[1] **Hiranya Jayawardana**

[2] **Madushi Hasara Pathmaperuma**

Suraksha: Spatio-Temporal Crime Forecasting and Micro-Location Analysis

Abstract: - Suraksha, a spatiotemporal crime prediction system, designed to elevate crime prevention with precise insights, empowering law enforcement for a safer tomorrow. Utilizing vast datasets, machine learning, and GIS, it forecasts crime hotspots by incorporating Chicago's extensive crime statistics. Addressing both precision and ethical considerations, Suraksha achieves RMSE values of 0.0874 (latitude) and 0.0602 (longitude), marking a leap in predictive policing. This pioneering approach aims to transform public safety by proactively combating crime, ensuring community well-being through innovative data-driven strategies.

Keywords: Random Forest, Chicago Crime Dataset, Crime Prediction, Location Micro-Analysis.

I. INTRODUCTION

As crime escalates in complexity, threatening human wellbeing and hindering societal progress, the need for innovative crime prediction methods becomes crucial. Approximately six out of seven people on the planet, or 83%, now reside in high-crime countries [1]. Imagine living in a city where law enforcement serves a proactive role but has the insight to see illegal activity coming before it happens. For law enforcement, it's like having a crystal ball that keeps them one step ahead. This project aims to tackle the challenge by developing a spatiotemporal crime hotspot detection and prediction system, leveraging Chicago's extensive, open-source crime dataset. By employing advanced data analytics and machine learning techniques, the system will identify and predict areas at high risk of criminal activity. For individuals, it serves as a critical guide in selecting safer neighborhoods, enhancing personal safety and community well-being. The initiative focuses on collecting and analyzing historical crime data, incorporating geographic and temporal variables to unveil patterns and predict future hotspots. Utilizing Geographic Information Systems (GIS) for spatial analysis and machine learning algorithms for temporal trend examination. The main goal is to deliver rapid and accurate findings so that law enforcement officers may anticipate criminal activity and identify regions that are prone to illegal activities.

Plan of the Paper The study presents "Suraksha," a revolutionary crime prediction algorithm that makes use of Chicago's immense crime dataset. The first section of the paper introduces the paper and highlights the urgent need for advanced crime prediction methods. The second section reviews the literature and highlights the development of crime prediction models and their methodologies. The third section presents the proposed framework, which reveals Suraksha's architecture. The fourth section describes the experimental setup, which explains the methodological flow. The fifth section presents the results, which demonstrate Suraksha's effectiveness. Suraksha forecasts crime hotspots with high predictive accuracy by combining machine learning and advanced analytics.

II. LITERATURE REVIEW

The authors, R. M. Aziz and P. Sharma [6] have suggested using a data-driven strategy to extract insightful information from India's crime statistics. The suggested method builds several regression models on top of various regression techniques. According to this study, with an adjusted R squared value of 0.96 and a MAPE value of 0.2, it is shown that Random Forest Regression (RFR), which predicts total IPC-related crime, fits the data relatively well.

The authors, Varvara and Sergey [5], aim to evaluate the challenge of estimating the quantity of crimes in various parts of the city. Three prediction models are being compared. Using Saint-Petersburg as a case study. The study's authors highlight the importance of feature selection strategies in boosting model accuracy. The best strategy turned out to be gradient-boosting. According to the authors in the future, temporal analysis will probably be used to forecast the frequency of crimes as well as their exact dates. Utilizing cross-validation, the authors calculated the average values of the R2 and MAE measures, which came out to be 0.9.

This research study builds and evaluates an ARIMAX-Transfer Function Model to forecast motorcycle theft rates, according to authors Azhari and Pradita Eko Prasetyo Utomo. The root mean square error and MAPE accuracy levels are 6.68 and 32.30, respectively. Forecasting results for the six periods ahead are produced using the best single input transfer function model for motor vehicle theft cases in the Yogyakarta area police from January to June 2016 [4].

[1] University of Central Lancashire, Preston, England. hirujayawardana18@gmail.com

[2] Universal College Lanka (UCL), Sri Lanka. madushi.pathmaperuma@ucl.lk

* Corresponding Author Email: hirujayawardana18@gmail.com

Using a variety of techniques, the research paper's authors, Koshe Ahana, Mankani Neha, and Sayyed Shafiya [2], conducted a criminal analysis and prediction. Following testing of several machine learning models, including K-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression, the following results were obtained: Random Forest 31% and Decision Tree 30%. Although the authors conducted thorough research, their work still overlooks several critical aspects: demographic and economic factors and the handling of temporal dynamics in crime prediction.

Ying-Lung, Meng-Feng, and Liang-Chih [3] focused their research on one of Taiwan's major cities. In order to improve model prediction performance, the study attempts to compile and combine geographic variables with machine-learning techniques. The authors noted that DNN-tuning showed optimal prediction performance and contrasting impacts. It beat other models in terms of F1 score, even though it had lesser precision in comparison to its greater recall ability. This work stimulates improvements to grid-based crime prediction models. The study is centered on Taiwan's Taoyuan. When used on larger datasets, deep learning models, like DNNs, may become computationally costly and experience scalability problems.

The authors, Ashly Thomas et al. [7], consider a few widely utilized analytical and prediction techniques in their study. The authors next examined the accuracy level by classifying the approaches into the following groups: spatiotemporal, statistical, and neural network approaches. An analysis has been conducted by taking into account the accuracy of the proposed output. By comparing with the actual scenarios, the diversity of crimes taken into consideration and the quality of the dataset are employed. They have concluded from the aforementioned analysis that they can achieve better prediction outcomes by utilizing regression modeling and LSTM. Which, on average, matches the real-world circumstances 90% of the time. For improved outcomes, time series models can also be compared to LSTMs. Regression and LSTM modeling are found to be useful, yet their interpretability may be limited by their complexity. Despite providing insightful information in their study work, the authors did not take into account aspects like interpretability and transparency since these models are used in practical settings.

A study by Samah, Eman M, and Hamdy [14] examined and explored a number of variables that affect criminal activity. The study examined techniques used to forecast future crimes and examined the results. Classification algorithms such as NB, KNN, Decision Tree, Random Forest, Linear Regression, Logistic Regression, and SVM form the foundation of the suggested crime prediction model. Four real datasets were used to test these algorithms: The United States, Egypt, Los Angeles, and Chicago datasets. The Egypt dataset was mostly taken from the website Zabatak.com. It is evident by looking at Samah Samir's research results that temporal dynamics in crime patterns were not taken into explicit consideration in this study. Incorporating temporal factors for more precise predictions could be the subject of future research.

III. PROPOSED FRAMEWORK

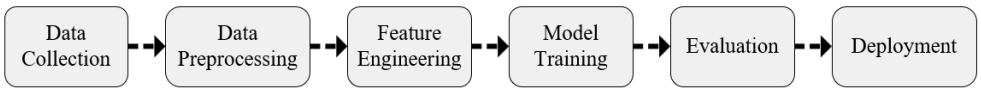

Figure 01

The overall concept of the proposed model, Suraksha, is illustrated in Figure 01 by graphically representing the entire procedure as a series of three primary phases. They are data preprocessing, where data is cleaned and prepared; model building, selecting and training the best algorithm; and model evaluation, using metrics to assess predictions. Each phase is designed to enhance the next, ensuring a smooth and effective machine learning process. And as for the dataset used, this study used a Kaggle dataset which compose of 22 attributes as shown in table 01 and 1,132,978 entries on Chicago crimes from 2017-2021, including numerical and categorical data. The proposed approach uses Random Forest which is an ensemble learning-based machine learning. It is a member of the supervised machine learning algorithm class. It supports both classification and regression problems. This ensemble technique inherently assesses model accuracy without the need for external cross-validation [8]. Due to this reason it is highly effective and efficient for a wide range of predictive modeling challenges. The proposed model has employed RF classification—a categorical variable—to predict the kind of crime in the suggested technique. It has been demonstrated by research investigations that the Random Forest Classifier combines predictions from several decision trees to determine the final class [9]. The Random Forest technique, applied for both classification and regression, predicts continuous outcomes by averaging outputs from numerous decision trees, enhancing accuracy and reliability [10]. The proposed approach uses RF regression to estimate the location of the crime, especially the latitude and longitude, which are continuous variables.

Table 01

ID	ID is a special number that only the records have.
Case Number	An incident number assigned by the Chicago Police Department is unique.
Date	Date of the incident that took place.

Block	The event occurred at the partially obscured address.
IUCR	The Illinois Uniform Crime Reporting code.
Primary Type	The IUCR code's main description.
Description	The IUCR code's supplementary description.
Location Description	An explanation of the location of the incident that occurred.
Arrest	Shows whether someone was arrested.
Domestic	Domestic Violence Act is indicated.
Beat	It is the smallest police geographic area.
District	Shows the police district in which the event took place.
Ward	The incident's ward (district of the City Council)
Community Area	Identifies the locality where the event took place.
FBI Code	Shows the categorization of the offense according to the National Incident-Based Reporting System (NIBRS) of the FBI.
X coordinate	The location's x coordinate on the State Plane Illinois East NAD 1983 projection is where the incident happened.
Y Coordinate	The location's y coordinate on the State Plane Illinois East NAD 1983 projection is where the incident happened.
Updated On	Date and time when the record last updated
Year	The year the incident occurred
Latitude	The latitude at where the event happened
Longitude	The location's longitude where the event happened.
Location	The incident's location in a format that permits this data portal's production of maps and other geographic functions.

IV. Experimental Setup

A. Data Preprocessing

The project's initial phase involved collecting Chicago crime statistics from 2017 to 2021 from CSV files. This dataset is crucial for predictive modeling as it enables to carry out micro analysis of crime hotspots and identify crime types which is essential for accurate spatiotemporal crime forecasts. The next step is data preprocessing. It involves arranging data for mining and optimizing model quality. It includes preprocessing to address missing values, inconsistent formatting, and irrelevant information, thus improving data quality and making it suitable for predictive modeling [11].

B. Feature Selection

Feature engineering is the act of turning raw data into features that help machine learning prediction models better understand the underlying issue and produce better results when the models are applied to data that has not yet been seen [18]. The feature engineering process of the suggested approach is implemented in a few phases. First, any columns that are considered to be irrelevant are manually deleted from the dataset. The proposed approach then transforms the "Date" column into date-time objects and extracts features such as the year, month, day, hour, and day of the week. This method splits a single date-time column into numerous features that help capture temporal trends. In essence, feature selection in the proposed approach combines manual judgment decisions (to eliminate irrelevant features) and automated adjustments.

The suggested model Suraksha is developed with Random Forest classification and regression, both ensemble learning approaches. They work by generating a large number of decision trees during training. Then these techniques output the class, which is the mode of the classes (classification) or the mean prediction (regression) of each tree [19]. The model includes parameters such as n_estimators = 100 to indicate the number of trees in the forest and random_state = 42 to ensure that the findings are reproducible.

The "n_estimator" option determines the number of trees in the forest. Increasing this quantity can boost the model's performance, but it also increases the computational cost and training time. The "random_state" argument guarantees that the splits created are repeatable. Specifying a constant value ensures that the same set of random integers is generated each time the code is executed [18].

C. Temporal and Geographical Analysis

The approach begins with transforming the 'Date' column from string to datetime format, then extracting year, month, day, and time to analyze crime trends over time. This temporal analysis, crucial for understanding how crime patterns evolve with time. Additionally, geographical analysis of the 'Location Description' column identifies frequent crime sites. Together, these insights support spatiotemporal crime predictions, guiding resource allocation and community safety.

D. Model Building

The approach uses two separate pipeline models: a pipeline consisting of a Random Forest Classifier to predict crime categories, and another pipeline consisting of a Random Forest Regressor to estimate crime locations. The model was trained exclusively on data spanning 2017 to 2020, with no exposure to 2021 data during its training phase. A popular technique called "holdout validation" involves setting aside a certain amount of data to assess the model's capacity [12]. The project's visualization involves charting actual vs. expected latitude and longitude values over time. We can quickly assess the correctness of the model and identify trends with the use of visualization.

E. Evaluation Metrics

It's critical to utilize relevant measures that can quantify the machine learning model's capacity to predict outcomes objectively while assessing its performance. Using discrete labels such as "THEFT", "ASSAULT," and so on, the random forest classifier was trained to predict categorical outcomes, or crime categories. The sklearn metrics module's "accuracy score" function is used to assess the classifier's performance.

$$\text{Accuracy} = \frac{\text{Total number of predictions made}}{\text{Number of correct predictions}} \quad (1)$$

On the other hand, the random forest regressor forecasts continuous results. However, various evaluation metrics are needed depending on the latitude and longitude of the crimes. The RMSE is determined for both latitude and longitude forecasts using the formula "mean_squared_error."

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}i)^2} \quad (2)$$

y_i : This is the actual value

$\hat{y}i$: This is the predicted value

n : This is the number of predictions

The RMSE for latitude and longitude of the model is mentioned in section 5 Results and Discussion. The RMSE is a commonly used metric to assess how inaccurate a model is in forecasting quantitative data. For the year of 2021 test set, predictions are produced after training. The accuracy of the crime type model is used to evaluate its performance, and the root mean squared error (RMSE) for both latitude and longitude is used to evaluate the performance of the location model. These measures aid in understanding how well the algorithms forecast the types and locations of crimes.

V. Experimental Results

A. Graphs

As show in figure 02, it visualizes a heatmap that shows the frequency of crimes by weekday and month. Next, Seaborn is used to create the heatmap. The 'cool-warm' color scheme of the graph is used to visually differentiate between crimes with greater and lower rates. The precise number of offenses is shown in annotations on the heatmap, which facilitates the identification of certain patterns. For instance, what days and months have greater crime rates combined? This graphic does a good job of highlighting the temporal trends in crime.

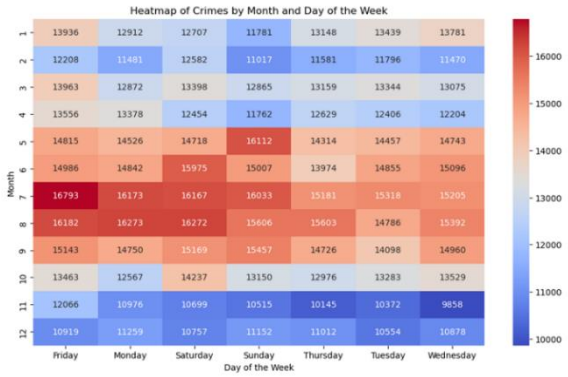

Figure 02

The figure 03 is a correlation heatmap. It visualizes a heatmap that shows the frequency of crimes by weekday and month. Next, Seaborn is used to create the heatmap. The 'cool-warm' color scheme of the graph is used to visually differentiate between crimes with greater and lower rates. The precise number of offenses is shown in annotations on the heatmap, which facilitates the identification of certain patterns. For instance, what days and months have greater crime rates combined? This graphic does a good job of highlighting the temporal trends in crime.

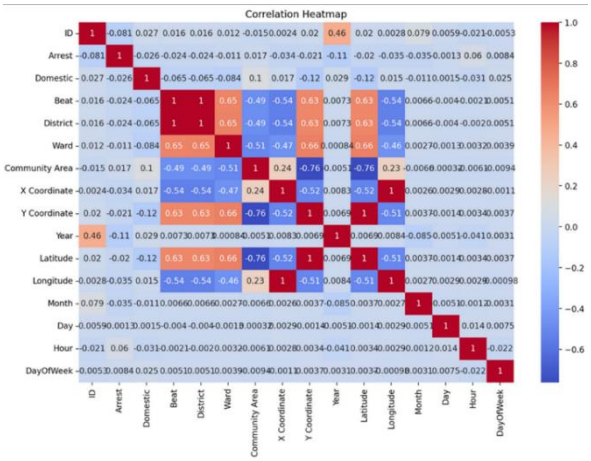

Figure 03

Figure 04 and Figure 05 shows a User-Centered Interface design which is one of the main goals of the proposed approach. Through this interface the user inputs specified date for prediction at the beginning of the procedure. The Suraksha system predicts crime types and locations and displaying results on an interactive map with a heatmap overlay with the help of pretrained models. It enhances law enforcement with insightful crime trend analysis, informed decisions, and user-focused, interactive crime forecasting tools. One of the advantages of using Suraksha for law enforcement is the ability to gain awareness through a thorough grasp of crime trends and hotspots. Accessibility to previous data analysis also facilitates well-informed decision-making. The application's use of dynamic and interactive illustrations for data interpretation improves the analytical process overall and increases user engagement.

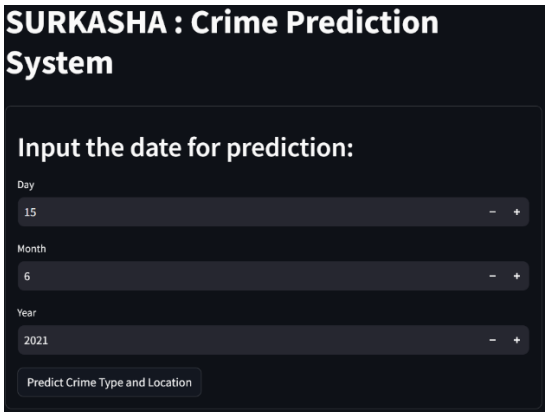

Figure 04

Figure 05

B. Feature Engineering

Feature engineering is a critical yet labor-intensive aspect of machine learning systems [17]. The model presents a collection of temporal features, a location feature, and a few additional variables that are associated with criminal behavior, hence improving the predictive performance of machine learning algorithms. According to the correlation heatmap in figure 02, the attribute “District” has the highest association with geographical coordinates.

Time Features: Parsing the Date column to extract the year, month, day, hour, and day of the week. These temporal traits are important because they reflect trends and patterns throughout time.

Location Features: Location features are used in spatial data analysis. Including this feature in the analysis helps to simplify the model by restricting it to a set of particular geographical categories, rather than only considering location as a continuous variable.

Additional features: The suggested model additionally includes binary properties such as "Arrest" and "Domestic". Binary characteristics are easy for models to comprehend since their binary nature makes them simple to include into a model.

The features (X) for the model are chosen by removing particular columns from the dataset (primary type, latitude, and longitude). The target variables are the geographic coordinates and the type of criminal activity. Conducting feature engineering research and looking into adding new variables or features that could reflect the distinctive qualities of different places and time periods is one of the main goals of the proposed approach.

C. Results and Discussion

The suggested model calculates the standard deviation of both latitude and longitude values. The standard deviation is a measurement of how widely distributed the numbers in a data set are [15]. In this context, it indicates how far the crime scenes depart from the mean (average) position. The model was able to obtain a value of 0.0880 for latitude standard deviation and 0.0603 for longitude standard deviation. The proposed model also calculates ranges. It is computed using the range of actual latitude and longitude data. It accomplishes this by calculating the maximum latitude and subtracting the minimum latitude. This provides the range of values from least to greatest, indicating how widely distributed the locations are throughout both dimensions [16]. This model was able to obtain a latitude of 0.3779 and a longitude of 0.4152. Finally, the model computes the RMSE value. This is a standard method for measuring a model's inaccuracy in predicting quantitative data. The model outputs RMSE values of 0.0874 for latitude and 0.0602 for longitude. The findings show that the model's location predictions varied by these little amounts from the actual data. These measures aid in understanding how well the algorithms forecast the types and locations of crimes. As future development we may include real-time data into a spatiotemporal crime prediction project, which would greatly improve its overall usefulness, accuracy, and responsiveness.

VI. Conclusion

Crime prediction is currently one of the most popular social trends. The goal of crime prediction is to lower the frequency of crimes. It achieves this by forecasting the kind of crime that could happen in the future [2]. A significant leap ahead in the use of machine learning and spatiotemporal analytics for crime prevention and prediction has been made with the development of Suraksha. The gathering of data, preprocessing, and use of Random Forest algorithms for both classification and regression tasks form the foundation of the suggested model. This initiative has shown potential for giving law enforcement organizations an effective tool to forecast and reduce crime. The depiction of crime patterns and hotspot forecasts are examples of experimental outcomes. The outcomes also demonstrate how well the system can see trends and predict criminal activity in the future. When putting technological advances into practice for a spatio-temporal crime prediction model, we must also take sensitive areas like public safety and privacy into account. Because of this, the project highlights how important ethical issues and user-centered design are. Through the development of a collaborative approach that incorporates knowledge from criminology, data science, and law enforcement, Suraksha raises the bar for predictive policing technology. Future work for the project will involve the integration of real-time data and the exploration of scalable frameworks. This could potentially contribute to increasing its applicability in various urban environments and attempting to increase its potential to protect individuals. The Random Forest Regressor produced Root Mean Squared Errors (RMSE) of 0.0874 and 0.0602 for latitude and longitude predictions, respectively, highlighting the predictive accuracy of the model. This accuracy shows how reliable the model is in predicting locations, underscoring its potential as a vital instrument for public safety and crime prevention.

Furthermore, when working with more data characteristics and bigger dimensions, the efficiency and efficacy of spatial crime prediction models can be improved. This can be done by applying feature selection, feature extraction, and cross-validation approaches. By using these methods, researchers may improve the capacity of spatial crime prediction models to identify and extend significant patterns and trends, as well as more successfully handle data sparsity difficulties [13].

VII. Acknowledgment

I express my gratitude to my Supervisor, Dr. Madushi Pathmaperuma, for dedicating her significant time to evaluate and offer feedback on my research. I also like to express my gratitude to Kaggle for providing crime datasets.

References

[1] "The Global Organized Crime Index 2023 - World | ReliefWeb," reliefweb.int, Sep. 29, 2023.
[2] Koshe Ahana Hemant, Mankani Neha Haresh, and Sayyed Shafiya Majid, "A Design to Predict and Analyze Crime," International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 5, pp. 2527–2536, May 2022.
[3] Y.-L. Lin, M.-F. Yen, and L.-C. Yu, "Grid-Based Crime Prediction Using Geographical Features," ISPRS International Journal of Geo-Information, vol. 7, no. 8, p. 298, Jul. 2018.
[4] Azhari and Pradita Eko Prasetyo Utomo, "Prediction the Crime Motorcycles of Theft using ARIMAX-TFM with Single Input," Oct. 2018.
[5] V. Ingilevich and S. Ivanov, "Crime rate prediction in the urban environment using social factors," Procedia Computer Science, vol. 136, pp. 472–478, 2018.
[6] R. M. Aziz, P. Sharma, and A. Hussain, "Machine Learning Algorithms for Crime Prediction under Indian Penal Code," Annals of Data Science, Jul. 2022.
[7] Thomas and N. V. Sobhana, "A survey on crime analysis and prediction," Materials Today: Proceedings, Feb. 2022.
[8] L. Wang, Z.-P. . Liu, X.-S. . Zhang, and L. Chen, "Prediction of hot spots in protein interfaces using a random forest model with hybrid features," Protein Engineering Design and Selection, vol. 25, no. 3, pp. 119–126, Jan. 2012.
[9] M. Khan, A. Ali, and Y. Alharbi, "Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques," Complexity, vol. 2022.

[10] H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, no. 3, pp. 1–21, Mar. 2021.
[11] S. Vieira, R. Garcia-Dias, and W. H. Lopez Pinaya, "Chapter 19 - A step-by-step tutorial on how to build a machine learning model," *ScienceDirect*, Jan. 01, 2020.
[12] J S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Feb. 2016.
[13] Y. Du and N. Ding, "A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods," *ISPRS international journal of geo-information*, vol. 12, no. 6, pp. 209–209, May 2023
[14] S. Zahran, E. Mohamed, and H. Mousa, "Detecting and Predicting Crimes using Data Mining Techniques: Comparative Study," *IJCI. International Journal of Computers and Information*, vol. 8, no. 2, pp. 57–62, Dec. 2021.
[15] X. Wan, W. Wang, J. Liu, and T. Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC Medical Research Methodology*, vol. 14, no. 1, Dec. 2014.
[16] D. K. Lee, J. In, and S. Lee, "Standard Deviation and Standard Error of the Mean," *Korean Journal of Anesthesiology*, vol. 68, no. 3, p. 220, May 2015.
[17] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," *IEEE Xplore*, Mar. 01, 2016.
[18] Adelson de Araujo, N. Cacho, Leonardo, C. Vieira, and J. Borges, "Towards a Crime Hotspot Detection Framework for Patrol Planning," Jun. 2018.
[19] J. Borges *et al.*, "Feature engineering for crime hotspot detection," *IEEE Xplore*, Aug. 01, 2017.