Methodology

Our data consists primarily of numeric variables. Our response variable, C_Rate (crime rate) and Inflation are continuous quantitative variables, whilst most of the others are discrete variables. NIA (Neighbourhood Improvement Areas) is a binary variable, taking on the value of 1 if a neighbourhood had been flagged by the city of Toronto, or 0 if it has not.

Our data was obtained as multiple different data sets. Thus in order to prepare it for further analysis it was tidied up (irrelevant columns were removed, missing values were treated, format was made uniform to facilitate consolidation into a single dataset, etc). Any missing values I encountered were found in datasets where the location of a facility or resource was unavailable. This was due to the fact that resources with a missing value for location were online resources, thus as a result I removed those resources completely from the individual datasets. This is because I'm looking at how the availability of different resources in different neighbourhoods and their individual demographics affect their crime rates. If a resource is equally accessible regardless of location it is essentially a constant and doesn't contribute to our analysis in any way.

We have classified our data into stratified groups (neighbourhoods), our goal will be to see what traits or factors these groups contain that contribute to their crime rate (a trait). Thus, we are looking to assess the association of the traits within a pre classified group, and are not necessarily looking to compare groups to each other. We will entertain two scenarios:

1. Analyzing feature values and crime rate levels measured in the same year
2. Analyzing feature values taking three years prior to their corresponding neighbourhood crime rate level.

We will compare the various models we fit onto both situations to see which one is most useful in classifying a neighbourhoods crime rate level based on sociodemographic information and resources available.

The algorithms and models we will use in our analysis are:

1. Learning Vector Quantization Model
2. Decision Trees
3. Permutation Feature Importance
4. Multinomial Logistic Regression

As our dataset is small it is vital that our model has low complexity to prevent overfitting. Thus we will aim to reduce dimensionality as much as possible and ensure that our models use few parameters. We will also use ensemble methods where we use multiple classifiers to help fit our data to help counter this issue as well. Further, we will try to counter imbalanced class levels in our dataset by using the synthetic minority oversampling technique (SMOTE) to increase the size of our dataset and reduce the number of minority class levels. We will run our models on both our original data and our data after implementing SMOTE and compare model efficacy.