## Capstone-Project-4: Exploring Crime Analysis with LAPD Leveraging Machine Learning for Public Safety

PROJECT REPORT

**Data Science with Python Programming**

**By**

**Team -4**

**INDUSTRIAL PROJECT BASED LEARNING**



**Department of Computer Science and Engineering**

**Accredited by NBA**

**Geethanjali College of Engineering and Technology**

**(UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi)

Cheeryal (V), Keesara (M), Medchal.Dist.-501 301.

**Team Details :**

**Team Leader:**

Shivanoor Vignesh 21R11A0597

**Team Members:**

G.Sandhya Rani 21R11A0569

R.Dhathri 21R11A0545

Hasini 21R11A0518

A.Pooja 21R11A0501

PSBN Sriya 21R11A0595

N Sai Charan 21R11A0588

MD.Akbar Ali 21R11A0586

# ABSTRACT

Crime is a complex phenomenon that affects communities worldwide, posing significant challenges for law enforcement agencies and policymakers. The Los Angeles Police Department (LAPD) often faces the task of predicting crime before it happens to better safeguard the community. The LAPD has a historical record of over a million incidences. Statistical analysis of some of these incidences can aid the LAPD in crime prevention by identifying patterns and trends. This capstone project delves into an in-depth analysis of LAPD crime data from 2020 to 2024, utilizing machine learning techniques to extract actionable insights for public safety enhancement in Los Angeles. The study comprehensively investigates temporal patterns, spatial distributions, crime type relationships, victim demographics, severity analysis, case statuses, and location details. Geospatial analysis is employed to visualize crime hotspots, enabling precise interventions and resource allocation. Ultimately, this endeavor contributes to bolstering public safety efforts and cultivating a safer environment across Los Angeles.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S No** | **Title** | **Page No** |
|  | **ABSTRACT** | 2 |
| **1** | **INTRODUCTION** | 4 |
| **2** | **PROBLEM STATEMENT** | 5 |
| **3** | **OBJECTIVES** | 5 |
| **4** | **Literature Survey** | 6 |
| **5** | **METHODOLOGY** | 7 |
|  | **5.1 Data Source** | 7 |
|  | **5.2 Data Pre processing**  5.2.1. Missing entries  5.2.2. Removal of Redundant Columns  5.2.3. Replacing the null values with respective suitable values  5.2.4. Date and Time Data Processing  5.2.5. Creation of Mapping Dictionaries | 9 |
|  | **5.3 Exploratory Data Analysis** | 14 |
|  | **5.4. Comparison of Various Models**  5.4.1. Gradient Boosting Classifier Model  5.4.1.1. Gradient Boosting Classifier Model with random state  5.4.2. Random Forest Classifier Model  5.4.3. Support Vector Machine (SVM) Classifier Model  5.4.4. Random Forest Classifier Model for Hotspots for crime | 24 |
|  | **5.5. Pickle File Creation** | 28 |
|  | **5.6. Flask setting up** | 29 |
| **6** | **Final Implementation** | 30 |
| **7** | **Conclusion** | 35 |
| **8** | **Future Scope** | 35 |

# INTRODUCTION

Crime has always been an existing factor in every city but is an even bigger issue in larger cities such as Los Angeles. Crimes are the most serious threat to the human society and is increasing now-a-days in most of the cities in various parts of the world. Data mining algorithms plays an important role in predicting the number and type of crime events likely to take place in the future. Also, various other technologies and hi-tech methods may help Law Enforcement agencies to track the pattern of crimes and predict the probability of crime incidents. Though predictions cannot be 100% accurate, yet the probability for its occurrence can be detected.

The primary goal of the Los Angeles Police Department (LAPD) is to safeguard the lives and properties of the Los Angeles community, with a key objective of reducing crime rates each year. However, understanding the limits of crime reduction and identifying the factors that influence crime occurrence remains a significant challenge. This capstone project aims to analyze LAPD crime data from 2020 to 2024 to uncover actionable insights that can support law enforcement agencies and policymakers in their efforts to combat crime and improve community well-being.

The study will focus on various key variables, including the Department Report Number (DR\_NO), the date and time of crime occurrence and reporting, geographical location, crime type, modus operandi, victim demographics, and weapon usage. By examining these variables, we aim to identify trends, correlations, and factors influencing crime occurrence in Los Angeles. This dataset serves as a valuable resource for analyzing crime patterns, victim demographics, and spatial-temporal trends, thereby informing targeted interventions and resource allocation strategies to enhance public safety.



Specifically, the analysis will focus on temporal patterns, spatial distributions, and crime type relationships. Temporal analysis will uncover seasonal variations and daily/weekly patterns in crime occurrence, while spatial analysis will identify high-crime areas and potential contributing factors. Crime type relationships will help understand the interconnectedness of different types of criminal activity.

Through this comprehensive analysis, this project seeks to contribute to evidence-based decision-making processes aimed at enhancing crime prevention strategies and fostering a safer and more secure environment for all residents of Los Angeles. By identifying crime patterns, predictors, and relationships, we hope to provide actionable insights that can help the LAPD better prepare for and respond to criminal activity, ultimately improving public safety and community well-being.

# LITERATURE SURVEY

# PROBLEM STATEMENT

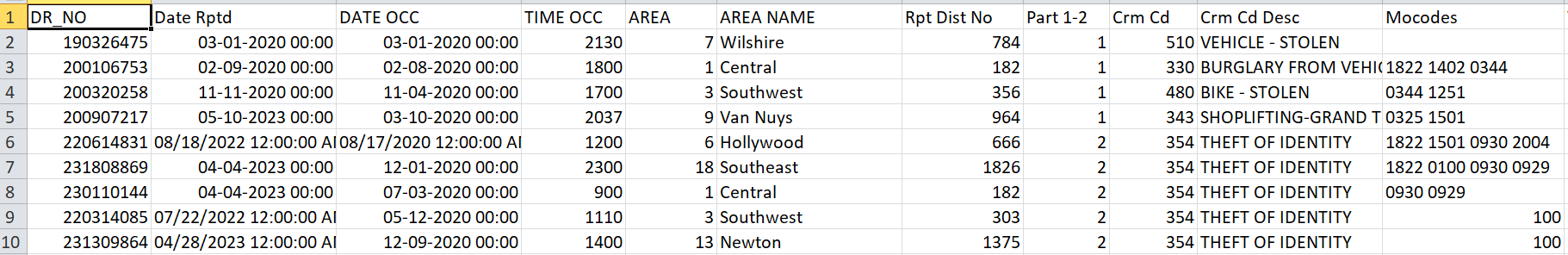
Crime in Los Angeles poses a significant threat to community safety and well-being, impacting the quality of life for residents. Despite efforts by law enforcement agencies, understanding the intricate dynamics of crime remains a challenge. Crime is a multifaceted challenge that presents significant obstacles for law enforcement agencies and policymakers worldwide. In Los Angeles, the Los Angeles Police Department (LAPD) grapples with the daunting task of predicting and preventing crime to safeguard the community effectively.

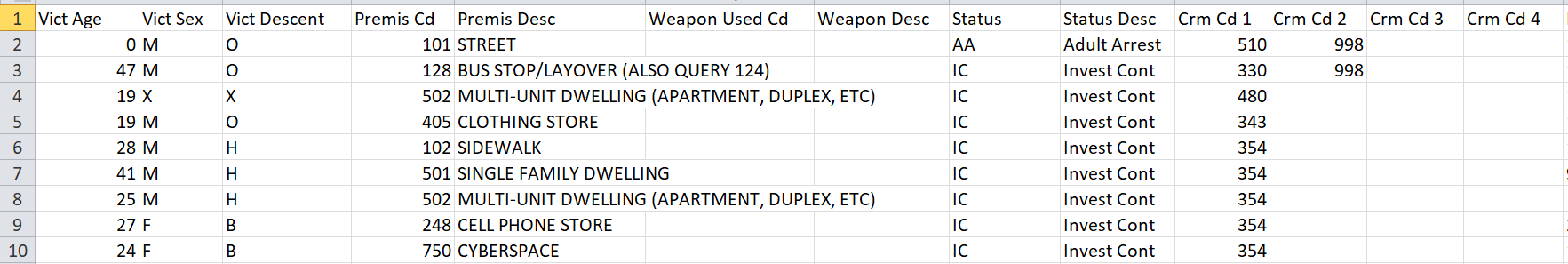
# 4.OBJECTIVES

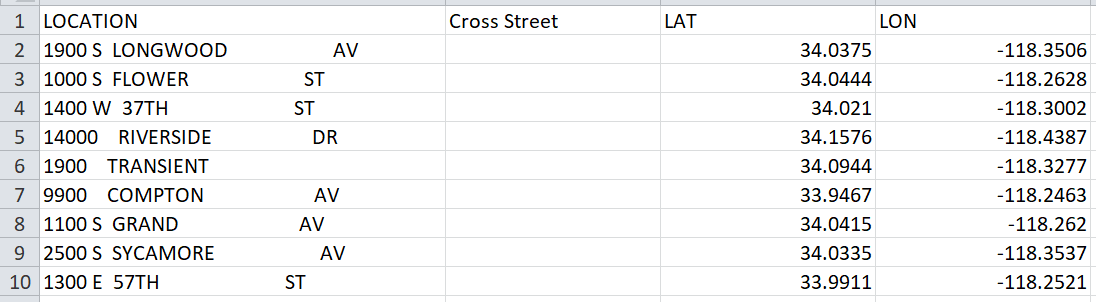
* To uncover insights into crime patterns, victim demographics, spatial-temporal trends, and factors influencing crime severity.
* To provide actionable insights for law enforcement agencies and policymakers to enhance crime prevention strategies and improve public safety in Los Angeles.
* Analyze crime trends over time: Identify seasonal variations (e.g., property crimes might increase during summer) or daily/weekly patterns.
* Explore spatial patterns: Investigate high-crime areas and identify potential factors (e.g., demographics, socio-economic indicators).
* Understand crime type relationships: Analyze how different crime types might be linked or occur together.

**5.METHODOLOGY**

**5.1 Data Source**







This data is extracted from a US Government website Data.gov.

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

The daily crime occurrence in Los Angeles state is recorded and updated on the website starting from 2020 till the latest.

Columns/ Variables - 28

Rows/ Observations – 925721

* **Brief description of the data source**

1. DR\_NO: The Department Report Number is represented as DR\_NO which is generated for every crime that is reported to the Los Angeles department.
2. Date Rptd: The date when the crime was reported to the police department is recorded as Date Rptd and the date lies between 2020 to 2024.
3. DATE OCC: The actual date when the crime occurred is recorded under this column.
4. TIME OCC: The time the crime occurred is represented in a 24-hour format.
5. AREA: The Los Angeles Police Department has 21 community police stations within the geographical location sequencing from 1-21.
6. AREA NAME: The 21 divisions also have an area name in Los Angeles
7. Rprt Dist. No: Code that represents a sub-area within a Geographic Area this is usually prefixed by area code
8. Part code: Indicate whether the crime is serious or less offensive.
9. Crm Cd: Indicates the crime committed.
10. Crm Cd Desc: Defines the Crime Code provided.
11. Mocodes: Modus Operandi code provides additional details about the crime.
12. Vict Age: Indicates the age of the victim.
13. Vict Sex: F: Female M: Male X: Unknown
14. Vict Descent: Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian
15. Premise Cd: Type of structure the crime took place in (vehicle, building, parking lot, etc.)
16. Premise Desc: Describes premise Cd.
17. Weapon Used Cd: Code for Type of weapon used in crime.
18. Weapon Desc: Description of the weapon.
19. Status: Status code of the case
20. Status Desc: Status of the case description
21. Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4: Additional status codes associated with the case.
22. LOCATION: Crime location
23. Cross Street: Cross street from the crime location
24. LAT: The latitude of the location is recorded under this column.
25. LON: The longitude of the location is recorded under this column

**5.2 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) serves as a crucial initial phase in the machine learning workflow, aiming to comprehensively examine and comprehend the data's properties before engaging in modeling endeavors. Through EDA, analysts seek to elucidate fundamental data characteristics, including distribution, inter-variable correlations, and discernible patterns or irregularities. This preliminary exploration is pivotal as it furnishes a profound insight into the dataset, facilitating informed decisions regarding feature manipulation, data preprocessing, and model selection.

By undertaking EDA, researchers can pinpoint and rectify any missing or erroneous data, outliers, or inconsistencies, thus ensuring a more robust foundation for subsequent machine learning tasks.

### 6.ALGORITHMS

Predicting Crime Type

* Random Forest
* Support Vector Machine (SVM)
* Gradient Boosting

Crime Severity Prediction

* Random Forest
* Support Vector Machine (SVM)
* Gradient Boosting

High-Crime Area Identification

* Random Forest
* Support Vector Machine (SVM)
* Gradient Boosting

### 7.Implementation

# 7.1 Predicting Crime Type

# 7.1.1 Random Forest:

# 

# The dataset is split into features (X) and target (y). Only the first 5000 rows of data are considered (X[:5000] and y[:5000]).

# The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibilit . The dataset was divided into training (80%) and testing (20%) sets.

# A Random Forest classifier is initialized with 100 trees (n\_estimators=100) and a random seed of 42. Model Training: The classifier is trained on the training data

# The accuracy of the model is evaluated using the score method, which calculates the accuracy on the test data (X\_test, y\_test).

# The computed accuracy is approximately 27.4%.

# 7.1.2 Support Vector Machine (SVM):

# 

# The dataset is split into features (X) and target (y). Only the first 5000 rows of data are considered (X[:5000] and y[:5000]).

# The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.

# The code initializes a Support Vector Machine (SVM) classifier using the SVC class from sklearn.svm. The classifier is configured to use the radial basis function kernel with regularization parameter C=1.0 .A random seed of 42 is specified for reproducibility .

# The SVM classifier is trained on the training data.

# The computed accuracy is approximately 20.7%.

# 7.1.3 Gradient Boosting:

# 

# The dataset is split into features (X) and target (y). Only the first 5000 rows of data are considered (X[:5000] and y[:5000]).

# The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.

# The code initializes a Gradient Boosting Classifier using the GradientBoostingClassifier class. The classifier is configured with

# Number of estimators= 100

# A learning rate=0.1

# Maximum tree depth=3

# A random seed = 42.

# The Gradient Boosting Classifier is trained on the training data.

* The computed accuracy is approximately 24%.

**Overall Observations:**

Based on the observed accuracies, the Random Forest classifier demonstrated the highest accuracy among the three methods evaluated. The Random Forest Classifier effectively classified instances in the synthetic dataset, indicating its suitability for classification tasks.

So the below model is built based on this classifier.

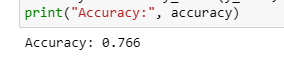
**7.2 Crime Severity Prediction**

**7.2.1 Random Forest:**



* + The code begins by preprocessing the dataset by replacing values in the "Vict Sex" column with numeric representations (2 for Female, 1 for Male, 0 for Unknown, and 3 for other cases).
  + Similarly, the "Vict Descent" column values are replaced with numeric representations based on a predefined mapping dictionary.After preprocessing, the dataset is filtered to include only selected columns .
  + The dataset is split into training and testing sets. The testing set size is set to 20% of the total data. The dataset was divided into training (80%) and testing (20%) sets.
  + A Random Forest classifier is initialized with 100 estimators. The classifier is trained on the training data.
  + The Random Forest classifier achieves a high accuracy of approximately 99% on the test data.

**7.2.2 Support Vector Machine (SVM):**



* + The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.
  + A Support Vector Machine (SVM) classifier is initialized with a radial basis function (RBF) kernel using the SVC class.
  + The classifier is configured with a regularization parameter C=1.0, auto-scale gamma parameter, and a random seed of 42.
  + The SVM classifier is trained on the training data.
  + The SVM classifier achieves an accuracy of approximately 76.6% on the test data.

**7.2.3 Gradient Boosting:**

* + The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.
  + A Gradient Boosting Classifier is initialized using the GradientBoostingClassifier class.
  + The classifier is configured with number of estimators=100, a learning rate= 0.1, a maximum tree depth =3.
  + The Gradient Boosting Classifier achieves an exceptionally high accuracy of approximately 99.9% on the test data.

**Overall Observations:**

The Random Forest classifier achieves an accuracy of 99%, the SVM classifier achieves an accuracy of 76.6%, and the Gradient Boosting classifier achieves an accuracy of 99.9%.

The best method among the three classifiers appears to be the Gradient Boosting Classifier, which achieved the highest accuracy of 99.9%.

So the below model is built based on this classifier.

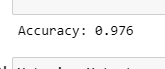
**7.3 High-Crime Area Identification**

**7.3.1 Random Forest:**



* + Only the first 5000 rows of data are considered for analysis.
  + The target variable (y) is defined as the 'AREA' column from data, limited to the first 5000 rows.
  + The dataset is split into training and testing sets, 80% of the data is allocated for training and 20% for testing. The random state is set to 42 for reproducibility.
  + A Random Forest classifier is initialized with 100 estimators and a random state of 42.
  + The accuracy of the Random Forest classifier on the test data is computed and printed, yielding an accuracy of approximately 97.9%.

**7.3.2 Gradient Boosting:**



* + Only the first 5000 rows of data are considered for analysis.
  + The dataset is split into training and testing sets with 80% of the data allocated for training and 20% for testing.
  + A Gradient Boosting Classifier is initialized with the following hyperparameters:
    - Number of boosting stages to be run. Number of estimators=100.
    - The shrinkage parameter to control the contribution of each tree. Learning rate is 0.1.
    - Maximum depth of the individual trees is 3.
    - A random seed for reproducibility, set to 42.
  + The calculated accuracy is printed, which is approximately 97.6%.

**7.3.3 Support Vector Machine (SVM):**



* + Only the first 5000 rows of data are considered for analysis.
  + The dataset is split into training and testing sets with 80% of the data allocated for training and 20% for testing.
  + An SVM classifier with a radial basis function (RBF) kernel is initialized with the following hyperparameters:
    - kernel: The type of kernel used for the SVM. Here, it's set to 'rbf'.
    - C: The regularization parameter, set to 1.0.
    - gamma: The kernel coefficient for 'rbf', set to 'scale'.
    - random\_state: A random seed for reproducibility, set to 42.
  + The calculated accuracy is printed, which is quite low at approximately 8.2%.

**Overall observations:**

Both Random Forest and Gradient Boosting classifiers perform well on the dataset, achieving accuracies of around 97.9% and 97.6%, respectively. However, the SVM classifier performs poorly, with an accuracy of only 8.2%. Therefore, Random Forest and Gradient Boosting classifiers seem to be more suitable for this particular task. The Random Forest Classifier effectively classifies instances in the dataset, demonstrating its utility in classification tasks. So the below model is built based on this classifier.