Stochastic Modelling and Computational Sciences

INSIGHTS INTO MOLESTATION CASES: A CLUSTERING PERSPECTIVE ON CLASSIFICATION OF CASES OF MOLESTATION AGAINST WOMEN IN INDIA

Mrs. Anjum A. Patel[1] and Dr. Pooja Kulkarni[2]
[1]Research Scholar and [2]Assistant Professor, Department of Computer Science and Technology, Vishwakarma University, Pune Maharashtra, India

ABSTRACT

Now a day in India, Women safety and Children safety becomes a major issue as they are living with fear of violence because of increasing rate of crime. Today, in the current global scenario, Women and children are facing lot of challenges. Crime against women and minor in India has increased 277% in past five years [1]. The rate of increase in crime depends upon many factors. There is need to analyze these factors and identify the predominant factors that can help curb the crime rate. The crime rate can be reduced, if we can identify the locations where the crime rate is above threshold limit, which is defined as 'Dark Spot'. The Dark Spot is identified with the K-means clustering method, with a K set to 4; the value of K was de- termined using the Elbow method. In order to decrease in molestation against women and minor, we need accurate and timely information such as location, age group of victim to take action against crime. In addi- tion to this we have also emphasised the importance of the additional parameters like a) Growth in popula- tion, b) female literacy rate, and c) working status of male and female – that affects crime against woman Hence, methodological approach for analyzing and identifying type and pattern of crime is required. This analysis of data will help police department and concerned authorities to take precautionary action to re- duce molestation against women and minor. This paper is divided into three sections: In first section; overview of problem defined with literature review. Second part elaborates methodologies applied for bet- ter results and Third section represents the analysis using statistical methods and regression analysis. The factors affecting to this analysis are age group, location and type of crime. Data mining is one of the best approaches to find out previously unknown, unstructured data and useful information by using different algorithms like regression, k-means, and apriori algorithm to find out trends, pattern and association of crimes at dark spot.

Keyword: Crime, Molestation, data mining, Statistical analysis, regression, Pattern, Dark Spot.

I) INTRODUCTION

Now a day in India, **Women safety and Children safety** becomes a major issue as they are living with fear of violence because of increasing rate of crime. Today, in the current global scenario, Women and children are facing lot of challenges. We hear the news of women harassments and children kidnapping rather than their career achievements. Crime against women is not confined to a specific culture, area , re- gion ,country, or to particular groups of women within a society. The crime against women takes many forms – physical, sexual, psychological and economic. The roots of crime against women lie in persistent discrimination against women. The number of rapes reported each year in Delhi has more than tripled in the last five years, registering an increase of 277% from 572 in 2011 to 2,155 in 2016 (data released by Delhi police). If we consider the pan India scenario then a total of 34,651 rape cases were reported in 2015 as per the report published in Indian express [2]. Furthermore, 69 rape cases were reported in Chandigarh alone in 2016; out of 69 reported cases, 41 victims were minors, i.e. below the age of 18 (NCRB 2016) [2].

The NCRB report published in 2016 high lights that crime against children has seen a continuously increasing trend over the past 3 years. The NCRB 2016 report shows a 13.6% increase in crime against minors with several cases crossing the mark of one lac. More than one lac cases reported in 2016 are distributed over different types of crimes against minors as follows: (a) Kidnapping & Abduction accounted for 52.3% of the cases, (b) Cases Reported un- der POCSO 34.4%, and (c) others. The subsequent report of 2017 from NCRB reported over 1.2 lakh cas- es against minors. In 2018,Majority of cases under crimes against women out of total IPC crimes against women were registered under 'Cruelty by Husband or His Relatives' (31.9%) followed by

‘Assault on Women with Intent to Outrage her Modesty’ (27.6%), ‘Kidnapping & Abduction of Women’ (22.5%) and ‘Rape’ (10.3%. The crime rate registered per lakh women population is 62.4 in 2019 in comparison with 58.8 in 2018. ‘Cruelty by Husband or His Relatives’ (30.9%) followed by ‘Assault on Women with Intent to Outrage her Modesty’ (21.8%), ‘Kidnapping & Abduction of Women’ (17.9%) and ‘Rape’ (7.9%). A total of 4,05,861 cases of crime against women were registered during 2019, showing an increase of 7.3% over 2018 (3,78,236 cases).(NCRB 2019).

According to poll conducted by Thompson Reuters, India is “Most dangerous country “in the world of women and the worst country for women among the G20 countries.[3] So with the aim of reducing such incidents of molestation against women and minor, analysis of crime against women and minor will give more precision to result and predicting **dark spots**.

With the statistics of reporting of crimes against women and minor as shown in Table1, Cases under Crime against Women have reported increase of 2.9% in 2016 over 2015, 12% increase in 2017 over 2016, and 7.3% increase in 2019 over 2018. So, there is urgent need to find accurate and timely infor- mation to react to women crime such as identifying the age group those who are mostly involved in crime and location or area where incidents happen most frequently with respect to type of crime. Analysis can be made with age group of girls, who are the main target of criminals. Again there is need to identify public areas considered as dark areas which have high probability of crime so that it can prevent such accidents. Thus it is very much important to **analyze the data and develop tools & techniques that can help the concerned authorities to suitable measures to diminish increasing molestation against women and Minors.** For this, we are analyzing clustering technique which can help us to find best technique to classi- fy molestation against women and minors.

Data mining uses different techniques and algorithms on large data set. The data source of the crime anal- ysis is from NCRB and police force of state, so it is heterogeneous type of data and technique used on this large set of data is Machine learning algorithms. Machine learning is divided into two areas as 1] **Super- vised learning**: In this learning method, we can train our model with labelled data and model learns from seen results. Regression and classification are types of supervised learning.[2] **Unsupervised learning**: It is the training of machine using unlabelled information and allows algorithm to act on information without guidance. Clustering and association are types of unsupervised learning algorithms. Most of the research- ers used statistical techniques[17],[18] and data mining using machine learning algorithms for analysis of crime patterns[13,20,21,22].

Year	2005	2010	2011	2012	2013	2014	2015	2016	2017	2018	2020
Rape	18K	22K	24K	25K	33.7 K	36.7K	36.7K	39k	38K	38K	28K
Assault on women	34K	40.5K	43K	45K	70.7 K	82K	82.4K	84.7 K	86K	89K	85K
Dowry Deaths	6.8K	8.4K	8.6K	8K	8.3K	8.5K	7.6K	7.6K	7.6K	7.2K	6.9K
Cruelty by husband or his Relatives	58K	94K	99K	100 K	1190 K	1230 K	1130 K	1110 K	1400 k	1300 K	1100 K
Crime against children	14K	26.7K	33K	38K	58K	89.4K	94K	1060 K	1290 K	1400 K	1228 K

Table 1: Crime Statistics 2005-2020(Reference: NCRB report)

Crime against women is not confined to a specific culture, region, state, country, or particular groups of women within a society. The crime against women takes many forms – physical, sexual, psychological, and economic. The roots of crime against women lie in persistent discrimination against women. Even after several legislative and other measures by center and state governments Crimes against women like rape, kidnapping and abduction, homicide for dowry, torture, molestation, sexual harassment, and the im- portation of girls are not declining. Dowry violence, rape, and attempt to rape by an intimate partner are the most common form of crime experienced by women in India. Women who experience such crimes suf- fer a range of physical and mental health problems and their ability to participate in public life comes into the dark. So there is a need to find out measures to eliminate or at least reduce the crime against women in India. Proper reporting of rape and assault cases, effective law enforcement agencies, exemplary punish- ment, Zero intolerance against rape cases, an effective and efficient Indian Police System, creating legisla- tive awareness amongst women, and proper training of women can be some measures that can deal with reducing crime patterns.

The status of Women in India both historically and socially has been one of respect and admiration. In In- dian culture, women are considered a symbol of the goddess. Rabindra Nath Tagore once said, *"Oh Lord why you have not given a woman the right to conquer her destiny why does she have to wait head bowed, By the Roadside, waiting with tired patience, hoping for a miracle in the morrow."[34]* This statement clearly shows the disappointing image of the Indian legislature. The Constitution of India not only grants equality to women but also empowers the State to adopt measures of positive discrimination in favor of Women for neutralizing the cumulative socio-economic educational and political disadvantages faced by them. Indian legislation system divided crime against women into two sections: 1. Indian Penal Code (IPC),2.Crime under Special and Local Laws (SLL).

THE CRIME UNDER INDIAN PENAL CODE (IPC):

i. Rape (Section 375,376, 376 A 376 B,376 C,376 D IPC)

ii. Acid Attack (Section 326 A and 326 B)

iii. Outraging the modesty of a woman or Assault on woman (Section 354)

iv. Kidnapping and abduction for specified purpose (Section 363-373 IPC)

v. Homicide for dowry, Dowry death, or their attempts. (Section. 302/304-B IPC).

vi. Cruelty caused by husband or his relatives.(Section 498-A –IPC).

vii. Insulting the Modesty of woman (Section. 509 IPC)

viii. Importation of girls (Up to 21 years of age) (Sec. 366-B IPC)

THE CRIMES UNDER THE SPECIAL AND LOCAL LAWS (SLL)- GENDER SPECIFIC LAWS

i. Immoral Traffic (Prevention) Act, 1956. .

ii. Dowry Prohibition Act, 1961 .

iii. Indecent Representation of Women (Prohibition) Act, 1986.

iv. Commission of Sati (Prevention) Act, 1987.

II) LITERATURE REVIEW

The author Waiter R. Borg, **“The literature in any field forms the foundation upon which all future work will be built.”** Without understanding the past we cannot think of new things in the field of re- search. If we want to do some new work on a subject, we must know the past of that subject. A literature survey is an essential prerequisite to actual planning and execution of any research topic which helps re- searcher to decide their objectives and aim concern to that topic.

Keeping the same concern in mind many researchers and developers have come up with several data mining algorithms which have been compared by researchers using various real-life applications.

Author Kadir Ozkan from turkey has studied concepts of data mining which was being used in digital crime investigation. He suggested if we use data mining and data warehouses with pre-planning and deci- sion making it is worth effective investigation. In this paper, two methods are discussed pre-crime imple- mentation and post-crime implementation [2]. After learning process of crime, data mining may supply solutions on how to find evidences by using their location file name etc. However the paper "**Managing Data Mining in digital crime investigation"** elaborates the effective use of data mining in crime investi- gation. But paper doesn't discuss the

methods and research design in investigating crime digitally.

The paper entitled" Research on Current Female Crime Control and Prevention Strategies by Wang Peiying in the year 2011 focuses on the main reasons for increasing women crimes in China are overall negligence of women's survival and education, development and economic rights, and women's own ignorance and disregard of their rights[4]. The characteristics and causes of female crimes in China are analyzed first and then appropriate strategies have been proposed with the aim to reduce female crimes. Analysis on parameters using any algorithm is not discussed in this paper.

In the paper entitled **"An Enhanced Algorithm to Predict a Future Crime using Data Mining"**,**Malathi A, et Coimbatore**, proposed an efficient approach towards prediction of Future Crimes us- ing data mining named as MV where they first identified missing values and computed using KNN.[5] Clustering technique K means is then applied to identify clusters of crime trend of city which is then intro- duced as high crime zone or low crime zone. Type of crime and age group is not to be considered Kadhim B et al. from Kufa University Iraq proposed a framework for the crime and criminal data analysis in the paper entitled " **A Proposed Framework for Analyzing Crime Data Set Using Decision Tree and Simple K-Means Mining Algorithms".** This paper focuses on crime pattern trends, forecasting and mapping criminal networks. Authors have used decision tree and simple K-means algorithm for data clus- tering[6] Variables like criminal's age, income, offence type and marital status is used for analysis. WE- KA is used for analyzing huge data set of criminals. Here we can find research gap in analyzing which type of crime are carried out by which age of person. The research focuses on trend patterns of criminals. Crime pattern specific region or area is not to be considered.

Chintan Shah et in 2013 analyzed data records of Breast Cancer patients using Decision tree, Naïve Bayes and KNN with the help of WEKA (Waikato Environment for Knowledge Analysis),which is an open source software, have been compared for the prediction of cancer. It has been concluded that Naïve Bayes is a superior algorithm compared to the two others[9]. We are analyzing data with K-mode and Baysian Network which will give better results.

Abba Babakura, Md Nasir Sulaiman et compare two different classification algorithms namely - Naïve Bayesian and Back Propagation (BP) for predicting "Crime Category" for distinctive states in USA. The result from the analysis demonstrated that Naïve Bayesian calculation outperformed BP calculation and attained the accuracy of 90.2207% for group 1 and 94.0822% for group 2. This clearly indicates that Naïve Bayesian calculation is supportive for prediction in diverse states in USA. compare two algorithm Back propogation and Naïve Bayes [11]. In this paper, authors have applied 10 cross-fold method on crime data set using WEKA on Crime attributes having labels 'Low", 'Medium' and High. So they only have considered the severity of crime but the location is not taken into consideration.

The paper entitled "**Performance Comparison of Data Mining Techniques to Analyse Crime against Women"** by Aarati Bansal in the year 2015 focused on comparison of three prominent data mining tech- niques (Decision Trees, Apriori and K-NN) for analyzing crimes against women [13] which shows Deci- sion tree is better than other two techniques. The elapsed time for decision tree is the minimum. Apriori Algorithm is also one of the good techniques. The accuracy of both decision tree and Apriori is same(76%). However, the performance of K-NN Algorithm (72%) is less in comparison to Apriori and Decision tree with the given training set. In this paper, author has worked on available data with the pa- rameter location only. We are using Bayesian algorithm and K- mode clustering technique on different parameters like Age group, Age of Victim and Accused, location, qualification etc. which can find out dark spot area .

In the paper **"Efficient clustering for crime analysis (2017)"**, Authors have studied K means algorithm is efficient algorithm for crime analysis. They have used data set of USA and UK for the analysis and get results accurate from 90- 95%. This paper discusses model of the crime analysis on grouping by patterns based on rate of criminal activities on the regional basis. Distortion analysis is being used for better re- sults[18]. Author only represent the cluster in particular region and analysis is based only on type of crime. Model suggests which type of crime happens where. Other attributes like age of victim, date ,time, gender of victim is not considered. If we consider these attributes then we can analyze causes of crime so that it will help to curb the crime.

In the paper entitled **"An Efficient Approach towards Crimes against Women using Time Series Algorithm** ",**Mayank Motwani , Rachit Mathur** et applied time series algorithm on crime data set for bet- ter accurate predicting and extracting patterns that occur frequently within a dataset to obtain useful hidden information. In this research, Time Series Algorithm is used to uncover and understand the underlying pat- terns in the court's records from their data in various sections [20]. This study doesn't correlate pattern and results with location type of crime and age group of victim.

Chaya Chauhan and Smriti Sehgal (Amity university, Uttar Pradesh) in their paper entitled, **A Review: crime analysis using data mining techniques and algorithms** discussed a comparison of algorithms like ID3, Z , Naïve bayes , Apriori etc. Authors have concluded that ID3 algorithm is more reasonable and more effective classification rules during analysis of experimental data. Classification techniques give more than 90% accuracy using Bayes theorem. [29]

All these studies mentioned above shows that clustering techniques like Naïve bayes, Apriori, KNN, deci- sion tree, K means can be used for analysis of molestation against women. For better results and accuracy K mean or K mode algorithm is best as it forms good quality clusters and we can analyze the data attrib- ute-wise. However, in proposed model we are computing the data using clustering and association algo- rithm so predictive model can be developed which can be used by government authorities to find out '**Dark Spots'.** With identification of dark spots police authority can apply some safety measures to curb the crime. We are working on the attributes like type of crime, location and age group which are forbidden by most of the researchers.

III) OBJECTIVES

i) To collect and analyze secondary data of Molestation against women and minor and find rela- tion/ association between variables.

ii) Define hypothesis as:

H0: There is no any relation between crime patterns of molestation and age group, location and type of crime.

H1: There is strong relationship between crime pattern against women and minor and age_ group and location.

H2: There is relationship between type of crime and age_group

H3: There is relationship between location and type of crime

iii) To analyze Women Molestation cases and compare using Machine learning models

iv) Design and develop a Predictive model to identify Dark spot areas using Machine learning al- gorithms/Techniques.

IV) RESEARCH METHODOLOGY

A) Model Flow

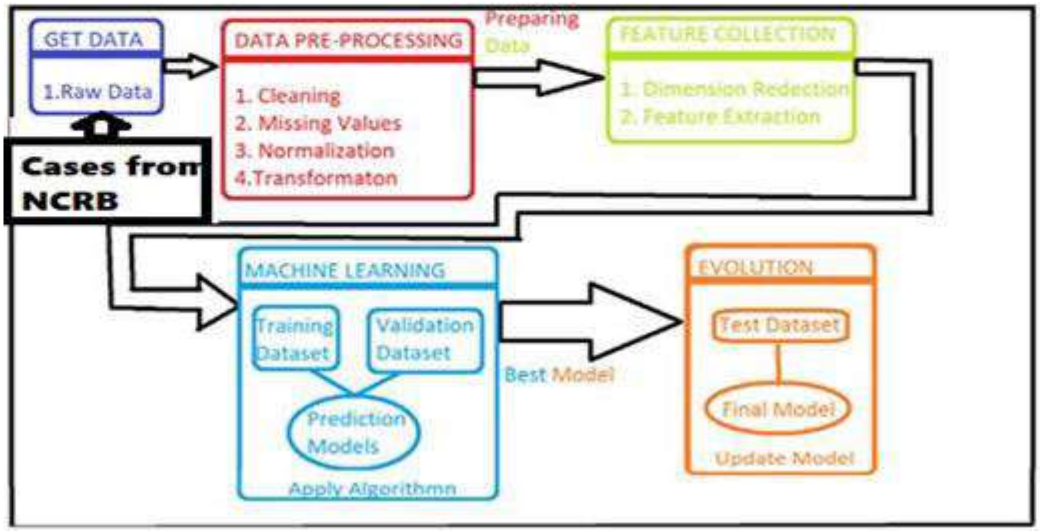


Fig 1.1: Machine Learning Model flow

As illustrated in Figure 1.1, we have obtained unprocessed data from NCRB. In the model of Ma- chine Learning, learning involves preprocessing. Formatting, cleansing, and sampling are involved. Using pre-processing, this research reduces unneeded data for a better-processed result. The appli- cation of Machine Learning algorithms, such as Regression, Classification, and Clustering, can play a crucial role in data prediction.

In the feature extraction phase, NCRB and census data are transformed into features that are suita-

ble for modelling. Data preprocessing and feature extraction have a substantial effect on model ef- ficacy.

By analyzing the model, the data set aids in anticipating the outcome of the NCRB's future data. This instance of over fitting is caused by i. the presence of absent data, ii. the small size of the training set, and iii. the complexity of the classifiers. In order to avoid over fitting, the test set is the data used to evaluate the model predicted by the training set [9]. If the number of features equals or exceeds the number of observations contained in a dataset, this can likely result in over fitting. To circumvent this issue, we employ dimension reduction techniques. During our feature collection procedure, feature extraction can reduce the number of features in a dataset by generating new fea- tures from the existing features, whereas feature selection eliminates unnecessary features. To summarize the concept, the following Figure 1.1 shows the outline of the framework for our research problem

B) Data Collection

The dataset used to perform this model is real and authentic. It is obtained from The National crime record bureau (NCRB) website uploaded by authority of Indian government. We have considered data from 2001 to 2020. This dataset contains a numerical data with different attributes as State wise reported crime against women molestation, crime wise and city wise. This analysis suggested us a 4 types of crimes are happened with very high range. Moreover, We also analyzed the region wise secondary type of data of Pune city from 2016 to 2019 i.e. 4 years. This study mainly emphasizes on the following attributes: Num- ber of cases reported (NoC),type of Crime(ToC),Location (LoC) and age group.

C) Statistical Analysis

i. Chi Square Test

Data were analyzed using Chi-Square test and P-vale below 0.05 was considered as statistically significant results.

The Chi-squared test is also written as χ^2 **test** and it **is** hypothesis test also known as goodness of fit. It is used to determine whether there is difference between expected frequencies and observed frequencies. For a given data we try to fit some probability distribution. Since there are several probability distributions, which probability distribution will fit is a question. In such cases we use Chi squared test for checking the compatibility of the theory and experiment. Hence, we want to test fitting of the probability distribution to given data is proper or not. Here we compare the expected and ob- served frequencies. So we write H_0 : There is no significant difference between observed and expected frequencies. H_1 : There is significant difference between observed and expected frequencies.

Let O_i (i=1,2,.........k) be the observed frequencies and E_i (i=1,2......k) be the corresponding expected frequencies.

$$\sum_{i=1}^{k} Oi = N = \sum_{i=1}^{k} Ei$$

P= Number of parameters estimated for fitting the probability distribution.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

It has Chi squared distribution with k-p-1 degrees of freedom. Degrees of freedom is a parameter of the Chi squared Distribution, In this case critical region at level of significance α is χ^2 k-p-1 > χ^2 k-p-1, α where χ^2 k-p-1, α is the table value corresponding to degrees of freedom k-p-1 and level of significance α.

1. We can apply this test if expected frequencies are greater than or equal to 5 and total of cell fre- quencies is sufficiently large. (Greater than 50).
2. When expected frequencies of a class is less than 5, the class is merged into neighbouring class along with its observed and expected frequencies until total of expected frequencies becomes ≥ 5. This procedure is called 'pooling the classes. In this case k is the number of class frequencies after pooling.
3. If any of the parameter is not estimated while fitting a probability distribution or obtaining ex- pected frequencies the value of p is zero.

If any data containing r rows and s columns is called r x s contingency table. O_{ij} is called observed fre- quency corresponding to (i,j)th cell. i= 1,2.......r , j=1,2.......................................s.

$$N= \sum_{i=1}^{r} \sum_{j=1}^{s} Oij = \text{Total observed frequency}$$

(Ai) $= \sum_{j=1}^{s}$ Oij = Total observed frequency in the ith row, i= 1,2, ...r.

(Bj) $= \sum_{i=1}^{r}$ Oij = Total observed frequency in the jth column, j= 1,2......s

If We wish to test: H_0: Two attributes A and B are independent against H_1 : Two attributes A and B are not independent. Under the hypothesis of independence of attributes the expected frequencies corresponding the given observed frequencies are obtained as eij = $\frac{(Ai)(Bj)}{N}$, i= 1,2…….r , j=1,2…………….s.

Using above formula, we can find all expected frequencies. If H0 is true, the statistic

$$\chi^2 = = \sum_{i=1}^{r} \quad \sum_{j=1}^{s} \frac{(Oij - eij)\text{\^{}}2}{eij}$$

It has X^2 distribution with (r-1) (s-1) d.f. We reject H0 at l.o.s. α if χ^2 (r-1) (s-1) ≥ χ^2 (r-1) (s-1), α and accept H0 otherwise.

- (Level of significance (l.o.s.) is probability of rejecting H0 when it is true.
- d.f : degrees of freedom

In our paper, we considered Ho: there is no relationship between location and type of crime and H1: there is strong relationship between location and type of crime. As we explained earlier, chi square test is best for finding a relation between attributes so we apply χ^2 to test independency on loc and Toc. Toc is divided into 4 classes as 1. Assault 2. Rape 3. kidnapping 4. Cruelty by hus- band. Loc is location, which is divided into 30 regions in Pune city police chowki). Following are the observations of χ^2 test .

	Table of Observed Values					
		Type of Crime(ToC)				
	Location (LoC)	1	2	3	4	
		Assault	**Rape**	**Kidnap- ping**	**Cruelty by husband**	**Total**
1	**Loc1**	62	23	60	50	195
2	**Loc2**	18	16	32	23	89
3	**Loc3**	35	8	27	23	93
4	**Loc4**	20	3	20	19	62
5	**Loc5**	33	11	44	36	124
6	**Loc6**	24	8	23	10	65
7	**Loc7**	34	20	36	30	120
8	**Loc8**	49	5	32	29	115
9	**Loc9**	53	62	66	58	239
10	**Loc10**	27	11	23	20	81
11	**Loc11**	30	7	30	27	94
12	**Loc12**	16	9	15	12	52
13	**Loc13**	46	36	51	46	179
14	**Loc14**	43	25	52	46	166
15	**Loc15**	13	13	27	22	75
16	**Loc16**	49	15	34	31	129
17	**Loc17**	45	23	48	36	152
18	**Loc18**	32	4	30	29	95
19	**Loc19**	34	17	44	37	132
20	**Loc20**	39	31	54	49	173
21	**Loc21**	90	31	56	44	221
22	**Loc22**	39	22	41	35	137
23	**Loc23**	125	64	153	143	485
24	**Loc24**	43	10	46	41	140

25	**Loc25**	163	72	151	125	511
26	**Loc26**	44	12	40	34	130
27	**Loc27**	84	75	113	94	366
28	**Loc28**	34	28	47	40	149
29	**Loc29**	9	6	12	10	37
30	**Loc30**	68	34	68	51	221
	Total	1401	701	1475	1250	4827

Table(A) : Observed Values (LoC versus ToC)

	Table of Expected Values					
		Type of Crime(ToC)				
	Location(LoC)	1	2	3	4	
		Assault	**Rape**	**Kidnapping**	**Cruelty by hus- band**	**Total**
1	**Loc1**	17.995	6.6755749	17.414543	14.51212	56.5973
2	**Loc2**	5.2244	4.6438782	9.2877564	6.675575	25.8316
3	**Loc3**	10.158	2.3219391	7.8365444	6.675575	26.9925
4	**Loc4**	5.8048	0.8707272	5.8048477	5.514605	17.995
5	**Loc5**	9.578	3.1926663	12.770665	10.44873	35.9901
6	**Loc6**	6.9658	2.3219391	6.6755749	2.902424	18.8658
7	**Loc7**	9.8682	5.8048477	10.448726	8.707272	34.8291
8	**Loc8**	14.222	1.4512119	9.2877564	8.417029	33.3779
9	**Loc9**	15.383	17.995028	19.155998	16.83406	69.3679
10	**Loc10**	7.8365	3.1926663	6.6755749	5.804848	23.5096
11	**Loc11**	8.7073	2.0316967	8.7072716	7.836544	27.2828
12	**Loc12**	4.6439	2.6121815	4.3536358	3.482909	15.0926
13	**Loc13**	13.351	10.448726	14.802362	13.35115	51.9534
14	**Loc14**	12.48	7.2560597	15.092604	13.35115	48.1802
15	**Loc15**	3.7732	3.773151	7.8365444	6.385333	21.7682
16	**Loc16**	14.222	4.3536358	9.8682411	8.997514	37.4413
17	**Loc17**	13.061	6.6755749	13.931635	10.44873	44.1168
18	**Loc18**	9.2878	1.1609695	8.7072716	8.417029	27.573
19	**Loc19**	9.8682	4.9341206	12.770665	10.73897	38.312
20	**Loc20**	11.319	8.997514	15.673089	14.22188	50.2119
21	**Loc21**	26.122	8.997514	16.253574	12.77067	64.1436
22	**Loc22**	11.319	6.3853325	11.899938	10.15848	39.7632
23	**Loc23**	36.28	18.575513	44.407085	41.50466	140.768
24	**Loc24**	12.48	2.9024239	13.35115	11.89994	40.6339
25	**Loc25**	47.31	20.897452	43.8266	36.2803	148.314
26	**Loc26**	12.771	3.4829086	11.609695	9.868241	37.7315
27	**Loc27**	24.38	21.768179	32.79739	27.28278	106.229
28	**Loc28**	9.8682	8.1267868	13.641392	11.6097	43.2461
29	**Loc29**	2.6122	1.7414543	3.4829086	2.902424	10.739
30	**Loc30**	19.736	9.8682411	19.736482	14.80236	64.1436

Table (B)- Expected Values

Degree of freedom = 87

Level of significance, alpha = 0.05

$\chi2$ calculated value = 8377.9272

Tabular value for 87, $\chi2$ for 0.05 = 109.77

Thus, $\chi2$ calculated > $\chi2$ tabular (Yes), Therefore, the Null hypothesis is rejected. Then the alternate hypoth- esis is there is strong relationship between type of crime (ToC) and Location (LoC) is accepted. This analysis indicates that if we analyze the location-wise type of crime irrespective of data of crime then dark spots can be predicted. In other words, prediction does not depend upon date of crime but depends upon location where it happened. Thus the following table summarizes percentage of type of crime location-wise. We have applied association rule mining on data of pune city location-wise crime happened due respect to type of crime.

Risk Zone	Type of Crime	Percentage of crime se- verity to total crimes in all location	Location with high risk
High Risk	Assault	29%	Hadapsar, Yerwada, Chatushrungi, Kondhawa Wanworie
Medium	Rape	15%	Kondhawa, Hadapsar, Yerwada, Wanworie
High Risk	Kidnapping	31%	Yerwada, Hadapsar,Kondhawa, wanworie
Medium	**Cruelty by husband**	**26%**	**Yerwada, Hadapsar, Kondhawa,**

Table C- Percentage of severity of crime in pune city

Table –C shows two zones viz. High and Medium by following the severity of crime that happened in a particular region in Pune city. The present study recorded 31% of crimes of Kidnapping happened in yerwada, hadapsar, kondhawa and wanoworie regions, and 29% of crimes of Assault happened in Hadapsar, yerwada, chatushrungi, kondhawa and wanowarie, Cruelty by husband type of crime recorded as 26% and 15% of Rape recorded in kondhawa, Hadapsar,yerwada and wanworie. If we apply this analysis with the data of month then more predictive rules can be identified.

In this research, we have tested the relationship between location and type of crime against women. In this paper, we have observed that,ToC i.e. Assault with outrage modesty, Rape, Kidnapping and cruelty by hus- band and his relatives can define the relationship with location. several cases are depending upon location in which it happens. We need to find whether this is the only attribute or are there any other attributes that can suffice to type of crime.

D) **Tools Used for Data Mining**

a) Weka

Is software with a collection of machine learning algorithms for data mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualiza- tion. It is also well-suited for developing new machine learning schemes. It shows you various rela- tionships between the data sets, clusters, predictive modeling, visualization etc. In our paper, we have analyzed data using K-means algorithm as well as the regression algorithm.

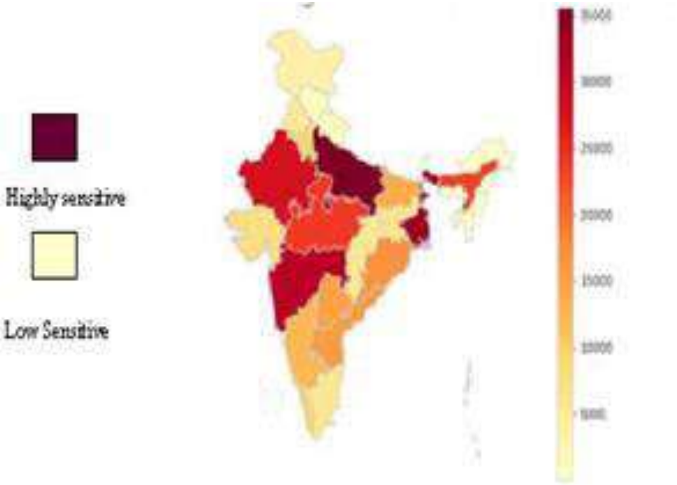
b) Python Libraries

Python is a scripting language, which uses an object-oriented approach. We can analyze and visualize data with python libraries. However, we can use python for predictive modeling.

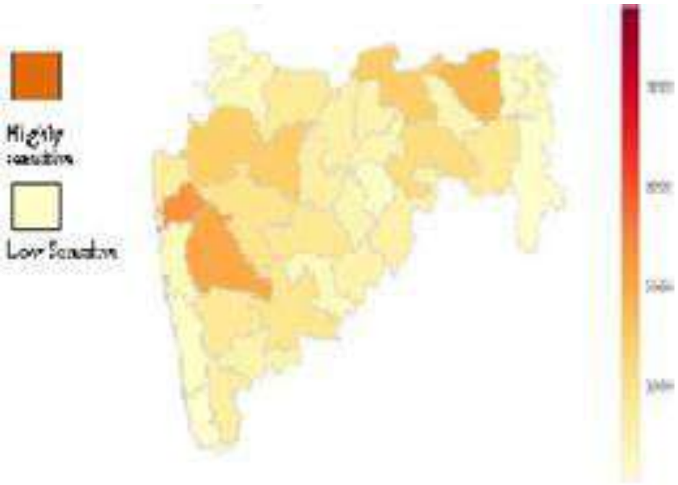
E) Discussion and Conclusion

In this research, we have tested relationship between location and type of crime against women. In this pa- per,we have observed that, ToC i.e. Assault with outrage her modesty, Rape, Kidnapping and cruelty by hus- band and his relatives can define the relationship with location. Number of cases are depending upon location in which it happens. We need to find whether this is the only attribute or are there any other attributes which can suffice to type of crime.

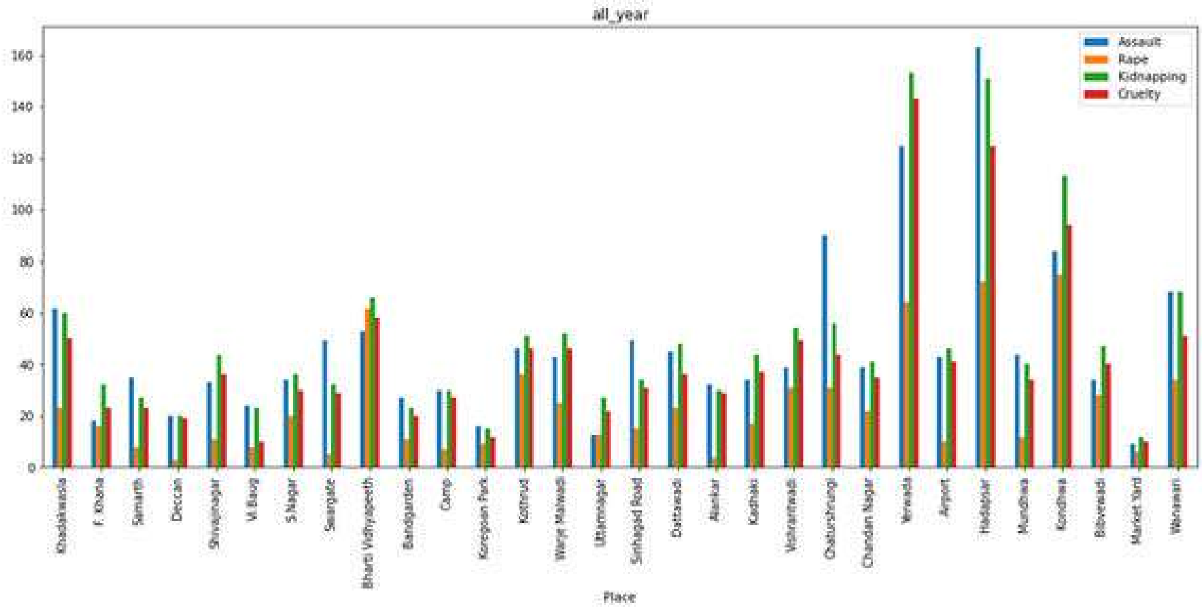
Initially, we analyzed records of crime against women and minor of all States of India with reference to NCRB data and analysis of qualitative data suggested to us that **Uttar Pradesh** and **Maharashtra** is highly sensitive area about women molestation considered **dark spot,** and Tamilnadu, Sikkim, Uttarakhand Himachal Pradesh are low sensitive areas as shown in Graph 1 below. However, in the analysis of district-wise crime against woman molestation in Maharashtra highly sensitive areas are **Pune**, **Nagpur** and **Thane** which are considered as Dark spots. So, in this paper, we concentrated on data of Pune city to predict high-risk are- as responsible for crime against women as shown in Graph 2. Moreover, Graph3 is showing location-wise crime details of four types of crime viz. Assault, Rape, Kidnapping and abduction and cruelty by husband. Graph 4 is showing crime pattern from 2000 to 2020; With this analysis we can predict headwise crime analysis increasing from 2020 for the type of crime Cruely by husband and his relatives and Kidnapping ab- duction followed by Assault .agegroup wise analysis of crime is shown in Graph 4 where age group of 18-30 years and 14-18 years girls are seen as targeted group for recorded crimes.



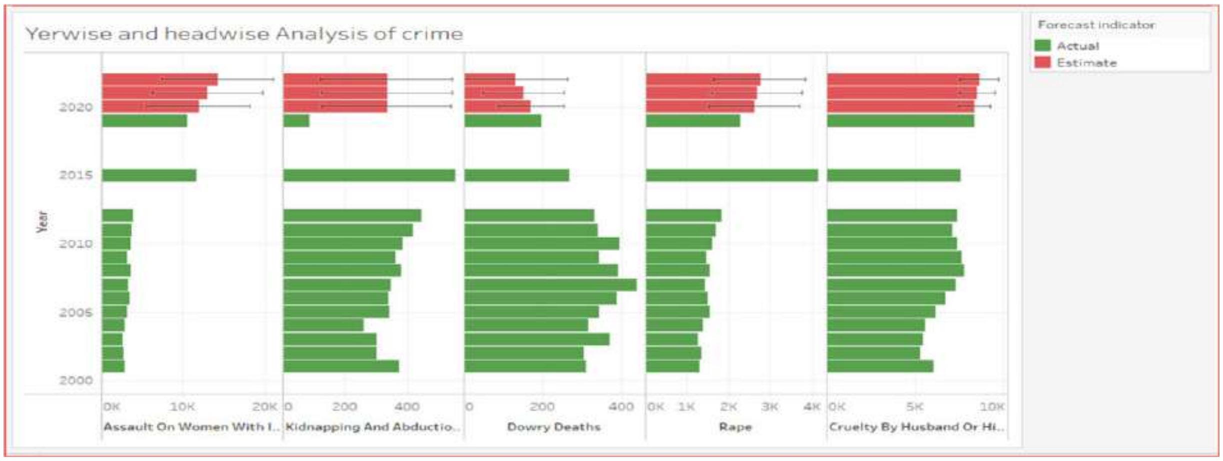
Graph 1: Highly sensitive areas in India against women molestation



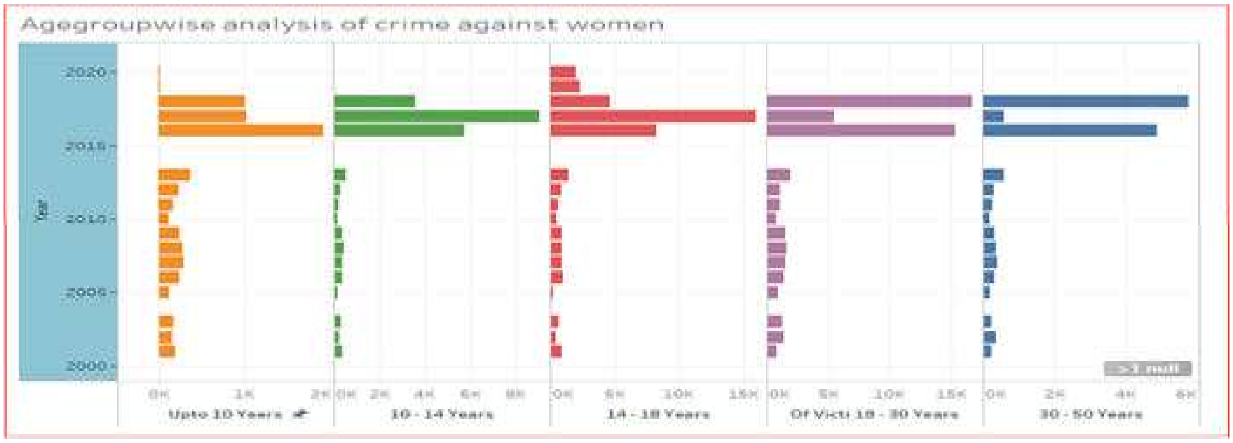
Graph2 : Highly sensitive areas in Maharashtra against women molestation



Graph 3: Crime details in pune city for last 4 years



Graph4: Crime Patterns in India from 2000-2020



Graph5: Age group-wise Analysis of crime against women

II. REGRESSION

Regression is a supervised learning technique that permits the discovery of correlations between variables and the prediction of continuous values based on these variables. When the output is continuous, the problem is referred to as a regression one —For Example, predicting a person's weight, age, or salary, temperature, weather forecasting etc. In regression, input variables is mapped to continuous output. Classification is the process of predicting the discrete labels of the input. Regression, on the other hand, is concerned with the prediction of continuous values. Regression is divided into two main categories: Simple Linear and Multiple Regression. In simple linear regression, a straight line is drawn to define the relationship between two varia- bles. In contrast, Multiple regression encompasses multiple variables. It can also identify the dis- tribution trends based on the available data or historic data. Here, the task of regression is to pre- dict a "Dark spot" in the given study area.

There are different variables/ attributes operating in regression, including a dependent variable— the main variable that you are trying to understand—and an independent variables—factors that may have an influence on the dependent variable. In order to make regression analysis work in our problem, you must collect all the relevant data of CAW. It can be presented on a graph, with an x-axis and a y-axis.

The main reasons we use regression analysis:

1. To determine the relationship between two or more attributes,
2. To understand how one variable change when another change, and
3. To predict and identify the Dark spots to curb the incidents by considering the NCRB attributes and dataset.

There are many different kinds of regression analysis. For the purpose of this research work, we will look at two: linear regression and multiple regression.

Statistics	Location	Assault	Rape	Kidnapping	Cruelty by husband
Count	31	4827	4827	4827	4827
Mean	15.5	90.387	45.226	95.161	80.645
Max	31	1401	701	1475	1250
Min	1	9	3	12	10
Std dev	8.803	245.358	123.32	258.275	219.041

Table D: Statistics of Attributes

Time taken to build model: 0.07 seconds

inst#, actual, predicted, error

inst#, actual, predicted, error

1	95	95.054	0.054	9	81	81.082	0.082
2	130	130.014	0.014	10	511	510.558	-0.442
3	52	52.121	0.121	11	239	238.938	-0.062
4	4827	4821.383	-5.617	12	137	137.021	0.021
5	173	172.988	-0.012	13	115	115.021	0.021
6	124	124.026	0.026	14	37	37.139	0.139
7	149	149.016	0.016	15	166	165.987	-0.013
8	485	484.6	-0.4				

Correlation Coefficient	1
Mean Absolute Error(MAE)	0.4695
Root Mean Squared Error(RMSE)	1.4597

Relative Absolute Error (RAE)	0.1189
Root Relative Squared Error(RRSE)	0.12

Figure 1.2: Attributes with Correlation Coefficient

As per the analysis observed in above model, prediction results are reasonably good. Perfor- mance of the above model can be measured with the analysis of Mean Absolute Error, Root Mean Squared error. Here RMSE is the standard deviation of predicted results. RMSE value is

1.46 which is 10% greater than the MAE i.e. 0.46. This interprets that, data has a linear relation- ship and which found correct. Model prescribes reasonably good results but only thing is that if we provide more training data then better possible results will be predicted. Model summarizes that, Type of Crime) (ToC) is associated with location and negatively associated with date.

III) K-MEANS ALGORITHM ANALYSIS

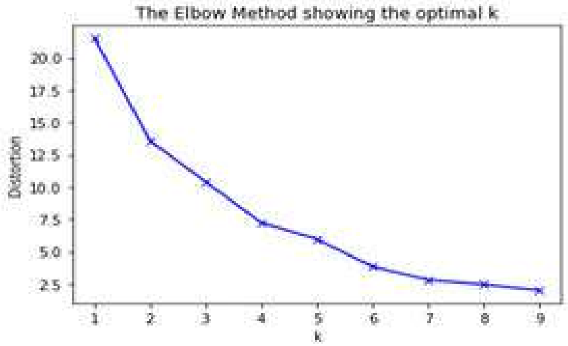
K means is the type of unsupervised learning applied on unlabelled data. This algorithm computes

centroids and iterates till we find optimal centroid. Data points are assigned to a cluster in such a way that sum of the squared distance between the cluster's centroid and data points at a minimum.

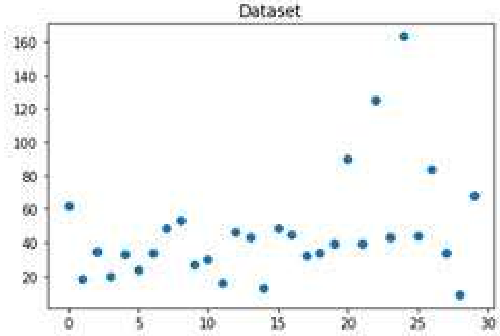
Working of K Means algorithm

1. Specify number of clusters K
2. Initialize centroids and select K data points randomly for the centroids
3. Repeat the process until there is no change to the centroids.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the centroid.
6. Calculate average of the all-data points that belong to each cluster.

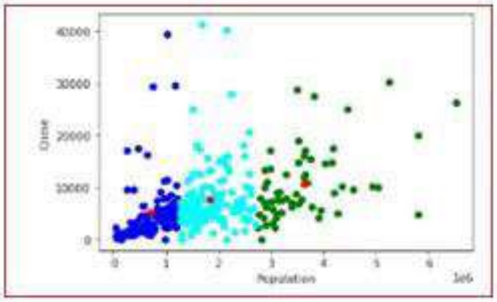
Basically, K means is used to identify homogeneous groups on which we can build supervised Model. In our study, we are trying to identify clusters of crimes happening in particular area, which can predict dark spot in Pune city. Here, we are finding the areas where similar types of crime are registered so we can apply precautionary measures to curb the crime against women. In below Graph 4: We have used Elbow method and considered value of K= 4 where distortion starts decreasing in a linear fashion. and Graph 5 shows clusters of our dataset.

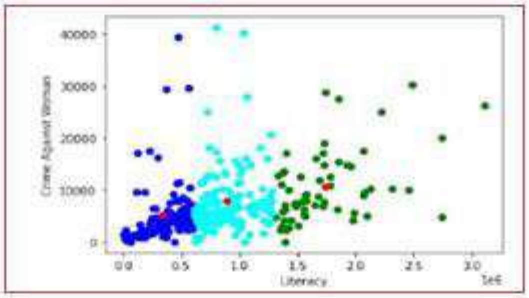


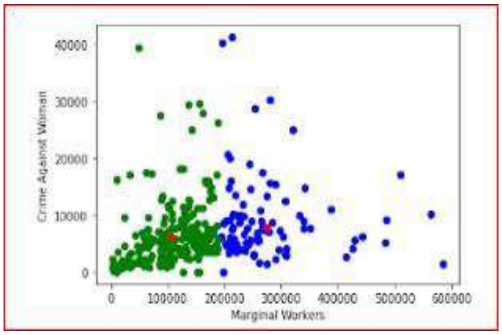
Graph 4: Elbow method showing optimal K

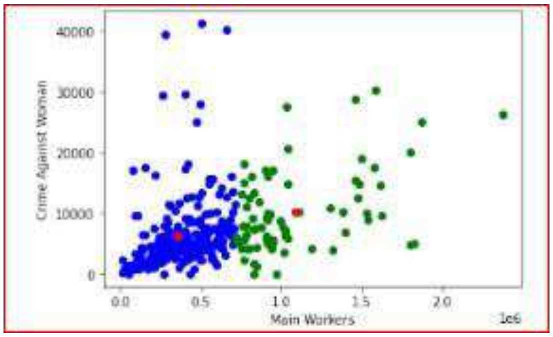


Graph 5: Clusters using K Means









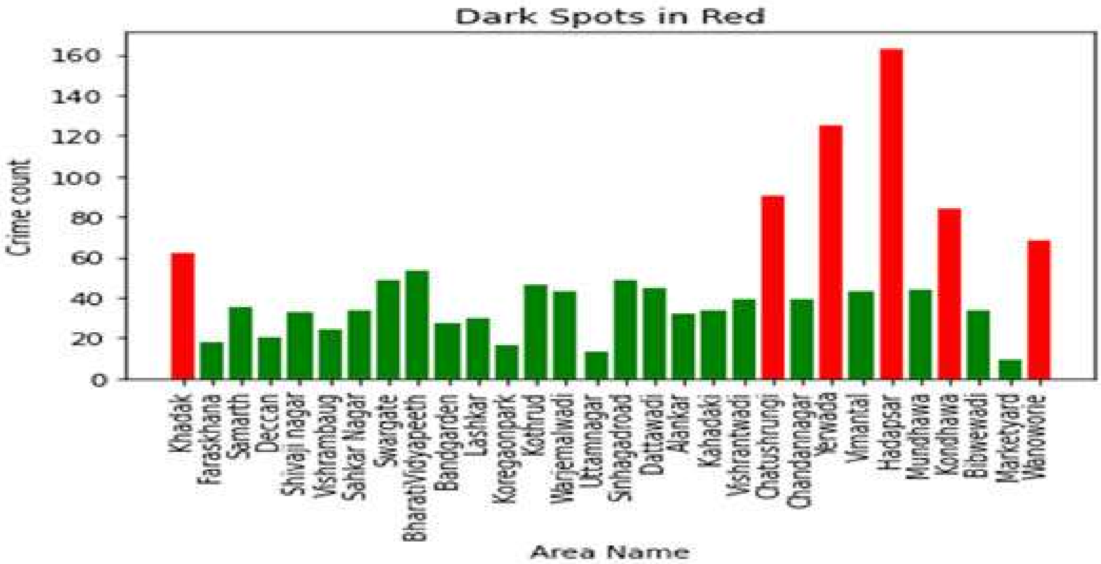
The Elbow approach was used to generate four clusters, and the characteristics of these four clusters are out- lined in the table that can be found above. A comprehensive investigation into each cluster unearths composite variables of crime committed against women, often known as multi-variate variables. Following an in-depth examination of each cluster, it was discovered that composite variables can be classified from cluster C-0 to cluster C-3 according to the sub-attributes of increase in population, Literacy rate, working status (Marginal, Nonworkers, Main workers), and places with high crime rates and locations with low crime rates respectively. We have conducted an analysis of all 568 districts in India, using the dataset provided by the NCRB. Cluster 0 comprises 216 districts, Cluster 1 comprises 238 districts, Cluster 2 comprises 106 districts, and Cluster 3 comprises eight districts.

The following is a summary of composite variables, as presented in the table that can be found above:

Cluster C-0 is made up of 216 districts, 127 of which have a lower crime rate due to a low growth in popula- tion (i.e. 58.80%), and 89 of which have a higher crime rate due to a high growth in population (i.e. 41.20%). Cluster C-0 is divided into two groups.

In Cluster C-1, a total of 238 districts have been taken into consideration; out of these, 120 districts are demonstrating a low crime rate (50.42%) whereas 118 districts are demonstrating a higher crime rate (49.58%).

Cluster C-2 examines 106 districts, of which 71 districts have a low crime rate with a percentage of 66.98%, while 35 districts have a higher crime rate with a percentage of 33.02%.



There are a total of eight districts that are taken into consideration for the fourth cluster C-3, of which five districts have an exceptionally low crime rate (62.50%) and three districts have an exceptionally high crime rate (37.50%).

In a similar vein, the composite variable literacy rate in all clusters shows a crime rate of 55.65% in low crime regions with higher female literacy and a crime rate of 44.35% in high crime regions with low female literacy.

V. RESULTS AND DISCUSSION

In this research, looking to the current Indian scenario, safety of women and minor is taken into consideration. As per analysis of records of National Crime Records Bureau from 2001 to 2020 statistics show that approximately 40% of female reported rape victims were minors and 95% knew the rapist. Nearly 3,38,954 persons were convicted in crime against women and only 1,67,042 persons were acquitted (NCRB report 2016) similarly, 90,098 persons were arrested in rape cases and only 17,047 were acquitted. As per records most of the cases are pending so our motive is to develop precautionary measures for preventing women mo- lestation. We use clustering techniques for analyzing data of crime against women in India. As per result **Ut- ter Pradesh** is at highest rank and **Sikkim** is at lowest rank in women molestation. **Maharashtra** stood at 4th Rank in women molestation. However, in **Pune city** we have predicted dark spot as Hadapsar, Kondhawa, Yerwada, Chatushrungi and wanowarie. With the analysis on molestation against women and minor we are considering age group(adult/minor), Location, type of crime etc. In further research, we can add more attrib- utes like time of an incident (Mor/Aft/Evening/Night). In addition to this research, we can consider some secondary attributes such as education of victim, literacy rate of region, law and jurisdiction etc. This study represents societal approach toward the current security of woman. It concentrates on two attributes such as location and type of crime. This paper reports the use of NCRB data to develop IMS (I aM secure) model to curb the crime against women and minor by predicting dark spots in city. In the Indian scenario, this study is very much important for better security of woman which extends to their safe life. At the same time, it will also help in taking suitable measures to mitigate increasing crime rate against woman.

REFERENCES

[1] Crimes against Women ,http: //ncrb.gov.in/ StatPublications/ CII/CII2016/pdfs/Crime% 20Statistics%20-%202016.pdf

[2] Kadir Ozkan Police Crime Laboratory in Istanbul, Vatan Cd. A. Blok Kat:7, Fatih, Istanbul, Turkey “**Managing data mining at digital crime investigation**” Elsevier-2004

[3] T.Millo, Om.P Murthy "**Female victims of violent crimes and abuses:An overall view of Indian Scenario**" , International Journal of Medical Toxicology and Legal Medicine · January 2006

[4] Wang Peiying, Research on **Current Female Crime Control and Prevention Strategies** *(ISBN: 978- 1-61284-109-0/11) 2011.*

[5] Malathi A , Santosh Saboo coimbature India , An Enhanced Algorithm to Predict a Future Crime using Data Mining

[6] Kadhim B. Swadi Al-Janabi, "**A Proposed Framework for Analyzing Crime Data Set Us ing Decision Tree and Simple K-Means Mining Algorithms**", Journal of Kufa for Math ematics and Computer Vol.1, No.3, may , 2011, pp.8- 24

[7] Anish Gupta, Vimal Bibhu, Md. Rashid Hussain,"**Security Measures in Data Mining"**

[8] Dr: Zakaria Suliman Zubi, Ayman Altaher Mahmmud, "**Using Data Mining Techniques to Analyze Crime patterns in the Libyan National Crime Data",** Recent Advances in Image, Audio and Sig nal Processing(2018)

[9] Chintan Shah and Anjali g. Jivani, Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction, IEEE International Conference on

[10] Nazlena Mohamad Ali, Masnizah Mohd , Hyowon Lee, Alan F. Smeaton, Fabio Crestani and Shahrul Azman Mohd Noah, **"Visual Interactive Malaysia Crime News Retrieval System"**

[11] Abba Babakura, Md Nasir Sulaiman and Mahmud A. Yusuf Faculty of Computer Science and In formation Technology,Universiti Putra Malaysia (UPM)-**Improved Method of Classification Algo rithms for Crime Prediction,** *2014 International Symposium on Biometrics and Security Technol- ogies (ISBAST)*

[12] Jeroen S. de Bruin, Tim K. Cocx, Walter A. Kosters, Jeroen F. J. Laros and Joost N. Kok Lei den Institute of Advanced Computer Science (LIACS) Leiden University,The Netherlands, "**Data Mining Approaches to Criminal Career Analysis,** https://www.researchgate.net/publication / 220764807.

[13] ***Aarti Bansal ,***M. Tech, Department of Computer Engineering, Punjabi University,Punjab, India "**Performance Comparison of Data Mining Techniques to Analyse Crime against Women",** ***In-ternational Journal Of Scientific Research AndEducation Volume*** ||3||Issue||9||Pages-4494-4512||October-2015|| ISSN (e): 2321-7545

[14] Elise Clougherty, John Clougherty, Xiaoqian Liu, and Donald Brown University of Virginia," **Spa tial and Temporal Analysis of Sex Crimes in Charlottesville, Virginia"** 2015 IEEE Systems and Information Engineering Design Symposium

[15] Dr. B. Umadevi, M. Snehapriya**, "A Survey on Prediction of Heart Disease Using Data Mining Techniques"** IJSR ISSN (Online): 2319-7064

[16] Ahlaq Hussain bhat, University of Kashmir, "Nature of various types of crime" International Journal in Management and Social Science (Impact Factor- 3.25) Vol.03 Issue-04,

[17] Supreet Kaur1, Amanjot Kaur Grewal *1Research Scholar, Punjab Technical University, Dept. of CSE, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib,Punjab, India,"* **A review paper on data mining classification techniques for detection of lung cancer**", International Re- search Journal of Engineering and Technology (IRJET)-2016.

[18] Malarvizhi S, Siddique Ibrahim, "**An Efficient Clustering for Crime Analysis"** International Jour nal of Innovative Research in Computer and Communication Engineering, March 2017.

[19] Uma Sheokand Research Scholar Department of Public Administration Punjab University Chandi garh " Crime against women Problem and suggestion:Case study in India. International Journal in Management and Social Science http://www.ijmr.net.in Page 218

[20] Mayank Motwani, Pratha Purwar, Rachit Mathur Aatif Jamshed- **An Efficient Approach towards Crimes against Women using Time Series Algorithm** (*International Journal of Computer Appli- cations (0975 – 8887) Volume 179 – No.34, April 2018*

[21] S Prabakaran and Shilpa Mitra 2018 *J. Phys.: Conf. Ser.* **Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning,** National Conference on Mathematical Techniques and its Applications (NCMTA 18) IOP Publishing IOP Conf. Series: Journal of Physics: Conf. Series **1000** (2018) 012046.

[22] Rajan Singh, Bramah Hazela *Department of Computer Science & Engineering, Amity University Uttar Pradesh, "***A Survey on Machine Learning And Data Mining Methods And Applica- tions",** *IOSR Journal of Computer Engineering (IOSR-JCE) e- ISSN: 2278- 0661,p-ISSN: 2278- 8727, Volume 20, Issue 5, Ver. I (Sep - Oct 2018).*

[23] Huda Kutrani, Saria Eltalhi **," Cardiac Catheterization Procedure Prediction Using Machine Learning and Data Mining Techniques",** IOSR Journal of Computer Engineering (IOSR-JCE)e- ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 1, Ver. I (Jan - Feb 2019), PP 86-92

[24] Jeroen S. de Bruin, Tim K. Cocx, Walter A. Kosters, Jeroen F. J. Laros and Joost N. Kok Leiden Institute of Advanced Computer Science (LIACS) Leiden University, The Netherlands "**Data Min- ing Approaches to Criminal Career Analysis**

[25] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann publications, pp. 1-39, 2006.

[26] Veepu Uppal and Gunjan Chindwani (2013) **, "An Empirical Study of Application of Data Min ing Techniques in Library System",** *International Journal of Computer Applications (0975 – 8887) Volume 74– No.11, July 2013.*

[27] Prof. Basavaraj Chougula, Archana Naik, Monika Monu, Priya Patil and Priyanka Das

[28] Remya George, Anjaly Cherian.V, Annet Antony, Harsha Sebestian, Mishal Antony and Rosemary Babu.T, ―An Intelligent Security System for Violence against Women in Public Places‖, ISSN: 2249 – 8958 International Journal of Engineering and Advanced Technology (IJEAT), Volume-3, Issue-4, April 2014.

[29] Chhaya Chauhan, SmritiSehgal Department of Computer Science and Engineering Amity University Uttar Pradesh, India. **A Review: crime analysis using data Mining techniques and algorithms,** International Conference on Computing, Communication and Automation (ICCCA2017)

[30] Julia weibe, **"Rape in India and Its Consequences, Did the Delhi Gang Rape Case Lead to any Changes in India?" University of Ruhr.**

[31] Fuyuan Cao, Jiye Liang, Deyu Li Liang Bai Chuangyin Dang **,"A dissimilarity measure for the k- Modes clustering algorithm"** Knowledge-Based Systems 26 (2012) 120– 127, Elsevier

[32] Smart Girls Security system, Department of Electronics and telecommunication KLE's College of Engineering and Technology Belgaum India, ISSN 2319 – 4847 International Journal of Application or Innovation in Engineering & Management (IJAIEM)

[33] Women Empowerment Index Report by Hindustan times under NHFS survey.

[34] Devesh Kanishk, Meshram Neelam National Law Institute University, Bhopal Online published on 28 June, 2017 “Crime against women and present status of law” Article DOI : 10.5958/0974-4533.2015.00010.x VIDHIGYA: The Journal of Legal Awareness Year : 2015, Volume: 10, Issue: 2 First page : (30) Last page : (39) Print ISSN : 0973-3825. Online ISSN : 0974-4533.