iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73585

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

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 VIII Aug 2025- Available at www.ijraset.com

Predictive Modeling using Data Mining Techniques for Crime Analysis

Bhavana Jakkula[1], G. Praveen Babu[2]
[1]*Post Graduate Student, M. Tech (Data Sciences) Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad*
[2]*Associate Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad*

Abstract: ***This paper presents a crime analysis and prediction system using data mining and machine learning techniques to interpret historical crime data and forecast future incidents. It integrates classification algorithms (Naive Bayes, Decision Trees, SVM), clustering (K-Means), linear regression, and association rule mining (Apriori) to identify crime patterns, hotspots, and trends. Advanced models like Artificial Neural Networks (ANNs) and ensemble methods (Random Forests) enhance prediction accuracy. Visualization tools such as heat maps and trend graphs aid in interpreting data and guiding law enforcement decisions. This system serves as a decision-support tool, enabling proactive crime prevention and smarter policing for safer communities.***
Keywords: Classification algorithms (Navie Bayes, Decision Trees, SVM), Clustering (K-Means), linear regression, association rule mining (Apriori), Advanced models like Artificial Neural Network (ANNs), ensemble method (Random Forests).

I. INTRODUCTION

The rise in crime rates and complexity has challenged traditional crime analysis methods, which struggle with large-scale and dynamic data. As urbanization increases, manual approaches are no longer sufficient for detecting evolving patterns or predicting future incidents. This paper leverages big data, machine learning, and data mining techniques to enhance crime analysis. It employs classification (Naive Bayes, Decision Trees, SVM), clustering (K-Means), association rule mining (Apriori), and regression analysis to identify patterns and forecast crime trends. Advanced models like neural networks and ensemble learning improve predictive accuracy, while visualization tools such as heat maps support intuitive interpretation and decision-making. The system aims to support proactive law enforcement and contribute to safer communities.

II. LITERATURE SURVEY

Crime analysis remains a critical aspect of public safety and law enforcement operations. Traditional approaches often rely on manual review of records and statistical analysis, which are time-consuming, lack scalability, and fail to uncover hidden patterns in vast, heterogeneous datasets. With the advent of data mining and machine learning techniques, crime prediction and pattern recognition have been significantly enhanced. These intelligent systems allow for proactive policing by identifying crime-prone areas, predicting potential criminal events, and recommending preventive actions in real-time.

Khushabu et al. [1] applied the *K-Means clustering algorithm* to group crimes based on similarity in type and frequency. This unsupervised learning approach helped identify high-risk zones and recurring crime trends, enabling targeted intervention strategies for crime prevention.

Benjamin Fredrick David [2] conducted a *comprehensive survey of crime prediction techniques,* emphasizing the use of both supervised and unsupervised machine learning models. The study particularly focused on *text analysis and hybrid models,* highlighting their potential to improve prediction accuracy and calling for deeper exploration of model combinations.

Deepika et al. [3] analyzed Indian crime data using *Random Forest, Neural Networks, and K-Means clustering.* Their research demonstrated that ensemble and deep learning techniques yield high predictive accuracy and advocated the use of *AI-powered bots* to automatically identify crime-prone areas, enhancing real-time response capabilities.

Tushar Sonawane et al. [4] explored *crime pattern recognition through clustering and correlation techniques,* integrated with *data visualization tools* such as heatmaps and pie charts. Their study emphasized the value of linking crime data with demographic factors, facilitating easier interpretation and informed planning.

Rajkumar et al. [5] introduced a *deep learning-based framework* leveraging *neural networks* to predict crime hotspots. Their model incorporated *social media data* to improve real-time crime detection and proposed the extension of such systems to broader datasets for comprehensive surveillance and analysis.

Ginger Saltos and Mihaela Coacea [6] focused on *open crime datasets* to apply various data mining models for prediction. Their work demonstrated how *publicly available data* can be used effectively to build reliable models for identifying potential threats.
Shiju et al. [7] developed a *decision support system* for crime prediction using *classification algorithms*. Their study emphasized practical deployment, offering tools for law enforcement to forecast and react proactively to criminal activity patterns.

III. OBJECTIVE

The primary goal of this paper is to develop an intelligent crime analysis and prediction system using data mining and machine learning to help law enforcement understand crime patterns, forecast incidents, and make informed prevention decisions. By analyzing historical data, the system identifies crime hotspots and predicts potential threats based on type, location, time, and frequency. It integrates advanced techniques classification (Naive Bayes, Decision Trees, SVM), clustering (K-Means), association rule mining (Apriori), regression, and neural networks for comprehensive analysis. Visualization tools like heat maps and dashboards ensure accessible insights for both analysts and decision-makers. The system promotes a shift from reactive to proactive policing and is designed to be scalable, adaptable, and real-time capable.

IV. METHODOLOGY

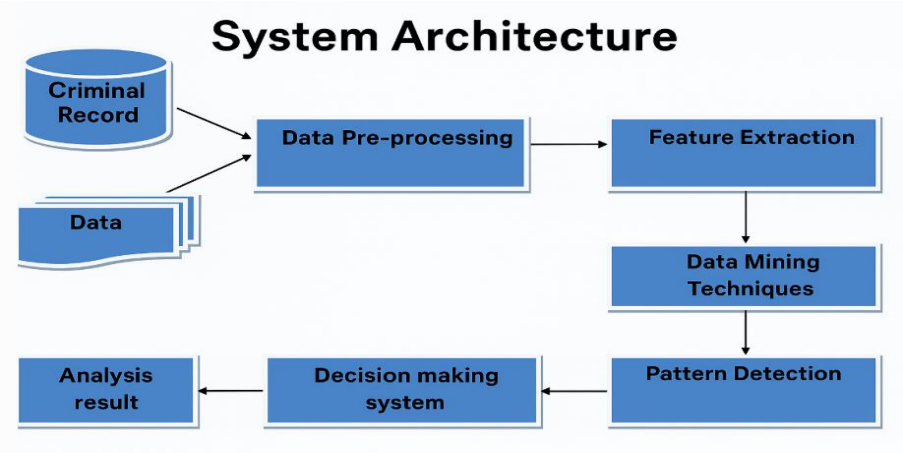


Figure 1: System Architecture for Crime Pattern Detection and Decision Support using Data Mining Techniques

The proposed crime analysis system is a multi-stage data pipeline that begins with collecting structured and unstructured crime data from various sources and processes it through several layers. These include pre-processing, feature extraction, data mining, pattern detection, and decision-making, eventually leading to a final analytical result. Each stage plays a crucial role in transforming raw crime records into actionable insights that support law enforcement and policymaking.

1) Input Layer: Criminal Record and Data Collection

The first and most fundamental phase in crime analysis systems is data collection. This layer gathers criminal records from multiple sources including police First Information Reports (FIRs), surveillance systems, forensic databases, online public records, and government open data portals. These datasets typically consist of structured information such as crime ID, type of crime, location (latitude/longitude or address), time and date, suspect and victim demographics, and unstructured data like narrative descriptions. The data gathered may also include social media feeds and web reports, which can give contextual background to incidents. The primary challenge at this stage is the heterogeneity of data formats, structures, and completeness vary significantly, making it essential to standardize the input before further processing. Ensuring data integrity and minimizing errors during this phase is crucial for downstream accuracy in crime pattern recognition and predictive modeling [2][8].

2) Data Pre-Processing Layer

Once the data is collected, it must undergo rigorous preprocessing to make it suitable for analysis. Preprocessing involves several key steps such as cleaning (removing or correcting missing, noisy, or inconsistent data), integration (combining data from different sources into a coherent dataset), transformation (normalizing or standardizing data), and reduction (dimensionality reduction to

focus only on relevant features). For example, missing crime timestamps may be imputed using statistical techniques, while redundant entries such as duplicate reports are removed. Categorical attributes like crime type or area are encoded numerically, and date fields are reformatted for easier manipulation. The goal is to ensure a clean, consistent, and noise-free dataset that can effectively support machine learning algorithms and statistical analysis. This phase is vital as even small inconsistencies can significantly degrade model performance [3][10].

3) Feature Extraction Layer

In this layer, significant attributes or features are extracted from the pre-processed data to improve model learning and prediction. Feature extraction transforms raw inputs into informative and non-redundant data, enhancing the efficiency and accuracy of predictive models. For example, date fields can be split into day of the week, hour of the day, or weekend vs. weekday to uncover temporal crime trends. Location data is used to derive spatial clusters using geospatial analysis. Textual descriptions of crimes are processed using Natural Language Processing (NLP) techniques like tokenization, stemming, and TF-IDF (Term Frequency–Inverse Document Frequency) vectorization to quantify text for classification. The quality of features directly impacts the success of the analysis, as well-engineered features can highlight hidden patterns and improve classifier accuracy significantly [2][4].

4) Data Mining Techniques Layer

The *Data Mining Techniques layer* is the most critical and analytical phase of the crime analysis system, where various machine learning algorithms are applied to discover patterns, classify data, and predict future crimes. *Decision Trees* are implemented due to their straightforward structure and ease of interpretation, allowing investigators to trace the logic behind crime classification and hotspot identification [4].

Naive Bayes is particularly effective for probabilistic analysis, especially in datasets dominated by textual information, such as crime descriptions [2]. *Support Vector Machines (SVM)* are used when crime categories are clearly separated, offering high classification accuracy through optimal hyperplane selection [3]. For grouping similar incidents or identifying geographical crime concentrations, *K-Means Clustering* is applied, which is especially useful in spatial pattern recognition [1][4]. Additionally, the *Apriori Algorithm* is utilized to find frequent itemsets and association rules, revealing crime co-occurrence patterns such as burglaries often occurring alongside vehicle thefts [4]. To capture complex, non-linear relationships, *Neural Networks* are integrated into the system, proving beneficial for modeling vast and diverse crime datasets [3][5]. To further enhance accuracy and mitigate issues like overfitting or missing data, *Ensemble Learning techniques* like *Random Forests* are employed, combining the strengths of multiple models for more reliable and robust predictions [3][5][6].

5) Pattern Detection Layer

In this phase, the insights derived from data mining are used to detect temporal and spatial crime patterns. Analysts can identify repeat crime areas (hotspots), peak crime hours, seasonal crime surges, and recurring criminal behaviours. For instance, if thefts are repeatedly observed near a specific marketplace on weekends, this pattern can be flagged for preventive action. Visualization tools like heatmaps, trend lines, and cluster maps are often employed to make patterns more interpretable for stakeholders. These findings serve as a basis for understanding criminal behaviour and help law enforcement agencies to deploy resources strategically [4][5].

6) Decision-Making System

This layer converts analytical results into actionable decisions. It supports law enforcement officers, policy-makers, and crime prevention units in planning and decision-making. By integrating the predictions and identified patterns, this system can recommend targeted interventions such as increasing patrols in high-risk areas, placing surveillance equipment at crime-prone spots, or alerting authorities to potential repeat offenders. Decision support systems may also generate alerts for emerging crime trends or sudden spikes in specific crime types. With real-time analytics dashboards and automated reporting, this layer enables informed, data-driven responses to improve public safety [5][7][11].

7) Analysis Result / Output Layer

The final output layer presents result in a user-friendly format for interpretation and action. Outputs include visual dashboards, graphs, time-series analyses, heatmaps, and predictive analytics reports. These are tailored for different stakeholders: police departments can use them for operational planning, city planners for infrastructural safety improvements, and researchers for policy analysis. Interactive visualizations help in understanding crime trends, comparing areas over time, and measuring the impact of

interventions. Advanced systems can integrate GIS for map-based outputs, enabling spatial awareness and effective geographical monitoring of crime [6][10][12].

V. RESULTS AND ANALYSIS

A. *Results*

1) *Navie Bayes*

Naive Bayes Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	8704
1	0.00	0.00	0.00	8742
2	0.00	0.00	0.00	8753
3	0.00	0.00	0.00	8930
4	0.00	0.00	0.00	8814
5	0.00	0.00	0.00	8584
6	0.00	0.00	0.00	8704
7	0.00	0.00	0.00	8797
8	1.00	1.00	1.00	8863
9	0.08	0.67	0.14	8884
10	0.61	1.00	0.76	8644
accuracy			0.12	1179468
macro avg	0.07	0.12	0.07	1179468
weighted avg	0.07	0.12	0.07	1179468

- Accuracy: 0.11989642788104467
- Accuracy Percentage: 11.99%

Analysis: The Naive Bayes model achieved only 11.99% accuracy, with precision, recall, and F1-scores near zero for most classes. This poor performance shows it cannot capture complex patterns in the crime data, likely due to its feature-independence assumption and data imbalance.

2) *Decision Tree*

Decision Tree Classification Report:

	precision	recall	f1-score	support
0	0.42	0.35	0.38	5475
1	0.84	0.97	0.90	5423
2	0.32	0.13	0.19	5287
3	0.88	0.99	0.94	5558
4	0.83	0.98	0.89	5325
accuracy			0.68	162924
macro avg	0.64	0.68	0.65	162924
weighted avg	0.64	0.68	0.65	162924

- Accuracy: 0.6777577275294002
- Accuracy Percentage: 67.78%

Analysis: The Decision Tree model achieved 67.78% accuracy with balanced precision, recall, and F1-scores, especially for classes 1, 2, and 4 (over 0.80), showing its strength in capturing complex, non-linear decision boundaries.

3) *SVM (Support Vector Machine)*

SVM Classification Report:

	precision	recall	f1-score	support
0	0.51	0.41	0.45	480
1	0.56	0.37	0.45	480
2	0.61	0.86	0.71	480
3	0.76	0.36	0.49	480
4	0.49	0.80	0.61	480
accuracy			0.56	2400
macro avg	0.58	0.56	0.54	2400
weighted avg	0.58	0.56	0.54	2400

- Accuracy: 0.5604166666666667
- Accuracy Percentage: 56.04%

Analysis: The SVM achieved 56.04% accuracy, with F1-scores around 0.45–0.60. While better than Naive Bayes, it lags behind Decision Tree, likely due to parameter sensitivity and dataset imbalance or high dimensionality.

4) *Linear Regression*

Regression Evaluation Metrics:

- Mean Absolute Error (MAE): 16.26
- Mean Squared Error (MSE): 395.56
- Root Mean Squared Error (RMSE): 19.89
- R^2 Score: 0.15

Analysis: Moving on to regression analysis, Linear Regression was applied to predict continuous outcomes. The evaluation metrics include a Mean Absolute Error (MAE) of 16.26, Mean Squared Error (MSE) of 395.56, Root Mean Squared Error (RMSE) of 19.89, and a very low R^2 Score of 0.15. These results suggest a poor fit of the linear model to the data, indicating that linear relationships do not capture the underlying patterns in the dataset well.

5) K-Means (Clustering)

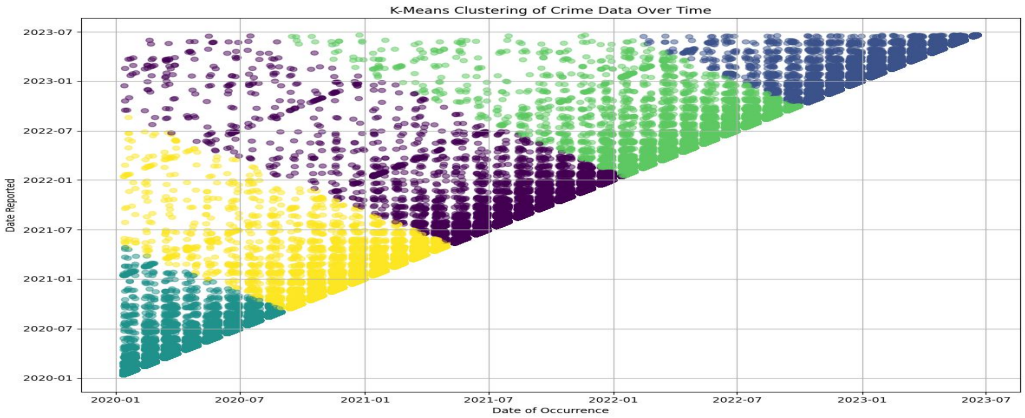


Figure 2: K-Means Clustering of Crime Incidents by Date of Occurrence and Date Reported

Analysis: K-Means clustering revealed clear patterns in crime data, with heatmaps showing distinct clusters by date and area, helping identify hotspots and trends, though no accuracy metric applies in unsupervised learning.

6) Apriori Algorithm (Association rule)

🔗 Association Rules (Apriori):

	antecedents	consequents	support	confidence	lift
0	(Assault)	(Urban)	0.4	1.0	1.25
1	(Urban)	(Assault)	0.4	0.5	1.25
2	(Urban)	(Night)	0.4	0.5	1.25
3	(Night)	(Urban)	0.4	1.0	1.25

Analysis: The Apriori Algorithm was used for mining association rules from the data. The derived rules such as {Urban} → {Assault} and {Night} → {Urban} with a confidence of 1.0 and lift of 1.25 suggest strong co-occurrence patterns between urban locations and assault crimes, particularly during night hours. These insights can be valuable for predictive policing and resource allocation.

7) Neural Network

📨 Neural Network Classification Report:

	precision	recall	f1-score	support
0	0.90	0.85	0.88	3393
1	0.90	0.93	0.91	4687
accuracy			0.90	8080
macro avg	0.90	0.89	0.89	8080
weighted avg	0.90	0.90	0.90	8080

🔹 Final Accuracy: 0.8976485148514851
🔹 Accuracy Percentage: 89.76%

Analysis: Among all models, the Neural Network classifier demonstrates exceptional performance with a final accuracy of 89.76%. Precision, recall, and f1-scores for both classes are high, indicating balanced and accurate predictions. This showcases the model's ability to learn complex, non-linear patterns in the dataset, particularly after using techniques like feature scaling, balancing, and appropriate model tuning.

8) Random Forest

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.78	0.81	0.79	6324
1	0.83	0.81	0.82	6324
2	0.90	0.87	0.88	6324
3	0.88	0.82	0.85	6324
4	0.79	0.85	0.82	6324
accuracy			0.83	31620
macro avg	0.83	0.83	0.83	31620
weighted avg	0.83	0.83	0.83	31620

Accuracy: 0.832764073371284
Accuracy Percentage: 83.28%

Analysis: Random Forest achieved 83.27% accuracy with high precision and recall, offering robust, well-generalized performance through its ensemble approach. Overall, the Neural Network performed best, followed by Random Forest and Decision Tree, while Naive Bayes and Linear Regression performed poorly. Together, classification, regression, clustering, and association analysis provide a strong framework for understanding and predicting crime patterns.

9) Visualization using Heatmap

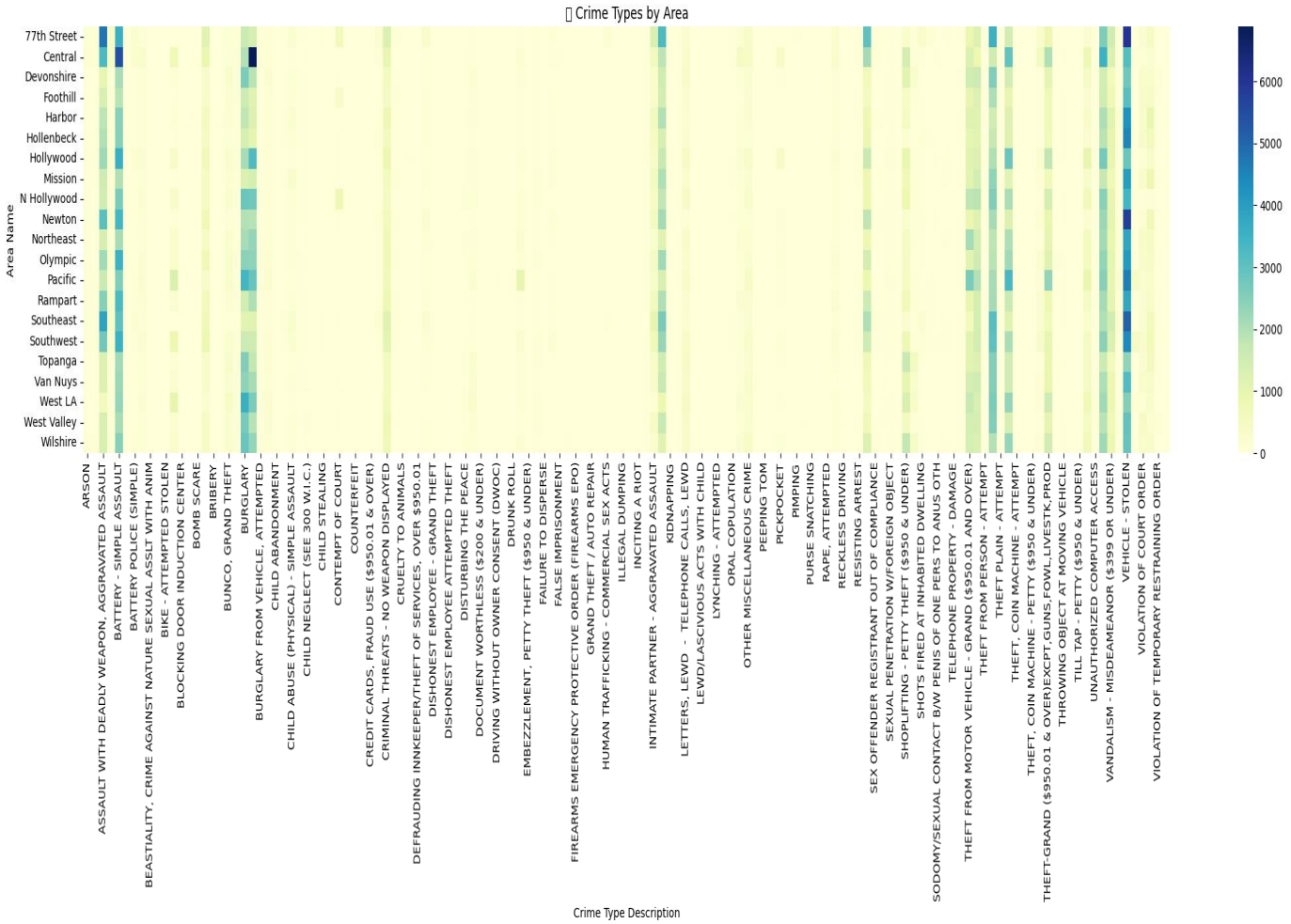


Figure 3: Crime Types Distribution by Area

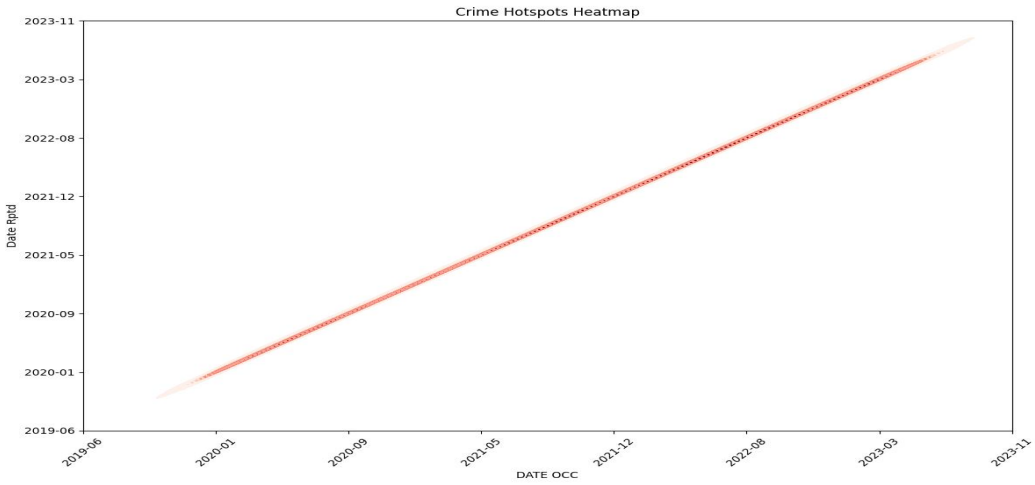


Figure 4: Temporal Crime Hotspots by Date of Occurrence and Date Reported

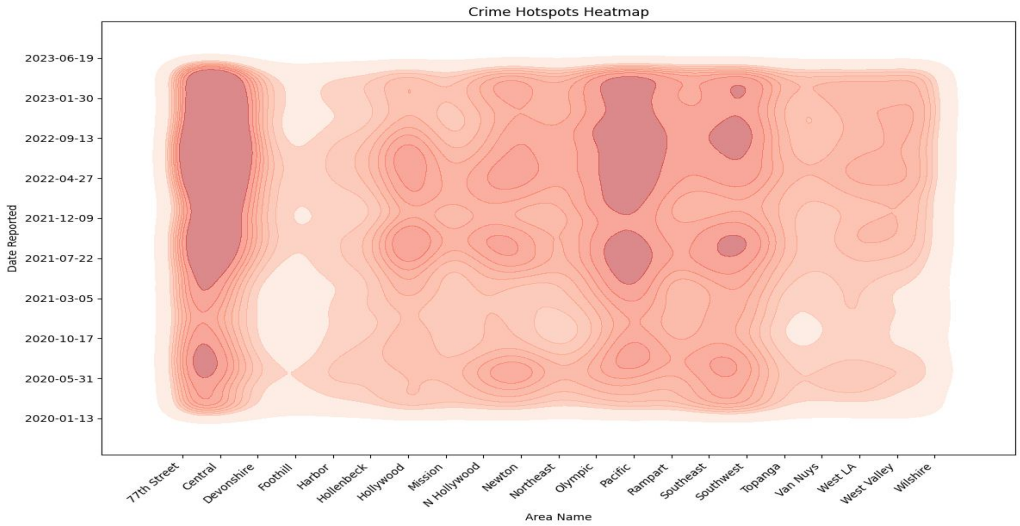


Figure 5: Spatial-Temporal Crime Intensity Across Areas

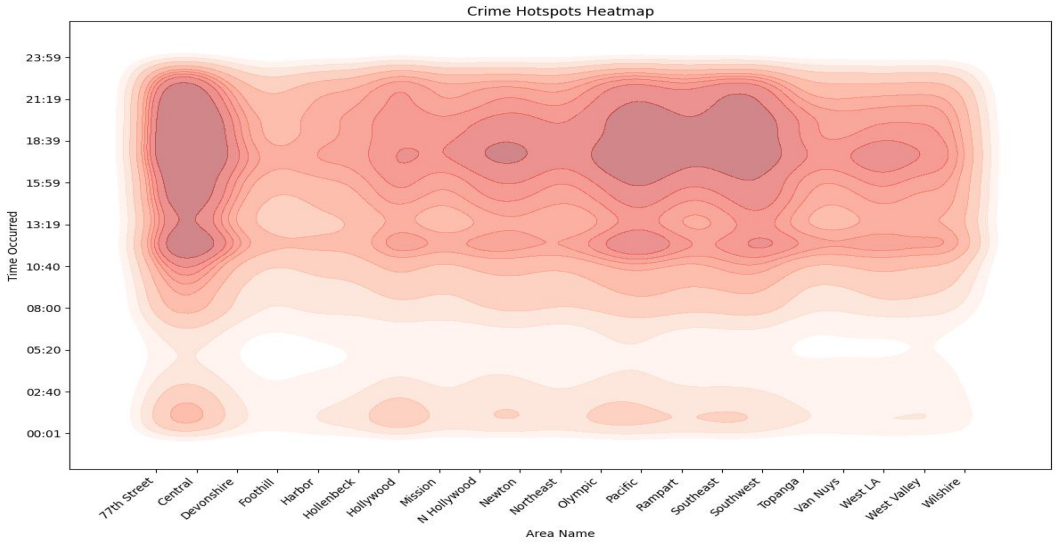


Figure 6: Crime Occurrence Patterns by Area and Time of Day

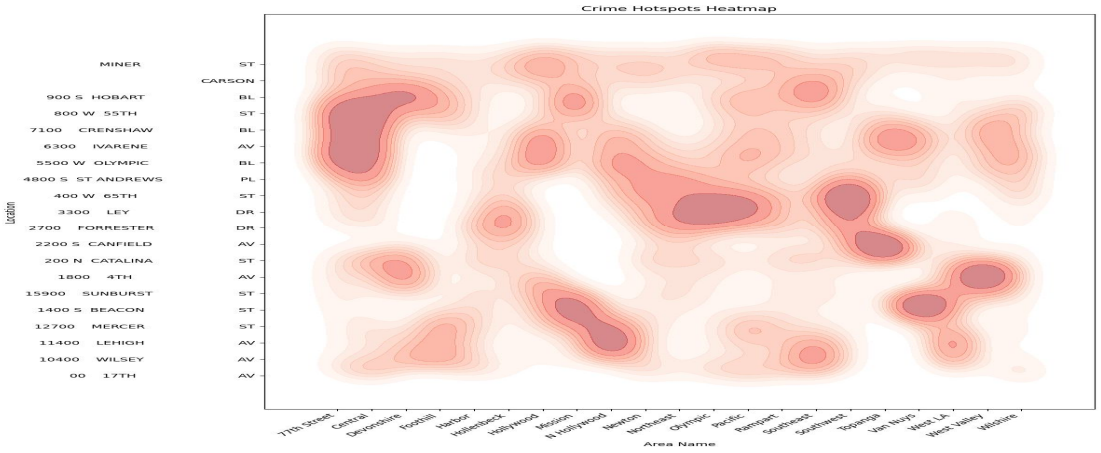


Figure 7: Crime Hotspots by Specific Locations and Areas

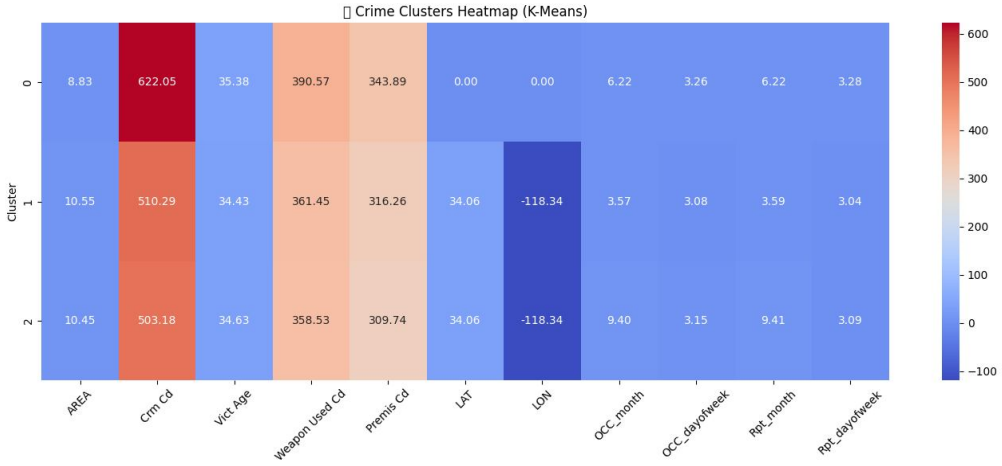


Figure 8: K-Means Cluster Distribution of Crime Attributes

Analysis: The final section focuses on Visualization using Heatmaps, which serves as an essential tool for exploring crime data spatially and temporally. The first heatmap shows crime types by area, with clear visual density gradients indicating which regions experience higher frequencies of specific crime categories. Lighter colours indicate higher counts, allowing quick identification of hotspots. Subsequent heatmaps illustrate crime hotspots over time, both by date and hour of the day. These time-based heatmaps reveal recurring patterns, such as spikes in crimes during certain months or late-night hours, which can aid in strategic deployment of law enforcement resources. The contour-style heatmaps also depict intensity variations across area names and time, providing a more detailed understanding of crime concentration. Lastly, a K-Means clustering heatmap visualizes grouped crime clusters by area and crime count. This final visualization helps reinforce earlier clustering insights by showing which areas consistently fall into high-risk clusters. Overall, these heatmaps not only complement the model-based results but also offer actionable insights for crime prevention and urban safety planning.

VI. CONCLUSION AND FUTURE SCOPE

This paper demonstrates the effective use of data mining in law enforcement by analyzing historical crime data to uncover patterns, identify hotspots, and predict future crimes. It employs various techniques—classification (Naive Bayes, Decision Trees, SVM), linear regression, clustering (K-Means), association rule mining (Apriori), neural networks, and ensemble learning for accurate and robust analysis. Visualization tools like heatmaps enhance result interpretation, aiding strategic planning and proactive crime prevention. Overall, the system offers a scalable, intelligent approach that transforms raw data into actionable insights, supporting smarter and safer policing.

A. Future Scope

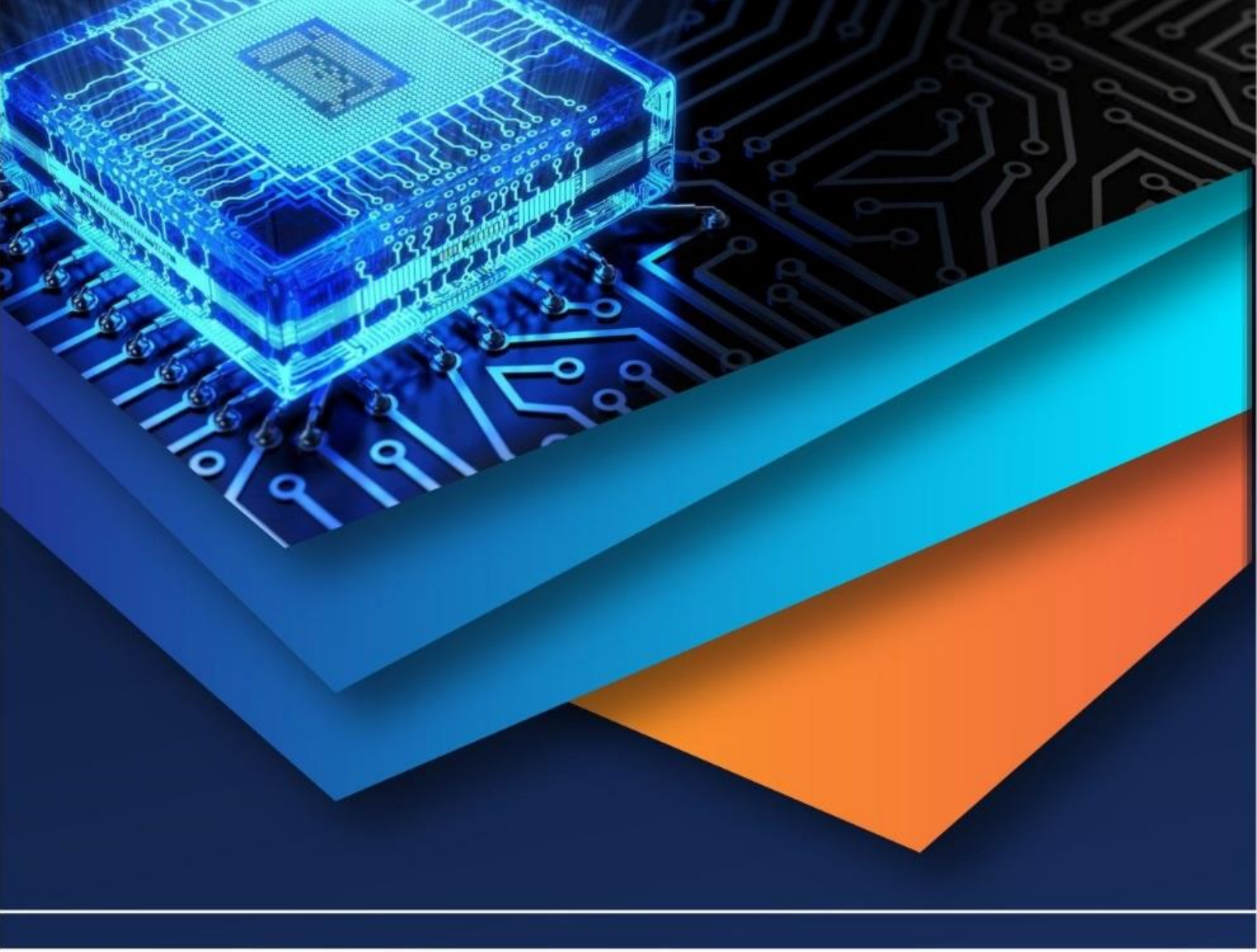
Future enhancement of this paper may involve integrating real-time crime data, incorporating geospatial and social media inputs, and applying deep learning models like LSTM for improved time-based predictions. Developing mobile and web applications and considering contextual factors such as weather and local events can also contribute to more accurate and timely crime forecasting.

VII. ACKNOWLEDGEMENT

I would like to express my sincere appreciation to G. Praveen Babu sir for his consistent support and encouragement throughout the course of this project. I am also deeply grateful to the authors of the research papers referenced in this study, whose insightful work has significantly informed and enhanced the quality of this research.

REFERENCES

[1] Khushabu A. Bokde, Tisksha P. Kakade, Dnyaneshwari S. Tumasare, Chetan G. Wadhai B.E Student, Crime Detection Techniques Using Data Mining and K-Means, International Journal of Engineering Research & technology (IJERT) ,2018.
[2] H. Benjamin Fredrick David and A. Suruliandi, Survey on crime analysis and prediction using data mining techniques, ICTACT Journal on Soft computing, 2017.
[3] Deepiika k.K, Smitha Vinod, Crime analysis in india using data mining techniques, international journal of Engineering and technology, 2018.
[4] Tushar Sonawanev, Shirin Shaikh, rahul Shinde, Asif Sayyad, Crime Pattern Analysis, Visualization And prediction Using Data Mining, Indian Journal of Computer Science and Engineering (IJCSE), 2015.
[5] RajKumar.S, Sakkarai Pandi.M, Crime Analysis and prediction using data mining techniques, International Journal of recent trends in engineering & research,2019.
[6] Ginger Saltos and Mihaela Coacea, An Exploration of Crime prediction Using Data Mining on Open Data, International journal of Information technology & Decision Making,2017.
[7] Shiju Sathyadevan, Devan M.S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing (IEEE) 2014.
[8] Sarpreet kaur, Dr. Williamjeet Singh, Systematic review of crime data mining, International Journal of Advanced Research in computer science, 2015.
[9] Ayisheshim Almaw, Kalyani Kadam, Survey Paper on Crime Prediction using Ensemble Approach, International journal of Pure and Applied Mathematics,2018.
[10] Dr. M. Sreedevi, A. Harha Vardhan Reddy, ch. Venkata Sai Krishna Reddy, Review on crime Analysis and prediction Using Data Mining Techniques, International Journal of Innovative Research in Science Engineering and technology ,2018.
[11] K.S.N. Murthy, A.V.S. Pavan kumar, Gangu Dharmaraju, international journal of engineering, Science and mathematics, 2017.
[12] Hitesh Kumar Reddy Toppy Reddy, Bhavana Saini, Ginika Mahajan, Crime Prediction &Monitoring Framework Based on Spatial Analysis, International Conference on Computational Intelligence Data Science (ICCIDS 2018).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY