International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

A Survey on Analysis and Prediction of Crime Hotspots using Stacked Generalization Approach

Aman Dwivedi
Student of B. E.
BNM Institute of Technology
Bengaluru
aman19is003@bnmit.in

Ayushi Lodha
Student of B. E.
BNM Institute of Technology
Bengaluru
ayushi19is008@bnmit.in

Mamatha Jadhav
Assistant Professor
BNM Institute of Technology
Bengaluru
mamsdalvi@gmail.com

Abstract: ***Ensemble learning is a collaborative decision-making process that aggregates the predictions of learnt classifiers to generate new examples. This paper suggests a practical, real-world approach called the assemble-stacking based crime prediction technique (SBCPM), which is based on algorithms for making accurate forecasts of crime. In comparison to past investigations, the proposed method showed a better predictive potential and yielded classification findings that were more accurate when applied to the testing data. The recommended method suggests that the ensemble model's prediction accuracy is higher than that of the individual classifiers, which is helpful for making prospective predictions about crimes.***

Keywords: crime prediction Decision tree, random forest, gradient boosting, forecasting

I. Introduction

This section includes a brief introduction about ensemble learning method that aggregates the predictions of trained classifiers in order to create new instances. According to preliminary investigation, ensemble classifiers are both conceptually and empirically more reliable than single part classifiers. Finding a suitable configuration for a certain dataset remains a challenge despite the fact that numerous ensemble methods are offered. There are several prediction-based theories that have dealt with the issue of machine learning for predicting crimes in India. It becomes challenging to determine the dynamic nature of crimes. In order to reduce crime rates and prevent criminal activity, crime prediction is used. The assemble-stacking based crime prediction method (SBCPM) is an effective authentic method that is proposed in this paper. According to what percentage of potential future violence in crimes, crime predictions are typically made using machine learning approaches. Many years have been spent on this research, but it has only used a small dataset and some limiting algorithms.

With the aid of empirical machine learning analysis and the additional contributions stated in this part, this research asserts its novelty. Even though machine learning models are frequently used in crime prediction, there are many areas where these new techniques developed in the field of artificial intelligence have not been thoroughly investigated and have significant drawbacks. The most popular methods that have claimed achieving accuracy in machine learning classifiers include the Random Tree Algorithm, K-Nearest Neighbor (KNN), Bayesian model, Support Vector Machine (SVM), and Neural Network.

Stacked generalization, commonly referred to as "stacking," is a machine learning technique that entails training many models and then combining their predictions to produce a final prediction. The models in stacked generalization are divided into two or more levels. A series of base models are trained on the training data in the first layer, and their predictions are then fed into a meta-model in the second layer of models, which makes the final prediction. Utilizing the predictions from the base models as extra features, or meta-features, for the meta-model is the idea behind stacked generalization. As a result, the meta-model is able to improve upon the advantages and disadvantages of the basic models and predict outcomes with greater accuracy.

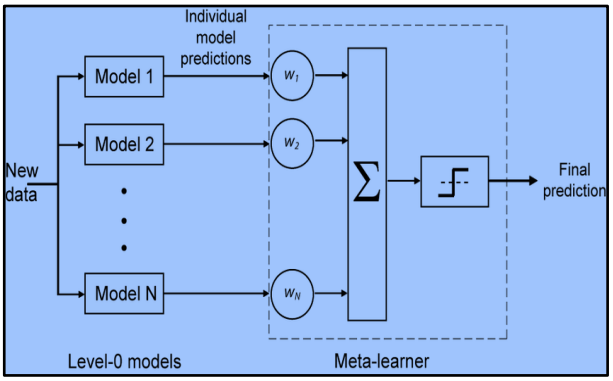


Figure 1: Stacked generalization

Figure1 shows a block schematic of stacked generalization. To establish how much each level-0 model should be trusted to make the final prediction based on the w_k weights of the level-1 model, stacking, integrating a straightforward meta-learner

method, was used. As a result of the research, the following contributions were made:

• A multi-level approach known as the crime stack is introduced and evaluated for crime prediction using a real-time database.
• Machine learning and ensemble learning-based techniques are combined to create crime stack.
• The accuracy is increased by employing the theory of piling to estimate violent crime, and stack generalization is employed to minimize error rate.

II. Literature Summary

A recognized and acknowledged element of contemporary society is the survey. It is one of the ways societies stays informed and a way to bring major issues of growing size and complexity into focus in order to gain perspective and a benchmark for comparison. A survey is a map rather than a precise plan that provides an overview of a field, differentiating it from a type of study that entails a microscopic investigation of a turf.

It is necessary to plan the survey before it is carried out. An essential component of the study is the literature review. It serves as a starting point for the development of project ideas into concepts and, eventually, theories. In order to analyze crime data, predict crime, identify criminals, and identify crime hotspot areas, researchers have proposed a number of data mining techniques. Here, a few of the papers are discussed.

According to XU ZHANG et al. [1], the model integrating built environment covariates has a stronger predictive effect than the initial model created solely using historical crime data. This paper takes historical data of public property crime from 2015 to 2018 from a large coastal city in the southeast of China as research data to assess the predictive power between several ML algorithms. The models are calibrated using historical crime data, and built environment variables, such as road network density and poi, are incorporated to the predictive model as covariates. The Hit Rate indicator is used to evaluate the prediction results of different machine learning models before and after adding covariates. S.

Mahmud, M. et al. [2] used different clustering approaches of data mining to analyze the crime rate of Bangladesh and use K-nearest neighbor (KNN) algorithm to train their dataset. Finally, they used the forecast rate to find out their safe route to the destination. The author reduced criminal activity and identified the crime zone by using the broken window theory, deep learning algorithms, random forest, and naive Bayes. The deep learning-tuned model offered 0.87% of the best accuracy. Regression and classification techniques offered by machine learning can be used to forecast crime rates. To better fit the crime data, WAJIHA SAFAT et al. used several machine learning algorithms, including logistic regression, support vector machine (SVM), Nave Bayes, k-nearest neighbors (KNN), decision tree, multilayer perceptron (MLP), random forest, and extreme Gradient Boosting (XGBoost), as well as time series analysis by long-short term memory (LSTM) and autoregressive integrated moving average (ARIMA) model. Using various machine learning algorithms on crime datasets from Chicago and Los Angeles, they enhanced the predicted accuracy for crimes. For Chicago, XGBoost performed the best, while for Los Angeles, KNN performed the best. The study also gives a visual summary using exploratory data analysis to depict crime kinds and counts, as well as anticipated crime rate and crime density areas for the following five years. An effective assemble-stacking based crime prediction method (SBCPM) based on SVM algorithms was proposed by Sapna Singh Kshatri et al. In order to predict crime data, Babakura et al. compared Nave Bayesian and Back Propagation, and Yadav et al. used several machine learning techniques. In order to analyze crime data, Sivaranjani et al. employed K-means clustering, hierarchical clustering, and DBSCAN clustering. Zhe Li et al. [3] gave forecasts of crimes in China.

To extract knowledge and find hidden links among the data, Hitesh Kumar Reddy et al. [4] employed raw datasets. To assist crime analysts in analyzing crime networks using interactive visualizations, a framework for crime prediction and monitoring based on spatial analysis is presented. Since safety measures are based on the type of crime, it is also vital to consider this. By using a variety of interactive visualizations, the initiative assists crime analysts in their analysis of these criminal networks. Algorithms for machine learning can help law enforcement agencies fight crime and protect people from harm.

To increase the effectiveness of the forecast, J. Vimala Devi and K. S. Kavitha [5] applied the adaptive DRQN model. The DRQN model uses reinforcement learning to find the ideal state value and update the reward function. Multi-agent-based Markov Decision Process (MDP) is used to update the state. In order to increase the effectiveness of learning, the reward function is added to the model. Data imbalance and overfitting are drawbacks of the current crime prediction models. In this study, a reliable reward-based crime prediction model is proposed.

The number of clusters and model of the clustering process were used by Simon Kojo Appiah et al. [6] to pinpoint crime hotspots. The total within-cluster variance of the K-means, BIC value, and LRT were used to get the best segmentation value. To assess the concentrations of these illicit activities,

two initialized model-based clustering approaches were investigated. The suggested E-M algorithm with semi-supervised K-clustering initialization proved effective in pinpointing crime hotspots, giving crucial data for programmes aimed at preventing crime.

III. Methodology

The stacked generalization approach, often referred to as stacked ensemble learning, is a well-liked machine learning strategy that mixes numerous models to increase the precision and reliability of predictions. Stacked generalization can be a helpful method for combining many data sources and creating more precise models for crime prediction when it comes to analyzing and predicting crime hotspots.

Here is a possible approach to employing layered generalization for crime hotspot prediction:

1) Data Collection and Preprocessing: Gather and preprocess data from a variety of sources, including crime incident reports, demographic information, geographic data, weather data, and social media data. This include preparing the data for machine learning algorithms, addressing missing values, and cleansing the data.

2) Model selection: Choose a group of fundamental models that are suitable for the objective of predicting crimes. You may, for instance, make use of gradient boosting, neural networks, decision trees, random forests, logistic regression, or decision trees. In terms of their underlying algorithms, hyperparameters, and feature inputs, pick models that are different.

3) Model Training: Use a subset of the available data to train each base model using the preprocessed data, and then assess its performance using a hold-out validation set. To choose the ideal hyperparameters for each model, use cross-validation.

4) Model Stacking: Using a meta-model, such as a logistic regression or a neural network, model stacking combines the predictions of the base models. Using the validation set as input, this meta-model is trained using the outputs of the underlying models. The meta-model learns how to balance each base model's predictions to arrive at the final forecast.

5) Performance Evaluation: Assess the stacked model's performance on a different test set that was not used during the model selection or training phases.

6) Deployment and Interpretation: Interpret the model's findings to learn more about the elements that influence crime hotspots. Use the model to predict crime hotspots in real-time, and keep an eye on how it performs over time to make sure it continues to be precise and pertinent.

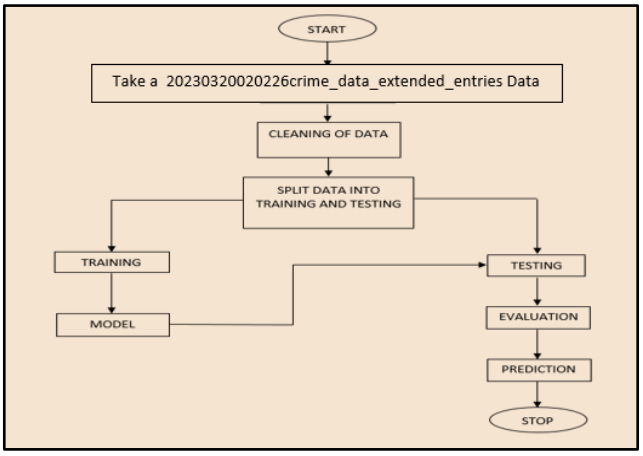


Figure 2 Block Diagram of Proposed System

There are numerous machine learning techniques for predicting crime hotspots. Decision Tree and gradient boosting are two examples of machine learning techniques. In our comparative analysis of machine learning algorithms for detecting crime hotspots, we employed the suggested and determined the optimal method for diagnosis. At this level, we have implemented each method separately and used the corresponding dataset before combining the results to calculate accuracy. Figure 2 shows the block diagram of the proposed system.

IV. System Design and Development

A. Architectural Diagram

The process of identifying a group of hardware and software components, as well as how they interact, in order to create the framework for a computer system's development is known as architectural design. This framework is created by looking at the software requirement document and constructing a model for supplying implementation details, which are used to specify the system components and their inputs, outputs, functions, and interaction with one another.

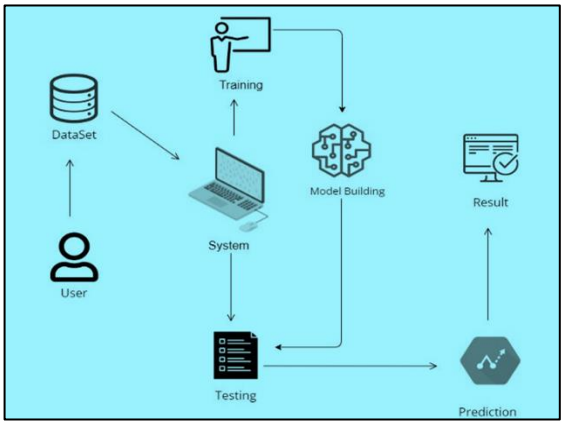


Figure 3 Architectural Diagram

As seen in Figure 3, a system's architecture, components, modules, interfaces, and data must be defined in order to accurately predict the places where crimes are most likely to occur. Based on the analysis of crime statistics from numerous sources, including police reports, social media, and CCTV footage. Pre-processing, which involves removing frames from CCTV footage and resizing them to a standard format, is done to the gathered data. This pre-processed data is then utilized to train and test models utilizing several methods, such as Random Forest, Decision Tree, and Gradient Boosting, to precisely forecast crime hotspots. By contrasting the anticipated hotspots with actual crime incidences, the system's accuracy is assessed.

B. Input/Output Design

The process of designing the input and output elements of an information system is known as input-output design. The input component is in charge of gathering data and information from diverse sources, while the output component makes the processed data and information meaningfully available to the users.

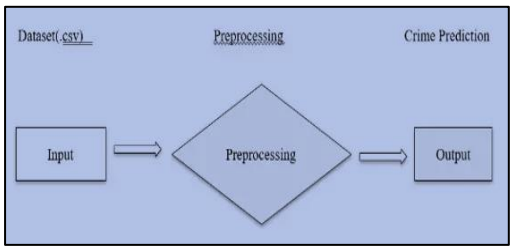


Figure 4. Input/Output Design

Fig 4.2 shows the input/output design of the proposed work.

Input: We import the crime dataset. The raw dataset is fed into our machine. A .csv file contains the dataset. After loading the data, it proceeds to pre-process

Pre-processing: Pre-processing is done on raw datasets. The dataset pre-processing is done to convert raw data into clean data.

Output: We apply Random Forest, Decision Tree, and Gradient Boosting algorithm. Finally, crime hotspots prediction is done.

C. Algorithms

This section provides a comprehensive explanation of the concepts of Random Forest, Decision Tree, and Gradient Boosting algorithms. Later, the proposed models using these algorithms are explained along with the crucial libraries used throughout the project.

1) Decision Tree

Decision-making processes are openly and visually represented using decision trees. The edges are divided into branches based on a condition or internal node, and they are depicted upside down with their root at the top. The technique used to determine whether a passenger lived or died is known as learning decision tree from data. Similar to decision trees, regression trees forecast continuous quantities such as house price. Choosing the right features, creating the right conditions for splitting, and understanding when to quit are all important aspects of growing a tree.

2) Random Forest

Machine learning methods for solving classification and regression issues include the random forest algorithm. It is learned using either bootstrap aggregation or bagging and consists of several decision trees. By using the average or mean of the results from different trees, it makes predictions. It eliminates decision tree algorithms' restrictions, lessens overfitting, and improves precision. It offers a practical method for dealing with missing data, is capable of making a valid prediction without hyper-parameter adjustment, and addresses the problem of overfitting in decision trees.

3) Gradient Boosting

A potent machine learning approach called gradient boosting is utilized to reduce bias and variance errors. It works with both continuous and categorical target variables and contains a fixed base estimator called Decision Stump. When employed as a classifier or regressor, the cost function is Mean Square Error (MSE) and Log loss.

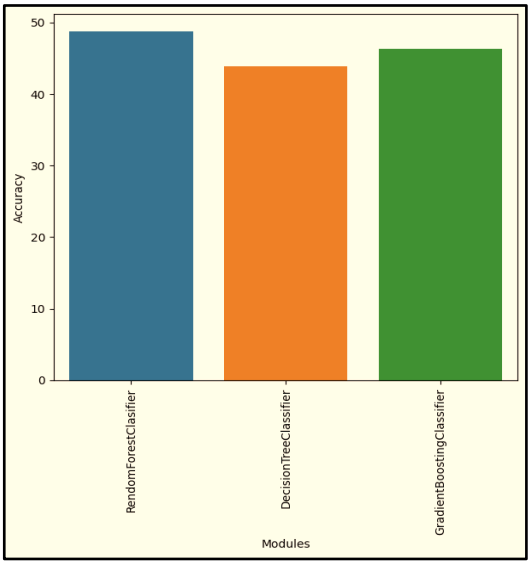


Figure 5 Accuracy of algorithms used

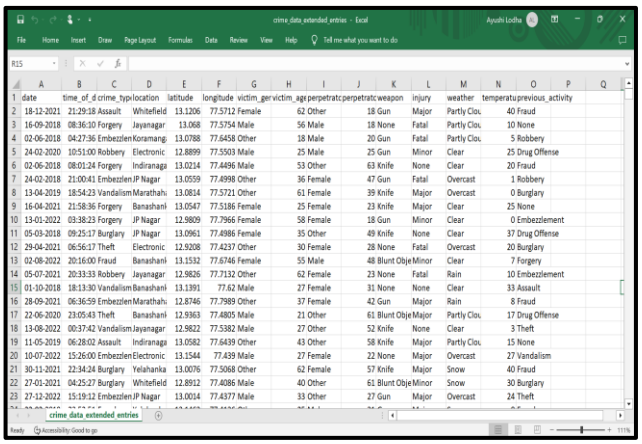
Figure 5 shows the accuracy of the algorithms obtained after applying stacked generalization approach on them.

V. Implementation

Every project's implementation phase is crucial since it entails carrying out the concepts and plans that were developed during the planning phase. The project team will put the necessary steps into action during implementation to turn the project into a usable good or service. This could involve creating, testing, and deploying the software or system, among other things. The implementation phase must be meticulously planned and managed if the project is to be completed on time, within budget, and to the required quality standards. Delays, cost overruns, and poor outcomes can occur from inefficient implementation phase management. To ensure successful execution, it is essential to have a transparent project strategy. The dataset contains the information on various factors related to criminal incidents:

1. Date: The date on which the crime occurred.
2. Time_of_day: The time of day when the crime occurred.
3. Crime_type: The type of crime that was committed.
4. Location: The location where the crime occurred.
5. Latitude: The latitude of the location where the crime occurred.
6. Longitude: The longitude of the location where the crime occurred.
7. Victim_gender: The gender of the victim.
8. Victim_age: The age of the victim.
9. Perpetrator_gender: The gender of the perpetrator.
10. Perpetrator_age: The age of the perpetrator.
11. Weapon: The type of weapon used, if any.
12. Injury: The nature and extent of the injuries sustained by the victim.
13. Weather: The weather conditions at the time of the crime.
14. Temperature: The temperature at the time of the crime.
15. Previous_activity: The previous activity of the victim or perpetrator prior to the crime.

This dataset can be used for a variety of purposes, such as analyzing crime patterns and trends, identifying risk factors for certain types of crimes, and developing crime prevention strategies. It's crucial to remember that the sources of the data and the methods used to gather them may have an impact on the data's correctness and dependability. A sample of the dataset utilized for the project is shown in Figure 6.



date	time_of_d	crime_typ	location	latitude	longitude	victim_ger	victim_age	perpetratc	perpetratc	weapon	injury	weather	temperatu	previous_activity
18-12-2021	21:29:18	Assault	Whitefield	13.1206	77.5712	Female	62	Other		18 Gun	Major	Partly Clou	40	Fraud
16-09-2018	08:36:10	Forgery	Jayanagar	13.068	77.5754	Male	56	Male		18 None	Fatal	Partly Clou	10	None
02-06-2018	04:27:36	Embezzlen	Koramang	13.0788	77.6458	Other	18	Male		20 Gun	Fatal	Partly Clou	5	Robbery
24-02-2020	10:51:00	Robbery	Electronic	12.8899	77.5503	Male	25	Male		25 Gun	Minor	Clear	25	Drug Offense
02-06-2018	08:01:24	Forgery	Indiranaga	13.0214	77.4496	Male	53	Other		63 Knife	None	Clear	20	Fraud
24-02-2018	21:00:41	Embezzlen	JP Nagar	13.0559	77.4998	Other	36	Female		47 Gun	Fatal	Overcast	1	Robbery
13-04-2019	18:54:23	Vandalism	Marathahi	13.0814	77.5721	Other	61	Female		39 Knife	Major	Overcast	0	Burglary
16-04-2021	21:58:36	Forgery	Banashank	13.0547	77.5186	Female	25	Female		23 Knife	Major	Clear	25	None
13-01-2022	03:38:23	Forgery	JP Nagar	12.9809	77.7966	Female	58	Female		18 Gun	Minor	Clear	0	Embezzlement
05-03-2018	09:25:17	Burglary	JP Nagar	13.0961	77.4986	Female	35	Other		49 Knife	None	Clear	37	Drug Offense
29-04-2021	06:56:17	Theft	Electronic	12.9208	77.4237	Other	30	Female		28 None	Fatal	Overcast	20	Burglary
02-08-2022	20:16:00	Fraud	Banashank	13.1532	77.6746	Female	55	Male		48 Blunt Obje	Minor	Clear	7	Forgery
05-07-2021	20:33:33	Robbery	Jayanagar	12.9826	77.7132	Other	62	Female		23 None	Fatal	Rain	10	Embezzlement
01-10-2018	18:13:30	Vandalism	Banashank	13.1391	77.62	Male	27	Female		31 None	None	Clear	33	Assault
28-09-2021	06:36:59	Embezzlen	Marathahi	12.8746	77.7989	Other	37	Female		42 Gun	Major	Rain	8	Fraud
22-06-2020	23:05:43	Theft	Banashank	12.9363	77.4805	Male	21	Other		61 Blunt Obje	Major	Partly Clou	17	Drug Offense
13-08-2022	00:37:42	Vandalism	Jayanagar	12.9822	77.5382	Male	27	Other		52 Knife	None	Clear	3	Theft
11-05-2019	06:28:02	Assault	Indiranaga	13.0582	77.6439	Other	43	Other		58 Knife	Major	Partly Clou	15	None
10-07-2022	15:26:00	Embezzlen	Electronic	13.1544	77.439	Male	27	Female		22 None	Major	Overcast	27	Vandalism
30-11-2021	22:34:24	Burglary	Yelahanka	13.0076	77.5068	Other	62	Female		57 Knife	Major	Snow	40	Fraud
27-01-2021	04:25:27	Burglary	Whitefield	12.8912	77.4086	Male	40	Other		61 Blunt Obje	Minor	Snow	30	Burglary
27-12-2022	15:19:12	Embezzlen	JP Nagar	13.0014	77.4377	Male	33	Other		27 Gun	Major	Overcast	24	Theft

Figure 6. Dataset used in prediction

VI. Results

This section summarizes the findings of the proposed project, interprets them, and places them in the appropriate context. The data obtained, any statistical analyses conducted, and any conclusions reached from the analysis are normally summarized in the outcomes section. Tables and figures should be utilized to support the findings, which should be presented in a clear and succinct manner. Following the results part is the discussion section, which gives the author a chance to explain the findings, come to some conclusions, and offer suggestions.

Crime hotspot prediction is the task of identifying areas where crimes are likely to occur in the future. It is an important application of machine learning in criminology and law enforcement, as it can help prevent crimes and allocate resources more effectively. This section consists of the work done during the project. Figure 7 shows the home page of the application.

Figure 7. Home Page

Figure 8. About Page

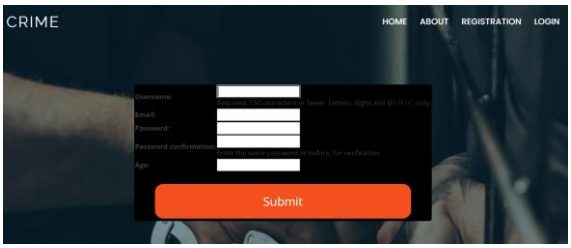


Figure 9. User Registration Page

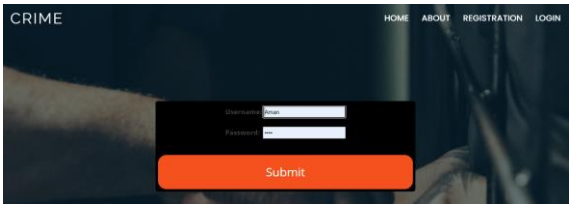


Figure 10. Login Page

Figure 8 shows the about page where a user can get information regarding the algorithms used and also about the dataset. Figure 9 shows the user registration page where user can create a new account and access the application by entering details like name, mail ID, age and password. Figure 10 shows the login page where a user can login to the application once he/she registers.

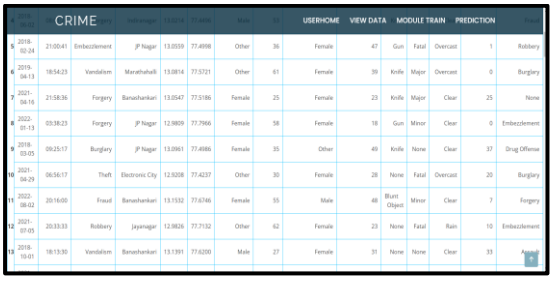


Figure 11. View Data Page

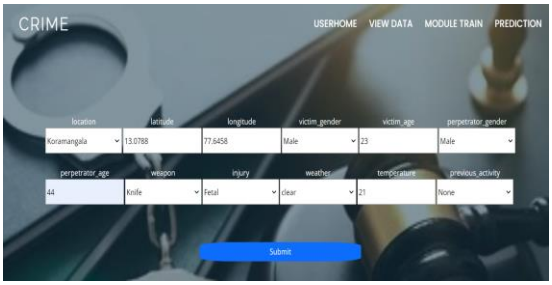


Figure 12. Prediction Page

Figure 11 shows the dataset which is being used in the project. Figure 12 shows the prediction page where user needs to fill the fields and submit so as to get the crime hotspots.

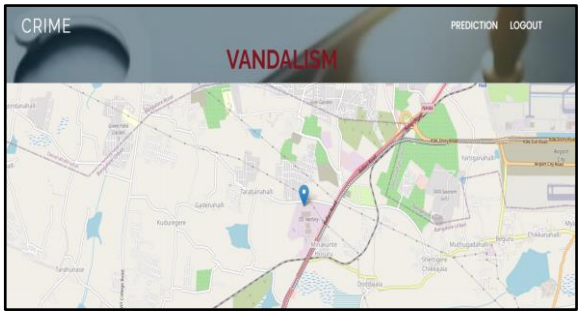


Figure 13. Output Page

Figure 13 shows the output page after the user inputs the data on prediction page, the output is in the form of pin which indicates the crime hotspot.

VII. Conclusion

Crime hotspots were examined and predicted using the layered generalization method using decision tree, random forest, and gradient boosting machine learning algorithms. In terms of accuracy, precision, and recall, the decision tree algorithm performed best. Location, latitude & longitude, weather, and prior crime incidences were discovered to be the factors most crucial in predicting crime hotspots. The outcomes of this project may be helpful in creating methods for crime prevention and more efficient resource allocation for law enforcement. The objective of the project is to create a user-friendly application that uses machine learning models to forecast crime hotspots, including Decision Tree, Random Forest, and Gradient Boosting.

VIII. Future Work

The potential for analyzing and predicting crime hotspots in the future using machine learning algorithms is encouraging. The following prospective research and development fields are listed:

1. Expansion to additional cities: To enhance efforts at crime prevention and law enforcement, crime hotspot analysis and prediction can be expanded to further cities or areas. To build a comprehensive model for forecasting crime hotspots, more datasets from other cities can be gathered and examined using machine learning techniques.
2. Incorporating more data sources: By including new data sources in the study, such as social media and weather data, the accuracy of crime hotspot prediction can be increased.
3. Hybrid algorithm development: To increase the accuracy and precision of crime hotspot prediction, hybrid algorithms that combine the strengths of various machine learning algorithms can be created.
4. Real-time prediction: Machine learning algorithms may be used to anticipate crime hotspots in real-time, allowing

law enforcement to react rapidly to prospective criminal situations.

5. Integration with other technologies: Combining machine learning algorithms with other technologies, such as CCTV cameras and drones, can improve the precision and efficiency of crime hotspot prediction.

In conclusion, there are many potential directions for the future development and use of machine learning algorithms for identifying crime hotspots, and more study and innovation in this field could lead to a considerable improvement in the general public's safety and security.

References

[1] Zhang, Xu, et al. "Comparison of machine learning algorithms for predicting crime hotspots." *IEEE Access* 8 (2020): 181302-181310.

[2] Mahmud, Sakib, Musfika Nuha, and Abdus Sattar. "Crime rate prediction using machine learning and data mining." *Soft Computing Techniques and Applications*. Springer, Singapore,2021.

[3] Safat, Wajiha, Sohail Asghar, and Saira Andleeb Gillani. "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques." *IEEE Access* 9 (2021): 70080-70094.

[4] Kshatri, Sapna Singh, et al. "An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach." *IEEE Access* 9 (2021): 67488-67500.

[5] Toppi Reddy, Hitesh Kumar Reddy, Bhavna Saini, and Ginika Mahajan. "Crime prediction & monitoring framework based on spatial analysis." *Procedia computer science* 132 (2018): 696-705.

[6] Yang, Bo, et al. "A spatio-temporal method for crime prediction using historical crime data and transitional zones identified from nightlight imagery." *International Journal of Geographical Information Science* 34.9 (2020): 1740-1764.

[7] Appiah, Simon Kojo, et al. "A model-based clustering of expectation–maximization and K-means algorithms in crime hotspot analysis." *Research in Mathematics* 9.1 (2022): 1-12.

[8] Vimala Devi, J., and K. S. Kavitha. "Adaptive deep Q learning network with reinforcement learning for crime prediction." *Evolutionary Intelligence* (2022): 1-12.