CRIME HOTSPOT PREDICTION IN OTTAWA USING MACHINE LEARNING

Algonquin College of Applied Arts and Technology

Murk Asad
Student
asad0022@algonquinlive.com

Shara Khandakar
Student
khan0611@algonquinlive.com

Anam Vakil
Student
vaki0005@algonquinlive.com

Maksuda E Elahi
Student
elah0010@algonquinlive.com

ABSTRACT

This project analyzes criminal offences reported in the City of Ottawa between 2016 and 2022 to identify spatial and temporal crime hotspots and evaluate the feasibility of predicting offence categories using machine learning techniques. Using a dataset of over 260,000 records from the Ottawa Police Service, the study applied extensive data cleaning, feature engineering, and exploratory data analysis to uncover trends across divisions, neighborhoods, census tracts, and time periods. Categorical inconsistencies were standardized, temporal attributes were decomposed, and outliers were removed to ensure analytical accuracy. Three machine learning models: Decision Tree, Random Forest, and Neural Network were trained to classify crime types based on spatial, temporal, and contextual features. Among all models, Decision Trees achieved the highest accuracy (above 99.9% across multiple train–test splits), demonstrating superior performance for this dataset. The results confirm that Ottawa's crime patterns exhibit identifiable spatial concentrations and predictable temporal structures, and that machine learning approaches can support crime prediction and hotspot identification to inform policing, planning, and public safety decision-making.

1. INTRODUCTION

This project investigates criminal offences within the City of Ottawa using historical crime data [1]. The primary objective is to identify spatial and temporal crime hotspots and explore whether machine learning models can help predict crime patterns based on time, location, and contextual attributes. By combining exploratory data analysis (EDA) with classification algorithms, the project aims to generate insights that could support law enforcement resource allocation, community safety initiatives, and evidence-based policy decisions.

2. BUSINESS UNDERSTANDING

Understanding where and when crimes are more likely to occur has strong practical significance. For police, hotspot analysis can guide patrol planning and rapid response strategies. For city planners, crime distribution can inform decisions about lighting, surveillance, and public infrastructure. For communities and policymakers, the results provide a data-driven foundation for prevention programs and public safety interventions.

3. DATA SOURCING AND PREPARATION

The dataset used in this project is the 'Criminal Offences (2016–2022) dataset published by the Ottawa Police Service through the City of Ottawa Open Data portal [1]. The data contains more than 260,000 records and approximately 15 fields, including spatial coordinates (X, Y), offence descriptions, premise types, divisions, sectors, and temporal information such as occurrence date, time, and weekday. The dataset was downloaded as a CSV file and loaded into Python using the Pandas library.

Before any modeling, substantial data preparation was required to ensure consistency and analytic quality. The main preparation steps included cleaning categorical labels, correcting data types, engineering time-based features, dropping redundant columns, and addressing outliers.

The first step of the analysis involved importing the Criminal_Offences_.csv file into a pandas Data Frame. This allowed the dataset to be manipulated, cleaned, and analyzed using Python. Loading the data makes it possible to examine the structure and characteristics of the crime records.

4. DATA EXPLORATION

All analysis was conducted in Python using the pandas, NumPy, Matplotlib, Seaborn, and scikit-learn libraries. EDA was used to understand distributions and patterns, while machine learning was used to model and predict crime-related outcomes. Before conducting advanced analysis, an initial high-level exploration of the dataset was performed to understand its overall structure and content.

4.1. Data Size

The output of df.shape confirmed that the dataset contains 260,910 rows and 15 columns, indicating a sufficiently large and robust dataset suitable for machine learning applications and hotspot analysis.

4.2. Summary Statistics - Numeric vs Categorical Features

The function df.describe(), provided descriptive statistics for numeric fields, such as X, Y, and Year. These statistics included the mean, minimum, maximum, and quartile values. Notably, the coordinate fields **X** and **Y** fall within tight ranges, which is expected given that all recorded incidents occurred within the Ottawa region.

4.2.1. Numeric Features

For numeric variables, describe () was utilized to quickly assess central tendencies, variability, and potential outliers. This aided in validating that the data appeared consistent with expected geographic and temporal values.

Overall, this preliminary exploration offered essential context and confirmed that the dataset was structurally sound and ready for further preprocessing and detailed analysis, as shown in Table 1.

TABLE 1. Statistics of Numeric Features

Statistic	X	Y	FID	Year
count	260,910	260,910	260,910	260,910
mean	-8,426,629	5,682,776	130,455.5	2019.11
std	10,334.49	9,044.15	75,318.37	1.995
min	-8,497,819	5,618,838	1	2016
25%	-8,430,626	5,677,851	65,228.25	2017
50%	-8,426,004	5,686,083	130,455.5	2019
75%	-8,422,116	5,688,956	195,682.75	2021
max	-8,378,081	5,704,508	260,910	2022

4.2.2. Categorical Features

A summary of all categorical (object-type) columns was generated using df.describe(include="object"). This step provided an overview of the dataset's non-numeric fields.

- OffSummary – high-level offence category
- Neighbourr – neighbourrhood name
- Sector – policing sector
- Division – division name
- Weekday – day of the week
- Date/time fields (e.g., *ReportDate*, *OccurDate*) that were still stored as text rather than in datetime format

This analysis revealed several inconsistencies within the categorical variables. Notable findings include:

- Inconsistent labels within the Division field, where entries appear in mixed formats such as *"Central"*, *"West"*, *"East"* alongside *"CENTRAL DIVISION"*, *"EAST DIVISION"*, and *"WEST DIVISION"*.
- Variation in Sector labels, with values recorded both as *"11"* and *"Sector 11"*, indicating non-standardized formatting.

These inconsistencies directly informed the data cleaning strategy, emphasizing the need to standardize categorical labels—particularly for the Sector and Division fields—to ensure accurate grouping, analysis, and visualization. A sample of these categorical columns is presented in Table 2 to illustrate the observed formatting issues.

TABLE 2. Statistics of Few Categoric Features

Statistic	Report Date	Off Summary	Prim Violat	Neighbor
count	260,910	260,910	260,910	260,910
unique	2,557	6	134	114
top	2022/02/18 00:00:00 +00	Crimes Against Property (2000)	Theft $5000 and Under	Stittsville
freq	206	173,885	73,1130	218,742

4.3. Data Types and Missing Values

A review of the dataset using df.info() displayed in Table 3 confirmed that no missing values were present in any column. This finding substantially simplifies the preprocessing phase, as no imputation strategies are required for the data cleaning process.

During this assessment, it was also observed that several columns had data types that did not align with their conceptual meaning:

- Columns that are inherently categorical, such as CensusTra, were stored as numeric values.
- Several date and time fields, including *ReportDate*, *ReportTime*, *OccurDate*, and *Occur_Time*, were stored as object (string) types.

These fields will be converted to their appropriate data types (categorical or datetime) in subsequent preprocessing steps to ensure accurate analysis and modeling

TABLE 3. Information of All Features

#	Column	Non-Null Count	Dtype
1	X	260,910 non-null	float64
2	Y	260,910 non-null	float64
3	FID	260,910 non-null	int64
4	Year	260,910 non-null	int64
5	ReportDate	260,910 non-null	object
6	ReportTime	260,910 non-null	object
7	OccurDate	260,910 non-null	object
8	Occur_Time	260,910 non-null	object
9	Weekday	260,910 non-null	object
10	OffSummary	260,910 non-null	object
11	PrimViolat	260,910 non-null	object
12	Neighbourh	260,910 non-null	object
13	Sector	260,910 non-null	object
14	Division	260,910 non-null	object
15	CensusTra	260,910 non-null	object

These preliminary checks provided a strong understanding of the dataset’s size, completeness, and quality, and helped identify the areas that required data cleaning later.

4.4. Imbalanced Target Variable

To assess the distribution of the target variable named *OffSummary* that represents crime categories within the dataset, a bar chart in Figure 1 visualizes it. This visualization highlights the frequency of each offence type and reveals a significant imbalance across categories. The most common offence summaries are summarized in Table 4.

TABLE 4. Total Count Per Crime Category

Offence Category	Count
Crimes Against Property (2000)	173,885
Crimes Against Person (1000)	47,383
Other Criminal Code Offences (3000)	38,457
Traffic Offences	719
Controlled Drugs and Substances Act (4000)	443
Cannabis Act (4900)	23

Understanding this distribution is critical for identifying the dominant crime types in Ottawa and for ensuring appropriate analytical and modeling considerations. Figure 1 presents the crime summary count visually. The distribution clearly demonstrates the following:

- Crimes Against Property (2000) constitute the majority of all recorded offences.
- Crimes Against Person (1000) and Other Criminal Code Offences (3000) appear frequently, though significantly less than property-related offences.
- Categories such as Traffic Offences, Controlled Drugs and Substances Act (4000), and Cannabis Act (4900) represent a very small proportion of the records.

Overall, the dataset exhibits substantial class imbalance, with property-related offences and broader criminal code categories dominating the distribution. This imbalance should be taken into account during further analysis, particularly in any predictive modeling efforts.

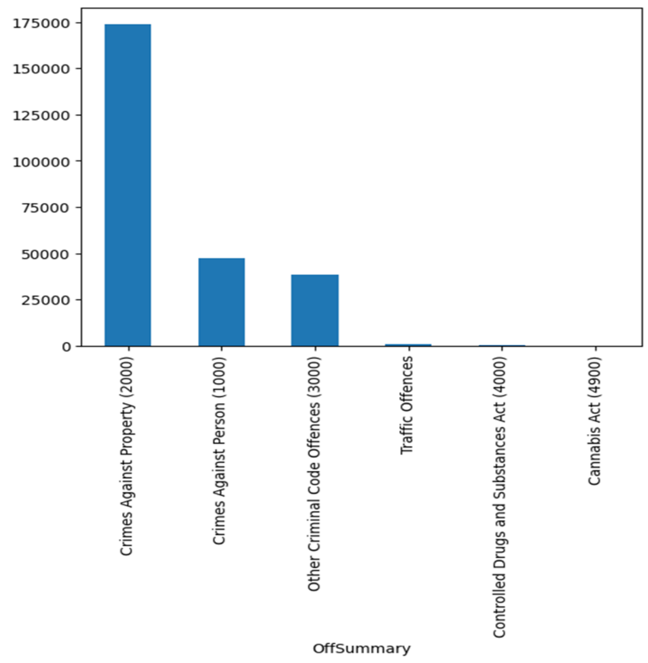

FIGURE 1. Count Per Crime Category

4.5. Crime by Division

To examine how different offence categories are distributed across divisions, a cross-tabulation of division versus offence type was created, and the results were visualized using a clustered bar chart. This approach allows for a clear comparison of crime patterns across Central, East, and West divisions. Before standardizing the labels, the division column contained values as shown in TABLE 5.

TABLE 5. Total Count Per Division

Division	Incidents
Central	87,084
West	47,229
East	46,204
CENTRAL DIVISION	37,841
EAST DIVISION	21,944
WEST DIVISION	20,608

After grouping similar labels (e.g., "Central" + "CENTRAL DIVISION"), the approximate total counts by division are displayed in Table 6. Note that this step is yet to be performed in feature engineering. Table 6 is displayed for reference.

TABLE 6. Division Count After Grouping

Division	Total Incidents
Central	124,925
East	68,148
West	67,837

The analysis in Figure 2 shows that Central clearly has the highest volume of reported crime, while East and West divisions report similar but noticeably lower counts. Across all divisions, Crimes Against Property (2000) is the dominant offence category, followed by Crimes Against Person (1000). These findings highlight Central Ottawa as a major hotspot region, whereas East and West experience moderate but still substantial crime levels. The clustered bar chart further reinforces these spatial patterns and provides a clear visual summary of division-level crime distribution.

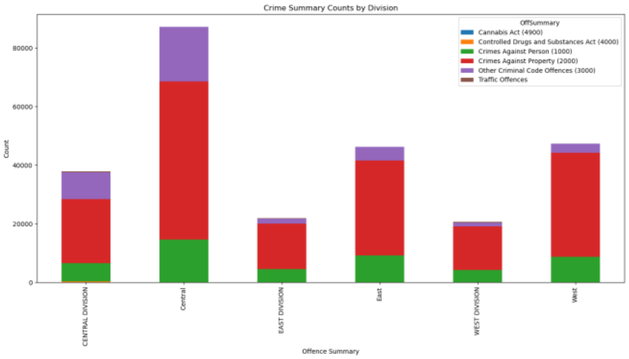

FIGURE 2. Crime Summary Count by Division

4.6 Crime by Census Tract

To analyze crime at a finer geographic level, incidents were grouped by census tract, and the number of incidents per tract was calculated. The counts were then sorted to identify the top tracts, and a plot was generated to visualize these micro-hotspots. Some census tract IDs appear in both float and integer formats, highlighting the need to treat this field as categorical or string rather than numeric. The top census tracts by incident count are shown in Table 7.

TABLE 7. Census Tract Count

Census Tract	Crime Count
5050054	8,915
5050048.0	8,569
5050048	7,886
5050037.0	6,488
5050054.0	4,906

The analysis in Figure 3 indicates that a small number of census tracts account for a disproportionately high number of incidents, identifying clear micro-hotspots. These tracts likely correspond to high-activity urban zones such as commercial centers or densely populated residential areas. Recognizing duplicated-looking IDs, like 5050048 and 5050048.0, underscores the importance of cleaning and consistently treating the CensusTra field for accurate hotspot mapping and granular spatial analysis.

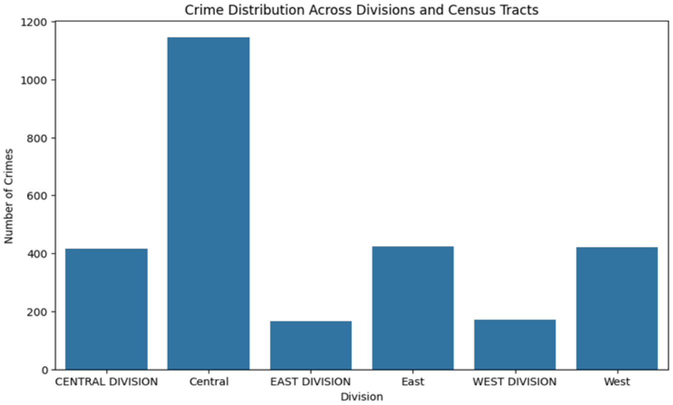

FIGURE 3: Crime Across Divisions

4.7. Annual Crime Trends

The temporal distribution of reported crimes was examined by grouping incidents by year, and in some cases by both year and division, with line charts plotted to visualize trends. The total number of reported incidents per year in the dataset is summarized in TABLE 8.

TABLE 8: Number of Years of Data in Dataset

Year	Crime Count
2016	33,027
2017	34,434
2018	38,912
2019	42,623
2020	31,521
2021	37,423
2022	42,970

The analysis displayed in Figure 4 shows a clear upward trend in reported crime from 2016 to 2019, followed by a noticeable dip in 2020, which coincides with the COVID-19 pandemic and associated lockdowns that include limited movement and public activity. Crime then rebounded in 2021 and reached a new peak in 2022. Further examination by year and division indicates that this overall pattern is consistent across divisions, with the Central division consistently reporting the highest volumes each year. These findings suggest that while crime trends are influenced by broader city-level conditions, spatial patterns remain relatively stable over time.

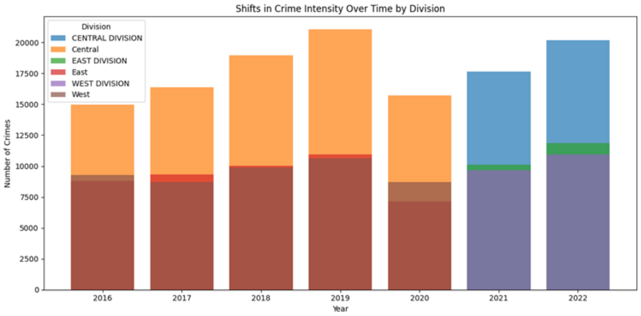

FIGURE 4: Crime Over time and By Division

4.8. Neighborhood-Level Crime Patterns

The distribution of reported incidents across neighborhoods was analyzed to identify areas with the highest concentration of crimes. The top neighborhoods by incident count are summarized in Table 9, and a corresponding plot was generated to visually highlight these trends.

TABLE 9: Crime Per Neighborhood (Top 6)

Neighbourhood	Crimes
Stittsville	218,742
Centretown	6,434
Byward Market	2,081
Lowertown	1,865
Overbrook - McArthur	1,592
New Barrhaven - New Development - Stonebridge	1,525

Similarly, in Figure 5, it can be seen that Stittsville dominates as the most frequent neighbourhood label by a significant margin. This may be influenced by the way the dataset is geomasked or aggregated, but it nonetheless indicates a strong concentration of reported incidents under this label. Other central and inner-city neighbourhoods, such as Centretown, Byward Market, and Lowertown, also report high counts, reflecting their status as busy urban areas. These findings help to highlight neighbourhood-level hotspots, providing valuable insights for targeted community policing and local interventions.

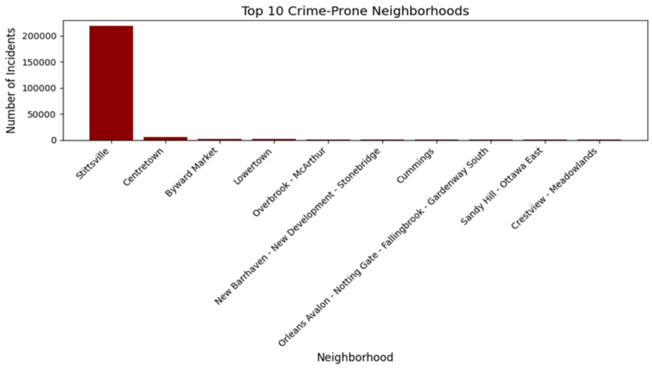

FIGURE 5: Crime Per Neighborhood (Top 10)

4.9. Temporal Patterns: Weekday Analysis

The distribution of reported crimes across weekdays was examined to understand temporal patterns and identify which days experience higher or lower crime volumes. The analysis revealed that midweek days have the highest number of reported incidents, while weekends record comparatively fewer crimes. Wednesday, Tuesday, and Thursday emerged as the days with the most criminal activity, whereas Sunday had the lowest count, followed by Saturday. This pattern is contrary to the common assumption that crime peaks on weekends, suggesting that in Ottawa, criminal activity is more influenced by routine weekday activities such as work, commuting, and school rather than weekend nightlife. The total counts for each weekday are summarized in Table10.

TABLE 10: Total Crimes on Weekdays

Weekday	Crime Count
Wednesday	41,212
Tuesday	40,898
Thursday	40,444
Friday	39,844
Monday	39,469
Saturday	31,154
Sunday	27,889

A bar chart in Figure 6 was also plotted to visually highlight the differences in crime frequency across the days of the week, making it easier to identify trends briefly.

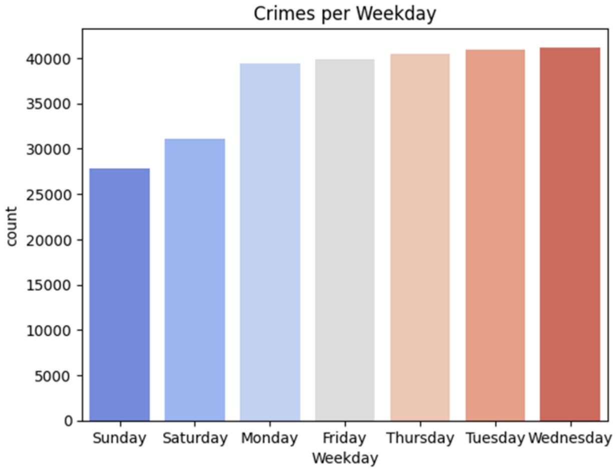

FIGURE 6: Total Crimes on Weekdays

4.10. Verifying Spatial Accuracy of Crime Records

To ensure that all incidents in the dataset were geographically valid, the coordinates were verified to fall within the Ottawa region. The X (longitude-like) values ranged from approximately -8,497,819 to -8,378,081, while the Y (latitude-like) values ranged from about 5,618,838 to 5,704,508. Minimum and maximum thresholds were defined to represent Ottawa's spatial extent, and the points were plotted on a scatter plot, with points inside the bounding box marked in one color and those outside in another. The resulting visualization in Figure 7 shows that all points fell within the expected Ottawa region, with no significant clusters outside the defined boundaries. This confirmation of spatial validity ensures that the dataset is suitable for hotspot mapping and other location-based analyses.

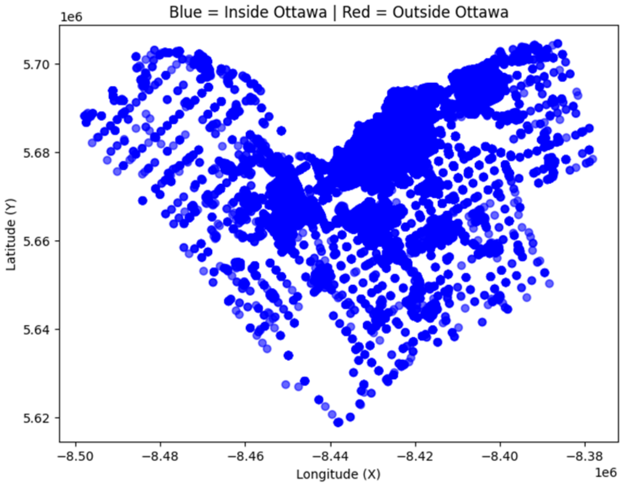

FIGURE 7: Verifying Ottawa's Data

5. DATA CLEANING

The dataset was thoroughly examined to ensure completeness, accuracy, and readiness for analysis. An initial check for missing values showed that all columns contained zero missing entries, eliminating the need for imputation or row removal. Duplicate records were also assessed, and the result confirmed that no duplicates were present. While the 'Sector' column initially appeared free of outliers, further examination following feature engineering such as standardizing labels and converting data types revealed anomalies, including an invalid entry ("408") and records labeled as "False calls." These outliers were removed to ensure that the dataset accurately reflected real and geographically valid crime records.

6. FEATURE ENGINEERING

6.1. Date and Time Feature Extraction

To enable detailed temporal analysis, the original date and time columns were decomposed into multiple features. For example, 'ReportDate' was split into 'ReportDate_Year', 'ReportDate_Month', and 'ReportDate_Day', while 'ReportTime' was split into 'ReportTime_Hour', 'ReportTime_Minute', and 'ReportTime_Second'. Equivalent splits were performed for 'OccurDate' and 'OccurTime', producing features such as 'OccurDate_Year', 'OccurDate_Month', 'OccurDate_Day', 'OccurTime_Hour', and 'OccurTime_Minute'. These derived features were later used for exploratory data analysis (EDA) and as inputs to machine learning models. Splitting these columns helps the algorithm identify which part of the date-time information contributes most to predicting crimes. Once the new columns were created, the original date and time columns were dropped to reduce redundancy while preserving all relevant temporal information as shown in TABLE 11.

TABLE 11: Date-time Feature Transformation

Original Column	Extracted Components	Description
ReportDate	ReportDate_Year, ReportDate_Month, ReportDate_Day	Date converted to datetime, then split into year, month, day
ReportTime	ReportTime_Hour, ReportTime_Minute, ReportTime_Second	Time converted to datetime (H:M:S), then split into hour, minute, second
OccurDate	OccurDate_Year, OccurDate_Month, OccurDate_Day	Date converted to datetime, then split into year, month, day
OccurTime	OccurTime_Hour, OccurTime_Minute, OccurTime_Second	Time converted to datetime (H:M:S), then split into hour, minute, second

These newly created columns also allow for analysis such as identifying which times of day crimes are more likely to occur as shown in Figure 8.

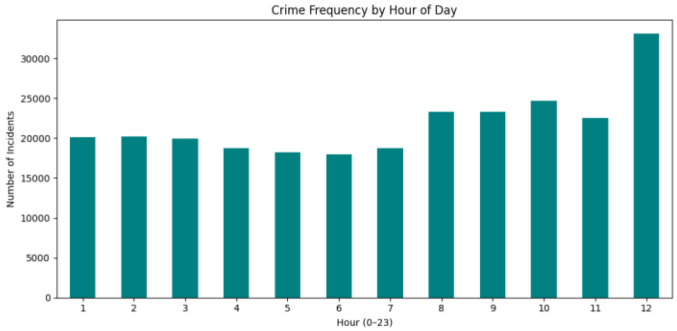

FIGURE 8: Crimes Per Hour Time

6.2. Grouping and Consolidation of Categories

The 'Division' column contained variations such as 'East' and 'East Division'. These were merged into unified labels (e.g., 'East') to avoid artificial fragmentation in the analysis. Before transformation, counts per division were as follows: Central: 87,084, West: 47,229, East: 46,204, CENTRAL DIVISION: 37,841, EAST DIVISION: 21,944, WEST DIVISION: 20,608. After transformation, the counts were consolidated into Central: 124,925, East: 68,148, and West: 67,837, as shown in Table 12.

TABLE 12: Division Feature Transformation

Before transformation	After transformation
Central	CENTRAL
CENTRAL DIVISION	
East	EAST
EAST DIVISION	
West	WEST
WEST DIVISION	

6.3. Categorical Encoding for Machine Learning

Machine learning algorithms require numerical input, so categorical columns such as Weekday, OffSummary, PrimViolat, Neighbourh, Sector, Division, and CensusTra were encoded using Label Encoder, which assigns a unique numeric value to each category. Although this approach can introduce minor bias due to higher numeric labels, it was preferred over one-hot encoding because many of these columns contain three or more categories, which would otherwise generate 15–20 additional columns. The original values of the target variable were also preserved before encoding to maintain a reference for interpreting the numeric labels. The 'Sector' column, which contained logically identical values written differently (e.g., '11' versus 'Sector 11'), was cleaned and standardized using the to_numeric function to ensure consistent numeric representation.

6.4. Outlier Detection and Removal

The 'Sector' column contains logically identical values recorded in different formats, such as '11' versus 'Sector 11'. These were cleaned and standardized into a single consistent format for each sector to ensure that grouping and aggregations would treat them as the same category. The to_numeric function was then used to convert the text-based values into numbers.

During the cleaning process, two types of outliers were identified in the 'Sector' column. First, 12 records were labeled as 'False calls', representing incidents where no actual crime occurred. These records were removed as they do not correspond to real offences. Second, one record contained an invalid sector code ('408') that does not match any known Ottawa sector. This record was also removed as an outlier. Removing these outliers ensures that subsequent hotspot analyses reflect real and geographically valid crimes. A violin plot can be used to visualize these outliers in the sector feature, as shown in Figure 9.

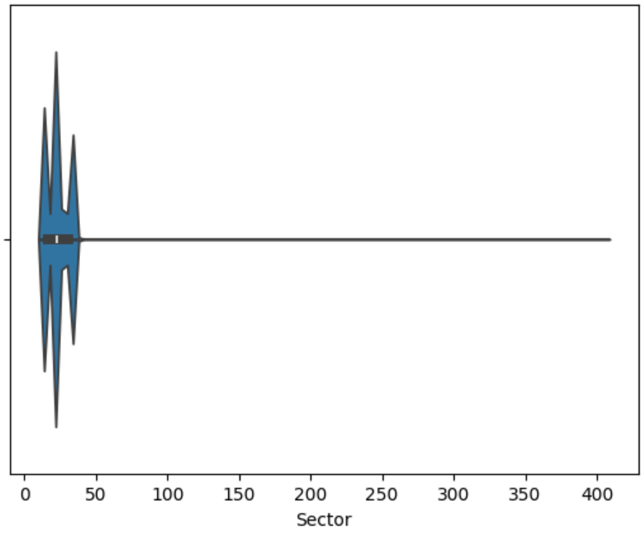

FIGURE 9: Outlier Detected (408)

6.5. Drop Irrelevant features

Columns like FID and CensusTra, which represent record identifiers and do not provide predictive value, were dropped to streamline the dataset.

6.6. Data after Transformation

After all feature engineering and cleaning steps, the dataset consisted of 21 columns and approximately 260,897 records. Key columns included the spatial coordinates (X, Y), temporal features (split year, month, day, hour, minute, second), encoded categorical variables, and the target variable for prediction. The cleaned dataset was then ready for machine learning model training and analysis, displayed in Table 13.

TABLE 13: Features After Transformation

#	Column	Non-Null Count	Dtype
1	X	260,897 non-null	float64
2	Y	260,897 non-null	float64
3	FID	260,897 non-null	int64
4	Year	260,897 non-null	int64
5	Weekday	260,897 non-null	int64
6	OffSummary	260,897 non-null	int64
7	PrimViolat	260,897 non-null	int64
8	Neighbourh	260,897 non-null	int64
9	Sector	260,897 non-null	Int64
10	Division	260,897 non-null	int64
11	CensusTra	260,897 non-null	object

12	ReportDate_year	260,897 non-null	int32
13	ReportDate_month	260,897 non-null	int32
14	ReportDate_day	260,897 non-null	int32
15	ReportTime_hour	260,897 non_null	int32

7. MODEL TRAINING AND EVALUATION

This section describes the development, training, and evaluation of various machine learning models applied to predict crime occurrences. The process includes selecting appropriate data splits, implementing different modeling techniques, tuning hyperparameters, and assessing performance through accuracy scores. Key features that contribute most to model predictions are also identified to provide insights into the factors influencing crime patterns. The following analysis summarizes model performance and highlights the most effective approaches for predictive accuracy.

7.1. Normalization

For data preprocessing, MinMaxScaler was applied to normalize the features. This was particularly important because variables such as latitude, longitude, and weekdays existed on significantly different scales, which could adversely affect model performance.

7.2. Data Split

The dataset was divided using multiple partition ratios 80:20, 70:30, and 85:15 to evaluate model performance under different train–test distributions. The 70:30 split was later chosen for Neural Network training because it corresponded to the best-performing model (Decision Trees) based on the accuracy comparison.

7.3. Machine Learning Model Training

Three primary models were developed and evaluated. For the Decision Tree classifier, both default and tuned versions were used, with hyperparameters [2] set to max_features=10 and max_depth=3. The Random Forest model was trained in its default form as well as with hyperparameter tuning, using n_estimators=100, max_depth=4, and max_features='sqrt'. The Neural Network was implemented using both default and tuned parameter settings, with hidden_layer_sizes=(3), batch_size=50, max_iter=200, and random_state=123. ReLU activation was applied in the hidden layer by default, and softmax was automatically used in the output layer for multiclass classification [3]. The Neural Network was specifically trained on the 70:30 split, as this partition corresponded to the best-performing model, Decision Trees, based on the accuracy table.

7.4. Feature Importance – Obtaining Best Features

The top five features contributing most to predicting crimes are identified in Figure 10 as PrimViolat, X and Y (latitude and longitude), Sector, Division, and the Time the crime occurred. These features were found to have the greatest impact on model performance and were critical for understanding the factors influencing crime predictions.

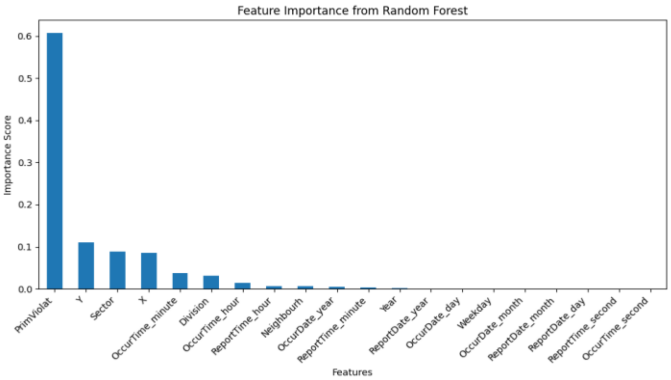

FIGURE 10: Importance Per Feature

8. MODEL COMPARISON AND RESULTS

The accuracy table summarized the performance of all models across the different splits. Decision Tree models consistently achieved the highest scores, exceeding 0.999 across all splits, making them the top-performing models. Random Forest models showed strong but slightly lower performance, with default scores around 0.993 and reduced accuracy for tuned and cross-validated versions. Neural Network models, trained on the 70:30 split, achieved accuracy of approximately 0.990 for the default model and 0.906 for the tuned model. These results highlighted Decision Trees as the most effective model, which guided the choice of the 70:30 split for Neural Network training and further analysis, shown in Table 14.

TABLE 14: Model Accuracies

#	Data Partition	Model Type	Accuracy Score
1	70-30	Decision Tree	0.999949
2	80-20	Decision Tree	0.999943
3	85-15	Decision Tree	0.999923
4	70-30	Neural Network (Default Model)	0.990942
5	70-30	Random Forest	0.993625

6	85-15	Random Forest	0.993484
7	80-20	Random Forest	0.993369
8	80-20	Random Forest Tuned	0.826524
9	70-30	Random Forest Tuned	0.826000
10	85-15	Random Forest Tuned	0.825961
11	85-15	Random Forest CV	0.825358
12	80-20	Random Forest CV	0.824102
13	70-30	Random Forest CV	0.820289
14	70-30	Decision Tree Tuned	0.787901
15	80-20	Decision Tree Tuned	0.782484
16	85-15	Decision Tree Tuned	0.759806
17	70-30	Neural Network (Tuned Model)	0.905711

9. CONCLUSION

This project demonstrated that historical crime data can effectively reveal meaningful patterns in both space and time, enabling reliable prediction of crime categories within Ottawa. The exploratory analysis identified Central Ottawa as the most prominent hotspot, with specific census tracts and neighborhoods showing consistently elevated incident counts. Temporal patterns further revealed that crime peaks during weekdays and varies significantly by hour of the day. After rigorous preprocessing and feature engineering, machine learning models were applied to the cleaned dataset. Decision Trees emerged as the most accurate and stable model, achieving near-perfect performance across all evaluation splits, followed by Random Forest and Neural Networks with slightly lower accuracy. These findings indicate that crime behavior in Ottawa follows predictable structures that can be captured through supervised learning methods.

From an applied perspective, the insights from this study can support law enforcement agencies in optimizing patrol allocation, improving situational awareness, and targeting interventions in high-risk locations and time periods. City planners and policymakers may also use these findings to guide infrastructure enhancements, community safety programs, and resource distribution. Future work could integrate additional contextual data such as weather, socio-economic indicators, or event schedules, apply advanced geospatial clustering techniques, and evaluate more sophisticated machine learning algorithms to further enhance predictive capability and hotspot detection.

10. REFERENCES

[1] City of Ottawa, Open Data Criminal Offenses Dataset, https://open.ottawa.ca/datasets/ottawa::criminal-offences-/about (accessed Oct 20, 2025).

[2] Tejashree Ganesan, How to Tune a Decision Tree, https://www.geeksforgeeks.org/machine-learning/how-to-tune-a-decision-tree-in-hyperparameter-tuning/#google_vignette (accessed Nov 1, 2025).

[3] “What Activation Function to use in Multiclass Classification Neural Networks”, https://www.reddit.com/r/learnmachinelearning/comments/1f0upgy/what_activation_function_to_use_in_multiclass/ (accessed Nov 8, 2025).

11. ABOUT THE AUTHORS

Murk Asad received M.Sc. degree in Information Technology from National University of Sciences and Technology (NUST), Islamabad, Pakistan, and a Postgraduate Certificate in Artificial Intelligence from George Brown College, Toronto. She has professional experience as a Business Analyst in the supply chain domain and is currently pursuing a Graduate Certificate in Business Intelligence Systems Infrastructure at Algonquin College, with academic and professional interests in data analytics, machine learning, and advanced database systems.

Shara Khandakar holds a bachelor's degree in Economics with a major in Finance and is currently pursuing a Graduate Certificate in Business Intelligence Systems Infrastructure at Algonquin College, Ottawa. Her interests include machine learning, data visualization, database systems, and business analytics. She aims to work in the field of data engineering and applied machine learning.

Anam Vakil holds a Master of Science in Information Technology and is currently pursuing Business Intelligence System Infrastructure at Algonquin College in Ottawa. With over three years of work experience as a Data Analyst, her career focus has been to develop skills within the IT field, which include business analytics, data visualization, database systems and machine learning.

She aims to work in the field of data engineering and applied machine learning.

Maksuda Elahi is currently studying Business Intelligence Systems Infrastructure at Algonquin College, developing strong skills in SQL, Python, and data visualization to interpret and transform complex data. Drawing on her experience in software QA, she brings precision and analytical depth to every project. She aims to create meaningful insights that support impactful decision-making in BI and analytics.