



Crime Examination and Forecasting Using Machine Learning

Gayatri Mali, Mahima Kakad, Anchal Shukla, Rohit Patil and
Dipamala Chaudhari

EasyChair preprints are intended for rapid
dissemination of research results and are
integrated with the rest of EasyChair.

May 31, 2023

International Conference on Communication and Information Processing
(ICCIP-2023)

Available on: Elsevier-SSRN

Crime Examination and Forecasting using Machine Learning

Gayatri Mali¹, Mahima Kakad², Anchal Shukla³, Rohit Patil⁴, Prof. Dipamala Chaudhari⁵

⁵Assistant Professor, Department of Computer Science and Engineering, Nutan College of Engineering and Research, Talegaon Dabhade, Pune, MH, India

^{1,2,3,4}Student, Department of Computer Science and Engineering, Nutan College of Engineering and Research, Talegaon Dabhade, Pune, MH, India

Abstract

Crimes are a social irritation and cost our society deeply in several ways. Any research that can help in solving crimes quickly will pay for itself. About 10% of the criminals commit about 50% of the crimes. The system is trained by feeding previous year record of crimes taken from legitimate online portal of India listing various crimes such as murder, kidnapping and abduction, dacoits, robbery, burglary, rape and other such crimes. As per data of Indian statistics, which gives data of various crime of past 10 years (2001-2010) a regression model is created and the crime rate for the following years in various states can be predicted. We have used supervised machine learning technique on the crime records for knowledge discovery and to help in increasing the predictive accuracy of the crime. This work will be helpful to the local police stations in crime suppression.

Keywords- SVM (support vector machine), Crime prediction, Regression, Machine learning ©

2023 – Authors.

1. Introduction

Crime rates are rising steadily, and crime patterns are evolving all the time. The behaviors in the crime pattern are so challenging to explain. This essay demonstrates how promoting social development might help reduce crime. The objective is to present a thorough examination of theory and research about the prevention of crime in society and to put into practice various data analysis methods that address the relationships between crime and its pattern. The data for the project are collected from the Kaggle open source name of dataset is property stolen and recovery. Data preparation was carried out once the data were transferred to.csv format.

The following methods were used to process the dataset:

- SVM (support vector machine)

2. Architecture and Working

The information includes several more variables in addition to the number of Area Name, year, and subgroup name and other attributes. In this case, SVM (Support vector machine) techniques are largely used for crime analysis and to discover hidden patterns of crime in India.

| ID | Area_Name | Year | Sub_Group_Name | Auto_Theft_Stolen |
|----|---------------------------|------|---------------------------|-------------------|
| 1 | Andaman & Nicobar Islands | 2001 | 1. Motor Cycles/ Scooters | 4 |
| 2 | Andhra Pradesh | 2001 | 1. Motor Cycles/ Scooters | 2725 |
| 3 | Arunachal Pradesh | 2001 | 1. Motor Cycles/ Scooters | 27 |
| 4 | Assam | 2001 | 1. Motor Cycles/ Scooters | 205 |
| 5 | Bihar | 2001 | 1. Motor Cycles/ Scooters | 946 |
| 6 | Chandigarh | 2001 | 1. Motor Cycles/ Scooters | 281 |
| 7 | Chhattisgarh | 2001 | 1. Motor Cycles/ Scooters | 790 |
| 8 | Dadra & Nagar Haveli | 2001 | 1. Motor Cycles/ Scooters | 12 |
| 9 | Delhi | 2001 | 1. Motor Cycles/ Scooters | 3185 |
| 10 | Goa | 2001 | 1. Motor Cycles/ Scooters | 152 |

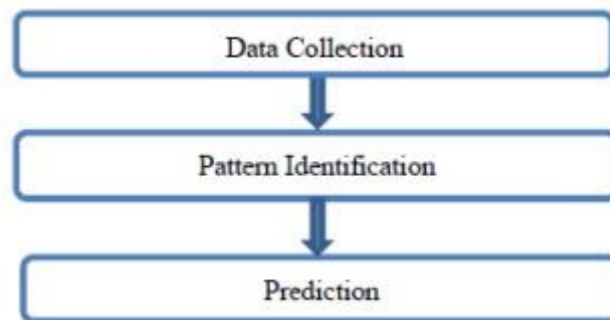
The data mining techniques used are as follows: -

SVM (support vector machine): - A classifier that divides the data set into categories by creating an ideal hyperplane between the variables. I choose this classifier because of how adaptable it is in terms of the variety of kernelling functions that may be used, and since this model can provide a high prediction rate.

3. Proposed Work

It can be divided into 4 parts:

- a. Data processing and cleaning
- b. Random sampling
- c. Train model
- d. Estimate the performance



3.1 *Data processing and cleaning:*

We must prepare the data in the proper format for analysis and cleaning in this phase. One of the methods could be required to convert the variables.

1. Uniformity or gradual normalization
2. Management of values that are missing

3.2 *Random sampling training sample:*

Model will be developed on this sample. 70% or 67% of the data goes here. Test Sample: Model performances will be validated on this sample. 30% or 33% of the data goes here.

3.3 *Train models:*

Verify the algorithm's underlying premises. Create a model and train it using the data that is currently available, then evaluate its performance in terms of error, accuracy, etc.

3.4 *Estimate the performance:*

Utilizing the test sample, evaluate the model's performance in terms of accuracy, precision, and other metrics.

4. Implementation

The initial stage of information gathering involves gathering data from previously accessible/current datasets via web sources. These datasets are combined to create a single dataset that will be the subject of study.

4.1 *Data collection*

The data set that was gathered to forecast crimes is divided into a training set and a test set. Typically, the Training set and Test set are divided into 7:3 ratios. The Support vector classifier (SVC) created data model is applied to the training set, and test set prediction is done based on the test result accuracy.

4.2 *Data Preprocessing*

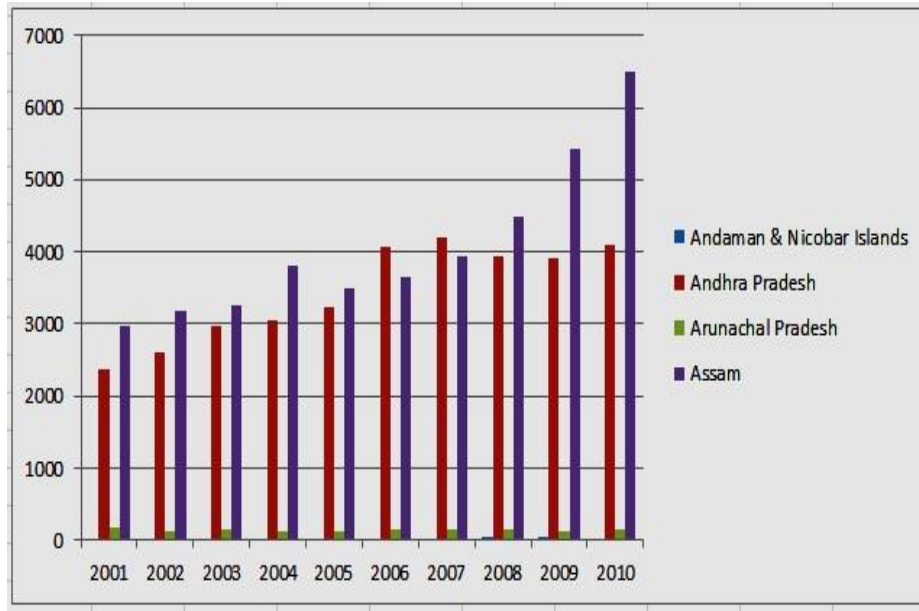
This procedure contains steps to eliminate any null or infinite values that can impair the system's accuracy. Formatting, cleaning, and sampling are the primary procedures. Data that is missing or needs to be fixed can be removed or fixed using the cleaning procedure. When adequate data are utilized in sampling, the algorithm's execution time may be shortened. The preprocessing is carried out in Python.

This study employed Tabular software and the Python module Scikit-learn (sklearn) to preprocess the dataset.

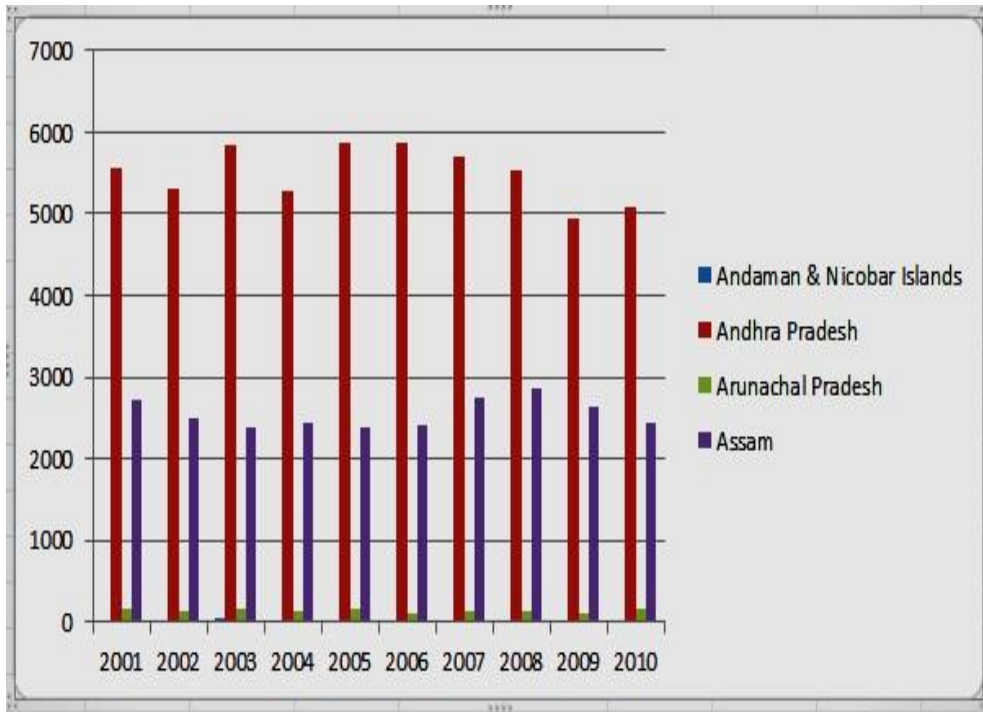
1. The properties in the dataset are divided into two groups: those with text values and those with numeric values. The text characteristics in the dataset for this study must be immediately translated into a numeric value to train the model. The NumPy package for Python is used to do this translation.

2. In our dataset, attributes of the string type include "Year," "Area Name," "Subgroup name" etc. This study assigned numerical values to certain skills using Python.

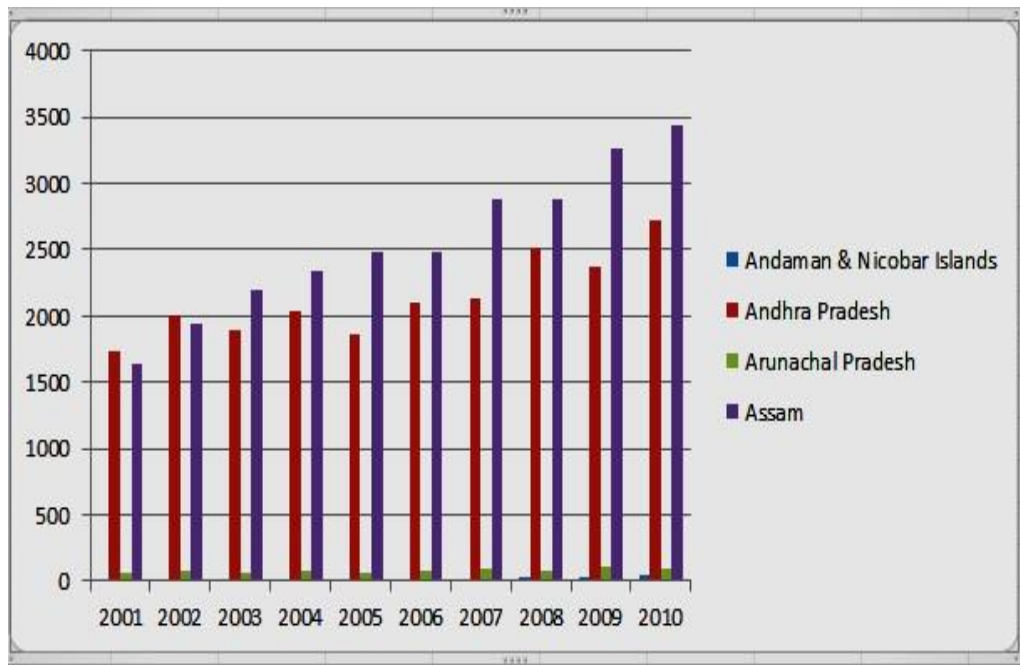
Kidnapping



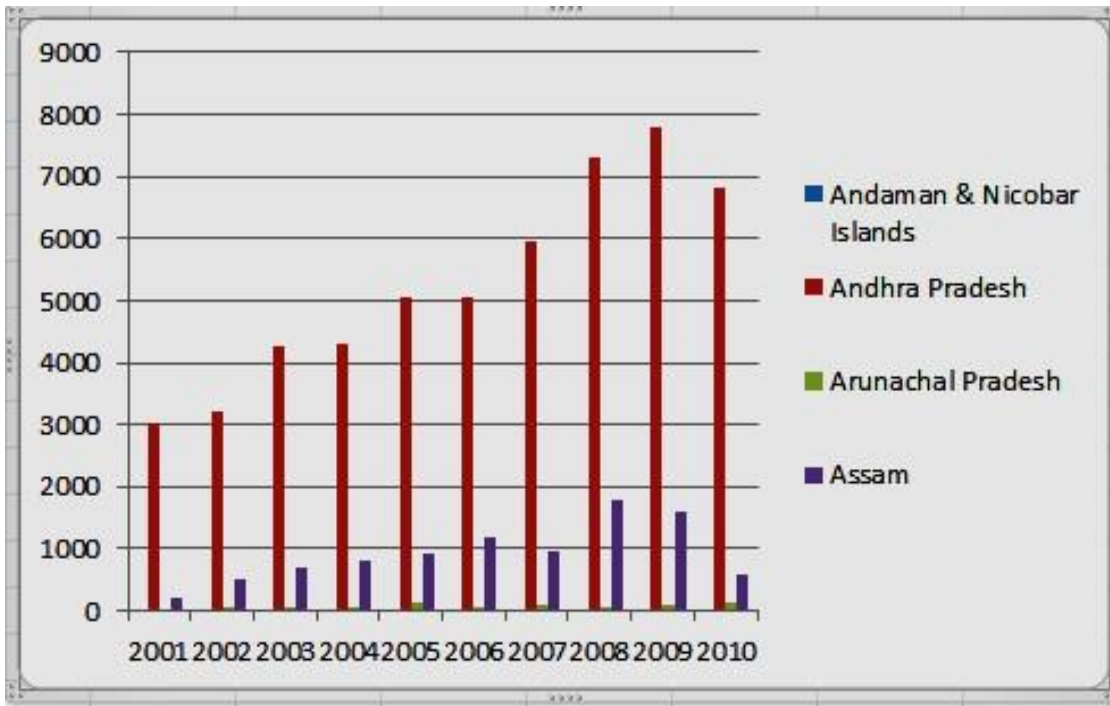
Murder



Rape



Theft



All the above graphs Shows criminal activities happening since 2001 to 2010. The y axis on the graph shows the number of crime rates and x axis on the graph shows years over the time period from 2001 to 2010.

4.3 Feature selection

Features selection is done which can be used to build the model. The attributes used for feature selection are year, Area name ,sub group name ,k a cases reported.

Table 1. Attributes that are used from the datasets

| |
|--------------------|
| Year |
| Area Name |
| Sub group |
| K A cases reported |
| Location |

4.4 Training

This technique randomly splits the dataset into training and test data in the proportions. Then any algorithm is enclosed. Then, we include our training data into this algorithm so that the computer may learn from it. The training portion is now over

4.5 Prediction

A NumPy array's additional feature dimensions as well as the predict function, which accepts an array as input and returns a projected target value as output. So, either 0 or 1 is the expected goal value. By comparing the actual values of the test set with the projected values, it is possible to determine the test score, which is calculated as the ratio of the number of correctly predicted outcomes to the total number of predictions made.

Result

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.96 | 1.00 | 0.98 | 66 |
| 1 | 1.00 | 0.98 | 0.99 | 198 |
| accuracy | | | 0.99 | 264 |
| macro avg | 0.98 | 0.99 | 0.99 | 264 |
| weighted avg | 0.99 | 0.99 | 0.99 | 264 |

Accracy : 98.86363636363636%

Conclusion

A systematic approach to identifying crime is crime analysis and prediction. This system can predict and visualize crime-prone areas by predicting regions with a high probability of crime occurrence. We can extract previously unknown, useful information from unstructured data using the concept of data mining. The accuracy of our project is 98.8%

Future Scope

Our future research direction will focus on:

- 1) Training the model on a larger data set from the deep web to perform a comparative analysis of information embedded in the surface web (Twitter) and darknet forums using SVM.
- 2) Collecting and annotating more data from surface web deep web which will be made publicly available.

Acknowledgements

The authors are grateful to Department of Computer Engineering, Nutan college of Engineering and research, DBatu University for the permission to publish this research.

References

1. N. Kanimozhi, N. V. Keerthana, G. S. Pavithra, G. Ranjitha and S. Yuvarani, "CRIME Type and Occurrence Prediction Using Machine Learning Algorithm," 2021 International journal on Artificial Intelligence and Smart Systems (ICAIS)2021
2. McClendon, Lawrence & Meghanathan, Natarajan. (2015). Using Machine Learning Algorithms to Analyze Crime Data. Machine Learning and Applications: An International Journal. 2. 1-12. 10.5121/mlaj.2015.2101
3. W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117
4. A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve and N. Lal, "Crime Prediction Using K-Nearest Neighboring Algorithm," 2020 International journal on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020
5. Kim, Suhong & Joshi, Param & Kalsi, Parminder & Taheri, Pooya. (2018). Crime Analysis Through Machine Learning
6. Wadhai, Chetan & Kakade, Dnyaneshwari & Bokde, Khushabu & Bhattacharya, Prof. (2018). Crime Detection Technique Using Data Mining and K-Means. International Journal of Engineering Research and. V7. 10.17577/IJERTV7IS020110.
7. Mahmud, Sakib & Nuha, Musfika & Sattar, Abdus. (2021). Crime Rate Prediction Using Machine Learning and Data Mining. 10.1007/978-981-15- 7394-1_5. R
8. Ch. Mahendra, G. Nani Babu, G. Balu Nitin Chandra, A. Avinash, Y. Aditya (2020). Crime Rate Prediction. Journal of Engineering and Science
9. Prof. Shivaprasad More, Sakshi Mench, Saloni Kuge, Hafsa Bagwan (2021) Crime prediction using machine learning approach.
10. Alkesh Bharati, Dr Sarvanaguru RA. K Crime Prediction and Analysis Using Machine Learning, (2018). International Research Journal of Engineering and Technology (IRJET)
11. Ms. Vrushali Pednekar, Ms. Trupti Mahale, Ms. Pratiksha Gadhave, Prof. Arti Gore Crime Rate Prediction using KNN (2018) International Journal IJRITCC
12. 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>
13. 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> Journal of Pharmaceutical Negative Results | Volume 13 | Special Issue 07 | 2022 7594
14. Masita Masila Abdul Jalil Fatihah Mohd Noor Maizura Mohamad Noor A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection (2017)

15. Prasad Dhore, Lalit Wadhwa, Pankaj Shinde, Deepak Naik, Sanjeevkumar Angadi “Proficient exploration of malnourishment with machine learning by CNN procedure”, “Journal of northeastern university”, 0