

FACULTY OF SCIENCE & TECHNOLOGY

MSc Data Science and Artificial Intelligence

MAY 2024

Forecasting Criminal Activity: Anticipating Future Incidents Through Predictive Analysis

by

Hanina Kombanezhuth Abdulrahman

Faculty of Science & Technology

Department of Computing and Informatics

Individual Masters Project

**Abstract**

Urban law enforcement agencies encounter formidable challenges in predicting and preventing crimes stemming from resource constraints and the dynamic nature of criminal activities. Leveraging predictive analytics techniques and historical crime data presents an opportunity to address these challenges effectively. This project focuses on developing and evaluating machine learning models for crime prediction, aiming to enhance public safety in urban areas.

Our project employs a diverse range of machine learning algorithms, including Decision Trees, Random Forests, Logistic Regression, K-nearest neighbours, Neural Networks, and LSTM-based Recurrent Neural Networks. By analysing patterns of past incidents, such as crime types, locations, and occurrence times, these models enable the proactive deployment of law enforcement resources to deter criminal activities.

One of the critical challenges in utilising historical crime data is the presence of class imbalance within crime datasets. To mitigate this issue, our project employs the Random Over Sampler technique. This balances the distribution of different crime types in the training data, thereby improving the fairness and performance of the predictive models. Furthermore, we rigorously evaluate the performance of the developed models using metrics such as accuracy, precision, recall, and F1 score. The most effective predictive models for various crime prediction scenarios are identified through systematic comparison and analysis. This empowers law enforcement agencies to make informed decisions regarding resource allocation and intervention strategies.

Overall, this project contributes to advancing the field of crime prediction by developing reliable and fair predictive models, thereby enhancing public safety and well-being in urban areas.

|  |
| --- |
| **Dissertation Declaration**  I agree that, should the University wish to retain it for reference purposes, a copy of my dissertation may be held by Bournemouth University normally for a period of 3 academic years. I understand that once the retention period has expired my dissertation will be destroyed.    **Confidentiality**    I confirm that this dissertation does not contain information of a commercial or confidential nature or include personal information other than that which would normally be in the public domain unless the relevant permissions have been obtained. Any information which identifies a particular individual's religious or political beliefs, information relating to their health, ethnicity, criminal history, or sex life has been anonymised unless permission has been granted for its publication from the person to whom it relates.    **Copyright**    The copyright for this dissertation remains with me.    **Requests for Information**    I agree that this dissertation may be made available as the result of a request for information under the Freedom of Information Act.    **Signed**: Hanina Kombanezhuth AbdulRahman .    Name: Hanina Kombanezhuth AbdulRahman    Date: 17/05/2024    Programme: MSc Data Science and Artificial Intelligence |

**Original Work Declaration**

This dissertation and the project that it is based on are my own work, except where stated, in accordance with University regulations.

**Signed**: Hanina Kombanezhuth AbdulRahman .

Name: Hanina Kombanezhuth AbdulRahman

Date: 17/05/2024

**Acknowledgments**

First, I would like to express my heartfelt gratitude to my Project Supervisor, Prof. Dr. Yan Gong, for the continuous support, guidance, motivation, and valuable feedback provided at crucial times. Your insights and suggestions were instrumental in the successful completion of this project.

I would also like to thank all the lecturers who imparted their knowledge and guidance during the initial semesters. Their teachings laid a solid foundation that significantly contributed to the completion of this project.

Lastly, I am deeply thankful to my family and friends for their unwavering support and encouragement throughout this journey. Your belief in me has been a constant source of strength.

TABLE OF CONTENTS

[1 INTRODUCTION 1](#_Toc166834213)

[1.1 Problem Definition 2](#_Toc166834214)

[1.2 Aims and Objectives 2](#_Toc166834215)

[1.2.1 Aims 2](#_Toc166834216)

[1.2.2 Objectives 3](#_Toc166834217)

[1.3 Research Questions 3](#_Toc166834218)

[2 LITERATURE REVIEW 5](#_Toc166834219)

[2.1 Machine Learning in Crime Forecasting 5](#_Toc166834220)

[2.2 Deep Learning in Crime Prediction 6](#_Toc166834221)

[3 METHODOLOGY 7](#_Toc166834222)

[3.1 Data Collection 7](#_Toc166834223)

[3.2 Data Preprocessing 8](#_Toc166834224)

[3.2.1 Oversampling 8](#_Toc166834225)

[3.2.2 Scaling Features 8](#_Toc166834226)

[3.2.3 Encoding 9](#_Toc166834227)

[3.3 Data Splitting 9](#_Toc166834228)

[3.4 Model Selection 10](#_Toc166834229)

[3.4.1 Decision Tree Classifier 10](#_Toc166834230)

[3.4.2 Random Forest Classifier 11](#_Toc166834231)

[3.4.3 Logistic Regression 11](#_Toc166834232)

[3.4.4 K-Nearest Neighbour 12](#_Toc166834233)

[3.4.5 Neural Network 12](#_Toc166834234)

[3.4.6 Recurrent Neural Network 12](#_Toc166834235)

[3.5 Training and Testing 13](#_Toc166834236)

[3.6 Optimal Model Selection 13](#_Toc166834237)

[4 EVALUATION AND RESULTS 15](#_Toc166834238)

[4.1 Evaluation Metrics 15](#_Toc166834239)

[4.1.1 Accuracy 15](#_Toc166834240)

[4.1.2 Precision 15](#_Toc166834241)

[4.1.3 Recall 15](#_Toc166834242)

[4.1.4 F1 Score 15](#_Toc166834243)

[4.1.5 Confusion Matrix 15](#_Toc166834244)

[4.2 Dataset Analysis 16](#_Toc166834245)

[4.3 Prediction Results 19](#_Toc166834246)

[5 DISCUSSION 24](#_Toc166834247)

[6 FUTURE WORKS 26](#_Toc166834248)

[7 CONCLUSION 27](#_Toc166834249)

[REFERENCES 28](#_Toc166834250)

[APPENDIX A : PROJECT PROPOSAL 29](#_Toc166834251)

[APPENDIX B: RESEARCH ETHICS CHECKLIST 32](#_Toc166834252)

[APPENDIX C: 34](#_Toc166834253)

LIST OF FIGURES

[Figure 1: Flowchart of the processes in the methodology 7](#_Toc166834581)

[Figure 2: Flowchart of Decision Tree 10](#_Toc166834582)

[Figure 3 : Crime occurrence over the years. 16](#_Toc166834583)

[Figure 4 : Types of figures in different Neighbourhood 17](#_Toc166834584)

[Figure 5 : Analysis of crime over the years 17](#_Toc166834585)

[Figure 6 : Plot Showing Crime Incident by Hour 18](#_Toc166834586)

[Figure 7 : Crime occurrence over Month 18](#_Toc166834587)

[Figure 8 : Confusion Matrix of Decision Tree 19](#_Toc166834588)

[Figure 9 : Confusion Matrix of Random Forest 20](#_Toc166834589)

[Figure 10 : Confusion Matrix of Logistic Regression 21](#_Toc166834590)

[Figure 11 : Confusion Matrix of KNN 22](#_Toc166834591)

[Figure 12: Confusion Matrix of Neural Network 22](#_Toc166834592)

[Figure 13 : Comparison of all algorithm 23](#_Toc166834593)

LIST OF TABLES

[Table 1 : Explanation of variables used in Dataset 8](#_Toc166834666)

[Table 2 : Train-Test Dataset 9](#_Toc166834667)

[Table 3 : Table showing performance of Decision Tree 19](#_Toc166834668)

[Table 4 : Table showing Performance of Random Forest 20](#_Toc166834669)

[Table 5 : Table showing the Performance of Logistic Regression 21](#_Toc166834670)

[Table 6 : Performance of KNN 21](#_Toc166834671)

[Table 7 : Performance of Neural Network 21](#_Toc166834672)

[Table 8 : Performance Of RNN 22](#_Toc166834673)

[Table 9 : Accuracy of previously calculated model 24](#_Toc166834674)

[Table 10 : Accuracy of newly calculated model 25](#_Toc166834675)

# INTRODUCTION

Crime is a persistent and constantly evolving societal challenge whose complexity makes it difficult to decipher the underlying patterns and behaviours (Bogomolov et al. 2014). Ensuring community safety is a top priority, prompting governments to take proactive measures to lower crime rates. Such actions are essential for fostering economic growth and improving the quality of life. Crime analysis, an integral component of criminology, involves examining behavioural patterns to pinpoint indicators of criminal activity. However, crime prevention is challenging due to the wide range of crime types, motives, consequences, handling approaches, and prevention strategies (Dakalbab et al. 2022). Crime, which is categorised into different types such as kidnapping, theft, murder and rape, poses a significant threat to public safety. Law enforcement agencies use information technologies to collect crime data, but crime prediction remains daunting. The emergence of advanced analytical tools, especially machine learning and deep learning methods, offers a promising way to fill gaps in existing detection mechanisms (Kim et al. 2018).

Machine learning, in which computers make autonomous decisions, is used in various fields, including crime prediction. Using machine learning techniques enables the analysis of high-dimensional and complex crime data, the detection of patterns, the prediction of future trends and the visualisation of findings. Traditional crime analysis methods are inadequate for such complex data, necessitating the development of sophisticated tools for predicting and analysing crime. Deep Learning (DL) is a branch of machine learning that mimics how the human brain operates. This method utilises artificial neural networks with multiple layers and various types, such as pooling, convolutional, fully connected, and dropout layers, to emulate brain functionality (Jenga et al. 2023).

Vancouver is one of Canada's most populous, ethnically diverse, multicultural urban centres. In 2017, the overall crime rate in Vancouver decreased by 1.5%, yet vehicle break-ins and theft remained significant concerns. To address this, the Vancouver Police Department (VPD) recently implemented a crime predictive model specifically targeting property break-ins. Following the introduction of this model, the city saw a 27% reduction in residential break-ins (Kim et al., 2018).

This paper explores various methodologies to effectively forecast the incidence of specific crimes within defined temporal and spatial contexts in Vancouver. Utilising classification techniques such as decision trees, random forests, logistic regression, K-nearest neighbours, and neural networks, our study focuses on feature extraction and trend analysis using crime data analogous to ours. The aim is to enhance crime prediction accuracy, thereby enabling proactive crime prevention and law enforcement strategies.

## Problem Definition

In urban areas, law enforcement agencies face significant challenges in effectively predicting and preventing crimes due to limited resources and the unpredictable nature of criminal activities. However, leveraging predictive analytics techniques and historical crime data presents an opportunity to address these challenges. Predictive models can forecast potential crime hotspots and times by analysing patterns of past incidents—such as the types of crimes committed, their locations, and the times at which they occur. This enables law enforcement to deploy officers proactively to these areas, deterring criminal activities and enhancing public safety.

Another challenge in utilising historical crime data for predictive analytics is the presence of class imbalance within crime datasets. Certain crime s may be significantly less frequent than others, leading to imbalanced data distributions. This imbalance can hinder the effectiveness of machine learning models, causing biased predictions and poor performance.

Furthermore, the mere construction of predictive models is insufficient in crime prediction. It is essential to rigorously evaluate their performance to determine their reliability and suitability for real-world deployment. Metrics such as accuracy, precision, recall, and F1 score provide valuable insights into how well the models perform across different crime types and prediction scenarios. Systematically evaluating and comparing the performance of various machine learning models allows law enforcement agencies to identify the most effective tools for crime prediction. This empowers them to make informed decisions and allocate resources efficiently to prevent and combat crime, ultimately enhancing public safety and well-being in urban areas.

## Aims And Objectives

### Aims

The project primarily uses machine learning models for crime prediction, including Decision Trees, Random Forests, Logistic Regression, K K-nearest neighbours, Neural Networks, and an LSTM-based Recurrent Neural Network. By training these models on historical crime data, the goal is to forecast future occurrences of specific types of crime. Through this aim, the project seeks to provide law enforcement agencies with valuable insights into potential crime hotspots and times, enabling proactive measures to prevent and combat criminal activities.

Another aim of the project is to tackle the class imbalance challenge in the crime dataset. By employing the Random Over Sampler technique, the project aims to balance the distribution of different crime types in the training data. This ensures that each crime type is adequately represented, improving the fairness and performance of the machine learning models trained on the dataset. Addressing class imbalance is crucial for building reliable predictive models that can accurately generalise to real-world crime scenarios.

The project aims to evaluate the performance of the trained machine learning models using various metrics such as accuracy, precision, recall, and F1 score. This evaluation process helps assess the effectiveness of each model in predicting different types of crime. By comparing the performance of these models, the project aims to identify the most effective predictive models for various crime prediction scenarios. This aim ensures that law enforcement agencies can make informed decisions regarding resource allocation and proactive intervention strategies based on the reliability and suitability of the predictive models.

### Objectives

The project's first objective revolves around developing robust predictive models for crime prediction. This involves gathering and preprocessing historical crime data to ensure its quality and consistency. Various machine learning algorithms, such as Decision Trees, Random Forests, Logistic Regression, K K-nearest neighbours, Neural Networks, and LSTM-based Recurrent Neural Networks, are implemented and trained on the pre-processed data. Cross-validation techniques are utilised to optimise model hyperparameters and prevent overfitting. Subsequently, the performance of each model is evaluated using metrics like accuracy, precision, recall, and F1 score. The project aims to identify the most effective predictive models tailored to different crime prediction scenarios based on these evaluations.

The second objective focuses on addressing class imbalance within the crime dataset. Analysing the distribution of different crime types reveals the extent of class imbalance, where certain crimes are significantly less frequent than others. To mitigate this issue, the RandomOverSampler technique is applied to balance the distribution of crime types in the training data. The effectiveness of class imbalance mitigation is verified by comparing the distribution of classes before and after resampling. Furthermore, the impact of class imbalance handling on the performance of predictive models is systematically assessed through evaluation.

Lastly, the project aims to evaluate the performance of predictive models comprehensively. A robust evaluation framework is developed to assess the predictive capabilities of each model. Metrics such as accuracy, precision, recall, and F1 score are calculated to gauge model performance across different crime prediction scenarios. The strengths and weaknesses of various

predictive models are identified through comparative analyses. The project provides actionable insights and recommendations based on performance evaluations to guide decision-making processes within law enforcement agencies. Additionally, continuous monitoring and refinement of predictive models based on real-world feedback and evolving crime patterns ensure their relevance and effectiveness over time.

## Research Questions

1. What is the main objective of crime prediction research?
2. How effectively is the crime prediction model processing and forecasting criminal incidents using Machine learning techniques?
3. what are the advantages and drawbacks of the proposed approach?
4. What evaluation criteria are used to assess the performance of the machine learning models?

# literature Review

Much research has been conducted to reduce crime, and numerous crime prediction algorithms have been developed. The kind of data utilised, and the qualities chosen for prediction affect how accurate the forecast is.

## Machine Learning in Crime Forecasting

Traditional machine learning models have demonstrated effectiveness in crime prediction. A range of models, including decision trees, support vector machines, logistic regression, and random forests, have been employed to analyse crime data and discern patterns useful for predicting criminal activity.

(Kim et al., 2018) conducted a study on utilising machine learning (ML) techniques to predict crime based on crime data spanning the last 15 years in Vancouver, Canada. The ML-based crime analysis encompassed data collection, classification, pattern identification, prediction, and visualisation. Additionally, the study implemented K-nearest neighbour (KNN) and boosted decision tree algorithms to analyse the crime dataset. The researchers examined a dataset comprising 560,000 crime records from 2003 to 2018. Their study explored the effectiveness of K-nearest neighbour (KNN) and boosted decision tree algorithms for crime prediction tasks. Although the predictive accuracy achieved by these machine learning models ranged from 39% to 44%, which is relatively modest, the researchers proposed that further improvements could be attained by refining both the algorithms and the crime data, tailoring them to the specific requirements of various applications.

(Bandekar and Vijayalakshmi 2020) Designed and analysed machine learning (ML) algorithms to mitigate crime rates in India. They applied ML techniques to a substantial dataset to identify patterns and relationships among various factors. The study centred on forecasting future crimes by leveraging the spatial data of past incidents. Researchers examined the dataset using Bayesian neural networks, the Levenberg-Marquardt algorithm, and a scaled algorithm. Notably, the scaled algorithm outperformed the other methods. Statistical techniques such as correlation analysis, analysis of variance, and graphical representations were employed to assess its efficacy. Findings revealed a significant reduction in crime rates by 78%, implying an accuracy level of 0.78.

In a study done by (Lin et al. 2018), they focused on developing a crime prediction model employing the decision tree algorithm, specifically J48. Recognised for its efficacy in law enforcement and intelligence analysis contexts, J48 shows promise in reducing crime rates. It is regarded as one of the most efficient ML algorithms for crime prediction. The J48 classifier was implemented using the WEKA toolkit and trained on a pre-processed crime dataset. Experimental results demonstrated that the J48 algorithm accurately predicted unknown crime categories with an impressive accuracy of 94.25287%. Such high accuracy instils confidence in the system's reliability for future crime predictions.

## Deep learning in Crime Prediction

Deep learning approaches rely on complex neural networks and extensive datasets for crime prediction. These methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), excel at learning intricate patterns and relationships within data. Deep learning models can automatically extract features from raw data, eliminating the need for manual feature engineering. However, they often require substantial computational resources for training and may be more challenging to interpret than traditional machine-learning models. Despite these challenges, deep learning has shown promising results in various domains, including crime prediction, by effectively leveraging its ability to handle large-scale and high-dimensional data.

(Kang and Kang 2017) a feature-level data fusion technique based on a deep neural network (DNN) is introduced to predict crime occurrences accurately by effectively integrating multi-model data across various domains and environmental context information. The research primarily aimed to predict forthcoming crimes by analysing the spatial patterns of prior incidents. Various techniques were deployed for data analysis, including Bayesian neural networks, the Levenberg-Marquardt algorithm, and a scaled algorithm. Notably, the scaled algorithm exhibited superior performance compared to the others. Statistical evaluations, such as correlation analysis, analysis of variance, and graphical representations, were employed to gauge its effectiveness. Comparative evaluations reveal that while SVM and KDE models achieved 67.01% and 66.33% accuracy, respectively, the proposed DNN model exhibits remarkable accuracy, reaching 84.25%.

The (Poppe 2007 )authors explored the feasibility of crime mapping using satellite imagery as an alternative to manual data collection, which is costly and time-consuming. They investigated the use of deep learning techniques to predict crime rates directly from raw satellite images. Their research entailed training a deep Convolutional Neural Network (CNN) using satellite images from over 1 million crime incident reports spanning 15 years provided by the Chicago Police Department. The best-performing model achieved a remarkable 79% accuracy in forecasting crime rates based solely on raw satellite images. To ensure the reliability of their findings, they conducted a reusability assessment by applying the trained models developed for Chicago to predict crime rates in Denver and San Francisco. The results, when compared to maps generated from years of data collected by the respective police departments, exhibited 72% and 70% accuracy, respectively.

# METHODOLOGY

Predictive modelling has been utilised to forecast outcomes, leveraging its capacity to construct models and generate predictions. This methodology encompasses a range of algorithms from both machine learning and deep learning domains, enabling the exploration of data properties during the training phase to facilitate prediction generation. These algorithms include Random Forest, Decision Tree Classifier, Logistic Regression, K Nearest Neighbour, Neural Network, and Recurrent Neural Network (RNN).

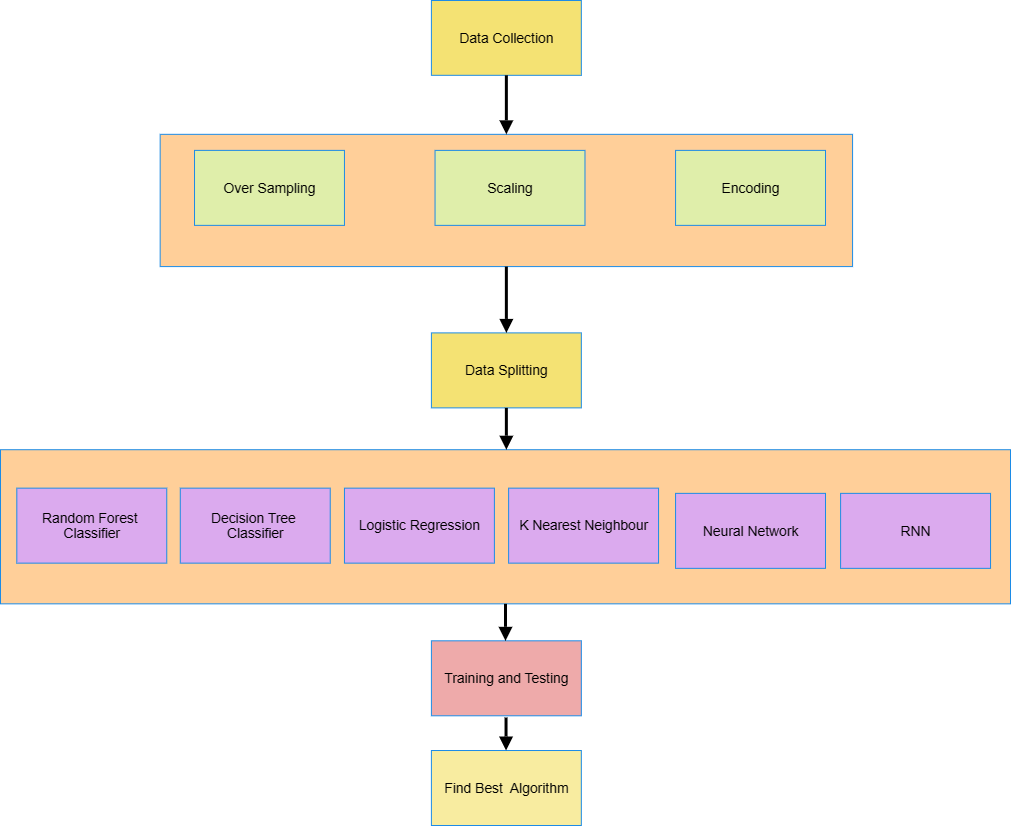


Figure 1: Flowchart of the processes in the methodology

## Data Collection

The crime dataset from Vancouver was sourced from Kaggle [Crime in Vancouver (kaggle.com)](https://www.kaggle.com/datasets/wosaku/crime-in-vancouver), which furnishes details regarding the nature of crimes committed along with the corresponding timestamps and locations. For this project, data from 2003 to 2017 comprised 530,652 records. Each record in our dataset corresponds to a specific crime occurrence and includes features outlined in Table I.

|  |  |
| --- | --- |
| Variables | Explanation |
| TYPE | the type of crime that occurred |
| YEAR | the year in which the crime took place |
| MONTH | month of the year in which the crime occurred, ranging from 1 to 12, representing January through December. |
| DAY | The day of the month when the crime occurred. It ranges from 1 to 31. |
| HOUR | The hour of the day when the crime occurred, ranging from 0 to 23, represents each hour of a 24-hour day. |
| MINUTE | The minute of the hour when the crime occurred ranged from 0 to 59. |
| HUNDRED\_BLOCK | the block or street address where the crime occurred |
| NEIGHBOURHOOD | denotes the neighbourhood or locality where the crime occurred. |
| X | represents the X-coordinate or longitude of the crime's location, used in Cartesian coordinate systems |
| Y | denotes the Y-coordinate or latitude of the location where the crime occurred. |
| Latitude | Geographic coordinates are measured in degrees. |
| Longitude | Geographic coordinates are measured in degrees. |

Table 1 : Explanation of variables used in Dataset

## Data Preprocessing

Data pre-processing involves methods to remove infinite or null values from data that might affect the model's performance. This step converts the data set into an understandable format by handling class imbalance using oversampling, scaling numerical features, and encoding categorical features.

### Oversampling

Random Oversampling increases the number of instances in the minority class by randomly duplicating existing instances until the class distribution is balanced. Class imbalance is when one class has significantly fewer instances than other classes. In this dataset, it’s a type of crime. Applying Random Oversampling creates a more balanced dataset where every kind of crime has a similar number of instances, allowing the model to learn effectively from all classes and make more accurate predictions.

### Scaling Features

Feature scaling, particularly standardisation using the Standard Scaler, is applied to numerical features in the dataset to ensure that all features contribute equally to the model training process. It ensures that numerical features like "X", "Y", "Latitude", and "Longitude" contribute equally to the model training process. Scaling these features with a mean of 0 and a standard deviation 1 normalises their range, preventing larger-scale features from dominating the learning process. This normalisation eliminates biases towards features with larger numerical values and makes them suitable for training machine learning models that require standardised input features.

### Encoding

Encoding categorical variables is crucial in preparing the data for machine learning models. Certain features like "NEIGHBOURHOOD" and "HUNDRED\_BLOCK" contain categorical information in our dataset. To achieve this, we employed two encoding techniques; one-hot encoding was applied to the "NEIGHBOURHOOD" column, creating binary vectors to represent each category independently. This approach ensures that the model does not misinterpret the categorical data as having an ordinal relationship. Additionally, label encoding was used for the "HUNDRED\_BLOCK" column, assigning a unique integer to each category. These encoding methods enable the machine learning algorithms to interpret and utilise the categorical data accurately during training, ultimately leading to more reliable predictions.

## Data Splitting

After the data cleaning and pre-processing stages, the dataset is partitioned into training and testing sets using an 80:20 ratio. This means that 80% of the data is allocated for training the models, while the remaining 20% is reserved for evaluating the model's performance. Table 1 explains how the dataset is divided or partitioned for use in the study. A consistent ratio is maintained across all models developed in this study.

|  |  |
| --- | --- |
| Train Dataset | 1,230,400 |
| Test Dataset | 307,601 |

Table 2 : Train-Test Dataset

Python's sci-kit-learn library offers a convenient 'test-train split' function, which facilitates the partitioning of the dataset into training and testing subsets. This function ensures that the split is randomised, helping to prevent any bias in the training or testing data. By leveraging this functionality, the dataset can be efficiently divided for training and evaluating the models' performance.

## Model Selection

This process trains different machine learning models, including a Decision Tree, Random Forest, Logistic Regression, K Nearest Neighbours, Neural Network and Recurrent Neural Network.

### Decision Tree Classifier

Decision trees are practical tools in machine learning used to classify and predict outcomes. They resemble trees, with each branch representing a test on a feature, and the endpoints hold labels for different classes. Figure 2 illustrates the functioning of a decision tree. Decision trees help assess the probability of a record belonging to a category or assigning records to the correct class.

Figure 2: Flowchart of Decision Tree

In our model, we train the Decision Tree Classifier on the training data (Xtrain, ytrain). During training, the algorithm learns patterns and relationships in the data, constructing a tree-based model to represent these patterns. Subsequently, we evaluate the model's performance using the testing data (Xtest, ytest) and metrics such as accuracy, precision, recall, and F1-score. The Decision Tree algorithm recursively partitions the feature space, creating decision rules at each node to optimise classification. This enables the model to predict by traversing the tree from the root node to the leaf nodes, where final class labels are assigned.

### Random Forest Classifier

Random Forest Classifier is a popular machine learning algorithm that constructs multiple decision trees during training and outputs the class based on the mode of the classes predicted by individual trees. It creates many decision trees and merges them to achieve a more accurate and stable prediction. Each tree is built from a sample drawn with replacement from the training set (bootstrapping), and at each node, only a random subset of features is considered for splitting. This randomness and diversity among the trees contribute to improved generalisation performance and reduced overfitting compared to logistic regression.

In our model, the algorithm constructs many decision trees during training, each utilising a random subset of the training data and features. These decision trees collectively form a "forest" and collaborate to make predictions. Once trained, the model undergoes evaluation using the testing data to gauge its performance metrics like accuracy, precision, recall, and F1-score. By aggregating the predictions of individual trees, the Random Forest algorithm produces more robust and accurate predictions than a single decision tree model. Using ensemble learning principles and decision trees, our model effectively forecasts crime outcomes based on the input features.

### Logistic Regression

Logistic Regression is a statistical method mainly used for binary classification tasks, predicting between two possible outcomes. It calculates the probability of an input belonging to one of the classes based on its features. Despite its name, it's not used for regression but for classification. During training, the model learns coefficients for each feature, indicating their influence on the class probability. These coefficients are adjusted to minimise the difference between predicted probabilities and actual labels. Logistic Regression is valued for its simplicity and efficiency, especially when dealing with linear relationships in data. However, it may struggle with nonlinear data or irrelevant features.

In our model, the algorithm learns the relationships between the input features and the target variable by estimating the probabilities of different classes. Logistic Regression models the likelihood that a given input belongs to a particular class using a logistic function, which ensures that the output is between 0 and 1. Once trained, the model is evaluated using the testing data to assess its performance metrics. Logistic Regression is particularly useful for binary classification tasks, where it estimates the probability that an instance belongs to a specific class, making it suitable for predicting categorical outcomes like crime occurrences in our dataset. By fitting a logistic curve to the data, Logistic Regression provides insights into the likelihood of different outcomes, aiding the classification and prediction process.

### K-Nearest Neighbour

K Nearest Neighbours (KNN) is a method where the classification or prediction of a new data point is based on the majority vote of its neighbours. It's simple to understand: if most of its neighbours belong to a specific class, then the new data point will likely belong to that class. However, the downside is that it can be slow with large datasets and sensitive to the choice of neighbours. It's a good starting point for many classification tasks due to its simplicity and easy implementation.

In our model, KNN learns from the training data by storing all instances after dividing the data into training and testing sets. When predicting outcomes for new data, KNN finds the closest neighbours based on distance and assigns the most common class among them as the prediction. This straightforward method is effective for classification tasks without needing a complex model. We evaluate KNN's performance using testing data to understand its accuracy, precision, recall, and F1 score. KNN offers a simple yet powerful way to predict categorical outcomes, making it a key component of our crime prediction model.

### Neural Network

A neural network is a computational model inspired by the structure and function of the human brain. It consists of interconnected nodes, called neurons, organised into layers. Each neuron receives input signals, processes them using an activation function, and passes the output to the next layer. Neural networks can learn complex patterns and relationships in data through training, where the network adjusts its weights based on the input data and desired outputs. They are widely used for classification, regression, and pattern recognition tasks. Neural networks offer flexibility and can handle structured and unstructured data, making them suitable for various machine learning and artificial intelligence applications.

Initially, the dataset undergoes preprocessing and is split into training and testing subsets. Subsequently, the NN model is trained on the training data, where it learns the underlying patterns and relationships in the data. This learning process involves adjusting the weights and biases of interconnected nodes, allowing the NN to capture complex patterns in the input data. Once trained, the model is evaluated using the testing data to assess its performance metrics. NNs excel in handling nonlinear relationships and capturing intricate patterns in data, making them well-suited for various classification tasks. The NN is crucial in accurately predicting crime outcomes in our model by leveraging its ability to learn from vast amounts of data and discern intricate patterns

### Recurrent Neural Network

Recurrent Neural Network is a neural network designed to process sequential data by retaining information in memory. Unlike traditional feedforward neural networks, which process input data in a single pass, RNNs have loops that allow information to persist over time. This loop structure enables RNNs to effectively model sequences, making them suitable for tasks like language translation, time series prediction, and speech recognition. One of the critical features of RNNs is their ability to handle variable-length input sequences, making them versatile for processing data with temporal dependencies. However, RNNs can suffer from the vanishing gradient problem, where gradients become increasingly small during training, affecting the model's ability to learn long-range dependencies. Despite this limitation, RNNs have been widely adopted in natural language processing, speech recognition, and other sequential data analysis tasks.

The Recurrent Neural Network (RNN) is a specialised architecture for handling sequential data in our model. Where it learns to capture temporal dependencies and patterns in the sequential input data. This training process involves iterating over data sequences, updating internal state representations, and adjusting model parameters to minimise prediction errors. Once trained, the RNN is evaluated using the testing data (Xtest, ytest) to assess its performance metrics such as accuracy, precision, recall, and F1-score. RNNs are particularly effective in analysing time series data and sequences, making them well-suited for natural language processing, speech recognition, and sequential prediction tasks like crime outcome prediction in our model. The RNN can effectively capture long-term dependencies and temporal patterns in sequential data through its recurrent connections and memory cells, enabling accurate predictions in our crime prediction model.

## Training and Testing

The model is first trained using the provided training data. The model learns from the training samples during training to minimise errors and improve accuracy. Once trained, the model's performance is evaluated based on its ability to predict outcomes on unseen data accurately. This evaluation involves assessing various performance parameters specific to the problem at hand. We train models such as Decision Trees, Random Forests, Logistic Regression, and others using the training data in our code. After training, we evaluate each model's performance by calculating metrics like accuracy, precision, recall, and F1 score on a separate test dataset. This process helps us determine how well each model generalises to new, unseen data.

## Optimal Model Selection

After training each model, including Decision Tree Classifier, Random Forest, Logistic Regression, K Nearest Neighbour, Neural Network, and Recurrent Neural Network (RNN), their performance metrics, such as accuracy, precision, recall, and F1-score, are compared. The model exhibiting the highest performance across these metrics best predicts crime outcomes in the given dataset. This rigorous selection process ensures that the chosen model effectively captures the underlying patterns and relationships in the data, leading to accurate predictions on unseen instances.

# EVALUATION AND RESULTS

## Evaluation Metrics

Several metrics are commonly employed to evaluate classification models used for prediction and classification.

### Accuracy

This measure measures the accuracy of the model's predictions and is calculated as the ratio of correctly classified instances to the total number of cases.

### Precision

Precision Indicates the proportion of correctly predicted positive instances out of all the cases predicted as positive. It focuses on the relevance of positive predictions.

### Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances out of all actual positive instances. It emphasises the model's ability to capture positive cases.

### F1 Score

The harmonic mean of precision and recall balances the two metrics. It is useful when there is an uneven class distribution or when false positives and negatives must be considered.

### Confusion Matrix

The confusion matrix is a fundamental tool in evaluating the performance of a classification algorithm. It provides a comprehensive summary of the model's predictions compared to the ground truth across different classes. The matrix consists of four sections: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Each row represents the actual class, while each column represents the predicted class. This allows us to visually inspect how well the model performs correctly and incorrectly classified instances for each class.

## Dataset Analysis

Upon delving into the dataset, our analysis reveals a consistent trend: Figure 3 shows that Theft from Vehicles is the most prevalent type of crime, maintaining a concerning occurrence level across the years. This enduring presence underscores the need for targeted interventions and proactive measures to address this persistent issue effectively.

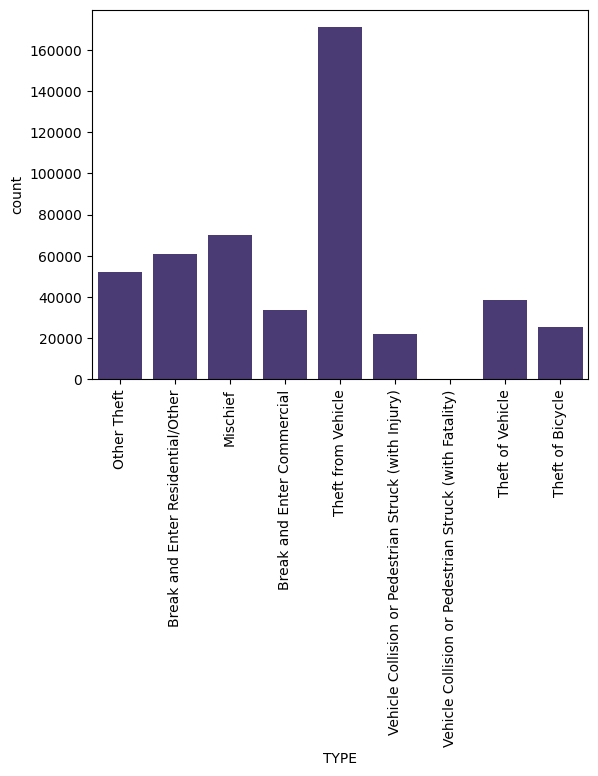


Figure 3 : Crime occurrence over the years.

Even after examining the data across various neighbourhoods, the dominance of Theft from Vehicle incidents remains apparent, indicating the widespread nature of this crime and the urgency for comprehensive solutions(Figure 4).

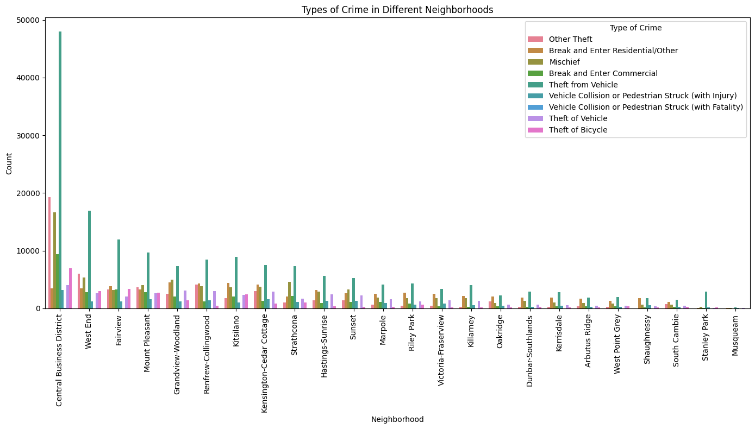


Figure 4 : Types of figures in different Neighbourhood

Furthermore, our trend analysis reveals intriguing patterns in Vehicle theft cases. Figure 5 shows a noticeable decline post-2004, followed by increased incidents since 2011. This fluctuation suggests potential shifts in criminal behaviour or enforcement efforts that warrant further investigation. While The distribution of crimes over a day also unveils intriguing insights. Crime rates gradually increase from the afternoon, peaking around 6:00 PM before tapering off (figure 6). This temporal pattern hints at potential factors influencing daily criminal activity and underscores the importance of time-sensitive policing strategies.

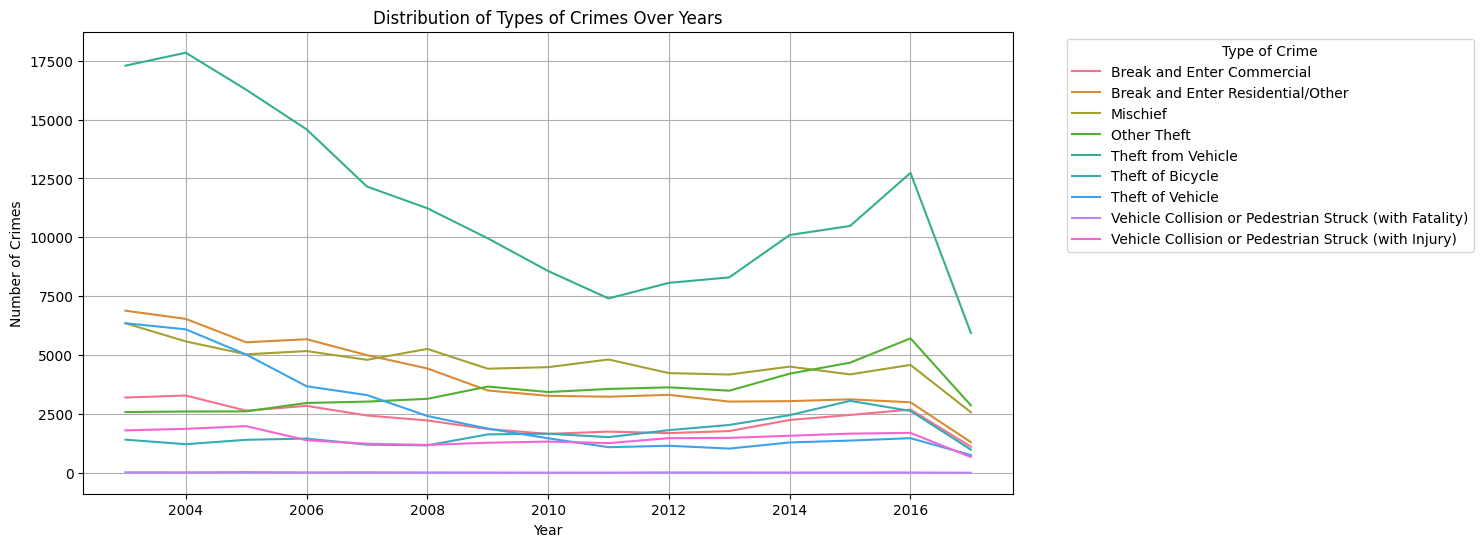


Figure 5 : Analysis of crime over the years

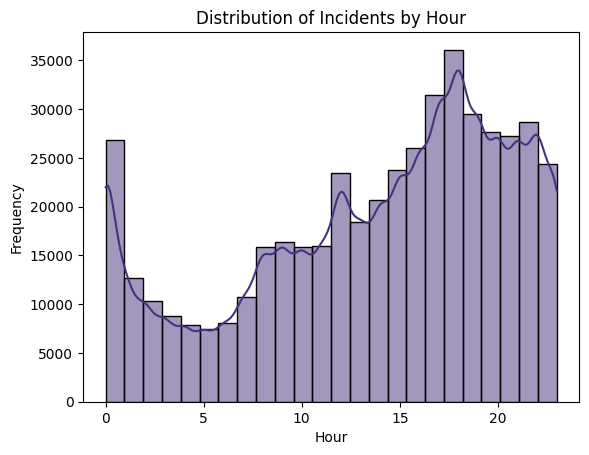


Figure 6 : Plot Showing Crime Incident by Hour

However, our examination of crime incidents by month, Figure 7, fails to unveil any discernible relationship, indicating that seasonal variations alone may not account for fluctuations in crime rates. This complexity underscores the multifaceted nature of crime dynamics and emphasises the need for nuanced approaches to crime prevention and enforcement.

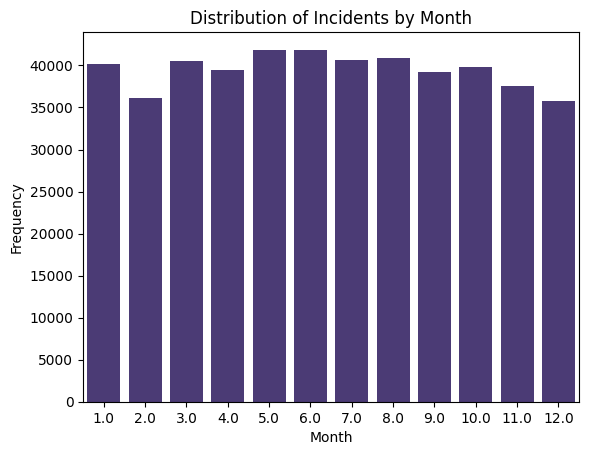


Figure 7 : Crime occurrence over Month

These dataset analysis results offer actionable insights for law enforcement agencies and policymakers. They provide valuable information to prioritise resource allocation, tailor intervention strategies, and foster collaboration across communities to combat and mitigate crime effectively.

## Prediction Results

The Decision Tree model achieved a commendable accuracy of 90.24%, showcasing its ability to classify dataset instances accurately. With a precision of 89.91% and a recall of 90.24%, it demonstrates a balanced performance in identifying positive instances and capturing the most actual positive cases. The F1 score of 89.38% further reinforces its balanced performance. Despite its relatively shorter computation time of 24.5 seconds, the Decision Tree model is robust and efficient in classifying instances, making it suitable for practical crime prediction applications. and confusion matrix and performance of the Decision tree are shown in Figure 8 and Table 3, respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Decision Tree | 90.24% | 89.91% | 90.24% | 89.38% |

Table 3 : Table showing performance of Decision Tree

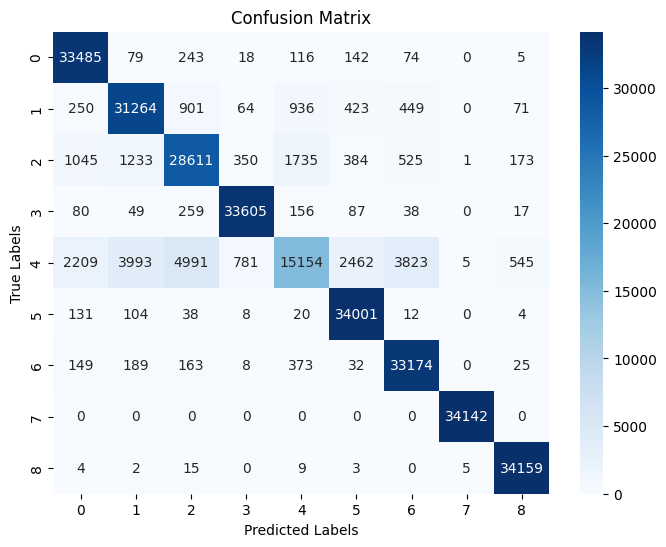


Figure 8 : Confusion Matrix of Decision Tree

The Random Forest model outperformed the Decision Tree model, achieving an accuracy of 91.84%. It showed high precision (91.63%) and recall (91.84%), indicating its ability to accurately identify positive instances and capture the most actual positive cases. The F1 score of 91.35% reflects its balanced performance across both metrics. Despite a longer computational time of approximately 6 minutes and 59 seconds, the Random Forest model proved efficient and effective in classification tasks, making it a strong choice for predictive modelling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Random Forest | 91.84% | 91.63% | 91.84% | 91.35% |

Table 4 : Table showing Performance of Random Forest

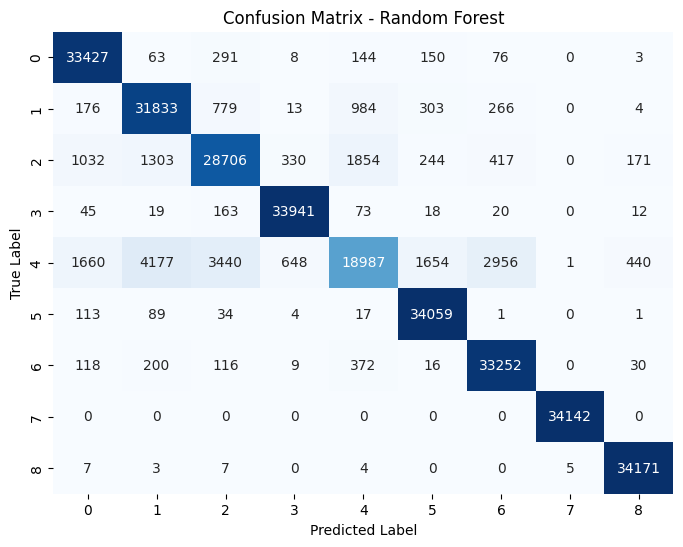


Figure 9 : Confusion Matrix of Random Forest

In contrast, the Logistic Regression model exhibited lower performance metrics, with an accuracy of only 20.36%. Table 5 shows that Logistic Regression has a notably low precision, recall, and F1 score. Despite its shorter computational time of around 1 minute and 2 seconds, the Logistic Regression model's effectiveness in classification tasks appears limited, suggesting its unsuitability for this dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Logistic Regression | 20.36% | 19.22% | 20.36% | 16.22% |

Table 5 : Table showing the Performance of Logistic Regression

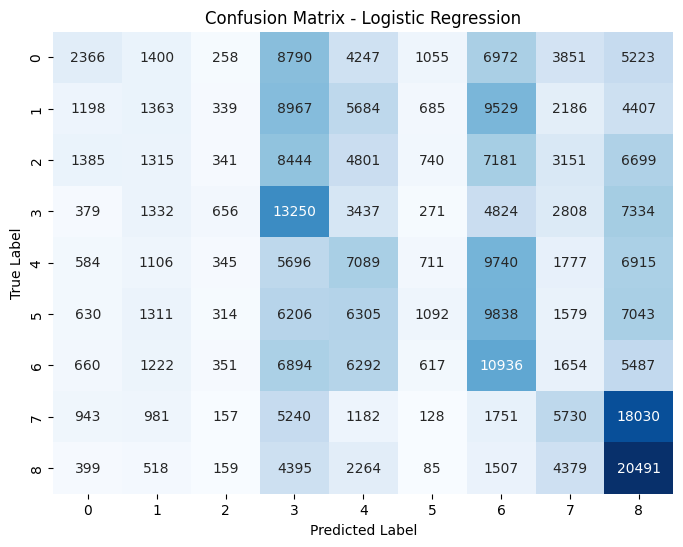


Figure 10 : Confusion Matrix of Logistic Regression

While the Decision Tree and Random Forest models exhibit strong performance and efficiency in classifying instances, the Logistic Regression model lags behind due to its lower accuracy and effectiveness. The K Nearest Neighbours model, Table 6, while not as accurate as the Decision Tree and Random Forest models, offers simplicity and interpretability, making it suitable for specific use cases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| KNN | 45.62% | 44.38% | 45.65% | 44.71% |

Table 6 : Performance of KNN

The Neural network model achieved an accuracy of approximately 28.30%, indicating its ability to classify instances correctly. A precision of about 47.78% demonstrates a relatively high ratio of correctly predicted positive instances out of all the cases predicted as positive. The recall score, calculated at approximately 18.85%, indicates the model's capability to capture a significant proportion of positive instances. The F1 score, approximately 27.03%, reflects a balanced performance across precision and recall metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Neural Network | 28.30% | 47.78% | 18.85% | 27.03% |

Table 7 : Performance of Neural Network

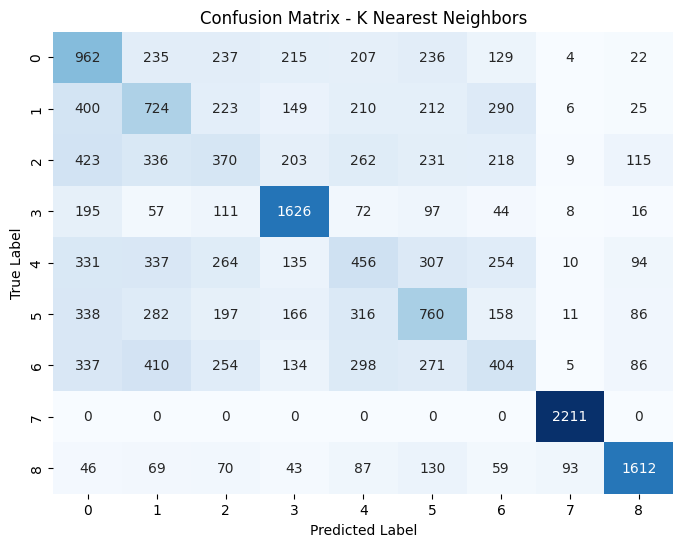


Figure 11 : Confusion Matrix of KNN

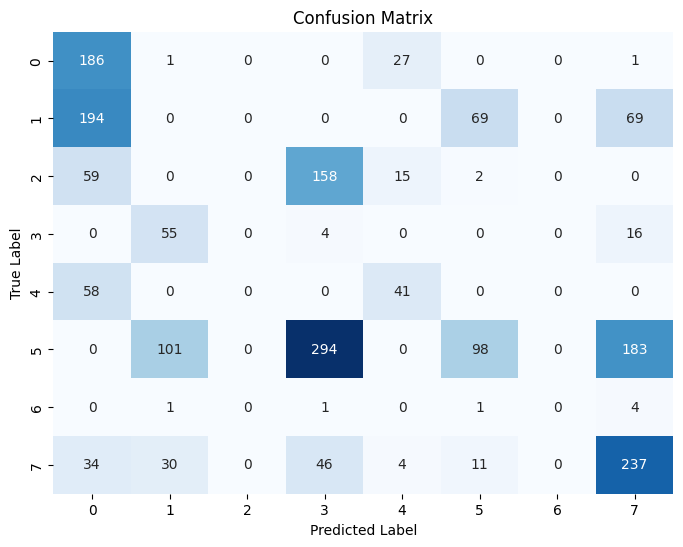


Figure 12: Confusion Matrix of Neural Network

On the other hand, the RNN model achieved an accuracy of about 50.95%, indicating its capability to make correct predictions in approximately half of the instances. A precision of approximately 62.72% showcases a higher ratio of correctly predicted positive instances out of all the cases predicted as positive compared to the overall accuracy. However, the recall score, measuring around 33.88%, indicates a lower proportion of actual positive instances correctly identified by the model. The F1 score, balancing precision and recall, is approximately 43.75%, reflecting the model's overall performance across both metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| RNN | 50.70% | 62.65% | 30.75% | 41% |

Table 8 : Performance Of RNN

In summary, the Random Forest model shines with its high accuracy, recall, and efficiency, positioning it as a robust and reliable choice for classification tasks. Conversely, the RNN model, with its higher precision but lower recall, presents a nuanced trade-off between accuracy metrics. Figure 13 compares all 4 Machine learning algorithms and 2 deep learning algorithms. The selection between these models hinges on the application's specific requirements, considering factors such as accuracy, precision, recall, computational efficiency, and interpretability to determine the optimal choice for the task.

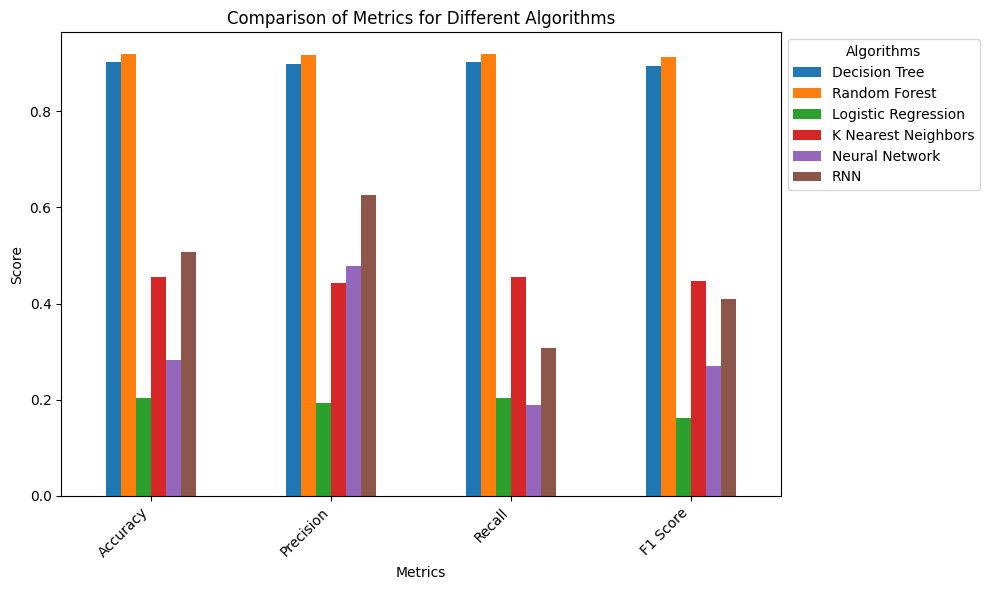


Figure 13 : Comparison of all algorithm

# DISCUSSION

The project delves into predictive modelling techniques to forecast crime outcomes, leveraging machine learning and deep learning algorithms. An in-depth analysis of a dataset encompassing details of crimes committed in Vancouver over a specific timeframe aims to develop models capable of accurately predicting crime occurrences based on various features such as crime type, location, and time.

The dataset provides valuable insights into crimes' nature, neighbourhood distribution, and temporal patterns. For instance, the data highlights "Theft from a Vehicle" as the most prevalent type of crime, underscoring its significance within the dataset. Moreover, examining crime incidents by hour reveals a gradual increase in crime rates from the afternoon to the evening, followed by a decline.

Various machine learning and deep learning algorithms are evaluated in exploring predictive models, including Decision Trees, Random Forests, Logistic Regression, K K-nearest neighbours, and Recurrent Neural Networks (RNNs). Evaluation metrics such as accuracy, precision, recall, and F1-score provide comprehensive insights into each model's performance.

The Random Forest model emerges as a strong contender among traditional machine learning algorithms, exhibiting high accuracy and balanced performance metrics. Its ability to accurately classify instances and capture actual positive cases positions it as a promising candidate for practical crime prediction applications.

The study by (Kim et al., 2018) on machine-learning-based crime prediction in Vancouver, employing two data-processing approaches and implementing predictive models like KNN and boosted decision trees, achieved 39% to 44% of crime prediction accuracies. While these results provide valuable insights, our model's performance may vary depending on the dataset and algorithms employed.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| KNN | 39% |
| Boosted Decision Tree | 44% |

Table 9 : Accuracy of previously calculated model

To assess our model's effectiveness compared with the study's findings on machine-learning-based crime prediction in Vancouver. While comparing Table 9, which shows the accuracy calculated by (Kim et al., 2018) and Table 10, it is clear that while the study achieved prediction accuracies ranging from 39% to 44%; our model may exhibit varying levels of accuracy depending on the specific dataset and the algorithms employed. If our model demonstrates higher prediction accuracies than the range reported in the study, our approach is more effective in predicting crime types using the given dataset. Conversely, if our model's accuracies fall within or below the reported range, there may be room for improvement in our methodology or data processing techniques.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Decision Tree | 90.24% |
| Random Forest | 91.84% |
| Logistic Regression | 20.36% |
| KNN | 45.62% |
| Neural Network | 28.30% |
| RNN | 50.7% |

Table 10 : Accuracy of newly calculated model

Ultimately, our crime prediction model significantly enhances law enforcement effectiveness by providing proactive insights into crime patterns and hotspots. By accurately anticipating crime occurrences, law enforcement agencies can strategically allocate resources, prioritise interventions, and develop targeted prevention strategies, ultimately bolstering community safety and security.

Our crime prediction model significantly enhances law enforcement effectiveness by providing proactive insights into crime patterns and hotspots. By accurately anticipating crime types and locations, law enforcement can strategically deploy resources, prioritise interventions, and develop targeted prevention strategies. This proactive approach optimises resource allocation, enhances community safety, and empowers authorities to focus investigative efforts where they are most needed. Ultimately, our model strengthens law enforcement capabilities, enabling more efficient crime prevention and response efforts to safeguard communities.

# FUTURE WORKS

A potential avenue for further research involves developing hotspot prediction models to supplement existing crime prediction efforts. These models would aim to pinpoint areas at heightened risk of criminal activity by considering various factors such as socio-economic conditions, environmental variables, and demographic characteristics. Moreover, there's potential for exploring dynamic hotspot prediction models that can adapt to changing crime patterns and contextual shifts over time. Integrating predictive analytics into decision-making tools for law enforcement agencies could also be explored. Future work can significantly improve crime prevention strategies and enhance public safety by embracing innovative technologies and adopting interdisciplinary approaches.

# CONCLUSION

In this paper, we used analysis and prediction to investigate Vancouver-related crimes. Analysis is done on the relationships between internal variables and the incidence of crime. Machine learning and deep learning algorithms are used to predict crime, like decision tree classifiers, random forest classifiers, logistic regression, k closest neighbour, neural networks, and radius- and crime-specific neural networks. The forecast had a high degree of accuracy, demonstrating the effectiveness of our contribution. Machine learning model random forest showed more prediction accuracy than other machine learning algorithms and deep learning algorithm

One of the main challenges tackled in this project is the class imbalance within crime datasets, which can lead to biased predictions and poor model performance. Through techniques like Random Over Sampling, we aim to balance the distribution of different crime types in the training data, improving the fairness and effectiveness of our predictive models.

Each model's performance is rigorously evaluated using accuracy, precision, recall, and F1 score metrics. This evaluation process helps us assess our models' effectiveness and provides valuable insights for law enforcement agencies to make informed decisions regarding resource allocation and proactive intervention strategies.

Furthermore, addressing class imbalance and continuously monitoring and refining our predictive models based on real-world feedback is essential to ensure their relevance and effectiveness over time. By providing actionable insights and recommendations, this project aims to contribute to the ongoing efforts of law enforcement agencies in combating crime and improving public safety in urban areas.

# REFERENCES

Bandekar, S. R. and Vijayalakshmi, C., 2020. Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India. *Procedia computer science* [online], 172, 122–127. Available from: http://dx.doi.org/10.1016/j.procs.2020.05.018.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F. and Pentland, A., 2014. Once upon a crime: Towards crime prediction from demographics and mobile data. *arXiv [cs.CY]* [online]. Available from: http://arxiv.org/abs/1409.2983 [Accessed 17 May 2024].

Dakalbab, F., Abu Talib, M., Abu Waraga, O., Bou Nassif, A., Abbas, S. and Nasir, Q., 2022. Artificial intelligence & crime prediction: A systematic literature review. *Social sciences & humanities open* [online], 6 (1), 100342. Available from: http://dx.doi.org/10.1016/j.ssaho.2022.100342.

Jenga, K., Catal, C. and Kar, G., 2023. Machine learning in crime prediction. *Journal of ambient intelligence and humanized computing* [online], 14 (3), 2887–2913. Available from: http://dx.doi.org/10.1007/s12652-023-04530-y.

Kang, H.-W. and Kang, H.-B., 2017. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one* [online], 12 (4), e0176244. Available from: http://dx.doi.org/10.1371/journal.pone.0176244.

Kim, S., Joshi, P., Kalsi, P. S. and Taheri, P., 2018. Crime analysis through machine learning. *In*: *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE.

Lin, Y.-L., Yen, M.-F. and Yu, L.-C., 2018. Grid-based crime prediction using geographical features. *ISPRS international journal of geo-information* [online], 7 (8), 298. Available from: http://dx.doi.org/10.3390/ijgi7080298.

Poppe, R., 2007. Vision-based human motion analysis: An overview. *Computer vision and image understanding: CVIU* [online], 108 (1–2), 4–18. Available from: http://dx.doi.org/10.1016/j.cviu.2006.10.016.

# APPENDIX A : PROJECT PROPOSOL

|  |  |
| --- | --- |
| **Degree Title:**  **MSc Data Science and Artificial Intelligence** | **Student’s Name:**  Hanina Kombanezhuth Abdulrahman |
| **Supervisor’s Name:**  Yan Gong |
| **Project Title/Area:**  Forecasting criminal Activity: Anticipating future incidents through predictive analysis. |

**Section 1: Project Overview**

|  |
| --- |
| * 1. **Problem** **definition:**   The rising occurrence of criminal activities globally poses a substantial and multifaceted threat to societies, individuals, and the overall welfare of communities. Crimes endanger public safety and have wide-reaching consequences, impacting social cohesion, individual reputations, and the economic prosperity of regions. Effectively addressing and alleviating the adverse effects of criminal incidents require innovative solutions that harness advanced technologies and analytical approaches. In this context, the current challenge is to develop a robust crime prediction model that utilises historical crime data to anticipate criminal activities, giving special attention to critical features such as date, time, location, and crime type.   * 1. **Project description:**   This project aims to develop a robust crime prediction model employing a machine learning algorithm for law enforcement, enabling them to predict and manage criminal trends, which helps them to enhance public safety and resource allocation and improve crime prevention effectiveness. The tool is expected to strengthen and address security challenges more predictably, which allows us to avoid and be aware of crimes likely to happen. The evident demand for such tools arises from the observed escalation in crimes in our day-to-day lives, highlighting the necessity for an apparent crime prediction model.   * 1. **Background:**   The increase in crime data recording coupled with data analytics resulted in the growth of research approaches aimed at extracting knowledge from crime records to understand criminal behaviour better and ultimately prevent future crimes. While many of these approaches make use of clustering and association rule mining techniques, fewer approaches are focusing on predictive models of crime ((Saltos & Cocea, 2017) Using crime datasets requires different strategies for the varying types of data that describe illicit activity. provide a survey of crime prediction efforts wherein various machine learning methods have been applied to multiple types of datasets: criminal records, social media, news, and police reports (Falade, Adesola; Azeta, Ambrose; Oni, Aderonke; Odun-ayo, Isaac;, 2019) Several AI techniques have been widely studied to reduce or prevent crime and ensure the safety of people in different countries. These machine-learning models can be used to predict future crimes, their attributes, etc. Machine learning will enable police departments to optimise their resources by finding hotspots based on time, type, or other factors (Mary Shermila, A; Bellarmine, Amrith Basil; Santiago, Nirmala;, 2018). Accurate real-time crime predictions help to reduce the crime rate but remain a challenging problem for the scientific community as crime occurrences depend on many complex factors. In this work, various visualising techniques and machine learning algorithms are adopted for predicting the crime distribution over an area (ToppiReddy, Hitesh Kumar Reddy; Saini, Bhavna; Mahajan, Ginika;, 2018)   * 1. **Research Questions** * How effectively is the crime prediction model processing and forecasting criminal incidents using Machine learning techniques? * Which ML algorithm will better predict criminal activities considering the various external factors influencing crime rates?   1. **Aims and objectives:**   This project aims to develop an advanced crime prediction model that Utilizes historical crime data to implement machine learning algorithms that accurately predict and forecast criminal activities based on temporal and spatial parameters, including time, date, coordinates, and crime type. Also, develop a visualisation component that effectively represents crime hotspots, allowing law enforcement agencies to identify geographical areas with a higher likelihood of criminal incidents. This visualisation aims to facilitate strategic resource allocation for crime prevention. |

**Section 2: Artefact**

|  |
| --- |
| **2.1 What is the artefact that you intend to produce?**  An innovative crime prediction model utilising machine learning algorithms in Python, aimed at accurately forecasting and managing criminal activities based on historical data encompassing temporal and spatial parameters such as time, date, coordinates, and crime type.  **2.2 How is your artefact actionable (i.e., routes to implementation and exploitation in the technology domain)?**  It will be helpful for law enforcement agencies to look after the possibility of crime and prevent it by precautions. Visualising it on the map can help them to identify crime hotspots. |

**Section 3: Evaluation**

|  |
| --- |
| **3.1 How are you going to evaluate your project artefact?**  The evaluation strategy involves a thorough analysis of the crime prediction model. We'll utilise performance metrics like accuracy, precision, recall, and F1 score to quantitatively measure the model's accuracy in forecasting crimes based on historical data. Additionally, we'll test the artefact with diverse datasets covering a range of crime scenarios, locations, and timeframes to gauge its generalisation across different situations  .**3.2 How does this project relate to your MSc Programme and degree title outcomes?**  This project is related to MSc Data Science and Artificial Intelligence by enabling me to knowledge gained by practical application of data analysis, mining and machine learning algorithms to analyse and interpret the data. This project allows me to deploy proficiency in predictive models, handling large datasets, etc.  **3.3 What are the risks in this project, and how are you going to manage them?**  First, gathering and analysing crime data is made more difficult by the enormous and varied nature of crime data worldwide. Finding the pertinent factors impacting criminal behaviour is essential. There is an additional degree of risk involved in testing and evaluating several AI algorithms because figuring out which variables interact in the correct order to predict crime is not an easy task. To reduce these risks and guarantee project success, a balance between algorithmic efficacy, timely execution, and valid data must be struck. |

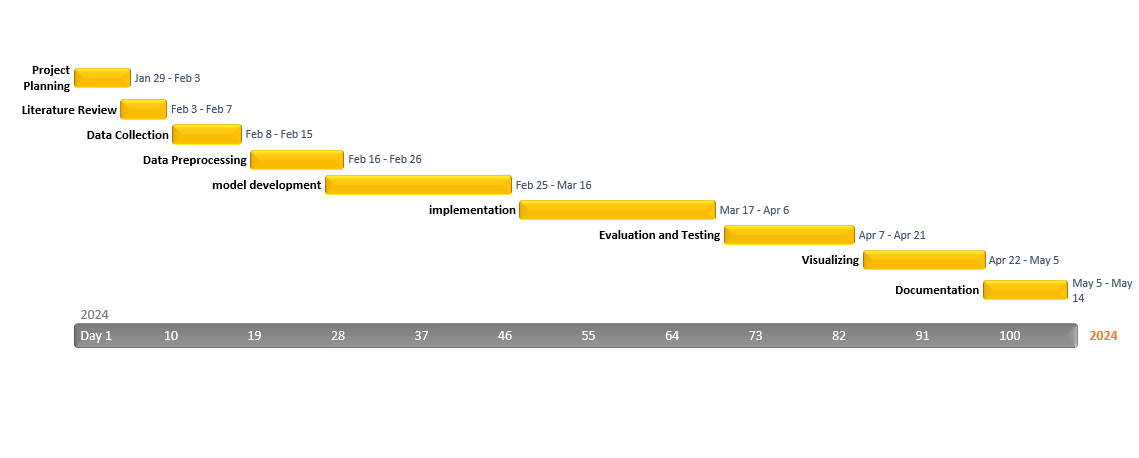
**Section 4: References**

|  |
| --- |
| **4.1 Please provide references if you have used any.**  Falade, A., Azeta, A., Oni, A. and Odun-ayo, I., 2019. Systematic literature review of crime prediction and data mining. *Review of computer engineer studies* [online], 6 (3), 56–63. Available from: http://dx.doi.org/10.18280/rces.060302.  Mary Shermila, A., Bellarmine, A. B. and Santiago, N., 2018. Crime data analysis and prediction of perpetrator identity using machine learning approach. *In*: *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE.  Saltos, G. and Cocea, M., 2017. An exploration of crime prediction using data mining on open data. *International journal of information technology & decision making* [online], 16 (05), 1155–1181. Available from: http://dx.doi.org/10.1142/s0219622017500250.  ToppiReddy, H. K. R., Saini, B. and Mahajan, G., 2018. Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science* [online], 132, 696–705. Available from: http://dx.doi.org/10.1016/j.procs.2018.05.075. |

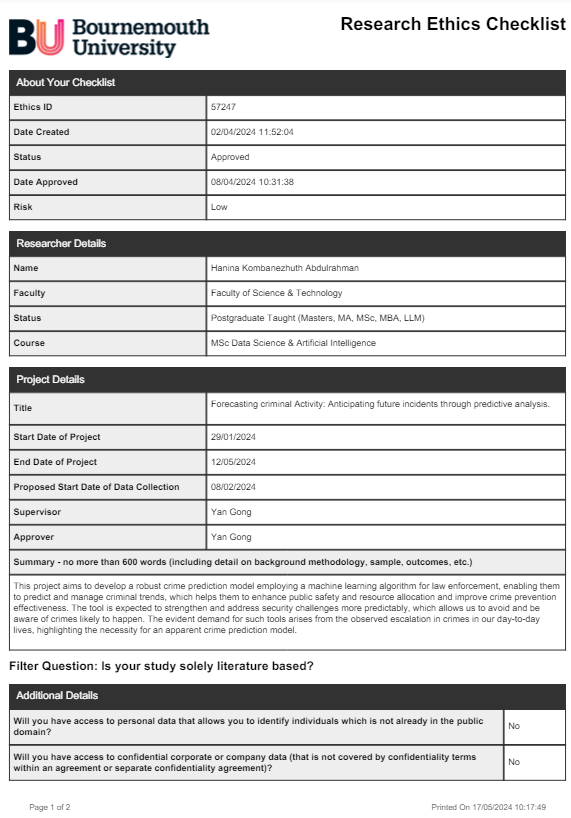
**Section 5: Academic Practice and Ethics**

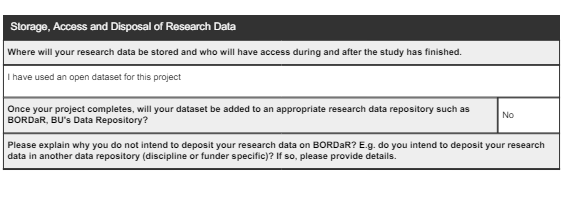
|  |  |
| --- | --- |
| **5.1 Have you made yourself familiar with, and understand, the University guidance on referencing and plagiarism?** | **Yes** |
| **5.2 Do you acknowledge that this project proposal is your own work and that it does not contravene any academic offence as specified in the University’s regulations?** | **Yes** |

**Section 6: Proposed Plan (please attach your Gantt chart below)**



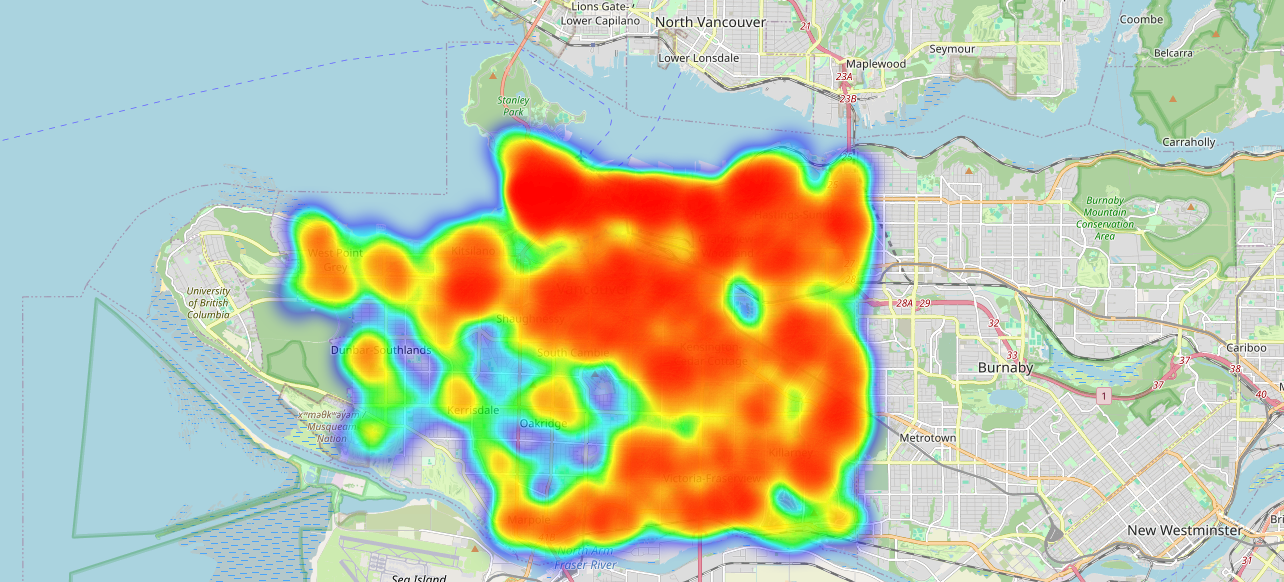
# APPENDIX B : RESEARCH ETHICS CHECK LIST



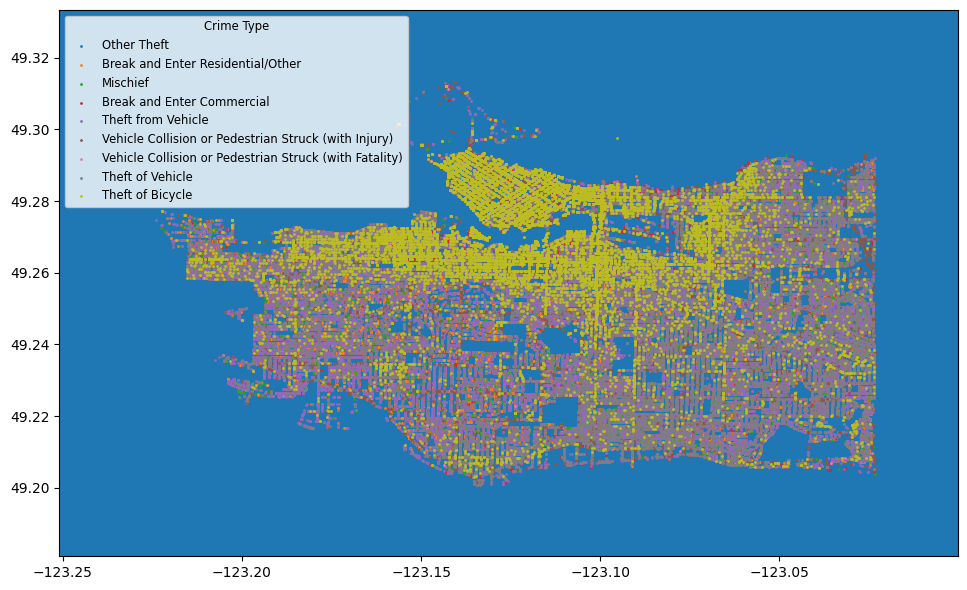


# APPENDIX C: DATA VISUALISATION

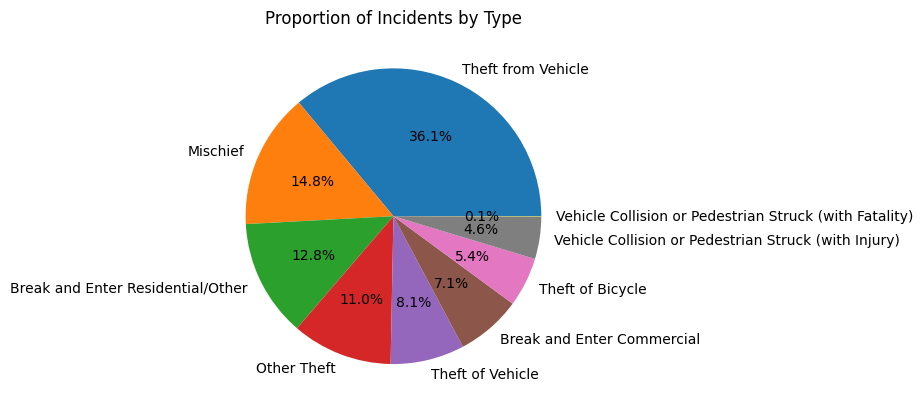
Geographic data for vehicle theft incidents in Vancouver in 2017 creates a Folium map centred on Vancouver and overlays a heatmap depicting the spatial distribution of these incidents.



Utilises GeoPandas to plot crime incidents on a world map, categorising them by crime type and assigning a unique colour to each category, enabling visual analysis of crime distribution across geographic regions. Additionally, it sets boundaries to focus the map on a specific area of interest, improving the readability and interpretation of the plotted data.



A pie chart to display the proportion of different incident types within the dataset, with percentage labels for each slice and a title indicating "Proportion of Incidents by Type".



A heatmap is a data visualisation technique that uses colour to represent the values of a matrix. Checking correlation between columns using the heatmap

