**Crime Detection**

**Introduction**

In recent years, rapid urbanization has significantly increased the global urban population, creating a growing demand for secure, hospitable, and sustainable urban environments. As cities continue to expand—often absorbing surrounding suburbs and rural areas—managing this urban growth has become an increasingly complex challenge for administrative bodies. Overpopulation in urban centers has compelled governments to implement smart city initiatives aimed at improving infrastructure management and addressing key issues such as security, sustainability, and development.

While these smart city initiatives hold great promise for enhancing quality of life, they also introduce various challenges—chief among them being public safety. The increasing complexity of urban life has prompted researchers to explore crime patterns and their correlation with socioeconomic factors such as education levels, family structures, and human behavioral characteristics. Understanding these relationships is essential to improving security in evolving metropolitan areas.

In this context, data mining has emerged as a powerful field that combines methods from machine learning and database systems to identify patterns in large datasets. With its wide range of applications—including future healthcare, market basket analysis, education, manufacturing, and crime investigation—data mining offers valuable tools for analyzing crime characteristics. Specifically, in the domain of public safety, data mining techniques can support crime detection efforts, enabling authorities to proactively identify threats and foster safer societies.

This report aims to achieve higher accuracy and create effective insights in the field of crime detection using machine learning and business analytics methods by working on 2 data sets.

**Literature Review**

Existing crime prediction methods can generally be categorized into two main approaches: traditional machine learning techniques and modern deep learning-based methods. Each approach offers its own unique strengths and challenges. According to a comprehensive survey conducted by one of the key limitations in this field is the limited scale and quality of publicly available datasets, which restricts model generalization and real-world applicability. To overcome this issue,proposed a big data research framework based on machine learning that can better support future studies and guide researchers in developing more robust crime prediction systems Yin (2022).

Another study emphasizes the importance of spatial-temporal dependencies in crime data, highlighting that crimes tend to cluster in both space and time. However, most machine learning models struggle to effectively capture this dependency. To address this limitation, the use of spatiotemporal lag variables has proven beneficial. An empirical study conducted in Dallas between 2014 and 2018 revealed that incorporating such variables significantly improved model accuracy. The findings also indicated that crime predictors exhibit strong nonlinearity over time and space, with tree-based models (e.g., Random Forest, XGBoost) outperforming linear models. Moreover, interpretable models provided meaningful insights into how various variables influence crime risk, assisting both researchers and law enforcement in developing targeted prevention strategies (Deng, He, & Liu, 2023).

Other research efforts have explored a wide range of machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Naïve Bayes, k-Nearest Neighbors (KNN), Decision Trees, Multilayer Perceptrons (MLP), Random Forest, and eXtreme Gradient Boosting (XGBoost). In the domain of time series analysis, models such as Long Short-Term Memory (LSTM) networks and Autoregressive Integrated Moving Average (ARIMA) have also been employed to predict crime trends. For instance, an empirical study covering crime data from Chicago and Los Angeles demonstrated that LSTM and ARIMA models were effective in capturing temporal crime patterns. The results indicated a general decline in crime rates in Los Angeles and seasonal fluctuations in both cities. These predictive insights provide valuable information for identifying crime hotspots and optimizing police resource allocation (Safat, Asghar, & Gillani, 2021).

Furthermore, data mining has been identified as a critical component in the crime analysis process. As a multidisciplinary field at the intersection of machine learning and database systems, data mining helps uncover hidden patterns and relationships within large datasets. Its applications in crime detection have shown promise in identifying meaningful crime characteristics and designing effective strategies to combat criminal activity. Research on data mining applications in crime investigation suggests that these methods can significantly enhance situational awareness and decision-making processes (Prabakaran & Mitra, n.d.).

Finally, a study by Palanivinayagam et al. (2023) highlighted the importance of feature generation in improving the accuracy of crime detection systems. They developed a methodology to extract key attributes such as time zones, crime probability, and hotspot vulnerability. When applied to standard datasets (e.g., San Francisco crime data), their approach achieved a remarkable accuracy of 97.5% using the Naïve Bayes algorithm. This finding underscores the potential of intelligent feature engineering in enhancing machine learning performance in crime prediction tasks (Palanivinayagam et al., 2023).

**Methodology**

**About Dataset**

1. **Cambridge Crime Data**

This dataset comprises crime incidents reported in the City of Cambridge, as featured in the Cambridge Police Department’s Annual Crime Reports, spanning from 2009 to 2024. The data provides detailed information about various crime types and their occurrences across different neighborhoods in Cambridge. (Cambridge Crime Data)

Dataset Details:

* File Number: A unique identifier for each crime report (Text).
* Date of Report: The date when the crime incident was reported (Floating Timestamp).
* Crime Data Time: The specific time when the crime incident occurred (Floating Timestamp).
* Crime: The type of crime committed (Text).
* Reporting Area: A numerical identifier for the community area where the crime occurred (Number).
* Neighborhood: The name of the neighborhood where the crime was reported (Text).
* Location: The street information indicating the approximate location of the crime (Text).

The dataset includes records of over 40 different types of crimes. However, certain crime types are excluded to maintain confidentiality and protect privacy rights. It is important to note that the addresses provided do not represent the exact locations of the crimes but are approximations within 100-block ranges.

1. **Crime Dataset**

This dataset provides detailed information about criminal incidents, capturing various characteristics of both the offenders and victims. It includes records of crimes along with demographic details such as age, gender, race, and the status of the individuals involved. The data also contains information on the disposition of the case (whether it was closed or open) and the nature of the crime. (Crime-data)

The dataset covers a wide range of crime categories such as theft, vandalism, violence, sexual crimes, and drug/weapon-related offenses. This allows for an in-depth analysis of criminal activities, their impact on different demographics, and potential correlations between various factors such as age, gender, and the type of crime committed.

Columns:

* Disposition: The current status of the case (Closed or Open).
* OffenderStatus: The status of the offender (e.g., ARRESTED).
* Offender\_Race: The race of the offender (e.g., BLACK, WHITE, ASIAN, etc.).
* Offender\_Gender: The gender of the offender (MALE or FEMALE).
* Offender\_Age: The age of the offender (numerical value).
* PersonType: Type of person involved in the case (e.g., VICTIM, REPORTING PERSON, INTERVIEW).
* Victim\_Race: The race of the victim (e.g., BLACK, WHITE, ASIAN, etc.).
* Victim\_Gender: The gender of the victim (MALE or FEMALE).
* Victim\_Age: The age of the victim (numerical value).
* Victim\_Fatal\_Status: Indicates if the victim’s injuries were fatal or non-fatal.
* Report Type: Type of report filed (e.g., Supplemental Report, Incident Report).
* Category: The category of crime (e.g., Theft, Vandalism, Violence, etc.).

**Data Preprocessing**

* 1. **Cambridge Crime Dataset**

In the Cambridge Crime dataset, the exploratory analysis began with a bar graph showing the 10 most common types of crime and revealing the concentration in several basic categories (Fig 1).

metin, ekran görüntüsü, yazı tipi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 1)

Two word clouds were created, one to visualize which neighborhoods reported the most crimes, and the other to highlight the most frequent types of crimes (Fig 2), (Fig 3)

metin, yazı tipi, tipografi, tasarım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 2)

metin, yazı tipi, ekran görüntüsü, tipografi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 3)

For spatial analysis, a random sample of 3,600 addresses was geocoded to obtain latitude and longitude data, and unsuccessful searches were removed. Using Folium, these locations were mapped with color-coded markers by crime type, providing a clear and interactive view of both spatial distribution and crime diversity throughout the city (Fig 4).

metin, harita, atlas içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 4)

* 1. **Crime Dataset**

The dataset was initially evaluated for completeness and consistency through descriptive statistics (Fig 5).

metin, ekran görüntüsü, yazı tipi, siyah içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 5)

Various visual analyses were conducted to gain an initial understanding of the dataset's structure and internal distribution.

The distribution of offender ages revealed a concentration in the 30–35 age range with a slight rightward skew, indicating that most crimes were committed by individuals within this age group (Fig 6).

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, plan içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 6)

A gender-based comparison highlighted a significant imbalance, emphasizing that male offenders vastly outnumbered females (Fig 7).

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 7)

In terms of racial distribution, the majority of offenders were identified as Black, as demonstrated through proportional visualization (Fig 8).

metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 8)

Furthermore, an analysis of crime categories in relation to case disposition showed that certain types of crimes had a significantly higher likelihood of being closed, indicating patterns in law enforcement response or procedural outcomes (Fig 9).

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Fig 9)

the frequency distribution of the "Neighborhood" variable was calculated to identify the number of records associated with each region Map-Reduce. Additionally, entries with missing neighborhood values (NaN) were quantified, revealing 8 instances of incomplete data. (Source code)



**Modelling and Evaluation**

* 1. **Crime Dataset**

In the modeling phase, the objective was to predict the Disposition of criminal cases—whether they would be classified as open (1) or closed (0). A key challenge encountered during this phase was the severe class imbalance, as visualized in the class distribution plot, where closed cases significantly outnumbered open ones. To mitigate this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data, generating synthetic instances of the minority class to balance the dataset and improve model generalization.

Three different classification algorithms were employed to evaluate the performance on the resampled data:

* Logistic regression

Logistic regression is a simple yet powerful classification algorithm used for decision-making problems where the outcome variable is categorical. It is a statistical method for analyzing a dataset in which there is one or more independent variables that determine the outcome. The main objective is to find the best-fitting model that describes the relationship between the dependent variable and the independent variables.

The logistic regression model estimates the probability that a given input point belongs to a particular class by using the following equation:

logit(p) = ln(p / (1 - p)) = β₀ + β₁X₁ + β₂X₂ + β₃X₃ + ... + βₖXₖ

* K nearest neighbors

K nearest neighbors (KNN) is a basic and simple supervised machine learning methods that can be used for both regression and classification tasks. In a regression task, its core step is to find the k nearest training samples to the predicted sample, and the outcome of the predicted sample is the average of the labels of the k nearest training samples. The KNN method is inefficient due to the needs to calculate the distance between all training and prediction samples separately and to find the top k nearest training samples to the prediction samples. To improve the training efficiency, this study uses the K-D tree to find the nearest neighbor samples. (Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models)

* Random forest

Random forest (RF) is a representative integrated machine learning approach that uses d ecision trees as a base predictor . There are two ways to interpret the term ‘random’. First, the sample features are selected randomly. That is, some, but not all, features are chosen randomly to serve as training features for the decision tree. Second, the training samples count is determined randomly. In other words, training samples for constructing each decision tree are chosen randomly from the training dataset. Eventually, the term “forest” refers to generating multiple regression trees for the samples and combining the predictions of these trees to obtain the result by voting. In addition, RF is insensitive to multicollinearity, relatively robust to missing and unbalanced data, and yields reasonable predictions. (Yin, J.)

4.2 Evaluation Metrics

Our evaluation metrics include accuracy, precision, and recall. The outputs of all classifiers are binary; hence, the following definitions are used for classification outcomes:

• True Positive (TP): when a crime event is predicted as a crime event

• True Negative (TN): when a non-crime event is predicted as a non-crime event

• False Positive (FP): when a non-crime event is predicted as a crime event

• False Negative (FN): when a crime event is predicted as a non-crime event

Accuracy is defined as the quality of correctness in predictions. It is calculated using the following formula:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision explains how many of the predicted positive cases are actually positive. It is given by:

Precision = TP / (TP + FP)

Recall measures how many of the actual positive cases were correctly predicted by the classifier. It is given by:

Recall = TP / (TP + FN) (An Optimized Machine Learning and Big Data Approach to Crime Detection)

# Results

## Random Forest

Accuracy: 0.9149

Classification Report:

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score |
| Class 0 | 0.96 | 0.95 | 0.95 |
| Class 1 | 0.27 | 0.31 | 0.29 |

Confusion Matrix:

|  |  |
| --- | --- |
| 1192 | 61 |
| 52 | 23 |

## Logistic Regression

Accuracy: 0.6777

Classification Report:

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score |
| Class 0 | 0.97 | 0.68 | 0.8 |
| Class 1 | 0.1 | 0.6 | 0.17 |

Confusion Matrix:

|  |  |
| --- | --- |
| 855 | 398 |
| 30 | 45 |

## KNN Classifier

Accuracy: 0.8087

Classification Report:

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score |
| Class 0 | 0.96 | 0.84 | 0.89 |
| Class 1 | 0.12 | 0.36 | 0.18 |

Confusion Matrix:

|  |  |
| --- | --- |
| 1047 | 206 |
| 48 | 27 |

Among the models tested, Random Forest showed the highest overall accuracy and balanced performance for the majority class, though it struggled with recall for the minority class. Logistic Regression had the highest recall for the minority class (open cases) but at the cost of overall accuracy. KNN provided a moderate balance between the two, making it suitable in scenarios where trade-offs between class performance are acceptable. These results highlight the challenge of handling imbalanced data and the importance of choosing models based on specific objectives (e.g., identifying open cases vs. overall reliability).

**References**

Cambridge Crime Data: [www.kaggle.com/datasets/melissamonfared/cambridge-crime-data-2009-2024/data](http://www.kaggle.com/datasets/melissamonfared/cambridge-crime-data-2009-2024/data)

Crime Data: <https://www.kaggle.com/datasets/fatmanur12/crime-data>

Deng, Y., He, R., & Liu, Y. (2023). *Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models*. Information Sciences, **646**, 119414. <https://doi.org/10.1016/j.ins.2023.119414>

Palanivinayagam, A., Gopal, S. S., Bhattacharya, S., Anumbe, N., Ibekwe, E., & Biamba, C. (2023). *An optimized machine learning and big data approach to crime detection*. Wiley Online Library. <https://doi.org/10.1002/cpe.7686>

Prabakaran, S., & Mitra, S. (n.d.). *Survey of analysis of crime detection techniques using data mining and machine learning*. SRM Institute of Science and Technology.

Safat, W., Asghar, S., & Gillani, S. A. (2021). *Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques*. IEEE Access, **9**, 74411–74426. <https://doi.org/10.1109/ACCESS.2021.3078117>

Source code: <https://github.com/SalihBirdal2434/CrimeDetection>

Yin, J. (2022). *Crime prediction methods based on machine learning: A survey*. Computers, Materials & Continua, **73**(3), 5795–5811. <https://doi.org/10.32604/cmc.2023.034190>