# Chicago Crime Analysis

Aayusha Shrestha, Bishwas Kafle, Suman Astani

STAT 6440 : Data Mining

Dr. Shuchismita Sarkar

April 22, 2023

## Table of Contents

**Abstract.......................................................................................... 1**

**1. Introduction..........................................................................................1**

1.1 Motivation..........................................................................................1

1.2 Objective..........................................................................................1

1.3 Data..........................................................................................2

**2. Methodology..........................................................................................4**

2.1 K-Nearest Neighbor (KNN)..........................................................................................4

2.2 Decision Tree.......................................................................................... 4

**3. Results..........................................................................................5**

3.2 Exploratory Analysis.......................................................................................... 5

3.3 Predictive Analysis.......................................................................................... 8

3.3.1 KNN..........................................................................................9

3.3.2 Decision Tree.......................................................................................... 9

3.3.3 Overall Result..........................................................................................11

**4. Conclusion.......................................................................................... 11**

**5. Limitations and Future Work..........................................................................................12**

5.1 Limitation of the analysis..........................................................................................12

5.2 Future work..........................................................................................12

**6. References..........................................................................................15**

## Abstract

This paper examines the arrest decision in Chicago, utilizing data from the Chicago Data Portal for analysis. Two different methods, namely K-nearest neighbor and decision tree, were employed to predict the likelihood of an arrest based on the type of crime. The performance of both models was evaluated and the decision tree model was found to have the highest accuracy, achieving 85.62% accuracy in predicting arrests.

## 1. Introduction

### 1.1 Motivation

Crime is a complex social problem that law enforcement agencies in Chicago have been grappling with, given the high crime rate in the area. To address this challenge, machine learning has emerged as a valuable tool for predicting arrest decisions. In this study, we aim to explore the opportunities and challenges associated with using machine learning or predicting arrest decisions in Chicago. By analyzing data from law enforcement agencies and leveraging machine learning techniques we seek to shed light on the potential benefits and limitations of using this approach in the context of crime prediction and arrest decision-making.

### 1.2 Objective

Our objective behind this study is to build machine learning models capable of predicting whether a person will be arrested or not, evaluate and compare the performances of different data mining algorithms in terms of their predictive accuracy, sensitivity, and specificity and identify the opportunities and challenges for predicting arrests, including issues related to data quality, feature selection, and model interpretability.

By addressing these objectives, we aim to contribute to the understanding of using machine learning for predicting arrests and provide insights into the opportunities and challenges associated with this approach.

### 1.3 Data

The dataset used for this project is the “Crimes-2001 to Present” publicly available in the Chicago Data Portal. It contains information about the reported crime incidents in the city from 2001 to the present day. The dataset includes details such as the type of crime committed, the location where the crime occurred, the date and time of the crime, and the arrest status of the offender. It also includes information about the community area, ward, and police district where the crime was reported [4].

A description of the variables in the dataset is in the table below:

| **Variable Name** | **Description** |
|---|---|
| ID | Unique identifier for each reported crime incident |
| Case Number | Unique identifier for each reported crime incident, including the year it was reported |
| Date | The date and time the crime was reported to the police |
| Block | The block address where the crime occurred |
| IUCR | The Illinois Uniform Crime Reporting code for the type of crime committed |
| Primary Type | Primary description of the crime committed |

| | |
|---|---|
| Description | A more detailed description of the type of crime committed |
| Location Description | The location type where the crime occurred |
| Arrest | Whether an arrest was made in connection with the crime (true/false) |
| Domestic | Whether an arrest was classified as domestic-related |
| Beat | The police beat where the crime occurred |
| District | The police district where the crime occurred |
| Ward | The ward where the crime occurred |
| Community Area | The community area where the crime occurred |
| FBI Code | The Federal Bureau of Investigation code for the type of crime committed |
| X-Coordinate | The x-coordinate of the location where the crime occurred |
| Y-Coordinate | The y-coordinate of the location where the crime occurred |
| Year | The year the crime was reported to the police |
| Updated On | The date the record was last updated in the data portal |
| Latitude | The latitude of the location where the crime occurred |
| Longitude | The longitude of the location where the crime occurred |

| | |
|---|---|
| Location | The location of the crime, in latitude and longitude coordinates |

Though the dataset has 22 features, we shall drop irrelevant features and consider only features that seem necessary in predicting the response variable. Regarding the response variable, it's important to note that in our crime prediction project, the primary focus was predicting the likelihood of an arrest, rather than the type of crime. Therefore, the models and analysis were geared toward predicting the binary outcome of "Arrest" based on the independent variables.

## 2. Methodology

In order to predict the arrest decision, a variety of classification methods have been used, and their performances, shall be evaluated.

### 2.1 K-Nearest Neighbor (KNN)

K-nearest neighbor (KNN) is a non-parametric supervised model which can be used for both classification and regression. The algorithm classifies new data points based on the nearest neighbor data points. The idea of KNN is to identify *k* records (neighbors) in the training dataset that are similar to the new record to be classified. It will assign the class of the majority of the neighbors. [1]

### 2.2 Decision Tree

A decision tree is a supervised machine-learning algorithm used to categorize or make predictions based on how a previous set of questions were answered [2]. The idea is to create a model that predicts the value of target variables by learning simple decision rules inferred from the data features. The decision is made by splitting the root node of a feature into child nodes until reaching terminal nodes with all observations belonging to the same class. The predictive model uses recursive partitioning until a decision is made to classify a new observation [3].

## 3. Results

### 3.1 Data Cleaning and Preprocessing

To prepare the data for analysis, we first dropped data with missing values to ensure data integrity and accuracy. Features deemed irrelevant or redundant for the analysis were filtered out to reduce noise and improve the efficiency of the data analysis process. The original “Data” variable was separated into three distinct variables - “Day”, “month” and “Weekday” to enable more granular analysis and extraction of relevant insights based on temporal patterns. Similar crime types were merged into a single category to reduce the number of categories and simplify the analysis. By implementing these data cleaning and preprocessing techniques, the dataset was refined to ensure the data was ready for further analysis and modeling.

### 3.2 Exploratory Analysis

During the data exploration phase, a meticulous preliminary analysis was conducted to identify pertinent features that could be utilized for constructing a predictive model. This comprehensive analysis uncovered intriguing trends and patterns within the data, shedding light on valuable insights that could inform the development of an effective predictive model.

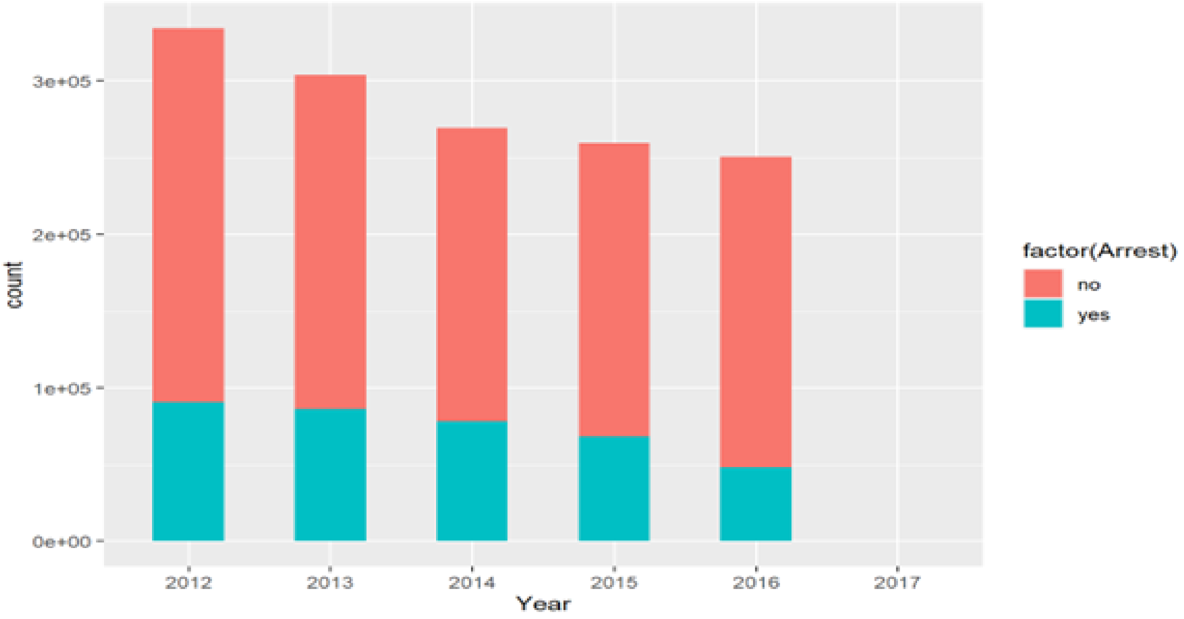

*Fig 1 : Bar graph showing the proportion of arrests for different years*

The figure above presents crime statistics for Chicago, encompassing the number of crimes and arrest rates. The visual representation provides a comprehensive overview of the occurrence of crimes and arrests from 2012 to 2017. Notably, it is evident from the graph that both arrest rates and crime rates have declined steadily over the years, indicating a downward trend in criminal activities in the region.

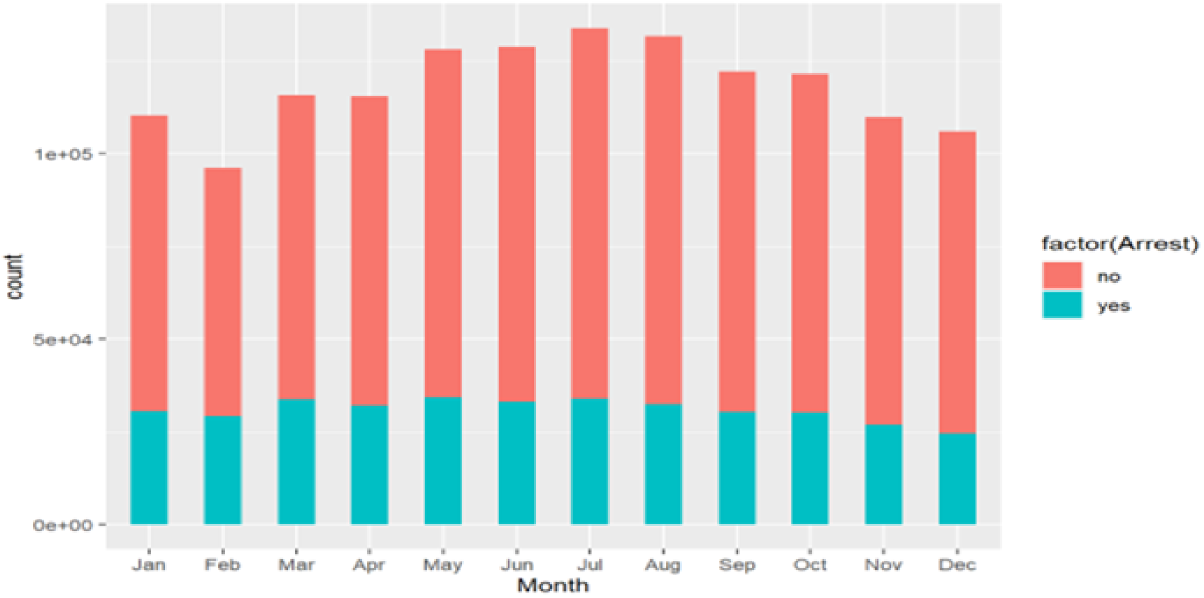

*Fig 2: Bar graph showing the proportion of arrests for different months*

The figure above illustrates the crime statistics for Chicago, capturing the number of crimes and arrest rates on a monthly basis over the period of 2012-2016. The visual representation provides a comprehensive summary of crime occurrences and arrests throughout the months. Notably, the graph reveals a discernible seasonable influence on crime rates, indicating that crime incidents may fluctuate based on the time of year. This insight underscores the importance of considering temporal patterns when analyzing crime data and devising effective strategies for crime detection and prevention.

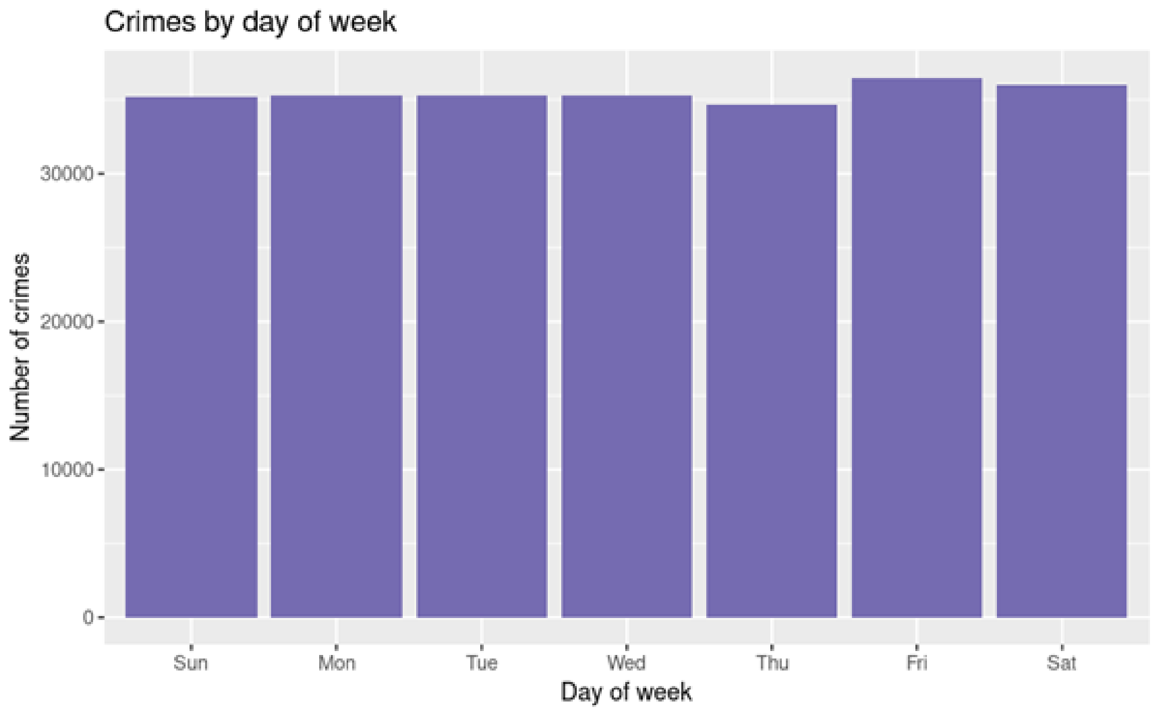

*Fig 3: Crime incident count distribution for different days of the week*

The visualization depicts the occurrences of crimes on different days of the week in Chicago. It reveals that Friday has the highest number of crimes, while Sunday has the lowest of crimes.

## 3.3 Predictive Analysis

### 3.3.1 KNN

For the prediction task, we employed the KNN algorithm which is a simple yet effective method that stores all available cases and classifies new cases based on a majority vote of its k nearest neighbors. This algorithm allows for the grouping of unlabeled data points into distinct categories.

Since all our explanatory variables were categorical, we created dummy variables to represent them. This step was crucial as it ensured that the values for each variable were on a consistent scale. Specifically, we converted the values to 1 if the feature was present and 0 if absent, transforming all values to a common scale. (The selection of explanatory variables was based on the pre-processing conducted during the exploratory analysis. A total of 7 variables were chosen for inclusion in the predictive model. These variables were deemed relevant for the prediction task. The inclusion of these variables in the model was based on their potential to contribute to accurate predictions and insights into the crime detection and prevention process).

The KNN algorithm was then applied to the training dataset, and the results were verified on the test dataset. To achieve this, we divided the dataset into two portions, with 80% of the data used for training and 20% for testing. This allowed us to evaluate the performance of our model on unseen data and assess its predictive accuracy.

All of the features were then converted into binary vectors to ensure equal representation and individual weights for each possible choice of each feature. Preprocessing was performed on all categorical variables to map them to their corresponding binary feature vector representations.

To train the model, the knn() function was utilized. The function identifies the k nearest neighbors using Euclidean distance, with the value of k being set to 5. This function returned a factor value of arrest labels for each observation in the test data, allowing for the prediction of arrest likelihood for the test dataset based on the trained model.

| | |
|---|---|
| Accuracy | 0.8372 |
| Kappa | 0.5051 |

### 3.3.2 Decision Tree

For building and visualizing decision trees, we utilized the “rpart” package. The decision tree model was specifically designed for classification purposes, as indicated by setting the “method” parameter to “class”. In order t control the tree’s complexity and prevent overfitting, we set the “minsplit” parameter to 15, which determined the minimum number of observations required to split a node, and the “minbucket” parameter to 5, which specifies the number of observations required in a terminal node.

To visualize the tree, we used the “prp” function from the “rpart.plot” package, which provided various formatting options. For instance, the “type”, “extra”, “under”, “split.for” and “varlen” parameters were used to customize the appearance of the tree, such as the type of the plot, additional graphical elements, placement of split labels, font size for splits, and variable name length displayed on the tree.

The combination of the “rpart” package for model building and the “rpart.plot” package for visualization allowed us to create an interpretable and visually appealing representation of the decision tree model in our analysis.

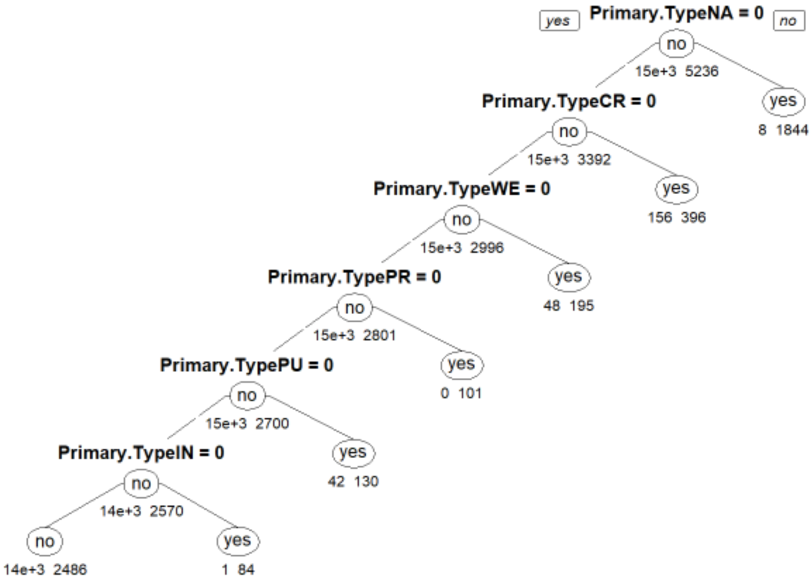

*Fig 4 : Decision Tree*

| | |
|---|---|
| Accuracy | 0.8562 |
| Kappa | 0.5627 |

### 3.3.3 Overall Result:

Based on the comparison of the results from the confusion matrix, our analysis suggests that the decision tree model performed slightly better than the KNN model, by a small margin.

| MODELS | ACCURACY | KAPPA | SENSITIVITY | SPECIFICITY |
|---|---|---|---|---|
| KNN | 0.8372 | 0.5051 | 0.4580 | 0.9718 |
| DECISION TREE | 0.8562 | 0.5627 | 0.5008 | 0.9807 |

## 4. Conclusion

In this project, we utilized data mining techniques to perform predictive tasks on crime data, specifically focusing on predicting arrests for a given type of crime in a specific location. Before training the models, we conducted pre-processing on the data to identify relevant features for building the predictive models. We applied the techniques we learned during the course to build the models and weigh the results of model evaluation indicating that the decision tree performed slightly better than KNN.

During our analysis, we observed recurring patterns in crime counts over months, with summer months exhibiting the highest crime rates. While this pattern did not have a significant impact on the prediction of arrests, it did influence the prediction of the number of crimes in a given day at a specific location. The type of crime also played a significant role in predicting arrests and crime frequencies in Chicago.

These predictive tasks can be further expanded to include more extensive patterns in crime prediction if additional information about the victims and offenders becomes available. This has the potential to enhance the accuracy and effectiveness of the predictive models in crime detection and prevention efforts.

# 5. Limitations and Future Work

## 5.1 Limitation of the analysis

- Data Quality: The dataset we utilized had numerous issues, including missing values, inconsistent data, and duplicate values. As a result of these issues, many of our models we tried struggled to converge i.e reach stable and optimal performance during training. These convergence challenges hindered the model's ability to achieve satisfactory accuracy and reliability in its predictions.
- Imbalanced Data: In our dataset, some crime types were grouped separately, even though they were similar. This led to redundant and overlapping categories that impacted the accuracy and consistency of our models. This could have also led to the loss of information and could have introduced noise in the modeling process.

## 5.2 Future work

- Exploring additional models: In addition to the models mentioned, further investigation can be conducted on other advanced machine learning models, such as Gradient Boosting, Support Vector Machines, or Deep Learning Models. This can help improve the accuracy and performance of the crime prediction system.
- Feature Engineering: Continuously refining and expanding the feature set used in the crime prediction model can enhance its predictive capabilities. Feature engineering techniques such as feature selection, feature extraction, and feature creation can be

applied to identify more relevant and informative features that can contribute to better crime prediction accuracy.

- Fairness and Bias Mitigation: Addressing fairness and bias concerns in crime prediction models is crucial to ensure ethical and responsible use. Techniques such as re-sampling. Reweighting or adversarial training can be employed to mitigate bias in the model's predictions, and fairness-aware evaluation metrics can be used to assess the fairness of the model across different demographic groups.
- Real-time Prediction: Developing a real-time crime prediction system can enable law enforcement agencies to proactively respond to crime incidents and prevent crimes from occurring. Implementing real-time data processing, streaming analytics, and event-driven architectures can enable the model to make predictions in real time, providing timely and actionable insights to law enforcement agencies.
- Crime Hotspot Prediction: Focusing on predicting crime hotspots can help allocate law enforcement resources more effectively. Spatial-temporal data analysis, clustering algorithms, and geospatial techniques can be applied to identify crime hotspots and predict their potential evolution over time, aiding in proactive crime prevention strategies.
- Criminal Network Analysis: Analyzing criminal networks and understanding their structure and dynamics can provide valuable insights for law enforcement agencies. Network analysis techniques such as social network analysis, graph theory, and link analysis can be applied to identify key actors, relationships, and patterns within criminal networks, aiding in identifying and disrupting criminal activities.

- Improve code efficiency: Optimizing code efficiency through techniques such as algorithmic optimization, parallel processing, and model deployment optimizations can enhance the model’s performance and enable it to handle larger datasets or real-time prediction scenarios more effectively.

## 6. References

[1] Sarkar, S. (n.d). Data Mining. *Course Materials.* Retrieved from https://bgsu.instructure.com/courses/1388522/files/100009833/download?download_frd=1

[2] "Master's in Data Science" https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/#:~:text=A%20decision%20tree%20is%20a,that%20contains%20the%20desired%20categorization

[3] Kumar, Gokul S. (2020). Decision Trees: A step-by-step approach to building DTs, 2020. Towards Data Science. Retrieved from Decision Trees: A step-by-step approach to building DTS| by Gokul S Kumar | Towards Data Science

[4] "Chicago Data Portal" https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2