Data Selection Proposal

i. Dataset

This project will try to model the recidivism of offenders released from jail based on multiple factors such as initial offense class and type, characteristics of the offender and release type. I will be using this dataset https://www.kaggle.com/slonnadube/recidivism-for-offenders-released-from-prison which has about 26000 records in the first file and 6718 records in the second file which only considers part of the factors.

ii. Methodology

a) Data Preprocessing

The data has been clearly labeled by multiple factors mentioned above, and each factor has different categories, so I do not have to do much data processing. I plan to use 50% of data as training set, 20% as validation set and the rest 30% as test set.

b) Machine Learning models

I will try implementing the Gradient Boost model and Random Forest model as the former is good for making predictions and the latter is good for classification of multifields. The method I am approaching will be similar to "Prediction algorithm for crime recidivism" (Julia, A. Luis, C. et Thomas, T. 2015)

c) Final conceptualization

I want to showcase my project through an academic poster, but there is the possibility of integrating it into an app used by judge to evaluate the "recidivism score" of offenders.

Reference

Julia, A. Luis, C. et Thomas, T. *Prediction algorithm for crime recidivism,* 2015, retrieved from http://cs229.stanford.edu/proj2015/250_report.pdf.