Spatiotemporal Analysis and Predictive Modeling of Community Safety Indicators in Toronto

Feng Qiu
AMOD 5430: Data Visualization
Trent University, Peterborough, Ontario, Canada
Student ID: 0854002

***Abstract*—This paper examines 474,819 offence and victim records from the Toronto Police Service Community Safety Indicators open dataset. The records belong to five categories: Assault, Auto Theft, Break and Enter, Robbery, and Theft Over. I excluded the partial 2026 data, leaving 464,571 records reported from 2014 through 2025. The exploratory analysis compares annual counts, occurrence hour, weekday, premises type, neighbourhood, and approximate coordinates. I also trained two classifiers using records from 2014 through 2023, used 2024 for model selection, and kept 2025 as a test set. The random forest reached 59.2% accuracy and a macro F1 score of 0.424 on 43,033 test records. A majority class baseline reached 58.2% accuracy but only 0.147 macro F1. The small difference in accuracy matters: the model learned some useful class patterns, but it still performed poorly on Robbery and Theft Over. The maps and model describe police reported records at approximate locations. They do not measure all victimization or the risk posed by a person or neighbourhood.**

***Index Terms*—exploratory data analysis, geospatial visualization, public safety, random forest, temporal validation, Toronto**

I. Introduction

Toronto's police open data contains dates, categories, premises, neighbourhoods, and approximate coordinates. Those fields are useful, but a file with nearly half a million rows is hard to read directly. Annual charts show changes over time. Hourly plots and maps show patterns that disappear in a citywide total.

These views can also be misleading. A count map is not a neighbourhood risk map because it has no population, visitor, traffic, or land use denominator. A classification model can look accurate simply by predicting Assault, which is the largest class. The row itself also needs attention. Several rows may share an event ID because the data is stored at an offence or victim level.

This project follows the submitted proposal and studies the five Community Safety Indicator (CSI) categories in Toronto. It asks how the records vary by category, year, weekday, hour, and premises type. It also compares their spatial patterns and tests whether time and approximate place can help classify a recorded occurrence. The prediction task is narrow. It assigns a category to an existing record; it does not predict whether or where a future incident will happen.

The analysis keeps the same unit from the source file, removes fields that would give away the target, and tests the models on a later year. I report macro F1 as well as accuracy because the classes are uneven. The submitted code includes a small sample and can reproduce every table and figure in the paper.

II. Previous work

Hotspot mapping uses past records to estimate where incidents are concentrated. Chainey, Tompson, and Uhlig compared several mapping methods and found that their predictive performance depended on the method and crime type [1]. Kernel density estimation performed well in their tests. Their results are a good reason to map the categories separately. An aggregate map may hide a pattern that is visible for one type of offence.

Some predictive policing studies address a different problem. Mohler *et al.* tested short term hotspot forecasts in randomized field trials [2]. This paper does not attempt that kind of forecast. It uses a retrospective classification experiment to see how much information is present in the contextual fields of a record. A correct label does not mean that the model could predict a future event or improve a police response.

The source also affects what can be concluded. Statistics Canada describes the Uniform Crime Reporting Survey as a standardized collection of crimes reported to police. It notes that reporting, police procedure, and enforcement practice can affect comparisons [3]. Events that are never reported are absent. One event may also contain several offences or victims. For that reason, this paper uses the word *records* instead of treating every row as one unique incident.

I compare logistic regression with a random forest. Logistic regression gives a simple linear reference. Random forests can model nonlinear relationships and interactions among place, time, and premises [4]. Both models are implemented with scikit-learn [5]. The comparison is meant to measure class separation under a chronological test, not to propose an operational policing tool.

III. Methodology

A. Data source and unit of analysis

The Toronto Police Service (TPS) publishes the Community Safety Indicators Open Data file through its Public Safety Data Portal [6]. The downloaded CSV had 474,819 rows and 31 columns, with report years from 2014 to early 2026. There were 414,033 unique event IDs. Another 60,786 rows appeared beyond one row per event ID, and 6,518 event IDs had records

TABLE I
FIELDS USED IN THE ANALYSIS

Role	Fields
Target	CSI category
Time	Report year; occurrence month, day of week, day of year, hour
Place	Division, 158-neighbourhood, premises type, location type
Coordinates	Approximate WGS84 longitude and latitude
Audit only	Event ID, occurrence year

in more than one CSI category. The analysis keeps the source unit and treats a row as an offence or victim record.

Table I lists the fields used by the pipeline. The target is `CSI_CATEGORY`. I left out the offence description and UCR code fields. Those fields define, or nearly define, the category and would leak the answer to the model.

B. Cleaning and scope

The script checks the required column names before it loads the file. It strips spaces from categorical values, converts numeric fields, and labels missing categories as `Missing`. I treated coordinates outside a broad Toronto box ($-80.1 \leq$ longitude ≤ -78.8 and $43.4 \leq$ latitude ≤ 44.0) as missing for modeling and removed them from the maps. This left 467,867 rows with plausible coordinates. The other 6,952 rows had invalid or suppressed coordinates.

The download included 10,248 records reported in 2026, but the year was not complete. Including them would create an artificial decline at the end of the annual chart. I used 464,571 records reported from 2014 through 2025. Occurrence time supplies the hourly and seasonal variables, while report year controls the train and test split. A record with a delayed report stays in the year when TPS recorded it.

C. Exploratory data analysis

The first figure combines four views. It shows class frequency and annual counts, then compares hour and premises type within each class. The last two panels use percentages within a category. Without that normalization, the much larger Assault class would dominate the plot. The annual panel uses record counts rather than population rates because the project data has no exposure denominator. I use counts here and do not describe them as rates.

A separate heatmap compares the weekday distribution within each category. Its columns follow the calendar from Monday through Sunday, and every cell is a percentage of the records in that category. A common scale from 0 to 18 percent keeps colour differences comparable across the five rows.

For the spatial view, I grouped valid coordinates into hexagonal cells and used a logarithmic colour scale. Each category has its own panel and scale. This makes both dense and sparse cells visible, but it also means that colours should not be compared as absolute counts across panels. TPS provides privacy protected coordinates. The plots only show broad concentrations. They do not locate exact addresses.

D. Predictive experiment

The model uses occurrence month, day of week, premises type, location type, division, neighbourhood, and approximate coordinates. I converted occurrence hour and day of year into sine and cosine pairs. The conversion keeps 23:00 close to 00:00 and late December close to early January. The pipeline fills missing coordinates with the median learned from its training data.

Both models use class weights. Logistic regression uses standardized numeric fields and one hot encoded categories. The random forest has 220 trees, a maximum depth of 24, a minimum leaf size of four, square root feature sampling, and balanced bootstrap weights. It uses categorical mappings learned from the training set and gives unseen values a separate code.

The training set contains 374,136 records reported from 2014 through 2023. I used 47,402 records from 2024 to choose the model with the better macro F1, then tested that choice on 43,033 records from 2025. A majority class prediction is included as a baseline. The reported measures are accuracy, balanced accuracy, macro F1, weighted F1, and the precision and recall for each category. All random operations use seed 5430.

IV. RESULTS

A. Category and time patterns

Assault accounts for 248,662 records, or 53.5% of the analysis data (Fig. 1(a)). Break and Enter has 83,110 records (17.9%), followed by Auto Theft with 76,843 (16.5%), Robbery with 39,640 (8.5%), and Theft Over with 16,316 (3.5%). A model that always predicts Assault starts with high accuracy, even though it never recognizes the other four categories.

The total increased from 32,536 records in 2014 to 49,672 in 2023, then fell to 47,402 in 2024 and 43,033 in 2025. The categories did not move together. Auto Theft rose from 3,703 records in 2014 to 12,592 in 2023, then dropped to 7,363 in 2025. Assault fell in 2020, rose to 25,806 in 2024, and remained close to that level in 2025. Robbery fell from 3,786 records in 2014 to 2,288 in 2021. It partly recovered before falling to 2,566 in 2025. These are changes in reported record counts, not estimates of crime prevalence.

The hourly curves in Fig. 1(c) are also different. Auto Theft peaks at 22:00, when 8.8% of its records occur. Robbery is highest around 20:00 and 21:00, at about 7.0% per hour. Break and Enter has more overnight activity. Assault is spread more evenly and rises in the evening. Theft Over has sharp peaks at midnight and noon. Exact occurrence time is not always known in administrative data, so those two spikes may partly reflect rounded or default times.

The weekday heatmap adds a pattern hidden by the hourly totals (Fig. 2). Assault is more common on Saturday and Sunday within its own records, at 15.6% and 15.4%. Auto Theft is lower on Saturday and Sunday, at 12.6% and 12.9%, and reaches 15.4% on Thursday. Friday contains 16.3% of Break and Enter records and 16.9% of Theft Over records. Sunday

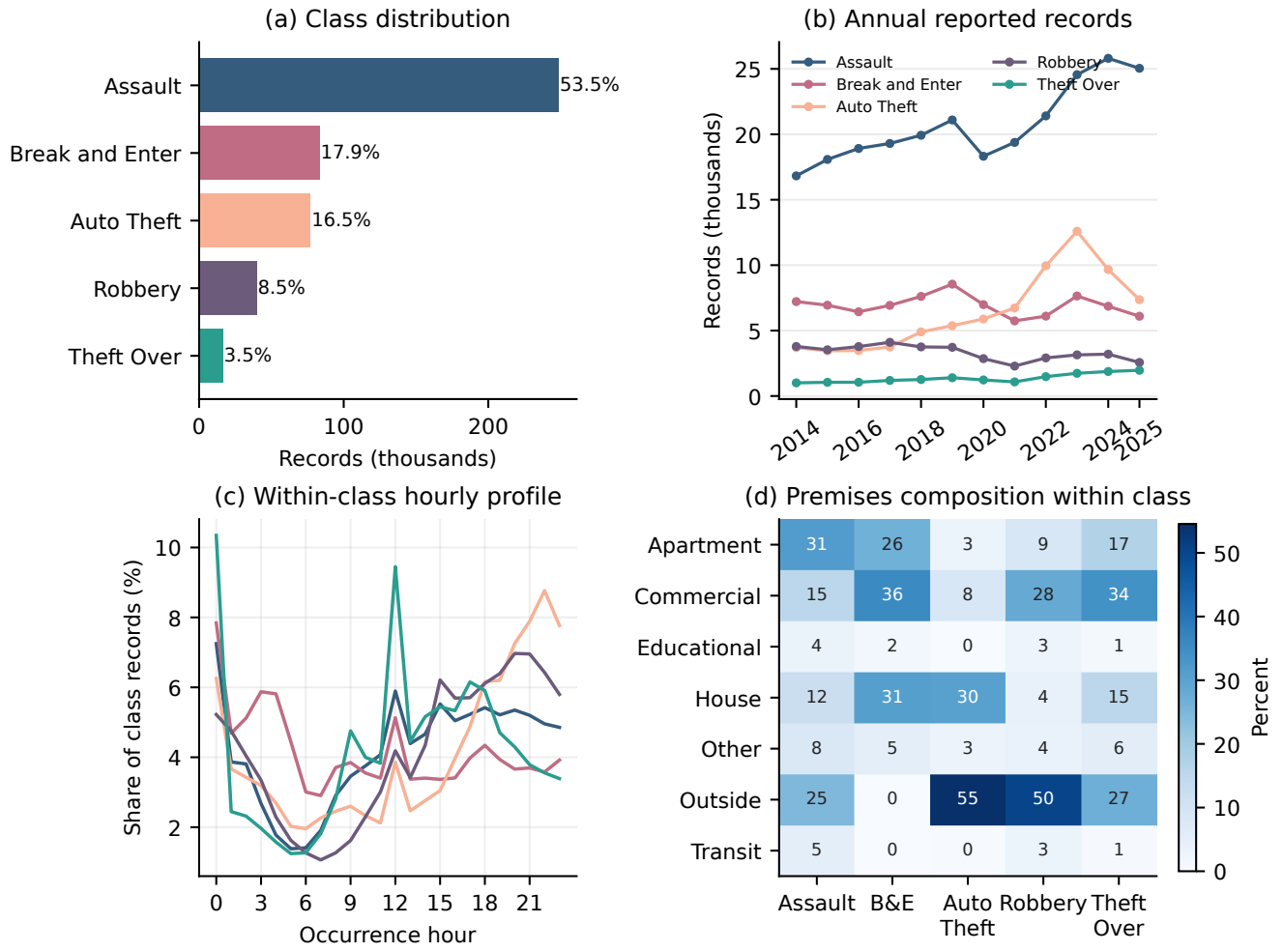

Fig. 1. Exploratory views for 464,571 offence and victim records reported from 2014 through 2025. Panels (c) and (d) use percentages within each category. Midnight and noon peaks may contain rounded or default times.

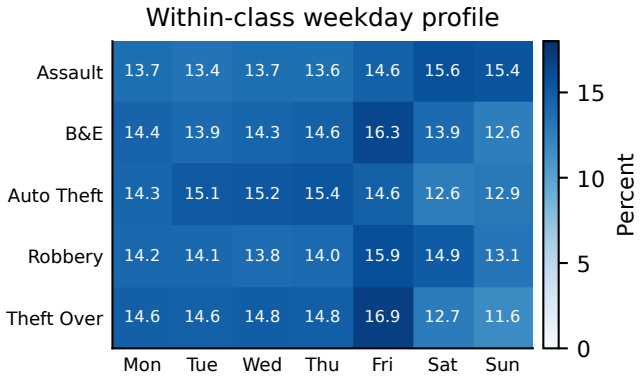

Fig. 2. Weekday composition within each CSI category. Cells are row percentages, and all categories use the same colour scale.

has the lowest Theft Over share at 11.6%. These percentages describe the weekly composition of each category, not rates per day or estimates of risk.

B. Premises and spatial patterns

Premises type separates several categories in Fig. 1(d). Outside locations make up 54.5% of Auto Theft records and 50.1% of Robbery records. For Break and Enter, houses and apartments together account for 57.3%. Commercial locations account for 35.5% of Break and Enter and 33.9% of Theft Over. Assault is less concentrated in one premises group. Apartments account for 31.5%, outside locations for 24.5%, and commercial locations for 15.3%.

Figure 3 shows a dense central area in the overall map, but the category panels are not identical. Assault, Break and Enter, and Robbery are concentrated near downtown. Auto Theft spreads farther west and northwest. Moss Park has the most Assault and Robbery records, the Annex has the most Break and Enter records, and West Humber-Clairville has the most Auto Theft and Theft Over records. These are raw neighbourhood counts. They do not account for population or activity, and several rows can belong to one event.

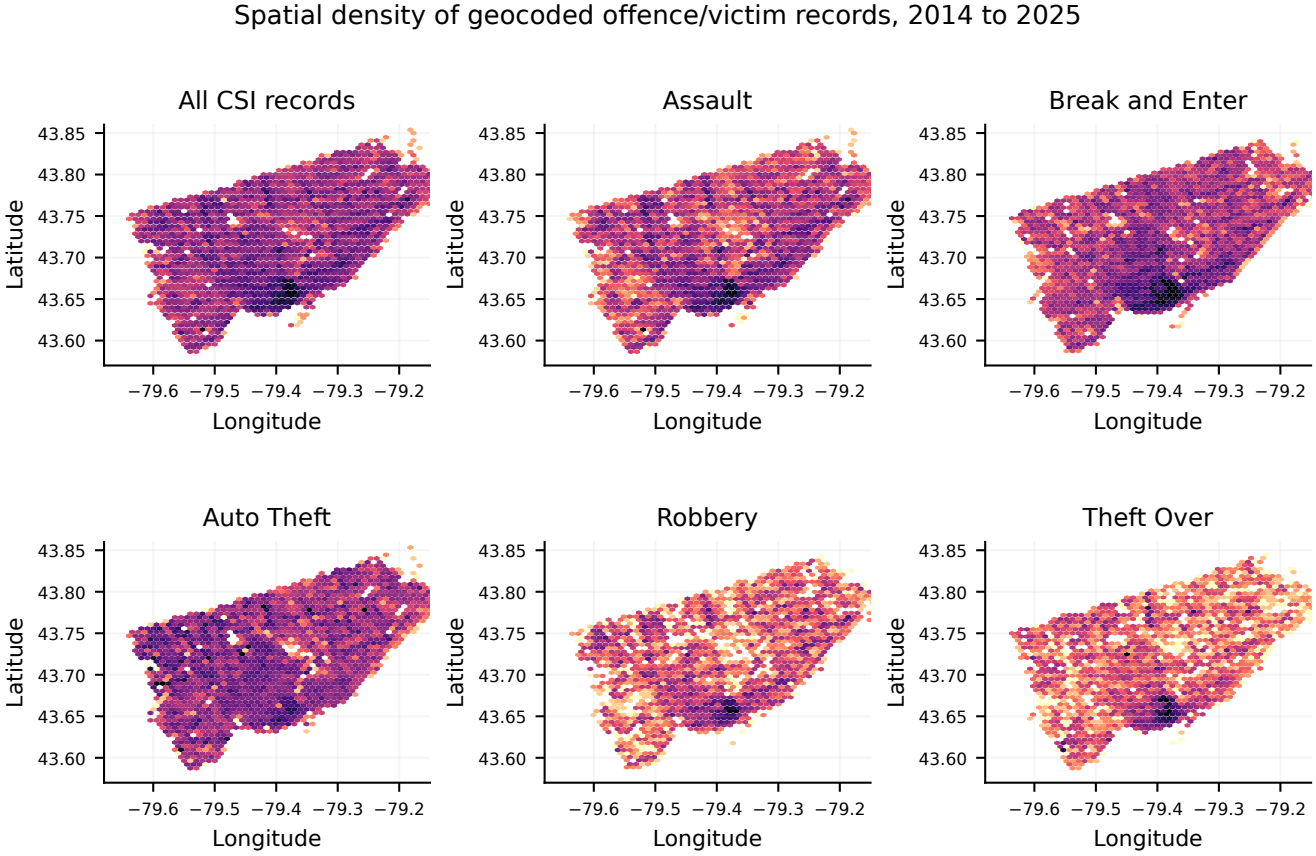

Fig. 3. Log scaled hexagonal density for rows with plausible approximate coordinates. Each panel has a separate density scale. The coordinates are privacy protected and are not exact addresses.

TABLE II
2025 CHRONOLOGICAL TEST PERFORMANCE

Model	Acc.	Bal. Acc.	Macro F1	Wtd. F1
Majority	.582	.200	.147	.428
Logistic	.444	.440	.368	.480
Random forest	**.592**	**.452**	**.424**	**.601**

C. Predictive performance

Table II reports the results for the 2025 test set. The majority baseline reaches 58.2% accuracy because Assault is so common. Its balanced accuracy is 0.200 and its macro F1 is 0.147. Class weighting changes the tradeoff for logistic regression. Its accuracy falls to 44.4%, but its macro F1 rises to 0.368. The random forest had the best 2024 validation macro F1. On the 2025 data it reaches 59.2% accuracy, 0.452 balanced accuracy, and 0.424 macro F1. Its weighted F1 is 0.601.

The random forest improves accuracy by only one percentage point over the majority baseline. Its more useful improvement is the macro F1 score, which shows that it recognizes some minority class records instead of predicting Assault every time. Even so, the result is modest.

The confusion matrix in Fig. 4(a) gives the clearest picture of where the model works. Recall is 0.64 for Assault, 0.60 for Break and Enter, and 0.68 for Auto Theft. It falls to 0.24 for Robbery and 0.10 for Theft Over. Theft Over is usually assigned to one of the three larger categories. Robbery is often confused with Assault. The class F1 scores are 0.709 for Assault, 0.493 for Break and Enter, 0.589 for Auto Theft, 0.206 for Robbery, and 0.124 for Theft Over. This is not reliable enough for operational category assignment.

Location type has the largest impurity importance in Fig. 4(b), followed by longitude and latitude. The day of year variables and premises type also contribute. That ranking matches the differences visible in the EDA. Because impurity importance tends to favour continuous fields and categories with many values, the bars do not show causal effects.

V. Limitations and Ethical Considerations

The data only includes events known to and recorded by police. Reporting habits, police procedure, and coding may change over time. The row count can also include several offences or victims under one event ID, so it should not be compared directly with a unique incident count.

The maps have no exposure denominator. A busy commercial area may have more records partly because more people and vehicles pass through it. TPS also protects location privacy,

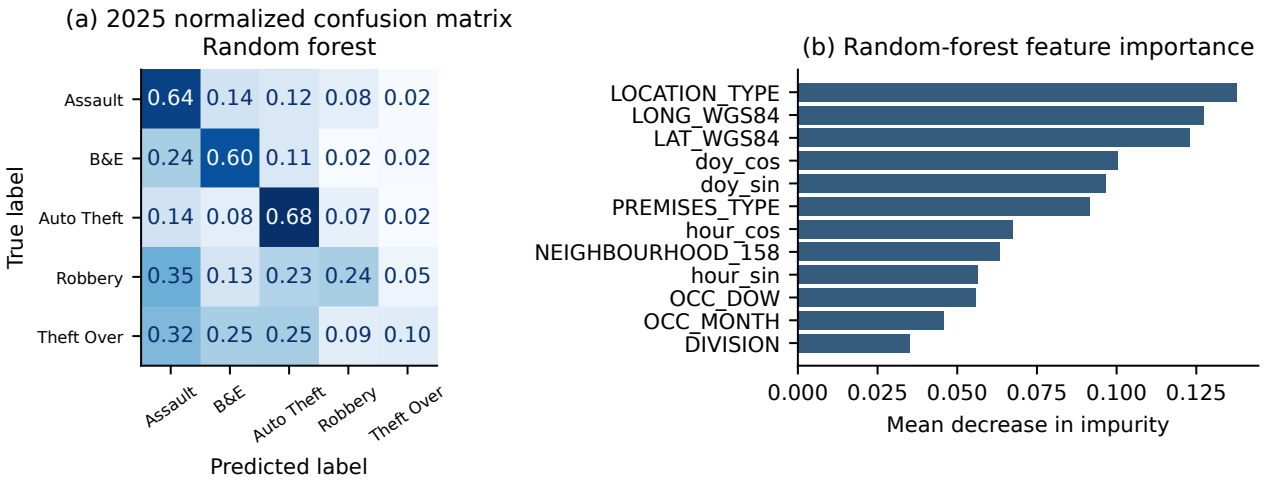

Fig. 4. Random forest results on the 2025 test set. The confusion matrix uses row percentages. Importance is the mean decrease in impurity, not a causal effect.

and some coordinates are missing. The maps avoid street labels and exact points for that reason. Rounded occurrence times may also explain part of the midnight and noon peaks.

The classifier sees contextual fields after a record exists. It is not a crime forecast. Its poor recall for the two rarest classes, along with the small accuracy gain over the baseline, rules out practical use in its current form. The outputs are suitable for aggregate course analysis. They should not be used for surveillance, neighbourhood ranking, or decisions about individuals. A follow up analysis should divide counts by a relevant population or activity measure and test whether the results remain stable from one year to the next. Without those checks, policy use would be premature.

VI. Conclusions

The five CSI categories do not follow one common pattern. Total reported records increased through 2023 and were lower in 2024 and 2025. The weekday and hourly profiles differ, and premises type gives an especially clear separation between Auto Theft, Break and Enter, and Robbery. The maps add another difference: Auto Theft is more dispersed than the categories concentrated near downtown. The random forest learned some of these patterns, but not well enough for dependable classification.

The model result needs a cautious reading. A 59.2% accuracy sounds reasonable until it is compared with the 58.2% majority baseline. Macro F1 gives a more complete account of the improvement. The confusion matrix makes the weakness hard to miss: recall is poor for Robbery and worse for Theft Over. The paper is best read as a description of police reported records. The code and sample data allow the analysis to be repeated, including the chronological test, without turning the model into a claim about future crime.

References

[1] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Security Journal*, vol. 21, no. 1-2, pp. 4–28, 2008.

[2] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham, "Randomized controlled field trials of predictive policing," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[3] Statistics Canada, "Police-reported crime statistics in canada, 2024," The Daily, Jul. 2025, [Online]. Available: https://www150.statcan.gc.ca/n1/daily-quotidien/250722/dq250722a-eng.htm.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[6] Toronto Police Service, "Community safety indicators open data," Public Safety Data Portal, 2026, accessed: 2026-07-21. [Online]. Available: https://data.tps.ca/datasets/TorontoPS::community-safety-indicators-open-data/about.