Python Assignment Report

Ayush Meena 220268

April 13, 2024

Github code link : `https://github.com/AYushM25/CS253_Python_Assignment`

1 Methodology

1.1 Data Preprocessing Steps

1.1.1 Reading Data :

The code reads training and testing data from CSV files using 'pd.read_csv()' function.

1.1.2 Removing Unnecessary Columns :

Columns like 'ID', 'Candidate', and 'Constituency ∇' are dropped from both training and testing data.

1.1.3 Converting Numerical Values :

The function 'convert_to_lakhs()' is defined to convert numerical values to lakhs. It handles different formats like 'Crore+', 'Lac+', 'Thou+', and 'Hund+'. This function is applied to columns 'Total Assets' and 'Liabilities' in both training and testing data.

1.1.4 Label Encoding :

The target variable 'Education' is label encoded using LabelEncoder() from sklearn.preprocessing.

1.1.5 One-Hot Encoding :

Categorical variables 'Party' and 'state' are one-hot encoded using pd.get_dummies() function.

1.1.6 Splitting Data :

The data is split into training and testing sets using 'train_test_split()' from sklearn.model_selection.

1.2 Feature Engineering :

Not present

1.3 Identifying outliers :

Not present

1.4 Dimensionality reduction techniques :

Not used

1.5 Normalization, standardization, or transformation used :

Not used

2 Experiment Details

2.1 Data Insights

This plot is showing the percentage distribution of parties having candidates with criminal records > 5

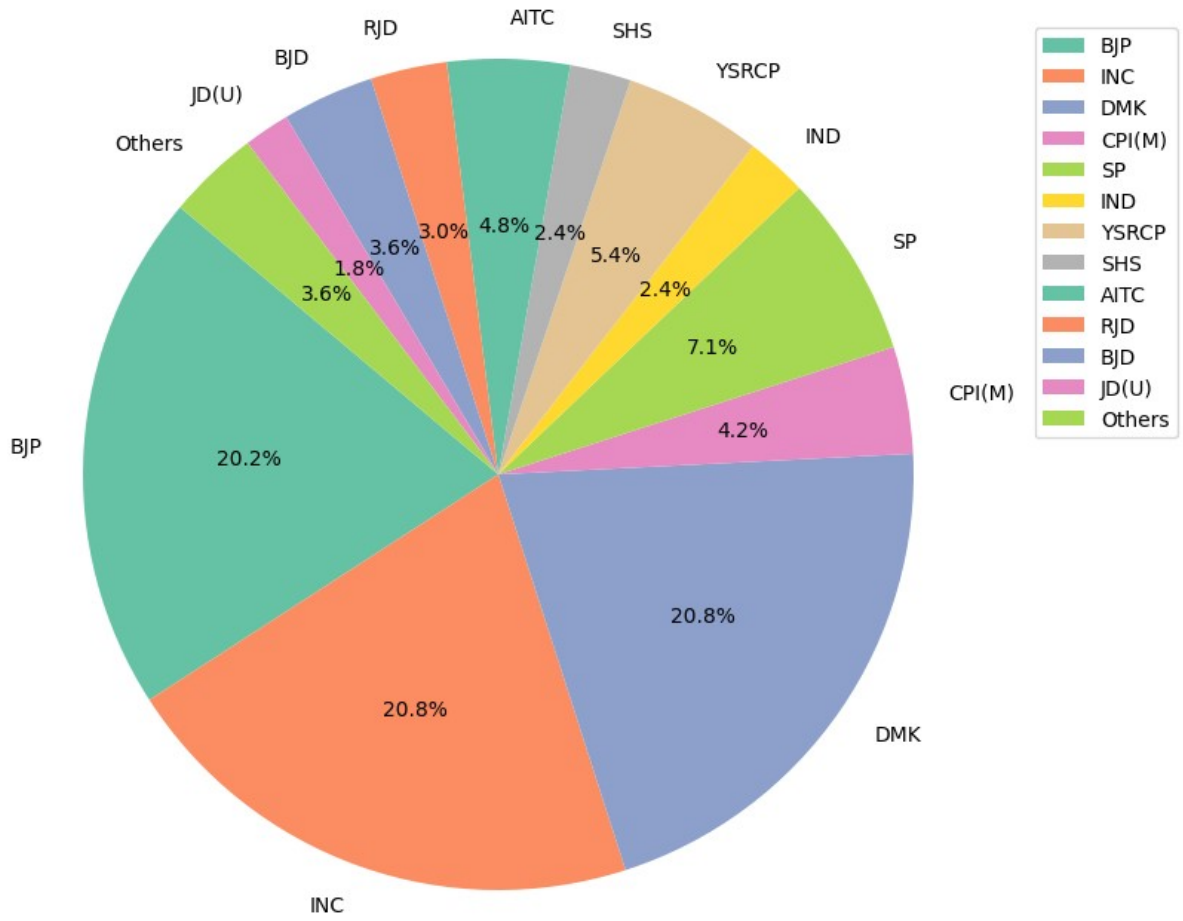

Figure 1: Parties having candidates with most criminal records

This plot is showing the percentage distribution of parties having candidates with total assets > median of total assets (11.5 Lac+)

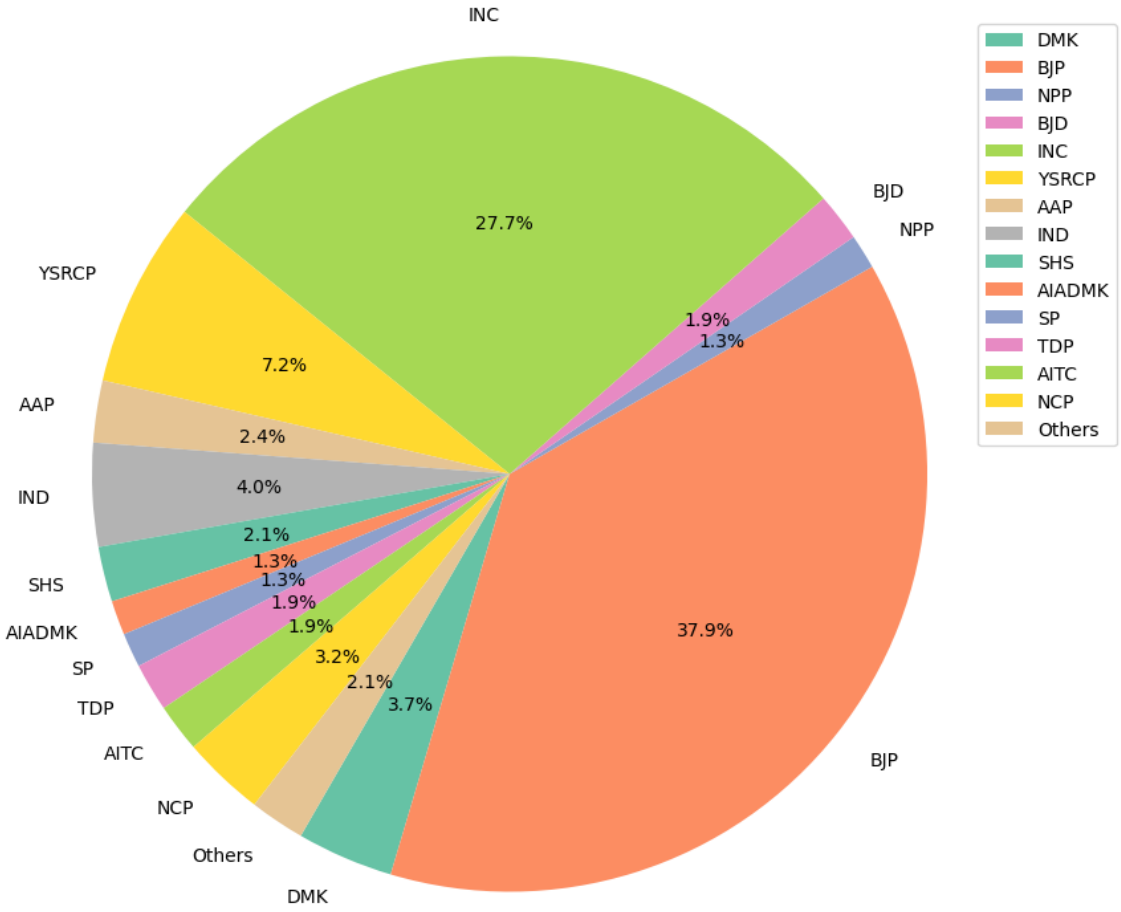

Figure 2: Percentage Distribution of Parties having wealthy candidates

2.2 Models Used

Model : Bernoulli Naive Bayes Classifier

Reason : Bernoulli Naive Bayes suits one-hot encoded data due to its assumption of independent binary features. It's effective when features lack correlation and represent categorical variables as binary indicators.

Hyperparameters	Details
Alpha	Used to handle zero probabilities in the computation of likelihoods.
Binarize	Useful when dealing with continuous-valued features that need to be converted into binary features
Fit Prior	A boolean parameter that determines whether to learn class prior probabilities or not.
Class Prior	This parameter specifies the prior probabilities of the classes.

3 Results

F1 score : 0.24420 (public), 0.25405 (private)

Leaderboard Rank : 54 (public), 56 (private)

4 References

- **Scikit-learn** : https://scikit-learn.org/stable/
- **Pandas** : https://pandas.pydata.org/docs/
- **Youtube** : https://www.youtube.com/channel/UCh9nVJoWXmFb7sLApWGcLPQ
  https://www.youtube.com/watch?v=RHydX4lBVsE&ab_channel=NANDINISHARMA
- **Geeksforgeeks** : https://www.geeksforgeeks.org/bernoulli-naive-bayes/
- **Kaggle** : https://www.kaggle.com/learn/intro-to-machine-learning
- **ChatGPT** :https://chat.openai.com/