CSE/ECE 343: Machine Learning Project Progress Report

Niteen Kumar(2022336)
niteen22336@iiitd.ac.in

Pratham Mittal(2022373)
pratham22373@iiitd.ac.in

Satyam(2022462)
satyam22462@iiitd.ac.in

Sachin Maurya(2022424)
sachin22424@iiitd.ac.in

Abstract

- ***CrimeCast*** *is a machine learning project aimed at predicting crime categories based on comprehensive data, including date, time, location, and victim demographics.*
- *Identify patterns within the data to classify crimes into different categories like:*
  - *Property Crimes*
  - *Violent Crimes*
  - *Crimes against Persons*
- *Enable proactive resource allocation and improve public safety by forecasting crime trends.*
- *Assist in preventing criminal activities by enhancing law enforcement strategies.*

1. Introduction

The increasing crime rates in urban areas pose a significant threat to public safety. Law enforcement agencies often face challenges in resource allocation and crime prevention due to the unpredictable nature of criminal activities. Traditional methods of crime monitoring are reactive and inefficient, relying on post-crime responses rather than proactive strategies.

This project, ***CrimeCast****, aims to address this issue by using machine learning techniques to predict crime categories based on available data. By identifying patterns in factors such as time, location, and demographics, CrimeCast seeks to enable law enforcement agencies to forecast crime trends, allocate resources effectively, and develop more proactive crime prevention strategies. Github*

2. Literature Survey

2.1. Paper 1

Title and Author:
The SKALA Approach of the State Office for Criminal Investigation of North Rhine-Westphalia by Kai Seidensticker, Katharina Schwarz (2022)

Objective:
To utilize predictive policing, risk terrain modelling, and time series analysis to forecast crime categories and enhance police strategic planning in North Rhine-Westphalia.

Key Findings:
The study implemented various models, including predictive policing with decision trees and random forests, risk terrain modelling, and time series analysis (ARIMA). Random forest models were particularly effective in identifying spatial crime patterns and predicting crime hotspots.

Model Relevance:
CrimeCast aims to predict crime categories using similar ML techniques and data-driven approaches as highlighted in the SKALA. Incorporating random forest models will enhance our ability to identify crime patterns and forecast crime hotspots.

2.2. Paper 2

Title and Author:
Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention by Neil Shah, Nandish Bhagat, and Manan Shah

Objective:
To explore and implement machine learning (ML) and computer vision techniques for crime prediction and prevention, aiming to improve the accuracy and efficiency of law enforcement efforts.

Key Findings:
The study employed various ML algorithms, including K-nearest neighbor (KNN), decision trees, and deep neural networks (DNN). The DNN model, which integrated multimodal data from several domains, achieved an accuracy of 84.25% in predicting crime occurrences.

Model Relevance:
CrimeCast aims to utilize similar ML techniques, such as decision trees, as demonstrated in this study. By adopting these proven methods, we aim to improve the accuracy and robustness of our model in categorizing crime data.

3. Dataset

3.1. Dataset Details

- *The dataset encompasses reported crime incidents across various districts, highlighting essential details regarding victims, premises, weapons used, time, and other relevant factors.*
- *The dataset reveals a predominance of property crimes across various districts, often occurring in residential and commercial settings.*

Field Name	Description
Location	Street address of the crime incident.
Cross_Street	Cross street of the rounded address.
Latitude	Latitude coordinates of the crime incident.
Longitude	Longitude coordinates of the crime incident.
Date_Reported	Date the incident was reported.
Date_Occurred	Date the incident occurred.
Time_Occurred	Time the incident occurred in 24-hour military time.
Area_ID	LAPD's Geographic Area number.
Area_Name	Name designation of the LAPD Geographic Area.
Reporting_District_no	Reporting district number.
Part 1-2	Crime classification.
Modus_Operandi	Activities associated with the suspect.
Victim_Age	Age of the victim.
Victim_Sex	Gender of the victim.
Victim_Descent	Descent code of the victim.
Premise_Code	Premise code indicating the location of the crime.
Premise_Description	Description of the premise code.
Weapon_Used_Code	Weapon code indicating the type of weapon used.
Weapon_Description	Description of the weapon code.
Status	Status of the case.
Status_Description	Description of the status code.
Crime_Category	The category of the crime (Target Variable).

Table 1. Field Descriptions of Crime Incident Dataset

3.2. Data preprocessing techniques

- ***Feature Selection:*** *To enhance model performance and interpretability by selecting the most relevant features while removing unnecessary columns (e.g., Cross_Street).*
- ***Missing Values:*** *Identified and addressed missing entries in critical fields (e.g., Victim Age, Weapon Used) using techniques like imputation or removal.*
- ***Converting Data Types:***
  - *Converted Date_Reported and Date_Occurred columns to datetime format for accurate datetime operations.*
  - *Converted Time_Occurred to a readable time format for better interpretability.*
  - *Converted Part 1-2 to a categorical column to improve model performance.*
- ***Encoding Categorical Variables:*** *Converted categorical fields (e.g., Area_Name, Status) into numerical formats using one-hot encoding or label encoding to facilitate model training.*
- ***Feature Engeenerring :*** *Developed additional variables (e.g., Date_Reported) to enhance analysis depth and predictive modeling.*

3.3. Exploratory Data Analysis (EDA)

Basic Descriptive Statistics:

- *Numerical Features:*
  - *Calculated key statistics such as mean, median, and standard deviation to summarize the distribution of numerical variables.*
  - *Assessed the range and interquartile range (IQR) to identify potential outliers in the data.*
- *Categorical Features:*
  - *Determined the count of occurrences for each category to understand the frequency distribution.*
  - *Identified unique values within categorical columns to assess diversity and potential redundancy.*

Histogram for numerical features:

- *Latitude & Longitude: Incidents are geographically clustered, suggesting they occur within a small area.*
- *Time_Occurred: Crime activity varies throughout the day, with spikes at specific times.*
- *Area_ID & Reporting_District_no: The distribution of incidents across areas and districts is fairly balanced, with some areas reporting more frequent incidents.*

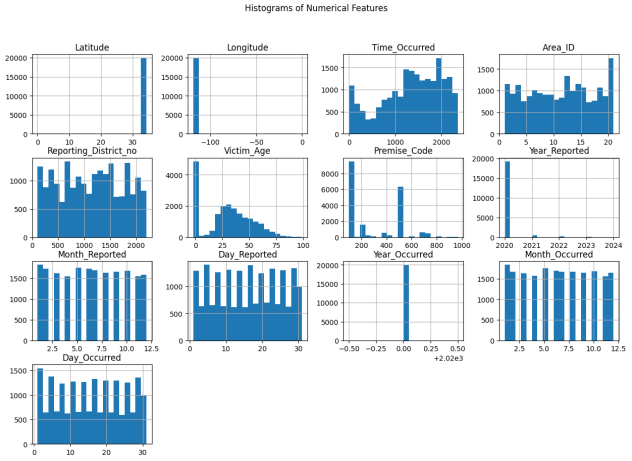

Figure 1. Histogram for numerical features

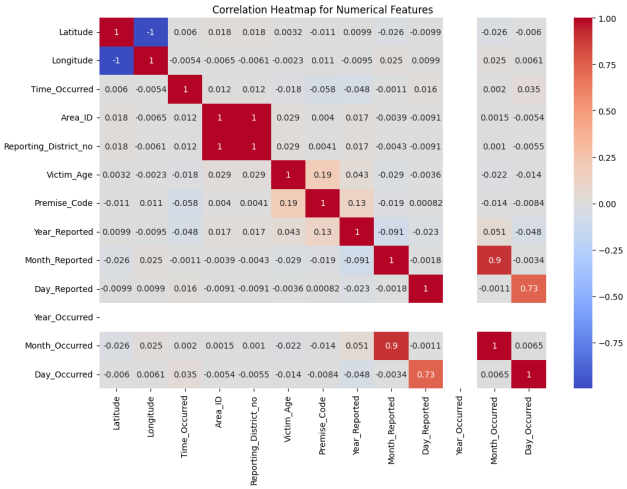

Figure 2. Heat Map

- *Victim_Age: The majority of victims are younger (20-40 years old), with fewer incidents involving older individuals.*

- *Premise_Code: Certain premises show notably higher incident rates, indicating vulnerability to crime.*

- *Year_Reported & Year_Occurred: Most incidents are recent, primarily from 2023.*

- *Month & Day (Reported/Occurred): The distribution of incidents is uniform across months and days, with no significant patterns by date or season.*

Heat Map: *The heat map helps us for visualizing patterns, relationships, and correlations within a dataset by highlighting areas of importance or clustering.The features having positive correlation are Victim Age and Premise Code.These variables are highly interrelated with each other.*

Box plot for Numerical Features : *The box plot summarizes the distribution of a dataset by showing key statistics distribution, central value, spread, and presence of outliers in a dataset. The features that contain outliers are Latitude, Longitude, Victim Age, and Year Reported. These points are significantly far from the rest of the data points.*

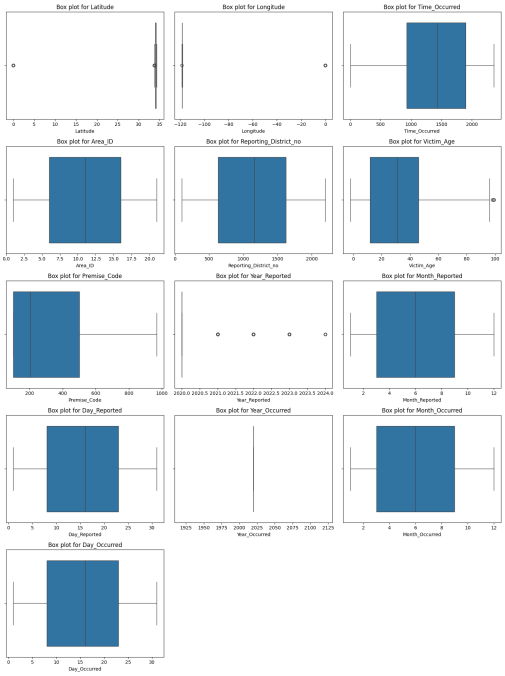

Figure 3. Box plot for Numerical Features

4. Methodology and Model Details

4.1. Feature Extraction

The features used in the model were extracted from the dataset, and the preprocessing involved handling categorical variables, missing values, and scaling numerical features. Some key aspects of the feature extraction process included:

- *Categorical variables such as* `Crime Type`, `Location`, *and* `Time of Crime` *were encoded using one-hot encoding.*

- *Missing values were handled using imputation techniques to ensure that the dataset remained robust for training the models.*

- *Numerical features were scaled using Min-Max scaling to normalize the data for the machine learning algorithms.*

4.2. Exploratory Data Analysis (EDA)

We began our analysis with an extensive Exploratory Data Analysis (EDA) to gain insights into the dataset and understand the distributions and relationships between variables. We visualized the data using various plots, which helped us identify patterns and trends.

4.3. Model Training

We trained three different machine learning models to predict crime categories based on the given features. These models were:

- ***Naive Bayes Classifier:*** *This probabilistic classifier is based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It performed well with categorical data, providing a strong baseline model.*
- ***Random Forest Classifier:*** *This ensemble model is based on decision trees and aggregates the results from multiple trees to improve accuracy and reduce overfitting. It handled both categorical and numerical features effectively.*
- ***XGBoost Classifier:*** *An optimized version of gradient boosting that builds a sequence of decision trees, each correcting the errors of the previous one. It is particularly good at handling complex patterns in the data.*
- ***SVM Classifier:*** *A supervised machine learning algorithm used for classification and regression tasks that finds the hyperplane best separating data into classes by maximizing the margin.*
- ***MLP Classifier:*** *A type of artificial neural network consisting of an input layer, hidden layers, and an output layer, used for solving complex problems through backpropagation and non-linear transformations.*

4.4. Cross Validation and Model Evaluation

To evaluate the performance of the models and ensure their generalizability, we implemented cross-validation. Specifically, we used k-fold cross-validation with $k = 10$, which allowed us to assess the model's performance across different subsets of the data.

5. Results and Analysis

These results demonstrate that the Random Forest and XGBoost models outperform the Naive Bayes classifier, indicating their effectiveness in the task of categorizing crime categories.

Table 2. Cross-Validation Results for Different Models

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	83.55	83.12	84.34	82.47
Naive Bayes	33.90	79.00	34.00	42.00
XGBoost	87.10	87.00	85.32	86.88
SVM	86.18	85.68	85.45	72.60
MLP	86.10	85.00	86.17	86

Table 3. classwise accuracies using xgboost

Class	Precision	Recall	F1-Score	Support
0	0.65	0.62	0.63	32
1	0.68	0.55	0.61	374
2	0.75	0.81	0.78	267
3	0.44	0.11	0.18	35
4	0.94	0.92	0.93	2303
5	0.83	0.92	0.87	989

6. Conclusion

In this project, we successfully categorized crime categories using machine learning models. Among the models, XGBoost demonstrated the highest accuracy at 88.74%, outperforming all the other models. The results indicate that advanced ensemble methods like XGBoost are more effective in handling the complexity of the dataset, providing robust and accurate predictions.

6.1. Project Timeline

6.1.1 Completed Tasks

- *Data collection, preprocessing, and feature identification.*
- *Exploratory data analysis (EDA).*
- *Model selection, training, and initial evaluation.*
- *Model optimization and final evaluation.*
- *Final project report and final presentation.*
- *Feedback integration.*

6.2. Contribution

- ***Sachin:*** *Data collection, preprocessing, feature identification, documentation, training, and web development.*
- ***Niteen:*** *Exploratory Data Analysis (EDA), feature engineering, and optimization.*
- ***Pratham:*** *Model selection, training, and initial evaluation.*

- ***Satyam:*** *Model selection, optimization, and final evaluation.*

7. References

References

[1] Zhu, S., Cheng, J., Wang, Y., & Zhou, X. (2021). Crime Analysis and Prediction Based on Urban Heterogeneous Data. Visual Computing for Industry, Biomedicine, and Art, 4(1), 1-10. https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z

[2] Martinez, R., Morales, A., Vega, A., & Garcia, L. (2022). Machine Learning Models for Crime Prediction and Categorization. MDPI Proceedings, 18(1), 39. https://www.mdpi.com/2673-4591/18/1/39

[3] B. McGlohon, K. C. E. Albright, & H. W. H. Lee. (2018). Crime and Punishment: Understanding and Predicting Crime in Urban Environments. arXiv preprint arXiv:1801.02858. https://ar5iv.labs.arxiv.org/html/1801.02858v1