CIND 820 Big Data Analytics Project

Crime Prediction Using Classification Approaches

Student Name: Mariela Marmanillo Mendoza

Student Number: 501188993

Supervisor: Tamer Abdou

Date of Submission: February 22, 2024

Ryerson University

Table of Contents

Abstract

Crime in Toronto has been relatively low compared to other larger cities in North America; however, there is a sensation that incidents are increasing more and more in Toronto, leading to an unsafe feeling among the population. Every day and every week, the news shows incidents of robberies, break-and-enters, and assaults in public places and private places where, before, no one ever knew about that type of event. Even when, in most cases, only material losses are experienced (for example, a car theft or theft of artifacts and jewelry in break-ins), fear and insecurity are growing among the population.

The intention of this study is to analyze crime occurrences in Toronto from 2014 to 2023, which are included in the database provided by the Toronto Police Service, where incidents have been reported since 2014 by time, date, and location – the nearest road intersection node to protect the privacy of parties involved in the occurrence. The Major Crime Indicators (MCI) in the database include Assault, Break and Enter, Auto Theft, Robbery, and Theft Over, excluding sexual violations, homicides, and shootings. According to the description of the database, the data collected has been determined through a police investigation as founded, meaning that the offence did occur or was attempted to occur. The database has 372,899 observations and was last updated on January 11, 2024 (Toronto Police Service, 2024).

In this project, the research questions are the following:

1. What is the comparative assessment of accuracy among various machine learning classification algorithms, including decision tree, Naïve Bayes and K-nearest Neighbour models, when analyzing crime incidents in Toronto?

2. Which information of a crime occurrence allows for more accurate predictions: the temporal patterns—time of the day, day of the week, or month-- or the spatial patterns--location, premises, or neighborhood?
3. How the results obtained from this research contribute to understand historical crime patterns and develop a plan with better-targeted interventions to improve community safety.

By answering these questions, we could help decision-makers understand historical crime patterns and develop a plan with better-targeted interventions to improve community safety.

I propose developing a predictive model that accurately classifies MCIs based on historical crime occurrence data in the dataset described above. Supervised machine learning method, such as the decision tree algorithm for classification tasks, is a suitable tool in this database labeled with categorical MCI features. Its performance will be compared with Naïve Bayes and K-nearest Neighbours models. In this case, the models will be trained and tested on a dataset that contains the desired categorization (The Chang School of Continuing Education, Class Notes, 2023).

Literature Review, Data Description, and Approach

Literature Review

Several studies have focused on the analysis of crime trends as it has become a significant issue affecting almost all countries around the world. All research in this area pretend to provide useful insights for city planners and law enforcement agencies in developing effective crime prevention plans.

Crime prediction can be reached using different methods, including statistical, visualization, unsupervised learning, and supervised learning methods. Machine learning algorithms are widely used in crime prediction (Saeed and Abdulmohsin, 2023). According to Safat et al. (2021), crime prediction involves assessing the accuracy of past reported crimes, while forecasting involves predicting future crime trends.

Numerous research projects have been conducted to predict crime types, crime rates, and crime hotspots using datasets from different areas in major cities around the world (Safat et al., 2021). Machine learning techniques have proven effective in forecasting spatial crime data (Saeed and Abdulmohsin, 2023). For this project, various research studies that explore the relationship between crime occurrences and different machine learning techniques were reviewed. One of them was carried out for Toronto using the same database as this project, while others focused on cities in the United States, and only one in India.

Using the Major Crime Indicators (MCI) database provided by the Toronto Police Service, Esfahani and Esfahani (2023) studied crime rate trends in Toronto from 2014 to 2022. They considered the temporal scales such as year, month, week, day and hour to predict the number of crime incidents per month, and forecast them for 2023 and 2024 using a deep learning method (neural network model). This method was compared with naïve and weighted moving average models. In order to evaluate the models' performance and accuracy in predicting crime occurrences, three metrics--Mean Square Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE)—were used. The results showed that deep learning model outperformed the other two, concluding that crime events have increased in Toronto from 2014 onwards, with this trend expected to continue throughout 2024.

The Naïve model is a simple method that forecast time series by assuming that the future observations will be the same than the last value and does not contemplate further changes of features in the data or futures adjustments, while the weighted moving average model considers relevant observations and assign weights to them based on certain characteristics to improve forecasting accuracy (Esfahani and Esfahani, 2023).

Almanie et al. (2015) analyzed crime datasets for the cities of Denver, Colorado and Los Angeles, California, to predict crime types on specific dates (month, date and time) and at a particular location in a data-mining model. The study combined the Denver dataset with the neighborhood demographics dataset existed for that city. Data transformation, discretization, as well as an apriori algorithm were applied in order to identify recurring patterns and improve model accuracy, while Naïve Bayesian and decision tree classification methods were used for crime-type prediction. These supervised learning algorithms were built using Scikit-Learn library tool, applying a 5-fold cross validation strategy on both models to compare accuracy. The Naïve Bayes model shown the best performance in crime prediction, with accuracies of 51% and 54% for Denver and Los Angeles respectively, while the decision tree classifier showed accuracies of only 42% and 43%.

Vaquero Barnadas (2016) used k-nearest neighbours (KNN), Parzen windows, and artificial neural networks on the San Francisco, California crime dataset to determine which of them works best solving the category classification problem. The dataset contains 39 different crime types or categories. Due to the huge amount of data, the k-means algorithm was applied for clustering and reducing the size of the database. The study focused on classification and prediction, with the neural network algorithm proving to be the most accurate.

Safat et al. (2021) compared different machine learning methods for crime prediction using the large crime datasets for Los Angeles, California and Chicago, Illinois. The methods include logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbors (KNN), Decision Tree, multilayer perceptrom (MLP), Random Forest and XBoost algorithms for prediction accuracy. Also, a deep learning model named long-short term memory (LSTM) was applied for time-series analysis, and the autoregressive integrated moving average (ARIMA) model for forecasting crime rate and crime density areas. The LSTM evaluation was reported using the performance metrics Root Mean Squared Error (RMSE) and MAE. The datasets contained 35 and 39 types of crime, with the results showing that XGBoost outperformed other models with accuracies of 94% and 88% in Chicago and Los Angeles, respectively, followed by KNN with 88% and 89%.

Hussain and Aljuboori (2022) explored crime report dataset for Boston, Massachusetts to predict incident crime types using decision tree, logistic regression and Naïve Bayes classification models. The results showed that the decision tree classifier performed better than the other machine learning techniques, concluding that crime location alone is enough to build an accurate model, as crime amount and type are strongly related to location.

According to Saeed and Abdulmohsin (2023), supervised learning approaches are predominantly used in crime prediction studies, with logistic regression being the most robust method. In their study, logistic regression algorithm was used to predict the correlation between burglar crimes and various factors, including time of day, day of week, barriers, connectors, and repeated victimization. However, this model was ineffective when applied to large geographical areas.

Kumar et al. (2020) applied the k-nearest neighbours (KNN) model to predict crime in the city of Indore, Madhya Pradesh, India. The research goal was to predict the type of crime likely to occur in a particular area based on the date, time and location. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used as reference metrics to compare the results with those of a previous study that determined a different optimal value for k.

The examined studies aim is focused on comparative study between supervised learning algorithms using classification techniques of machine learning to predict the crime type in different cities with unique features. All researchers' intention is to use the information for criminal investigations and crime prevention strategies. Table 1 summarizes the classifying method used by those research projects and the method that showed the best performance. For this project, we are using the three more commonly used classification algorithms: Decision Tree, Naïve Bayes and KNN.

Table 1. Summary of classification techniques used by previous studies

Study	City	Methods	Best Method
Almanie et al. (2015)	Denver, CO & Los Angeles, CA (USA)	Naïve Bayesian and Decision Tree	Naïve Bayes
Vaquero Barnadas (2016)	San Francisco, CA (USA)	K-nearest Neighbours (KNN), Parzen Windows, and Artificial Neural Networks	Neural Network
Kumar et al. (2020)	Indore, MP (India)	K-nearest Neighbours (KNN)	KNN
Safat et al. (2021)	Los Angeles, CA & Chicago, IL (USA)	Logistic Regression, Support Vector Machine, Naïve Bayes, K-nearest Neighbors (KNN), Decision Tree, Multilayer Perceptrom, Random Forest and XBoost Algorithms	KNN followed by XBoost
Hussain and Aljuboori (2022)	Boston, MA (USA)	Decision Tree, Logistic Regression and Naïve Bayes	Decision Tree
Esfahani and Esfahani (2023)	Toronto, ON (Canada)	Neural Network (Deep Learning), Naïve Bayes and Weighted Moving Average	Neural Network

All the studies reviewed address cases similar to the one that will be carried out in this project, including attributes related to the temporality of the crime such as time, day, month, year, and the location of the crime scene, which includes latitude and longitude. However, the categories or labels assigned for the types of crimes are not the same as the categories in our dataset, with the exception of the Esfahani and Esfahani (2023) study, which uses the same Major Crime Indicators (MCI) database as in this project. This discrepancy arises because the police department of each city assigns its categories depending on the characteristics of the impact zone. Furthermore, Esfahani and Esfahani (2023) combine deep learning methods with machine learning methods, whereas in this project, only use machine learning classification algorithms are suggested to be used. Therefore, it can be stated that we are not replicating any study exactly.

Data Description and Exploratory Data Analysis (EDA)

The dataset used in this study includes all Major Crime Indicators (MCIs) retrieved from the Toronto Police Service public safety data portal. It is published in accordance with the Municipal Freedom of Information and Protection of Privacy Act, which means that the Toronto Police Service has taken care of the privacy of every person involved in the reported crime occurrences; hence, no personal information is provided nor the exact location of the crime scene is revealed for any reason (Toronto Police Service, 2024). Furthermore, before downloading the database, full knowledge of the Open Government License for Ontario was taken, which literally indicates that the user is free to "Copy, modify, publish, translate, adapt, distribute or otherwise use the Information in any medium, mode or format for any legal purpose" (Ontario, 2013).

The dataset is available here: https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about. It is updated on a regular basis. For this study, it was downloaded on

January 22, 2024, reflecting the last upload to the database as of January 11, 2024. The selected dataset contains 372,899 observations since 2014. According to the description, the collected data has been determined through a police investigation as founded, which means that the offence did occur or was attempted to occur.

The MCIs include Assault, Break and Enter, Auto Theft, Robbery, and Theft Over, while excluding sexual violations, homicides, and shootings. Appendix 1 shows the indicators glossary. The information is collected at the crime and/or victim level; therefore, several observations may have the same identification number EVENT_UNIQUE_ID associated with the different MCIs used to categorize the incident (Toronto Police Service, 2024). Appendix 2 shows the dictionary of attributes.

The exploratory data analysis (EDA) report for the selected dataset was prepared using the ydata-profiling library in Python. The following link https://github.com/marielamarmanillo/CIND820 allows access to the repository for this project where all the files including Python code are stored. The following graph (Figure 1) shows an overview of the dataset statistics, noteworthy indicating a small number of missing values, representing less than 0.1% of the total values, which suggests the dataset's quality. Additionally, there are 31 attributes in this dataset, categorized into four types: numeric, text, date-time, and categorical.

Figure 2 shows key considerations regarding the studied data. First, the feature OBJECTID is highlighted for its unique values and uniform distribution. This is expected, as the OBJECTID attribute serves as a simple enumeration of dataset observations, starting from 1. This variable is omitted from the Toronto Police Service column description, so it will be dropped during data cleaning and preprocessing.

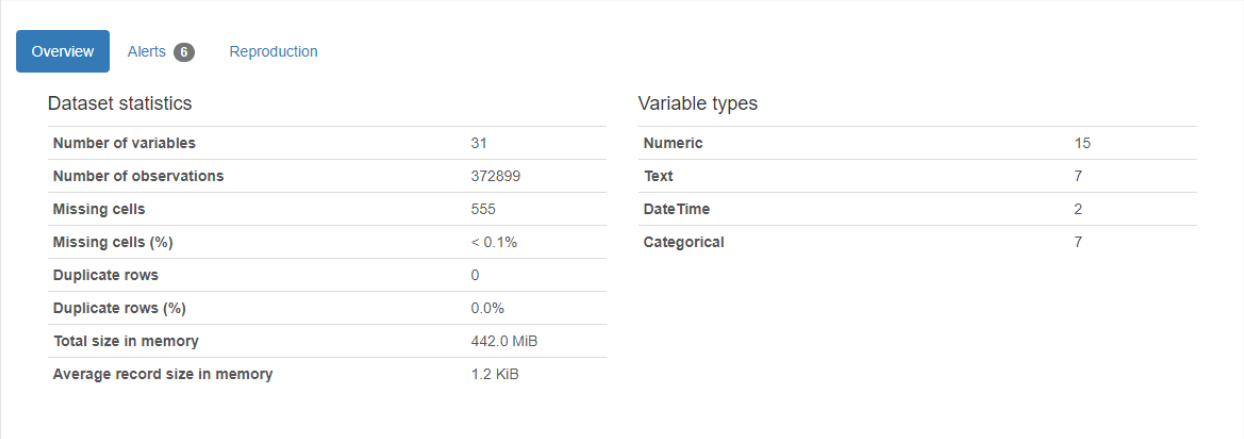

Figure 1. Dataset Overview from *ydata-profiling* Report

The REPORT_HOUR and OCC_HOUR variables include some percentage of zeros due to the reporting protocol of the Toronto Police Service. In their 24-hour time format, crimes reported or occurred within the first hour of the day are denoted as 0, corresponding to the time range from 12:00 am to 12:59 am.

The longitude (LONG_WGS84) and latitude (LAT_WGS84) variables, which both together indicate the coordinates of the nearest intersection where the crime occurred, show a small percentage of zero values. Specifically, there are 5,750 zero values in each column. As these instances involve both longitude and latitude being assigned zero at the same time, these rows will be dropped. Given the relatively small number of observations, omitting these rows does not significantly impact the overall analysis of the dataset.

Figure 3 exhibits the overview of the categorical variable of interest, Major Crime Indicators. It has 5 categories or labels as described earlier in this section. No missing values were found. The events categorized as Assault constituting more than half of the total crime occurrences (53.1%). Following Assault, Break and Enter is the next significant category, representing 18.8% of the

total. Auto Theft and Robbery are the next two notable categories, making up 15.7% and 9.1% of the total. Finally, Theft Over is the least frequent category, accounting for only 3.3%.

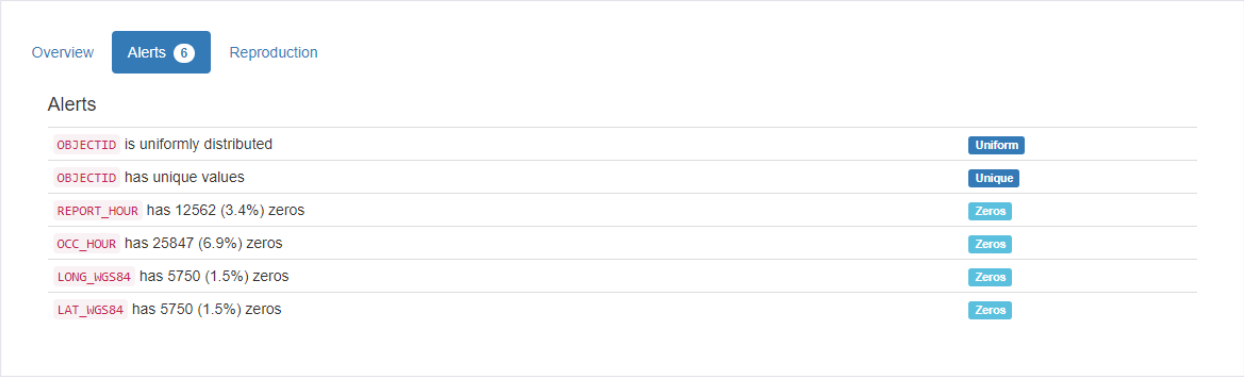

Figure 2. Dataset Alerts from *ydata-profiling* Report

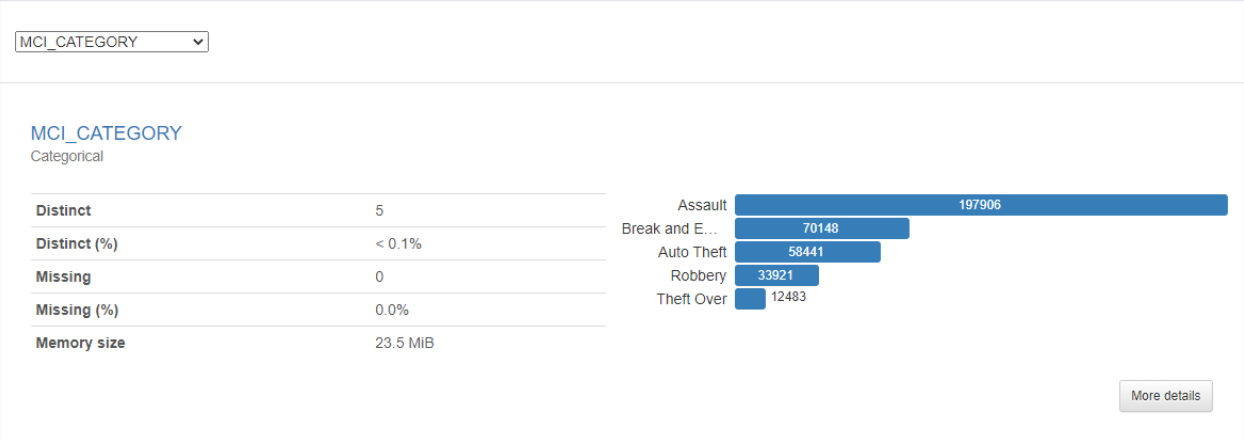

Figure 3. MCI Categorical Variable Overview from *ydata-profiling* Report

Figure 4. Year of Crime Occurrence Overview from *ydata-profiling* Report

The variable containing the year in which the crime occurs ranges from 2000 to 2023. However, based on the histogram shown in the Figure 4, 99.5% of the data falls within the years 2014 to 2023, while only 0.4% falls between 2000 and 2013. Additionally, there is no data for 111 records, representing 0.1% of the total data in this attribute. Therefore, records for years with insignificant data will be omitted from the analysis.

```
DataFrame info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 372899 entries, 0 to 372898
Data columns (total 31 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   X                  372899 non-null  float64
 1   Y                  372899 non-null  float64
 2   OBJECTID           372899 non-null  int64
 3   EVENT_UNIQUE_ID    372899 non-null  object
 4   REPORT_DATE        372899 non-null  object
 5   OCC_DATE           372899 non-null  object
 6   REPORT_YEAR        372899 non-null  int64
 7   REPORT_MONTH       372899 non-null  object
 8   REPORT_DAY         372899 non-null  int64
 9   REPORT_DOY         372899 non-null  int64
 10  REPORT_DOW         372899 non-null  object
 11  REPORT_HOUR        372899 non-null  int64
 12  OCC_YEAR           372788 non-null  float64
 13  OCC_MONTH          372788 non-null  object
 14  OCC_DAY            372788 non-null  float64
 15  OCC_DOY            372788 non-null  float64
 16  OCC_DOW            372788 non-null  object
 17  OCC_HOUR           372899 non-null  int64
 18  DIVISION           372899 non-null  object
 19  LOCATION_TYPE      372899 non-null  object
 20  PREMISES_TYPE      372899 non-null  object
 21  UCR_CODE           372899 non-null  int64
 22  UCR_EXT            372899 non-null  int64
 23  OFFENCE            372899 non-null  object
 24  MCI_CATEGORY       372899 non-null  object
 25  HOOD_158           372899 non-null  object
 26  NEIGHBOURHOOD_158  372899 non-null  object
 27  HOOD_140           372899 non-null  object
 28  NEIGHBOURHOOD_140  372899 non-null  object
 29  LONG_WGS84         372899 non-null  float64
 30  LAT_WGS84          372899 non-null  float64
dtypes: float64(7), int64(8), object(16)
```

Figure 5. Dataset Structure

Project Approach

Figure 6 represents the workflow and overall methodology of this project.

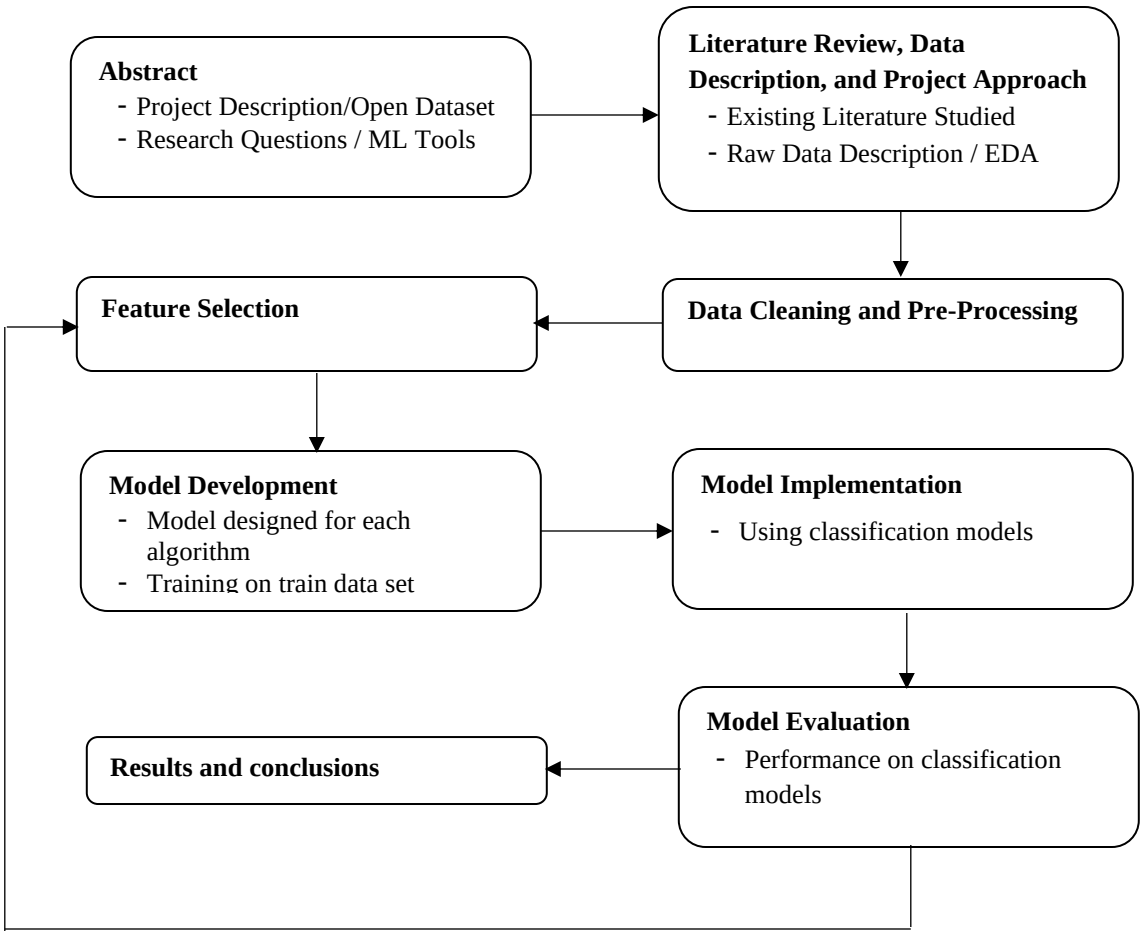

Figure 6. Project Methodology Workflow

Data Cleaning and Pre-Processing

Cleaning the selected dataset involved eliminating duplicate rows, dropping some variables, omitting missing values, and filtering the observations that were recorded on years other than the period 2014 to 2023. Thus, the study only focused on analyzing data from January 1, 2014 to December 31, 2023, spanning 10 years of crime data.

- Removing Duplicate Rows

  According to the dataset description provided by the Toronto Police Services, some crime events were reported by different individuals involved in the incident, such as the victim, the

victimizer, and witnesses. As a result, there were multiple reports made by different people about the same event. To address this, only one observation of each event was retained. A total of 47,920 duplicate rows were identified and removed based on the offence identifier (EVENT_UNIQUE_ID). After removing these duplicates, the dataset was left with 324,979 rows and 31 columns.

- Dropping Variables

The columns X and Y were excluded from the analysis due to the absence of descriptions or information regarding their relevance. They appear to contain internal information managed by the Toronto Police Service. The removal of the variable OBJECTID was addressed earlier in section Data Description.

Additionally, variables corresponding to report date, report year, report month, report day, report day of the year, report day of the week, report hour, as well as the division where the crime was reported were also removed. This decision was made because this project's main focus is on the times, dates and location of the crime occurrence rather than the times and dates of when the incidents were reported.

The crime occurrence date variable (OCC_DATE) was also omitted from the analysis because it embraces information contained in other columns corresponding to occurrence year, occurrence month, etc. Similarly, the column related to location type (LOCATION_TYPE) was also removed as the variable PREMISE_TYPE is a categorical one and provides more informative details about the type of location compared to LOCATION_TYPE, which offers overly broad information about the crime scene.

Moreover, the columns related to the Uniform Crime Reporting (UCR) system code and extension were excluded from the analysis. This information is primarily collected by police agencies to generate historical records of crime and traffic for Statistics Canada's surveys. Additionally, the column OFFENCE was removed because it duplicates information already provided in the MCI categories column. Despite some offences may include more detailed characteristics such as the weapon used, this redundancy was believed unnecessary.

Furthermore, the old identifiers and names of neighborhood structures in Toronto (HOOD_140, NEIGHBOURHOOD_140) were also dropped. This decision was made because the information from the new structure represented in columns HOOD_158 and NEIGHBOURHOOD_158 was kept for the analysis of this project. In total, 18 attributes out of the 31 presented in the raw dataset were removed based on the explanations provided above. The new data subset has 324,979 rows and 13 columns.

- Omitting Missing Values

After dropping some variables to clean the dataset, 88 missing values were identified in each of the columns corresponding to the year, month, day, day of the year, and day of the week of occurrence. Interestingly, the same row is the one with data missing for each of the above-mentioned columns. Given that only 88 rows out of the 324,979 total contain missing values, it was decided to omit those rows at this stage; therefore, the new data subset has 324,891 rows and 13 columns.

- Dropping Rows with Crimes that Occurred in years other than the period 2014-2023

Since this study covers the analysis of crime events from January 1, 2014 to December 31, 2023, the crime data observations with years of occurrence from 2000 to 2013 were removed from the dataset. In this case, 1,171 were eliminated and the cleaned dataset shows 323,720 rows and 13 columns.

After the described arrangements, all relevant attributes to be used in this study do not include missing values. Furthermore, object variables such as month of crime occurrence, day of the week, premises type, and MCI category were converted to category data type. Similarly, year, day of occurrence, and day of the year underwent conversion from float to integer data type. The new data types are shown in Figure 7.

```
EVENT_UNIQUE_ID        object
OCC_YEAR                int64
OCC_MONTH            category
OCC_DAY                 int64
OCC_DOY                 int64
OCC_DOW              category
OCC_HOUR                int64
PREMISES_TYPE        category
MCI_CATEGORY         category
HOOD_158               object
NEIGHBOURHOOD_158      object
LONG_WGS84            float64
LAT_WGS84             float64
```

Figure 7. Variables Data Type

Preliminar Analysis

After cleaning the data, this section presents an analysis of crime trends and patterns. It aims to identify key information such as the most common types of crimes committed in Toronto, trends in crime rates over time, frequently occurring crimes, and crime hotspots based on location, among other factors.

In Toronto, the average number of crime incidents in the last 10 years was 32,372 per year, 2,698 per month, and 89 per day. Table 2 shows the count of different types of crimes (Assault, Auto Theft, Break and Enter, Robbery, and Theft Over) for each year from 2014 to 2023, along with the total sum of all crimes for each year. It is noteworthy that overall, crime rates in Toronto have shown a sustained upward trend, except for the years during the pandemic (2019 to 2021), where there was a significant decrease. Subsequently, there was a resumed increasing and sharp trend in crime rates, as shown by the trend line in Figure 8.

However, this has not been a trend observed in all types of crimes in this study. Figure 9 shows the trend described above, which is very pronounced in assaults and break-ins; however, car thefts have increased significantly even during the pandemic. These three cases deserve a lot of attention as they are having the most impact. Regarding robbery, cases decreased during the pandemic, but post-pandemic they have not increased to pre-pandemic levels. Finally, theft over cases have remained constant.

Table 2. Crime Counts by Category and Year (2014-2023)

MCI_CATEGORY	Assault	Auto Theft	Break and Enter	Robbery	Theft Over	Total
OCC_YEAR						
2014	14080	3462	7101	2911	999	28553
2015	14942	3068	6817	2763	1016	28606
2016	15793	3029	6262	2918	1012	29014
2017	16225	3324	6780	3109	1158	30596
2018	16727	4268	7503	2965	1257	32720
2019	17645	4782	8412	2848	1324	35011
2020	15473	5113	6816	2133	1167	30702
2021	16138	5900	5545	1727	1043	30353
2022	18043	8714	5868	2125	1374	36124
2023	20018	10912	7217	2335	1559	42041

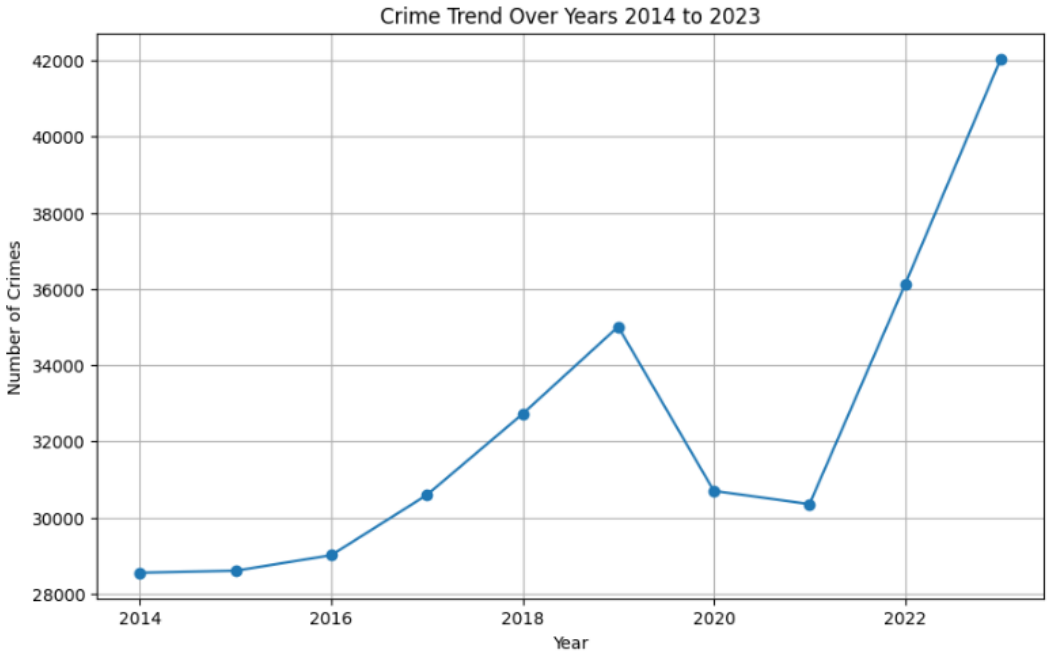

Figure 8. Crime Trend in Toronto (2014 to 2023)

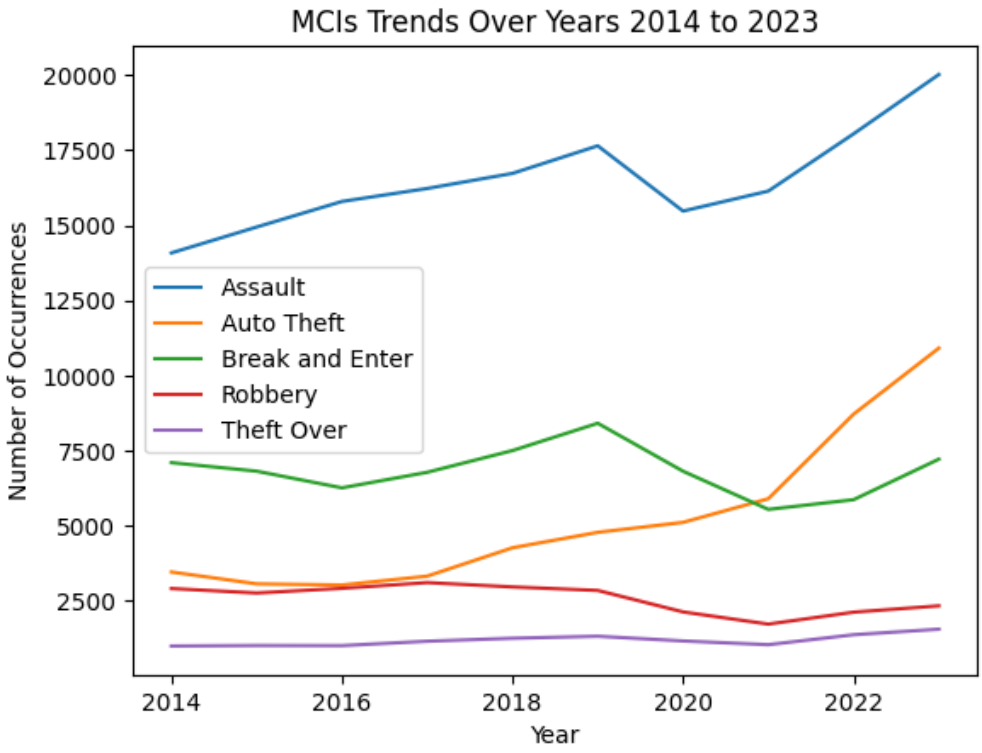

Figure 9. Crime Trends per Major Crime Indicators (2014 to 2023)

Crime occurs most frequently in late winter, late spring, and early summer, and it begins again during the holiday season corresponding to Halloween, Christmas, and New Year's Eve (Figure 10). The first few days of any month are considered to have the greatest number of incidents, and the last few days are considered to have the fewest. It should be noted that the observed peak occurs on the first day of each month (Figure 11). Observing by day, the crime peak is observed at midnight. It is then low during the day, and another peak occurs at noon. Subsequently, it begins to gradually increase from 4 p.m. onwards until night (Figure 12). Additionally, more crimes occur on Sundays than on any other day of the week (Figure 13).

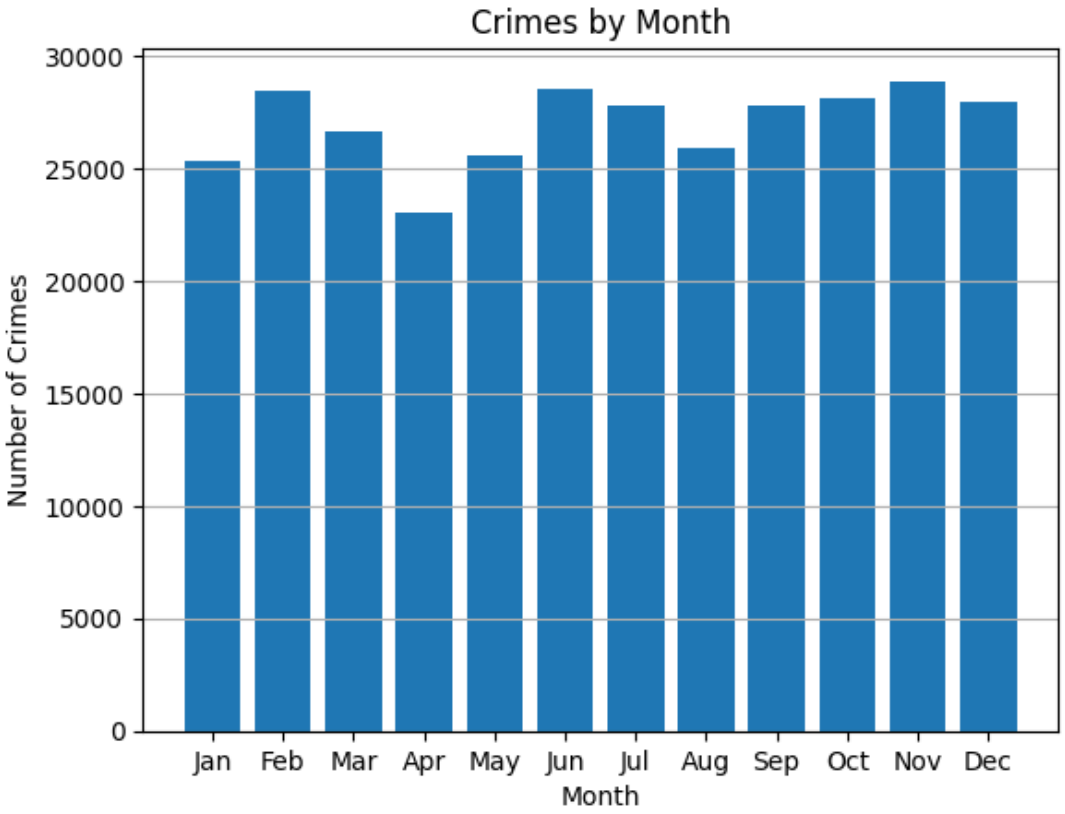

Figure 10. Crimes by Month

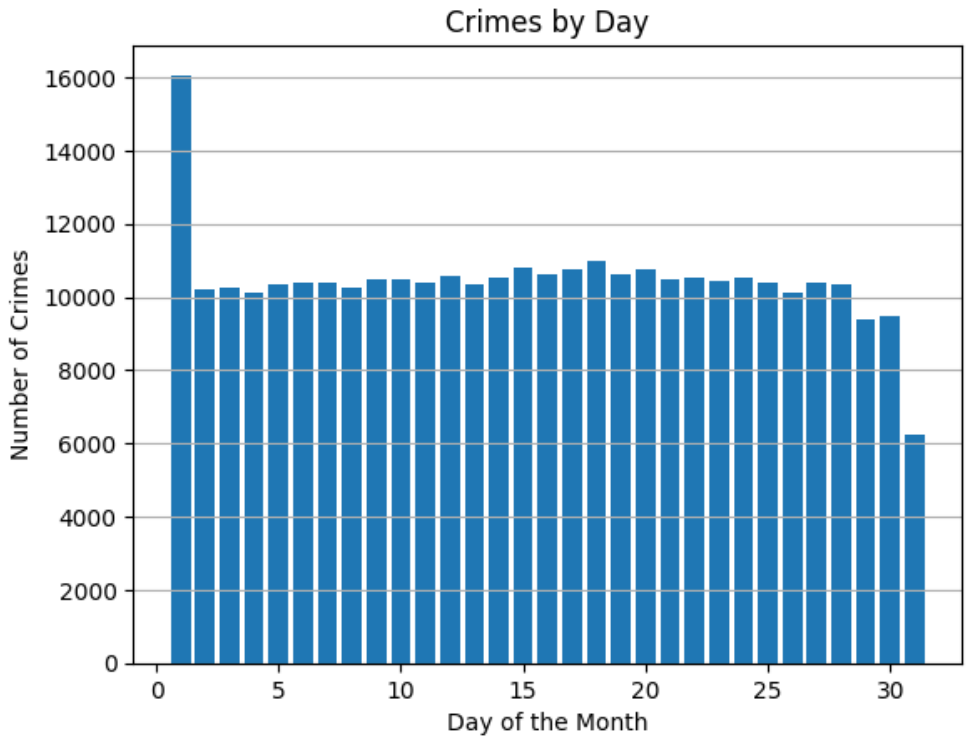

Figure 11. Crimes by Day

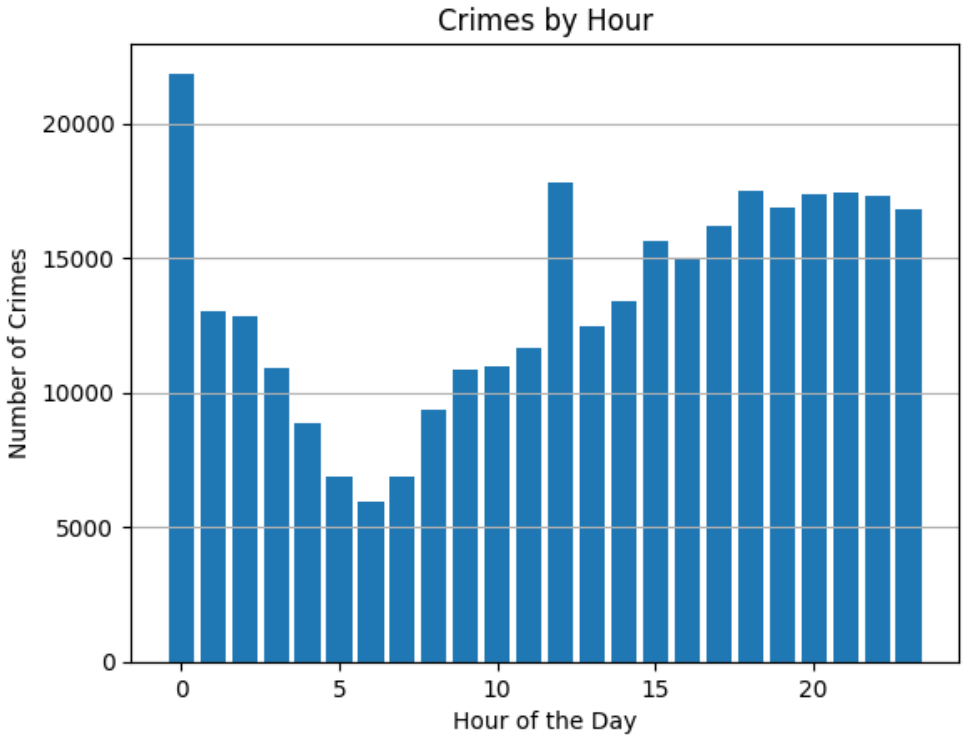

Figure 12. Crimes by Hour

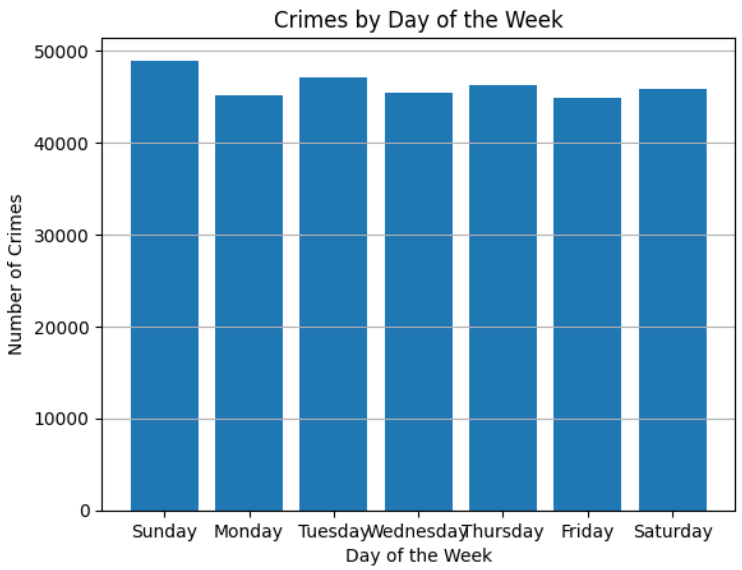

Figure 13. Crimes by Day of the Week

Over the past 10 years, most crimes occurred outdoors, followed by apartments, commercial establishments and homes. To a lesser extent, they occur in transportation facilities and educational centers.

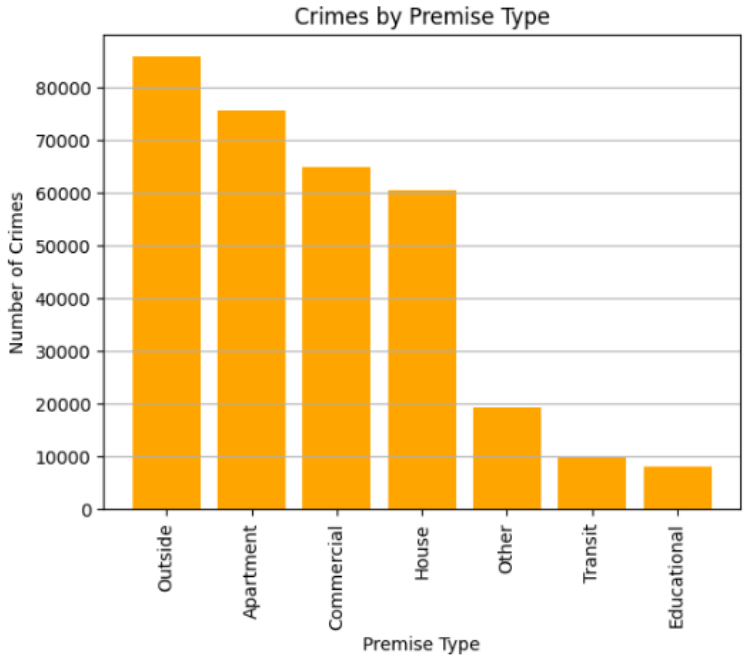

Figure 14. Crimes by Type of Premises

According to Toronto's new neighborhood structure, the most dangerous neighborhood is West Humber-Clairville, followed by Moss Park and Downtown Yonge East, listed as the other two most dangerous (Figure 15). While among the least dangerous neighborhoods are Lambton Baby Point, Woodbine-Lumsden and Guildwood (Figure 16).

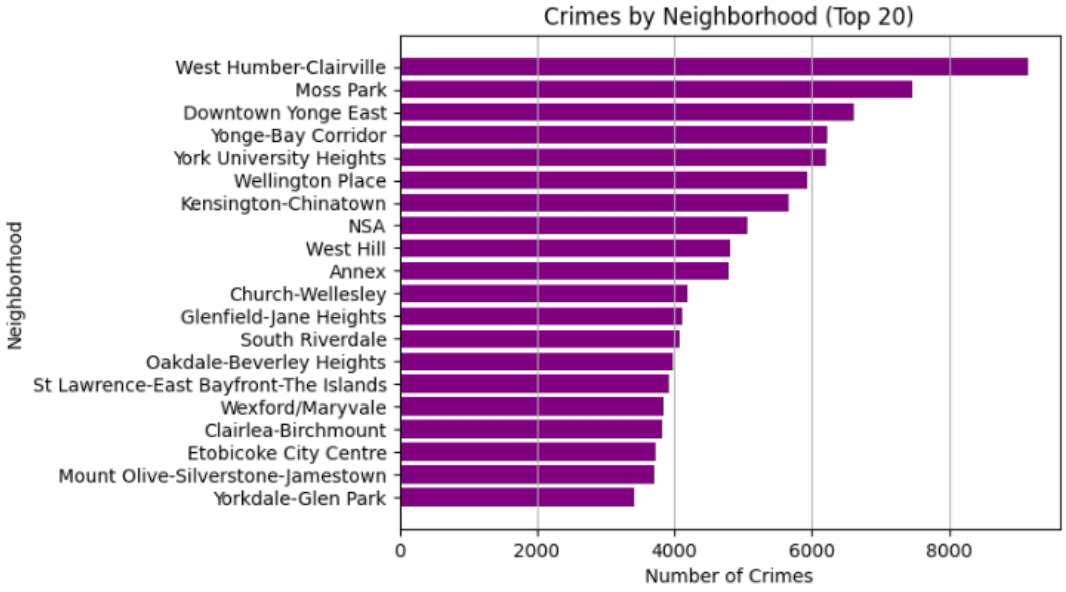

Figure 15. Most Dangerous Neighborhoods

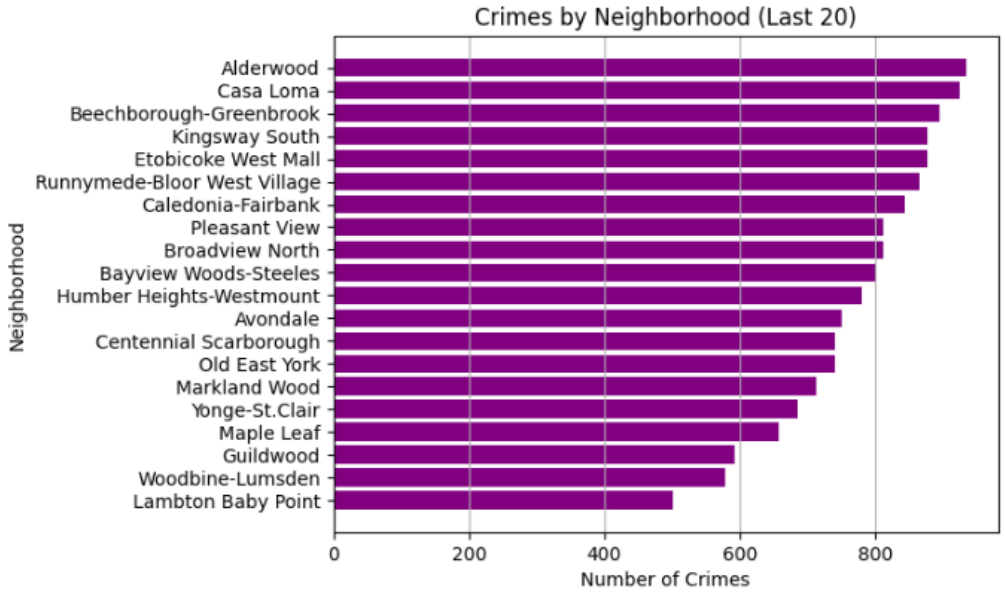

Figure 16. Least Dangerous Neighborhood

Feature Selection

The goal of feature selection is a data preprocessing technique for selecting the best subset of variables prior to building a machine learning model. It helps to remove irrelevant variables to optimize model construction (Gupta, 2023).

There are a wide variety of feature selection techniques, including filter-based, wrapper-based, and hybrid/embedded families. In this project, specific filter-based feature selection techniques, such as information gain, are employed. Filter-based techniques evaluate features independently of the model that will be built and are less prone to overfitting. This characteristic fits the needs of this study, as three different classification models will be used. Moreover, these techniques are faster and less computationally expensive, although they do not necessarily capture correlations or interactions among features (Gupta, 2023).

Common variants of filter-based feature selection techniques include information gain (capturing the information gain of each variable with respect to the target variable), chi-square test (evaluating the best chi-square scores), and Fisher's score (assessing the best Fisher's scores), among others. In this project, the information gain feature selection method is utilized. This method was chosen due to its effectiveness in ranking variables according to their information gain with respect to the outcome, aligning with the project's goals.

The dataset is split into 70% training set and 30% test set before applying feature selection techniques, which is applied only on the training set. This approach helps prevent data leakage, ensures accurate model evaluation, and improves computational efficiency. According to Table 3, The features with higher information gain values are: type of premises (19.4%), longitude (16.8%), and latitude (16.7%), which are more informative or influential for predicting the types of major

crime indicators. However, the variables for day, day of the week, day of the year, and month of the crime occurrence provide 0% of information gain.

Table 3. Feature Selection Results

```
Feature: PREMISES_TYPE, Information Gain: 0.19377085369393576
Feature: LONG_WGS84, Information Gain: 0.16833888733371882
Feature: LAT_WGS84, Information Gain: 0.16748923099888247
Feature: EVENT_UNIQUE_ID, Information Gain: 0.07707029018597567
Feature: HOOD_158, Information Gain: 0.0577919978068957
Feature: NEIGHBOURHOOD_158, Information Gain: 0.0575189186478382
Feature: OCC_HOUR, Information Gain: 0.03032288906964764
Feature: OCC_YEAR, Information Gain: 0.015416627516639192
Feature: OCC_DAY, Information Gain: 0.0029653741734669836
Feature: OCC_DOW, Information Gain: 0.002338467822204038
Feature: OCC_DOY, Information Gain: 0.0023066494171612995
Feature: OCC_MONTH, Information Gain: 0.0021381320730617936
```

In this case, we will retain the features that were manually selected previously, as the filter-based feature selection technique does not yield conclusive or divergent results.

Prior to model building, the class imbalance issue should be addressed in the training set. Table 4 shows that the training dataset has class imbalances within the MCI category. The data is heavily skewed to the assault class with more than 50% of the observations pertaining to this class within MCI.

Table 4. Class imbalance

```
Class Distribution:
Assault            115506
Break and Enter     47852
Auto Theft          36807
Robbery             18076
Theft Over           8363
```

Brownlee (2021) states that the synthetic minority oversampling technique (Smote) effectively addresses class imbalance on classification datasets. Smote consists of oversampling the minority class without adding any new information to the model but synthetizing new examples from the

existing ones in the dataset. Therefore, given the class imbalance in our dataset, the SMOTE technique is utilized in this project to get class-balanced training set and prevent our machine learning models from performing poorly (Table 5).

Table 5. Class Balance

New Class Distribution	
Assault	115506
Auto Theft	115506
Break and Enter	115506
Robbery	115506
Theft Over	115506

Model Implementation

Since this is a classification problem, Decision Tree, K-nearest Neighbours, and Naïve Bayes algorithms will be used.

Decision Tree Algorithm

This classification algorithm was attempted to be implemented on our training set; however, the output is showing some mistakes at certain points, and the confusion matrix doesn't make much sense. It will be thoroughly reviewed shortly.

Appendix

Appendix 1. Major Crime Indicators (MCI) Glossary

MCI	Definition
Assault	The direct or indirect application of force to another person, or the attempt or threat to apply force to another person, without that person's consent.
Auto Theft	The act of taking another person's vehicle (not including attempts).
Break and Enter	The act of entering a place with the intent to commit an indictable offence therein.
Robbery	The act of taking property from another person or business by the use of force or intimidation in the presence of the victim.
Theft Over	The act of stealing property in excess of $5,000 (auto theft is excluded).

Source: Toronto Police Service (2024)

Appendix 2. Attributes Dictionary of MCIs Dataset[1]

Field	Attribute Name	Description
1	EVENT_UNIQUE_ID	Offence Number
2	REPORT_DATE	Date Offence was Reported (time is displayed in UTC format when downloaded as a CSV)
3	OCC_DATE	Date Offence Occurred (time is displayed in UTC format when downloaded as a CSV)
4	REPORT_YEAR	Year Offence was Reported
5	REPORT_MONTH	Month Offence was Reported

[1] Dataset description retrieved from Toronto Police Service (2024) did not include attributes X, Y and OBJECTID.

6	REPORT_DAY	Day of the Month Offence was Reported
7	REPORT_DOY	Day of the Year Offence was Reported
8	REPORT_DOW	Day of the Week Offence was Reported
9	REPORT_HOUR	Hour Offence was Reported
10	OCC_YEAR	Year Offence Occurred
11	OCC_MONTH	Month Offence Occurred
12	OCC_DAY	Day of the Month Offence Occurred
13	OCC_DOY	Day of the Year Offence Occurred
14	OCC_DOW	Day of the Week Offence Occurred
15	OCC_HOUR	Hour Offence Occurred
16	DIVISION	Police Division where Offence Occurred
17	LOCATION_TYPE	Location Type of Offence
18	PREMISES_TYPE	Premises Type of Offence
19	UCR_CODE	UCR Code for Offence
20	UCR_EXT	UCR Extension for Offence
21	OFFENCE	Title of Offence
22	MCI_CATEGORY	MCI Category of Occurrence
23	HOOD_158	Identifier of Neighbourhood using City of Toronto's new 158 neighbourhood structure
24	NEIGHBOURHOOD_158	Name of Neighbourhood using City of Toronto's new 158 neighbourhood structure
25	HOOD_140	Identifier of Neighbourhood using City of Toronto's old 140 neighbourhood structure
26	NEIGHBOURHOOD_140	Name of Neighbourhood using City of Toronto's old 140 neighbourhood structure
27	LONG_WGS84	Longitude Coordinates (Offset to nearest intersection)
28	LAT_WGS84	Latitude Coordinates (Offset to nearest intersection)

Retrieved from: Toronto Police Service (2024)

References

Almanie, T., Mirza, R., & Lor, E. (2015). Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots. *International Journal of Data Mining & Knowledge Management Process*, 5(4), 1-20. DOI:10.5121/ijdkp.2015.5401

Brownlee, J. (2021, March 16). *Smote for imbalanced classification with python. MachineLearningMastery.com.* https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Esfahani, H. N. and Esfahani, Z. N. (2023). *Exploring crime rate trends and forecasting future patterns in Toronto city using police mci data and deep learning*. https://doi.org/10.21203/rs.3.rs-3806294/v1

Gupta, A. (2023, December 21). *Feature selection techniques in Machine Learning (updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

Hussain, F. S., & Aljuboori, A. F. (2022). A Crime Data Analysis of Prediction Based on Classification Approaches. *Baghdad Science Journal*, 19(5), 1073. https://doi.org/10.21123/bsj.2022.6310

Kumar, A., Verma, A., Shinde, G., Sukhdeve, Y. & Lal, N. (2020). *Crime Prediction Using K-Nearest Neighboring Algorithm*, 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 1-4, 10.1109/ic-ETITE47903.2020.155.

Ontario (2013). *Open Government Licence - Ontario*. (Published 2013, June 18. Updated 2023, May 10). https://www.ontario.ca/page/open-government-licence-ontario

Saeed, R. & Abdulmohsin, H. (2023). A study on predicting crime rates through machine learning and data mining using text, *Intelligent Systems*, 32(1), 20220223. https://doi.org/10.1515/jisys-2022-0223

Safat, W., Asghar, S., & Gillani, S.A. (2021). Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *IEEE Access*, 9, 70080-70094.

The Chang School of Continuing Education (Fall Term, 2023). *Module 10.1 – Classification vs Clustering*. Module 10. Statistical Learning Methods. CMTH642 - Data Analytics: Advanced Methods.

Toronto Police Service (2024, January 22). *Major Crime Indicators Open Data*. Toronto Police Service - Public Safety Data Portal. Retrieved January 22, 2024, from https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about

Vaquero Barnadas, M. (2016, November 14). *Machine learning applied to crime prediction*. Handle Proxy. http://hdl.handle.net/2117/96580