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. 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 collects Crime Records of India (2001 - 2019) for the kind of people who are victimized the most (like women, children, senior citizens, etc.) and analyses the crime trend District-wise throughout the country with the help of some visualization tools. 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. The metrics used are Confusion Matrix, Accuracy score, F-measure, Precision score, and Recall 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 [1]. 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 [1]. 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 [2]. 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.

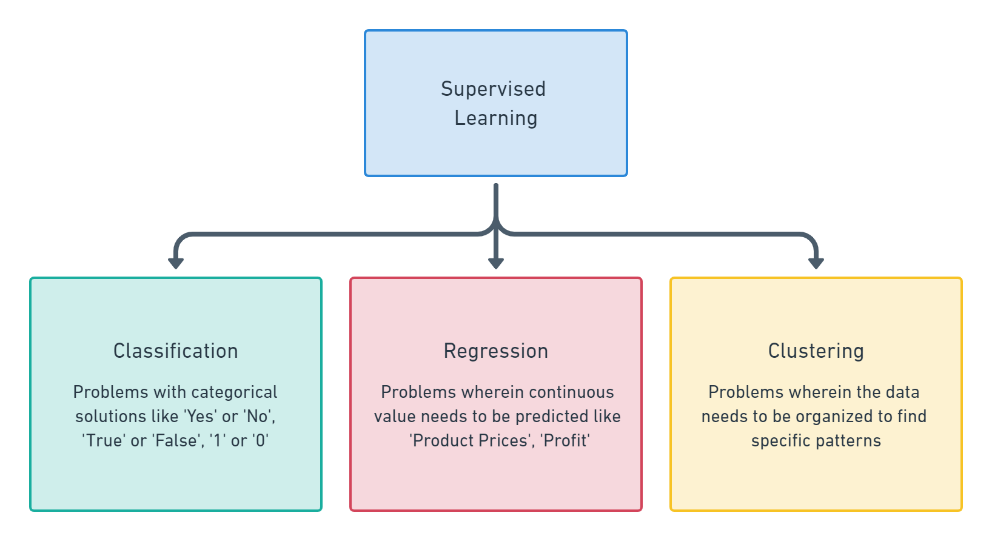


Fig. 1. – Classification of Supervised Learning

Further, the sections are organized in the manner of first Related Work, which discusses other research works on a similar topic and the main difference between those work and in this. After this comes the Methodology section, which is the essential portion of the paper as it shows all the trends and implementations. This section is classified further into three sub-parts, which are Data Collection, Data Visualization, and Crime Prediction. In these three sub-parts, the work done is using Tableau for the Visualization part [3], and Python programming language along with various modules like Sci-Kit Learn, Pandas, NumPy, and Sci-Kit Fuzzy, is used for Crime Prediction [4][5][6][7]. Then comes the Conclusion section, which crisply summarizes the findings and observations done in the paper. The last section of the research paper is for all the references, which are just for listing out all the references throughout the work.

1. Related Work:
   1. Spatial and geo-graphical analysis

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. For example, one of the researchers has 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 in the UK based on the POI layers from OpenStreetMap [8]. 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 [9]. The author has done spatial analysis on India, but of 99 Cities, and the data used contains 14 different types of crime, and the data is of 1971 [10]. This work is different as Tableau’s shapefiles were integrated with the district-wise data of 2014, for a better demonstration of data (shown in fig. 2, fig. 3, fig. 4 & fig. 5).

* 1. Data Mining Algorithms

Researchers have pursued algorithms for Crime Prediction, such as Naïve Bayes and Decision Tree in two papers, one by Sathyadevan, S. (2014, August) [11], which is to predict crime-prone areas based on a data set of India and another one which is to predict Crime Category using a data set of USAs [12]. Another example research is for Naïve Bayes, Linear Regression and Decision Tree algorithms in which a regression model is built using a data of India from 2001-2014, written by Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April) [13]. Also, the clustering algorithm K-means is used to identify the crime patterns [14], whereas Fuzzy C-Means Clustering is used in this work.

Using a constructive approach in classification, Decision Tree, Naïve Bayes, and Random Forest are implemented for the same data set for better comparison. This comparison is done based on some accuracy metrics like F-measure, precision score, recall score, and confusion matrix. Among these three algorithms, regression models for all three are built and then the best one is concluded which is Random Forest Algorithm.

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

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 [15][16][17][18]. The data sets used are from a period of 2001 – 2019. The study has been done on various parameters like based on the type of crimes, the place of occurrence of crime, crime against different kinds of people, and State/UT-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. Other preprocessing techniques used is 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.

* 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. *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. 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 [19]. 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, arson, child or domestic abuse, kidnapping, rape, and statutory rape [20]. Heat-map visualization of personal crimes throughout India is shown in Fig.2. 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).

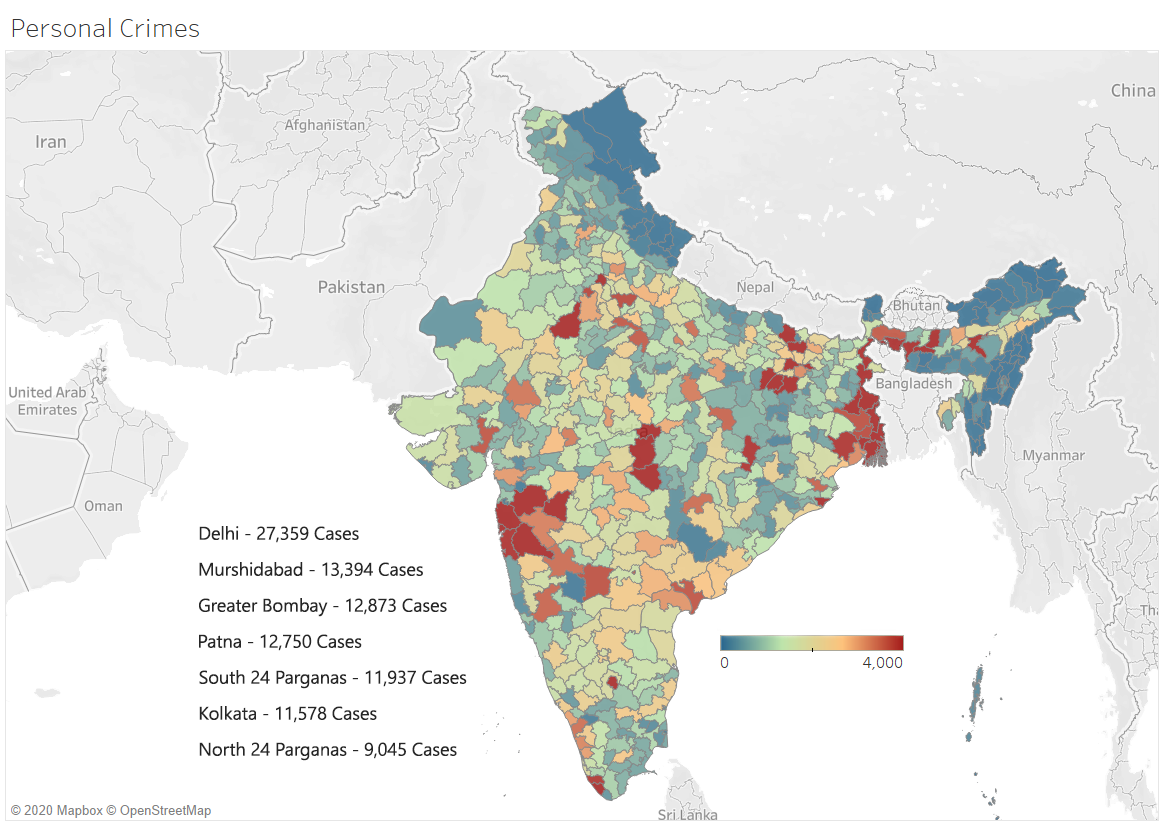


Fig.2. - Visualization of Personal Crimes in 2014 using Heat-Map.

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 [20]. Heat-map visualization of property crimes throughout India is shown in Fig.3. 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).

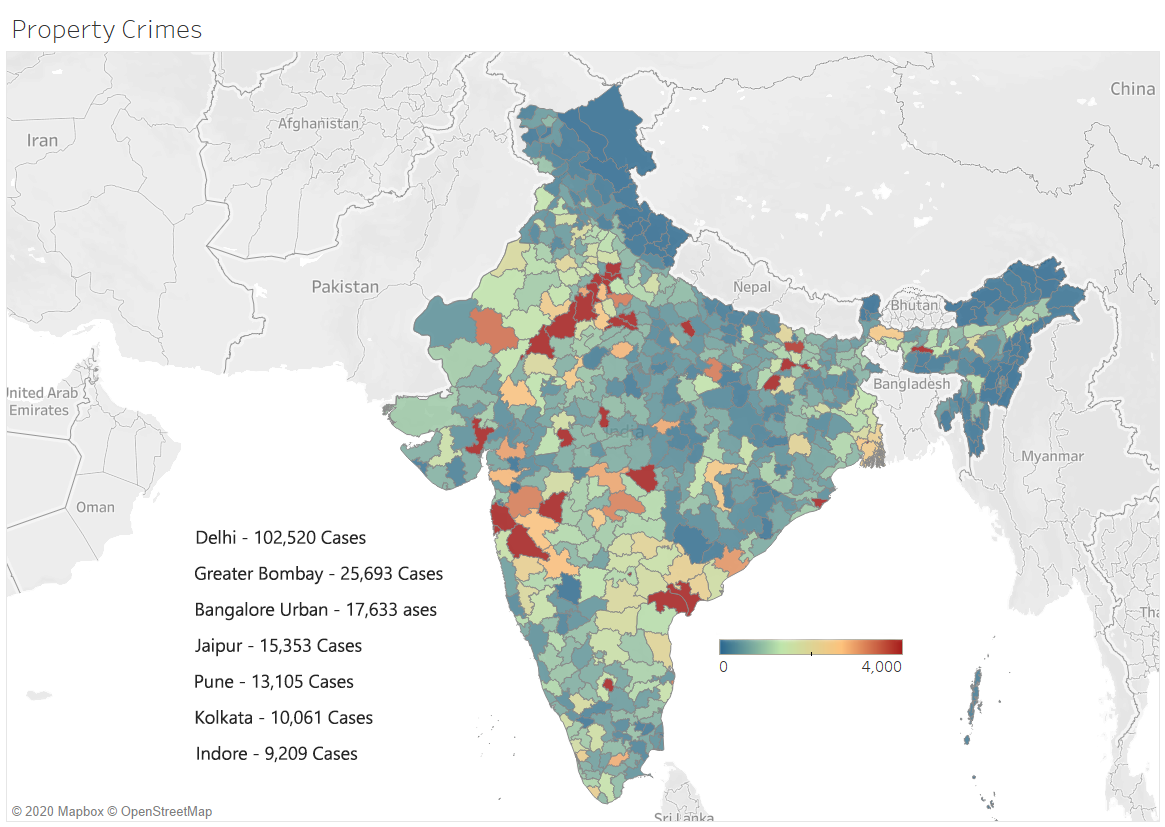


Fig.3. - Visualization of Property Crimes in 2014 using Heat-Map.

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.4. 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).

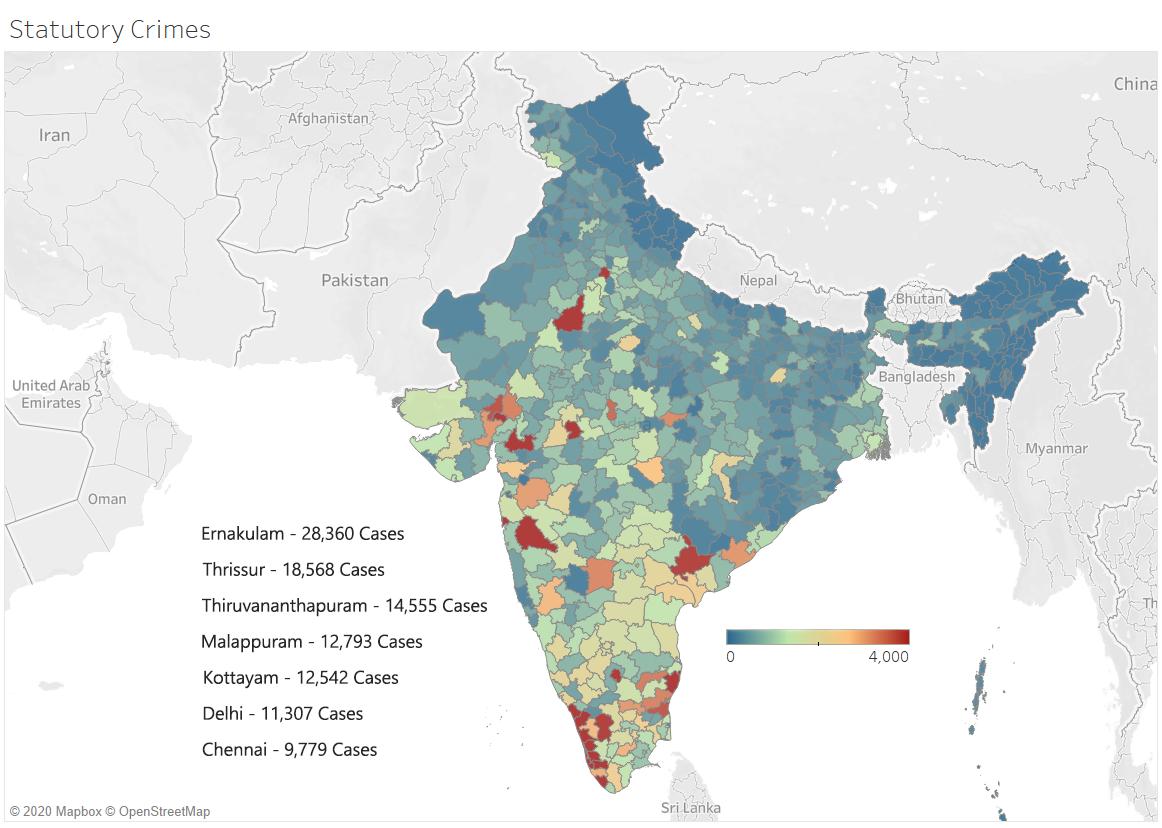


Fig.4. - Visualization of Statutory Crimes in 2014 using Heat-Map.

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 [20]. 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.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, 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).

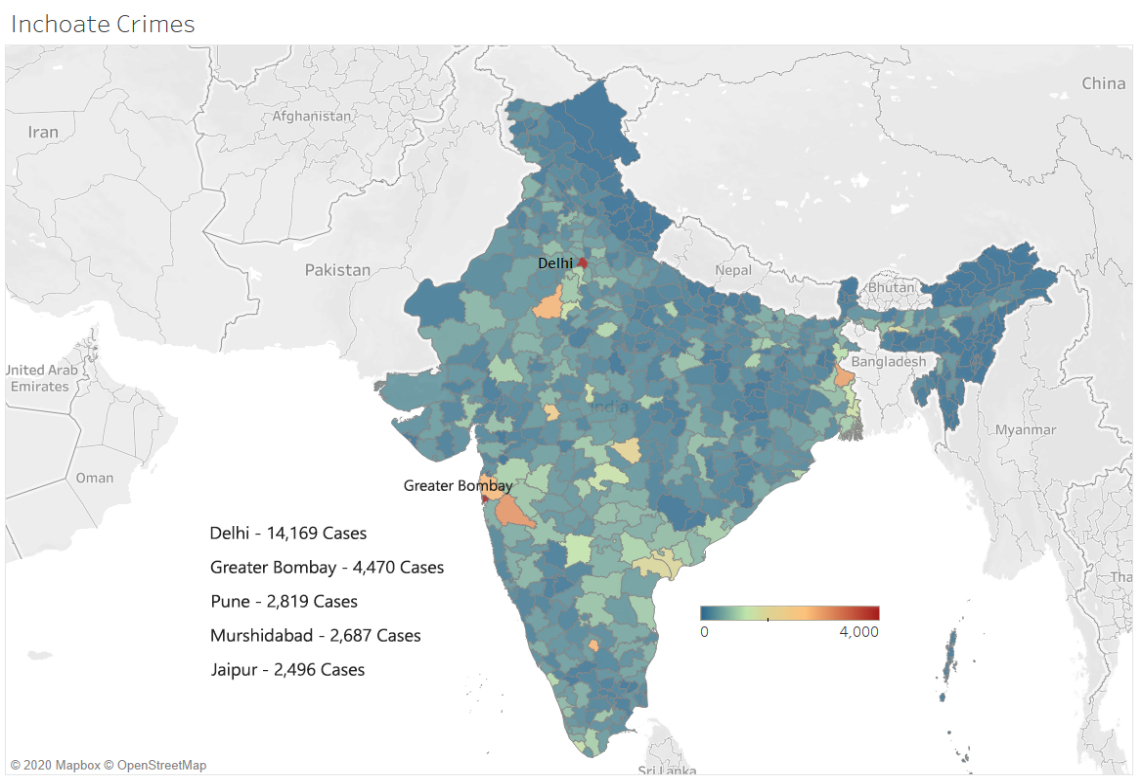


Fig.5. - Visualization of Inchoate Crimes in 2014 using Heat-Map.

* + 1. *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 191.53% increase in crime rate, noticed in the area graph fig. 7. 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.6. 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.

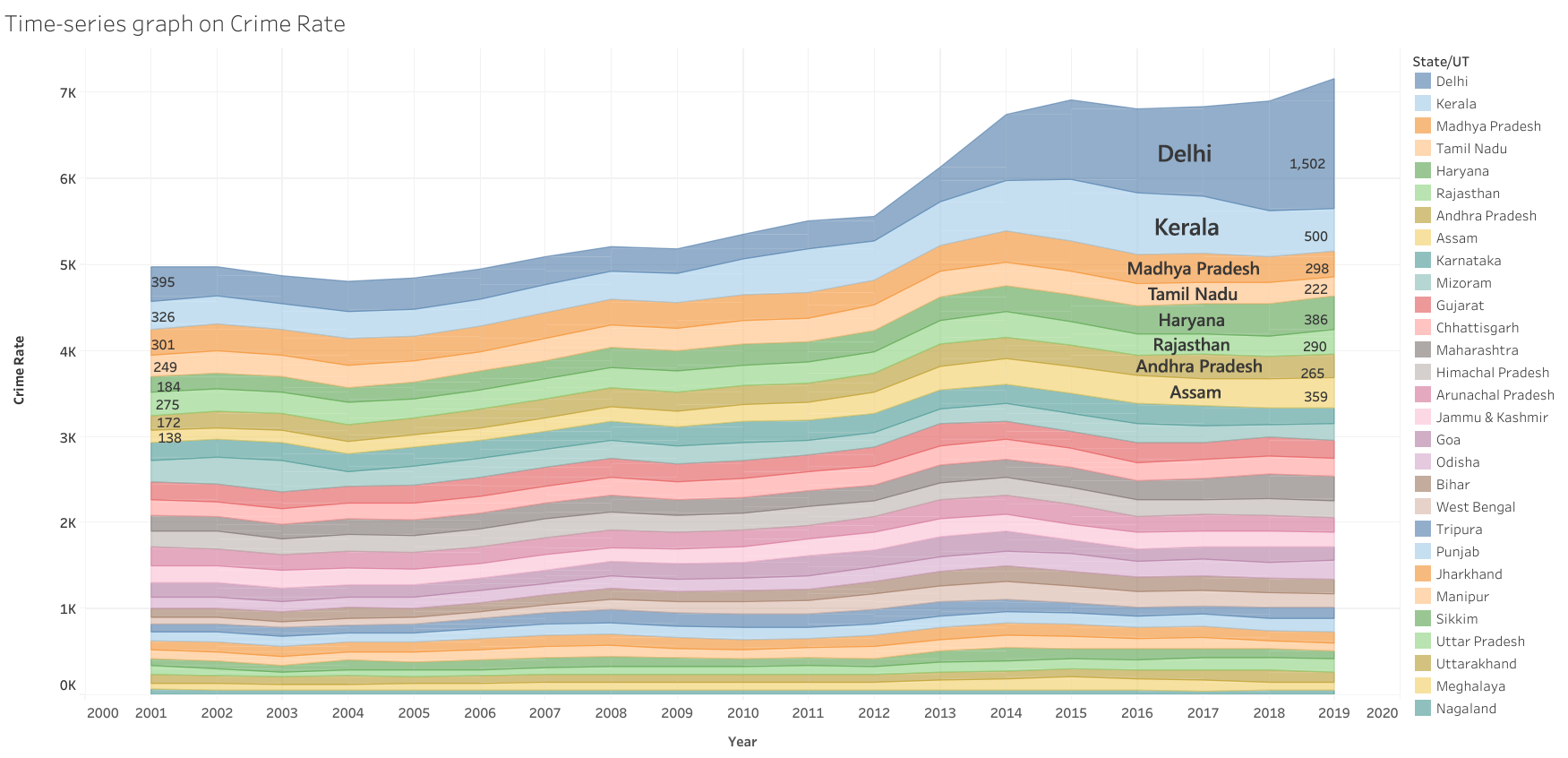


Fig.6. – Crime Rate visualization on the States of India and Delhi (Capital of India)

(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.

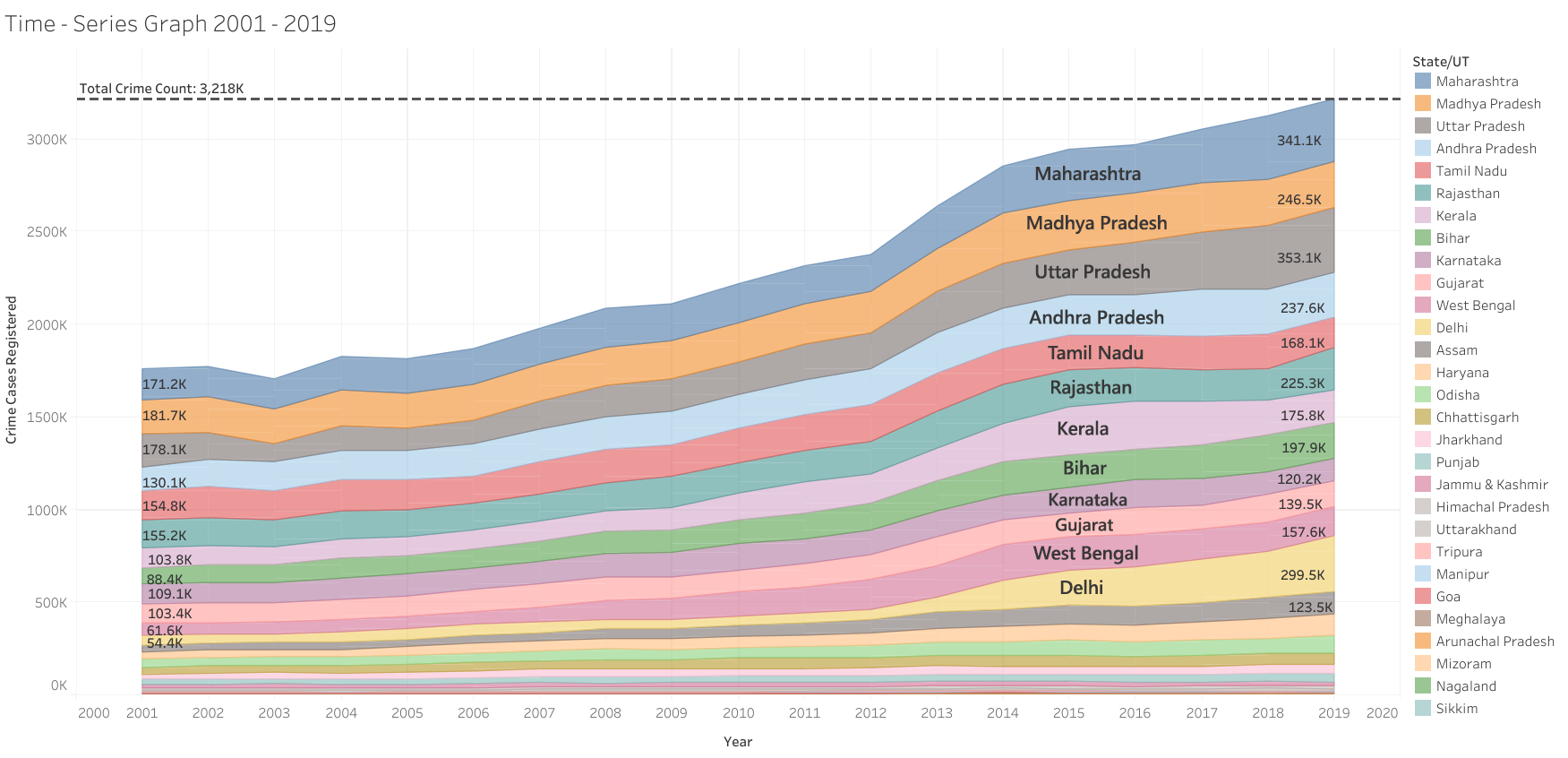


Fig.7. – Total Crime Cases visualization on the States of India and Delhi (Capital of India)

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

* + 1. *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 better comparison. These property crimes which are utilized are dacoity, theft, robbery, burglary, and other offenses in which property is lost. In Figure 8 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).

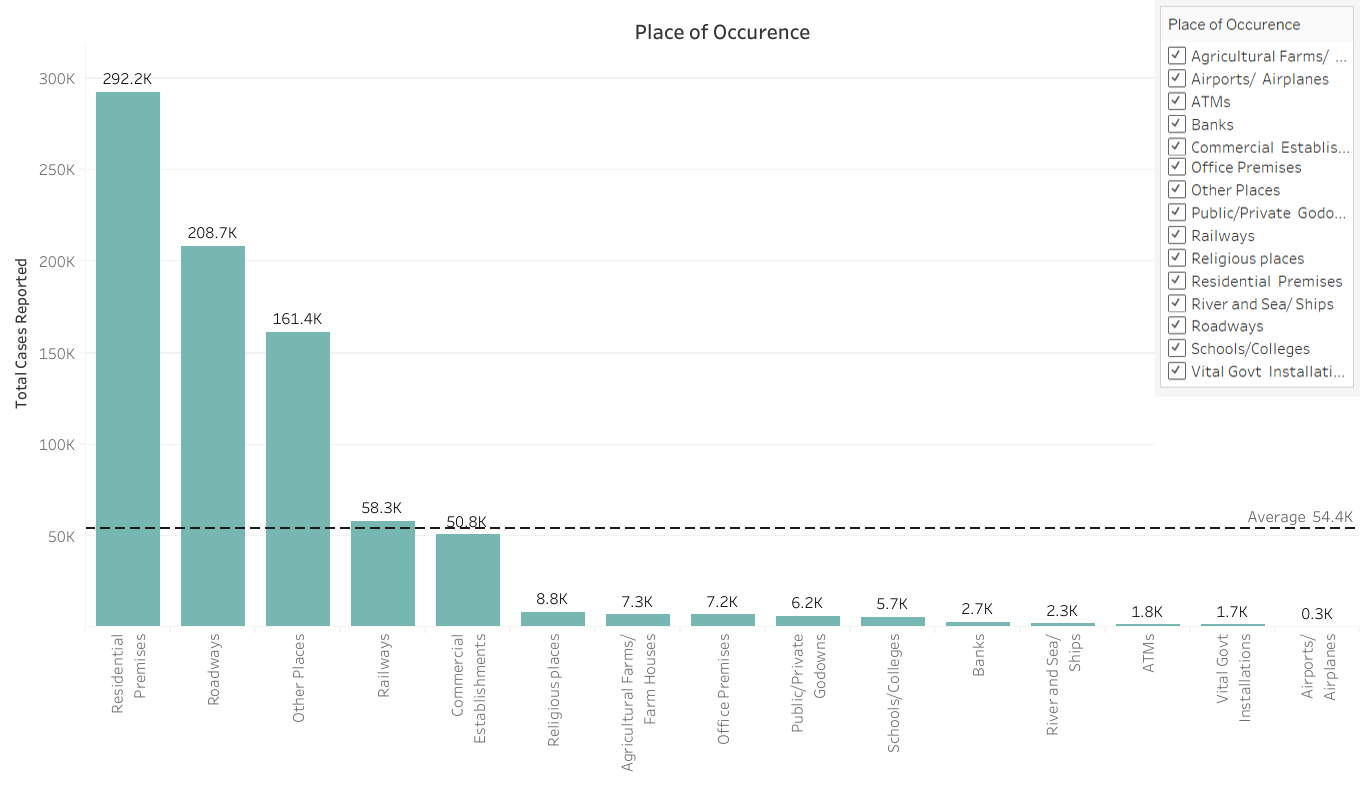


Fig.8. – Visualization based on place of occurrence 2019

* + 1. *Crime against some specific kinds of people who get the most targeted*

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 parameters 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 9 and it’s been done for three consecutive years (2017, 2018 & 2019). From this, we can calculate 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 list.

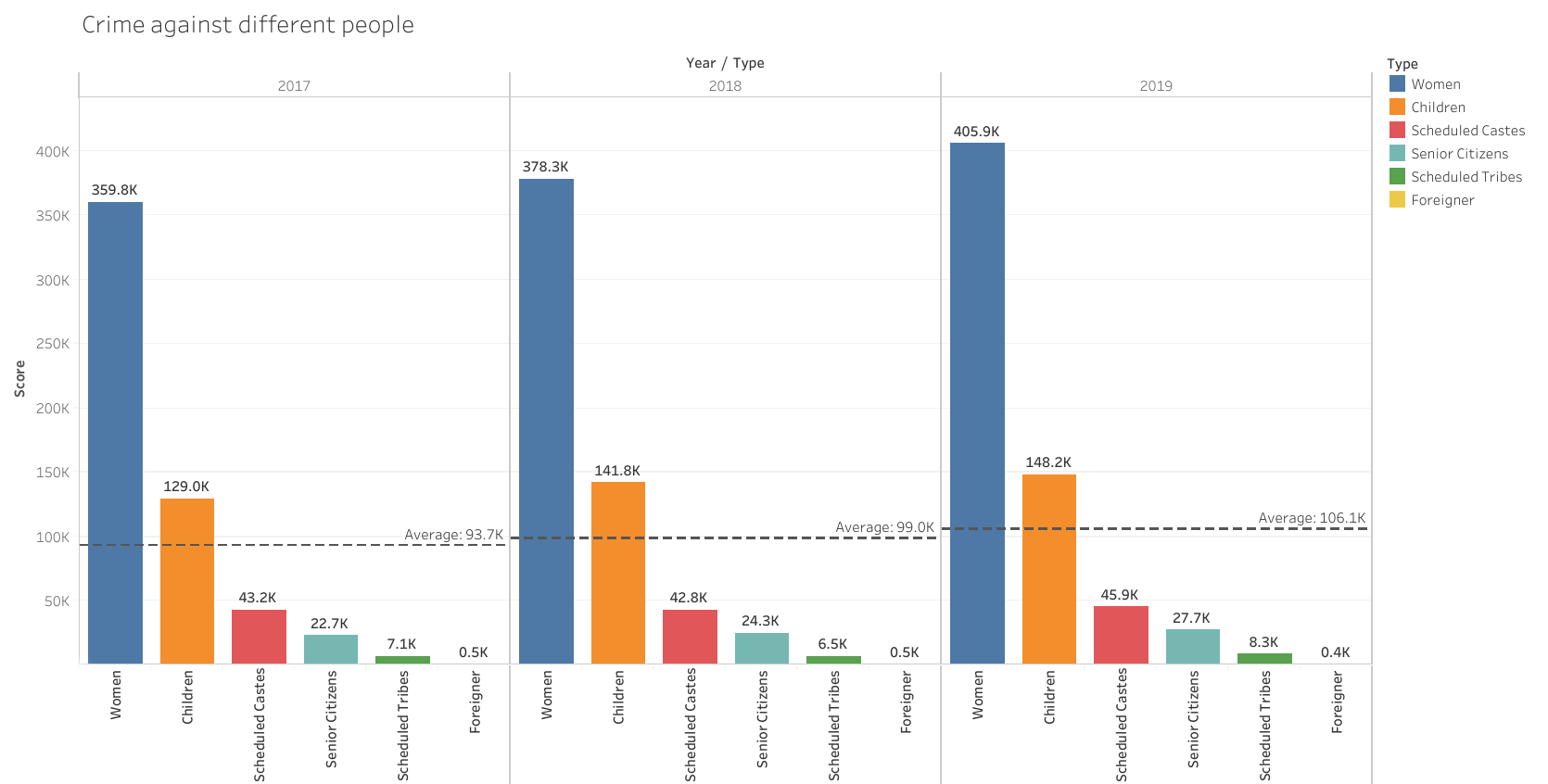


Fig.9. – Visualization based on different kinds of people

* 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 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.

* + 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 [21].

* 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 [22] and improved by Bezdek [23]. 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.

Where, ‘*m*’ is the fuzziness index which is greater than 1, ‘*N*’ is the number of data points, ‘*C*’ is the number of centroids, ‘*mij*’ represents the membership of *ith* data to *jth* cluster centroid, ‘*xi’* is the *i*th of *d*-dimensional measured data, ‘*cj’* is the d-dimension centre of the cluster, and ‘’ is the Euclidean Distance between ith data point and jth cluster centre.

Following are the steps in algorithm [24][21]:

**Step 1:** Initialize *U* = [*mij*] matrix, *U (0)*,

**Step 2:** At *k-step*, calculate the centres vectors *C (k)* = [*cij*] with *U (k)*,

**Step 3:** Update *U (k), U (k+1)*,

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

**Step 5:** The Fuzzy partitioning [24] is carried out through an iterative optimization of the objective function shown in Eq. (1), with the update of membership *uij* and the cluster centers *cj* by using Eq. (3) and Eq. (2)

**Step 6:** This iteration will stop when.

where, *e* 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 *Jm*.

So, these were the steps of the algorithm. To partition the clusters there’s 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 [16], 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 [7], as shown below,

*skfuzzy.cluster.cmeans(data\_set, n\_centers, 2, error=0.005, maxiter=1000, init=None)*

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 [25], which is shown in Fig. 12. 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. 11 through a line graph for better analysis of the highest value. It can be seen that 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’s why we have shown the representation of data points with 5 clusters formed in Fig. 12.

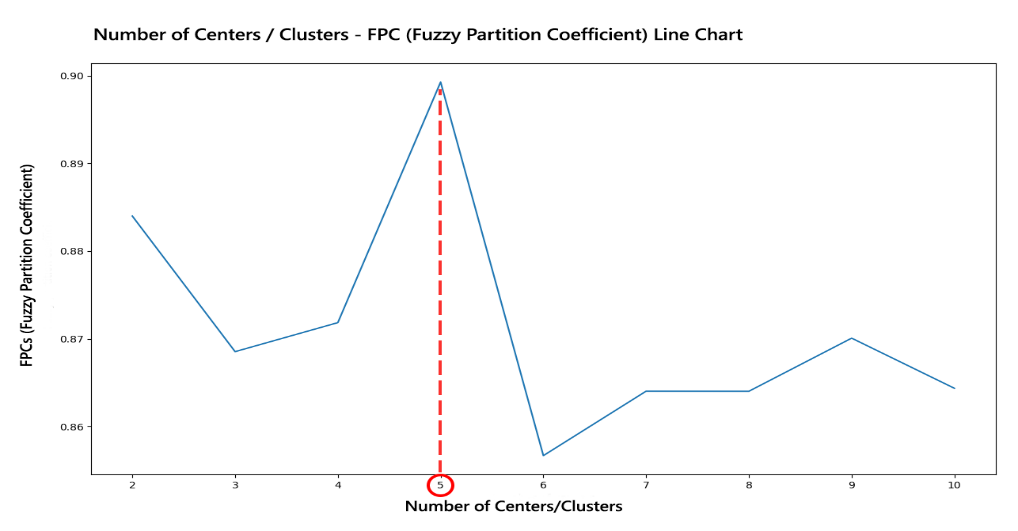


Fig. 11. – Fuzzy Partitioning Coefficient for Population – Crime Rate Chart

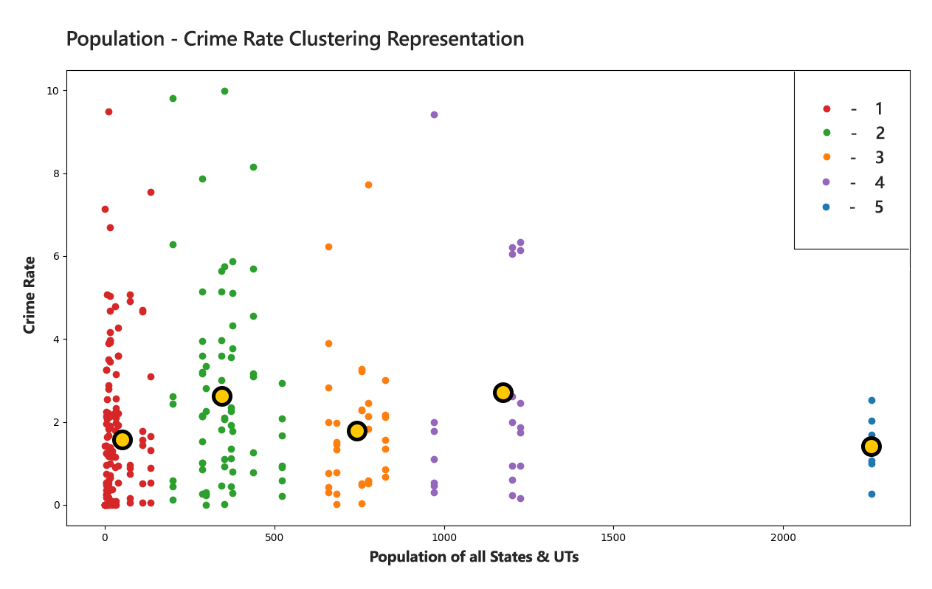


Fig. 12. - Population – Crime Rate Fuzzy C-Means Clustering

By carefully looking at Fig. 12, 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’s just one cluster (the blue one) towards the right. In fact, in the blue cluster, there’s 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’s 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.

The reason the clustering can be said to be better than K-Means clustering algorithm as K-Means is a Hard Clustering type algorithm and Fuzzy is not, which means that with every iteration the data points can belong to more than one clusters. This is the reason that Fuzzy C-Means is also known as *‘Soft K-Means Algorithm’.*

* + 1. *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 has to 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 [26]. It is better shown in Fig. 13 below, where TP is True Positive, TN is True Negative, FN is False Negative, and FP is False Positive.

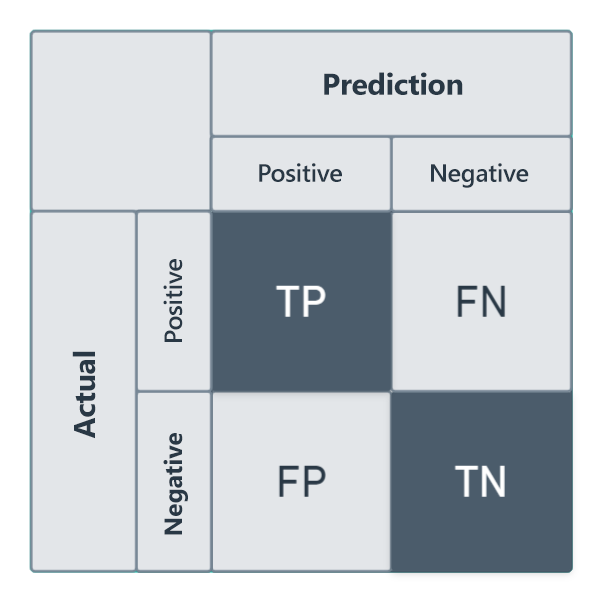


Fig. 13. – Confusion Matrix

* ***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 [27][28]. The score is given in percentage.

* ***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) [27][28].

* ***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) [27][28].

* ***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.

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:

1. ***Naïve Bayes Algorithm***

This algorithm is based upon the Bayes Theorem [29], 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,



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 [30].

1. ***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 [30]. This algorithm was discovered to predict the target column, after splitting the data set into random training and test sets.

1. ***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 [31]. 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 [16]. 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 [4]. 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 taken into account 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 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.

**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 | 90.83% | 87.55% | 90.83% | 88.81% |

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. In Decision Tree while implementing the arguments were,

*DecisionTreeClassifier(criterion="entropy", random\_state=0, max\_depth=3, min\_samples\_leaf=5)*

For Naïve Bayes, *GaussianNB* was implemented and for Precision Score, Recall Score, and F1 Score average was ‘*weighted*’, same as in other algorithms [32]. In Random Forest Classifier 5 Decision Trees were assigned,

*RandomForestClassifier(n\_jobs=5, random\_state=1000)*

Confusion Matrix for Random Forest Classifier is shown below (in Table 4) 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.

**Table 4** Confusion Matrix for Random Forest

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 5 | 1 | 0 | 0 | 0 |
| 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 8 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 14 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 9 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 31 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |

1. 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 it can be seen in Fig. 12 that five clusters were formed with an FPC of more than 0.89 and lesser than 0.90.

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. 2, Fig. 3, Fig. 4, and Fig. 5). From both the Time-Series Area Graphs (Fig. 6 and Fig. 7) 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. 8), 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. 9), 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%.

1. References:

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

[2] 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.

[3] Tableau Software is an interactive data visualization tool developed by Sales Force, Retrieved from <https://www.tableau.com/>

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

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

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

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

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

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

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

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

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

[19] Heat-Map Definition, Retrieved from <https://en.wikipedia.org/wiki/Heat_map>

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

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

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

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

[24] A Tutorial on Clustering Algorithms, Available at: <https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm>

[25] Matplotlib Module in Python, Retrieved from <https://matplotlib.org/>

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

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

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

[29] Joyce, James (2003), ["Bayes' Theorem"](https://plato.stanford.edu/archives/spr2019/entries/bayes-theorem/), in Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 ed.), Metaphysics Research Lab, Stanford University, retrieved 2020-01-17

[30] 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.

[31] 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.

[32] 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.