Assaying the Statistics of Crime Against Women in India using Provenance and Machine Learning Models

Geetika Bhardwaj, Dr. R. K. Bawa
Department of Computer Applications, Punjabi University, Patiala, India

Abstract—**Now-a-days, the surging of crime against women is occurring at a startling rate in India. According to the National Commission for Women, there was a 46% increase in reports of crimes against women in the initial months of the year 2021 in comparison with the same period in 2020. However, to handle this problem, the need of the hour is to fetch relevant and timely information about the various types of crime taking place and make specific predictions based on the existing information to safeguard women from future predictable contingencies. AI and Machine learning mechanisms have become a powerful tool in predicting the crime rate in India under various crime categories by analyzing the crime patterns, crime–centric areas, and the comparative study of various crime categories. Hence, from 2001 to 2019, a women's crime-based dataset from NCRB has been used in this paper, which included various crime sub-categories, for instance; molestation, sexual harassment, rape, kidnapping, dowry deaths, cruelty to family, importation of girls, immortal traffic, sati prevention act, and others. To acquire a better understanding of the data, a framework has been created which makes use of provenance and machine learning algorithms on the dataset, which has been grouped based on several factors such as distribution of cases convicted or reported every year, safest and un-safest states for women in India, etc. Different machine learning algorithms, such as gradient boosting and its many versions, Random forest, and many more, have been used on the dataset. Their performances are evaluated using various metrics such as accuracy, recall, precision, F1 score, and root mean error square.**

Keywords—Crime against women; provenance; scalar techniques; machine learning techniques; decision tree; random forest; gradient boosting; XgBoost; CatBoost; LightGBM

I. Introduction

The overall crime rate in India is increasing at a steady pace. Crime cannot be predicted as it is either efficient, coincidental, or goes unreported. Various modern advancements and hi-tech procedures enable criminals to carry out their crimes when it comes to cybercrime. As far as women's crimes are concerned, it has been seen that a girl can be a victim or target of a crime from the moment she is born or even before. According to crime statistics, authority violations against women, for example, chain grabbing, sex abuse, child abuse, assault, and murder, are rapidly increasing [1]. Hence, keeping that in view, the establishment of the gender equality principle was done in the Constitution of India, and to uphold and implement the Constitutional mandate, the state has created several laws as well as has taken various actions to guarantee equal rights, eradicate social injustice, and prohibit multiple forms of violence and massacres [2].

India's National Crime Records Bureau records various incidents showing that crime against women increased by 6.4%, and it occurs every three minutes [3]. The reports also revealed that in 2011 the number of reported crimes against females was more than 228,650. In the year 2015, reported incidents were more than 300,000. It can be deduced that there was a rise of 44% in felonies against the members of fair sex.7.5% of females residing in the state of West Bengal in India account for 12.7% of reported crimes against women [4]. The female population of Andhra Pradesh accounts for 7.3% of total India's women population, and 11.5% of all reported crimes against women were from this state. Hence to compile all the information, a graphical analysis has been shown in Fig. 1 to define the rate of women crimes in India in different states for 2018. It is also important to note that exact figures on the breadth of case occurrences are difficult to obtain because many cases go unreported. This is mainly due to the potential reporter's fear of scorn or embarrassment and tremendous pressure not to jeopardize the family's honor [5].

A crime investigation system should be able to rapidly and efficiently identify crime patterns to detect and act on future crime trends to work on such crimes. Various law enforcement agencies and state governments should take significant steps to reduce such crimes and promote a secure environment for women [6]. Multiple scholars have been working on detecting women-related crime worldwide in the field of research. Data Analytics and machine learning have contributed notably to detecting and preventing crime, providing a basis for crime analysis. They are also acknowledged as a relatively new and highly sought-after area of research. In reality, law enforcement agencies are seeking the support of data mining and AI techniques to help deter crime and ensure law enforcement. In a nutshell, data mining is a branch of the multidisciplinary subject of database knowledge discovery. The input in the process of Data mining is the unprocessed data, which is transformed into information that is used to produce precise projections and applied to real-life circumstances (through inference and analysis). Multiple techniques, such as scientific and statistical machine learning algorithms, have recently been used in the recognition of images and speech, detection of medical ailments, and classification [7] and are additionally designed to estimate crime rates for a particular year formed on statistical information of crime against females.

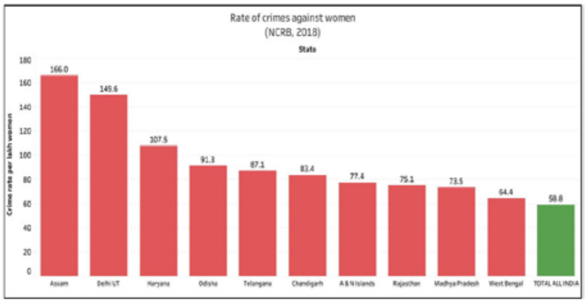


Fig. 1. Rate of Crimes against Women in the Year 2018.

In a nutshell, it can be said that studying the crime data helps us to solve criminal cases and is taken as a first step to prevent the1 crime via which we can reduce or deter criminals and their activities. But depending on raw data is utterly erroneous as it contains plenty of noise, incomplete or missing values, etc. to confuse the investigation team so that they cannot afford either forecast its future features or catch the culprit. Hence, to work on such destructed data, we want to create a decision-making system that can analyze the crime data, find the data uncertainty, and remove the errors.

The essential part and main motive behind this research work is data provenance which is a novelty and is used to train the system so that we can work on the uncertainty of the crime data against women. Provenance aids in getting into the origin of data and various transformations made on the data over time. Suppose the analysis is done based on data obtained from a particular source that no longer exists or is unavailable to us; in that case, the data provenance generated might be the guiding light to establish the uncertainty in the data.

Hence, the contribution of the research for analyzing the statistics of women-based crimes in India using crime data provenance is as follows:

1) The information is gathered from NCRB and Kaggle between 2001 to 2019, which is further pre-processed to remove NAN or missing values.

2) Later feature scaling techniques such as Min-Max, Standard scalar, and PCA is being applied to scale or generalize the values.

3) At the end, the scaled data is further evaluated using various evaluation parameters such as accuracy, rmse, R2, mse, precision, recall, etc to find the best technique for predicting the crime.

The study has been divided into several sections where Section I has been already defined as an Introduction; Section II defines the related work. The proposed model is demonstrated in Section III, which includes information about datasets, libraries, data collection, data provenance, data pre-processing, feature scaling, algorithms, and evaluative parameters. Section IV mentions the results, whereas Section V concludes the study.

II. Related Work

Many scholars have investigated crime control concerns and proposed various crime prediction algorithms. The qualities of the techniques and the dataset, along with the challenges, have been used as a benchmark (Table I) to determine the accuracy of the prediction and generate the research gap. In [7], the authors examined and found the key factors influencing crime in certain parts of the country. They employed clustering, a graphical representation based on determining which areas have the greatest and minor crimes. In addition, the authors have implemented the Community Detection Algorithm. The authors presented a method for predicting crime without human involvement [8] using computer vision and machine learning technologies. The paper employed rectified linear units (ReLU) and convolutional neural networks (CNN) to identify weapons in images, such as knives or guns. This aided in confirming the occurrence of a crime and identifying the event's site. The results looked highly accurate, with roughly 92% accuracy for a test set. In [9], the authors developed a provenance capturing mechanism that aims to trace digital evidence's transformation to bolster the trustworthiness of digital evidence when the incident response takes place on the affected system. In their work, the data provenance was recorded simultaneously in both the systems, i.e., the incident response system and the affected system.

Likewise, in [10], the authors employed a variety of machine learning algorithms on data of criminal cases in India to find patterns in criminal activities in a particular geographical location. The aim was to reduce the number of pending criminal cases by classifying them based on their crime patterns to solve them faster. Different machine learning algorithms were applied, and a comparison was made based on several parameters to find the best algorithm to solve this problem efficiently. It had been concluded that the Random Forest Classification method was best suited to predict the desired results after classification. In [11], the authors created a prediction model that can be used for the prognosis of crime rates accurately, and they tested the accuracy of six different types of Machine learning algorithms on crime data, which included Linear Regression, SVM, KNN, decision trees, Nave Bayes and CART (Classification and Regression Tree). The authors of [12] described how the frequency of crimes and crime features in India, such as rape, sexual assault, and kidnapping, may be examined using machine learning models in the investigation process.

Similarly, in [13], the authors worked on several machine learning algorithms predicting crimes. They devised a new framework to combine two machine learning algorithms, i.e., XGBoost and TF-IDF, to improve the results of various algorithms used in text mining to predict crimes. The improvement of data accuracy to strengthen the accuracy of crime prediction overweighs the optimization process of algorithms. The inaccuracy of crime data might result in an inaccurate prediction of a specific category of crime that has its basis in the historical data. Training a good classification model is imperative in improving the accuracy of data, which in turn forms the basis for analyzing and accurately predicting crime. The text-based classification of theft crime data based on the two algorithms, i.e., XgBoost and TF-IDF, were also used to get a logical and error-free classification effect of data. This was a constructive trial at a machine learning algorithm for data mining of police-related data and quintessential for predicting crime.

TABLE I. ANALYSIS OF THE PREVIOUS WORK

Ref	Dataset	Techniques	Outcomes	Limitation
[8]	Dataset containing criminal tools	CNN, ReLu	Accuracy = 90.02%	High computational cost
[11]	Real world data	KNN	Mean = 0.33	Accuracy needed to be improved
[34]	Crime reports	Visionary system, natural language processing	The system detected the crime based on analysing analytic provenance trails.	Multiple techniques needed to be incorporate for achieving the better results
[30]	NCRB data	Linear Regression	Acc = 83%	Limited dataset

[13]	data collected from 2009 to 2019	Tf-IDF, XGBoost model	Prec = 92.3%
			Rec = 91.6%
			F1 score = 91.9%
				Values of performance metrics needs to be enhanced

[31]	Data collected from past 10 years	Logistic Regression	Acc = 80.769%	Less scalability
		Random forest	Acc= 76.923%	
		Naïve Bayes	Acc = 80.769%	
		Decision tree	Acc= 76.923%	
[33]	Crime dataset	Huber Regression	Results obtained in the form of predictions and score of each state where women crime has taken place	Design complexity

No research on data provenance classification in criminal data has been conducted. Scientific data is kept in databases. Therefore, provenance management solutions were created with that in mind. No system has taken the provenance of crime data and its implementation concerns and challenges into account. Various women-based crimes have been targeted using multiple machines and deep learning techniques. Still, it has been discovered that researchers have encountered specific issues, either in terms of detection or system performance. A few models, such as logistic regression, KNN, Relu, and SVM, had a complex design, used a limited dataset, or needed improvement in their performance. As a result, the research's primary motivation is to fill the gaps to forecast a better system for detecting crime utilizing data provenance.

III. PROPOSED SYSTEM

A framework has been designed to analyze the statistics of women-based crimes in India, named the Crime Data Provenance Framework, as shown in Fig. 2. This framework aims to design a methodology to use the provenance of the given crime data in the form of annotations (a subtext or metadata which provides additional information about the actual data) to create crime classes that are used further for the classification and prediction of crime data. There are a variety of crime sub-categories in this framework which has become the basic building block, and prediction models have been applied. The results have been compared and analyzed. The framework has only been implemented for a single category of crime, i.e., crime against women. The various steps in the framework are categorized into several sub-sections wherein each sub-section elucidates the purpose and tasks performed in each step.

A. Data Collection

The data has been accumulated from the government website named National Crime Record Bureau (NCRB) [35] and Kaggle [36]. The website was established on 11th March 1986. The general aim was to empower Indian Police with Information Technology and Criminal Intelligence to make law enforcement more productive. The data from 2001 to 2015 was in .xls format, and the rest from 2015 to 2019 was in pdf format. Some books helped to understand how crime provenance trends can be represented in a textual form that could later become part of Provenance. The first book, named 'Violence Against Women in India' prepared by International Center for Research on Women, aimed at elaborating on the trends and patterns in crime against women, and the second book was 'Tackling Violence Against Women: A Study of State Intervention Measures' which was a report prepared by Bhartiya StreeShakti, a voluntary, autonomous, apolitical organization committed to empowering women, families and the society at large. The report aimed to study different aspects of crime against women by taking perspectives from other professionals like Lawyers, Police Personnel, Public Prosecutors, and Medical Officials [12].

At a fleeting glance, the data seemed to have a lot of dimensions as there were a lot of crime subcategories. There were several crime categories as shown in Fig. 3 and each category had further subcategories. Each crime subcategory had different crime status categories. For instance, our crime category was CRIME AGAINST WOMEN. Further, it had subcategories such as cruelty by husband and relatives, kidnapping of women and girls, Indecent Representation of Women (Prohibition) Act, 1986, and many more.

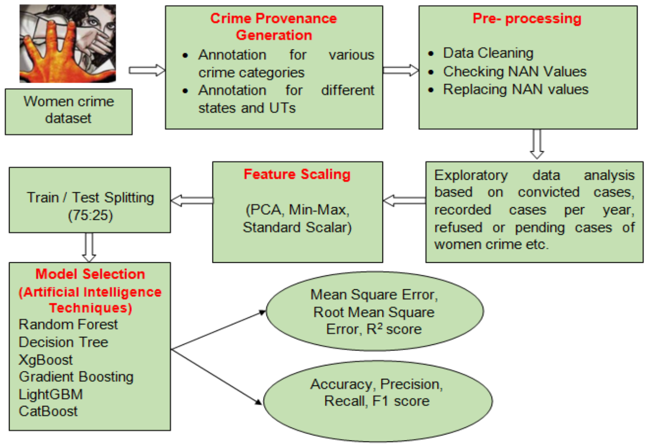

Women crime dataset
Crime Provenance Generation
• Annotation for various crime categories
• Annotation for different states and UTs
Pre- processing
• Data Cleaning
• Checking NAN Values
• Replacing NAN values
Exploratory data analysis based on convicted cases, recorded cases per year, refused or pending cases of women crime etc.
Feature Scaling
(PCA, Min-Max, Standard Scalar)
Train / Test Splitting (75:25)
Model Selection (Artificial Intelligence Techniques)
Random Forest
Decision Tree
XgBoost
Gradient Boosting
LightGBM
CatBoost
Mean Square Error, Root Mean Square Error, R² score
Accuracy, Precision, Recall, F1 score

Fig. 2. Crime Data Provenance Framework.

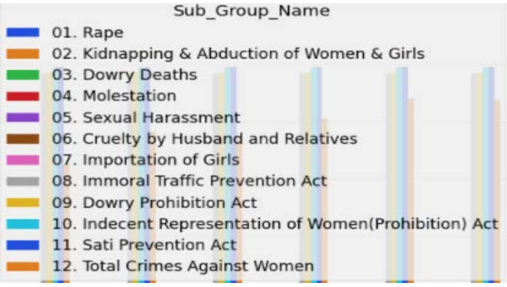


Fig. 3. Various Crime Categories under Crime against Women.

B. *Provenance Generation*

The crime against women has been subcategorized based on physical or mental abuse, or both caused to women under various circumstances with varying intentions. Different laws are formulated to deal with cases for each crime category, and they have been listed along with their annotations as a small sample data in Table II.

In addition to this, the number of cases for different categories of crime reported at various places has also been organized and segregated according to various States and Union territories. As we know, India has 28 States and 8 Union Territories, and the data has been compiled by the number of cases of different crime categories reported in these places. The states and union territories have been annotated to be used as Provenance for crime prediction purposes. Table III consists of the name of a state and a Union Territory and their annotations as a sample.

TABLE II. SAMPLE OF ANNOTATIONS FOR DIFFERENT CRIME CATEGORIES

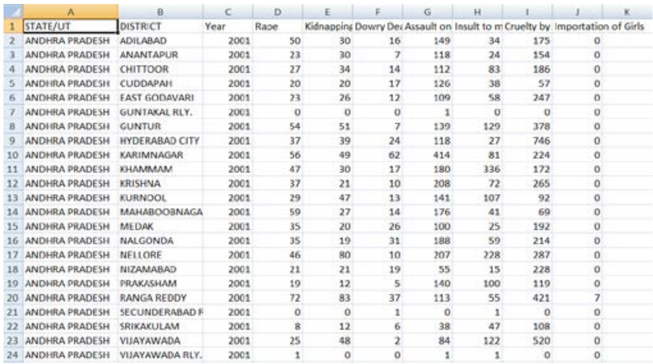
Case Sub-categories	Crime Act/Section	Case Code (Annotations)
Rape	Section 376 IPC	CSR-376
Kidnapping & Abduction of Women & Girls	Sec. 363-373 IPC	CSK-363-373
Commission of Sati Prevention Act 1987	NA	CSSP-1987
Protection of Women from Domestic Violence Act 2005	NA	CSDV-2005

TABLE III. SAMPLE OF ANNOTATIONS FOR DIFFERENT STATES AND UNION TERRITORIES

Name of the State/Union Territory	State/UT Code(Annotations)
Arunachal Pradesh	SCAR
Puducherry	UTPU
Himachal Pradesh	SCHP
Jharkhand	SCJH
Karnataka	SCKA
Kerala	SCKE
Madhya Pradesh	SCMP
Maharashtra	SCMH
Manipur	SCMA
Meghalaya	SCME
Mizoram	SCMI

C. *Crime Data Pre-processing Phase*

The Preprocessing of data is necessary to clean it and make it suitable for a machine learning model, which improves the effectiveness and precision of the machine learning model. Loading libraries and setting up the platform is the prerequisite to initializing the process of Data pre-Processing Fig. 4. Several Python libraries, such as Matplotlib, Numpy, Sklearn, Itertools, SimpleImputer, Seaborn, Maths, and Pandas, have been loaded to perform specific functions [14-16].



	A	B	C	D	E	F	G	H	I	J	K
1	STATE/UT	DISTRICT	Year	Rape	Kidnapping	Dowry Dea	Assault on	Insult to m	Cruelty by	Importation of Girls	
2	ANDHRA PRADESH	ADILABAD	2001	50	30	16	149	34	175	0	
3	ANDHRA PRADESH	ANANTAPUR	2001	23	30	7	118	24	154	0	
4	ANDHRA PRADESH	CHITTOOR	2001	27	34	14	112	83	186	0	
5	ANDHRA PRADESH	CUDDAPAH	2001	20	20	17	126	38	57	0	
6	ANDHRA PRADESH	EAST GODAVARI	2001	23	26	12	109	58	247	0	
7	ANDHRA PRADESH	GUNTAKAL RLY.	2001	0	0	0	1	0	0	0	
8	ANDHRA PRADESH	GUNTUR	2001	54	51	7	139	129	378	0	
9	ANDHRA PRADESH	HYDERABAD CITY	2001	37	39	24	118	27	746	0	
10	ANDHRA PRADESH	KARIMNAGAR	2001	56	49	62	414	81	224	0	
11	ANDHRA PRADESH	KHAMMAM	2001	47	30	17	180	336	172	0	
12	ANDHRA PRADESH	KRISHNA	2001	37	21	10	208	72	265	0	
13	ANDHRA PRADESH	KURNOOL	2001	29	47	13	141	107	92	0	
14	ANDHRA PRADESH	MAHABOOBNAGA	2001	59	27	14	176	41	69	0	
15	ANDHRA PRADESH	MEDAK	2001	35	20	26	100	25	192	0	
16	ANDHRA PRADESH	NALGONDA	2001	35	19	31	188	59	214	0	
17	ANDHRA PRADESH	NELLORE	2001	46	80	10	207	228	287	0	
18	ANDHRA PRADESH	NIZAMABAD	2001	21	21	19	55	15	228	0	
19	ANDHRA PRADESH	PRAKASHAM	2001	19	12	5	140	100	119	0	
20	ANDHRA PRADESH	RANGA REDDY	2001	72	83	37	113	55	421	7	
21	ANDHRA PRADESH	SECUNDERABAD F	2001	0	0	1	0	1	0	0	
22	ANDHRA PRADESH	SRIKAKULAM	2001	8	12	6	38	47	108	0	
23	ANDHRA PRADESH	VIJAYAWADA	2001	25	48	2	84	122	520	0	
24	ANDHRA PRADESH	VIJAYAWADA RLY.	2001	1	0	0	1	1	0	0	

Fig. 4. Preprocessing of Dataset.

D. *Exploratory Data Analysis*

The pre-processed data has been categorized into various categories, including the most unsafe and safe states for women in India from 2001 to 2013, as displayed in Fig. 5. The classification of different categories of women-based crimes, distribution of cases convicted per year, and distribution of cases reported per year (i.e., from 2001 to 2010) has been displayed in Fig. 6 and Fig. 7, which show that the convicted cases in terms of the total number of crimes against women range from 25000 to 35000 while as in the year 2010.

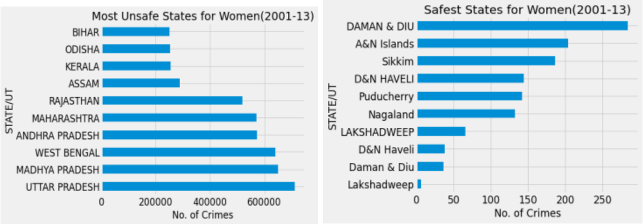


Fig. 5. Most Unsafe and Safe States for Women in India.

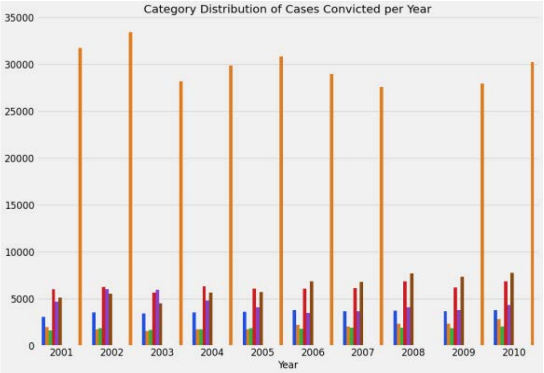


Fig. 6. Convicted Cases of Crime against Women in India.

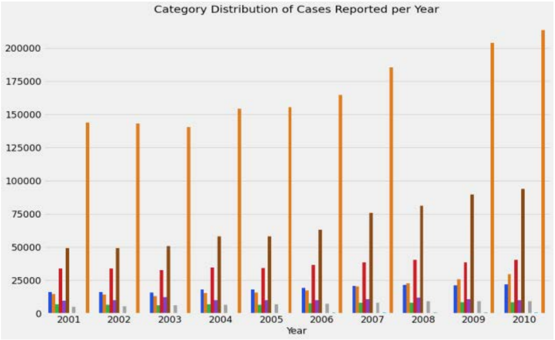


Fig. 7. Reported Cases of Crime against Women in India.

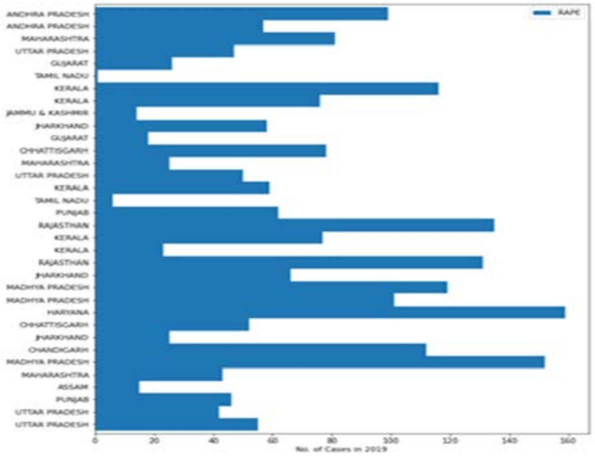


Fig. 8. Statewise Records of Rape Cases in 2019.

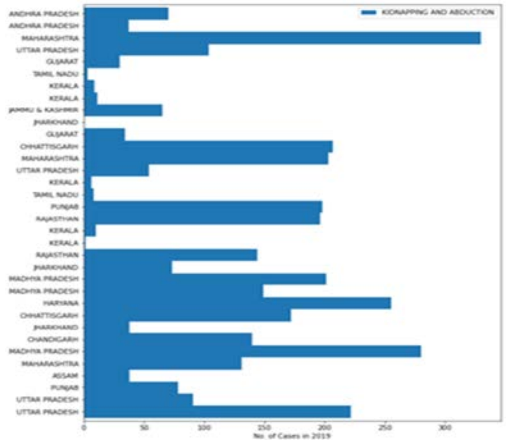


Fig. 9. Statewise Records of Kidnapping and Abduction in 2019.

In addition, Fig. 8 and Fig. 9 explore the data for the most recent year, 2019, the statistics displaying states that fell most heavily in crimes such as rape, the modesty of women, cruelty by husbands, kidnapping, and so on. The state of Haryana had the most rape cases, followed by Madhya Pradesh, while Tamil Nadu had the least. Similarly, Maharashtra ranked first in kidnapping and abduction, while Tamil Nadu and Kerala tied for last place. Dowry deaths have decreased dramatically across the country, with at least five states reporting the lowest number of women affected by dowries. Women's molestation has also reduced in several states, including Gujarat, Jammu & Kashmir, and many other conditions. Uttar Pradesh, Andhra Pradesh, and other states have many offenses involving spouse or relatives' maltreatment.

On the other hand, Andhra Pradesh ranks first among all states regarding girl importation as we know that the input to all these attributes is in numerical data. Hence, standardizing it using various scaling techniques mentioned in the next section is essential.

E. Feature Scaling

Feature scaling is the final stage in machine learning data processing. It is a method for standardizing the independent variables in a dataset within a given range. Multiple scaling techniques can be used here, but the one given priority is the one that offers more optimized results after normalizing the data [17-21]. Hence in this section, the scaling techniques such as Min-Max Scalar, Principal component analysis, and Standard Scalar have been used to showcase the performance of machine learning models such as decision tree, gradient boosting, and its many versions and random forest. These models have been applied to the dataset taken from various women-based crimes like cruelty by husbands, rape cases, an insult to modesty, kidnapping and abduction cases, dowry deaths, and importation of girls and are shown graphically in Fig. 10 to 12.

1) Min Max: The entire data is scaled between 0 and 1. To calculate min-max, the formula is shown in Eq. (1):

$$x_{scaled} = \frac{(x - x_{min})}{(x - x_{max})} \tag{1}$$

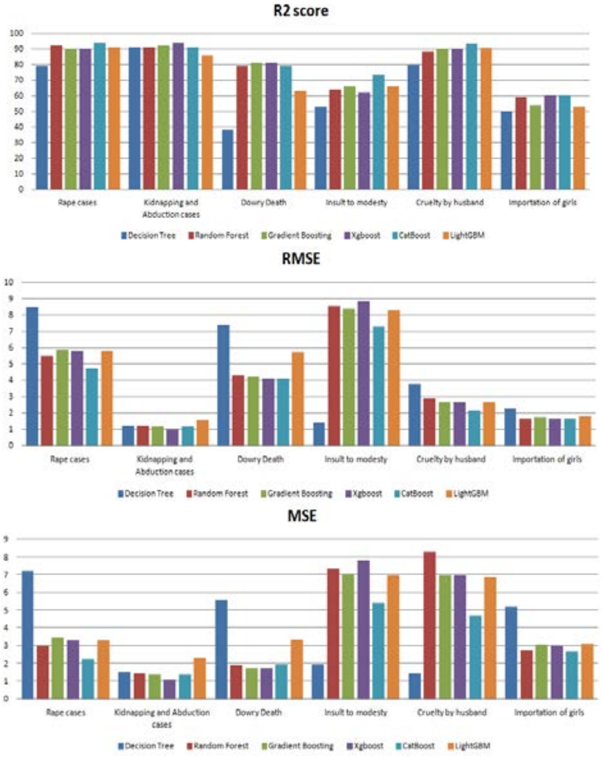


Fig. 10. Analysis of Algorithms using Min-Max Scalar.

On assaying Fig. 10, it can be said CatBoost worked well for importaton of girls, cruelty by husband, insult to modesty, and rape cases by 60, 93.4, 73.4, and 94 R^2 score, respectively while as XgBoost obtained great score in dowry death cases and kidnapping by 81 and 94 R^2 score respectively. On the other hand, for both mean square error and root mean square error best values have been obtained by CatBoost and XgBoost as compared to the other algorithms.

1) Standard scalar: It scales the values in such a way where the standard deviation or variance is 1 and mean is 0. The formula is shown in Eq. (2):

$$x_{scaled} = \frac{(x-mean)}{std\_dev} \qquad (2)$$

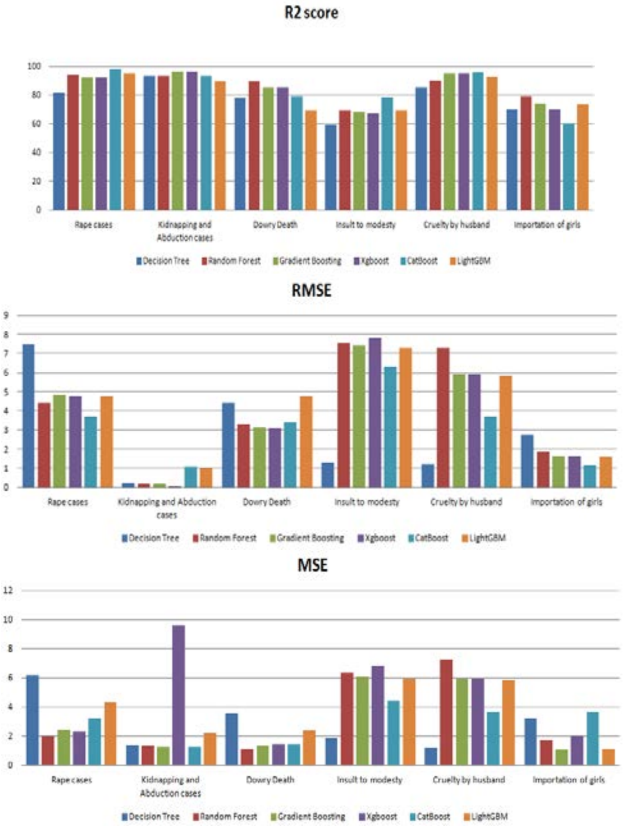


Fig. 11. Analysis of Algorithms using Standard Scalar.

The analysis in Fig. 11 determines that CatBoost worked well for the importation of girls, cruelty by husbands, an insult to modesty, and rape cases with 79, 93.4, 78.4, and 98 R2 scores, respectively. At the same time, XgBoost obtained a great score in dowry death cases and kidnapping with 89 and 96 R2 scores, respectively. On the other hand, for root mean square error and mean square error, the best values have been obtained by CatBoost and XgBoost as compared to the other algorithms.

2) Principal component analysis: A statistical technique in determining interrelations among a set of variables. The conversion of correlated variables to a set of uncorrelated variables takes place.

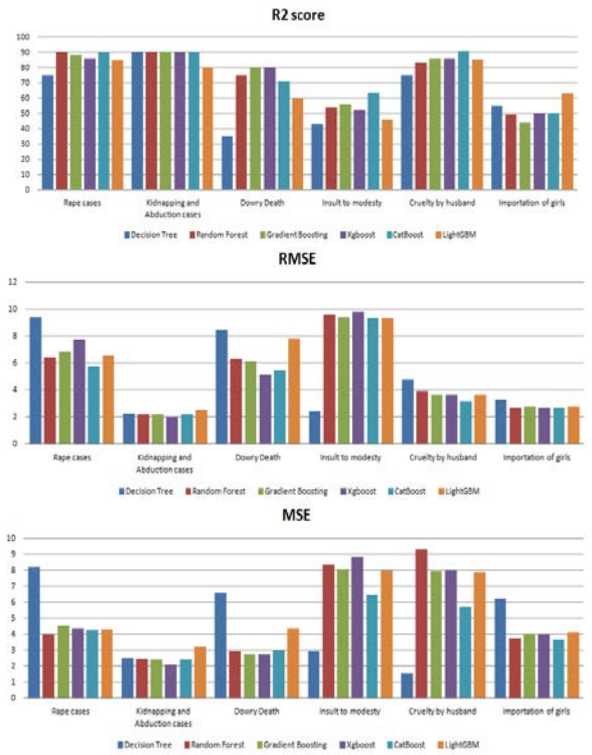


Fig. 12. Analysis of Algorithms using Principal Component Analysis.

On assaying Fig. 13, it can be said that LightGBM, CatBoost, XgBoost, and Random Forest gave better results for the women crime dataset by 63, 90.4, 80, 90 R2 scores, 2.76, 31.6,20.3,64.6 root mean square error value for data such as importation of girls, cruelty by husband, dowry, and rape, respectively.

The Scaling technique that gave the best results after normalizing the data was Standard Scalar. After feature scaling, the dataset has been divided into two halves. The first half consists of 75% of the training dataset, and the remaining half has a 25% testing dataset on which application of machine learning algorithms has been made.

F. Model Selection

Various machine learning models have been applied on the women crime dataset. A variety of models have been used on the dataset as one particular model cannot give accurate results as each model has its own characteristics and working technique. The output produced by each one of them is compared to get the best results. Features of each one of them have been discussed subsequently.

1) Decision tree: It is a representation of every all probable solutions to a problem/decision which is based on particular specifications [20]. It is organized in a tree-like form with two nodes: the Leaf Node and the Decision Node and Leaf nodes keep track of the outcomes of those decisions and have no other branches, on the other hand Decision nodes can have multiple branches depending upon the number of decisions made [21]. The features of the given dataset are used to make judgments or run tests. The values of decision tree can be calculated by computing the impurity of a node and its entropy which are shown in Eq. (3) and Eq. (4):

$$I_G(\mathrm{n}) = 1 - \sum_{i-1}^{J}(pi)^2 \tag{3}$$

where J is the count of classes present in the node and p is the distribution of the class in the node.

$$Entropy = \sum_{i=1}^{c} -(pi)^{*}log_2(pi) \tag{4}$$

where C is the number of classes present in the node and p is the distribution of the class in the node.

2) XgBoost: It is based on the concept of gradient boosting. It also makes use of decision trees. It is a custom, parallelized tree building algorithm which contains several decision trees. It provides features like efficient handling of missing data, and automatic feature selection [18].

3) Gradient boosting: Gradient Boosting is one of the most powerful techniques to construct predictive models. It trains many models in parallel. Every new model minimizes the loss function using the Gradient Descent Method. It is a greedy algorithm, and overfitting of the training dataset can happen quickly. It benefits from regularization methods by penalizing various parts of the algorithm and improving the algorithm's performance by reducing overfitting [22].

4) Random forest: It uses the concept of many decision trees to solve a compounded problem and helps improve the model's performance [23]. A random forest has the following characteristics. It combines several decision trees from distinct subsets of the provided dataset and averages them to increase the dataset's predicting accuracy. The number of trees and precision has a linear relationship, which helps to avoid over-fitting. It is divided into two phases: In the first phase, a random forest is created by mixing N decision trees, and in the second phase, predictions are made for every tree formed in the former step. The Random Forest algorithm is solved by Eq. (5) to Eq. (8):

$$RFfi_{\mathrm{i}} = \frac{\sum_{\mathrm{j}eall\,treas} normfi_{ij}}{T} \tag{5}$$

$$normfi_{\mathrm{i}} = \frac{fi_i}{\sum_{jeall\,features} fi_j} \tag{6}$$

$$fi_{\mathrm{i}} = \frac{\sum_{\mathrm{jinode\ j\ splits\ on\ feature\ i}} ni_j}{\sum_{keallnodes} ni_k} \tag{7}$$

$$Ni_j = W_jC_j - W_{left(j)}C_{left(j) - W_{right(j)}C_{right(j)}} \tag{8}$$

Here ni_{i} means importance of node j, W_{i} = weighted number of samples reaching node j, C_{i} = the impurity value of node j, left(j)= child node from left split on node j, right(j) = child node from right split on node j, fi_{i} = the importance of feature i, $RFfi_{\mathrm{i}}$ = the importance of feature i calculated from all trees in the Random Forest model, *normfi* = the normalized feature importance for i in tree j, T = total number of trees [19].

5) CatBoost: CatBoost is an open-source algorithm which uses gradient boosting on decision trees. It allows the use of non-numerical. It uses a combination of one-hot encoding and an advanced mean encoding [24].

6) LightGBM: Microsoft created LightGBM, a free and open-source distributed gradient boosting platform for machine learning. It's a decision tree-based gradient boosting framework that improves a model's efficiency while minimizing the utilization of memory. It uses a split strategy which is leaf-wise rather than level-wise, and consequently produces complex trees, which is the primary factor in attaining higher levels of accuracy [25].

G. Evaluation Metrics

Various evaluation metrics have been used such as accuracy; F1 score, precision, recall, root mean square error, and R score to test the performance of the applied machine learning models. Each one of them has been discussed briefly below:

Accuracy: The percentage of predicted values that match actual values is calculated by accuracy [25] and is shown in Eq. (9).

$$Acc = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ negative + False\ Positive + False\ Negative} \tag{9}$$

F1 Score: It is used to assess the authenticity of a test. The Harmonic Mean of precision and recall is the F1 Score which has ranges between 0 and 1. It tells you how accurate and how robust our classifier is [26]. It is calculated by Eq. (10).

$$F1\ Score = 2\frac{Precision*Recall}{Recall+Precision} \tag{10}$$

Precision: It is the number of positive classifiers divided by the total number of positives both true as well as false number of correct positive outcomes. [25]. It is calculated by Eq. (11).

$$Precision = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive} \tag{11}$$

Recall: The number of rational positive results divided by total relevant samples [25]. It is calculated by Eq. (12).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{12}$$

Root mean square error: Euclidean distance uses to indicate how far predictions differ from true measured values [27]. It is one of the most widely used criteria for assessing the accuracy of predictions [28]. The root mean square error is calculated by Eq. (13).

$$RMSE = \sqrt{\sum_{i=1}^{j} \frac{II_{x(j)-x(j)}II^2}{n}} \tag{13}$$

where n denotes the number of data points, x(j) refers to the j-th measurement, and x(j) belongs to the prediction which corresponds to the measurement.

R^2 score: It is known as coefficient of determination. It is the proportion of the variance in the dependent variable that is forecasted from the independent variable(s) [13] and is calculate by Eq. (14).

$$R^2 = 1 - \frac{SS_{RES}}{SS_{err}} \tag{15}$$

where SS_{RES} is the residual sum of squared errors and SS_{err} is the total sum of squared errors.

IV. Results

The performance of the algorithms based on the various parameters such as accuracy, recall, root mean square error, precision, and F1 score has been computed during testing phase as shown in Table IV along with the discussion.

TABLE IV. Computation of Various Metrics for Machine Learning Models

Algorithms	**Metrics**				
	Accuracy	***Recall***	***RMSE***	***Precision***	***F1 Score***
Decision Tree	72	71	2.31	77	71
Random Forest	**92**	**85**	**0.91**	**86**	**85**
Gradient Boosting	74	80	2.25	79	80
Xgboost	88	78	1.35	78	78
CatBoost	79	84	2.11	84	85
LightGM	80	84	1.98	85	85

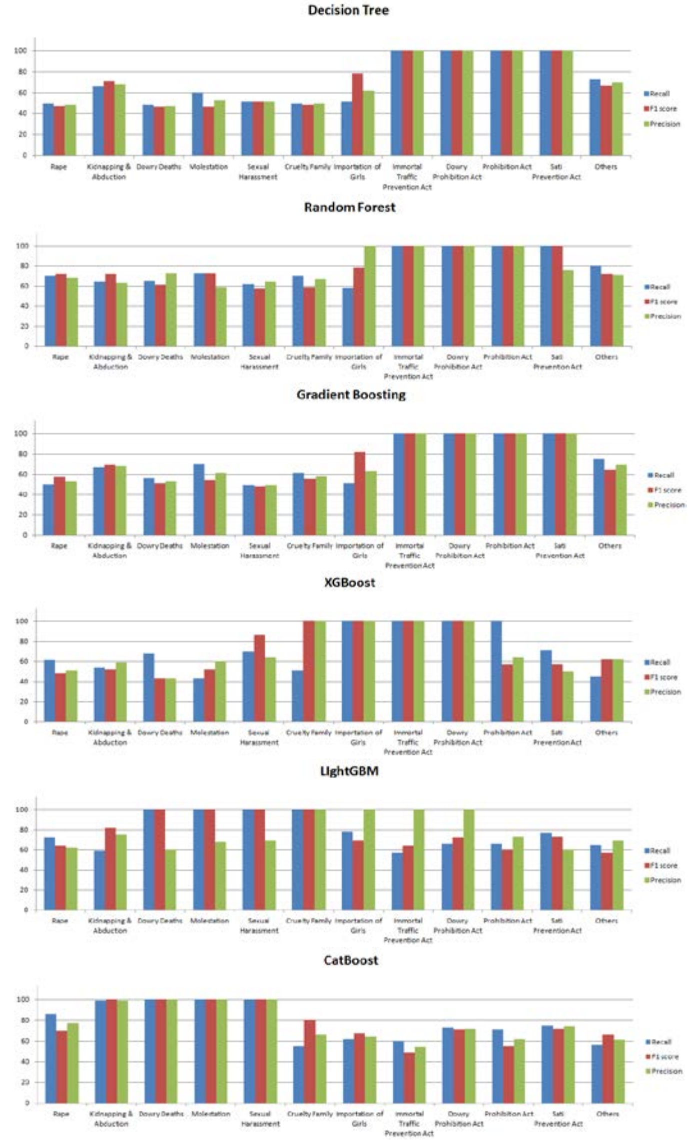


Fig. 13. Evaluation of Various Machine Learning Models using Three Different Metrics

On analyzing it can be said that random forest obtained the highest accuracy by 92%, recall by 85% and precision by 86%, F1 score by 85% on comparison with other algorithms. Besides this, each woman's recall, precision, and F1 score based on criminal activities such as dowry deaths, sexual harassment, importation of girls, prohibition act, sati prevention act, etc., have also been computed shown in Fig. 13. These parameters have been calculated to analyze the work of each learning model, as mentioned.

V. Discussion

In this research paper, the proposed system checks the credibility of the women based crime data by applying various machine learning models. Their performances have been evaluated using various evaluation metrics such as accuracy, recall, precision, F1 score, and root mean error square, with random forest achieving the highest accuracy of 92%, recall of 85%, the precision of 86%, F1 score of 85%, and the root mean square error value of 0.91 is obtained which is highest when compared to other algorithms. In addition to this, the work of various researchers in analyzing women's crime using different datasets has been also considered and compared with our study to understand our research work in a more efficient manner as shown in Table V.

TABLE V. Comparative Analysis of Previous Work with our Work

Ref	**Techniques**	**Dataset**	**Accuracy (%)**
[29]	Random forest	Crime in India dataset	86
[30]	Linear regression	National Crime Records Bureau (NCRB) crime data	83
[31]	Naïve Bayes	Collection of data from 2001 to 2012	81
[32]	KNN	Primary data	77
Our study	**Random forest**	**National Crime Record Bureau from year 2001 to 2019**	**92**

After comparing, it has been concluded that our study has achieved a great result in terms of accuracy for the National crime record bureau (2001-2019) with 92%, while linear regression has obtained less accuracy by 83% on applying the data that has been collected from the same repository. Overall, KNN has obtained the lowest accuracy value by 77% while working on the primary data.

VI. Challenges Faced during Research Work

During this research, we faced specific challenges while collecting the data, which is the most important step for crime-based research. Like this, the three main problems that we came across were:

1) No uniformity in Data: The data collected initially from 2001 to 2015 had a different data format and many parameters related to the case status. Still, the data from 2015 onwards was more compressed and less comprehensive. In addition, many crimes go unreported, and no one has a record of them.

2) Varying data formats: The data from 2001 to 2016 was in excel format and had a different format and parameters, and data from 2016 onwards was in pdf format and had a different set of parameters. A lot of time was taken to pre-process data into a desired, standardized format.

3) Size of the crime data: The data was huge as there are 11 sub-categories of crime against women, and each crime sub-category case status is stored. There are 13 types of case status. All this information is stored for 27 states and 8 Union Territories. The records taken were for 15 years. As the data size was huge, rectifying the problems with the data was an uphill task. Along with it, the data taken from books was compatible for analysis.

4) Missing values in crime data: There were a lot of missing values in the data, so the data cleaning was time-consuming. Hence, the SimpleImputer technique was used to handle missing data.

VII. CONCLUSION

The study focuses on the investigation of women's crime statistics such as rape, abduction, dowry deaths, molestation, sexual harassment, cruelty to family, importation of girls, immortal traffic, dowry prohibition act, prohibition act, sati prevention act, and others from 2001 to 2019 in various Indian states using data provenance and machine learning models. It has been discovered that Random Forest proved to be the best approach when compared to the others. Moreover, the proposed system demonstrates that the algorithms used for machine learning perform effectively in analyzing various crimes against women across states and union territories. Hence, the same technique can also be applied to search for information about other crimes in states with a higher crime rate. In addition, deep learning algorithms might be utilized to improve prediction accuracy while dealing with complex criminal cases in India.

REFERENCES

[1] Das, Priyanka, and Asit Kumar Das. "Crime analysis against women from online newspaper reports and an approach to apply it in dynamic environment." 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). IEEE, 2017.

[2] Hagan, Frank E. Crime types and criminals. Sage, 2009.

[3] Balasubramanian, t. “Violence against women’s in India”. Journal of shanghai jiaotong university, 876-892, 2020.

[4] Saravanan, Parthasarathy, et al. "Survey on crime analysis and prediction using data mining and machine learning techniques." Advances in Smart Grid Technology. Springer, Singapore, 2021. 435-448.

[5] Mittal, Mamta, et al. "Monitoring the impact of economic crisis on crime in India using machine learning." Computational Economics 53.4 (2019): 1467-1485.

[6] Shah, Neil, Nandish Bhagat, and Manan Shah. "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention." Visual Computing for Industry, Biomedicine, and Art 4.1 (2021): 1-14.

[7] Braga, Anthony A., et al. "Hot spots policing of small geographic areas effects on crime." Campbell Systematic Reviews 15.3 (2019): e1046.

[8] Nakib, Mohammad, et al. "Crime scene prediction by detecting threatening objects using convolutional neural network." 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). IEEE, 2018.

[9] Englbrecht, Ludwig, et al. "Enhancing credibility of digital evidence through provenance-based incident response handling." Proceedings of the 14th International Conference on Availability, Reliability and Security. 2019.

[10] Telugu M.,Vaddemani Sai M., K V Sai S. ,G. Shriphad R. Crime Data Analysis Using Machine Learning Models”, IJAST, vol. 29, no. 9s, pp. 3260 – 3268, 2020.

[11] Tamilarasi, P., and R. Uma Rani. "Diagnosis of crime rate against women using k-fold cross validation through machine learning." 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2020.

[12] Anjali, “ Women empowerment and constitutional provisions”. In Legalserviceindia.com.

[13] Qi, Zhang. "The text classification of theft crime based on TF-IDF and XGBoost model." 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, 2020.

[14] Shermila, A. Mary, Amrith Basil Bellarmine, and Nirmala Santiago. "Crime data analysis and prediction of perpetrator identity using machine learning approach." 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018.

[15] Amponsah, Wellington, and Parvinder Kaur. "Exploratory Data Analysis And Crime Prevention Using Machine Learning: The case of Ghana.".

[16] Mishra, Shivani, and Suraj Kumar. A comparative study of crimes against women based on Machine Learning using Big Data techniques. No. 4376. EasyChair, 2020.

[17] Tamilarasi, P., and R. Uma Rani. "Predict the Crime Rate Against Women Using Machine Learning Classification Techniques." Data Science and Its Applications. Chapman and Hall/CRC, 2021. 295-313.

[18] Durgapal, Vartika Hari. "Crime Based Evaluation of GPS Network Using Machine Learning Techniques." 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2022.

[19] Prabakaran, S., and Shilpa Mitra. "Survey of analysis of crime detection techniques using data mining and machine learning." Journal of Physics: Conference Series. Vol. 1000. No. 1. IOP Publishing, 2018.

[20] Ivan, Niyonzima, et al. "Crime Prediction Using Decision Tree (J48) Classification Algorithm." (2017).

[21] Patel, Nisarg P., et al. "Fusion in Cryptocurrency Price Prediction: A Decade Survey on Recent Advancements, Architecture, and Potential Future Directions." IEEE Access 10 (2022): 34511-34538.

[22] Ghankutkar, Surabhi, et al. "Modelling machine learning for analysing crime news." 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2019.

[23] Mistry, Chinmay, et al. "MedBlock: An AI-enabled and blockchain-driven medical healthcare system for COVID-19." ICC 2021-IEEE International Conference on Communications. IEEE, 2021.

[24] Kumar, Yogesh, Komalpreet Kaur, and Gurpreet Singh. "Machine learning aspects and its applications towards different research areas." 2020 International conference on computation, automation and knowledge management (ICCAKM). IEEE, 2020.

[25] Rawat, Romil, et al. "Analysis of darknet traffic for criminal activities detection using TF-IDF and light gradient boosted machine learning algorithm." Innovations in electrical and electronic engineering. Springer, Singapore, 2021. 671-681.

[26] Sachin Bhardwaj Yogesh Kumar, M. K. P. K. R. Recent Trends of Data Mining in the Field of Intrusion Detection System. Journal of Critical Reviews, 7, 2360–2365,2020.

[27] Parekh, Raj, et al. "DL-GuesS: Deep Learning and Sentiment Analysis-based Cryptocurrency Price Prediction." IEEE Access 10 (2022): 35398-35409.

[28] McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyze crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015): 1-12.

[29] Sarah, J., Danny, A., Deen, J., Dongre, L., Chitransh, V. and Ramchandhani, H. Analysing Crimes of Indian Datasets Based on Machine Learning Methods. In vol 12, issue 8, 2415-2435, 2021.

[30] Ravi Teja, K., et al. "Analysis of Crimes against Women in India Using Machine Learning Techniques." Communication Software and Networks. Springer, Singapore, 2021. 499-510.

[31] Prasad, D. S., Rachit Sharma, and V. Anbarasu. "Analysis and Prediction of Crime against Woman Using Machine Learning Techniques." Annals of the Romanian Society for Cell Biology 25.6 (2021): 5183-5188.

[32] Mahmud, Sakib, Musfika Nuha, and Abdus Sattar. "Crime rate prediction using machine learning and data mining." Soft Computing Techniques and Applications. Springer, Singapore, 2021. 59-69.

[33] Sonal Singh. "Leveraging ML to Predict Crime Against Women" International Journal of Engineering Research & Technology (IJERT) ,Volume 11, Issue 01 (January 2022),

[34] Stoffel, Florian, et al. "VAPD: A Visionary System for Uncertainty Aware Decision Making in Crime Analysis." Symposium on Visualization for Decision Making Under Uncertainty at IEEE VIS 2015. 2015.

[35] https://ncrb.gov.in/crimeinindiatablecontents?field_date_value[value][year]=2019&field_select_additional_table_ti_value=All&items_per_page=All.

[36] https://www.kaggle.com/datasets/khanmohammadanas/district-wise-crimes-in-india.