MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Analysis of Crime in a Country Using Data Mining and Analysis

By

Maria Lisa D Silva (200953012)

Vedika Malviya (200953102)

INTRODUCTION

Crime is a behavioural disorderliness which is a result of societal, economical, and environmental factors. Currently, analysis of crime is becoming increasingly important and has grown to be one of the most popular and widely accepted disciplines in the world. There is an increase in violence against women each day, and it can take several forms. Rape, Sexual harassment, Dowry, abduction, etc., are some of the most committed crimes against women. These crimes result in life-long trauma or deaths. Safety and Security of all, not just women, is an absolute right. To control the rising crime rate, it is necessary to extract and analyse all relevant data on a regular basis and take the necessary steps to ensure the safety of all. Government plays a vital role in laying down rules and regulations that protect the rights of all the country's citizens.

The generation of this model was done by using the crime data from India for a duration of 10 years, from 2001 to 2010. Then a clustering algorithm is selected based on the comparison of the clustering methods used in this paper.

LITERATURE SURVEY

1. Rasoul Kiani et al. [1] aims on categorizing the clustered crimes on the grounds of occurrence frequency during different years. The Genetic Algorithm (GA) is used for optimizing Outlier Detection operator parameters using the RapidMiner tool. K-means is used as the clustering/partitioning algorithm for large datasets as it is simple and has less computational complexity. But the ability to determine the number of clusters by the user, affectability from outlier data, etc. Decision tree learning, neural network, KNN, Naïve Bayes method, and support vector machine are different algorithms that are used for classification. GA gives an optimal solution. Accuracy and classification error with optimization of the Outlier Detection operator's parameters are 91.64% and 8.36% respectively and 85.74% and 13.26% without optimization.

2. H. Benjamin Fredrick David at al. [2] intends to carry out a survey on supervised learning and unsupervised learning techniques that have been applied to criminal identification . He presents the survey on Crime analysis using the K-Means Clustering algorithm and crime prediction using several Data Mining techniques . The supervised learning and clustering algorithms were used to identify the crime patterns which are used to commit crimes knowing the fact that each crime has certain patterns . Naïve Bayes was used to predicting possible suspects from the criminal records . The system is a multi-agent system made with managed Java Beans. It makes it easy to encapsulate the requested entities in work into objects and returns it to the bean for exposing properties . The accuracy of prediction based on optimized parameters is 91.64% and the non-optimized parameter is 85.74%. The Generic Algorithm optimizes the parameters.

3. Devendra Kumar Tayal et al. [3] proposes a design and implementation method for crime detection and identification systems for Indian cities using data mining methods such as data extraction (DE), data pre-processing (DP), clustering, Google map representation, classification, and WEKA implementation . The JAVA-based graphical user interface for crime verification of K-Means results was used. This in turn would reduce the cost and time of crime investigation . WEKA verifies high accuracy of 93.62% and 93.99% in the formation of two crime clusters using selected crime attributes. Investigating agencies can utilize the proposed data mining tool to ease their crime investigation process . Data privacy needs to be improved.

4. A. Malathi et al. [4] aims at developing analytical data (from 2006 to 2010) mining methods that can systematically address the complex problem related to various forms of crime . Obtained clusters with different clustering algorithms which are then compared to select the most appropriate one giving crime prediction in both space and time to help police departments in tactical and planning operations. The best of the 4 distance measures- Euclidean, Manhattan, Mahalanobis, and Pearson is considered . Clustering and Apriori were used. Working with a large dataset proved to be slow, inefficient, and

used a lot of resources which were tackled upon enhancement. The KNN Imputation method is enhanced by proposing a new distance metric and by using LVQ (Learning Vector Quantization) approaches. Combined with generalized relevance learning to perform the classification and missing value treatment simultaneously producing a speed and accuracy efficient model.

5. Rizwan Iqbal et al. [5] applies classification to a dataset with criminal records to predict the ‘Crime Category’ for different states of the United States of America. The two different classification algorithms used are Decision Tree and Naïve Bayesian and the data mining tool used is WEKA. The results from both the algorithms will be evaluated on three performance measurements- Accuracy, Precision, and Recall values for Naïve Bayesian are 70.8124%, 66.4%, and 70.8%, respectively , and 83.9519%, 83.5%, and 84% respectively for Decision Tree. Hence, Decision Tree outperformed Naïve Bayesian and manifested higher performance. This tool is of great advantage to law enforcing agencies to effectively fight crime and the war against terrorism.

6. Shiju Sathyadevan et al. [6] suggested the Apriori algorithm for discovery of the trends and patterns in crime. This algorithm is also used to determine association rules highlighting general trends in the database. This paper has also proposed the naïve Bayes algorithm to construct the model by training crime data. After testing, the result showed that the Naive Bayes algorithm showed 90% accuracy .

7. Khushabu A. Bokde et al. [7] mainly focuses on methods like Clustering,crime Analysis and Clustering by K-means algorithms. Some of the purposes of crimal analysis are Extraction of crime patterns by crime analysis and, based on available criminal information, crime recognition. Clustering means dividing a set of data or objects into several clusters. Thereby cluster is composed of a set of similar data the same as a group. K-means is the simplest and most commonly used portioning algorithm among the clustering algorithms in scientific and industrial software . The methods used are Crime, Clustering, K-Means Algorithm. The primary intent of this paper is to categorise clustered crimes based on occurrences frequently during different years. Data mining is used in terms of analysis, investigation, and discovery of patterns for crime occurrences. From the accuracy result, crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis . From the accuracy result, crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis .

8. Ayisheshim Almaw et al. [8] suggests various algorithms for sorting in problem resolving that are nominated based on the applicable requirements in crime data prediction. One technique may provide a better accuracy value than other techniques for solving particular problems. Some papers introduced combined different models to achieve better performance which overcomes the individual models called ensemble learning . The methods used are Crime prediction, naïve Bayes, J48 , Artificial network. This paper investigates several data mining algorithms and ensemble learning which are applied to

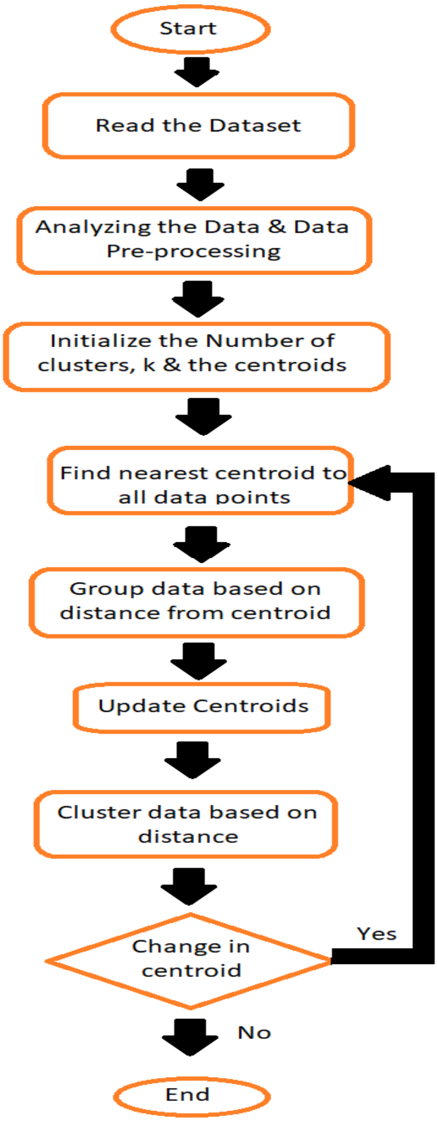
crime data mining . This paper delivers a reasonable investigation of data mining techniques and ensemble classification techniques for the discovery and prediction of upcoming crimes. A Survey is conducted so that crime forecasting can be improved by the use of efficient data collection and data mining strategies . In future, it will predict what kind of crime might occur next in a particular district within a specific period and identify the season and time factor at which crimes occur more frequently .

9. Ginger Saltos et al. [9] uses the systematically recorded crime data from the police of many years. In the last decades, there has been a surge of Open Crime Data and apps or web-based applications displaying crime statistics on maps, both from official sources, such as police UK and other sources using the same official data. This paper investigates and predicts many types of crime and discusses their applicability. The methods used are Crime prediction, Data mining, Open data, regression, decision trees, instance-based learning. This paper explores models predicting the frequency of several types of crimes. This paper uses three kinds of algorithms: instance-based learning, Decision trees, and regression . The experiments were conducted using the SCIAMA High-Performance Computer Cluster at the University of Portsmouth and the WEKA software. Further experiments can be conducted to investigate other aspects, such as the time frame for prediction, the amount of data necessary for reliable prediction models, and predictive models for particular types of crime .

10. Benjamin Frederick David. H et al. explains that criminology is a procedure utilized to identify the characteristics of crime and crime identification and by using these data mining algorithms, he was able to build crime reports and help identify criminals much more efficiently than any human being ever could . While escaping the crime scene, these criminals leave behind some traces and evidences that can be utilized as clues to identify them as criminals. This process is used to determine criminals based on clues or information provided by the local population . This document uses a few crime analysis methods: text contents and the NPL-based techniques, and crime patterns. The methods used are Criminology, Criminal Analysis, Criminal Prediction, and Data Mining. Utilizing the existing data sets, the extraction of new information is forecasted. Offenders can also be speculated based on the crime data. The main objective of this work is to study the techniques of supervised learning and unsupervised learning that have been applied for the identification of criminals . The quantitative analysis generated outputs showing an growth in the level of classification accuracy due to the usage of GA for optimization of the parameters. In the future, there is a proposal to apply another classification algorithm on criminal data to improve the accuracy of the predictions.

METHODOLOGY

K-Means Algorithm:


Start
Read the Dataset
Analyzing the Data & Data Pre-processing
Initialize the Number of clusters, k & the centroids
Find nearest centroid to all data points
Group data based on distance from centroid
Update Centroids
Cluster data based on distance
Change in centroid
Yes
No
End

1. **Importing the dataset (read the dataset) :**
   Crime Analysis dataset was collected from Kaggle for implementation of k-means algorithm to analyse the data. The code was written and executed in Jupyter Notebook interface. Some of the libraries used in this project are numpy, pandas, matplotlib, seaborn, sklearn, etc.
   After importing the dataset, we compare analyse the dataset.

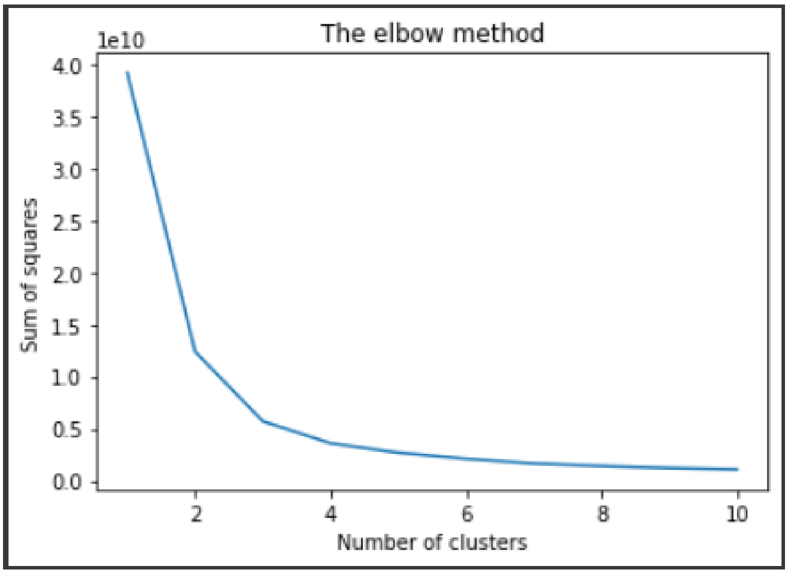
2. **Analysing the Data and Data preprocessing:**

   Following this, the first step towards creating the model is Analysing the data and the Data Pre-processing step. Data pre-processing is a process that consists of data cleaning, data integration and data transformation . This step checks for duplicate values, null values and replaces them with average value, converts nominal values into numeric values, and removes unwanted attributes according to the requirement of the algorithm.

3. **Initialize the number of clusters , k and the centroids:**

   For a given set of objects , clustering is a process of discovery of classes, wherein the objects are organized into clusters and the classes are unknown in advance.

   In this step, the Elbow method was utilized to find the optimal number of clusters, k needed for the dataset. This step is necessary to pool the dataset to generate meaningful results for analysis.



   Attributes taken in to consideration are:

   Cases_Chargesheeted , Cases_Convicted, Cases_Reported.

   Along with other attributes :

   Area_Name, Year, Group_Name

   We then find the random centroids for the given number of clusters after scaling the data.

4. **Find nearest centroid to all datapoints and group the data based on this distance from the centroid:**
   We use Euclidean distances to find this distance between the centroid and each datapoint.

$$d(m,n)= \mathrm{sqrt}((n_1-m_1)^2 + (n_2-m_2)^2)$$

This will result in the formation of k predefined clusters.

5. **Update the Centroids and cluster based on the distances:**

   We now compute the mean of every cluster as a new centroid and update the centroid with this mean value of each cluster.

   Now repeat the previous step with a new center for the cluster.

   Repeat these steps until the new centroid values are the same as the old centroid values and until the maximum number of iterations is complete.

6. **Visualization:**

   The clusters are systematically visualized using the Scatter plot for the above implemented k-means algorithm.

DBSCAN algorithm:

The DBSCAN algorithm is an unsupervised data clustering algorithm. The main objective of DBSCAN algorithm is to identify and locate high-density regions separated by low-density areas in the form of clusters formed on the basis of distance measurement . Therefore, the clusters formed by DBSCAN may acquire any shape. Two main parameters for DBSCAN are eps and min_samples. Eps examines the farthest distance between 2 samples that are to be taken into consideration in a cluster formed in the DBSCAN algorithm. The min_samples parameter refers to the minimum number of samples that are to be considered as a point in a single cluster . It is suitable for data which contains the clusters of similar density .

1.Data collection
Spatial data bases and whether the data contains outliers and noise.

2. Data recovery
Data recovery in algorithm code.

3. Pre-processing and data cleaning
It comprises of correcting the data types of the columns, changing the date and time to the 24-hour standard format, splitting of the time ,date, month , and year and further storing this data individually for analysis, etc. The clean data can later be preserved in local storages to bypass latency each time the user runs the application .

4. Data visualization
The Web application contains a number of visualisations to assist the user in data analysis and visualisation. Some of them are bar charts, line plots, pie charts, maps,

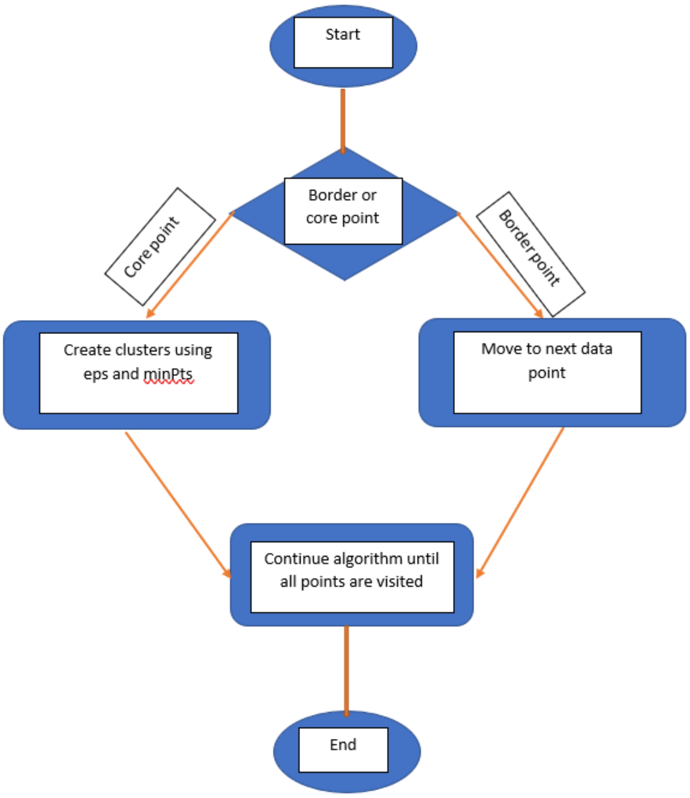
etc. The previously mentioned visualization techniques helps in easy understanding and interpretation of the huge data set in a very effective way.

5. Data analysis

It includes identifying a criminal from their case number, analysing the places in the city that were affected the most, comprehending when the intensity of execution of a crime is most likely, type of crimes committed in different places, and further analysing the foci of crime throughout the city.

6.Clustering

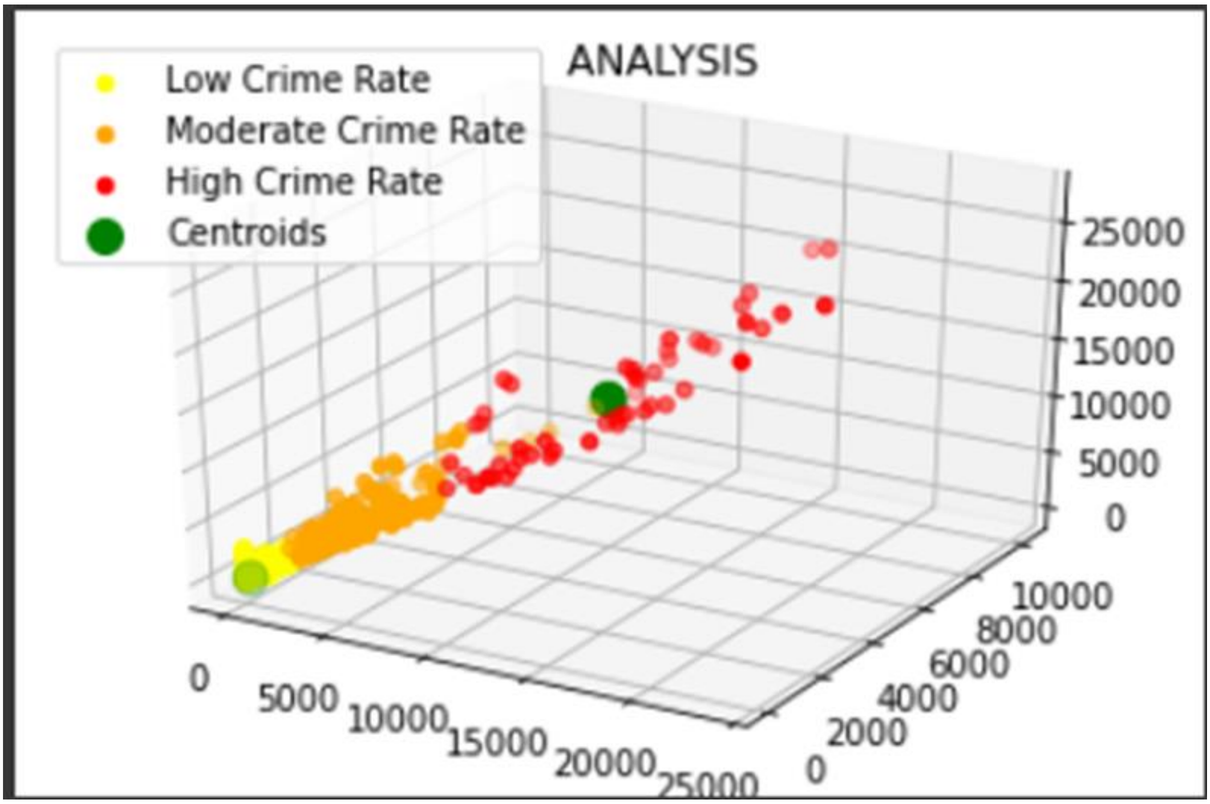
DBSCAN, a type of clustering algorithm, helps to detect the clusters to which each type of crime belongs to on the basis of location. It groups the closest points in the data together and accordingly group them together to form a cluster . This algorithm has the ability to identify the dense region in the given data set.


Start
Border or core point
Core point
Border point
Create clusters using eps and minPts
Move to next data point
Continue algorithm until all points are visited
End

RESULTS AND DISCUSSION

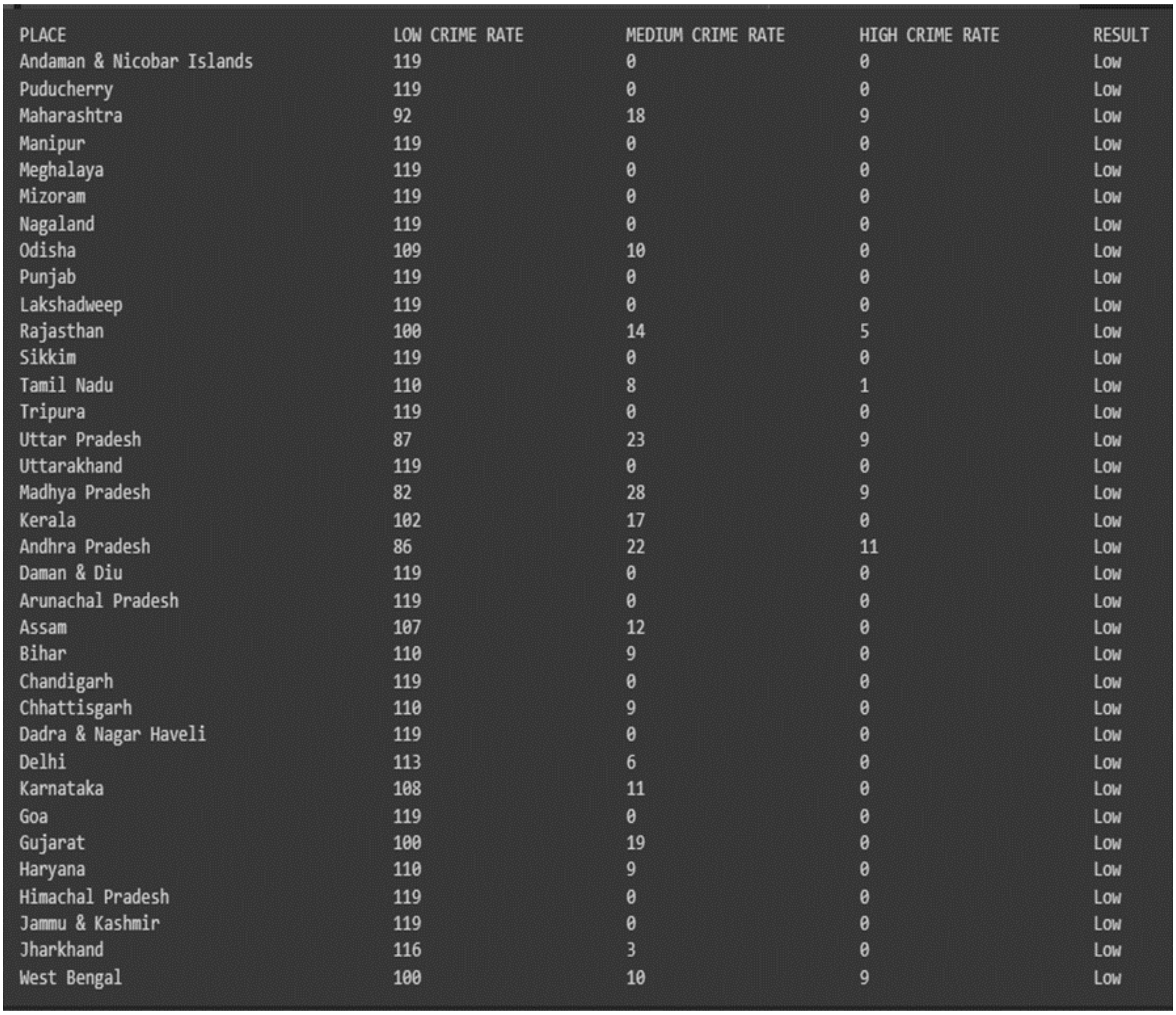
K-Means:

The optimal number of clusters as found out by the elbow method is 3.Hence that dataset can be divided into three clusters namely, low crime rate areas, Moderate crime rate areas and High crime rate areas



On comparing the results from this clustering with different attributes like Area_Name, Group_Name, and Year, we observe that the trend in the Crimes that take place in India follow a certain pattern. We take into consideration the above mentioned attributes for further analysis as follows.

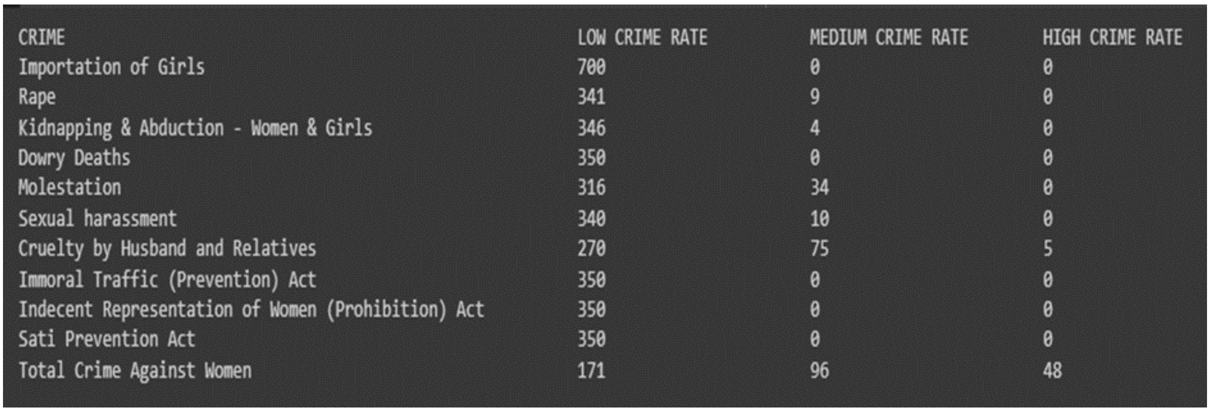
1.Area-wise Pattern:



```
PLACE                       LOW CRIME RATE    MEDIUM CRIME RATE    HIGH CRIME RATE    RESULT
Andaman & Nicobar Islands   119               0                    0                  Low
Puducherry                  119               0                    0                  Low
Maharashtra                 92                18                   9                  Low
Manipur                     119               0                    0                  Low
Meghalaya                   119               0                    0                  Low
Mizoram                     119               0                    0                  Low
Nagaland                    119               0                    0                  Low
Odisha                      109               10                   0                  Low
Punjab                      119               0                    0                  Low
Lakshadweep                 119               0                    0                  Low
Rajasthan                   100               14                   5                  Low
Sikkim                      119               0                    0                  Low
Tamil Nadu                  110               8                    1                  Low
Tripura                     119               0                    0                  Low
Uttar Pradesh               87                23                   9                  Low
Uttarakhand                 119               0                    0                  Low
Madhya Pradesh              82                28                   9                  Low
Kerala                      102               17                   0                  Low
Andhra Pradesh              86                22                   11                 Low
Daman & Diu                 119               0                    0                  Low
Arunachal Pradesh           119               0                    0                  Low
Assam                       107               12                   0                  Low
Bihar                       110               9                    0                  Low
Chandigarh                  119               0                    0                  Low
Chhattisgarh                110               9                    0                  Low
Dadra & Nagar Haveli        119               0                    0                  Low
Delhi                       113               6                    0                  Low
Karnataka                   108               11                   0                  Low
Goa                         119               0                    0                  Low
Gujarat                     100               19                   0                  Low
Haryana                     110               9                    0                  Low
Himachal Pradesh            119               0                    0                  Low
Jammu & Kashmir             119               0                    0                  Low
Jharkhand                   116               3                    0                  Low
West Bengal                 100               10                   9                  Low
```

From the above table, we observe that the all the major cities in India fall into the category of low crime rate, i.e., most of the areas ,on the basis of the three different attributes mentioned before, on clustering with k=3 fall into this category.

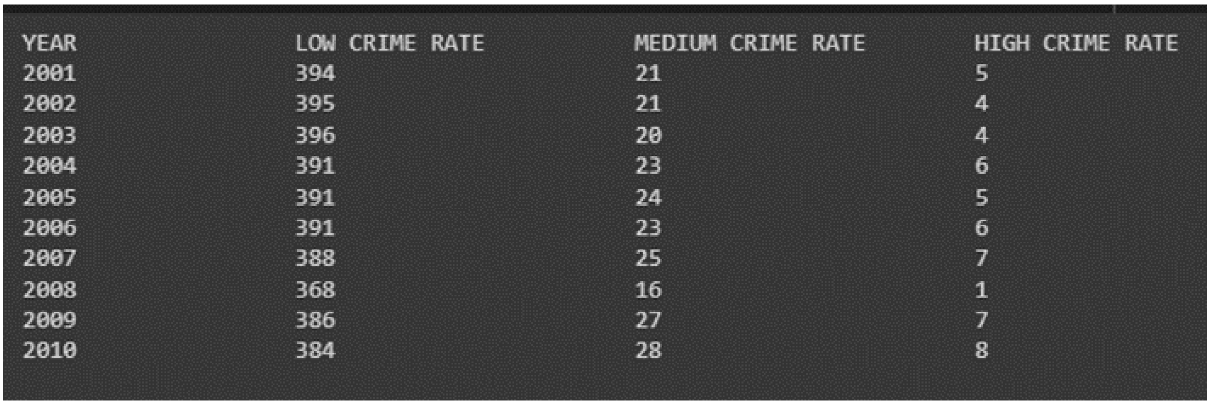
2.Crime Type pattern:



```
CRIME                                                  LOW CRIME RATE    MEDIUM CRIME RATE    HIGH CRIME RATE
Importation of Girls                                   700               0                    0
Rape                                                   341               9                    0
Kidnapping & Abduction - Women & Girls                 346               4                    0
Dowry Deaths                                           350               0                    0
Molestation                                            316               34                   0
Sexual harassment                                      340               10                   0
Cruelty by Husband and Relatives                       270               75                   5
Immoral Traffic (Prevention) Act                       350               0                    0
Indecent Representation of Women (Prohibition) Act     350               0                    0
Sati Prevention Act                                    350               0                    0
Total Crime Against Women                              171               96                   48
```

This table shows the prominent crimes in each cluster i.e., in cluster 0(Low crime rate), Importation of Girls is the most prominent crime. Similarly, in cluster 1(moderate crime rate) and 2(high crime rate), Other Crimes (total crime against women) are prominent.

3. Year-wise pattern:



```
YEAR                LOW CRIME RATE           MEDIUM CRIME RATE        HIGH CRIME RATE
2001                394                      21                       5
2002                395                      21                       4
2003                396                      20                       4
2004                391                      23                       6
2005                391                      24                       5
2006                391                      23                       6
2007                388                      25                       7
2008                368                      16                       1
2009                386                      27                       7
2010                384                      28                       8
```

The number of crimes each year, from 2001 to 2010, in each of the groups is shown in the table above. 2008 has the least number of crimes, while 2010 has the maximum number of crimes.

<u>Silhouette Score for K-Means:</u>

The efficiency of unsupervised learning algorithms such as the K-means algorithm can be determined by measuring the performance of the classification problem using a value called the Silhouette Score.

Silhouette Score,S, for each group can measure how similar an object is to its own cluster or ‘cohesion’ in contrast to the other clusters or ‘separation’. The maximum value is for Silhouette score is 1 and the minimum value is -1.

A high value of S implies that the object taken into consideration matches well to its own cluster and matches poorly to its neighboring clusters.

$$S = (q-p)/\max(p, q)$$

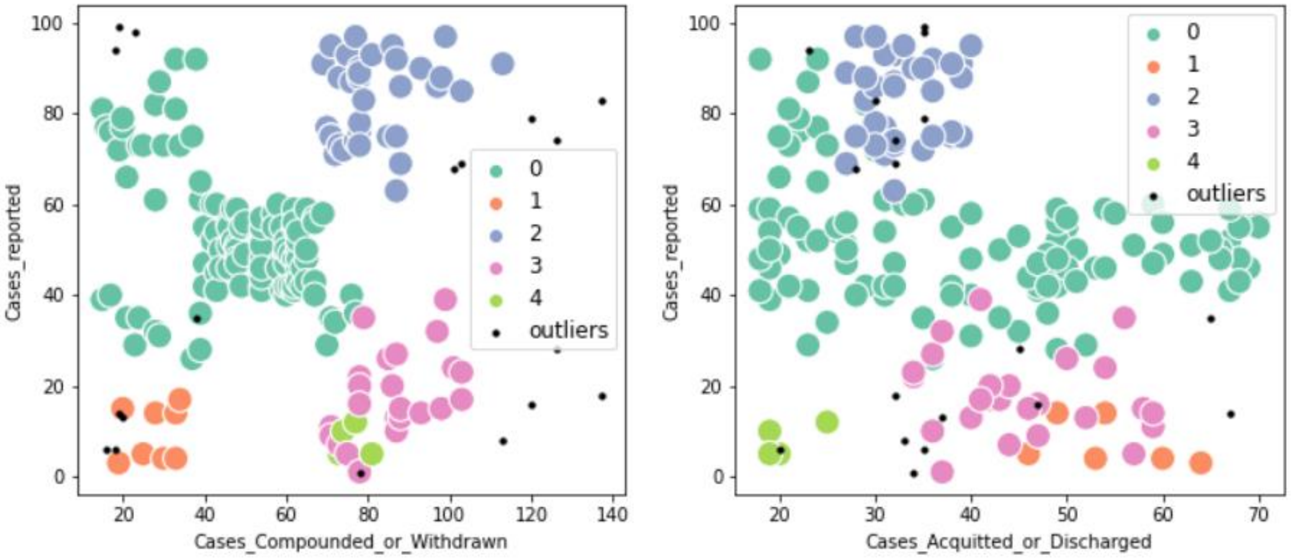
Here, ‘p' is the mean intra-cluster distance of each sample and 'q' is the mean distance of nearest cluster for each sample.

We obtained a Silhouette score of 0.7444622007021535 which is close to 1. It denotes that the data point is very compactly packed within the cluster to which it belongs to and is far away from all the other clusters. Hence the results obtained from k-means clustering are valid and accurate.

DBSCAN:

It'll fit X_train in the DBSCAN algorithm with eps 12.5 and min_sample 4. Later, it'll generate a DBSCAN_dataset from X_train and generate a 'Cluster' column using clustering.labels_.

We will use value_counts() to visualize the distribution of clusters and convert it to a data frame.



Finally, we understand that K-Means proves to be more efficient than DBSCAN for the given dataset due to the enormous size of the dataset.

CONCLUSION

This paper focuses on crime analysis and crime mapping of the Indian Crime data set. It helped us analyse big data on crimes against women to identify the patterns, trends and hotspots of crime. This is a major issue encountered by law enforcement and intelligence gathering organizations in India.

Our tool, which applies the K-Means clustering algorithm to the dataset describing crimes against women, will allow agencies to efficiently and inexpensively cleanse, characterize, and analyse crime data to identify patterns and trends in the different regions of the country based on the intensity, year of occurrence and type of crime. We aim at further optimizing this tool to efficiently cluster the data based on the age of the victims and survivors, population of any given area along with its literacy rate. Furthermore, we understood how the DBSCAN algorithm used for clustering will help reduce crime by assisting law enforcement agencies in the decision making process and authorities in crime investigations. This algorithm cannot be used for large data and will consume a lot of memory for samples that are enormous. In the future, more optimized algorithms other than DBSCAN may be used to increase the efficiency with which the application loads .

REFERENCES

[1] Kiani, Rasoul, Siamak Mahdavi, and Amin Keshavarzi. Analysis and prediction of crimes by clustering and classification. *International Journal of Advanced Research in Artificial Intelligence* 4.8 (2015): 11-17.

[2] David, H., and A. Suruliandi. SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES. *ICTACT journal on soft computing* 7.3 (2017).

[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] Malathi, A., and S. Santhosh Baboo. Evolving data mining algorithms on the prevailing crime trend–an intelligent crime prediction model. *Int J Sci Eng Res* 2.6 (2011).

[5] 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.

[6] Shiju Sathyadevan, Devan M.S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing (IEEE) 2014.

[7] Khushabu A.Bokde, Tisksha P.Kakade, Dnyaneshwari S. Tumasare, Chetan G.Wadhai B.E Student, Crime Detection Techniques Using Data Mining and K-Means, International Journal of Engineering Research & technology (IJERT) ,2018.

[8] Ayisheshim Almaw, Kalyani Kadam, Survey Paper on Crime Prediction using Ensemble Approach, International journal of Pure and Applied Mathematics,2018.

[9] Ginger Saltos and Mihaela Coacea, An Exploration of Crime prediction Using Data Mining on Open Data, International journal of Information technology & Decision Making,2017.

[10] H.Benjamin Fredrick David and A.Suruliandi,Survey on crime analysis and prediction using data mining techniques, ICTACT Journal on Soft computing, 2017.