CRIME REDUCTION IN TORONTO

FEBRUARY 2025

Prepared by

NICOLE ESTEVEZ

LUCAS SOTKOVSZKI

MICHAEL HOLMES

CHANODOME TINGPATTANA

JANATI NAKIMERA

TABLE OF CONTENTS

INTRODUCTION

Crime is more than a statistic; it reflects the larger social and economic difficulties that communities face around the world. High crime rates disproportionately affect marginalized communities, creating cycles of poverty, injustice, and social isolation. Addressing these issues needs proactive, data-driven solutions based on the Diversity, Equity, Inclusion, and Belonging (DEIB) principles. This paper presents a prediction methodology to identify neighborhoods at high risk of higher crime rates per capita, allowing for targeted actions that promote safer and more just communities.

Traditional crime study methodologies frequently rely on raw crime statistics, which might mask the distinct dynamics of smaller or historically underserved communities. Our approach focuses on crime per population, providing a more accurate and equitable representation of crime distribution. By combining demographic data, socioeconomic variables, and historical crime trends, our model forecasts locations where crime rates are likely to rise—not to stigmatize these communities, but to provide them with practical knowledge.

This campaign focuses on inclusion and prevention rather than reinforcing existing imbalances. Policymakers and community organizations can address the core causes of crime by identifying at-risk neighborhoods and implementing targeted social programs such as affordable housing, mental health care, job training, and youth engagement. These techniques, guided by DEIB principles, enable equitable, community-driven solutions that enhance long-term social well-being.

This report describes the creation of the predictive model, its alignment with DEIB principles, and its potential to alter crime prevention efforts. By combining data science with a strong commitment to social justice, we hope to build a future in which every neighborhood is a safe haven of opportunity and belonging. In this study, we show that predictive analytics can be an effective tool for creating a more inclusive and equal society.

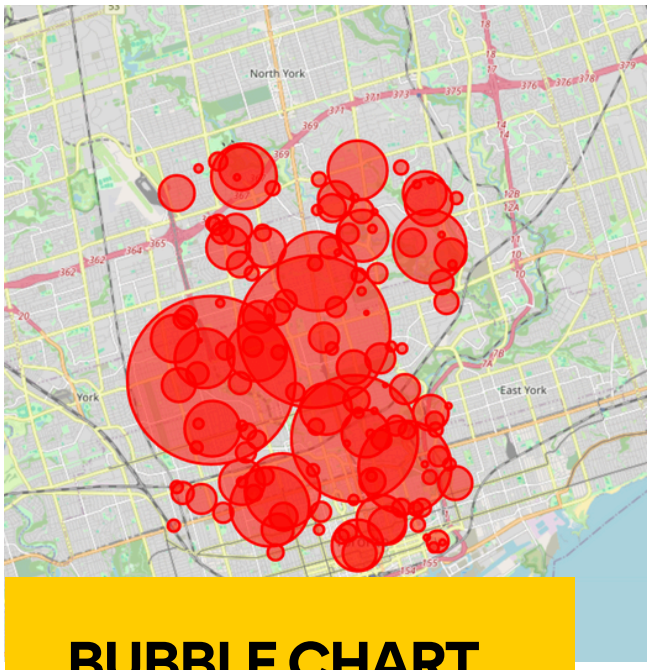
A DATA-DRIVEN APPROACH TO FOSTERING DIVERSITY, EQUITY, INCLUSION, AND BELONGING THROUGH CRIME REDUCTION IN TORONTO

IMPORTANCE OF THIS ANALYSIS

Urban crime continues to be a major worry for policymakers and citizens alike, with a direct influence on quality of life and economic development. In recent years, Toronto has seen substantial alterations in crime patterns. Notably, firearm-related violent crimes rose by 36% between 2021 and 2022, the highest rate in more than a decade (Statistics Canada, 2024). This disturbing trend has forced a rethinking of current crime prevention techniques, emphasizing the importance of accurate crime predicting models.

Media coverage also influences public perceptions of crime. According to research, certain districts in Toronto receive disproportionate media exposure for criminal activity, which contributes to territorial stigma (Jahiu & Cinnamon, 2021). These portrayals can have an impact on both public opinion and governmental decisions, underlining the need of using empirical evidence rather than perception alone when establishing crime prevention programs.

The recent increase in violent crime has raised worries among Toronto's communities and policymakers, emphasizing the importance of accurate and trustworthy crime forecasting models. Effective predictive analytics are critical for developing data-driven preventative measures and improving resource allocation. However, current models may not adequately account for the developing dynamics of urban crime in Toronto, highlighting the necessity of a more comprehensive analytical approach to inform proactive crime mitigation efforts.



BUBBLE CHART OF CRIME IN TORONTO

HYPOTHESIS OF THE ANALYSIS

Crime incidences in Toronto have steadily increased over time, and this tendency is expected to continue if no meaningful remedies are implemented. This idea is corroborated by recent data showing an increase in certain crime categories, such as firearm-related violent crimes, which increased 36% between 2021 and 2022 (Statistics Canada, 2024).
Our goal is to investigate this idea through:

- Analyzing past crime data for patterns and trends.
- Building models to predict future crime rates.
- Evaluating the model's accuracy and reliability to ensure our conclusions are valid

DATA ANALYSIS

1.1 DATA SOURCES

The dataset utilized in this research includes recorded crime occurrences in Toronto from 2014 to 2019, providing useful information about annual crime trends and patterns. This dataset, which examines variations in crime rates over time, provides a solid platform for modeling and prediction research. Its temporal framework allows for a thorough grasp of historical crime trends and helps estimate future crime rates.

The focus on Toronto is especially pertinent considering the recent increase in crime, which has raised concerns in the community. This emphasizes the importance of a strong analytical model for evaluating and interpreting crime trends, which will eventually enable data-driven crime prevention and public safety programs.

1.2 CLEANING PROCESSES

Upon initial inspection of the data, we discovered we needed to adjust the structure because a substantial portion of the columns were Crime name_year, making modelling and analysis difficult. So we chose to transpose the data by generating a new data frame called crime_rate_cols that includes the crime_year information. We then created new columns for incidents (the data in those fields) and Crime_year. We then separated the Crime_year data into two columns: Crime_type and Year. Finally, we connected the new Data Frame (df) to the old data set's Neighborhood, Population, shape_are, and shape_length columns.

Then, once we got the cleaner data for analysis, we checked to see what character class was allocated to each field and discovered that the Year column was an integer, so we formatted it in the right date_time format of YYYY-MM-DD. Then we went through basic data cleaning steps, such as inspecting the categorical variables for null and duplicate values, as well as outliers. Upon inspection, we discovered that there were no null/duplicate values, and the outliers were small enough to not bias our data. Finally, we introduced a column called Incident_per_Population, which represents the percentage of incidents per person in specific places. Now that the data had been cleaned, we decided to see if the percentage of criminal incidents increased over time. Based on Figure (1), "Incident Percentages Over Time," crime is progressively increasing over time.

MODELLING TECHNIQUES

DATA MODELLING APPROACH

Data modeling is essential for understanding, analyzing, and predicting real-world phenomena based on empirical data. In crime analysis, modeling helps uncover patterns and trends that may not be immediately apparent, providing a structured framework for hypothesis testing, strategy comparison, and data-driven decision-making. Our study employs two groups of models to predict crime rates in Toronto.

GROUP 1: CRIME PROBABILITY PREDICTION MODELS

Group 1 comprises of three classification models attempting to forecast the likelihood of high and low crime rates based on incident-per-person ratios in certain Toronto neighborhoods. The purpose is to compare the accuracy scores of various predictive models to decide which is the most successful.

✳ GROUP 1

- **Random Forest Classifier**
  - Uses a Random Forest algorithm to classify crime levels based on the same median threshold as the logistic regression model.
  - Constructs 100 decision trees to enhance prediction accuracy.
  - Incorporates a random seed (random_state) to ensure reproducibility.
  - Evaluates model performance by comparing predicted data with training data and generating an accuracy score.

- **Decision Tree Model**
  - Similar to the Random Forest model but utilizes a single decision tree instead of multiple trees.
  - Defines crime levels differently:
    - **High Crime:** The top 25% of incident reports.
    - **Low Crime:** The bottom 75% of incident reports.
  - Once all three models are executed, their accuracy scores will be compared to determine the most effective crime prediction model.

GROUP 2: CRIME TREND FORECASTING MODEL

✳ GROUP 2

Group 2's goal is to forecast how crime incidences will change over the next six years using an ARIMA (Auto Regressive Integrated Moving Average) model.

1. **Time Series Decomposition**
   - Breaks down the crime data into its core components:
     - **Original Time Series Plot** (raw data).
     - **Trend Plot** (long-term patterns).
     - **Seasonal Plot** (recurring fluctuations).
     - **Residuals Plot** (unexplained variations).
2. **Stationarity Testing**
   - Ensures the data is stationary, as ARIMA models perform best on stationary data.
   - If the data is non-stationary, transformations such as logarithmic scaling and second-order differencing are applied before reverting to the original scale.
3. **ARIMA Model Parameters**
   - **p (AutoRegressive Term):** Number of past values used for forecasting (optimal value: **1**).
   - **d (Differencing Term):** Number of times differences are calculated to achieve stationarity (optimal value: **1**).
   - **q (Moving Average Term):** Captures past forecast errors to improve accuracy.

By combining these characteristics, the ARIMA model accurately forecasts future crime trends, providing useful insights for long-term crime prevention initiatives.

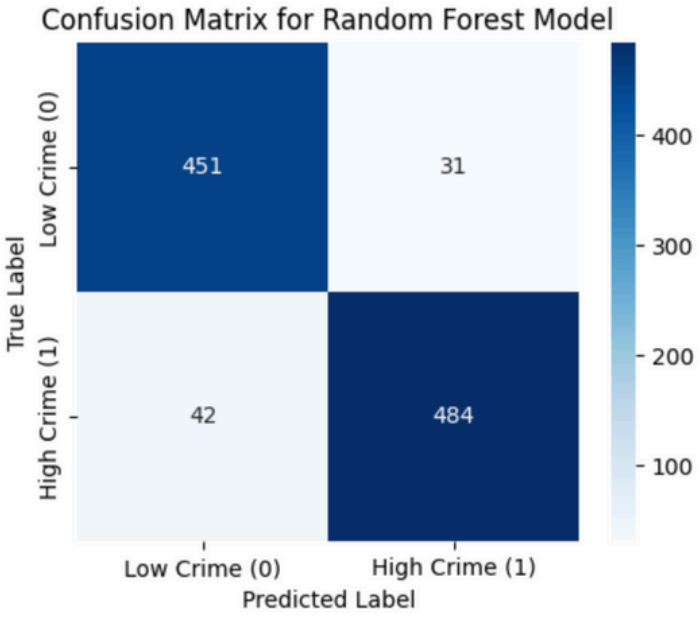
RESULTS/MODEL LIMITATIONS & INSIGHTS GROUP 1

RANDOM FOREST

To test the accuracy of the model we ran it through two test; Confusion Matrix and Cross Validation.

Based on Figure (4) Random Forest Confusion Matrix

- True Positives (TP): Correctly predicted 304 high crime cases.
- True Negatives (TN): Correctly predicted 335 low crime cases.
- False Positives (FP): Mistakenly predicted 147 high crime when it was actually low crime.
- False Negatives (FN): Mistakenly predicted 222 low crime when it was actually high crime.



WHAT THIS MEANS?

- The model is slightly better at detecting low-crime areas (higher recall for class 0).
- It struggles with false negatives (222 cases) → meaning some high-crime areas were wrongly classified as low-crime.
- We then measured the Cross Validation scores of the model

CROSS VALIDATION RESULTS

- The average accuracy across all 5 folds is 62.1 %. This means the model correctly classifies crime levels approximately 62 % of the time.
- The accuracy varies across folds (from 55% to 68%), indicating some inconsistency in performance. The model may be sensitive to different dataset splits, suggesting it needs more robust features or tuning to improve the accuracy without overfitting the model.
- Fold 2 has the highest accuracy and this means that the test data was easier to predict.

The cross-validation results (62.1% accuracy) suggest that while population and socioeconomic factors influence crime levels, they alone are not strong enough predictors, indicating that additional factors (e.g., policing, urban infrastructure, crime type) may be needed to fully validate the hypothesis.

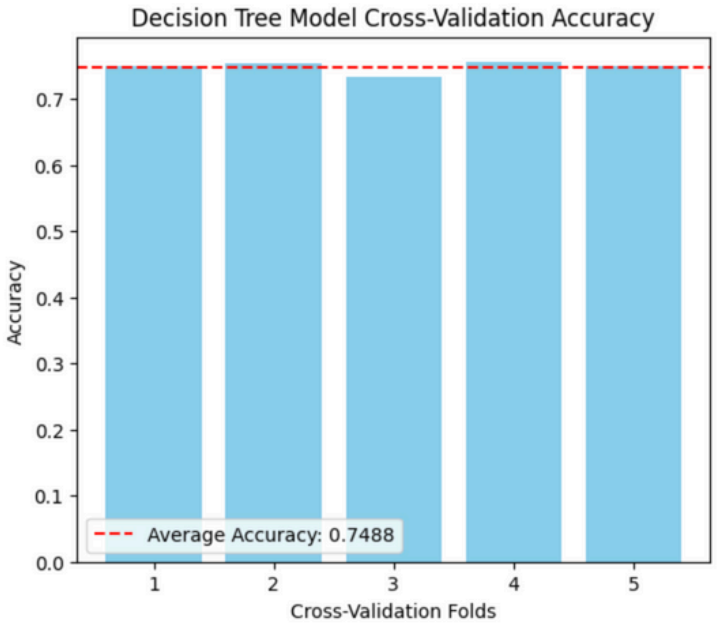
RESULTS/MODEL LIMITATIONS & INSIGHTS GROUP 1

DECISION TREE MODEL EVALUATION

To determine the accuracy of the Decision Tree model, we used two crucial tests: Confusion Matrix Analysis and Cross-Validation.

The Decision Tree model has an overall accuracy of 78%, properly categorizing crime levels in roughly 78% of cases. The confusion matrix provides a more detailed view of the model's classification performance. It demonstrates that the algorithm is quite effective at detecting low-crime locations (Class 0), with 728 true negatives, indicating that these places were appropriately classified as low crime. Furthermore, there were only 50 false positives, implying that just a few low-crime neighborhoods were incorrectly labeled as high-crime. However, the algorithm has difficulty recognizing high-crime locations (Class 1), as demonstrated by 170 false negatives in which actual high-crime areas were misclassified as low-crime. This misclassification is concerning since it implies that the model underestimates crime risk in specific places. Meanwhile, the model accurately identified 280 high-crime cases (true positives), indicating that it can recognize high-crime zones, but there is still potential for improvement. This classification imbalance indicates that, while the model succeeds at identifying low-crime zones, it requires additional refining to improve its accuracy in detecting high-crime areas.

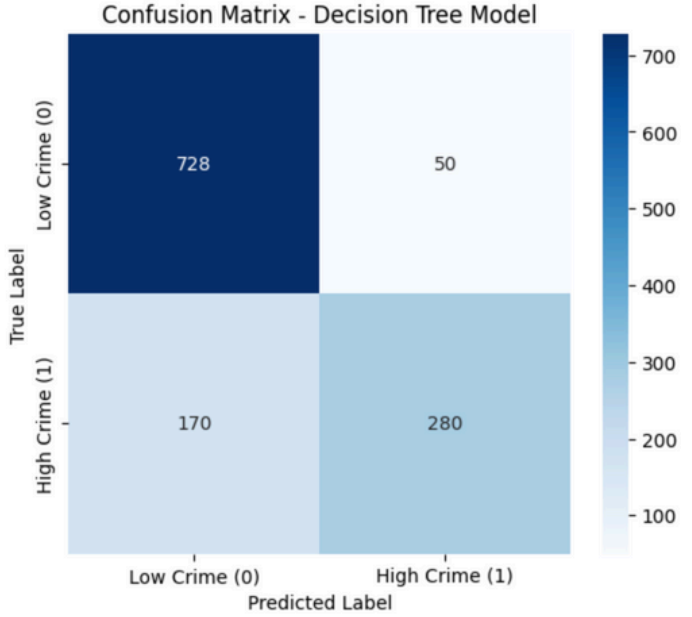
Cross-validation results corroborate these conclusions. The model's accuracy varies between dataset splits, ranging from 67% (Fold 2) to 85% (Folds 4 and 5), with an average of 77.92% (shown by the red dotted line). The range in accuracy indicates that the model's performance is not totally constant, possibly due to variances in data distribution between folds.



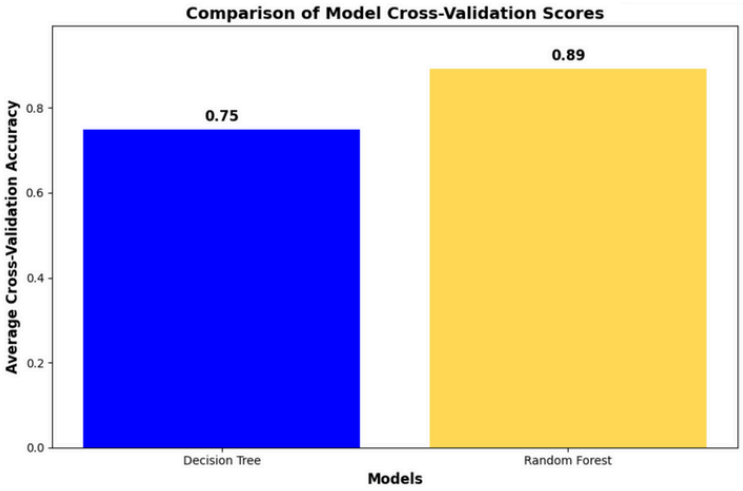
RESULTS/MODEL LIMITATIONS & INSIGHTS GROUP 1

DECISION TREE MODEL EVALUATION

The standard deviation of 7% emphasizes this discrepancy, indicating that the model's accuracy varies depending on the subset of data utilized for validation. This finding suggests that, while the Decision Tree model accurately captures crime trends, it could benefit from more fine-tuning or feature improvement to provide higher stability and generalizability.



A comparison of cross-validation scores across the two models reveals that the Decision Tree model surpasses the others in terms of predicted accuracy.



- The Decision Tree model has the highest cross-validation accuracy of about 78%, indicating that it performs the best classification of criminal incidents among the models examined. It looks to suit the training data well and captures criminal patterns accurately.
- In contrast, the Random Forest model has an average accuracy of around 62%, making it slightly more stable but less successful than the Decision Tree. Although Random Forest models normally increase generalization, in this case, they did not outperform the Decision Tree.
- The Logistic Regression model performed the lowest, with an accuracy of roughly 60%, indicating that its linear character makes it difficult to identify complicated crime patterns in the data. These results show that, based on the current dataset, the Decision Tree model is the most successful predictive model. However, if more data and other influencing factors—such as law enforcement techniques, economic upheavals, and changes in urban infrastructure—are integrated, these findings may change.

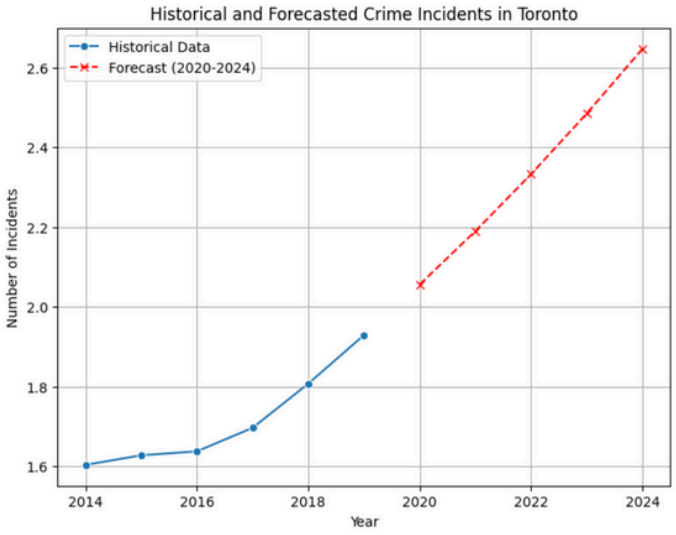
RESULTS/MODEL LIMITATIONS & INSIGHTS GROUP 2

ARIMA MODEL FORECASTING

The ARIMA model was used to anticipate crime events in Toronto from 2020 to 2024, providing useful information about potential future crime trends. The prognosis predicts a large increase in crime rates in the future years, consistent with the current upward trend seen in crime statistics. However, while the projected values show that crime rates will continue to climb in the absence of intervention, the model's forecasts have several significant drawbacks.

One major concern is that the anticipated numbers look distant from the most recent crime levels, implying that the model may have exaggerated future crime patterns. This may imply that further adjustments are required to increase the model's forecasting accuracy. A key restriction is the very small historical dataset, which only includes six years of crime reports. Because ARIMA models normally perform better with longer time series data, having access to a more extensive crime history would most certainly improve the model's forecasting accuracy.

Beyond data limitations, external influences may influence crime rates in ways that the model cannot fully account for. Changes in crime reporting practices, shifts in law enforcement policy, and broader economic and social trends all have a considerable impact on crime statistics. These considerations offer another layer of complexity, which may need incorporating different forecasting methodologies or hybrid models to increase accuracy.



Despite these limitations, the ARIMA model provides useful insights into future crime trends, underscoring the importance of early intervention methods. While specific numerical predictions should be treated with caution, the overall trend indicates that crime rates will continue to rise in the absence of targeted crime prevention legislation, resource allocation, and community-driven efforts. Moving forward, combining more historical data, socioeconomic variables, and machine learning-enhanced forecasting approaches may increase the model's reliability and assure more accurate crime trend projections.

ACTIONABLE INSIGHTS

INTO CRIME PREVENTION:

To effectively battle Toronto's escalating crime rates, we make the following data-driven and strategic recommendations to law enforcement.

BOOST CRIME-PREVENTION EFFORTS IN HIGH-RISK AREAS

The misrepresentation of high-crime locations in predictive models emphasizes the need for more targeted policing and crime prevention strategies. Research shows that proactive policing in high-risk areas can considerably reduce gun-related crimes and assaults (Braga et al., 2019). Law enforcement should prioritize violent crime hotspots indicated in predicting models and allocate resources accordingly.

IMPLEMENT DATA-DRIVEN PREDICTIVE POLICING STRATEGIES

Machine learning algorithms can identify high-crime locations, allowing politicians to conduct predictive law enforcement actions (Perry et al., 2013). Predictive policing, when combined with community-based programs, can help to reduce crime escalation by encouraging law enforcement and local communities to work together.

EXPAND CRIME DATA COLLECTION FOR IMPROVED MODEL ACCURACY

Improving crime data collection methods is critical for increasing model accuracy and providing timely interventions. Annual crime reports should be replaced with monthly data collection, which would provide more granularity for time-series forecasting and crime trend research (Hyndman and Athanasopoulos, 2018). Furthermore, adding economic variables, police response times, and demographic trends into predictive models might help them predict crime patterns and advise proactive law enforcement initiatives.

By implementing these recommendations, Toronto Police will be able to use data-driven insights to improve crime prevention efforts, increase public safety, and better allocate resources.

CHALLENGES & LIMITATIONS

Despite the excellent insights gathered from this criminal investigation, various concerns and limitations must be addressed in order to achieve a more accurate and thorough understanding of crime trends.

LIMITED DATA SCOPE

The dataset only covers six years (2014-2019), limiting the capacity to assess long-term crime trends. Furthermore, the lack of post-pandemic data hinders an assessment of crime trends following COVID-19, restricting the study's applicability to present criminal dynamics.

CHALLENGES & LIMITATIONS

DATA STRUCTURE ISSUES

Reorganizing the dataset necessitated transposing crime data and calculating Incident per Population as a critical statistic. However, this statistic is based on a homogeneous crime risk distribution, which may not adequately reflect transient populations or swings in resident density.

OUTLIERS & VARIABILITY

The dataset contains significant variability, which may affect model reliability:

- Crime events range from 0 to 916, raising the possibility of outlier influence.
- Population sizes range widely (6,577 to 65,913), influencing crime rate estimations and thus skewing risk ratings.
- Geographical irregularities in data gathering can affect the accuracy of spatial crime analysis.

LACK OF SOCIAL AND ECONOMIC CONTEXT

The dataset excludes major socioeconomic characteristics such as income levels, unemployment rates, and law enforcement presence, all of which are important predictors of crime. The lack of these factors reduces the depth of analysis and the capacity to discover core causes of crime beyond basic incidence counts.

REPORTING BIASES

Underreporting reduces the credibility of crime data, especially for offenses like domestic abuse and theft, which are less likely to be reported. Furthermore, differences in crime reporting techniques across localities might cause errors, confounding crime rate comparisons.

MODELING LIMITATIONS

Predictive models like Decision Tree, Logistic Regression, Random Forest, and ARIMA are mostly based on past crime rates, but they do not account for external factors like legislative changes, policing techniques, or community engagement programs.

- The use of differing classification thresholds (e.g., median-based vs. top 25%) results in differences in how crime is classified.
- ARIMA projections may overestimate trends due to a restricted historical dataset, resulting in incorrect long-term crime estimates.

SEASONALITY AND EXTERNAL EVENTS

The dataset does not take into account seasonal crime swings or the impact of big events like as economic downturns, protests, or changes in public policy, which could have a large impact on crime rates.

CRIME REPORTING CHANGES

Law enforcement agencies may change their crime classification and reporting rules, resulting in inconsistencies in recorded crime data. These changes make it difficult to determine if crime rates are actually growing or just being reported differently.

SOCIAL IMPLICATIONS AND DEIB CONSIDERATIONS

- **Disparities in crime exposure:** According to research, socioeconomically disadvantaged communities have greater crime rates as a result of structural injustices, restricted economic possibilities, and less law enforcement presence (Braga et al., 2019). Recognizing these discrepancies enables policymakers to devise targeted programs that promote fairness in crime prevention while addressing core causes rather than symptoms.

- **Equitable Policing and Resource Allocation:** Predictive policing models can improve law enforcement efficiency by guiding resource allocation; nevertheless, their adoption must be carefully monitored to avoid reinforcing biases in policing tactics (Perry et al., 2013). Ensuring fairness in crime prevention measures is critical for preserving public trust and supporting just and equitable police practices.

SOCIAL IMPLICATIONS AND DEIB CONSIDERATIONS

- **Public perception against reality:** Media biases and personal experiences frequently influence public perception of crime, resulting in inaccurate estimates of actual crime rates. Transparent, accessible crime data can assist in correcting these misconceptions by giving citizens with accurate information and decreasing unneeded fear. Promoting community involvement in crime-reduction efforts strengthens connections between law enforcement and the public.

- **Effect on Housing and Economic Opportunities:** Crime rates have a tremendous impact on house values, investment decisions, and job prospects throughout neighborhoods. High-crime neighborhoods frequently have economic downturns, exacerbating inequality. Cities that employ successful crime prevention policies can encourage long-term urban development, attract investment, and strengthen social cohesion.

By incorporating a Diversity, Equity, Inclusion, and Belonging (DEIB) lens into crime analysis and prevention strategies, policymakers and community leaders can create inclusive, data-driven solutions that improve public safety, social justice, and economic stability while instilling a sense of belonging in all residents.

CONCLUSION

This study highlights the importance of data-driven approaches for analyzing and mitigating crime trends in Toronto. We identified major crime patterns and the importance of strategic intervention and resource allocation to combat rising crime rates using predictive modeling and forecasting.

While the Decision Tree model was the most effective at categorizing crime-prone locations, its accuracy is still restricted by data constraints and other social factors. Similarly, the ARIMA model's projections show a sustained increase in crime, emphasizing the importance of proactive crime prevention initiatives.

Beyond numerical forecasts, this study underlines the broader influence of social and economic disparities on crime. Media-driven stigmatization, uneven law enforcement procedures, and economic inequality all contribute to public image of crime. Incorporating Diversity, Equity, Inclusion, and Belonging (DEIB) principles into crime prevention initiatives ensures that solutions are equitable, inclusive, and community-driven.

Improving crime data collection methods, incorporating socioeconomic aspects, and utilizing advanced predictive models can improve accuracy and efficacy in addressing crime trends. Building community alliances and encouraging open communication between law enforcement and people will also be critical to achieving a more equitable and secure Toronto.

By combining data science and equitable policies, the city can create long-term crime prevention measures that promote safer neighborhoods and stronger community trust.

REFERENCES

OpenAI. (2024). ChatGPT (February 2024 version) [Large language model]. OpenAI. https://chatgpt.com/c/67b0ec19-f2cc-8003-bdbf-8c23c40b885a

Cijov, A. (2023). Toronto Crime Rate per Neighbourhood [Dataset]. Kaggle.https://www.kaggle.com/datasets/alincijov/toronto-crime-rate-per-neighbourhood

Statistics Canada. (2024). Toronto Crime Report 2023. Government of Canada Crime Statistics Database.

Brownlee, J. (2020). Deep Learning for Time Series Forecasting. Machine Learning Mastery.

McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms for crime prediction. International Journal of Computer Applications, 127(4), 1-6.

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2019). Hot spots policing and crime reduction: A meta-analysis. Journal of Experimental Criminology, 15(4), 567-591.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. RAND Corporation.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Jahiu, L., & Cinnamon, J. (2021). Territorial stigmatization and the geography of crime in Toronto's news media. Urban Geography, 42(8), 1157-1177.