"Crime Classification and Pattern Detection in Chicago Using Machine Learning Models"

Anabel Aguila Ramirez
Lewis University
Romeoville, IL. United States
anabelaguilarramir@lewisu.edu

***Abstract*— This project presents a machine learning approach to classifying crime data using supervised models. The dataset, obtained from the Chicago Police Department's CLEAR system (2024), contains crime reports categorized into broader crime types. Several machine learning models were implemented, including Logistic Regression (with Polynomial Features), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Decision Trees, and Random Forest for classification. Additionally, Principal Component Analysis (PCA) was used for dimensionality reduction, and K-Means and DBSCAN were applied for clustering. Among the classification models, MLP achieved the highest accuracy, while DBSCAN effectively identified crime hotspots. This study provides insights into crime classification and pattern recognition using machine learning, offering valuable information for crime trend analysis and law enforcement decision-making.**

Keywords— Crime Classification, Machine Learning, Supervised Learning, Clustering, PCA, Multi-Layer Perceptron, K-Means, DBSCAN.

I. INTRODUCTION

The goal of this project is to build a machine learning system that classifies crimes into different categories based on a structured dataset. To achieve this, multiple machine learning techniques were implemented, including parametric models, neural networks, support vector machines (SVM), tree-based models, and clustering methods. The performance of these models was evaluated using standard classification metrics such as precision, recall, F1-score, and log loss.

Understanding crime patterns through data analysis can provide useful insights for law enforcement agencies and policymakers. By applying both supervised learning models for classification and unsupervised learning models for pattern recognition, this project aims to develop an efficient crime classification system.

II. DATASET

A. Dataset Description

The dataset used in this project was obtained from the Chicago Police Department's CLEAR system (2024). Initially, it contained 251,616 observations and 17 attributes. After preprocessing, the dataset was refined to 85,981 observations with 8 attributes, improving data quality and ensuring better classification and clustering performance.

B. Preprocessing and Feature Selection

To prepare the dataset for machine learning, the following preprocessing steps were applied:

- Feature Removal: Eliminated unnecessary attributes such as Case #, Date of Occurrence, Block, IUCR, Location, X coordinate, Y coordinate, Latitude, Longitude.
- Handling Missing Values: Removed records with incomplete or missing data.
- Categorical Encoding: Converted categorical variables into numerical values using Label Encoding.
- Feature Scaling: Standardized numerical attributes using StandardScaler().
- Class Balancing: Adjusted the dataset to reduce bias in classification models.
- Dimensionality Reduction: Principal Component Analysis (PCA) was used for visualization in clustering techniques.

After feature selection, the final dataset consisted of 8 key attributes, described below.

TABLE I. DATASET ATRIBUTES TABLE

Feature Name	Description
SECONDARY DESCRIPTION	*More specific classification of the crime type.*
LOCATION DESCRIPTION	*The location where the crime occurred.*
ARREST	*Whether an arrest was made (**0 = No, 1 = Yes**).*
DOMESTIC	*Whether the crime was domestic-related (**0 = No, 1 = Yes**).*
BEAT	*A numerical identifier representing the police beat.*
WARD	*The political ward in which the crime took place.*
FBI CD	*Classification code assigned by the FBI for crime types.*
CATEGORY (Target Variable)	*Crime category (Violent Crime, Property Crime, Drug-Related Crime, Sex-Related Crime, Fraud & Financial Crime, Public Order Crime and Other Offense.).*

[a.] Table shows description of each feature

After preprocessing, to ensure reliable model evaluation, the dataset was randomly split into 80% training data and 20% test data. The training data was used to train machine learning models, while the test data was reserved for evaluating model performance.

III. Models Implemented

This study utilized a combination of supervised learning models for crime classification and unsupervised clustering techniques to analyze crime patterns. The classification models aimed to assign crimes to predefined categories, while the clustering techniques helped identify hidden patterns and crime hotspots within the dataset.

A. Supervised Learning Techniques

1. Logistic Regression: Starting point to evaluate the effectiveness of linear classification methods. Applied Polynomial Features (degree=3) to capture non-linear relationships, used class_weight='balanced' to handle class imbalances, and applied regularization with C=0.5 to prevent overfitting.
2. Multi-Layer Perceptron (MLP): The model was designed as a deep learning system to recognize complex patterns in crime data. It had two hidden layers (16 and 8 neurons), learning_rate='adaptive', early stopping, and dropout regularization to prevent overfitting.
3. Support Vector Machine (SVM): Implemented Radial Basis Function (RBF) Kernel to handle non-linear data patterns with C=1 to balance accuracy and simplicity. Probability estimates were turned on (probability=True) to interpret classification results better.
4. Decision Tree: Useful model for interpretability and ability to handle mixed data types. The model was optimized with a maximum depth of 5 to avoid overfitting and a minimum of 15 samples per split.
5. Random Forest: The model used ensemble learning to improve classification accuracy. It combined 200 decision trees (n_estimators=200), each limited to a maximum depth of 6 to prevent overfitting. Additionally, it applied class_weight='balanced' to handle uneven class distributions in the data.
6. K-Nearest Neighbors (KNN): Applied to evaluate similarity-based classification, using 11 neighbors (n_neighbors=11) and distance-based weighting.

B. Unsupervised Learning Techniques (Clustering)

1. K-Means Clustering: Used to group similar crimes by optimizing clusters with the Elbow Method, selecting K=5 as the optimal number of clusters. PCA was applied to reduce dimensions and visualize the clusters.
2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Implemented to identify crime hotspots and anomalies, using eps=1.5 and min_samples=10. Unlike K-Means, DBSCAN does not require a predefined number of clusters and is effective in detecting dense crime regions.

IV. Results

This section presents the performance of the implemented supervised classification models and unsupervised clustering techniques, evaluated using standard machine learning metrics. The results include accuracy, precision, recall, F1-score, confusion matrices, and clustering evaluation metrics.

A. Supervised Learning Model Performance

1. Logistic Regression:

Logistic Regression Report:

	precision	recall	f1-score	support
0	0.89	0.91	0.90	659
1	0.92	0.80	0.86	1841
2	0.69	0.66	0.67	2445
3	0.61	0.48	0.54	4559
4	0.42	0.68	0.52	895
5	0.32	0.56	0.41	254
6	0.79	0.86	0.82	6543
accuracy			0.71	17196
macro avg	0.66	0.71	0.67	17196
weighted avg	0.72	0.71	0.71	17196

Fig. 1. *Logistic Regression Report*

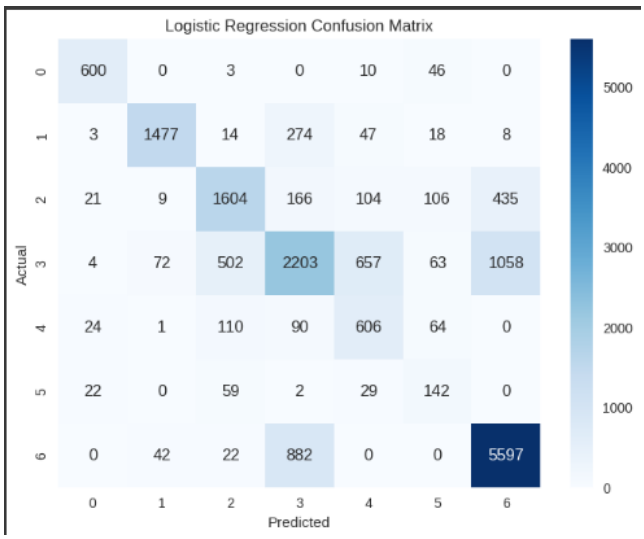

Fig. 2. *Logistic Regression Confusion Matrix*

Logistic Regression achieved 71% accuracy, but its confusion matrix revealed significant misclassification, particularly in categories 3 and 5, suggesting that crime categories were not linearly separable. The classification report indicated moderate precision and recall, with an overall F1-score of 0.71.

2. Multi-Layer Perceptron (MLP)

MLP Report:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	659
1	1.00	1.00	1.00	1841
2	0.93	0.93	0.93	2445
3	0.99	1.00	0.99	4559
4	0.91	0.86	0.88	895
5	0.78	0.81	0.79	254
6	0.99	0.99	0.99	6543
accuracy			0.98	17196
macro avg	0.94	0.94	0.94	17196
weighted avg	0.98	0.98	0.98	17196

MLP Log Loss: 0.1115

Fig. 3. *MLP Report*

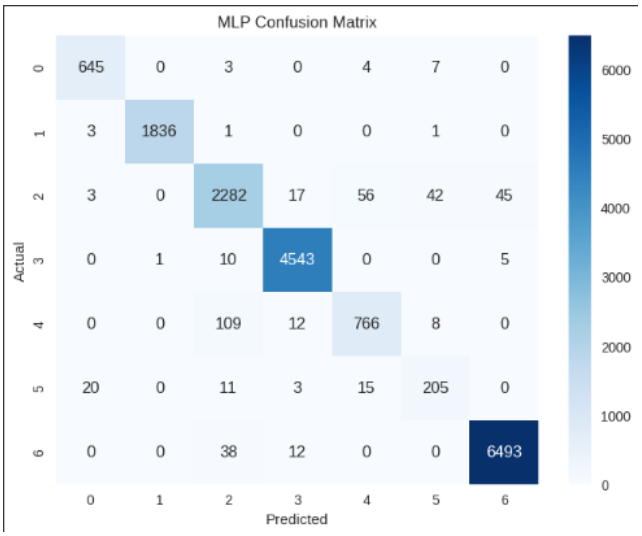

Fig. 4. *MLP Confusion Matrix*

The MLP model achieved the highest accuracy (98%), outperforming all other classifiers. It successfully learned complex patterns in the dataset, with minimal misclassification, as indicated by the confusion matrix. The classification report demonstrated consistently high precision, recall, and F1-score across all crime categories, confirming its effectiveness.

3. Support Vector Machine (SVM):

SVM Report:

	precision	recall	f1-score	support
0	0.94	0.93	0.94	659
1	0.94	0.99	0.96	1841
2	0.84	0.75	0.79	2445
3	0.92	0.87	0.89	4559
4	0.76	0.63	0.69	895
5	0.75	0.65	0.70	254
6	0.90	0.99	0.94	6543
accuracy			0.90	17196
macro avg	0.87	0.83	0.85	17196
weighted avg	0.89	0.90	0.89	17196

SVM Log Loss: 0.2965

Fig. 5. *SVM Report*

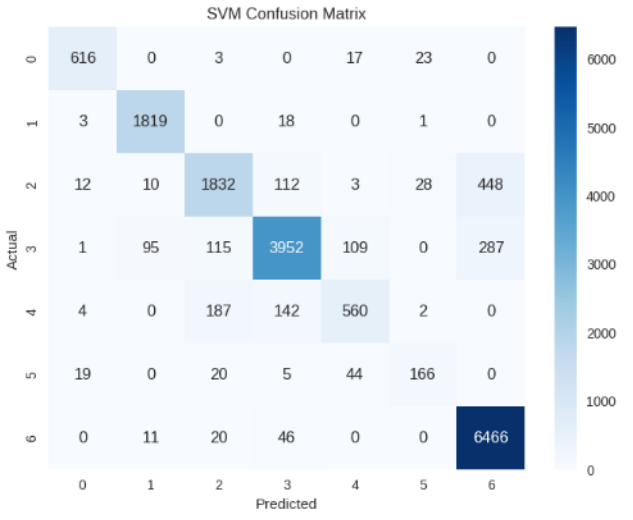

Fig. 6. *SVM Confusion Matrix*

SVM achieved 90% accuracy, showing strong classification capabilities but requiring higher computational resources. The confusion matrix sometimes indicated misclassification in categories 3 and 5, leading to a slightly reduced F1-score. The log loss was 0.2965, suggesting that while SVM performed well, its predictions were not as accurate as MLP's.

4. Decision Tree:

Decision Tree Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	659
1	1.00	1.00	1.00	1841
2	0.74	0.87	0.80	2445
3	1.00	0.84	0.92	4559
4	0.88	1.00	0.94	895
5	0.80	0.97	0.88	254
6	0.98	0.99	0.99	6543
accuracy			0.94	17196
macro avg	0.91	0.95	0.93	17196
weighted avg	0.95	0.94	0.94	17196

Fig. 7. *Decision Tree Report*

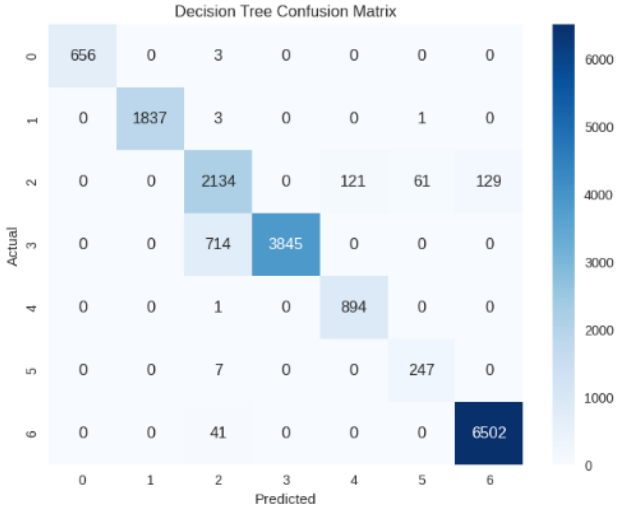

Fig. 8. *Decision Tree Confusion Matrix*

The Decision Tree model performed well, achieving 94% accuracy, with high precision, recall, and F1-score. The confusion matrix showed that most crimes were correctly classified, with minimum misclassifications. The model was easy to understand and explain, which made it a good choice for classifying crimes effectively.

5. Random Forest:

Random Forest Report:

	precision	recall	f1-score	support
0	0.89	0.97	0.93	659
1	0.96	0.99	0.98	1841
2	0.93	0.80	0.86	2445
3	1.00	0.94	0.97	4559
4	0.92	0.88	0.90	895
5	0.78	0.97	0.87	254
6	0.92	0.99	0.96	6543
accuracy			0.94	17196
macro avg	0.91	0.93	0.92	17196
weighted avg	0.94	0.94	0.94	17196

Fig. 9. *Random Forest Report*

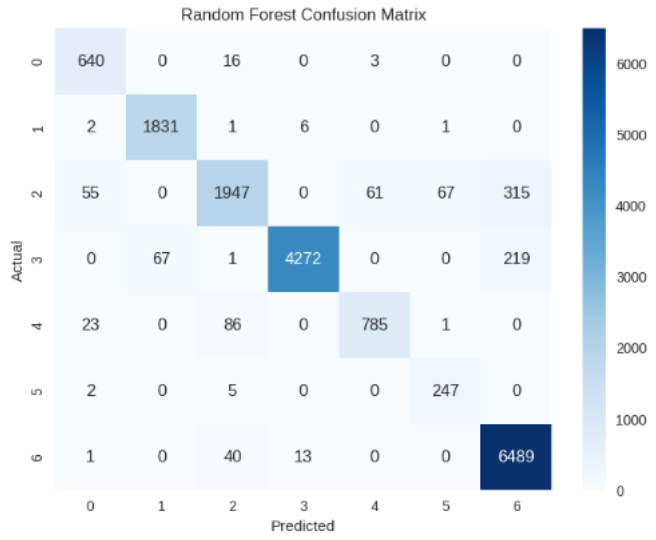

Fig. 10. *Random Forest Confusion Matrix*

The Random Forest model had 94% accuracy, like the Decision Tree, but it avoided overfitting by combining multiple trees. It made very few mistakes, as shown in the confusion matrix, and the classification report highlighted its strong and reliable performance for analyzing crime data.

6. K-Nearest Neighbors (KNN):

KNN Report:

	precision	recall	f1-score	support
0	0.95	0.96	0.95	659
1	0.98	0.97	0.98	1841
2	0.92	0.88	0.90	2445
3	0.95	0.97	0.96	4559
4	0.89	0.80	0.84	895
5	0.79	0.71	0.75	254
6	0.96	0.99	0.98	6543
accuracy			0.95	17196
macro avg	0.92	0.90	0.91	17196
weighted avg	0.95	0.95	0.95	17196

Fig. 11. *KNN Report*

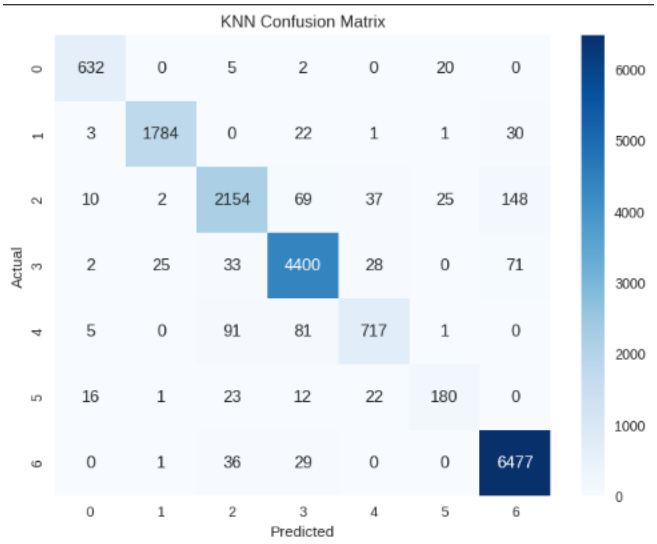

Fig. 12. *KNN Confusion Matrix*

KNN achieved 95% accuracy, showing that crime categories formed clear clusters. It made very few mistakes, as seen in the confusion matrix, and the classification report showed high precision, recall, and F1-scores for all categories, confirming its strong performance.

B. Unsupervised Clustering Performance

1. K-Means Clustering:

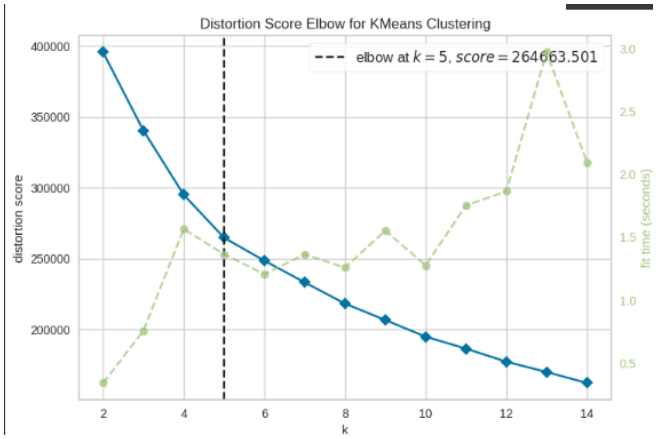

Fig. 13. *Elbow Method for K-Means Optimization*

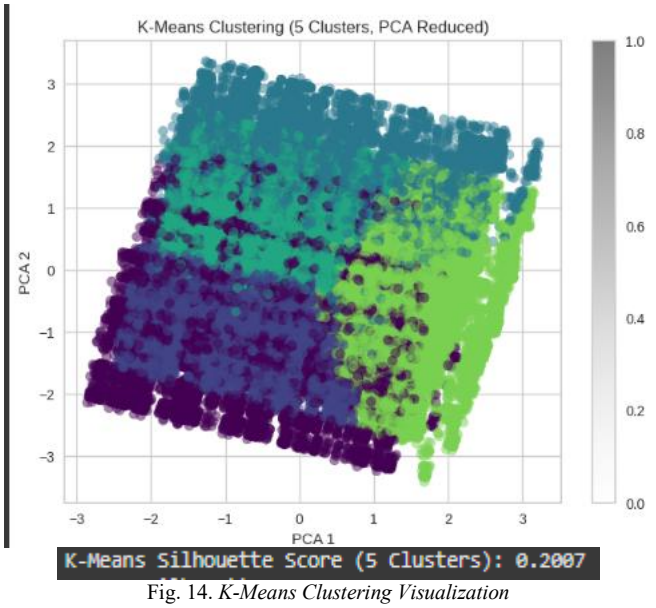

Fig. 14. *K-Means Clustering Visualization*

K-Means was used to find hidden crime patterns, with 5 clusters chosen as the best number using the Elbow Method. However, the low silhouette score of 0.2007 showed that the clusters were not very distinct, meaning crime types had a lot of overlap in their features.

2. DBSCAN:

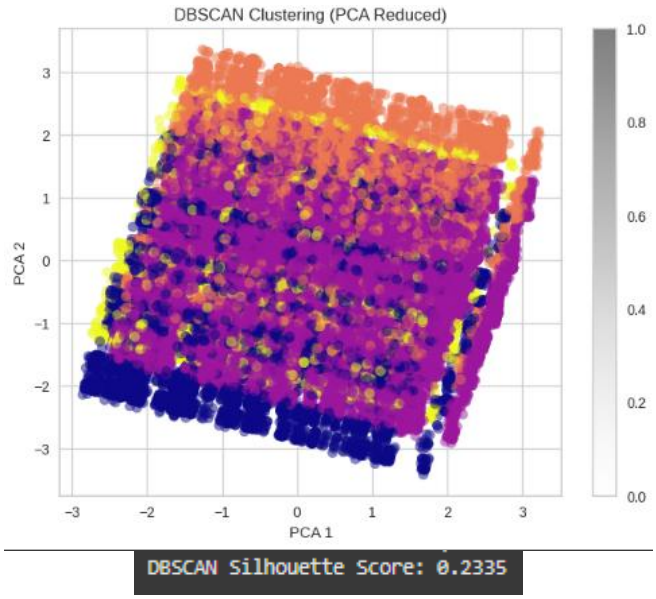

Fig. 15. *DBSCAN Clustering Visualization*

DBSCAN was implemented to detect crime hotspots and outliers. Unlike K-Means, DBSCAN does not require the number of clusters to be set in advance. It successfully found areas with a high concentration of crimes and got a better silhouette score (0.2335) than K-Means, making it more effective for spotting crime patterns.

V. Conclusion

The project shows that machine learning models are highly effective for classifying crimes and spotting patterns in crime data. Based on the results, the Multi-Layer Perceptron (MLP) stood out among supervised models, achieving the highest accuracy of 98%, thanks to its ability to learn and understand complex relationships in the data. On the other hand, when it comes to clustering, DBSCAN performed better than K-Means, with a higher silhouette score of 0.2335 compared to 0.2007 for K-Means. This indicates that DBSCAN was more effective at finding crime hotspots and detecting unusual patterns in the data.

These findings can help law enforcement agencies make smarter decisions by using data. They could apply these insights to predict future crimes, allocate resources more effectively, and analyze trends to improve public safety measures. Although the results are promising, there are areas for future improvement. Enhancing deep learning models by exploring other architectures can improve the accuracy and performance to become even more powerful tools for crime analysis, aiding law enforcement agencies in making data-driven decisions to improve public safety.

References

[1] IBM, "Supervised Learning: What it is, How it Works, and Examples," IBM Think, 2024 https://www.ibm.com/think/topics/supervised-learning.

[2] Scikit-Learn Documentation, "Supervised Learning," Scikit-Learn, 2024. https://scikit-learn.org/stable/supervised_learning.html.

[3] Scikit-Learn Documentation, "Unsupervised Learning," Scikit-Learn, 2024. https://scikit-learn.org/stable/unsupervised_learning.html.

[4] U.S. Government, "Crimes - One Year Prior to Present," *City of Chicago Data Portal*, 2025. https://data.cityofchicago.org/Public-Safety/Crimes-One-year-prior-to-present/x2n5-8w5q/about_data.

[5] N. Shah, N. Bhagat, and M. Shah, "Crime Forecasting: A Machine Learning and Computer Vision Approach to Crime Prediction and Prevention," *Journal of Machine Learning Applications*, vol. X, no. Y, 2021. https://www.semanticscholar.org/paper/Crime-forecasting%3A-a-machine-learning-and-computer-Shah-Bhagat/5d1797987692b0be3e84abee0c3a37d1b7b50913