Analysis and Prediction of Crime Against Women

[1]Asmi Patel, [2]Dhwani Shah, [3]Namrata Poojary, [4]Priya Mishra, [5]Alvina Alphonso
[1,2,3,4]Student,[5]Assistant Professor
Information Technology, St. Francis Institute of Technology, Mumbai,India

ABSTRACT: *The main purpose of this project is to analyze the crimes occurring district wise and predict the further crime trends using the past twelve years of crime records. The data is released by the National Crime Records Bureau (NCRB) and is made available on the Open Government Data Platform India (OGD) . This data can be used to gain various insights about the different crime trends occurring in the country. The types of crimes in the dataset are molestation, rapes, dowry, exploitation, kidnapping, cruelty by relatives etc. K-means clustering data mining approach will be done to obtain crime trends and linear regression will be done to predict crime rates.*

Keywords: *K-means, Linear Regression, Types of crimes, Crime Analysis, Prediction, Visualization, scatter plot, correlation matrix*

I. Introduction

India is a vast country with diverse societies. Position of women has been of great importancesince ancient times in Indian culture. Unfortunately current scenario depicts a different story.According to National Crime Records Bureau , crime against women has significantly increased in recent years such as Dowry, Kidnapping, Insult to modesty, Rape, Assault, Importation of girls, Cruelty by relatives etc.

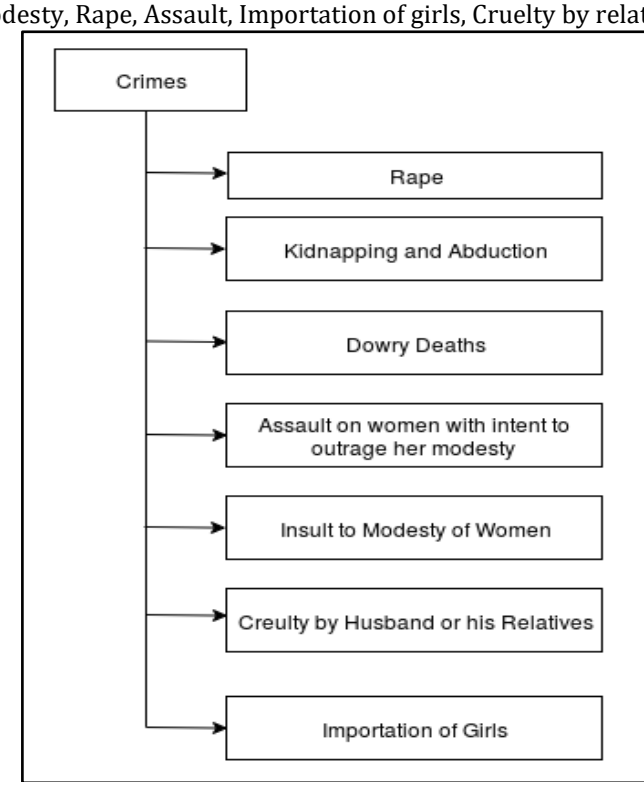

fig 1. crime hierarchy

It has become the topmost priority to the administration to enforce law and order to reduce this increasing rate of the crime against women. In country like India crimes against women has been always a serious issue leading to various legal norms and measures against it. With each passing year crime reports are generated leading to huge amount of data. Such data can be used to generate analysis and statistics which will help government and non-government organizations to initiate certain schemes and policies

accordingly. In 2012, Uttar Pradesh reported the highest cognisable crime rate of 455.8 among States of India, while Nagaland recorded lowest rates 47.7. The rates were calculated by National Crime Records Bureau as the number of incidents per 100,000 of the population. The crime rate differ in each and every state, and hence the measures to be taken also has to be different. This paper primarily focuses on analysis of crime against women which will help in order to generate predictions as an outcome this may help to generate valuable informationfrom the existing data. The initial phase includes data acquisition provided by the government and processing it further for analysis and prediction.

II. Literature Review

Priyanka das et al. has proposed InfoMap algorithm which is used to detect communities, that displays changing trends of various crimes in the different states in India. This paper proposes a graph based approach to find crime similarity between states and to predict the crime patterns of the states. Graph based clustering is used to create a weighted graph and Cosine Similarity is used to calculate the crime similarities among the states. The prediction of the overall crime trend is done by combining the graphs and applying InfoMap algorithm.

Anant Joshi et al. proposed crime analysis using k-means algorithm. This paper includes the process of co-relation discovery and discovery of patterns among fields. In this paper, the Rapid Miner software tool is used to carry out cluster analysis. The purpose of this paper is to analyze the crime which consists of theft, homicide and various drug offences which also include suspicious activities, noise complaints and burglar alarm. K-means clustering data mining approach has been used on a crime dataset from New South Wales region of Australia to generate clusters of homogenous elements. Analysis of these clusters is done using k-means where identification of k is done using silhouette measuring.

Sunil Yadav et al. proposed k-means algorithm to create clusters according to high low values. Patterns, forecast trends, determining relationships, mapping crime trends and possible suspects are discovered with the help of Weka tool and R tool. Result of the k-means is an input to the Apriori algorithm. The association among a number of attributes is discovered with the help of Apriori algorithm. The association between the suspect arrested during the year and the person acquitted in the same year is shown in the result. The data acquisition and data staging were the two biggest hurdle in the project. Various measures can be carried out in order to reduce the crimes.

Shiju Sathyadevan et al. proposed Apriori algorithm which is used to identify the trends and patterns in crime. The algorithm determines association rules which results in highlighting general crime trends in the database. The crime pattern for a particular place is the output of this phase. This paper has also proposed the Naïve Bayes algorithm. This algorithm classifies new article into a crime type to which it fits the best. The prediction has been implemented using the decision tree concept. The crime prone areas are graphically represented using a heat map, which indicates level of activity, usually darker colors to indicate low activity and brighter colors to indicate high activity.

III. Methodology

Modules that are included in the project are:

1. Data Pre-processing
2. Clustering
3. Linear Regression
4. Data visualization

K-means clustering technique is used for crime analysis since it is suitable for clustering large data sets, due to computational complexity. The clustered results makes it easy to identify crime trend over years and it can be used to design precaution methods for future.

Following is the System flow diagram:

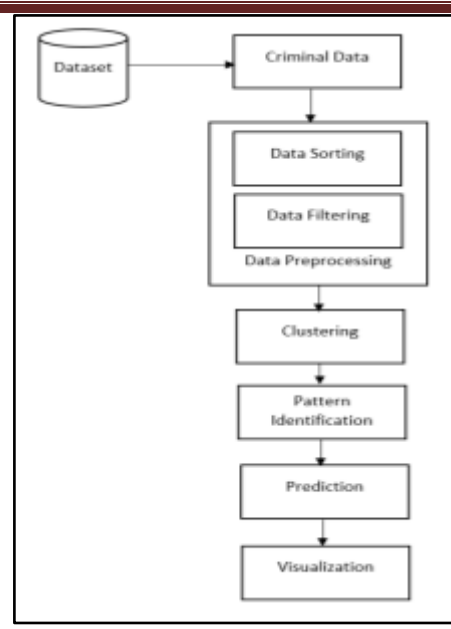

fig 2. system block diagram

DATA PRE-PROCESSING: Data Pre-processing is done in order to pre-process the data which consist of incomplete attributes and null values. The dataset is acquired from the government website www.data.gov.in. The data is provided by the National Crime Records Bureau . The dataset contains crime information of all 27 states and union territories from the year 2001-2012. Crimes such as Dowry, Kidnapping, Insult to modesty, Rape, Assault, Importation of girls, Cruelty by husband or his relatives are included in the dataset. In Order to manage data successfully missing values must be either removed or filled with the mean value. In Machine Learning models are based on Mathematical equation hence every categorical value must be transformed to numeric value. Further, the dataset is segregated into training and testing set into 80:20 ratio.

K-MEANS CLUSTERING: For clustering, k-means clustering algorithm is used which is unsupervised technique. The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters and the data set. The dataset is a collection of features for each data point. The dataset consist of Statesand union territories along with its districts, seven different crimes such as rape,dowry deaths, assault, insult to modesty, importation of girls, kidnapping, Crueltyby husband and his relatives. The algorithms starts with initial estimates for the centroids, which can either be randomly generated or randomly selected from the data set. So here the number of clusters, k=3. Three different clusters are formedbased on the number of crimes in thatparticular place. The districts with highestnumber of crime rate will be grouped together in a cluster, similarly the one withlowest and average number of crimes. The algorithm then iterates between two
steps:

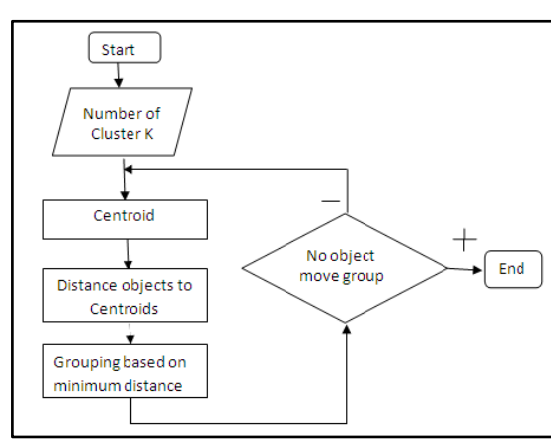

fig 3. k-means algorithm [8]

(a) Data assignment step: Each centroid defines one of the clusters. In this step,each data point is assigned to its nearest centroid, based on the squared Euclideandistance. Since the data is present in the form of 1d array the y coordinatebecomes zero and the distance can be calculated by just finding thedifference between the centroids and the dataset values.
(b) Centroid update step: In this step, the centroids are recomputed. This is doneby taking the mean ofall data points assigned to that centroid's cluster. Thealgorithm iterates between steps one and two until a stopping criteria is met(i.e., no data points change clusters, the sum of the distances is minimized, orsome maximum number of iterations is reached). Here the maximum numberof iteration is kept one. This algorithm is guaranteed to converge to a result.The result may be a local optimum (i.e. not necessarily the best possibleoutcome), meaning that assessing more than one run of the algorithm withrandomized starting centroids may give a better outcome.

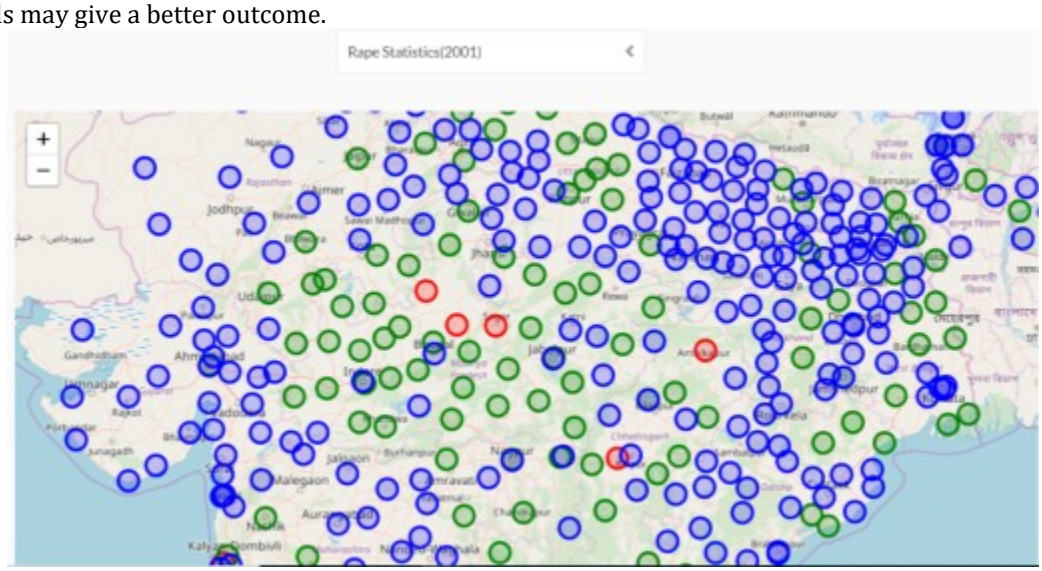

fig 4. k-means clusters

LINEAR REGRESSION: In order to forecast the crime rate for future years, linear regression technique is being used. This technique consists of a dependent and an independent variable. The linear regression line has an equation of the form y = mX + c, where m is the slope of the line, c is the coefficient of the line, X is the independent variable and y is the dependent variable. Here, the independent variable (X) is Year and the dependent variable (Y) will be the rate of specific crime from the dataset. The core idea is obtaining a line that best fits the data which can be used to predict any new feature value. The best fit line is the one for which total prediction error (all data points) are as small as possible. This best fitting line is called the regression line.

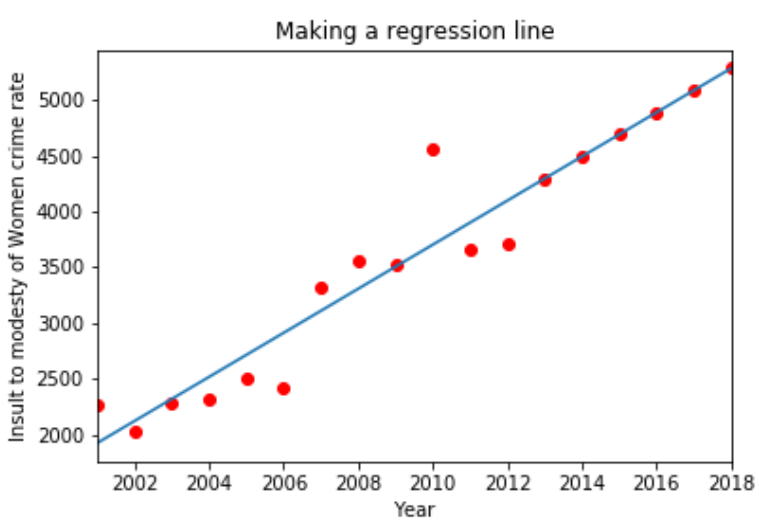

fig 5. linear regression

DATA VISUALIZATION: Data visualization is essential for representing insights from data in a graphical manner. With the large amount of data in dataset, one of the greatest challenge is to easily communicate the hidden patterns and findings in an easy and understandable manner. To visualise the data, there are many visualisation techniques available. Some of techniques that we are utilizing for the project are line chart, bar chart, correlation matrix, scatterplot.

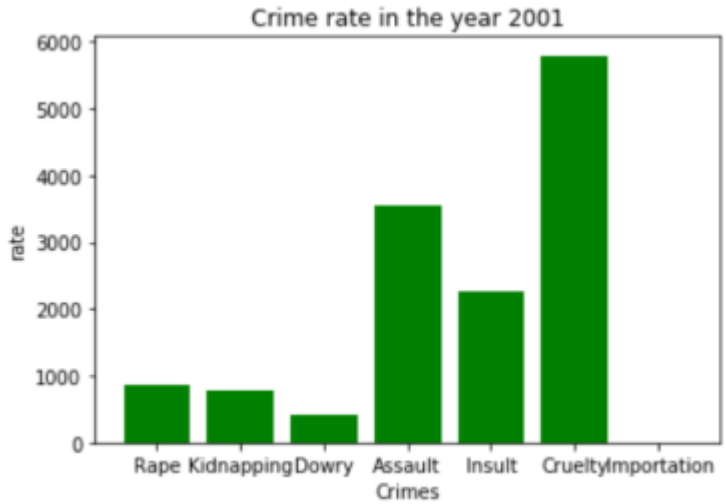

fig6. data visualization using bar chart

	Rape	Kidnapping and Abduction	Dowry Deaths	Assault on women with intent to outrage her modesty	Insult to modesty of Women	Cruelty by Husband or his Relatives	Importation of Girls
Rape	1.000000	0.548460	0.373808	0.650259	0.160520	0.610098	0.039514
Kidnapping and Abduction	0.548460	1.000000	0.486531	0.327858	0.158379	0.599610	0.011755
Dowry Deaths	0.373808	0.486531	1.000000	0.260285	0.201940	0.345819	0.101380
Assault on women with intent to outrage her modesty	0.650259	0.327858	0.260285	1.000000	0.330311	0.480329	-0.004923
Insult to modesty of Women	0.160520	0.158379	0.201940	0.330311	1.000000	0.269082	-0.005557
Cruelty by Husband or his Relatives	0.610098	0.599610	0.345819	0.480329	0.269082	1.000000	0.017730
Importation of Girls	0.039514	0.011755	0.101380	-0.004923	-0.005557	0.017730	1.000000

fig 7. correlation matrix

IV. Conclusion

Patterns related to the crimes in specific regions can be identified which can help concerned authorities to take preventive measures. Mapping by area in relation to the type of violence helps to prepare better strategies to prevent specific violence. The trends observed will help in better decision making and reducing the future crime against women. Analysis and prediction of crime by clustering and regression will help to identify the regions where specific crimes are more frequent as well as the crime rate, which in turn can be used by concerned authorities to focus on those areas.

V. Future Work

The proposed project mainly focuses on emphasizing the crime rates around a particular country (India). So as per the future vision other part of the world can be taken into consideration. Right now it is specifically bound to a single gender (Women). Other genders can be included further. There are certain crimes and cases which are unheard and unregistered around the globe and if they are taken into consideration the accuracy of the crime rates can be improved. Currently the dataset consist of only seven crimes but this can expanded to include many more crimes in the future.

VI. Reference

1. Priyanka Das, Asit Kumar Das.2017.Behavioral Analysis of Crime against Women using a Graph Based Clustering Approach.IEEE International Conference on Computer Communication and Informatics (ICCCI): 1-5.
2. Anant Joshi, A. Sai Sabitha, Tanupriya Choudhury. 2017. Crime Analysis using k-means Clustering. Third International Conference on Computational Intelligence and Networks (CINE): 33-39
3. Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma, Nikhilesh Yadav. 2017. Crime Pattern Detection, Analysis & Prediction. IEEE International Conference on Electronics, Communication and Aerospace Technology (ICECA): 225-230.
4. S. Sathyadevan, D. M. S, and S. G. S. 2014. Crime analysis and prediction using data mining. First International Conference on Networks Soft Computing (ICNSC): 406–412.
5. U. Thongsatapornwatana. 2016. A survey of data mining techniques for analyzing crime patterns. Second Asian Conference on Defence Technology (ACDT): 123–128.
6. Chhaya Chauhan, Smriti Sehgal. 2017. A review: Crime Analysis using data mining techniques and algorithms. International Conference on Computing, Communication and Automation (ICCCA): 21-25.
7. Javed Anjum Sheikh, Iqra Shafique, Madiha Sharif, Syeda Ambreen Zahra, Tuba Farid. 2017.IST: Role of GIS in Crime Mapping and Analysis. International Conference on Communication Technologies (ComTech): 126-131.
8. X. Yang, Y. Wang, D. Wu, A. Ma. 2010. K-Means basedclustering on mobile usage for social network analysis purpose. Sixth International Conference on Advanced Information Management and Service (IMS): 223-228
9. S. Sivaranjani, Dr. S. Sivakumari, Aasha. M. 2016. Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches. International Conference on Emerging Technological Trends (ICETT):1-6.
10. Zhang Ying. 2016. Analysis of Crime Factors Correlation based on data mining techniques. International Conference on Robots & Intelligent System (ICRIS).
11. A. Bansal. 2015.Performance Comparison of Data Mining Techniques to Analyse Crime against Women. International Journal Of Scientific Research And Education (IJSAE): 4494-4512.
12. A. Babakura, Md N. Sulaiman and M. A. Yusuf. 2014.Improved Method of Classification Algorithms for Crime Prediction. International Symposium on Biometrics and Security Technologies (ISBAST): 250-255.
13. Chung-Hsien Yu, Max W. Ward, Melissa Morabito, and Wei Ding. 2017.Crime Forecasting Using Data Mining Techniques. 11th IEEE International Conference on Data Mining Workshops: 779-786.
14. Aman Kumar, Nikhil Tiwari, Prakhar Gupta, Dr. S. N. Rajan. 2017.Women Crime Prediction. International Research Journal of Engineering and Technology (IRJET).
15. National Crime Bureau Record. [Online]. Available: https://data.gov.in