Crime Analysis & Prediction

Using Data Mining

**Abstract**

Crime is a violation of laws set forth by the state to maintain social order and stability in society. Data Mining of crime data trends helps in advanced enforcement measures. Due to the substantial increase in the crime rate, the application of data mining techniques can be beneficial for achieving insights on crime patterns, which will help law enforcement prevent crime with proper crime prevention strategies. This paper concentrates on Data Visualization and Prediction, using Machine Learning Algorithms to discover the patterns. This present work collects Crime Records of India for the type of people who are victimized the most (like women, children, senior citizens, etc.) and analyses the crime trend district-wise throughout the country by applying various classification techniques. Analyzing the crime would be much easier by using the predictive models developed in this work using complex Machine Learning Algorithms. Algorithms implemented are Fuzzy C-means Clustering, Random Forest, Naïve Bayes, and Decision Tree, used to calculate accuracy rate, recall rate, F-measure & precision score.

**Keywords**

*Data Mining • Classification • Data Visualization • Crime Prediction • Fuzzy C-means Clustering • Random Forest • Naïve Bayes • Decision Tree*

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. Therefore, every crime violates the law, but not every violation of the law is considered a crime. Like breaches of contract and other civil law will not lie among the category of crime, it will lie under the category of “offenses” or in “infractions”. 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. There’s 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.) [1]. 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 experience without explicitly programming it. Machine Learning is classified into three main components, Supervised Learning, Unsupervised Learning, and Reinforcement Learning. In Supervised Learning, the machine learns by using labeled data as the input to predict future events. In Unsupervised Learning, the model is trained using unlabeled data without any guidance and used for Association & Clustering. The last one, Reinforcement Learning is a learning method that interacts with the environment by producing actions and discovering errors or rewards. In this work Supervised Learning has been done and parts of Unsupervised Learning as well. Supervised Learning is further classified into sub-parts as shown in Fig. 1.

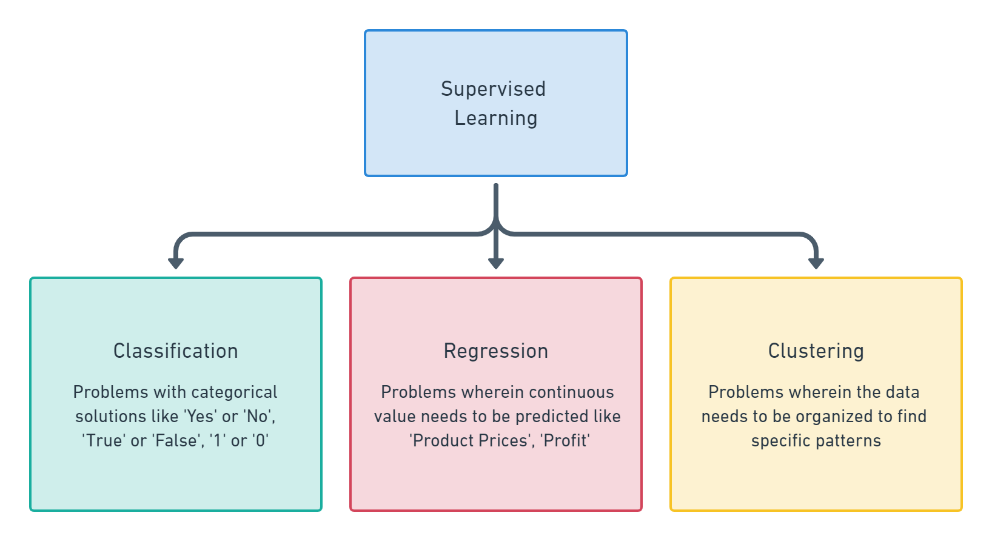


Fig. 1. – Classification of Supervised Learning

In this paper, the raw crime data sets used are confirmed and verified by the NCRB (National Crime Records Bureau), which proves its authenticity and assurance [2][3][4]. The data sets used are from a period of 2001 – 2019 with various kinds of parameters like based on place of occurrence, State/UT – wise, type of crime, etc. For the Data Visualization part of the paper, Tableau has been used as the visualization tool and for the Data Prediction part, Python programming language is used along with some known modules like Sci-Kit Learn, Pandas, NumPy, and Sci-Kit Fuzzy [5][6][7][8].

1. Related Work:

Crime Analysis and Prediction is an important activity that can be done and optimized using various ideas and techniques. Research work in this Crime domain is being done before and considered to be an important and practical one as well. This research work is quite different because analysis and prediction are done on unique factors, using some complex algorithms and utilization of the latest tools is being done. ToppiReddy, H. K. R. (2018) has done spatial analysis on a data set of the United Kingdom of the year 2015-2017 [9]. In this work, the latest data set of Crime in India 2019 is used, which is not yet used in any research work. Other than that spatial representation is done district-wise using the shapefile of each district and then represented through a heat map of India (shown in fig. 2, fig. 3, fig. 4 & fig. 5).

Another paper is of Sathyadevan, S. (2014, August), which is on crime prediction using Naïve Bayes and Decision Tree [10]. In this paper, these algorithms are used as well but to compare and show that the accuracy and precision of the Random Forest Classification algorithm are higher. The last related work, which was referred to is of David, H., & Suruliandi, A. (2017), in which K-Means clustering is done on the crime data set of 2015 [11]. This work uses and implements Fuzzy C-Means Clustering, which is an advanced version of K-Means, and the main difference is that each point can belong to more than one centroid or clusters.

1. Methodology:

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

* 1. Data Collection & Preprocessing

Multiple data sets have been utilized for Data Mining, and the data sets used are reliable, real, and verified by the NCRB (National Crime Records Bureau of India) [12]. The data sets are acquired by the official website of the Indian Government [3][4]. The study has been done on various parameters like based on the type of crimes, the place of occurrence of crime, and state-wise as well. In this phase, 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.

* 1. Data Visualization

Data Visualization is a graphical representation of data for those charts, graphs, tables, and maps are the elements used. 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 consuming lesser time as well. 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 in Tableau, which is a software built by an MNC (Multi-National Company) called Salesforce.

As crimes occur in an area, analysis based on the place of occurrence (such as Railways, Residential Area, etc.) or based on the kind of citizens which get targeted the most (such as women, children, senior citizens, etc.) can help a lot in understanding. Data Visualization plays an essential role in this for better demonstration & understanding.

* + 1. Module 1: Heat Map representation of different types of crime

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. 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.

1. Personal Crimes

Personal crimes are those that result in physical or mental harm to another individual. Further, classified into two categories, forms of homicide and other violent crimes. Where the physical damage to another individual is so severe that it causes death, 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 and battery, child or domestic abuse, kidnapping, rape, and statutory rape [13]. Heat-map visualization of personal crimes throughout India is shown in Fig.1. From this, we can analyze that criminal cases are 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).

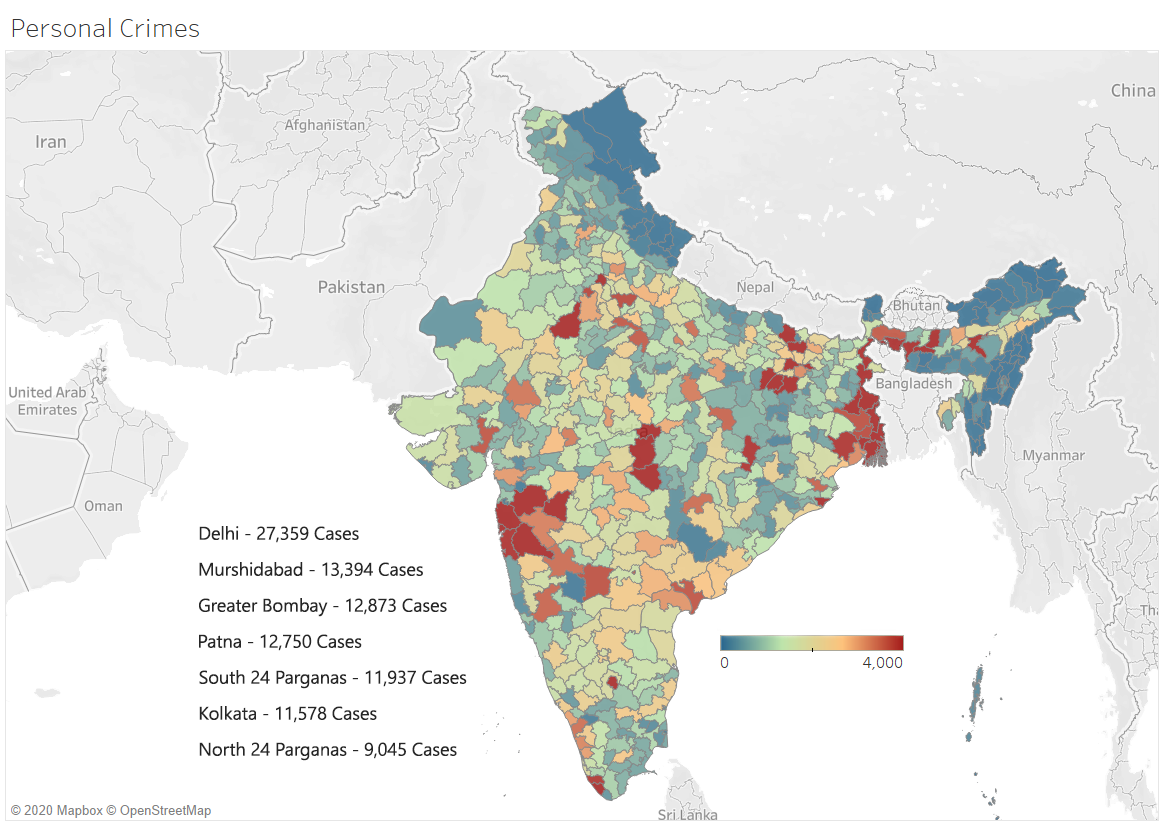


Fig.1. - Visualization of Personal Crimes in 2014 using Heat-Map.

1. Property Crimes

Property crimes mean involvement or interference with 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. Many property crimes are theft crimes, which include arson, burglary, dacoity, larceny, auto theft, and trespassing [13]. Heat-map visualization of property crimes throughout India is shown in Fig.2. From this, we can analyze that 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).

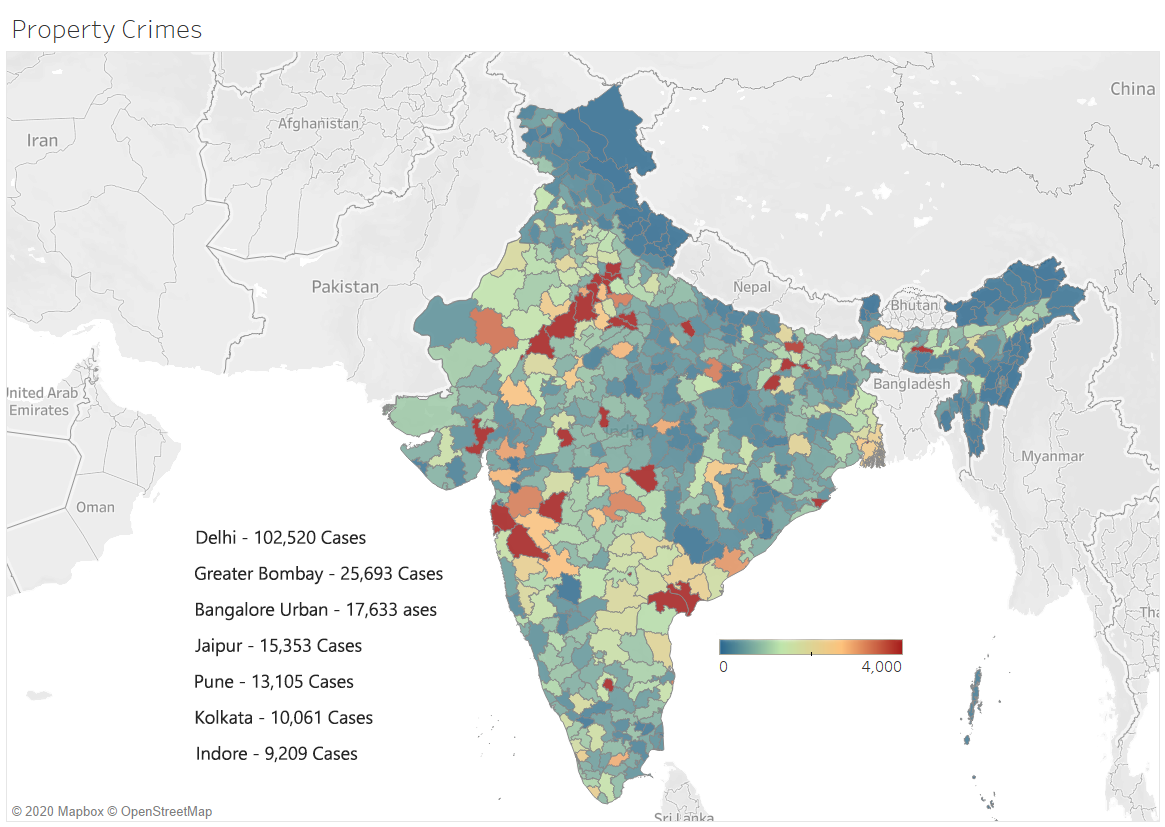


Fig.2. - Visualization of Property Crimes in 2014 using Heat-Map.

1. 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 minor 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.3. From this, we can analyze 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.3 is up to 4,000 which is the same as in Fig.1, Fig.2, and Fig.4 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).

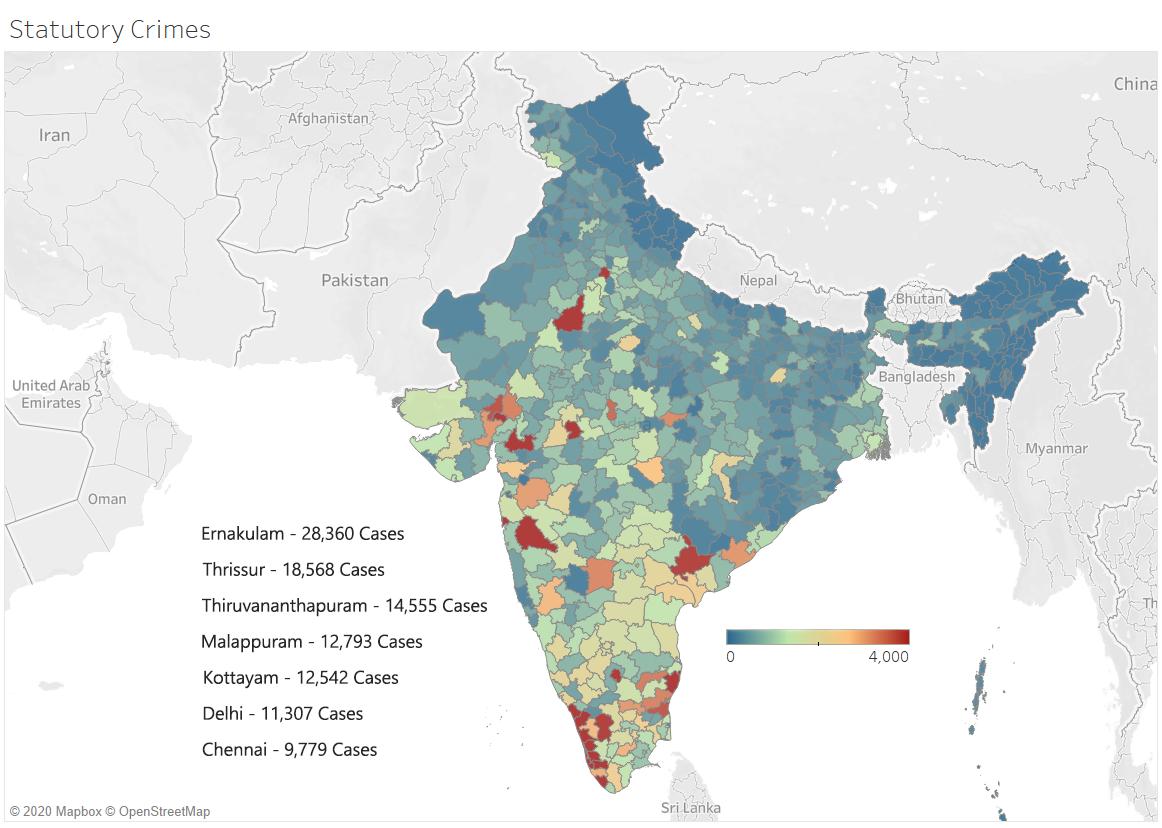


Fig.3. - Visualization of Statutory Crimes in 2014 using Heat-Map.

1. 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 [13]. The most common inchoate offenses are an attempt, solicitation, aiding and abetting, and conspiracy. 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 as it can be of the same degree as that of the underlying crime, or it can be a lot less severe too. Heat-map visualization of inchoate crimes throughout India is shown in Fig.4. 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, we can analyze that inchoate criminal 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).

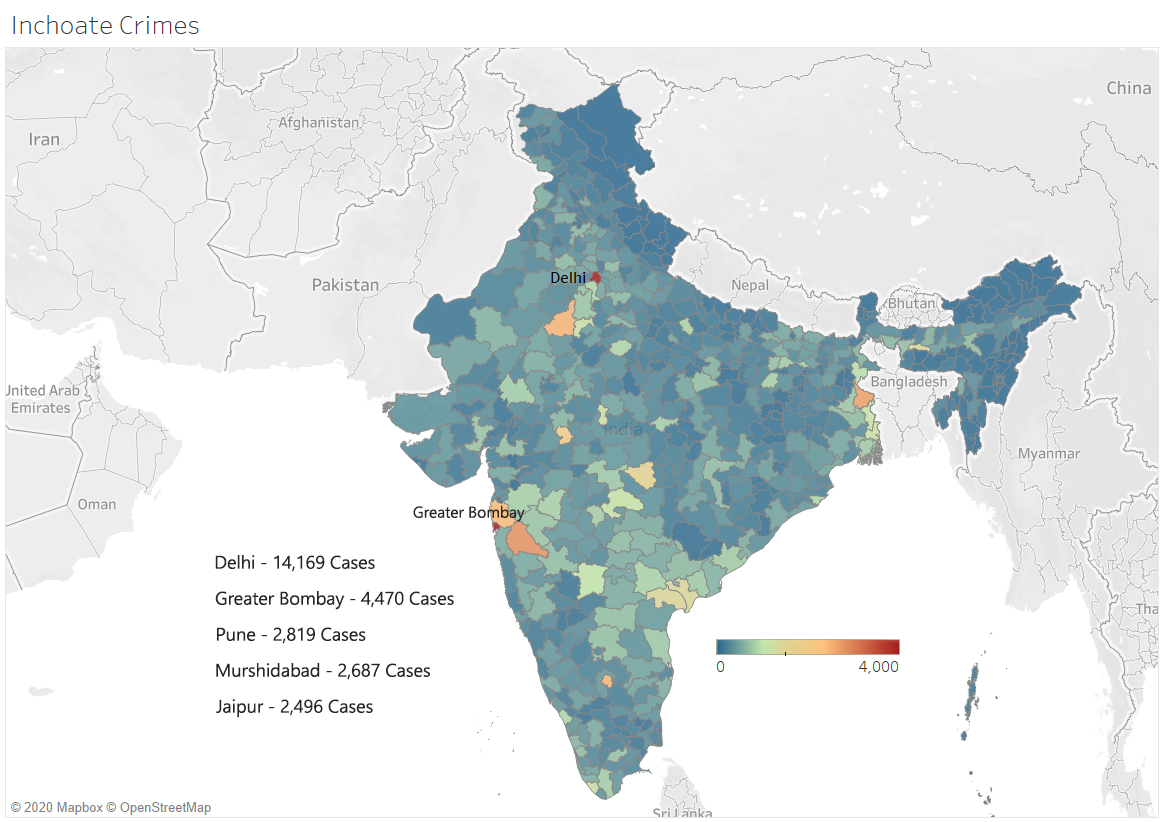


Fig.4. - Visualization of Inchoate Crimes in 2014 using Heat-Map.

* + 1. Module 2: Visualization of Crime State-Wise from 2001 – 2019

Crime in India has increased a lot 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 130% increase in crime rate. Data Visualization is done on two criteria, which are crime rate and total crime cases through time-series area graph representation. Time-series visualization of all states and the capital of India which is Delhi based on crime rate has been shown in Fig.5. From this, we can analyze 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 formula for crime rate is shown below.

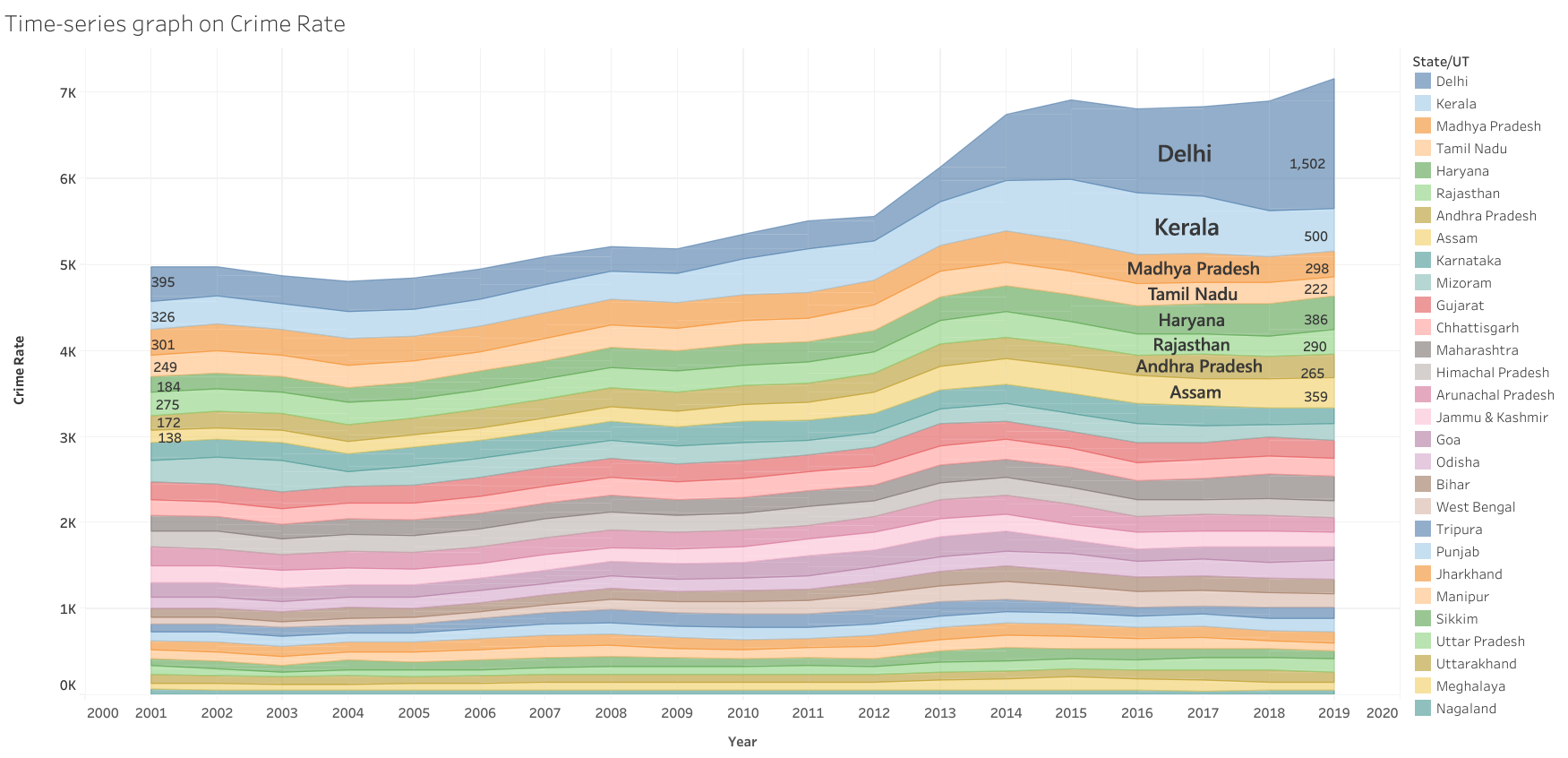


Fig.5. – Crime Rate visualization on the States of India and Delhi (Capital of India)

(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.6. 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 ½ of the total criminal cases.

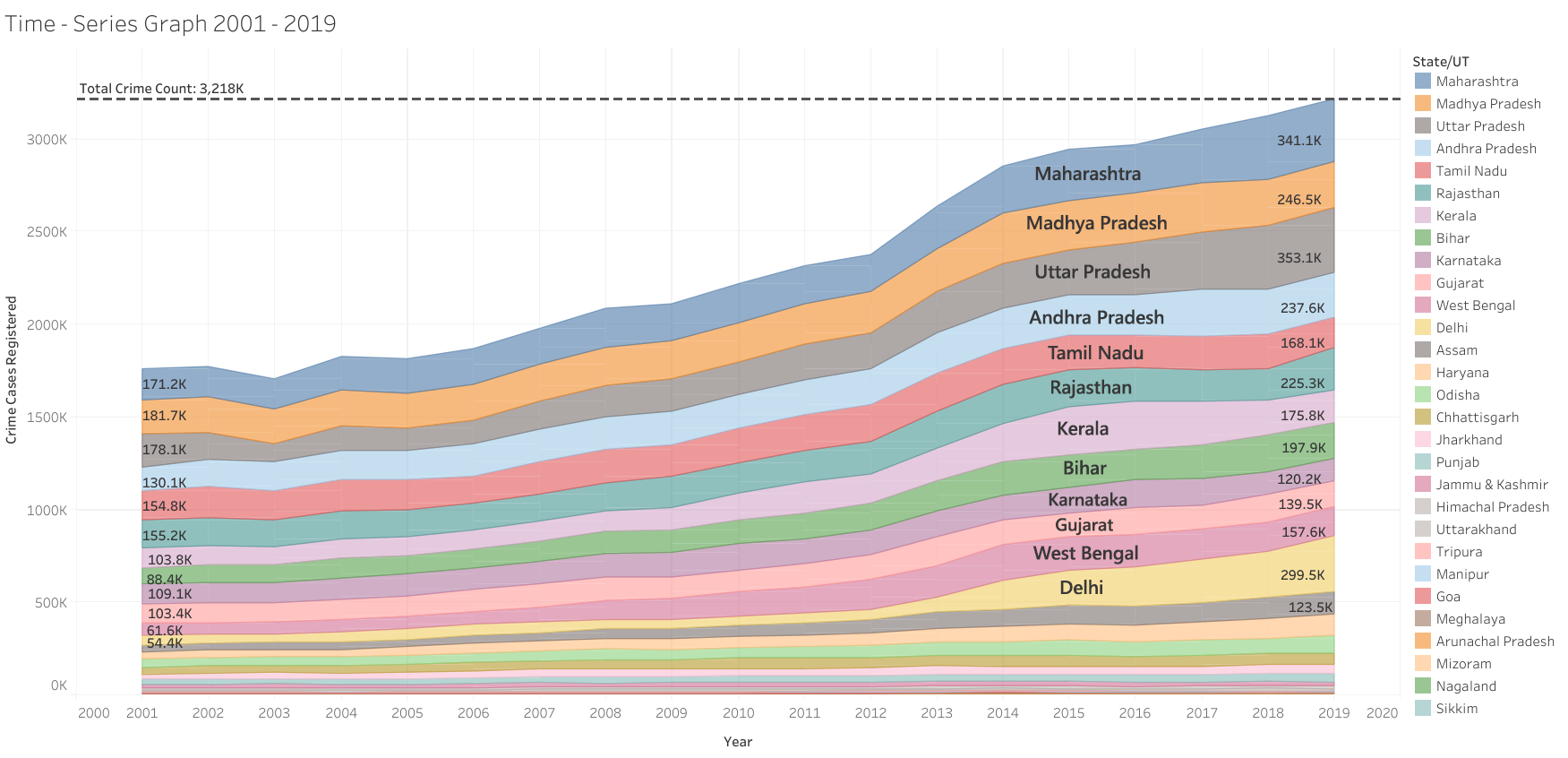


Fig.6. – Total Crime Cases visualization on the States of India and Delhi (Capital of India)

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

* + 1. Module 3: Visualization of Crime based on the place of occurrence

This section concentrates on the type of areas where most of the crimes get executed and in what kind of places should an individual be cautious and aware. In this analysis, not all the crimes have been included, most of the property crimes have been included for judgment. These property crimes which are utilized are dacoity, theft, robbery, burglary, and other offenses in which property is lost. In Figure 7 the analysis has been depicted, from which we can analyze 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).

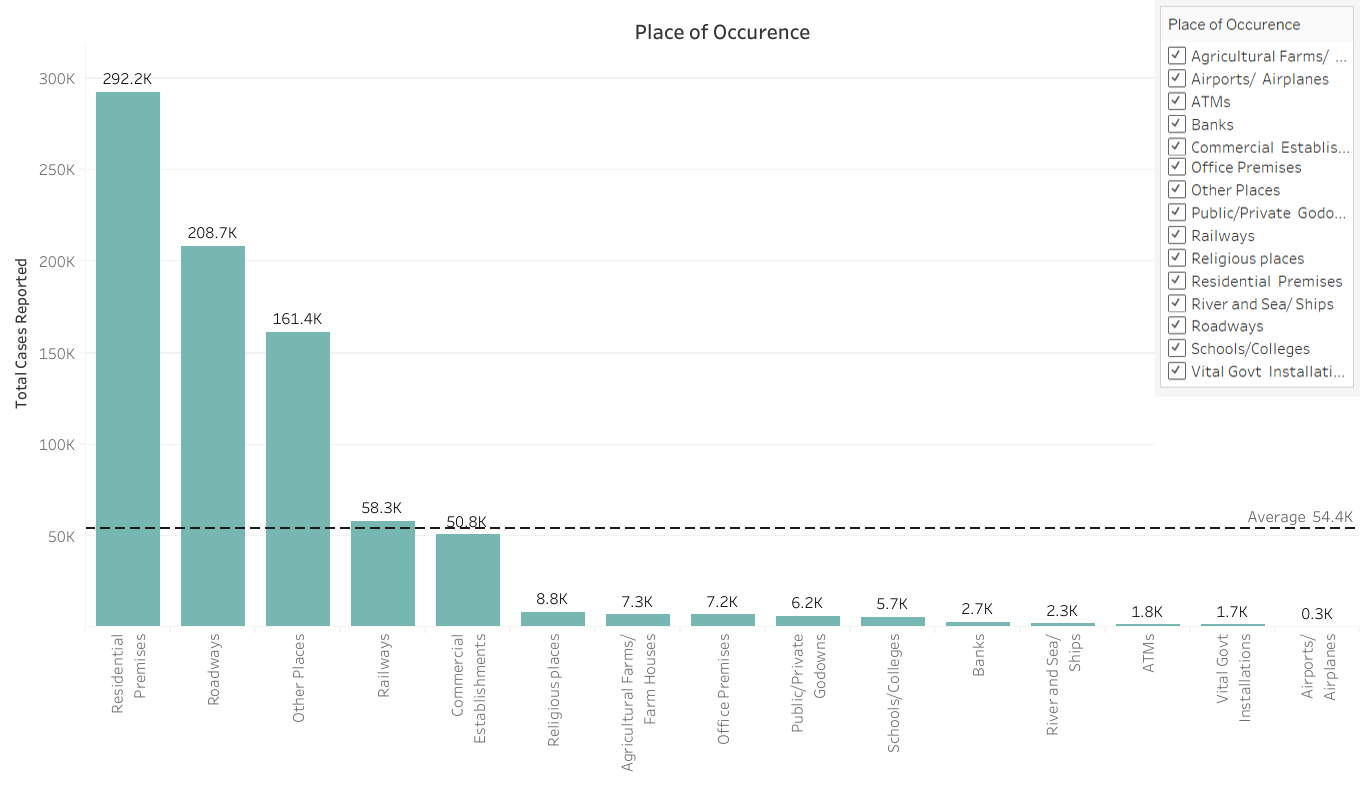


Fig.7. – Visualization based on place of occurrence 2019

* + 1. Module 4: Crime against different kinds of people

This section is crucial as it helps to understand the situation in India based on 2019’s data that what kind of citizens or non-citizens of India are mostly targeted or victimized by criminals. The type of people we are taking into account for this analysis are Women, Children, Senior Citizens, Scheduled Castes, Scheduled Tribes, and Foreigners. This analysis is shown below in figure 8 and it’s been done for three consecutive years (2017, 2018 & 2019). From this, we can analyze that the average number of cases has increased by 13.2% from 2017 – 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 list.

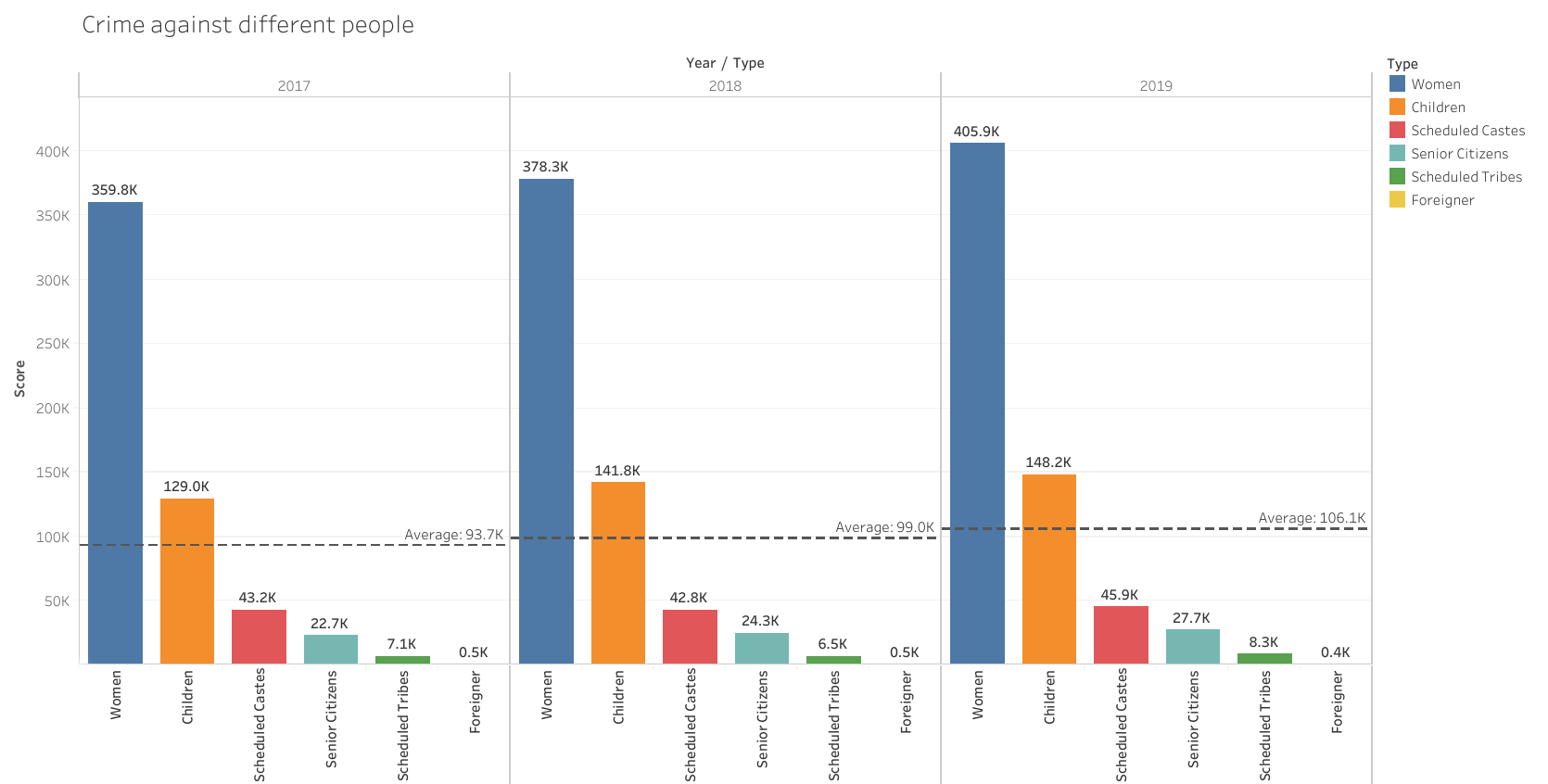


Fig.8. – Visualization based on different kinds of people

* 1. 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 [5][6][7][8].

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 the theory of all the algorithms including their respective formulas. Second sub-part is to depict the results and patterns obtained after implementation of the algorithms.

* + 1. Theory
       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 [14].

* 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.

* + - 1. Classification

1. *Naïve Bayes Algorithm*
2. *Decision Tree Algorithm*
3. *Random Forest Algorithm*
   * 1. Experimental Results
        1. Fuzzy C-Means Clustering
        2. Classification
4. Conclusion:
5. Future Scope:
6. References:

[1] Summarized analysis and definition of Crime in India at: <https://en.wikipedia.org/wiki/Crime_in_India>

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

[3] Crime in India 2017 – 2019 data set retrieved from, <https://ncrb.gov.in/en/crime-india-2019-0>

[4] Crime in India 2001 – 2014 data set retrieved from, <https://data.world/rajanand/crime-in-india>

[5] Scikit Learn Module in Python, Retrieved from <https://scikit-learn.org/stable/>

[6] Pandas Module in Python, Retrieved from <https://pandas.pydata.org/>

[7] Numpy Module in Python, Retrieved from <https://numpy.org/>

[8] Scikit Fuzzy Module in Python, Retrieved from <https://pypi.org/project/scikit-fuzzy/>

[9] ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science*, *132*, 696-705.

[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] David, H., & Suruliandi, A. (2017). SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES. *Ictact journal on soft computing*, *7*(3).

[12] Brief about the National Crime Records Bureau of India, <https://en.wikipedia.org/wiki/National_Crime_Records_Bureau>

[13] Types of Criminal Offenses, Retrieved from <https://www.justia.com/criminal/offenses/>

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

[1] Nath, S. V. (2006, December). Crime pattern detection using data mining. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops* (pp. 41-44). IEEE.

[2] 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.

[3] Tayal, D. K., Jain, A., Arora, S., Agarwal, S., Gupta, T., & Tyagi, N. (2015). Crime detection and criminal identification in India using data mining techniques. *AI & society*, *30*(1), 117-127.

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

[5]

[6] <https://en.wikipedia.org/wiki/Data_analysis>

[7] https://ncrb.gov.in/en/crime-india-2019-0

[8] <https://data.gov.in/catalog/crime-india-2015>

[9] <https://data.world/rajanand/crime-in-india>

[10] <https://www.tableau.com/>

[11] <https://scikit-learn.org/stable/>

[12] <https://en.wikipedia.org/wiki/National_Crime_Records_Bureau>

[13] https://ncrb.gov.in/en

[14] <https://desktop.arcgis.com/en/arcmap/10.3/manage-data/shapefiles/what-is-a-shapefile.htm>

[15] <https://www.justia.com/criminal/offenses/>

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

[17] ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science*, *132*, 696-705.

[18] Das, P., & Das, A. K. (2019). Application of classification techniques for prediction and analysis of crime in India. In *Computational Intelligence in Data Mining* (pp. 191-201). Springer, Singapore.

David, H., & Suruliandi, A. (2017). SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES. *Ictact journal on soft computing*, *7*(3).