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, 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 et al., 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 et al., 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 et al. (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.

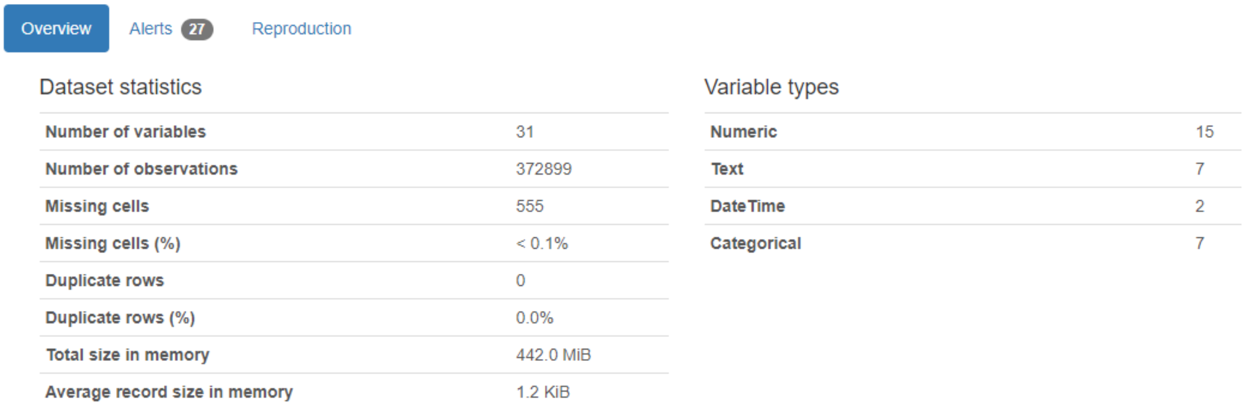
Data Description and Exploratory Data Analysis

The dataset used in this study includes all Major Crime Indicators (MCI) 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 used to create the profiling report. The following EDA graph shows an overview of the data statistics, 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 more than 30 attributes in this dataset, categorized into four types: numeric, text, date-time, and categorical. For this project, only attributes related to the crime scene location, time of day, day of the week, and month of occurrence will be utilized.



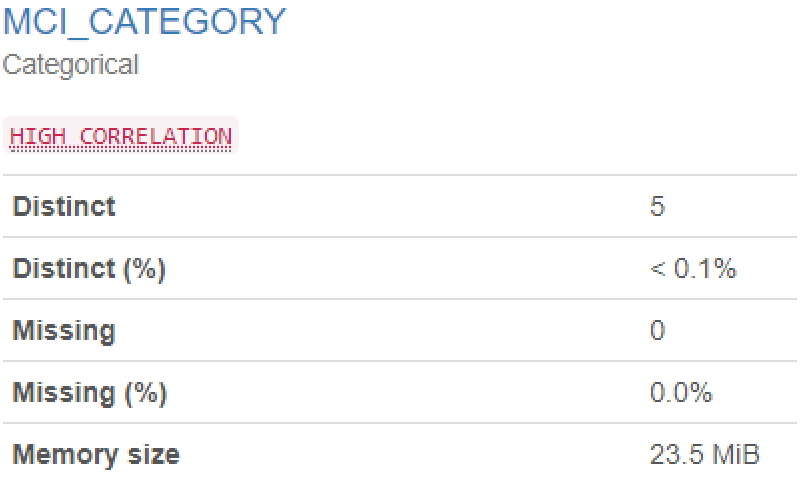
Dataset statistics

Number of variables	31
Number of observations	372899
Missing cells	555
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	442.0 MiB
Average record size in memory	1.2 KiB

Variable types

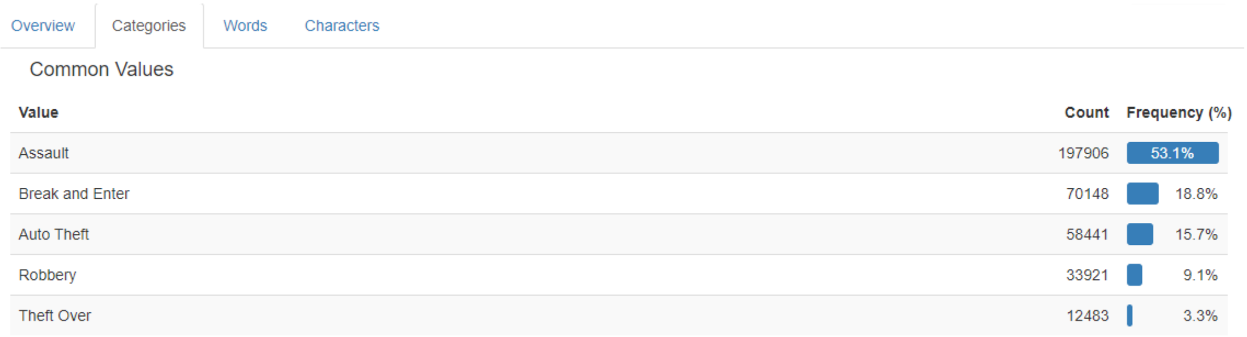
Numeric	15
Text	7
DateTime	2
Categorical	7

The categorical attribute overview is shown in the graph below. No missing values were found. Additionally, there are 5 labels as expected, with events categorized as Assault constituting more than half of the total crime occurrences. All relevant attributes to be used in this study do not include missing values.



MCI_CATEGORY
Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	23.5 MiB

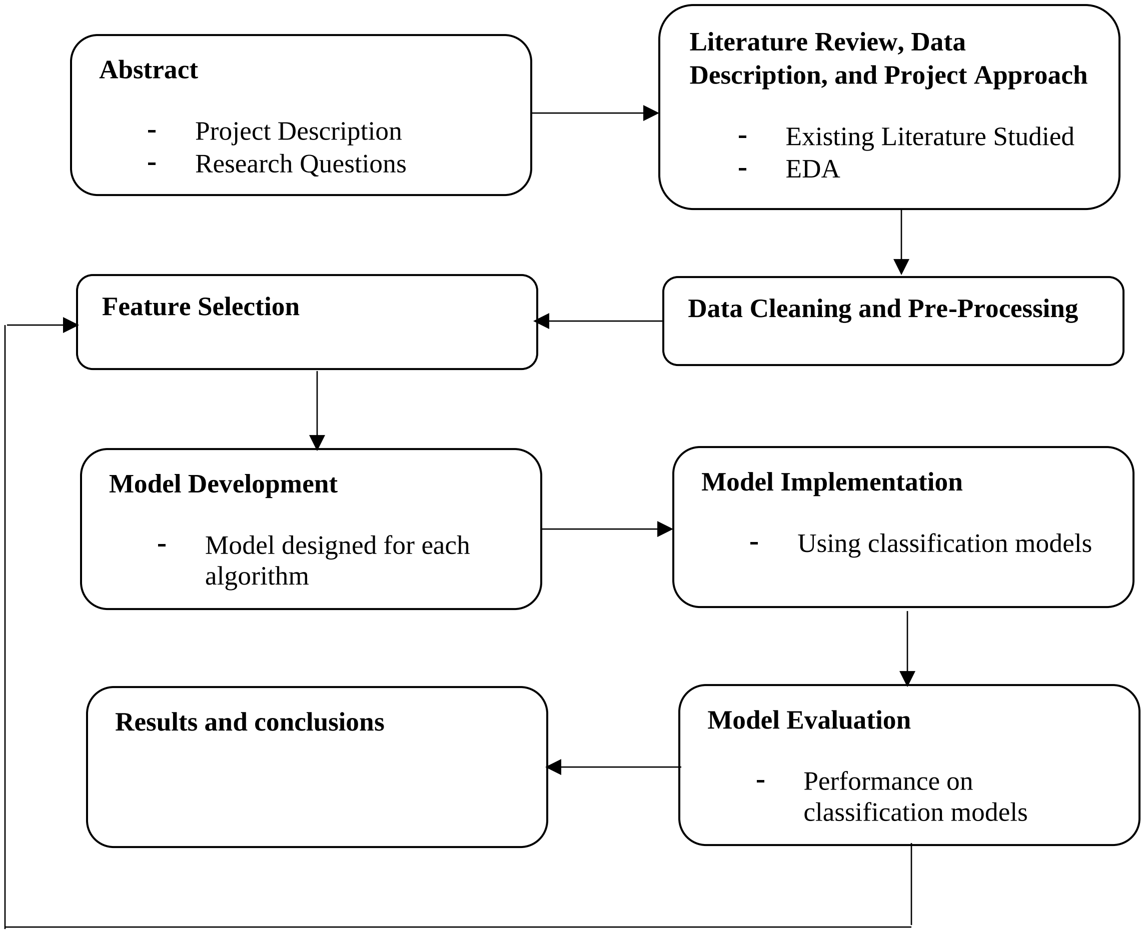


Common Values

Value	Count	Frequency (%)
Assault	197906	53.1%
Break and Enter	70148	18.8%
Auto Theft	58441	15.7%
Robbery	33921	9.1%
Theft Over	12483	3.3%

Project Approach

The following chart represents the workflow and overall methodology of this project.


Abstract
- Project Description
- Research Questions
Literature Review, Data Description, and Project Approach
- Existing Literature Studied
- EDA
Data Cleaning and Pre-Processing
Feature Selection
Model Development
- Model designed for each algorithm
Model Implementation
- Using classification models
Model Evaluation
- Performance on classification models
Results and conclusions

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

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
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

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

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