Journal of Applied Security Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/wasr20

A Study of Lightweight Approaches to Analyze Crime Conditions in India

Bhavna Saini, Dinesh Kumar Saini, Sumit Srivastava & Mayank Aggarwal

To cite this article: Bhavna Saini, Dinesh Kumar Saini, Sumit Srivastava & Mayank Aggarwal (2021): A Study of Lightweight Approaches to Analyze Crime Conditions in India, Journal of Applied Security Research, DOI: 10.1080/19361610.2021.2006031

To link to this article: https://doi.org/10.1080/19361610.2021.2006031

Published online: 02 Dec 2021.

JOURNAL OF APPLIED SECURITY RESEARCH
https://doi.org/10.1080/19361610.2021.2006031

A Study of Lightweight Approaches to Analyze Crime Conditions in India

Bhavna Saini[a], Dinesh Kumar Saini[b]*, Sumit Srivastava[a]*, and Mayank Aggarwal[a]*

[a]Department of Information Technology, Manipal University Jaipur, Jaipur, India; [b]Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur, India

ABSTRACT

Crime prediction is an effort to cut down the crime rate and criminal activities in the society. This paper concentrates on monitoring the crime trends in India using data from 2001 to 2019. The work analyzes the crime condition in India by following a top-to bottom approach and implementing visualization and machine learning algorithms. Heat map visualization of India is done to focus on specific crime types for all states. Furthermore, various demography of human population like age-groups, sex, caste, etc., are considered to have a larger perspective of the issue. The model also suggest proactive measures based on these findings.

KEYWORDS

Crime; data visualization; prediction; classification

Introduction

Crime is an intentional action violating the criminal code imposed by the governing or administering authority, for which an individual or a group of individuals can get punished. Crimes in India are broadly classified into two major categories: Cognizable crimes and Non-Cognizable crimes. A cognizable crime can be investigated directly by police station in-charge without orders of magistrate and are broadly classified into IPC (Indian Penal Code) crimes and SLL (Special and Local Laws) (Bhatnagar, 1990; Mangoli & Tarase, 2009). On the other hand, non-cognizable crimes cannot be investigated without the permission of magistrate. Figure 1 displays the broad classification of crimes in India.

In India, the crimes are so rampant that in about an hour, a total of 187 cognizable IPC (Indian Penal Code) crimes and 443 SLL (Special and Local Laws) crimes get committed (Sharma, 2015). There is an annual increase of 1.6% in the registration of cases (50, 74,635 cases) and more than one-fifth of all registered crime cases (10, 50,945) were classified as violent crimes

CONTACT Mayank Aggarwal mayank2aggarwal@gmail.com Department of Information Technology, Manipal University Jaipur, Jaipur, Rajasthan, India.
*These authors contributed equally to this work.

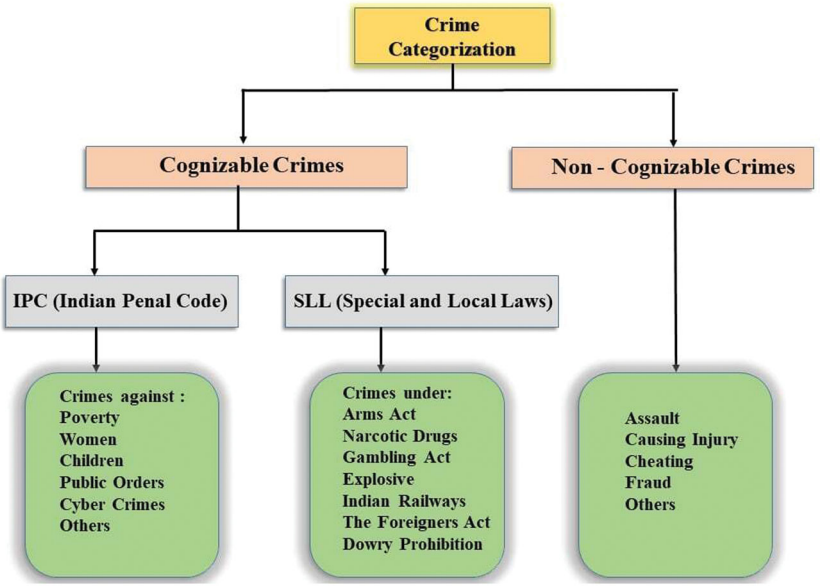

Figure 1. Broad classification of crimes.

(e.g. murder, kidnapping, assault, death by negligence, etc.). The increase in crime rate per 100,000 population has increased from 383.5 in 2018 to 385.5 in 2019. These figures can be reduced if preventive measures are introduced after proper analysis and prediction of crime data. The conventional process of analysis includes the study of crime reports and then discovering unique patterns, series, trends, and gradients through machine learning data mining. Machine learning is a component of artificial intelligence (AI) in which models are trained and tested, so that, the model can learn and improve on its own based on previous experience (Onan, Bal, et al., 2016). Machine learning can prove helpful to find out the minute details using geographical information and other related factors.

One of the most popular research areas where geographical information plays an important role is crime prediction. Early identification of the areas having high risk of criminal events, can help in taking proactive measures. In today's world, security has become the most essential aspect due to the increase in crime rate, especially in India. The primary objective of crime analysis is to estimate the probability of any mishap happening in the country.

Top-to-bottom approach

As per the literature review most of the previous work done in the area of crime analysis and prediction focuses on solving a particular problem. Such

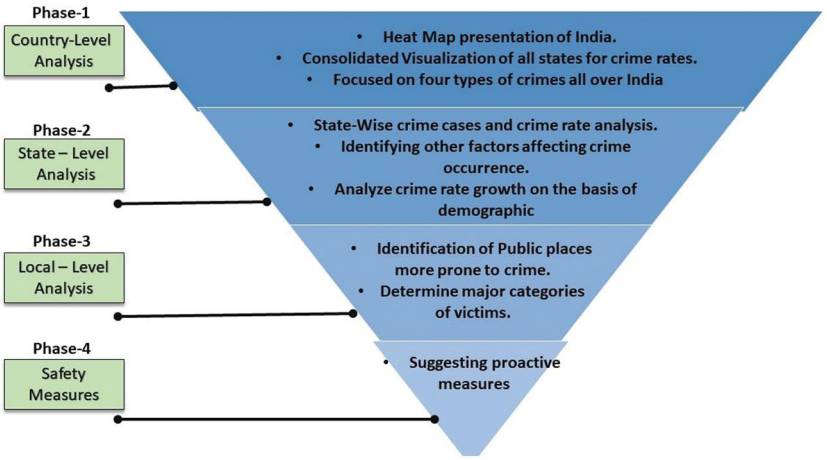

Figure 2. Lightweight top-bottom approach.

as crime category prediction, identifying risky areas, spatial temporal analysis and many more.

In this paper, a top to bottom approach has been followed to have an in-depth crime analysis of India. As shown in Figure 2, this work starts with a broader view of problems i.e. it provides an overview of different types of crime at country level. It analyzes crime conditions of all states/UT of India using a single map visualization. A consolidated view helps to understand the criticality of a particular crime at a country wide level. In the second phase, the problem gets narrowed down to crime rate analysis at state level. Here, the work focuses on figuring out the crime cases and crimes rates state-wise. Further, it identifies the crime rates of a particular state based on its population density. In the third phase, the problem further narrows down to public places which are more prone to crime.

This phase also focuses on identifying the categories of victims and warning them to take precautions beforehand via apps. An early detection of crime can help in taking proactive measures such as increase in policing, CCTV surveillance, etc. at such places. These phases all together form a complete framework to analyze the crime conditions in India and to take preventive measures at state level as well as at local level.

The work proposes a lightweight framework, which helps in identifying the crime types, crime rate, risky areas, and probable victims of different States/UT of India. Moreover, it also works on identifying the other factors related to particular kind of crime and suggesting proactive measures to minimize the problem. To make this framework work in real time with less computations and reliable accuracy, implementation of lightweight machine learning algorithms is done.

The main concern behind this is to allow law enforcement agencies to anticipate the crime rate and its origin as well. In short, this study depicts

the feasibility of applying geospatial methods and data mining techniques to predict crimes using crime data of India from 2001 to 2019. The other sections of this paper are structured as follows: Section "Related work" discusses the related work done in the field of crime prediction and analysis. It also gives a brief overview of the techniques applied previously in the area of spatial crime analysis. The next Section "Methodology" discusses the methodology used in this study along with all the trends and implementations. This section is divided into three sub-parts, which are Data Collection, Data Visualization, and Crime Prediction.

Related work

Machine Learning algorithms have been implemented in various major aspects of crime prediction. This includes the identification of criminals, analyzing hotspots, types of crimes, and many more (Hajela et al., 2020; Wang et al., 2018). A detailed review of crime analysis using data mining techniques was done by Hassani (Hassani et al., 2016). The study discusses mainly five types of taxonomies, and the emphasis was done on mod els that implemented neural networks, support vector machines (SVM), and decision trees. Further, Shamsuddin et al. (2017) reviewed the work done on four crime predictive methods which are fuzzy theory, multivariate time-series, artificial neural networks, and support vector machines (SVM). One more detailed study done by Hardyns and Rummens brings out three criteria for the evaluation of predictive policing (Hardyns & Rummens, 2018). These criteria include the first effect of predictive implementations to actual crime rates, the second, the costs relative to the replaced methods, and the last one was the correctness of the prediction. Researchers have pursued various mining algorithms for crime prediction, such as naive Bayes and decision tree. Sathyadevan (2014), proposed a model to predict crime-prone areas based on a data set of India. In another work, the author focuses on Crime Category prediction using a data set from USA (Iqbal et al., 2013). Yadav proposed a regression model based on naive Bayes, linear regression, and decision tree using crime data of India from 2001 to 2014 (Yadav et al., 2017).

Spatiotemporal crime forecasting tools have received much attention in recent years from academics, private companies, law enforcement, and police departments, as it is an effective visualizing technique. Traditionally, spatial analysis was done for some countries but not for India, and if done then, it's not projected well enough based-on district or city data. Various researchers have worked on low and high-risk crime prediction using the geotagged crime events and point of interest (POI) data. This was done for the four urban areas of the UK, based on the POI layers from

OpenStreetMap (Cichosz, 2020). Another work analyses the spatial relationships between crime occurrences, demographic data, socio-economic, and environmental variables, together with geo-located Twitter messages and their "violent" subsets. The dataset used was collected from Chicago crime data (Ristea et al., 2020). The author has done spatial analysis in India, but of 99 Cities, and the data used contains 14 different types of crimes of year 1971 (Dutt & Venugopal, 1983). Our work is different as Tableau's shapefiles were integrated with the district-wise data of 2014, for a better demonstration of data, which is discussed in the next section.

A crime prediction model based on spatial data was proposed by Bernasco and Elffers, where two spatial outcomes were compared in terms of modeling, spatial movement, and distribution (Bernasco & Elffers, 2010). Various other spatial methods were also implemented in this work such as spatial filtering, weighted regression, and multi-level regression with spatial dependence. Catlett et al. worked on a spatio temporal algorithm for crime prediction in urban areas. The large cities were divided into subparts on the basis of density which further helped in forecasting number of crimes (Catlett et al., 2019). In another approach, the author identified the low population density areas and demonstrated an imbalance aware machine learning application to identify burglary risk areas (Kadar et al., 2019). Rummens et al. focused on spatiotemporal crime forecasting, which proposes a combined predictive model like risk terrain Model, consisting of LR and MLP, therefore concluding crime hotspots that have a risk of more than 20% (Rummens et al., 2017). Further, Araujo et al. proposed a well-defined framework followed by feature-engineering dependent methods (AraúJo et al., 2018). In this work, visualization of crime data is done spatially using a time-series graph along with a forecasting model. Another work of Huang et al. is somewhat different than others as usually spatial mapping is done using the number of crimes or the crime rate, however, this category of crime is fore casted using the binary classification problem (Huang et al., 2018). Zhuang et al. (2017) carried out the crime prediction based on spatial analysis, but it is not using the traditional machine learning algorithms. The author used deep learning, which presents a much more complex internal structure that proved useful. The deep learning architecture used in this work is Long Short-Term Memory (LSTM) architecture, which is compared with the Recurrent Neural Network (RNN), and Gated Recurrent Unit (GRU).

Methodology

For optimum and organized analysis of crimes in India, various visualization techniques and machine learning algorithms have been implemented. Classification of the analysis has been done below in three sub-parts.

Data collection and preprocessing

In this paper, the crime data sets used are confirmed and verified by the NCRB (National Crime Records Bureau), which proves its authenticity and assurance (The National Crime Records Bureau of India Website, n.d.). NCRB is a nodal organization to collect, accumulate and circulate the crime data of India in the form of annual reports. The flow of crime information starts from FIR and moves to District Crime Record Bureau (DCRB). Further this data is collectively moved to State Crime Record Bureau (SCRB) and finally to NCRB.

The data sets used in this work lies in the period of 2001−2019. The study has been done on various parameters such as type of crime, the place of crime occurrence, categories of people affected by crime, and State/UT where the crime happened. In this section, the history of crimes from the year 2001–2019 has been considered. In the pre-processing phase, removal of inconsistent data (such as missing values, redundant information, etc.), joining two or more data sets constructively, and transformation of data as required for the visualization and prediction of crime has been done. Other preprocessing techniques are used for the heat-map, for which the district-wise data is joined with India's geographical shape file to obtain the accurate shapes of all the districts or cities.

Data visualization

Data Visualization is a graphical representation of data using charts, graphs, tables, and maps etc. This technique is imperative as it allows us to see the trends and patterns in the data more clearly and effectively, which results in a better understanding of the data. These data visualization tools and techniques come to use even more when dealing with Big Data to analyze it and make data-driven decisions. In this study of criminal activities, the software used for Data Visualization is Tableau.

While performing crime analysis, there are various factors which needs to be consider such as the place of occurrence (such as Railways, Residential Area, etc.), age and gender of citizens which get targeted the most (such as women, children, senior citizens, etc.) (Alves et al., 2018). This study also works on these factors and aims to identify the areas where the citizens suffer a lot. For this, a geographical visualization of the previous data is done. Data Visualization plays an essential role in this for better demonstration understanding. It also helps in determining the various hidden facts which cannot be interpreted using tabular data.

Heat map visualization

This module uses a district-wise crime data set of the year 2014 and a shape file for all the districts of India. Visualization done on this dataset is based on crime types represented through a heat-map of India. A Heat-Map is a kind of data visualization techniques in which the variation in colors by hue or intensity, depicts obvious visual cues to the reader for better understanding of the affected areas. The classification of crime is in four parts, as shown below. The main thing about the analysis done in this section is that the scale taken for analysis is the same for all four crime types, which is from 0 to 4,000 cases. In previous research works based on crimes in India, the analysis used to be quite basic, in which circles were shown on top of each state and union territory, and that helped to judge the number of cases in that state based on the circle size. To take a different approach, initially district-wise crime data has taken and then the Latitudes and Longitudes for all the districts are hardcoded. The Tableau software needed to know the exact coordinates of each city in the country. Geospatial data containing the shapefiles for all the districts were then used and joined with the pre-processed crime data and, its merging point was by district name and coordinates. Transformation of the dataset like this allowed us to depict the data, giving an outlining for all the districts, which results in allowing us to trace the cases based on the color's darkness.

Personal crimes. It consists of crimes that bring physical or mental harm to another individual which are further classified into two categories, forms of homicide and other violent crimes. Sometimes these personal crimes, where the physical damage to another individual is so severe that it causes death. Hence, the defendant can be charged with homicide (e.g., murder, manslaughter, or homicide using vehicle). Conversely, violent crimes, which are also very severe, include assault, arson, child or domestic abuse, kidnapping, rape, and statutory rape. Heat-map visualization of personal crimes throughout India is shown in Figure 3. The graph shows that such type of cases is very high in areas close to Delhi, Haryana, Rajasthan, Bihar, West Bengal, Maharashtra, Madhya Pradesh, and Kerala, which is approximately more than 4,000 cases. Some of the cities with the highest cases are Delhi (27,359 cases), Murshidabad (13,394 cases), Greater Bombay (12,873 cases), Patna (12,750 cases), South 24 Parganas (11,937 cases), Kolkata (11,578 cases), North 24 Parganas (9,045 cases), Muzaffarpur (8,648 cases), and Pune (8,301 cases).

There are various other factors which are directly or indirectly associated with personal crimes. Literacy rate is among one of those factors, which intensifies the occurrence of such events. Using heat map visualization, areas that are more prone to personal crimes can be identified and actions for improving the literacy rate can be taken.

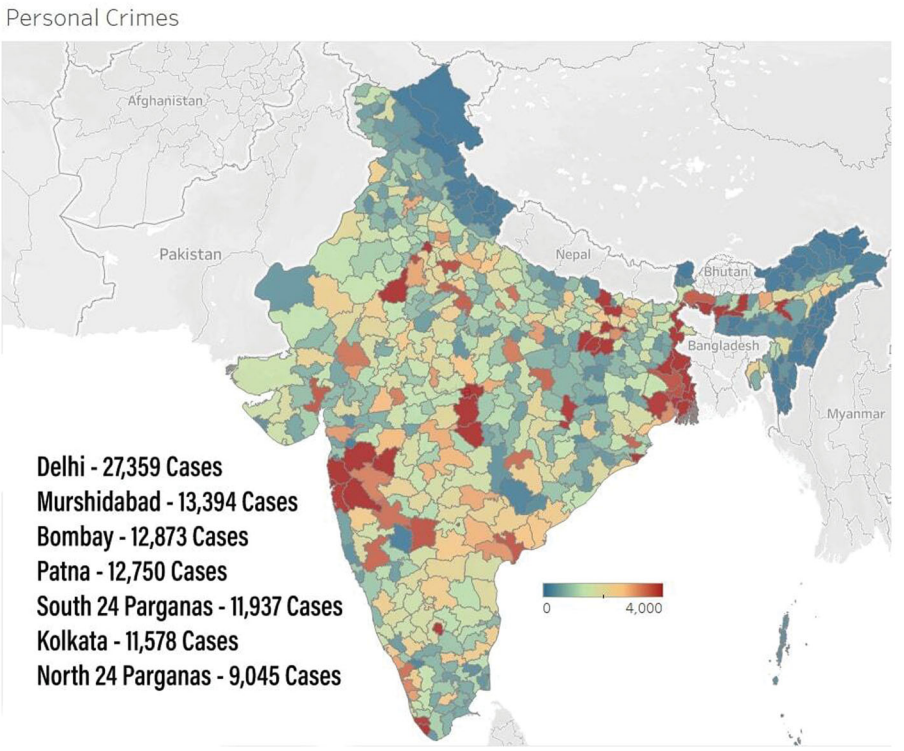

Figure 3. Personal crimes.

Property crimes. Property crimes mean intrusion or trespassing in the property of another with- out any consent of that individual. The main purpose usually is to obtain money, property, or some other benefit. It might involve force, or threat of force if we take robbery or extortion as examples. Property crimes include crimes like arson, burglary, dacoit, larceny, auto theft, and trespassing. Heat-map visualization of property crimes throughout India is shown in Figure 4. The result shows that these criminal cases are very high in areas close to Rajasthan, Haryana, Delhi, Uttar Pradesh, Bihar, Maharashtra, Andhra Pradesh, and Bangalore (Karnataka), which is approximately more than 4,000 cases. It also depicts that criminal activity is slightly on the higher side in North-West India. Some of the cities with the highest cases are Delhi (102,520 cases), Greater Bombay (25,693 cases), Bangalore Urban (17,633 cases), Jaipur (15,353 cases), Pune (13,105 cases), Kolkata (10,061 cases), Indore (9,209 cases), and Thane (9,023 cases).

Price inflation in a state or country exacerbate the property crimes. When rise in prices of goods occur, the purchase power decreases, which further results in various property crimes. Proactive measures at state government level can reduce the same.

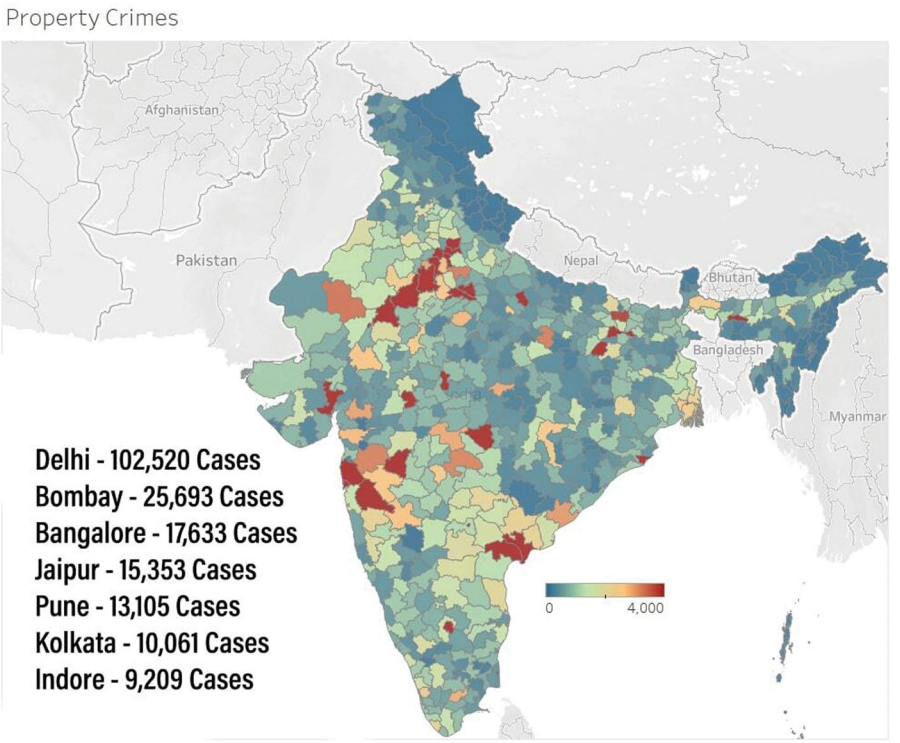

Figure 4. Property crimes.

Statutory crimes. Statutory Crimes include those crimes which are made illegal by-laws passed by a governing body, like the legislature. Three significant types of statutory crimes are alcohol-related crimes, drug crimes, traffic offenses, and financial or white-collar crimes. Statutory crimes are violations of a specific state or federal statutes. These crimes are prohibited by statute because society hopes to deter individuals from engaging in them. Some examples of statutory crimes are juveniles in possession of alcohol, underage driving, selling alcohol to minors, and public intoxication. Heat-map visualization of statutory crimes throughout India is shown in Figure 5.

From this, it can be depicted that the statutory crime hotspots are around Delhi, Gujarat, Maharashtra, Karnataka, Andhra Pradesh, Tamil Nadu, and Kerala, which is approximately more than 3,000 cases. The map-scale in Figure 5 is up to 4,000 which is the same as in Figures 3 and 4 as well, which is done intentionally for better comparison among them. It further depicts that criminal activities are slightly higher toward the South and South-west of India. Some of the cities with the highest cases are Ernakulam (28,360 cases), Thrissur (18,568 cases), Thiruvananthapuram (14,555 cases), Malappuram (12,793 cases), Kottayam (12,542 cases), Delhi (11,307 cases), Chennai (9,779 cases), Kolkata (6,412 cases), and Greater Bombay (6,402 cases).

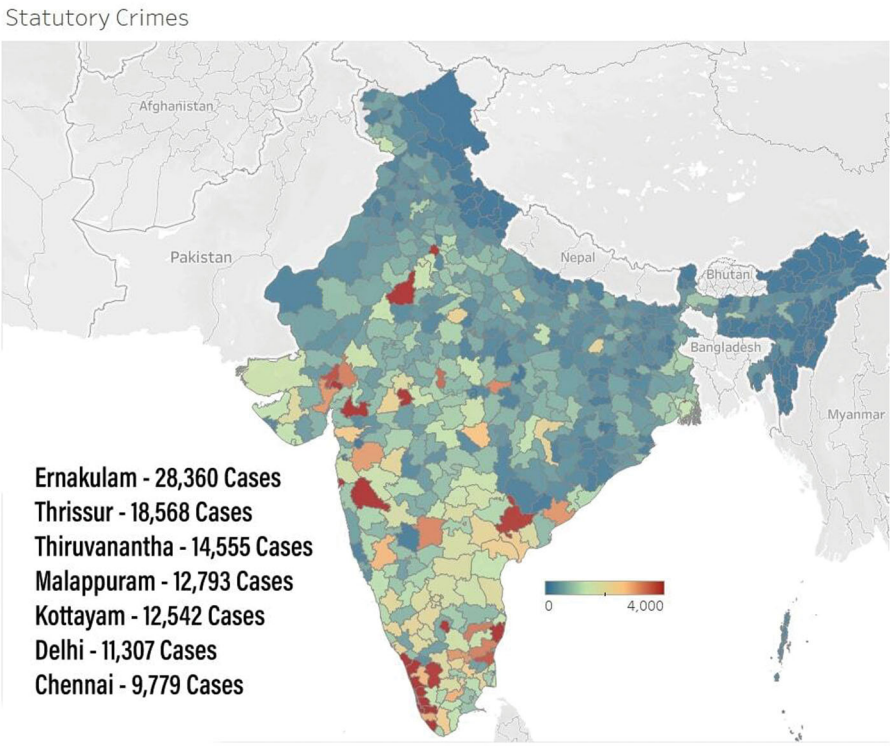

Figure 5. Statutory crimes.

Inchoate crimes. Inchoate crimes, preliminary crimes, or incomplete crimes refer to those crimes that were initiated but not completed and act as an assist to another crime. The most common inchoate offenses include attempt, solicitation, conspiracy, aiding and abetting. It's an inchoate crime if the individual takes a "substantial step" toward the completion of the crime, to be found as guilty. Like if a person is simply intending to or hoping to commit an offense, then it's not considered as inchoate. Punishment for an inchoate crime varies a lot. Some- times it happens to be of same degree as that of the underlying crime or can be a lot less severe too. Heat-map visualization of inchoate crimes throughout India is shown in Figure 6.

The scale used for the following visualization is the same as other heat-map visuals, which is 0 to 4,000 cases. This is done for better reasoning and comparison among other maps. From this visualization, it can be analyzed that inchoate crime activities are very high in areas close to Delhi (capital of India), Rajasthan, Maharashtra, Andhra Pradesh, and West Bengal, which is approximately more than 1,000 cases. Some of the cities with the highest cases are Delhi (14,169 cases), Greater Bombay (4,470 cases), Pune (2,819 cases), Murshidabad (2,687 cases), and Jaipur (2,496 cases).

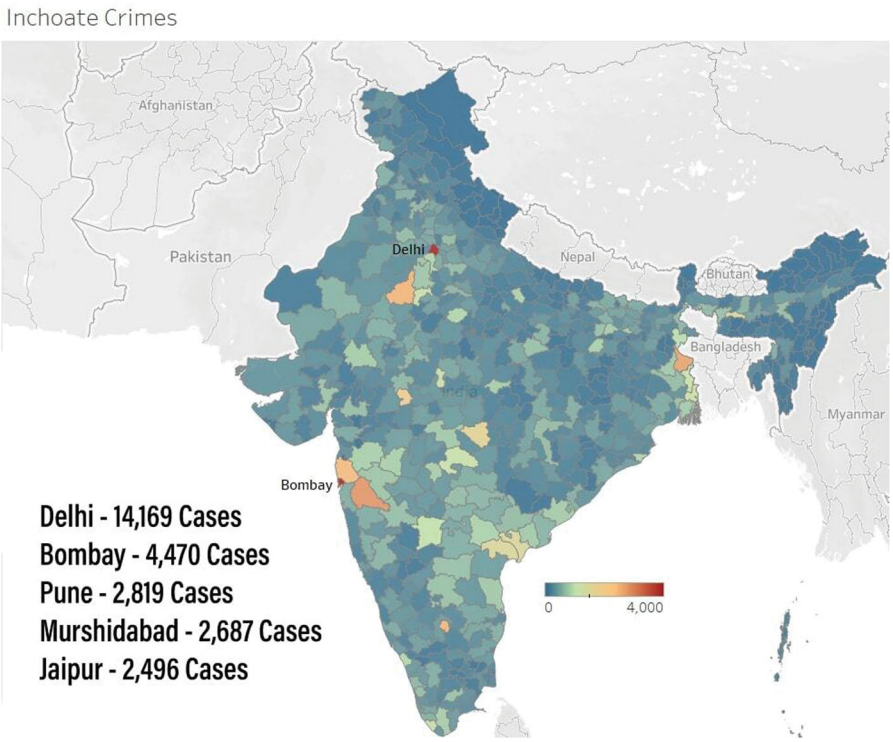

Figure 6. Inchoate crimes.

Crime rate insight through time-series graph

India is facing the upsurge in crime rate each year, and it's not on the verge of dropping even by the slightest. In 2001 the criminal cases were approximately around 17.7 lakhs, and now in the year 2019, it's approximately 51.6 lakhs, which accounts for about 191.53% increase in crime rate. The same is illustrated in the area graph as shown in Figure 7. This time –series area graphs highlight the major two criteria, which are crime rate and total crime cases for all states of India.

The analysis displays that the crime rate is maximum in Delhi, Kerala, Mad- Madhya Pradesh, Tamil Nadu, Haryana, Rajasthan, Andhra Pradesh, and Assam. It's noticeable that the crime rate steeped from 2012 mainly in Delhi, one of the reasons for this can be the drastic increase in population over there.

The crime rate has been calculated by dividing the number of incidences with that of projected population as shown in equation 1.

$$\text{Crime Rate} = \frac{\text{Number of Incidences}}{\text{Projected} - \text{Population}} \tag{1}$$

Now, analysis based on total crime cases from 2001 to 2019 is shown in Figure 7. From this, we can analyze that the criminal activity is the highest

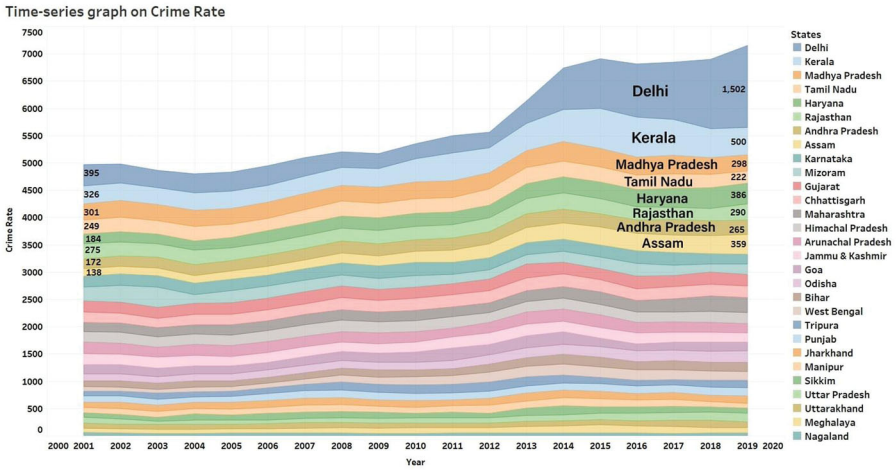

Figure 7. Crime rate visualization in different states of India.

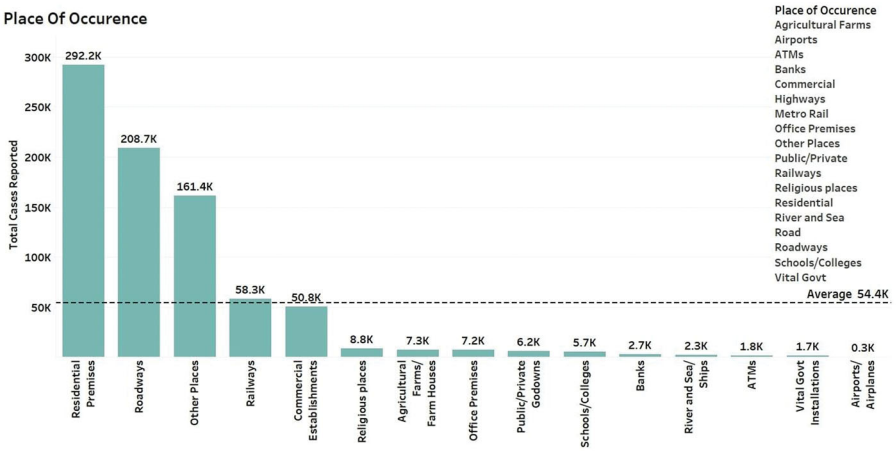

Figure 8. Analysis based on place of occurrence.

in Maharashtra (171.2–341.1 K), Madhya Pradesh (181.7–246.5 K), Uttar Pradesh (178–353.1 K), Andhra Pradesh (130.1–237.6 K), Tamil Nadu (154.8–168.1 K), Rajasthan (155.2–225.3 K), and Kerala (103.8–175.8 K), account to about more than 50% of the total criminal cases shown in Figure 8.

Crimes based on place of occurrence

This section concentrates on determining the areas where most of the crimes got executed and can occur in future also. The analysis can help the law enforcement as well as individuals to be more cautious and aware. This section mainly focuses on property crimes which can further be categorized as dacoit, theft, robbery, burglary, and other offenses in which property is

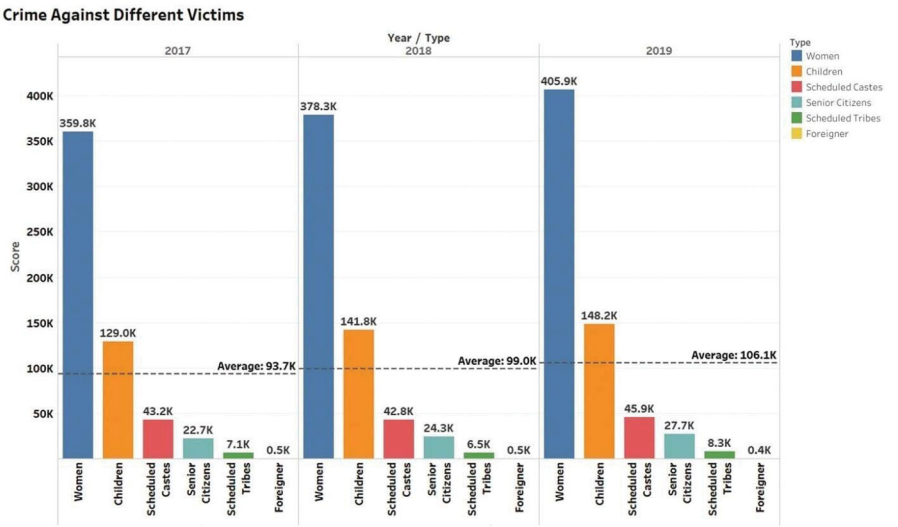

Figure 9. Analysis based on different kinds of victims.

lost. Heat map on these categories helps in concentrating on some specific crime types. The bar chart shown in Figure 8 is a kind of an extension to this, which makes people and the concerned authorities know about the kind of places to focus on, such as Residential Premises, Roadways, Railways, etc.

Victim based analysis

This section is crucial as it helps to understand the situation of citizens or non-citizens of India, which are mostly targeted or victimized by criminals. This work is using year 2019 data and is mainly focusing on Women, Chil dren, Senior Citizens, Scheduled Castes, Scheduled Tribes, and Foreigners. This analysis shown in Figure 9 has been done for three consecutive years (2017, 2018, 2019). The graph shows that the average these people. It is also noticeable that the highest number of cases is toward women, and it still has increased quite enough in three years only, which is about 12.8% to be precise. The second-highest number of crimes are toward children, which require the most protection as they are the most vulnerable people amongst the lists.

Crime analysis

This section concentrates on supervised learning through Classification, Clustering, and Regression of data. This exploration and extraction will be done using Data Mining Machine Learning algorithms as it will help us find unique trends or patterns in the data which is not easily noticeable

through Data Visualization alone. Python is the programming language used for this purpose and to build and train some predictive models. These libraries are Sci-Kit Learn, Pandas, NumPy, and Sci-Kit Fuzzy. As the aim of this work is to have a lightweight framework, the machine learning algorithms used are random forest, decision tree, and naive Bayes. Further, the algorithm with the highest performance has been referred. The highest performance in this aspect refers to a high accuracy rate, precision score, F-measure, and recall score. The data set used for all three is the same for better and accurate comparison. The process is divided below into two sections, first sub-part is for Fuzzy C-means clustering in which everything from the theory to the implementation is shown. The second sub-part is of all three classification techniques utilized in this work. They are grouped as the comparison is being done among the three to find out the most effective one. Fuzzy C-means and Classification are separate as they are not related to one another and comparison are not possible among them.

Fuzzy C-means clustering algorithm

Fuzzy C-means (FCM) is a type of Supervised Clustering algorithm, for which knowing about clustering approaches is necessary. It is the segregation of data points into several partitions, based on characteristics and attributes of the data points, so that similar kind of data points are in the same cluster. The objective of these approaches is to isolate the data points and assign them to a cluster. There are three types of clustering, which are hard, soft, and overlapping (Yamini, 2019).

- Hard Clustering—Every data object can belong to only one cluster.
- Soft Clustering or Fuzzy Clustering—Every data object can belong to two

or more clusters, but to a certain degree.

- Overlapping Clustering or Multi-View Clustering.

Every data object belongs to more than one cluster, usually contains hard clusters. In most of the research works, K-means clustering is used, as it is a hard clustering method. In this study, the datasets used are overlapped, hence FCM is adopted. FCM comes under the category of Soft Clustering, which means that the data points in can belong to two or more clusters as well. This algorithm is developed by Dunn (1973) and improved by Bezdek et al. (1984). It is also known as soft K-means as the main difference among these two is that in K-means is a hard-clustering type algorithm whereas FCM is of soft. This algorithm works by assigning each data

Table 1. Details of dataset attributes for classification.

Attributes	Description
Region	8 Regions of Indian State/UT
State/UT	28 States and 7 Union Territories
Population	Mid-year projected population (in lakhs)
Density	State-wise density count based on census
Crime type	Types of crimes occurred such as murder, riots
Crime cases	Registered criminal cases or criminal incidences
Crime rate	Crime cases per population in lakhs

object membership corresponding to each cluster centroid based on the Euclidean Distance between them. After each iteration, the membership of each data object is updated based on the minimization formula shown below.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} (\mu_{ij})^m \ ||x_i - c_j||^2 \tag{2}$$

where "m" is the fuzziness index which is greater than 1, "N" is the number of data points, "C" is the number of centroids, "ij" represents the membership of ith data to jth cluster centroid, "x_i" is the ith of d-dimensional measured data, "c_i" is the d-dimension center of the cluster, and "$||x_i\ c_i||^2$" is the Euclidean Distance between ith data point and jth cluster center.

The steps of fuzzy C partitioning is shown in Algorithm 1. To partition the clusters there is a certain metric used, which is the Fuzzy Partition Coefficient (FPC), and it tells us that how cleanly our data is described by a certain model. The FPC is defined on the range from 0 to 1, with one being the best. The higher is the FPC value, the cleaner becomes the partitioning of Clusters or Centroids. The data set used for FCM Clustering is of Violent Crimes in India 2019, and the data set is of around 250 tuples (or rows). The attributes or parameters are shown in Table 1, which are used in the data set, among which the algorithm is implemented with the parameter State/UT-Wise mid-year projected population (in lakhs) on the x-axis and Crime Rate on the y-axis (Figure 10).

Attributes such as State/UT, and Crime Type are in the form of string, and to make the algorithm work values in these columns or fields are factorized and then the refactored data set is added to the C-means function provided by the Sci-Kit Fuzzy Python Library, which returns the FPC (Fuzzy Partitioning Coefficient), centers of the clusters, and the cluster membership array, through which we plot a scatter graph using the Matplotlib Library, which is shown in Figure 11. The plot is shown only of the chart in which the FPC was the highest, to calculate that the FPC is calculated for each value ranging from 2 to 9 of the centroids or the clusters. The range starts from 2 centroids as it cannot be starting from 1 as in

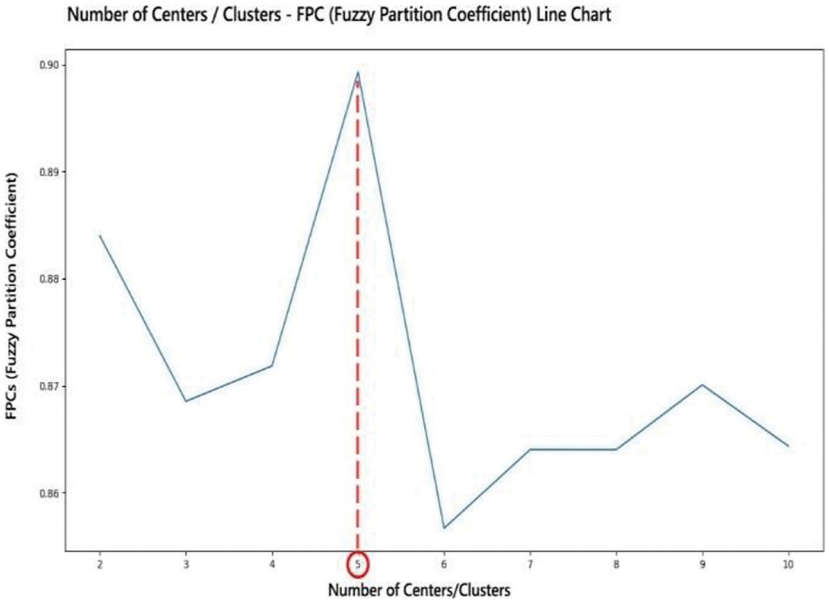

Figure 10. Fuzzy partitioning coefficient for population—crime rate chart.

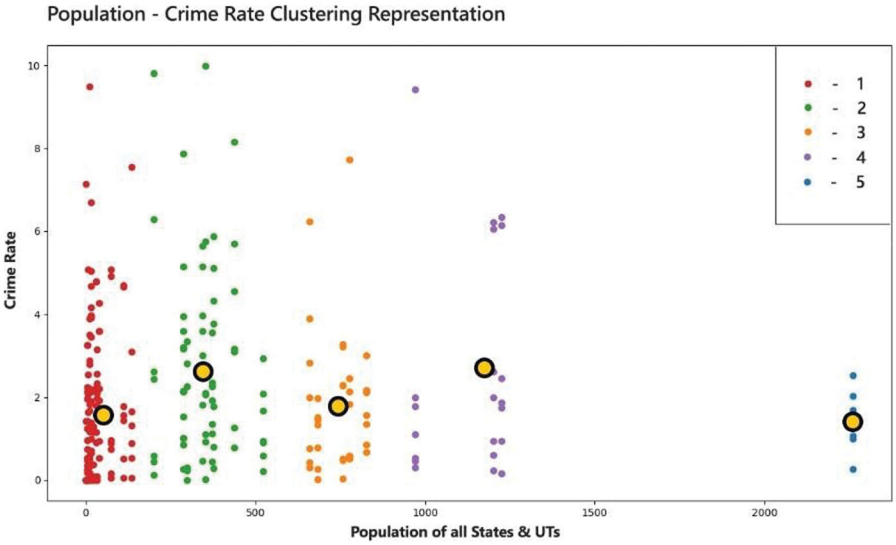

Figure 11. Population—crime rate Fuzzy C-means clustering.

the case of 1 the FPC value would always be at 1 which is the highest. The result of which is shown in Figure 10 through a line graph for better analysis of the highest value. There is a spike in FPC when there are 5 clusters, which also means that the data points were cleanly partitioned when there were 5 clusters or centroids. That is why we have shown the representation of data points with 5 clusters formed in Figure 10.

By carefully looking at Figure 11, it can be stated that the first four clusters (the red one, the green one, the orange one, and the purple one) are in the less than 1,300 population part and, there is just one cluster (the blue one) toward the right. In fact, in the blue cluster, there is just one State or Union Territory, and that is Uttar Pradesh, which is the most highly populated state of India in 2019 with the value of 23.79 crores, and that is one of the reasons its crime rate is up to 3.00. Among the left 4 clusters, the red cluster is between 0–198 population (in lakhs) and crime rate up till 9.73. This cluster includes about 17 of the States and Union Territories in it. After this is the green cluster ranging from 199 to 523 population (in lakhs) and crime up till 10.00, and this cluster contains 10 States and Union Territories. Next comes the orange cluster ranging from 659 to 826 in population and the crime rate is up to 7.92, and this result is for about five States. The last one is the purple cluster which depicts the remaining three three States (West Bengal, Maharashtra, and Bihar) whose population is between 971 and 1,225 in lakhs, and the crime rate is up to 9.4 with West Bengal having the highest for crime type Attempt to commit Murder.

Classification models

In Data Mining and Machine Learning, classification refers to a predictive model where a class label or target label is assigned, which is to be achieved by a given set of input data (Onan, Korukoğlu, et al., 2016). At first, the model is trained using the given data, and then the data for which prediction must be made is tested (Onan, 2017). In this research, the model is created by using a part of data for training and the rest for prediction, and as we have the desired target values for the rest of the data set, using which we can calculate some parameters which help verify the performance of the model. These parameters are listed below:

- Confusion Matrix

Confusion Matrix (also known as Error Matrix) is a kind of table that helps in better judgment and visualization of the performance of a Data Mining Algorithm, usually the algorithm is of supervised learning (Visa, et al., 2011).

- Accuracy Score

After the resultant or the predicted value is calculated through the respective Data Mining algorithm, comparison is done based on the closeness between the predicted value and the targeted value which we keep just

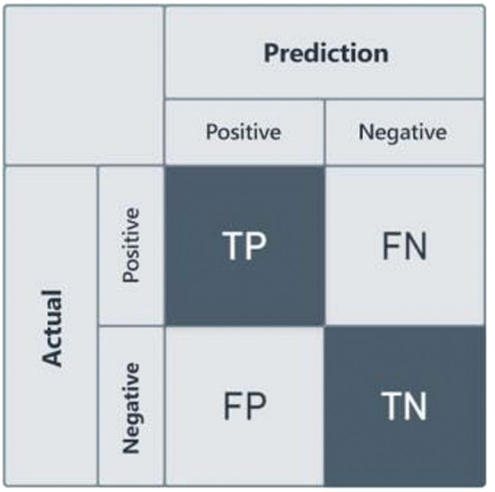

Figure 12. Confusion matrix.

to check out the score (Onan & Toçoğlu, 2021; Yerpude, 2020). The score is given in percentage.

- Precision Score

$$\text{Accuracy Score} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}} \tag{3}$$

Precision Score is another metric used to check the efficacy and performance of the algorithm (Figure 12). It is a good measure to determine when the values of False Positive are high. For instance, in email spam detection, a False Positive means that a non-spam email (Actual Negative) is identified as spam (Predicted spam).

- Recall Score

$$\text{Precision Score} = \frac{\text{True Positive}}{\text{True Positive} + \text{True Positive}} \tag{4}$$

Recall Score is also a metric used to check the efficiency and performance of the algorithm. It calculates how many of the Actual Positives our model captures through labeling it as Positive (True Positive).

- F1 Score

$$\text{Precision Score} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{5}$$

F1-Score or F1-Measure is another accuracy testing metric which depends on the values of Precision Score and Recall Score both. F1 Score might be a better metric if you seek a balance between the precision and recall score.

$$\text{F1 Score} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{6}$$

This research has taken three Classification Supervised Learning algorithms, among which the best algorithm for this particular purpose of crime prediction will be concluded. This comparison of accuracy and performance will be done based on the accuracy metrics which we just talked about earlier in this section. The three algorithms are:

Naive Bayes algorithm. This algorithm is based upon the Bayes Theorem (Joyce, 2003), in which he describes the probability of an event, based on prior knowledge of conditions that might be related to it. Naive Bayes Classification is a supervised learning classifier that returns a set of classes, instead of a single output. The classification is thus given by the probability that an object belongs to a class. This approach is mainly used for *i*.

Decision tree algorithm. It is another Supervised Classification Algorithm that uses root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. This algorithm was discovered to predict the target column, after splitting the data set into random training and test sets.

Random forest algorithm. Random forest is also known as the more accurate version of a decision tree as it takes multiple trees (decision trees) into account and produces the mean result which is useful in balancing the biased data (Aytuğ, 2018). Each Decision Tree in it individually classifies the data set and then the algorithm chooses the classification commonly chosen by the greatest number of individual trees.

The data set used for Supervised Classification is of Crime in India 2019 and consists of around 500 tuples of data. Among which 75% of the data (around 384 tuples), which is randomly sorted is used just to train the model, the rest 25% of the data (around 120 tuples) is used for prediction and then calculating the accuracy metrics (Accuracy Score, F1 Score, Recall Score, Precision Score, and Confusion Matrix), for which the functions are already provided by the Sci- Kit Learn Machine Learning Python Library. All the attributes or parameters are shown in Table 1. Amongst the table shown above the target field or class label is Region, the rest of it are the

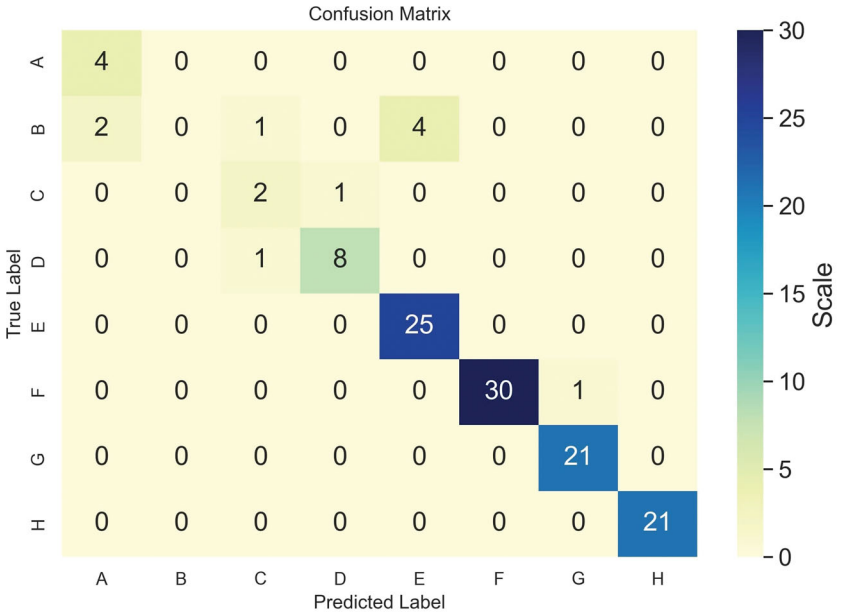

Figure 13. Confusion matrix for random forest.

attributes or the given data. In the data set, the number of tuples and the attributes are the same for all three classification algorithms as it will make the comparison process convenient and smoother. It will even help in reaching the conclusion faster.

The algorithms in which this data set is tested are: • Naive Bayes • Decision Tree • Random Forest

Below are the results of all three algorithms for 4 performance metrics.

From this, we can easily analyze that the least accurate Classifier based on these attributes and data is naive Bayes, and then it is decision tree and at the last it's Random Forest Classifier with the highest accuracy, precision, recall and F1 Score. For Naive Bayes, GaussianNB was implemented and for Precision Score, Recall Score, and F1 Score average was "weighted," same as in other algorithms (Onan & Korukoğlu, 2017). Confusion Matrix for Random Forest Classifier is shown in Figure 13 as its Classification was the most accurate in getting the Regions of India, and it will help verify the values.

The matrix is 9×9 as the States and Union Territories are classified into 9 regions. This result of random forest was formed by the decision tree which came out from the average of 5 different approaches of Decision Trees. The final tree is shown in Figure 14. The terminologies used in this are:

Entropy – It is a measure of the randomness in the information being processed. Samples – A set of inputs paired with a label, which is the correct output (also known as the Training Set). Values – It is the values of all the variables

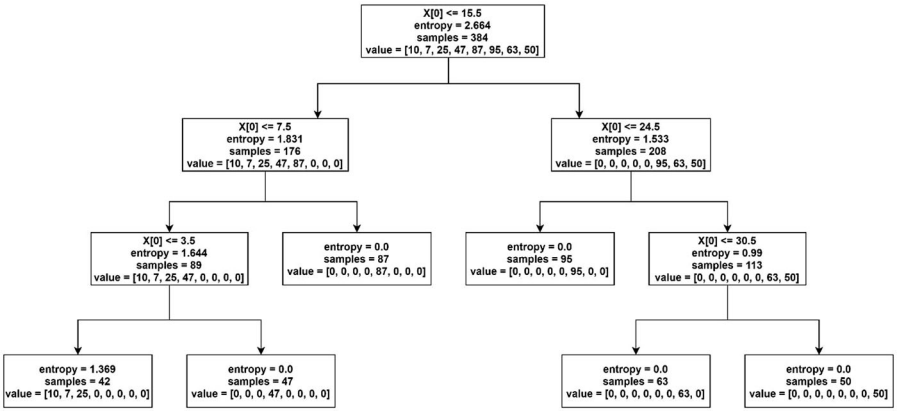

Figure 14. Classification using random forest.

The *i*-th element of each array holds information about the node *i*. Node 0 is the tree's root. Some of the arrays only apply to either leaf or split nodes. In this case, the values of the nodes of the other type are arbitrary. For example, the arrays feature, and threshold only apply to split nodes. The values for leaf nodes in these arrays are therefore arbitrary. These trees are built in a top-down approach. As you can see in the tree Figure 14 that on each branch- division the values array gets split up too and then the value of the sample is the total sum of values array, which also changes with respect to the change in the values array. Entropy plays a major role here; it is best described as the measure of disorder or uncertainty and the goal of machine learning models and Data Scientists, in general, is to reduce uncertainty. Its Mathematical formula is:

$$E(S) = \sum_{i=1}^{C} -p_i \log_2 p_i \tag{7}$$

where "P_i" is simply the frequency probability of an element/class "I" in our data. For simplicity's sake let's say we only have two classes, a positive class, and a negative class. Therefore "I" here could be either + or (−). So if we had a total of 100 data points in our dataset with 30 belonging to the positive class and 70 belonging to the negative class then "P+" would be 3/10 and "P−" would be 7/10. The target entropy is as close to 1 as possible which means it is at the maximum disorder at that point, so, in the Tree formed we can see that the entropy comes out to be 1.369 in the leaf node which is the closest value possible in this scenario.

The classification algorithms used in these models are decision tree, naive Bayes, and random forest. After training using 75% of the data and then testing on the rest, it can be easily observed that the best results came out in Random Forest Classifier with an accuracy of 91.67% as shown in Table 2. The use case in this study is of India, but that doesn't restrict the

Table 2. Performance of various classification algorithms for crime prediction.

Classifier	Accuracy score (%)	Precision score (%)	Recall score (%)	F1 score (%)
Naive Bayes	85.41	84.73	85.41	85.05
Decision tree	88.63	86.76	88.63	87.60
Random forest	91.67	89.21	91.67	89.21

researchers to use the built model and then train it for any other respective countries. To achieve this for the Classification section will require slight data pre-processing, in that seven major parameters are necessary—Region, States, Population in thou sands, District, Crime Cases, and Crime Type. In this, the targeted class is Region which means that the sections of a country help segregate the States based on their region. In most of the datasets, this data isn't provided, for that joining of two datasets, one of regions and states of a country, and another would be the state-wise data consisting of the remaining 6 classes. This will get your data ready to insert in the model and train and test it accordingly. For the clustering part also, a slight adjustment in the data is required, in which a dataset of State and Crime Type Wise is required of any country which consists of a population of each state and Crime Rate for all the Crime Types in each State. Just after doing all this will get the model ready to execute for other countries.

Conclusion

This paper concentrates on monitoring the trends in India using data from 2001 to 2019 and then identifying the best possible methods to analyze the region of a crime based on the parameters such as type of crime, crime cases, district, etc. This work proposes a lightweight top to bottom approach which analyses the crimes, state-wise and further narrows down it to local level. Initially a geo-spatial representation is done district-wise, based on four major classifications of crime (Personal, Property, Inchoate and Statutory). Next, some in-depth analysis using state-wise time series analysis on crime is done. It looks for public locations prone to crime in India and then analyze various demography of human population, like age-groups, sex, caste, etc., which makes it different in terms of vast approaches considered to provide a larger perspective of the issue with a crisp and on point explanation. In the next phase, FCM clustering is implemented on a dataset with state-wise entries for some specific known crimes across the country. This algorithm has been used in very few research works in general, and in studies related to criminal activities, the count is even lesser. Doing all this visualization and clustering helps understand the issue well and helps identify different trends which weren't looked at before. The classification algorithms used in these models are decision tree, naive Bayes, and random forest. After training using 75% of the data and then testing

on the rest, it can be easily observed that the best results came out in Random Forest Classifier with an accuracy of 91.67%. This model will help the police as well as law administrators to understand the crime scenario and take proper safety measures on time.

References

Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and Its Applications*, *505*, 435–443. https://doi.org/10.1016/j.physa.2018.03.084

AraúJo, A., Cacho, N., Bezerra, L., Vieira, C., & Borges, J. (2018, June). Towards a crime hotspot detection framework for patrol planning. In *2018 IEEE 20th International Conference on High Performance Comput ing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (pp. 1256–1263). IEEE.

Aytuğ, O. N. A. N. (2018). Sentiment analysis on Twitter based on ensemble of psychological and linguistic feature sets. *Balkan Journal of Electrical and Computer Engineering*, *6*(2), 69–77.

Bernasco, W., & Elffers, H. (2010). Statistical analysis of spatial crime data. In *Handbook of quantitative criminology* (pp. 699–724). Springer.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bhatnagar, R. R. (1990). *Crimes in India: Problems and policy*. Ashish Publishing House.

Catlett, C., Cesario, E., Talia, D., & Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data- driven approach and experiments. *Pervasive and Mobile Computing*, *53*, 62–74. https://doi.org/10.1016/j.pmcj.2019.01.003

Cichosz, P. (2020). Urban crime risk prediction using point of interest data. *ISPRS International Journal of Geo-Information*, *9*(7), 459. https://doi.org/10.3390/ijgi9070459

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57.

Dutt, A. K., & Venugopal, G. (1983). Spatial patterns of crime among Indian cities. *Geoforum*, *14*(2), 223–233. https://doi.org/10.1016/0016-7185(83)90020-9

Hajela, G., Chawla, M., & Rasool, A. (2020). A clustering based hotspot identification approach for crime prediction. *Procedia Computer Science*, *167*, 1462–1470. https://doi.org/10.1016/j.procs.2020.03.357

Hardyns, W., & Rummens, A. (2018). Predictive policing as a new tool for law enforcement? Recent developments and challenges. *European Journal on Criminal Policy and Research*, *24*(3), 201–218. https://doi.org/10.1007/s10610-017-9361-2

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(3), 139–154. https://doi.org/10.1002/sam.11312

Huang, C., Zhang, J., Zheng, Y., & Chawla, N. V. (2018, October). Deep-Crime: attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1423–1432).

Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanah- Madliravi, N. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, *6*(3), 1–4225. https://doi.org/10.17485/ijst/2013/v6i3.6

Joyce, J. (2003). Bayes' theorem. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019 ed.). Metaphysics Research Lab, Stanford University.

Kadar, C., Maculan, R., & Feuerriegel, S. (2019). Public decision support for low population density areas: An imbalance-aware hyper-ensemble for spatio-temporal crime prediction. *Decision Support Systems*, *119*, 107–117. https://doi.org/10.1016/j.dss.2019.03.001

Mangoli, R. N., & Tarase, G. N. (2009). Crime against women in India: A statistical review. *International Journal of Criminology and Sociological Theory*, *2*(2), 292–302.

Onan, A., Bal, V., & Yanar Bayam, B. (2016). The use of data mining for strategic management: A case study on mining association rules in student information system. *Croatian Journal of Education - Hrvatski Časopis za Odgoj i Obrazovanje*, *18*(1), 41–70. https://doi.org/10.15516/cje.v18i1.1471

Onan, A. (2017). Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*, *46*(2), 330–348. https://doi.org/10.1108/K-10-2016-0300

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25–38. https://doi.org/10.1177/0165551515613226

Onan, A., & Toçoğlu, M. A. (2021). Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts. *Computer Applications in Engineering Education*, *29*(4), 675–689. https://doi.org/10.1002/cae.22252

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232–247. https://doi.org/10.1016/j.eswa.2016.03.045

Ristea, A., Al Boni, M., Resch, B., Gerber, M. S., & Leitner, M. (2020). Spatial crime distribution and prediction for sporting events using social media. *International Journal of Geographical Information Science*, *34*(9), 1708–1732. https://doi.org/10.1080/13658816.2020.1719495

Rummens, A., Hardyns, W., & Pauwels, L. (2017). The use of predictive analysis in spatio-temporal crime forecasting: Building and testing a model in an urban context. *Applied Geography*, *86*, 255–261. https://doi.org/10.1016/j.apgeog.2017.06.011

Sathyadevan, S. (2014, August). Crime analysis and prediction using data mining. In *2014 First International Conference on Networks Soft Computing (ICNSC2014)* (pp. 406–412). IEEE.

Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017, May). An overview on crime prediction methods. In *2017 6th ICT International Student Project Conference (ICT-ISPC)* (pp. 1–5). IEEE.

Sharma, S. (2015). Caste-based crimes and economic status: Evidence from India. *Journal of Comparative Economics*, *43*(1), 204–226. https://doi.org/10.1016/j.jce.2014.10.005

The National Crime Records Bureau of India Website. n.d. https://ncrb.gov.in/en

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *MAICS*, *710*, 120–127.

Wang, S., Wang, X., Ye, P., Yuan, Y., Liu, S., & Wang, F. Y. (2018). Parallel crime scene analysis based on ACP approach. *IEEE Transactions on Computational Social Systems*, *5*(1), 244–255. https://doi.org/10.1109/TCSS.2017.2782008

Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April). Crime pattern detection, analysis prediction. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)* (Vol. 1, pp. 225–230). IEEE. https://doi.org/10.1109/ICECA.2017.8203676

Yamini, M. P. C. (2019). A violent crime analysis using fuzzy c-means clustering approach. *ICTACT Journal on Soft Computing*, *9*(3), 1939–1944.

Yerpude, P. (2020). Predictive modelling of crime data set using data mining. *International Journal of Data Mining Knowledge Management Process (IJDKP)*, *7*, 43–58.

Zhuang, Y., Almeida, M., Morabito, M., & Ding, W. (2017, August). Crime hot spot forecasting: A recurrent model with spatial and temporal information. In *2017 IEEE International Conference on Big Knowledge* (ICBK) (pp. 143–150). IEEE. https://doi.org/10.1109/ICBK.2017.3