**Crimes in Chicago 2020**

**Prepared for:**

*Prof. Khaled Mohamed, Eng. AlAmir Hassan*

**Prepared by:**

1. *Mahmoud Amgad Tolba - 202001143*
2. *Bahaa Shamoon Atia - 202001572*
3. *Arwa Khaled Noureldin - 202000849*
4. *Silvana Nageh Mansy - 202001817*
5. *Michael Hany Adel - 211001760*

**Contents**

[**Introduction** 2](#_Toc135389386)

[**Objectives** 3](#_Toc135389387)

[**Data Description** 3](#_Toc135389388)

[**Data Cleaning** 7](#_Toc135389389)

[**Exploratory Data Analysis (EDA):** 8](#_Toc135389390)

[**Modelling:** 12](#_Toc135389391)

[**Results:** 12](#_Toc135389392)

[**Conclusion:** 16](#_Toc135389393)

[**Learning outcomes:** 17](#_Toc135389394)

# **Introduction**

Crime has been a prevalent anti-social trait in the human society since time immemorial and it continues to be so even today. Analysing many aspects of crime and identifying criminal trends are highly important. Therefore, having a criminal information system that can analyse a lot of data quickly is essential. Also, one of the most efficient techniques to explore, analyse, find patterns, and forecast future crimes in a vast quantity of data is data analysis by using classification. Therefore, we aimed to work on a dataset that can help us build our model.

# **Problem definition**

Crime is a pressing issue in urban areas, security agencies throughout the world are working hard to lower these crimes, but the volume of crime data is growing quickly, making it challenging to handle such a large amount of data. As we mentioned in the introduction, we aim to build a machine learning model that can help law enforcement agencies, policymakers, and community stakeholders to handle this issue and save much effort and time.

# **Objectives**

Accurate forecasting of crime rates can provide valuable insights for law enforcement agencies, policymakers, and community stakeholders. Therefore, we aim to achieve some objectives that may lead to reduce the crime rate in such areas and assist the stakeholders to handle this issue:

First, we want to present the data that we have as it is and study it well so that we can know what interests us from this raw data and what will help us achieve our main goal. This part will be covered in the section of **data description**.

Second, after displaying the raw data it will be clear to know what interests us and what columns we will need to work on, therefore, we will clean our data and remove the unnecessary parts. This will be covered in the section of **data cleaning.**

Third, the section of **exploratory data analysis** is very important, as we will have some visualisations and correlations which will enable us to know the relations between our data and graph them, and have some useful information that will help us to reach our main goal which is building a machine learning model.

The primary objective of this project is to develop crime rate prediction models for the city of Chicago. These models should enable us to predict crime levels based on various features such as the month, day, hour, and district of the reported crime incidents. This objective will be covered in the **model** section.

# **Data Description**

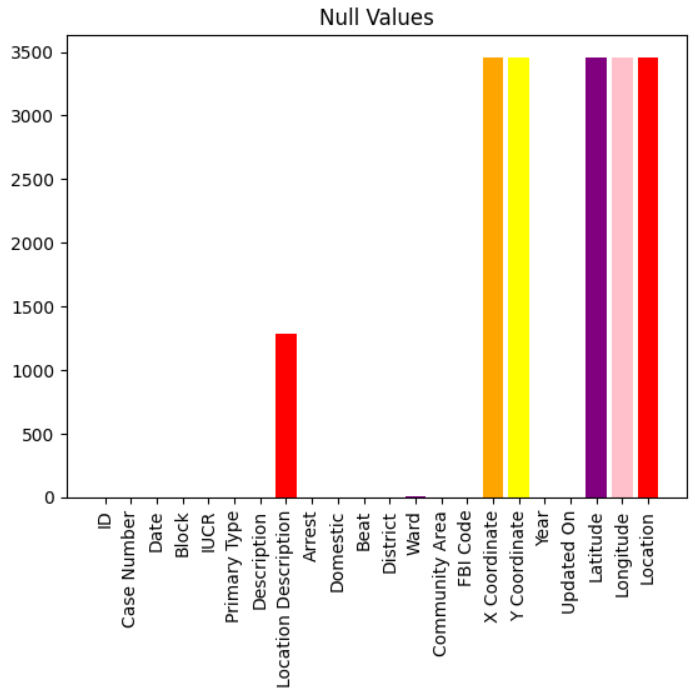
The Chicago Crime 2020 Dataset provided by the Chicago Data Portal and Kaggle website will be used for our project. The Chicago Police Department has registered numerous criminal cases daily since 2001 and has made this data available publicly.

* **Data source:** Chicago crimes in 2020 uploaded from Kaggle: <https://www.kaggle.com/datasets/salikhussaini49/chicago-crimes?select=Chicago_Crimes_2020.csv>

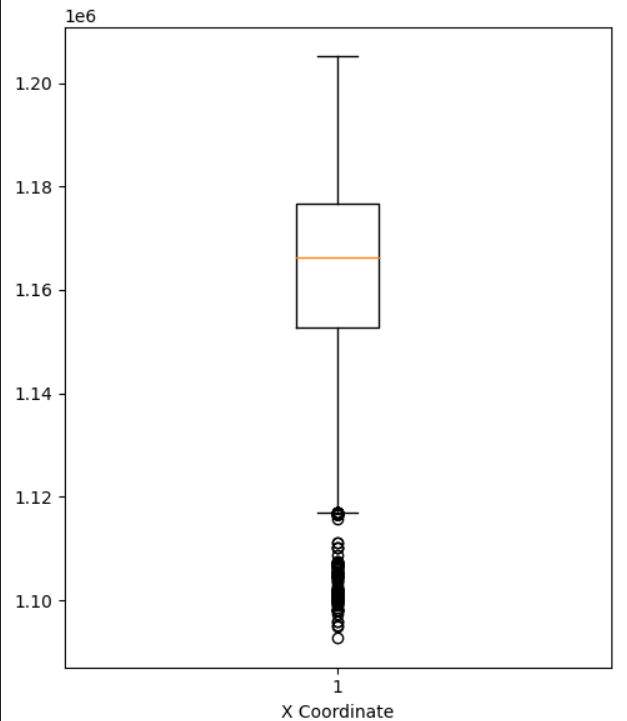
Also it could be found on Chicago police department:

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

* **Size and format:** Our dataset size is 51.66MB as CSV file format, it includes 211922 rows and 22 columns. The data types varies between: (int, float, strings, dates, Time)
* **Data type:** our dataset is mixed between categorical, numerical. There are exactly 12 numerical variables represented in ['ID', 'Arrest', 'Domestic', 'Beat', 'District', 'Ward', 'Community Area', 'X Coordinate', 'Y Coordinate', 'Year', 'Latitude', 'Longitude'] columns, and 10 categorical variables represented in ['Primary Type', 'IUCR', 'Case Number', 'FBI Code', 'Updated On', 'Date', 'Block', 'Description', 'Location', 'Location Description'] columns.
* **Attributes**:
* 1. ID → Unique identifier.
* 2. Case Number: Chicago Police Division Record Number.
* 3. Date: Date of the incident.
* 4. Block: The abbreviated address for criminal activity.
* 5. IUCR: Internal Uniform Chicago Crime Reporting Code.
* 6. Primary Type: Type of crime
* 7. Description: Little more details about the criminal activity
* 8. Location Description: Location where the crime occurred.
* 9. Arrest: Boolean indicating whether the arrest was made.
* 10. Domestic: Indicates whether the incident was domestic-related.
* 11. Beat: Indicates the Beat where the incident occurred.
* 12. District: Indicates Police District where the incident occurred.
* 13. Ward: The City Council District where the incident occurred.
* 14. Community Area: The Community where the incident occurred.
* 15. FBI Code: Indicates the crime classification code.
* 16. X Coordinate: The X coordinates the location of the incident.
* 17. Y Coordinate: The Y coordinate of the location of the incident
* 18. Year: Year the accident occurred.
* 19. Updated On: The date of the last update of the record.
* 20. Latitude: Latitude of the location of the incident
* 21. Longitude: Longitude of the location of the incident.
* 22. Location: Latitude and longitude saved as a tuple.
* **Missing values:** Our data contains total 18578 represented in [location description, X coordinate, Y coordinate, latitude, longitude, location] missing value or null value which represents 0.39847422422143314% of our dataset.



* **Outliers:** there were 820 outliers in our data in the X coordinate, so we calculated them and showed them in plots using the boxplot and scatterplot.



A picture containing text, diagram, plot, number

Description automatically generated

* **Data quality:** our data was retrieved from a reliable source is reliable which is Chicago police department, but this doesn’t prevent having some errors or human mistakes in the data entry. Therefore, we worked on have our data cleaned in the data cleaning section before we perform any process on it.

# **Data Cleaning**

As showed before our data came from a reliable source and it had some data entry errors such as missing values or some outliers, but we handled them, so the graphs and the results in the visualization section was on a cleaned data and we can rely on them:

* **Handling missing values:** Our –data contains total 18578 represented in [location description, X coordinate, Y coordinate, latitude, longitude, location] missing value or null value which represents 0.39847422422143314% of our dataset. As we notice that the percentage of the messing data and null values are so small, so we decided to drop these values, Since, these features are not direct numeric values, we can't use summary statistical functions to fill in the missing values.
* **Handling duplicates:** We tested for the duplicates using python functions and we found no duplicates in our dataset. Every crime has key identifier so this guarantee that there will be no duplicates.
* **Handling outliers:** there were 820 outlier in our data in the X coordinate, so we calculated them and showed them in plots using the boxplot and scatterplot, and using the IQR test (inter quartile range) we dropped them from the dataset and created new cleaned dataset.

After cleaning:

A picture containing text, diagram, plot, map

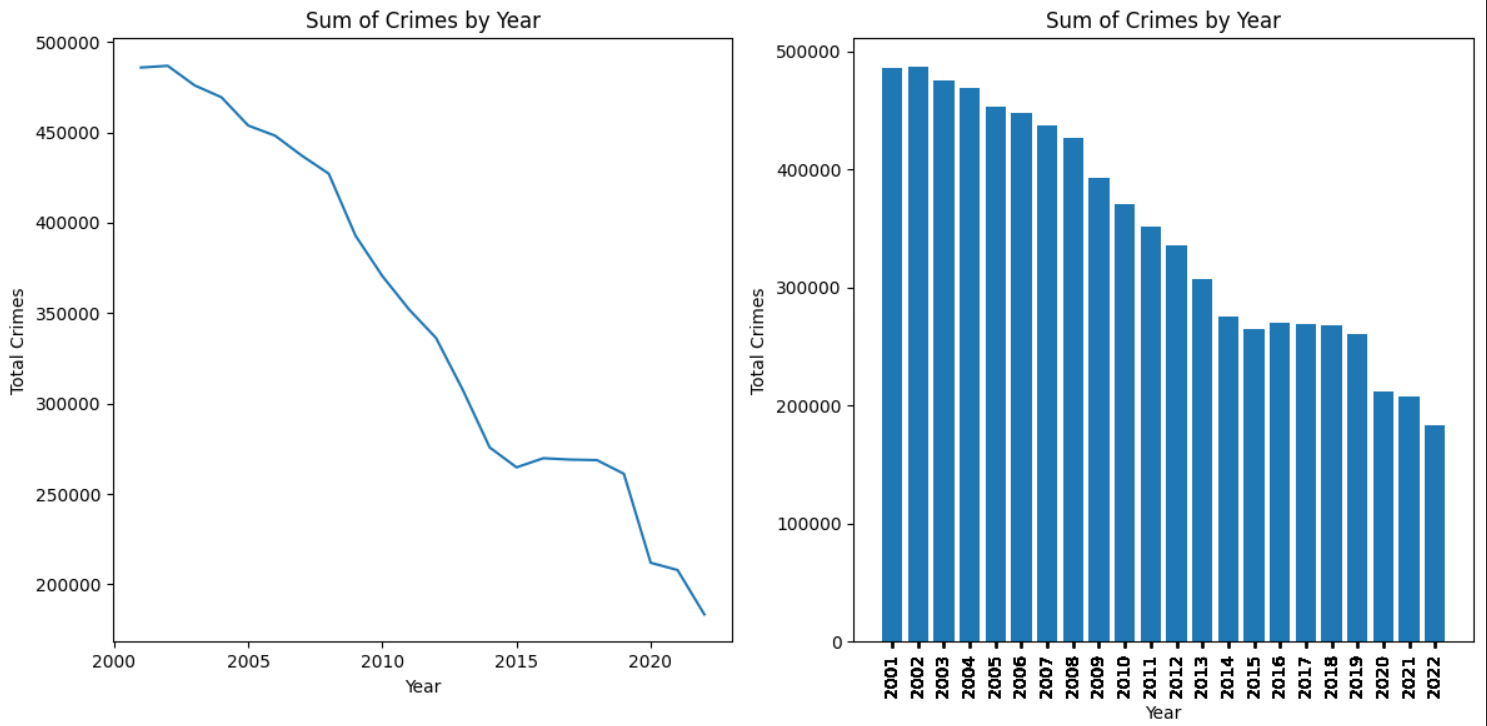
Description automatically generated

* **Data transformation:** In our dataset we had some columns that was string like: Time column we did sliced them and created new columns that include the hours, minuets, seconds separately.
* **Data validation:** To validate our data we had to perform some functions and have our data cleaned from outliers, duplicates, and missing values. Then we used graphs to show the data after cleaning.

# **Exploratory Data Analysis (EDA):**

This part focuses on the exploratory data analysis (EDA). The objective of the EDA is to gain insights into the crime dataset and understand the distribution and trends in crime rates in the city of Chicago over time. In this section, we present a summary of the dataset, including key statistics and visualizations, to provide a comprehensive overview of the crime data. Additionally, we used 2 extra datasets just to explore how the rate of crimes has changed since 2001 until the present. we investigate relationships and correlations between different variables to uncover potential patterns and factors influencing crime rates. By delving into these aspects, we aim to develop a deeper understanding of the dataset and lay the foundation for further analysis and modelling in our research paper.

* **Visualizations:**
  1. This shows how the crime rate decreases since 2001 till now:



* 1. The distribution of crimes per month 2020:

A picture containing text, screenshot, font, number

Description automatically generated

* 1. This plot shows the most frequent emergencies and crimes in Chicago:

A picture containing text, screenshot, font, parallel

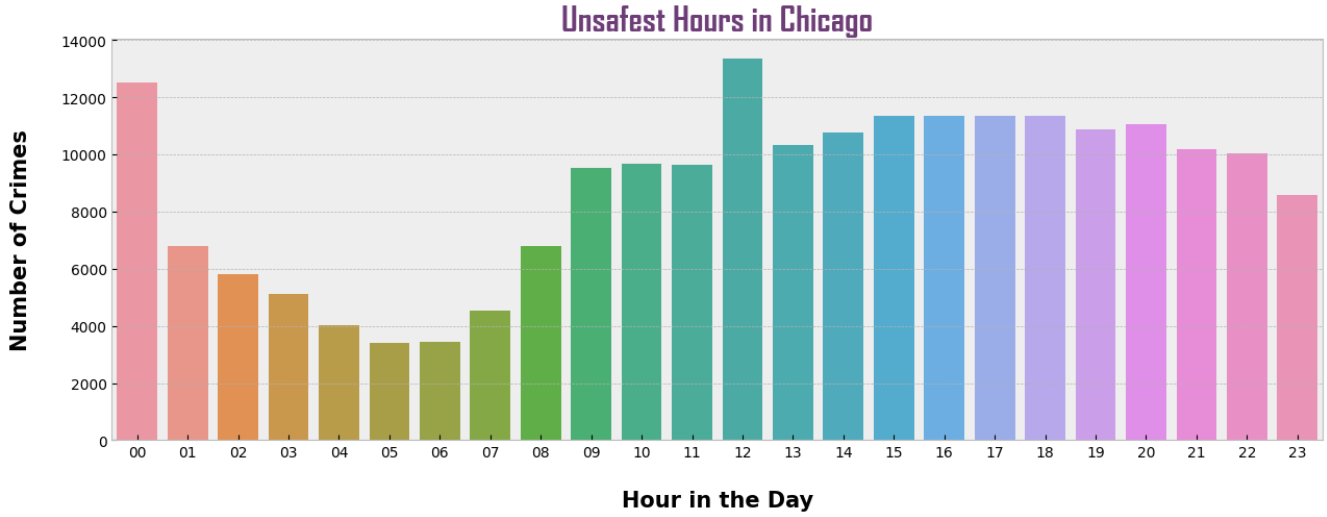
Description automatically generated

* 1. Here are the places that most of the crimes happen in:

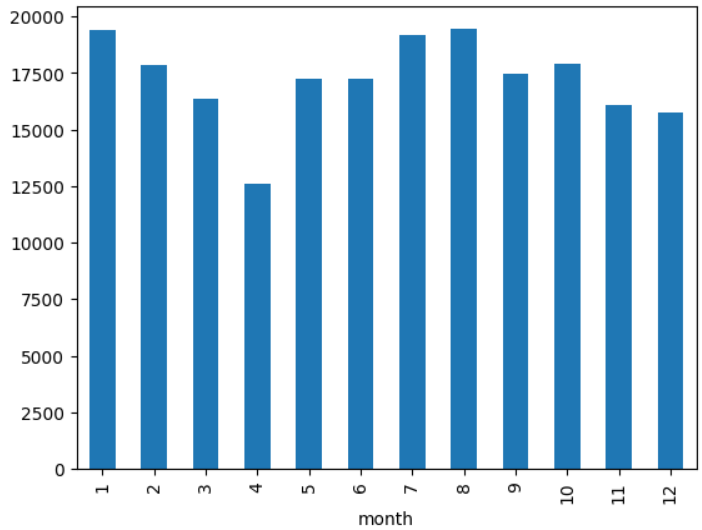
A picture containing text, screenshot, parallel, diagram

Description automatically generated

* 1. This shows the unsafe hours in Chicago:



* 1. This graph shows crimes per month in Chicago 2020:



* 1. Here we show number of crimes per day in the month in Chicago

A picture containing text, screenshot, line, plot

Description automatically generated

* 1. In this plot we were exited to find the top sold drugs in Chicago:

A picture containing text, screenshot, font, design

Description automatically generated

* 1. We wanted to see how crime rate per day was since 2001 till now:

A picture containing text, plot, screenshot, diagram

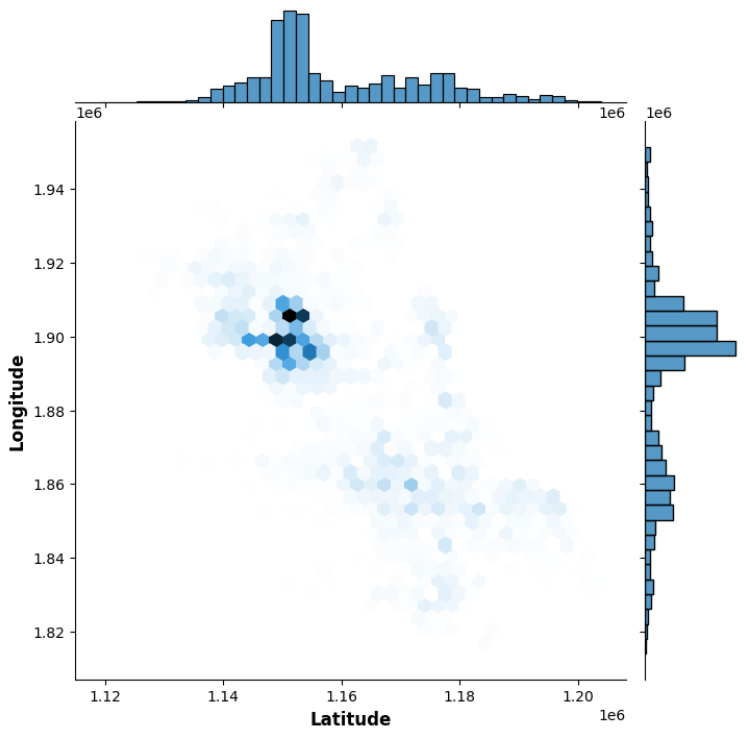
Description automatically generated

* 1. Also we wanted to see how crime rate per month was since 2001 till now:

A picture containing text, plot, diagram, screenshot

Description automatically generated

* 1. This coming plots shows the concentration of any needed crime per location for example(theft and weapon violence):



A picture containing text, screenshot, diagram, map

Description automatically generated

* 1. The coming plot shows the number of arrests for criminals per month in Chicago so we can observe the performance of the police:

A graph of blue and yellow bars

Description automatically generated with low confidence

* **Correlations:** We used heatmaps to show the corelation between specific locations and the frequency for the dangerous crimes like murder or weapon violence.



# **Modelling:**

Coming to the most important section of our project documentation which is the modelling section. We built two predictive models that aims to forecast crime rates in Chicago, and we will have a summary of a comparison between them and see if there is essential difference between them.

The models utilize a decision tree and random forest classifiers to predict crime levels based on various features such as the month, day, hour, and district of the reported crime incidents. The predictive models provide insights into the crime rates in Chicago based on historical crime data from 2020. By considering temporal factors (month, day, hour) and the location (district) of reported crimes, the models predict crime rate levels (low, medium, high) for different time intervals and areas. This information can be valuable for law enforcement agencies, policymakers, and city planners to allocate resources effectively and implement targeted crime prevention strategies.

# **Methodology and results:**

Firstly, the coming points will show the decision tree classifier model specification and details separately.

* **Model methodology:** our model methodology is divided into three levels [ data pre-processing, crime rates analysis, evaluation and prediction] we will talk about every point separately as follows:

1. Data pre-processing:
   * + The initial dataset is filtered to include only relevant columns, namely 'Date', 'Block', 'Location Description', 'District', and 'Primary Type'.
     + The 'Date' column is converted to a datetime format using the time\_convert() function, which extracts the date and time components separately.
     + New columns are created to extract additional temporal information such as the hour, day of the week, and month number from the 'Date' column.
     + The dataset is then grouped by month, day, district, and hour, and the count of crimes is calculated.
2. Crime rate analysis:
   * + The grouped dataset is sorted in descending order based on the district.
     + The average number of crimes per day, district, and hour is computed to understand the overall crime rate.
     + We used the supervised classification approach to assign crime rate levels (0: low, 1: medium, 2: high) based on predefined thresholds.
     + The dataset is modified to include the assigned crime rate levels.
3. Evaluation and prediction:
   * + The distribution of the assigned crime rate levels is analysed to determine the proportion of each category.
     + The decision tree classifier is trained on the modified dataset, with crime rate levels as the target variable.
     + The dataset is split into training and testing sets using a 75:25 ratio.
     + The model's accuracy is evaluated using the accuracy score, and predictions are made on the testing set.
     + A confusion matrix is generated to assess the performance of the model in predicting crime rate levels.

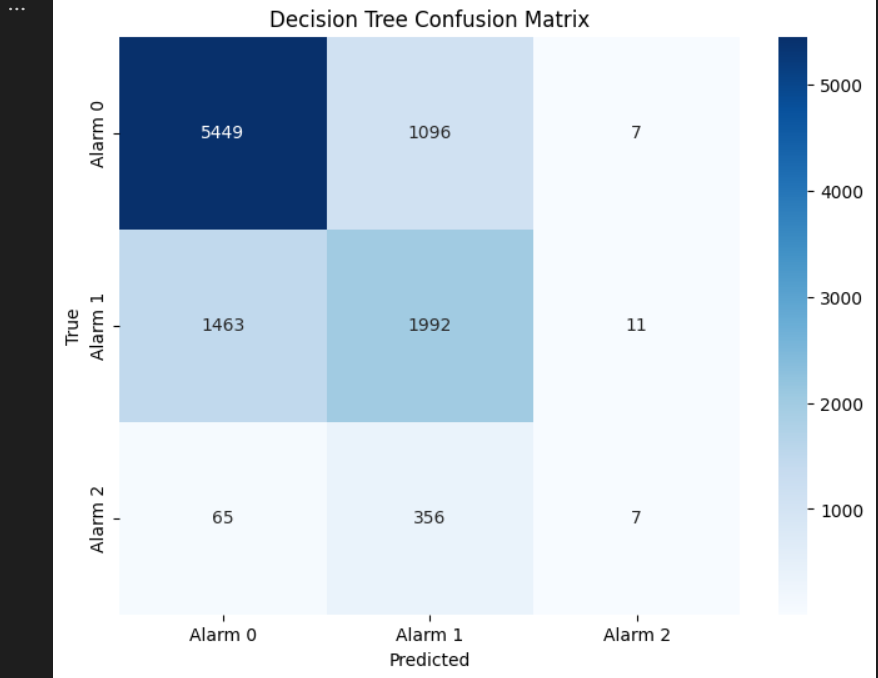
* **Model results:** based on the previous methodology and the process our model make to deal with our data, we can calculate some results and some statistics to evaluate our model. The evaluation includes metrics specific to classification tasks and provides insights into the model's performance.

1. Performance Metrics:
   * + Accuracy: our model’s accuracy depends on some factors for example the maximum depth of the tree; for example, when we let the model work on unlimited maximum depth the accuracy was nearly 63%, but when we changed the maximum depth of the decision tree to 10 the accuracy raised to nearly 71%. (The accuracy metric measures the overall correctness of the predictions, indicating the percentage of correct predictions out of the total number of predictions made.)
     + Precision: Precision is a measure of the model's ability to correctly predict positive instances (high crime rate) out of all instances predicted as positive. It provides an understanding of how precise the model is in identifying high crime rate cases.
     + Recall quantifies the model's ability to identify all positive instances correctly. In this context, it indicates how well the model predicts high crime rate instances.
     + The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance, considering both precision and recall.
   1. Confusion matrix:
      * The confusion matrix provides a detailed breakdown of the predicted and actual crime rate levels. It displays the number of true positives, true negatives, false positives, and false negatives.

A screenshot of a computer

Description automatically generated with medium confidence

* **Visualizations:**
  + - Decision tree heatmap:

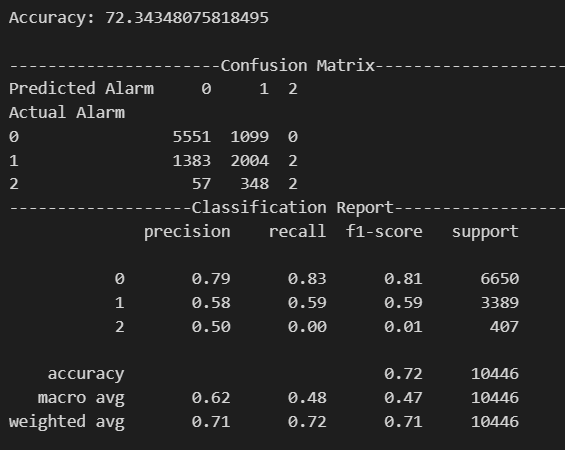


* + - Decision tree graph:

A screenshot of a computer

Description automatically generated

Secondly, the random forest classifier algorithm, it is very likely to the decision tree classifier model, it has the same methodology as the decision tree classifier, but it works with different algorithm. Both are kinds of supervised classification techniques.



And this is the confusion matrix of the random forest algorithm:

A picture containing text, screenshot, diagram, number

Description automatically generated

**Key results:** The random forest algorithm gives more accurate results, but it has higher time complexity, also the accuracy of the random forest algorithm is not that different from the decision tree algorithm; as we saw the accuracy of the decision tree algorithm was 71.36703044227455% , and the random forest algorithm was 72.34348075818495%.

# **Conclusion:**

In conclusion, our project's objectives were covered by analysing historical crime data, performing exploratory data analysis, and creating predictive models. In this project we successfully developed a crime rate prediction model for the city of Chicago. Through the utilization of machine learning techniques, specifically decision tree and random forest classifiers, we were able to construct models that forecasted crime levels based on various temporal and spatial features.

The exploratory data analysis section provided valuable insights into the dataset, including summary statistics and visualizations that depicted the distribution and trends in crime rates over time. we were able to observe how crime rates have changed over the years, contributing to a deeper understanding of the dynamics of crime in Chicago.

Overall, this project not only developed a crime rate prediction model but also provided valuable insights into the dataset, contributing to the body of knowledge in the field of crime analysis. The findings and methodologies employed in this project can serve as a foundation for future research and the development of more comprehensive crime prevention strategies.

# **Learning outcomes:**

It is important at the end of our project to mention some learning outcomes and what did we learned from it.

1. In the part of data acquisition: we learned how to retrieve the data that fits with our objectives and meet our needs.
2. In the part of data description: we improve our knowledge on how to show the data and its attributes, this improved our knowledge of the functions and libraries we used.
3. coming to the data cleaning section: this section improved our understanding to the importance of data cleaning process and how much it is important part in the pre-processing process, as it makes a very huge difference in the results that we get.
4. The exploratory data analysis section was the most important section, as it deepens our understanding to the data and how to extract the important statistics and corelations between its attributes and gave us chance to deepen our knowledge of the plots and graphs and when to use them.
5. In the modelling part we learnt more about machine learning techniques and widen our information about the classification and clustering types.