Crime

by Mayan Aagarwal

Submission date: 01-Feb-2021 11:08AM (UTC+0530)
Submission ID: 1498819470
File name: Research_Paper_2_2.docx (2.63M)
Word count: 7561
Character count: 39512

Predictive Analysis of Crimes in Developing World

Abstract

Crime is a violation of laws set forth by the state to maintain social order and stability in society. Due to the substantial increase in the crime rate, the application of data mining techniques helps in achieving unique crime patterns and trends, which will further help law enforcement to come up with proper crime prevention strategies. This paper concentrates on Data Visualization and Prediction, using Data Mining and Machine Learning Algorithms with the help of Tableau and Python. This present work focuses on crime records of India (2001 - 2019) for the people who are victimized the most (like women, children, senior citizens, etc.). The crime trends are analyzed district-wise throughout the country with the help of visualization. Analyzing the crime would be much easier by using the predictive models developed in this work, which can be used in the future as well. The Data Mining algorithms implemented are Fuzzy C-means Clustering, Random Forest Classifier, Naïve Bayes Gaussian Classifier, and Decision Tree Classifier, which shows its accuracy and efficacy through various accuracy metrics.

Keywords

Crime, Data Visualization, Prediction, Classification

1. Introductions:

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)[1][2]. On the other hand, non-cognizable crimes cannot be investigated without permission of magistrate. Figure 1 displays the broad classification of crimes in India.

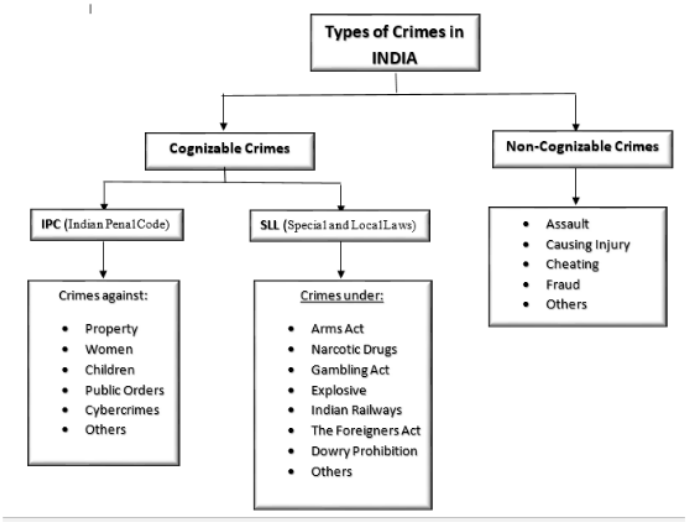

Fig.1. – Crimes in India

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 [3]. There is an annual increase of 1.6% in the registration of cases (50,74,635 cases) and the increase in crime rate per 100,000 population has increased from 383.5 in 2018 to 385.5 in 2019. More than one-fifth of all registered crime cases (10, 50,945) were classified as violent crimes (e.g. – murder, kidnapping, assault, death by negligence, etc.). 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 inclinations 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. In this paper Supervised Machine Learning Algorithms are used which is a sub-part of it and the other is Unsupervised one [4]. 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. This study includes three primary objectives, which are:

- To visualize and analyse the crime patterns and trends in India based on various factors. This will help in better readability and identification of related facts and figures.
- To discover new trends and patterns in the crime data using unsupervised learning algorithms like Fuzzy C-Means Clustering.
- To generate a prediction model, which will help to identify the factors impacting crime rate the most.

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 2001-2019.

The other sections of this paper are structured as follows: Section 2 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. Next section 3 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.

2. Related Work:

Data mining 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 [5][6]. A detailed review of crime analysis using data mining techniques was done by Hassani [7]. The study discusses mainly five types of taxonomies, some of which are also described in Chen et al. (2004). Among these five types, the emphasis was done on models that implemented neural networks, support vector machines (SVM), and decision trees. Further, Shamsuddin [8] 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 [9]. 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 Naïve Bayes and Decision Tree. Sathyadevan, S. (2014, August) [10], 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 of USAs [11]. Yadav proposed a regression model based on Naïve Bayes, Linear Regression, and Decision Tree using crime data of India from 2001-2014 [12].

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 is done for some countries but not in India, and if done then, it's not projected well enough based on district or city-wise. 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 [13]. Another work analyses the spatial relationships between crime occurrences, demographic, socio-economic, and environmental variables, together with geo-located Twitter messages and their 'violent' subsets the data used is of Chicago [14]. The author has done spatial analysis in India, but of 99 Cities, and the data used contains 14 different types of crime, and the data is of 1971 [15]. 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 [16]. Various other spatial methods were also implemented in this work such as spatial filtering, weighted regression, and multi-level regression with spatial dependence. Catlett C. 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 [17]. In another approach the author identified the low population density areas and demonstrated an imbalance aware machine learning application to identify burglary risk areas [18]. Rummens et al. (2017) focused on spatiotemporal crime forecasting, which proposes a combined predictive model like the Risk Terrain Model, consisting of LR and MLP, therefore concluding crime hotspots that have a risk of more than 20% [19]. Further, Araujo et al. (2018) proposed a well-defined framework followed by feature-engineering dependent methods [20]. In this work, visualization of crime data is done spatially using a time-series graph along with a forecasting model. This paper of Huang et al. (2018) 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 forecasted using the binary classification problem [21]. Zhuang et al. [22] 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).

3. 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.

3.1. Data Collection & 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 [23]. NCRB is a nodal organization to collect, accumulate and circulate the crime data of India in the form of annual report. 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 based on the types of crimes, the place of occurrence of crime, a crime against different kinds of people, and State/UT-Wise as well. 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 used are for the heat-map, for which the district-wise data is joined with India's geographical shapefile to obtain the accurate shapes of all the districts or cities.

3.2. 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.) [24]. This study also works on these factors and aims to identify the areas where the citizens suffers 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.

3.2.1. District-wise Crime Cases using Heat Map

This module uses a district-wise crime data set of the year 2014 and a shapefile for all the districts of India. Visualization is done on this data set 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 color 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.

(i) 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 Fig.2. 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).

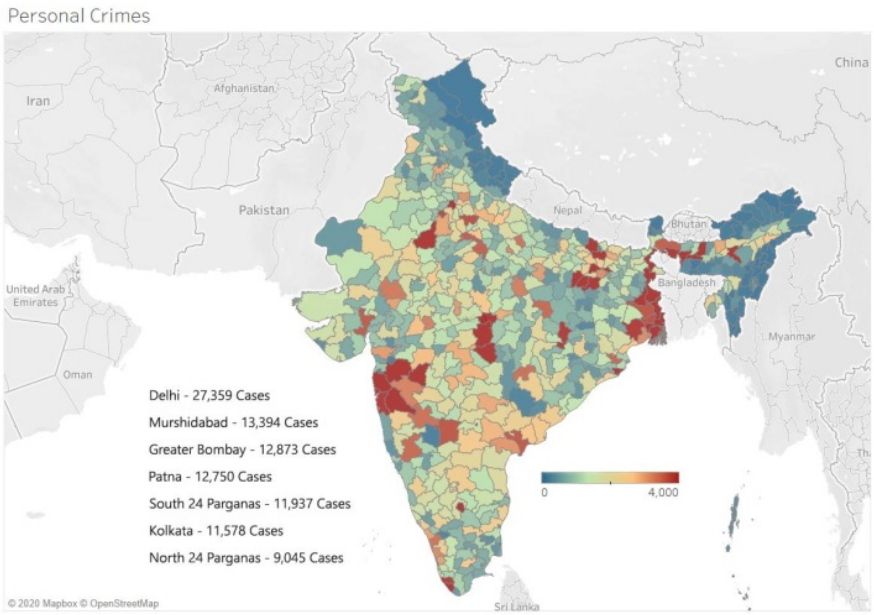

Fig.2. - Visualization of Personal Crimes in 2014 using Heat-Map.

(ii) Property Crimes

Property crimes mean intrusion or trespassing in the property of another without 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, dacoity, larceny, auto theft, and trespassing. Heat-map visualization of property crimes throughout India is shown in Fig.3. 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).

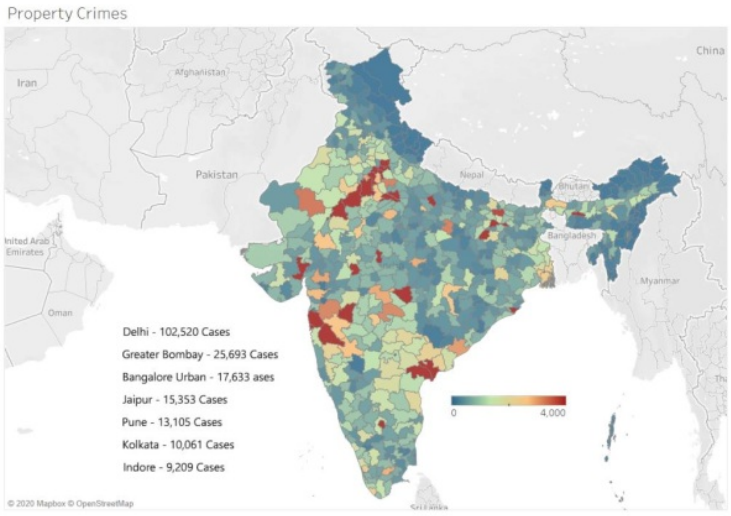

Fig.3. - Visualization of Property Crimes in 2014 using Heat-Map.

(iii) 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 Fig.4. 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 Fig.4 is up to 4,000 which is the same as in Fig.2, Fig.3, and Fig.5 as well, which is done intentionally for better comparison among them. It further depicts that criminal activities are slightly higher towards 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).

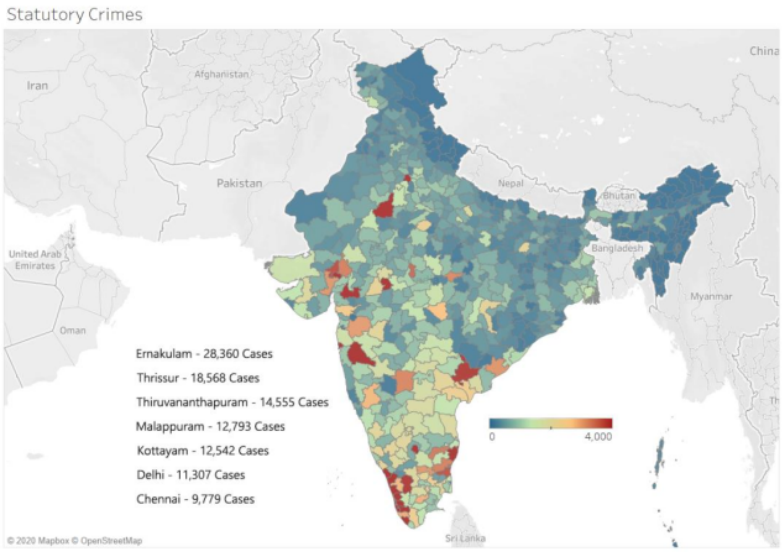

Fig.4. - Visualization of Statutory Crimes in 2014 using Heat-Map.

(iv) 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" towards 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. Sometimes 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 Fig.5. 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).

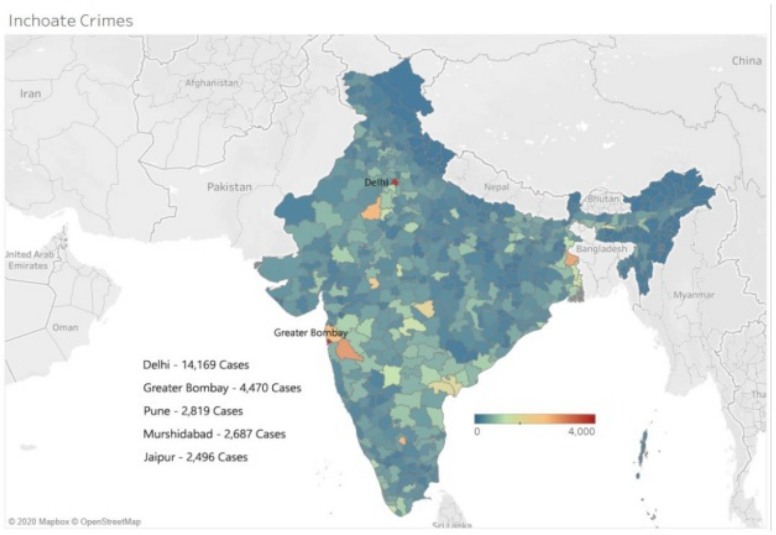

Fig.5. - Visualization of Inchoate Crimes in 2014 using Heat-Map.

3.2.2. 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 fig. 6 and fig 7. These 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, 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.

$$Crime\ Rate = \frac{Number\ of\ Incidences}{Projected\ Population\ (in\ lakhs)}$$

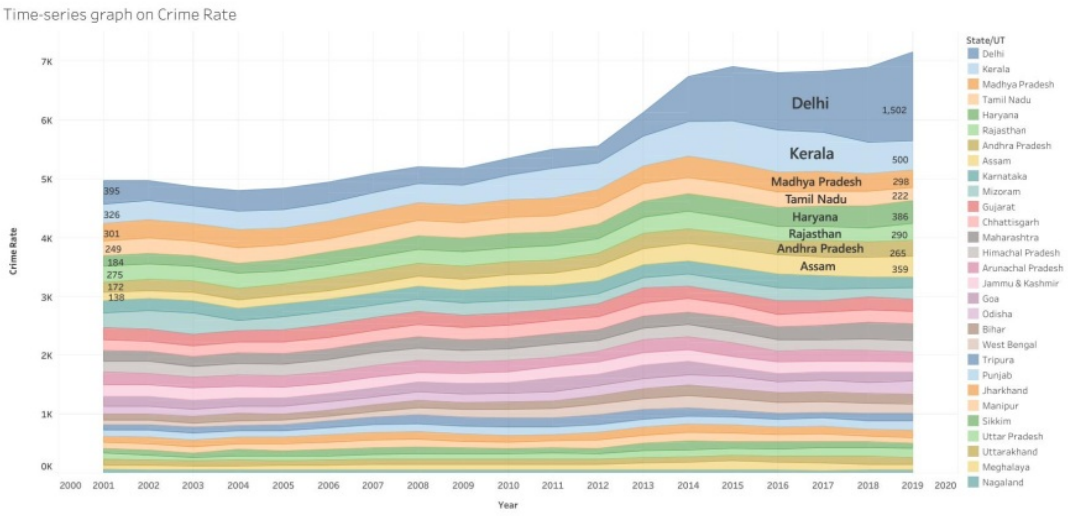

Fig.6. – Crime Rate visualization on the States of India and Delhi (Capital of India)

(a) State names are from 2001. (b) Telangana and Andhra Pradesh as considered as one from 2014 as Andhra Pradesh was divided into two.

Now, analysis based on total crime cases from 2001 – 2019 is shown below in Fig.7. From this, we can analyze that the criminal activity is the highest in Maharashtra (171.2K – 341.1K), Madhya Pradesh (181.7K – 246.5K), Uttar Pradesh (178.K – 353.1K), Andhra Pradesh (130.1K – 237.6K), Tamil Nadu (154.8K – 168.1K), Rajasthan (155.2K – 225.3K), and Kerala (103.8K – 175.8K), account to about more than 50% of the total criminal cases shown in the figure.

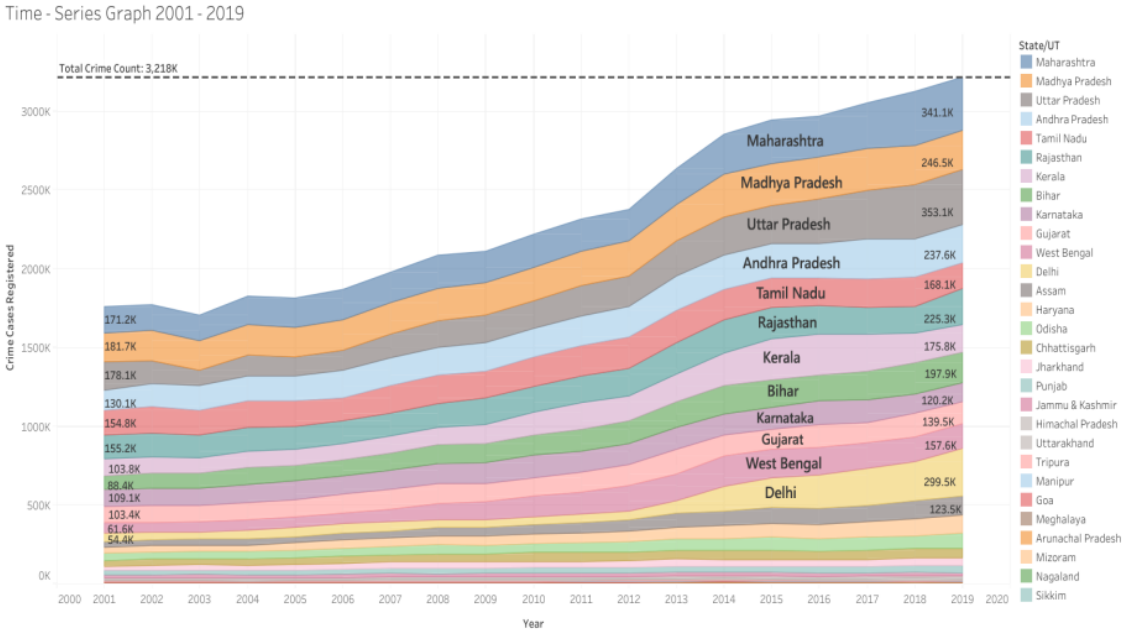

Fig.7. – Total Crime Cases visualization on the States of India and Delhi (Capital of India)

(a) State names are from 2001. (b) Telangana and Andhra Pradesh as considered as one from 2014 as Andhra Pradesh was divided into two.

3.2.3. 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 enforcements 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 lost. Figure 8 depicts that the greatest number of criminal activities happens in Residential Premises (292.2K), Roadways (208.7K), Other Places (Places other than the ones which are listed have 161.4K cases), Railways (58.3K) and Commercial Establishments (50.8K).

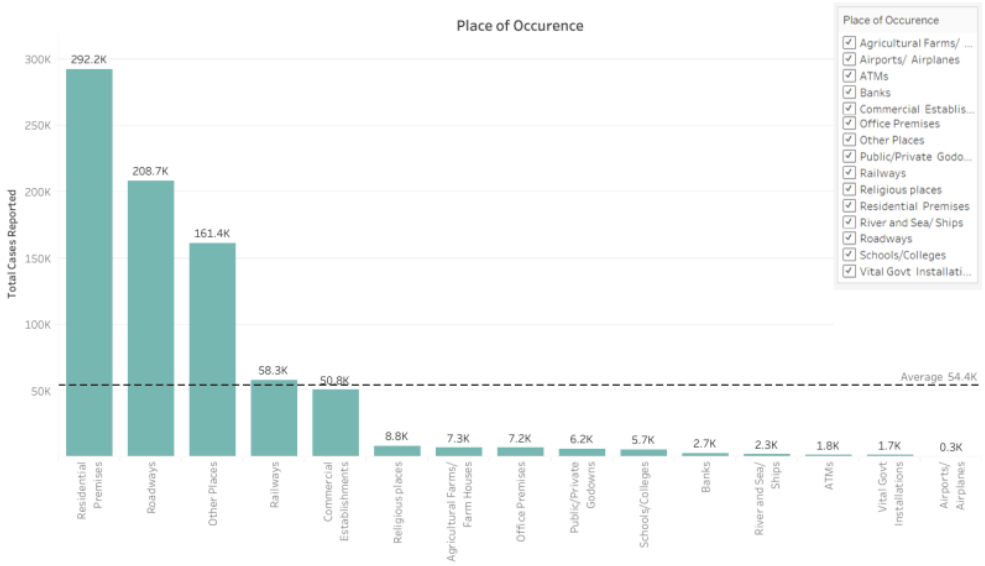

Fig.8. – Analysis based on place of occurrence 2019

3.2.4. 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, Children, 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 number of cases has increased by 13.2% from 2017 to 2019 on these people. It's also noticeable that the highest number of cases is towards 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 towards children, which require the most protection as they are the most vulnerable people amongst the lists.

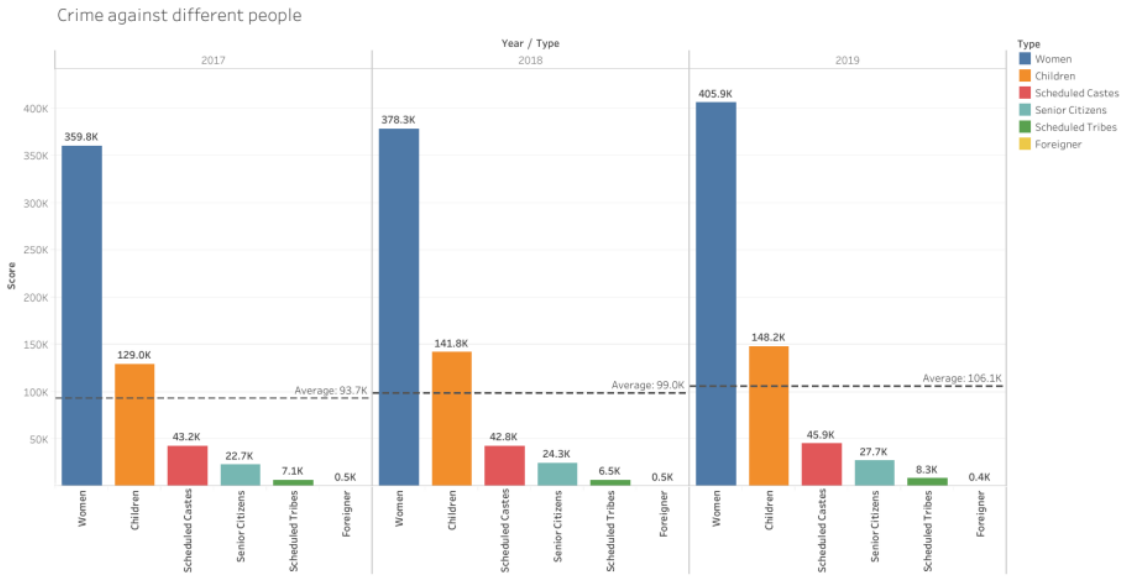

Fig.9. – Visualization based on different kinds of people

3.3. Crime Prediction

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. Some reputed and known Python libraries are used for the following to manipulate and read Big Data. These libraries are Sci-Kit Learn, Pandas, NumPy, and Sci-Kit Fuzzy.

Random Forest, Decision Tree, and Naïve Bayes Classification/Regression algorithms are applied as they are similar and to further find out the algorithm with the highest performance. 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.

3.3.1. 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 [25].

- 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 which usually contains hard clusters.

Fuzzy C-Means (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 [26] and improved by Bezdek [27]. 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 object membership corresponding to each cluster centroid based on the Euclidean Distance between them. After each iteration, the membership of each data objects are updated based on the minimization formula shown below.

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

Where, 'm' is the fuzziness index which is greater than 1, 'N' is the number of data points, 'C' is the number of centroids, '$\square_{ij}$' represents the membership of i^{th} data to j^{th} cluster centroid, 'x^i' is the i^{th} of d-dimensional measured data, 'c^j' is the d-dimension centre of the cluster, and '$\left\|x_i - c_j\right\|^2$' is the Euclidean Distance between ith data point and jth cluster centre.

Following are the steps in algorithm [31][28]:

Step 1: Initialize $U = [\square_{ij}]$ matrix, $U^{(0)}$,

Step 2: At *k-step*, calculate the centres vectors $C^{(k)} = [c_{ij}]$ with $U^{(k)}$,

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} . x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}} \quad -(2)$$

Step 3: Update $U^{(k)}$, $U^{(k+1)}$,

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad -(3)$$

Step 4: If *||U(k+1) - U(k)||* < □, then STOP; otherwise return to step 2.

Step 5: The Fuzzy partitioning is carried out through an iterative optimization of the objective function shown in Eq. (1), with the update of membership u_{ij} and the cluster centers c_j by using Eq. (3) and Eq. (2)

Step 6: This iteration will stop when.

$$max_{ij} \left\{ \left| \mu_{ij}^{(k+1)} - \mu_{ij}^{(k)} \right| \right\} < \varepsilon,$$

where, □ is a termination criterion between 0 and 1, whereas *k* is the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

So, these were the steps of the algorithm. 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.

Table 1 Details of the collected and pre-processed data

Attributes	Description
State/UT	There are 28 states and 7 union territories, Ladakh and Jammu & Kashmir are considered as one state as Ladakh became a Union Territory in late 2019.
Population	States/UTs wise population is in this attribute and the population used is a Mid-Year Projected Population (in lakhs).
Density	States/UTs wise density is in this attribute and the count is based on Census 2011 of India
Crime Type	Crime Types which consist some of the violent crimes, and those are Murder, Rape, Riots, Robbery, Arson, Attempt to Commit Murder, and Dowry Deaths
Crime Cases	Criminal Cases Registered or Criminal Incidences that occurred.
Crime Rate	Crime Rate is Cases per population in lakhs

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 Fig. 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 the case of 1 the FPC value would always be at 1 which is the highest. The result of which is shown in Fig. 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 Fig. 11.

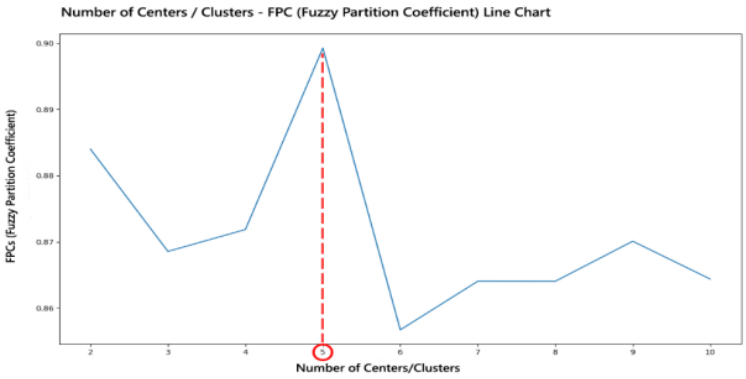

Fig. 10. – Fuzzy Partitioning Coefficient for Population – Crime Rate Chart

Fig. 11. - Population – Crime Rate Fuzzy C-Means Clustering

By carefully looking at Fig. 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 1300 population part and, there is just one cluster (the blue one) towards 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 199-523 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 659-826 in population and the crime rate is up till 7.92, and this result is for about 5 States. The last one is the purple cluster which depicts the remaining three 3 States (West Bengal, Maharashtra, and Bihar) whose population is between 971-1225 in lakhs, and the crime rate is up till 9.4 with West Bengal having the highest for crime type *Attempt to commit Murder*.

3.3.2. Classification

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. At first, the model is trained using the given data,

and then the data for which prediction must be made is tested. 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 which helps in better judgement and visualization of the performance of a Data Mining Algorithm, usually the algorithm is of supervised learning [28]. It is better shown in Fig. 12 below, where TP is True Positive, TN is True Negative, FN is False Negative, and FP is False Positive.

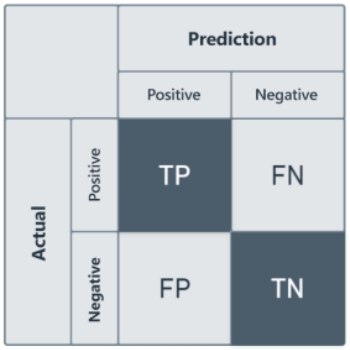

Fig. 12. – Confusion Matrix

- ***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 to check out the score [29][30]. The score is given in percentage.

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

- ***Precision Score***

Precision Score is another metric used to check the efficacy and performance of the algorithm. 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).

$$Precision\ Score = \frac{True\ Positive}{True\ Positive\ + False\ Positive}$$

- ***Recall Score***

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 capture through labeling it as Positive (True Positive).

$$Recall\ Score = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- ***F1 Score***

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.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

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:

(a) Naïve Bayes Algorithm

This algorithm is based upon the Bayes Theorem [31], in which he describes the probability of an event, based on prior knowledge of conditions that might be related to it. The mathematical formula is shown below,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Naïve 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 its ease in implementation and precise results [32].

(b) 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.

(c) *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 [33]. 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 of the attributes or parameters are shown below in the table.

Table 2 Details of the collected and pre-processed data

Attributes	Description
Region	Indian States/UTs are divided into 8 regions which are Arabian Sea, Bay of Bengal, Northern, Northeastern, Central, Eastern, Western, and Southern.
State/UT	There are 28 states and 7 union territories, Ladakh and Jammu & Kashmir are considered as one state as Ladakh became a Union Territory in late 2019.

Population	States/UTs wise population is in this attribute and the population used is a Mid-Year Projected Population.
Crime Type	Crime Types which consist of most of the crimes are considered, which are Murder, Rape, Hurt, Kidnapping and Abduction, Riots, Grievous Hurt, Dowry Deaths, Deaths due to negligent driving/act, Theft, Dacoity, Robbery, Offenses against the State, Incidence of Rash Driving, and Other IPC (Indian Penal Code) Crimes.
Crime Cases	Criminal Cases Registered or Criminal Incidences that occurred.
Crime Rate	Crime Rate is Cases per population in lakhs

Amongst the Table shown above the target field or class label is Region, the rest of it are the 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 makllle 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.

Table 3 Experimental Results of all three classifiers

Classifier	Accuracy Score	Precision Score	Recall Score	F1 Score
Naïve Bayes	85.41	84.73	85.41	85.05%
Decision Tree	88.63%	86.76%	88.63%	87.6%
Random Forest	91.67%	89.21%	91.67%	89.21%

From this, we can easily analyze that the least accurate Classifier based on these attributes and data is Naïve 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 Naïve Bayes, *GaussianNB* was implemented and for Precision Score, Recall Score, and F1 Score average was '*weighted*', same as in other algorithms [39]. Confusion Matrix for Random Forest Classifier is shown below (in Figure 13) as it's Classification was the most accurate in getting the Regions of India, and it will help verify the values. The matrix is 9x9 as the States and Union Territories are classified into 9 regions.

	Arabian Sea	Bay of Bengal	Central	Eastern	North-eastern	Northern	Southern	Western
Arabian Sea	4	0	0	0	0	0	0	0
Bay of Bengal	2	0	1	0	4	0	0	0
Central	0	0	2	1	0	0	0	0
Eastern	0	0	1	8	0	0	0	0
North-eastern	0	0	0	0	25	0	0	0
Northern	0	0	0	0	0	30	1	0
Southern	0	0	0	0	0	0	21	0
Western	0	0	0	0	0	0	0	20

Fig. 13. – Confusion Matrix for Random Forest

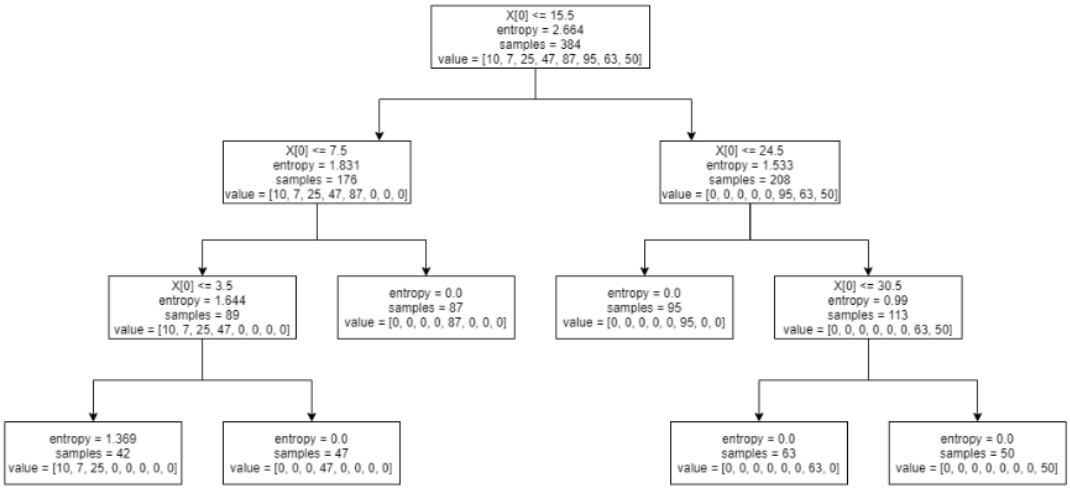

Fig. 14. – Crime Prediction Model based on Random Forest

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 the Figure 13 above. 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.

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 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:

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

Where 'Pi' is simply the frequentist 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.

4. Conclusion:

This paper concludes for the classification part that the Random Forest Classification model gives the most balanced and reliable results concerning Accuracy Score, Precision Score, Recall Score, and F1 Score. Random Forest

Classifier was the most abled classifier among Random Forest, Naïve Bayes, and Decision Tree. The target field in these three classifiers was the same, which is the 'Region' class label, which categorizes each State and Union Territory to any one of the eight regions (as mentioned in Table 2), which are based on the coordinates of the state on the Indian Map. For the Clustering part, we can conclude that Fuzzy C-Means Clustering is a soft version of K-Means, which shows that a data point or object can belong to two or more than two clusters, and this gets updated with every iteration. FCM is implemented for the parameter State/UT-wise Projected Mid-year Population 2019 to Crime Rate, which also concludes that five clusters were formed with an FPC of more than 0.89 and lesser than 0.90, as shown in Fig. 9 and the cluster projection is shown in Fig. 10.

From the Data Visualization Section (3.3), we can also conclude that Delhi, Greater Bombay (now Mumbai), and Bangalore Urban is noticed frequently in the top 10 in the district wise heat map of India (Fig. 1, Fig. 2, Fig. 3, and Fig. 4). From both the Time-Series Area Graphs (Fig. 5 and Fig. 6) combined, it can be concluded that Madhya Pradesh, Tamil Nadu, Andhra Pradesh, and Rajasthan have a high Count of Crime Cases and high Crime Rate too. In the Bar Graph based on Place of Occurrence (Fig. 7), it can be seen that most of the criminal activities have occurred in Residential Premises and Roadways. From the Bar Graph Representation of crimes on the kinds of people who get targeted the most from 2017-2019 (Fig. 8), it can be concluded that crime against women has increased by 12.8% in just 3 years and for children, it has increased by 14.8%.

5. References:

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

[2] Bhatnagar, R. R. (1990). Crimes in India: problems and policy. New Delhi: Ashish Publishing House.

[3] Sharma, S. (2015). Caste-based crimes and economic status: Evidence from India. Journal of comparative economics, 43(1), 204-226.

[4] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.

[5] Rosenthal, S. S., & Ross, A. (2010). Violent crime, entrepreneurship, and cities. Journal of Urban Economics, 67(1), 135-149.

[6] Hajela, G., Chawla, M., & Rasool, A. (2020). A Clustering Based Hotspot Identification Approach For Crime Prediction. Procedia Computer Science, 167, 1462-1470.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction. Indian Journal of Science and Technology, 6(3), 4219-4225.

[12] 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.

[13] Cichosz, P. (2020). Urban Crime Risk Prediction Using Point of Interest Data. ISPRS International Journal of Geo-Information, 9(7), 459.

[14] 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, 1-32.

[15] Dutt, A. K., & Venugopal, G. (1983). Spatial patterns of crime among Indian cities. Geoforum, 14(2), 223-233.

[16] Bernasco, W., & Elffers, H. (2010). Statistical analysis of spatial crime data. In Handbook of quantitative criminology (pp. 699-724). Springer, New York, NY.

[17] 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.

[18] 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.

[19] Rummens, A., Hardyns, W., & Pauwels, L. (2017). The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. Applied geography, 86, 255-261.

[20] 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 Computing 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.

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

[22] 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.

[23] The National Crime Records Bureau of India Website, https://ncrb.gov.in/en

[24] 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.

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

[26] Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

[27] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.

[28] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion Matrix-based Feature Selection. MAICS, 710, 120-127.

[29] Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India.

[30] Yerpude, P. (2020). Predictive Modelling of Crime Data Set Using Data Mining. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 7.

[31] Joyce, James (2003), "Bayes' Theorem", in Zalta, Edward N. (ed.), The Stanford Encyclopedia of Philosophy (Spring 2019 ed.), Metaphysics Research Lab, Stanford University, retrieved 2020-01-17

[32] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[33] Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

[34] Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B., & Zhang, H. (2014). Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. PloS one, 9(1), e86703.

Crime

ORIGINALITY REPORT

5%	6%	1%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	towardsdatascience.com Internet Source	1%
2	scikit-learn.org Internet Source	1%
3	www.saagtechnolegalconsultants.com Internet Source	1%
4	Hitesh Kumar Reddy ToppiReddy, Bhavna Saini, Ginika Mahajan. "Crime Prediction & Monitoring Framework Based on Spatial Analysis", Procedia Computer Science, 2018 Publication	1%
5	aut.researchgateway.ac.nz Internet Source	1%
6	millerlaw.law Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		