Aishwarya Hiremath

Professor Dr. Krishna Bathula

Capstone Project

23 Aug 2024

Crime Spotter: Intelligent Crime Hotspot Detection

Abstract

This capstone project focuses on developing machine learning (ML) models for crime hotspot detection and spatial-temporal analysis using the New York City Police Department (NYPD) arrest dataset from NYC Open Data. The project hypothesizes that advanced ML techniques can predict crime hotspots with greater accuracy than traditional methods, leveraging comprehensive arrest data. ML algorithms machine learning algorithms, including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Gradient Boosting Classifier, Logistic Regression will be implemented and evaluated using metrics such as accuracy, precision, recall, F1-score, and Confusion Matrix. The significance of this research lies in its potential to revolutionize crime prevention strategies by enabling law enforcement agencies to allocate resources more effectively, enhance public safety measures, and potentially reduce crime rates in urban environments. This project aims to contribute to the field by demonstrating the superior performance of ML-based approaches in crime prediction compared to conventional methods.

Introduction

In modern urban environments, the effective prediction and prevention of crime remain paramount challenges for law enforcement agencies worldwide. Leveraging advancements in

machine learning (ML) this project aims to address these challenges by developing robust models for crime hotspot detection and spatial-temporal analysis using the NYPD arrest dataset from NYC Open Data. The primary goal is to equip law enforcement agencies with actionable insights to allocate resources strategically and enhance proactive crime prevention measures.

The hypothesis driving this research is that ML models, when trained on comprehensive arrest data, can accurately predict crime hotspots, and analyze temporal trends with higher precision than traditional statistical methods. By harnessing the wealth of data provided by the NYPD arrest dataset, this project proposes to implement a range of ML algorithms, including K-Nearest Neighbor (KNN),Support Vector Machine (SVM), Naive Bayes (NB), Gradient Boosting Classifier, Logistic Regression These models will be rigorously evaluated using metrics like accuracy, precision, recall, F1-score and Confusion Matrix to quantify their effectiveness in crime prediction.

The significance of this project lies in its potential to revolutionize crime prevention strategies. Law enforcement agencies, urban planners, and policymakers stand to benefit significantly from the insights generated. By accurately identifying crime hotspots and understanding temporal patterns, authorities can allocate resources more efficiently, enhance public safety measures, and potentially reduce crime rates. Furthermore, the methodology and findings of this research could serve as a blueprint for similar initiatives in other urban settings facing similar challenges.

Previous research in crime prediction has predominantly utilized traditional statistical methods or basic ML techniques. While these approaches have provided valuable insights, they often struggle with the complexity and non-linearity of crime patterns. This project differentiates itself by leveraging advanced ML techniques capable of learning intricate spatial and temporal

dependencies within crime data. By comparing the performance of the ML models against historical crime data, this research aims to demonstrate superior accuracy and reliability in hotspot prediction and trend analysis.

In summary, this project proposes a novel approach to crime hotspot detection using ML, aiming to enhance law enforcement capabilities and urban safety measures. By building upon existing methodologies and integrating state-of-the-art techniques, this research seeks to make significant strides in predictive crime analytics, ultimately contributing to safer and more secure urban environments.

Literature Review

The study [1] uses historical crime data and built environment covariates like Points of Interest (POIs) and road network density to predict crime hotspots. Among the algorithms tested, the LSTM model outperformed others such as KNN, random forest, SVM, naive Bayes, and CNN, especially when incorporating built environment data.

The paper [2] addresses the challenge of optimizing police force distribution to combat crime in cities with limited resources, using the Chicago Crime Dataset as a case study. The research focuses on identifying "crime hotspots"—regions with heightened criminal activity—at specific times and dates. The problem is modeled as a multi-class classification task with class imbalance, and various supervised machine learning algorithms, including Logistic Regression, Naive Bayes, K Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, and Gradient Boosting Trees (GBT), are evaluated. Among these, GBT emerges as the most effective model, demonstrating high accuracy and superior performance in correctly identifying high-risk areas while minimizing misclassification. The study highlights the potential of data science and machine learning in enhancing public safety by facilitating the efficient allocation of

police resources, with a particular emphasis on its applicability to regions like India with low police-to-population ratios.

The paper [3] presents a comparative study of crime data from New York City's five boroughs—Brooklyn, Bronx, Manhattan, Queens, and Staten Island—focusing on the year 2019. It explores the use of machine learning techniques, including Decision Tree, Multivariate Linear Regression, and k-Nearest Neighbors (kNN), to predict crime hotspots and types of criminal offenses. The study demonstrates that Decision Tree and kNN models are highly effective in predicting the borough where a crime occurred, while Multivariate Linear Regression excels in predicting the type of crime. By automating the detection of crime patterns, the proposed models aim to enhance the allocation of police resources and improve public safety. The results underscore the potential of data-driven methods to inform crime prevention strategies, with future work suggesting the integration of additional data sources to further refine these predictions.

The paper [4] presents an enhanced Naïve Bayes model for crime prediction by integrating Recursive Feature Elimination (RFE) to improve its performance. The study utilizes crime data from the Chicago Police Department's CLEAR dataset to demonstrate the effectiveness of this approach. The authors compared the original Naïve Bayes model with the enhanced version and found that the RFE-enhanced model significantly improved accuracy from 72.55% to 96.6%. The enhanced model also performed competitively with, and in some cases outperformed, other tree-based algorithms such as Random Forests and Extremely Randomized Trees. The study concludes that feature selection methods like RFE can significantly boost the predictive power of simple probabilistic models like Naïve Bayes, making them viable alternatives to more complex algorithms in crime prediction.

The paper [5] explores the use of spatio-temporal data mining techniques, specifically Heat Maps and Hot Spot analysis, to understand crime patterns in Rawalpindi, focusing on Crimes Against Persons and Crimes Against Property. By incorporating socio-economic factors and geographical information, the study reveals that Crimes Against Persons are more prevalent in low socio-economic, residential areas, while Crimes Against Property are concentrated in high socio-economic, commercial areas. Heat Maps highlighted regions of high and low crime density, whereas Hot Spot analysis identified areas of significant crime concentration. The findings suggest that crime patterns are influenced by socio-economic status and land use, providing valuable insights for law enforcement agencies to tailor crime prevention strategies effectively.

The paper [6] explores the use of deep neural networks (DNNs) to predict criminal behavior based on historical arrest records, focusing on handling imbalanced data through data augmentation and weighted loss functions. It employs a deep fully connected neural network to classify crime types and predict future criminal activity at the individual level, using a 5-year look-ahead window for labeling. The study compares different methods, including a dynamic window approach and a fixed window model, finding that while the fixed window model achieves high accuracy in predicting crime occurrence and levels, the dynamic window approach provides better general insights. The research demonstrates the potential of neural networks in crime prediction and suggests further exploration of data augmentation and modeling techniques to enhance results.

The paper [7] explores using machine learning to predict crime hotspots in Saudi Arabia by analyzing extensive crime data with various classifiers. It finds that Factor Analysis of Mixed Data (FAMD) is more effective than Principal Component Analysis (PCA) for feature selection, achieving higher accuracy in crime prediction. The Naïve Bayes classifier outperforms others with an accuracy of 97.53% using FAMD features, while Deep Learning shows lower accuracy due to

data limitations and underfitting. The study highlights the effectiveness of advanced machine learning techniques in analyzing crime data to identify and predict crime patterns and suggests future improvements such as integrating risk terrain modeling and expanding the dataset to include additional factors like education and weather.

The paper [8] explores crime analysis in Manila City, Philippines, using data mining techniques to identify hotspots and predict future crime trends. Historical crime data from 2012-2016 was geocoded and analyzed using ArcGIS to generate a heat map highlighting high-crime areas like Tondo, Malate, and Ermita. The Apriori algorithm was employed for association rule mining to uncover patterns and relationships in crime occurrences, revealing that certain crimes are more likely in specific districts under conditions. Time series forecasting methods—Linear Regression, Gaussian Processes, Multilayer Perceptron (MLP), and Sequential Minimum Optimization Regression (SMOreg)—were used to predict future crime counts, with MLP providing the most accurate forecasts. The study concludes that data mining can significantly enhance crime prevention strategies and suggests future work to refine predictive algorithms and expand the study to other cities.

The paper [9] explores the application of big data analytics to predict and prevent crime in San Francisco using the city's publicly available crime data from 2003 to 2018. The study employs exploratory data analysis (EDA) to identify crime patterns, trends, and correlations, and then uses various machine learning models—including Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbor, and Multinomial Logistic Regression—to predict crime types in different districts. It addresses the challenge of data imbalance by refining preprocessing methods and leveraging Apache Spark for efficient data handling. The results show that Random Forest achieved the best performance with a Log Loss score of 2.276, demonstrating the efficacy of the

proposed methodologies in improving crime prediction accuracy and supporting smart city initiatives for better resource allocation and crime prevention.

The paper [10] investigates crime prediction using machine learning and deep learning techniques on datasets from Chicago, San Francisco, and Boston. Four approaches were tested: XGBoost, CatBoost, TabNet, and a hybrid model combining 1D CNN with XGBoost. Results showed that XGBoost achieved the highest accuracy for the Chicago and San Francisco datasets, while the hybrid model excelled in datasets with fewer features, and Bagging proved most effective for Boston. The study demonstrated that feature-rich datasets significantly enhance prediction accuracy, with XGBoost providing superior results overall. The findings underscore the importance of dataset quality and feature selection in improving crime prediction models.

Problem Statement:

How can machine learning (ML) be utilized to accurately predict crime hotspots and analyze spatial and temporal trends in criminal activity using the NYPD arrest dataset from NYC Open Data.

Methodology

1. Data Collection and Loading:

The primary dataset for this project is the NYPD arrest dataset from NYC Open Data, comprising 63,621 rows and 20 columns of comprehensive information about arrests made by the NYPD. This dataset includes details such as arrest date, offense descriptions, location, and demographic information about perpetrators, though some columns, like 'PD_CD', 'KY_CD', and 'LAW_CAT_CD', have null values, ensuring a robust foundation for analysis and modeling. The

dataset, with its auto-incrementing index and detailed arrest records across various boroughs, was chosen for its richness and depth in criminal activity data. Previous research utilizing this dataset involved a comparative study of New York City crime data from 2019, employing classification models such as Decision Tree and kNN to analyze relationships between crime types, boroughs, and other variables, focusing on predicting future crime trends based on historical data.*2. Inspecting Data:*

Upon loading the dataset, an initial inspection is conducted to understand its structure, size, and data types. This inspection includes checking for missing values, assessing the distribution of numerical features, and identifying potential issues such as data discrepancies or anomalies. Inspecting the data ensures its quality and suitability for further analysis and model development. It helps in identifying any preprocessing steps needed to clean and prepare the data for modeling.

3. Renaming and Deleting Unwanted Columns:

Columns in the dataset are renamed for clarity and ease of understanding. Unnecessary columns that do not contribute to the predictive task or contain redundant information are identified and deleted. Renaming columns improves readability and clarity, making it easier to work with the dataset during analysis and modeling. Removing unnecessary columns reduces noise and simplifies the dataset, focusing only on relevant features for crime hotspot prediction.

4. Creating Features:

a. Feature Extraction from Date Column

In analyzing the 'Arrest_Date' column, several temporal features can be extracted to enhance the understanding of criminal activity trends. Extracting the year of each arrest provides insight into annual trends, allowing for year-over-year comparisons and long-term patterns identification. Capturing the month name helps in discerning seasonal variations

in criminal activity, highlighting potential correlations with factors like weather or cultural events. Extracting the day of the month facilitates daily pattern analysis, uncovering any recurring trends or anomalies within shorter timeframes. Additionally, determining the day of the week (e.g., Monday, Tuesday) enables weekday-specific analysis, which can reveal variations in criminal behavior based on weekly cycles or societal routines. These extracted features collectively support comprehensive temporal analysis, enabling the identification of patterns and trends in criminal activity across various time scales.

b. Description and Class

Standardizing offense descriptions according to the official NY Penal Code documentation is crucial for ensuring consistency and clarity in understanding the nature of each offense. This approach helps in reducing ambiguity and ensuring that analysts and models alike interpret criminal activities based on standardized definitions. Additionally, categorizing offenses into felony, misdemeanor, violation, or infraction categories, as defined by the severity outlined in the NY Penal Code, facilitates effective modeling and analysis. This classification system ensures that the severity of offenses is appropriately considered, aiding in the development of more accurate predictive models and analytical insights in crime hotspot detection.

c. Target Variable "Indicator":

New column "No_days," indicating the number of days since each crime occurred relative to the most recent date in the dataset, and "Indicator," designed to identify crime hotspots based on the type of crime. For Felony crimes, if the crime occurred within the last 8 days, it is marked as RED; if between 8 days and 1 month ago, it is ORANGE; for crimes between 1 to 3 months ago, the indicator is YELLOW; and for crimes over 3 months

old, it is LIGHT BLUE, indicating serious crimes that should phase out after 6 months. Misdemeanor crimes are categorized as ORANGE if occurring within 8 days, YELLOW between 8 days and 1 month, and LIGHT BLUE between 1 to 3 months. Violation crimes receive a YELLOW indicator if within 8 days and LIGHT BLUE between 8 days and 1 month, while Infraction crimes are marked LIGHT BLUE within 8 days. The indicator helps prioritize recent and severe crimes for law enforcement agencies, guiding resource allocation and crime prevention strategies effectively.

5. Handling Duplicates and Outliers:

Ensuring there are no duplicates and that all values fall within the desired range confirms the dataset's integrity and prepares it for accurate analysis. This meticulous preprocessing step is crucial as it ensures the reliability of subsequent analyses and models, safeguarding against erroneous conclusions or biased outcomes based on flawed data inputs.

6. Exploratory Data Analysis (EDA):

In the statistical analysis phase of the project, distributions, correlations, and trends within the dataset were explored to gain insights into crime patterns. Various features of the data were examined for their distribution, and significant relationships between variables were identified. Visualization tools, including histograms and maps, were employed to illustrate spatial and temporal patterns of crime. Histograms revealed the distribution of crime types and maps depicted spatial patterns, while temporal patterns were visualized to observe trends over time, aiding in the identification of crime hotspots.

To provide a clearer understanding of the findings, the following graphs illustrate the various aspects of the data analysis.

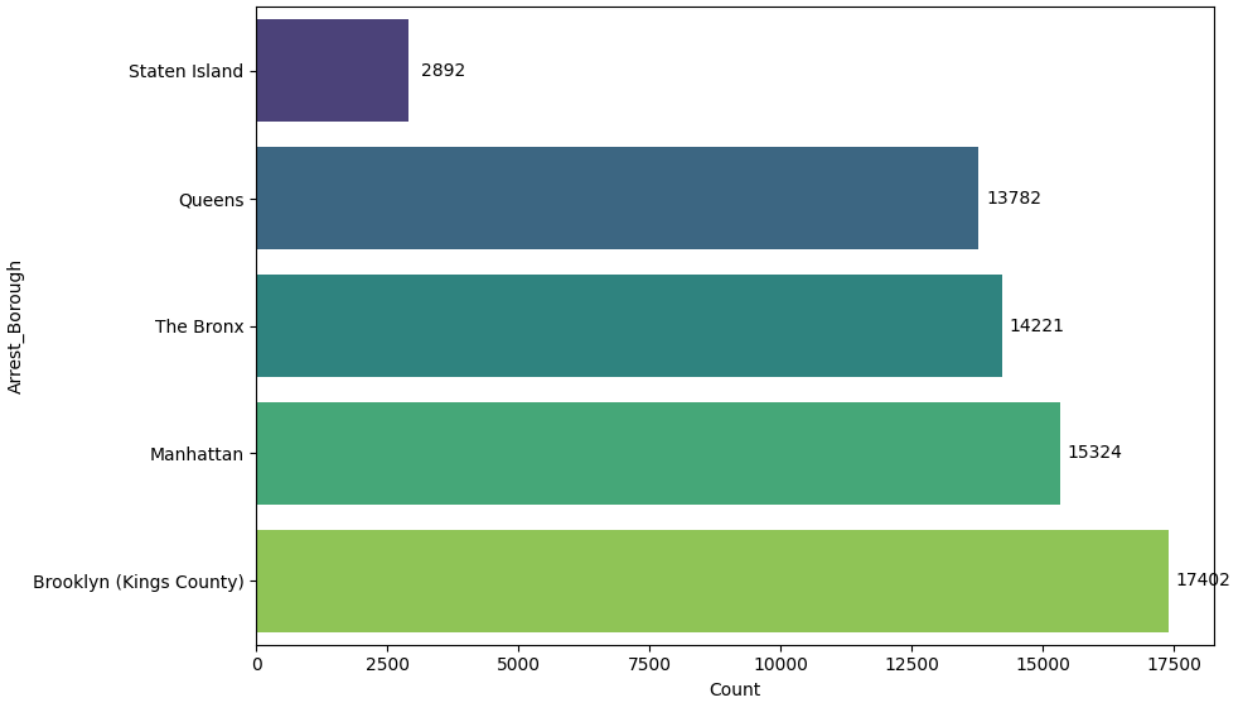

Figure 1: Distribution of Arrest Borough

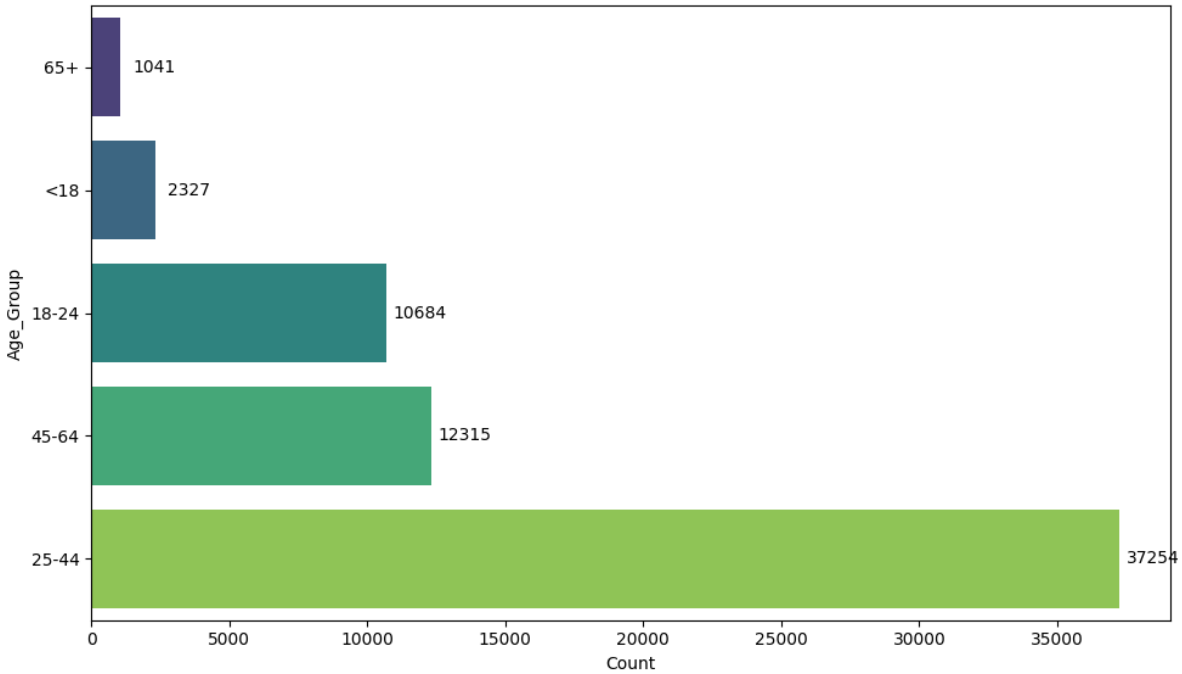

Figure 2: Distribution of Age Group

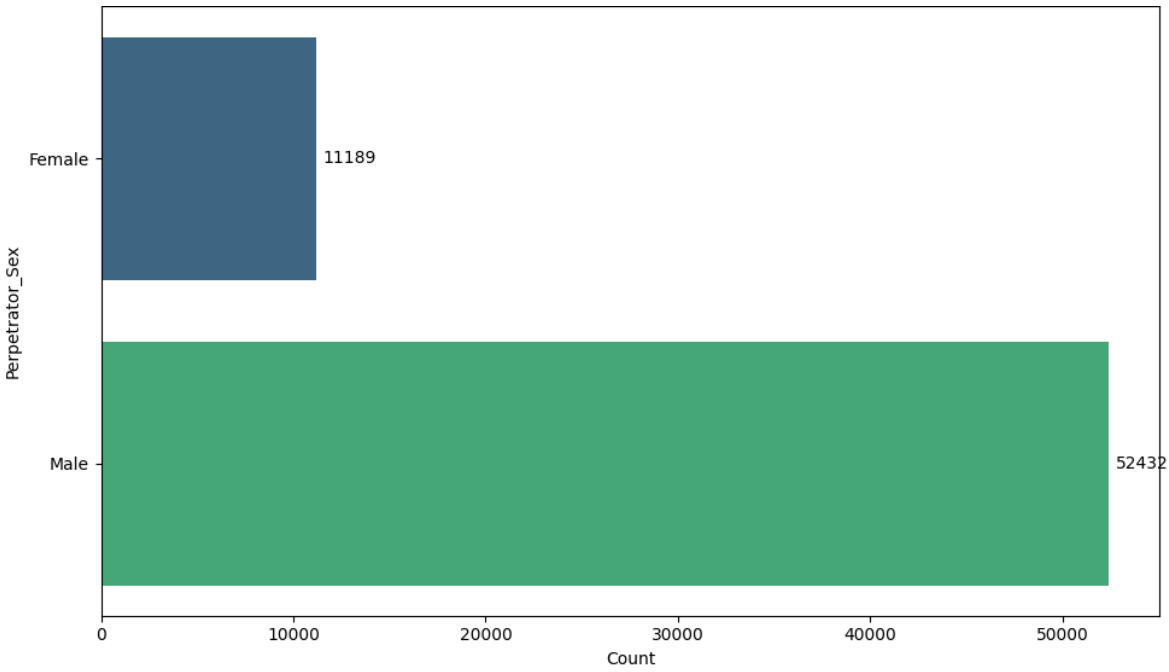

Figure 3: Distribution of Perpetrator sex

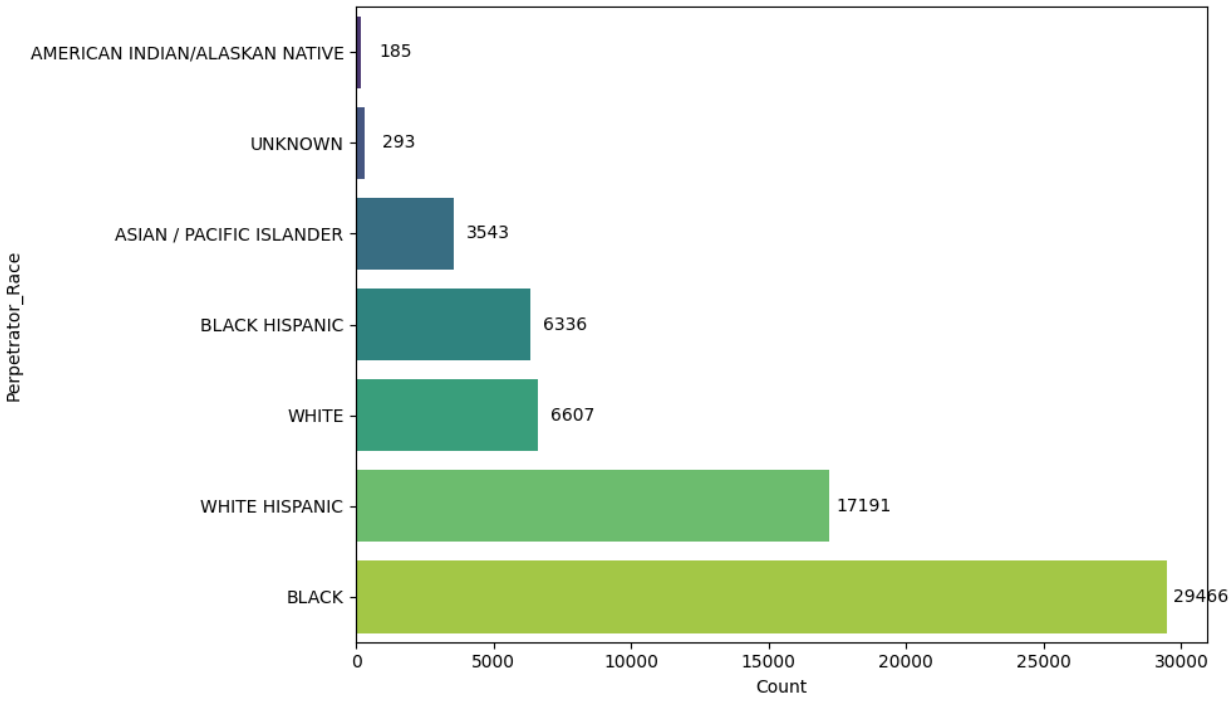

Figure 4: Distribution of Perpetrator race

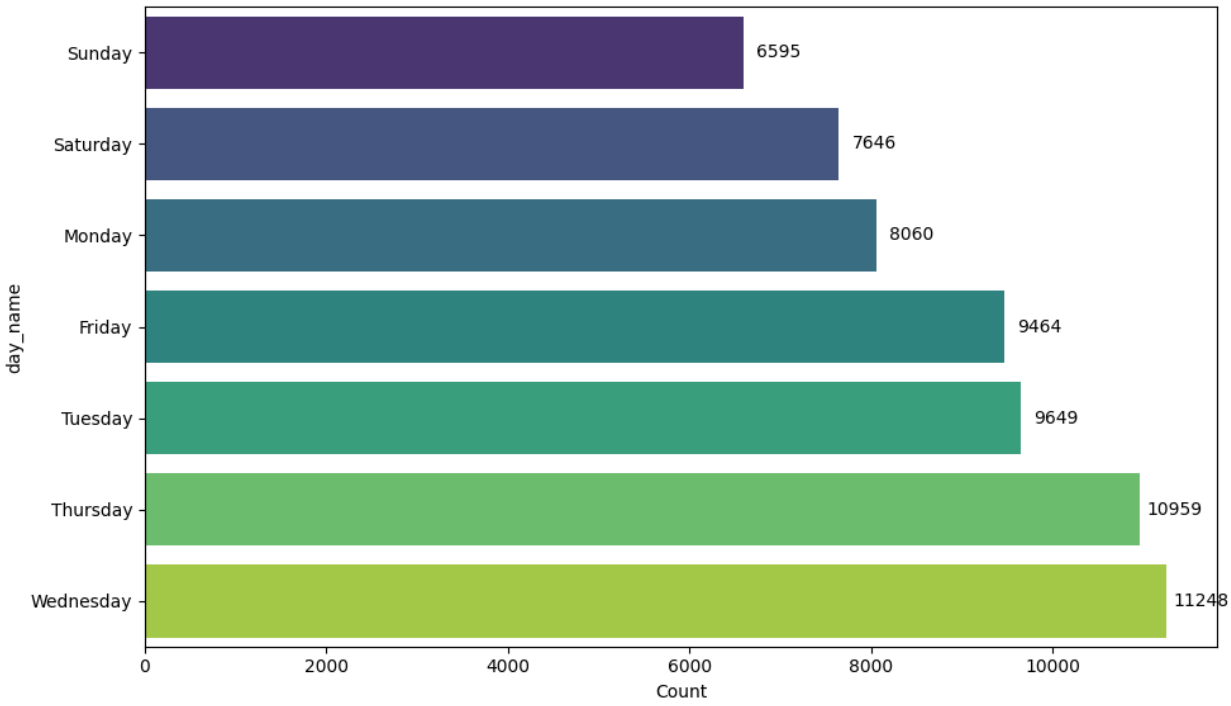

Figure 5: Distribution of day

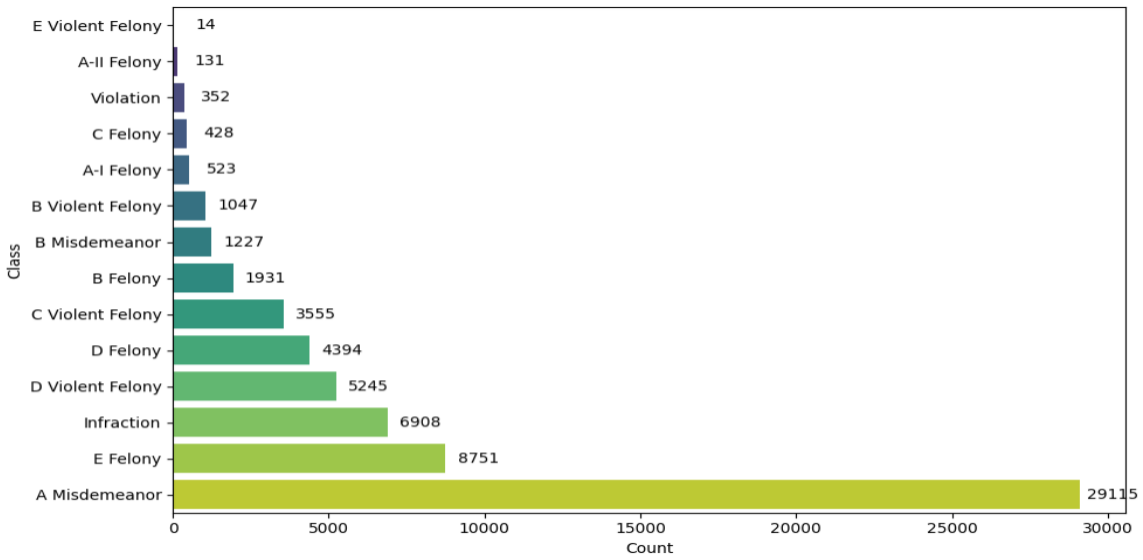

Figure 6: Distribution of Class

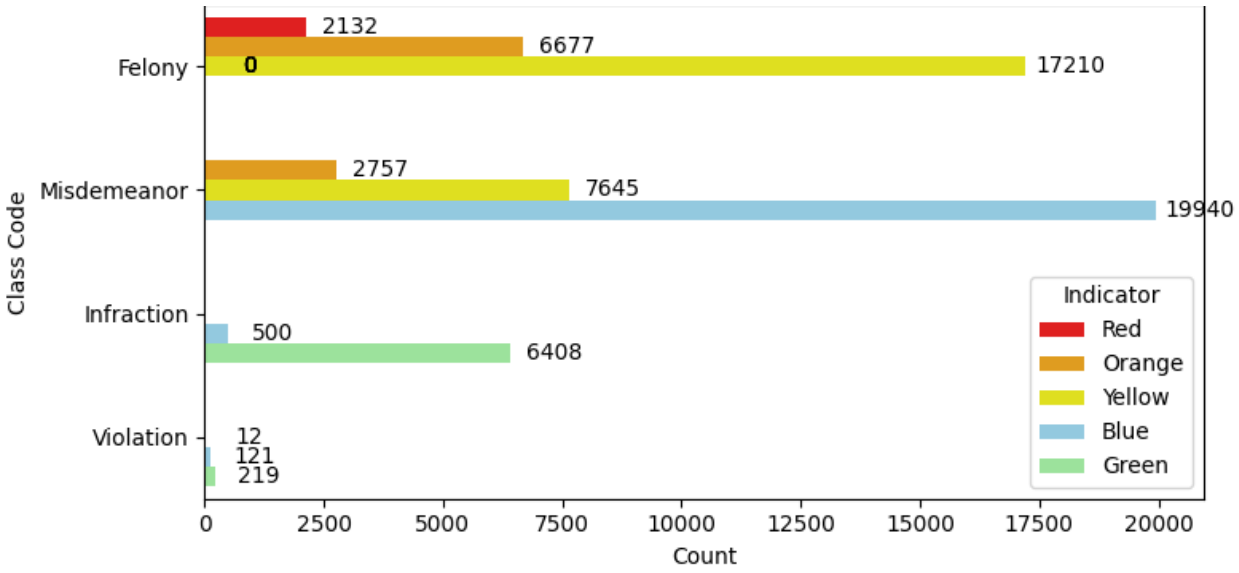

Figure 7: Crime Severity Indicators by Class Code

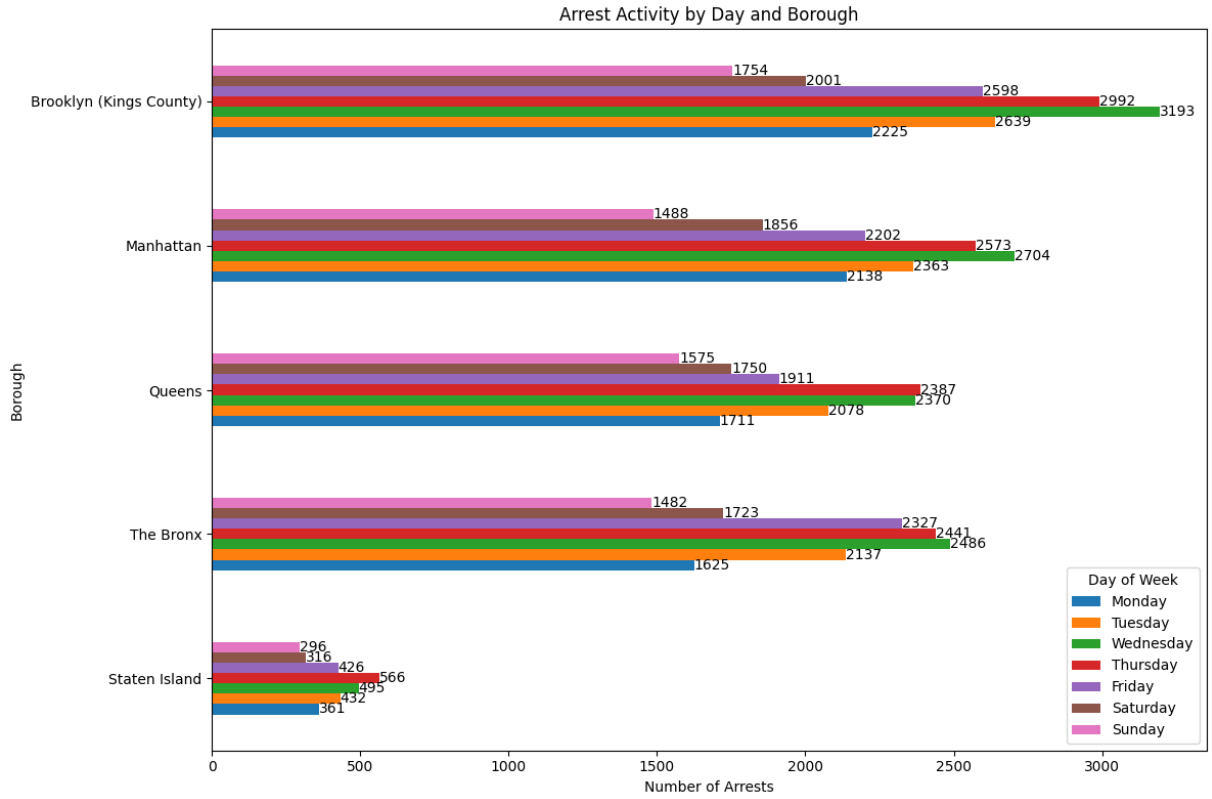

Figure 8: Arrest activity by day and borough

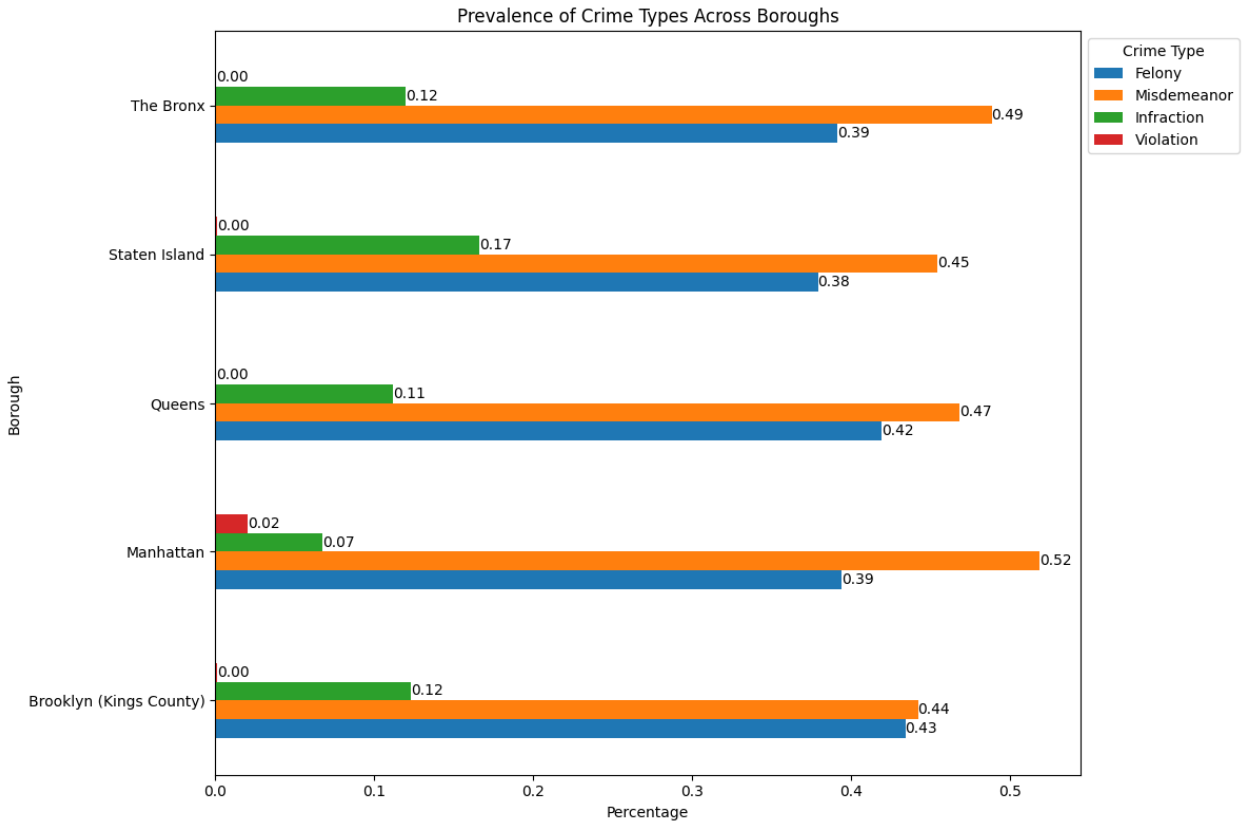

Figure 9: Prevalence of Crime types across borough

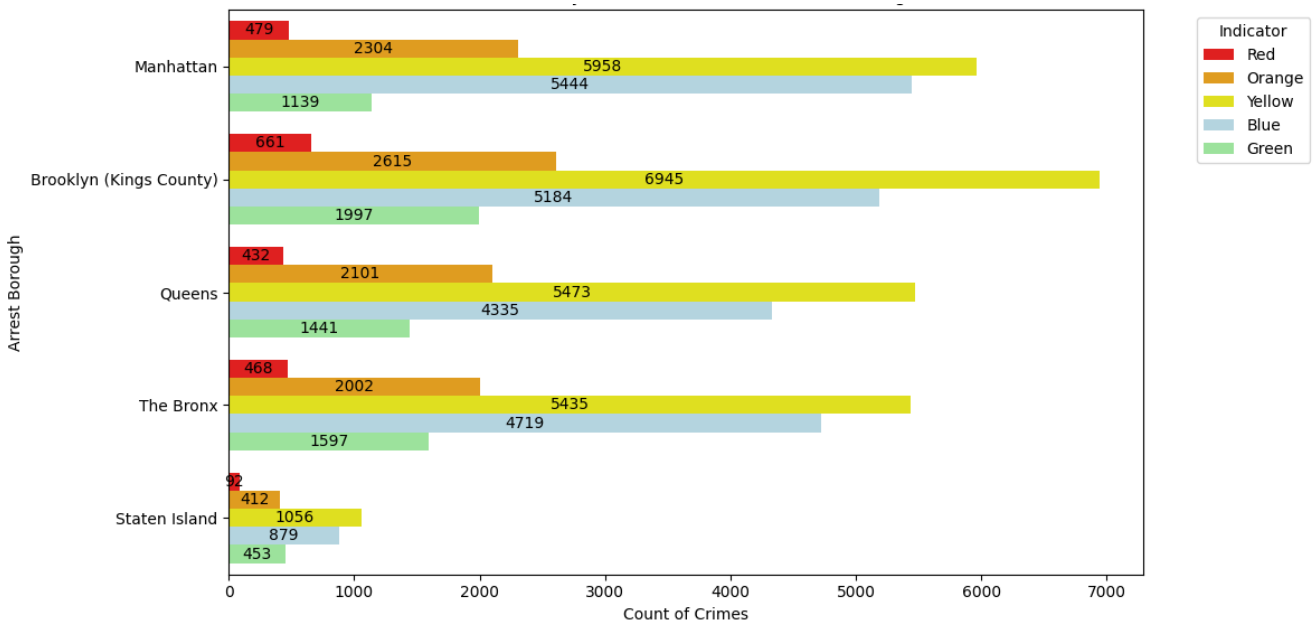

Figure 10: Crime Severity Indicators across different Boroughs

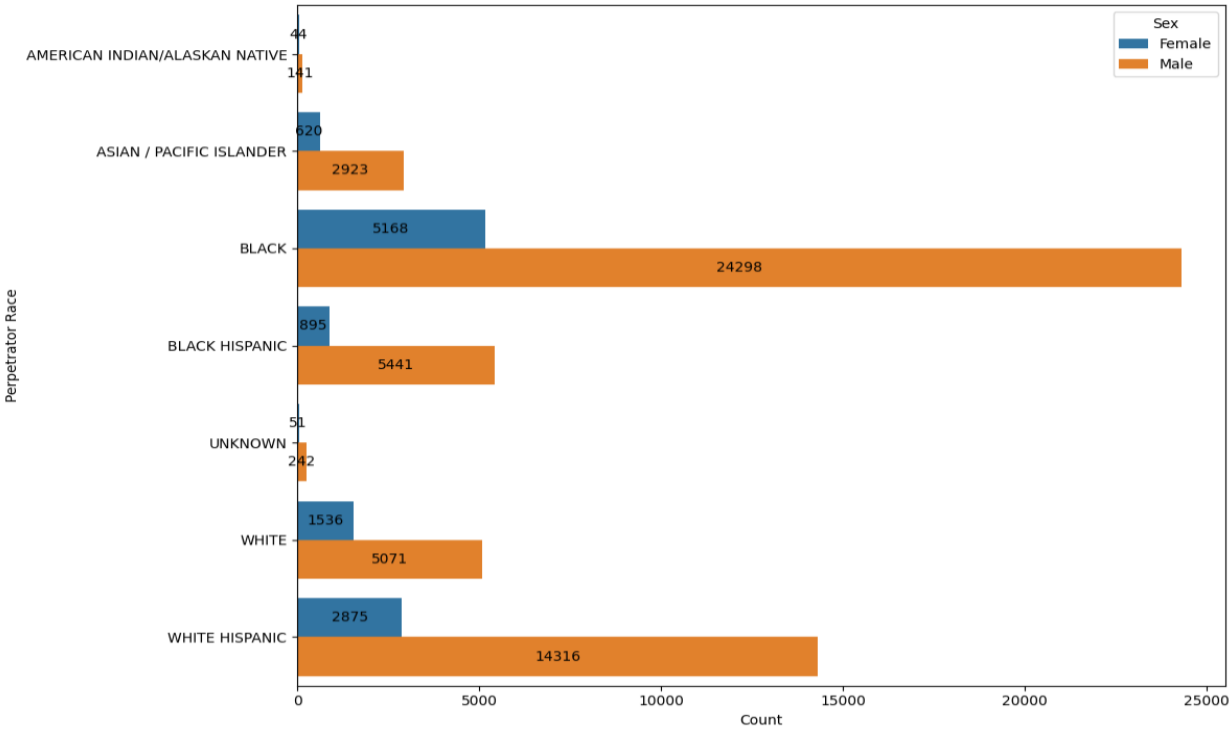

Figure 11: Distribution of Perpetrator demographics across races and sexes

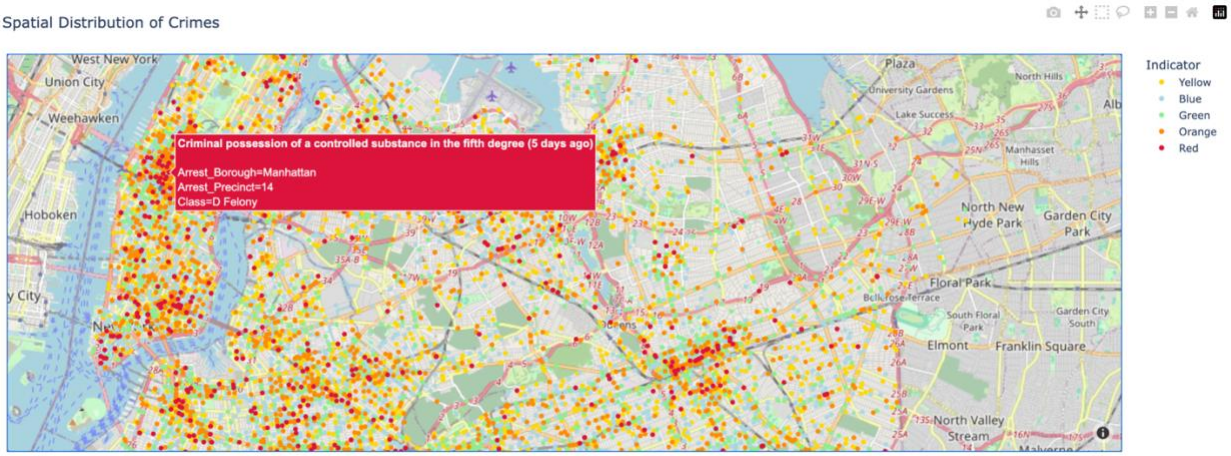

Figure 12: Spatial Distribution of Crimes

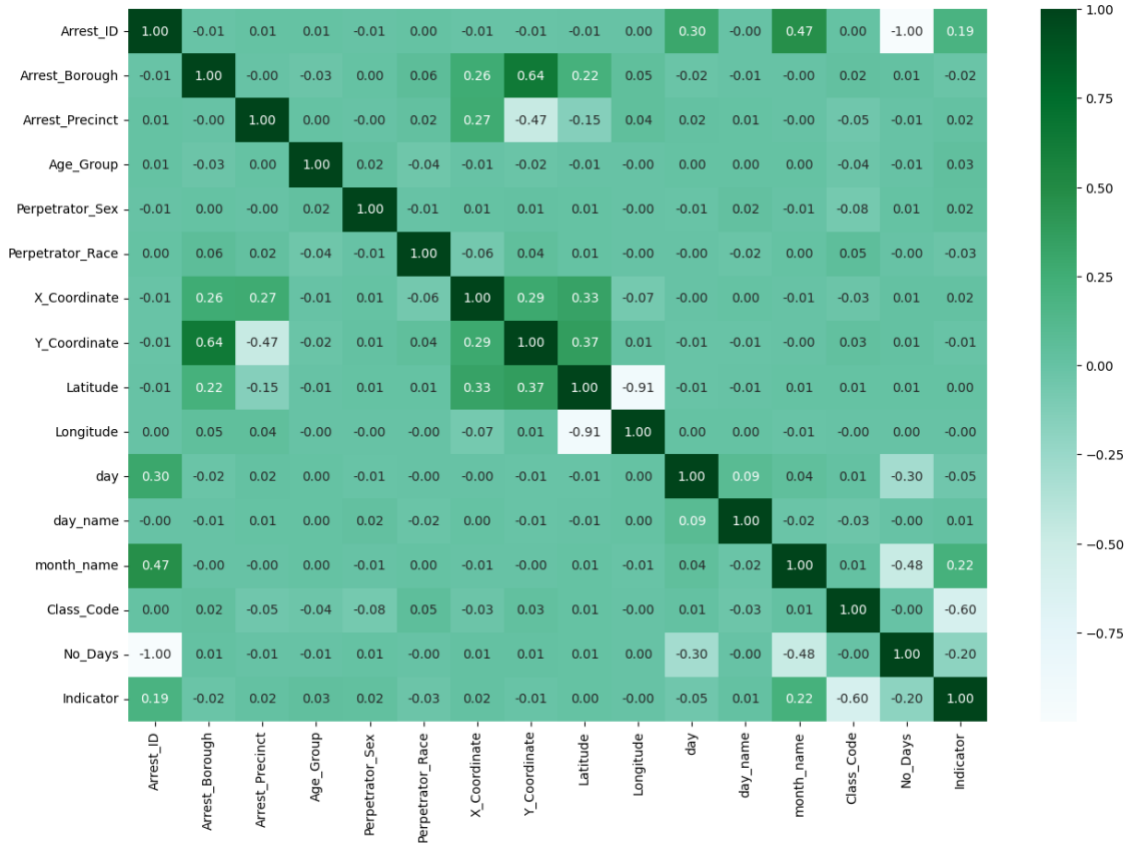

Figure 13: Co-relational matrix

7. Stratified K-Fold Cross-Validation:

In machine learning, especially with imbalanced datasets, it's crucial to ensure that our model is trained and tested on representative samples. Stratified K-Fold is an advanced cross-validation technique that addresses this by preserving the proportion of each class label across all folds. This ensures that each fold in the cross-validation process maintains the same class distribution as the original dataset, leading to more reliable and unbiased model evaluation.

8. Model Selection and Training:

Various machine learning (ML) models are selected based on their suitability for crime hotspot prediction:

- K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Gradient Boosting Classifier, Logistic Regression.

9. Model Evaluation and Optimization:

The performance of each model is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Cross-validation techniques are employed to optimize model performance and prevent overfitting. Regularization techniques are applied for Support Vector Machines (SVM) and Gradient Boosting Classifiers to further address overfitting.

10. Model Comparison and Selection:

Models are compared based on their evaluation metrics to identify the most effective approach for crime hotspot prediction and trend analysis. The strengths and weaknesses of each model are analyzed to inform the selection of the final model that best meets the project objectives.

This methodology ensures a systematic approach to leveraging data-driven insights for crime prevention and resource allocation in urban environments.

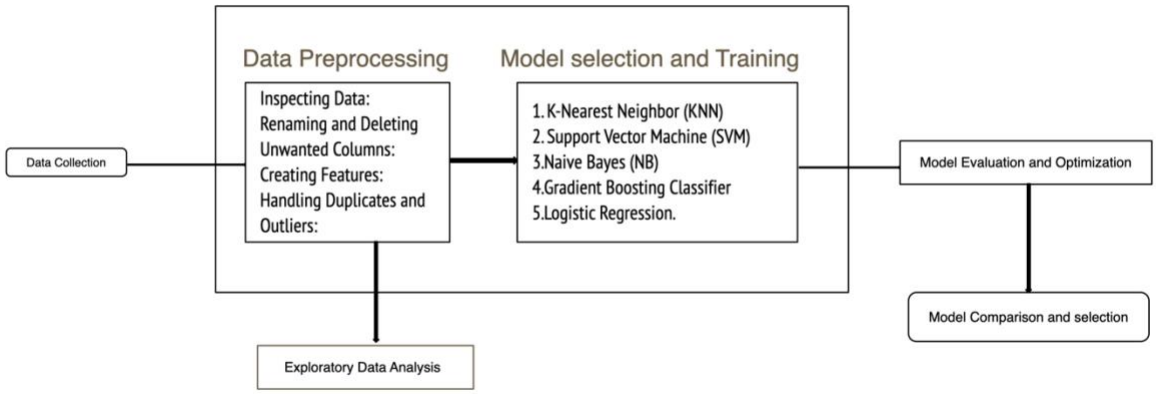

Figure 14: Methodology

RESULT

1. *Model Performance:*

To ensure the robustness and reliability of our model evaluations, we employed Stratified K-Fold Cross-Validation. This technique was crucial for maintaining the balance of class distributions across all folds, allowing us to fairly assess each model's performance on the diverse classes in our dataset. The following summarizes the key results for each model based on the average validation accuracy across all folds:

● Gaussian Naive Bayes: Achieved an average validation accuracy of 0.77.

● Support Vector Machine (SVM): After applying regularization techniques, the Support Vector Machine (SVM) achieved an average validation accuracy of 0.91.

● K-Nearest Neighbors (KNN): Achieved an average validation accuracy of 0.97.

● Gradient Boosting Classifier: Achieved a perfect average validation accuracy of 1.00. The model's performance was flawless, correctly classifying all instances. While this result is impressive, it also indicates that the model may be overfitting to the training data, leading to potentially poor generalization on unseen data.

● Logistic Regression: Achieved an average validation accuracy of 0.97.

Model	Accuracy
1. Gaussian Naive Bayes	77%
2. SVM	91%
3. KNN	97%
4. Gradient Boosting Classifier	100%
5. Logistic Regression	97%

Table 1: model names and their accuracies.

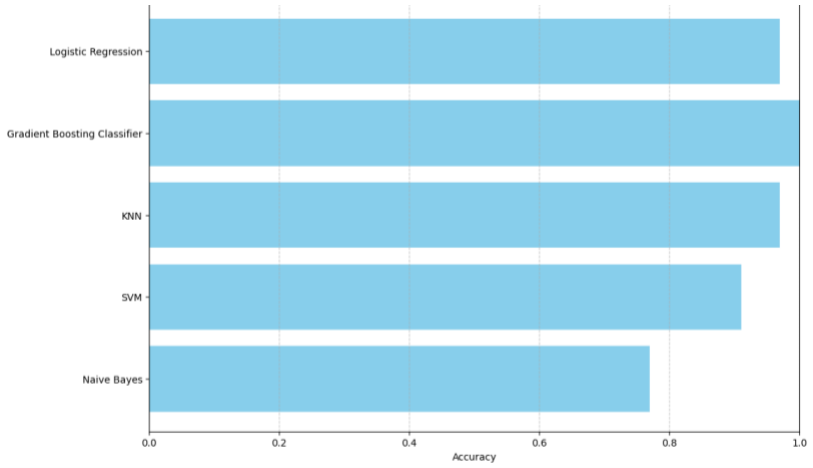

Figure 15: Model Accuracies Comparison

2. *classification reports Analysis:*

Here's a summary of the classification reports for various models:

Naive Bayes:

Table 2: Classification Report for Naive Bayes

	precision	recall	f1-score	support
Blue	0.66	0.97	0.79	20561
Green	0.93	0.97	0.95	6627
Orange	0.93	0.33	0.49	9434
Red	0.53	1.00	0.69	2132
Yellow	0.89	0.69	0.78	24867
accuracy			0.77	63621
macro avg	0.79	0.79	0.74	63621
weighted avg	0.82	0.77	0.75	63621

- The model performs very well on the green and yellow classes.
- Orange has high precision but very low recall, meaning the model is good at predicting Orange when it does, but it misses many true orange instances.

- Red has perfect recall but lower precision, indicating that while it identifies all red instances, it also misclassifies other classes as Red.
- Blue has good recall but lower precision, meaning it's often correctly identified but with more false positives.

Support Vector Machine (SVM):

- The model performs exceptionally well on the blue, green, and yellow classes.
- Orange shows moderate performance, with a decent balance of precision and recall.
- The model struggles significantly with the red class, failing to predict any instances correctly.
- The high overall accuracy is mainly due to the strong performance on the larger classes (Blue, Green, and Yellow), but the red class requires attention and improvement.

Table 3: Classification Report for Support Vector Machine (SVM)

	precision	recall	f1-score	support
Blue	1.00	0.96	0.98	20561
Green	0.92	0.99	0.95	6627
Orange	0.74	0.69	0.71	9434
Red	0.00	0.00	0.00	2132
Yellow	0.89	0.99	0.94	24867
accuracy			0.91	63621
macro avg	0.71	0.73	0.72	63621
weighted avg	0.88	0.91	0.89	63621

K-Nearest Neighbors (KNN):

Table 4: Classification Report for K-Nearest Neighbors (KNN)

	precision	recall	f1-score	support
Blue	0.99	0.98	0.99	20561
Green	0.99	0.99	0.99	6627
Orange	0.92	0.93	0.93	9434
Red	0.92	0.85	0.88	2132
Yellow	0.97	0.97	0.97	24867
accuracy			0.97	63621
macro avg	0.96	0.95	0.95	63621
weighted avg	0.97	0.97	0.97	63621

- The model demonstrates exceptional performance across all classes, particularly Blue, Green, and Yellow, with very high precision, recall, and F1-scores.
- Orange and Red also perform well, though Red has slightly lower recall, indicating a few missed instances.
- The overall metrics show a well-balanced model with high precision and recall across the board, leading to an excellent accuracy of 97%.

Gradient Boosting:

- The model achieves a perfect accuracy of 100% across all classes, with flawless precision, recall, and F1-scores, indicating it has perfectly learned the patterns in the dataset.
- Despite applying regularization, the perfect scores suggest potential overfitting, meaning the model might not generalize well to new, unseen data, highlighting the need for further validation on independent datasets.

Table 5: Classification Report for Gradient Boosting

	precision	recall	f1-score	support
Blue	1.00	1.00	1.00	20561
Green	1.00	1.00	1.00	6627
Orange	1.00	1.00	1.00	9434
Red	1.00	1.00	1.00	2132
Yellow	1.00	1.00	1.00	24867
accuracy			1.00	63621
macro avg	1.00	1.00	1.00	63621
weighted avg	1.00	1.00	1.00	63621

Logistic Regression:

Table 6: Classification Report for Logistic Regression

	precision	recall	f1-score	support
Blue	0.98	0.96	0.97	20561
Green	0.96	1.00	0.98	6627
Orange	0.94	0.94	0.94	9434
Red	1.00	0.98	0.99	2132
Yellow	0.97	0.97	0.97	24867
accuracy			0.97	63621
macro avg	0.97	0.97	0.97	63621
weighted avg	0.97	0.97	0.97	63621

- The model shows outstanding performance across all classes, particularly in the green, red, and yellow classes, with near-perfect precision, recall, and F1-scores.
- Blue and Orange also perform very well, though Blue has slightly lower recall compared to the other classes.
- The high precision and recall across the board indicate that the model is highly reliable and accurate in predicting all classes.

3. Confusion Matrix Analysis:

The confusion matrices provided further insights into the model performance:

- Gaussian Naive Bayes:
    - Blue: The model correctly classified 19,939 instances, with 500 misclassified as other classes, and 120 misclassified as Orange.
    - Green: All instances were correctly classified as Green, with 6,405 instances correctly identified and 222 misclassified as other classes.
    - Orange: The model correctly identified 3,099 instances as Orange but misclassified many others across different classes.
    - Red: All instances were correctly classified, with 2,132 instances.
    - Yellow: Similarly, 17210 instances were correctly identified, with the rest misclassified.

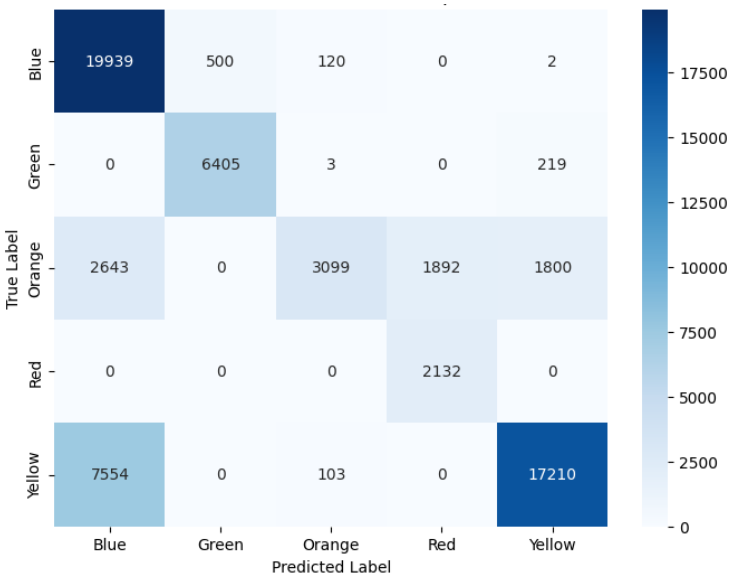

Figure 16: Confusion Matrix for Gaussian Naive Bayes

- SVM:
    - The model performs exceptionally well on the blue, green, and yellow classes.
    - Orange shows moderate performance, with a decent balance of precision and recall.
    - The model struggles significantly with the red class, failing to predict any instances correctly.
    - The high overall accuracy is mainly due to the strong performance on the larger classes (Blue, Green, and Yellow), but the red class requires attention and improvement.

Figure 17: Confusion Matrix for SVM

- KNN:

  - Blue: The model correctly classified 202,337 instances, with 324 misclassified as other classes, and 228 misclassified as yellow.
  - Green: The model correctly identified 6,583 instances, with a small number of misclassifications.
  - Orange: The model correctly classified 8,794 instances as Orange, with minor misclassifications, including 160 instances as Red and 480 as Yellow.
  - Red: The model correctly identified 1,808 instances but misclassified 324 as Orange.
  - Yellow: The model successfully classified 24,215 instances as Yellow, with only a small number of misclassifications.

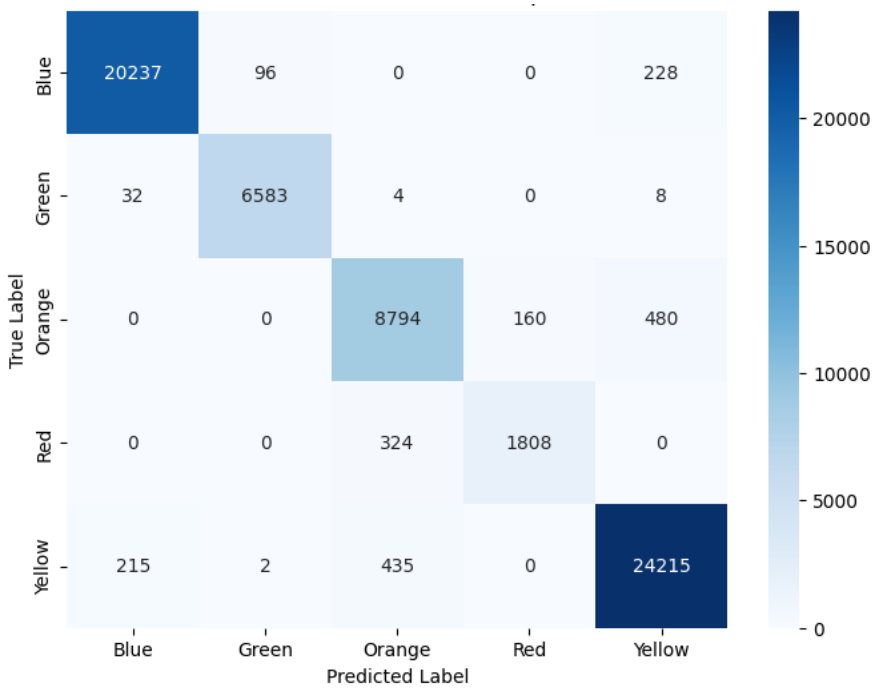

Figure 18: Confusion Matrix for KNN

- Gradient Boosting Classifier:

    - Blue: The model perfectly classified all 20,561 instances as Blue, with no misclassifications.
    - Green: All 6,627 instances were correctly classified as Green, with no errors.
    - Orange: The model correctly classified all 9,434 instances as Orange, with no misclassifications.
    - Red: All 2,132 instances were correctly identified as Red, with no misclassifications.
    - Yellow: The model perfectly classified all 24,867 instances as Yellow, with no errors.

Given the perfect average validation accuracy of 1.00 and final fold accuracy of 1.00, coupled with a confusion matrix showing no misclassifications, it strongly suggests that the model may be overfitting. Overfitting occurs when a model performs exceptionally well on the training and validation data but may not generalize effectively to new, unseen data. The model's flawless performance on both validation and final fold data implies it might have learned noise or specific patterns from the training data rather than generalizable features.

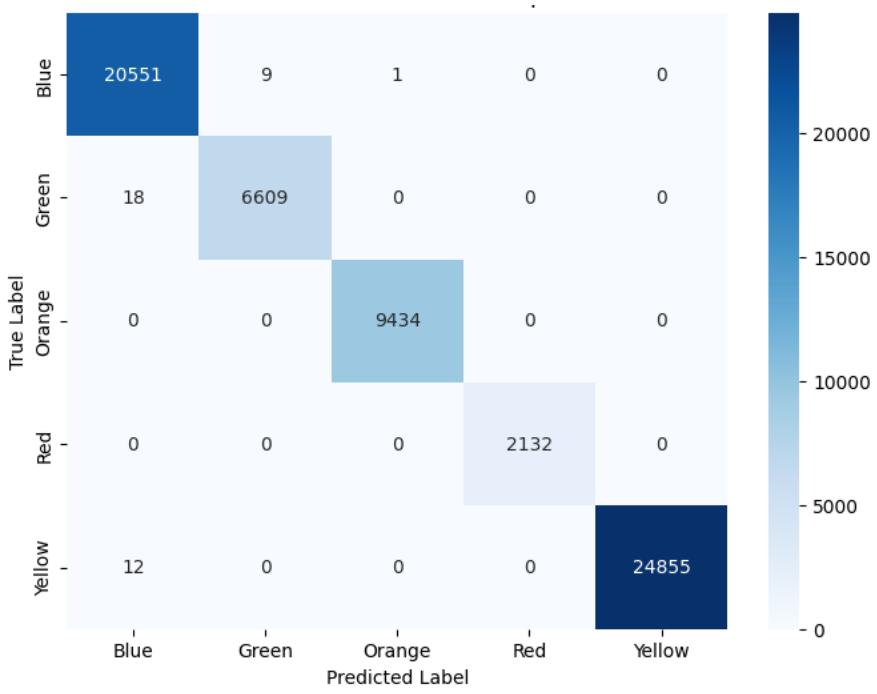

Figure 19: Confusion Matrix for Gradient Boosting Classifier

● Logistic Regression:

- Blue: The model correctly classified 19,729 instances as Blue, with some misclassified as Green (241), Orange (261), and Yellow (330).
- Green: The model perfectly identified 6,626 instances as Green, with no misclassifications into other classes.
- Orange: The model correctly classified 8,890 instances as Orange, with minimal misclassification to Blue (94) and Yellow (450).
- Red: The model identified 2,081 instances as Red, with only 51 instances misclassified as Orange.

- Yellow: The model accurately classified 24,232 instances as Yellow, with misclassifications to Blue (353), Green (3), and Orange (279).

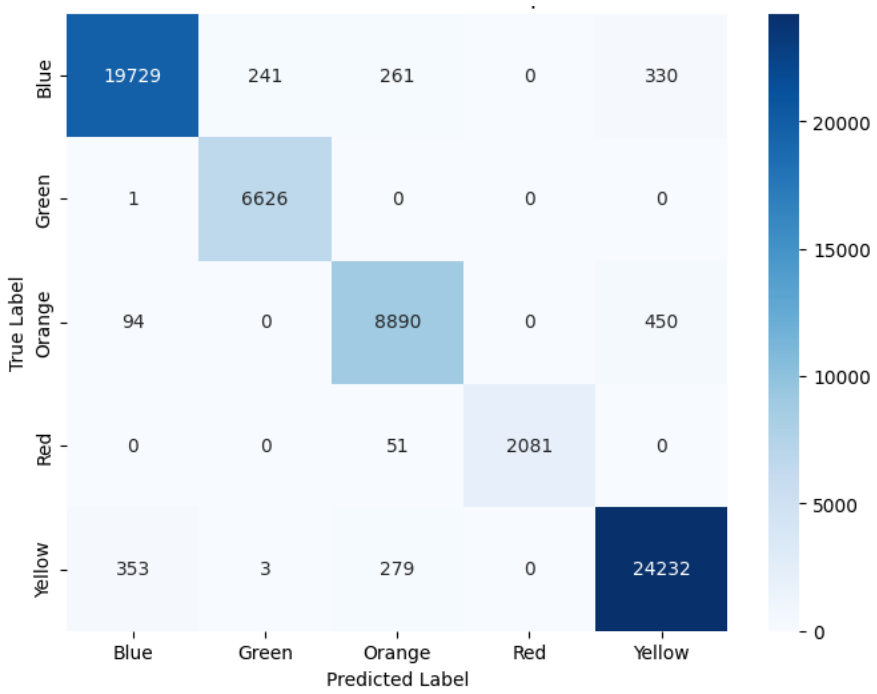

Figure 20: Confusion Matrix for Logistic Regression

Limitation:

a) Potential Overfitting of Models: The Gradient Boosting Classifier exhibited near-perfect accuracy and showed signs of overfitting even after applying regularization. This could lead to poor generalization when applied to new data, reducing its effectiveness in real-world scenarios.

b) Dependence on Historical Data: The reliance on historical arrest data may limit the models' ability to capture emerging crime patterns or shifts in criminal behavior, potentially hindering accurate future hotspot predictions.

Conclusion:

The analysis of crime hotspot detection using machine learning models revealed that certain models, particularly KNN and Logistic Regression, performed well, achieving high accuracy and minimal misclassifications. However, model like Gradient Boosting Classifier, while demonstrating near-perfect accuracy, showed signs of overfitting even after applying regularization. This overfitting could limit their ability to generalize to new data, making KNN and Logistic Regression more reliable choices for practical applications in crime hotspot detection.

Future work could focus on further refining these models to balance accuracy with generalization, perhaps by exploring additional features that could enhance model performance. Moreover, the inclusion of more diverse data sources and real-time data could improve the robustness of the hotspot detection system, making it more applicable in dynamic and evolving urban environments like New York City.

Overall, this study demonstrates the potential of machine learning in enhancing public safety through effective crime hotspot detection, providing valuable insights for law enforcement and public safety agencies.

Works Cited

1. XU ZHANG, LIN LIU, LUZI XIA, AND JIAKAI JI. "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots". October 2, 2020.

2. Yadhunath Ramshankar, Srikanth Srivenkata, Sudheer Arvind and Palaniswamy Suja. "Identification of Criminal Activity Hotspots using Machine Learning to Aid in Effective Utilization of Police Patrolling in Cities with High Crime Rates". IEEE, 12 March 2020.

3. Marcus Pinto, Hsinrong Wei, Kiyatou Konate, and Ida Touray. "Delving into Factors Influencing New York Crime Data with the Tools of Machine Learning".ACM, NY, USA. 01 October 2020

4. Sphamandla May, Omowunmi Isafiade, and Olasupo Ajayi. "An Enhanced Naïve Bayes Model for Crime Prediction using Recursive Feature Elimination". Association for Computing Machinery New York, NY, United States. 25 February 2022

5. Malik Sunia, Afzal Hammad, Siddiqi Imran, and Majeed Awais. "Analyzing Socio-economic and Geographical factors for Crime Incidents using Heat maps and Hotspots". 22 November 2016

6. Soon Ae Chun, Venkata Avinash Paturu, Shengcheng Yuan, Rohit Pathak, Vijay Atluri and Nabil R. Adam. "Crime Prediction Model using Deep Neural Networks". ACM, NY, USA. June 2019

7. Alsaqabi Anadil, Aldhubayi Fatimah, and Albahli Saleh. "Using Machine Learning for Prediction of Factors Affecting Crimes in Saudi Arabia". ACM, NY, USA. 11 June 2019

8. Charlie S. Marzan, Maria Jeseca C. Baculo, Remedios de Dios Bulos, and Conrado Ruiz, Jr. "Time Series Analysis and Crime Pattern Forecasting of City Crime Data". ACM, NY, USA. 10 August 2017.