iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XI Month of publication: November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75725

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Next-Gen Crime Analytics Using Big Data & Machine Learning

Ankita Mathur[1], Rainy Jain[2]
[1, 2]Professor, Department of BSCIT& CS Shankar Narayan College of Arts & Commerce

Abstract: With the rapid growth of digital data, crime analysis has entered a new era driven by Big Data and Machine Learning technologies. The increasing volume, variety, and velocity of crime-related information offer powerful opportunities to uncover hidden patterns, detect trends, and predict future incidents. Next-generation crime analytics aims to transform traditional policing methods by leveraging large-scale datasets, intelligent algorithms, and advanced visualization tools.

This research focuses on the application of Big Data processing and Machine Learning techniques to analyze major crime patterns, identify hotspots, and build predictive models for proactive crime prevention. Using methods such as classification, clustering, regression, and anomaly detection, crime records are processed to extract meaningful insights. Machine Learning algorithms enable accurate predictions of potential crime occurrences, while Big Data platforms support handling massive and unstructured datasets with high efficiency.

The study also presents an integrated visualization framework that helps administrators, police departments, and policymakers understand spatial and temporal crime trends. By combining predictive analytics, real-time data processing, and interactive dashboards, the system provides an intelligent solution for modern law enforcement.

Keywords: Big Data Analytics, Machine Learning, Crime Prediction, Crime Pattern Analysis, Predictive Modeling, Crime Hotspot Detection, Data Visualization.

I. INTRODUCTION

India is a vast nation with diverse socio-cultural backgrounds, and the position of women has historically held great significance. However, recent statistics from the National Crime Records Bureau (NCRB) indicate a concerning rise in crimes against women. This growing trend has made it essential for the administration to adopt advanced, technology-driven approaches to maintain law and order. Traditional criminology—focused on the scientific study of crime and criminal behavior—now requires the integration of Big Data and Machine Learning to handle the increasing complexity of crime patterns.

Crime datasets maintained by police departments across the country contain massive volumes of structured and unstructured information. These datasets often include incomplete, inconsistent, and varied data formats, which make manual analysis difficult and time-consuming. To transform this raw data into meaningful intelligence, modern crime analytics must incorporate scalable Big Data platforms and intelligent Machine Learning models.

Next-generation crime analytics aims to identify hidden relationships, detect crime patterns, forecast criminal behavior, and assist law enforcement agencies in proactive decision-making. An ideal analytical system should quickly analyze large crime datasets, recognize patterns efficiently, and support future crime prediction with high accuracy.

However, several challenges exist in the current scenario:

Key Challenges

A. Exponential Growth of Crime Data:

The size of crime records is increasing rapidly, requiring advanced Big Data techniques to store, process, and analyze information efficiently.

B. Inconsistent and Incomplete Data:

Crime reports often contain missing or unstructured data, making it challenging for conventional analytical methods to extract reliable insights.

C. Complexity in Crime Investigation:

The multifactor nature of crimes leads to longer investigation times, demanding automated systems that can reduce manual effort through intelligent analysis.

These challenges have motivated the development of a next-generation crime analysis framework powered by Big Data analytics and Machine Learning. This research focuses on building analytical models that can systematically address various complexities of crime and assist police departments with actionable intelligence.

Main Goals of the Project :

1) Detect and analyze crime patterns using large-scale datasets.
2) Provide actionable insights to help authorities design effective crime-prevention strategies.
3) Identify recurring crime behaviors and prevent similar incidents in the future through predictive analytics.

Specific Objectives

- Develop an efficient data cleaning pipeline capable of handling large, inconsistent datasets and preparing them for advanced analysis.
- Enhance and apply classification models using Machine Learning to predict future crime occurrences based on historical trends.
- Design anomaly detection algorithms to identify sudden or unusual changes in crime behavior, enabling early intervention.

II. LITERATURE SURVEY

Akoglu, Leman, and Christos Faloutsos ,Since crime is a growing concern in every part of the world it is very essential to find techniques to reduce it and also enable the police officials to easily catch the culprits. There are many approaches in solving crimes faster and a lot of researches are going on to find the best technique in data mining. The authors of this paper developed a new tool to track the culprits. Two algorithms Data Association and Back Propagation NN Classifier are used to analyse the data stored in the database. In order to extract criminal relations from an incidents summary and to create a group of suspects two approaches are used; With the help of BPN Classifier and Data Association algorithm the network is partitioned into subgroups and the interaction pattern is studied. The results prove that BPN Classifier is very accurate in identifying the crime patterns and also for future predictions. Bengio, Yoshua, Nicolas Boulanger- Lewandowski, and Razvan Pascanu [2], ACDCI (Crime Detection and Criminal Identification) technique was used to fasten the process of detecting the crimes in our Indian cities. In this technique the criminals were identified based on features like suspects name, sex, origin, facial features, crime reason, location, weapon used, etc. It had six main modules data extraction, pre-processing, clustering, map representation, and classification and WEKA tool. K-means algorithm was used for crime detection and it generated two clusters of crime. The KNN classification was used for criminal identification. The combination of k-means and KNN helped in improving the filtration for large databases. Bergstra, James, and Yoshua Bengio [3], The authors focused on the day-to-day factors rather than the causes for crime occurrences like the culprit’s background or political enmity. The proposed system can predict the regions with high crime occurrences and also visualize those regions. The system will help the investigating officials to resolve crimes faster. The steps followed in this approach are data collection, classification, pattern prediction and visualization. Bayes theorem is used for classification and by using this algorithm the news articles were trained and the model was built. Apriority algorithm helps in finding the frequent patterns of a particular region. The system developed predicts crime regions in India on a particular day. Boureau, Y-Lan, Jean Ponce, and Yann LeCun [4], The paper concentrated on analysing the approaches between Computer Science and Police department as one of the main application of data mining. Pattern detection technique has been implemented and suggestion for future prediction is also included. K-means algorithm is used for clustering and this will help in identifying the patterns of crime and hence, will help in solving crimes faster. In order to increase the accuracy of prediction semi-supervised technique is used. The crimes are represented using Geo spatial spots. Based on the selection of time range, type of crime and geographical region the results are shown graphically. Chang, Yi-Chun [5], The authors used algorithms like Naïve Bayesian, K-nearest Neighbour and Neural Networks (Multilayer-Perceptron) and proved that it is better than Decision tree and Support Vector Ma-chine. Two different feature selection methods are tested on the dataset. Comparison of algorithms are carried out on the basis of Area under Curve (AUC). The Chi-square feature selection technique is used for improving the performance of data mining re-salts. KNN gives better results by using Chi-square feature selection technique. The dataset chosen is categorised into two different types

III. EXISTING SYSTEM

In the field of modern crime analytics, detecting cybercrimes—especially SQL-based attacks and web intrusions—has become an essential requirement. As digital crime continues to rise alongside physical crime, law-enforcement agencies and cyber-security units face increasing challenges in identifying sophisticated attack patterns. To address these challenges, researchers have proposed an enhanced query-based layered analytical approach that aligns with Big Data–driven crime detection frameworks.

The proposed system operates on the foundation of knowledge-based crime pattern detection. A centralized knowledge repository is developed that stores frequently occurring cyberattack signatures, SQL injection patterns, and malicious query templates. This repository is continuously updated using probabilistic learning mechanisms that analyze historical cybercrime datasets. Such an approach not only captures known attack behaviors but also improves the system's ability to identify emerging digital crime patterns.

In crime analytics, the detection workflow begins with the validation of incoming digital evidence—such as suspicious queries, logs, and access-requests. The system applies advanced query-analysis methods that examine structural features, keywords, user-access behavior, and past incident logs. Using the enhanced layered approach, if any analyzed pattern matches previously identified attack signatures stored in the knowledge base, the event is immediately categorized as a potential cybercrime. Automated warnings are then triggered to assist cyber-forensics teams and law-enforcement agencies.

Additionally, keyword-driven filtering mechanisms significantly increase detection speed. Critical cybercrime-related indicators—such as malicious operators, high-risk commands, and abnormal access sequences—are pre-indexed using Big Data techniques. This enables fast clustering, classification, and anomaly detection across massive datasets commonly found in national and state-level cybercrime monitoring systems.

A. DRAWBACKS AND CHALLENGES IN EXISTING DIGITAL CRIME DETECTION SYSTEMS

Although query-based approaches provide improved cybercrime detection, several limitations still exist in traditional crime-analytics systems:

1) Inconsistent Crime Data Quality:

Digital crime logs often contain missing fields, unstructured formats, and noise. This significantly affects the accuracy of predictive models and hinders real-time monitoring.

2) Limited Analytical Flexibility:

Many existing cybercrime-detection systems rely on fixed rules or predefined fraud signatures. This restricts the system from identifying new, unknown, or evolving cyberattack patterns, reducing the overall intelligence capability.

3) Cross-Region Crime Data Variability:

Cybercrime data collected from different states or countries often follow different formats and standards. This inconsistency prevents accurate comparison, integrated crime mapping, and nationwide cyber-trend prediction.

4) Historical Data Gaps:

For many years, digital crime data may be incomplete or missing, especially for regions lacking proper reporting infrastructure. This leads to unreliable trend analysis and inaccurate long-term crime forecasts.

IV. PROPOSED SYSTEM

The proposed system introduces a predictive crime-analytics application designed to reduce crimes against women by leveraging machine learning—specifically, the Linear Regression model. The primary objective of this framework is to forecast crime rates across different cities and assist users, law-enforcement agencies, and administrative authorities in making informed decisions based on data-driven insights.

In this application, the system analyzes historical crime datasets and identifies temporal–spatial patterns to estimate the likelihood of future crime occurrences. Before traveling to a particular city, a user can check the predicted crime intensity and choose safer routes or destinations accordingly. This empowers women and general travelers with real-time safety insights, improving preventive action rather than reactive responses.

From an operational perspective, the system includes multiple stakeholders: users, managers, administrators, and police authorities. The manager updates city information, ticket availability, and travel schedules. Users book tickets based on predicted safety levels. Once a booking is confirmed, the manager verifies availability and finalizes the reservation. Police authorities are granted access to view confirmed travelers and their respective routes, enabling enhanced surveillance and proactive security arrangements.

The admin oversees the entire ecosystem—adding or removing managers, monitoring user activity, and ensuring accurate data integration for crime prediction. By combining crime forecasting with travel-management workflows, the system supports safer mobility and improves coordination between citizens and law-enforcement agencies.

A. ADVANTAGES

1) Predicting Crime Before It Occurs The system allows early detection of high-risk zones using machine learning, enabling individuals and authorities to take preventive safety measures.
2) Understanding Crime Patterns By analyzing large-scale datasets, the model reveals patterns in location, time, and crime type, helping administrators and police departments identify emerging threats and deploy resources effectively.

V. METHODOLOGY

The proposed application aims to reduce crime against women by applying advanced data mining and machine learning techniques to analyze crime trends across various cities. Designing such a predictive crime-analytics system introduces several technical challenges, particularly when dealing with large-scale, heterogeneous datasets. The system incorporates multiple transformation operations on incoming crime data, and the order in which these operations are performed significantly affects computational efficiency, accuracy, and overall system cost.

One of the major challenges lies in determining the optimal sequence of data transformations. Since the search space of possible transformation combinations is extremely large, exploring all combinations is impractical, especially in real-time environments. Therefore, the optimization process must remain lightweight, fast, and capable of producing near-optimal transformation paths without affecting system performance.

Another challenge emerges from the tradeoff between performance goals and computational cost. The system must dynamically adjust data-processing strategies based on varying crime volumes, user demands, and prediction requirements. Thus, the planner must intelligently balance accuracy, cost-efficiency, and runtime overhead while operating in an online environment.

To address these challenges, the proposed framework employs the K-Means algorithm as a core clustering technique and implements the complete system across four integrated modules, enabling efficient crime pattern detection, hotspot identification, and crime forecasting.

VI. ALGORITHMS USED

A. Linear Regression

Linear Regression is used to model the relationship between crime occurrences and influential factors such as location, time, and demographic variables. By fitting a linear equation to historical data, the system predicts crime values for future time periods. The algorithm estimates unknown parameters (weights) using methods such as least mean squares. Linear Regression is widely used due to its simplicity, interpretability, and suitability for predicting numerical crime trends. Variants include simple regression, multiple regression, and pace regression—each effective for high-dimensional datasets.

B. Decision Tree

Decision Trees serve both classification and prediction purposes in this system. Through a hierarchical structure of attribute-based splits, the algorithm identifies patterns and classifies crime types into relevant categories. Decision Trees are beneficial for crime prediction due to their interpretability and ability to model nonlinear relationships. They automatically learn crime decision rules from historical data and offer high transparency, which is essential for law-enforcement analysis.

C. K-Means Algorithm

K-Means is one of the most widely used clustering algorithms and forms a core part of this crime-analytics framework. It partitions the crime dataset into clusters by identifying groups with similar characteristics—such as crime hotspots, time-based patterns, or location-based crime densities. The algorithm's efficiency, low computational cost, and scalability make it suitable for large datasets typically involved in national crime databases.

1) *Advantages of K-Means:*

- Simple and easy to implement
- Highly scalable for large datasets
- Guarantees convergence
- Efficient and adaptable to new incoming data

2) Disadvantages of K-Means:

- Requires manual selection of the number of clusters
- Sensitive to initial centroid values
- Performs poorly when clusters vary in shape, size, or density

VII. IMPLEMENTATION

The proposed crime-analytics framework utilizes an advanced multi-stage methodology designed to identify crime patterns, classify incidents, and predict future crime occurrences using Big Data and Machine Learning techniques. The overall process consists of five major phases: data collection, classification, pattern identification, crime prediction, and visualization. Each phase contributes significantly to building an intelligent crime-analysis and decision-support system.

1) Data Collection

Data collection forms the foundation of the crime-analysis process. Crime-related information is gathered from multiple online sources such as verified crime-report websites, digital newspapers, public datasets, and blogs. Since most of the collected data is unstructured or semi-structured, it is stored in a flexible database environment that supports object-oriented design. This allows easy manipulation, preprocessing, and transformation of diverse data types for subsequent analysis. Big Data platforms (e.g., Hadoop, Spark) can also be integrated to manage large-scale datasets efficiently.

2) Classification

In this phase, the system applies the Naïve Bayes classifier, a supervised machine learning algorithm based on probabilistic modeling. Naïve Bayes generates a probability distribution across multiple crime categories rather than producing a single output, making it highly effective for multi-class crime classification.

Naïve Bayes is preferred due to:

- its simplicity
- fast execution
- minimal memory requirements
- excellent performance on small or moderate-sized training datasets

Crime types such as vandalism, murder, robbery, burglary, sexual assault, and gang rape are used as training categories. The classifier estimates probabilities based on historical occurrences. However, a known challenge with Naïve Bayes is encountering zero probability for unseen events (e.g., when $P(C|D) = 0$), which requires smoothing techniques like Laplace smoothing.

3) Pattern Identification

Pattern identification involves detecting frequently occurring crime combinations, trends, and correlations. For this purpose, the Apriori algorithm is used to extract association rules from the crime dataset. Apriori identifies crime patterns by determining which crime combinations frequently occur together in specific regions or time periods.

These discovered rules assist police officials in making informed decisions such as deploying additional CCTV surveillance, increasing security patrols, installing alarm systems, or monitoring high-risk zones. Pattern identification strengthens predictive policing by providing actionable intelligence.

4) Crime Prediction

The fourth phase focuses on predicting the type and intensity of crime likely to occur in a specific location within a given time frame. The prediction model uses key crime attributes, including:

- Month of occurrence
- Day of the week
- Time of the day
- Geographical location

Crime prediction is achieved using a combination of classification algorithms that identify potential hotspots and forecast crime tendencies. Commonly used techniques include:

- K-Nearest Neighbor (K-NN)
- Decision Trees (J48)

- Support Vector Machine (SVM)
- Neural Networks
- Naïve Bayes and Ensemble Learning techniques

In addition to these, Linear Regression is used to establish a mathematical relationship between crime frequencies and influencing variables. The regression line is computed as:

Y = aX + b,

where Y is the predicted crime rate, X is the independent variable, b is the slope, and a is the intercept. The slope is calculated as b = r(sx/sy) and the intercept as a = My – bMx.

This hybrid use of machine learning algorithms results in more accurate crime forecasting and helps in designating hotspots and cold spots across different regions.

5) *Visualization*

The final step is visualization, where crime patterns, predictions, and hotspots are graphically represented using heatmaps and geospatial charts. A heatmap visually indicates crime activity intensity:

- Darker colors represent low crime activity
- Brighter colors represent high crime concentration

Visualization tools significantly improve decision-making by presenting complex crime data in an intuitive and interactive manner. Law-enforcement agencies can instantly monitor hotspots, analyze time-based crime trends, and allocate resources more effectively.

VIII. CONCLUSION

As a future extension of this research, multiple advanced directions can be explored to further improve the accuracy, reliability, and scope of crime prediction systems. One of the primary objectives is to integrate additional machine learning classification models—such as Random Forest, Gradient Boosting, XGBoost, Logistic Regression, and Deep Learning architectures—to enhance predictive accuracy. By comparing multiple algorithms within an ensemble-learning framework, the system can automatically select the best-performing model for different crime categories and geographical regions, thereby improving overall performance.

Another important direction for future work is the incorporation of socio-economic variables, particularly neighborhood income levels, employment rates, educational status, and population density. Integrating income-related data may reveal strong correlations between economic conditions and crime occurrences. Identifying such relationships can assist policymakers, law-enforcement agencies, and social organizations in understanding the deeper socio-economic root causes of crime. This would help in designing targeted interventions for high-risk communities.

Additionally, the extension of this research to include datasets from multiple new cities and states, along with their demographic profiles, will make the system more robust and more widely applicable. Adding diverse datasets such as age distribution, gender ratio, literacy rate, migration patterns, and urban infrastructure details can help generate a more comprehensive and generalizable prediction model. Studying multiple cities also enables comparative crime analysis, which can reveal unique regional trends, hotspot shifts, and behavioral characteristics of criminal activity.

Overall, these extensions will significantly improve the analytical depth, accuracy, and scalability of the proposed crime-prediction framework, making it more suitable for real-world deployment in modern smart-city surveillance and policing systems.

REFERENCES

[1] Akshay, R., & Kumar, P. (2021). Crime prediction using machine learning and data mining techniques. International Journal of Computer Applications, 175(23), 1–6.

[2] Bhardwaj, R., & Gupta, S. (2020). Analyzing crime patterns using K-means clustering. International Journal of Advanced Research in Computer Science, 11(2), 45–52.

[3] F. A. Thabtah. (2007). A review of naïve Bayes classifiers for educational data. Journal of Machine Learning Research, 10(1), 1–15.

[4] Lin, Y., & Brown, D. (2016). Crime prediction using regression and spatial analysis. IEEE Transactions on Information Forensics and Security, 11(3), 543–557.

[5] Liu, H., &Motoda, H. (2007). Data mining: Concepts and techniques. Morgan Kaufmann Publishers.

[6] Mohler, G., Short, M., & Bertozzi, A. (2011). Self-exciting point process modeling of crime. Journal of the American Statistical Association, 106(493), 100–108.

[7] Singh, R., & Kaur, G. (2019). Comparative study of classification algorithms for crime prediction. International Journal of Engineering and Technology, 8(4), 280–287.

[8] Tan, P.-N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.

[9] Xu, J., & Chen, H. (2005). Criminal network analysis and visualization. Communications of the ACM, 48(6), 100–107.

[10] Zhang, Z., & Zhao, L. (2017). Crime forecasting using machine learning approaches. Procedia Computer Science, 122, 451–457.