جامعة التقنية
والعلوم التطبيقية
University of Technology
and Applied Sciences
نزوى
Nizwa

Crime Data Analysis and Prediction

Students:

- Sohaib Al-Adawi 76s198
- Mohammed Al-Hadidi 26s2053

Date: [14/5/2025]

Table of Contents

1. Abstract

This project focuses on analyzing and predicting crime occurrences using a dataset of reported incidents. The process includes data cleaning, preprocessing, and modeling using machine learning techniques. Two main objectives are addressed: predicting whether an event is criminal and detecting crime hotspots using clustering techniques. The Random Forest classifier is used for classification tasks, and HDBSCAN is applied for spatial clustering. Key steps include handling missing values, encoding categorical variables, and scaling features. Evaluation metrics such as the classification report and the silhouette score are used. The project demonstrates the effectiveness of ML in crime prediction and geographical crime pattern detection.

4. Introduction

Background

Urban crime is a persistent issue affecting public safety. Identifying patterns in crime data helps law enforcement and urban planners to allocate resources efficiently.

Objectives

- Predict whether a reported incident is criminal.
- Identify geographical crime hotspots.

Importance

Real-time crime prediction can support proactive policing, and hotspot detection can guide surveillance deployment.

Methodology Overview

- Clean and preprocess the dataset.
- Use Random Forest for crime occurrence prediction.
- Use HDBSCAN for hotspot detection.
- Evaluate model accuracy and clustering quality.

5. Data Collection and Preprocessing

Dataset

- Dataset: `Crime_Dataset.csv`
- Fields include: Date, Latitude, Longitude, Category, Resolution, etc.

Preprocessing Steps

- **Duplicate Removal:** All duplicated rows were removed.
- **Missing Value Handling:** 'NONE' values were replaced with NaN, then dropped.
- **Column Renaming:** Latitude and Longitude column names corrected.
- **Time Features:** Extracted Hour, DayOfWeek, Month, and Year from Dates.
- **Label Encoding:** Categorical columns encoded using `LabelEncoder`.
- **Target Variable:** Binary target `Crime_Occurred` created.
- **Scaling:** Applied `StandardScaler` to numerical features.

6. Methodology

Classification Model

- **Algorithm:** Random Forest Classifier
- **Task:** Predict `Crime_Occurred` (binary classification)
- **Feature Selection:** All numerical and encoded categorical features.
- **Train-Test Split:** 80:20 ratio
- **Hyperparameter Tuning:** Used `GridSearchCV` with a parameter grid:
    - `n_estimators`: [50, 100]
    - `max_depth`: [10, 20, None]
    - `min_samples_split`: [2, 5]
    - `min_samples_leaf`: [1, 2]

Clustering Model

- **Algorithm:** HDBSCAN
- **Features Used:** Latitude, Longitude

- **Filtering:** Kept data within San Francisco's geographical bounds.
- **Clustering Objective:** Identify geographical crime hotspots.

Evaluation Metrics

- **Classification:**
  - `classification_report` (precision, recall, F1-score)
  - `roc_auc_score` (via GridSearchCV)
- **Clustering:**
  - Silhouette Score (ignoring noise points)

7. Model Evaluation

Classification Model

- **Algorithm Used:** Random Forest
- **Classification Report Output:** Provided precision, recall, and F1-score.
- **ROC-AUC Optimization:** Best parameters identified using GridSearchCV.
- **Feature Importance:** Top features include Latitude, Longitude, and time-based variables.

Clustering Model

- **Silhouette Score:** Reported score for non-noise points, indicating clustering quality.
- **Visualization:** Latitude vs. Longitude plot showing clusters with color-coded IDs.

8. Analysis and Conclusion

Key Findings

- Random Forest classifier accurately predicts crime occurrences based on location and time features.

- Feature importance shows that geographical coordinates and time (hour, day) are strong indicators of crime likelihood.

- HDBSCAN effectively identifies spatial clusters of crime, useful for pinpointing hotspots.