Crime Prediction Using Ensemble Approach

ARJUN K[1], SUCHETHA N V [2], PANCHAMI B S[3]

[1]*Department of AIML, Canara Engineering College, Bantwal, 574219 and Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India*

[2]*Department of CSE, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire-574240 and Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India*

[3]*Student, Department of CSE, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire and affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India.*

Abstract - In today's context, preventing crime is increasingly vital to safeguard communities and ensure public safety. Similar to how vaccinations shield children from diseases, a proactive approach to crime prevention aims to mitigate risks before they escalate. This involves not only educating the public and enhancing awareness but also implementing efficient policing strategies and employing technology-driven predictive models. By analysing historical crime data alongside geographic and demographic factors, this study employs advanced analytical techniques. These methods aim to uncover correlations between socio-economic conditions, environmental factors, and criminal activities. By integrating machine learning algorithms and statistical models, the research strives to enhance the accuracy of crime prediction. Ultimately, the findings seek to empower law enforcement agencies with actionable insights. This enables them to adopt pre-emptive measures, fostering a safer environment and bolstering community resilience against crime.

Indexed Terms- Crime prevention, Crime prediction models, Machine learning techniques, Socioeconomic factors, Geospatial analysis, Demographic factors, Predictive analytics, Statistical modelling, Risk assessment, Crime prevention strategies

I. INTRODUCTION

Identifying and predicting crime systematically is made possible through crime analysis and prediction technologies. These tools enable the identification of high-risk areas for criminal activity and the forecasting of potential incidents. The use of sophisticated tools and modern technology by criminals has contributed to an increase in daily crime rates. While certain crimes like arson and burglary show reported increases according to Crime Record Bureau data, others such as domestic violence or murder may not exhibit the same trends. Collating criminal information from various sources such as websites, blogs, and news outlets forms the basis of a comprehensive crime database. Data mining techniques leverage large volumes of data to generate insightful crime reports, identifying areas most affected and aiding in the swift apprehension of offenders.

The process of data mining involves extracting patterns and valuable information from extensive datasets, making it a powerful tool for criminal analysis. Leveraging data repositories to analyse and predict crime patterns becomes crucial in tackling India's rising crime rates, posing significant challenges to law enforcement and intelligence agencies. This project aims to conduct in-depth analyses of relevant crime trends and statistical data to enhance crime prevention strategies and improve community safety.

Within the realm of data science and machine learning, crime prediction focuses on forecasting criminal activities based on historical data and relevant variables. Predictive models analyse patterns and trends to provide law enforcement with actionable insights, aiming to allocate resources more effectively and enhance crime prevention and response strategies. Advanced algorithms and data analytics play a pivotal role in adopting proactive measures to address public safety concerns and enhance community security.

The use of ensemble techniques in crime prediction represents an innovative approach to improving the accuracy and reliability of crime forecasting. Ensemble methods combine the strengths of multiple predictive models to overcome the limitations of individual approaches, offering a robust framework for capturing the complex nature of criminal behaviour and spatial-temporal patterns. Among these methods, the random forest algorithm stands out for aggregating predictions from numerous decision trees trained on different data samples, effectively capturing diverse factors influencing crime occurrences.

In their paper titled "Crime Prediction using Machine Learning and Deep Learning: A Systematic Review and Future Directions" [1], the authors explore models leveraging both machine learning and deep learning techniques to predict various types of crimes and examine their interrelationships. This systematic review highlights the integration of advanced computational methods in crime prediction research, aiming to enhance predictive accuracy and inform law enforcement strategies.

"Ensem_SLDR: Classification of Cybercrime using Ensemble Learning Technique" [2] introduces an innovative approach combining machine learning and natural language processing (NLP) to categorize cybercrime complaints. By aligning these complaints with relevant legal sections, this model provides a robust framework for handling text-based cybercrime data, offering practical insights for law enforcement and legal authorities dealing with cyber threats.

"Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning" [3] delves into empirical studies utilizing machine learning algorithms on datasets from major urban centres. This research focuses on classifying and forecasting criminal activities, utilizing diverse analytical methods to extract meaningful patterns and trends crucial for proactive policing and crime prevention measures.

"Empirical Analysis for Crime Prediction Using Stacked Generalization: An Ensemble Approach" [4] proposes an ensemble learning strategy by combining J48 and C4.5 classifiers. This stacked generalization approach enhances the predictive power of crime forecasting models, integrating multiple classifiers to capture complex relationships within crime data and improve overall accuracy.
"An effective & secure mechanism for phishing attack detection using ML approach" [5] presents a comparative study of machine learning classifiers for detecting phishing attacks. The research emphasizes the efficacy of neural networks in distinguishing malicious URLs, providing a secure mechanism to combat cyber threats and protect users from phishing attempts.

"Crime Analysis and Prediction using Machine Learning" [6] explores the application of various machine learning techniques in crime analysis. This study investigates decision trees, random forests, support vector machines, and neural networks, analysing their effectiveness in predicting crime patterns based on demographic, environmental, and social media data.

"Crime forecasting using spatio-temporal pattern with ensemble learning" [7] employs ensemble methodologies to analyse spatiotemporal crime trends. By integrating data mining techniques and predictive analytics, this research enhances criminological understanding and forecasting capabilities, aiding law enforcement agencies in pre-emptive crime management strategies.

Comparative Analysis of the Related Work
The table 1 discusses the comparative analysis of the current systems in light of the suggested proposal.

Table 1 Comparative Analysis

Sl. No	Author(s)	Algorithms/Techniques	ccuracy
1.	Varun Mandalapu et al. (2023)	Logistic regression, support vector machines, convolution neural network and decision trees.	99%
2.	Hemakshi Pandey et al. (2022)	SVM, Logistic regression, Decision tree, and Random Forest	96.55%
3.	Wajiha Safat et al. (2021)	Random forest and LogisticsRegression.	89%

4.	Asghar and Saira Andleeb Gillani et al. (2021)	Random Tree algorithm, K- Nearest Neighbor (KNN), Bayesian model, Support Vector Machine (SVM) and Neural Network.	99.5%
5.	Sapna Singh Kshatri et al. (2021)	Naïve Bayes algorithm.	78.05%

These were the research papers that we looked over in order to have better comprehend the issue. Many studies have looked into the use of machine learning to predict crimes. A crossover model based on the combination of the J48 and C4.5 classifiers is one approach. Another study shows how powerful deep learning algorithm like CNN and RNN are for predicting criminal activity based on a range of data sources. One approach uses data mining techniques and ensemble classification, while the other focuses on using NLP technology to handle text-based concerns about cybercrime.

II. METHODOLOGY USED

The development of a Crime Prediction model follows a systematic process designed to enhance the accuracy and usability of crime forecasts. This involves multiple stages, each crucial for ensuring the reliability and effectiveness of the model. By integrating advanced data analytics and machine learning techniques, the model aims to provide actionable insights to law enforcement and the public. The following steps outline the methodology used:

a) Data Collection: The process of gathering and measuring information from different sources. The data is collected and stored in a way that makes sense.

b) Data pre-processing: Organizing the selected data by formatting, cleaning, andsampling.

- Formatting: The process of transforming data into a consistent and suitable format foranalysis.
- Data cleaning: It involves deleting or correcting missing data.
- Sampling: Finding the important information in the bigger dataset by doing an analysis on sample of all the data.

c) Features Extraction: Features selection Features that can be utilized to construct the model are chosen. Block, Location, District, Community area, dates, crime description,and day of week are the attributes that are considered in the feature selection process.

d) Evaluation Model: A selection of features that can be used to construct the model is made. The block, location, district, community area, dates, crime description, and day of week are the attributes that are considered in the feature selection process.

e) Final Interface: Providing the public and law enforcement organizations with an easy-to-use interface so they may obtain crime forecasts and associated data.

System Design

Architecture of the Proposed System

The dataset must be gathered, cleaned, and pre-processed as the first stage. The dataset is then trained with machine learning classification techniques. Each model's correctness is determined, and the web application will be implemented using the approach that yields the best results. After integrating the various components of the project – Prediction model, website content analyzer, and blacklist, it is deployed as the final plugin application.

There could be complete gaps in the data or missing numbers in particular columns. The context and the quantity of missing the data must be taken into consideration while choosingthe optimal course of action. Data from several sources may have uneven formatting. For instance, names may have different capitalization or dates may be inputted in several formats. Analytical consistency is ensured by standardizing formats. You will have to determine whether they are accurate or need to be removed. Errors in manual data entry andtypes are frequent in this system. Data profiling tools and data validation guidelines might help reduce these inaccuracies. Figure 1 shows the architecture of the proposed system.

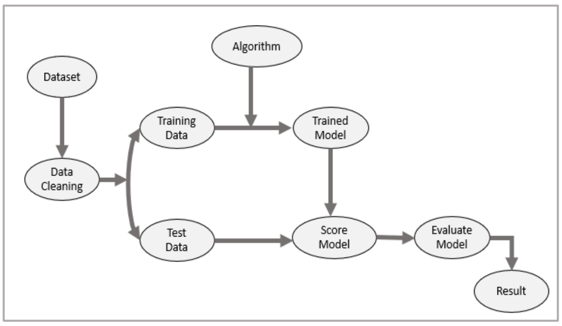


Figure 1: Architecture of the proposed system

System Flowchart

A system flowchart is a way of depicting how data flows in a system and how decisions are made to control events. Figure 2 depicts the system flowchart.

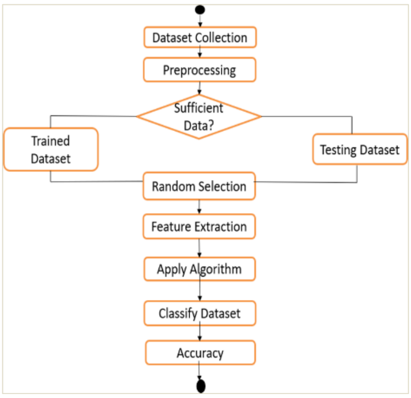


Figure 2: System Flowchart

The dataset goes through extensive cleaning and pre-processing after it is uploaded. This include addressing inaccurate or missing data, standardizing formats, and eliminating outliers or superfluous entries. Making sure the dataset is clean guarantees that the analysisthat follows is founded on accurate and consistent data.

The flowchart depicts a simplified data-cleaning process designed to optimize data for machine learning. Data cleaning is essential in machine learning because it ensures the quality and accuracy of data used to train models. The adage, "garbage in, garbage out" applies here: if you use inaccurate or incomplete data, you'll likely get unreliable results. The process can be broken down into several steps. First, data is collected from various sources. Either human or automated procedures could be used for this. In this situation, it is important to think about what data you will need in order to achieve your analytical objectives and guarantee quality right away. Prior to commencing the primary cleaning cycle, the data could go through the preparation. For missing values, this could mean utilizing more advanced methods, deleting the rows that contain the missing data points, or filling them in with an average value. Also, to make them understandable to machine learning algorithms, categorical variables like hair color could be transformed into a numerical format.

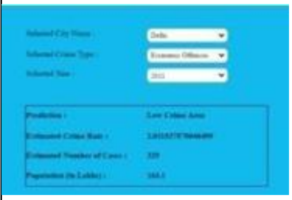
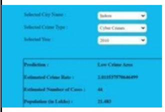
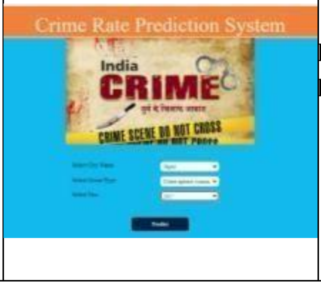
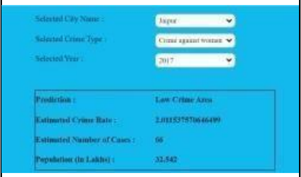
This is typically done through one-hot encoding, which creates a separate binary feature for each category. Additionally, features in your data may be scaled if they have different units.Scaling ensures all features contribute equally during model training. The core of the processis an iterative loop where data is cleaned and assessed. A random subset of the data might bechosen for cleaning, especially for large datasets, to make the process more manageable. In this case, feature extraction is also applicable. To do this, more features must be created from the given data, which can significantly improve a model's functionality. If you have client information that includes purchase history, for instance, you might create a new feature that displays a customer's total spending. After the data has been cleansed, a machine-learning model is trained using it. Based on the task at hand, a different algorithmwill be chosen. The data is then categorized using the model; given the diagram's context, this could entail classifying the data into multiple groups. After the model has been put intopractice, its performance is assessed using metrics like as accuracy, precision, and recall. To decide whether there is sufficient data, there is a decision point. This might be based on another performance parameter or a predefined threshold for model accuracy. The procedure comes to the conclusion when there is enough data, at which point the cleaned data may be utilized to train the final machine learning model. To improve the performance of the model, more cleaning or preprocessing may be applied to the data. To sum up, data cleaning is an iterative process that can necessitate going back to previous phases until the data is judged to be of a high enough quality for machine learning.

III. RESULTS AND DISCUSSION

System Testing

In this crime prevention project, comprehensive system testing is essential to ensure the reliability and accuracy of the predictive models. The system testing phase involves validating the integration of various components, such as data ingestion, preprocessing, machine learning algorithms, and visualization tools. The goal is to confirm that the entire system works seamlessly to provide actionable insights for law enforcement. Table 2 illustrates three critical test cases designed to evaluate the system's performance.

Table 2: Unit test cases

Test case number	Input	Stage	Expected behavior	Observed behavior	Status P=Pass F=Fail
1	ar, place, crime 	Input page	Crime rate is predicted	 Selected City Name : Delhi Selected Crime Type : Economic Offences Selected Year : 2015 Prediction : Low Crime Area Estimated Crime Rate : 2.0115375706466499 Estimated Number of Cases : 339 Population (in Lakhs) : 163.1	P
2	ar, place, crime 	nput Page	Crime rate is predicted	 Selected City Name : Indore Selected Crime Type : Cyber Crimes Selected Year : 2010 Prediction : Low Crime Area Estimated Crime Rate : 2.0115375706466499 Estimated Number of Cases : 44 Population (in Lakhs) : 21.483	P
3	ar, place, ime  Crime Rate Prediction System India CRIME CRIME SCENE DO NOT CROSS Predict	Input Page	Crime rate is predicted	 Selected City Name : Jaipur Selected Crime Type : Crime against women Selected Year : 2017 Prediction : Low Crime Area Estimated Crime Rate : 2.0115375706466499 Estimated Number of Cases : 66 Population (in Lakhs) : 32.542	P

Result Analysis

The main aim of the project was to predict the crime using machine learning algorithms. Table shows the analysis that was performed on the four models with the different trainingand testing sizes. It was found that ensemble model was the most accurate in all the cases.

Table 7.2: Analysis of the four algorithms

Training Size	Testing Size	Accuracy			
		RF	SVM	K-NN	Ens
70%	30%	0.887	0.850	0.901	0.8874
80%	20%	0.9257	0.9342	0.9254	0.9375

Figure 3 shows the bar graph for the accuracy of the four algorithms where the trainset size was 0% and the test set size was 20%

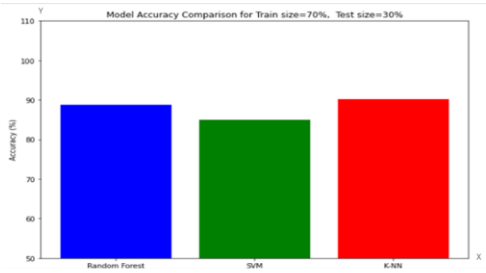


Figure 3: Graph analysis of the first set

CONCLUSION

By carefully analyzing several machine learning techniques, we have made great progress in the field of crime prediction. Through a thorough examination and evaluation of algorithmic correctness, we have discovered a potent predictive model that makes use of a wide range of the machine learning strategies. This model has the potential to be a very useful tool for intelligence services, security services, and law enforcement.

This algorithm is the best choice for crime prediction jobs because it was carefully chosen after extensive testing on representative samples and trained on a large reservoir of historical crime Compared to earlier approaches, our effort has demonstrated a notable improvement in the accuracy and effectiveness of crime prediction by utilizing machine learning. This model has the potential to be a very useful tool for intelligence services, security services, and law enforcement.

This development has the enormous potential to completely transform law enforcement tactics by enabling the preventative actions to lower crime rates and improve public safety. Our machine learning model's proven skills not only improve prediction accuracy but also provide decision-makers with useful insights from thorough data analysis. As we march forward, this model will serve as a lighthouse of innovation, advancing the use of data-driven strategies to prevent crime and promote social welfare.

Scope for Future Work

In the future, we see a number of viable paths to further improve the resilience and effectiveness of our ensemble-based crime prediction model. Initially, investigating new ensemble methods like dynamic ensemble selection and meta-learning may provide fresh perspectives on enhancing prediction accuracy. These approaches improve the model's resistance to changing crime patterns and environmental variables by allowing it to dynamically adjust and integrate the strengths of distinct base learners based on the properties of the input data.

Furthermore, the chance to develop a more responsive and adaptable crime prediction system is attractive because it involves utilizing online learning techniques and integrating real-time data. Through repeatedly improving the model's predictions and adding new data to it frequently, we can make sure that it stays Through iteratively improving the model's predictions and adding new data on a regular basis, we can make sure the model stays tuned to changing crime dynamics and new threats in practical situations. Investigating interpretability and explainability strategies for ensemble models that can also improve transparency and reliability by enabling stakeholders to comprehend the underlying causes of predictions and supporting policymakers' and law enforcement officials' well- informed decision-making.

REFERENCES

[1] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach," in IEEE Access, vol. 9, pp. 67488- 67500, 2021, doi: 10.1109/ ACCESS.2021.3075140.

[2] Pandey, H., Goyal, R., Virmani, D., & Gupta, C. (2022). Ensem_SLDR: Classification of cybercrime using ensemble learning technique. International Journal Computer Network and Information Security,15(1),81.

[3] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning" in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[4] V. Mandalapu, L. Elluri, P. Vyas and N. Roy, "Crime Prediction Using Machine Learningand Deep Learning: A Systematic Review and Future Directions ", in IEEE Access, vol. 11, pp. 60153-60170, 2023, doi: 10.1109/ACCESS.2023.3286344.

[5] Du, Y.; Ding, N. A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods. ISPRS Int. J. Geo-Inf.

2023, 2, 209.

[6] Llaha, "Crime Analysis and Prediction using Machine Learning," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2020, pp. 496-501.

[7] Yu, Chung-Hsien, et al. "Crime forecasting using spatio-temporal pattern with ensemble learning." Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II 18. Springer International Publishing, 2014.

[8] Lamari, Yasmine, et al. "Predicting spatial crime occurrences through an efficient ensemble-learning model." ISPRS international journal of geo-information 9.11 (2020): 645.

[9] Zaidi, Nur Ain Syahira, et al. "A classification approach for crime prediction." Applied Computing to Support Industry: Innovation and Technology: First. International Conference, ACRIT 2019, Ramadi, Iraq.

[10] Hajela, Gaurav, Meenu Chawla, and Akhtar Rasool. "A clustering based hotspot identification approach for crime prediction." Procedia Computer Science 167 (2020):1462-1470.

[11] A. Almaw and K. Kadam, "Crime Data Analysis and Prediction Using Ensemble Learning," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 1918-1923, doi: 10.1109/ICCONS.2018.8663186.