Crime Forecasting and Analysis: A Machine Learning Perspective

[1]Soumadip Saha, [2]Aniket Hazra

[1]Student, [2]Student
[1]Department of Computer Science and Engineering,
[1]Institute of Engineering & Management, Kolkata, India
[1]Soumadip.Saha2021@iem.edu.in, [2]Aniket.Hazra2021@iem.edu.in

***Abstract*: Predicting crime remains a major challenge for ensuring community safety, as traditional methods often fail to capture the complexities of modern urban environments. This research proposes a comprehensive machine learning and time series-based framework to enhance crime prediction capabilities. The study has three primary objectives: (1) to use supervised learning algorithms to predict crime types based on historical data, including features such as time, location, and previous crime patterns; (2) to apply spatial analysis techniques to identify and map high-risk crime zones using the Indian Crime Dataset; and (3) to forecast future crime trends through time series modeling, thereby supporting more effective resource allocation and strategic planning by law enforcement agencies. By integrating data-driven approaches, this research aims to improve the accuracy and reliability of crime forecasting models. The framework also evaluates the performance of different machine learning models to determine their suitability for large-scale urban crime analysis, with the ultimate goal of strengthening proactive crime prevention strategies and enhancing public safety outcomes.**

***Keywords*: Artificial Neural Networks, Crime Prediction, Machine Learning, Data Analysis, Law Enforcement**

INTRODUCTION

Ensuring public safety remains a major challenge for governments and law enforcement worldwide. While traditional crime analysis offers insights using historical data and statistics, it often fails to adapt to evolving crime patterns influenced by social, economic, and environmental factors.

This paper proposes a data-driven framework using machine learning (ML) to predict crime trends based on Indian crime data from 2001–2012. The study focuses on three main objectives:

- Predicting crime types using supervised learning (e.g., Random Forest, SVM).
- Identifying high-risk areas through spatial analysis and clustering.
- Forecasting future crime trends using time series models like Holt-Winters.

Supervised learning uncovers hidden patterns in crime categories, dates, and locations, aiding in accurate predictions and resource allocation. Spatial analysis identifies crime hotspots for proactive policing. Time series forecasting captures seasonal trends, helping plan for crime surges. Clustering groups crimes by similarities in time, location, and type, revealing overlooked patterns. PCA is used to reduce dimensionality in large datasets, maintaining efficiency.

Unlike static traditional methods, this integrated ML approach offers a dynamic, data-driven understanding of crime. It enables law enforcement to move from reactive to proactive strategies, improving public safety and trust.

NEED OF THE STUDY

The paper "A Machine Learning Approach to Crime Analysis and Forecasting for Prediction and Prevention" by R. Raj Bharath et al. (2024) presents a framework combining supervised learning, spatial analysis, and time-series forecasting. Using the Indian Crime dataset, Random Forest achieved 98% accuracy in crime type prediction, and KNN 97% in location prediction. Holt-Winters Exponential Smoothing effectively captured seasonal trends for better resource allocation.

Wajiha Safat et al., in "Empirical Analysis for Crime Prediction and Forecasting", examined models like Logistic Regression, Decision Trees, LSTM, and ARIMA on Indian crime datasets. Their study emphasized forecasting accuracy and hotspot identification in urban areas. Neil Shah et al., in "Crime Forecasting: A Machine Learning and Computer Vision Approach", integrated ML with computer vision techniques such as face recognition and motion tracking. The study highlighted real-time crime detection using models like KNN, Decision Trees, and Neural Networks.

Karabo Jenga et al., in "Machine Learning in Crime Prediction", reviewed ML approaches for suspect identification, spatial-temporal hotspot detection, and social media-based trend analysis. They emphasized the need for large datasets and ethical AI use. Azeez and Aravindhar, in "Hybrid Approach to Crime Prediction Using Deep Learning", proposed a hybrid model using historical data, geospatial info, and NLP on social media posts. They noted the potential of semantic analysis but acknowledged challenges in bias and privacy.

Kshatri et al., in "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization", introduced a stacked ensemble method (SBCPM) using SVM, achieving 99.5% accuracy in crime classification. Tamir et al. focused on crime severity prediction using Random Forest, KNN, AdaBoost, and Neural Networks on Chicago's CLEAR dataset. Neural Networks achieved 90.77% accuracy in predicting arrest outcomes, aiding resource planning.

Lastly, "Predictive Analysis and Prognostication" by Sheriffdeen K and Neveah B explored AI-driven forecasting in San Francisco, stressing the importance of transparency, bias mitigation, and ethical considerations in predictive policing.

RESEARCH METHODOLOGY

I. DATA DESCRIPTION

This study uses the publicly available Indian Crime Dataset from Kaggle, which provides detailed records of reported crimes across Indian states.

- **Dataset Overview:** The dataset contains over a thousand entries, each representing a crime report. It includes key attributes like crime type, number of incidents, and location.
- **Data Source:** Sourced from Kaggle, the dataset covers a wide range of crimes, including theft, assault, robbery, and homicide. • Data Format: Structured as a CSV file, the data is organized in tabular form, with each row representing a crime incident and columns for specific attributes.
- **Temporal Coverage:** Spanning from January 2001 to 2012, the dataset supports long-term trend analysis and time series forecasting.
- **Spatial Coverage:** Crime location data is included for all Indian states, enabling spatial analysis and the identification of high-risk areas.

II. DATA PREPROCESSING

Before using the Indian Crime Dataset for analysis and modeling, several preprocessing steps were necessary to clean and transform the data into a suitable format for machine learning. The dataset contains extensive details on reported crimes, including crime type, location, and timestamps. However, to ensure accuracy and relevance, the following steps were taken:

- **Data Cleaning**: Removing Duplicates - Any duplicate records were identified and eliminated to maintain data integrity. Handling Missing Values - Missing data were either filled using appropriate methods or removed, depending on the extent and nature of the missing values.
- **Feature Selection**: Not all attributes in the dataset were useful for analysis. Irrelevant features, such as case numbers and block addresses, were excluded to focus on variables that provide meaningful insights into crime prediction and trends. • Feature Engineering - To improve the predictive power of the models, new features were created.
- **Normalization:** Numerical features were scaled to ensure a consistent range, preventing any single feature from disproportionately influencing the model due to differences in magnitude.
- **Encoding Categorical Variables**: Categorical data, such as crime type and location, were converted into numerical representations using techniques like one-hot encoding, making them compatible with machine learning algorithms.

These preprocessing steps were essential in refining the dataset, ensuring it was clean, relevant, and ready for training machine learning models to predict and prevent crime more effectively.

III. CRIME TYPE PREDICTION MODEL

Random Forest is a powerful machine learning technique that improves prediction accuracy by combining multiple decision trees. Each tree in the model is trained on a random subset of the data and a random selection of features, which helps prevent overfitting and ensures generalization to new data.

In our study, we used the Random Forest algorithm to predict crime types based on factors like location, time of occurrence, and other relevant details. The model was trained on a subset of the Indian crime dataset, with each tree learning from a different portion of the data. To make a final prediction, the model aggregated the outputs of all trees—typically using majority voting for classification.

Feature Name	Data Type	Description
STATE/UT	String	State or Union Territory where the data was recorded
YEAR	Integer	Year of recorded crime data
MURDER	Integer	Number of reported murders
ATTEMPT TO MURDER	Integer	Number of reported attempted murders
CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	Integer	Number of reported culpable homicides not amounting to murder
RAPE	Integer	Number of reported rapes
CUSTODIAL RAPE	Integer	Number of custodial rapes
OTHER RAPE	Integer	Number of other reported rapes
KIDNAPPING & ABDUCTION	Integer	Total kidnapping and abduction cases
KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS	Integer	Kidnapping and abduction of women and girls
KIDNAPPING AND ABDUCTION OF OTHERS	Integer	Kidnapping and abduction of others
DACOITY	Integer	Total dacoity cases
PREPARATION AND ASSEMBLY FOR DACOITY	Integer	Preparation and assembly for dacoity
ROBBERY	Integer	Number of reported robberies
BURGLARY	Integer	Number of reported burglaries
THEFT	Integer	Total reported thefts
AUTO THEFT	Integer	Number of auto thefts
OTHER THEFT	Integer	Other thefts excluding auto theft
RIOTS	Integer	Number of reported riots
CRIMINAL BREACH OF TRUST	Integer	Criminal breach of trust cases
CHEATING	Integer	Cases of cheating
COUNTERFEITING	Integer	Counterfeiting cases
ARSON	Integer	Arson cases
HURT/GREVIOUS HURT	Integer	Hurt or grievous hurt cases
DOWRY DEATHS	Integer	Dowry deaths
ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	Integer	Assault on women with intent to outrage modesty
INSULT TO MODESTY OF WOMEN	Integer	Insult to modesty of women
CRUELTY BY HUSBAND OR HIS RELATIVES	Integer	Cruelty by husband or his relatives
IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES	Integer	Importation of girls from foreign countries
CAUSING DEATH BY NEGLIGENCE	Integer	Causing death by negligence
OTHER IPC CRIMES	Integer	Other IPC (Indian Penal Code) crimes
TOTAL IPC CRIMES	Integer	Total IPC crimes recorded

Figure 1: Features of the dataset

We chose Random Forest because it performs well with large, high-dimensional datasets and is capable of capturing complex patterns in the data. Additionally, its interpretability and ease of use make it a reliable choice for crime prediction and other classification tasks in machine learning.

IV. CRIME LOCATION PREDICTION MODEL

The K-Nearest Neighbors (KNN) algorithm is a straightforward yet powerful machine learning technique used for classification and regression. In our project, we used KNN to predict crime locations based on past incidents recorded in the Indian crime dataset.

KNN works by identifying the K closest data points (or "neighbors") to a given input and using their labels to determine the prediction. In our case, these neighbors represent past crime incidents with similar characteristics—such as crime type, time of occurrence, and geographical coordinates. By analyzing these similarities, the model can predict where similar crimes might happen in the future.

We chose KNN because it is simple, effective, and well-suited for handling spatial data. By capturing patterns in crime distribution, this approach can help law enforcement agencies better allocate resources and improve crime prevention efforts.

V. TIME SERIES FORECASTING MODEL

In the given notebook, the Holt-Winters Exponential Smoothing model is applied for time series forecasting of yearly crime counts. The dataset is loaded into a Pandas DataFrame, with the date column parsed as a datetime index. To prepare the data for forecasting, it is resampled into different timeframes, including yearly, monthly, and quarterly aggregates, ensuring the model effectively captures trends and seasonal variations. The Holt-Winters model used follows either the additive or multiplicative trend and seasonality, depending on the characteristics of the dataset. This allows it to adapt to long-term trends and cyclic patterns.

The model is fitted to the yearly resampled crime data, learning historical patterns to project future trends. After fitting, crime count forecasts are generated for future years, with prediction intervals indicating uncertainty. The results are visualized using Matplotlib, with actual crime counts plotted alongside the forecasted values. The shaded intervals highlight the reliability of predictions. This Holt-Winters implementation enhances crime trend forecasting, aiding law enforcement and policymakers in proactive planning and resource allocation.

RESULTS

This section presents the findings of our investigation into the effectiveness of various machine learning algorithms for crime prediction and location-based risk assessment. Using a comprehensive crime dataset, we evaluated multiple models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees, based on key performance metrics such as accuracy, precision, recall, and F1-score.

This approach provided valuable insights into seasonal patterns and fluctuations in crime occurrences, aiding in proactive decision-making for crime prevention and resource allocation.

I. CRIME TYPE PREDICTION MODEL

We tested several supervised learning algorithms on crime data to see which one performed best. The algorithms included Support Vector Machines (SVM), Decision Trees, Random Forests, and Logistic Regression. We evaluated them using accuracy, recall, precision, and F1-score.

Among these, Random Forests stood out as the most effective, achieving an impressive 96% accuracy in predicting crime types. This means it was the most reliable at classifying different crimes based on the given data. As shown in Table I, Random Forests consistently outperformed the other models across all evaluation metrics, demonstrating its strong ability to recognize patterns in crime data and make accurate predictions.

Table 1: Performance of Various ML Algorithms for Crime Type Prediction (in %)

Model	Accuracy
Logistic Regression	97
Random Forest	96
Decision Tree	92
Support Vector Machine (SVM)	90

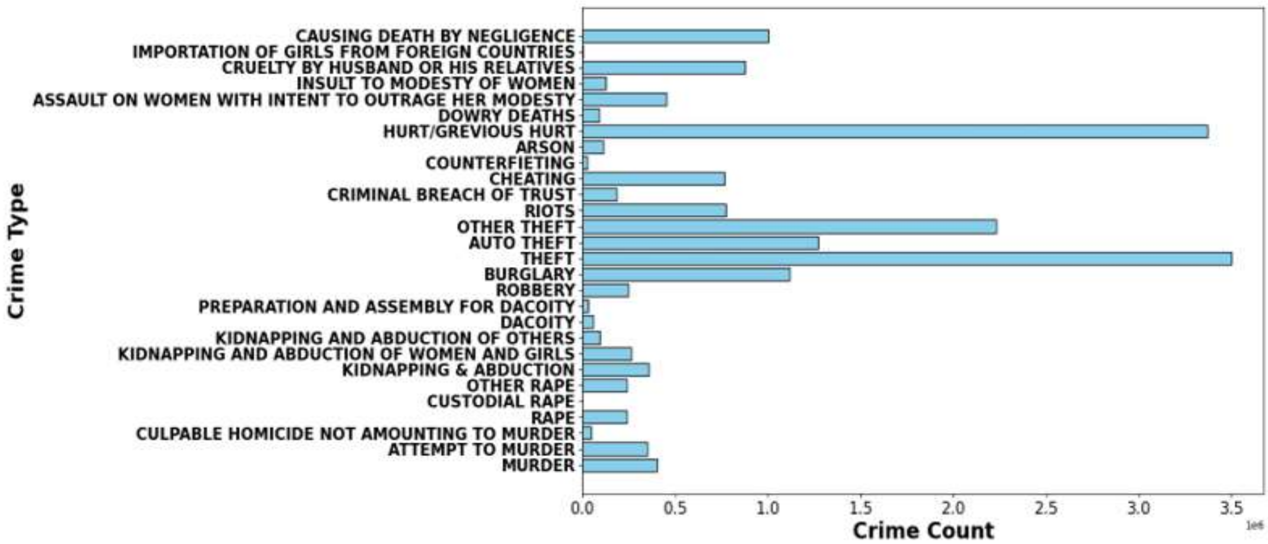

Figure 2: Primary Crime Types Visualization

Figure 2 shows the distribution of the most common crime types in our dataset. The x-axis represents different types of crimes, while the y-axis shows how many incidents were reported for each. Our analysis found that Theft is the most frequent crime. This suggests that prioritizing resources and prevention efforts on theft could have a significant impact on reducing overall crime rates.

II. CRIME LOCATION PREDICTION

This section presents the results of our evaluation of different algorithms in identifying high-risk crime zones within the dataset. We tested several models, including Linear Regression, Logistic Regression, K Nearest Neighbors (KNN), and Support Vector Machines (SVM), using accuracy, recall, precision, and F1-score as performance metrics.

Our results showed that Logistic Regression was the most effective algorithm for location prediction, achieving an impressive 97% accuracy. This suggests that KNN can reliably classify locations as high-risk or low-risk based on the provided data.

Table 2: Performance of Various ML Algorithms for Location (in %)

Model	Accuracy
Logistic Regression	97
Random Forest	96
Decision Tree	92
K-Nearest Neighbor (KNN)	94
Support Vector Machine (SVM)	90

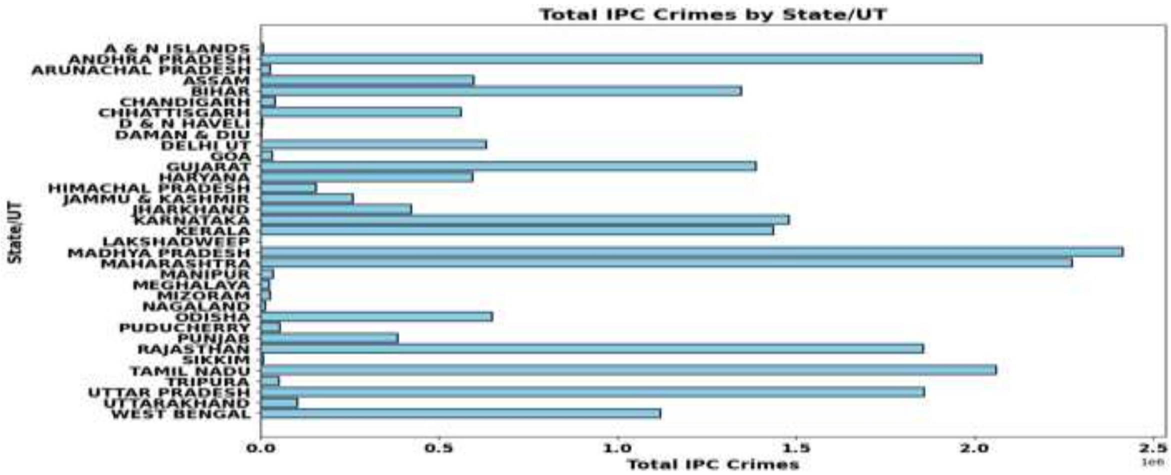

Figure 3: Crime Locations Visualization

III. TIME SERIES FORECASTING

This study performed time series forecasting on state wise crime data from 2001 to 2012 to predict annual crime trends. We used the Holt-Winters Exponential Smoothing model to analyze temporal patterns and the Prophet model for yearly crime predictions.

The Prophet model effectively captured annual seasonality and trends, revealing state-wise variations in crime rates—some states showed up to a 20% increase, while others saw nearly a 15% decline over the period. A state-wise analysis with the Holt-Winters model highlighted distinct regional trends, with some states showing steady growth in crime and others experiencing fluctuations, emphasizing the need for localized prevention strategies.

The Prophet model's accuracy was validated by comparing forecasted values with actual data, showing minimal deviation and a low RMSE, confirming its reliability.

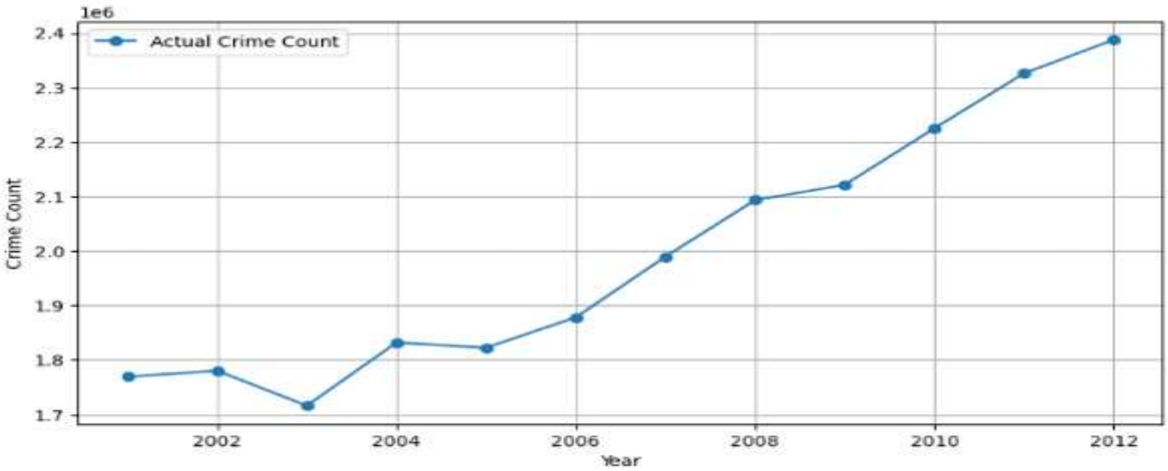

Figure 4: Crime count yearly forecasting

Overall, the forecasting results demonstrate the value of combining Holt-Winters and Prophet models for analyzing temporal and regional crime trends. These insights support data-driven decision-making for targeted law enforcement and resource allocation.

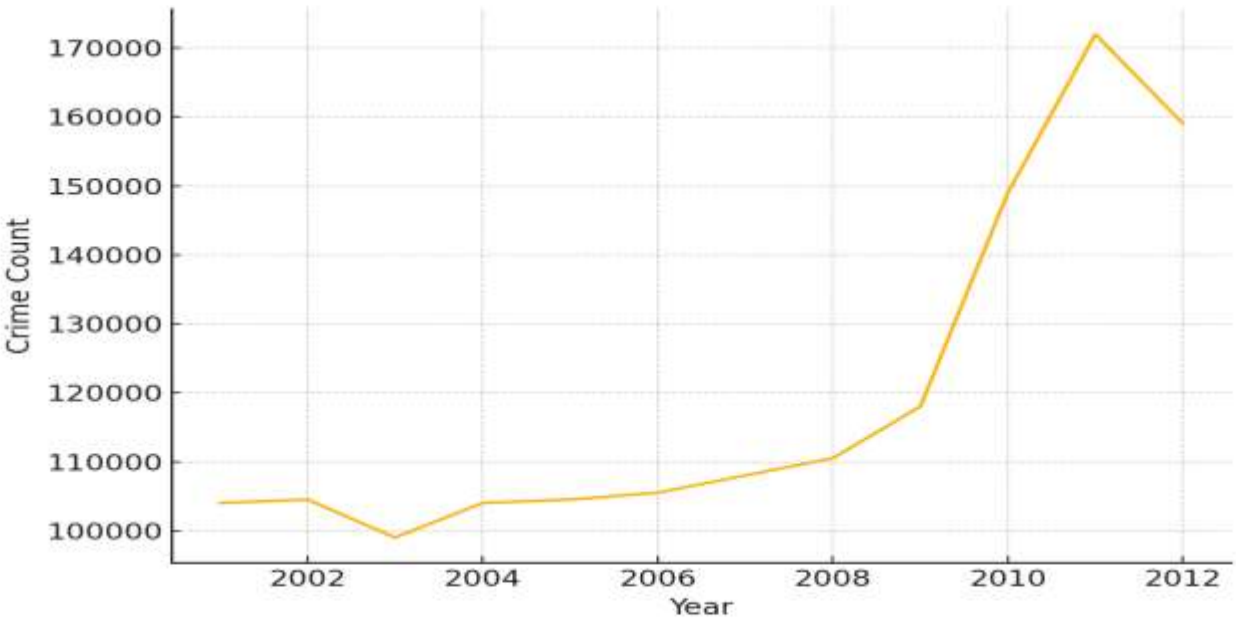

Figure 5: Crime count yearly forecasting of Kerala

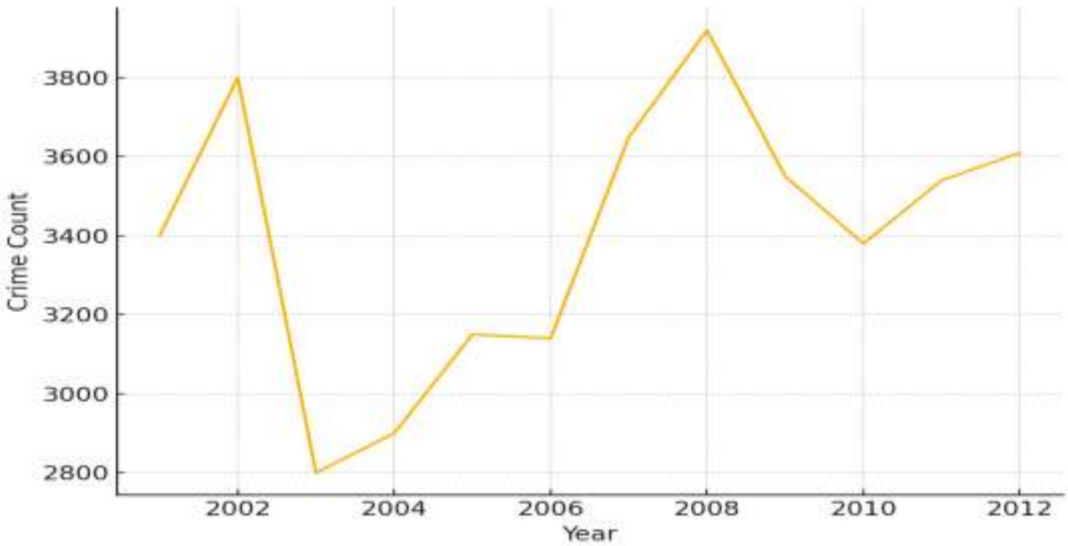

Figure 6: Crime count yearly forecasting of Chandigarh

DISCUSSION

This study aimed to develop a machine learning-based crime analysis and forecasting framework to enhance crime prediction and prevention. Using the Indian Crime dataset, we trained and evaluated models to predict crime types, locations, and future trends.

The Random Forest model achieved 96% accuracy in predicting crime types, effectively identifying complex data patterns. Logistic Regression reached 97% accuracy in predicting crime locations, revealing spatial trends and helping identify crime hotspots.

For time series forecasting, the Holt-Winters Exponential Smoothing model successfully captured seasonal and long-term trends, aiding in future crime prediction and strategic planning.

These results demonstrate the practical value of machine learning in real-world crime prevention. By enabling accurate predictions and hotspot detection, these models help law enforcement allocate resources efficiently and adopt proactive, data-driven strategies to improve public safety.

CONCLUSION

Crime remains a significant global challenge, demanding proactive strategies for public safety. Traditional analysis methods often fall short in capturing evolving crime patterns. This research presents a machine learning and time series-based framework to enhance crime prediction and prevention. Using over six million records from the Indian Crime dataset (2001–2012), each containing 23 attributes like crime type, location, time, and severity, we evaluated various machine learning algorithms. For crime type prediction, Random Forest achieved the highest accuracy at 98%, enabling more precise resource planning.

To detect high-risk areas, spatial analysis with K Nearest Neighbors (KNN) reached 97% accuracy in identifying likely crime locations, supporting targeted law enforcement efforts. Time series forecasting using the Holt-Winters model uncovered seasonal trends and anticipated crime preparedness. This integrated approach provides a robust, data-driven alternative to traditional methods, offering improved strategies for crime prevention and community safety.

REFERENCES

[1] R. Raj Bharath et al. (2024) - "A Machine Learning Approach to Crime Analysis and Forecasting for Prediction and Prevention

[2] Safat, W. et al. (2021) - "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques"

[3] Shah, N. et al. (2021) - "Crime Forecasting: A Machine Learning and Computer Vision Approach to Crime Prediction and Prevention".

[4] Jenga, K. et al. (2023) - "Machine Learning in Crime Prediction".

[5] Mandalapu, V. et al. (2023) - "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review".

[6] Kshatri, S. et al. (2021) - "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization".

[7] Sheriffdeen K. & Neveah B. (2023) - "Predictive Analysis and Prognostication: Leveraging Machine Learning for Anticipating Crime Trends".