Visualization and Factor Analysis For Crime Trends

Abinaya Vina
Dept. of Networking and Communication
SRM University
Kattankulathur, India
av2473@srmist.edu.in

Sanjai
Dept. of Networking and Communication
SRM University
Kattankulathur, India
@srmist.edu.in

Sharmila
Dept. of Networking and Communication
SRM University
Kattankulathur, India
@srmist.edu.in

Farthika
Dept. of Networking and Communication
SRM University
Kattankulathur, India
@srmist.edu.in

Dr. D Saveetha
Dept. of Networking and Communication
SRM University
Kattankulathur, India
saveethd@srmist.edu.in

Abstract — Urban crime is a major concern for residents and law enforcement agencies. Traditional crime prediction methods often rely on past data and basic statistical models, but these approaches struggle to accurately predict future crimes because they can't handle large amounts of data. This creates problems while identifying and responding to future crime patterns. Therefore, our focus is on applying advanced techniques to analyze crime patterns in metropolitan areas, to improve the accuracy and reliability of crime predictions.

Keywords — Urban Crime, Crime Prediction, Crime Patterns, Metropolitan Areas

1. INTRODUCTION

As mentioned above, effective prediction and prevention methods are becoming increasingly important for maintaining public safety. Traditional approaches often fail to handle large datasets, leading to less accurate predictions. To address these challenges, our project focuses on utilizing big data and advanced machine-learning techniques to analyze crime patterns in metropolitan areas. By considering diverse data sources, such as crime reports, demographic information, and environmental factors, we aim to uncover deeper insights into the factors that cause criminal activity.

The significance of this topic extends beyond law enforcement, as accurate crime prediction can enhance public safety, optimize resource allocation, and support informed decision-making in urban planning. Furthermore, our project aligns with the United Nations Sustainable Development Goal 16, which promotes peace, justice, and strong institutions. By analyzing the crime in metropolitan areas, we are contributing to building a safer community.

We address several key aspects, including data integration and management, spatial-temporal analysis, real-time data processing, and the ethical considerations of using big data for crime prediction. These elements are critical for developing a scalable system. By overcoming the limitations of traditional methods, our project has the potential to significantly impact urban safety and security.

2. RELATED WORK

In 2024 research was conducted for crime prediction using HDLCP (Hybrid Deep Learning Methodology for Crime Prediction). The study aimed to overthrow traditional methods of data mining and use a more modern and efficient approach, the HDLCP model. The HDLCP model mainly uses the rubrics of the Decision Tree approach. The model guarantees to improve accuracy for crime prediction. The rubrics

for the model's performance include factors such as accuracy, sensitivity, specificity, and loss ratio. Additionally, the HDLCP model assists with data cleaning and visualization by using an integrated data-cleaning algorithm. The only challenge faced in the study was the various scenarios that could be processed through the model. [1]

To look into older related works, Yadav, R. and Kumari Sheoran, S used the Auto Regression Technique in 2018 to implement crime prediction. The study focuses on using ARIMA (Auto Regressive Integrated Moving Average). ARIMA is used to forecast the time series data that is given in the dataset. This paper combines both autoregression and moving average processes. The study mainly focuses on using the mentioned statistical model to predict crime trends. Furthermore, the study shows the use of the "R" tool. In summary, the R tool is used to aid in the following tasks; data management, modeling, error reduction, and visualization. [2]

The Hubei University of Technology uses an algorithm called Random Forest to predict the areas where criminal activity is happening in a large volume. The study focuses on crime spots in San Francisco, and it is predicted based on spatial factors of the random forest algorithm. Although the study focuses on San Francisco, the application is versatile, meaning the application would be able to adapt to other regions or cities, making it a scalable solution for crime prediction. A challenge faced is the effectiveness of the model depends on the availability and quality of detailed spatial and demographic data. [3]

The paper, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," has used 150 articles to analyze machine learning algorithms and deep learning algorithms that have been used for crime prediction. The paper explores various methodologies, datasets, and trends or behaviors for crime predictions. The study also aims to explore future developments and potential gaps. The study acts as a valued reference for researchers by exploring and creating a reference for all the possible methodologies and machine learning algorithms, that exist. [4]

B. Sivanagaleela and S. Rajesh's study focuses on using the Fuzzy C-means clustering algorithm. The study identifies crime-prone areas using the above algorithm to successfully analyze and predict future crime occurrences based on historical data, aiding law enforcement in reducing crime rates. The advantages pinpointed in the study are; time efficiency and data-driven decision-making. The fuzzy clustering technique processes data quickly, providing timely insights into crime trends. Followed by which the analysis assists law enforcement agencies in planning preventive measures based on accurate data. [5]

A recent study that was done in 2024 focuses on the criminal activity that has happened in Maryland in the USA. The major methodologies used are crime mapping and predictive analysis. The study focuses on a timeline of 2016-2020. The machine learning algorithms used are the K Neighbors Classifier, Random Forest Classifier, and Logistic Regression. The Random Forest Classifier, Logistic Regression, and K Neighbors Classifier displayed an accuracy of 52.55%, 43.86% and 38.39%. The study aims to prevent crimes before they happen. [6]

In 2018, a study introduced an N-ensemble learning technique to improve crime prediction accuracy. The experiment was divided into three categories: base classification models (Naive Bayes, J48, Random Tree), ensemble learning models (1-ensemble and 3-ensemble), and statistical data visualization. The Random Tree model achieved the highest accuracy of 82.02%. The 1-ensemble model outperformed the 3-ensemble, scoring 81.61%. Crime patterns were analyzed based on time, month, and season, showing peak crime occurrences during summer afternoons (3:00 PM to 6:59 PM) and a decrease during winter mornings. [7]

Once more, in 2024, a study focused on addressing crimes against women through a data-driven approach. It aimed to analyze crime rates, predict trends, and implement an emergency alert system for real-time assistance. The analysis used the K-Means clustering algorithm to identify crime-prone areas, while DBSCAN, Agglomerative Clustering, and Mean Shift predicted emerging crime trends. The

system's efficiency was evaluated using the Silhouette Score. The study demonstrated the effectiveness of these models in providing early alerts and fostering safety for women in high-risk areas. [8]

In 2024, a study aimed to predict crime occurrences using machine learning algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). The research focused on crime analysis using data from Kaggle to predict the types, timings, and locations of crimes. By employing KNN and SVM, the study achieved improved accuracy compared to previous models. The project's goal was to provide better insights into crime patterns, enabling authorities to respond more effectively. This approach enhances understanding and anticipation of crime trends. [9]

Lastly, a study implemented the K-Nearest Neighbor (KNN) algorithm to predict crime hotspots by analyzing historical crime data and incorporating social and environmental factors. The KNN model effectively identifies potential crime-prone areas, aiding law enforcement and urban planners in creating targeted crime prevention strategies. The study highlighted KNN's ability to reveal underlying crime patterns, although its accuracy depends on data quality and relevant pattern detection. Further research is suggested to expand the approach's scope in crime prediction. [10]

3. PROPOSED MODEL

1. Data Collection and Preprocessing

The first step in implementing "Visualization and Factor Analysis For Crime Trends" is to gather the required datasets. The chosen dataset is taken from a public safety portal that provides available open data by the Toronto Police Service [11]. Specifically, this study will focus on analyzing the dataset under 'Major Crime Indicators Open Data' between 2014 and 2017. The datasets are stored in CSV in a structured format.

Once the dataset is imported and loaded from the Toronto dataset, the data must cleaned. A series of protocols are followed; firstly, the null values are removed, and the row with missing values is also removed. Secondly, statistics models like Z-scores and IQR are used to handle the outliers. Lastly, the data must be standardized, which means the date and time formats must be regulated.

The final step for preprocessing the data is implementing feature selection. The study requires choosing relevant columns, such as year, month, day, crime type (MCI_CATEGORY), division, and location. The selected columns or fields are as follows; 'PREMISES_TYPE,' 'OCC_YEAR,' 'OCC_MONTH,' 'OCC_DAY,' 'OCC_DOY,' 'OCC_DOW,' 'OCC_HOUR,' 'DIVISION,' 'HOOD_158'. The last step in feature selection is to use the technique of one-hot encoding. This ensures the categorical variables are encoded.

2. Exploratory Data Analysis

Exploratory data analysis, or EDA, consists of two major steps; descriptive statistics and visualization. Descriptive statistics are used to calculate the means, medians, and other summary statistics to understand criminal distribution. Furthermore, the study used statistical tests like the Chi-square test to understand the relation between the crime type and location. Visualization is a process used to plot criminal activity based on the selected columns in the previous step. The crime trends from 2014-2017 are shown using a bar graph but a line graph can also be used.

3. Predictive Modelling

Although this step is optional, it helps elevate the accuracy and prediction. Step one is data splitting; to divide the data into training and testing sets to create a 75/25 split. Later the model must be chosen, any appropriate model can be used, for example - Random Forest, Gradient Boosting, or Logistic Regression. This paper uses the Random Forest algorithm to predict crime occurrence based on the input variables. Next, ensure the chosen models are trained with the dataset. Once the model is trained, it must be evaluated. Evaluate the model performance using accuracy, precision, recall, and F1 score. Fine-tune hyperparameters to optimize model performance.

4. Visualization of Results

This study has used a confusion matrix and classification reports to visualize the results. This is done to visualize predicted vs. actual crime occurrences. Later, the crime rate changes per year using a bar chart with specific ranges on the y-axis (e.g., 30,000–35,000) are shown. The bar graph and the confusion matrix highlight significant crime categories like assault, robbery, auto theft, etc. Finally, analyze and interpret the conclusions.

4. RESULTS

The following plot, "Number of Crimes per Year (2014-2017)," represents a visual fluctuation in the crime rates over the four years in Toronto. The below bar graph shows the yearly increase in crime numbers. After analyzing the bar graph, it is notable that the cime trends are steadily rising from just over 31,000 in 2014 to more than 33,500 in 2016, before slightly coming to a stable form in 2017. The bar grah helps analysts and city officials to understand crime trends, this could be crucial for allocating resources or analyzing the causes behind the shifts in the crime rates.

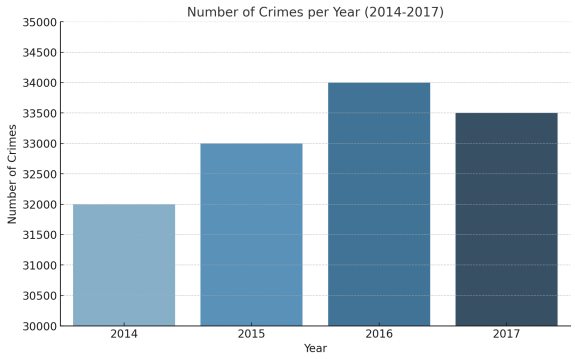

Figure 1: Number of Crimes per Year (2014-2017)

The result is followed by the accuracy of the numeric encoded model. The below figure titled “Accuracy for Numeric Encoded Model” is the output for the representation of the performance across different crime categories that have used numeric encoding. From the figure below, it is observed that the model performs best on Assault with a high recall of 0.85, while Theft Over has the lowest performance, showing a precision of 0.14 and a recall of 0.02. The overall accuracy of 0.65 highlights predictive power for common crime types like Auto Theft and Break and Enter.

Accuracy for Numeric Encoded Model: 0.6465354051728501

	precision	recall	f1-score	support
Assault	0.67	0.85	0.75	52112
Robbery	0.64	0.34	0.45	9056
Break and Enter	0.58	0.43	0.50	18413
Theft Over	0.14	0.02	0.04	3297
Auto Theft	0.60	0.53	0.56	15907
accuracy			0.65	98785
macro avg	0.53	0.43	0.46	98785
weighted avg	0.62	0.65	0.62	98785

Figure 2: Accuracy for Numeric Encoded Model

The second output displayed below represents the "Accuracy for One Hot Encoded Model," which helps display the model performance using one-hot encoding. With an accuracy of 0.66, it shows an improvement over the previous model, particularly in Assault predictions, where recall reaches 0.88. However, predictions for Robbery and Theft remain challenging, with Robbery showing better precision but lower recall. This plot helps identify the effects of encoding techniques on the classification performance of various crime categories.

Accuracy for One Hot Encoded Model: 0.6638052335880954

	precision	recall	f1-score	support
Assault	0.68	0.88	0.76	52112
Robbery	0.75	0.32	0.45	9056
Break and Enter	0.63	0.45	0.52	18413
Theft Over	0.15	0.01	0.02	3297
Auto Theft	0.63	0.54	0.58	15907
accuracy			0.66	98785
macro avg	0.57	0.44	0.47	98785
weighted avg	0.65	0.66	0.64	98785

Figure 3: Accuracy for One Hot-Encoded Model

The last output that was analyzed is the "Accuracy for Balanced Class Weight Model." This output highlights the model's attempt to handle imbalanced data. Despite adjustments, the overall accuracy of 0.65 remains similar to the numeric encoded model. Performance on Assault remains consistent, while Robbery and Theft Over still suffer from lower recall. This plot serves to illustrate how class weighting does not significantly improve predictions for less frequent crime types but maintains strong predictive power for more common categories like Assault.

Accuracy for Balanced Class Weight Model: 0.6459381485043276

	precision	recall	f1-score	support
Assault	0.67	0.86	0.75	52112
Robbery	0.65	0.35	0.46	9056
Break and Enter	0.59	0.42	0.49	18413
Theft Over	0.14	0.02	0.04	3297
Auto Theft	0.61	0.52	0.56	15907
accuracy			0.65	98785
macro avg	0.53	0.43	0.46	98785
weighted avg	0.62	0.65	0.62	98785

Figure 4: Accuracy for Balanced Class Weight Model

The study was able to visualize the confusion matrix after the outputs. Figure 5 is titled "Confusion Matrix" and is the visual breakdown of the classification model's accuracy in predicting five major crime categories: Assault, Robbery, Break and Enter, Theft Over, and Auto Theft. The matrix highlights correct predictions along the diagonal, with notable performance in predicting Assault crimes accurately. It also provides insights into where the model misclassifies, such as in the Theft Over category, where prediction accuracy is lower. This visualization helps stakeholders identify areas where the model performs well and where further refinement is needed.

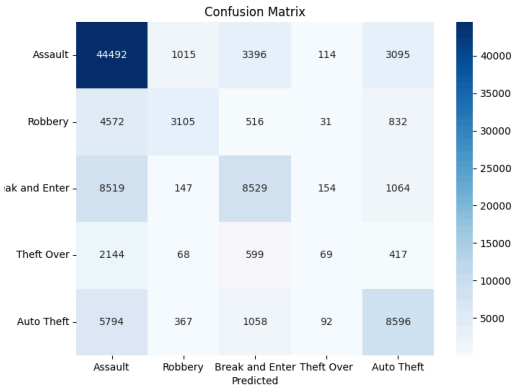

Figure 5: Confusion Matrix

5. CONCLUSION AND FUTURE SCOPE

This project aimed to predict crime occurrences in Toronto using machine learning models, specifically focusing on various encoding techniques to assess their impact on model performance. The results showed that the Random Forest model achieved moderate accuracy, with better performance in predicting Assault cases, but faced challenges in predicting less frequent crimes like Theft Over and Robbery. The use of One-Hot Encoding offered slight improvements in accuracy over Numeric Encoding, but incorporating balanced class weights did not significantly enhance the results. These findings highlight the complexity of predicting crime due to data imbalances and the diverse nature of criminal activities.

Looking forward, there are several potential areas for improvement. One key aspect is addressing the class imbalance through data augmentation techniques, such as SMOTE, to better predict underrepresented crime categories. Incorporating additional features, such as socio-economic factors, weather conditions, or proximity to law enforcement, could provide a more comprehensive dataset, improving predictive accuracy. Additionally, exploring deep learning models, such as neural networks or recurrent neural networks, could capture more complex patterns within the data. Another promising direction is integrating real-time predictions to assist law enforcement in making proactive decisions and allocating resources effectively. Finally, adding geospatial data for hot spot analysis could provide further insights into crime-prone areas, helping to refine predictions based on location-specific trends. These enhancements could significantly improve the model's capability to predict crimes and support law enforcement efforts in crime prevention and resource management.

6. REFERENCES

[1] Tamilselvi, M., V. K, and P. T., "HDLCP: Experimental Analysis and Development of Hybrid Deep Learning Methodology for Crime Scenario Assessment and Prediction," IEEE, IC-CGU, 2024, pp. 1-6, doi: 10.1109/ic-cgu58078.2024.10530836.

[2] Yadav, R., and S. Kumari Sheoran, "Crime Prediction Using Auto Regression Techniques for Time Series Data," IEEE, ICRAIE, 2018, pp. 1-5, doi: 10.1109/icraie.2018.8710407.

[3] Yao, S., Y. X., and D. Z., "Prediction of Crime Hotspots Based on Spatial Factors of Random

Forest," IEEE, ICCSE, 2020, pp. 1-6, doi: 10.1109/iccse49874.2020.9201899.

[4] Mandalapu, V., R. M., and S. R., "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," IEEE Access, 11, pp. 60153-60170, 2023, doi: 10.1109/access.2023.3286344.

[5] Sivanagaleela, B., and S. Rajesh, "Crime Analysis and Prediction Using Fuzzy C-Means Algorithm," IEEE, ICOEI, 2019, pp. 1-6, doi: 10.1109/icoei.2019.8862691.

[6] Thomas, B.A., and S. Raja, "Crime Mapping and Predictive Analysis of Crimes in Maryland, USA," IEEE, IConSCEPT, 2024, pp. 1-6, doi: 10.1109/iconscept61884.2024.10627834.

[7] Almaw, A., and K. Kadam, "Crime Data Analysis and Prediction Using Ensemble Learning," IEEE, ICICCS, 2018, pp. 1918-1923, doi: 10.1109/ICCONS.2018.8663186.

[8] Julian, A., and V. K., "Analysis and Prediction of Crimes Against Women," IEEE, INOCON, 2024, pp. 1-5, doi: 10.1109/INOCON60754.2024.10512112.

[9] K. T. M., L. T. N., M. Ithihas, N. R. Shetty, A. H. N., and S. Hebbar, "Crime Type and Occurrence Prediction Using Machine Learning," IEEE, ICAIT, 2024, pp. 1-5, doi: 10.1109/ICAIT61638.2024.10690652.

[10] V. K., R. K. S., V. R. R., N. Mekala, S. P. Sasirekha, and R. Reshma, "Predicting High-Risk Areas for Crime Hotspot Using Hybrid KNN Machine Learning Framework," IEEE, ICIRCA, 2023, pp. 848-852, doi: 10.1109/ICIRCA57980.2023.10220738.

[11] Open data (no date) Toronto Police Service Public Safety Data Portal. Available at: https://data.torontopolice.on.ca/pages/open-data (Accessed: 25 October 2024).