Neighbourhood Crime Insights (NCI): ML-Based Crime Analysis

CSCI6409: Process of Data Science
Project Report

Viren Joshi (viren.joshi@dal.ca)
Samruddhi Mulay (sm824555@dal.ca)
Shreya Rao (sh316686@dal.ca)

Abstract

This project aims to analyze crime and socio-economic data across Nova Scotia, detecting community-level risk profiles. Using unsupervised clustering methods, particularly HDBSCAN, we identified high and low-crime clusters characterized by factors such as income, education, age, and population density. The project also highlights a decrease in overall crime rates alongside economic improvements. The dataset was created by combining Nova Scotia Open Crime Data with demographic attributes from the Canada Census data for 2016 and 2021, including median income, education levels, age distribution, and visible minority status. After preprocessing and exploratory data analysis, we performed clustering to group locations based on both crime incidence and socio-economic similarity. The resulting clusters revealed distinct regional profiles ranging from low-crime, highly educated rural zones to high-crime urban areas with significant socio-economic challenges. Notably, our comparative analysis between 2016 and 2021 data demonstrated a general decline in average crime rates alongside improvements in economic indicators. These insights underscore the impact of education and income on public safety and provide a data-driven foundation for targeted policy interventions. Our work sets the stage for real-time monitoring applications and broader comparative studies across Canadian regions.

1 Introduction

Crime remains a pressing concern in modern urban planning and governance, impacting not only public safety but also the economic and social well-being of communities. Accurate and timely crime analysis is crucial for law enforcement agencies, policymakers, and local governments to allocate resources efficiently, design effective prevention strategies, and foster safer environments for residents. Traditional crime analysis methods often focus on raw incident data, overlooking the broader socio-economic contexts that contribute to criminal behavior.

This project seeks to bridge that gap by utilizing machine learning techniques to analyze the relationship between crime and socio-economic indicators across Nova Scotia. Using open-source datasets from the Nova Scotia Crime Statistics [1] and Statistics Canada Census (2016 and 2021) [2],[3], we aim to uncover latent patterns that distinguish high-risk from low-risk areas and investigate how these patterns have evolved over time.

Our primary objective is to develop a data-driven model that can identify clusters of neighborhoods sharing similar crime and socio-economic characteristics. The results show that including socio-economic data helps make crime analysis more useful and can be applied in real-time tracking or comparing crime across provinces.

2 Problem Statement

The integration of socio-economic data with crime statistics has emerged as a promising approach to gain deeper insights into the root causes and geographic distribution of crime, over the recent years. However, traditional crime statistics lack the socio-economic context needed to explain shifts in crime patterns. This project integrates demographic and economic indicators with crime data to help public safety authorities identify high-risk areas, understand contributing factors, and predict future trends for proactive, data-driven policy decisions.

3 Literature Review

3.1 Predicting Crime through Spatio-Temporal Data Analysis

Recent advancements in data science have enabled the development of robust models for crime prediction by leveraging spatio-temporal patterns. Studies such as those analyzing 12-year crime datasets from San Francisco demonstrate the efficacy of machine learning techniques, including decision trees, k-nearest neighbors (KNN), and ensemble methods like Random Forest and XGBoost [4]–[6]. While simpler algorithms like KNN and decision trees often struggle with precision and class imbalance, ensemble methods significantly improve accuracy by aggregating predictions across multiple models [6], [7]. For instance, Random Forest, combined with under-sampling techniques to address data imbalance, has proven effective in mitigating biases toward majority crime classes [7]. Performance metrics such as log-loss further refine these models by penalizing misclassifications, ensuring reliability in imbalanced datasets [6]. These approaches highlight the potential of supervised learning to forecast crime hotspots and trends, enabling law enforcement to allocate resources proactively and disrupt cyclical criminal activity in specific regions [8].

3.2 Socio-Economic Indicators and Crime Dynamics

The interplay between socio-economic factors and crime is a critical theme in criminological research. Studies consistently link poverty, unemployment, and income inequality to elevated crime rates, particularly in socio-economically disadvantaged urban areas [6]. For example, in Southern Ontario's intermediate-sized cities, lower income levels and the presence of alcohol or cannabis stores correlate strongly with property crime, suggesting that economic vulnerability and specific land-use types create opportunities for criminal activity. However, the relationship is nuanced: while income inequality (measured via the Gini index) has shown a negative correlation

with crime in Kitchener-Waterloo-Cambridge, other cities exhibit no universal trends, underscoring the context-dependent nature of these factors. Additionally, homelessness, inadequate community resources, and industrial land use further compound risks, emphasizing the need for localized policy interventions such as progressive taxation or social welfare programs [6]. These findings reinforce the value of integrating census variables such as education, income, and employment into crime prediction models to capture systemic vulnerabilities beyond purely spatial or temporal trends. Guindon's MA thesis explores these exact themes in-depth, showing links between property crime and the presence of alcohol/cannabis stores, variations in income inequality, and urban vulnerabilities [9].

3.3 Clustering Techniques in Crime Analytics

Clustering algorithms have become powerful tools for segmenting urban landscapes into crime-prone zones based on socio-spatial characteristics. The application of k-means clustering, particularly when paired with dimension-reduction techniques like UMAP, has demonstrated superior performance in identifying geographically coherent profiles. For example, in Lima, Peru, the k-means with UMAP achieved high Calinski-Harabasz (2,683.29) and Silhouette scores (0.58), delineating clusters aligned with commercial, residential, and historical areas [5]. The adaptability of this method is evident in its response to freight trip generation (FTG) models, where linear FTG, introducing greater data variance, enabled finer-grained cluster differentiation (e.g., k = 12 clusters) [5]. In contrast, density-based approaches like HDBSCAN, while less statistically robust in some metrics (Calinski-Harabasz: 425.59), excel at identifying noise and natural cluster boundaries without predefined cluster counts, offering complementary insights in heterogeneous urban environments [4]–[6]. Such techniques align with crime analytics goals, enabling researchers to isolate high-risk zones shaped by socio-economic disparities or logistical patterns, thereby informing targeted policing strategies and urban planning interventions.

By synthesizing spatio-temporal prediction models, socio-economic contextualization, and clustering-driven spatial profiling, the literature underscores the transformative potential of integrated data-driven approaches in crime analytics. However, challenges such as model explainability, regional variability in socio-economic impacts, and algorithmic trade-offs between granularity and generalizability remain critical areas for further exploration.

4 Methodology

This section outlines the approach used to collect, preprocess, analyze, and cluster the data for identifying crime patterns across Nova Scotia. The

methodology was designed to ensure reproducibility, accuracy, and relevance to real-world applications.

4.1 Dataset Description

4.1.1 Dataset Overview

The dataset used in this study combines crime incident reports from the Nova Scotia Open Crime Data Portal [1] and socio-economic attributes from Statistics Canada's 2016 and 2021 Census datasets [2], [3]. The merged dataset spans multiple years and includes over 10,000 records at the community level, with each record representing the number and rate of crimes per 100,000 residents in a given area and year. Key attributes include geographic identifiers such as city, community name, and postal code, alongside socio-economic indicators such as median income, education level, age distribution, and percentage of visible minorities.

4.2 Data Preprocessing

The data preprocessing phase was essential to ensure the quality and consistency of the dataset before applying machine learning algorithms. This stage involved handling missing values, feature selection, encoding categorical variables, and normalizing the data to improve model performance and clustering accuracy. The steps followed were:

1. **Handling Missing Data**: Proper imputation and filtering were applied as needed to preserve data integrity and ensure meaningful analysis. This included the following:

   (a) *Dropping Columns with Excessive Missing Values*: Columns with more than 90% missing values were removed, as they contributed little to no analytical value and introduced noise. This included fields from the raw socio-economic dataset that were either too specific or inconsistently reported.

   (b) *Row Removal for Essential Fields*: Rows missing key identifiers, such as geographic region or crime type, were excluded, since they were necessary for spatial analysis and labeling.

   (c) *Imputation of Categorical Fields*: Mode imputation was used for categorical variables (e.g., visible minority status or education bracket) to maintain consistency and preserve the overall distribution across regions.

2. **Feature Engineering**: To improve the model's ability to detect patterns, we derived additional variables that provided richer context.

(a) *Combining Census and Crime Data*: Socio-economic indicators from 2016 and 2021 were linked to crime data by geographic location, enabling year-over-year comparison and temporal trend analysis.

(b) *Crime Rate Normalization*: Raw crime counts were converted into standardized rates per 100,000 residents to ensure comparability across regions with varying population sizes.

(c) *Time-Based Aggregation*: Crime data was aggregated at the annual level to align with Census reporting, smoothing out short-term fluctuations and focusing on structural patterns.

3. **Feature Encoding**: Encoding techniques like label encoding and one-hot encoding were employed, depending on the specific use case.

   (a) *Categorical Encoding*: Categorical variables such as education level and minority status were encoded using one-hot encoding to allow for proper inclusion in clustering algorithms.

   (b) *Feature Scaling*: All numerical features were standardized using z-score normalization. This ensured that variables like income, education, and age were on a comparable scale, preventing any single feature from disproportionately influencing clustering results.

4.3 Dataset Instance

A single instance of data in this project is the number and rate of a crime incident per 100,000 residents that occurred in a specific location in Nova Scotia in a particular year, as shown in Table 1.

Geography	**Year**	**Violations**	**Incidents**	**Rates**
Pictou County, Nova Scotia, Royal Canadian Mounted Police, rural [12010]	2010	Total breaking and entering [210]	104	467.79

Table 1: Single Instance of Data

The key features of the dataset can be seen in Table 2.

Feature	Data-Type	Description
Geography (Location)	Text	The location where the data was collected.
Year	Text/Number	The calendar year for the reported data.
Violations	Comma Separated Categorical values	The type or category of criminal offense recorded.
Incidents	Number	The total number of reported cases of the specified violation.
Rates	Number	The crime rate per 100,000 population.

Table 2: Key Features of Dataset.

4.4 Clustering Algorithms

To uncover underlying spatial and socio-economic patterns in the dataset, we applied a set of clustering algorithms. These models allowed us to group communities with similar crime and demographic profiles, helping identify both high-risk zones and low-crime outliers.

We initially employed **K-Means** due to its computational efficiency and its ability to partition data into distinct clusters based on feature-space proximity. In line with standard methodology, the optimal number of clusters was identified using the Elbow Method. However, our observations revealed that while K-Means produced well-defined segments, its assumption of spherical cluster shapes limited its effectiveness. Specifically, it struggled to capture the complex, non-spherical structures commonly present in spatial data characterized by heterogeneous geographic and demographic patterns.

To address these limitations, **DBSCAN** was subsequently applied. Unlike K-Means, DBSCAN identifies clusters based on point density rather than distance to centroids, making it well-suited for detecting irregularly shaped crime hotspots and isolating outlier or noise regions. It operates by locating dense regions of points separated by sparser zones, governed by two parameters: `eps` (the maximum distance between neighboring points) and

`min_samples` (the minimum number of points required to form a cluster). While DBSCAN demonstrated improved flexibility in capturing complex spatial structures, it encountered challenges in areas with highly variable point densities, such as, differences between urban and rural zones, leading to inconsistent clustering outcomes in some cases.

To mitigate the density sensitivity limitations observed with DBSCAN, we employed **HDBSCAN**—a hierarchical density-based clustering algorithm. Unlike DBSCAN, which requires a fixed `eps` parameter, HDBSCAN constructs a hierarchy of clusters and identifies the most stable ones based on density persistence. This characteristic enabled us to adapt the clustering process to spatial contexts with varying point densities. As a result, HDBSCAN proved effective for our dataset, which encompassed both densely populated urban neighborhoods and sparsely distributed rural communities.

5 Experiment

This component of the study focuses on identifying meaningful crime clusters across Nova Scotia using unsupervised machine learning techniques. The objective is to group neighbourhoods based on shared crime characteristics and socio-economic profiles. In this way, we seek to uncover insights into the underlying factors contributing to elevated crime levels in certain regions, and to examine how demographic and economic conditions influence these spatial patterns.

Preprocessing was a crucial step in our pipeline. First we manually included geographic coordinates - latitude and longitude, to capture spatial variation. We then integrated 2016 Census data and 2021 Census data with the crime data based on geography. This temporal alignment helped us maintain socio-economic context when analyzing trends over time. The datasets were cleaned by imputing missing values and ensuring consistency in data types. Categorical values were encoded, and derived metrics such as crime rate per capita were computed to better reflect localized crime intensity. All numeric features were standardized using z-score normalization to ensure comparability during clustering. This preprocessing step was essential since variables like income and population could otherwise disproportionately influence distance calculations. After cleaning and filtering, the dataset included incident counts, crime rates per 100,000 residents, and census-based attributes such as median income, average income, population density, education levels (no diploma, high school, post-secondary), age group proportions, and visible minority status. Geographic coordinates (latitude and longitude) were also retained to support spatial clustering.

or the clustering task, we experimented with K-Means and DBSCAN; however, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications

with Noise) was ultimately selected due to its ability to handle clusters of varying densities, as detailed in Section 4.4. We tuned the model's hyperparameters using a grid search and set the minimum cluster size to 3, allowing smaller but distinct neighbourhoods to form their own clusters. Euclidean distance was used as the distance metric, which was appropriate since all features were normalized. To check the model's stability and reliability, we ran it multiple times. The quality of the clustering was measured using the Silhouette Score, which shows how well each point fits within its cluster compared to others. Our model achieved a Silhouette Score of 0.87, indicating strong internal consistency and clear separation between clusters.

Visualization also played a key role in validating the clustering results. Using Folium maps and scatter plots, we verified that the clusters corresponded to real-world geographic patterns. High-crime clusters emerged in areas such as Cape Breton and Annapolis Royal, characterized by low income, limited education, and higher population density. In contrast, low-crime clusters like Enfield and Amherst showed strong socio-economic indicators such as high income or post-secondary education levels. Some surprising cases also emerged- Berwick, for instance, exhibited low income but no reported crime, possibly due to strong community cohesion or underreporting.

The presence of outliers, grouped into Cluster-1 by HDBSCAN, is particularly noteworthy. These areas do not align clearly with any dominant cluster pattern, suggesting a mix of socio-economic characteristics or transitional dynamics. Their ambiguous positioning warrants further investigation, as they may reflect regions undergoing change or exhibiting complex and layered crime behaviours.

Overall, from our experiment, we demonstrate that clustering algorithms, particularly HDBSCAN, is a powerful algorithms for revealing hidden structure in complex crime and demographic data. The combination of strong quantitative performance and meaningful geographic insights validates our approach and lays the foundation for developing targeted policy interventions and community safety strategies.

6 Results

The clustering results provided significant insights into how socio-economic and geographic factors relate to crime patterns across Nova Scotia. Using HDBSCAN, we identified a diverse set of clusters that represented neighbourhoods with similar crime levels and demographic compositions. These clusters were then analyzed both quantitatively and qualitatively to interpret their underlying characteristics.

6.1 Exploratory Data Analysis

6.1.1 Cluster Performance and Distribution

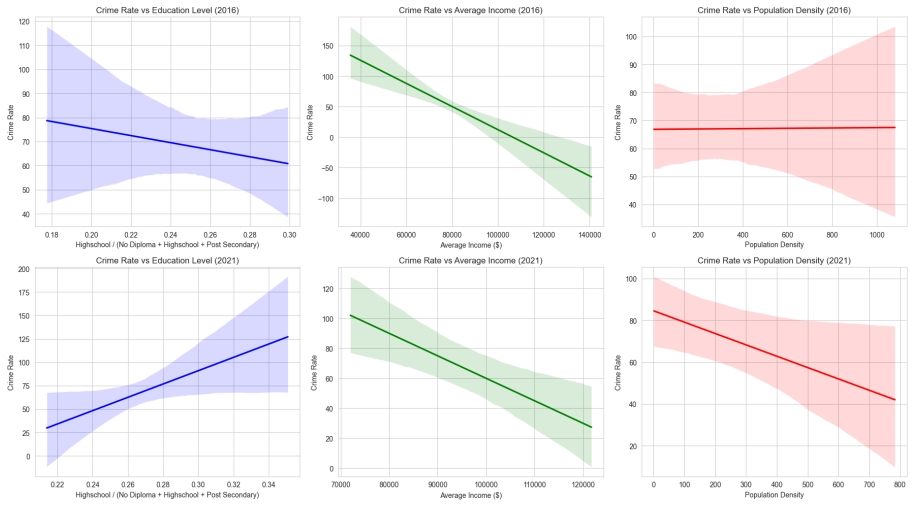

Figure 1: Crime Rate VS Socio-Economic Factors

From figure 1, we observed varying relationships of these demographical factors with crime upon initial investigation. Therefore, we decided to leverage clustering for better analysis of the data. The HDBSCAN model produced 21 distinct clusters along with one noise cluster (Cluster -1), which captured outlier data points that didn't fit well into any group of points. The Silhouette Score of 0.87 validated the quality of these clusters, indicating strong intra-cluster similarity and distinct inter-cluster separation. This score suggests that the chosen features and clustering configuration allowed for meaningful groupings that reflect real differences between communities. Figure 2 shows the map of Nova Scotia spreading across the provinces, representing different clusters as the output of HDBSCAN.

The clusters varied not only in size but also in socio-economic makeup and crime rates. For instance, Cluster 2, which predominantly included regions in Cape Breton, exhibited higher crime rates and was characterized by low median income, high population density, and moderate education levels. In contrast, Cluster 20, which included communities like Enfield in East Hants, showed virtually no crime. These areas were generally rural, wealthier, and had a higher percentage of post-secondary educated residents.

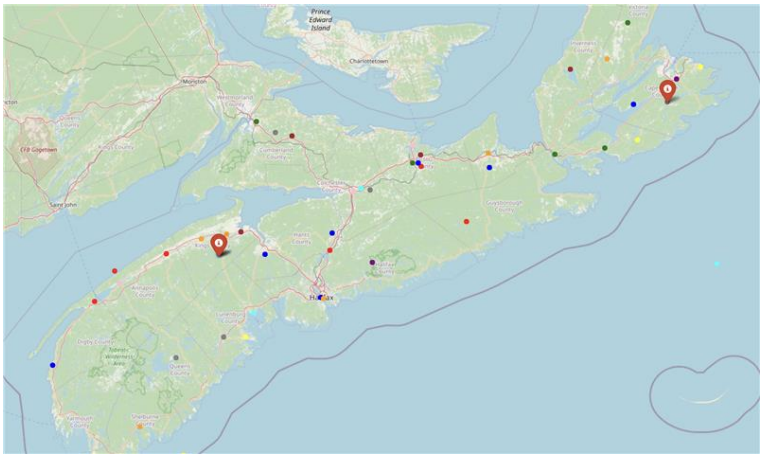

Figure 2: Visualization of crime clusters using HDBSCAN clustering.

6.1.2 High-Crime vs Low-Crime Cluster Insights

High-Crime Clusters: Cluster 15 (Annapolis Royal) was notable for high crime rates in small rural areas with an aging population, low education, and low income. These communities often lacked economic opportunities, which appeared to correlate with elevated crime activity. Cluster 2 (Cape Breton Region) consisted of larger urban centers that also struggled with poverty and under-resourced public infrastructure, contributing to persistent moderate-to-high crime levels.

Low-Crime Clusters: Cluster 24 (Amherst) was a small, aging town with no reported incidents during multiple years in the dataset. Despite not being economically wealthy, strong community cohesion or possible underreporting might explain the lack of recorded crime. Cluster 22 (Berwick) stood out as a low-income area with surprisingly zero recorded crime, yet its residents were highly educated. This contradicted the assumption that low income always correlates with high crime, suggesting that education might serve as a protective buffer.

6.2 Socio-Economic and Temporal Observations

An analysis of the socio-economic variables across clusters revealed clear patterns: Median income increased by approximately 9% between the 2016 and 2021 census-aligned clusters, indicating a modest improvement in economic conditions province-wide. The proportion of low-income households decreased by 43%, particularly in rural and suburban clusters, suggesting a potential link between economic uplift and reduced crime in certain regions. However, the average population age increased, which may reflect the aging demographics of the province. This shift appeared to coincide with a decline

in post-secondary education enrollment possibly due to youth migration or pandemic-related disruption as seen in figure 3. Interestingly, some clusters with moderate or low income still showed zero or very low crime, which points to the potential role of community resilience, social networks, or even data gaps like underreporting. These "anomalous" clusters provide a valuable direction for further qualitative investigation.

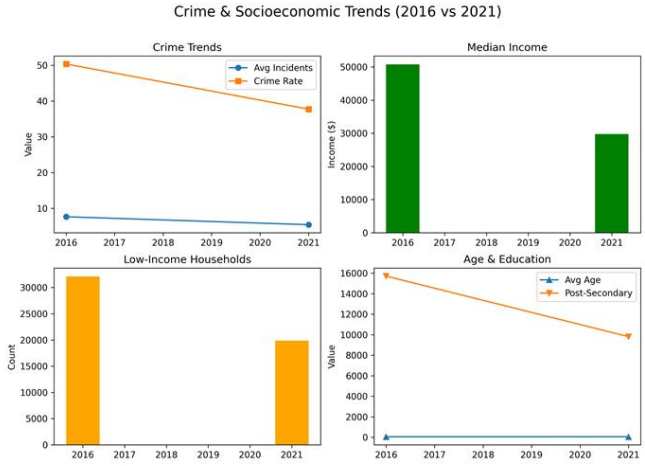

Figure 3: Comparative analysis of crime and socioeconomic indicators.

6.3 Impact of Spatial Clustering

The spatial nature of our clustering approach allowed us to map the distribution of crime-prone areas throughout Nova Scotia. We observed that crime was not concentrated solely in urban centers but was also present in some remote, economically stressed rural areas. This spatial dispersion highlights the importance of localized interventions rather than one-size-fits-all provincial strategies.

Furthermore, the inclusion of noise points by HDBSCAN (Cluster -1) helped us identify outlier neighbourhoods-either transitional or with conflicting socio-economic profiles-that didn't fit clearly into any one cluster. These areas may represent regions undergoing socio-economic change, policy interventions, or emerging crime trends.

7 Conclusion

This project set out to explore and analyze crime patterns across Nova Scotia by combining historical crime statistics with detailed socio-economic indicators from the Canadian Census. By leveraging unsupervised machine learning,

specifically HDBSCAN clustering, we were able to uncover community-level insights that traditional analyses might overlook. The results clearly indicate that socio-economic factors such as income, education level, population age, and geographic density significantly influence crime trends.

High-crime clusters tended to correlate with low income, low educational attainment, and aging populations, particularly in urban or densely populated areas like parts of Cape Breton and Annapolis Royal. On the other hand, low-crime clusters often included rural regions with high educational levels and stronger economic indicators. Interestingly, the presence of clusters with low income but minimal crime suggests that factors like social cohesion, community resilience, or even underreporting could also be influencing crime outcomes.

The performance of the HDBSCAN model was validated both quantitatively-via a high silhouette score of 0.87-and qualitatively through geographical and demographic validation of the resulting clusters. The clustering model successfully grouped neighbourhoods not just by proximity, but by shared socio-economic realities, enabling a data-driven way to understand where resources might be most needed.

This analysis underscores the value of integrating social data with crime statistics to go beyond surface-level mapping and toward actionable insights. The ability to identify high-risk zones, anomaly clusters, and socio-demographic patterns offers an opportunity for policymakers and community planners to make informed, targeted interventions.

8 Future Work

While this project provided meaningful results, it also opened up several avenues for further development and expansion. One major direction would be the integration of real-time data streams, such as 911 emergency call records or police incident reports. These data sources, when combined with the existing historical dataset, could enhance the temporal dimension of our analysis and enable real-time risk monitoring and response.

Another valuable addition would be to develop a fully interactive web-based dashboard using tools like Dash or Tableau. This would allow users like municipal leaders, law enforcement agencies and citizens to visualize crime clusters, demographic profiles, and changes over time. Such a dashboard could include drill-down capabilities by region, year, and crime type, empowering stakeholders to explore the data and generate customized insights.

From a modeling perspective, future work could explore explainable AI techniques, such as SHAP (SHapley Additive exPlanations), to further interpret the influence of individual features on the clustering results. Additionally,

supervised classification models could be developed to predict the likelihood of a region falling into a high- or low-crime cluster, based on evolving census indicators.

Lastly, the scope of the project could be extended beyond Nova Scotia to other Canadian provinces or cities. A comparative analysis of urban vs rural crime trends, or inter-provincial socio-economic drivers of crime, could provide broader insights and contribute to national-level safety policy recommendations.

Overall, the foundation built through this project sets the stage for a scalable, adaptable, and socially aware crime analytics system that can evolve alongside growing data sources and community needs.

References

[1] Government of Nova Scotia, *Crime statistics (incidents and rates for selected offences)*, Accessed: 2025-04-15, 2024. [Online]. Available: https://data.novascotia.ca/Crime-and-Justice/Crime-Statistics-Incidents-and-rates-for-selected-/m862-kmjy/about_data.

[2] Statistics Canada, *Census profile, 2016 census of population*, Accessed: 2025-04-15, 2016. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm.

[3] ——, *Census profile, 2021 census of population*, Accessed: 2025-04-15, 2021. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E.

[4] D. Dipakkumar Pandya, G. Amarawat, A. Jadeja, S. Degadwala, and D. Vyas, "Analysis and prediction of location based criminal behaviors through machine learning," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, 2022, pp. 1324–1332. DOI: 10.1109/ICECAA55415.2022.9936498.

[5] A. Regal, J. Gonzalez-Feliu, and M. Rodriguez, "A spatio-functional logistics profile clustering analysis method for metropolitan areas," *Transportation Research Part E: Logistics and Transportation Review*, vol. 179, p. 103 312, 2023, ISSN: 1366-5545. DOI: https://doi.org/10.1016/j.tre.2023.103312. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1366554523003009.

[6] K. Taha, "Empirical and experimental insights into data mining techniques for crime prediction: A comprehensive survey," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 2, Feb. 2025, ISSN: 2157-6904. DOI: 10.1145/3699515. [Online]. Available: https://doi.org/10.1145/3699515.

[7] E. Cesario, P. Lindia, and A. Vinci, "Multi-density crime predictor: An approach to forecast criminal activities in multi-density crime hotspots," *Journal of Big Data*, vol. 11, no. 1, p. 75, 2024, ISSN: 2196-1115. DOI: 10.1186/s40537-024-00935-4. [Online]. Available: https://doi.org/10.1186/s40537-024-00935-4.

[8] F. Ersöz, T. Ersöz, F. Marcelloni, and F. Ruffini, "Artificial intelligence in crime prediction: A survey with a focus on explainability," *IEEE Access*, vol. 13, pp. 59 646–59 674, 2025. DOI: 10.1109/ACCESS.2025.3553934.

[9] E. D. Guindon, "Unveiling urban vulnerabilities: Exploring property crime patterns and risk factors in intermediate-sized cities in southern ontario," English, Ph.D. thesis, Carleton University, 2024. [Online]. Available: https://repository.library.carleton.ca/files/g158bj58m.

[10] V. Upadhyay and D. Rathod, "Location-based crime prediction using multiclass classification data mining techniques," in *IOT with Smart Systems*, T. Senjyu, P. Mahalle, T. Perumal, and A. Joshi, Eds., Singapore: Springer Nature Singapore, 2022, pp. 619–626, ISBN: 978-981-16-3945-6.