International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

AI-Based Women's Safety Risk Prediction Using GIS, Crime Data, and Machine Learning

Shreya Raut[1], Sakshi Tarapure[2], Pragati Bhenki[3]

[1,2,3]*MSc Computer Applications, Semester III (2026–2027), School of Information Technology, Indira University, Pune, India*

Under the guidance of Dr. Kavita Dhakad

Asst. Prof., School of Information Technology, Indira University, Pune, India

***Abstract*— Women's safety is a serious social concern, yet most existing safety technologies react after an incident has begun. In this paper, an AI-based system is proposed to estimate the risk level of an area. The risk level is classified as Low, Medium or High from criminal activity data using machine learning algorithms and GIS visualisation techniques. A synthetic dataset containing 5,000 records and 12 attributes was used. After feature engineering, 13 input features, including location, crime type, population density and time, were used to train four classifiers: Logistic Regression, Decision Tree, Random Forest and XGBoost. The data was split using an 80/20 stratified train–test split. XGBoost achieved the highest accuracy (81.9%) and a macro F1-score of 0.81, while Decision Tree and Random Forest each achieved 81.6% accuracy. Population density, hour of the day and geographic location were more important than the other features. The records were also mapped using Folium as an interactive map based on risk levels. The study is a research aid and not a guarantee of safety.**

***Keywords*—crime data, Folium, GIS, machine learning, Random Forest, risk prediction, women's safety, XGBoost**

I. Introduction

Crimes such as harassment, stalking, assault, kidnapping and domestic violence affect women's daily life, travel decisions and access to public spaces; routine activity theory explains such crime as arising where offenders, targets and weak guardianship meet in space and time [26]. Historical crime records, location information and open datasets make it possible to study where, when and what kind of incidents occur, and to estimate the risk associated with a location.

This paper presents a study of crime datasets containing attributes such as location, time, crime type and population density. Machine learning algorithms are used to classify each record as Low, Medium or High risk, and the records are mapped using GIS. Four different algorithms have been implemented using Python [21].

A. Existing System and Need for System

Women's safety remains an important social concern. Many safety tools are designed to respond once an incident has started, rather than helping people understand risk beforehand. This paper presents a framework that groups an area's risk as Low, Medium or High by combining historical crime records, machine learning and GIS visualisation. The limitations of existing approaches are as follows:

1. *Reactive safety technologies:* mobile applications, GPS tracking, wearables and IoT devices mainly give alerts after an incident has started, instead of predicting risk before it happens. These tools are therefore reactive rather than proactive.
2. *GIS analysis:* many GIS studies map incidents and hotspots, but they do not place locations into clear risk categories. This lack of classification limits their practical use.
3. *General crime prediction:* machine learning studies usually address crime in general, not women's safety specifically.
4. *Lack of integration:* mapping, prediction and evaluation are presented separately.

The need therefore is for a women-specific, location-based framework that uses multiple features (geography, time, crime type, population density), compares several models using class-wise metrics, and offers interpretability through feature importance and maps.

B. Scope of Work

The scope covers (i) data understanding and pre-processing, (ii) exploratory data analysis, (iii) training Logistic Regression, Decision Tree, Random Forest and XGBoost, (iv) evaluation with accuracy, precision, recall, F1-score and confusion matrices, and (v) GIS visualisation and prediction for a new case.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

Real-time incident reporting, emergency dispatch and any guarantee of actual safety are out of scope. The results may be useful to women travellers, students, working professionals, police departments, urban planners and government agencies.

C. Operating Environment – Hardware and Software

Hardware:

- A dual-core or better processor is needed.
- 8 GB or more of RAM is recommended.
- A few GB of storage space is required.
- An active internet connection is needed.

Software:

- Google Colab with the Python 3 programming language is required.
- Pandas and NumPy are used for working with data.
- Scikit-learn is used for pre-processing, Logistic Regression, Decision Tree, Random Forest and evaluation metrics.
- XGBoost is used for advanced boosting algorithms.
- Matplotlib and Seaborn are used for plotting graphs.
- Folium is used for creating an interactive map using Esri World Street Map tiles.
- Windows 10/11 or another operating system with a modern browser, such as Google Chrome, is recommended.

II. Literature Review

Previous research provides useful support for spatial and time-based crime forecasting. Systematic reviews [1], [4] emphasise the importance of validating models carefully and using both location and time features. Hotspot mapping has been used to identify spatial crime patterns [25], and predictive-policing field trials have also been reported [28]. Reviews of AI-based crime prediction [5], [11] and comparisons of machine learning methods [27] provide reasons to test multiple classifiers, including Decision Tree, Random Forest and boosting methods. Research focused on women's safety has combined GIS and statistical modelling to study crimes against women in West Bengal [2], as well as spatial clustering and machine learning in Bhopal [3]. Reviews of safety technologies in India [6], assault detection [7] and IoT-based systems [8] describe solutions that often focus on devices or responses. Video-based deep learning work [9], by contrast, relies on surveillance footage rather than structured crime records. Other studies explore safe-route prediction [10], crime-type prediction with risk scores [12], forecasting methods [13] and multidimensional crime indices for crimes against women [14].

Research gap: prior work has examined safety technologies, GIS hotspots and machine learning prediction, but these components are not always combined in one framework. Some safety systems mainly respond to incidents, GIS studies often focus on historical patterns, and general crime models may not address women's safety directly. This project brings together historical crime data, contextual features and GIS to classify risk as Low, Medium or High and make the results easier to interpret.

III. Proposed System

A. Proposed System

The proposed framework integrates historical crime data, machine learning classification and GIS visualisation to give an understandable risk category for each record. Its workflow has six stages, as shown in Table I.

TABLE I
Workflow Stages of the Proposed Framework

Stage	Description
1. Dataset	Load the crime dataset (CSV) with location, date, time, crime type, victim gender, population density, police station and risk level.
2. Pre-processing	Convert Date and Time, extract Year, Month, Day of Week, Hour and Time Period, handle missing values and encode categorical features.
3. Model training	Train Logistic Regression, Decision Tree, Random Forest and XGBoost on the 80% training split.
4. Evaluation	Compare accuracy, precision, recall, F1-score and confusion matrices on the 20% test split.
5. GIS visualisation	Plot all records on a Folium map by latitude and longitude, coloured by risk level.
6. Prediction	Predict the risk level and class probabilities for a new case.

1) Feasibility Study: The project is feasible from the following viewpoints.

- *Technical:* the project can be built using only free Python libraries and Google Colab; no special hardware is required.
- *Economic:* all tools are free or open source, so the cost of development is very low.
- *Operational:* a simple Low, Medium or High label and a colour-coded map help citizens, police and planners to understand the results easily.

B. Objectives

The objectives of this study are as follows:

1. To analyse historical crime and contextual data (location, time, crime type, population density).

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

2. To pre-process the data and engineer temporal and categorical features.
3. To compare Logistic Regression, Decision Tree, Random Forest and XGBoost for predicting Low, Medium and High risk.
4. To evaluate the models using accuracy, precision, recall, F1-score and confusion matrices.
5. To identify the most influential features using Random Forest feature importance.
6. To visualise the geographic distribution of risk on an interactive GIS map.
7. To demonstrate risk prediction, with probabilities, for a new case.

C. Research Requirements

- *Dataset:* a CSV file holding 5,000 records and 12 attributes: Crime_ID, Date, Time, State, District, Latitude, Longitude, Crime_Type, Victim_Gender, Population_Density, Police_Station and Risk_Level, which is the target column to be predicted.
- *Tools:* Python running in Google Colab with pandas, NumPy, scikit-learn, XGBoost, Matplotlib, Seaborn and Folium.
- *Knowledge base:* classification, data pre-processing and encoding, evaluation metrics and basic GIS visualisation guide the analysis.

IV. Dataset and Exploratory Data Analysis

A. Dataset

The dataset contains 5,000 rows and 12 columns (listed in Section III-C). It covers 8 states and 23 districts, with 10 crime types, 230 police stations and 3 risk levels (Low, Medium and High). Latitude ranges from 9.82 to 28.82 and longitude from about 72.44 to 88.48. Mean population density is about 9,079 (minimum 1,946, maximum 32,835). Female victims account for 4,852 records. There were no missing values and every Crime_ID is unique, so no records were removed.

B. Exploratory Data Analysis

TABLE II
Risk Level Distribution

Risk Level	Records	Percentage
Medium	2,808	56.16%
Low	1,135	22.70%
High	1,057	21.14%
Total	**5,000**	**100%**

Medium is the largest class, so class-wise precision and recall are reported alongside accuracy [22].

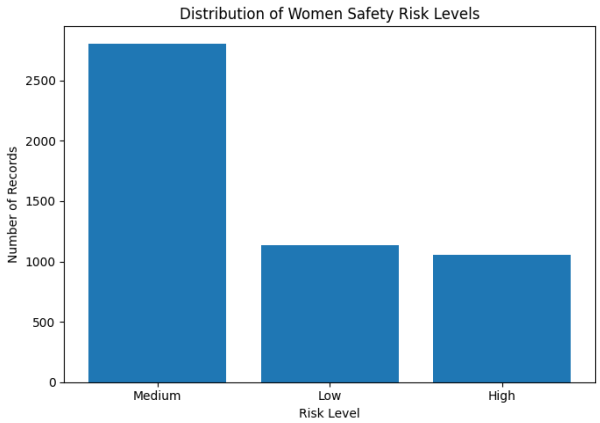


Fig. 1. Distribution of women's safety risk levels

TABLE III
Crime Type vs Risk Level (Number of Cases)

Crime Type	High	Low	Medium	Total
Cyber Harassment	49	201	371	621
Domestic Violence	171	33	263	467
Eve Teasing	109	107	349	565
Harassment	124	156	439	719
Kidnapping	76	0	61	137
Public Intimidation	37	192	286	515
Robbery	168	23	227	418
Sexual Assault	161	0	85	246
Stalking	108	117	307	532
Theft	54	306	420	780

Sexual Assault and Kidnapping have no Low-risk records, while Domestic Violence, Robbery and Sexual Assault have high shares of High risk.

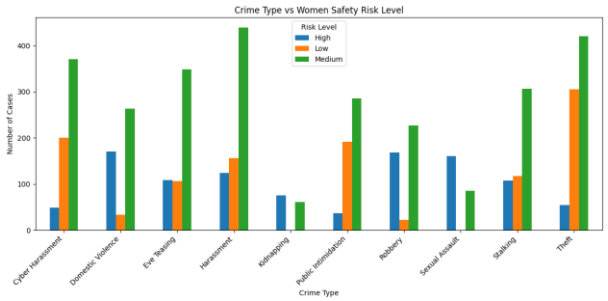


Fig. 2. Crime type vs women's safety risk level

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

TABLE IV
TIME PERIOD VS RISK LEVEL (NUMBER OF CASES)

Time Period	High	Low	Medium	Total	High-risk share
Morning (05:00–12:00)	69	403	526	998	6.9%
Afternoon (12:00–17:00)	58	468	608	1,134	5.1%
Evening (17:00–21:00)	466	185	929	1,580	29.5%
Night (21:00–05:00)	464	79	745	1,288	36.0%

High risk is concentrated in the evening and night.

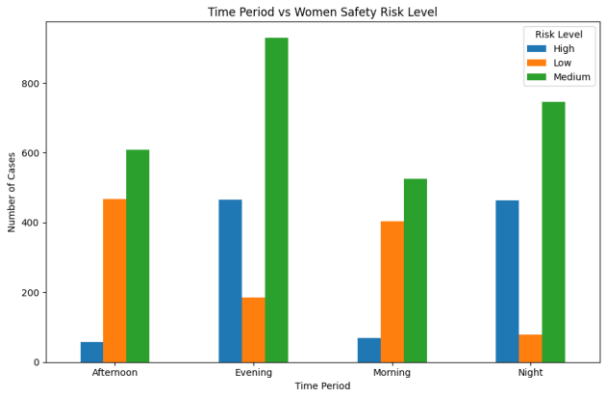


Fig. 3. Time period vs women's safety risk level

TABLE V
STATE-WISE RISK DISTRIBUTION (NUMBER OF CASES)

State	High	Low	Medium	Total	High-risk share
Delhi	297	9	334	640	46.4%
Gujarat	174	39	321	534	32.6%
Karnataka	25	252	281	558	4.5%
Maharashtra	192	323	716	1,231	15.6%
Rajasthan	44	233	369	646	6.8%
Tamil Nadu	90	149	323	562	16.0%
Telangana	110	110	242	462	23.8%
West Bengal	125	20	222	367	34.1%

These counts describe the supplied dataset and should not be read as official state crime rates.

V. RESULT AND DISCUSSION

A. Research Methods

1) Feature Selection: Risk_Level is the target. Crime_ID was dropped because it carries no predictive information. Date was replaced by Year, Month and Day_of_Week, and Time was replaced by Hour and Time_Period, because the EDA showed that risk changes considerably with the time of day. The other features were kept as they were. This gave 13 input features: State, District, Latitude, Longitude, Crime_Type, Victim_Gender, Population_Density, Police_Station, Year, Month, Day_of_Week, Hour and Time_Period.

2) Pre-processing: All the steps were combined in one scikit-learn ColumnTransformer and Pipeline, so that the same transformations are applied to the training data, test data and new cases. Time_Period is defined as Morning (05:00–11:59), Afternoon (12:00–16:59), Evening (17:00–20:59) or Night (all other hours). A median imputer was used for the numerical features and a most-frequent imputer for the categorical features. The categorical features State, District, Crime_Type, Victim_Gender, Police_Station and Time_Period were one-hot encoded. The XGBoost labels were encoded as High = 0, Low = 1 and Medium = 2. The data was split 80/20 with stratification (random_state = 42), which gave 4,000 training records and 1,000 test records.

3) Models: Each model is a pipeline of pre-processing and a classifier. The settings used are given in Table VI.

TABLE VI
SETTINGS OF THE CLASSIFICATION MODELS

Model	Settings
Logistic Regression [20]	max_iter = 1000, random_state = 42
Decision Tree [19]	max_depth = 10, random_state = 42
Random Forest [18]	n_estimators = 200, random_state = 42, n_jobs = -1
XGBoost [17]	n_estimators = 200, max_depth = 6, learning_rate = 0.1, eval_metric = mlogloss, random_state = 42

B. Experiment

The experiment loaded the data, created the date and time features and then split the data 80/20. The four pipelines were trained on the 4,000 training records, and predictions were made on the 1,000 test records. The predictions were compared with the labels using accuracy, weighted precision, recall and F1-score, class-wise reports and confusion matrices. Random Forest importance values were used to identify the key features. All records were then plotted on a Folium map (High = red, Medium = orange, Low = green), and XGBoost was used to predict the risk for a new case.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

C. Results

Metrics were computed with scikit-learn's confusion matrix and classification report [15], [16]. The overall performance of the four models is given in Table VII and compared in Fig. 4.

TABLE VII
OVERALL PERFORMANCE ON THE TEST SET (1,000 RECORDS, WEIGHTED AVERAGES)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	75.70	76.11	75.70	75.14
Decision Tree	81.60	81.73	81.60	81.48
Random Forest	81.60	81.89	81.60	81.35
XGBoost	81.90	81.85	81.90	81.85

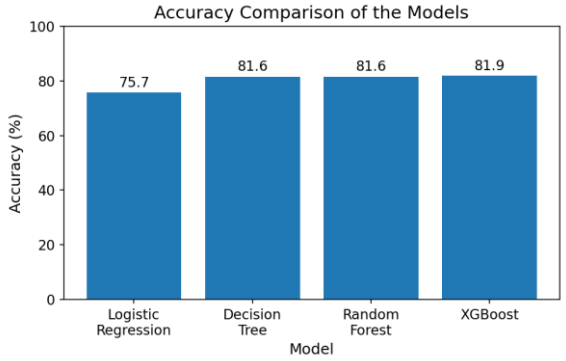


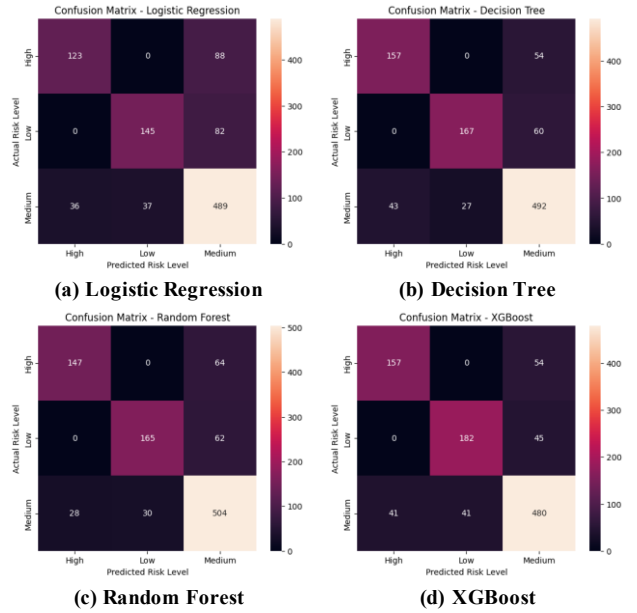
Fig. 4. Accuracy comparison of the machine learning models

TABLE VIII
CLASS-WISE PERFORMANCE

Model	Class	Precision	Recall	F1-Score
Logistic Regression	High	0.77	0.58	0.66
	Low	0.80	0.64	0.71
	Medium	0.74	0.87	0.80
Decision Tree	High	0.79	0.74	0.76
	Low	0.86	0.74	0.79
	Medium	0.81	0.88	0.84
Random Forest	High	0.84	0.70	0.76
	Low	0.85	0.73	0.78
	Medium	0.80	0.90	0.85
XGBoost	High	0.79	0.74	0.77
	Low	0.82	0.80	0.81
	Medium	0.83	0.85	0.84

Macro-average F1-scores are 0.72 (Logistic Regression), 0.80 (Decision Tree), 0.80 (Random Forest) and 0.81 (XGBoost). XGBoost achieved the highest accuracy and macro F1 and the best Low-risk recall (0.80), and was selected as the final model, although its margin over Decision Tree and Random Forest is only 0.3 percentage points. Logistic Regression is about 6 points less accurate, suggesting a non-linear relationship between features and risk; it also raised a convergence warning because numerical features were not scaled.

1) Confusion Matrix (XGBoost): The system correctly classified 157 High, 182 Low and 480 Medium records, making 819 out of 1,000 correct. Most mistakes come from the Medium class, and High and Low records are rarely mixed up with one another. Of 211 actual High-risk records, 54 were predicted as Medium, which matters because under-estimating risk is the more serious error in a safety application. The confusion matrices of all four models are shown in Fig. 5.



(a) Logistic Regression **(b) Decision Tree**

(c) Random Forest **(d) XGBoost**

Fig. 5. Confusion matrices of the four models (rows = actual, columns = predicted)

2) Feature Importance: Random Forest importance values rank Population_Density, Hour, Longitude and Latitude as the four most important features (Table IX and Fig. 6).

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

TABLE IX
TEN MOST IMPORTANT FEATURES (RANDOM FOREST)

Rank	Feature	Importance
1	Population_Density	0.1305
2	Hour	0.0917
3	Longitude	0.0780
4	Latitude	0.0774
5	Month	0.0445
6	Day_of_Week	0.0394
7	Time_Period = Night	0.0301
8	Crime_Type = Sexual Assault	0.0285
9	Time_Period = Afternoon	0.0284
10	Year	0.0231

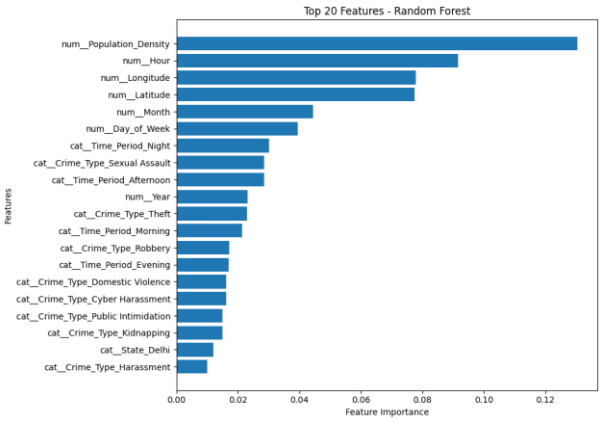


Fig. 6. Top 20 features by Random Forest importance

3) Prediction for a New Case: For State = Maharashtra, District = Pune, Latitude = 18.5204, Longitude = 73.8567, Crime_Type = Harassment, Victim_Gender = Female, Population_Density = 10,000, Year = 2026, Month = 9, Day_of_Week = Sunday and Hour = 21 (Night), XGBoost predicted High risk, with probabilities of 83.02% High, 16.95% Medium and 0.03% Low.

D. Discussion

XGBoost gave the best overall balance. The three tree-based models clearly outperform the linear baseline. Time and location matter a lot: Hour, coordinates and population density rank highly, which matches what was seen in the EDA, where evening and night records are often labelled High risk.

The machine learning model and the GIS map work well together. The map shows where the labelled risk is concentrated, while the model classifies records into the same three risk levels. This combination helps to make sense of the data and to spot patterns.

1) Limitations and Responsible Interpretation:

- Results depend on the quality and origin of the synthetic Risk_Level labels; the model learns to predict existing labels.
- A random train/test split does not prove performance on new cities or future periods.
- Some actual High-risk records are predicted as Medium.
- Feature importance does not imply causation, and risk classes are model outputs, not definitive statements about a place.

VI. ANNEXURES

A. Annexure 1: Source Code

The implementation was written in a Google Colab notebook in seven steps: (1) importing libraries and loading the CSV file; (2) creating date/time features and building the Time_Period; (3) selecting features and splitting the data 80/20; (4) building the ColumnTransformer pipeline with imputation and one-hot encoding; (5) training the Logistic Regression, Decision Tree and Random Forest models; (6) training XGBoost with label encoding; and (7) making a prediction for a new case. The full code is given in the project report.

B. Annexure 2: Output Report

The output report consists of the risk distribution table, class-wise classification reports, confusion matrices for all four models, the Random Forest feature importance chart and the model comparison tables presented in Section V.

C. Annexure 3: Dashboard (GIS Risk Map)

The dashboard is an interactive Folium map on Esri World Street Map tiles, with each record plotted by latitude and longitude and coloured red (High), orange (Medium) or green (Low), as shown in Fig. 7.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

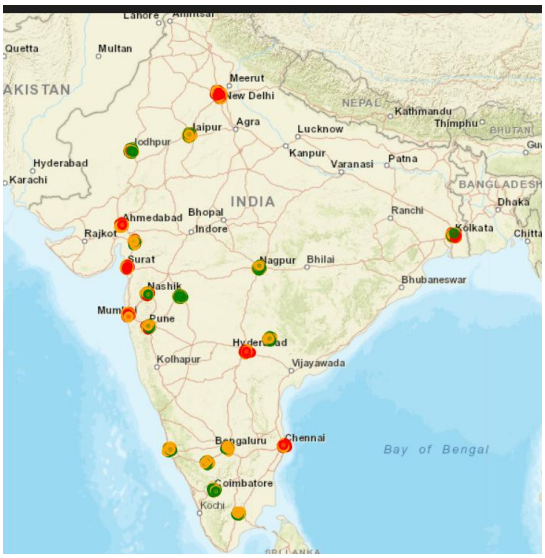


Fig. 7. GIS risk map of the crime records coloured by risk level

VII. Future Enhancements

- Validate the model on real, officially published crime data and on unseen cities and later time periods.
- Add feature scaling to the Logistic Regression model and optimise hyperparameters using cross-validation.
- Handle class imbalance to reduce under-estimation of High-risk cases using techniques such as class weights or resampling methods such as SMOTE [23].
- Include clustering and hotspot analysis, provide model-agnostic explanations using SHAP [24], and incorporate richer contextual features such as street lighting, nearby police stations and public transport.

VIII. Conclusion

This study showed that historical crime data, machine learning and GIS can be combined into a single, interpretable workflow for women's safety risk prediction. XGBoost achieved the best results, with an accuracy of 81.9% and a macro F1-score of 0.81. The Decision Tree and Random Forest models followed closely, each reaching an accuracy of 81.6%. The Logistic Regression model performed lower, with an accuracy of 75.7%. Population density, time of day and location were the strongest predictors, and evening and night records were far more often High risk. Because the data is synthetic and the labels are model-dependent, the framework should be treated as a research aid, to be validated on real data before any practical use.

REFERENCES

[1] Kounadi et al. 2020. A systematic review on spatial crime forecasting.

[2] Biswas and Das Chatterjee. 2025. Crime Prediction Using GIS and Statistical Modelling: Crime Against Women in West Bengal.

[3] Sahu and Kumar. 2024. Spatial Clustering and Machine Learning for Crime Prediction: A Case Study on Women Safety in Bhopal.

[4] Du and Ding. 2023. A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods.

[5] Dakalbab et al. 2022. Artificial intelligence & crime prediction: A systematic literature review.

[6] Naaz and Basthikodi. 2026. Comprehensive review on women safety technologies: Indian perspective.

[7] Shyam Sundar and Thangamuthu. 2025. Towards Intelligent Safety: A Systematic Review on Assault Detection and Technologies.

[8] The Role of IoT in Woman's Safety: A Systematic Literature Review. 2023. *IEEE Access*.

[9] Crime Prediction Against Women in Surveillance Videos Using Deep Learning Models: A Review. 2025.

[10] Sejal et al. 2025. Women Safety Platform with Safe Route Prediction Using Crime Data.

[11] Mandalapu et al. 2023. Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions.

[12] Smart Policing Technique With Crime Type and Risk Score Prediction Based on Machine Learning for Early Awareness of Risk Situation. *IEEE Access*.

[13] Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *IEEE Access*.

[14] A Hybrid Machine Learning and Regression Approach for Validating a Multi-Dimensional Crime Index in the Context of Crime Against Women. *IEEE Access*.

[15] scikit-learn developers. Confusion matrix documentation. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

[16] scikit-learn developers. Classification report documentation. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

[17] Chen, T., and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[18] Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[19] Quinlan, J. R. 1986. Induction of Decision Trees. *Machine Learning*, 1(1), 81–106. https://doi.org/10.1007/BF00116251

[20] Tolles, J., and Meurer, W. J. 2016. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, 316(5), 533–534. https://doi.org/10.1001/jama.2016.7653

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[22] Sokolova, M., and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[23] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

[24] Lundberg, S. M., and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765–4774. https://arxiv.org/abs/1705.07874

[25] Chainey, S., Tompson, L., and Uhlig, S. 2008. The utility of hotspot mapping for predicting patterns of crime. *Security Journal*, 21(1–2), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

[26] Cohen, L. E., and Felson, M. 1979. Change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), 588–608. https://doi.org/10.2307/2094589

[27] Dileep, A., Ramalakshmi, K., Venkatesan, R., Sundar, G. N., Nancy, G., and Shirly, S. 2024. A Comparative Analysis of Machine Learning Algorithms for Crime Rate Prediction. In *2024 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 732–736. https://doi.org/10.1109/ICAAIC60222.2024.10575332

[1] Mohler, G. O., Short, M. B., Malinowski, S., et al. 2015. Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710