International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Extraction and Analysis of Crime Data to Prevent Society using Machine Learning Approaches

Dr. Pradeep Kumar[1], Maryada Purushottam[2]

[1]University Professor of Mathematics, Bhimrao Ambedkar University, Muzaffarpur
Principal, S. N. S College, Motihari
Contact: 9031349667
Email: *dr.pkumar15[at]gmail.com*

[2]Ph.D Scholar (Computer Science), Bhimrao Ambedkar University, Muzaffarpur (Bihar)
Contact: 9716718702
Email: *maryada1990[at]gmail.com*

Abstract: ***The era of digitalization and computers is upon us. The vast coverage of the digital world has completely changed the way we do and look at large and small parts of our lives. Most of our processes objectives and future planning is now based on technology. Being new to this technology driven and dominated world, we are now forced to rely on machine learning that will lead us to enhance our processes and adapt to the new culture. Provide the best suggestion to the citizens to choose the right residential location and for the police departments to solve crime through the dataset.***

Keywords: Digitalization, Information Technology, Machine learning, Police departments, Crime related dataset.

1. Introduction

The main motive behind designing, and employing this system is the randomized functionality/reactions of the domain and the lack of knowledge of the user regarding the domain. This system would help in enhancing user capabilities and thus in optimizing the results/outcome. It is the perfect solution to the problem of 'lack of knowledge' which is very common in the present - day scenario due to the continuous development of existing fields and the emergence of newer fields (Fayyad et al., 1996).

This will help the citizens make better decisions in selecting a safe and secure place for their residence. It will provide safeguard to the citizens from the criminal activities which the state is facing. Through this, the government agencies can also keep in check the criminal activities (Ozkan, 2004).

1.1 Problem Overview

This system seeks to provide citizens with the most suitable advice when selecting a residential location and police department for the purpose of combating crime through the use of datasets. To begin the process of problem - solving, a criminal data set from various locations of India is fed into the project and analyzed. Subsequently, the user is asked to provide their preferred location, and a machine learning system is then used to monitor the data and identify the most suitable living place based on its lowest crime rate, as reported by Gupta et al. (2008). The system will then make a prediction based on the information provided as a dataset for government agencies, using a machine learning approach.

1.2 Significance and Applications of Problem in Real World

The purpose of the system is to reduce the crime rates by eliminating the workload on the police department as the system automatically does the work for them, for example, the places where the crime rates are high, but the government are taking no actions against it and the places where the crime rates are low, but the government are over - protected on those places (Malathi and Baboo, 2011). Some other real - world examples are as follows:

- This project is the study of the previous and current criminal data and analyzes the crime rate at any place.
- Government agencies can also take advantage of the system for finding preventive measures and taking the necessary steps to nullify the crime.
- Project study the field of crime and analyses, for providing the suggestion to the people for selecting the safe place for their living.
- The project will help the associated authorities to know/target the place and use the developmental tool for making the people aware and stop them from doing so.

1.3 Objective

This paper is all about creating a system that uses machine learning algorithms to figure out crime rates in places and make sure everyone is safe. The goal is to make a system that takes the crime data and comes up with the best solutions based on that data.

2. Features

2.1 Location Selection

The safest location can be easily found by the people according to the lowest crime rate of the place from the dataset provided to the system (Brown, 1998). The data will provide information about the location of the location so that the people and authorities can take precautionary measures to prevent the situations (like theft, robbery, fraud, kidnapping etc.).

2.2 Minimization of crime

With the help of the system the government agencies, police departments and security organizations can easily detect the problem of the places where the crime rate is high as compared to the others and minimize it by taking necessary decisions (Wang et al., 2012).

2.3 Public awareness

According to the saying that precaution is better than cure, this can apply to this also. For decreasing the crime rate of the place first we have to educate the people by educating them. This can be done differently according to the face of the crime (Chau et al., 2002).

2.4 Decision making

Decision making can be done at different levels, at authority level and personal level. Authority or security department can take the right decision to handle the situation and the citizens can take precautions at their personal level by restricting themselves from going to those places (Gupta and Rana, 2019).

3. Dataset Description

The dataset consists of 2500 total instances and 8 attributes for communities, 6 predictive, 2 non predictive in each instance belonging to different states in India. The states are represented in the form of the number. Attributes include information across a variety of crimes and distinguish them on their nature such as murder, rape, robbery, and assault. The complete details of all the attributes can be obtained from the machine learning repository website www.kaggle. com (Gupta and Rana, 2019a).

3.1 Pre - processing

There are a few techniques practiced for data pre - processing. The techniques like data cleaning, discretization and data transformation, and feature selection are employed for this purpose. Which leads to reducing some noises, incomplete and inconsistent data. The result from the preprocessing step is then followed by the data mining algorithm. For the first step, the goal for data cleaning is to decrease noise and handle missing values and then we perform data normalization, discretization, and data type transformation (Gupta and Rana, 2019b).

3.2 Feature Selection

Relevance analysis or feature selection is used to find the attribute upon which we consider the problem and remove the irrelevant or redundant attributes. Feature selection has several objectives such as enhancing model performance by avoiding overfitting in the case of supervised classification (Gupta et al., 2021). For attribute selection, two mechanisms were used to select the final set of attributes:

- The Golden Standard or manual selection of attributes is based on human understanding and intellect.
- Using the Chi - square test to detect the correlated attributes.

3.3 Selected Classifiers

After preprocessing and feature selection phases, we must find a suitable algorithm that will help in getting the result. Here what we do is perform a few machine learning algorithms on the dataset to get the goal and then we compare the result and opt for the best one among them. As we have cut down the numbers of attributes meaningfully which will give us more precise data for building the data mining models (Vaishnav et al., 2021). In order to qualitatively predict the crime status from the quantitative data, as mentioned above, the following machine learning algorithms are being used.

- Using a supervised learning technique, Naive Bayesian classifiers can estimate the likelihood of a given tuple dependency to a certain class. This classifier is relatively simple to build and can be applied to large data sets with ease.
- Support vector machines (SVMs) are supervised learning algorithms that may be used for classification or regression. The SVM aim in a two - class learning problem is to find the optimal classification function to discriminate between members of the two classes in the training data (Chatterjee, 2021).

3.4 Challenges and Limitations

- **Data extraction:** Data extraction is the act or process of obtaining data from unstructured data sources for further processing or storage (also known as data migration). Thus, the input into the intermediate extraction system is frequently followed by data modification and optionally metadata addition before export to the next stage in the data pipeline (Kumar and Dhiman, 2021). In this work, data is extracted from several sources (security agencies, police departments, defense authorities etc.).
- **Predictive analysis:** It is a subset of advanced analytics that is used to forecast unknown future occurrences. To produce predictions, predictive analysis employs a variety of approaches from data mining, statistics, modeling, machine learning, and artificial intelligence.
- **Classification and generalization of data:** The process of classifying and categorizing data into numerous types, forms, or any other separate class is known as data classification. Data classification allows for the separation and classification of data based on data set needs for a variety of corporate or personal goals. It is mostly a data management procedure. Because we have a broad domain of criminal occurrences, the procedure of categorization or category division is necessary.
- **Inconsistency:** If the same data is saved in two files in different formats or if data must be matched between files, there is inconsistency.
- **Coordination:** As we have data of too many places so the correlation or the coordination of the data is necessary.

3.5 Technology to be used

- **Machine learning tools:** Machine learning is a technique that enables computers to automatically learn and improve based on their experiences without being explicitly programmed. Machine learning is concerned

with the creation of computer programmes that can access data and utilize it to learn on their own. There are several machine learning algorithms that lead to the output decision tree, clustering, regression, and classification.

- **Implementation Language:** Python is a general - purpose interpreted, interactive, object oriented, and high - level programming language which complements machine learning very well and it is more compatible with the language.

4. Results and Discussions

Evaluation on two selected classification algorithms on two different sets of features was conducted by comparing the findings on precision and accuracy. Precision shows that the proportion of data is classified correctly. Accuracy is the percentage of instances which is classified correctly by classifiers. From Figure 1, the property stolen type of crime is shown according to the area. And, as we can see that the crime rate in Maharashtra, Delhi, Gujarat is the highest, so the police department must take care of those areas to reduce the crime rate instead of other areas. The type of property stolen crimes are shown in the figure which includes burglar - property, criminal breach of trust - property, Dowry property, other heads of property, robbery - property, theft - property.

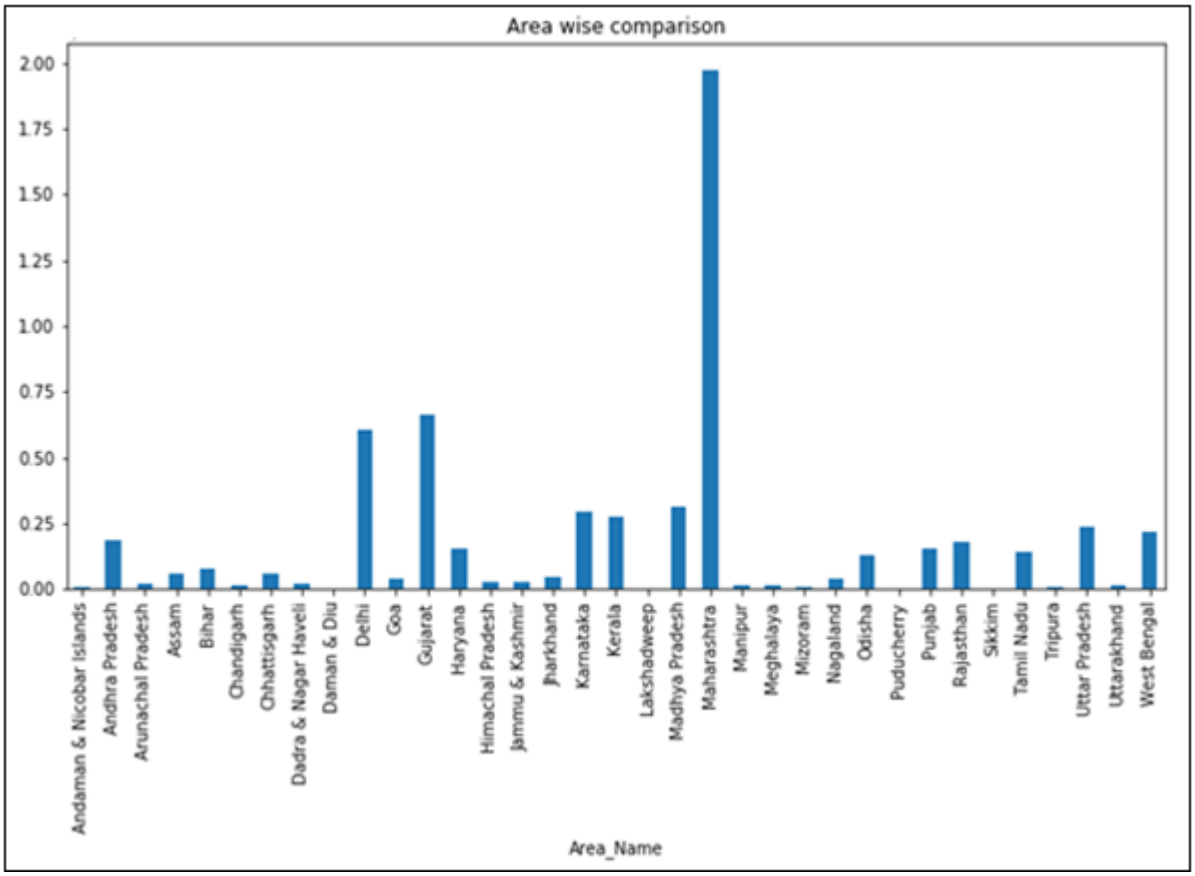

Figure 1: The property stolen type of crime is shown according to the areas of India.

5. Conclusion

The aim of this study is to classify the given dataset into two types of variables which are dependent (critical) and independent (non - critical) variables. In this regard, we used two classification algorithms by combining two different ways of feature selection techniques, manually and Chi - square, to determine more accurate classifiers. At present, the crimes in different states of India are increasing day by day because of this, people feel insecure and find their society inappropriate. There are different types of crime and because of that, the security personnel face difficulty in handling those. This system will identify and focus on the highest committed crime at the location. The system will be applicable at different levels, a citizen can find their own perspective by finding a secure place for their livelihood and the security department can apply this system for making the place secure by handling criminals because of this crime rate can predictively be minimized.

6. Future Work

Currently, the scope of this project is limited to data available from police departments. Soon, we will broaden the data source available from only the police departments to NIA, CBI, CID and many more departments as well. This time only common people and police departments use the system but, in future, the system will also be available to corporate offices, businessmen so that they can find a suitable place for their business enhancement.

References

[1] Brown, D. E. (1998, October). The Regional Crime Analysis Program (ReCAP): a framework for mining data to catch criminals. In SMC'98 Conference Proceedings.1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218) (Vol.3, pp.2848 - 2853). IEEE.
[2] Chatterjee, I. (2021). Artificial intelligence and patentability: review and discussions. International Journal of Modern Research, 1 (1), 15 - 21.
[3] Chau, M., Xu, J. J., & Chen, H. (2002, May). Extracting meaningful entities from police narrative reports. In Proceedings of the 2002 annual national conference on Digital government research (pp.1 - 5).
[4] Fayyad, U. M., Djorgovski, S. G., & Weir, N. (1996). Automating the analysis and cataloging of sky surveys. Advances in knowledge discovery and data mining (pp.471 - 493).
[5] Gupta, M., Chandra, B., & Gupta, M. P. (2008). Crime data mining for Indian police information system. Computer society of India, 40 (1), 388 - 397.
[6] Gupta, V. K., & Rana, P. S. (2019a). Activity assessment of small drug molecules in estrogen receptors using multilevel prediction models. IET systems biology, 13 (3), 147 - 158.
[7] Gupta, V. K., & Rana, P. S. (2019b). Toxicity prediction of small drug molecules of androgen receptor using multilevel ensemble model. Journal of bioinformatics and computational biology, 17 (05), 1950033.
[8] Gupta, V. K., & Rana, P. S. (2019c). Toxicity prediction of small drug molecules of aryl hydrocarbon receptor using a proposed ensemble model. Turkish Journal of Electrical Engineering & Computer Sciences, 27 (4), 2833 - 2849.
[9] Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID - 19 confirmed, death, and cured cases in India using random forest model. Big Data Mining and Analytics, 4 (2), 116 - 123.
[10] Kumar, R., & Dhiman, G. (2021). A comparative study of fuzzy optimization through fuzzy number. International Journal of Modern Research, 1 (1), 1 - 14.
[11] Malathi, A., & Baboo, S. S. (2011). Enhanced algorithms to identify change in crime patterns. International Journal of Combinatorial Optimization Problems and Informatics, 2 (3), 32 - 38.
[12] Ozkan, K. (2004). Managing data mining at digital crime investigation. Forensic science international, 146, S37 - S38.
[13] Vaishnav, P. K., Sharma, S., & Sharma, P. (2021). Analytical review analysis for screening COVID - 19 disease. International Journal of Modern Research, 1 (1), 22 - 29.
[14] Wang, Y., Chen, F., & Qu, X. (2012). Research and Application of Large - Scale Data Set Processing Based on SVM. Journal of Convergence Information Technology (JCIT), AICIT, 7 (16), 195 - 200.

Author Profile

Maryada Purushottam, Ph.D Scholar (Department of Computer Science), BRA Bihar University, Muzaffarpur, Bihar.

Dr. Pradeep Kumar, received Ph. D in 1995 from BRA Bihar University, Muzaffarpur, Bihar, Working as an university professor & head of the Department of Mathematics in M. S College, Motihari P. G college, two UGC sponsored MRP performed on the topics “Parametrized Post Newtonian Formalism: Gravitation & Application” (2001 - 03) and “Super Energy of Gravitational Waves “ (2011 - 13) worked, as PhD guide, published several research papers, recently “Gravitational Spiral Waves” vol.2 issue 6th June 2013, “Intensive Energy of Gravitational Waves”, vol.2 issue 7th July 2013, “Quantum Effects Gravitation”, vol.2 issue 10th October 2013, “Gravitational Lens” vol.3 issue 9th September 2014, “Generation of Gravitational Waves”, vol.4 issue 10 October 2015, and “Gravitational Radiation: A Stimulating Flux of Energy, volume 5 issue 12 December 2016.