# Crime Data Analysis and Prediction Team 16:

## Diya Baldota - NUID: 002747966 Veenadharini Shukla - NUID:002704948

## Datasource link: <https://www.kaggle.com/datasets/chaitanyakck/crime-data-from-2020-to-present?resource=download>

## <https://www.kaggle.com/datasets/chaitanyakck/crime-data-from-2020-to-present/download?datasetVersionNumber=11>

## **Content**

* Introduction
  + - Background
    - Motivation
    - Aim
* Methodology
  + - Data Acquisition and Pre-Processing
      * Data acquisition
      * Data exploration and understanding
      * Data Cleaning
      * Removing irrelevant or redundant data
      * Handling Missing Values
      * Data validation and quality assurance
      * Documentation and reproducibility
      * Additional steps for specific columns
    - Exploratory Data Analysis
      * Crime Occurred by Day of the Week
      * Crime occurred over the Years
      * Types of Crimes Committed
      * Gun-Related Crimes over the years
      * Top 20 Crime Occurrences by Premises with most frequent time
      * Word cloud of most frequent terms in Weapon Desc column
      * Victim Sex distribution
      * Age distribution by top 20 crimes
      * Heat Map for area wise distribution
    - Feature Engineering
    - Model Selection
      * Logistic Regression
      * Random Forest Classifier
      * Decision Tree Classifier
      * Gaussian Naive Bayes
    - Time series analysis
      * Prophet Model
      * Arima Model
* Description of the Dataset
* Conclusion
* References

### **Introduction:** **Background:**

### As crime continues to plague many cities around the globe, law enforcement agencies are constantly searching for innovative ways to address this issue. Los Angeles, a city that has experienced a steady increase in various forms of crime, such as theft, assault, and homicide, is at the forefront of this battle. These crimes not only endanger lives and property but also place a heavy burden on the city's resources and public safety infrastructure.

### In response to the growing crime rates, law enforcement agencies have been forced to reevaluate their traditional approaches to crime prevention. Historically, police departments operated in a reactive manner, responding to crimes after they had occurred. While this approach was helpful in apprehending offenders and solving cases, it had minimal impact on preventing future incidents, leaving communities vulnerable to ongoing criminal activity.

### Recognizing the limitations of reactive policing, law enforcement agencies have turned to technology and data-driven methods to improve their crime prevention strategies. One of the most significant advancements in this area has been the development of machine learning algorithms capable of analyzing and categorizing data from past crimes.

This targeted approach not only helps to deter potential offenders but also increases the likelihood of apprehending criminals in the act.

**Motivation:**

The motivation behind this project stems from the pressing need to address crime in Los Angeles and make meaningful contributions towards creating safer and more secure communities. Crime has significant social, economic, and psychological impacts on individuals, families, and neighborhoods. It not only poses threats to public safety but also affects the overall well-being and quality of life of residents.

Traditional approaches to crime prevention and reduction have relied on reactive measures such as increased law enforcement presence and stricter sentencing policies. However, these approaches have limitations in effectively addressing crime and its underlying causes. This is where the power of machine learning and data analysis can make a difference.

Ultimately, our motivation is driven by the desire to contribute to the ongoing efforts of creating safer and more secure communities for all residents of Los Angeles. We believe that our project has the potential to make a positive impact and contribute to the overall well-being and safety of the community.

**Aim:**

Analyzing the LA Crime dataset: We will thoroughly analyze the LA Crime dataset to gain a comprehensive understanding of crime patterns and trends in Los Angeles. This will involve exploring the dataset to identify the types of crimes that are most prevalent in different areas of the city, examining the distribution of crimes across different time periods, and identifying any notable patterns or trends in the data.

Identifying underlying factors contributing to crime: We will conduct in-depth data analysis to identify the underlying factors that contribute to crime in Los Angeles. This may include examining socio-economic, demographic, and environmental factors that have been shown to be correlated with crime rates. By understanding these underlying factors, we aim to provide insights that can help law enforcement agencies and policymakers develop more effective strategies for crime prevention and intervention.

Predicting likelihood of crimes occurrence: Using machine learning techniques, we aim to develop predictive models that can estimate the likelihood of certain crimes occurring in different areas. This can provide valuable information for law enforcement agencies to allocate resources effectively, such as deploying patrols or implementing targeted interventions during high-risk periods and areas, in order to prevent crimes and increase public safety.

Time series analysis for crime trend prediction: We have utilize time series analysis techniques to predict crime trends over time. By analyzing historical crime data, we aim to identify any recurring patterns or trends in crime occurrences, such as seasonal variations or long-term trends. These predictions can assist law enforcement agencies and policymakers in their strategic planning and resource allocation for crime prevention and reduction efforts.

Overall, our aim is to provide data-driven insights that can support evidence-based decision-making for crime prevention and intervention in Los Angeles. By leveraging the power of data science and machine learning,.

### **Methodology:**

* **Data acquisition**: We downloaded the LA Crime dataset from Kaggle, which included two csv files named:

1. Crime\_Data\_from\_2010\_to\_2019
2. Crime\_Data\_from\_2020\_to\_Present

We have combined these files into a single dataframe called crime\_data and saved it in a csv file - Crime\_Data.

* Table

  Description automatically generatedA picture containing text, receipt, screenshot

  Description automatically generated**Data exploration and understanding:** We performed an initial exploration of the dataset to understand its structure and contents. This involves checking the dimensions of the dataset, examining the data types and formats of each variable, and gaining insights into the meaning and interpretation of each attribute.

We checked the shape of the data and understood that there are 2444416 rows and 29 columns. The 29 columns are listed such as DR\_NO, Date Rptd, DATE OCC, TIME OCC, AREA, AREA NAME, Rpt Dist No, Part 1-2, Crm Cd, Crm Cd Desc, Mocodes, Vict Age, Vict Sex, Vict Descent, Premis Cd, Premis Desc, Weapon Used Cd, Weapon Desc, Status, Status Desc, Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4, LOCATION, Cross Street, LAT, LON, and AREA.

Text

Description automatically generated with medium confidence

* **Data Cleaning:** We found out the null values counts in the columns for data cleaning purposes.

Table

Description automatically generated

* **Removing irrelevant or redundant data:** We identified and removed the columns that were duplicated – For e.g. AREA column was repeated twice in the dataset. Hence kept only relevant columns for further analysis.



* **Handling missing values:** We will assess the extent of missing values in the dataset for columns such as AREA, Mocodes, Vict Sex, Vict Descent, Premis Cd, Premis Desc, Weapon Used Cd, Weapon Desc, Status, Crm Cd 1, Crm Cd 2, Crm Cd 3, and Crm Cd 4.

We will apply appropriate techniques to handle missing values, such as imputation or deletion, depending on the nature and amount of missing data.

1. **Area and Area Name:** For the ‘Area’ and ‘Area Name’ columns, we first created a dictionary mapping each unique Area value to its corresponding Area No value. Then we filled missing values in the ‘Area’ column of ‘crime\_data’ using the fillna method. The map method is used to map ‘Area Name’ values to their corresponding ‘Area’ values in the area\_dict dictionary, effectively filling the missing ‘Area’ values based on the ‘Area Name’ values.

Graphical user interface, text, application

Description automatically generated

1. **Weapon Used Cd:** We filled the null values in this column with ‘0’ since these were codes and in numerical format. Hence indicating ‘0’ for ‘No Weapon’ value in Weapon Desc. The identical no. of nulls in both the columns indicated this.
2. **Weapon Desc:** We filled the null values in this column with ‘No Weapon’.

Graphical user interface, text, application

Description automatically generated

1. **Cross Street:** We filled the null values in this column with ‘Unknown’ since it was a geographical field which wasn’t mentioned.

Graphical user interface, text

Description automatically generated

1. **Crm Cd 1:** We created a dictionary that maps unique "Crm Cd Desc" values to their corresponding "Crm Cd 1" values in a DataFrame called "crime\_data". We then filled missing values in the "Crm Cd 1" column of "crime\_data" using the "Crm Cd Desc" values and the created dictionary. Finally, we checked for any remaining null (missing) values in the "Crm Cd 1" column using the isnull method and the sum method. This helps to ensure that all missing values in the "Crm Cd 1" column had been successfully filled using the mapping in the created dictionary.

Graphical user interface, text, application

Description automatically generated

1. **Crm Cd 2:** We filled the missing values with ‘0’
2. **Crm Cd 3:** We filled the missing values with ‘0’
3. Text

   Description automatically generated**Crm Cd 4:** We filled the missing values with ‘0’
4. **Status:** We tried to create a Dictionary for Status and Status Desc. However there were only “UNK” values of Status Desc mapped for the missing values in Status column. Hence we filled it with “NA”

Graphical user interface, text, application

Description automatically generated

1. Graphical user interface, application

   Description automatically generated**Vict Descent:** For this column, we had a character ‘-‘ in the values. Hence we first replaced it with Nan and then filled all the null values with ‘U’ which indicates Unknown Descent.
2. **Vict Sex:** For this column we replaced the ‘-‘ and ‘H’ values with ‘X’ which indicates Unknown and also the null values with the same.

A picture containing graphical user interface

Description automatically generated

1. Graphical user interface, text, application

   Description automatically generated**Mocodes:** We replaced the null values with ‘0’
2. **Premis Cd and Premis Descp:** We filled missing values in the "Premis Cd" and "Premis Desc" columns in a DataFrame called "crime\_data" using a dictionary mapping unique "Premis Desc" values to their corresponding "Premis Cd" values. We then checked for remaining missing values in both columns and filled them with a default value of 'Unknown'.

Graphical user interface, text, application, email

Description automatically generated

After cleaning the data, we did a final check on the null count values and they were all 0 indicating all the null values had been removed.

Table

Description automatically generated

* **Data validation and quality assurance:** We validated the Vict Age column to make sure that invalid dates and negative dates are accounted for. We found no such ages in the data.

Graphical user interface, text, application

Description automatically generated

We checked if our Dates are in the DateTime format and also further formatted the TIME OCC values to 24 hour format (HH:MM)

* **Documentation and reproducibility:** We have thoroughly documented all the steps taken during the data preprocessing and cleaning process, including any assumptions, decisions, or transformations made. This will ensure the reproducibility of our analysis and provide transparency for future reference or validation.
* **Additional steps for specific columns:**For columns with categorical variables, such as Vict Sex, Vict Descent we carefully handled any discrepancies or inconsistencies in the data, such as standardizing and understanding the naming conventions or encoding categorical variables into numerical representations for further analysis.

We also extracted the Day OCC, Month OCC and Year OCC columns from the Date OCC column for doing Visualizations. We extracted similar columns from the Date Rptd.

By following these steps, we tried to ensure that the LA Crime dataset is thoroughly cleaned and preprocessed, providing reliable and accurate data for our analysis and helping us derive meaningful insights and patterns in crime in Los Angeles.

**Exploratory data analysis:**

1. Crime Occurred by Day of the Week:Chart, bar chart

   Description automatically generated

The code generates a bar chart that displays the number of crimes reported for each day of the week. The x-axis represents the days of the week, such as Monday, Tuesday, etc., while the y-axis represents the counts of crimes reported.

Through this visualization, we can gain insights into the distribution of crimes reported by day of the week, helping us identify which days may have higher or lower crime occurrences. This information can be valuable for resource allocation, scheduling patrols, and identifying potential patterns or trends in crime reporting.

1. Crime occurred over the Years:Chart, line chart

   Description automatically generated

Created a line plot using seaborn's lineplot() function to show the trend of crime occurrences over the years, with years on the x-axis and the number of crimes on the y-axis.Through this visualization, we can observe the trend of crime occurrences over the years, which can help us identify patterns, trends, or anomalies in crime data.

This information can be useful for understanding the overall crime situation and making informed decisions related to law enforcement, crime prevention, and resource allocation.

1. Types of Crimes Committed:

Chart

Description automatically generated

Created a vertical bar chart using Seaborn's barplot() function to visualize the top 20 types of crimes by count from the 'crime\_data' DataFrame.

This visualization helps to understand the distribution of different types of crimes based on their counts. It provides insights into the top 20 common types of crimes committed, as well as their relative frequencies. This information can be useful for law enforcement agencies, policymakers, and other stakeholders to prioritize resources, target interventions, and make informed decisions related to crime prevention, public safety, and resource allocation.

1. Gun-Related Crimes over the years:

Chart, line chart

Description automatically generated

For this visualization, we first extracted the gun\_crimes, i.e all the rows where-in there is a word ‘gun’ in the weapon desc field. Then we tried to map it’s occurrence over the years. We have further analyzed that the gun related crimes reached it’s lowest in 2013 and also declined during the post covid period.

So here we have identified peaks or declines in gun-related crimes over the years, which can be useful for understanding the effectiveness of measures or interventions related to gun control and crime prevention.

1. Top 20 Crime Occurrences by Premises with most frequent time:

Chart, histogram

Description automatically generated

Here we tried to analyse the top 20 premises with the highest crime count and further also mapped which were the most peak times at which the crimes occurred there.

This visualization enables comparison of the most frequent time of day for crimes across different premises, which can be useful for understanding temporal patterns of crime activity and planning targeted interventions.

1. Word cloud of most frequent terms in Weapon Desc column

A picture containing logo

Description automatically generated

Here we have provided a visual representation of the most frequent terms in the 'Weapon Desc' column of the 'crime\_data' DataFrame, allowing for quick identification of commonly occurring weapon types.

This visual helps in understanding the types of weapons involved in the reported crimes and their frequency, which can provide insights into crime trends and patterns.

Can be used for exploratory data analysis and visualization of weapon-related information in the 'crime\_data' DataFrame to support decision-making, resource allocation, and law enforcement strategies related to weapons used in crimes.

1. Victim Sex distribution:

Chart, pie chart

Description automatically generated

We have created a pie chart visualization to display the distribution of the 'Vict Sex' feature in the 'crime\_data' DataFrame. The pie chart is created using the 'plot.pie()' method of the 'value\_counts()' output for the 'Vict Sex' feature, with additional parameters for autopct, startangle, title, axis, and legend customization.

Helps in understanding the gender distribution of crime victims, which can provide insights into patterns or trends related to victimization based on gender. We have plotted Male(M), Female(F), Non-Binary (N) as well as the Unknowns (X).

1. Age distribution by top 20 crimes:

Chart

Description automatically generated

We have provided a visual representation of the distribution of victim age for the top 20 crime categories in the 'crime\_data' DataFrame using a box plot.

This helps in identifying potential age-related patterns or trends in crime victimization for the top 20 crime categories.

It allows for comparison of the distribution of victim age across different crime categories, which can provide insights into potential age-based differences in crime types.

1. Map

   Description automatically generatedHeat Map for area wise distribution

* Here we have provided a visual representation of crime data on a Folium map for the Los Angeles area.
* The heat map layer represents the density of crimes based on the coordinates (latitude and longitude). The higher density areas shown with darker colors.
* The cluster layer groups nearby crimes into clusters, and clicking on a cluster displays a pop-up message with the crime description ('Crm Cd Desc') for each individual crime in that cluster.
* The layer control allows users to toggle the visibility of different layers (heat map and cluster) on the map, providing flexibility in exploring the crime data in different ways.

Useful for visualizing the spatial distribution of crime data and identifying patterns or trends in crime hotspots or clusters in the Los Angeles area, which can be helpful for crime analysis, resource allocation, and decision-making.

**Feature engineering:**

Chart

Description automatically generated

Correlation Matrix

* We selected the features based on the correlation matrix, by running it with ‘AREA ’ as the target matrix.
* Based on the output we created a new dataframe ‘crime\_data\_area’.
* We then performed feature engineering on the 'crime\_data\_area' dataset.
* Then we defined a function called 'map\_time\_slot' that maps a given time range to a numerical code and made a column ‘Time Slot’ from it.
* The 'Time Slot' column is added to the 'crime\_data\_area' dataframe based on the hour and minute of the 'Time' column, mapped to numerical codes using the 'map\_time\_slot' function.
* Standardizing data types and formats: We have ensured that the data types and formats of each variable are consistent and appropriate for their respective meanings and interpretations. This involved converting and checking the data types of the columns DATE RPTD, DATE OCC and TIME OCC standardizing date formats, or transforming categorical variables into appropriate numerical representations.
* Handling outliers or anomalies: We identified and addressed any outliers or anomalies in the dataset. This involve using statistical techniques or domain knowledge to detect and correct or remove outliers from the dataset.

**Model Selection and Evaluation:**

**1. Logistic Regression:**

****

* The code defines a logistic regression model using the 'LogisticRegression' function from the 'sklearn.linear\_model' module
* The model is trained on the training data using the 'fit' method
* The model is used to make predictions on the testing data using the 'predict' method
* The accuracy of the model is evaluated using the 'accuracy\_score' function from the 'sklearn.metrics' module
* Logistic regression is a simple and efficient model that can be used for binary classification problems

1. **Random Forest Classifier:**

**Graphical user interface, text, application

Description automatically generated**

* The code defines a random forest classifier model using the 'RandomForestClassifier' function from the 'sklearn.ensemble' module
* The model is trained and evaluated using 10-fold cross-validation using the 'cross\_val\_score' function from the 'sklearn.model\_selection' module
* Random forest is a powerful model that can handle both numerical and categorical data, and can also handle missing values and outliers well.

1. **Decision Tree Classifier:**

**Graphical user interface, text, application

Description automatically generated**

* The code defines a decision tree classifier model using the 'DecisionTreeClassifier' function from the 'sklearn. tree' module
* The model is trained and evaluated using 10-fold cross-validation using the 'cross\_val\_score' function from the 'sklearn. model\_selection' module
* Decision tree is a simple and interpretable model that can handle both numerical and categorical data and can also handle missing values well.

1. **Gaussian Naive Bayes:**

**Graphical user interface, text, application

Description automatically generated**

* The code defines a Gaussian Naive Bayes classifier model using the 'GaussianNB' function from the 'sklearn. naive\_bayes' module
* The model is trained on the training data using the 'fit' method.
* The model is used to make predictions on the testing data using the 'predict' method.
* The accuracy of the model is evaluated using the 'accuracy\_score' function from the 'sklearn. metrics' module
* Gaussian Naive Bayes is a simple and efficient model that can handle both numerical and categorical data, and can also handle missing values well.

Graphical user interface, application

Description automatically generated

Therefore, the logistic regression model has the highest accuracy score, followed by the Gaussian Naive Bayes model, the average accuracy score with Gaussian Naive Bayes, and finally the average accuracy score with DecisionTreeClassifier.

Based on the accuracy scores alone, the Logistic Regression model appears to be the best model for your analysis, as it has the highest accuracy score of 0.972.

## **Time series analysis:**

### Prophet Model

Chart

Description automatically generated

* We used Facebook Prophet, a time series forecasting tool, to make predictions on the number of crimes occurring in Los Angeles over time.
* The first step is to group the data by the date of occurrence and count the number of crimes for each date. This creates a dataframe with two columns: 'ds' for date and 'y' for the count of crimes.
* Next, the Prophet model is initialized and fit to the data.
* The resulting plot shows the predicted values for the count of crimes over time, along with upper and lower confidence intervals. The x-axis represents time, and the y-axis represents the predicted count of crimes.

### Graphical user interface Description automatically generatedGraphical user interface Description automatically generatedArima model

* Here we have conducted a time series analysis on crime data using the ARIMA (Autoregressive Integrated Moving Average) model.
* The first step is to group the crime data by the date of occurrence and count the number of crimes for each date.
* Next, the column names are changed to 'ds' and 'y'. This is because the ARIMA model expects the time series data to have these column names.
* Then, an ARIMA model is fitted to the time series data using the statsmodels package. The order of the ARIMA model is set to (1, 1, 1), which means that the model includes one autoregressive term, one differencing term, and one moving average term.
* We then predict the crime for the next year.
* The first plot shows both the actual crime counts and the forecasted values. The second plot shows only the actual crime counts.
* The predictions for the future are based on the fitted ARIMA model, which takes into account the historical data and patterns in the time series.

### **Description of the Dataset:**

The LA Crime dataset contains crime data from 2010 to 2019, as well as data from 2020 to the present. The dataset includes information such as the type of crime, the location of the crime, and the date and time of the crime. It is a large and comprehensive dataset that provides valuable insights into crime trends in Los Angeles. The size of our dataset is 619.48MB.

These are some relevant columns from the dataset :

1. DR Number: A unique identification number assigned to each crime report.
2. Date Reported: The date that the crime was reported to law enforcement.
3. Date Occurred: The date that the crime occurred.
4. Time Occurred: The time that the crime occurred.
5. Area ID: The identification number of the LAPD patrol area in which the crime occurred.
6. Area Name: The name of the LAPD patrol area in which the crime occurred.
7. Reporting District: The identification number of the LAPD reporting district in which the crime occurred.
8. Crime Code: A code that identifies the type of crime that was committed.
9. Crime Code Description: A description of the type of crime that was committed.
10. MO Codes: Modus operandi codes, which are used to classify the method used by the perpetrator in committing the crime.
11. Victim Age: The age of the victim.
12. Victim Sex: The gender of the victim.
13. Victim Descent: The race or ethnicity of the victim.
14. Premise Code: A code that identifies the type of location where the crime occurred.
15. Premise Description: A description of the type of location where the crime occurred.
16. Weapon Used Code: A code that identifies the type of weapon used in the crime.
17. Weapon Description: A description of the type of weapon used in the crime.
18. Status Code: A code that identifies the current status of the investigation.
19. Status Description: A description of the current status of the investigation.
20. Location: The exact location where the crime occured.

**Conclusion:**

In summary, the code uses four different machine learning models to predict the geographic area of the crime. Each model is trained and evaluated using different methods and hyperparameters. The models used are logistic regression, random forest classifier, decision tree classifier, and Gaussian Naive Bayes classifier. These models have their own strengths and weaknesses and may be more suitable for different types of datasets and problems. It is important to evaluate multiple models and choose the best one based on their performance on different evaluation metrics.

Based on the accuracy scores alone, the Logistic Regression model appears to be the best model for your analysis, as it has the highest accuracy score of 0.972.

**References:**

<https://www.kaggle.com/datasets/chaitanyakck/crime-data-from-2020-to-present?resource=download>

<https://data.lacity.org/d/63jg-8b9z/visualization>

<https://github.com/SlicerBX/crime-data-analysis-and-prediction-in-LA/blob/main/Crime_Prediction_and_Classification_in_LA_Master.ipynb>

<https://stackoverflow.com/questions/48085110/no-module-named-folium-plugins-python-3-6>

<https://www.researchgate.net/publication/224346385_Forecasting_crime_using_the_ARIMA_model>

### 