TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background

The increasing need to decline crime activity is essential for promoting peace and justice, aligning with the Sustainable Development Goals (SDGs) aimed at fostering inclusive societies and effective institutions. This paper utilizes machine learning techniques to predict crime activities, which can significantly contribute to cost reduction and resource optimization in law enforcement and public safety initiatives. The exploratory data analysis examined criminal data in Vancouver. Next, the data was analyzed using a time-series approach to predict future crime trends for the following years. The dataset analysed originates from the Vancouver Open Data Catalogue, comprising 881,242 records spanning from 1 January 2003, to 24 November 2023.

1.2 Problem Statement

The prediction of crimes in Vancouver is restricted by the lack of accurate and efficient techniques to analyze historical crime data. This affects the capability of the law enforcement agencies and public safety initiatives, resulting in higher costs to control the crime rate and inability to prevent future crimes. The usage of machine learning techniques to analyze and predict crimes can help to solve these problems and provide actionable insights for improving public safety.

1.3 Project Objectives

The following are the objectives that this research aims to achieve:

1. To identify trend, patterns, and correlations in Vancouver's crime data from 2003 to 2023 through exploratory data analysis.
2. To model crime prediction using machine learning techniques such as Random Forest Classifier, Decision Tree Classifier, XGBClassifier, and KNeighborsClassifier, along with time-series forecasting to predict future crime trends.
3. To evaluate the performance of the models by comparing their accuracy in predicting crime categories and trends, with a focus on improving prediction accuracy for law enforcement applications.

CHAPTER 2: METHODOLOGY

2.1 Data Collection

This dataset contains 8 features as shown in the table below.

Features	Description
TYPE	The type of crime activity.
YEAR	4 digit field that indicate the yearof the reported crime activity occurred.
MONTH	Numerical field that indicates which month of the year the crime activity occurred.
DAY	Numerical field that indicates the day of the month the crime activity occurred.
HOUR	Numerical field that indicates the hour time in 24 hours format the crime activity occurred.
MINUTE	Numerical field that indicates the minute the crime activity occurred.
HUNDRED_BLOCK	Generalized location of the crime activity occurred.
NEIGHBOURHOOD	The Neighbourhood name within the city of Vancouver based on the census tract (CT) concept.
X	The Longitude coordinate values in UTM Zone 10 format.
Y	The latitude coordinate values in UTM Zone 10 format.

2.2 Data Cleaning

In this section, the process and methodologies employ to ensure the quality of our dataset are outlined. This includes column renaming, removing null values, transformation, and handling outliers.

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y
0	Break and Enter Commercial	2012	12	14	8	52	NaN	Oakridge	491285.0000	5.453433e+06
1	Break and Enter Commercial	2019	3	7	2	6	10XX SITKA SQ	Fairview	490612.9648	5.457110e+06
2	Break and Enter Commercial	2019	8	27	4	12	10XX ALBERNI ST	West End	491004.8164	5.459177e+06

Figure 1: Before Data Cleaning

	type	year	month	day	hour	minute	address	neighbourhood	longitude	latitude	date	time	dayofweek
1	Break and Enter Commercial	2019	3	7	2	6	10XX SITKA SQ	Fairview	490612.9648	5.457110e+06	2019-03-07	02:06:00	Thursday
2	Break and Enter Commercial	2019	8	27	4	12	10XX ALBERNI ST	West End	491004.8164	5.459177e+06	2019-08-27	04:12:00	Tuesday
3	Break and Enter Commercial	2021	4	26	4	44	10XX ALBERNI ST	West End	491007.7798	5.459174e+06	2021-04-26	04:44:00	Monday

Figure 2: After Data Cleaning and Transformation

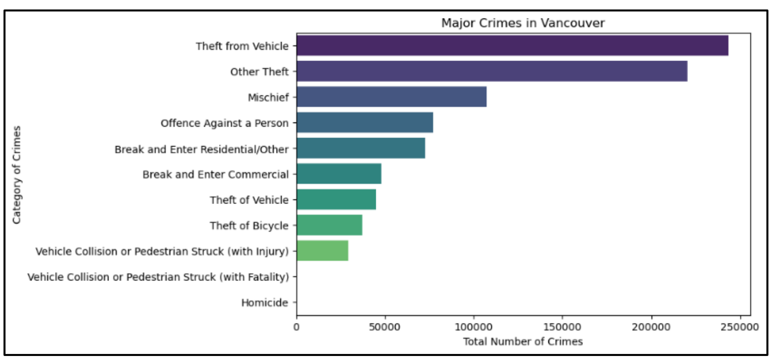
All column has been renamed for easier analysis. Data transformation includes adding new column such as date column which combine month and day column while time column are combination of hour and minute.

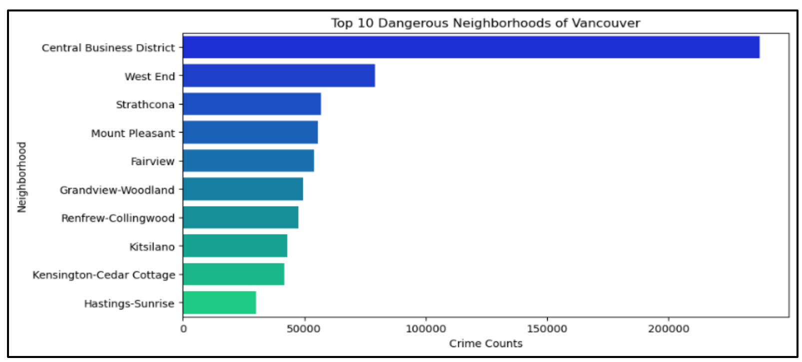
Removing Outliers

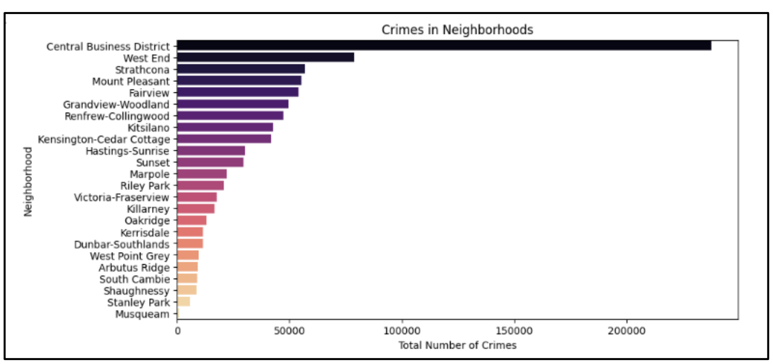
To deal with outliers, Interquartile Range (IQR) method has been used. First, the first Quartile (Q1) and the third Quartile (Q3) for related time column (Year, Month, Hour, Minute) was calculated. The IQR formula is "IQR=Q3-Q1". The acceptance value is within the range of Upper and Lower Boundary. The value outside these bounds was effectively filtered out to avoid skew of our analysis.

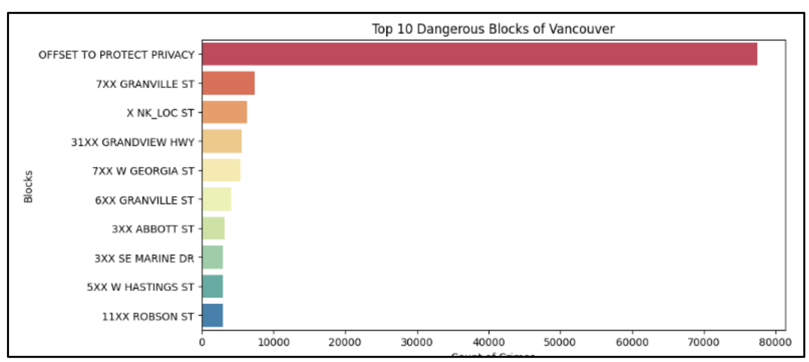
2.3 Exploratory Data Analysis

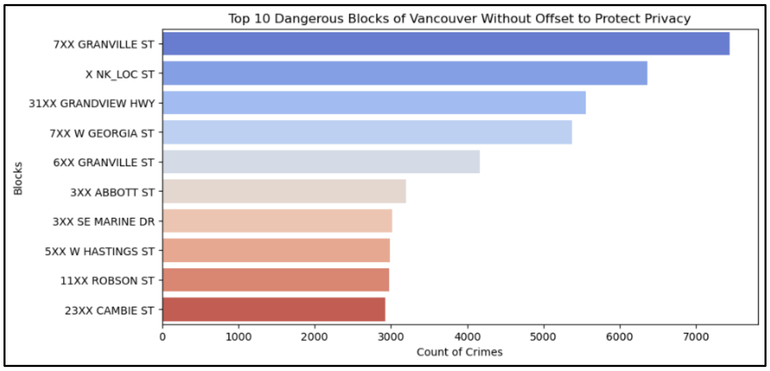
In this section, exploratory data analysis has been conducted to gain insight into the crime patterns in Vancouver from 2003 to 2023. Various aspects of the dataset have been examined including major types of crimes, neighborhoods distributions and trends and pattern over time. The aim of this EDA is to identify significant patterns and correlations that can inform public safety strategies and also to provide a clear understanding of overall crime activity in Vancouver for further investigation.



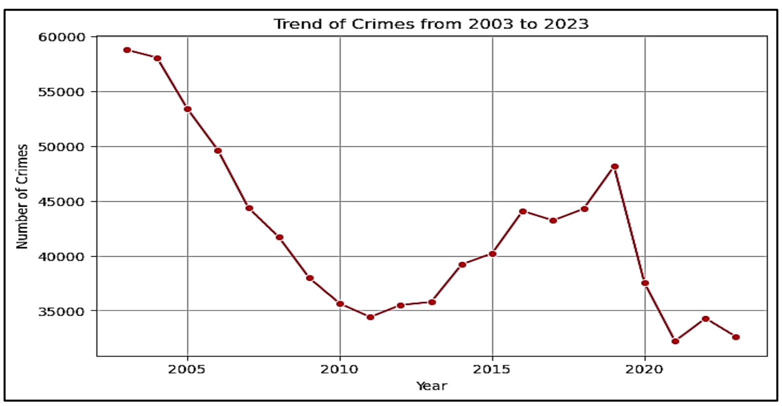


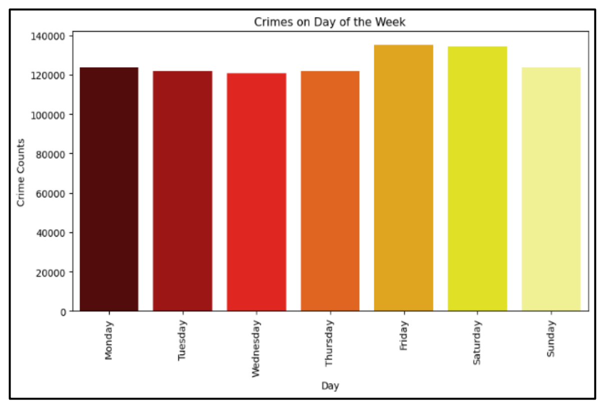


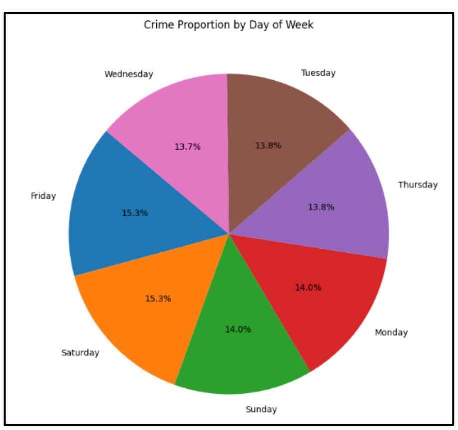


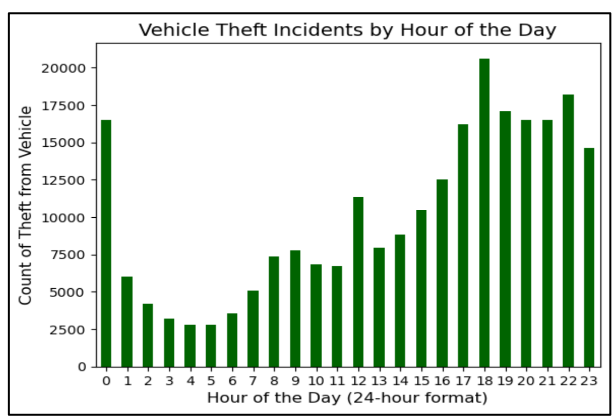


The exploratory data analysis (EDA) of Vancouver's crime data from 2003 to 2023 reveals key patterns and insights. Theft from vehicles emerges as the most common crime, followed by other thefts and mischief, highlighting the prevalence of property-related offenses. Neighborhood analysis shows the Central Business District as the highest-crime area, followed by the West End and Strathcona, with lower crime rates in residential areas like Musqueam. Further examination identifies Granville Street as the most dangerous block, consistently ranking highest in crime counts both with and without privacy offsets, alongside other high-crime blocks like Robson Street and Hastings Street. These findings pinpoint hotspots and trends that can help inform targeted safety measures and resource prioritization for crime prevention.

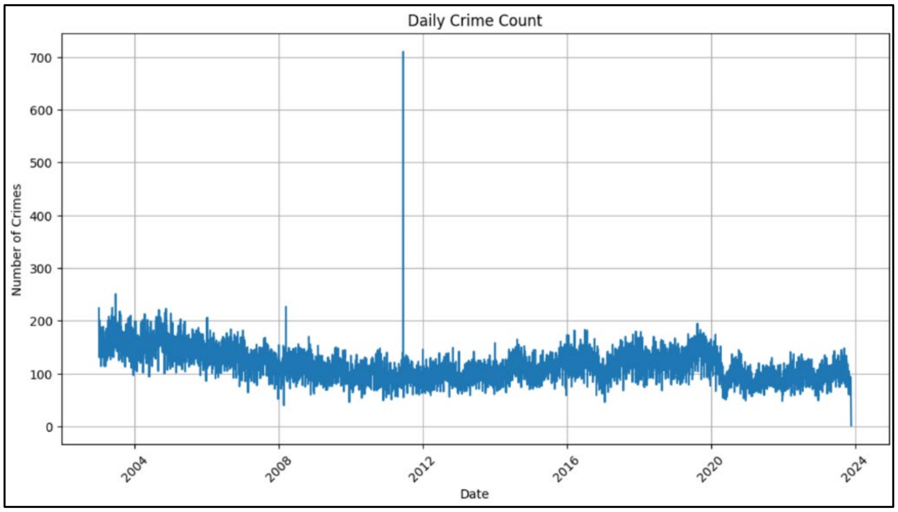




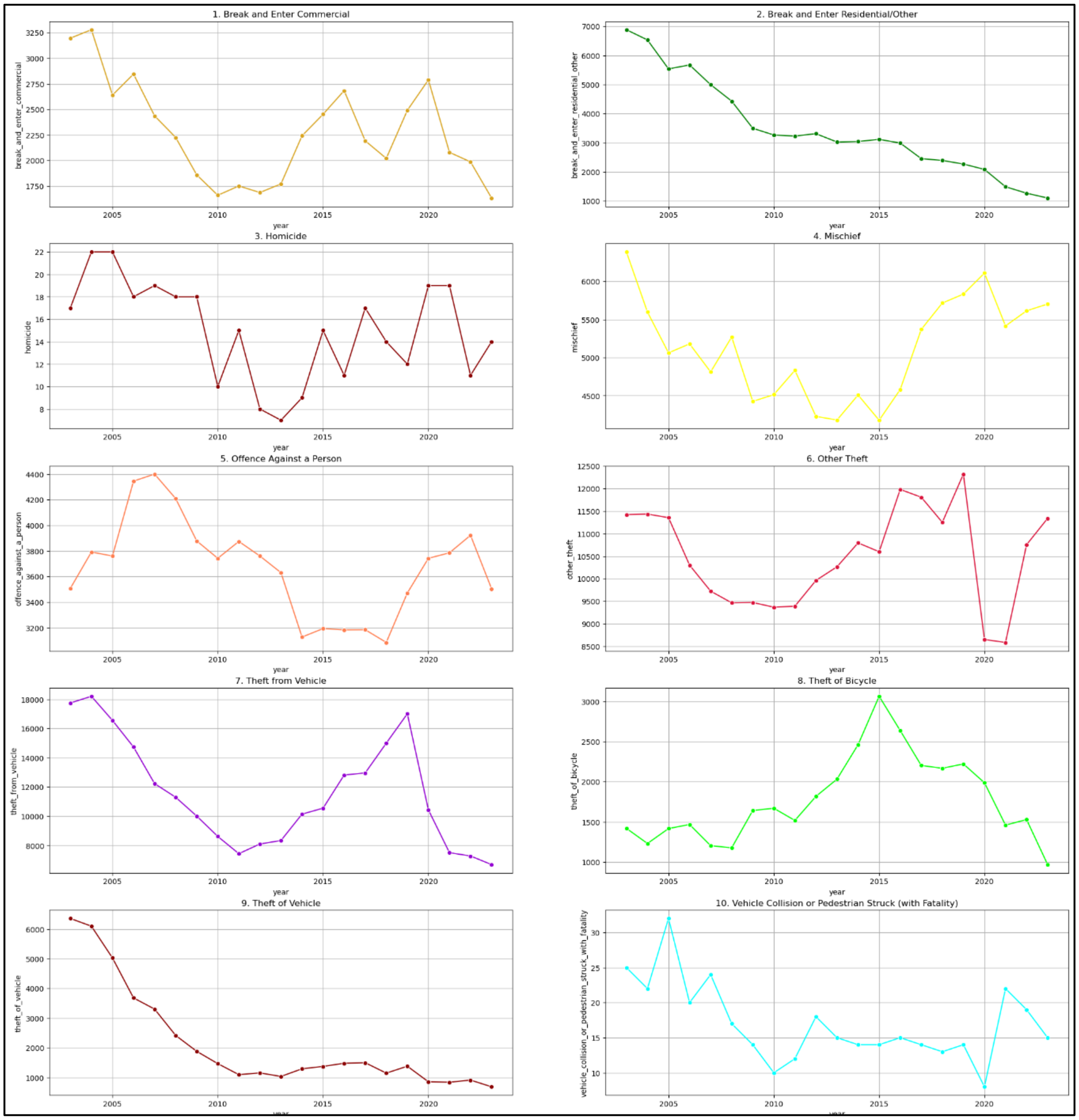




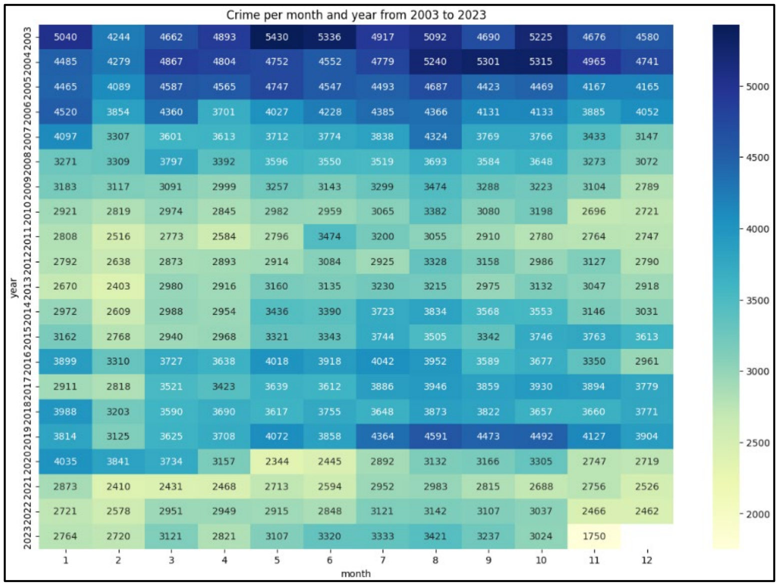
The analysis reveals a general decline in crime incidents over time, with occasional fluctuations. Fridays and Saturdays show the highest crime activity, while Wednesdays and Tuesdays have the lowest. The proportion of crimes is fairly distributed throughout the week, with slight peaks on weekends. Vehicle theft incidents are most frequent during evening hours, peaking between 6 PM and 9 PM, and are least common in the early morning hours, around 4 AM. These patterns highlight key temporal trends in criminal activity.

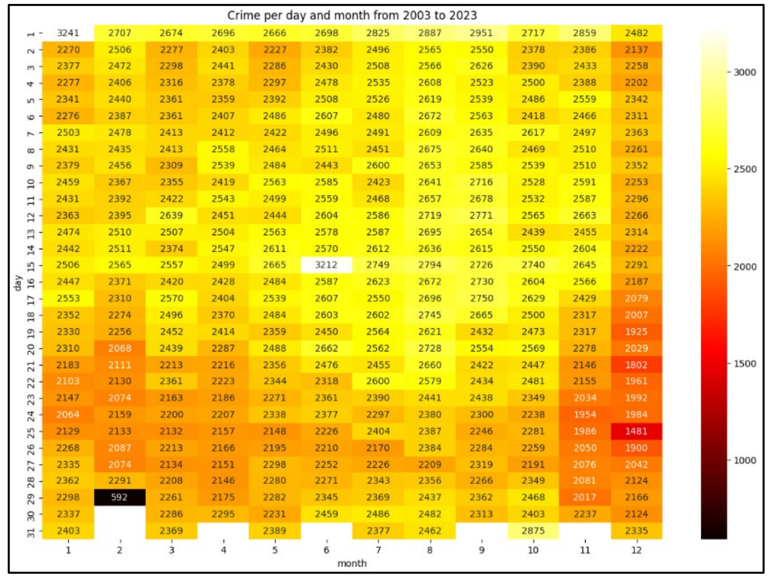


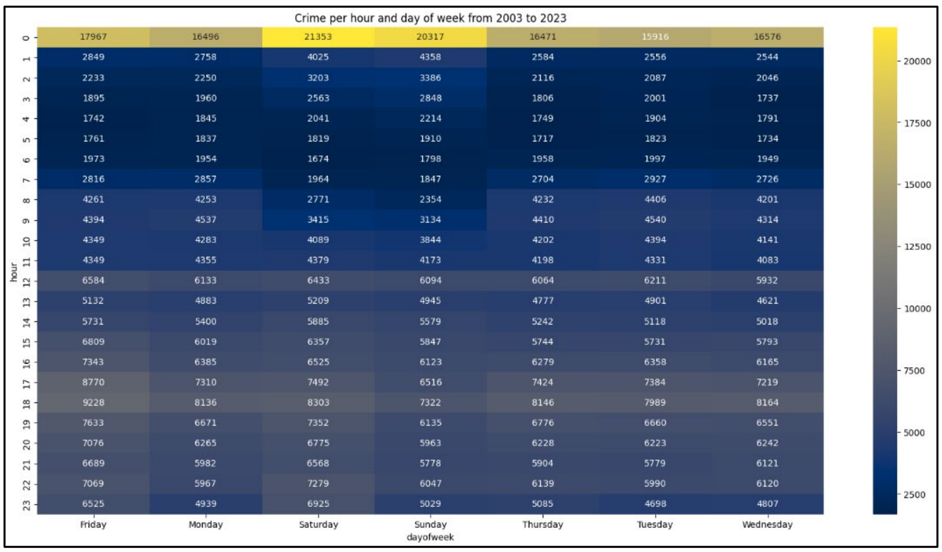
The time series graph of daily crime counts depicts overall variability, with a sharp spike indicating an isolated high-crime day. A significant spike in 2011, where daily crimes exceeded 700, suggests an anomaly likely tied to a specific event, such as a riot or protest. After 2012, crime counts stabilize with smaller fluctuations, though slight increases in variability appear in recent years. This trend reflects the impact of crime reduction strategies while emphasizing the need to investigate anomalies and consider external factors influencing crime patterns.



This distribution of each crime type from 2003 to 2023 allows for identifying which crime categories are most prevalent, their trends over time, and any shifts in patterns.







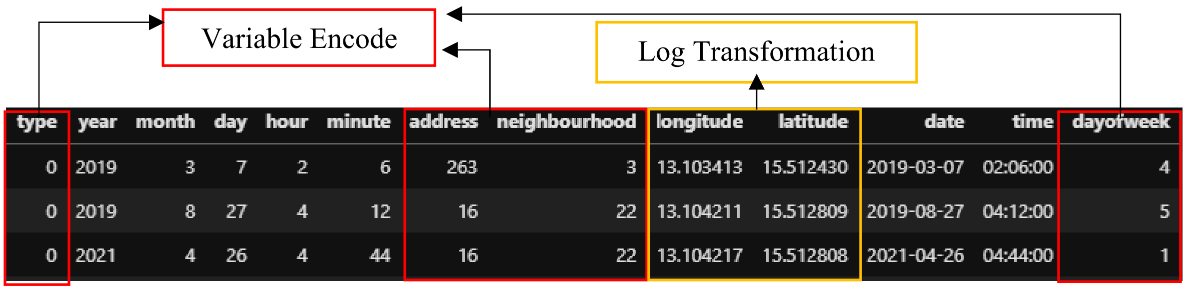
The first heatmap, showing crime by month and year, highlights long-term trends and seasonal patterns, such as summer spikes or yearly declines. The second heatmap, breaking down crime by day and month, reveals intra-month variations, like higher rates at month-ends or anomalies in February. The final heatmap, detailing crime by hour and day, shows increased activity during late nights and on Fridays and Saturdays. Together, these visualizations provide insights into temporal trends, seasonal variations, and daily patterns, guiding effective crime prevention efforts. Correlating these trends with factors like weather or population growth could enhance findings.

2.4 Data Modelling

This section is categorized into 2 sub parts which is Feature Engineering, Algorithm/Model Used and evaluation.

2.4.1 Feature Engineering

Feature Engineering was applied to enhance in quality, and to improve the model performance when the machine learning model are applied. There are several key techniques use in this phase:



type	year	month	day	hour	minute	address	neighbourhood	longitude	latitude	date	time	dayofweek
0	2019	3	7	2	6	263	3	13.103413	15.512430	2019-03-07	02:06:00	4
0	2019	8	27	4	12	16	22	13.104211	15.512809	2019-08-27	04:12:00	5
0	2021	4	26	4	44	16	22	13.104217	15.512808	2021-04-26	04:44:00	1

- Variable Encoding: It’s a process of encoding or converting categorical variables into numerical values.
- Log Transformation: Log transformation is applied to longitude and latitude data using formula of log(x+1).
- Chi-Square Independence Test: Chi Square Independence test is a statistical method to identify the strength of one feature and the target output variable.

2.4.2 Algorithm/Model Used

In this section, before any modelling and algorithm was applied, the datasets were split into 80% training data and 20% testing data. The type of algorithm use for different type of model (Classification & Forecasting) are mentioned as below.

Classification Model :

- Random Forest
- Decision Tress
- XGBoost
- K-Nearest Neighbour

Forecasting Model : Holt-Winter Time Series Forecasting

Evaluation Metrics: Metrics use to evaluate the model are precision, recall, F1 score and support.

2.5 Data Interpretation

This section presents and interprets the results of the classification models and time series forecasting applied in this project. The primary objectives were to classify crime types based on location data (longitude and latitude) and to predict future crime trends using a time-series approach. The classification task aimed to identify which features contributed most to accurately predicting crime types, utilizing models such as Random Forest Classifier, Decision Tree Classifier, XGBClassifier, and KNeighborsClassifier. After determining the crime categories, the data was further analyzed using time-series forecasting to predict crime trends for the upcoming years. The following subsections will detail the results from the classification models, evaluate their performance, and present the insights from the time-series analysis.

Classification Report:

	precision	recall	f1-score	support
0	0.38	0.28	0.32	9514
1	0.37	0.34	0.35	14548
2	0.00	0.00	0.00	50
3	0.38	0.27	0.32	21654
4	1.00	1.00	1.00	15392
5	0.63	0.68	0.66	44006
6	0.50	0.65	0.56	48594
7	0.21	0.08	0.11	7507
8	0.22	0.12	0.15	9050
9	0.19	0.05	0.07	64
10	0.88	0.80	0.84	5839
accuracy			0.55	176218
macro avg	0.43	0.39	0.40	176218
weighted avg	0.53	0.55	0.53	176218

Accuracy: 55.08%

Classification Report : Random Forest Classifier

Classification Report:

	precision	recall	f1-score	support
0	0.35	0.32	0.34	9514
1	0.36	0.39	0.37	14548
2	0.00	0.00	0.00	50
3	0.37	0.30	0.33	21654
4	1.00	1.00	1.00	15392
5	0.63	0.68	0.66	44006
6	0.51	0.62	0.56	48594
7	0.23	0.07	0.10	7507
8	0.22	0.09	0.13	9050
9	0.10	0.05	0.06	64
10	0.86	0.80	0.83	5839
accuracy			0.55	176218
macro avg	0.42	0.39	0.40	176218
weighted avg	0.53	0.55	0.53	176218

Accuracy: 54.96%

Classification Report: Decision Tree Classifier

Classification Report:

	precision	recall	f1-score	support
0	0.35	0.02	0.04	9514
1	0.37	0.08	0.13	14548
2	0.00	0.00	0.00	50
3	0.42	0.03	0.06	21654
4	1.00	1.00	1.00	15392
5	0.56	0.67	0.61	44006
6	0.39	0.81	0.53	48594
7	0.42	0.02	0.03	7507
8	0.31	0.00	0.00	9050
9	0.00	0.00	0.00	64
10	0.49	0.15	0.23	5839
accuracy			0.50	176218
macro avg	0.39	0.25	0.24	176218
weighted avg	0.49	0.50	0.41	176218

Accuracy: 49.62%

Classification Report: XGBClassifier

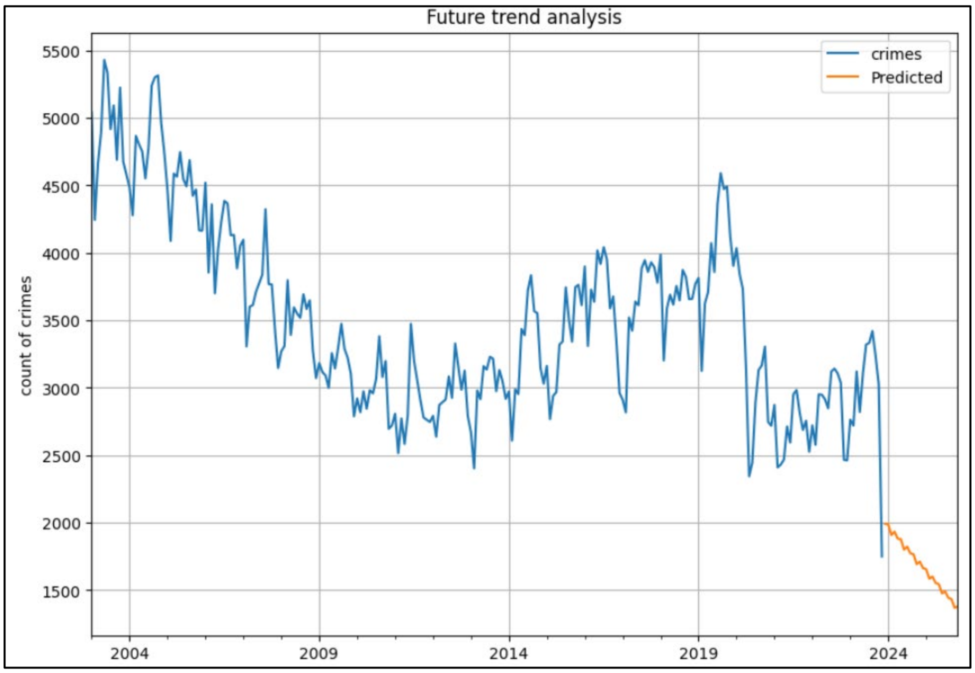
Classification Report:

	precision	recall	f1-score	support
0	0.29	0.33	0.31	9514
1	0.33	0.41	0.36	14548
2	0.00	0.00	0.00	50
3	0.34	0.33	0.33	21654
4	1.00	1.00	1.00	15392
5	0.62	0.64	0.63	44006
6	0.50	0.56	0.52	48594
7	0.22	0.06	0.10	7507
8	0.21	0.08	0.12	9050
9	0.00	0.00	0.00	64
10	0.85	0.73	0.79	5839
accuracy			0.52	176218
macro avg	0.39	0.38	0.38	176218
weighted avg	0.51	0.52	0.51	176218

Accuracy: 52.32%

Classification Report: KNeighborsClassifier

Classification models were used to predict crime types based on location data, with varying accuracies. The RandomForestClassifier performed best (55.08%), balancing precision and recall for common crimes but struggling with rarer ones. The DecisionTreeClassifier followed closely at 54.96%, showing similar performance. The XGBClassifier achieved 49.62%, excelling in common categories but underperforming for rare crimes, while the KNeighborsClassifier reached 52.32%, effective for common crimes but sensitive to noise. These models highlight the importance of spatial data in understanding crime patterns and the challenge of predicting less frequent crimes.



The data was analyzed using a time-series approach, specifically the Holt-Winters forecasting method, to predict future crime trends in Vancouver. The analysis reveals a promising outcome, as the future crime trend is forecasted to decline significantly, indicating a reduction in crime rates in the coming years. This is a positive sign for the city, suggesting that Vancouver may experience less crime in the future, which could be attributed to effective strategies or societal changes. Maintaining this trend would require continued efforts in crime prevention and proactive measures to sustain the observed decline

2.6 Plan for Reproducible Research

Reproducible research is a system for documenting and publicizing the findings of an impact evaluation. Reproducibility allows other researchers to analyse the same data and obtain the same results as the original study, thus strengthening the original study's conclusions. It is a core principle of scientific inquiry that emphasizes transparency and openness in research procedures and data. It ensures that others can independently validate and expand on study findings. The key components of reproducible research include clear & detailed documentation, accessible data & code, and open-source tools and platforms. It is critical to encourage academics to publish reproducible research since the path to research findings is equally essential as the findings themselves.

Reproducible research brings many benefits to everyone involved in the research industry. When research is replicable, it fosters trust and confidence in the results. Independent verification by additional researchers increases the evidence while lowering the likelihood of bias or errors. This improved trust in study findings is critical for furthering scientific understanding and informing decision-making. Besides, this also speeds up scientific progress by allowing researchers to build on previous work more effectively.Reproducible research encourages collaboration among researchers. By exchanging data and code, academics may easily cooperate on projects, share ideas, and build on each other's efforts. This collaborative setting can result in more original and impactful research findings. Not only that, making research replicable can help discover and address problems. When researchers attempt to duplicate or share their own findings with others, they may come upon flaws or inconsistencies that were previously unknown. This can lead to improved research technique and higher overall study quality.

There are multiple tools and techniques that we have used to ensure the research is reproducible. One of them is the version control systems like Git. They are critical for tracking changes in code and data across time. They enable researchers to simply revert to prior versions, cooperate on code development, and maintain track of all modifications made to study materials. They also help to create a well-structured repository with clear commit messages. Another common practice to produce reproducible research that we have done is to write clear and concise documentation throughout the project in an interactive notebook such as Jupyter Notebook. Both this method is exactly what our team has done for this project as it enables researchers to mix code, data, and visualizations in one document.

This facilitates the dissemination of research findings while also allowing others to interact with the analysis and replicate the results. The notebook is then shared on the GitHub platform for anyone to access. The link to the github and the jupyter notebook can be obtained from the reference section.

In summary, reproducible research aims to create a trustworthy and transparent research ecosystem in which discoveries may be independently validated and expanded upon.

2.7 Deployment of Data Product

To operationalize a machine learning model for predicting criminal activity based on historical data, a data product was developed and deployed. This deployment process involved creating an intuitive web application designed to forecast potential criminal activities based on user-provided geographic inputs in the UTM coordinate system (easting and northing, Zone 10). The application integrates a trained Random Forest model with an accuracy of 55.08% and is built using the Flask web framework.

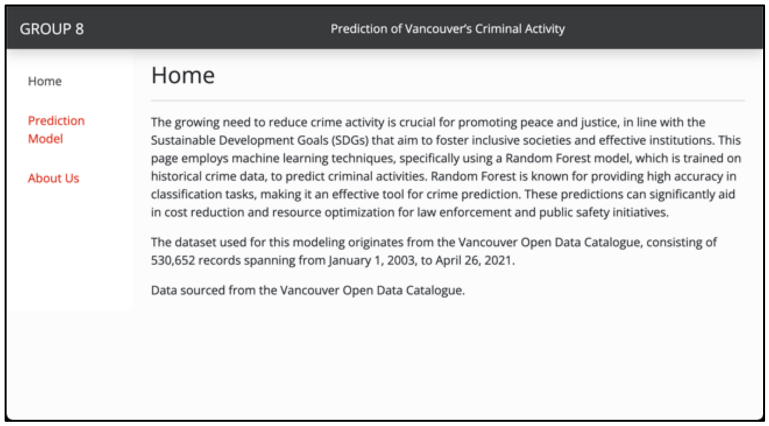
The web application features an interactive input form, allowing users to enter UTM coordinates and select the type of crime. Input validation ensures all required fields are completed accurately. Once inputs are submitted, the Random Forest model processes the data and makes predictions based on historical crime data. The model categorizes outputs into specific crime types, which are dynamically displayed on the webpage, providing users with immediate feedback and actionable insights.

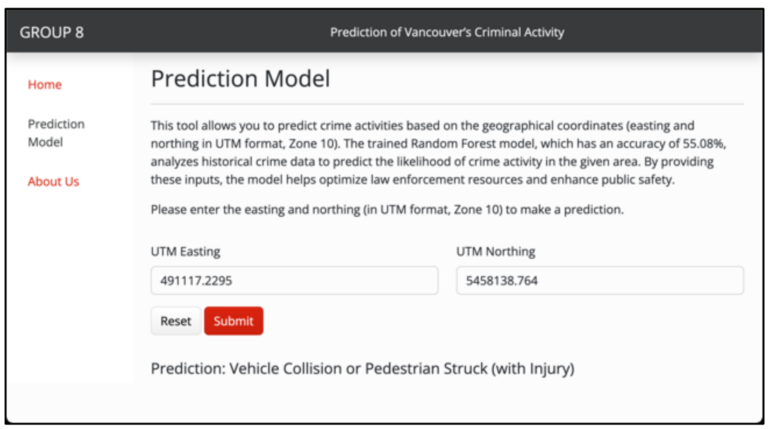
From a technical perspective, the backend of the application is powered by the Flask framework, which handles routing, input processing, and integration with the machine learning model. A Python script preprocesses the user inputs and feeds them into the trained Random Forest model for predictions. For the frontend, Bootstrap 5 was utilized to create a responsive and user-friendly interface, with stylized input forms enhancing usability.

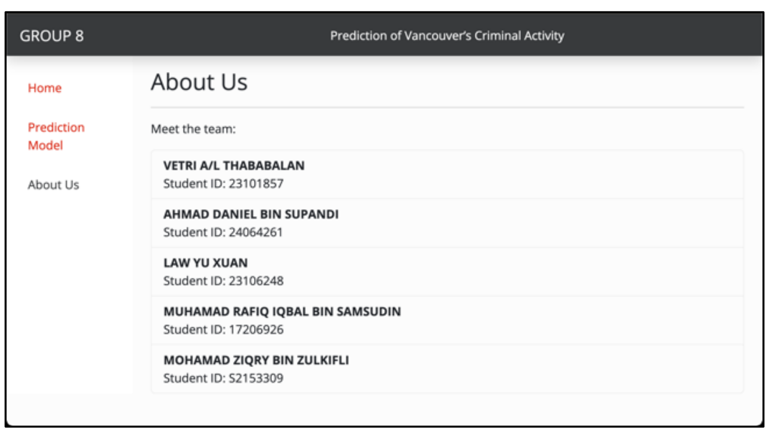
The deployment of the model involved serializing the trained Random Forest model into a `model.pkl` file for efficient loading during runtime. When user inputs are received, the model processes the data and predicts the corresponding crime type. The application is hosted locally for demonstration purposes and can be accessed via `http://127.0.0.1:5000/`.

To enhance accessibility and scalability, the application can be deployed on cloud platforms such as AWS, Azure, or Heroku.

This web application serves as a practical implementation of machine learning for public safety, enabling predictions of potential criminal activities and supporting the optimization of law enforcement resources. Below is the screenshot of the deployed model.







CHAPTER 3: RESULTS AND DISCUSSION

3.1 Insights and Conclusions

The analysis of Vancouver's crime data revealed key insights to support public safety and law enforcement strategies. Theft from vehicles was the most common crime, followed by other theft and mischief. Crime rates generally declined from 2003 to 2010 but rose sharply in 2019 before dropping after 2020, likely due to COVID-19 restrictions limiting public mobility. Temporal analysis showed crime peaked on weekends, especially Fridays and Saturdays, with the highest activity occurring between 6:00 PM and midnight, indicating a need for targeted law enforcement during these times.

Machine learning models were evaluated, with the Random Forest Classifier performing best, achieving 55.08% accuracy with strong precision and recall for high-frequency crime categories. The Decision Tree Classifier and XGBClassifier achieved accuracies of 54.96% and 49.62%, respectively, but struggled with less frequent crimes. The KNeighborsClassifier had moderate performance, with 52.32% accuracy, but was sensitive to data noise, highlighting the importance of addressing data imbalance for better prediction accuracy. Time-series forecasting using the Holt-Winters method predicted a continued decline in crime rates, suggesting a positive outlook for Vancouver's safety. Seasonal trends also showed higher crime rates in the summer months, potentially influenced by environmental factors, offering insights for proactive planning by law enforcement. While the models provided valuable insights, challenges such as data imbalance and predicting less frequent crimes remain. Future research should integrate additional factors like socioeconomic data, weather patterns, and public events to improve prediction accuracy. Including more granular spatial data, such as longitude and latitude, could further enhance the analysis.

In conclusion, this project demonstrates the effectiveness of machine learning and time-series forecasting in crime prediction. By incorporating data-driven approaches, law enforcement can make informed decisions and design targeted interventions, helping

REFERENCES

Catlett, C., Cesario, E., Talia, D., & Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive and Mobile Computing*, *53*, 62-74. https://doi.org/https://doi.org/10.1016/j.pmcj.2019.01.003

Cesario, E., Lindia, P., & Vinci, A. (2024). Multi-density crime predictor: an approach to forecast criminal activities in multi-density crime hotspots. *Journal of Big Data*, *11*(1), 75. https://doi.org/10.1186/s40537-024-00935-4

Hajela, G., Chawla, M., & Rasool, A. (2020). A Clustering Based Hotspot Identification Approach For Crime Prediction. *Procedia Computer Science*, *167*, 1462-1470. https://doi.org/10.1016/j.procs.2020.03.357

Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M., & Sarker, I. H. (2020, 2020//). Crime Prediction Using Spatio-Temporal Data. Computing Science, Communication and Security, Singapore.

Khan, M., Ali, A., & Alharbi, Y. (2022). Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques. *Complexity*, *2022*, 1-13. https://doi.org/10.1155/2022/4830411

Patterson, E. (2018). Poverty, income inequality, and community crime rates. *Criminology*, *29*, 755-776. https://doi.org/10.1111/j.1745-9125.1991.tb01087.x

Schleimer, J. P., Pear, V. A., McCort, C. D., Shev, A. B., De Biasi, A., Tomsich, E., Buggs, S., Laqueur, H. S., & Wintemute, G. J. (2022). Unemployment and Crime in US Cities During the Coronavirus Pandemic. *Journal of Urban Health*, *99*(1), 82-91. https://doi.org/10.1007/s11524-021-00605-3

Sundaram, S., Perumal, S., Parthiban, D., & Revathy, D. (2024). Crime Hotspot Analysis Using Six Layered Deep Recurrent Neural Networks 1. 2022. https://doi.org/10.48047/intjecse/V14I8.309

Venter, Z. S., Shackleton, C., Faull, A., Lancaster, L., Breetzke, G., & Edelstein, I. (2022). Is green space associated with reduced crime? A national-scale study from the Global South. *Science of The Total Environment*, *825*, 154005. https://doi.org/https://doi.org/10.1016/j.scitotenv.2022.154005

Yu, C. H., Ward, M. W., Morabito, M., & Ding, W. (2019, 11-11 Dec. 2011). Crime Forecasting Using Data Mining Techniques. 2011 IEEE 11th International Conference on Data Mining Workshops,

APPENDIX 1 - Critical Analysis Table

Reference	Objective	Method and Algorithm (Accuracy Achieve)	Observation	Limitation
(Khan et al., 2022)	Study proposes a crime classification model by analyzing and comparing three known prediction classification algorithms.	Naïve Bayes (65.82%) Random Forest (63.43%) Gradient Boosting (98.5%)	This study categorized their crime as violent and not violent type of crime. The classification focuses on **location and not toward time analysis.**	More temporal analysis can be perform to determine the number and intensity of criminal activity using **time series analysis**
(Sundaram et al., 2024)	**Crime Rate forecasting** based on given location using Recurent Neural Networks. Crime are combination of Vehicle, Residential Theft.	RNN (89 %)	The RNN consist of six dense layer . Predict crime rate in term of location.	It does not classify the **type of crime** happening and. Futher improvement can be done through investigate time feature more deeply.
(Hajela et al., 2020)	Spatiotemporal **crime hotspot classification technique** based on machine learning	K-Means Clustering (85 % -87%)	It is concluded that crime shows a geographical pattern in space and time. It is shown that hotspot can improve crime prediction accuracy	Other approach of analysis **time-series** where smartly split it into training validation can be propose.
(Hossain et al., 2020)	To use supervised machine learning to forecast crime rate base on **location and time.**	Decision-Tree (31.17%) KNN (28.50%) Random Forest (73.89%)	Use feature extraction to divide time **into early, late morning, afternoon and night.** Its able to predict crime rate base on hour of day and the segmentation techniques use.	The location type use in this paper are area of which police department crime was reported.
(Yu et al., 2019)	Develop a reliable method to classify crime hotspots, specifically **residential burglaries**, using data mining techniques.	SVM (84.38 %)	The model only **using 10 months of datasets.** The challenge that they indicated that we may face in the future could be to locate the best point.	The prediction can be improves if specific location data **(longitude and latitude)** is available instead of area
(Catlett et al., 2019**)**	Develop a model to detech high-risk crime regions in urban areas and to reliably **forecast the crime trends** in each regions specifically in Chicago.	Classification Model DBSCAN. Forecasting Model (ARIMA= 81.34).	The model use dataset that consist crime from 2001 to 2016 in Chicago	The prediction models only detect high risk area and deal with crime as **general type of crime**.
(Ingilevich & Ivanov, 2018)	The aim is to compare different approaches of regression to forecast robbery crime rate in different areas of the city of Saint Petersburg, Rusia	Linear Regression (R^2= 0.9) Gradient Boosting(R^2= 0.9)	The study reveals the best accuracy to predict crime rate was by gradient boosting model.	To avoid model overfitting, feature selection techniques could be use.