Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. ICMASD (2025)

Evaluation of Various Machine Learning Algorithms for Crime Prediction

Mankaranjit Singh[1], Kamal Malik[1]

[1] Department of Computer Science & Engineering, CT University, Ludhiana

Article History:

Received: 12-12-2024

Revised: 25-01-2025

Accepted: 05-02-2025

Abstract: Crime poses a challenge to every nation's jurisdiction and administration. Thus, computerized crime forecasting and prediction may contribute toward making cities more secure. However, creating accurate and fast predictions about criminal activity is challenging. This is due to the incapability of humans that they can't process large amounts of data and information. Thus, in the present scenario, machine learning algorithms are utilized in crime prediction models to analyse big data and find crime patterns based on various factors. In this paper, we have evaluated the various machine learning algorithms, namely, KNN, NN, RF, and NB for crime prediction. In the proposed model, the same dataset is trained and tested for different machine learning algorithms and find out which algorithm is effectively predicting the crime. In addition, pre-processing of the dataset is done to remove inconsistencies in the dataset and select the appropriate features using the correlation matrix. Further, the Chicago dataset is used for evaluation purposes and the code is designed and simulated with the help of Python and google colab software. Finally, the various performance metrics are determined for the crime prediction model and find out that NN outperforms over other machine learning algorithms.

Keywords: Chicago, Crime, KNN, Machine Learning, Neural Network, Prediction, Random Forest.

Abbreviations:

ML: Machine Learning

NB: Naïve Bayes

KNN: K-nearest Neighbour

RF: Random Forest

NN: Neural Network

A: Accuracy

P: Precision

R: Recall

1. Introduction

Criminal activity has become a serious societal issue due to its negative impact on human lives, safety, and economy. In the past few years, crime data has become more accessible, which has allowed experts

to create models that can predict crime [1]. Based on past crimes, the government and other responsible officials can take steps to stop crimes before they happen. To keep society safe from crimes, it would be helpful to understand the reasons behind crime predictions so that suitable preventative measures may be planned. It is difficult to reach conclusions from the crime data since it is both large and unorganized. Because of this, machine learning (ML) methods have the potential to minimize effort in the current context by rapidly evaluating vast volumes of data to uncover criminal patterns. This is made possible by the rising movement toward technology and developments in artificial intelligence (AI) [2]. The next section provides an overview of machine learning and its many categories.

Machine learning (ML) is widely recognized as the newest and most popular technology because it enables systems to automatically learn from experience and improve without special programming. Figure 1 [3] illustrates how machine learning algorithms are primarily categorized into four types. These are semi-supervised learning, reinforcement learning, unsupervised learning, and supervised learning.

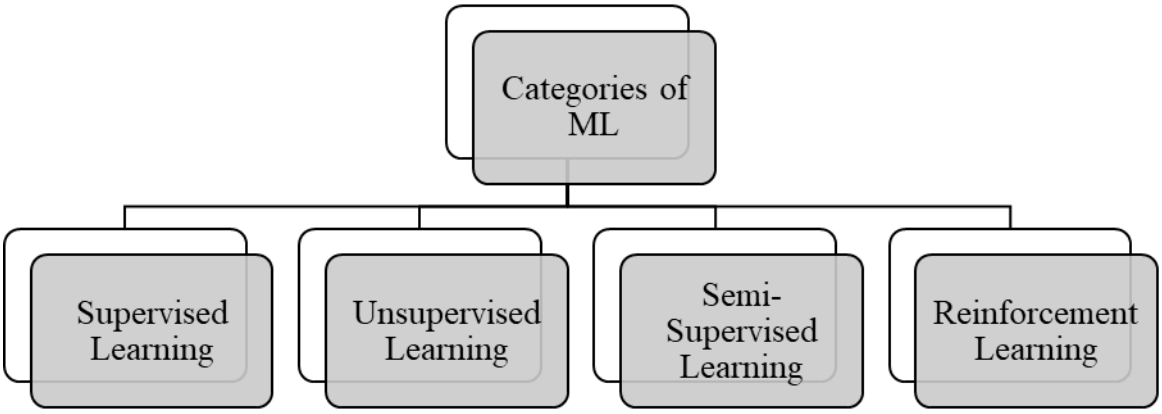

Figure 1: Categories of the ML

The following provides a brief summary of every type of learning strategy and how it might be used to address real-world issues [3].

- Supervised: Supervised learning is a machine learning technique that involves using example input-output pairs to develop a function that links an input to an output. The method generates a function from a set of labelled training examples and training data. This type of learning is called task-driven learning, and it happens when clear goals need to be met from a set of sources [105]. The two most popular supervised tasks are “regression,” which fits the data, and “classification,” which divides the data.
- Unsupervised: Unsupervised learning, also known as data-driven processing, analyses unlabelled information without the need for human supervision. This is often used for exploratory reasons, groupings in findings, generative feature extraction, and significant trend and structure identification. In unsupervised learning, the most common duties are finding association rules, estimating density, learning features, reducing the number of dimensions, finding outliers, and so on.
- Semi-supervised: Semi-supervised learning is a combination of the previous supervised and unsupervised approaches which operates on both labelled and unlabelled data [41, 105]. Therefore, it

is situated the middle of "supervised" and "unsupervised" learning. In real life, semi-supervised learning is helpful since labelled data may be limited in some situations while unlabelled data are common. A semi-supervised learning model's main goal is to make predictions that are more accurate than those made with just the labelled data from the model.

- Reinforcement: Reinforcement learning is a form of machine learning technique that allows software agents and computers to automatically analyse the best behaviour in a certain context or environment in order to enhance its efficiency. This method is known as an environment-driven approach. With the help of environmental activists, this kind of reinforcement learning aims to take action that will maximize reward and reduce risk.

The main motive of this research is to evaluate the various machine learning algorithms for crime prediction. In this research, NB, KNN, RF, and NN algorithms are taken into consideration. Besides that, the crime prediction dataset is inconsistent so pre-processing is done on it for removing the unwanted attributes, missing values, finding the appropriate features, and splitting in the training and testing dataset. The simulation evaluation of the proposed model is done for Chicago dataset and various performance metrics are determined. The result shows that the NN algorithm is outperformed over the other algorithms.

The paper is outlines into six sections. Section 1 gives a background information, followed by why machine learning algorithms are gained popularity in the crime prediction models. Section 2 shows the related work is done in the crime prediction models. Section 3 explains the proposed methodology in this section the relevant dataset, machine learning algorithms, and performance metrics. Section 4 explains the proposed crime prediction model. Section 5 shows the simulation results are performed for Chicago dataset using the various performance metrics. Finally, the paper is concluded and future aspects are defined to enhance the proposed model in Section 6.

2. Related Work

In this section, related work is shown to understand how the researchers utilized the various machine learning algorithms to design crime prediction model in the literature.

The authors, Safat et al., (2021), designed two models, the first model is employed for crime predicted whereas second model is employed for crime forecasting. In their research, a number of machine learning algorithms are utilized and evaluated using the various performance metrics. The simulation evaluation of the crime prediction model is done using various performance metrics such as accuracy, precision, recall, and F1-score. On the other side, the crime forecasting model is evaluated using RMSE and MAE parameter. The result shows the LR outperforms over other algorithms in the crime prediction model and forecasting model shows that the crime rate in the Los Angeles increases in the future over the Chicago.

3. Proposed Methodology

In order to understand the proposed crime prediction model, in this section, dataset, machine learning algorithms, and performance evaluation metrics are explained.

3.1 Dataset: In this study, the criminal record of Chicago city dataset is taken because of data availability and higher crime rate over the US. This dataset contains the crime information from the

year 2001 to 2019. In this research, 7002821 instance are taken into consideration out of 7019734 because some of the dataset instance is inappropriate.

3.2 Machine Learning Algorithms: In this section, a detailed description of the machine learning algorithms is given which are employed in the crime prediction model.

- **NB:** The Naive Bayes technique of conditional probabilities relies on the Bayesian theorem. It determines probability by the counting of frequently occurring values [1]. The following is an overview of Naive Bayes:

1. A basic classifier for a classification method
2. Most appropriate for past data and forecasting
3. Analyse the link between attributes and class instances using the classification approach.
4. A technique for supervised learning that can resolve probabilistic and categorical issues
5. A well-liked method of classification for text classification.

In 1995, Naive Bayes algorithm was first used. It is referred to by several names in the machine learning and data mining communities, including independence Bayes and simple bases. This classifier is widely utilized in many applications, including ensemble prediction and sentiment classification models. The Naive Bayes classifier must compute two sorts of values from the dataset. These are conditional probabilities and class probabilities. The following equation describes the Bayesian classifier's approach:

$$P\left(\frac{C}{X}\right) = P\left(\frac{X}{C}\right)\frac{P(C)}{P(X)}. \quad (1)$$

In this case, P(C-X) represents the maximum posterior hypothesis, P(X) is evidence, P(C) is the prior, and P(X-C) is the probability of the hypothesis.

- **KNN:** This is a well-known machine learning method that works well even with simple, old, or noisy training data. However, there is a drawback as well. For example, since it saves all states while computing distances, it requires a significant quantity of memory space. The KNN algorithm's stages are as follows [1]:

1. The first step is to identify the parameter k. The number of nearby neighbors to a particular place is represented by this parameter. For example, suppose k=2. In this instance, the two nearest neighbors will be used to determine classification.
2. Using distance functions, new data is added to the sample data set by calculating its distance from existing data.
3. The k closest neighbors of the associated distances are evaluated. Based on the attribute values, it is allocated to the class of k neighbors or neighbors.
4. The chosen class corresponds to the predicted observation value for estimation. As a result, the new information has labels.

- **RF:** The random forest method is a supervised classification system. It is applicable to both classification and regression issues [1]. The algorithm's goal is to enhance classification value by creating several decision trees throughout the classification phase. The random forest method selects the highest score from a group of independent decision trees. Our ability to generate exact results grows

as the number of trees increases. The primary distinction between the random forest algorithm and the decision tree method is that the process of locating the root node and dividing the nodes is random.

- **NN:** For classification and prediction purposes, neural networks (NNs) are an excellent machine learning approach [1]. Because neural networks learn, they can solve arbitrarily complicated problems. NNs may learn via two different mechanisms: unsupervised learning, which extracts patterns directly from the data, and supervised learning, which needs historical data with known outcomes. The NN needs to be taught to learn, and there are different training methods for each type of NN learning. For controlled learning, backpropagation is the most common, and for unsupervised learning, self-organizing maps (SOM) are the most common.

Every NN has two layers: an output layer that specifies the intended classification or prediction result, and an input layer that specifies the variables supplied to the NN for learning. One or more hidden layers are included in both supervised learning and hybrid models, and each layer has a weighted link that connects it entirely to the previous layer. This form of neural network learns by detecting the error of a training forecast from the actual value. And then sending out this error backward across the network to alter the weights of the links to better align the forecast with the true output value.

3.3 Performance Metrics: Next, Table 2 gives a detailed description of the performance metrics are evaluated for the proposed model.

Table 2 Performance Metrics

Parameter	Equation
A	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
P	$Precision = \frac{TP}{TP + FP}$
R	$Recall = \frac{TP}{TP + FN}$
F1-Score	$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$

4. Proposed Crime Prediction Model

In this research, crime prediction model is designed using the various machine learning algorithms to find out which algorithm outperforms over others. The flowchart of the proposed crime prediction model is shown in Figure 1.

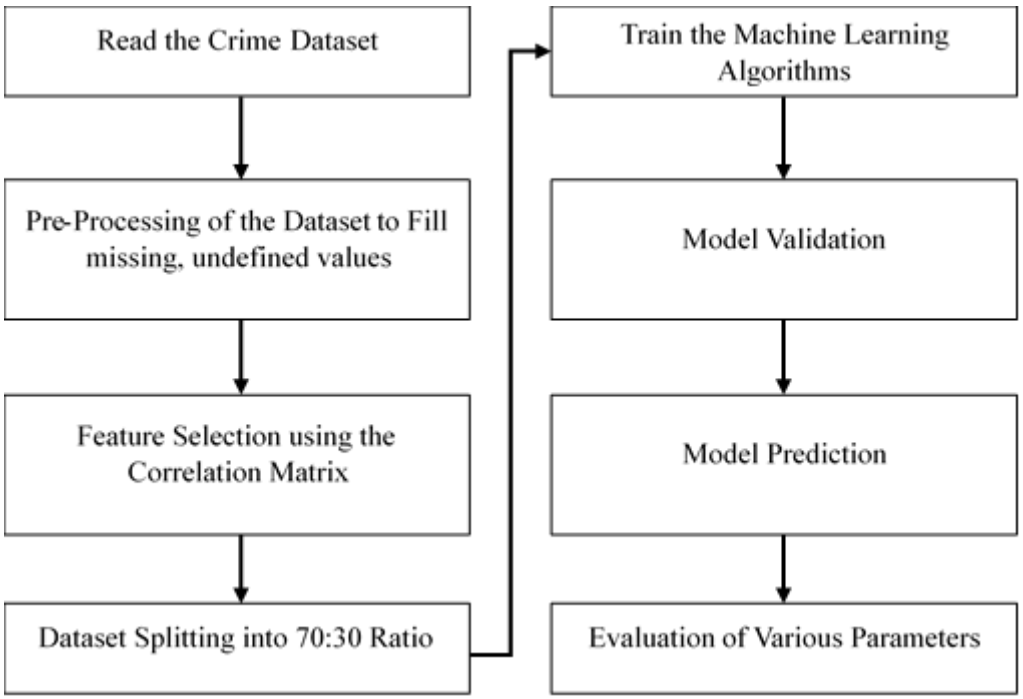

Figure 1 Flowchart of the Proposed Crime Prediction Model

Initially, in the proposed model, the standard dataset is read and pre-processing is done to fill the missing and undefined values of the attributes are presented in the dataset. Further, feature selection is performed to select the most appropriate data by decreasing the inputs for analysis and processing purposes. In other terms, it is referred as the procedure to select the attributes, subset of the dataset to construct the model. In this research, correlation matrix is used for feature selection. Next, the dataset is split into 70:30 ratio. The 70% dataset is used for train the model whereas 30% dataset is used for validate the model. Moreover, the machine learning model predict the crime based on the various attributes are chosen in the feature selection. Finally, the performance metrics are evaluated for evaluation purposes.

5. Simulation Results

In this section, the simulation evaluation of various machine learning algorithms for crime prediction is shown. The crime prediction model is designed and simulated in the google colab software. Further, Table 2 shows the simulation setup configuration is defined for machine learning algorithms during simulations.

Table 2 Simulation Setup Configuration for the Machine Learning Algorithms

Algorithms	Parameter Value
NB	Alpha:0.1
KNN	n-neighbour:5 Weight: Distance
RF	n-estimator:50

	Max Depth:15 Min Sample Leaf:1
NN	Hidden Layer Size: 50 Alpha: 0.1 Maximum Iteration: 1000

Finally, Table 3 shows the various performance metrics are determined for the various machine learning algorithms. The results show that the neural network algorithm achieves the highest accuracy, precision, recall, and F1-score over the NB, KNN, and RF algorithm.

Table 3 Performance Metrics for Different Machine Learning Algorithms

Parameter	**Naïve Bayes**	**KNN**	**RF**	**NN**	**Proposed Model**
Accuracy	0.91	0.92	0.88	0.94	0.94
Precision	0.91	0.92	0.89	0.94	0.94
Recall	0.91	0.92	0.88	0.94	0.94
F1-Score	0.90	0.92	0.87	0.94	0.94

6. Conclusion and Future Scope

In this paper, we have designed crime prediction model using the various machine learning algorithms such as NB, KNN, RF, and NN. Besides that, pre-processing of the dataset is done using the correlation matrix to find the appropriate features from it. Further, the dataset is split into 70:30 ratio. The 70% dataset is utilized for train the machine learning algorithm and 30% dataset is utilized for validate the model in the testing phase. The simulation evaluation shows that the neural network algorithm achieves the highest accuracy over the other ML algorithms. In the future, we will enhance the performance of the neural network by finding the optimal weight values of it using the metaheuristic algorithms. Further, we will validate the robustness of the proposed model by evaluating on the different datasets.

References

Introduction

[1] Y. Rayhan and T. Hashem, "AIST: An Interpretable Attention-Based Deep Learning Model for Crime Prediction," *ACM Transactions on Spatial Algorithms and Systems*, vol. 9, no. 2, pp. 1–31, Apr. 2023, doi: 10.1145/3582274.

[2] F. Dakalbab, M. A. Talib, O. A. Waraga, A. B. Nassif, S. Abbas, and Q. Nasir, "Artificial intelligence & crime prediction: A systematic literature review," *Social Sciences & Humanities Open*, vol. 6, no. 1, p. 100342, Jan. 2022, doi: 10.1016/j.ssaho.2022.100342.

ML

[3] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science*, vol. 2, no. 3, Mar. 2021, doi: 10.1007/s42979-021-00592-x.

KNN and RF

[4] A. Sayli and S. Başarir, “S Sampling Techniques and Application in Machine Learning in order to Analyse Crime Dataset,” *European Journal of Science and Technology*, no. 38, pp. 296-310, Jun. 2022, doi: 10.31590/ejosat.1115323.

NB

[5] M. Khan, A. Ali, and Y. Alharbi, “Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques,” *Complexity*, vol. 2022, pp. 1–13, Feb. 2022, doi: 10.1155/2022/4830411.

NN

[6]S. Walczak, “Predicting Crime and Other Uses of Neural Networks in Police Decision Making,” *Frontiers in Psychology*, vol. 12, Oct. 2021, doi: 10.3389/fpsyg.2021.587943.

Crime Dataset

[7] Crimes - 2001 to present - Dashboard. (n.d.). City of Chicago | Data Portal. https://data.cityofchicago.org/stories/s/Crimes-2001-to-present-Dashboard/5cd6-ry5g