International Journal for Modern Trends in Science and Technology, 9(02): 42-46, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0902007
Available online at: http://www.ijmtst.com/vol9issue02.html

Crime data analysis and prediction using Machine Learning

Dr.D.Suneetha[1] | P.Harshitha[2] | S.Jhansi Sri[2] | K.Sri Lasya[2]| P.Sai Pavan[2]

[1]Professor & HOD, Department of CSE, NRI Institute of Technology, India
[2]B.Tech Student, Department of CSE, NRI Institute of Technology, India

To Cite this Article
Dr.D.Suneetha, P.Harshitha, S.Jhansi Sri, K.Sri Lasya, P.Sai Pavan, Crime data analysis and prediction using Machine Learning. International Journal for Modern Trends in Science and Technology 2023, 9(02), pp. 42-46. https://doi.org/10.46501/IJMTST0902007

Article Info
Received: 30 December 2022; Accepted: 01 February 2023; Published: 03 February 2023.

ABSTRACT

To reduce the prevalence of crime in our society is an important goal. Analysis of criminal activity follows a systematic method for uncovering and analyzing trends and patterns. Researching the causes of crime, taking into account relevant circumstances, establishing causality, and identifying effective strategies for crime prevention are all very important. Using clustering approaches based on occurrences and frequency, this study aims to categorize and differentiate between different types of crime. Crime trends may be analyzed, investigated, and checked using data mining. The criminal data in this project is analyzed using a clustering method; the collected information is grouped together with the help of the K-Means algorithm. Once we've categorized and clustered relevant data, we may use it to make a crime prediction. The suggested approach can identify high-crime locations and those with lower crime rates.

KEY WORDS: crime analysis, prediction analysis, machine learning, decision trees, pattern detection.

1. INTRODUCTION

The most serious danger that humanity faces is that posed by criminals. There are a great number of crimes that occur at consistent intervals of time. It's possible that it's becoming worse and more widespread at an alarmingly quick pace. There is criminal activity everywhere, from the smallest hamlet to the most populous city. There are many distinct types of criminal offenses, including robbery, homicide, rape, assault, battery, false imprisonment, abduction, and homicide. [1][2]. Because the number of crimes is rising, there is an urgent need for significantly accelerating the process of solving the cases. The pace at which criminal actions are committed has accelerated, and it is the duty of the police department to bring these criminal activities under control and bring them to a lower level. [3][4]. As a result of the enormous quantity of crime data that is now available, the primary challenges faced by the police department are those of crime prediction and criminal identification. There is a need for technologies that might make the process of addressing cases much quicker. The goal would be to hone a predictive model via the process of training [5][6]. The training would be carried out with the help of the training data set, which would then be verified with the help of the test dataset. The construction of the model will be carried out using a more efficient method in accordance with the correctness. For the purpose of crime prediction, both

the K-Nearest Neighbor (KNN) classification and alternative algorithms will be used. The visualization of the dataset is done in order to conduct an analysis of the possible criminal acts that have taken place in the nation [33] [34 [35].

This study enables law enforcement agencies in Chicago to enhance the accuracy with which they forecast and identify criminal activity, which in turn leads to a reduction in the city's overall crime rate [7], [8]. There has been a significant growth in the number of machine learning algorithms, which, when applied to historical data, have made crime prediction possible. Using machine learning models, the purpose of this study is to conduct an analysis and make predictions on criminal activity in states [47] [48]. It focuses on developing a model that may assist in the estimation of the number of crimes committed in a certain state according to the nature of the offense [9][10].

In the context of this study, several machine learning models, such as K-NN and boosted decision trees, will be used in order to make predictions about criminal behavior [11], [12]. Area It is possible to have a better understanding of the pattern of crimes by doing careful geographical research. It is possible to improve the ability of law enforcement authorities to identify and forecast criminal activity by making use of a number of different visualization methods and plots [30], [31]. This will assist indirectly to lower the rates of crime, and it may also help to strengthen the security in regions where it is essential to do so. Because criminals are busy and work within their familiar environments, it is possible to anticipate illicit acts. After they have been successful, they attempt to do the same crime again in conditions that are quite similar [13][14].

2. LITERATURE SURVEY

2.1 Investigating Criminal Activity Through the Lens of Machine Learning Algorithms

Data mining and machine learning have rapidly become indispensable tools for the investigation and prevention of criminal activity. In this investigation, we make use of WEKA, a piece of open-source data mining software, to carry out a comparative study between the violent crime patterns derived from the Communities and Crime Unnormalized Dataset that was made available by the repository at the University of California, Irvine, and the actual crime statistical data for the state of Mississippi that has been made available by neighborhoodscout.com. These two sets of information are compared in order to determine whether or not there is a correlation between the two sets of information. On the Communities and Crime Dataset, we used the same limited number of characteristics to develop the Linear Regression, Additive Regression, and Decision Stump algorithms. Among the three algorithms that were chosen, the one that performed the best overall was the linear regression method. The purpose of this research is to demonstrate the efficacy and precision of the machine learning algorithms that are used in the process of data mining analysis in terms of forecasting patterns of violent crime.

2.2. Conducting statistical investigations based on findings from machine learning programs

Data mining and machine learning have rapidly become indispensable tools for the investigation and prevention of criminal activity. In this investigation, we make use of WEKA, a piece of open-source data mining software, to carry out a comparative study between the violent crime patterns derived from the Communities and Crime Unnormalized Dataset that was made available by the repository at the University of California, Irvine, and the actual crime statistical data for the state of Mississippi that has been made available by neighborhoodscout.com. These two sets of information are compared in order to determine whether or not there is a correlation between the two sets of information. On the Communities and Crime Dataset, we used the same limited number of characteristics to develop the Linear Regression, Additive Regression, and Decision Stump algorithms. Among the three algorithms that were chosen, the one that performed the best overall was the linear regression method. The purpose of this research is to demonstrate the efficacy and precision of the machine learning algorithms that are used in the process of data mining analysis in terms of forecasting patterns of violent crime.

2.3. Prediction and Investigation of Criminal Activity Using many forms of machine learning

The prevention of crime is a very essential endeavor since it is one of the most pervasive and serious problems that our society faces. There are a significant number of crimes that are perpetrated on a regular basis. This necessitates keeping a record of every crime

committed and preserving a database containing that information so that it may be referred to in the future. The present challenge that is being dealt with is the upkeep of accurate crime datasets and the analysis of these datasets to assist in the prediction and resolution of future crimes. This project's goal is to examine a dataset that contains a large number of crimes in order to make a prediction about the kind of crime that could occur in the future dependent on a number of different factors. For the purpose of this research, we will be making predictions about the criminal activity based on the Chicago crime data set utilizing the methods of machine learning and data science. The statistics on crimes were taken from the official website of the Chicago Police Department. It includes details on the crime, such as the date, time, latitude, and longitude of the incident, as well as a description of the place. The data will be preprocessed before the training of the model, and then feature selection and scaling will be carried out thereafter in order to ensure that the accuracy obtained is as high as possible. For the purpose of crime prediction, both the K-Nearest Neighbor (KNN) classification and a number of other algorithms will be evaluated, and the one that demonstrates the highest level of accuracy will be selected for further training. The dataset will be shown via the use of graphical representations of many different scenarios, such as determining at what time of day the highest crime rates occur or during which month the highest number of criminal offenses are committed. This project's overarching objective is to provide some kind of an understanding about how machine learning may be used by law enforcement organizations in order to identify, anticipate, and solve crimes at a much accelerated pace, which would ultimately result in a lower overall crime rate. This is not limited to Chicago; in fact, depending on the availability of the dataset, it might be utilized in other states or perhaps other nations.

The Chicago data set is used for the purpose of crime prediction, in which a variety of machine learning models are utilized. [15] In this study, a comparison of several models, such as KNN, Naive Bayes, and SVM, is carried out. It is clear that the accuracy of prediction shifts based on the dataset as well as the characteristics that are prioritized for inclusion [16] [17 [18]. According to the findings of the study, the accuracy of prediction is 78% for KNN, 64% for GaussianNB, and 31% for SVC. Auto regressive integrated Moving average models were used in the development of machine learning algorithms for the purpose of predicting the patterns of criminal activity in metropolitan regions. [19][20]. Identifying and examining the recurrent nature of criminal activity is a significant challenge in the field of criminology. Knowing how to read and interpret datasets is another crucial idea in this scenario. It is important for us to have precise forecasts in order to avoid squandering our efforts on erroneous indications. [21] [22][23]. Additionally, a strategy was suggested for categorizing the crime rate as either high, medium, or low. None of them have categorized the many types of criminal activity that may occur and the likelihood of those crimes occurring. [24][25][26]. Analysis and forecasting of criminal behavior is an essential task that may be improved by making use of a variety of different methods and procedures. A significant amount of effort in the form of study is performed in this field by a variety of researchers. The scope of the work that has been done so far is confined to identifying crime hotspots via the use of datasets [27][28][29].

3. PROPOSED SYSTEM

The random forest method is used here in the system that was presented by us in order to get excellent results and improved accuracy when compared to the algorithms described above or those that already exist. For increased accuracy, we make advantage of random forest . Random forest is a widely used and highly effective supervised machine learning algorithm that is capable of performing both classification and regression tasks. It achieves this by constructing a large number of decision trees during the training phase and then outputting a class that is either the mode of the classes (during classification) or the mean prediction (during regression) of the individual trees . Random decision forests are an alternative to decision trees, which have a tendency to overfit to the data in their training sets. Rainfall, perception, production, and temperature are the data sets that are taken into account in order to build a random forest, which is a collection of decision trees built by taking into account two-thirds of the records in the datasets . In order to classify the data accurately, these decision trees are applied to the records that are

still outstanding. The random forest method has an accuracy score of 80.6%. Using the undersampled data, Adaboost decision tree successfully classified criminal activities based on the time and location. With a accuracy of 81.93%, it was able to outperform other machine learning algorithms.

4. RESULTS

```
In [9]: crime.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1456714 entries, 0 to 1456713
Data columns (total 23 columns):
Unnamed: 0              1456714 non-null int64
ID                      1456714 non-null int64
Case Number             1456713 non-null object
Date                    1456714 non-null object
Block                   1456714 non-null object
IUCR                    1456714 non-null object
Primary Type            1456714 non-null object
Description             1456714 non-null object
Location Description    1455056 non-null object
Arrest                  1456714 non-null bool
Domestic                1456714 non-null bool
Beat                    1456714 non-null int64
District                1456713 non-null float64
Ward                    1456700 non-null float64
Community Area          1456674 non-null float64
FBI Code                1456714 non-null object
X Coordinate            1419631 non-null float64
Y Coordinate            1419631 non-null float64
Year                    1456714 non-null int64
Updated On              1456714 non-null object
Latitude                1419631 non-null float64
Longitude               1419631 non-null float64
Location                1419631 non-null object
dtypes: bool(2), float64(7), int64(4), object(10)
memory usage: 230.2+ MB
```

Figure 1: Information on Crime

Figure 2: Crimes per Month

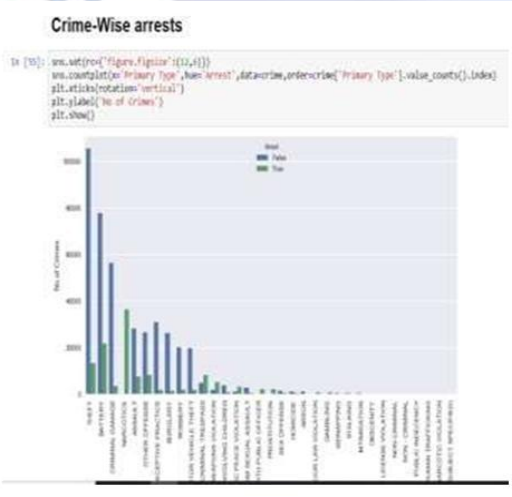

Figure 3: Crime wise Arrest

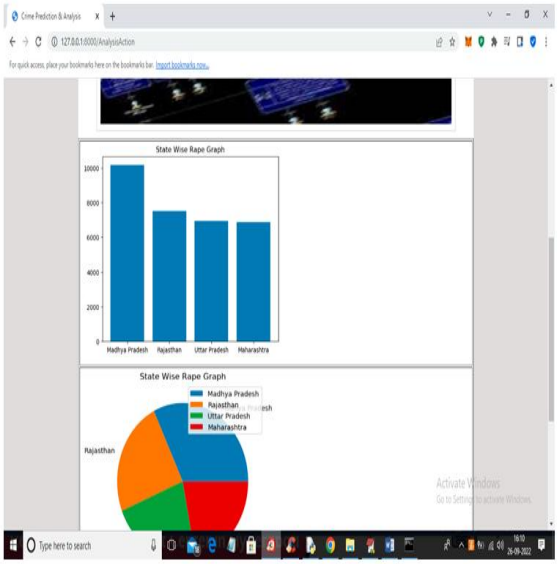

Figure 4: State wise Crime Graph

5. CONCLUSION

Throughout the course of the investigation, it became clear that the most fundamental aspects of criminal activity in a specific region include markers that may be used by machine learning agents to categorize criminal behavior given a place and a date. The learning agent has been able to conquer the challenge by oversampling and undersampling the dataset. This is despite the fact that the dataset contains categories that are unbalanced. Through the use of the experiments, it is possible to show that the unbalanced dataset benefited from the use of ENN undersampling. The Adaboost decision tree was able to accurately classify illegal behaviors based on the time and place using the data that was only partially sampled. It was able to surpass other machine learning algorithms with a level of accuracy that reached 81.93%. One of the most significant obstacles on the path to a better outcome is having courses that are unbalanced. Even though the machine learning agent was able to create a prediction model using just crime data, it is likely that adding a demographic dataset will assist to further enhance the outcome and consolidate it.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The`BigData'Revolution in Healthcare: Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Ofce, 2016.

[2] M. Chen, S. Mao, and Y. Liu, ``Big data: A survey,'' Mobile Netw. Appl., vol. 19, no. 2, pp. 171209, Apr. 2014.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak, ``Mining electronic health records: Towards better research applications and clinical care,'' Nature Rev. Genet., vol. 13, no. 6, pp. 395405, 2012.

[4] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, ``A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics,'' IEEE Trans. Intell. Transp. Syst., vol. 16, no. 6, pp. 30333049, Dec. 2015.

[5] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, ``Wearable 2.0: Enable human-cloud integration in next generation healthcare system,'' IEEE Commun., vol. 55, no. 1, pp. 5461, Jan. 2017.

[6] M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu, ``Smart clothing: Con- necting human with clouds and big data for sustainable health monitor- ing,'' ACM/Springer Mobile Netw. Appl., vol. 21, no. 5, pp. 825845, 2016.

[7] M. Chen, P. Zhou, and G. Fortino, ``Emotion communi-cationsystem,'' IEEE Access, vol. 5, pp. 326337, 2017, doi: 10.1109/ACCESS.2016.2641480.

[8] M. Qiu and E. H.-M. Sha, ``Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems,'' ACM Trans. Design Autom. Electron. Syst., vol. 14, no. 2, p. 25, 2009.

[9] J. Wang, M. Qiu, and B. Guo, ``Enabling real-time information service on telehealth system over cloud-based big data platform,'' J. Syst. Archit., vol. 72, pp. 6979, Jan. 2017.

[10] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, ``Big data in health care: Using analytics to identify and manage high-risk and high-cost patients,'' Health Affairs, vol. 33, no. 7, pp. 11231131, 2014.

[11] L. Qiu, K. Gai, and M. Qiu, ``Optimal big data sharing approach for tele- health in cloud computing,'' in Proc. IEEE Int. Conf. Smart Cloud (Smart- Cloud), Nov. 2016, pp. 184189.

[12] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, ``Health- CPS: Healthcare cyber-physical system assisted by cloud and big data,'' IEEE Syst. J., vol. 11, no. 1, pp. 8895, Mar. 2017.