CS253 – Assignment 3 Report

Chinmay Pillai

200298

Code Used (GitHub Link):

https://github.com/ChinmayPillai/CS253-Software_Development_and_Operations/blob/main/Assignment%203/Assign3.ipynb

Final F1 Score: 0.23

Feature Engineering:

A new feature called net_worth = Total Assets – Liabilities has been generated. This better represents the individual's financial capacity.

Features Used:

Since ID, Candidate Name and Constituency are unique features, they won't be useful in training and hence are dropped. Additionally, we don't expect to have any correlations between the party an individual belongs to and their education level, hence this feature is also dropped. The below plots also reflect the same. The feature net_worth is used instead of Total Assets and Liabilities since it better represents the individual's financial capacity.

The catagorical feature state and been one-hot encoded before being used for training.

The data provided also doesn't have any Nan or invalid data entries and hence data cleaning ins't necessary.

Numerical Features Used: net_worth, Criminal Cases

Catagorical Features Used: State

Target Variable:

Initially one-hot encoding was used on the catagorical target 'Education' as well. But, since different classes of education do lie on an scale, using Labe Encoding where the lower levels of education are given a lower integer value and the higher levels are given a higher value, better represents the target feature. The classes have been orders as - 'Others', 'Literate', '5th Pass', '8th Pass', '10th Pass', '12th Pass', 'Graduate', 'Graduate Professional', 'Post Graduate', 'Doctorate', where 'Others' maps to a label of 0, 'Literate' to 1 and so till 'Doctorate' to '9'.

Unique Education levels: ['8th Pass' '12th Pass' 'Post Graduate' 'Graduate Professional' 'Graduate'

'10th Pass' 'Others' 'Doctorate' 'Literate' '5th Pass']

Class Imbalance:

Graduate	531
Post Graduate	432
12th Pass	349
Graduate Professional	339
10th Pass	227
8th Pass	78
Doctorate	52
Others	28
Literate	14
5th Pass	9

Model Used:

After testing the data on DecisionTree, RandomForest, K-Nearest neighbours, Linear SVM etc.

The following the result on the initial test:

Logistic Regression:

F1-Score: 0.10

Accuracy: 0.23

K-Nearest Neighbors:

F1-Score: 0.17

Accuracy: 0.18

Decision Tree:

F1-Score: 0.17

Accuracy: 0.17

Random Forest:

F1-Score: 0.16

Accuracy: 0.17

Linear Support Vector Machine:

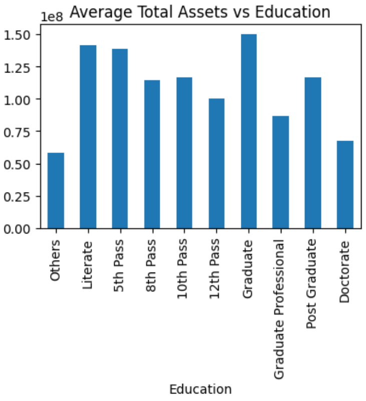
F1-Score: 0.06

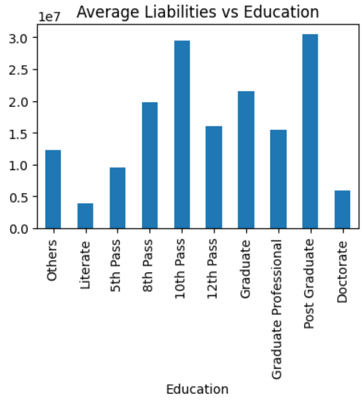
Accuracy: 0.12

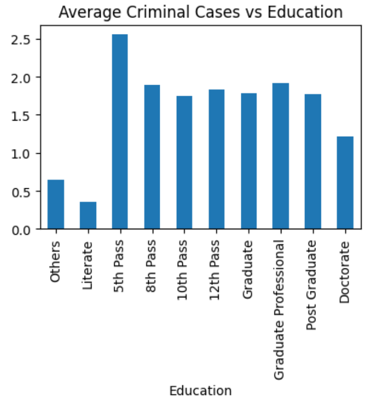
On further testing, Random Forest model was shown to give the best results and hence it was used to model the data.

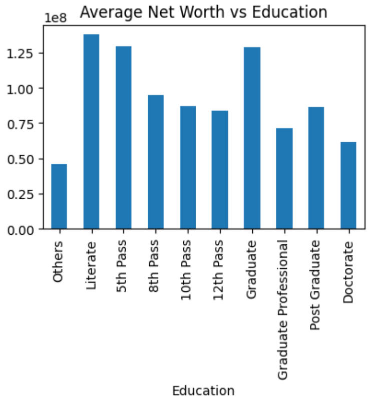
<u>Graphs used to obtain insight into the date:</u>

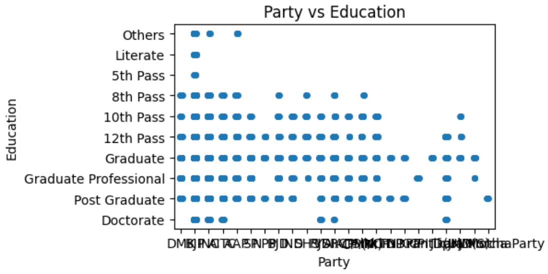
1. <u>Plots of each feature vs Education Level:</u>

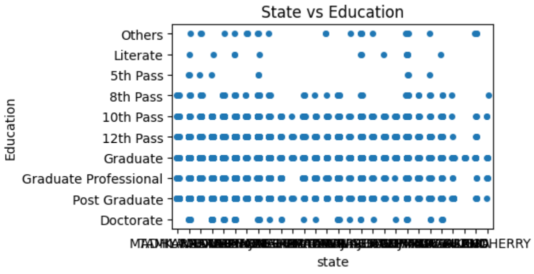




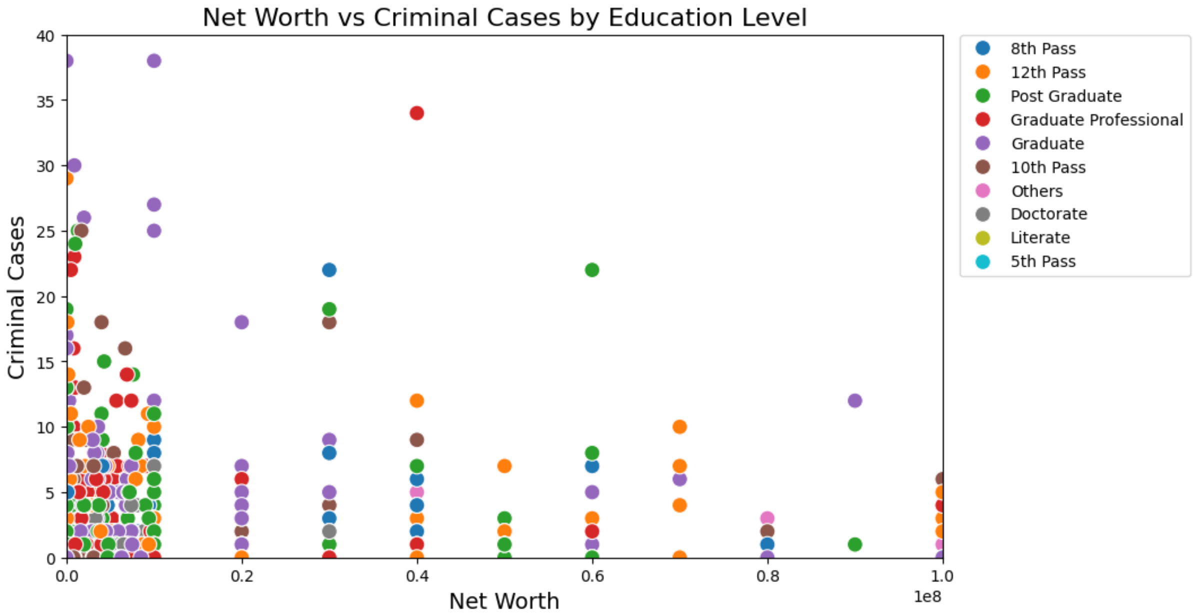








2. Scatter Plot of net_worth and Criminal Cases:



We do not observe any clear cluster formation for each Education class and hence clustering methods like Gaussian Mixture Models are not being considered to model the data.