Analysis and Prediction of Crime against women in India using machine learning algorithms.

Vinay Narayan Bhat[1], V Santhosh Kumar [2], Prof. Saravanan C [3]
PG Student[1,2], Assistant Professor[3]
[1,2,3] Department Master of Computer Applications, RV College of Engineering® , Mysore road, Bangalore, Karnataka, India.

***Abstract*:** In India, according to the National Commission for Women (NCW), 23,722 complaints of crimes against women were received in 2020. Crime detection and prevention is a very crucial work which is in the hands of police, law enforcement agencies and local government. Crime against women is increasing at an alarming rate in almost all parts of India. Women in the Indian society have been victims of humiliation, torture and exploitation. This paper focuses on analysis of various types of crimes against women in various states of India. Linear regression and random forest algorithm was used to predict the crime rate against various crime types against women. Linear regression algorithm gave better prediction accuracy rate compared to random forest algorithm. A web portal was implemented using flask architecture and MongoDB to store data at the backend.

KEYWORDS: Crime Analysis, Flask Architecture, Linear Regression, Random Forest Classifier.

I. Introduction

In India, according to the National Commission for Women (NCW), 23,722 complaints of crimes against women were received in 2020, the highest in the last six years. It is a big threat to humanity and in some parts of India, women are still treated as prisoners. Violence against women is perhaps as old as mankind. Not only in rural areas of India but also in urban areas women face a lot of problems like dowry, trafficking, acid attacks, miscarriage, kidnapping, and abduction of women. Crime against women is majorly happening because of the inefficient legal justice system, weak rules of law, crime prediction, and criminal identification. There is a need to analyse and predict the crime rate and the necessary steps to be taken further by the government officials to avoid the increase in threats and crime against women.

II. Literature Survey

B. Sivanagaleela, S. Rajesh[1] proposed the project to identify the crime areas based on the clustering technique. They stated that crime patterns identified are not static. So they have identified the crime areas and which type of crime occurred very frequently in which place using the fuzzy clustering technique.

Keerthi.R, Kirthika.B, Pavithraa.S , Dr. V.Gowri[2] The authors have built various analytical process which are data cleaning and processing, Eliminating missing value, exploratory analysis and finally building the model and evaluation to utilize the resources identify the hotspots of crimes and allocate vigilante resources such as policeman, police cars, weapons etc. reschedule patrols according to the vulnerability of a place. Through that they could avoid crimes ensure better civilization through avoiding happening crimes such as murder, rapes, thefts, drug, smugglings etc.

Lavanyaa, D. Akila[3] The authors have identified the Dominant part of the examination works in Crimes against women by utilizing the 'WEKA Tools' for their usage; in this way they acquired the better outcomes by utilizing MATLAB. Credulous bayes was frequently utilized and mainstream calculation to arrange and anticipate in the ground of information mining. The Next most utilized calculation is Apriori. The abnormal state of precision has been exploited by Naïve Bayes calculation. So as to investigate the better grouping calculations to foresee Crimes in different urban areas and nations, interconnected examine credentials comprise to be assembled. Investigation demonstrates the preeminent piece of utilizing tools and calculations. Regard we found that clustering in WEKA utensils, Euclidean distance calculation gives the improved exactness in the metropolitan urban areas violations rate to decrease and foresee

Priyanka Das, Asit Kumar Das, Janmenjoy Nayak , Danilo Pelsui[4] This paper demonstrates an unsupervised approach of extracting relations from newspapers based on criminological data. The proposed work demonstrates an unsupervised approach of extracting relations from newspapers based on criminological data. The proposed clustering technique identifies significant crime patterns that can help both in the criminology and criminal justice industry and eventually it will help the law enforcement agencies to analyze crime at a faster pace.

Bhajneet Kaur, Laxmi Ahuja[5] In this paper the authors have detected and predicted crime against women, using various data mining techniques by many researchers the authors have used the Indian Crime dataset and few used the techniques on US and England datasets. Most of the techniques used by the researchers are classification and clustering for crime pattern and detection. In the classification some authors depicted the Naïve bayes, some used decision trees and others used Bayesnet, J48, JRip & OneR. Correlation and regression techniques are also used for the crime analysis against women. In this paper review has been done on various techniques of data mining used for crime against women.

Based on the outcome of Literature survey, it has been observed most of the work has been implemented using Naïve Bayes Algorithm and K-Means Clustering Algorithm. Analysis and Prediction together has not been discussed together. There is no dedicated portal implemented to demonstrate the analysis and prediction of crimes against women in India.

III. Methodology

This section would demonstrate the methodology and the principle in an elaborating way. The objective of the paper is to analyze and predict the crimes against women in India for the years (2020-2024). Visualization of analysis and prediction of crime against women in India with the data gathered from National Criminal Records Bureau(NCRB) which was used in order to predict the rate of crime against women and its severity in the forthcoming years, in different areas of the country based on the previous year criminal records of India. Data Preprocessing plays a vital role before implementing the machine learning algorithms. From the Dataset acquired from the government portal there were undefined and unnamed features which needed to be dropped as they were not related and had no meaning to adding these features to the algorithms. The system uses linear regression Algorithm in major section i.e. Predicting the Crime Patterns with minimum accuracy rates as a factor. Linear Regression is the suitable Algorithm to predict the pattern of this data compared to other algorithms used for testing purposes such as Random Forest Algorithm, and found linear regression as the best fit for this proposed system with maximum accuracy in the results. The system is scalable as the dataset increases, the variation in the visualization of analysis and the accuracy of the prediction increases. As the increase in the system's workload, the system would be able to process and so the system is scalable to an apex.

Flask Architecture and Web application

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or the other components where pre-existing third-party libraries provide common functions. Flask supports extensions which will add application features implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and a number of other common framework related tools.

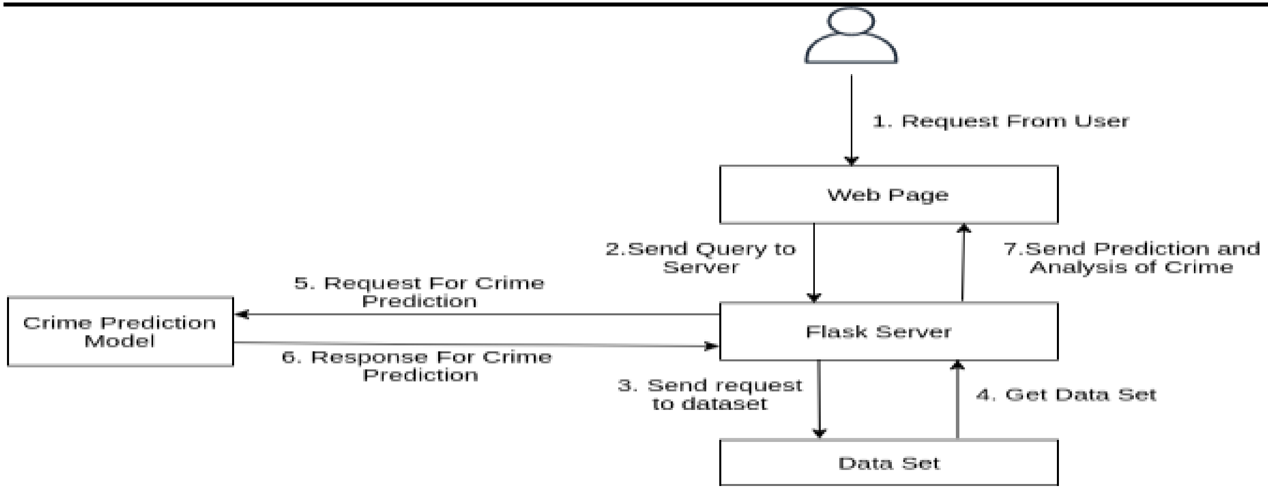

1. Request From User
Web Page
2.Send Query to Server
7.Send Prediction and Analysis of Crime
5. Request For Crime Prediction
Crime Prediction Model
Flask Server
6. Response For Crime Prediction
3. Send request to dataset
4. Get Data Set
Data Set

Figure 1: Block diagram of the Application

The Figure 1 describes the proposed system of the application. Python programming language is used for implementation. The user will request the analysis or prediction of crime against women in India using three inputs: year, state of India in which the user wants to visualize and the type of crime category. After fetching the inputs from the web page, the query values will be fetched to the server which will request for the crime prediction or analysis visual from the model which in turn would be fetched back to the user.

The system requires dataset of the Crime against women and the proposed system would present the supplied data in some visuals such as line graph and bar charts as output with the prediction of the crime against women in India for upcoming four years.

Data Definition

In the proposed work, the datasets have been collected from the National Crime Records Bureau (NCRB) crime data which provides data of various crimes for public use. The data collected for the work contains crime information State/Union territories wise crimes against women of all the 28 states (Andhra Pradesh inclusive of Telengana) and 7 union territories. Several crime types like ‘rape’, ‘kidnapping and abduction’, ‘dowry death’, ‘assault on women with intent to outrage her modesty’, ‘insult to the modesty of women’, ‘cruelty by husband or his relatives’, ‘importation of girls’ have also been gathers. The dataset was prepared with the above stated attributes of crimes against women in India. The dataset is converted to suitable format using data preprocessing techniques such as eliminating missing values, eliminating redundant data and data transformation.

Data Visualization

Crime type predictions are performed, for four years, for each state as well as all the states of India using the data from year 1990-2019. These predictions are displayed using simple visualization charts. These visualizations would provide the comparative study between the various crimes as well as between crimes in different region. The system is not supposed to change the information without permission, this implies the system is secure and accurate information is presented on the platform.

The following are machine learning algorithm implemented for the work

1. Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical procedure that's used for predictive analysis. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. The linear regression model provides a sloped line representing the connection between the variables. Mathematically, we can represent a linear regression as:

a. $y= a0+a1x+ \varepsilon$

b. Y=Dependent Variable (Target Variable)

c. X= Independent Variable (predictor Variable)

a0= intercept of the road (Gives a further degree of freedom)

a1 = rectilinear regression coefficient (scale factor to every input value).

ε = random error

d. The values for x and y variables are training datasets for Linear Regression model representation.

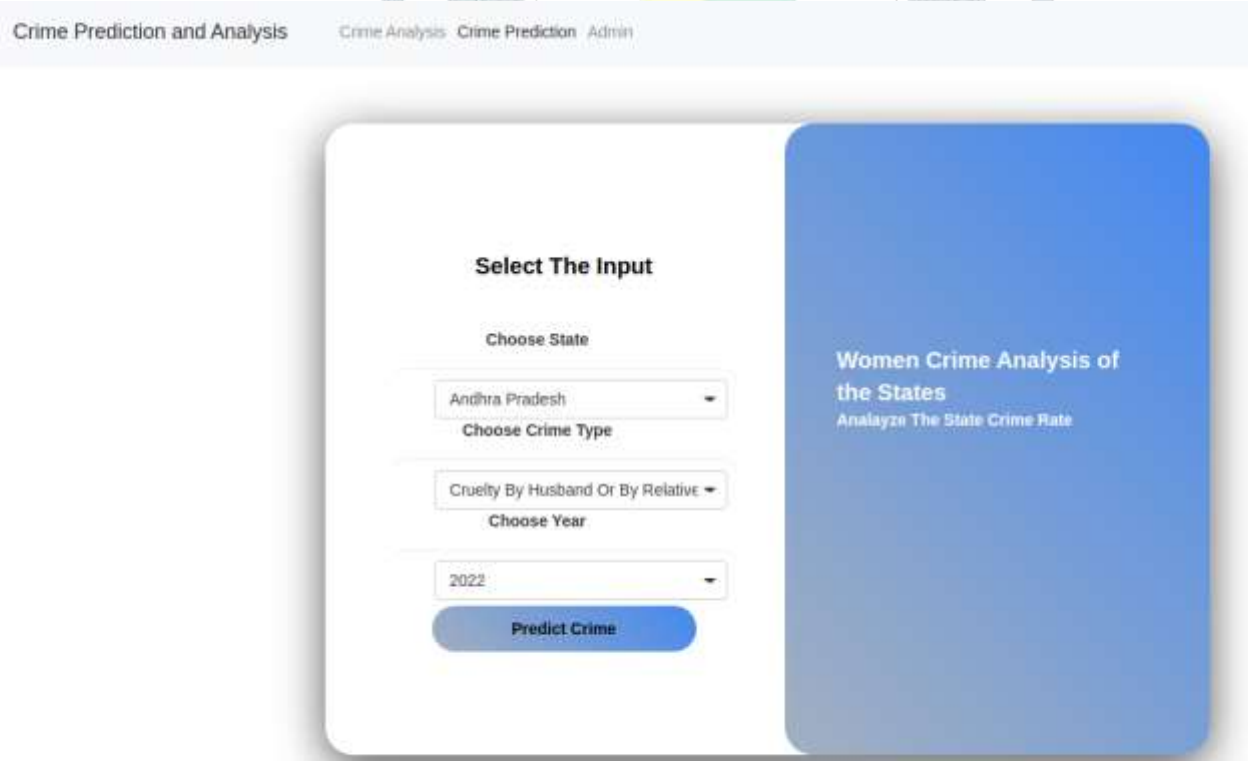
2. Random Forest: Random Forest is another classification method, also known as improved CART method (Classification and Regression Trees). It is based on ensemble learning and using that it makes a lot of classification trees. Every can be built using a single deterministic algorithm or they can be built using different algorithms. Their built depends on two factors. a. A best split is chosen at each node from a random subset. b. These trees are built using two-third to create the model and rest is employed to predict the accuracy.

IV. Implementation

In the proposed work the data is first pre-processed by removing the missing value and columns which are not needed for the analysis and prediction model. Then we are splitting the data into X and Y attributes, where Y contains column class and its corresponding row entities and X attribute contains all other column and their corresponding row entries with the help of sklearn_split_data library. The test and train data is split into 3:1 ratio in which the training data set size would be 0.80 of the total dataset size and the testing dataset size would be 0.20 of the total data set. After dividing the training and testing dataset, Machine learning algorithms Random Forest algorithm and Linear Regression which are compatible and relevant for our dataset was applied.

Random Forest was implemented to the train data which was 80% of the total dataset. Then , we tested the model using the test data which consists of 20% of the total dataset . Accuracy score of the predictive model was observed to be 0.7692307692307693 which was 76.923% accuracy.

Linear Regression was implemented on the prediction model and it was observed an accuracy of 83% percentage. Linear Regression gave better accuracy result when compared with Random Forest Algorithm.



Select The Input

Choose State

Andhra Pradesh

Choose Crime Type

Cruelty By Husband Or By Relative

Choose Year

2022

Women Crime Analysis of the States

Analayze The State Crime Rate

Figure 2 : Web Portal of the application

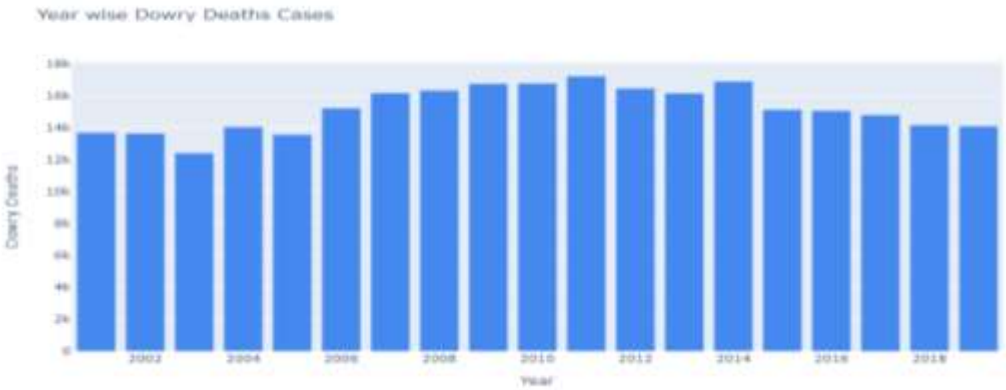


Figure 3 : Year wise Analysis of Crime (Dowry death Category) in India

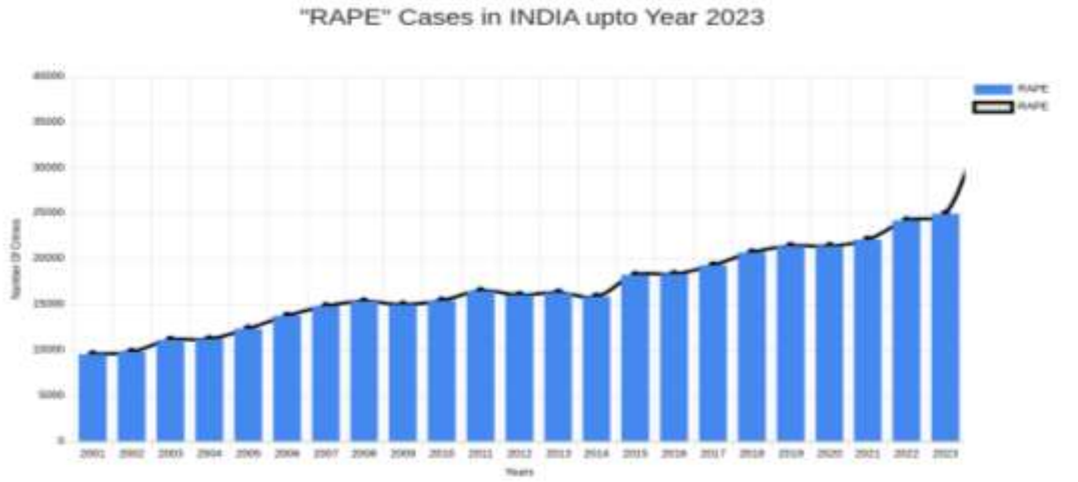


Figure 4 : Prediction of Crime (Rape case Category) in India up to year 2023

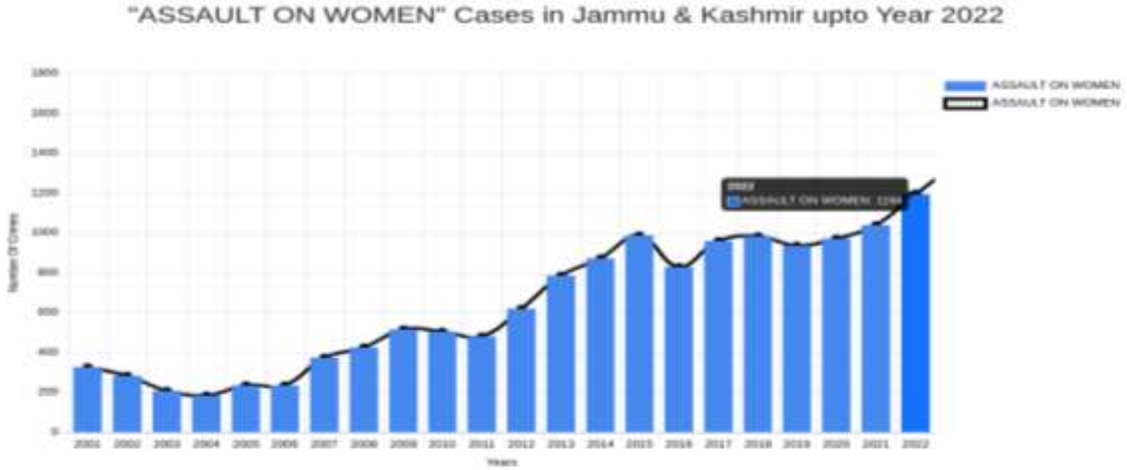


Figure 5 : Prediction of Crime (Assault on Women) in Jammu and Kashmir up to year 2022

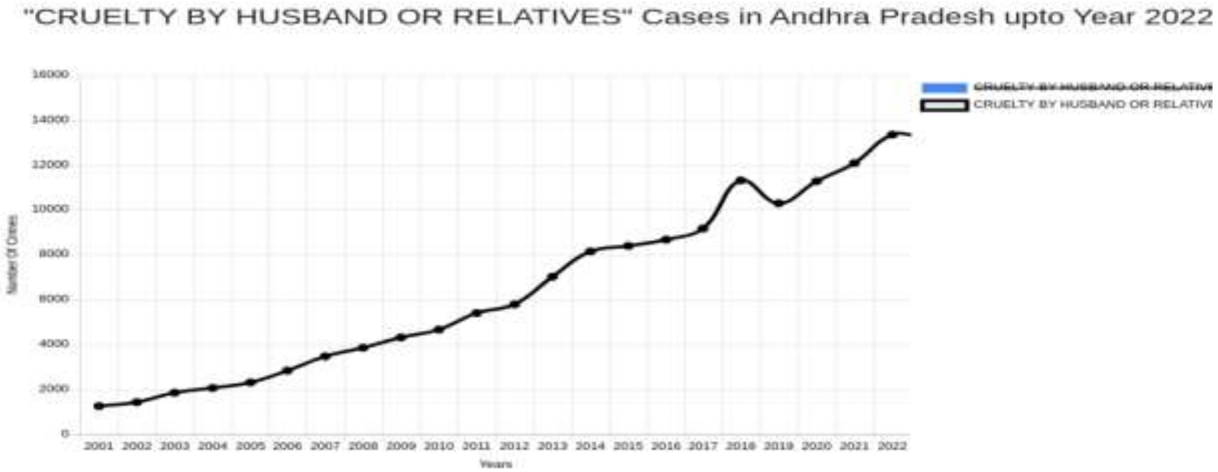


Figure 6 : Prediction of Crime (Cruelty by Husband or his Relatives against Women) in Andhra Pradesh up to year 2022

V. Conclusion

This application was successful in analyzing and predicting various crimes against women in India. Random Forest and Linear Regression predictive models was implemented on the data set and an accuracy of 76.923% and 83% of accuracy was obtained respectively. Linear Regression Algorithm showed better performance in terms of accuracy for prediction of crime rate against women in India. Police Department of a particular state can visualize the statistics and prediction result of various crime types and increase their Police force so that crime in that particular state is reduced and women can travel without any hesitation to that part of India. Gradually crimes rates will be decreased with help of police force and Government officials with implementing strict rules, hoping India to be better and safe place for women. As part of future work, analysis and prediction can be done district wise of each states in India and user statistics using the application can be implemented.

References

[1] Harpreet Kaur1 and Dr. Williamjeet Singh2, Systematic Review of Crime Data Mining, IJARCS International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017, ISSN No. 0976-5697.

[2] Vineet Pande, Viraj Samant, Sindhu Nair, Crime Detection using Data Mining, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 5 Issue 01, January-2016.

[3] Neeru Mago, Design and Implementation of Intelligent Crime Information and Analysis System (ICIAS) based on Crime Data Mining,IJARCSSE International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 1, January 2016 ISSN: 2277 128X.

[4] Shiju Sathyadevan1, Devan M.S2, Surya Gangadharan. S3, 1 2 Amrita Center for Cyber Security Amrita Vishwa Vidyapeetham, Amritapuri, Kerala, Amrita Vishwa Vidyapeetham Amritapuri, Kerala, India, 2014 First International Conference on Networks & Soft Computing, 978-1-4799-3486-7 114 2014 IEEE.

[5] Priyanka Das, Asit Kumar Das, "Crime analysis against women from online newspaper reports and an approach to apply it in dynamic environment", (ICBDAC) 2017 (IEEE Xplore :October 2017)

[6] Sunil Yadav, Meet Timbadia, AjitYadav, RohitVishwakarma, NikhileshYadav, "Crime pattern detection, analysis & prediction", (ICECA) 2017 (IEEE Xplore : December 2017)

[7] CharuNangia, D. P. Singh, Sabir Ali, "Built Environment and Crime Against Women", 2019 9th International Conference on Cloud Computing, Data Science & Engineering (IEEE Xplore : July 2019)

[8] P. Tamilarasi, R.Uma Rani, "Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning", (ICCMC) 2020 (IEEE Xplore : April 2020)

[9] Shraddha Ramdas Bandekar, C. Vijaylakshmi, "Design and Analysis of Machine Learning Algorithms for the reduction of crime rate in India", The 9th world Engineering Education Forum(Weef-2019)(Procedia Computer Science: January 2020)