

Predictive Analysis and Visualization of Crimes against Children in India Using XGBoost

Shivani Sagore¹, Devendra Singh Kaushal²

¹Department of Computer Science & Engineering, Jawaharlal Institute of Technology, Borawan, Khargone, Madhya Pradesh, India

²Department of Computer Science & Engineering, Jawaharlal Institute of Technology, Borawan, Khargone, Madhya Pradesh, India

Abstract - Crimes against children are a significant social and law-enforcement concern in India. The availability of historical crime records provides an opportunity to apply machine learning and data visualization techniques for identifying temporal and geographical patterns and supporting preventive decision-making. This paper presents a machine-learning-based framework for predictive analysis and visualization of crimes against children in India using National Crime Records Bureau (NCRB) data covering the period 2001–2023. The proposed framework consists of data preprocessing, exploratory data analysis, geographical visualization, feature preparation, model training, and predictive evaluation. Bar charts, line graphs, geographical maps, and heat maps are used to examine state-wise and temporal crime patterns. XGBoost is employed as the primary prediction algorithm because of its ability to model nonlinear relationships and complex patterns. Its performance is compared with Linear Regression and Random Forest using Accuracy, R^2 , Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The experimental results reported in the study show that XGBoost achieved an accuracy of 97.00%, an R^2 score of 0.970, MSE of 3.25×10^7 , MAE of 4,125, and RMSE of 5,701. Random Forest achieved 94.85% accuracy and an R^2 score of 0.949, whereas Linear Regression achieved 89.70% accuracy and an R^2 score of 0.897. The visualization analysis also indicates substantial regional variation, with Uttar Pradesh, Madhya Pradesh, Maharashtra, Rajasthan, and West Bengal appearing among the states with comparatively high reported crime incidence. The proposed framework can support evidence-based analysis, resource

allocation, hotspot identification, and child-protection planning.

Keywords: Crimes Against Children, NCRB, Machine Learning, XGBoost, Crime Prediction, Data Visualization, Random Forest, Linear Regression, Crime Hotspot Analysis, Geospatial Analysis.

1. INTRODUCTION

Crimes against children include kidnapping and abduction, trafficking, sexual offences, child labour, physical abuse, exploitation, and other forms of violence. Such offences can have serious physical, psychological, educational, and social consequences for children. The availability of digital crime records has created opportunities to examine historical patterns using computational techniques.

Traditional crime analysis generally focuses on descriptive statistics and historical reporting. Although these approaches are useful, they may not adequately reveal complex relationships and nonlinear trends in large datasets. Machine learning provides an alternative approach in which historical observations can be used to identify patterns and estimate future trends.

The present study focuses on crimes against children in India using NCRB data from 2001 to 2023. The research combines exploratory data analysis and visualization with machine-learning-based prediction. The dissertation reports that the number of recorded crimes increased from 9,838 in 2001 to 167,904 in 2023, although reported crime data should not automatically be interpreted as the actual prevalence of crime because under-reporting and differences in reporting practices affect the dataset.

The study therefore addresses two complementary objectives: first, understanding temporal and spatial patterns through visualization; and second, developing a predictive model capable of learning relationships in historical crime data.

2. RESEARCH OBJECTIVES

The major objectives of the study are:

- To analyze historical crimes against children in India using NCRB data from 2001–2023.
- To identify temporal trends in reported crimes.
- To visualize state-wise crime distributions using charts, geographical maps, and heat maps.
- To develop an XGBoost-based prediction model.
- To compare XGBoost with Linear Regression and Random Forest.
- To evaluate the models using multiple performance metrics.
- To identify regions with comparatively high reported crime incidence.
- To provide a data-driven framework that can support child-protection planning and resource allocation.

3. RELATED WORK

Crime prediction research has progressed from conventional statistical analysis and hotspot mapping toward supervised machine learning and deep-learning approaches. Previous studies have explored clustering, classification, regression, neural networks, spatial analysis, and ensemble learning for crime prediction.

The source study reviews work involving spatiotemporal forecasting, neural models, clustering, Random Forest, socio-economic variables, and geographical crime analysis. It also notes that important challenges remain in data quality, model interpretability, generalization, and the possibility of historical policing bias influencing predictive systems.

Indian crime-analysis research has also used clustering and machine-learning methods. One reviewed study analyzing Indian states and Union Territories for 2001–2012 applied k-means clustering and Google Map marker clustering and reported higher performance for

Random Forest than a neural-network model. Other studies investigated relationships between crime and socio-economic variables such as GDP, HDI, inflation, and unemployment.

However, the present work specifically concentrates on crimes against children and combines long-term NCRB data, visualization, geographical analysis, and XGBoost-based prediction.

4. DATASET AND PROBLEM DOMAIN

The study uses the NCRB dataset covering crimes against children in India for 2001–2023. The dataset contains state- and Union-Territory-level annual crime records and includes categories such as kidnapping and abduction, trafficking, POCSO-related sexual offences, child labour-related crimes, infanticide, abandonment, and other forms of abuse and exploitation.

4.1 Data Challenges

Crime prediction involving children has several important challenges:

under-reporting of incidents;

inconsistent reporting across states and agencies;

missing and inaccurate records;

fragmented data sources;

differences in data formats;

geographical and temporal variations;

privacy and ethical concerns.

The source dissertation specifically identifies under-reporting, fragmented data sources, missing values, data inaccuracies, and standardization as important challenges for predictive analysis.

Therefore, predictions generated from NCRB records should be interpreted as predictions of reported/recorded crime patterns, rather than direct estimates of the true incidence of crimes.

5. PROPOSED METHODOLOGY

The proposed framework consists of the following stages:

NCRB Data Collection → Data Preprocessing → Feature Engineering → Exploratory Data Analysis → Visualization → Train/Test Data → Model Training → Hyperparameter Tuning → Prediction → Performance Evaluation

5.1 Data Preprocessing

The preprocessing stage involves:

- identifying missing values;
- removing or correcting inconsistent observations;
- encoding categorical variables such as State/UT and crime type;
- organizing annual state-wise observations;
- preparing features for machine-learning models;
- dividing the data into training and testing subsets.

The source work describes missing-value handling, categorical encoding, and training/testing division as part of the model-development process.

5.2 Exploratory Data Analysis

EDA is performed to identify:

- annual crime trends;
- state-wise differences;
- growth patterns;
- high-incidence regions;
- temporal changes;
- geographical variation.

The study uses line plots, bar charts, geographical maps, and heat maps. The reported visualization includes state-wise graphs for 2021–2023, top-ten-state comparisons for 2014–2023, total growth from 2001–2023, geographical maps for 2020–2023, and a heat map for 2014–2023.

5.3 XGBoost Model

XGBoost is an ensemble learning algorithm based on gradient boosting. It sequentially builds decision trees, with subsequent trees attempting to reduce the errors generated by previous trees.

The prediction function can be represented as:

$$\hat{y}_i = \sum f_k(x_i)$$

where $\hat{y}_i$ represents the predicted value, x_i is the input feature vector, and f_k represents an individual decision tree.

The objective function combines prediction loss and model regularization:

$$Obj = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_k)$$

The source study describes this sequential tree-building and regularization mechanism as the basis of its XGBoost prediction approach.

5.4 Feature Engineering

The proposed framework uses historical crime-related information including:

- year;
- State/UT;
- historical crime counts;
- yearly trends;
- growth patterns;
- crime-category information.

The source methodology states that state-wise crime counts, yearly trends, growth rates, and historical patterns are extracted as relevant features for the XGBoost model.

6. VISUALIZATION AND TREND ANALYSIS

The longitudinal analysis demonstrates a strong upward trend in reported crimes against children.

The study reports that recorded cases increased from 9,838 in 2001 to 167,904 in 2023. A fitted linear trend is reported as:

$$y = 7824x - 2E+07$$

with:

$$R^2 = 0.897$$

The reported slope indicates an average increase of approximately 7,824 recorded cases per year over the study period.

The study also reports a temporary reduction around 2020, followed by an increase in the post-pandemic period. Such changes may reflect both changes in crime occurrence and changes in reporting, institutional processes, and social conditions.

6.1 Geographical Analysis

The geographical visualization shows considerable variation among states. The study identifies Uttar Pradesh, Madhya Pradesh, Maharashtra, Rajasthan, and West Bengal among states with comparatively high reported incidence during the study period.

The heat-map analysis similarly shows darker/high-incidence regions in parts of central and northern India, particularly Madhya Pradesh, Uttar Pradesh, Rajasthan,

and Maharashtra. However, population size, urbanization, reporting intensity, and socio-economic conditions can influence reported counts, so raw case counts should not be interpreted as risk rates without appropriate population normalization.

7. EXPERIMENTAL RESULTS

Three machine-learning models were compared:

Model	Accuracy (%)	R ²	MSE	MAE	RMSE
Linear Regression	89.70	0.897	1.82 × 10 ⁸	10,245	13,491
Random Forest	94.85	0.949	8.76 × 10 ⁷	6,854	9,360
XGBoost	97.00	0.970	3.25 × 10 ⁷	4,125	5,701

7.1 Performance Discussion

XGBoost produced the highest reported accuracy and R² value and the lowest MSE, MAE, and RMSE among the three evaluated models. The reported results therefore indicate that XGBoost provided the strongest predictive performance for the experimental dataset.

Random Forest performed better than Linear Regression but remained below XGBoost. Linear Regression produced the largest prediction errors, suggesting that a simple linear relationship was less suitable for representing the patterns captured by the experimental data.

8. DISCUSSION

The results indicate that ensemble tree-based models can effectively capture nonlinear relationships in historical crime records. XGBoost's reported performance suggests that it is a promising method for crime-trend prediction when historical state-wise and temporal information is available.

However, high predictive performance should not be interpreted as proof that the model can predict individual criminal events or identify individuals who will commit crimes. The model is intended for aggregate historical trend analysis and forecasting.

An additional limitation is that NCRB data represent reported cases. Under-reporting, differences in reporting practices, administrative changes, and changes in

legislation can influence observed trends. The source dissertation explicitly identifies these issues as challenges for crime prediction.

Ethical considerations are especially important because the research concerns children. Personally identifiable information should not be used in public analytical outputs, and predictions should not be used to discriminate against communities or individuals.

9. CONCLUSION

This paper presented a machine-learning-based framework for predictive analysis and visualization of crimes against children in India using NCRB data from 2001–2023. The framework integrates preprocessing, exploratory data analysis, geographical visualization, feature engineering, and machine-learning prediction.

The reported experimental results show that XGBoost achieved 97.00% accuracy and an R² score of 0.970, outperforming Random Forest and Linear Regression. It also produced the lowest reported MSE, MAE, and RMSE values.

The visualization analysis demonstrates substantial temporal and geographical variation in reported crimes against children. The results suggest that predictive analytics can support evidence-based planning, hotspot monitoring, and allocation of child-protection resources.

Nevertheless, the model should be considered a decision-support tool, not a replacement for law-enforcement judgment or social investigation. The interpretation of reported crime data must account for under-reporting, population differences, reporting practices, and other contextual factors.

10. FUTURE WORK

Future research can improve the proposed framework in several directions:

- Incorporate population-normalized crime rates instead of only absolute counts.
- Include socio-economic variables such as literacy, unemployment, poverty, urbanization, and population density.
- Develop district-level and police-station-level prediction models.

- Investigate time-series and spatiotemporal deep-learning approaches.
- Compare XGBoost with LightGBM, CatBoost, Random Forest, LSTM, GRU, and Transformer-based models.
- Apply explainable AI techniques such as SHAP to interpret model predictions.
- Develop real-time data integration with authorized government and law-enforcement systems.
- Evaluate model fairness and robustness across different geographical regions.
- Develop interactive dashboards for policymakers and child-protection organizations.
- Validate predictions using temporally separated test datasets to assess generalization.

The source dissertation similarly proposes incorporating demographic and socio-economic variables, continuously updated data, and more localized district/city/police-station-level forecasting as future extensions.

References

- [1] National Crime Records Bureau (NCRB), Crime in India, Government of India.
- [2] Chen, T., and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
- [3] Breiman, L., "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001.
- [4] Yu et al., research on crime-data aggregation and data-mining-based crime forecasting, as reviewed in the source dissertation.
- [5] Deepika K. K. and Smitha Vinod, research on crime analysis of Indian States and Union Territories using clustering and machine-learning approaches.
- [6] Manna and Biswas, research examining relationships between socio-economic variables and crime in India.
- [7] Catlett et al., research on spatiotemporal crime forecasting.
- [8] Wang et al., research applying ST-ResNet-based approaches to city-level crime prediction.
- [9] Stec and Klabjan, research incorporating external data sources for crime-count prediction.
- [10] Mohler et al., research addressing predictive policing and fairness considerations.
- [11] Mudgal et al., review of computational methods for crime identification, analysis, and prediction.
- [12] Government of India, Protection of Children from Sexual Offences (POCSO) Act, 2012.
- [13] National Crime Records Bureau, Crime in India reports and statistical datasets used for historical crime analysis.