International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Prediction & Classification of Crimes Against Women

[1]Ankit Agarwal, [2]C.Malathy, [3]Nilesh Mukherjee

Abstract--*Crime is one of the significant social-issue in a nation influencing open well-being and the most ideal approach to end violations is to keep it from occurring in any case by tending to its root and auxiliary causes. In this paper it is proposed to focus on the 11 different types and sub-types of crimes (rape, dowry deaths, cruelty by husband or his relatives, indecent representations of women, immoral trafficking, etc) related to women and is a major concern for their safety. Our work involves various tasks to be achieved which can be classified as : cleaning the data set, analysis and visualization of the data for better understanding , predicting the crime rate for the upcoming year using supervised machine learning algorithm and finally make a classification of top 3 crimes(with the help of plotly dash to make a simple yet effective web-interface for the user).*

Key words--*data set, analysis and visualization, machine learning algorithm, user-interface*

I. INTRODUCTION

Criminal law and sociology scholars have also been studying the underlying pattern of crime and it's relation to a region or an area's social or economic development , the characteristics of different group of people living there, family structure, level of education among other things. Different studies and researches have shown that significant concentration of crime happens at micro level of a region.

The standard of sex balance is revered within the Constitution of India in order to maintain and execute the Constitutional Mandate, the state has sanctioned different laws and brought measures expected to affirm equivalent rights, check social separation and different sorts of brutality and atrocities, though ladies could likewise be casualties of any broad violations like homicide, burglary and so forth. Exclusively the violations that square measure coordinated explicitly against ladies i.e, sexual orientation explicit are described as 'Wrongdoings against Women'.

The principal section of our work is to clean the informational collection which is finished by distinguishing the off base, fragmented, erroneous, immaterial or missing piece of the information and afterward altering, supplanting or erasing them as indicated by the need and afterward utilize the amended arrangement of information to make an appropriate perception so as to make the comprehension of information clear and simple. Second segment will be to use the linear regression algorithm to forecast the crime trends for every state in accordance of the crimes and we will be comparing the accuracy of the result obtained with that of K-Nearest Neighbour alone.

[1]*Computer Science & Engineering, SRM Institute of Science & Technology, Chennai, India, ankitagarwal_suresh@srmuniv.edu.in*
[2]*Computer Science & Engineering, SRM Institute of Science & Technology, Chennai, India, malathyc@srmist.edu.in*
[3]*Computer Science & Engineering, SRM Institute of Science & Technology, Chennai, India, nileshmukherjee_ashutosh@srmuniv.edu.in*

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

Finally, we will be using dash (Dash is a beneficial Python system for building web applications, composed over Flask, Plotly.js and React and we can convey our applications to servers and afterward share them through URLs) for making a user interface consisting of various features such as the crime trends, top 3 crimes and analysis of the crimes giving the suggestions for the betterment of women.

Rest of the paper is sorted out as: Section II of this paper is about various methodologies that have been utilized used for the work related to crime. Section III is all about how we are proceeding further with our work and gives a proper detail of every step that has been followed throughout. Section IV is about the evaluation of the data set , Section V is all about the results obtained lastly, Section VI finishes up the paper.

II. RELATED WORK

For the past years researchers and scholars across the world have been working on crime related work using various approaches such as big data, data mining, machine learning and many more to understand the pattern and minimize the crimes using the result obtained as a result of their work. In this section we will be briefly reviewing some of the work.

In [1] the authors have talked about Big Data based information diagnostic approaches. After gathering, planning, bringing in information, and Analysis utilizing the information various modules of Hadoop structure, perception and expectation will be through Naïve Bayes Machine learning calculation. RapidMiner instrument has been utilized to play out this and to check the precision of the model too.

In Prediction of Crime Rate Analysis [2] the authors have made a legitimate examination of informational collection utilizing different regulated calculation. Choice Trees or Random Forest calculations with exactness of 98%, the most elevated among the remainder of the calculation. The least which can be utilized will be SVM. 17.37 % of wrongdoing are vandalism and 16.38 % are robberies.

The authors of [3] have utilized the informational index for the year 2001 to 2014 for foreseeing the crime percentage utilizing different AI calculations. Creator have made two unique groups of classes for example group 1 : high number of individuals associated with the wrongdoing. Bunch 2: low number of individuals engaged with the wrongdoing and utilizing this group of class the creator is then attempting to picture the districts with high or low number of violations.

The authors of [4] have talked about an orderly methodology of Big Data Analytics and a few information mining and profound learning advancements. The author has utilized informational collection that was gathered for 3 distinct urban areas which are San Francisco, Philadelphia and Chicago. Utilizing the informational collection got, the creators are attempting to construct a model that can anticipate the crime percentage and the significant violations for the forthcoming years, the main 3 wrongdoings for San-Francisco is Theft, different offenses and non-criminal; for Chicago top 3 wrongdoings are robbery, battery and criminal harm and for Philadelphia the violations are for the most part different offenses, different ambushes and burglary. At last, the authors talk about how the neural model was outperformed by different models that have been utilized (Prophet Model and Keras).

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

In [5] the authors have utilized the informational collection for the Vancouver city for as long as decade. Also, their work for this paper is isolated into two distinct parts. Initial segment of the work is to apportion a one of a kind number to the area and the wrongdoing class when a wrongdoing happens in wherever. What's more, in the second methodology paired number stamped 1 was designated when a wrongdoing occurred in a day or in the week for a specific neighborhood and in the event that there was no wrongdoing submitted, at that point the number 0 was assigned lastly the expectation precision was between 39% to 44%.

Author of paper [6] have utilized two distinct informational collection for Denver, CO and Los Angeles. For Denver, the base help esteem was 0.0012, which compares to 277 total frequencies and for Los Angeles the base help esteem was 0.0018, which relates to 354 supreme frequencies. Accuracy of Gausian Bayes was 51% for Denver and 54% for Los Angeles whereas accuracy of Decision Tree classifier was 42% for Denver and 41% for Los Angeles.

In paper [8] the author use WEKA to lead a similar report between the fierce wrongdoing designs from the Communities. Furthermore, the dataset being utilized for the work was taken from the University of California-Irvine storehouse and the genuine wrongdoing factual information was taken for the territory of Mississippi. Also, the anticipated wrongdoing aggregate for the weka apparatus was 179 for homicide, 1076 for assault, 2485 for burglary and 4635 for attack. The outcome shows that the viability and exactness of the direct relapse

We are completely focusing on the crimes related to women in India and are trying to make a simple and effective user-interface for the crime trends across the year and make a classification of top 3 crimes and provide a complete analysis of the crimes and measures that should be taken .

III. METHODOLOGY

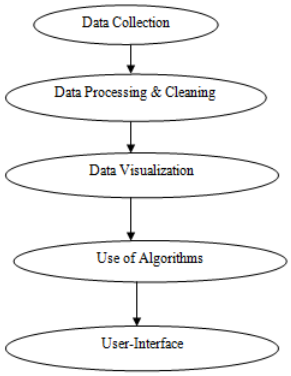

Fig 1

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

A. Data Processing & Cleaning

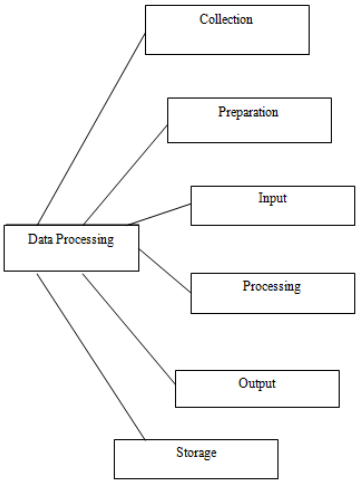

Fig 2

Data Processing is a method of converting data from a given source to a much usable form. Collection is the process of obtaining the dataset and various free sources of dataset can be Kaggle, data.gov.in, UCI a data repository.

Preparation is the process where the data collected from different sources are analyzed according to our need, this prepared data may not be in machine readable format, so we need to convert this data into machine readable form with the help of conversion algorithm. Processing is the stage where we make use of algorithms to perform the instructions provided over a huge amount of data and the results obtained are procured in a manner that it can be easily inferred by the user such as graphs and tables.

Having a dataset that is liberated from inconsistencies, for example, missing information, terrible information and copies, insignificant highlights and on occasion anomalies is the main essential advance towards Machine Learning work process. Information cleaning might be clarified as a procedure where we channel and adjust these abnormalities to investigate, comprehend and model effectively. The anomalies in a dataset is a diverse assortment and one can say that they may contain some key data since they are not quite the same as the fundamental gathering and then again they lose our perspective on the primary gathering since we need to watch so out of sight to see exceptions. Awful information just implies that any information focuses or values that shouldn't be in the dataset and this can be taken care of either by dropping or utilizing some savvy substitution.

There are numerous approaches to deal with information in python, for example,

Checking for NANs: This distinguishes both 'NANs' and 'Non pd.isnull(object)

Dropping information: Returns the information outline where any information focuses containing NANs have been expelled df.dropna(axis=0, how=’any’)

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

Replacing the data: Replace the values given in 'to_replace' with 'value'. df.replace(to_replace=None, value=None)

B. Data Visualization

As an individual we are utilized to effectively get a handle on the data from a diagrammatic portrayal. Henceforth, information representation has gotten mainstream as of late as it has the ability to show the outcomes in such a way,that can be effortlessly comprehended by any person. It is progressively being utilized as a device for exploratory information investigation before applying AI models. Not many of the plots that are accessible to us are: disperse plot, box plot, bar outline, line plot, appropriation plot and so on.

C. Use of Algorithms

For our paper we will utilize directed AI calculation and the term regulated AI can be characterized as the procedure where we have the information variables(x) and a yield variable (Y) and we utilize one of the calculations.

$Y = f(x)$ --------- 1

And the supervised machine learning is further divided into two parts:

a) Regression: When the yield variable is a genuine value ,such as 'rupee' or 'price of a car'.

b) Classification: When the yield variable is a class, such as 'will it rain' and 'will it not rain' or 'red' and 'blue'.

1) Linear Regression

Linear regression is mostly preferred when we are trying to find a relation between two continuous variables. One of the variable can be named as indicator or free factor and other is reaction or ward variable. Statistical relation is what linear regression looks for and not the deterministic relationship

Model with one predictor uses:

$b_o = \bar{y} - b_1\bar{x}$ ----------2

To calculate the error (if any) we can use:

$Error = \sum_{i=1}^{n}\left(actual_{output} - predicted_{ouput}\right) ** 2$ ---3

2) K-Nearest Neighbour (KNN)

This algorithm belongs to the category of pattern recognition methods. The strategy does not impose a priori any assumptions regarding the distribution from which the modelling sample is drawn. A training set with both positive and negative values is used for KNN By ascertaining the separation to the closest neighbouring preparing case, another example is arranged and the indication of this point will help in deciding the order of the example. The presentation of the kNN calculation is mostly affected by 3 variables: (1) the separation measure used to find the closest neighbours; (2) the choice standard used to get an order from the k-closest neighbours; and (3) number of neighbours used to characterize the new example.

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i) \wedge 2} \qquad \text{--------4}$$

D. User-Interface

About the technology being used:

Dash is a python system which is utilized for building web applications. It is composed over Flask, Plotly.js, and React.js and is perfect for building information perception applications with exceptionally custom UIs in unadulterated Python. Run abstracts away the entirety of the advances and conventions that are required to manufacture an intelligent electronic application, through several basic examples Dash applications are rendered in the internet browser. We can send our applications to servers and afterward share them through URLs. Run is naturally cross-stage and versatile prepared, since Dash applications are seen in the internet browser.

On completion of building the model using the algorithm we will be using dash to make an easy yet effective web-based application. This application will show the trends of various crimes against women for 2 decades and the user can easily select their state and type of crime.

IV. EVALUATION

Our dataset was obtained from various sites such as Kaggle and data.gov.in. The raw dataset had few irregularities such as missing data , null value and irrelevant data , using the methods mentioned in the section III(under data processing and cleaning)we removed these irregularities from the dataset.

Table 4.1: Dataset used

State	Crime head	Years
All The Indian State	1.Rape 2.Kidnapping & Abduction 3.Dowry Deaths 4.Assualt on women with the intent to outrage her modesty 5.Insult to the modesty of women 6.Cruelty by husband or his relatives 7.Importation of Girls from foreign country 8.Immoral traffic 9.Dowry 10.Indecent representation of women 11.Commission of sati	2001 to 2019

Irregularities that were removed were mainly under the crime head “Commission of Sati”,“ Indecent Representation of women”,” Importation of girls from foreign country” and all the irregularities were replaced by

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

0(zero). Finally, the rectified dataset was used for further process. We have used 70% of data for training and 30% for testing.

V. RESULTS

Data Visualisation

Figure 5.1: Crime head v/s Crime Rate year (2001)

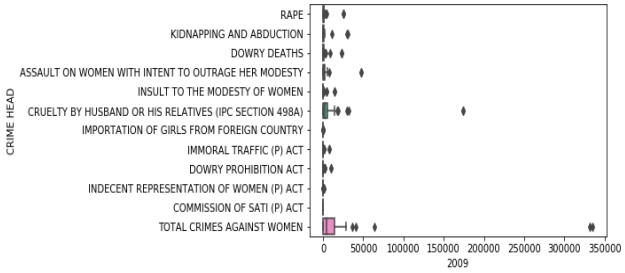

Figure 5.2: Crime head v/s Crime Rate year (2009)

Table 5.1: Change in crime for 2001 and 2009

Year	Top 3 Crime Head	Year	Top 3 Crime Head	Change
2001	Cruelty by husband Assault On Women Rape	2009	Cruelty by husband Assault On Women Rape	37.23% increase 11.72% increase 20.88% increase

The above table depicts the change in percentage of crimes for the year 2001 and 2009. For the year 2001 the exact number for the top 3 crime heads were 1,09,467, 42,244 ,20,446 respectively. And for the year 2009 it was near about 1,74,395 , 47,856, 25,845 respectively . The total crime for the year 2001 is 2,43,589 and for the year 2009 it is 3,35,336 and this shows that the overall crime rate has drastically increased from the year 2001 to 2009.

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

Results for the Model Built

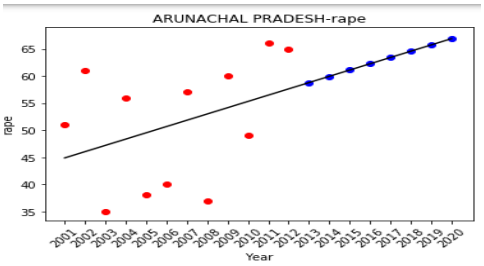

Figure 5.3: Graph representing rape crime head for Arunachal Pradesh

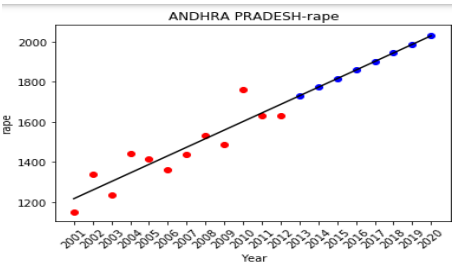

Figure 5.4: Graph representing rape crime head for Andhra Pradesh.

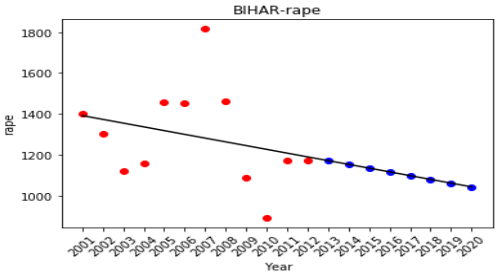

Figure 5.5: Graph representing rape crime head for Bihar

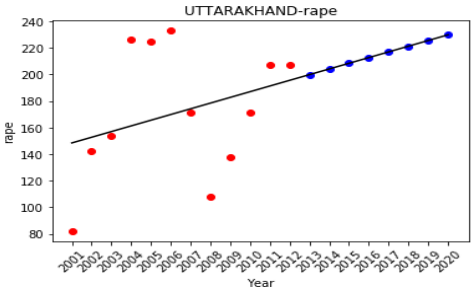

Figure 5.6: Graph representing rape crime head for Uttrakhand

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

Table 5.2: Variation of crime for year 2010 and 2020

Type Of Crime	Year 2010	Year 2020	% Change
1.Cruelty By Husband and his relative	1,80,413	1,81,058	0.35 % increase
2.Assault on women	50,024	49,376	1.31% decrease
3.Rape	27,074	27,184	0.40% increase
4.Kidnapping & Abduction	34,250	31,686	8.091% decrease
5.Dowry Deaths	23,280	22,799	2.109% decrease
6.Insult to the modesty of women	10,404	12,984	19.87% increase
7.Immportation of girls from foreign countries	81	-38	313.157% decrease
8.Immoral Trafficking	7,731	7,727	90.61% decrease
9.Dowry Prohibition Act	12,080	10,101	19.58% decrease
10.Indecent Representation Of Women	958	2,172	55.89% increase
11.Commission of Sati	0	0	No change
12. TOTAL	3,45,339	3,45,045	0.08% decrease

The reason for such increase in the crime can be:

1. No strict action has been taken against the culprits.
2. Police protection is not up to the mark.
3. Many cases of crimes are unreported.
4. The family member at times doesn't support the victim.
5. Assault on women deals with sexual assault, physical assault, etc. and the victims are threatened to not go to the police and this results in fear and the culprit doesn't face any consequences.

Results of the User Interface

Figure 5.7: Dropdown Menu for selecting the preferred state

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

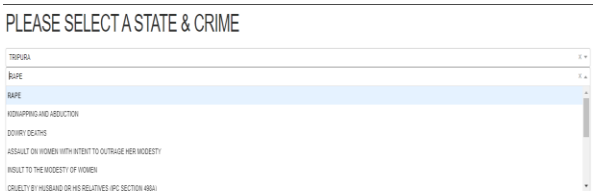

Figure 5.8: Drop down Menu for selecting the type of crime head

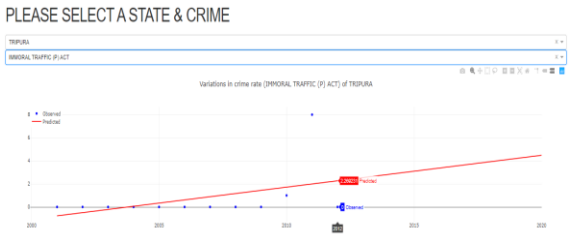

Figure 5.9: Graph for the selected attributes.

VI. CONCLUSION

This paper exhibited the utilization of supervised machine learning algorithm to build a model for predicting the crime rates for all the Indian States and was completely focused on crimes that are specifically related to women. And the result obtained was then implemented with the KNN algorithm for computing the accuracy of the result obtained and the accuracy of Linear Regression was far better than K-nearest Neighbour Algorithm. And the user interface designed in this paper will be helpful to the state police as well as common people.

REFERENCES

1. Pranay Jha, Raman Jha, Ashok Sharma," Behavior Analysis and Crime Prediction utilizing Big Data and Machine Learning", International Journal of Recent Technology and Engineering, May 2019.
2. Kirthika, Krithika Padmanabhan, Lavanya, Lalitha S.,"Prediction of Crime Rate Analysis Using Supervised Classification Machine Learning Approach", International Research Journal Of Engineering and Technology, March 2019.
3. Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav, "Wrongdoing Pattern Detection, Analysis and Prediction", ICECA, 2017.
4. Mingchen Feng , Jiangbin Zheng, Jinchang Ren, Amir Hussain , Xiuxiu Li , Yue Xi , and Qiaoyuan Liu , "Enormous Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data", IEEE, 2015
5. Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri, "Wrongdoing Analysis Through Machine Learning", Fraser International College, Simon Fraser University.
6. Tahani Almanie, Rsha Mirza and Elizabeth Lor, "Wrongdoing Prediction Based on Crime Types and utilizing spatial and worldly criminal hotspots", International Journal of Data Mining and Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015.
7. Sheena Rewari, Dr.Williamjeet Singh," System Review of Crime Data Analytics", IEEE, 2017.
8. Lawrence McClendon and Natarajan Meghanathan," Using Machine Learning Algorithms to Analyze Crime Data ", Machine Learning and Applications: An International Journal (MLAIJ), March 2015.
9. P. Chen, H. Yuan, and X. Shu, "Guaging wrongdoing utilizing the arima model," in Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference.
10. C.- H. Yu, M. Ward, M. Morabito, and W. Ding, "Wrongdoing guaging utilizing information mining methods," in Data Mining 2011 IEEE eleventh International Conference on, 2011.

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

11. B. Chandra, M. Gupta, and M. Gupta, "A multivariate time arrangement bunching approach for wrongdoing patterns expectation," in Systems, Man and Cybernetics, IEEE International Conference on, 2008.
12. S. V. Nath, "Wrongdoing design discovery utilizing information mining," in Web Intelligence and Intelligent Agent Technology Workshops, 2006. WIIAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on, 2006.
13. M. Tayebi, M. Ester, U. Glasser, and P. Brantingham, "Crimetracer: Activity space based wrongdoing area expectation," in Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014.
14. 14 H. Wang, D. Kifer, C. Graif, and Z. Li, "Crime percentage induction with enormous information," in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. ACM, 2016.
15. 15 C. Catlett, T. Malik, B. Goldstein, J. Giuffrida, Y. Shao, A. Panella, D. Eder, E. van Zanten, R. Mitchum, S. Thaler, and I. T. Encourage, "Plenario: An open information disclosure and investigation stage for urban science," IEEE,2014.
16. V. N. Gudivada, D. Rao, and V. V. Raghavan, "No SQL Systems for Big Data Management," IEEE,2014.
17. D. Keim, H. Qu, and K. Mama, "Enormous Data Visualization," IEEE, July 2013.

DOI: 10.37200/IJPR/V24I8/PR280264
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020