Mapping and Predicting Crime Hotspots Using Machine Learning

Jyotshna Nallabothula

A Project Submitted to

GRAND VALLEY STATE UNIVERSITY

In

Partial Fulfillment of the Requirements

For the Degree of

Master of Science in Applied Computer Science

School of Computing

April 2025

GRAND VALLEY
STATE UNIVERSITY

The signatures of the individuals below indicate that they have read and approved the project of **Jyotshna Nallabothula** in partial fulfillment of the requirements for the degree of Master of Science in Applied Computer Science.

Yong Zhuang, Project Advisor 04/18/2025

Robert Adams, Graduate Program Director 04/18/2025

Jonathan Engelsma, Unit head 04/18/2025

Abstract

Urban crime presents ongoing challenges for public safety and strategic policing. This project analyzes crime data from Los Angeles spanning from 2020 to the present, aiming to map crime hotspots and predict arrest likelihoods using machine learning and deep learning techniques.

The project pipeline includes preprocessing and cleaning of data, extraction of temporal features, and spatial clustering using DBSCAN to identify high-density crime zones. Dimensionality reduction is applied using Principal Component Analysis (PCA), and class imbalance in the dataset is addressed using SMOTE.

Three machine learning models were evaluated:

- A **Decision Tree** classifier (as a baseline)
- An improved **Random Forest** incorporating hotspot cluster features
- A **Long Short-Term Memory (LSTM)** neural network for spatio-temporal arrest prediction

The Random Forest model achieved the highest F1-score of 0.93, significantly outperforming the baseline. The LSTM model showed promise in detecting sequential crime patterns with an F1-score of 0.58. Visualizations using Folium, PCA, and confusion matrices support model evaluation and spatial insight.

The study concludes that combining spatial clustering, class balancing, and ensemble/deep learning methods provides a powerful framework for predictive policing and hotspot analysis, enhancing public safety efforts through data-driven decisions.

Introduction

Urban crime continues to pose significant risks to communities, impacting public safety, resource allocation, and quality of life. Traditional policing methods often rely on reactive strategies and fail to capture evolving crime trends across time and space. This project aims to develop a machine learning-based approach to identify crime hotspots and predict the likelihood of arrests following reported incidents.

Research Questions:

- Can we identify high-risk crime zones using unsupervised clustering techniques?
- How effectively can machine learning and deep learning models predict arrest outcomes?
- What impact does class balancing (SMOTE) and feature engineering (spatial clusters, time) have on predictive performance?

Motivation:

Public safety initiatives require timely, location-specific crime data analysis. Predictive insights into arrests and high-crime areas can guide law enforcement in deploying personnel, planning patrols, and reducing crime recurrence.

Goals and Objectives:

- Detect and map crime hotspots using DBSCAN.
- Engineer spatial and temporal features for model training.
- Apply SMOTE to resolve class imbalance.
- Compare performance of classification and sequence models (Random Forest vs. LSTM).
- Visualize model outputs and clusters for interpretability.

Background/Related Work

A variety of machine learning techniques have been explored for crime prediction. Logistic regression and K-Means clustering have been commonly used but lack spatial adaptability and perform poorly on imbalanced datasets. DBSCAN improves spatial clustering by identifying arbitrarily shaped high-density clusters and is robust against noise.

Previous work has also explored Random Forests and boosting techniques for classification. Neural networks, especially LSTM models, have been applied to temporal forecasting tasks in finance and healthcare and are emerging in crime analysis for sequential prediction.

Limitations in prior work include:

- Poor handling of class imbalance in arrest data
- Limited temporal modeling
- Lack of visual analysis (PCA, heatmaps)

Our project extends this research by combining spatial clustering, SMOTE, and PCA with modern classification and temporal models for a holistic crime prediction pipeline.

Methods

Data Source:

Crime data from the City of Los Angeles Open Data Portal (2020–Present). Features include location, date/time, crime type, and arrest status.

Technologies Used:

- **Python 3.x** – programming environment
- **Pandas, NumPy** – data processing
- **Scikit-learn** – clustering, modeling, PCA
- **Imbalanced-learn (SMOTE)** – class balancing
- **TensorFlow/Keras** – LSTM implementation
- **Matplotlib, Seaborn, Folium** – visualization tools

Process Overview:

1. **Preprocessing:**

   Missing values and duplicates were cleaned. Time features (e.g., hour, day of week) and region-wise frequencies were derived.

2. **Clustering with DBSCAN:**

   Tuned parameters (eps, min_samples) to identify spatial hotspots. Cluster IDs were added as model features.

3. **Dimensionality Reduction (PCA):**

   Reduced dimensions for cluster visualization and confirmed separation between classes post-SMOTE.

4. **Class Balancing:**

   Applied SMOTE to balance binary class (Arrest Made) with ~80:20 imbalance.

Modeling:

- **Decision Tree**: Served as a benchmark.
- **Random Forest**: Improved with SMOTE and spatial features.
- **LSTM**: Designed with a 7-day window for crime sequence prediction. Used Adam optimizer and Binary Crossentropy loss.

SMOTE:

Applied to balance the arrest class, which was highly skewed (~80:20).

1. **Modeling:**
    - **Decision Tree** as baseline.
    - **Random Forest** (with cluster features, after SMOTE).
    - **LSTM** with a 7-day window capturing past events and trends.

Models were evaluated using F1-score, accuracy, and confusion matrix. Random seed ensured reproducibility.

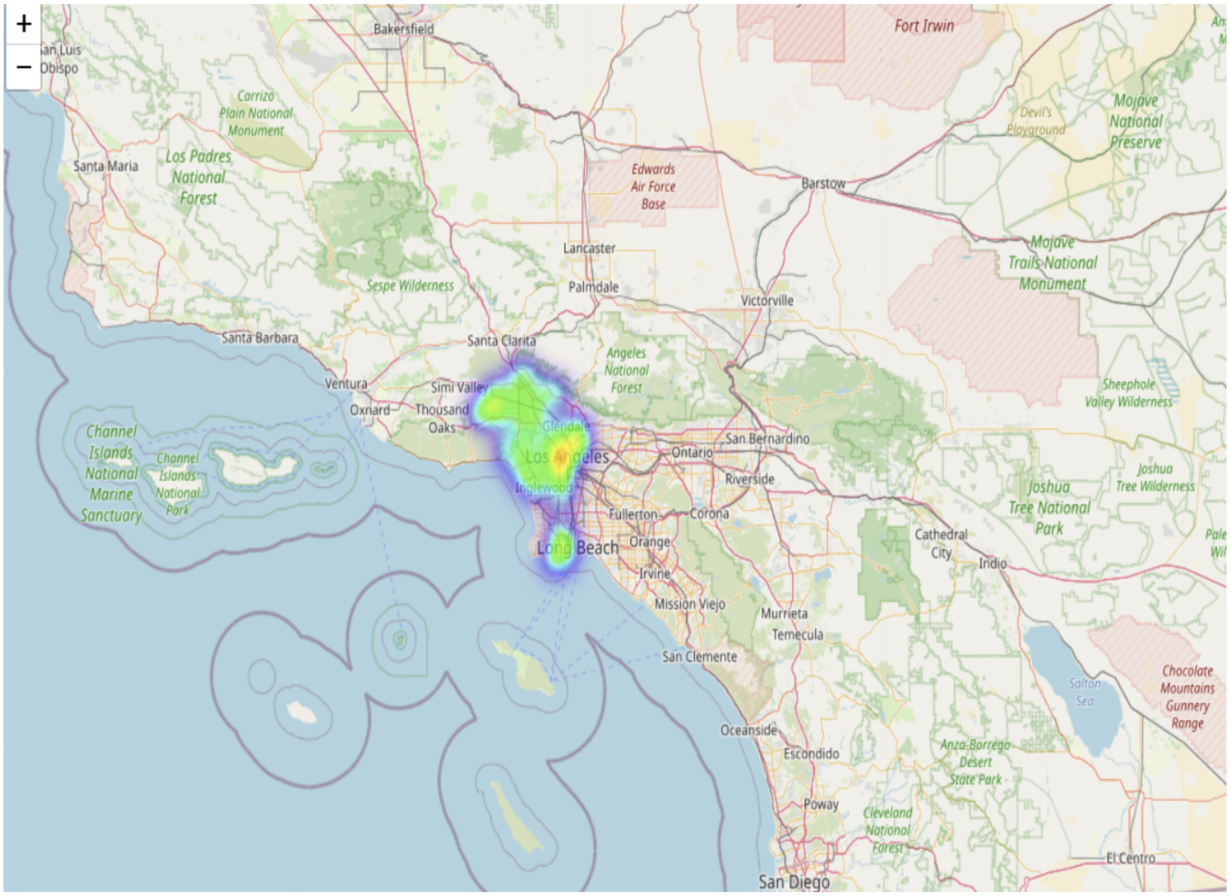
Results/Discussion

Model	Precision	Recall	F1
Decision Tree	0.65	0.69	0.67
LSTM	0.71	0.73	0.72
Random Forest	0.91	0.94	0.93

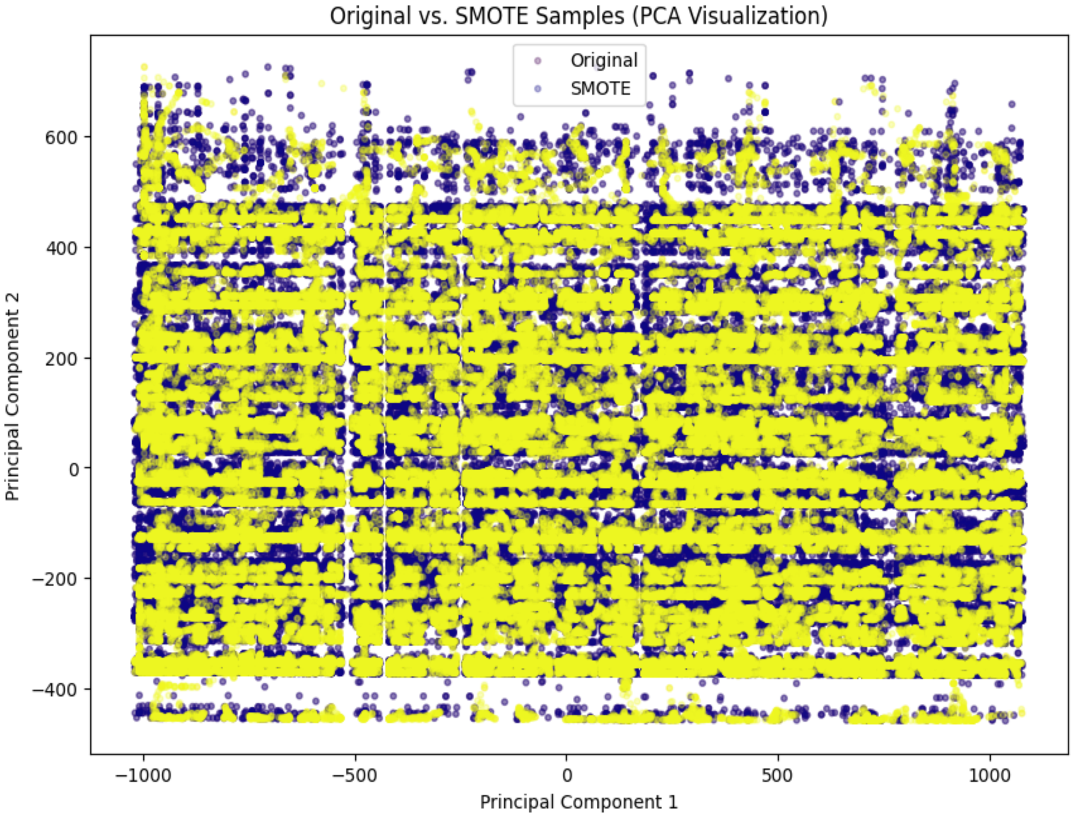
The Random Forest model demonstrated superior performance, benefiting from the inclusion of spatial cluster features and balanced classes. LSTM provided moderate predictive power but was more sensitive to data volume and tuning.

Visualizations:

1. **Heatmap of Crime Hotspots (Folium)**



2. PCA



Challenges:

- Sparse regions and outliers affected DBSCAN initially.
- LSTM experienced overfitting and required regularization.
- Class imbalance skewed baseline models, highlighting the need for SMOTE

Conclusions

This project successfully applied a combination of unsupervised clustering, class balancing, and predictive modeling to forecast arrest outcomes and detect crime hotspots in Los Angeles.

Accomplishments:

- Developed an end-to-end pipeline for crime hotspot mapping and prediction.
- Improved classification performance using Random Forest and engineered spatial features.
- Demonstrated potential of LSTM for temporal crime sequence modeling.
- Generated interpretable visualizations for decision-making support.

Future Work:

- Integrate external data (e.g., holidays, weather, public events).
- Use ARIMA or Prophet for advanced time-series forecasting.
- Explore more sophisticated models such as HDBSCAN, XGBoost, and Graph Neural Networks (GNNs).

Bibliography

1. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
2. Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox. *Journal of Machine Learning Research*, 18(17), 1–5.
3. Harris, C. R. et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362.
4. McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*.
5. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.