Crime Rate Prediction Using Machine Learning and Data Mining

Chapter · January 2021

DOI: 10.1007/978-981-15-7394-1_5

Abdus Sattar
Daffodil International University

Crime Rate Prediction Using Machine Learning and Data Mining

Sakib Mahmud, Musfika Nuha, and Abdus Sattar

Abstract Analysis of crime is a methodological approach to the identification and assessment of criminal patterns and trends. In a number of respects cost our community profoundly. We have to go many places regularly for our daily purposes, and many times in our everyday lives we face numerous safety problems such as hijack, kidnapping, and harassment. In general, we see that when we need to go anywhere at first, we are searching for Google Maps; Google Maps show one, two, or more ways to get to the destination, but we always choose the shortcut route, but we do not understand the path situation correctly. Is it really secure or not that's why we face many unpleasant circumstances; in this job, we use different clustering approaches of data mining to analyze the crime rate of Bangladesh and we also use K-nearest neighbor (KNN) algorithm to train our dataset. For our job, we are using main and secondary data. By analyzing the data, we find out for many places the prediction rate of different crimes and use the algorithm to determine the prediction rate of the path. Finally, to find out our safe route, we use the forecast rate. This job will assist individuals to become aware of the crime area and discover their secure way to the destination.

Keywords Crime · Numerous safety problem · Data mining · KNN (K-Nearest Neighbor) · Safe route

S. Mahmud · M. Nuha · A. Sattar (✉)
Department of Computer Science and Engineering, Daffodil International University, 102 Sukrabadh, Mirpur Road, Dhaka 1207, Bangladesh
e-mail: abdus.cse@diu.edu.bd

S. Mahmud
e-mail: sakib15-7962@diu.edu.bd

M. Nuha
e-mail: musfika15-7930@diu.edu.bd

S. Borah et al. (eds.), *Soft Computing Techniques and Applications*,
Advances in Intelligent Systems and Computing 1248,
https://doi.org/10.1007/978-981-15-7394-1_5

1 Introduction

In this era of modern world, our popularity is increasing and citification carries enormous general, financial and environmental, while presenting challenges in urban management issues such as traffic resource planning, environment and safe water quality, public policy and public safety services. In addition, represent the most crime rates in larger cities, crime reducing is becoming one of the most important social issues in enormous metropolitan areas as it affects people security issues, youngster growth and person socio-economic status. Crime rate forecast is a scheme that uses different algorithms to determine the crime rate based on prior information. For our daily purposes we have to go many places every day and many times in our daily lives we face numerous security issues such as hijacking, kidnapping, harassment, etc. In general, we see that we are searching for Google Maps when we need to go anywhere at 1st, Google Maps show that one, two or more ways to get to the destination, but we always choose the shortcut route, but we do not comprehend the path condition properly. Is it really safe or not that's why we are faced with many unpleasant circumstances; this research introduces the design and execution of a strategy based on past crime data and analyzes the crime rate in past areas at distinct moments; for this work, we use primary data those are collected from the people based on their previous crime problem. In our train information collection, we used different algorithms to figure out the highest precision between the KNN algorithms that provides the greatest precision. In this paper, we use different models and table to show the different types of crime rate, mostly working data from last 3 years of crime and showing the level of crime prediction in different issues described in Sect. 3. In Sect. 1 gives the introduction about the whole paper, Sect. 2 in literature review section describes the previous work on crime rates, Sect. 3 describes the general paper information in the Methodology section, this chapter discusses the dataset, data processing, crime analysis, crime rate prediction and describes the multiple algorithms that we used in the crime database assessment and lastly demonstrates the greatest precision of the forecast rate and model using python matplotlib and KNN algorithm. In Sect. 5 proves the result and discussion, Sect. 6 dicussed the conclusion about the total work, and finally acknowledgements and references are discussed at the end of this chapter.

2 Literature Review

For this paper we have studied the relationship between crime and different features in the criminology literature. Reduce the crime and detect the techniques of crime and stop the crime before the author uses different techniques. Use Z-Crime Tools and Advanced ID3 algorithm with data mining technology to predict criminal activity.

Identify the appropriate crime pattern and statistical analysis of hidden linked detection algorithm used data. Forensic Tool Kit 4.0 is used to remove research and visualization of data [1]. The uses the K-Means Clustering algorithm for unsupervised learning to determine the crime rate. The model was then analyzed, preprocessed and implemented to taste the set of information and trained the algorithm. K-Means Clustering algorithm provided more than 75% [2].

The author used broken window theory, deep learning algorithm, random forest and naïve Bayes to reduce criminal activity and detect the crime zone. Prepare the data frame to train the model for recognition of images, preprocessing of information and detection of crime hotspot. The model tuned with deep learning provides 0.87% of the best accuracy. Machine learning offers methods of regression and classification used to predict rates of crimes. The author uses multi-linear regression to find the link between dependent and independent variables.

K-Nearest neighbors is used for classification to single and multi-class variable. The Neural Network is used for the prediction's precision. The precision rate by using Neural Network the model accuracy is 60, 93 and 97% [2]. The Author presents a geographical analysis-based and self-regressive approach to automatically identify large danger urban crime areas and to represents crime patterns in each region reliably. Show the result of the algorithm of crime prediction system consisting of a collection of thick crime areas and a set of related crime forecaster. This operate primarily for the large region where the large amounts of individuals live and demonstrates that the suggested strategy achieves excellent precision over rolling time horizons in spatial and temporal crime forecasting. This paper's working process collects raw data the hotspot uses after splitting the data to create the new hotspot model and finally shows the predictive crime rate [3].

Shiju-Sathya devan proposed Apriori algorithm for the identification of criminal trends and patterns. This algorithm is also used to identify association rules in the database that highlight general trends. This paper also suggested the naïve Bayes algorithm by training crime data to create the model. The result showed after testing that the Naive Bayes algorithm gave 90% precision. K. Zakir-Hussain et al. used the methods of information mining to analyze criminal conduct. This paper proposed tool for analyzing criminal investigation (CIA). Within the law enforcement community, this instrument was used to assist resolve violent offenses. This study is about the various type of crime scene. Both from an investigative and a behavioral perspective, the analysis was done. It provided insight into the unknown criminals as well as recommendation for investigation and interview and trial strategies [4].

Classification is one kind unique methods of information mining used to classify each object in an information set into one of the predefined classes or groups, The idea is to define the use of the Criteria for the segmentation of the entire database, once this is done, individual datasets can then fall naturally into one or more groups. By means of classification, existing datasets can be easily understood and it also helps to predict how new individual datasets will behave based on the classification. Data mining generates models of classification by observing classified data and discovering a predictive pattern between those data. Naive Bayes is a classification algorithm used to predict that it works on the principle of Bayesian.

Table 1 Dataset table

Number	Details		
	Name	Type of columns	Descriptions
1	Person_Id	Value type	Person ID in dataset
2	Name	String	Victim person name
3	Year	Numeric	Crime occurs year
4	Ages	Numeric	The ages of the victim
5	Gender	String	Victims neuter
6	Time	String	Time when the crime has occurred
7	Victim area	String	Area where the crime has occurred.
8	Region	String	Region of the victim
9	Home town	String	Home town of the victim
10	Month	String	The year in which the crime has occurred

3 Research Methodology

3.1 Dataset

The crime dataset is extracted from primary data collection based on field work. This dataset consists of about 500 in 10 rows details. The key features such as Name, Years, Months, Crime Type, Crime Areas, Victim Genders, Victim Ages, Victim Areas, and Months are selected from the dataset as the system input features.

The characteristics Perpetrator Ages, Perpetrator Genders and Victims relation are select system's target variables (Table 1).

3.2 Preprocessing

Since unknown is not a value to be considered but an indicator of unfinished papers, we decided to remove unknown values from the information set, Dates and times was recorded as a times window in MM/DD/YY HH/MM format from the documents as direct matching of these dates are complicated by the classification system to match date and time with each other. Classification of the date was considered to classify a particular date into three groups as weekends, weekdays and unaware. This classification is based on the features of the date time windows (Fig. 1).

Explains the system's workflow. The workflow begins by extracting data from data collection, which is a dataset repository on different roles. The primary data will be preprocessed and transformed into a criminal data.

Four target variables are predicted.

i. Linear Regression is used to find the aged based crime accuracy.

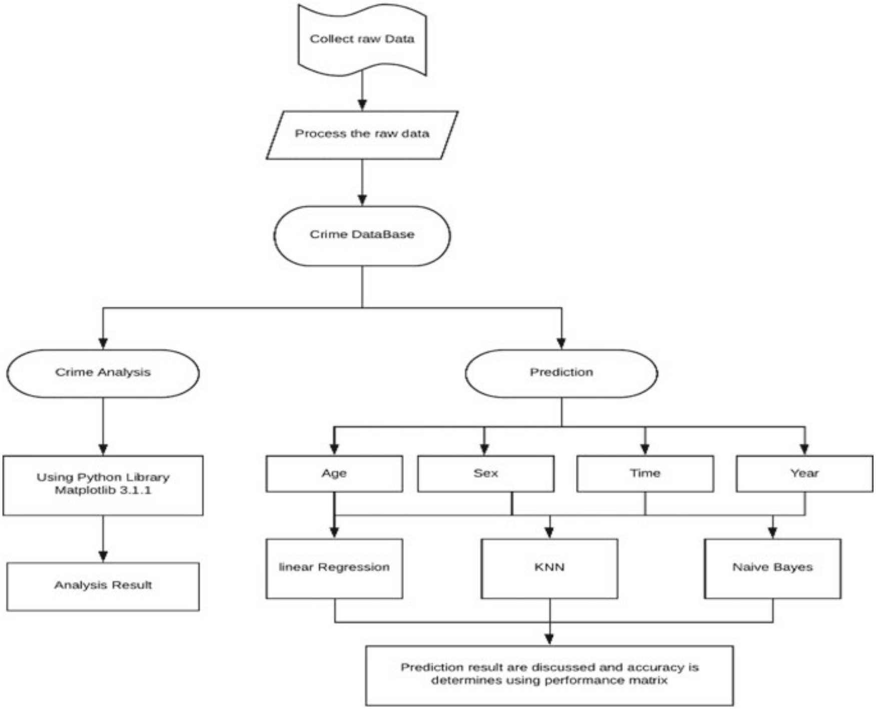


Fig. 1 Work flow diagram

ii. The sex is estimated using the classification of K- Nearest Neighbors.
iii. The gender of the perpetrator is estimated using the classification of K-nearest neighbors.
iv. The Final prediction rate for years based on age, sex, time, and year using K-NN algorithm.

3.3 Crime Dataset

See Table 2.

Few property features of dataset like months, types of crime, victim gender, age, area and the relationship are in the above data set are qualitative from. This qualitative information should be classified as in order 0 or 1 to apply the mathematical models for prediction.

Table 2 Male female identification by binary digit

Gender	Male	Female
	0	1

Table 3 Uses binary value to present the month status

Months	D_A	D_B
Jan–April	1	0
May–August	0	1
Sept–Dec	0	0

As follow the rules male contains 0 and female contains 1 in Table 1. Also adding dummy column by using the crime data in different month of the year in Table 3.

Table 3 adds $N - 1$ dummy columns in the given column for unique N values. The table above explains how data is preprocessed. Three months are selected from the Column state in the data table. In this case, in the column states, the number of unique values is $N = 3$. It is necessary to add dummy columns to categorize the data N s 1. Here are two types of dummy columns those are dummy A and dummy B.

1. 1 is assigned for column A and 0 for column B result between January to April.
2. If 0 is assigned for column A and 1 for column B result between May to August.
3. If both is assigned 0 for column A and column B result between September to December.

4 Algorithm

The domain contains many clustering algorithms. There is widespread use and acceptance of the K-means partitioning method [5]. Apart from the K-means strategy, the Linear regression algorithm [6] is the one we used because it enables consumers to determine the number of clusters based on those values Naïve Bayes is also pretend good result but above two are provides the best accuracy.

4.1 Linear Regression

Multi-linear regression is a sort of mathematical approach to finding a relation between the dependent variables (Victim age) and a set of independent variables those input values gathered from the crime spot. This methodology predicts the Era of the victims age values based on the input characteristics indicate in the metadata column. The multi-linear regression is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{1}$$

Here,

Y performs as the reliant variable
X performs the independent variable
β represent the coefficient formula function of regression.

Linear regression is used in the crime prediction situation to represent the most likely perpetrator age given the crime scenes.

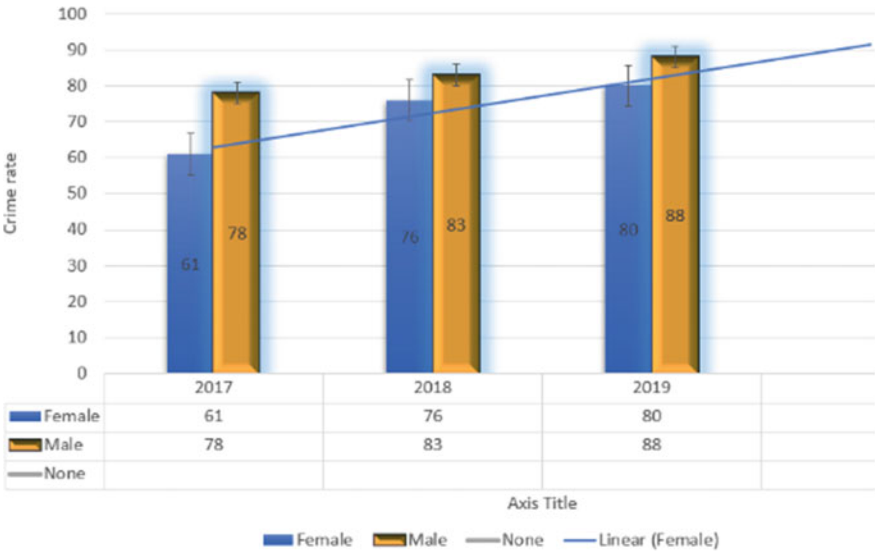
In this figure, the victim rate between male and female is shown, and the past information analyzes show that female victims are quickly increasing between males every day (Fig. 2).

To find out the monthly crime rate we used KNN algorithm at the start we find out each month crime rate for every year. This figure shows the individual month crime rate for three years. We also find out the crime rate which time is more crime occurs in the year, finding that rate divides the years into three dimensions for type 1 from January to April, type 2 may occur to August and type 3 from September to December, and showing that type 1 and type 3 occurred the most crime in those months.

By using the K-NN and Linear regression we find out the aged crime rate for 3 years. To find that rate we divide the age in three types those are Teenager, young and old (Fig. 3).

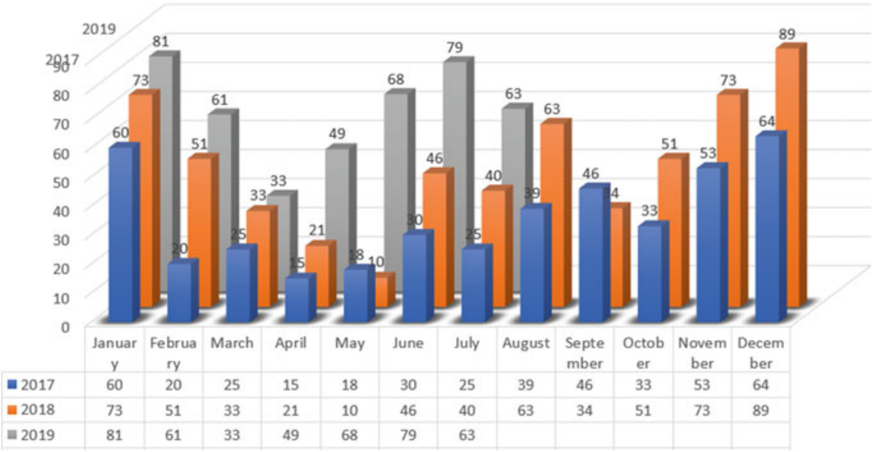
By using the K-NN and Linear regression we find out the aged crime rate for 3 years. To find that rate we divide the age in three types those are Teenager, young and old (Table 4).

In Dhaka city crimes rate we divide the city in two zone north and south and zone then plot the data by using KNN algorithm *X* axis for north and *Y* axis for south, finally show the result in Fig. 4 that is north cities people are more victim then south city in Dhaka (Fig. 5).



	2017	2018	2019	
Female	61	76	80	
Male	78	83	88	
None				

Fig. 2 Male versus female rate



	January	February	March	April	May	June	July	August	September	October	November	December
■ 2017	60	20	25	15	18	30	25	39	46	33	53	64
■ 2018	73	51	33	21	10	46	40	63	34	51	73	89
■ 2019	81	61	33	49	68	79	63					

Fig. 3 Male versus female rate

Table 4 Age with range

Age	Range
Teenager	13–19
Young	20–55
Old	56–100
Old	56–100

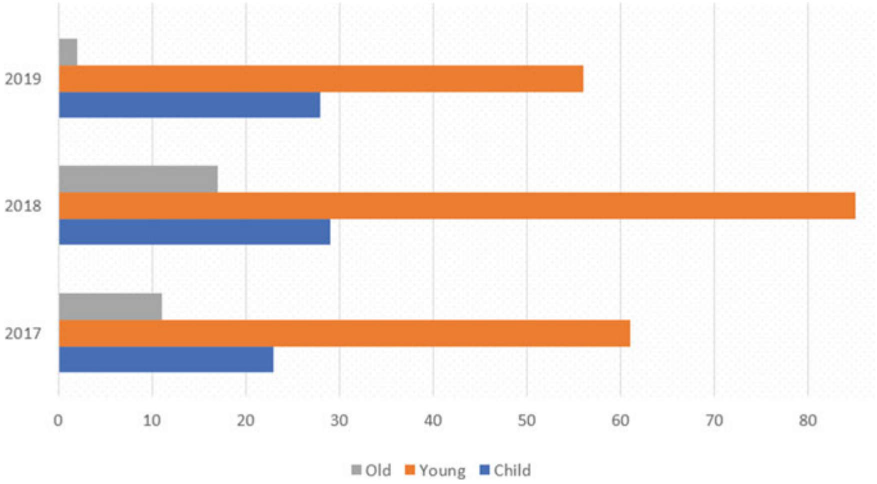


Fig. 4 Age based crime rate

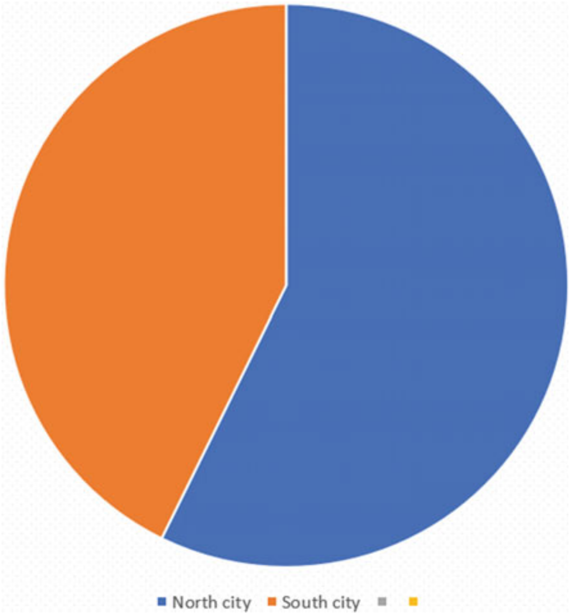


Fig. 5 dhaka city crime rate

4.2 K-Nearest Neighbors

K-nearest neighbors is used when the target variable must be classified in more than two classes. In this dataset, there are three classes of target variable perpetrator sex: male, female, and unknown. Similarly, three categories of young, old and kid are defined in age. To classify these target variables, K-nearest neighbors Classifier is used.

$$D(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{2}$$

Pseudo Code:

```
At first KNN Classifier (Data Entry)
Assign the Number of Cluster in $K$
Choose a set of $K$ instances to be cluster centers
Data points for each output
Calculate the Euclidean range
Assign next to the data point the cluster
Perpetually Calculate centroids and reassign the variables of the cluster
Repeat until you reach an appropriate cluster Result, Give back the clusters and their values.
```

4.3 Naïve Bayes

Naive Bayes classifiers are a collection of Bayes' theorem-based classification algorithms. It is not a single algorithm, but an algorithm family in which all of them share a common principle, each combination of features being classified is separate from each other. Bayes' theorem considers the probability of a case occurring considering the probability of another case occurring already. The theorem of Bayes is indicated as the following equation:

$$p(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

Naive Bayes algorithms are mostly used to determine emotions, delete spam, suggestions, etc. We are fast and easy to implement, but their biggest downside is the need for autonomous predictors. In most real-life situations, the predictors are dependent, hampering the output of the classifier. We do not use that algorithm to find the final result in our work for this problem, but in many previous cases its works are very good and given the best crime rate accuracy.

5 Result and Discussion

In this part summarizes the paper and make aware about the future crime based on algorithms and crime data set we find out the crime rate in various section like age based, male vs female, area based and monthly crime rates. The data sources and methods used to guide forecasting include various type crime statistics, survey of the general people data, literature reviews and statistical models that extrapolate crime trends into the future. Algorithms models that describe the behavior of observe past values can be used to forecast future crime trends by projecting a time series analysis of crime trends into the future. Any predictive model endeavor to show a relationship between certain predictor and a dependent variable. To ensure the greater accuracy those models must identify and predict the scope and nature of a number of factors that will influence crime and victimization in the future. This research paper about future crime rate predictions are much more specific and precise. The accuracy table of different algorithm accuracy (Table 5):

By referring to these 3 algorithms, we demonstrate that K-nearest neighbor provides the crime rate forecast system the greatest precision.

Table 5 Accuracy table

Year	Algorithm	Accuracy
2017, 18, 19	Linear	73.61403
2017, 18, 19	Naïve Bayes	69.5087
2017, 18, 19	KNN	76.9298

6 Conclusion

The sparsity of crime in many areas complicates the application of the prediction rate area-specific modeling. We used the Machine Learning algorithm in that work to create and test age, sex, year, moment, month prediction of crime. In that job we use three types machine learning algorithms Linear regression, Naïve Bayes and K-nearest neighbor among which we discover distinct precision in different instances some linear operates good and provides better precision but the general situation K-nearest neighbor provides the appreciated accuracy other than that's why we use K-nearest neighbor for our Crime Prediction scheme. By using these predict systems we will discover the stronger precision in the future and also by using this precision we will identify and discover the hot zone region in crime rate. In order to finish this job, we would like to use the CNN algorithm to analyze picture information and add the Google API for viewing the hot zone.

Acknowledgements We are very grateful to my Daffodil International University for offering us the chance to be part of the independent research study that contributes to this work being developed. Many thanks to Mr. Abdus Sattar, Assistant Professor, Department of Computer Science and Engineering at Daffodil International University for innumerable debates and feedback that helped me effectively finish the job.

References

1. Lin, Y., Chen, T., Yu, L.: Using machine learning to assist crime prevention. In: 2017 sixth IIAI International Congress on Advanced Applied Science (IIAI-AAI) (2017)
2. Munasinghe, M., Perera, H., Udeshini, S., Weerasinghe, R.: Machine learning based criminal short listing using modus operandi features (2015). https://doi.org/10.1109/icter.2015.7377669
3. Chauhan, C., Sehgal, S.: A review: crime analysis exploitation data processing techniques and algorithms, pp. 21–25 (2017). https://doi.org/10.1109/ccaa.2017.8229823
4. Anon: [online] Available at https://www.researchgate.net/publication/322000460_A_review_Crime_analysis_using_data_mining_techniques_and_algorithms. Accessed 30 Aug. 2019 (2019)
5. Kerr, J.: Vancouver police go high tech to predict and prevent crime before it happens. *Vancouver Courier*, July 23, 2017. [Online] Available https://www.vancourier.com/news/vancouver-police-go-high-tech-topredict-and-prevent-crime-before-it-happens-1.21295288. Accessed 09 Aug 2018
6. Marchant, R., Haan, S., Clancey, G., Cripps, S.: Applying machine learning to criminology: semi parametric spatial demographic Bayesian regression. Security Inform. **7**(1) (2018)