UNIVERSITY
OF WOLLONGONG
AUSTRALIA

# Crime Prediction Model

Fauzia Faria - 7557747
Namita Samant - 7382285
Adiba Hasin - 7331265
Varun Ravi Sankar - 6776000
Shruti Parmar - 7362262
Karishma Gosalia - 7361932

**Abstract**

This project explored the application of machine learning techniques for crime prediction using historical crime data. The research makes use of a number of machine learning algorithms, such as K-Nearest Neighbours (KNN), ridge regression, Linear Support Vector Classification (SVC), and Gaussian Naive Bayes, to help anticipate criminal activity in the South Australian region. Through comparative analysis, the study aims to identify the most effective model among these approaches.The outcomes of this project facilitated efficient resource allocation, allowing efforts to be focused on areas with a higher likelihood of criminal activities. The study also looks into additional data like demographics and the amount of people detained on remand for crimes that have already been committed. A vital aspect of the study's analytical methodology is the investigation of how these anti-crime tactics affect predictions of criminal activity. The project's findings support evidence-based decision-making in law enforcement,enabling authorities to allocate resources effectively, implement timely interventions, and develop proactive crime prevention strategies. By leveraging machine learning algorithms, this project sought to enhance the efficacy of crime prediction and prevention efforts, ultimately contributing to the reduction of crime rates and the improvement of overall public safety.

# Contents

# List of Figures

# 1 Introduction

To prevent crime and preserve public safety, crime prediction plays a critical role in law enforcement. Recent developments in artificial intelligence (AI) create new opportunities for reliable crime prediction models. While Brantingham and Brantingham's (2017) work highlights the role of environmental elements and routines in crime patterns, Kim and Jeong's (2021) study addresses the usage of AI-based ways to lower crime rates. To improve the accuracy and efficiency of current prediction algorithms, anti-crime measures must be incorporated.

Our goal in this project is to create a thorough crime prediction model for the South Australia region in Australia, with an emphasis on including deterrent practices to anticipate crime rates. We identified intricate links and patterns in the crime data by utilizing the Random Forest algorithm, a cutting-edge AI method. Our main goal is to develop a context-specific, highly precise crime prediction model that takes into consideration South Australia's particular peculiarities. Our model offers practical insights for crime prevention and promotes evidence-based decision-making by incorporating anti-crime characteristics such as community policing programs, law enforcement deployment tactics, and targeted preventive actions. In the section that follows, we looked at related research by Kim and Jeong (2021) and Brantingham and Brantingham (2017), summarized their main conclusions, and pointed out any gaps in the body of knowledge. We also went over the precise goals and objectives of our project, highlighting the significance of including anti-crime characteristics to enhance the precision and efficacy of crime rate projections in South Australia.

Crime is not just an abstract concept; it affects real people in our communities every day. It brings fear and distress to individuals, disrupts families, and undermines the social fabric of our society. Moreover, crime has far-reaching consequences, impeding economic growth and hindering the development of safe and prosperous neighborhoods. To tackle this multifaceted problem, accurate crime prediction is of utmost importance. It empowers law enforcement agencies and policymakers to make informed decisions, allocate resources efficiently, and proactively combat criminal activities. By leveraging a reliable crime prediction model, we can stay one step ahead of criminals and take preventive measures to protect our communities. The value of accurate crime prediction lies in its ability to guide law enforcement agencies in strategic planning and resource allocation. By identifying high-risk areas and vulnerable populations, authorities can concentrate their efforts where they are most needed. This proactive approach allows for the deployment of resources in a targeted and efficient manner, ensuring a timely response to potential crime hotspots. Moreover, accurate crime prediction enables the implementation of preventive measures that address the root causes of criminal behavior.

By understanding crime patterns and trends, we can develop evidence-based strategies to emphasize community engagement, promoting of social interventions to reduce crime rates. Our project aims to go beyond conventional crime prediction models by incorporating anti-crime measures. Measures like 'Taken into remand', reflect the number of offenders detained for offenses within a specific location. This integration aims to further improve the effectiveness of our models and potentially aid law enforcement in making informed decisions. Through the development of a precise and context-specific crime prediction model for South Australia, our project addresses the pressing issues of crime, contribute to public safety, and enhance the overall well-being of our community.

## 1.1 Description

Crime prediction is a fundamental aspect of modern law enforcement, enabling proactive strategies to prevent and respond to criminal activities effectively. In this project, we delve into crime prediction for South Australia, utilizing a comprehensive dataset sourced from the South Australian government's open data portal (Data.SA). This dataset spans from January 2010 to July 2023 and encompasses critical information such as the date of incidents, detailed offence descriptions, and offence counts.

Our primary data source, the South Australian crime dataset, serves as the cornerstone of our analysis. Before constructing predictive models, we meticulously prepared the data by undertaking a comprehensive data cleaning process. This involved thorough procedures to address missing values, ensuring that our dataset was complete and reliable. Furthermore, we meticulously encoded categorical variables, enabling the inclusion of qualitative information in our analysis. Dates were carefully structured , facilitating subsequent temporal analysis. Additionally, we performed data aggregation at several temporal and spatial levels to improve the prediction skills of our models. Through this procedure, we were able to produce useful features that improved the adaptability and accuracy of our predictive models.

Our project encompasses both regression and classification tasks, each employing distinct machine learning techniques.

**Regression**

**Principal Component Analysis (PCA):** We applied PCA to reduce dimensionality, enhancing model computational efficiency.

**Ridge Regression:** Utilizing historical data and engineered features, we employed Ridge Regression to predict crime counts.

**K-Nearest Neighbors (KNN) Regression:** KNN regression was employed to make predictions based on the similarity of crime patterns in the dataset.

**Classification**

**Random Forest Classifier:** This ensemble learning technique was deployed to classify criminal offences, offering high accuracy and robustness.

**Support Vector Classifier (SVC):** SVC was utilized for binary and multiclass classification tasks, capable of handling complex decision boundaries.

**Gaussian Naive Bayes:** We explored Gaussian Naive Bayes classifier's simplicity and efficiency for classification tasks, assuming normal feature distributions.

**Model Evaluation**

Critical to our project was the rigorous evaluation of model performance. We utilized a suite of evaluation metrics tailored to the specific task, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), F1-score, and accuracy. Cross-validation techniques ensured our models' generalization to unseen data.

Our project focuses on predicting crime rates in South Australia by harnessing machine learning techniques. With a solid foundation in data preprocessing and a diverse set of predictive models, we aim to provide law enforcement agencies with valuable insights to enhance public safety and resource allocation strategies. Our detailed model evaluations

and results contribute to the selection of the most suitable predictive models for crime prediction in South Australia. This project represents a significant step toward leveraging data and machine learning to improve crime prevention and intervention efforts in the region.

## 1.2 Aims

The aim of our proposed project is to accomplish several key objectives within the field of crime prediction. Our main goal is to develop a highly accurate crime prediction model by utilising a variety of advanced machine learning methods, particularly Random Forest, Gaussian Naive Bayes, Ridge Regression, Linear Support Vector Classification (SVC), and K-Nearest Neighbours (KNN). This model will leverage historical data and pertinent variables such as offense count, offense level description, post-code, while also incorporating anti-crime features, thus ensuring its effectiveness in forecasting crime rates. Our project seeks to assess the impact of various anti-crime measures through comprehensive analysis, enabling us to determine how these interventions contribute to the prediction of criminal activities. By providing evidence-based insights, our research outcomes will offer valuable guidance to law enforcement agencies, policymakers, and community stakeholders, facilitating targeted resource allocation and the implementation of effective crime prevention strategies. Through these collective efforts, our project aims to contribute significantly to proactive measures and community-driven initiatives that enhance public safety and mitigate crime.

# 2 Related Work

Crime prediction plays a vital role in proactive law enforcement and crime prevention strategies. Researchers have dedicated their efforts to exploring innovative approaches and methodologies to improve the accuracy of crime prediction. This literature review provides an overview of key studies that have contributed to the field, focusing on the use of artificial intelligence, pattern theory, self-exciting point processes, and geocoding accuracy.

One notable study by Kim and Jeong (2021) sought to reduce crime rates through the application of artificial intelligence (AI). By utilizing advanced algorithms like Multiclass Logistic Regression and Multiclass Neural Network, the researchers successfully predicted various types of crimes and the tools employed in those crimes. Their analysis, based on historical criminal records from Baltimore spanning 2011 to 2016, demonstrated the superior performance of Multiclass Logistic Regression in accurately predicting both crime types and tools. This study highlights the potential of AI-based approaches in crime prediction, providing law enforcement agencies with valuable tools to combat crime effectively.

Dakalbab et al.(2022) conducted a systematic literature review to explore the application of artificial intelligence (AI) in crime prediction. The study examined various perspectives such as the type and category of crimes, time of study, and the techniques employed in crime prediction. They identified 64 different machine learning (ML) techniques, including classification algorithms like, Naive Bayes and back propagation, and evaluated their accuracy and suitability for predicting crime categories.The strength of this study lies in its comprehensive analysis of existing research and the identification of ML algorithms used in crime prediction. However, limitations include potential bias in selected papers, the need for cautious interpretation of synthesized results, and the lack of conclusive recommendations on the best algorithm for crime prediction. Additionally, biases in training data

and their impact on reinforcing existing biases in law enforcement practices were acknowledged.

In addition to AI, Brantingham and Brantingham (1993) proposed a pattern theory of crime, emphasizing the influence of environmental factors, routines, and situational contexts on criminal behavior. Their research underscores the importance of understanding spatial-temporal characteristics and crime patterns for effective prediction and prevention. By considering the role of the environment and situational contexts, this theory contributes to the development of environmental criminology and situational crime prevention strategies, enabling law enforcement to address crime proactively. The strength of this study lies in its theoretical framework, which highlights the importance of understanding crime patterns and the spatial-temporal characteristics of criminal events. By considering these factors, law enforcement agencies and policymakers can develop targeted crime prevention strategies. However, a limitation of this study is its focus on theoretical concepts rather than practical implementation. Further research is needed to translate these ideas into actionable crime prediction models.

Ahishakiye et al. (2017) aimed to predict the likelihood of low, medium, or high violent crimes in counties using the Decision Tree (J48) classification algorithm. They achieved a high prediction accuracy of 94.25% out of 174 datasets, demonstrating the reliability of the J48 algorithm. The study highlighted the advantages of using decision tree algorithms like J48, including their relatively quick execution time and suitability for crime analysis and prediction by law enforcement agencies. However, limitations of this study include the use of a dataset from a single source, potentially limiting the generalizability of the findings. Also, only one classification algorithm was considered, neglecting other available algorithms, and external factors such as socio-economic and demographic variables were not taken into account, which could impact crime rates.

Furthermore, Mohler et al. (2011) introduced a self-exciting point process model for crime prediction that incorporates temporal dependencies and the contagious nature of crime. Their study demonstrated the model's efficacy in capturing spatial and temporal patterns, facilitating the prediction of crime hotspots and temporal trends. By accounting for the influence of past criminal events on future events in nearby locations, the self-exciting point process model enhances the accuracy of crime prediction and enables proactive law enforcement strategies. However, this study is limited by its reliance on historical crime data and assumptions about the underlying processes. External factors and changes in crime patterns over time may not be fully captured by the model, necessitating ongoing refinement and validation.

Accurate geocoding of crime incidents is another crucial aspect of precise crime prediction. Ratcliffe (2004) focused on the geocoding of crime data and introduced the concept of a minimum acceptable hit rate (MAHR) for evaluating the accuracy of crime prediction models. Establishing the MAHR as a quantitative criterion, the study emphasized the importance of reliable geocoding methods in enhancing the effectiveness of crime prediction algorithms. This research underscores the significance of geocoding accuracy in crime analysis and prediction efforts, enabling law enforcement to make informed decisions based on reliable location data. The study primarily focuses on the evaluation of geocoding methods rather than proposing new prediction techniques. Future research should explore the integration of reliable geocoding methods with advanced machine learning algorithms to enhance the precision and effectiveness of crime prediction models.

These existing research studies on one hand, because of their application of AI algorithms,

such as Multiclass Logistic Regression and Multiclass Neural Network, provide promising results in predicting crime types and tools. They offer law enforcement agencies valuable insights to allocate resources effectively and implement targeted crime prevention strategies. Additionally, the pattern theory of crime contributes to a comprehensive understanding of criminal behavior, enabling the development of proactive interventions based on environmental factors and situational contexts. However, it is important to acknowledge the limitations of these studies. While AI algorithms show promise, the reliance on historical data from specific locations, such as Baltimore, raises concerns about the generalizability of the findings to other cities or regions with different crime patterns. Moreover, the focus on specific algorithms may overlook the potential benefits of alternative approaches. Additionally, the studies primarily consider crime records and fail to account for external factors like socioeconomic status and demographic changes, which can influence crime rates. Ethical considerations related to the use of predictive policing systems are yet to be fully addressed, necessitating further research to explore potential biases or unintended consequences.

In the existing literature, there are gaps in terms of incorporating anti-crime features into crime prediction models and analyzing the impact of specific anti-crime actions on crime rates. Additionally, there is a need to bridge the gap between crime prediction research and evidence-based decision making for effective resource allocation and intervention implementation. Therefore, the proposed study aims to address these gaps by developing an accurate crime prediction model with anti-crime features, identifying influential anti-crime actions, and informing evidence-based decision making in crime prevention efforts.

# 3 Methodology

The proposed methodology for the crime prediction project involves a systematic approach to gather and analyze data, develop predictive models, and incorporate anti-crime actions. The following steps outline the methodology followed:

**Data Collection:** The primary data source for this project is from the South Australian government's open data portal. The collection includes crucial data such as the reported incident dates, in-depth descriptions of the offences, and numerical counts of these offences. The project focuses on South Australia (SA) as the geographical area of interest to analyse and predict crime patterns accurately.

**Data Pre-processing:** The collected crime data underwent a thorough pre-processing stage. This step involved cleaning the data, handling missing values, and transforming the data into a suitable format for analysis. Features such as crime type, location, time, and additional relevant variables are extracted and processed to capture meaningful patterns and trends.

**Feature Engineering:** Feature engineering is a crucial step in preparing the data for modeling. It involves creating new features or transforming existing features to enhance the predictive power of the model. Domain knowledge and insights from crime experts were incorporated to engineer informative features that capture important aspects of crime behavior and demographics specific to South Australia.

**Feature Selection:** A careful analysis of the collected data was conducted to identify the most significant features that contribute to crime prediction. Feature selection techniques, such as correlation analysis and statistical tests, were used to identify relevant predictors.

**Algorithm Selection:** Based on the specific requirements of the crime prediction problem and the characteristics of the dataset, we selected a range of proficient machine learning algorithms, including decision trees, random forests, K-Nearest Neighbors (KNN), ridge regression, Linear Support Vector Classification (SVC), and Gaussian Naive Bayes. The selection criteria were based on their demonstrated success in effectively handling classification tasks and their capacity to manage large-scale datasets efficiently.

**Model Development:** The selected machine learning algorithms were implemented to develop predictive models. The models were trained on the historical crime data, using appropriate evaluation metrics to assess their performance. Techniques such as cross-validation and hyperparameter tuning were employed to optimize the models.

**Software**

1. **Jupyter Notebook:** Jupyter Notebook was used as the primary coding environment for developing the crime prediction model. Jupyter Notebook provides an interactive and collaborative platform for writing Python code, visualizing data, and documenting the research process. It allows for easy experimentation, code execution, and visualization of results.
2. **Python:** Python, a widely-used programming language in data science and machine learning, is the primary language for implementing the crime prediction model. Python provides a rich ecosystem of libraries and frameworks for data manipulation, statistical analysis, and machine learning. Key libraries that will be used include NumPy, Pandas, Matplotlib, Seaborn, and scikit-learn.
3. **Google Drive:** Google Drive is utilized for data storage and collaboration. It provides a cloud-based storage platform where datasets, research papers, and other project-related files can be securely stored and accessed. Google Drive also facilitates easy sharing and collaboration among team members, allowing for seamless integration with Jupyter Notebook and other tools.
4. **Data Visualization Tools:** Various data visualization tools, such as Matplotlib and Seaborn, were used to create insightful visualizations of the crime data, model predictions, and evaluation metrics. These tools enable the representation of data patterns, trends, and relationships, enhancing the interpretation and communication of results.
5. **Machine Learning Libraries:** Scikit-learn, a popular machine learning library in Python, was employed for building and training machine learning models. Scikit-learn provides a wide range of algorithms, evaluation metrics, and preprocessing techniques, making it suitable for crime prediction tasks. Other specialized libraries for time series analysis or spatial data analysis may also be utilized based on the specific requirements of the project.
6. **Data Manipulation and Analysis Tools:** Libraries like NumPy and Pandas were utilized for data manipulation, preprocessing, and feature engineering tasks. These libraries offer efficient data structures, data cleaning functions, and powerful data manipulation capabilities, allowing for efficient data preprocessing and exploration.
7. **Version Control:** Git, a widely-used version control system, was employed to track changes in the project's codebase and collaborate with team members. Git provides version control functionalities, enabling efficient code management, collaboration, and the ability to revert to previous versions if necessary.

**Incorporating Anti-Crime Actions:** We included an anti-crime element to the algorithm to improve the models' predicted efficacy and accuracy. A dataset of people detained on remand at specific locations was obtained. Following that, this dataset was incorporated with the dataset on crime statistics to provide our model with further training and testing data. This comprehensive approach made it easier to pinpoint high-risk regions and improve narrowly focused efforts to prevent crime.

**Evaluation and Validation:** The developed models are evaluated using various performance metrics, such as accuracy, precision, recall, and F1 score. Furthermore, we use various classification algorithms to compare the performance. The models are validated on a separate set of data to assess their generalization capability and robustness. Comparison with existing crime prediction methods or historical crime data is also performed to gauge the effectiveness of the proposed approach.

**Project Management:** The project was managed using a structured approach, with defined timelines, milestones, and regular progress tracking. Collaboration among team members, utilized zoom meetings, and online collaboration tools. The methodology provides a comprehensive framework for the development of a crime prediction system that incorporates anti-crime actions. By following this methodology, we wished to leverage machine learning techniques to accurately predict crime patterns and enable proactive measures for crime prevention.

**Choosing the Proper Dataset:**

- **Identifying Relevant Datasets:** Exploring the government databases, open data initiatives, or crime statistics provided by law enforcement agencies to obtain crime-related data. For example, the Australian Bureau of Statistics (ABS) Crime Statistics dataset was able to provide information on various types of crimes reported in Australian cities.
- **Ensure Data Quality:** Cleansing and preprocessing the collected data, handling missing values, outliers, and inconsistencies.
- **Feature Engineering:** Selecting relevant features: Identifying potential features such as demographics, historical crime data, geographical features (e.g., location, proximity to certain establishments), and user-provided information (e.g., perceptions, experiences, or safety-related data).
- **Transform and engineer features:** Performing feature transformations, scaling, encoding categorical variables, and creating new derived features if necessary.
- **Splitting the dataset:** Dividing the dataset into training, validation, and testing subsets to evaluate model performance.
- **Select appropriate algorithms:** Exploring machine learning techniques suitable for crime prediction, such as logistic regression, decision trees, random forests, or neural networks.
- **Model training:** Training the selected models using the training dataset, adjusting hyperparameters and applying techniques like cross-validation to optimize model performance.
- **Model evaluation:** Assessing the trained models using the validation dataset, considering metrics such as accuracy, precision, recall, and F1-score.

- **Model selection:** Choosing the best-performing model based on evaluation results and considerations of interpretability, scalability, and computational efficiency.
- **Crime Probability Estimation:** Applying the selected model to the testing dataset for estimating crime probabilities.
- **Analyze prediction outcomes:** Evaluating the performance of the model by comparing predicted crime probabilities with actual crime occurrences.
- **Fine-tuning and optimization:** Refining the model if necessary, considering feedback from the evaluation phase.

**Challenges:**

- **Address potential challenges:** Identifying and addressing potential limitations, such as data availability, data quality issues, bias in user-provided information, or model interpretability.
- **Considering ethical considerations:** Ensuring privacy protection and compliance with legal and ethical guidelines when handling user-provided data.
- **Handle class imbalance:** Addressing the challenge of imbalanced crime data distribution by employing techniques such as oversampling, undersampling, or using specialized algorithms.
- **Interpretability and explainability:** Striving to enhance the interpretability of the developed model by employing techniques such as feature importance analysis, model-agnostic interpretability methods, or surrogate models.

## 3.1 Dataset

In this methodology, we aimed to utilize different sources to obtain information on various types of crimes reported in Australian cities. The datasets are preprocessed and integrated with other relevant features, such as demographics, geographical data, and user-provided information. By applying appropriate machine learning algorithms and model evaluation techniques, the developed model estimates crime probabilities based on the selected features.

The various datasets we meticulously explored for this application include:

1. **Australian Bureau of Statistics (ABS) Crime Statistics:** The ABS provides crime statistics at various levels, including national, state, and local government areas. You can access data on different types of crimes, such as burglary, assault, theft, etc.
2. **Australian Institute of Criminology (AIC) Datasets:** The AIC collects and maintains comprehensive crime-related datasets for research purposes. These datasets cover a wide range of crime topics, including crime victimization, offender characteristics, and crime patterns.
3. **New South Wales Bureau of Crime Statistics and Research (BOCSAR):** BOCSAR provides crime statistics for New South Wales, including data on different types of offenses, their locations, and trends over time. These datasets are publicly accessible and can be useful for analyzing crime patterns in specific areas of NSW.

Initially, our project revolved around the utilization of the aforementioned datasets; however, due to various constraints and the unavailability of the specific data we were seeking, we had to pivot in a different direction. Building upon this pivot, our project then centered around the predictive modeling of crime in South Australia. In light of the challenges faced in obtaining the initial datasets, we redirected our efforts toward leveraging a rich dataset obtained from the South Australian government's open data portal, Data.SA This dataset spans from January 2010 to July 2023 and comprises a detailed record of reported crime incidents. The quality and diversity of this input data are paramount, laying the foundation for rigorous analysis and model development.

**Data Source**

**Identification:** The genesis of our input data lies in reported crime incidents in South Australia, harnessed from a diverse array of law enforcement channels and reporting systems. These sources collectively ensure a comprehensive representation of the regional criminal landscape.

**Details:** Characterized as crime incident records, the data spans the geographic expanse of South Australia and is subject to periodic updates, integrating the latest incident reports for real-time relevance.

**Data Type**

The input data manifests in various types:

**Text:** Embracing suburb names and detailed offense descriptions.

**Numerical:** Encompassing variables like postcodes, offense counts, and a binary indicator denoting individuals taken into remand.

**Categorical:** Providing insights into offense levels through descriptive categorizations.

**Data Format**

The structured tabular format of the input data, housed in a CSV format, facilitates seamless accessibility, laying the groundwork for effective analysis.

**Data Schema Overview**

The data schema used in the project includes pivotal fields :

- **Reported Date:** Date of the incident report.
- **Suburb - Incident:** Suburb where the incident occurred.
- **Postcode - Incident:** Postcode of the incident location.
- **Offence Level 1 Description:** General category of the offense.
- **Offence Level 2 Description:** Subcategory of the offense.
- **Offence Level 3 Description:** Detailed description of the offense.
- **Offence count:** Number of offenses reported for the specific incident.
- **Taken to remand:** Binary indicator (1 for yes, 0 for no) of individuals taken into remand.

The Reported Date and Suburb - Incident fields serve as linchpin identifiers, anchoring the dataset and enabling the discernment of individual incident records.

**Data Quality Assessment**

The robust assessment of data quality extended to addressing missing values, outliers, and errors.

**Data Preprocessing Steps**

Preceding project utilization, a thorough preprocessing regimen unfolds, incorporating normalization, cleaning, and feature engineering. These steps collectively enhanced the quality and utility of the input data.

**Data Exploration Insights**

The exploratory data analysis undertaken yielded invaluable insights, unveiling trends, patterns, and anomalies within the dataset. Visualizations served as compelling tools, elucidating key findings, and paving the way for subsequent analytical endeavours.

**Data Volume**

The dataset, comprising an extensive 945,013 records over a span of 13 years, offers a substantial corpus for analysis. Considerations are judiciously made regarding the potential impact of this voluminous dataset on processing and analytical methodologies. The corresponding CSV file, containing the dataset, is approximately 104.401 megabytes in size.

## 3.2 Project Challenges & Solutions

The development and implementation of a crime prediction model raises significant ethical considerations that require thorough examination and attention. These deliberations encompass potential biases, privacy implications, and the need for stakeholder engagement. While developing the crime prediction model, our project faced several multifaceted challenges. These challenges span from data collection and preprocessing hurdles to the intricacies of machine learning model selection and evaluation. Navigating these challenges was essential to achieve the full potential of data-driven crime prediction for the benefit of public safety.

**Availability of high-quality crime data:** One of the most difficulties is the lack of high-quality crime statistics. Crime statistics can be difficult to get, and the data that is accessible may not be complete or reliable. Furthermore, the collection and use of crime data raises privacy and ethical considerations. These issues must be overcome if the full potential of machine learning and deep learning for crime prediction is to be realized.

**Solution:** Identifying a reliable dataset is crucial for the success of the project.In our research to find the appropriate data we came across number of data portals offering an extensive collection of publicly accessible datasets.These repositories are centralized collections consisting a wide array of datasets from various time periods. They have different data values which can be utilized in order to predict the future outcome. Also, there are various Government websites available that produces crime-based statistics. These portals provide a range of data at various levels, including local, state, and national, making them viable options for use in our research. These websites consist of the up-to-date information regarding the crimes happened at a specific area and additional details. These reports can be downloaded and used for the analysis. These data encourage transparency and accessibility

to the official government data. Considering the above we tried looking for different data repositories and came across quite a few which help us for our model.

**Data format:** Input data can be on any format such as in text format wherein all the legal reports, statements or any other activity related to the criminal offence can be included. Data could also be in any image format wherein images from the crime scenes or there may be pictures of any evidence related to the offence. Similarly, graphical data are the ones' which can help us to compare the relationships between the entities. In general, data be can of any form. But the one which we require must be in a form which is accepted by the model and can be useful for the prediction of the crime. Hence there is transformation required to convert these into a standardized form to make an accurate model with better decision making.

**Solution:** Data Integration is a technique which can be used for converting the data into the reliable format. It combines the data from different heterogeneous sources which might be in different formats and structures and then gives it a merged format. ETL is a technique which is used in data integration process to maintain the data in an efficient way. ETL which stand for Extract, Transform and Load consists of three phases, first one is the extraction phase wherein data is extracted from the different sources in a raw form. In this the structure of the data would remain the same and these can be extracted from various APIs, websites, or databases. Next is the Transform phase which does the data cleaning and processing, it converts the original raw data into a particular format which can be used for analysis. It includes the data mining process such as cleaning, removing duplicates and normalization. This phase includes feature engineering wherein new features are added in way how it is required for the model. During the Third phase, the data is uploaded into the database or any platform where it is required and can be accessed for the future analysis. We have used the same technique to convert our data into the same form. Since data was collected from different sources, we combined them using the same into the csv format.

**Data Interpretability:** Machine Learning may have thousands of parameters which seems to be very complex and important as well. This makes them extremely complicated and makes the understanding even more difficult. Machine Learning and Deep Learning models have number of layers, and their structure are so complex that understanding them becomes very difficult and adding more logic keeps making them more complex. Many machine learning models such as neural network are called as black box because it is very complicated and becomes hard to understand on how they reach to a particular prediction or the result. There is no link between the output predicted from the model and the input data given to these algorithms. Hence, lack of interpretability can be a major problem here. Hence to make the model work we need to focus more on the way the model work and build it in a way that I would be trusted.

**Solution:** The predictions of several base models (weak learners) are combined by ensemble learning techniques like Random Forest, Gradient Boosting, or AdaBoost to produce a stronger, more accurate model. Each base model contributes its forecasts, which are combined by the ensemble using several methods like voting, averaging, and weighted averaging. The goal is to create the ensemble in such a way that interpretability is preserved without degrading its prediction performance. Below are the ways we used to accomplish this:

A. **Transparent Model Selection:** Picked the models for the ensemble that can be un-

derstood from the start. For instance, we used decision trees or linear models can be employed as basis models because of their transparent structure and understandable decision rules, which make them simpler to interpret.

B. **Ensemble Feature Importance:** Calculated the total feature importance for each base model in the ensemble. The elements that are important for the ensemble's predictions were collectively understood through methods like averaging the feature significance scores from several models.

C. **Rule Extraction from Ensemble:** Following ensemble training, we tried drawing human-interpretable rules from the basis models' collective wisdom. Rule extraction techniques can convert the ensemble's behaviour into a list of clear rules or decision-making processes.

**Accurate Instance Labeling:**The goal of classifying instances or data points into predetermined groups or categories based on their attributes is a fundamental challenge in machine learning. Each class has a specific label or result that it matches to. As an illustration, the classes in a spam detection issue may be "spam" and "not spam." Finding the best decision boundary to properly divide the classes in the feature space is the difficult part. Depending on the type of data and the issue at hand, the decision boundary may be linear, nonlinear, or more complicated. Accurate classification becomes more challenging when the decision boundary is complex or when examples of many classes overlap. When the classes in the dataset have considerably different numbers of instances, the data distribution is unbalanced. In comparison to other classes, one or more may have an excessively high or low number of instances. For example, in a dataset for fraud detection, the proportion of fraud instances may be substantially smaller than that of non-fraudulent cases. As they try to reduce overall error, models trained on unbalanced data are often biased towards the majority class. As a result, the minority class is frequently projected incorrectly. The model may attain high accuracy by correctly predicting the majority class, creating a false sense of competence. On the minority class, which is typically of more importance (e.g., finding uncommon illnesses), it could, however, perform badly. Due to the small number of cases, models may not successfully learn the characteristics and patterns associated with the minority class, resulting in predicting performance for that class.

**Solution:** The performance of a model can be considerably impacted by the machine learning method chosen, particularly when working with unbalanced datasets. Different algorithms naturally handle class imbalances differently. Hence, we tried considering working on the number of algorithms and tried applying different techniques for the same. Below are some of them which we tried understanding so that we could modify our algorithm in a good way:

A. **Ensemble techniques:** For unbalanced datasets, ensemble techniques like Random Forests, AdaBoost, and Gradient Boosting are frequently successful. They pool the weaknesses of several weak students to produce a more robust model, potentially enhancing performance in the minority class.Support Vector Machines (SVMs) can be useful when balanced class weights are used, according to research. In unbalanced situations, SVM enables the establishment of class weights to penalize misclassifications differentially.Neural networks may successfully manage unbalanced data when they are created with the right topologies (e.g., employing class weights, weighted loss functions, or certain network structures).

B. **Cost-sensitive Learning:** Some algorithms provide cost-sensitive learning, where the cost of a misclassification can be changed to give the minority class greater weight.

C. **Anomaly Detection Algorithms:** Using specialized anomaly detection algorithms and treating the minority class as an anomaly detection task might be advantageous depending on the issue.Standard algorithms can occasionally be changed to better handle skewed data. For instance, depending on the significance of the minority class, changing the decision threshold in a binary classification issue might favor sensitivity or specificity.

D. **Individualized Loss Functions:** The model may be directed to concentrate on the minority class by creating unique loss functions that penalize misclassifications of the minority class more severely than those of the majority class.

E. **Hybrid strategies:** Combining different algorithms or tactics might be useful. Performance may be enhanced, for instance, by employing a group of classifiers, each of which has been tuned to handle unbalanced data.

F. **Experimentation and evaluation:**It's critical to test out various algorithms and methods to see how they perform on the particular unbalanced dataset. In light of the dataset's imbalance, evaluate them using suitable assessment measures. The secret is to choose or alter an algorithm that fits the dataset's features and overcomes the problems caused by class imbalance. To make an educated decision on the algorithm to be utilized, experimentation and a complete grasp of the problem are necessary. Additionally, cross-validation and hyperparameter adjustment are crucial for enhancing the algorithm's performance on unbalanced datasets.

**Accuracy:** Ensuring the fairness and accuracy of the prediction model is crucial to prevent unfounded allegations or wrongful convictions. It is imperative to conduct rigorous testing and validation procedures to verify the effectiveness of the model. Clear guidelines should be established to interpret and explain the model's outcomes, and human oversight should be incorporated to avoid overreliance on automation. Transparency in the algorithmic processes and data used is essential, allowing for critical examination while respecting confidentiality and security measures. Establishing accountability mechanisms, such as routine audits and impartial evaluations, promotes public trust and facilitates the identification and rectification of any biases or deficiencies in the system.

**Solution:** The comprehensive testing and validation of the model is the foundation of dependability in the field of predictive modelling. To do this, we have exposed our model to a wide range of datasets that reflect different real-world settings and scenarios pertinent to the predictive system's application. By doing this, the model's generalizability was evaluated, guaranteeing that it can function well in a variety of circumstances. Additionally, a thorough assessment of the model's performance was provided by using well-recognized measures including precision, recall, F1-score, and accuracy. We understood the process of cross-validation, a method for evaluating the model's consistency across several data subsets, which helped us in preventing overfitting and increase trust in the resilience of the model. The process of validation also includes the discovery and mitigation of bias. As eliminating these biases is crucial to preventing discrimination and ensuring fairness in predictions, additional vigilance is exerted to discover any biases connected to sensitive traits like race, gender, or ethnicity. The model was put through robustness testing to see

how well it holds up to adversarial assaults and edge situations that simulate real-world difficulties. It is crucial to improve the model's performance by careful parameter tweaking as well as to test its scalability and responsiveness under pressure, ensuring that it will work at its best even during times of high usage. The model develops and enhances its accuracy and application by iterative refinement and validation against known patterns or historical data, ultimately strengthening its efficacy and dependability in forecasting outcomes in crime prediction.

Addressing the challenges outlined in this section is critical for developing an effective and ethical crime prediction model. Overcoming data quality issues, standardizing data formats, enhancing model interpretability, ensuring accurate instance labeling, and prioritizing accuracy through rigorous testing and validation are essential steps in harnessing the power of data-driven crime prediction for the benefit of public safety.Facing these challenges head-on led us to discover solutions that not only improved our understanding of prediction models but also enhanced the model itself, ultimately contributing to public safety and upholding ethical standards.

## 3.3 Project Team

Our project development team consists of Six members. Each member has a role assigned to them based on their technical capabilities. Though we will all take an equal part in development and coding, we have nonetheless decided to appoint a specialist for each important aspect of the development lifecycle. All the tasks are divided amongst team members and every member will work independently on their assigned tasks.

| **Role** | **Name** |
|---|---|
| Team Leader/ Project Manager | Namita Samant |
| Machine Learning Specialists | Adiba Hasin |
| Criminologists/Domain Expert | Varun Ravi Sankar |
| Quality Assurance Specialist | Fauzia Faria |
| Documentation Manager | Karishma Gosalia |
| Data scientist | Shruti Parmar |

**Team Leader/Project Manager:** Namita, an accomplished developer with exceptional leadership skills, took on the role of Team Leader and Project Manager. In this capacity, she assumed responsibility for all technical and project management activities. Namita played a crucial role in guiding the team's technical direction, coordinating development efforts, and overseeing the successful execution of the project. Her expertise was instrumental in aligning the project with best practices and achieving its technical objectives.

**Machine Learning Specialist:**Adiba, a skilled machine learning specialist, was responsible for developing accurate crime prediction models that played a vital role in crime prevention, resource allocation, and decision-making. Her expertise in data analysis and model development ensured the precision and effectiveness of the project's machine learning systems, benefiting law enforcement efforts in safeguarding communities.

**Criminologists/Domain Expert:** Varun, a seasoned criminologist with a deep understanding of the domain, collaborated with the team to enhance the accuracy, relevance, and interpretability of crime prediction models. By combining his domain expertise with AI

and machine learning techniques, Varun provided valuable insights for law enforcement agencies and policymakers.

**Quality Assurance Specialist:**Fauzia, in her role as the Quality Assurance Specialist, conducted rigorous testing, verified data integrity, and assessed system performance and usability. Her meticulous efforts ensured the overall quality and reliability of the AI-based crime prediction model. Fauzia played a pivotal role in identifying and addressing any issues to optimize the system's performance, ultimately delivering a user-friendly and dependable tool for law enforcement and stakeholders.

**Documentation Manager:**Karishma, as the Documentation Manager, played a crucial role in managing all project-related documentation, ensuring that records and information were organized, maintained, and accessible. Her expertise in documentation helped in maintaining transparency and accountability throughout the project..

**Data Scientist:** Shruti, a skilled Data Scientist, leveraged her expertise in statistical analysis, machine learning, and data manipulation. Her role involved understanding the data, selecting relevant features, developing robust models, evaluating performance, and generating actionable insights to support evidence-based decision-making in crime prevention.

## 3.4 Plan and Timeline

The project followed a systematic and well-structured plan that ensured the successful development of a crime prediction model for South Australia.The initial phase involved project initiation, where objectives and research questions were defined, and a comprehensive review of relevant literature was conducted. Following this, data collection and pre-processing were performed to obtain clean and reliable crime data. Exploratory data analysis was then conducted to gain insights into the data and identify any patterns or trends. Feature selection and engineering were carried out to identify the most relevant predictors for crime prediction. The main focus was on implementing the Random Forest algorithm for model development and optimization. The model's performance was evaluated and validated using appropriate metrics, and anti-crime measures were incorporated to enhance the precision and effectiveness of the model. The results were thoroughly analyzed and interpreted, and a comprehensive project report was prepared, documenting the methodology, findings, and recommendations. The project concluded with a presentation to stakeholders, ensuring their engagement and the dissemination of key findings for informed decision-making.

We diligently attempted to adhere to our initial tentative timeline for the proposed project. However, we encountered necessary adjustments to task durations due to various factors, including evolving project needs, data availability fluctuations, resource constraints, and the inherent complexity of the project. Throughout the project, we remained adamant in monitoring our progress and flexibly adapted the timeline to these changing circumstances. This adaptability allowed us to stay responsive and agile in the face of challenges. To ensure effective communication among team members and stakeholders, we held two group meetings per week. This practice ensured that everyone remained informed, engaged, and aligned with project goals and progress.

| Task | Subtask | Duration (Weeks) |
|---|---|---|
| Project Initiation | Defining project objectives and scope | 2 |
| | Conducting a thorough review of relevant literature | |
| | Finalizing the research questions and hypotheses | |
| Data Collection and Pre-processing | Identifying and acquire relevant crime data for prediction model | 3 |
| | Cleaning and pre-processing the data, including handling missing values and outliers | |
| Exploratory Data Analysis | Performing descriptive analysis of the crime data | 2 |
| | Identifying patterns, trends, and correlations among variables | |
| | Visualizing the data using appropriate charts and graphs | |
| Feature Selection and Engineering | Selecting relevant features that have significant impact on crime prediction | 2 |
| | Engineering new features if necessary to enhance the predictive power of the model | |
| Model Development | Implementing the suitable algorithms like KNN, Ridge Regression, Linear SVC, and Gaussian Naive Bayes etc for crime prediction. | 4 |
| | Training the model using the cleaned and pre-processed data | |
| | Optimizing the model parameters through cross-validation | |
| Model Evaluation and Validation | Evaluating the performance of the crime prediction model using appropriate metrics (e.g., accuracy, precision, recall) | 4 |
| | Validating the model using a holdout dataset and through cross-validation | |
| Incorporation of Anti-Crime Measures | Identifying relevant anti-crime measures and variables | 3 |
| | Integrating these measures into the prediction model | |
| | Assessing the impact of these measures on the model's performance | |
| Results Analysis and Interpretation | Analysing the findings from the crime prediction model and anti-crime measures | 3 |
| | Interpreting the results in the context of South Australia's crime patterns and characteristics | |
| | Identifying key insights and implications for law enforcement agencies and policymakers | |
| Reporting and Documentation | Preparing a comprehensive project report summarizing the methodology, findings, and recommendations | 4 |
| | Creating visualizations and charts to effectively communicate the results | |
| | Documenting the code and procedures for future reference | |
| Presentation and Project Demonstration | Presenting the project findings to relevant groups such as faculty members and peers, including law enforcement agencies and policymakers | 2 |
| | Engaging in discussions and seek feedback on the proposed crime prediction model and its implications | |
| | Addressing any questions or concerns raised | |

Figure 1: Project Timeline

# 4 Outcomes

The proposed research aimed to achieve the following outcomes:

**1. Development of an Accurate Crime Prediction Model:**

- The primary outcome of this research is to develop an accurate crime prediction model that incorporates anti-crime features.
- The model will utilize advanced machine learning algorithms, such as K-Nearest Neighbors (KNN), ridge regression, Linear Support Vector Classification (SVC), and Gaussian Naive Bayes to predict crime rates based on historical data and relevant variables.
- By incorporating anti-crime features into the model, it aims to enhance the accuracy and effectiveness of crime rate predictions.

**2. Evaluation of Model Performance:**

- The research will assess the performance of the crime prediction model by evaluating metrics such as accuracy, precision, recall, and F1 score.
- This evaluation will provide an understanding of the model's effectiveness and its ability to make reliable predictions.
- The outcomes will help validate the proposed approach and provide recommendations for further improvements.

**3. Facilitating User-Centric Interface Development:**

- The research outcome establishes the fundamental groundwork necessary for the development of a user interface that effectively utilizes the insights derived from our comprehensive research.

The overall outcome of this research is to develop an accurate crime prediction model incorporating anti-crime features, evaluate its performance, and establish the groundwork for a user-centric interface based on research insights.

# 5 Results

## 5.1 Initial Data Analysis and Preprocessing

Data Exploration and Preprocessing The analysis begins with the loading of crime data from the `2010-23-data-sa-crime.csv` file. The dataset is then examined and preprocessed as follows:

**Pairplot and Boxplot:** Initial exploratory data analysis is conducted by visualizing relationships between features. A pairplot is generated to explore the pairwise relationships between 'Offence count,' 'Taken to remand,' and 'Postcode - Incident.' A boxplot is created to investigate the distribution of 'Offence count' across different 'Offence Level 1 Description' categories. The figure given below illustrates the box plot of "Offence Count" vs "Taken to Remand".

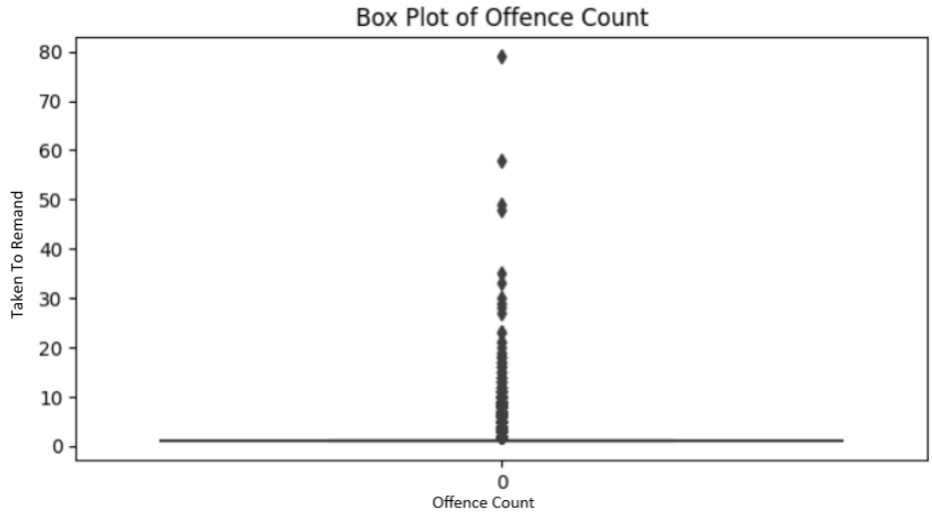


Figure 2: Box Plot

**Graph Insights:**

In Figure 2, the horizontal axis (x-axis) shows the "Offence count", while the vertical axis (y-axis) displays the number of individuals "Taken into Remand". We can see the distribution is focused in the first quartile and the median is towards the bottom. We generate some box plot for different crimes. From the results we understand how the outliers values are spread out.

The figure below represents the distribution of ‘Offence Count’ across different ‘Offence Level 1 Description’ categories.

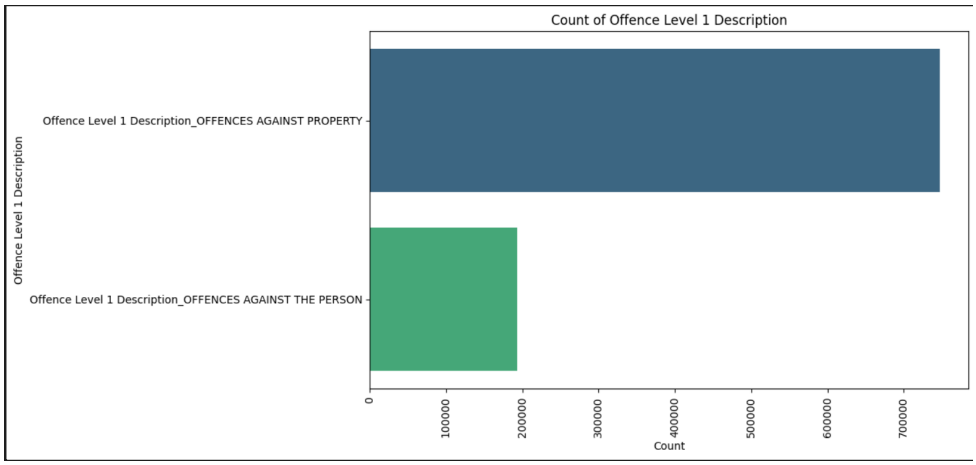


Figure 3: Box plot Of Detailed Offence Level Description

**Graph Insights:**

From Figure 3, we can conclude that the number of offence against property is higher then the offence against person.The number of offence against property is 70,0000 where as the number of offence against person is 20,0000. The x-axis represents the number of ”offence count” and y-axis represents the ”offence level 1 description”.The blue color represents the offence against property and the green color represents the offence against person.

**Handling Missing Values:** Missing values in the dataset are identified and dealt with by removing rows with missing data. The missing value counts for each column are printed to provide transparency about data quality.

**One-Hot Encoding:** Categorical columns are one-hot encoded to prepare the data for machine learning. This transformation is performed on ‘Suburb - Incident,’ ‘Offence Level 1 Description,’ ‘Offence Level 2 Description,’ and ‘Offence Level 3 Description’ columns.

**Data Type Conversion:** Data columns that should be numeric are converted to the appropriate data type, handling any potential data type errors.

**Feature Importance Visualization:** The decision tree model is visualized to display feature importance, providing insights into which features are most influential in predicting ‘Offence count.’

**Feature Selection with SelectKBest:** SelectKBest is used to select the top k features that are most informative for predicting ‘Offence count.’

**Gaussian Naive Bayes Classifier:** A Gaussian Naive Bayes classifier is trained with different max depth values to assess its accuracy using cross-validation. From the results of the plot we understood the impact of max depth on classifier accuracy. Which helped us to select the best classifier to get maximum accuracy.

**Train-Test Split and Prediction:** The final part involves splitting the data into training and testing sets, training a Gaussian Naive Bayes classifier, and making predictions. The model’s mean accuracy is reported.

**Classification:**

In order to evaluate various models by visualizing their performance we conducted classification task. Below are some of the main steps included:

**Classification Report and Confusion Matrix:** A classification report is generated to evaluate the classification model's performance on the test data. Additionally, a confusion matrix is created to visualize the model's true positive, true negative, false positive, and false negative predictions.

**Optimal Alpha for Ridge Regression:** Ridge Regression is performed with different alpha values, and the optimal alpha is determined. The R-squared scores for different max depths are visualized to assess the model's performance. optimal alpha refers to the smoothing parameter that provides the best fit to the data. This parameter controls the weight given to the most recent observation when updating the forecast.

| Optimal Alpha: 0.01 |
|---|
| R-squared Scores |
| 0.999 |

**Model Evaluation for Regression and Classification:** We evaluate models for both regression and classification and recorded their performance metrics. The chart below shows the comparison of K-NN Regression vs Ridge Regression.

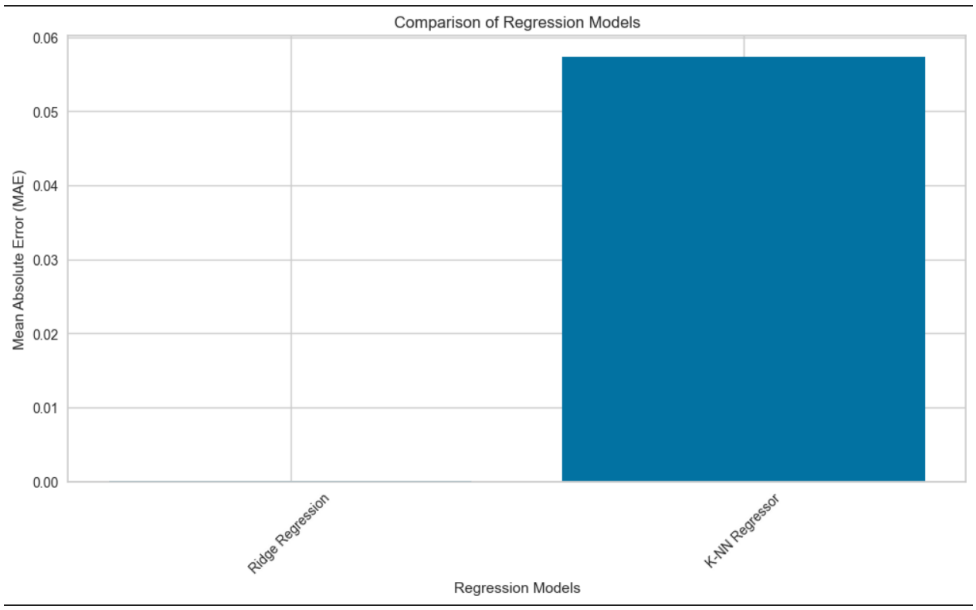


Figure 4: Comparison of Regression Models

**Chart Insights:**

Accordance with the chart Figure 4, we can conclude that, Ridge Regression model performs better than K-NN since the MAE (Mean Absolute Error) for K-NN is higher than Ridge Regression.

**R-squared ($R^2$) Score:** We calculated the R-squared score for each model. R-squared measures the proportion of variance in the dependent variable (Offence count) that is predictable from the independent variables. A higher R-squared indicates a better fit. Our R squared score for Ridge Regression is the following.

| R-squared Scores |
|---|
| 0.99 |

| Best Alpha | Best R-squared Score |
|---|---|
| 0.1 | 0.99 |

In general, an R-squared value of 0.99 is considered very high and indicates a strong relationship between the independent variable(s) and the dependent variable in a regression model. Our R-square value is 0.99 which is very high for Ridge Regression. Therefore, we can conclude that Ridge Regression is a better fit.

**Mean Absolute Error (MAE):** We calculated the Mean Absolute Error for regression models. MAE represents the average absolute difference between the predicted and actual values. A lower MAE is desirable as it indicates a more accurate model. MAE for K-NN and PCA is 0.0002645 which is pretty low hence K-NN and PCA are a good fit.

| Mean Absolute Error for KNN and PCA: | 0.0002645 |
|---|---|

**Classification Report:** For classification models, we generated a classification report. This report includes metrics like precision, recall, F1-score, and support for each class. It provides insights into the model's performance on different classes.

Classification Report for Random Forest Classifier:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 1.00 | 1.00 | 1.00 | 16511 |
| 2 | 0.99 | 0.99 | 0.99 | 1759 |
| 3 | 0.99 | 0.96 | 0.97 | 377 |
| 4 | 0.94 | 0.95 | 0.94 | 130 |
| 5 | 0.87 | 0.88 | 0.87 | 66 |
| 6 | 0.88 | 0.85 | 0.86 | 26 |
| 7 | 1.00 | 0.44 | 0.62 | 9 |
| 8 | 1.00 | 0.82 | 0.90 | 11 |
| 9 | 1.00 | 0.67 | 0.80 | 3 |
| 10 | 0.75 | 0.75 | 0.75 | 4 |
| 11 | 0.33 | 0.50 | 0.40 | 2 |
| 12 | 0.00 | 0.00 | 0.00 | 1 |
| 13 | 0.00 | 0.00 | 0.00 | 1 |
| 19 | 0.00 | 0.00 | 0.00 | 1 |
| accuracy | | | 1.00 | 18901 |
| macro avg | 0.70 | 0.63 | 0.65 | 18901 |
| weighted avg | 1.00 | 1.00 | 1.00 | 18901 |

Figure 5: Classification for Random Forest

**Figure Insights:**

In the classification report for the Random Forest, the accuracy is reported as 1.00, signifying 100 percent accuracy. Additionally, the report includes values for the macro-average and weighted average.

Classification Report for Support Vector Classifier:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 1.00 | 1.00 | 1.00 | 16511 |
| 2 | 1.00 | 1.00 | 1.00 | 1759 |
| 3 | 1.00 | 1.00 | 1.00 | 377 |
| 4 | 1.00 | 1.00 | 1.00 | 130 |
| 5 | 1.00 | 1.00 | 1.00 | 66 |
| 6 | 1.00 | 1.00 | 1.00 | 26 |
| 7 | 1.00 | 1.00 | 1.00 | 9 |
| 8 | 1.00 | 1.00 | 1.00 | 11 |
| 9 | 1.00 | 1.00 | 1.00 | 3 |
| 10 | 1.00 | 1.00 | 1.00 | 4 |
| 11 | 0.67 | 1.00 | 0.80 | 2 |
| 12 | 0.00 | 0.00 | 0.00 | 1 |
| 13 | 0.00 | 0.00 | 0.00 | 1 |
| 19 | 1.00 | 1.00 | 1.00 | 1 |
| accuracy | | | 1.00 | 18901 |
| macro avg | 0.83 | 0.86 | 0.84 | 18901 |
| weighted avg | 1.00 | 1.00 | 1.00 | 18901 |

Figure 6: Classification for Support Vector

**Figure Insights:**

Similarly, in the classification report for the Support Vector, an accuracy of 1.00, indicating 100 percent accuracy, is evident. The report also presents information on the macro-average and weighted average, providing a comprehensive overview of the model's performance.

Classification Report for Gaussian Naive Bayes:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.96 | 0.09 | 0.16 | 16511 |
| 2 | 0.05 | 0.04 | 0.05 | 1759 |
| 3 | 0.01 | 0.03 | 0.01 | 377 |
| 4 | 0.00 | 0.00 | 0.00 | 130 |
| 5 | 0.00 | 0.08 | 0.00 | 66 |
| 6 | 0.00 | 0.12 | 0.00 | 26 |
| 7 | 0.00 | 0.00 | 0.00 | 9 |
| 8 | 0.00 | 0.00 | 0.00 | 11 |
| 9 | 0.00 | 0.00 | 0.00 | 3 |
| 10 | 0.01 | 0.50 | 0.02 | 4 |
| 11 | 0.00 | 0.00 | 0.00 | 2 |
| 12 | 0.00 | 0.00 | 0.00 | 1 |
| 13 | 0.00 | 0.00 | 0.00 | 1 |
| 19 | 0.00 | 0.00 | 0.00 | 1 |
| accuracy | | | 0.08 | 18901 |
| macro avg | 0.07 | 0.06 | 0.02 | 18901 |
| weighted avg | 0.85 | 0.08 | 0.15 | 18901 |

Figure 7: Classification for Gaussian Naive Bayes

**Figure Insights:**

Finally, in the classification report for Gaussian Naive Bayes, an accuracy of 0.08, corresponding to 80 percent, is observed. This also includes values for both the macro-average and weighted average.

**Confusion Matrix:** We Visualized the confusion matrix for classification models. The confusion matrix shows the true positive, true negative, false positive, and false negative counts, helping to understand the model's ability to correctly classify instances.

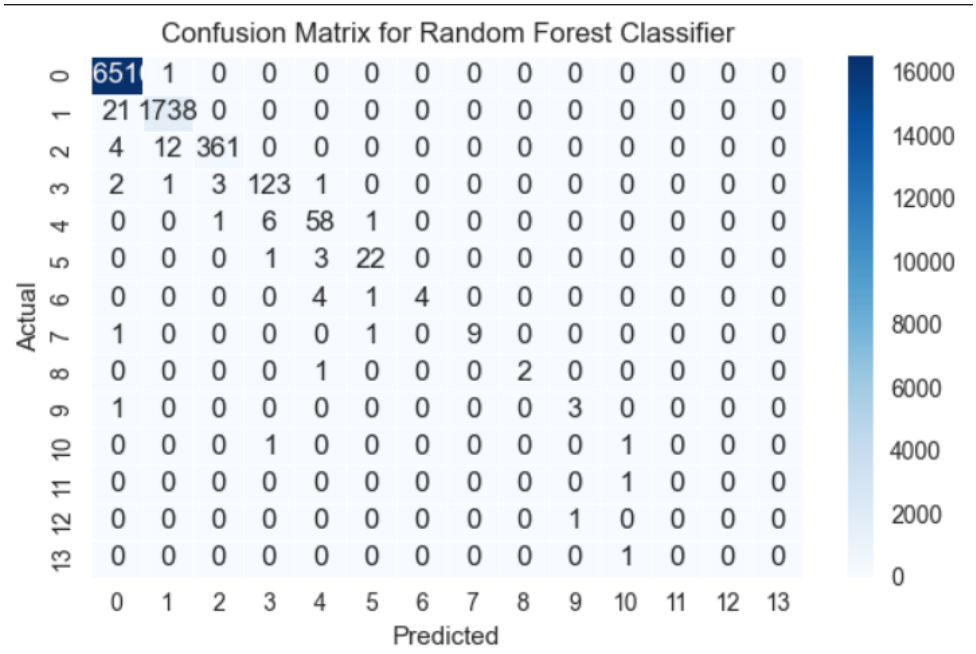


Figure 8: Confusion Matrix for Random Forest

**Figure Insights:**

Figure 8 displays the confusion matrix for the Random Forest Classifier. It effectively presents the counts of actual versus predicted values in Y axis and X axis respectively. Utilizing this confusion matrix, we determined the model's accuracy to be 97 percent.

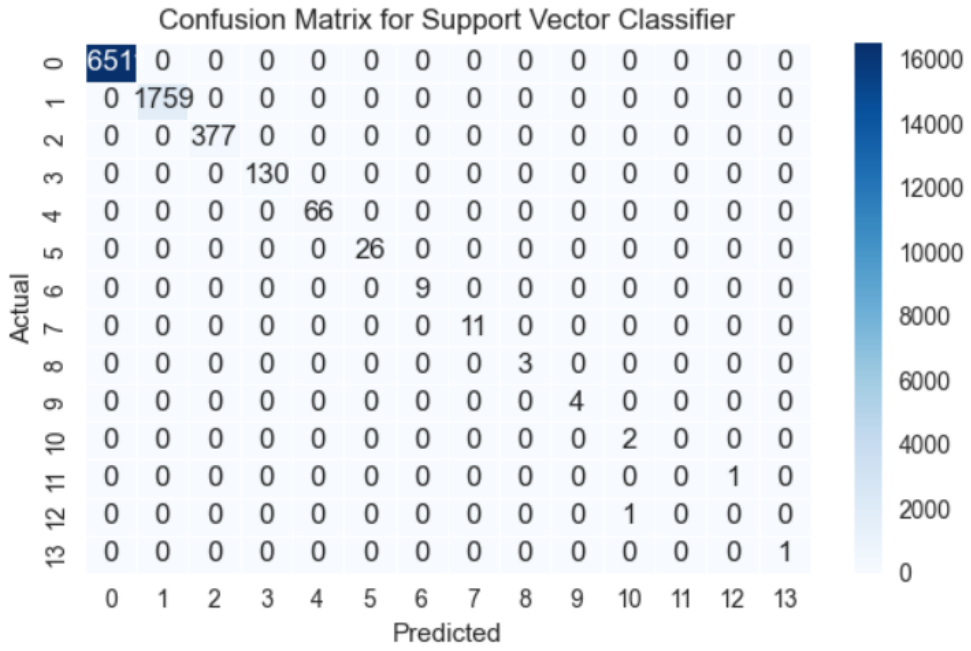


Figure 9: Confusion Matrix for Support Vector

**Figure Insights:**
Figure 9 represents the confusion matrix for Support Vector Classifier. Similarly to previous figure it shows the actual vs predicted count. From the confusion matrix we calculated the accuracy of the model which is 99 percent.

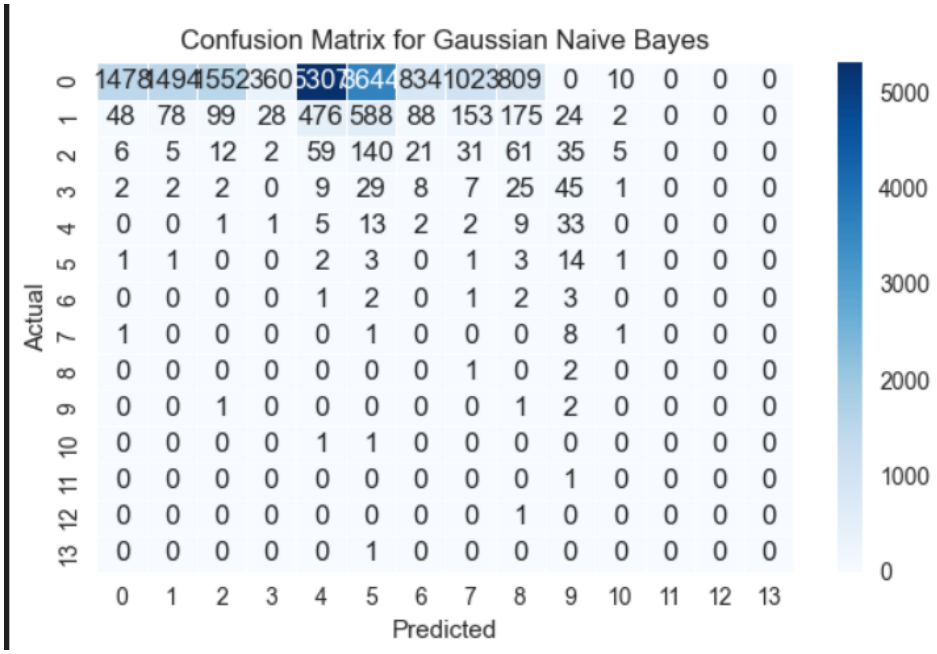


Figure 10: Confusion Matrix for Gaussian Naive Bayes

**Figure Insights:**
Finally Figure 10 represents the confusion matrix for Gaussian Naive Bayes classifier. We calculated the accuracy of the model which is 9.3 percent.

The analysis of the confusion matrices indicates that the Support Vector Classifier achieves the highest accuracy at 99 percent. The Random Forest Classifier also performs well, exhibiting an accuracy of 97 percent. In contrast, the Gaussian Naive Bayes Classifier demonstrates the lowest accuracy among the three classifiers, with a rate of 9.3 percent.

## 5.2 Prediction Improvement

At this phase, our objective is to validate and enhance the accuracy of our predictions. Specifically, we focus on the forecast for the year 2023 in three selected suburbs. Given the unique challenges of the COVID-19 years (2020-2022) and the significant disruptions to crime patterns, we chose to test our model on the year 2023. This decision was based on the anticipation that 2023 marked a return to more normal conditions, allowing us to evaluate the model's performance in a context closer to pre-pandemic circumstances. Figure 11 below illustrates the projected values for 2023 using the Linear Support Vector Machine regression model. Notably, the model consistently predicts values below 2.

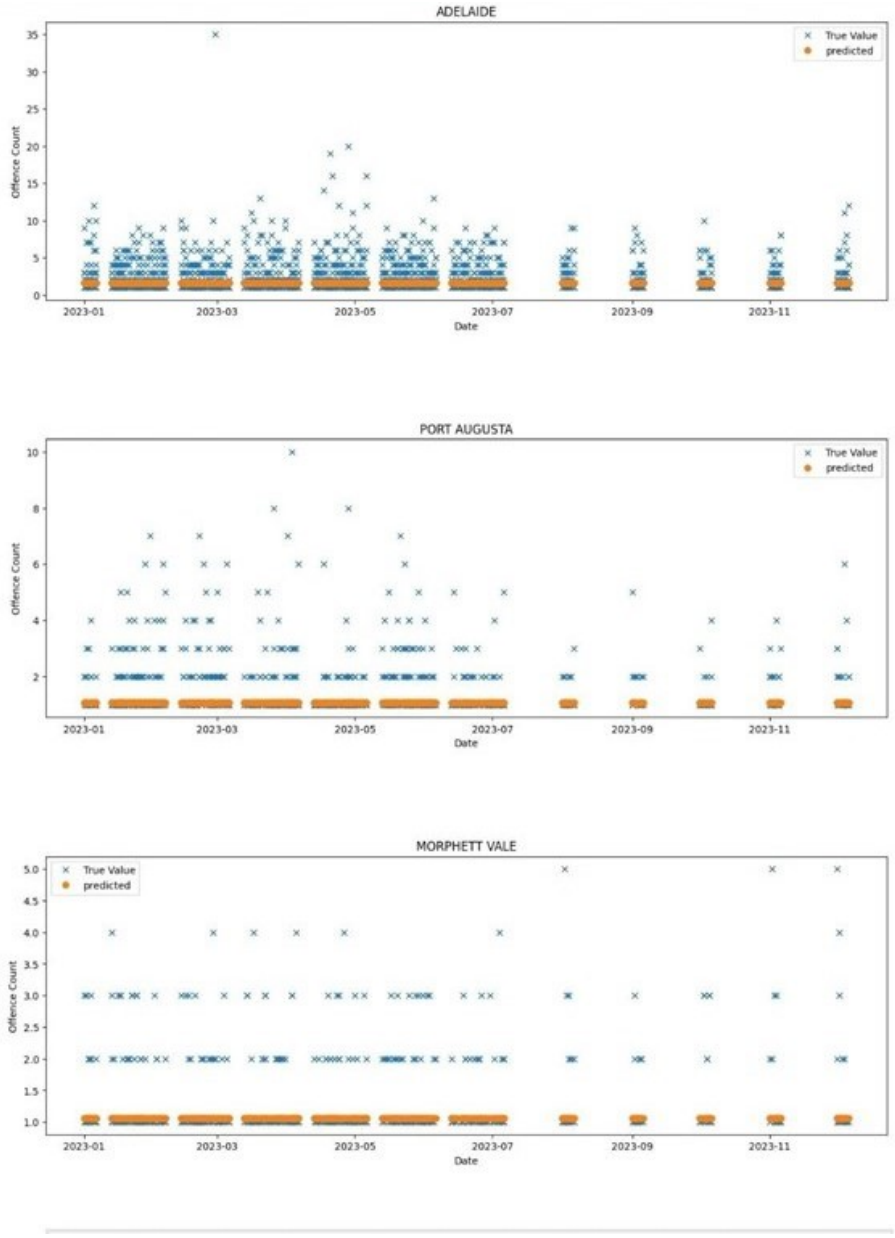


Figure 11: Linear Support Vector Machine

Subsequently, we forecast future data based on historical trends, yielding comparable outcomes. Consequently, we explore an alternative regression model, specifically the LightGBM, in an effort to enhance the accuracy of our predictions. LightGBM provides alot of flexibility and options to train a regression model, it uses boosting techniques to improve the efficacy of the model itself, by controlling the tree size, depth and nodes.

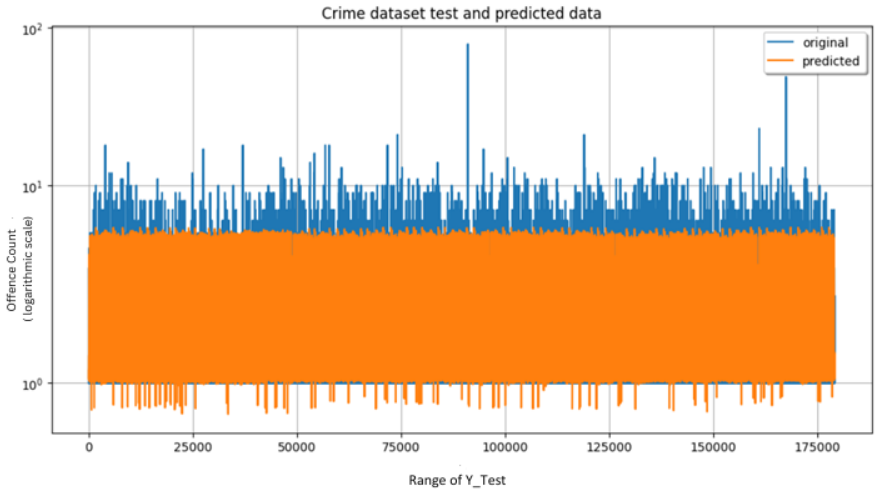


Figure 12: LightGBM Model

**Graph Insights:**

Figure 12 above is log scaled to show the prediction values compared with true values in the test set. The X-axis represents the range of y test, the Y-axis represents offence count.The count reflects the original values in y_test, simply mirroring the length of the y_test dataset. It's worth noting that including dates isn't feasible here, as the data from the training set is randomized, making it impractical to establish meaningful associations with date values. The model exhibits a modest improvement in the predictions with small spikes in the estimation count. However, it falls short of completely capturing the scale of the fluctuations in real-world values, even showing negative numbers. Despite these limitations, there is a slight positive trend in predicting crime rates. It's crucial to acknowledge that these predictions are based on past data, and unforeseen events could significantly influence the actual outcomes.

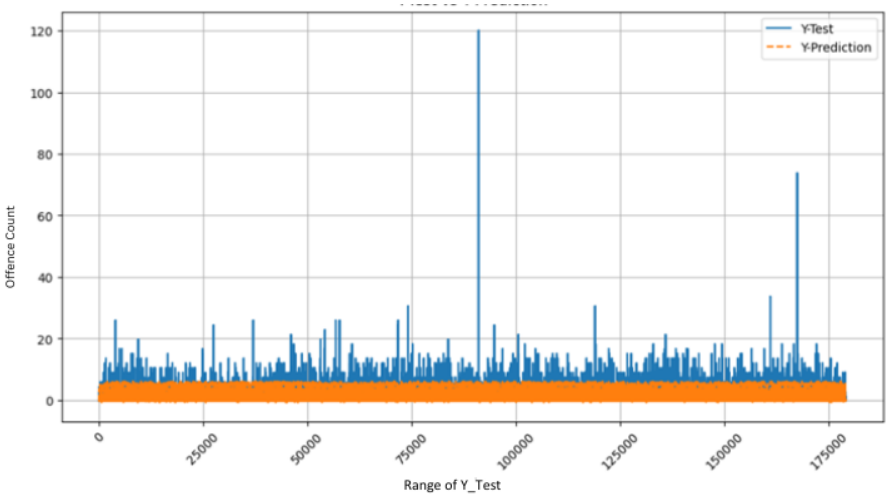


Figure 13: Linear scaled LightGBM

**Graph Insights:**

Figure 13 is similar to Figure 12 but displayed in a linear scale. In reality, values rarely reach such high levels, potentially distorting the chart. This is why we use logarithmic scaling to better observe the nuances of the LightGBM model. Without logarithmic scaling, we might overlook these finer details in the LightGBM model. Moving on to the predictions from the LightGBM model, the chart indicates a potential increase in crime compared to the current baseline. Looking into the future crime rates, the LightGBM model foresees a dynamic trend with both upward and downward fluctuations.

Finally we compare the predicted vs original values using Keras model for the first half of 2023

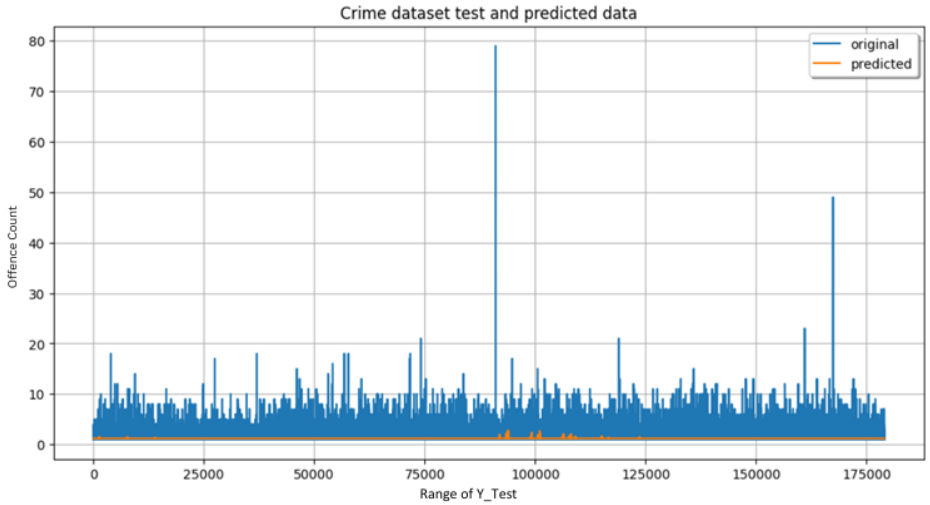


Figure 14: Keras Model

**Graph Insights:**

Figure 14, above, depicts the model output, affirming that the model is suitably fitted and performs well on the test set. However, given that this custom Keras neural network is tailored for crime rate estimation, we aim to assess its real-world performance. Figure 14 compares our custom Keras neural network's crime predictions for the first half of 2023 with actual data. Here, the X and Y axes represent the same parameters as in the previous Figures 12 and 13. While capturing the overall trend, the model notably diverges from reality during specific periods, especially during spikes and dips.

Navigating available 2023 Crime Data with the Keras Model shows varied results. While promising in following the general trend, noticeable disparities compared to actual data suggest potential areas for improvement.

## 5.3 Verification of Prediction

We trained the model using historical data from 2010 to 2022 and made predictions for the crime rates in the years 2023, 2019, and 2016. The graphical representations provide a detailed insight into the projections made for each of these years. Specifically, the illustrations highlight the forecasted values for the respective years (2023, 2019, and 2016).

Subsequently, we conducted a comparative analysis, contrasting the forecasted values with the actual results recorded in the first half of 2023. This iterative process of prediction and validation was similarly applied to the years 2019 and 2016, ensuring a comprehensive assessment of the model's accuracy against real-world data for multiple years.

```
data['Reported Date'] = pd.to_datetime(data['Reported Date'], format='%d/%m/%Y')
data['Year'] = data['Reported Date'].dt.year
X_train = data[data['Year'] <= 2022]['Year'].values.reshape(-1, 1)
y_train = data[data['Year'] <= 2022]['Offence count']
gradient_boosting = GradientBoostingRegressor(n_estimators=100, random_state=42)
gradient_boosting.fit(X_train, y_train)
years_to_predict = range(2016, 2024)
X_future = pd.DataFrame({'Year': years_to_predict})
future_predictions = gradient_boosting.predict(X_future)
actual_data = data[(data['Year'] >= 2016) & (data['Year'] <= 2024)]
for year in years_to_predict:
    plt.figure(figsize=(8, 6))
    plt.subplot(2, 1, 1)
    plt.bar(['Actual', 'Predicted'], [actual_data[actual_data['Year'] == year]['Offence count'].values[0],
                                      future_predictions[year - 2016]], color=['blue', 'red'])
    plt.title(f'Actual vs Predicted Offence Counts for {year}')
    plt.ylabel('Offence Count')
    plt.subplot(2, 1, 2)
    plt.plot(['Actual', 'Predicted'], [actual_data[actual_data['Year'] == year]['Offence count'].values[0],
                                       future_predictions[year - 2016]], marker='o', linestyle='-', color='green')
    plt.title(f'Comparison of Actual vs Predicted Offence Counts for {year}')
    plt.xlabel('Type')
    plt.ylabel('Offence Count')

    plt.tight_layout()
    plt.show()
```

Figure 15: Code snippet for the comparison of actual value vs predicted value

The following graphs represent the prediction comparison between actual value vs predicted value for year 2023, 2016 and 2019. For the depicted graph/bar chart, the offence count values represent incidents reported in a given suburb during a specific month for the given year for a specific type of offence.

The graph given below represents the prediction comparison of actual value vs predicted value for the first half of 2023 .

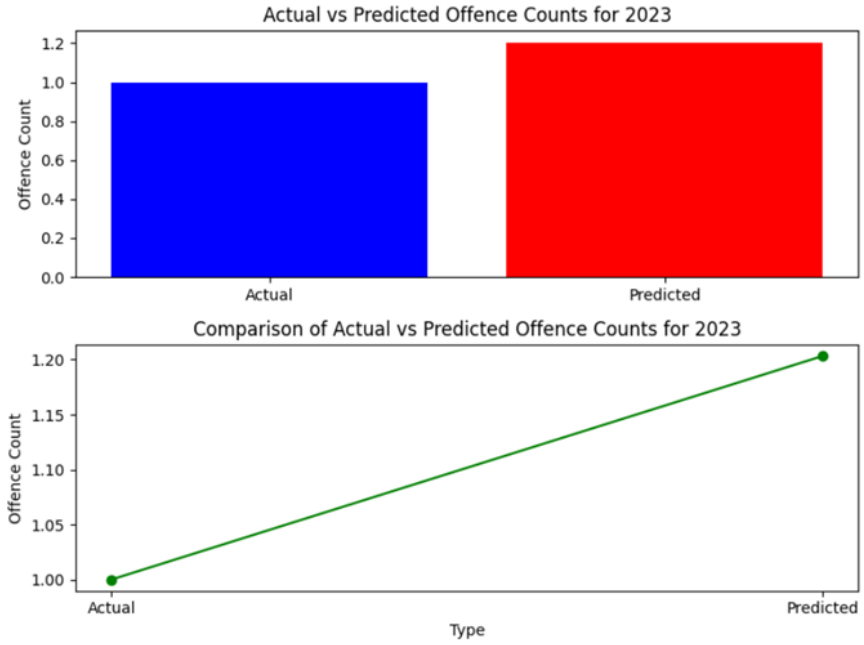


Figure 16: Comparison of Actual vs Predicated Offence counts for first half -2023

**Graph Insights:**

**Chart 1**

In the depicted chart, the actual crime rate is denoted by the blue colour, while the predicted crime rate is illustrated in red. The X-axis corresponds to different offense types, and the Y-axis signifies the count of offenses. The actual crime prediction stands at 1, and the forecasted crime rate is marginally higher at 1.2. This proximity indicates the model's commendable accuracy in predicting crime rates. The graph serves as a visual representation of the comparison between actual and predicted crime rates for the year 2023.

**Chart 2**

The second graph provides a clearer contrast between the actual and predicted values, showcasing an actual value of 1 compared to a predicted value of 1.2. The X-axis delineates offense types, while the Y-axis represents offense counts. This proximity to the actual crime rate indicates the model's precise predictive capabilities. The graph serves as a visual representation, offering a more distinct comparison between actual and predicted crime rates.

The graph given below represents the comparison of actual value vs predicted value for year 2019 .

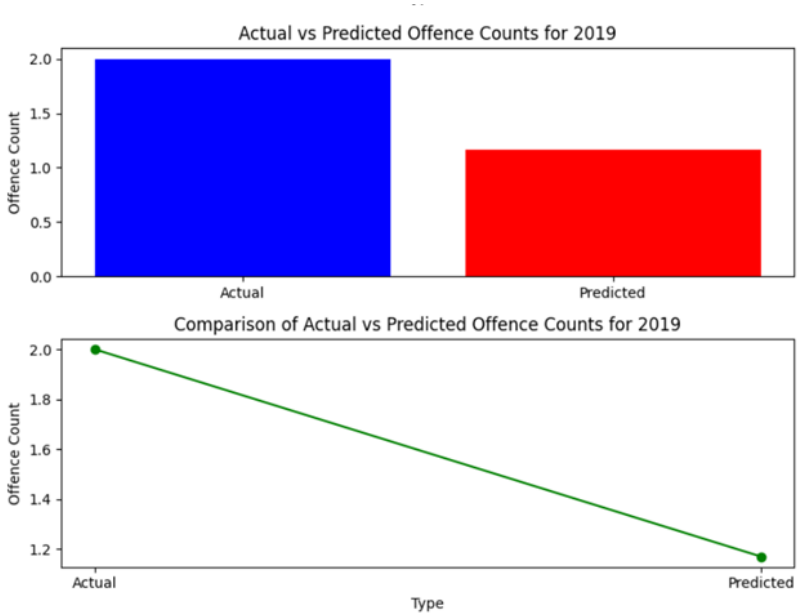


Figure 17: Comparison of Actual vs Predicated Offence counts for 2019

**Graph Insights:**

**Chart 1**

Similarly, in this visual representation, the actual crime rate is represented by the blue color, contrasting with the predicted crime rate shown in red. On the X-axis, various offense types are categorized, while the Y-axis indicates the count of offenses. The actual prediction for crime is 2, while the forecasted crime rate is 1. This striking similarity highlights the model's exceptional precision in foreseeing crime rates for the year 2019.

**Chart 2**

The second graphical illustration enhances the clarity of the comparison between actual and predicted values. It exhibits an actual value of 2 in contrast to a predicted value of 1. The X-axis categorizes offense types, and the Y-axis quantifies offense counts. representation provides a more vivid and nuanced comparison between actual and predicted crime rates.

Finally, the graph given below represents the comparison of actual value vs predicted value for year 2016.

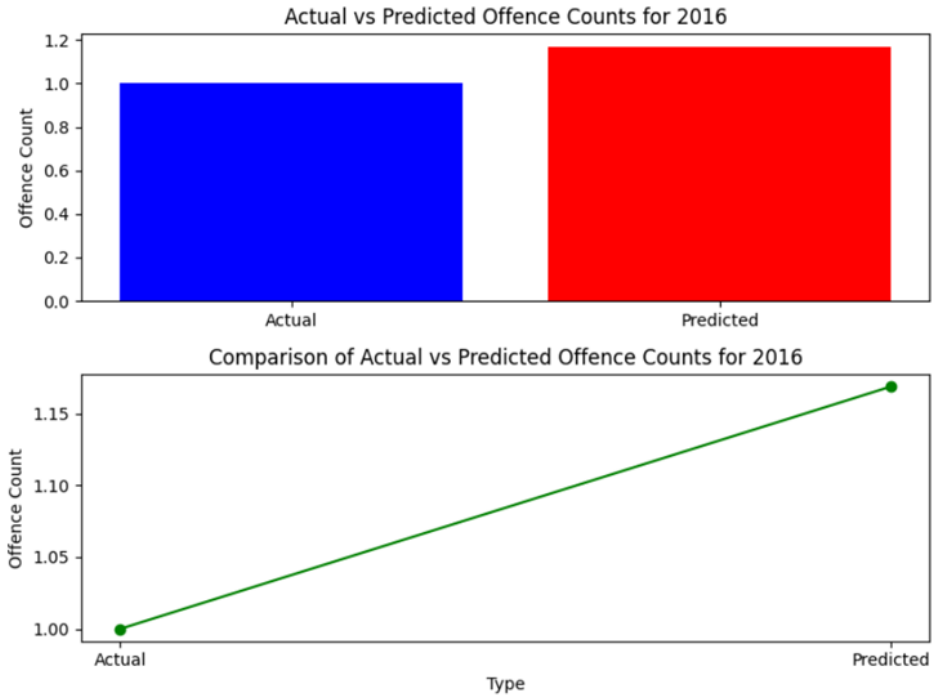


Figure 18: Comparison of Actual vs Predicated Offence counts for 2016

**Graph Insights:**

**Chart 1**

In this representation, the blue colour signifies the actual crime rate, while the red colour represents the predicted crime rate. The X-axis categorizes different offense types, and the Y-axis indicates the count of offenses. The actual crime prediction is 1, and the forecasted crime rate is slightly higher between 1.1 to 1.2. This closeness underscores the model's commendable accuracy in predicting crime rates, providing a visual depiction of the comparison for the year 2016.

**Chart 2**

The second graph offers a more distinct comparison between actual and predicted values, with an actual value of 1 compared to a predicted value of 1.1 to 1.2. The X-axis details offense types, and the Y-axis portrays offense counts. The proximity to the actual crime rate highlights the model's precise predictive capabilities. This graph serves as a visual representation, offering clarity in comparing actual and predicted crime rates.

Figure 19 below presents an alternative perspective on the comparison between actual and predicted values for the year 2016

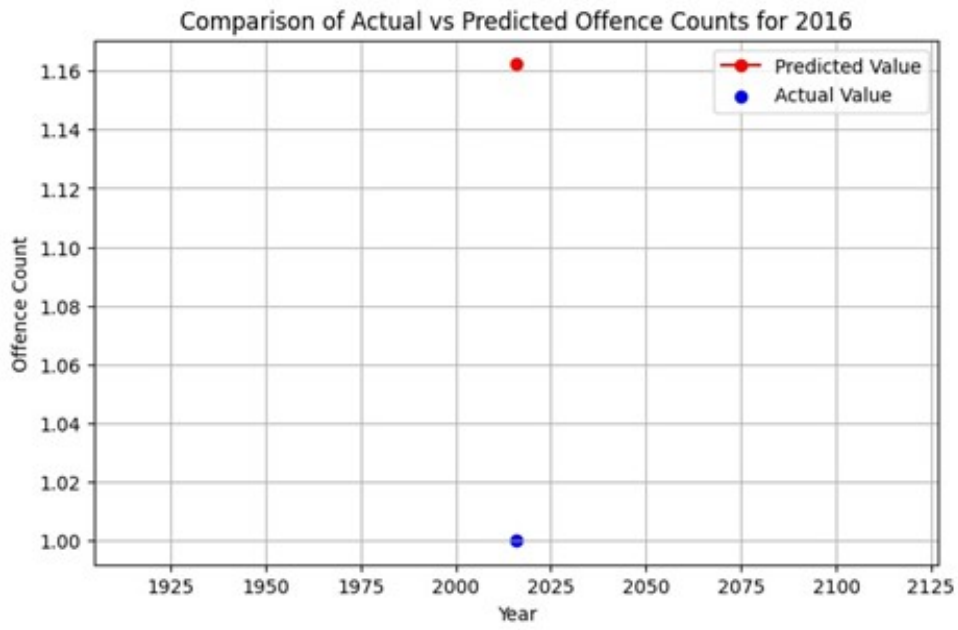


Figure 19: Comparison of Actual vs Predicated Offence counts for 2016

**Graph Insights:**

The Y-axis illustrates the offense count, while the X-axis denotes the year. In this representation, the actual values are depicted in blue, and the predicted values are represented in red. The minimal difference observed between these two values suggests a high level of accuracy in the prediction.

As an additional step in validating our model's predictions, For the suburb Birdwood, we forecasted the crime rate for the year 2023, yielding the following outcomes. Subsequently, we conducted a thorough examination, comparing the predicted value with the actual value to assess the model's accuracy. The observed crime rate stands at 31, and the forecasted value aligns closely at 31, indicating a high level of precision in the model's predictions.

```
Predicted Offence Count for 2023: 30
Actual Offence Count for 2023: 31
```

Figure 20: Comparison of Actual vs Predicated Offence counts for Suburb(Birdwood)

To show the comparison of actual data vs predicted data for the year 2023 we performed some additional steps to visualize the results we plotted the graph given below:

```
# x_ax = range(len(future_df_pca_Y))
plt.figure(figsize=(12, 6))
plt.plot_date(future_df.loc[future_df['Date'].dt.year == 2023, 'Date'], Y_scaler.inverse_transform(future_df_pca_Y.reshape(
plt.plot_date(future_df.loc[future_df['Date'].dt.year == 2023, 'Date'], Y_scaler.inverse_transform(future_df_pca_y_pred.resh
# plt.yscale('symlog')
plt.title("Crime dataset test and predicted data")
plt.xlabel('X')
plt.ylabel('Offence Count')
plt.legend(loc='best',fancybox=True, shadow=False)
plt.grid(True)
plt.show()
```

Figure 21: Code snippet for the comparison

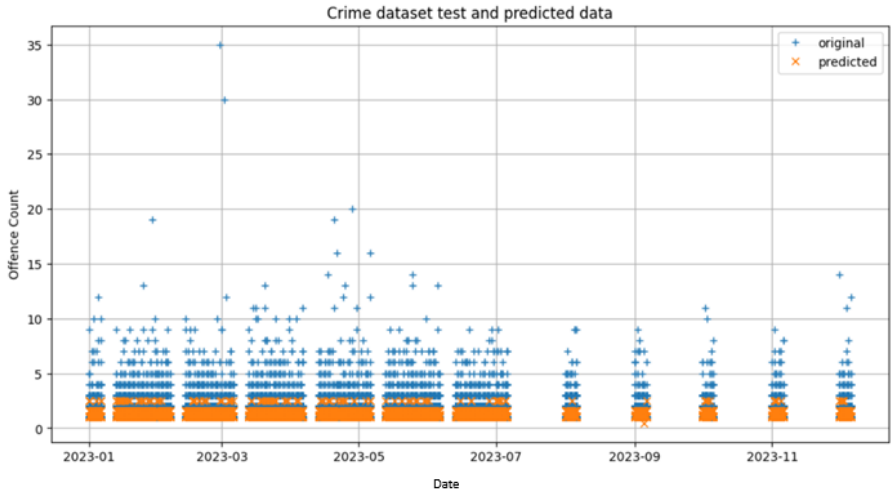


Figure 22: Original vs Predicted data Comparison

**Graph Insights:**

Figure 22 shows that comparison of original data for the year 2023 vs predicted data of the year 2023. Here the y-axis represents the number of offence count whereas x -axis represents the year–month. The blue color represents the original data and orange color represents the predicted data. The SVM model predicted a steady amount of crime in 2023, without any major increases or decreases. It paints a calmer picture, suggesting that crime predictions will stay below the usual levels. Compared to the past, the model indicates a potential decrease in crime throughout the year.

Upon contrasting the actual data with the predictions made by SVM and LightGBM models for the year 2023 individually, we proceeded to compare the outcomes of both models with each other.

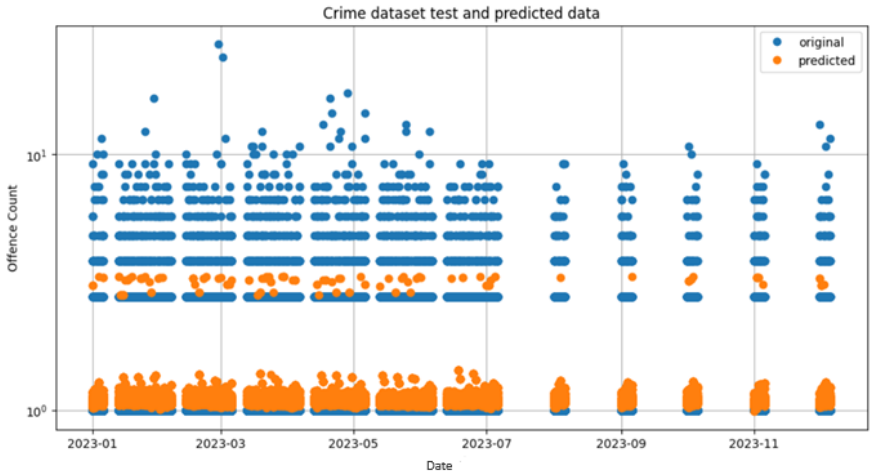


Figure 23: SVM vs LightGBM

**Graph Insights:**

Figure 23 compares crime predictions for 2023 from our LightGBM and SVM models with actual data. Both models show a similar overall trend, indicating stable crime rates throughout the year. However, the LightGBM model captures minor variations better, while the SVM model provides a smoother line. Overall, both models effectively capture the general trend, with LightGBM excelling in detecting smaller fluctuations.

Finally we compared original vs predicted data for the year 2023 with Keras model.


Figure 24: Keras Model

**Graph Insights:**

Figure 24 examines the 2023 crime predictions generated by Keras model in comparison to the actual data. While the model generally aligns with the real trend, it tends to underestimate certain peaks and overlooks some dips. The visualization offers insights into how Keras neural network navigates the 2023 crime data, showcasing its ability to capture the overall upward trend of the actual data, albeit at the expense of smoothing out some of the smaller fluctuations. Despite the model's proficient grasp of the general trend, it encounters difficulties in precisely capturing intricate nuances in the data, especially during periods of swift change. This underscores the inherent challenges in predicting complex real-world phenomena and underscores the necessity for continued refinement and exploration of the model's capabilities. Moving forward, it prompts a call to action for further optimization and in-depth exploration.

## 5.4 Predictions For Year 2024

We predicted the crime rates for the year 2024 for three different suburbs in South Australia. The graphs given below illustrates the predictions.

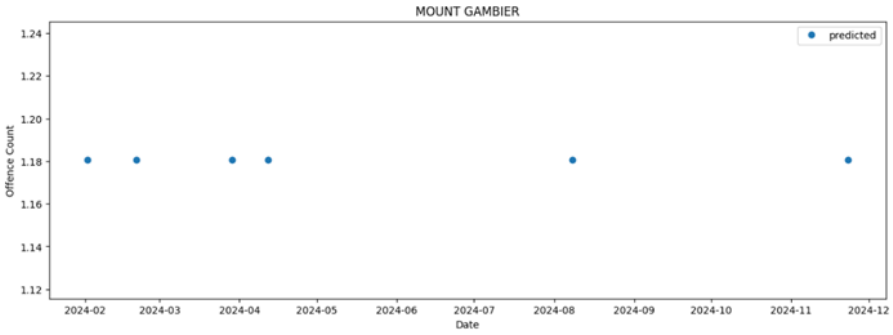


Figure 25: Mount Gambier

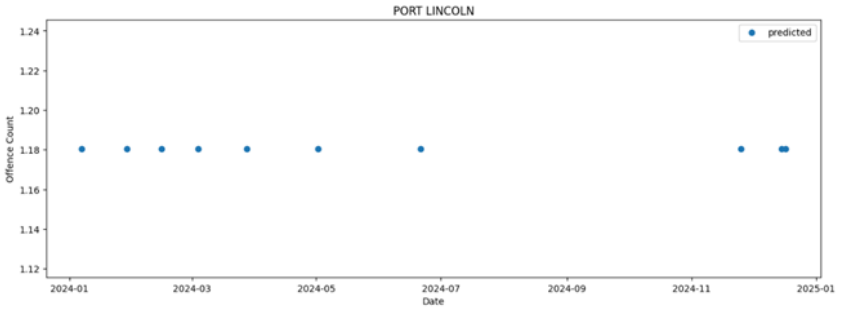


Figure 26: Port Lincoln


Figure 27: Salisbury

**Graph Insights :**

Figure 25-27 shows the crime rate prediction for 3 suburbs namely Mount Gambier, Port Lincoln, Salisbury respectively. The blue colour shows the prediction value for crime rates in particular suburb , Y-axis represents the number of offence count while X -axis represents the months of year 2024. As clearly visible above the crime rates value is on average 1.18 for all 3 suburbs where some of the months have high offence count whereas few months have low offence count.

We utilized the model to generate predictions for crime rates in three distinct suburbs—Adelaide, Enfield, and Kingswood. The following section presents the results of these predictions-

```
PS C:\Users\namit\open cv>  & 'C:\Python312\python.exe' 'c:\Users\namit\.vscode\extensions\ms-py
thon.python-2023.22.1\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '62404' '--
' 'C:\Users\namit\open cv\results10.py'
Enter the suburb: ADELAIDE
```

Figure 28: User Input

```
PS C:\Users\namit\open cv>  & 'C:\Python312\python.exe' 'c:\Users\namit\.vscode\extensions\ms-py
thon.python-2023.22.1\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '62404' '--
' 'C:\Users\namit\open cv\results10.py'
Enter the suburb: ADELAIDE
C:\Users\namit\open cv\results10.py:18: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/in
dexing.html#returning-a-view-versus-a-copy
  suburb_data['Date'] = pd.to_datetime(suburb_data['Date'])
Predicted Offence Count for ADELAIDE in 2024: 4655.30
PS C:\Users\namit\open cv>
```

Figure 29: Predicted Offence Count for Adelaide

In Figure 28, we see the desired user input is Adelaide. In Figure 29, we get the prediction result. The number of offence count in Adelaide in the year 2024 for offence against property and person is predicted to be 4655.30

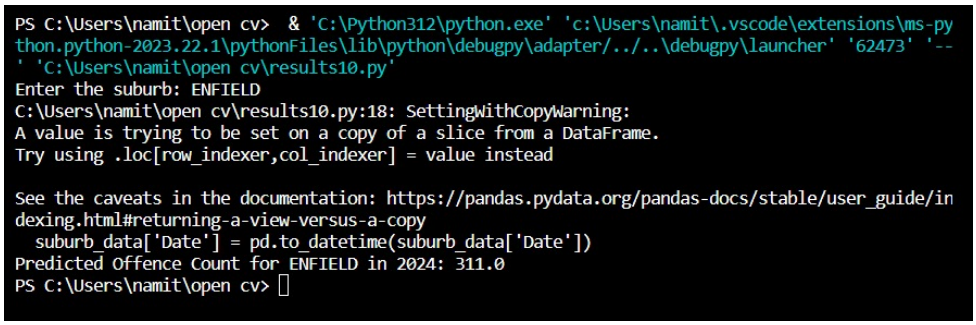


Figure 30: Predicted Offence Count For Enfield

Similiarly, for input Enfield, the predicted number of offence count in the year 2024 for offence against property and person is 311.0.

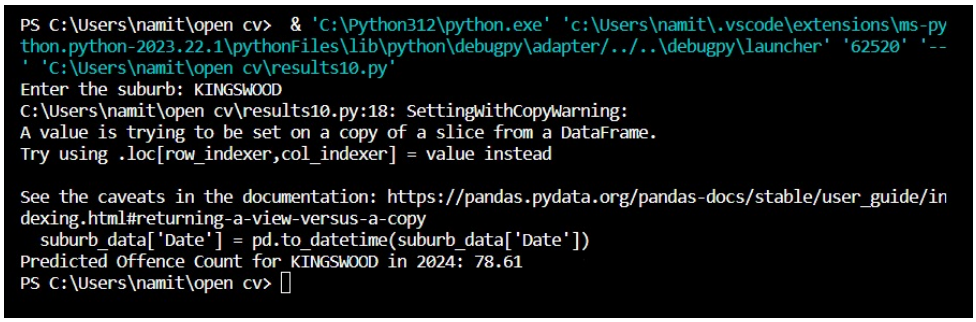


Figure 31: Predcited Offence Count for Kingswood

Lastly, for Kingswood, the number of offence count for offence against property and person is predicted to be 78.61.

Upon reviewing the predictions for various suburbs in South Australia, it is evident that Adelaide exhibits the highest count of offenses, while Kingswood, conversely, displays the lowest incidence of offenses. Which makes Kingswood seemingly safest suburb in South Australia.

## 5.5 Dashboard

Our prediction model is designed to provide accurate crime rate forecasts for any specified suburb in South Australia region. Users simply input the suburb name and select the desired year for predictions related to both property and personal crimes. This predictive tool aims to offer valuable insights and assist in strategic decision-making for safety and security measures. It's important to note that the data is not limited to South Australia; rather, it encompasses information from all states. Looking ahead, our goal is to extend this predictive tool to include crime predictions and enhance our understanding of crime rates for the entirety of Australia.

A dashboard brings together different views, providing an interface to explore a wide range of data points simultaneously. To understand crime dynamics in South Australia, we used

Tableau to integrate complex data, allowing us to explore crime rates in various suburbs and forecast future trends.

DASHBOARD LINK: Crime Rate Dashboard

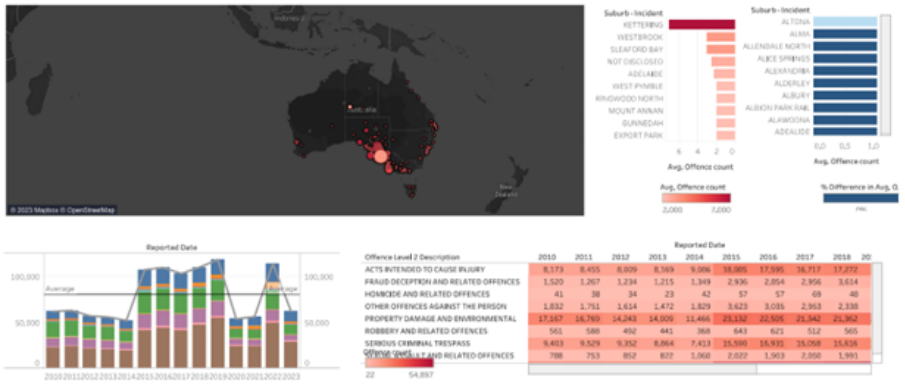

Figure 32: Dashboard for Crime Rates across Suburbs

**Key Findings**

**1. Temporal Trends:** Our analysis shows that 2019 had the highest reported crimes. This insight directs us to investigate the factors contributing to this spike in incidents.

**2. Suburb-specific Analysis:** "Kettering" stands out as the suburb with the highest reported crimes, prompting us to explore its unique socio-economic factors. On the other hand, "Export Parks" is the safest suburb, and understanding its safety profile is crucial.

**3. Prevalence of Offenses:** "Theft and Related" offenses consistently top the Level 2 category from 2010 to 2023. This highlights the need to address theft-related offenses by understanding their root causes.

**4. Practical Implications:** These insights go beyond statistics, offering practical significance. They provide valuable tools for developing strategies that prioritize safety. Law enforcement can use this information to implement proactive plans for a safer society.

**5. Predictive Accuracy and Future Applications:** The predictive model used in this analysis shows impressive accuracy, making it a robust tool for forecasting crime trends. This capability has promising applications in crime forecasting, empowering authorities to address emerging challenges proactively.

## 5.6 Limitation

There were a number of noticeable restrictions and difficulties that we ran into while working on the project. The availability and quality of historical crime data for our initial region of Australia was a serious limitation. Despite our best efforts, there were occasionally gaps and discrepancies in the data, which made it difficult to train and evaluate the models. A careful analysis and specialised procedures were also needed to overcome the class imbalance in crime statistics, where some crimes are much less common than others. Further-

more, the interpretability of the Random Forest algorithm posed a challenge when trying to explain the reasoning behind specific predictions, which is crucial for gaining the trust of law enforcement and policymakers.These drawbacks demonstrate the need for continued data collecting advancements and additional study of interpretive machine learning models for crime prediction. We suggest further developing our model to solve these drawbacks, with an emphasis on enhancing interpretability and reliability as well as adding an anti-crime component. The model's overall usefulness could be improved by this component, which could provide proactive suggestions and methods for preventing crime.

## 5.7 Future Work

As we wrap up this phase of our project, our team is optimistic about the continued development of our crime prediction model. We are excited to discover its practical application and the positive impact it can have on law enforcement efforts in South Australia. Looking ahead, there are several opportunities for future enhancements that we are keen to explore. Firstly, we wish to incorporate real-time data sources, resulting in a new era of efficiency and responsiveness. This will empower law enforcement agencies with the most up-to-date information, ensuring that they are equipped to respond promptly to evolving security concerns. Ethical considerations remain as a core value of our mission. We will continue to maintain the ethical use of our model by monitoring for bias and fairness, implementing measures to mitigate disparities in predictions, and aligning with best practices in responsible AI. We're also hopeful to develope an accessible interface fitted for law enforcement and policymakers, simplifying interactions with the model. Additiontally our vision extends beyond data and algorithms. We are eager to explore the potential of social media and open-source data, sources that can provide valuable information about local events, gatherings, and community sentiment that may impact crime rates. The contextual influences on crime can be better understood with the usage of such technique. Our team is excited to start this next stage of the project because it has the potential to enhance the capabilities of our crime prediction model, preferably not limited to South Australia.

# 6 Conclusion

In conclusion, we believe that our project has the potential to advance the development of a reliable crime prediction model for South Australia. We have shown how machine learning techniques can help with decision-making, resource allocation, and crime prevention. The model's performance indicators prove how well it predicts and locates crime hotspots, which helps law enforcement activities be more effective.As we move forward, we offer several recommendations to further enhance the model's impact. Firstly, we recommend the continued monitoring and refinement of the model to ensure its accuracy and reliability. Furthermore, prioritizing the integration of additional anti-crime features into the model is crucial, as the inclusion of such features significantly enhances the model's predictability by accounting for their impact. It is crucial to continue a solid cooperation with law enforcement organisations and subject matter specialists. The model will need to be adjusted, new difficulties will need to be addressed, and it will need to be relevant to real-world crime prevention initiatives. This will require collaboration and continual feedback.Additionally, it is crucial to continue a solid cooperation with law enforcement organisations and subject matter specialists. The model will need to be adjusted, new difficulties will need to be addressed, and it will need to be relevant to real-world crime prevention initiatives. This will require collaboration and continual feedback.

Our crime prediction model, we believe, has the potential to dramatically impact crime reduction and improve public safety in South Australia with further development and the deployment of anti-crime measures.We look forward to continuing this important work and making a positive contribution to our community's well-being.

# References


[1] Ahishakiye, E., Taremwa, D., Omulo, E.O. and Niyonzima, I., 2017. Crime prediction using decision tree (J48) classification algorithm. International Journal of Computer and Information Technology, 6(3), pp.188-195.

[2] Australian Bureau of Statistics. (n.d.). Crime and justice statistics. Retrieved from https://www.abs.gov.au/statistics/people/crime-and-justice

[3] Australian Institute of Criminology. (n.d.). Australian Institute of Criminology. Retrieved from https://www.aic.gov.au/

[4] Brantingham, P.L. and Brantingham, P.J., 2017. Environment, routine, and situation: Toward a pattern theory of crime. In Routine activity and rational choice (pp. 259-294). Routledge.

[5] Bureau of Crime Statistics and Research (BOCSAR). (n.d.). Retrieved from https://www.bocsar.nsw.gov.au/

[6] Dakalbab, F., Talib, M.A., Waraga, O.A., Nassif, A.B., Abbas, S. and Nasir, Q., 2022. Artificial intelligence and crime prediction: A systematic literature review. Social Sciences and Humanities Open, 6(1), p.100342.

[7] KIM, K.S. and JEONG, Y.H., 2021. A study on crime prediction to reduce crime rate based on artificial intelligence. Korea Journal of Artificial Intelligence, 9(1), pp.15-20.

[8] Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P. and Tita, G.E., 2011. Self-exciting point process modeling of crime. Journal of the American statistical association, 106(493), pp.100-108.

[9] Ratcliffe, J.H., 2004. Geocoding crime and a first estimate of a minimum acceptable hit rate. International Journal of Geographical Information Science, 18(1), pp.61-72.

[10] South Australia Crime Statistics Dataset. Retrieved from `https://data.sa.gov.au/data/dataset/crime-statistics`