Crime Rate Prediction using Machine Learning

Priyanshu Ladha, Nitya Patyal

Chandigarh University, India

Corresponding author: Priyanshu Ladha, Email: priyanshuladha7@gmail.com

A common problem in the world is crime, and predicting crime rates is an important element in providing and predicting crime rates is an important effective crime prevention and resource management. This paper examines the use of machine learning in prediction of crime rates in order to prevent crime and allocate resources more efficiently. This study uses dataset of crime statistics and demographic information for specific regions and applies various machine learning algorithms such as K-Nearest Neighbor, Support Vector Machine and Decision tree to classify given region as high, medium, and low crime rate region. Each algorithm is evaluated based on metrics such as accuracy, precision and recall. This study provides insight of machine learning potential in predicting crime and suggests future research options in this field. Ultimately, these findings could have important implications for crime prevention and resource allocation. Therefore, helping policy makers and law enforcement to accurately, efficiently forecast and reduce crime rate. Crime rates can change over time due to changes in social, economic, or political factors, and machine learning algorithms can adapt to these changes and make more accurate predictions. However, there are also potential ethical issues associated with using machine learning to predict crime rates. In addition, privacy and traceability issues may arise when models use sensitive data such as personal information or criminal records. This is a risk of bias or discrimination if the data used to train the model is not representative of the general population. The research emphasizes the value of interdisciplinary cooperation between data scientists and law enforcement agencies and shows the potential of machine learning in crime prediction.

Keywords: Crime Prediction, Machine Learning, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Accuracy, Precision, Recall, Datapoints, Hyperplane.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 359–368. DOI: https://doi.org/10.56155/978-81-955020-7-3-31

1. Introduction

Throughout the history, crime has been an issue that has plagued many societies, and the possibility of accurately predicting and preventing ever growing crime rates has been the main center of concern for the people, governmental bodies in maintaining law and order. Based on types of crime occurring on a major part of the year they can be divided into classes of robbery, theft, abduction, assault, rape, murder, suicide. Besides solving the number of cases being lodged day by day, it is essential to prevent crime that may happen in the future.[11] For this it is essential to analysis crime and carry out a detailed inspection of ongoing crime. In this paper, analysis of vast amount of data is done and meaningful results using machine learning algorithms has been derived. Classification algorithms of Machine Learning like K- Nearest Neighbor, Support Vector Machine and Decision Tree are being used to classify high, medium, and low crime rate areas.[14] The foremost goal of this research is to build an accurate model for prediction of rate of crime in given region. The Data is preprocessed to filter out outliers and missing data. Features are extracted to produce results of high accuracy. Machine learning algorithms are applied and then evaluated based on metrics like recall, precision, accuracy. The dataset has been taken from the official site of Kaggle.com which is then evaluated using jupyter notebook of anaconda with the help of python as a core tool to analyze data and make predictions using machine learning algorithms. Machine learning algorithms are used for crime rate prediction as it can analyze large and complex datasets to accurately predict crime rates in each area, enabling law enforcement and policy makers to identify areas of high risk and allocate resources more effectively. [4] They can adapt to changes in crime rates over time, making them more robust and effective than traditional methods of crime rate prediction. They can process large volumes of data quickly and efficiently, making them ideal for analyzing complex crime statistics and demographic information. They can automatically identify which features are most important for predicting crime rates, reducing the risk of bias and error in the model. By accurately predicting crime rates, machine learning algorithms can also helps law enforcement and policy makers to allocate resources more effectively, reducing the incidence of crime and improving public safety.

2. Background Study

For this paper we have considered the relationship between wrongdoing and diverse features in criminology writing. Typically named as crime determining wrongdoing determining alludes to the fundamental process of anticipating wrongdoings in recent time as they happen. Instruments are required to foresee a wrongdoing and their evolution to future bigger wrongdoing or crime. Authors Neil Shah, Nandish Bhagat and Manan Shah [5] carried out a comparative ponder between rough wrongdoing designs from the Communities and Wrongdoing. Un normalized dataset versus real wrongdoing measurable information utilizing the open-source information mining computer program Waikato Environment for Information Examination (WEKA). [10] Three calculations, specifically, direct relapse, added substance relapse, and choice stump, were executed utilizing the same limited set of highlights on communities and genuine wrong doing datasets. Tests were arbitrarily selected. The straight relapse calculation might handle haphazardness to a certain degree within the test tests and hence demonstrated to be the leading among all three chosen calculations. The scope of the venture was to demonstrate the proficiency and precision of ML calculations in foreseeing dangerous wrong doing designs and other applications, such as deciding criminal hotspots, making criminal profiles, and learning trends.[12]Authors within a long time 2014 and 2013, individually, anticipated wrong doing utilizing tKNN's calculation. Sunet al.[5] demonstrated that a better wrongdoing forecast precision can be obtained by combining the dark relationship examination based on modern weighted KNN (GBWKNN) filling calculation with the KNN classification calculation. Utilizing the proposed calculation, we were able to get an precision of roughly 67%. By differentiating, Shojaee et al. [6] partitioned wrongdoing information into two parts, specifically, basic, and non-critical, and connected a basic KNN calculation. [9] They accomplished a surprising exactness of around 87%.For a long time 2013 and 2015 wrongdoing is anticipated employing a decision tree calculation, respectively. In their ponder, Obuandike et al.[5] used the Zero R calculation alongside a choice tree but fizzled to achieving precision of above 60%.In expansion, Iqbal et al accomplished a shocking precision of 84% employing

a choice tree algorithm. In both cases, be that as it may, a little alter within the information might lead to an expansive alter within the structure. A novel wrongdoing location strategy called naïve Bayes was executed for wrong doing expectation and investigation.[8] Jangraand Kalsi [5]accomplished a shocking wrongdoing forecast exactness of 87% but might not apply their approach to datasets with an expansive number of highlights. By contrast, Wibowo and Oesman [5] accomplished a precision of 66% in predicting violations and fizzled to consider the computational speed, robustness, and scalability.Shiju-Sathya devan[6]proposed Apriori calculation for the distinguishing proof of criminal patterns and designs. This calculation is additionally utilized to distinguish affiliation rules within the database that highlight common patterns. This paper too recommended the naïve Bayes calculation by preparing wrong doing information to form the show. The result appeared after testing that the Credulous Bayes calculation gave 90% accuracy. K. Zakir-Hussain et al. [6] utilized the strategies of information mining to analyze criminal conduct. This paper proposed apparatus for analyzing criminal examination (CIA). Within the law authorization community, this instrument was utilized to help resolve rough offenses. Both an investigative anda behavioral perspective examination was done. Author Faisal Tareque Shohan [2] proposed five directed machine learning calculations specifically Choice Tree, Additional Tree, Arbitrary Timberland, Adaboost, and Extraordinary Angle Boost (XGBoost). Choice Tree, Choice Tree, Additional Tree, Arbitrary Woodland, and AdaBoost usage from Scikit Learn library bundle. The information is divided into two bunches of train and test with rates 90% and 100% respectively. Training set utilized for learning reason whereas test set is utilized for assessment of metrics. Information about previous researches done and author content can be inferred from Table 1.

Table 1. Prior research results with different authors' use of machine learning techniques together with the corresponding years of study completed

Authors	ML Technique used	Year
Neil Shah ,Nandish Bhagat and Manan Shah	Linear Regression, additive regression and decision stump	2021
Su et al	Modern weighted KNN(GBWKNN)	2014
Obuandike et al	ZeroR calculation alongside a choice tree	2015
Jangra and Kalsi	Naïve Bayes	2019
Shiju-Sathya devan	Apriori calculation	2019
K Zakir-Hussain et al	Information mining	2019
Faisal Tareue Shohan	Choice Tree, Additional Tree, Arbitary Timberland,Adaboost, and Extraordinary Angle Boost(GBoost)	2022

3. Concepts of Proposed System

A. Machine Learning Model

This type of model predicts the output value for a given set of features/attribute/properties. Machine Learning under Supervised learning comprises of Classification and Regression models. Classification which focuses on distributing entities in given dataset in particular classes based on their properties and Regression works with real time values like changing temperature, etc in environment.[13]

B. Types of Machine Learning Algorithms used

(Given the detailed use of algorithms used)

a. Decision Tree: This algorithm allows the formation of tree like flowchart where each leaf node depicts class and internal nodes represents test on entity (attributes). The attribute with highest information gain is selected as Root node and subsequent attributes based on decreasing information gain at each level. Formula for Information Gain is as:

Information Gain(S, A) = H(S) – H(S | A) (1)
where H(S) is the entropy for the dataset before any change (described above)
and H(S | A) is the conditional entropy for the dataset given the variable A.

b. Support Vector Machine (SVM): This most favored algorithm for classification as well as regression is used to separate datapoints in N-dimensioned spaces where a hyperplane (boundary line) is constructed to distribute each datapoints in its underlying classes. SVM can be used for Linear as well as Non-Linear functions using kernel trick. For example: Taking below two classes of Square and Circle datapoints (support vectors) boundary is constructed to separate both classes. A graphical representation of a Support Vector Machine (SVM) can be derived from the data presented in the below Figure 1.

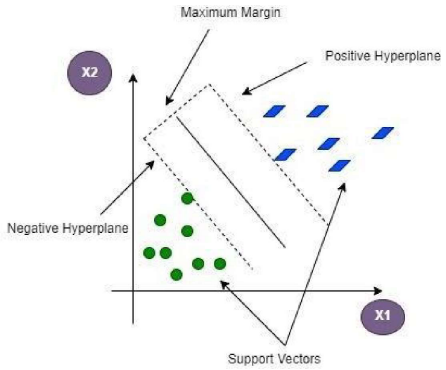

Figure 1. Sections of a Support Vector Machine

c. K- Nearest Neighbor (KNN): This algorithm follows the lazy learning approach which earlier just stores data at the time of training and then on new data predicts classes depending on similarity with the underlying data already present and do prediction based on that only. The below provided Figure 2 depicting an example of K- Nearest Neighbor.

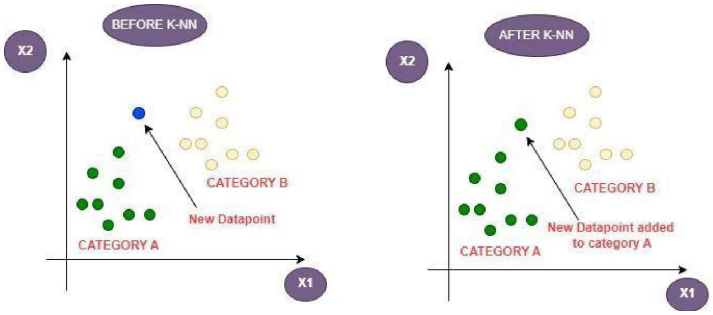

Figure 2. Illustration of KNN technique

C. Functional Diagram of Proposed Work

Below is the given block diagram of proposed system which showcases that how dataset is initially located through the means of secondary available data then preprocessed to train model and then test it on random data divided between training and testing data. The predicted outcome is tested on the grounds of accuracy, cross validation score, etc. meaningful feature extraction and new threshold is

further calculated to find best optimal model. The depiction of functional diagram of proposed work is given below in Figure 3.

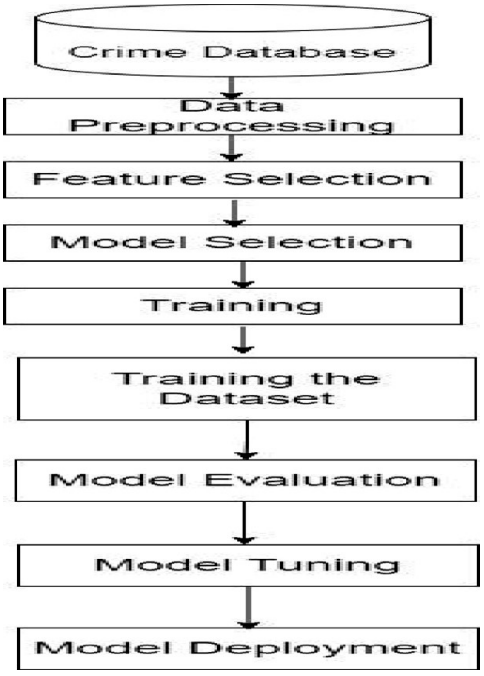

Figure 3. Flowchart for block diagram of proposed system

The methodology of this research can be broken down into the following steps:

a. Data Preprocessing: The dataset is preprocessed by cleaning the data, handling missing values, dropping irrelevant columns, and converting categorical variables into numerical variables using one-hot encoding.
b. Feature Selection: Feature selection is performed to identify the most important features that contribute the most in predicting crime rate. The Recursive Feature Elimination (RFE) method is used to select the top 15 features out of the original 122 features.
c. Model Building: Three different models are built and trained on the preprocessed data: Decision Tree, K-Nearest Neighbor and Support Vector Machine (SVM).
d. Model Evaluation: The performance of the models is evaluated using several metrics, for instance accuracy, score, precision, recall, etc.
e. Model Tuning: Hyper parameter tuning is performed for the Random Forest and SVR models to improve their performance. Grid search cross-validation is used to find the optimal hyper parameters for each model.
f. Model Deployment: The final model is deployed using streamlit, which is a library of Python especially used to deploy machine learning models. The deployed model can be accessed through a user interface, where users can input values for the selected features and get the corresponding results. [3]

4. Results

The results are evaluated based on analysis derived from the dataset as:

A. Crime Dataset

Figure 4. below is initial raw dataset accumulated from Kaggle.com contains a number of missing data and outliers which need to be processed before utilization.

	state	county	community	communityname	fold	population	householdsize	racepctblack	racePctWhite	racePctAsian	...
0	8	?	?	Lakewoodcity	1	0.19	0.33	0.02	0.90	0.12	...
1	53	?	?	Tukwilacity	1	0.00	0.16	0.12	0.74	0.45	...
2	24	?	?	Aberdeentown	1	0.00	0.42	0.49	0.56	0.17	...
3	34	5	81440	Willingborotownship	1	0.04	0.77	1.00	0.08	0.12	...
4	42	95	6096	Bethlehemtownship	1	0.01	0.55	0.02	0.95	0.09	...

Figure 4. Dataset before Cleaning

Figure 5. depicts the data after data pre-processing phases included dropping rows containing missing entries along with imputing outlier with mean to gain required results.

	state	communityname	fold	population	householdsize	racepctblack	racePctWhite	racePctAsian	racePctHisp	agePct12t21
0	1	Alabastercity	7	0.01	0.61	0.21	0.83	0.02	0.01	0.41
1	1	AlexanderCitycity	10	0.01	0.41	0.55	0.57	0.01	0.00	0.47
2	1	Annistoncity	3	0.03	0.34	0.86	0.30	0.04	0.01	0.41
3	1	Athenscity	8	0.01	0.38	0.35	0.71	0.04	0.01	0.39
4	1	Auburncity	1	0.04	0.37	0.32	0.70	0.21	0.02	1.00

5 rows × 104 columns

Figure 5. Dataset after Cleaning

B. Crime Visualization

The Visualization of data with graphs is an important aspect of data analysis and provides general head count or idea about various parameters greatly affecting the target variables which can also be seen in Figure 6.

a. Locations vs. violent crime observed per population:

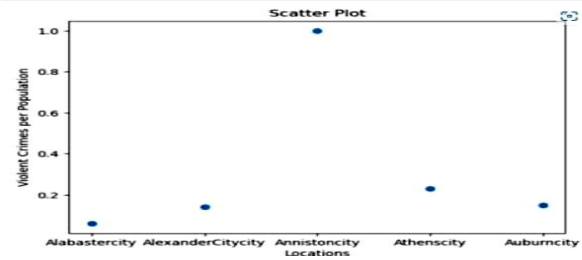

Figure 6. Scatter Plot for locations vs. violent crime per population

b. The below line plot in Figure 7. shows crime rate in specific region given below:

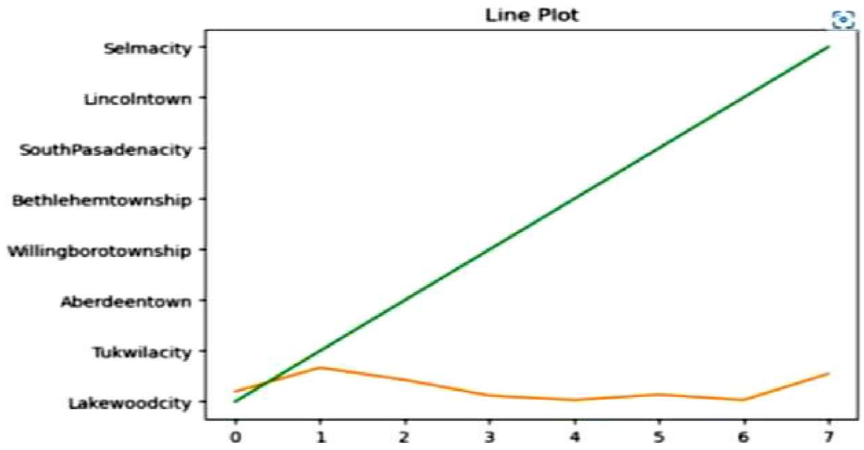

Figure 7. Line plot for crime Rate for specific region

C. Model Selection

The Model is then applied to dataset using feature of importance then evaluating the results based on accuracy, score, root mean square values, etc.

a. Use of Decision Tree: The decision tree evaluates results based on gini criteria based on which result is generated like in Figure 8.

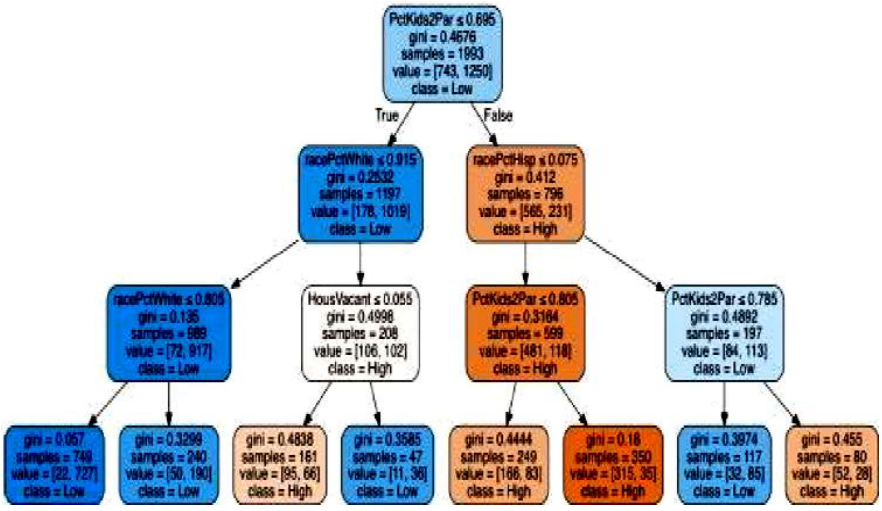

Figure 8. Decision Tree for predicting areas of high and low crime rate

b. Use of K Nearest Neighbor: Below graph in Figure 9 states accuracy analysis based on different values of K.

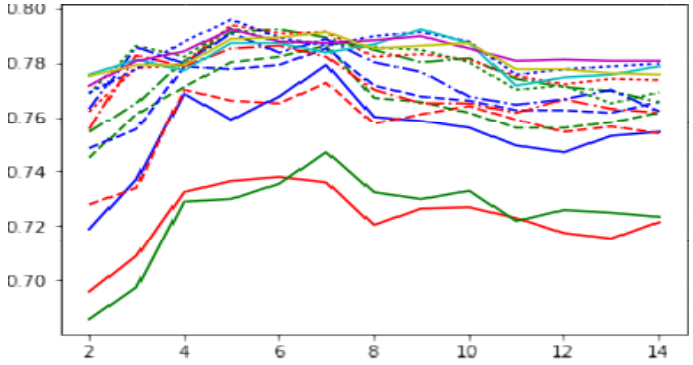

Figure 9. Plot for Accuracy vs. K value for KNN

The blue dotted line at X axis 5 and gives best accuracy at k =12.

c. Use of Linear SVM: In Figure 10. x-axis is the error penalty and y-axis is the accuracy

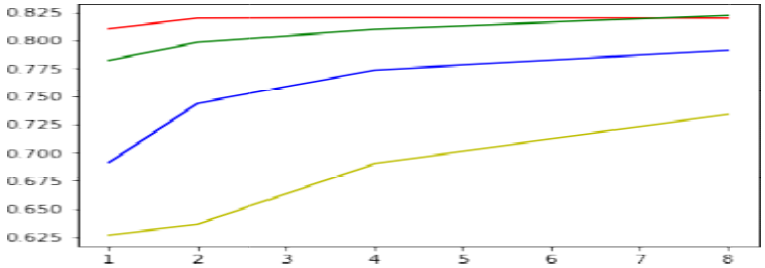

Figure 10. Plot for Linear SVM Accuracy vs Error penalty

Green gives best accuracy and for accuracy and best error penalty 8.

d. Use of Polynomial SVM: This graph in Figure 11. showcases the accuracy and error penalty in polynomial SVM.

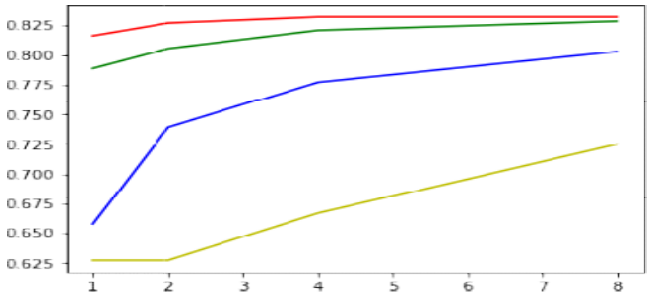

Figure 11. Plot for Polynomial SVM Accuracy vs Error penalty

Accuracy obtained after feature extraction and testing with provided Models:

Table 2. Table for representation of Algorithms applied, and their accuracy gained.

Algorithm	Accuracy
Decision Tree	80.596
KNN	84.27
SVM(Polynomial)	82.23
SVM(Linear)	84.67

From the above Table 2 it can be seen as Linear SVM and KNN gave the maximum accuracy in comparison to Decision Tree and Polynomial SVM. In comparison to Decision tree linear SVM improves accuracy this could be explain as Linear SVM is able to find linear separating optimal hyper plane for this dataset whereas the decision tree uses axis aligned planes to just split data in hierarchical fashion. On careful observation SVM has different predictive features than other models. [2]

5. Discussion

This research can be used a foundation to construct future embedded crime prediction system as an asset for providing security mechanisms in various applications like Google Map navigation systems where regions with higher crime rate can be marked unsafe and can help users in safe travel.[15] It can be used in Law enforcement systems, crime investigation departments, etc. detecting possible crime types, frauds, and reason for misfortune. It can also be used as a base for analyzing public mentality, perception and thought development over growing crime in different regions and its negative impact on building valuable human resources. Finally, crime prediction is an important and challenging task for law enforcement agencies. Machine learning techniques can be very useful in this regard, as the projects discussed here demonstrate. The use of multiple machine learning algorithms and feature selection techniques helps identify key factors that influence crime rates and create accurate predictive models. [4] The project shows that socioeconomic factors such as income, education and race can be strong predictors of crime rates. Also, we can know that certain types of crime are more likely to occur in certain geographic locations. Using predictive models can help law enforcement identify high- risk areas and deploy resources accordingly. However, it is important to note that machine learning models are not foolproof and may be subject to biases and inaccuracies. [7] It is important to continue to refine and improve these models while considering ethical and privacy implications. Overall, the project demonstrates the potential of machine learning in crime prediction and highlights the importance of interdisciplinary collaboration between data scientists and law enforcement agencies. [1]

References

[1] D. S. R. Alkesh Bharati, "Crime Prediction and Analysis Using Machine Learning," International Research Journal of Engineering and Technology (IRJET), 2018.

[2] A. U. A. M. I. a. M. S. A. Faisal Tareque, "Crime Prediction using Machine Learning with a Novel Crime Datatset," 2022.

[3] Yu, H., Liu, L., & Guo, H. , "Predicting crime hotspots using machine learning algorithms with feature selection", 2020

[4] Malathy, S., "Machine learning approaches for crime prediction: A case study in Indian cities.", 2017.

[5] [5] M. S. a. N. S. Nandish Bhagat, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention," Springer Nature

[6] A. Sattar, "Crime Rate Prediction Using Machine Learning and Data Mining," in Soft Computing Techniques and Applications , 2021.
[7] S. Walczak, "Predicting Crime and other uses of Neural Network in Police Decision Making", 2021.
[8] Jangra M, Kalsi S, "Crime analysis for multistate network using naïve Bayes classifier", 2019.
[9] Wibowo AH, Oesman TI, " The comparative analysis on the accuracy of k- NN, naïve Bayes, and decision tree algorithms in predicting crimes and criminal actions in Sleman regency", 2020.
[10] Obuandike GN, Isah A, Alhasan J, " Analytical study of some selected classification algorithms in WEKA using real crime data", 2015.
[11] K a rabo Jenga, Cagatay Catal2, Gorkem Kar1, "Machine learning in crime prediction ",2023
[12] B a m Bahadur Sinha, Tarun Biswas, "An Efficient Framework for Forecasting of Crime Trend Using Machine Learning Technique",2023
[13] S . S r i n i v asulu Raju, G. Narasimha Swamy, "Analysis and Prediction of Crime Using Machine Learning Techniques",2021
[14] Sun CC, Yao CL, Li X, Lee K," Detecting crime types using classification algorithms ", 2014
[15] Zhang, K., Wang, L., & Wang, Y. ,"Predicting crime hotspots using ensemble machine learning algorithms. Symmetry" 2019