Analysis of the Chicago Crimes over the last 6 years

Margarida Cardeano Pinheiro
Dept. of Computer Science
Faculty of Sciences of the University of Porto
Porto, Portugal
up201805012@edu.fc.up.pt

***Abstract*— Crime is an unfortunate and prevalent aspect of our society; its impact can be felt by both victims and perpetrators, as it has affected the lives of countless individuals. This work analyzes Chicago crimes over the past six years using the "Chicago Crime dataset report" from 2017 to 2023. We examine trends, crime hotspots, and develop predictive insights. By leveraging supervised machine learning techniques, including random forest, and ensemble methods, we aim to predict crimes based on time, location, and other parameters. This research contributes to understanding crime patterns, aiding law enforcement, policymakers, and urban planners in developing effective strategies for crime prevention and public safety.**

Keywords*— *crime analysis, chicago crimes, trends, hotspots, predictive modeling, supervised machine learning, random forest, ensemble methods, crime prevention, public safety.

I. INTRODUCTION

The analysis of crime patterns and trends is crucial for maintaining public safety and developing effective strategies to combat criminal activities.[1] In recent years, the city of Chicago has faced significant challenges regarding crime rates and the overall impact on its communities.[1,2] Understanding the dynamics of crime is essential for law enforcement agencies to allocate resources efficiently and enhance preventive measures.[2] This work focuses on analyzing the Chicago crimes over the past six years to gain insights into the nature of criminal activities, identify high-risk areas, and assist police officers in making informed decisions. By providing valuable and hard-to-obtain information, our objective is to support law enforcement personnel in implementing appropriate measures to mitigate crime effectively. By harnessing the power of data analysis and machine learning, we aim to contribute to the ongoing efforts to address crime-related issues in Chicago. The outcomes of this study can guide law enforcement agencies, policymakers, and community organizations in formulating targeted strategies to reduce crime rates, enhance public safety, and foster the well-being of residents throughout the city.

II. DATA UNDERSTANDING

The dataset used for the analysis of Chicago Crimes over the last 6 years comprises incidents reported from 2017 to 2022. It is sourced from the official records of the Chicago Police Department, providing reliable and accurate information. The dataset includes crucial details such as the date, type, description, and location of each crime. It offers valuable insights into crime patterns, helping researchers understand trends and inform evidence-based policies.

After conducting a preliminary examination, we discovered that there were some data elements that were not worth considering for analysis, such as the FBI code. After careful evaluation, we decided to exclude these data points as they did not directly contribute to the objectives and metrics of our study.

Additionally, we identified duplicated "id" and "case_number" entries within the dataset. Recognizing that such duplications could skew our results and compromise the reliability of our analysis, we performed data cleansing by removing all duplicate entries. This ensured the integrity and consistency of our dataset.

Furthermore, we found that the cases from April 2023 were not significant for our analysis. We determined that these cases lacked pertinent characteristics or patterns relevant to our study. Therefore, we made the decision to exclude the data pertaining to that specific month, focusing our analysis on the preceding months that held greater relevance and directly contributed to understanding crime patterns in Chicago.

III. DATA PREPARATION

In this section of the work, we focus on the crucial step of data preparation for analyzing Chicago crimes data. Data preparation involves several essential tasks, including cleaning the dataset, handling missing values, checking for duplicates, and organizing the data for further analysis. By following these steps, we ensure data integrity and create a reliable foundation for subsequent exploration and modeling.

The initial part of the process involves reading the dataset and performing basic cleaning tasks, such as standardizing column names and removing unnecessary columns. We then proceed to address missing values, eliminating rows with missing or empty values to ensure the quality of the dataset.

Next, we check for any duplicate entries based on specific columns and take necessary actions to maintain data integrity. Duplicate entries can distort analyses, so we remove rows with duplicate case numbers, resulting in a clean dataset with unique cases.

To facilitate temporal analysis, we convert the date column into a standardized format and create a new column to extract the year and month. This allows us to count the number of crimes per month, providing valuable insights into the temporal trends of crime in Chicago.

Visualizing the crime trends through a monthly plot helps us identify patterns and changes over time. We create a line plot that displays the crime count for each month, offering a clear visual representation of the evolution of crime in Chicago. Additionally, we subset the data to exclude crimes that occurred in specific months, saving the subsets for further analysis. This allows us to create a hidden set, which contains crimes from January, February, and March 2023, for evaluation or validation purposes. The original dataset is also updated to exclude these months, ensuring a comprehensive analysis excluding the hidden set.

IV. EXPLORATORY DATA ANALYSIS

In this part of the work, various data visualization techniques are employed to analyze and visualize the dataset after the data preparation. The visualizations aim to provide a clear and intuitive understanding of different aspects of the data, such as correlations between variables, crime types, locations, temporal patterns, and more.

A. *Crime and Arrests evolution*

From 2017 to 2019, the number of crimes in Chicago varied, with peaks in the summer months and valleys in the winter months. There was a sharp decrease in crime from the beginning of 2020, with only a slight increase in the summer. Since then, crime has never reached the levels of 2017-2019, but there has been a significant increase in crime by the end of 2022. The number of arrests closely mirrored the trend of crime rates, with the only exception being between 2017 and 2019 and from late 2020 until 2022, where it remained relatively constant.

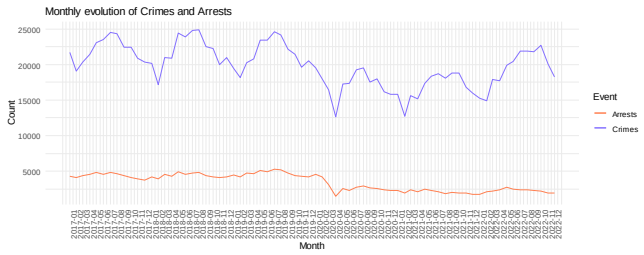

Fig. 1. Monthly evolution of crimes and arrests from 2017 to 2022.

B. *Number of crimes by ward and year*

The heatmap plot shows the number of crimes in different wards and years. Each tile represents a combination of ward and year, and the color intensity reflects the crime count. This visualization helps identify trends and patterns in crime across wards and over time.

The data reveals a decrease in the number of crimes in all wards from 2020 onwards. Additionally, the wards closest to the center of Chicago (wards 27, 28 and 42) have the highest crime rates.

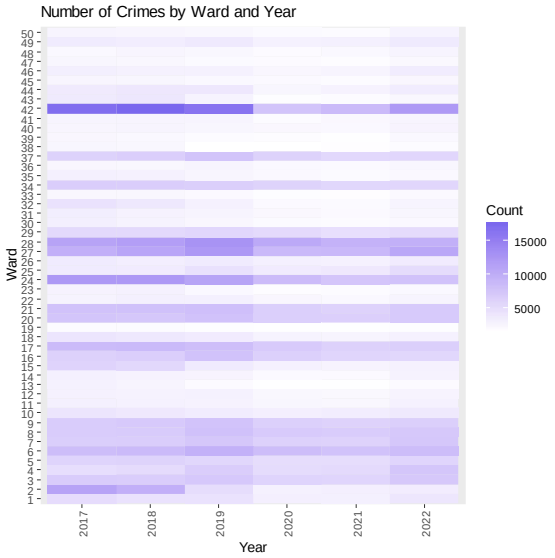

Fig. 2. Number of crimes by ward and year (2017 to 2022).

C. *Top 5 crime types*

The map obtained displays the top 5 crime types and their occurrences over the years. Each line represents a specific crime type, while the y-axis shows the number of occurrences. This visual representation allows for a comparison of occurrence patterns over time.

The graph indicates that despite a sharp decrease in robberies in 2020, it remains the most prevalent crime type, followed closely by assaults.

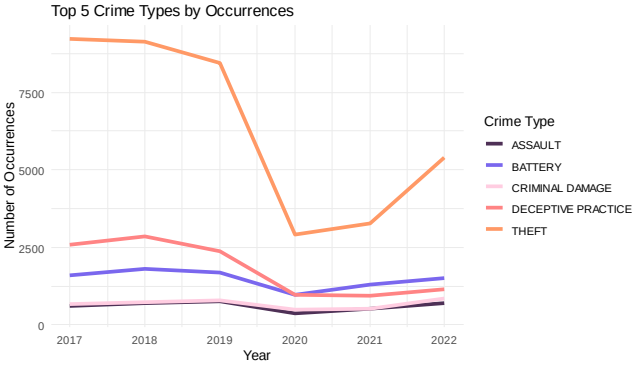

Fig. 3. Top 5 crimes types by occurrences from 2017 to 2022.

D. *Theft Crimes Per Ward*

The resulting map showcases markers indicating specific crime types, such as theft, and their respective locations. To enhance clarity when multiple crimes occur in close proximity, the markers are clustered together. Clicking on a marker reveals a detailed pop-up containing pertinent information about the crime, such as its primary type, date, description, and location details. Additionally, the map includes clearly delineated ward boundaries, with each ward labeled by its corresponding number. By overlaying ward boundaries, the map offers a comprehensive view of theft occurrences in Chicago, highlighting a higher concentration of such crimes in the city's urban center.

V. ASSOCIATION RULES

To perform the association rule analysis, we started by selecting columns ‘primary_type’, ‘location_description’, ‘time_of_day’, and ‘arrest’ from the original crimes’ dataset. Before generating the rules, we set the minimum support to 0.03 (3%) and the minimum confidence to 0.2 (20%). 125 rule(s) were created.

In association rule mining, rules are generated to discover interesting relationships or patterns in a dataset. Each rule consists of an antecedent (lhs) and a consequent (rhs), and the support, confidence, coverage, lift, and count values provide measures of the strength and significance of the rules.

Here are the interpretation of some rules that were found:

Rule [8] indicates that when "primary_type=NARCOTICS," the consequent "arrest=TRUE" occurs with a support of 0.0374 and confidence of 0.9931. This suggests that when the primary type is "NARCOTICS," there is a high probability (99.31%) of an arrest occurring.

Rule [12] shows that when the primary type is "MOTOR VEHICLE THEFT," the location description is likely to be "STREET" with a support of 0.0393 and confidence of 0.7385. This implies that when a motor vehicle theft occurs, it is common for it to happen on the street.

Rule [17] If a crime occurs in a STORE, there is a strong association with the primary_type being THEFT (support: 0.0413, confidence: 0.6102, lift: 2.7132).

Rule [42] and **[43]** indicate a strong association between the time of day being "MORNING" and the primary type being "THEFT". The support and confidence values suggest that "THEFT" incidents are more likely to occur in the morning.

These are just a few examples of the rules generated. Each rule captures a specific association or pattern between different variables in the dataset. The support value represents the frequency of the rule in the dataset, while the confidence value indicates the strength of the association between the antecedent and consequent. The lift value measures the deviation from independence, and the coverage value represents the proportion of instances covered by the rule.

VI. LINK ANALYSIS

Focusing on the "primary_type" and "ward" columns. We constructed a graph where each vertex represents a ward, and the edges between vertices indicate the connections between wards based on the similarity of crime occurrences. The vertex labels show the ward numbers.

We started by calculating the degree, closeness, and betweenness centrality measures for the graph. These measures provide information about the importance or centrality of each vertex (ward) in the network.

However, most importantly, we decided to focus our analysis on the community structure, to detect communities or clusters within the graph. The communities obtained are visualized by coloring the vertices with different colors. The resulting graph shows different communities as distinct groups or clusters:

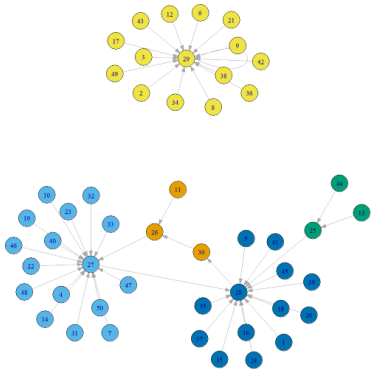

Fig. 4. Relationships between wards based on shared crime types.

By analyzing the crime data within each community, we can gain insights into any patterns, trends, or characteristics specific to those areas. In this case, the community information obtained can be useful for understanding the distribution of crime across different wards and identifying potential factors influencing crime rates in each community.

VII. RECOMMENDATIONS

We started by generating recommendations for the different wards, these recommendations are derived by analyzing the historical patterns of primary types of crimes in other wards that are similar to the target ward. The recommendations suggest primary types of crimes that have been frequently observed in the similar wards.

The x-axis represents the wards, and the color of each bar in the plot represents the recommended primary types of crimes for the corresponding ward. Different colors indicate different primary types of crimes.

The top-N Recs value (y-axis in the plot) represents the index of the recommendation within the top N list. In other words, the values ranging from 1 to 7 indicate the position of each recommended primary type of crime within the top 7 recommendations for each ward.

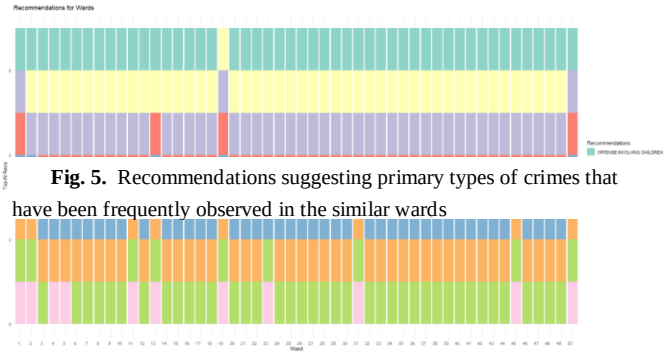

Fig. 5. Recommendations suggesting primary types of crimes that have been frequently observed in the similar wards

In summary, the plot presents a visual representation of the recommendations, enabling a side-by-side comparison of the suggested crime types that are likely to occur in different wards, providing insights into the crime patterns

and potential areas of focus for law enforcement or preventive measures.

VIII. MODELLING

The analysis of the machine learning models focused on predicting arrests in crime data using various predictor variables. The target variable of interest was 'arrest', and the predictors included 'block,' 'IUCR,' 'location_description,' 'beat,' 'district,' 'ward,' 'community_area,' 'year,' 'time_of_day,' 'quarter,' and 'week_day'.

To ensure reliable model evaluation, the data was divided into training and testing sets. Essential data preprocessing steps were performed, such as removing irrelevant columns and addressing class imbalance by creating a balanced dataset with equal counts of 'False' and 'True' arrests. Next, categorical features and the target variable were encoded using the LabelEncoder. This encoding transformation prepared the data for utilization by machine learning models.

For this analysis, four machine learning approaches were explored: Decision Tree, Random Forest, AdaBoost, and Bagging of Trees. These models were evaluated based on their performance in accurately predicting arrests.

Table 1. Model performance evaluation. Comparative results of the Models for the test set and the hidden set.

TEST SET

Approach	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.603494	0.60489	0.596833	0.600835
Random Forest	0.716492	0.747884	0.653171	0.697326
AdaBoost	0.699048	0.710328	0.67223	0.690754
Bagging Classifier	0.619107	0.621085	0.610931	0.615966

HIDDEN SET

Approach	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.647454	0.124021	0.310644	0.17727
Random Forest	0.814	0.234446	0.230112	0.232259
AdaBoost	0.751978	0.206522	0.361905	0.262976
Bagging Classifier	0.689167	0.117957	0.238095	0.615966

The test set results indicate that Random Forest performed the best among the evaluated machine learning approaches, achieving an accuracy of 71.65% with a balanced precision and recall. AdaBoost also showed promising results with an accuracy of 69.90% and a relatively balanced trade-off between precision and recall. The Decision Tree and Bagging Classifier models had lower accuracy and showed room for improvement. However, on the hidden set, Random Forest remained the top performer with an accuracy of 81.40%, although precision and recall were relatively low. AdaBoost showed improvement over the Decision Tree on the hidden set as well.

IX. CONCLUSION

In conclusion, the analysis of Chicago crimes over the past six years revealed some interesting patterns and trends. Through meticulous data preparation and exploration processes employed during this analysis, was observed a significant decrease in crimes in 2020, and they never returned to the high levels seen before that year. The COVID-19 pandemic, which caused lockdowns and altered societal dynamics, is one of the factors that contributed to the decline in crime in 2020.[3] These actions discouraged criminal behavior and broke up criminal networks.

Notably, findings revealed that various types of criminal activities were detected mostly occurring around Chicago's central districts than other regions within town limits. The larger population density, increased commercial activity, and improved accessibility in the city's heart all contribute to the concentration of crimes there. The greater crime rates seen may also be attributed to social and economic reasons including gang activity, inequality, and poverty.

Regarding the modeling results, we employed different approaches on both the test set and the hidden set. Among the tested models, the Random Forest algorithm achieved the highest accuracy on the test set, followed closely by AdaBoost. These models demonstrated reasonable precision, recall, and F1 scores, indicating their effectiveness in predicting arrest patterns.

However, when evaluating the models on the hidden set, the performance metrics dropped significantly across all approaches. This outcome suggests that the models may not generalize well to unseen data or that there were underlying complexities or shifts in the hidden set that the models struggled to capture. OneHotEncoding was not utilized due to the high cardinality of the categorical variables present (which had numerous categories), and the insufficient computational capacity available to handle such complexity.

Overall, our analysis and modeling provide valuable insights into the patterns and characteristics of Chicago crimes. The findings underscore the importance of ongoing efforts to address crime prevention and law enforcement strategies, particularly in the central areas of the city.

REFERENCES

DataSet: Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

[1] Papachristos, Andrew V. 48 YEARS of CRIME in CHICAGO. Yale ISPS, 9 Dec. 2013.

[2] William, P, et al. “Crime Analysis Using Computer Vision Approach with Machine Learning.” Mobile Radio Communications and 5G Networks, Edited by Nikhil Marriwala et al., Springer Nature Singapore, 2023, pp. 297–315.

[3] Yang, Mengjie, et al. “The Impact of COVID-19 on Crime” ISPRS International Journal of Geo-Information, vol. 10, no. 3, 1 Mar. 2021, p. 152.