Python Assignment Documentation

Introduction:

I have implemented a machine learning model to predict the Education levels of various candidates based on provided data of their party, criminal records, state etc.

Libraries used:

- Sklearn, Numpy and Pandas.
- Matplotlib

Data Pre-processing:

- Label encoding is performed on categorical features in 'X' and 'X_test' dataframes using LabelEncoder from scikit-learn. This converts categorical variables into numeric format.
- Standard scaling is applied to the 'Criminal Case' column of 'X' and 'X_test' dataframes using StandardScaler from scikit-learn. This ensures that all features have a mean of 0 and a standard deviation of 1.

Feature Engineering:

I have decided which features will be best suited to train the model and predict the education levels and with logical reasoning found out the three best identifiers as party, criminal case and state of the candidate.

Standardisation:

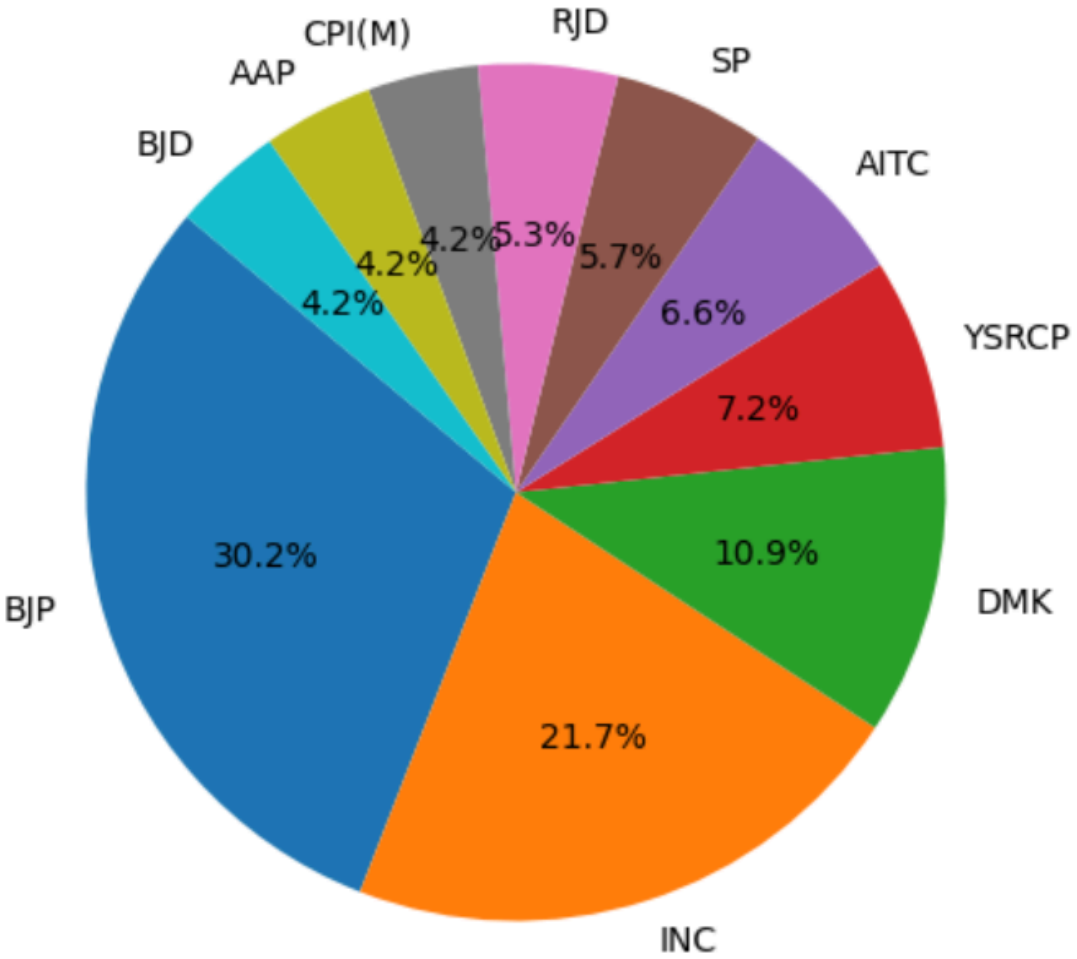
It is applied to the 'Criminal Case' column using StandardScaler to scale features to have a mean of 0 and a standard deviation of 1.

Model Implemented:

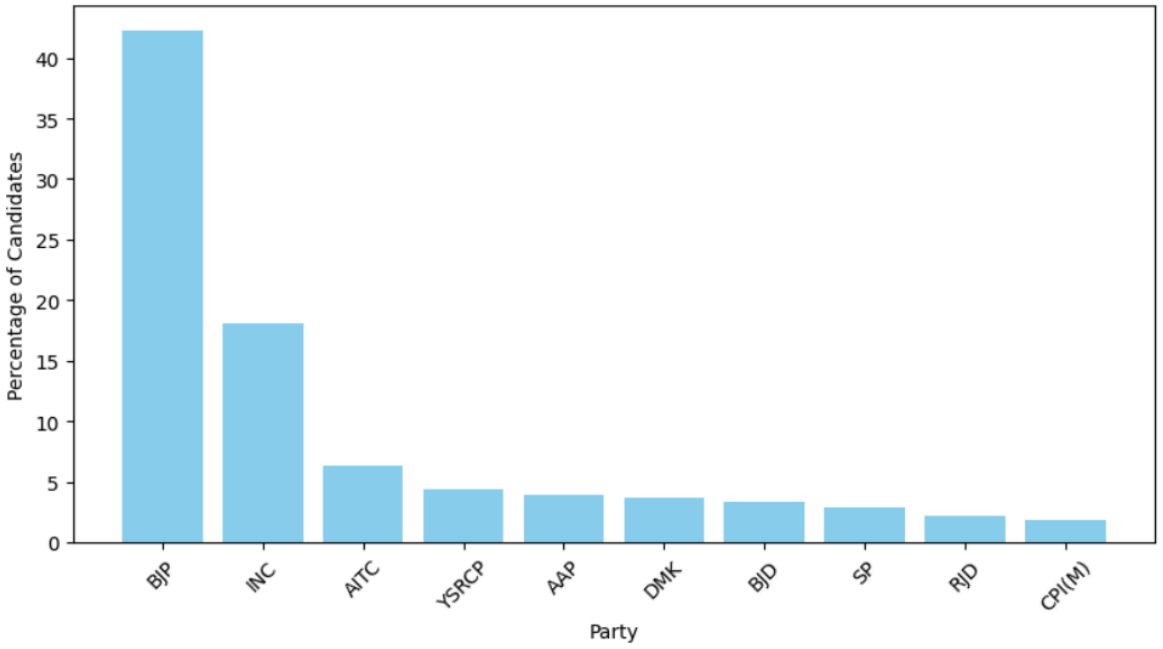
I have implemented RandomForestClassifier from sklearn.ensemble

Plots Obtained:

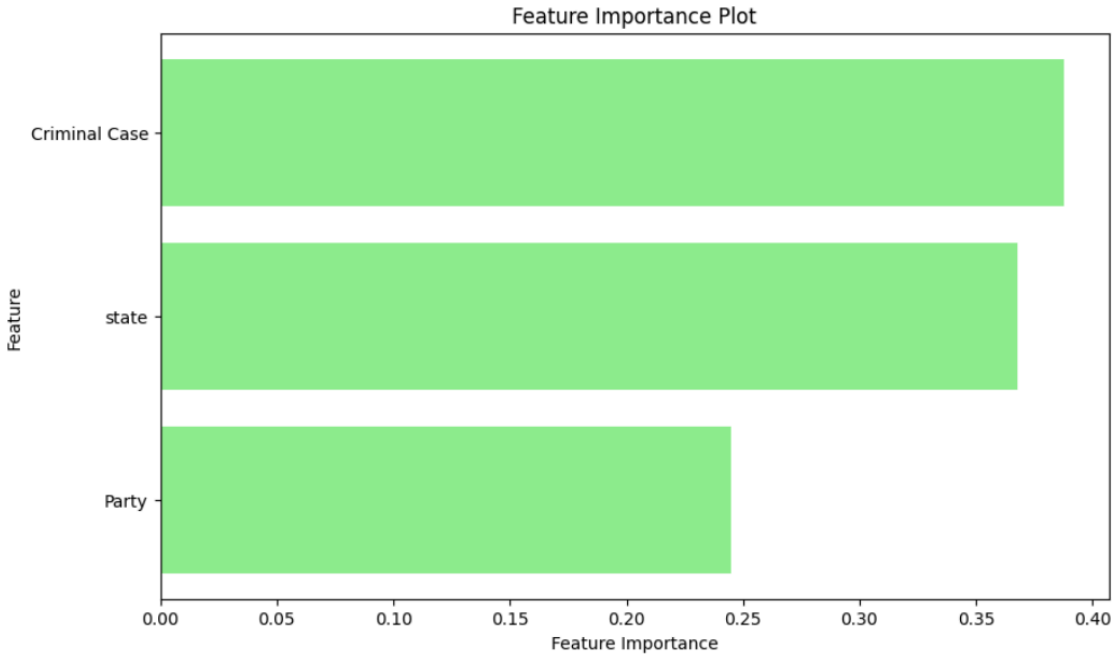
1) ***Parties vs Criminal Cases:*** The following bar graph shows the percentage distribution of the top 500 candidates based on the number of criminal cases among various parties. The top 10 parties in the list are shown.



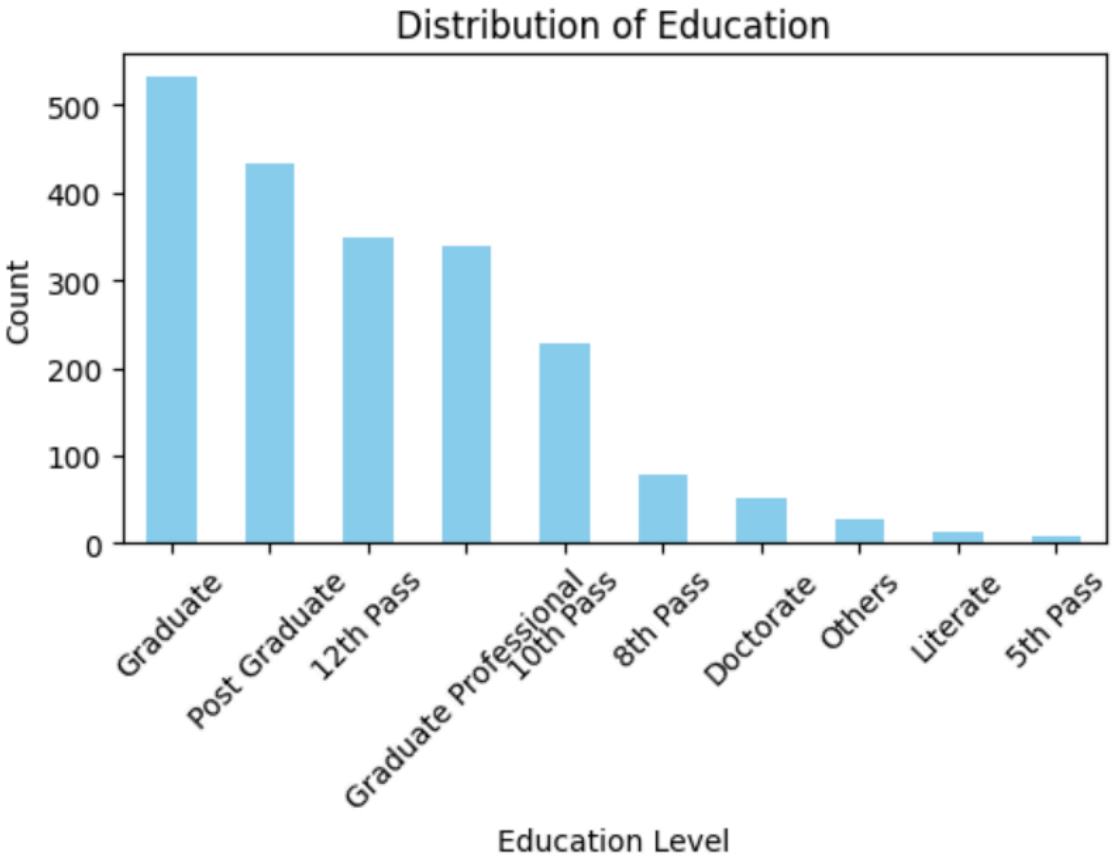
2) ***Parties vs Wealthy Candidates*****:** The following pie chart shows the percentage distribution of the top 1800 wealthy candidates among various parties. The threshold wealth is **20lac.** The top 10 parties according to the list are shown.



3) ***Feature Importance Plot:*** The following plot shows the importance of the features used for training the model. It shows that the parameter Criminal Case was the most significant one, followed by state and then party of the candidates.



4) ***Education Distribution Plot:*** The following plots show the distribution of education types amongst all candidates. It reveals that most candidates have completed graduation or post graduation.



F1 Score

My best F1 score was:
Public: 0.23460, Rank 118
Private: 0.22423, Rank 142

References

1) Tutorial provided in the discussions section of kaggle competition.
2) ChatGPT to improve performance and readability of code.