Assignment 3

Who's The Real Winner ?

Swayamsidh Pradhan

April 14, 2024

1 Github repo link

Click here!!

2 Preprocessing Steps

2.1 Data Description

The dataset consists of seven columns:

- **ID**: Unique identifier for each candidate.
- **Candidate Name**: Name of the candidate.
- **Constituency**: Constituency of the candidate.
- **State**: State in which the candidate is contesting.
- **Party**: Political party of the candidate.
- **Assets**: Non-numerical data representing candidate's assets.
- **Liabilities**: Non-numerical data representing candidate's liabilities.

2.2 Preprocessing Steps

To prepare the dataset for machine learning tasks, the following preprocessing steps were performed:

1. **Removing Irrelevant Columns**: Since the **ID**, **Candidate Name**, and **Constituency** columns contained almost unique values for each candidate and are not relevant for our analysis, they were dropped.
2. **One-Hot Encoding**: The **State** and **Party** columns were categorical variables. One-hot encoding was applied to convert them into numerical format suitable for machine learning algorithms.
3. **Converting Non-Numerical Data to Numerical**: The **Assets** and **Liabilities** columns contained non-numerical data. To utilize this information in our analysis, we converted them into numerical format using appropriate methods (e.g., extracting numerical values from strings).
4. **Scaling the Data**: Since the numerical features may have different scales, we applied StandardScaler to standardize the features and bring them to a similar scale.

3 Models Used

Table 1: Results for KNN models.

Family	Model	F1 Score	Runtime
KNN	KNN-17	0.237864	0.186352
KNN	KNN-18	0.233010	0.090526
KNN	KNN-25	0.228155	0.083889
KNN	KNN-12	0.228155	0.220816
KNN	KNN-27	0.223301	0.165576
KNN	KNN-26	0.223301	0.085786
KNN	KNN-24	0.223301	0.071045
KNN	KNN-15	0.221683	0.070096
KNN	KNN-16	0.221683	0.063123
KNN	KNN-28	0.221683	0.054734

Table 2: Results for Random Forest models.

Family	Model	F1 Score	Runtime
RF	RF-1000-100-10-10	0.270227	2.616872
RF	RF-1000-100-20-10	0.270227	2.173240
RF	RF-1000-100-20-10	0.270227	2.152584
RF	RF-1000-100-10-10	0.270227	2.166511
RF	RF-1000-100-30-10	0.263754	2.124736
RF	RF-1000-100-30-10	0.263754	4.993485
RF	RF-100-100-40-10	0.262136	0.213711
RF	RF-100-100-40-10	0.262136	0.344057
RF	RF-100-100-20-10	0.260518	0.220281
RF	RF-1000-100-40-10	0.260518	2.131024

[9]In the table, 'KNN-x' denotes K Nearest Neighbors with x neighbors.

[9]In the table, 'RF-a-b-c-d' denotes Random Forest with parameters: n_estimators=a, max_depth=b, min_samples_split=c, min_samples_leaf=d.

4 Data Analysis

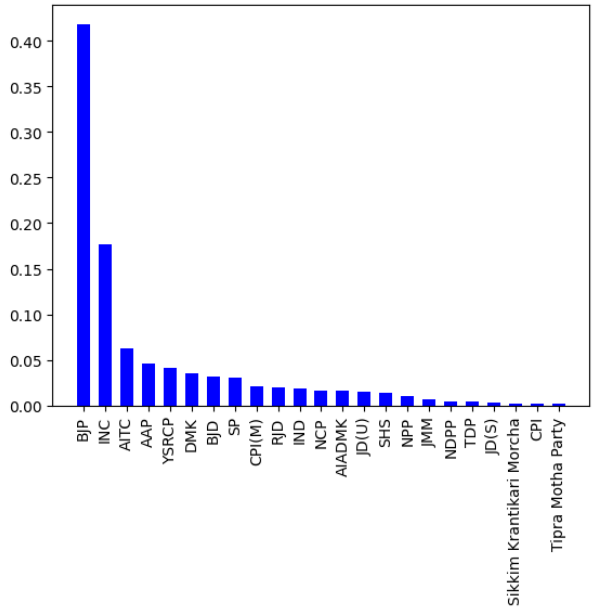

Figure 1: Percentage distribution of Candidates in Parties

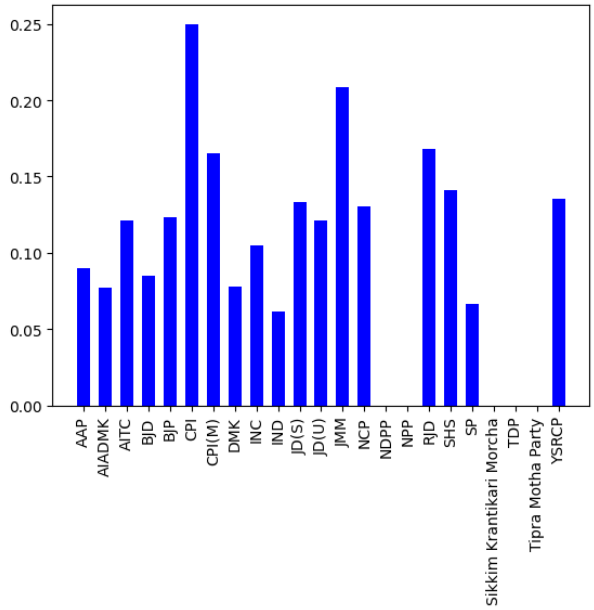

Figure 2: Percentage distribution of Criminal Candidates in Parties

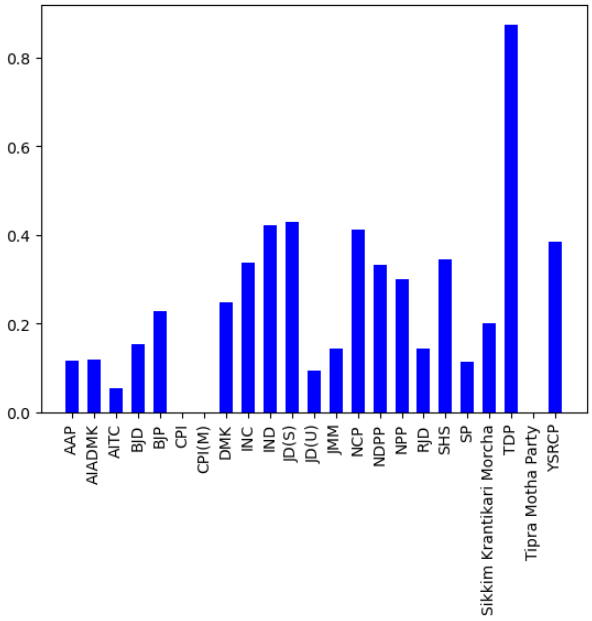

Figure 3: Percentage distribution of Wealthy Candidates in Parties

5 Results

	Rank	F1-score
Public	67	0.24881
Private	96	0.23572

References

[1] Aurélien Géron. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Sebastopol, CA.

[2] Predictive Modeling and Multiclass Classification. (n.d.). *Towards Data Science*. Retrieved from `https://towardsdatascience.com/predictive-modeling-and-multiclass-classification-a4d2c428a2eb`