Naïve Bayes Approach for the Crime Prediction in Data Mining

Mrinalini Jangra
Student, M-TECH CSE
Dav Institute of Engineering and Technology
Jalandhar, Punjab 144022, India

Shaveta Kalsi
Assistant professor, CSE
Dav Institute of Engineering and Technology
Jalandhar, Punjab 144022, India

ABSTRACT
Prediction analysis is the analysis in which future trends and outcomes are predicted on the basis of assumption. It is the analysis in which future trends and outcomes are predicted on the basis of assumption. Machine learning techniques and regression techniques are the two approaches that have been utilized in order to conduct predictive analytics. In the conducting predictive analytics, machine learning techniques are widely utilized and become popular as large scale datasets handled by it is effective manner and provide high performance. It provides the results with uniform characteristics and noisy data. The KNN is the popular technique which is applied in the prediction analysis. To improve accuracy of crime prediction technique of Naïve Bayes is applied in this research work. It is evaluated that Naïve Bayes give higher accuracy as compared to KNN for the crime prediction.

Keywords
Crime prediction, KNN, Naïve Bayes, prediction

1. INTRODUCTION
Prediction analysis is the analysis in which future trends and outcomes are predicted on the basis of assumption. Machine learning techniques and regression techniques are the two approaches that have been utilized in order to conduct predictive analytics [1]. In conducting predictive analytics, machine learning techniques are widely utilized and become popular as large scale datasets are handled by it in effective manner and provide high performance. It provides the results with uniform characteristics and noisy data [2]. In the several domains, innovative predictive models have been implemented for the betterment of old results which provide results which are justified and better. By the use of prediction technique the dependent and independent variables and their relationship can be analyzed. The use of this technique can be used in future to attain enhanced results. Data mining is the technique that is utilized in almost all kind of applications. It is the process of building the representative model in which observational data is implemented [3]. Two purposes is served by this model first, on the basis of the input variables, the output is predicted and second, it helps in understanding the relationship between the output variable and all the input variables. In the society, crime is the part which needs to be recovered since laws were first approved. It is the act in which all the rules abided by the law are forbidden or omitted by the criminals for which, punishments are imposed which is a long process [4]. It is not easy to predict crimes. It is the act in which criminal offense the law and made a miss happening that affects the stable life of society. In this appropriated or good theoretical understanding is required so that practical solutions can be provided for the prevention of crime in which all details of time and location is suitably provided. The past crime data has been utilized for the crime analysis in order to predict future crime locations and time. Crime prediction for future crime is a process that finds out crime rate change from one year to the next and projects those changes into the future. In this process both qualitative and quantitative methods has been utilized. In order to forecast crime, environmental scanning, scenario writing are some attributes using which future nature of criminal activity is identified [5]. Hence, in order to predict the future scope of crime detectable methods has been utilized. To develop crime rates a common method for developing forecasts is to project annual crime rate trends developed through time series models. The exploitation of crime data is involved in the crime analysis by which law enforcement is enabled in order to capture criminals and prevent crimes. In the crime analysis process, there are two main components that are usually present such as crime variables and crime matching [6].The crime characteristics uniquely are described by the parameters of crime variables. Hence, it is considered as the main subject of crime analysis process. The process of assigning crimes or criminals to the previously solved or unsolved crime incidents is known as crime matching. In order to identify, discover and predict crime patterns a systematic approach is utilized of crime analysis. All the data and information gathered from city police department is considered as the input of a crime analysis system. Nowadays, the traditional methods are obsolete due to the complexity and presence of large amount of previously stored data [7]. In order to get patterns of data, large amount of resources and human effort is required for this method. Therefore, to obtain desired results the gathered data is transformed into useful knowledge by the data mining using which above mentioned problems can be easily overcome. Various data mining techniques are applied such as clustering and classification techniques so as to get patterns of data, once the data is transformed into the useful knowledge. There are few standard crime prediction techniques. Centrography is the technique in which the descriptive statistics are inspected, utilized in measurements of population potential or population centrality [8]. The coordinates x and y are assigned to crime and the center of mass is measured as follows: the mean x-coordinate and the mean y-coordinate and the association of this pair is done with residence of criminal and are identified by this technique. Journey to Crime is based on the belief that crime is likely to occur somewhere near, it can be close to an offender's home and with crimes it follows a distance-decay function in order to stay away from offender's house. As per Routine Activity Theory, in order for a crime to occur, there is a requirement of the presence of three elements which can be stated as an offender, an easy target and lack of a system for the protection against crimes. Circle Theory is the method in which distances between crimes are measured and the selection of two distant crimes is done. After that a drawing of an extreme circle is done so that both the points on the circle can be located on the great circle

[9].The criminal's residence is the midpoint of this extreme circle which is an assumed location and the where the criminal operates is the area bounded by the great circle. It is easy to understand this computationally economical model. This model is easy to use and in order to master this technique a very little training is required. However, this method has some limitations as well.

2. LITERATURE REVIEW

Olivera Kotevska, et.al (2017) proposed a dynamic network model through which the service resilience to data loss could be improved [9]. The statistically significant shared temporal trends available across multivariate spatiotemporal data streams were recognized through this network model. The dataset was collected from the city-based crime reports filed in Montgomery County, USA to demonstrate the performance of proposed network model. The cities with higher crime rates achieved an average of 5.6% of improvement when the proposed model was applied. Based on the reduction of prediction error, all the optimal network connections were identified correctly by this model.

S Prabakaran, et.al (2018) presented in the huge dataset, the design and patterns by finding out by the present procedure in the data mining. In the convergence of machine learning and database framework, this method includes the different strategies [10]. In the various fields, this technique can be implemented such as future healthcare, market basket analysis, education, manufacturing engineering, crime investigation and many more. In order to process the crime characteristics, the crime investigation is an interesting application that helps the society for a better living. In this paper, they studied all the data mining techniques used in this domain. It becomes easy to design new strategies for crime prediction and analysis due to this study.

Prajakta Yerpude, et.al (2017) presented there is enhancement in the crime rates across the globe that needs to be minimized for which proper analysis is required in order to lower the crime rate to minimal. With the help of this analysis it becomes easier for police to take appropriate steps and solve the crime as soon as possible [11]. They implemented the data mining techniques on crime data in this paper in order to predict features by which high crime rates are affected. Necessary actions have been taken by the Crimes Record Bureau and Police Department on the basis of ranking of the features by which probability of occurrence of the crime can be decreased.

Shyam Varan Nath, et.al (2016) presented data mining technique by which all the issues related to crime detection problems are solved easily. The whole society suffered due to crimes as it causes social irritation. About 50% of the crimes are committed due to the 10% of the criminals. In order to design the proactive services we can use data mining technology by which there is reduction in the crime activities and they can be controlled easily and early [12]. The implementation of a framework of data mining which works with geospatial crime plot is quite easy. It also enhances the productivity of detectives and other law enforcement departments. It can also be used for the implementation of activities against terrorism.

Dr. Nevine Makram Labib, et.al (2015) presented data mining is the technique widely utilized in almost all applications due to which it becomes powerful in all domains. In the field of counter terrorism these tools has been widely utilized. They also discussed the important data mining applications utilized in predictive policing. The main objective of this paper was to propose and recommend a data-mining model by which the most important factors affecting crime incidence has been predicted [13]. On the basis of performed experiments, it was concluded that proposed model must be implemented as a system which can be put in use by police directly from police head quarters for various levels of decision making.

Lawrence McClendon et.al (2015) presented for the detection and prevention of the crime, an essential role is played by both data mining and machine learning. In this paper, open source data mining software WEKA has been utilized by them using which there is a comparative analysis of patterns of violent crimes from various communities the dataset provided by the university of California regarding occurrence of crime [14]. On the basis of performed experiments, it is concluded that performance of the linear regression algorithm is better as compared to other two. The effectiveness and accuracy of the machine learning algorithms used in data mining analysis is proved using which all the violent crimes are predicted.

Shiju Sathyadevan, et.al (2014) presented the systematic approach in this paper using which patterns and trends in crime has been identified and analyzed properly. With the help of proposed model, it becomes easy to predict high probability for crime occurrence using which prone areas can easily visualize. The process of solving crimes is speeding up due to the increase in the computerized systems, due to increase in the advent of computerized systems [15]. In order to develop a data mining procedure using which crimes can be solve at faster rate, this approach is between the computer science and criminal justice. They mainly focused on the criminal background of offender, political enmity etc are the factors for the occurrence of crime.

Rasoul Kiani, et.al (2015) presented there is influence of crime on the society as it affects their fundamentals it creates fear among all other citizens. The classification of clustered crimes on the basis of occurrence frequency during different years is the main objective of this paper. In terms of analysis, investigation and discovery of patterns data mining techniques has been utilized for the occurrence of different crimes [16]. In this paper, they assigned the weights to the features by which quality of the model is improved and low value of them is removed. For the optimization of Outlier Detection operator parameters the RapidMiner tool is used with Genetic Algorithm (GA).

Tahani Almanie, et.al (2015) presented the identification of spatial and temporal criminal hotspots are the main focus of this paper. Two different real-world crimes datasets for Denver, CO and Los Angeles, CA was analyzed in this paper using which comparison between the two datasets was done using a statistical analysis supported by several graphs. In this paper the use of Decision Tree classifier and Naïve Bayesian classifier was showed as it helps in prediction of potential of crime types [17]. It is also used to capture the factors by which the safety of neighborhoods is affected. The obtained results are utilized to raise awareness among people regarding the dangerous locations and also predict future crimes in a specific location within a particular time that help all agencies.

3. PROBLEM FORMULATION

Data mining is the process in which all the necessary and useful information is extracted for the analysis of data. Various types of data mining tools are present in the data mining that are used for the analysis of different types of data. Some of the applications that use data mining in order to analyze the collected information are decision making, market

basket analysis, production control, customer retention, scientific discovers and education systems .There are some databases in the data mining such as multimedia, object relational, relational and data ware houses studied in detailed. The large amount of collected data became a devastating process of extraction as it is not easy to handle such huge amount of data. The prediction analysis is the technique which can predict future possibilities from the current data. This research work is based on the crime prediction of the states in India. In the existing work, technique of KNN is applied for the prediction of crime in the states. It is analyzed that KNN classifier has less accuracy and higher execution time which can be reduced to improve efficiency.

4. RESEARCH METHODOLOGY

This research work is based on the crime prediction in India. In the existing approach technique KNN classifier is applied for the prediction analysis. A simple and effective non-parametric classification approach used in several scenarios is known as k-Nearest-Neighbor (KNN). The k-nearest neighbors are extracted for classifying data record t. the neighborhood of t is generated. The classification of t can be decided using majority of voting amongst the data records present in the neighborhood. The distance-based weighting can be either available or not. However, an appropriate value of k is required to be selected for the application of KNN. On the basis of this value only, the success of classification is dependent. Thus, the KNN approach is completely dependent on value of k. this value can be chosen in many different ways. However, running the algorithm several times with various k values is one simple manner. The value through which best performance is achieved is selected here. New instances can be classified at higher cost using kNN. Several applications avoid using this algorithm since it is a lazy learning method. Identifying the representatives such the complete training data for classification can be represented is one manner through which the efficiency of this approach can be improved. The training dataset is used to generate an inductive learning model. Further, to perform classification, this model is utilized. For classification, KNN is known to be the simplest and most effective method. Thus, the efficiency of KNN is required to be enhanced while the classification accuracy of this algorithm is preserved.

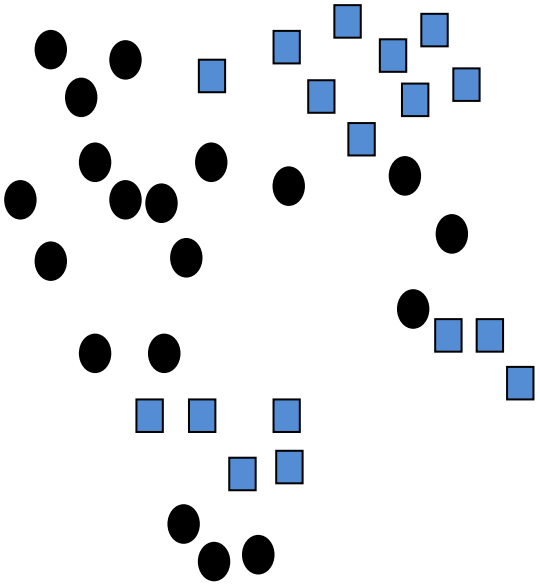

Fig. 1: The distribution of data points .

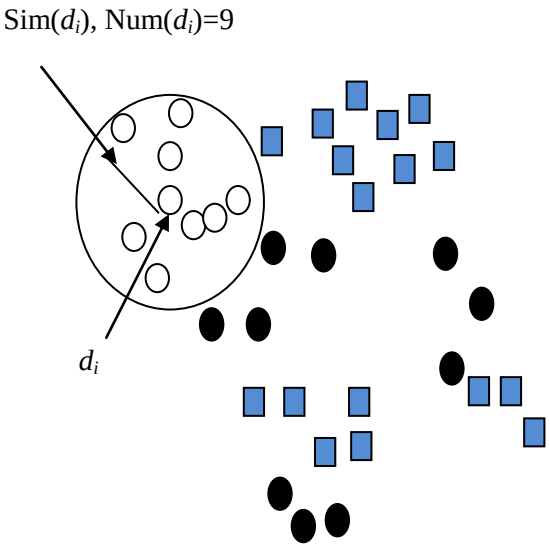

Fig. 2: The first obtained representative.

As shown in figure 1, there are 36 data points with two classes within a training dataset in a 2-D region. Several data points with similar class label are clearly shown closer to each other if Euclidean distance is utilized for measuring similarity. As shown in figure 2 for instance, in each local region, the central data point di is taken along with some extra information like Num(di) which provides the number of data points present within the local region. Within the local region di, the similarity of most distant data point is also used here which is denoted by Sim(di). The technique of KNN is replaced with the naïve Bayes classifier for the prediction analysis which increase accuracy of classification.

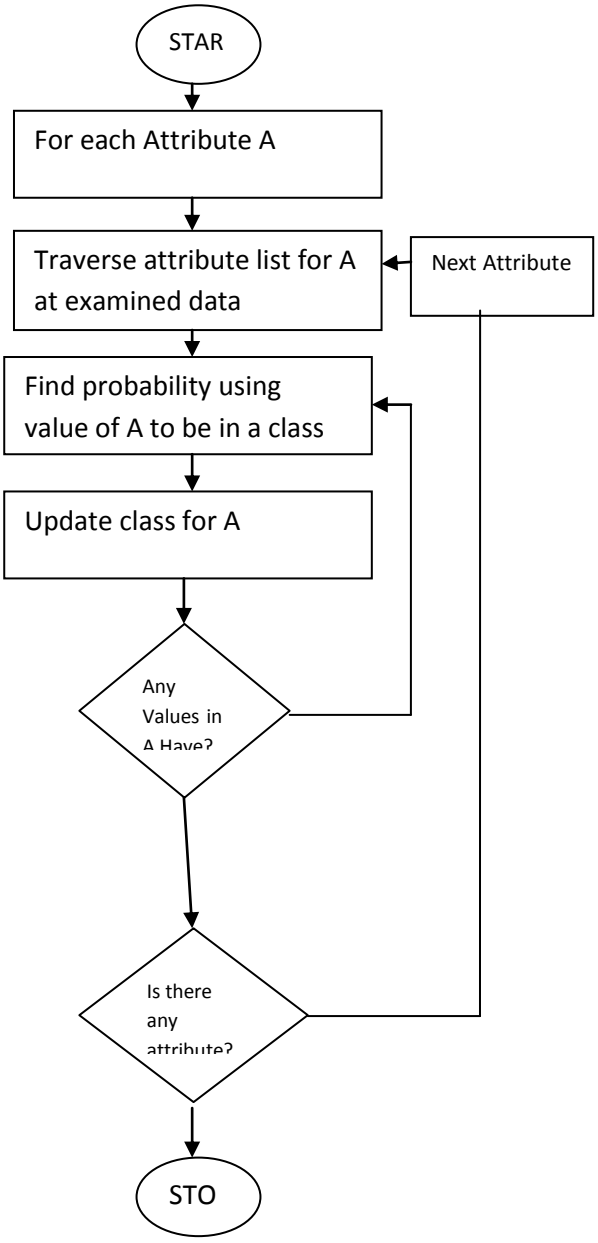

Fig 3: Proposed Methodology.

5. RESULTS AND DISCUSSIONS

This research work is related to crime prediction in India. The crime prediction is complex problem due to large number of attributes in the dataset. In the previous work KNN classifier was applied for the crime prediction. In the network, the naïve Bayes classifier is applied for the prediction analysis. The performance of the both classifiers are compared in terms of accuracy.

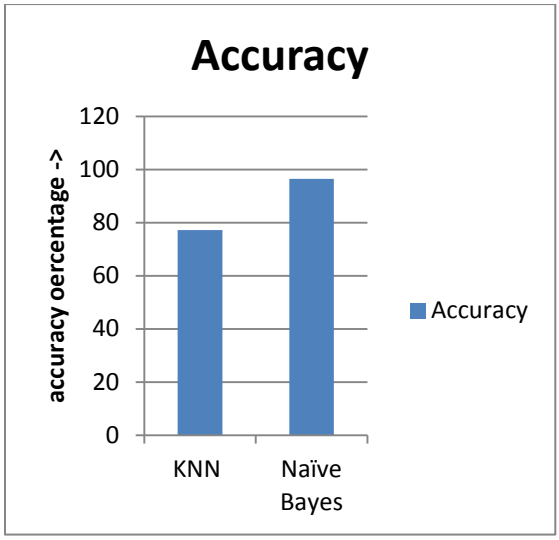

Fig 4: Accuracy comparison.

As shown in figure 4, the accuracy of proposed and existing scenarios are compared for the performance analysis. In the existing scenarios KNN classifier was applied for the crime prediction but in the proposed scenario naïve Bayes classifier is applied for the crime prediction. The accuracy of naïve Bayes classifier is high as compared to KNN for the prediction analysis.

Table 1: Accuracy comparison.

Parameter	KNN	Naïve Bayes
Accuracy	77.18	96.48

As shown in table 1, the accuracy of naïve Bayes and KNN is compared for the performance analysis. The naïve Bayes give higher accuracy as compared to naïve Bayes for the crime prediction.

6. CONCLUSION

In this work, it is concluded that crime prediction is the complex problem due to large number of attributes. In the previous research work, the approach of KNN is applied for the crime prediction analysis. In this work, the naïve Bayes technique is applied for the crime prediction. The proposed and existing

technique are implemented in python and it is analyzed that naïve Bayes give maximum accuracy of 96.48 percentage for crime prediction.

7. REFERENCES

[1] Bilmes, J. and Models, H. M.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, vol. 4, no 510, p. 126, 1998.

[2] Xinyu chen, Youngwoon Cho, and Suk young Jang, “Crime prediction using twitter sentiment and weather”, IEEE systems and information engineering design symposium, 2015.

[3] S.Yamuna, N.Sudha Bhuvneswari, "Data mining techniques to analyze and predict crimes", The International Journal of Engineering And Science (IJES)volume 1 Issue 2.

[4] Rasoul Kianin, Siamak Mahdavi, Amin Keshavarzi, "Analysis and prediction of crimes by clustering and classification", International Journal of Advanced Research in Artificial Intelligence (IJARAI) – volume 4 No. 8 2015.

[5] A.Malathi, Dr. S. Santhosh Baboo, "Evolving data mining algorithms on the prevailing crime trend – An intelligent crime prediction model", International Journal of Scientific & Engineering Research (IJSER) – volume 2 Issue 6 - June-2011.

[6] Yu-Yueh Huang, Cheng-Te Li, Shyh-kang Jeng, "Mining location-based social networks for criminal Activity prediction", Wireless and optical communication conference(WOCC) – 2015.

[7] Zakaria S. Zubi, Rema A. Saad,"Using Some Data Techniques for Early Diagnosis of Lung cancer",ISBN:978-960-474-273-8.

[8] Zakaria Suliman Zubi, Marim Aboajela Emsaed. 2010. Sequence mining in DNA chips data for diagnosing cancer patients. In Proceedings of the 10th WSEAS international conference on Applied computer science (ACS'10), Hamido Fujita and Jun Sasaki (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 139-151.

[9] Olivera Kotevska, A. Gilad Kusne, Daniel V. Samarov, Ahmed Lbath, and Abdella Battau, "Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City-to-City Network for Crime Analytics", IEEE Access, Volume: 5, 2017.

[10] S Prabakaran and Shilpa Mitra, "Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning", National Conference on Mathematical Techniques and its Applications (NCMTA 18), 2018.

[11] Prajakta Yerpude1 and Vaishnavi Gudur, "Predictive Modelling Of Crime Dataset Using Data Mining", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4, July 2017.

[12] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining", IEEE, 2016.

[13] Dr. Nevine Makram Labib, Brigadier-General Wael Kamal Arafa, "A Proposed Data Mining Model to Enhance Counter- Criminal Systems with Application on National Security Crimes", International Research Journal of Computer Science (IRJCS) ISSN: 2393-9842 Issue 7, Volume 2 (July 2015).

[14] Lawrence McClendon and Natarajan Meghanathan, "Using Machine Learning Algorithms To Analyze Crime Data", Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 2015.

[15] Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S, "Crime Analysis and Prediction Using Data Mining", IEEE, 2014.

[16] Rasoul Kiani, Siamak Mahdavi, Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.8, 2015.

[17] Tahani Almanie, Rsha Mirza and Elizabeth Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015.