Crime Prediction Analytics using Supervised Learning Algorithms

by

Kariman Karm Mousaa

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
December 2022

Abstract

Humanity has witnessed many crimes throughout the ages and there are many attempts to combat the increase in crime. Many scientists in all fields have worked on this problem and with technological development, we have been able to use machine learning to predict the type of crimes, and whether the type of crime depend on cities or on time or on the seasons or other factors, there are many questions that we want to answer using dataset. In this research paper, the types of crimes are classified based on several different factors using supervised learning algorithms.

Keywords

Supervised Machine Learning algorithms, crime prediction, classification Machine Learning algorithms.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Economic growth, tourism demand and material stability at the personal or international level are affected by the crime rate and the type of crimes. On a personal level, where the person decides to move or visit a place based on the extent of safety and the type of crimes and their rate in the city to which it will go, and On the international level, finding countries where tourism and the economy are declining significantly due to the crimes of terrorism, murder, theft and other crimes [13]. The tremendous development in a data mining approach in recent times has made us able to solve such problems, but our biggest challenge is dealing with data, as the number of crimes in a state of continuous increase and increase in the volume of data may lead to problems in analysis and storage.

1.2 Problem

Crime in relation to infringement is divided into two groups, it can be violent and non-violent crimes. Nonviolent crimes are said to be crimes without harm or force. A violent crime or a violent crime the person employed by the offender, or intimidation of the victim by a hostile act, and may or may not be carried out with munitions. Our study aims to predict the type of crime using a set of realistic crime datasets. Discover and analyze data, explore the relationship between features and types of crimes, find the most common crimes in each city, and compare datasets in terms of the algorithm's impact on them, the best results, and find the most common crimes in each of them. There are many problems with our interfaces, including the imbalance of types of crimes and dealing with data for Vancouver, it has different features and there are features that are almost the same.

1.3 Objective

The expectation of this paper is to build a good model, avoid overfitting, and create a useful insight that helps reduce crime and raise awareness of the trend. Find out the types of crime that are most prevalent in the city. Find out if there is a relationship between the type of crime and other factors such as time, place, or other factors and Comparison between Toronto and Vancouver dataset. That will help the police, the public and states in confronting these crimes.

1.4 Contributions

The key challenges in this thesis are listed as followings. We were able to improve the performance, and we were able to deal with the problem of weak features and extract other features from them that helped us to improve. We also used better algorithms from the previous work close to our data, and we compared them.

1.5 Organization of Thesis

The first chapter is an introduction to the definition of the problem, motives for working on that problem, and an overview of what has been achieved in the work. In the second chapter, we discuss previous work like our problems and the scientific background. In the third chapter, we discuss the scientific methodology, solutions, data sources, and work steps. In the fourth chapter we discuss the results and a comparison between the data and previous work, and in the fifth chapter we discuss a summary of everything that has been done in the work and future work.

Chapter 2

Background

Many techniques were used to solve this problem and finding this in detail in related work section. There are several techniques reported in scientific research that can be used to predict and classify in crime predict. As there are some ideas that extract data from the news or social media and analyze it, find insight, and know which crimes are prevalent at this time. There is also the identification of the type of crime through photos and videos. There is also unsupervised learning to crime Detection. In this research will be used supervised machine learning algorithms. After the data is processed, analyzed, and insight is extracted. A model uses supervised machine learning algorithms to predict the type of crimes, this is known as classification. We will use the best algorithms in related work and compare it with our work, they will be used and compared.

2.0.1 Supervised Machine Learning

It's a type of machine learning that labels data, it's a model of relationships and dependencies between the target prediction output and the input features. It's defined as "analyzes the training data and produces an inferred function, which can be used for mapping new examples' [6].it has two types of regression and classification.

1. Regression algorithms, it is supervised learning, The output is having continuous value. Such as linear regression, Lasso Regression, Random Forest Regressor, Support Vector Regression, Decision Tree etc.

2. classification, it is supervised learning. It's defined as "an important topic in machine learning, a function that weighs the input features so that the output separates one class into positive values and the other into negative values" [4], The output is having discrete value. Such as Random Forest, decision trees, Logistic regression, Support Vector machine, KNN etc.

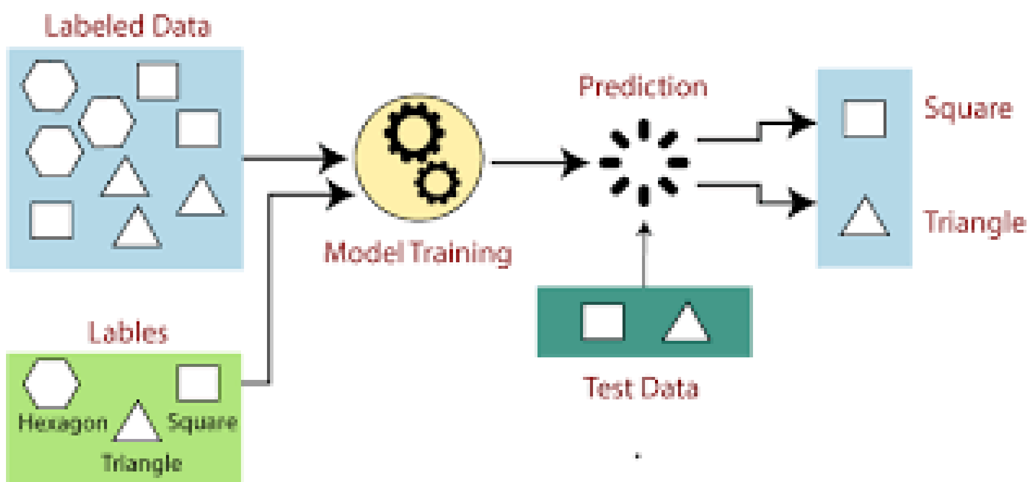

Figure 2.1: Classification [3].

2.0.2 Parametric Algorithms

These are machine learning algorithms. These are defined as "Assumptions can simplify the learning process but can also limit what can be learned. Algorithms that simplify the function to a known form are called parametric machine learning algorithms." [5]. It is completely dependent on mathematics and its total dependence on the number of parameters, regardless of the number of training samples. Its advantages are that it is simple and fast, and not need to have large data to work well. Its disadvantages are that it is used in minor problems, meaning limited, poor fit and constrain. Such as: Linear regression, logistic regression, perception, Naive Bayes, Simple Networks.

2.0.3 Non-Parametric Algorithms

These are machine learning algorithms. These are defined as "Algorithms that do not make strong assumptions about the form of the mapping function are called non-parametric machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data." [5]. It depends on the data, and for that it needs big data can reach millions of rows. This is an important thing. Even if the data is large and there is no prior knowledge about it, it will not be worried in terms of the right features. All of these are considered advantages. Its disadvantages remain, of course, that it is expensive, complex, also slow, needs large data and it can happen over-fitting. Like KNN, Decision Trees like CART AND C4.5, SVM, Random Forest (RF).

2.1 Related Work

The classification of crime is one of the important topics and has received attention in much research. There are several techniques reported in scientific research that can be used to predict and classify crime predict. There are some ideas that extract data from the news or social media and analyze it and find insight and know which crimes are prevalent currently [14]. There is also the identification of the type of crime through photos and videos [15]. There is also unsupervised learning to crime Detection [12]. Though the related works that we focused on are the articles that used structure data and focused on classification using supervised learning. In [16], the data was collected from law enforcement agencies. Relates to areas in which crime and offender data occurs. Several models have been used which are Artificial Neural Networks, Naive Bayes Classifier, Support Vector Machine, and the decision tree were used. They used recall, precision and F1. The best algorithm was the decision tree with F1 equal to (.761) and the least algorithm was the Support Vector Machine F1 equal to (.632). In [17], focusing on violent crimes in Brazil, but they were suffering from misleading information about violent crimes because of mistrust in the police, so they collected data from the Onde Fui Roubado website, which is Unofficial Records, and Official Records provided by Brazilian government. Models have been used which are The Support Vector Machines, Random Forest, k-NN, and the extreme Gradient Boosting. They used recall, precision, F1 and accuracy. The best algorithms are Random Forest and XGBoost and the least algorithms are k-NN and SVM. Scores differ from data to other data as well from year to other year. In [18], the data was collected from the police department of San Francisco. Unsupervised learning algorithms were used to feature engineering, select appropriate features, and use supervisor learning to predict the type of crimes. The best algorithm is Random Forest with F1 equal 0.190 and with Log loss equal 16.7994 and the least algorithm is Gradient Boosting with F1 equal 0.151 and with log loss equal 2.53044. In [11], data from Global data events that happened in Saudi Arabia, they have many problems with data such as missing values, and uneven data distribution, they used PCA and FAMD for dimensional reduction. The best algorithm is Naive Bayes. There is a comparison between related work and current work in the results.

Chapter 3

Methodology

We strongly believe that finding relationships between crime features can help us in predicting the type of crime. This research focuses on the main purpose that's crime types in Toronto and Vancouver. We tried to extract everything possible interesting repeatedly Patterns based on crime variables. Then, we analyzed and processed the data. Next, we present how we set up data mining. Then, we applied some classification models for predicting the types of crimes in a particular place during a certain time and other features such as neighbors. In this section, we explain how we set up our data sets. Models to achieve our goal.

3.1 Data Source

We use two datasets, where comparing different machine learning algorithms on different datasets and know the extent of their impact on each data set. Our datasets do not have the same values, not even Featured, and this may help us with how machine learning algorithms react to difference and featured data sets.

1. The first dataset for cities in Toronto, from the Open Data Catalog from the

	X	Y	Index_	event_unique_id	Division	occurrencedate	reporteddate	location_type	premises_type	ucr_code	...	occurrenceday
0	-8.859955e+06	5.424372e+06	110	GO-20141625305	D23	2014/03/02 05:00:00+00	2014/03/02 05:00:00+00	Single Home, House (Attach Garage, Cottage, Mo...	House	1430	...	2
1	-8.861110e+06	5.424036e+06	188	GO-20141272968	D23	2013/12/24 05:00:00+00	2014/01/03 05:00:00+00	Commercial Dwelling Unit (Hotel, Motel, B & B,...	Commercial	1610	...	24
2	-8.861120e+06	5.417043e+06	287	GO-20141284361	D23	2013/01/05 05:00:00+00	2014/01/05 05:00:00+00	Commercial Dwelling Unit (Hotel, Motel, B & B,...	Commercial	1430	...	5
3	-8.861463e+06	5.425856e+06	384	GO-20141292177	D23	2013/12/31 05:00:00+00	2014/01/06 05:00:00+00	Other Commercial / Corporate Places (For Profi...	Commercial	2120	...	31
4	-8.859522e+06	5.418688e+06	438	GO-20141297201	D23	2014/01/03 05:00:00+00	2014/01/07 05:00:00+00	Other Commercial / Corporate Places (For Profi...	Commercial	2120	...	3

Figure 3.2: Vancouver dataset

Public Safety Data Portal in the Toronto Police Service, that is 'MCI 2014 to 2022'. It has 836190 records in ten columns. The features are TYPE: type of crime, YEAR, MONTH, DAY, HOUR, MINUTE, HUNDRED BLOCK, NEIGHBOURHOOD, X and Y [1].

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y
0	Theft from Vehicle	2008	12	7	16	0	16XX GRANVILLE ST	Central Business District	490419.584	5.457877e+06
1	Theft from Vehicle	2009	4	6	2	0	16XX GRANVILLE ST	Central Business District	490419.584	5.457877e+06
2	Theft from Vehicle	2009	4	21	19	15	16XX GRANVILLE ST	Central Business District	490419.584	5.457877e+06
3	Theft from Vehicle	2009	4	24	18	30	16XX GRANVILLE ST	Central Business District	490419.584	5.457877e+06
4	Theft from Vehicle	2009	5	24	13	30	16XX GRANVILLE ST	Central Business District	490419.584	5.457877e+06

Figure 3.1: Toronto dataset

2. The second dataset for cities in Vancouver, from the Vancouver Police department website, is 'GeoDash Open Data.' It has 281692 records and thirty columns. The features are X, Y, Index, event unique id, Division, occurrence date, reported date, location type, premises type, ucr code, ucr ext, offence, reported year, reported month, reported day, reported day of year, reported day of week, reported hour, occurrence year, occurrence month, occurrence day, occurrence day of year, occurrence day of week, occurrence hour and MCI, Hood ID, Neighborhood, Long, Lat, and Object ID [2].

3.2 Data Prepossessing and Analysis

1. In Toronto data, we have 10 columns, it duplicated 29832 records, we remove the duplication and kept the first records. And it has missing values in four columns, but their number is very small compared to the size of the data, so it did not reach a quarter, so we removed it.

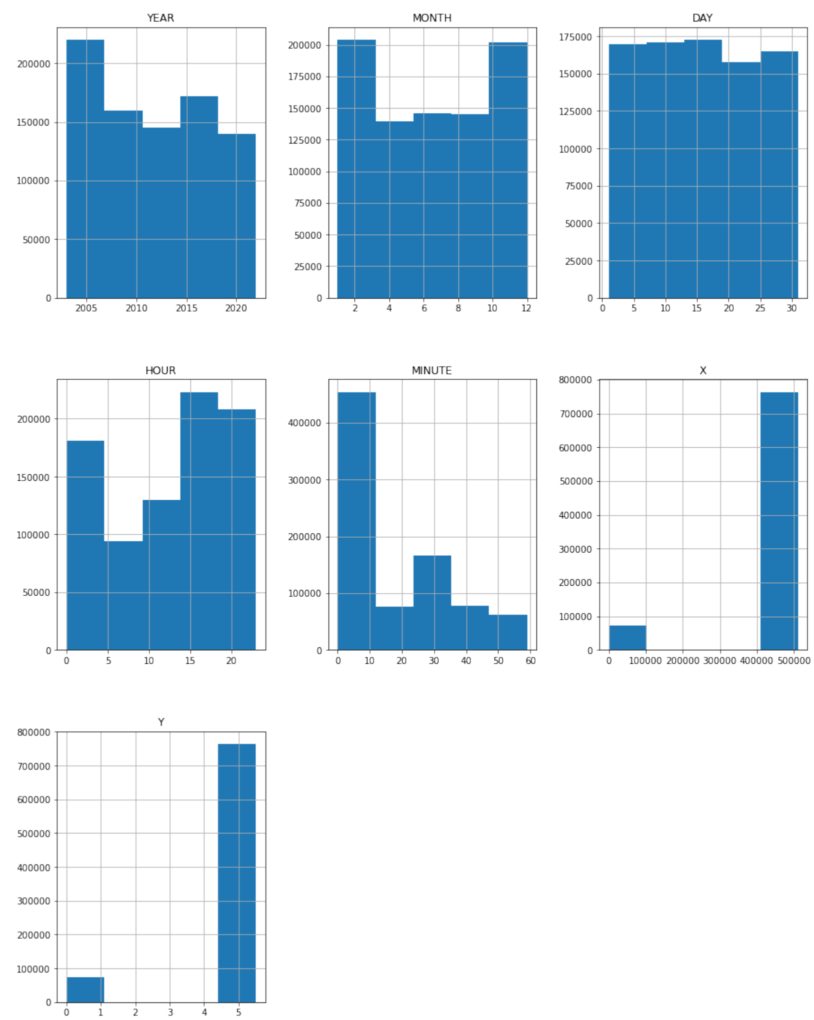

Figure 3.3: Histogram for Toronto features

2. There are many 10 classes such as Theft from Vehicle, Other Theft, Mischief, etc., but when we see the number of each class, we will find that there is no

balance here, as we will find that Theft from Vehicle acquires 234189 rows and Other Theft acquires 205014 rows, while Vehicle Collision or Pedestrian Struck (with Fatality) equals 334 rows and Homicide equals 301 rows, but this makes sense in real life that there is an imbalance in the types of crimes, we also note that the features are weak to help us predict types of crime.

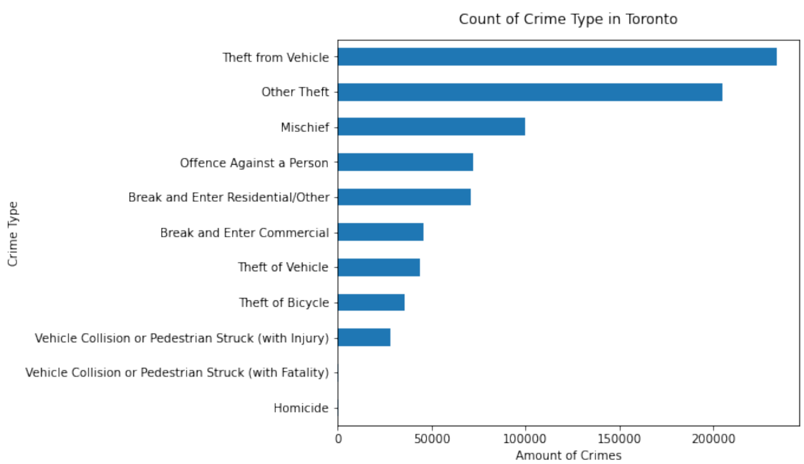

Figure 3.4: Crime distribution in Toronto

3. In Vancouver data, there isn't duplication or missing values. It has 30 records.

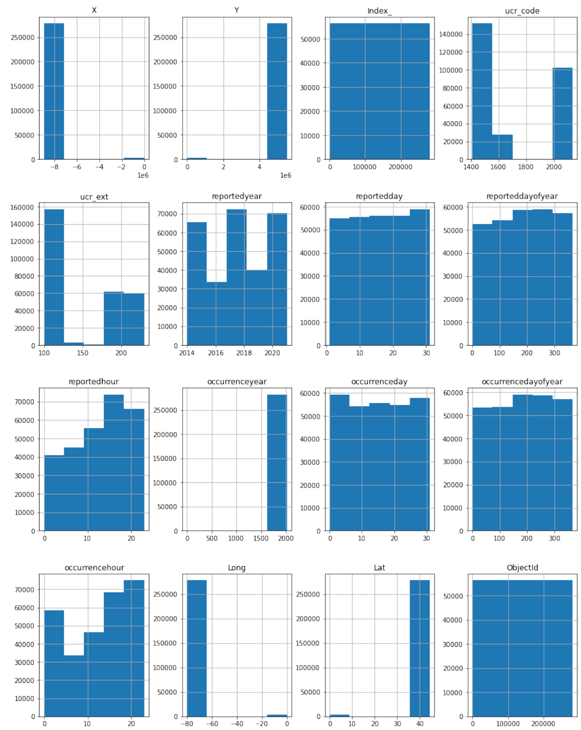

Figure 3.5: Histogram for Vancouver features

4. Class numbers of crime types is much higher compared to Toronto and this is due to the different data sources, but there is still an imbalance in the distribution of the class, as assault has 103627 rows, while robbery to steal firearm equals 1 row.

3.3 Feature Engineering

In Vancouver data, there are many useless features such as Object Id, Hood ID, event unique ID, Index and ucr code also here are columns with almost recurring values such as offence and MCI, and Long, x, Lat, and y. we delete the useless columns and keep one of the columns and delete the other in the duplicate columns, and then we create new columns holidays and weekends. The same happens with Toronto, though

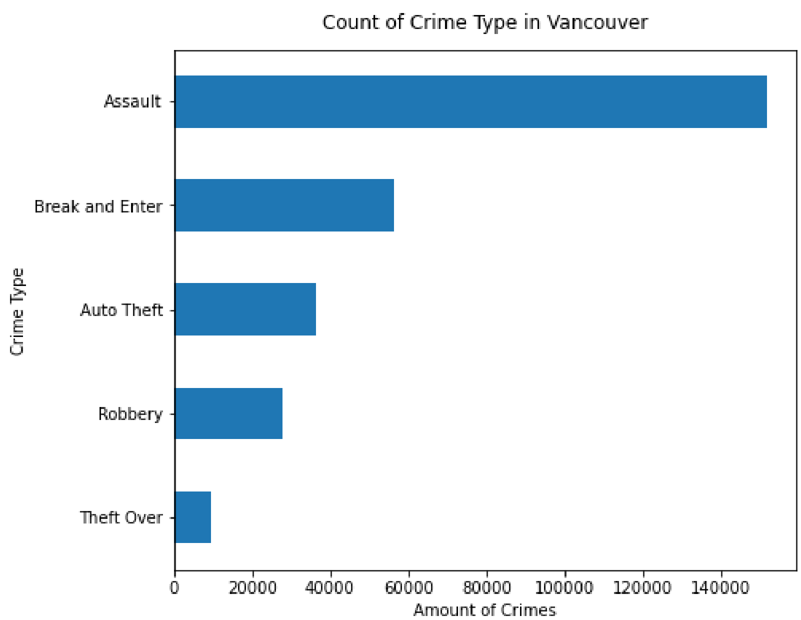

Figure 3.6: Crime distribution in Vancouver

there are no useless features or duplicate values, so we create columns for holidays and weekends and while retaining all the features.

3.4 Models

We used many algorithms for classification of crimes like K-Neighbors Classifier (KNN), Naive Bayes, random forest, and Extreme Gradient Boosting (XGBoost), these algorithms have been selected based on the related work.

3.4.1 K-Neighbors Classifier (KNN)

They are defined as algorithms that do not contain training parameters though keep training data and use them during the classification process for the test point. This type of algorithm has high performance in results because hypotheses are not built about the model function. We just need to have two things in the data: a way to calculate the distance between the data. Assume that data close to each other are similar and far from each other are not.[7] The algorithm done in the following steps: 1- We determine the value of the variable k, which expresses the number of neighbors. 2- Calculate the value of the distance between the new example and the examples in the dataset.3- We arrange the examples to get the neighbors based on the least distance that calculated in the previous step and take the number of k adjacent from them. 4- We define the class for the neighbors. 5- The most common class of the neighbors is the expected class. To find the distance between any two points by this is equation:

$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

3.4.2 Naive Bayes

The Naive Bayes algorithm is a technique based on the theory of the independence hypothesis among predictors. That is, the presence of a particular feature in a class that is unrelated to the existence of any other feature. The Naive Bayes algorithm is easy to construct and is especially useful for large data sets. Besides simplicity, Naive Bayes is known to outperform even other complex classification methods. Bayes theorem provides a method for calculating the posterior probability P(c — x) from P(c), P(x), and P(x — c) [10].

$$P(A|B) = P(B|A) * P(A)/P(B)$$

3.4.3 Random Forest

It builds decision trees on different samples and takes their majority vote for classification. The algorithm is done in several of the following steps: From the data set containing k number of records, the number of random records is taken. For each sample, individual decision trees are generated. And each decision tree will produce an output. The outcome taken based on majority vote or average rating[9].

3.4.4 Boosting

The boosting algorithm attempts to improve predictive capabilities by training a series of weak models, each of which can compensate for the weaknesses of its predecessors [8]. Adaptive Boosting (AdaBoost), it is a type of boosting and is a machine learning algorithm and can be used with many other types of learning algorithms to improve performance. On each iteration, AdaBoost will identify a misclassified data point, thereby increasing its weight (in other words decreasing the weight of the integer point) so that the next classifier will be more attentive to correcting it. Gradient Boost, one of the common boosting algorithms, this algorithm works by adding prediction models sequentially, and each model corrects the previous model, but instead of adjusting the weights of the examples at each step as is done in AdaBoost, this method tries to fit the new prediction model with the remaining errors from the previous model. Extreme Gradient Boosting (XGBoost), it is one of the extremely popular and widely used boosting algorithms simply because it is powerful. It is like the Gradient Boost algorithm, but it has some additional features that make it much more powerful; Where training is extremely fast and can be balanced or distributed across groups. After that, Turing the best algorithms. We talk in detail about the products, the best algorithms, and the disadvantages of each algorithm in our problems.

Chapter 4

Results

We note from the previous sections that accuracy was not the best performance in order to evaluate the model because the data has an imbalance in the type of crimes, so we used F1, classification report performance of each class. In the beginning, when we converted all categories to numbers using Label Encoder, the performance of the algorithms was extremely poor, the XGBoost gave the best result and it was forty-five%, but we improved by transforming Neighborhood in each of the data by One Hot Encoder method. Naive Bayes and KNN were bad even though they were doing well with the related work. KNN was taking a long time to run, and this is logical because it calculates the distance between each point, and it also needed scaling, and this may lose us some data. The decision tree was not good for both data: it was good at training(f1=99% and accuracy=99.8%) and testing(f1=76% and accuracy=90%) in Vancouver, it was good at training(f1=98% and accuracy=99%) and bad at testing(f1=44.4 % and accuracy=61.5%, there is overfitting) in Toronto. But the best algorithms were XGBoost, and Random Forest. Classes of types of crimes in Toronto are almost recurring, and we notice this clearly when according to the algorithms in the beginning, the performance was not good enough, so we regrouping the classes, and you will find that a comparison between before and after in table 4.1.

(1)Toronto		
Algorithms(F1)	Before Regrouping	After Regrouping
Decision Tree	0.4446	0.7425
Random Forrest	0.4782	0.80139
XGBoost	0.476	0.80928

Table 4.1: Before Regrouping After Regrouping in Toronto

In another attempt to improve, we have made Hyper-parameter tuning for KNN, Random Forest by Using Randomized Search, the best hyperparameters in KNN were number of neighbors= 15, weights is distance, algorithm is brute, and metric is Manhattan, before tuning KNN F1 was 35 %, and after f1= 37%. The best hyper-parameters in Random Forest were number of estimators= 200, and max of features is auto. Before tuning, Random Forest F1 was 47%, and after f1= 48%. Although XGBoost, and Random Forest performed the best, there were still some classes that could have been predicted and this is due to the imbalance as shown in the tables below. Random Forest in Vancouver table 4.2 in Toronto table 4.3 and table 4.4. To solve this problem, we did Oversampling using Smote, but it did not make a difference to the results. In the table below, compare the best results for all the algorithms in both Toronto, and Vancouver table 4.6.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	45497
1	0.88	0.96	0.92	10899
2	0.89	0.98	0.93	16934
3	0.87	0.70	0.77	8344
4	0.62	0.17	0.27	2777
accuracy			0.93	84451
macro avg	0.85	0.76	0.78	84451
weighted avg	0.92	0.93	0.92	84451

Table 4.2: Random Forest in Vancouver

	precision	recall	f1-score	support
0	0.58	0.33	0.42	13693
1	0.59	0.53	0.56	21421
2	0.00	0.00	0.00	88
3	0.54	0.43	0.48	29849
4	0.99	1.00	1.00	12723
5	0.82	0.91	0.86	61497
6	0.65	0.92	0.76	61131
7	0.56	0.19	0.28	10853
8	0.45	0.12	0.19	13130
9	0.00	0.00	0.00	85
10	0.79	0.64	0.70	8397
accuracy			0.70	232867
macro avg	0.54	0.46	0.48	232867
weighted avg	0.68	0.70	0.67	232867

Table 4.3: Random Forest in Toronto before regrouping

	precision	recall	f1-score	support
0	0.91	0.74	0.82	35114
2	1.00	1.00	1.00	12811
3	0.67	0.55	0.61	29849
5	0.84	0.91	0.88	61497
6	0.75	0.94	0.83	61131
7	0.93	0.64	0.76	23983
10	0.81	0.64	0.72	8482
accuracy			0.82	232867
macro avg	0.84	0.78	0.80	232867
weighted avg	0.82	0.82	0.81	232867

Table 4.4: Random Forest in Toronto after regrouping

Algorithms(F1)	Toronto	Vancouver
Decision Tree	0.7425	0.760319
Random Forrest	0.80139	0.7783
XGBoost	0.80928	0.7881
KNN	0.3474	0.45027
Naive Bayes	0.1913	0.3268

Table 4.5: Compare the best results of Toronto and Vancouver.

Algorithms(F1)	Toronto	Vancouver	[16]	[18]
Decision Tree	0.7425	0.760319	.761	-
Random Forrest	0.80139	0.7783	-	0.190
XGBoost	0.80928	0.7881	-	0.151
KNN	0.3474	0.45027	-	-
Naive Bayes	0.1913	0.3268	-	.154

Table 4.6: comparison between related work and current work in the results.

Chapter 5

Discussion

5.1 Summary

We used supervised learning on two datasets extracted from responses from two diverse sources, one for Vancouver and one for Toronto. Then we processed the data by deleting redundancy, missing data, and changing the data type for some features, and then extracting the useful features and deleting the useless features. The algorithms were selected based on the related work. The best algorithms were XGBoost, and Random Forest. Naive Bayes and KNN were worse. We have an imbalance in the data, so we used oversampling, although it didn't change much. The best evaluation was f1.

5.2 Future Work

It is good to deal with larger data, different cities, and data that contains stronger features that help us achieve the goal easily. One of the important things is to use the dashboard to display the analyses of cities to make it easier to see the results. Create a website or application in which all data for tourist or immigration areas are collected to help people avoid crimes.

5.3 limitations challenges

It was daunting in the beginning to know how to improve the data to be important features that help us in predicting types of crimes until I was able to create two columns with the existing features, but I needed stronger features. The part about regrouping crime types in Toronto was added and significantly improved performance. Although despite all attempts to solve the problem of imbalance until there are still some classes, they were not well predicted, so we need more data that contains these classes to make it easier to predict.

Bibliography

[1] Toronto dataset. toronto police service. URL https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-1/explore.

[2] vancouvar dataset. geodash open data.from the vancouver police department website. URL https://geodash.vpd.ca/opendata.

[3] classification image.supervised machine learning- javatpoint. URL data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAT4AixYoVK1asWLFixYr1C9T/AaNNHtiFqbDaAAAAAElFTkSuQmCC.

[4] Classification algorithm. sciencedirect., 2014. URL https://en.wikipedia.org/wiki/Supervised_learning.

[5] Parametric and nonparametric machine learning algorithms. machine learning mastery., 2020. URL https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms.

[6] Supervised learning. wikipedia., 2020. URL https://en.wikipedia.org/wiki/Supervised_learning.

[7] k-nearest neighbors. wikipedia., 2021. URL https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

[8] boosting. analyticsvidhya., 2021. URL `https://www.analyticsvidhya.com/blog/2021/04/best-boosting-algorithm-in-machine-learning-in-2021/`.

[9] Random forest. analyticsvidhya., 2022. URL `https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/`.

[10] Naive bayes classifier. wikipedia., 2022. URL `https://en.wikipedia.org/wiki/Naive_Bayes_classifier`.

[11] S. Albahli, A. Alsaqabi, F. Aldhubayi, H. T. Rauf, M. Arif, M. Mohammed, and H. Tayyab. Predicting the type of crime: Intelligence gathering and crime analysis. *Cmc -Tech Science Press-*, 66:2318–2341, 12 2020. doi: 10.32604/cmc.2021.014113.

[12] R. Bullibabu, G. Snehal, and P. Kiran. Detection of crimes using unsupervised learning techniques. *Indian Journal of Science and Technology*, 9, 05 2016. doi: 10.17485/ijst/2016/v9i17/92994.

[13] D. Du, A. Lew, and P. Ng. Tourism and economic growth. *Journal of Travel Research*, 55:454–464, 12 2016. doi: 10.1177/0047287514563167.

[14] S. Ghankutkar, N. Sarkar, P. Gajbhiye, S. Yadav, D. Kalbande, and N. Bakereywala. Modelling machine learning for analysing crime news. pages 1–5, 12 2019. doi: 10.1109/ICAC347590.2019.9036769.

[15] J. Karri, J. Bhargavi, and E. Ijmtst. Classification of crime scene images using the computer vision and deep learning techniques. *International Journal for Modern Trends in Science and Technology*, 8:1–05, 02 2022. doi: 10.46501/IJMTST0802001.

[16] O. Llaha. Crime analysis and prediction using machine learning. pages 496–501, 09 2020. doi: 10.23919/MIPRO48935.2020.9245120.

[17] U. Rosa Monteiro de Castro, M. Rodrigues, and W. Brandão. Predicting crime by exploiting supervised learning on heterogeneous data. 05 2020. doi: 10.5220/0009392005240531.

[18] D. Sardana, S. Marwaha, and R. Bhatnagar. Supervised and unsupervised machine learning methodologies for crime pattern analysis. *International Journal of Artificial Intelligence Applications*, 12:83–99, 01 2021. doi: 10.5121/ijaia.2021.12106.