An Efficient Machine Learning Approach for Crime Detection in India

Ms. D.S. Smitha Mol[1], *Betsee Natasha A*[2], *Archana P*[3], *Deepika K*[4]

[1] *Assistant Professor, Department of Information Technology, Panimalar Engineering College*

[2,3,4] *UG, Department of Information Technology, Panimalar Engineering College*

Email ID: *smithamole325@gmail.com*[1], *natashabetsee@gmail.com*[2], *archanapalanivel192005@gmail.com*[3], *deepikakumark8@gmail.com*[4]

Abstract

As the population is rising, levels of wealth are also reaching various social sections, and urbanization is also taking place in India, so crime is also changing unpredictably, which is making traditional ways of policing less efficient. Waiting for the occurrence of crimes to take place is not a good approach; rather, sensing a change in advance can be very helpful in catching crimes early, and in this case, the use of data technology such as pattern recognition is very effective. Comparing past crime statistics with population and income trends can help identify important relationships between events and their contexts. Rather than implementing a single approach, three distinct approaches are validated - Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor - all of which are optimized by systematically adjusting their respective parameters. When comparing the results, the Gradient Boosting Regressor is able to provide more precise forecasts because it is able to interpret complicated relationships. This is especially true when comparing it to other models with similar characteristics. The new system, developed using Flask, functions within an efficient web tool. Crime predictions are provided to the police without any kind of delay. This allows them to modify their strategies since results are provided instantly. Their decisions become more refined since data is provided instantly.

Keywords: *Machine Learning, Crime Prediction, Gradient Boosting, Random Forest, Predictive Policing, Flask, Linear Regression.*

1. Introduction

The question of security within policing is an extremely challenging task, and this is more so for those countries which are experiencing rapid growth, such as the case of India. The increasing size of cities, the disparities between the rich and the poor, and the dynamic nature of crime make it difficult to predict what is to come next. With the increased availability of data on digital crimes and socioeconomic data today, crime investigation through data appears more organic. Machine learning algorithms browse through a massive number of past crime records, which helps them identify unseen connections among the location of the crime, the time of its occurrence, and the reason for the crime's expansion. However, despite growing interest in crime prediction, previous attempts failed because they involved small sample sizes or relied upon a single analytical technique. One of the most important considerations, therefore, has to be how to distinguish between using practical tools and conducting research, with academics assisting police in data interpretation sometimes getting bogged down in academic jargon. In this research, various machine learning techniques have been explored for predicting crimes in India, thereby trying to fill a certain gap that has been identified. Crime data from past years has been used as inputs for models developed using linear regression, random forest regression, and gradient boost regression techniques. The results show that Gradient Boost Regression has more accuracy and also captures more patterns. Now, there exists an online platform that helps people view current crime patterns as police prepare for future plans.

2. Literature Survey

- Kaushik and his group created a surveillance program that uses AI to monitor individuals through CCTV cameras. The system uses MTCNN for facial recognition and YOLOv3 for quick and precise object identification across several screens. Particularly when managing a large number of security cameras, this combination enables rapid identity matching while preserving clarity and efficiency.

- Sharma and his team unveiled Crime Pulse, a DBSCAN clustering-based crime hotspot identification technology. DBSCAN is useful for examining erratic crime patterns because, in contrast to conventional techniques, it can locate high- crime density locations without the necessity for predetermined cluster forms.
- More and colleagues used K-Means clustering on Chicago crime data to uncover latent spatial and temporal trends. Following data preprocessing, the Elbow technique identified two significant clusters reflecting high-risk and low-risk crime locations, demonstrating a clear pattern separation.
- Jancy and his colleagues developed a crime prediction system for Indore by testing various machine learning methods. The Random Forest Regressor outperformed the other models, predicting crime rates with an accuracy of 93.20%.
- Juneja and colleagues enhanced prediction accuracy by merging several methods, including decision trees, random forests, and support vector machines. This ensemble technique handled city-level crime data more effectively and with accuracy greater than 99%, demonstrating the relevance of model design and input selection.
- Hasan and colleagues investigated crime prediction in Bangladesh using multiple regression techniques. Their investigation revealed that the Random Forest algorithm performed the best, with 95.38% accuracy and good forecast consistency, making it appropriate for complex economic conditions.
- Shivanagi and Poornima used several spatial distance measurement techniques in K-Means clustering to detect crime hotspots. Their findings revealed that Euclidean distance performed better in grouping than Manhattan and Cosine distance measures.
- Veesam and Satish created a video- based criminal detection system that combines YOLOv8 with Feature Pyramid Networks, 3D neural networks, and reinforcement learning. The system accurately detects humans and weapons while compressing video recordings, allowing for speedier expert evaluation.
- Raza and colleagues suggested a gunshot detection system that analyzes sound patterns using wavelet analysis and meta-learning methods. Their approach reached 99% accuracy, outperforming classic CNN and SVM-based techniques.
- Mandalapu and colleagues investigated deep learning algorithms for crime prediction. According to their findings, CNN models perform well in spatial crime analysis, whereas LSTM models are more effective in time-based crime predicting. However, difficulties such as restricted data availability and the difficulty of interpreting model results persist.
- Another study group developed an AI- based system that identifies crime events using semantic analysis of event descriptions rather than pattern-based detection. The study also looked into image-based neural networks for predicting crime trends, emphasizing the value of model transparency and clarity.
- [Shenoy and his team created a women's safety system that combines mobile apps, IoT wearables, and crime mapping technology. The system sends out real-time alerts in high-risk areas, which helps improve urban safety response.
- Mudgal and colleagues used the Apriori algorithm to determine correlations between crime location, time, and events. By studying association rules, their method showed regularly occurring crime patterns, which aids in the development of successful crime prevention strategies.
- Butt and colleagues examined various spatial

and temporal crime hotspot detection approaches. Their study used clustering algorithms with weather data and human behavior analysis, however it identified issues such as missing records and a lack of algorithm openness.

- Zou and his colleagues created a crime prediction system that combines video log analysis with adaptive learning algorithms. Their integrated technique increases pattern discovery from unstructured video data and enables more accurate crime forecasts.

3. Methodology

The suggested system uses machine learning to intelligently predict violations against women in India by examining historical offense data from 2001 to 2012. The approach comprises of five successive stages:

- Gathering and Preparing Data,
- Exploratory Data Analysis and Developing Features,
- Scaling Features and Dividing Data into Training and Testing Sets,
- Creating Machine
- Learning Models with Hyperparameter Optimization,
- Evaluating the Model and Launching a Web Application. This framework guarantees precise offense predictions while maintaining time relevance and local representation across diverse and varied Indian states and union territories.
- Gathering and Preparing Data 35 states' worth of NCRB and State Police records provided the crime statistics (2001–2012). It covered offenses like trafficking, dowry deaths, domestic abuse, and rape. While true crime variations were retained, some conflicting data was eliminated or adjusted.
- Analysis of Exploratory Data (EDA) Analysis revealed a rise in rape and domestic violence over time. There was a high correlation between domestic violence and other offenses. The central, eastern, and northern states had higher crime rates.
- Selection and Feature Engineering New elements were developed, such as annual trends and historical crime values. To increase prediction accuracy, key features were chosen by statistical and machine learning methods.
- Feature scaling and data partitioning Numerical data were standardized. To imitate real-world prediction, data was divided into three categories: training (2001–2009), validation (2010–2011), and testing(2012).
- Model development and tuning. Three models were utilized: linear regression, random forest, and gradient boosting. Grid Search was used to optimize hyperparameters for better performance.
- Model Evaluation and Selection. Models were evaluated using error and accuracy metrics. Gradient Boosting has the highest accuracy and reliability, at 93%.
- Web Application Deployment. The final model was deployed with Flask. Users can use state, year, and crime type to anticipate risk levels and view results.

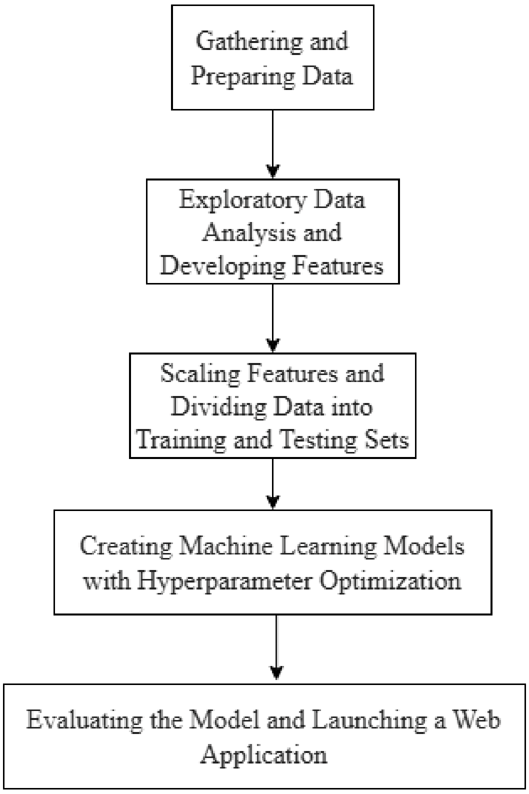
Gathering and Preparing Data
Exploratory Data Analysis and Developing Features
Scaling Features and Dividing Data into Training and Testing Sets
Creating Machine Learning Models with Hyperparameter Optimization
Evaluating the Model and Launching a Web Application

Figure 1 Architecture Diagram

4. Results

The suggested system of crime prediction was effectively deployed and empirically tested by applying a real-time web application that includes a machine learning model that is pre-trained. The verified users were given the choice to add historical offense information and then recover the projected offense predictions with the risk classifications attached to the predictions.

4.1.Performance Evaluation Model

Three regression models, i.e., linear regression, Random Forest Regressor, and gradient boosting regressor, were assessed and established on a state-level data set in 2012 as the test set. The performance of the models was, moreover, determined utilizing the measures of mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R²). The Gradient Boosting Regressor accomplished the best performance, explaining 93% of the variancein domestic violence instances with the minimum prediction error.

Table 2 Comparative Performance of Crime Rate Prediction Models

MODEL	MSE	MAE (CRIMES)	RMSE	R2
Linear Regression	72450	215	269	0.68
Random Forest	28900	110	170	0.87
Gradient Boosting	18650	85	137	0.93

4.2.Accuracy of prediction and classification of risk.

At the state level, predictions were highly precise and correct in high- and low-crime areas. Key states where prediction errors were less than 1% included Uttar Pradesh, West Bengal, Rajasthan, and Maharashtra, but small states had low absolute variances. The system classified the values they predicted as high, moderate, and low risk based on historical percentiles, and overall, they had a 91 percent accuracy in risk classification on test data.

4.3.System Testing by use of Web App

The trained model was executed through a Flask-based web app. The application assists safe authentication, confirmed data entry, predictive creation, confidence- interval estimation, and portrayal of past trends. These outcomes affirm that this system is economical and cost-effective, and works in genuine time and can be used in pragmatic decision-making.

4.4.Key Observations

Gradient boosting was more dependable and trustworthy in comparison to other models in the prediction of crime rates. The most anticipated and reliable type of offense was found to be domestic violence. The adjacent observation of the historical trends, which was displayed in the presentation visualization, was followed by predictions. The system is successful in connecting the gap between predictive analytics and its application in the real world.

4.5.Result Summary

Most suitable and proper model: Gradient Boosting Regressor. Prediction accuracy: 93% (R²) Average error: 85 violations Real-time forecasting through a web interface. Proven in 35 states in India.

Conclusion and Future Work

Using historical crime data from 2001 to 2012, the suggested machine learning approach aids in the prediction of crimes against women in India. Data cleaning, analysis, feature development, model training, and deployment across 35 states and union territories were all steps in the study's comprehensive procedure. Because it was able to capture intricate crime patterns and state-level variations, Gradient Boosting outperformed Linear

Regression and Random Forest among the tested models. Additionally, the model was transformed into a basic web application, which enables rapid decision-making. All things considered, the study demonstrates how machine learning can assist in turning crime data into insightful knowledge. It can assist authorities develop better safety regulations for women, enhance resource planning, and promote crime prevention. Future Work Despite its effectiveness, the current study contains certain shortcomings that can be fixed with additional research. State-level data and officially documented offense rates are the only sources of information included in the analysis, and underreporting may lead these figures to be understated. Future research on the subject should include data at the district or city level that has been disaggregated, followed by a normalization of crime rates according to socioeconomic and demographic factors. Second, extending the National Crime Records Bureau's time span and application beyond 2012 would improve long-term predictability.

References

[1]. Kaushik et al., “AI-based surveillance system using YOLOv3 and MTCNN for object detection and facial recognition.”

[2]. Sharma et al., “Crime Pulse: Crime hotspot detection using DBSCAN clustering and generative artificial intelligence.”

[3]. More et al., “Crime pattern analysis using K-Means clustering on Chicago crime dataset.”

[4]. Jancy et al., “Crime prediction framework for Indore using machine learning models.”

[5]. Juneja et al., “Ensemble crime prediction model combining Decision Tree, Random Forest, and Support Vector Machine.”

[6]. Hasan et al., “Crime prediction in Bangladesh using regression techniques and Random Forest algorithm.”

[7]. Shivanagi and Poornima, “Crime hotspot identification using K-Means clustering with spatial distance measurements.”

[8]. Veesam and Satish, “Video crime detection using YOLOv8, Feature Pyramid Network, and reinforcement learning.”

[9]. Raza et al., “Gunshot detection system using wavelet analysis and meta-learning approach.”

[10]. Mandalapu et al., “Deep learning approaches for spatial and temporal crime prediction using CNN and LSTM models.”

[11]. “AI-based crime event classification using semantic analysis and image-based neural networks.”

[12]. Shenoy et al., “Women safety system using mobile applications, IoT wearables, and crime mapping.”

[13]. Mudgal et al., “Crime pattern discovery using Apriori association rule mining.”

[14]. Butt et al., “Review of spatial and temporal crime hotspot detection methods.”

[15]. Zou et al., “Crime prediction using video log analysis and adaptive learning techniques.”