IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 5, September-October 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 5, September-October 2025 ||

DOI:10.15680/IJARETY.2025.1205004

Predicting Crime Categories and Their Frequency using Machine Learning Techniques

Raksha T

Student, Masters of Computer Applications, CMR Institute of Technology, Bengaluru, India

ABSTRACT: This project focuses on predicting crime trends using publicly available datasets to enhance public safety and assist law enforcement agencies in proactive decision-making. By analyzing historical crime records through advanced machine learning techniques, particularly the Random Forest Classifier, the system uncovers complex patterns and relationships between factors such as time, location, crime type, and frequency. The model's high prediction accuracy allows authorities to anticipate high-risk areas and times, optimize resource allocation, and implement targeted preventive strategies. Additionally, the system can be integrated into surveillance and monitoring platforms to provide real-time alerts, aiding rapid response and crime deterrence. This intelligent crime prediction framework demonstrates the transformative potential of AI in fostering safer communities through data-driven policing.

KEYWORDS: Crime Prediction, Machine Learning, Random Forest Classifier, Crime Trends, Public Safety, Predictive Policing, Law Enforcement, Crime Forecasting.

I. INTRODUCTION

Crime has always been a major concern for society, impacting public safety, economic growth, and overall quality of life. With the rapid growth of urban populations and increased availability of digital data, law enforcement agencies face the challenge of efficiently analyzing large volumes of crime-related information to prevent and control criminal activities. Traditional crime prevention methods often rely on manual analysis and reactive measures, which can be time-consuming, resource-intensive, and less effective in predicting future crime patterns

.

In recent years, advancements in data science and machine learning have opened new possibilities for proactive crime prevention. By leveraging historical crime datasets, modern predictive models can identify hidden patterns and forecast potential crime occurrences based on factors such as location, time, and type of crime. This predictive approach allows law enforcement authorities to allocate resources strategically, respond more effectively, and take preventive measures before crimes occur.

This project presents a Crime Prediction System that utilizes machine learning algorithms particularly the Random Forest Classifier to analyze past crime data and forecast future trends. The system processes features like geographical location, date, time, and crime category to determine highrisk areas and time periods. The proposed model achieves high accuracy in prediction, providing a valuable decision-support tool for police departments, policymakers, and urban planners.

By integrating advanced analytics into crime prevention strategies, this project demonstrates how technology can assist in creating safer communities, optimizing law enforcement efforts, and reducing crime rates through timely interventions.

II. LITERATURE REVIEW

Suhong Kim, she Singh Kalsi, Pooya
Taheri, and Param Joshi This study looks at forecasting crimes using neural networks. Two distinct techniques to analysis are used during this research to evaluate criminal activity information from Toronto during the past fifteen decades. By using artificial intelligence models for prediction, such as augmented choices and K-nearest-neighbor, it is possible to forecast homicide in London with a precision of thirty-nine percent to 43 percent.

Benjamin Fredrick David. H and A. Suruliandi
Crime disrupts societal harmony and leaves lasting physical and psychological impacts on victims. Traditional methods of crime data collection and analysis are no longer sufficient, as modern technology enables the extraction of dynamic

behavioral data from sources like geospatial tools, IoT devices, online platforms, and security systems. Advances in data integration, granular classification, and crime sequence identification have paved the way for improved frameworks. Deep learning, with its ability to learn from unstructured data, has transformed crime forecasting, predictive policing, and legal strategy development. This survey examines how large-scale data analysis, demographic evaluation, anomaly detection, trend classification, and neural network-based algorithms can enhance offense prediction and prevention.

Rohit Patil, MuzamilKacchi, PranaliGavali and Komal Pimparia
Violations cause societal annoyance as well as substantial costs that impact culture in a number of forms. Any study and expedites the process of investigating criminals can be financially beneficial. Roughly 10% of offenders are responsible for 50% of all murders. The technology is taught by supplying it criminal activity information through before, which is obtained from a reliable Indian web source. The data includes a variety of crimes, including sexual assault, burglaries, theft, and kidnapped and enslavement, killing someone, and many other comparable offenses Based on Indian statistics collected over the last 14 years (20012014), an equation based on regression is developed to forecast the crime rate in multiple states to the times to come.

Umair Muneer Butt, Sukumar Letchmunan, FadratulHafinaz Hassan, Mubashir Ali, AneesBaqir and Hafiz Husnain Raza Sherazi, this study conducts the first systematic literature review (SLR) on geospatial criminality hotspot identification and forecasting, analyzing 49 relevant works from major databases like IEEE, Science Direct, Springer, Scopus, and ACM. The review highlights the role of data mining and machine learning especially clustering in hotspot detection, crime trend prediction, and the integration of historical and spatial data to improve forecasting accuracy. It also identifies challenges and gaps in current methods, providing a foundation for advancing geospatial crime prediction research.

Nasiri, Zakikhani, Kimiya and Tarek Zayed, This study develops predictive models for external corrosion failures in underground gas transmission pipelines by incorporating both conventional and environmental/geographical factors. Using multiple regression analysis on historical U.S. pipeline data, models were created for the Great Plains and South East regions. Scenario-based analysis identified the most suitable model, achieving RMSE values of 0.04 and 0.07 with R^2 scores of 0.93 and 0.75. These models help estimate the critical time for potential failures, enabling costeffective maintenance planning and reducing reliance on expensive frequent inline inspections.

III. EXISTING SYSTEM

In the existing system, a free database is initially cleaned to remove redundant information and attributes. Deterministic models are then used to gather features from large datasets and identify behavioral patterns, forming the foundation for later classification processes. Deep neural networks are applied to extract crimerelated features, with outcomes computed using training and testing variables to forecast potential criminal incidents. This allows authorities to predict and prevent future unlawful activities. However, the system has limitations. Prior research suffers from low precision because the machine learning model relied heavily on metrics, leading to biased results where standard attributes received higher ratings. Additionally, areas with real-valued attributes and incomplete data are not wellsuited for grouping approaches, and classifier results often require adjustment by selecting the lowest possible threshold value.

Limitations of Existing System

The existing system struggles with low prediction precision because it relies heavily on metric-based evaluations, which cause bias by favoring standard attributes with higher ratings. It is also unsuitable for datasets containing real-valued attributes with incomplete information, making grouping approaches less effective. Additionally, classifier results require manual adjustment by choosing minimal threshold values, which can affect consistency and performance.

IV. PROPOSED SYSTEM

The proposed system enhances accuracy and efficiency by applying a machine learning-based operator and wrapper method to eliminate redundant or unnecessary attributes and reduce data complexity. The cleaned dataset is split into training and testing sets, which are then used to train the model. For easier classification, each crime type, decade, monthly income, time, date, and location are numerically encoded. A Random Forest Classifier is used to analyze the influence of each attribute. Crimes are labeled with historical and location details, enabling precise analysis of crime intensity over time and space. This method identifies the most common crimes while considering spatialtemporal

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 5, September-October 2025 ||

DOI:10.15680/IJARETY.2025.1205004

patterns, and the accuracy score is used to evaluate prediction effectiveness. The model is developed in Python, leveraging artificial intelligence and statistical analytics techniques.

The proposed system offers several advantages. It is highly effective in recognizing crime patterns because it incorporates both location and time variables. It also addresses the challenge of evaluating the individual impact of each attribute, handles missing data without requiring initialization, and achieves higher accuracy compared to other statistical forecasting models.

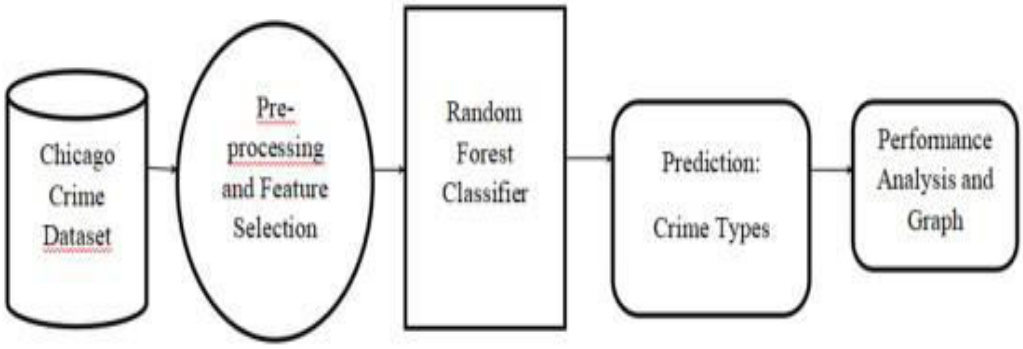

Fig:1 System Architecture

Advantages of the Proposed System

The proposed system is more effective in recognizing crime patterns by incorporating both spatial and temporal factors, making it well-suited for location- and time-based crime detection. It evaluates the influence of each attribute individually, avoids the need for initializing missing values, and produces more reliable results.

Furthermore, it achieves higher accuracy compared to conventional statistical models, improving crime forecasting and enabling better decision-making for preventive measures

V. MODULE DESCRIPTION

Data Collection: This is the initial and most critical step in building the crime prediction model. Data is gathered from reliable sources to ensure model accuracy. In this project, the dataset was obtained from Kaggle's "Predict Crime Rate in Chicago" repository, containing detailed records of incidents reported to the Chicago Police Department. The quality and volume of data directly impact the model's predictive performance.

Dataset: The dataset consists of 520 records with 23 attributes, including identifiers, date/time of the incident, location details, crime classification codes, and geographic coordinates. These features provide a comprehensive view of each reported crime, enabling effective analysis and modeling.

Data Preparation: Before training the model, the dataset undergoes preprocessing to improve quality. This includes removing duplicates, correcting errors, handling missing values, normalizing features, and converting data types where necessary. The data is randomized to eliminate ordering bias, and exploratory data analysis is performed to detect class imbalances or relationships among variables. The processed data is then split into training and testing sets.

Model Selection: The Random Forest Classifier was chosen for its robustness, ability to handle missing values, and high classification accuracy. This ensemble method combines multiple decision trees, each trained on random subsets of data, and uses a voting system to decide the final output. Its advantages include reduced overfitting risk, applicability to both classification and regression, and the ability to determine feature importance.

Analyze and Prediction: The model was trained using selected features Year, Month, Day, DayOfWeek, Minute, Second, Latitude, and Longitude to identify temporal and spatial crime patterns. Once trained, the model predicts the likelihood of a crime occurrence based on new input data, helping identify high-risk times and locations.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 5, September-October 2025 ||

DOI:10.15680/IJARETY.2025.1205004

Accuracy on Test Set: After training, the model was evaluated on the test dataset, achieving an accuracy of 80%. This performance indicates strong predictive capability for the chosen features and algorithm.

Saving the Trained Model: Once validated, the trained model is saved in a .pkl file format using Python's pickle library. This ensures the model can be deployed directly in a production environment without retraining, enabling real-time predictions in crime forecasting applications.

VI. RESULT

The proposed crime prediction system was evaluated using the Chicago crime dataset containing 520 records with 23 attributes. After preprocessing and feature selection, the Random Forest Classifier was trained and tested on a split dataset. The model focused on temporal and spatial variables Year, Month, Day, DayOfWeek, Minute, Second, Latitude, and Longitude to capture crime occurrence patterns.

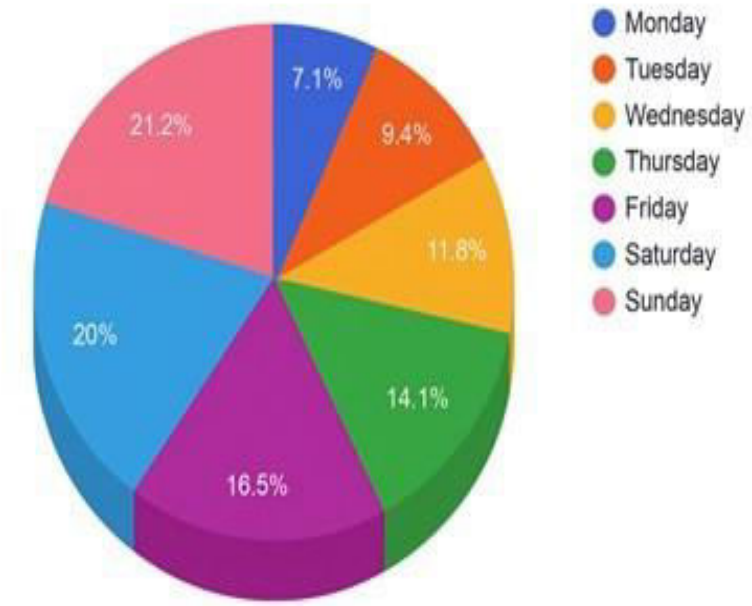

Fig. Piechart of crime in daywise

The system achieved an accuracy of 80% on the test dataset, indicating strong predictive capability for identifying potential crime occurrences based on historical data. The model successfully identified crime-prone time periods and locations, providing valuable insights for proactive law enforcement measures. Feature importance analysis revealed that spatial coordinates (latitude and longitude) and temporal variables (month, day, and day of the week) contributed most significantly to prediction accuracy.

These results demonstrate that the integration of spatial-temporal data with machine learning can effectively support crime forecasting and hotspot identification. Furthermore, the ability to save the trained model enables seamless deployment for real-time prediction, making the system applicable for practical crime prevention and resource allocation strategies.

VII. CONCLUSION

Throughout this study, multiple machine learning algorithms, like Multinominal NB with Stochastic NB, are used to tackle the challenge of handling the corresponding nominal distributed and genuine rated characteristics. The system is ideal for current predictions and does not need a lot of instructional period. Additionally, it addresses the issue of dealing having a constant objective group of elements, which the previous research was unable to accommodate. Consequently, Natural Forest Categorization could one day be used to forecast and identify the most common crimes. A few common measures are also used to compute how well the algorithm performs. The measures that are primarily

DOI:10.15680/IJARETY.2025.1205004

used in technique assessment contain median recall, precision, F 1 score, and correctness. By using statistical methods, precision level might have improved significantly.

REFERENCES

1. Suhong Kim, Param Joshi, Parminder Singh Kalsi,Pooya Taheri, “Crime Analysis Through Machine Learning”, IEEE Transactions on November 2018.
2. Benjamin Fredrick David. H and A. Suruliandi,“Survey on Crime Analysis and Prediction using Data mining techniques”, ICTACT Journal on Soft Computing on April 2012.
3. Shruti S.Gosavi and Shraddha S. Kavathekar,“A Survey on Crime Occurrence Detection and prediction Techniques”, International Journal of Management, Technology And Engineering , Volume 8, Issue XII, December 2018.
4. Chandy, Abraham, "Smart resource usage prediction using cloud computing for massive data processing systems" Journal of Information Technology 1, no. 02 (2019): 108-118.
5. Learning Rohit Patil, MuzamilKacchi, PranaliGavali and Komal Pimparia, “Crime Pattern Detection, Analysis & Prediction using Machine”, International Research Journal of Engineering and Technology, (IRJET) e-ISSN: 2395-0056, Volume: 07, Issue: 06, June 2020
6. Umair Muneer Butt, Sukumar Letchmunan, FadratulHafinaz Hassan, Mubashir Ali, AneesBaqir and Hafiz Husnain Raza Sherazi, “Spatio-Temporal Crime Hotspot Detection and Prediction: A Systematic Literature Review”, IEEE Transactions on September 2020.
7. Nasiri, Zakikhani, Kimiya and Tarek Zayed, "A failure prediction model for corrosion in gas transmission pipelines", Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, (2020).
8. Nikhil Dubey and Setu K. Chaturvedi, “A Survey Paper on Crime Prediction Technique Using Data Mining”, Corpus ID: 7997627, Published on 2014.
9. aCh, Thippa Reddy Gadekallu, Mustufa Haider Abdi and Abdulrahman Al-Ahmari, “Cmputational System to Classify Cyber Crime Offenses using Machine Learning”, Sustainability Journals, Volume 12, Issue 10, Published on May 2020
10. Hyeon-Woo Kang and Hang-Bong Kang, “Prediction of crime occurrence from multimodal data using deep learning”, Peer reviewed journal, published on April 2017.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com