Crime Tracking System and People's Safety in India using Machine Learning Approaches

Vishan Kumar Gupta*
CSE Department,
Graphic Era Deemed to be University, Dehradun, India
E-mail: vishangupta@gmail.com

Surendra Kumar Shukla
CSE Department,
Graphic Era Deemed to be University, Dehradun, India
E-mail: surendrakshukla21@gmail.com

Anupriya
CSE Department,
Graphic Era Deemed to be University, Dehradun, India
E-mail: anupriya@gehu.ac.in

Ramesh Singh Rawat
CSE Department,
Graphic Era Deemed to be University, Dehradun, India
E-mail: rsrawat06@gmail.com

Corresponding author: Vishan Kumar Gupta

Abstract
The era of digitization and computers has already arrived. The enormous coverage of cyberspace has changed the way of overall major and minor aspects of the life of doing and looking over things. Most of our processes aims and future planning are now partially or fully dependent on technology. Thus, being new to this technology-driven and dominated environment, we are forced to cling upon machine learning that will guide us to improve our processes and to adapt to this new culture. To provide the optimal suggestion to the citizen to select the best residential location as well as for the police department to tackle crime through the dataset.

Keywords- Machine learning; support vector machine; crime rates; police department; pre-processing, classifiers.

1. Introduction
The main motive behind designing, developing, and employing this system is the randomised 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 for making 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 Description & Overview

This project aims to provide the optimal suggestion to the citizen in selecting a residential location and police department to tackle crime through datasets. For the solution of the problem statement first, we will feed the criminal data set of the different locations of India and analyse it. After that, the user will give their preferable location and then with the help of a machine learning system will observe the data and predict the best living place according to the lowest crime rate (Gupta et al., 2008). Prediction will be made by the system over the information which has been provided as the dataset for the government agencies by using the machine learning method.

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 analyses the crime rate at any place.
- Government agencies can also take the advantage of the system for finding preventive measures and taking the necessary step 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

The objective of this paper is to develop and implement a system that utilizes machine learning algorithms for users such as police department and normal citizens to calculate the crime rates in places and provide safety. The goal of the paper is to create a system that will input the crime data and provide optimal solutions based on it.

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 situation of the location so that the people and authority can take the precautionary measure 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 literate 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 citizen can take precautions at their personal level by restricting themselves from going to those places (Gupta and Rana, 2019a).

3. Dataset Description

The dataset consists of 2450 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, 2019b).

3.1 Pre-processing

There are a few techniques is 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 pre-processing 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 performed data normalization, discretization, and data type transformation (Gupta and Rana, 2019c).

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 over fitting 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.

4. Selected Classifiers

After pre-processing 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).

5. 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, defence 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, modelling, machine learning, and artificial intelligence.

- **Classification and generalization of data**: The process of classifying and categorising 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.

- **Co-ordination:** As we have data of too many places so the co-relation or the co-ordination of the data is necessary.

6. Technology to be used

- **Machine learning tools used:** 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 utilise it to learn on their own. There are several machine learning algorithms that lead to the output decision tree, clustering, regression, and classification [10].
- **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.

7. 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.
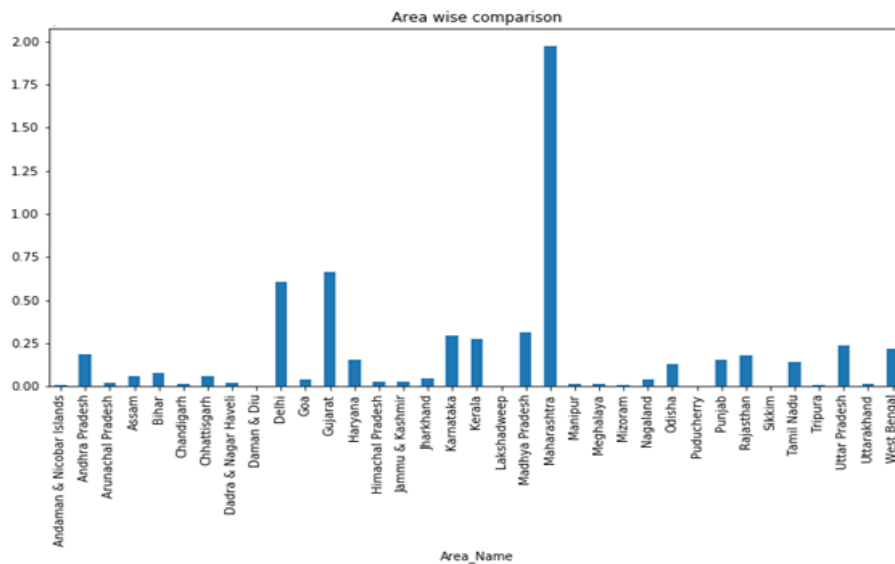

Figure 1: The property stolen type of crime is shown according to the areas of India.

From Figure 2, the crime in a particular state (ASSAM) is shown according to yearly basis, which shows the increase in crime in the year 2006 and 2009. So, this needs to be minimized or reduced and with the help of machine learning, we can predict that what type of crime is going to take place in which area and at which place.
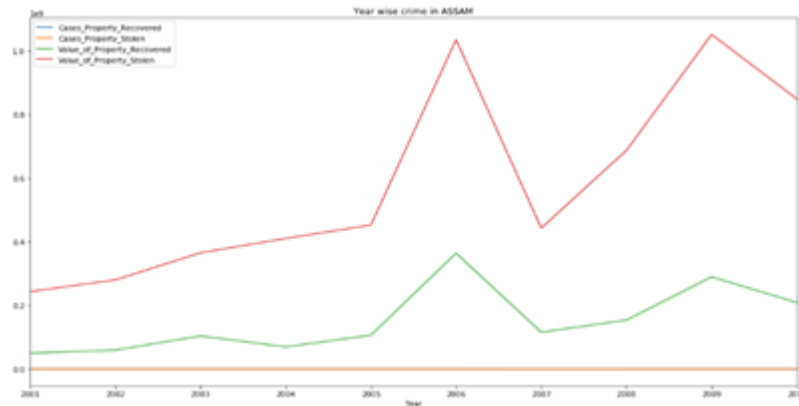
Figure 2: The crime in a particular state (ASSAM) is shown according to yearly basis.

8. 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.

9. 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.

quality of this work.

References

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.

Chatterjee, I. (2021). Artificial intelligence and patentability: review and discussions. *International Journal of Modern Research*, *1*(1), 15-21.

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).

Fayyad, U. M., Djorgovski, S. G., & Weir, N. (1996). Automating the analysis and cataloging of sky surveys. In *Advances in knowledge discovery and data mining* (pp. 471-493).

Gupta, M., Chandra, B., & Gupta, M. P. (2008). Crime data mining for Indian police information system. *Computer society of India*, *40*(1), 388-397.

Gupta, V. K., & Rana, P. S. (2019a). Activity assessment of small drug molecules in estrogen receptor using multilevel prediction model. *IET systems biology*, *13*(3), 147-158.

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.

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.

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.

Kumar, R., & Dhiman, G. (2021). A comparative study of fuzzy optimization through fuzzy number. *International Journal of Modern Research*, *1*(1), 1-14.

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.

Ozkan, K. (2004). Managing data mining at digital crime investigation. *Forensic science international*, *146*, S37-S38.

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.

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.