# New York City Crimes Detection using Machine Learning

Ahmed Ben Salem, Aymen Laabidi, Jaouhar Cherif and Hamza Faidi

***Abstract*— This research project centers on employing machine learning for crime detection in New York. The initiative involves a user-friendly web app enabling individuals to input personal data and select a location within the city. Through advanced machine learning algorithms, the system predicts potential criminal activities in the specified New York area. The paper covers the methodology, model selection, and web app implementation, considering ethical implications and societal impacts. Results underscore the efficacy of the approach in enhancing crime awareness and supporting decision-making for users and law enforcement in the context of New York.**

## I. INTRODUCTION

With the escalating challenges associated with urban security, leveraging technological advancements becomes imperative for enhancing crime detection and public safety. This research endeavors to address this need through the application of machine learning techniques in the context of New York. The project not only focuses on developing a robust crime prediction model but also integrates this capability into a user-friendly web application. This intersection of machine learning and user interaction aims to empower individuals to make informed decisions and assist law enforcement agencies in proactively managing and mitigating potential criminal activities. The following sections delve into the methodology, model selection, implementation of the web app, and the ethical considerations inherent in deploying such a system. The ultimate goal is to contribute to the broader discourse on utilizing cutting-edge technology for crime prevention and community well-being in urban environments, with a particular focus on the unique challenges presented by New York.

## II. LITERATURE REVIEW

Several algorithms for predicting crime have been proposed, with prediction accuracy contingent on the type of data employed and the attributes selected for prediction. In[1], crime prediction and classification were performed using data gathered from diverse websites and newsletters. The Naive Bayes algorithm and decision trees were employed, revealing superior performance by the former. In[2] conducted a comprehensive study on various crime prediction methods, including Support Vector Machine (SVM) and Artificial Neural Networks (ANN), concluding that no single method can universally address diverse crime dataset challenges. In[3] explored supervised and unsupervised learning techniques[4] on crime records, aiming to uncover connections between crime and crime patterns for knowledge discovery. This exploration contributes to enhancing the predictive accuracy of crime. Clustering approaches were implemented for crime detection, while classification methods were applied to crime prediction in [5].

## III. METHODOLOGY

The aim is to develop a robust machine learning model capable of accurately predicting offense descriptions, categorizing them into Personal, Property, Sexual and Drugs/Alcohol categories,

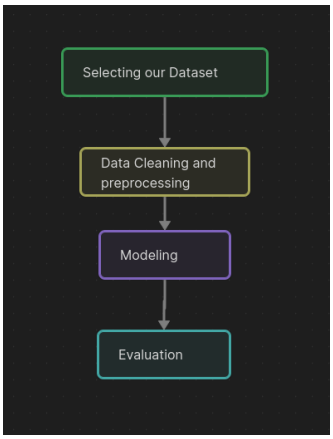



Fig. 1: Workflow

### A. Data Collection

Data collection is the process of gathering, measuring, and recording information on variables of interest for the purpose of research, analysis, or decision-making. It is a fundamental step in the research and analysis pipeline, and the quality of collected data directly influences the validity and reliability of any subsequent findings or insights.

### B. Data Cleaning and Preprocessing

Data cleaning involves identifying and rectifying errors in a dataset, including handling missing values, eliminating duplicates, addressing outliers, and standardizing data formats. It aims to enhance the accuracy and integrity of the dataset by ensuring consistency and eliminating discrepancies. On the other hand, data preprocessing focuses on transforming raw data into a format suitable for analysis or machine learning. This includes normalizing numerical features, encoding categorical variables, handling imbalances, performing feature engineering, and addressing issues specific to certain types of data, such as time series or text data. Both processes are integral to preparing data for meaningful analysis and building reliable machine learning models, contributing to the overall effectiveness and validity of the results

### C. Modeling

Gradient boosting algorithms have gained prominence in machine learning for their effectiveness in predictive modeling. Three notable implementations of gradient boosting—XGBoost, LightGBM, and CatBoost—stand out for their unique features and capabilities.

- **XGBoost (eXtreme Gradient Boosting)** is renowned for its efficiency, scalability, and regularization techniques. It has become a staple in machine learning competitions and real-world applications. XGBoost's key strengths lie in its ability to handle complex datasets, mitigate overfitting, and deliver high performance. It employs a gradient boosting framework that sequentially builds decision trees, continuously improving predictive accuracy.
- **LightGBM (Light Gradient Boosting Machine)** Developed by Microsoft, LightGBM is designed for distributed and efficient training. What sets LightGBM apart is its novel approach to handling large datasets using a histogram-based learning method. This enables faster training times and reduced memory usage, making LightGBM particularly suitable for scenarios where efficiency is crucial. The algorithm excels in capturing intricate patterns in data and is well-suited for applications in both research and industry.
- **CatBoost** CatBoost, short for Category Boosting, is a gradient boosting algorithm developed by Yandex. CatBoost is recognized for its ability to handle categorical features seamlessly without the need for extensive preprocessing. It incorporates a robust handling of categorical variables, making it user-friendly and efficient. CatBoost's optimizations, such as the implementation of ordered boosting and advanced strategies for dealing with overfitting, contribute to its competitive performance in various machine learning tasks.

### D. Evaluation

Model evaluation is the process of assessing the performance and effectiveness of a machine learning model based on its predictions or classifications. It is a crucial step in the model development pipeline, providing insights into how well the model is likely to perform on new, unseen data. The goal of model evaluation is to measure the model's accuracy, generalization ability, and suitability for the intended task.

## IV. Implementation

### A. Data Collection

**NYPD Complaint Data Historic** This dataset includes all valid felony, misdemeanor, and violation crimes re- ported to the New York City Police Department (NYPD) from 2006 to the end of 2019. The data contains 6901167 complaint and 35 columns including spatial and temporal information about crime occurrences along with their description and penal description.

### B. Data Cleaning and Exploratory data analysis

*1) Data Cleaning:* The dataset underwent thorough preprocessing to enhance its quality and suitability for analysis.

```
CMPLNT_NUM have  0.0  % missing values
CMPLNT_FR_DT have  0.008370073269449016  % missing values
CMPLNT_FR_TM have  0.0006133794151657293  % missing values
CMPLNT_TO_DT have  22.28987569993939  % missing values
CMPLNT_TO_TM have  22.228346077355578  % missing values
ADDR_PCT_CD have  0.027678746109353537  % missing values
RPT_DT have  0.0  % missing values
KY_CD have  0.0  % missing values
OFNS_DESC have  0.24064919055002115  % missing values
PD_CD have  0.08639704637365617  % missing values
PD_DESC have  0.08639704637365617  % missing values
CRM_ATPT_CPTD_CD have  0.002146827953080053  % missing values
LAW_CAT_CD have  0.0  % missing values
BORO_NM have  0.15947864794308964  % missing values
LOC_OF_OCCUR_DESC have  20.676802846693867  % missing values
PREM_TYP_DESC have  0.5368986693372525  % missing values
JURIS_DESC have  0.0  % missing values
JURISDICTION_CODE have  0.08639704637365617  % missing values
PARKS_NM have  99.60571204468879  % missing values
HADEVELOPT have  95.54802831103805  % missing values
HOUSING_PSA have  92.34179187806426  % missing values
X_COORD_CD have  0.22157053499080379  % missing values
Y_COORD_CD have  0.22157053499080379  % missing values
SUSP_AGE_GROUP have  62.403292109551096  % missing values
SUSP_RACE have  44.91506548016938  % missing values
SUSP_SEX have  46.6186501333653  % missing values
TRANSIT_DISTRICT have  97.79598719519356  % missing values
Latitude have  0.22157053499080379  % missing values
Longitude have  0.22157053499080379  % missing values
Lat_Lon have  0.22157053499080379  % missing values
PATROL_BORO have  0.09223692955554655  % missing values
STATION_NAME have  97.79598719519356  % missing values
VIC_AGE_GROUP have  20.937259080858613  % missing values
VIC_RACE have  0.004983707748221551  % missing values
VIC_SEX have  0.00393585124731343  % missing values
```

Fig. 2: Dataset before cleaning

Initial steps involved handling missing values by either dropping columns with significant null entries or imputing binary indicators for specific categorical variables. Subsequently, date and time columns were standardized by converting them to datetime objects and removing rows lacking essential temporal information. New variables, such as year, month, day, hour, and weekday, were derived from the incident dates to provide additional temporal insights.

Further cleaning efforts focused on addressing missing values and inconsistencies in categorical variables related to demographics. Unknown or missing values were appropriately imputed, ensuring a more complete representation of the dataset. Redundant or unnecessary columns were removed to streamline the dataset. Additionally, a new categorical column was introduced to classify crime types based on specific categories, facilitating a more structured and interpretable analysis.

These preprocessing steps collectively aimed at ensuring the dataset's integrity, improving its usability, and providing a solid foundation for subsequent exploratory data analysis and modeling.

*2) Exploratory data analysis:* Exploratory Data Analysis is a crucial phase in the data analysis process that involves examining and visualizing data to uncover patterns, relationships, and insights. By employing statistical and graph-

ical techniques, EDA aids in understanding the underlying structure of the dataset, identifying potential outliers, and informing subsequent analyses. It serves as a preliminary step to gain familiarity with the data, paving the way for more informed decision-making and hypothesis formulation.

In the following plots, we visually explore key patterns and trends revealed through the Exploratory Data Analysis, providing a comprehensive overview of the dataset's characteristics.

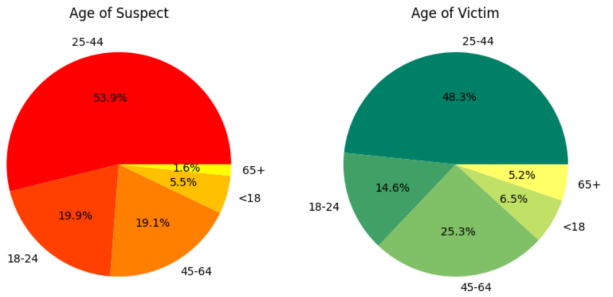



Fig. 3: Age of Suspect/Victim

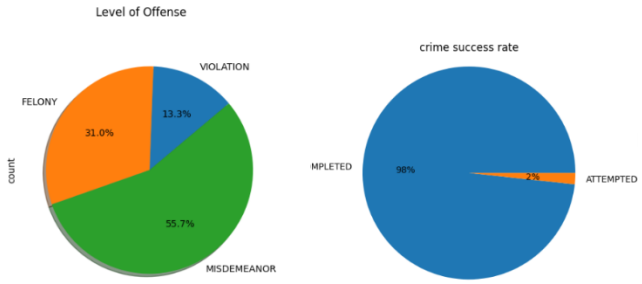



Fig. 4: level of offense / Crimes Success rate

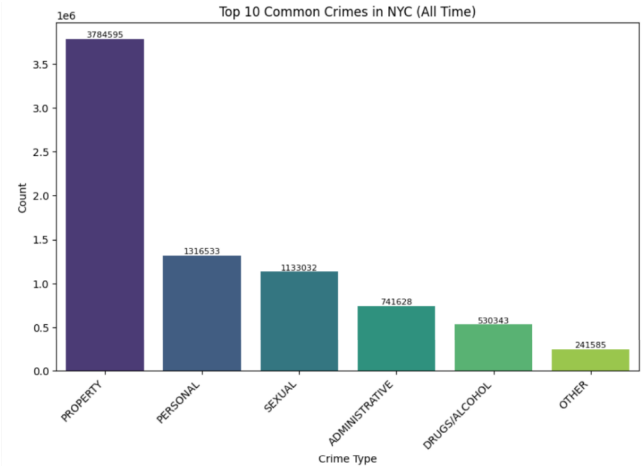



Fig. 5: Top Commun Crimes in New York

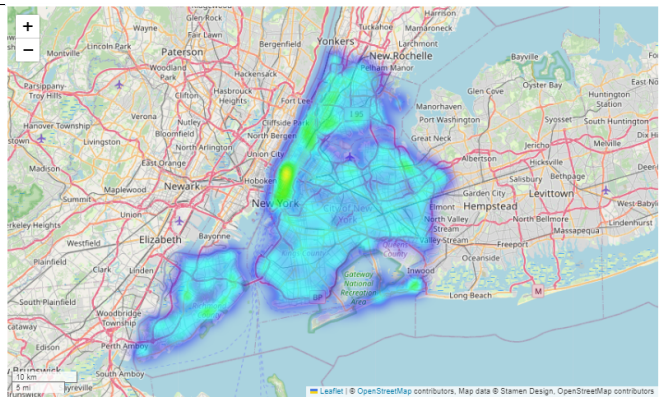


Fig. 6: Crimes Heatmap from NYC

### C. Data Preprocessing

Before transitioning to the modeling phase, several preprocessing steps were applied to refine the dataset. Instances associated with specific categories were filtered out, and the target variable underwent encoding to facilitate analysis. Balancing techniques were employed to address class distribution issues, ensuring a more representative dataset.

The feature selection process focused on key aspects, including temporal information, geographic coordinates, and relevant crime-related attributes. Binary classification within certain columns was established to simplify subsequent analyses. Exploratory correlation analysis was performed to understand relationships among selected features, visualized through a heatmap. Categorical variables were encoded, and boolean columns were transformed to streamline compatibility with machine learning algorithms.

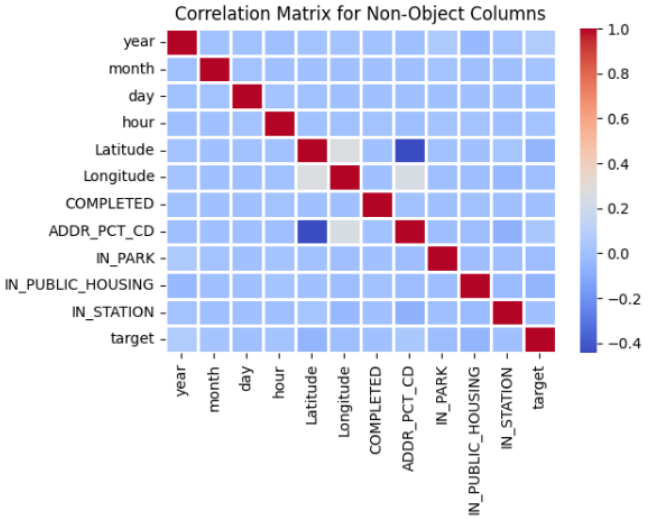



Fig. 7: numeric variables correlation matrix

### D. Modeling

In the training phase, the dataset is partitioned into training and testing sets, with approximately 15% reserved for testing to ensure robust generalization evaluation. Shuffling introduces randomness, and a specific random state is set for reproducibility. This division enables effective model training on one subset and testing on another, facilitating a comprehensive performance assessment. Hyperparameter tuning with Optuna was executed for each algorithm, enhancing model configurations and optimizing predictive capabilities.

The primary objective of the model is to classify and predict the likelihood of specific crimes occurring within categories such as 'DRUGS/ALCOHOL,' 'PROPERTY,' 'PERSONAL,' and 'SEXUAL.' Evaluation metrics will gauge the model's ability to discriminate between these crime types, offering valuable insights into its effectiveness in predicting and distinguishing among specific criminal activities.

### E. Evaluation and Metrics

- **ROC Curve:**

The ROC curve visually represents the trade-off between sensitivity (true positive rate) and specificity (true negative rate) across various threshold values. In crime prediction, it illustrates how well the model distinguishes between positive (occurrence of crime) and negative (non-occurrence of crime) instances. The area under the ROC curve (AUC-ROC) quantifies the model's overall performance, with a higher AUC indicating superior class discrimination.

- **Confusion Matrix:**

The Confusion Matrix breaks down predictions into true positives, true negatives, false positives, and false negatives. This matrix enables the assessment of precision, recall, and F1 score, providing nuanced insights into the model's performance.

- **Accuracy:** Measures overall correctness using the formula:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

- **Precision:** Quantifies accuracy of positive predictions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

- **F1 Score:** Balances precision and recall:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

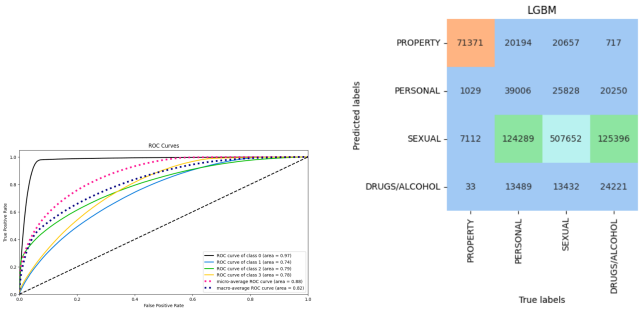
- **Recall:** Ratio of true positive predictions to total actual positives:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

These metrics comprehensively evaluate the effectiveness of our crime prediction model in the specified New York area.

### F. Obtained Results

*1) LightGBM:* This section encompasses the assessment of the LightGBM (LGBM) model.

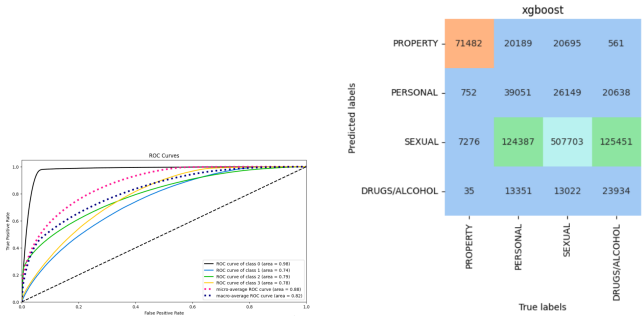



(a) ROC Curve for the LGBM Model

(b) Confusion Matrix for the LGBM Model

Fig. 8: Evaluation Metrics for the LGBM Model

*2) XGBoost:* This section encompasses the assessment of the XGBoost model.

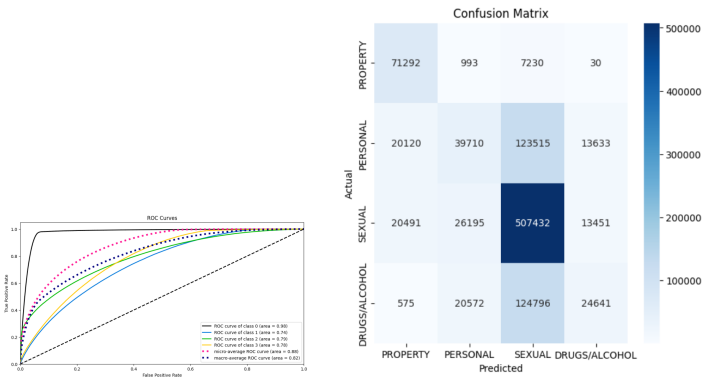



(a) ROC Curve for the XGBoost Model

(b) Confusion Matrix for the XGBoost Model

Fig. 9: Evaluation Metrics for the XGBoost Model

*3) CatBoost:* This section focuses on assessing the performance of the CatBoost model.




(a) ROC Curve for the CatBoost Model

(b) Confusion Matrix for the CatBoost Model

Fig. 10: Evaluation Metrics for the CatBoost Model

### G. Models Comparison

As observed, the performance of the three models is quite comparable. Nevertheless, it is noteworthy that the LightGBM model exhibited a slightly superior performance, particularly evident when comparing their confusion matrices.

TABLE I: Comparison of different models

| Model | Accuracy (%) | F1 Score |
|---|---|---|
| XGBoost | 61.2 | 59.64 |
| CatBoost | 63.38 | 61.29 |
| **LightGBM** | **64.6** | **65.31** |

## V. USER INTERFACE

After training and saving the model weights, we created a Streamlit and Folium-based web app for interactive crime prediction. Users provide input on gender, race, age, date, and time, and can select a location on the map, specifying a category like a park, public housing, or station. Destination selection is flexible, allowing users to click on the map or type its name.

This information is then transformed to match the model's input. We utilized various shapefiles to determine the police precinct and borough from coordinates. Subsequently, using the loaded model weights file, we make predictions regarding the type of crime. The predicted crime type, along with potential subtypes, is then sent back to the user. The web application is deployed using Streamlit, and you can access it here.

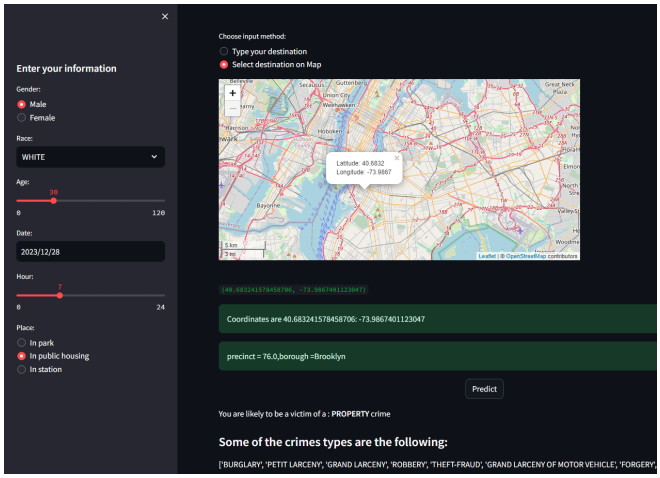



Fig. 11: Prediction when user selects map destination

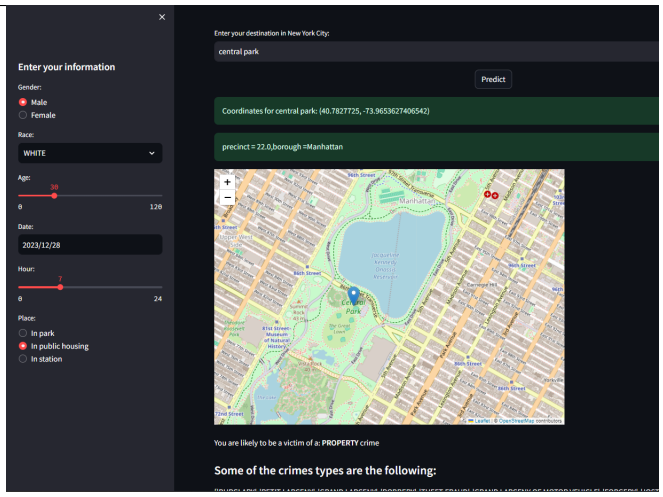



Fig. 12: Prediction when user inputs text destination

## VI. CONCLUSIONS

Predicting and preventing crime has emerged as a pivotal focus in contemporary society. The overarching goal of crime prediction is to mitigate the incidence of criminal activities by foreseeing the types of crimes likely to occur in the future. This study employs the Random Forest model for the analysis and prediction of crime patterns. The obtained results underscore the model's effectiveness, particularly when trained optimally, yielding commendable accuracy. It is worth noting that the choice of the most suitable model is contingent upon the specific characteristics of the dataset in use, emphasizing the significance of tailoring approaches to the unique attributes of the data for optimal predictive outcomes.

## REFERENCES


[1] Shiju Sathyadevan,Devan M. S.,Surya S Gangadharan, First,"Crime Analysis and Prediction Using Data Mining" International Conference on Networks Soft Computing (ICNSC), 2014.

[2] Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav,"Crime pattern detection,analysis and prediction,International Conference on Electronics, Communication and Aerospace Technology(ICECA), 2017.

[3] Amanpreet Singh,Narina Thakur,Aakanksha Sharma,"A review of supervised machine learning algorithms",3rd International Conference on Computing for Sustainable Global Development,2016

[4] Bin Li,Yajuan Guo,Yi Wu,Jinming Chen,Yubo Yuan,Xiaoyi Zhang,"An unsupervised learning algorithm for the classification of the protection device in the fault diagnosis system",in China International Conference on Electricity Distribution (CICED),2014

[5] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219- 4225, Mar. 2013.