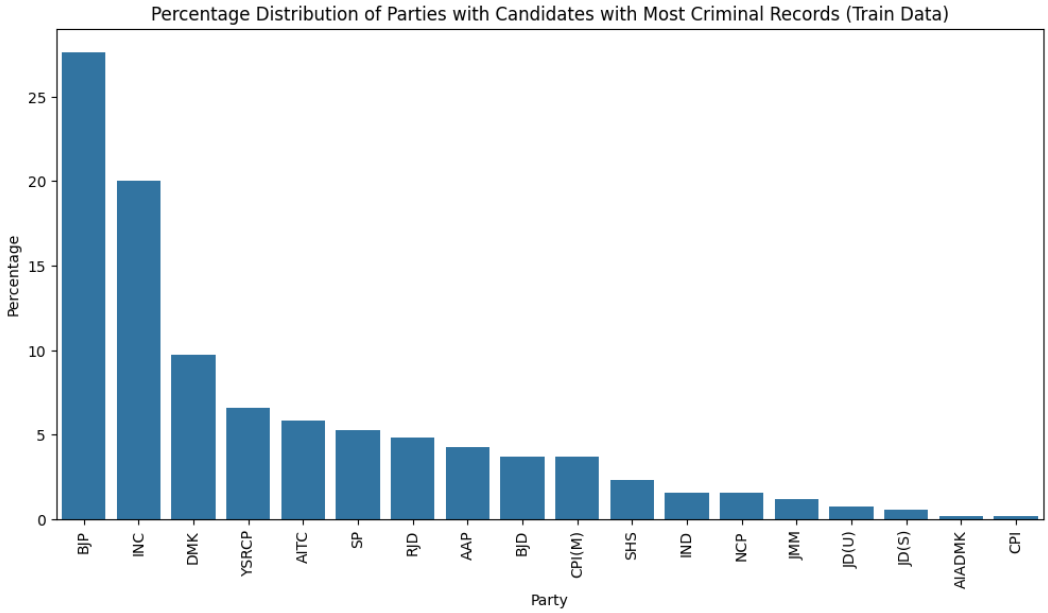


Figure 1: Train Data

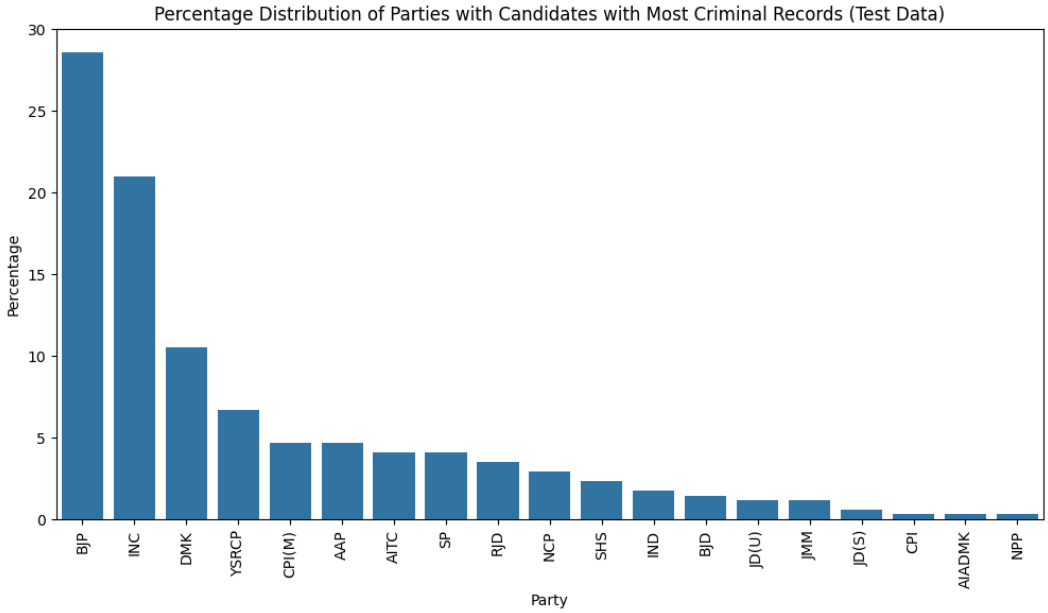


Figure 2: Test Data

2.2 Parties with the Most Wealthy Candidates

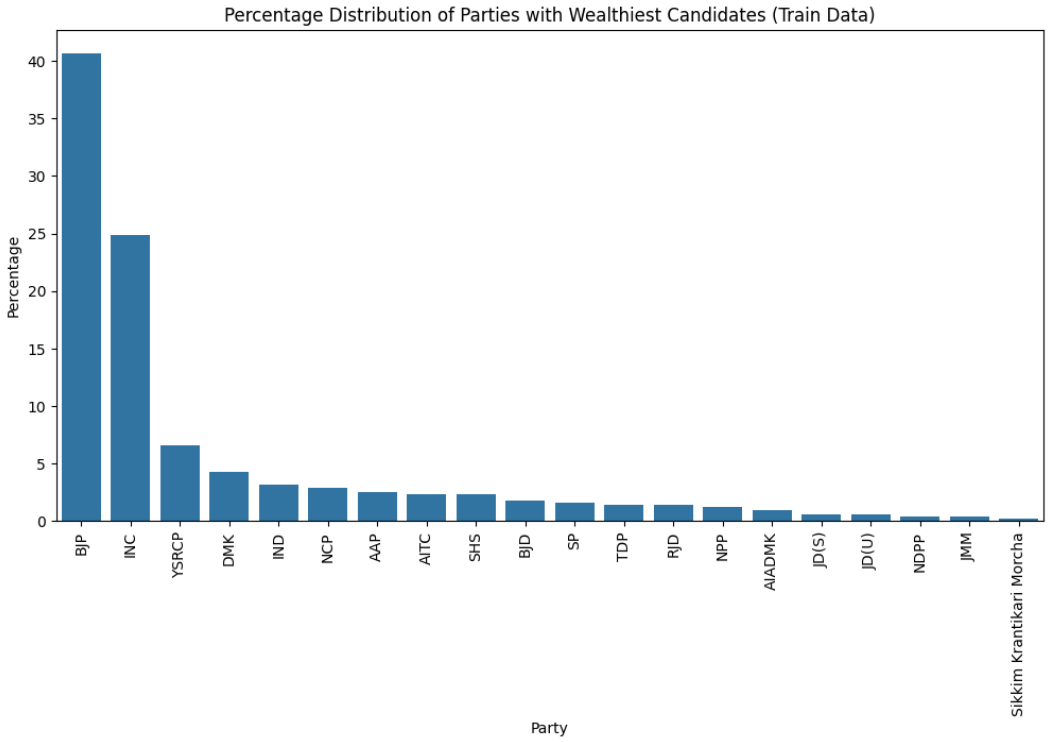


Figure 3: Train Data

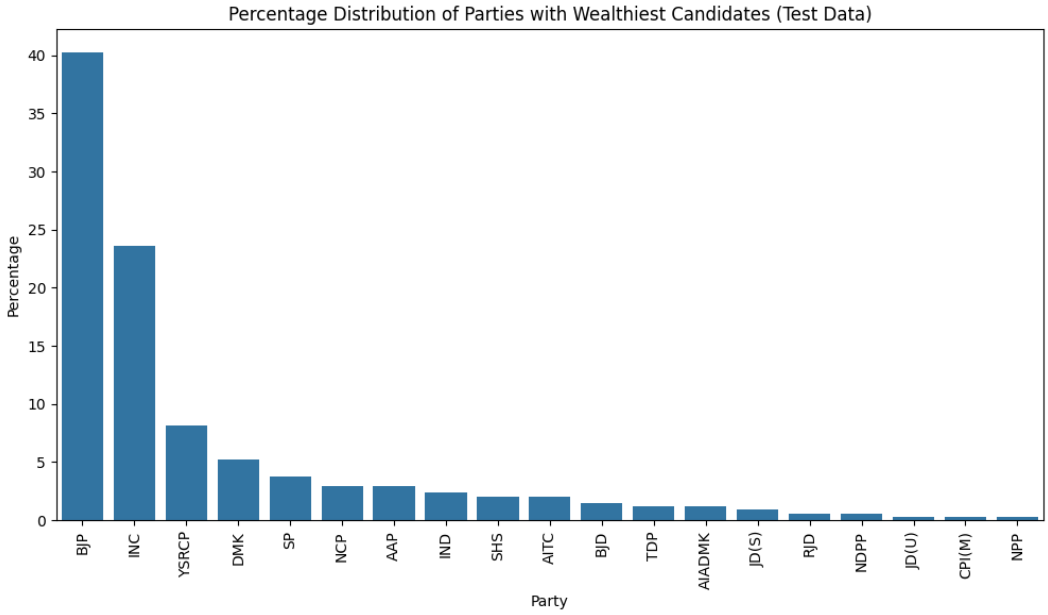


Figure 4: Test Data

2.3 Parties with differences in Candidates' Education level

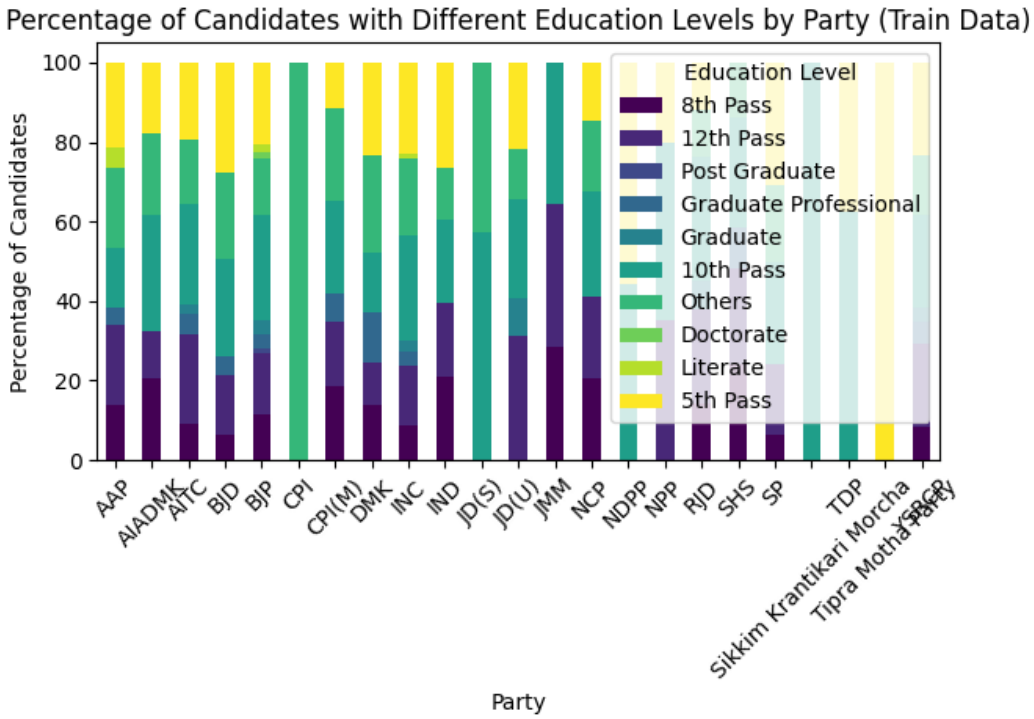
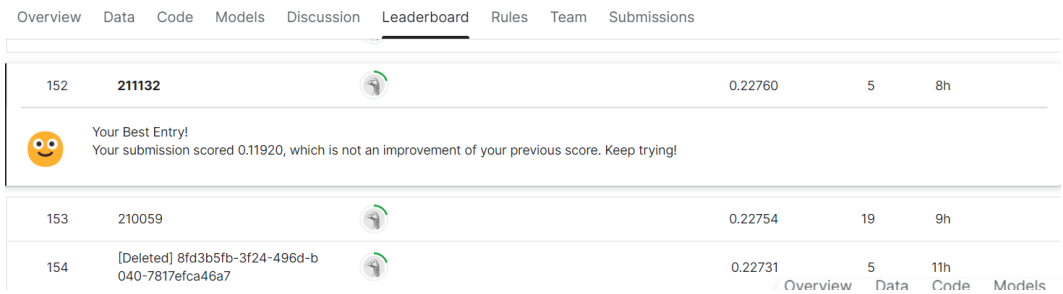


Figure 5: Train Data

Analysis of above 3 plots gives great insights about the elections and the party politics involved.

- From **Fig1** & **2**, we can see that parties like BJP and INC have greater number of candidates involved in criminal activities and how they could negatively affect the campaigning process in the country and also shows some atrocities which people might face in case these parties form the government.
- From **Fig3** & **4**, we could see that parties like BJP and INC have greater wealth or assets of the candidates posing answerability and scrutiny of their income sources as well as assets and liabilities owned. It also allows people to make informed decisions.
- From **Fig5**, we could see that how different parties have different set of candidates with wide spectrum of educational classification, even as low as just literate and thus emphasizes the need of education for all not just to make decisions but to understand various domains of administration,science,technology,sustainability and foreign affairs before making wise decisions affecting all.

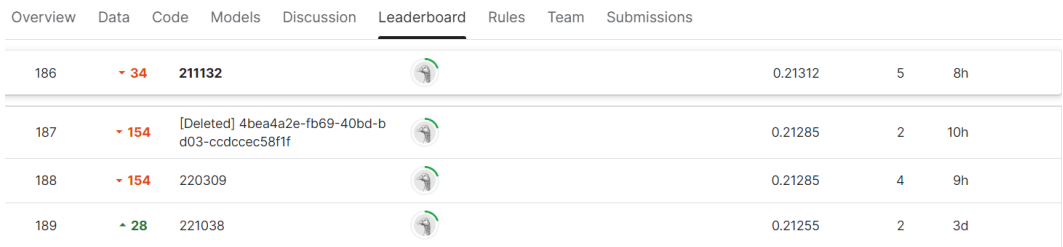
3 f-score and Results



Overview Data Code Models Discussion Leaderboard Rules Team Submissions

152	211132	0.22760	5	8h
	Your Best Entry! Your submission scored 0.11920, which is not an improvement of your previous score. Keep trying!			
153	210059	0.22754	19	9h
154	[Deleted] 8fd3b5fb-3f24-496d-b040-7817efca46a7	0.22731	5	11h

Figure 6: f-score on Public Leaderboard



Overview Data Code Models Discussion Leaderboard Rules Team Submissions

186	▾ 34	211132	0.21312	5	8h
187	▾ 154	[Deleted] 4bea4a2e-fb69-40bd-bd03-ccdccec58f1f	0.21285	2	10h
188	▾ 154	220309	0.21285	4	9h
189	▴ 28	221038	0.21255	2	3d

Figure 7: f-score on Private Leaderboard

4 References

Github - Classification

https://github.com/utpaldwivedi/CS253-assign/blob/main/Python/submit.py

Colab Notebook - Visualisation

https://colab.research.google.com/drive/1JQaCBhIF3Qp4jXc8PdAddYuJLtqES10p#scrollTo=cr_Ddq8UhtTP

Learning - ML

https://www.kaggle.com/learn/intro-to-machine-learning

AI - Improving Score

https://chat.openai.com/