Analysis of Crime Against Women Using Machine Learning Algorithms

Dr.K.Anuradha, Associate Professor, Department of MCA, Karpagam College of Engineering and Technology Coimbatore & India, anu_kce@yahoo.com

K.Janani, PG Student, Department of MCA, Karpagam College of Engineering and Technology Coimbatore & India, jananik319@gmail.com

Abstract: In worldwide Crime level is increasing every day. Offense can't be expected since it is either proficient or unplanned. Crime is a cost-effective trouble distressing life value and economic escalation. The particulars of how crime is performed revolutionize depending on the type of nation and society. Prior researches in crime prediction have originated that factors resembling education, deficiency, employment, and circumstances affect the crime rate. Some forms of violence are perpetrated or condoned by the state such as war rape; sexual violence and sexual slavery during conflict; forced sterilization; forced abortion; violence by the police and authoritative personnel; stoning and flogging. Many forms of Crime against women, such as trafficking in women and forced prostitution are often perpetrated by organized criminal networks. This paper analyses these crimes based upon the age groups and types of crime.

Keyword - Decision tree, Random forest regression, Linear Discriminant analysis, Linear regression, SVM (support vector machine)

I. Introduction

In worldwide Crime level is increasing every day. Offense can't be expected since it is either proficient or unplanned. Crime is a cost-effective trouble distressing life value and economic escalation. The particulars of how crime is performed revolutionize depending on the type of nation and society. Prior researches in crime prediction have originated that factors resembling education, deficiency, employment, and circumstances affect the crime rate. A few types of brutality are condoned by the state, for example, war rape; sexual violence and sexual slavery during struggle; constrained cleansing; constrained early termination; viciousness by the police and legitimate faculty; stoning and whipping. Many forms of Crime against women, for example, trafficking in women and forced prostitution are often perpetrated by organized criminal networks. This paper analyses these crimes based upon the age groups and types of crime.

II. Literature Survey

Puvi Prasad et al.,[1] used the data set from Kaggle which is time series dataset converted into supervised dataset in order to predict the number of crimes that may take place in future. For analyzing these crimes, Huber regression is used. It determines the loss score, that is how much the difference is going to between actual value and predicted value.

In [2], S.Lavanyaa and D.Akila compared the existing and proposed system with five different papers and compared the measurement of accuracy of crime predicted by different algorithms. The author came with a conclusion that clustering in WEKA utensils, Euclidean distance calculation gives exactness of crime in metropolitan and urban areas.

In paper[3], Dr. V. Anbarasu^{3} used three data mining algorithms namely Logistic regression, Decision tree, Naïve Bayes and Random forest to test the accuracy of the data set. In this paper visualization of data's is done to covert the textual data into visuals like graph, plot, pie chart etc for easy understanding of end user. In the conclusion the accuracy score was 0.807 in terms of 80% by logistic regression.

In paper[4],R.Devakunchari states that he objective of this paper is to identify the certain age groups and certain states of India which tend to commit each crime and also build a model which will help to predict the number of arrests that can be expected made for each crime, state wise and age-group wise.

In paper[5],OliSarkar proposed a crime analysis against women and as well as choose a method which is used to forecast the types of crime in different districts in West

DOI : 10.35291/2454-9150.2022.0298

Bengal for this analysis three major clustering technique is used namely K-means clustering, DBSCAN clustering and Agglomerative hierarchical clustering. After comparing with these four algorithms DBSCAN clustering gave the most appropriate results.

In paper [6],Dr.R.UmaRani analysed the crime rate against women using linear regression and Ordinal least square methods. Linear regression techniques are implemented to predict the crime rate. Ordinal least square also performed well in this process. Finally the accuracy value is predicted using R-square and the accuracy obtained is 98%.

In paper[7], Prof. Saravanan C^3 proposed work of this paper is to find the accuracy using two algorithms namely Random Forest and Linear regression. This paper uses flask architecture and web application. Flask is a micro web framework written in python which does not requires any tools or libraries. The output is visualized in the form of line graph and bar charts.

III. Analysis Of Project

Project Concept

This system is proposed for increasing transparency regarding the crime in the society and to help police officials, Government of India and the rest concerned authorities to take measures for minimizing the woman crime in our country.

A.Existing Sytsem

The existing work proposes an effective methodology for analysing crime against women in India. The work involves extracting crime reports from online newspaper articles and the documents consisting crime reports of various states and union territories of India are made to undergo several preprocessing techniques. Similarity has been measured among the words for selecting relevant features for crime trend analysis.Existing system is based on manual analysis. The data collected is inconsistent and have huge errors which is not properly pre-processed data. It provides a large ongoing training cost. The existing system does not provide any web application to predict and visualization of crime rate in particular states. There is no comparison between the different machine learning algorithms to increase the accuracy. This overall drawbacks are achieved in proposed system.

B.Proposed System

The project follows three different types of methods to process the data. Data pre-processing is the first step it is the most important techniques it is used to transform data into clear format. In this method it removes blank spaces, redundant data and rescale the data using standardization. The second step is propose K-fold cross validation. It is used to ensure that every observation from original dataset has the chance of appearing in training and testing phase. The third step is predicting the accuracy of crime rate using different machine learning algorithms such as decision tree, Linear regression, Linear discriminant algorithm, support vector machine and time series. This system is proposed to increase the transparency regarding the data in the society and to help the police and other official authorities to take measures for minimizing the rate of occurrence of crime against women in our country. The results are visualized in the form of graphs. For visualizations we have covered various demographic factors like population, literacy rates with respect to area.

Input Size limitations : It utilizes year wise dataset of women Crime types from 2001 to 2019, as the size of data will increase it would be reflected in the accuracy percentage of predictions for coming years.

Input Validation : The dataset given to the system checks for the numerical values in proper format and converted to suitable format techniques such as eliminating missing values, eliminating redundant data, data transformation, etc.

Input Dependency : The data is obtained from the official Indian Government Websites. Algorithms for the implementation of Visualization is also an important aspect on which the project is dependent. 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.

Major Inputs and Outputs : The System requires dataset of the Crime against women and Children, Indian Penal Code and Special and Local Laws Crimes.The proposed system would present the supplied data in some visuals such as Animate and barcharts as output with the prediction of the crime for coming four years with accuracy percentage.

IV. Methodology

This Project followed three different types of methods to process the data set. Data pre-processing is the first step. In this method removed the blank spaces, redundant data and rescaling the data using standardization. In the Second step, proposed k-fold cross validation. In the third step, predict crime rate and accuracy using different types of ML algorithms

- Liner regression
- Liner discriminant analysis
- Support vector machine
- Decision tree classifier

The system uses Linear regression Algorithm in major section i.e. Predicting the Crime Patterns with minimum

DOI : 10.35291/2454-9150.2022.0298

accuracy rates as factor. Linear Regression is the suitable Algorithm to predict pattern of this data. We had tried some other algorithms for testing purpose such as Random Forest Algorithm, and found Linear regression as the best fit for this proposed system with minimal accuracy in the results.

The following modules are there in the proposed work,

A. Data collection
B. Pre-processing and data cleaning
C. EDA(Export Data Analysis)
D. Model creation
E. Model Prediction

A.Data Collection

The data being used for this project are collected from Government websites of each state for a period from 2001 to 2021.The dataset contains 7 different types of crime occurred in different states of India. The dataset contains 10 columns and 10678 row entries. India's Crime Dataset from National Crime Records Bureau (NCRB)[8] and Open Government Data website[9]. The dataset includes following attributes,

- Rape
- Kidnapping & Abduction
- Dowry Death
- Assault on women with intent to outrage her modesty
- Insult to the modesty of women
- Cruelty by husband or relatives
- Total crime against women

B. Pre-processing and data cleaning

Data pre-processing refers to addition ,deletion or transformation of training data set. In this step I have transformed raw data that was collected into a form which is used in modelling. Before feeding the data into modelling process , following data pre-processing and data cleaning process is to be done,

- Checking null values.
- Removing redundant values.
- Splitting the data into training and testing sets.
- Making the data set in a same range neglecting the irrelevant data.

C. EDA(Exploratory Data Analysis)

EDA is a process which involves graphical representation and visualization to explore and analyze dataset by summarizing their main characteristics through visual methods. It also helps to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. The exploratory data analysis of our project includes,

Future women Crime Rate Prediction : This module of the architecture would predict the crime patterns for coming four years.

State-level Comparative Analysis : This module would present the comparative observations for all type of users, including the region wise comparison and crime type wise comparison.

Visualizations: This module is for representing comparative information in visuals for making the study easier and interesting and this will grab user's interests.

V. ALGORITHM ANALYSIS

The algorithms used in the proposed work includes,

- Decision tree.
- Random forest regression.
- Support vector machine.(SVM)
- Linear regression.
- Linear discriminant analysis.

DECISION TREE

In this project splitting the source set into subsets based on the attribute value other, This process is repeated on each derived subset in a recursive manner called recursive partitioning. In this model the data set is categorized and generalized. Data comes in record form:

(x,Y)=(x1, x2, x3,…., xk , Y)

The dependent variable , Y, is the target variable to generalize. The vector x is composed of the inputs variables x1, x2, x3 etc that are used for analysis.

RANDOM FOREST REGRESSION

In Random Forest Regression first we should import the libraries, the libraries used in this project is *NumPy*, *Matplotlib* and *pandas* libraries are imported. The next step is importing the crime data set in this I assigned independent variable (x) to the types of crime and the dependent variable(y) to the years. The next step is import the training set and import the Random forest regressor class and assign into variable regressor. Then I used .fit() to fit the x_train and y_train values to the regressor by reshaping.

SUPPORT VECTOR MACHINE

Support Vector Machine (Fig 1) is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

DOI : 10.35291/2454-9150.2022.0298

- The target variable has two values: Positive or Negative.
- These columns are the actual values of the target variable.
- These strings represent the predicted values of the target variable.

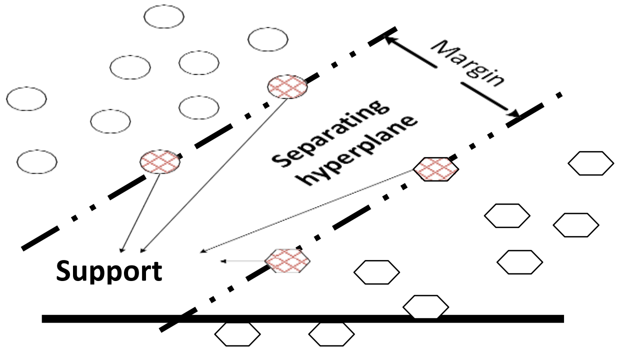


Fig 1. Support Vector Machine

LINEAR REGRESSION

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable. The linear regression model is simple and provides enough description of how the input affects the output. In this project I have predicted a variable Y (target variable) as a liner function of another variable X (input variable/features), given m training examples of the form (x1,y1), (x2,y2), …, (xn,yn), where xi ε X and yi ε Y. The form of hypothesis of linear regression can be expressed as

$h_\theta(x)= \theta0+ \theta1x1+ \theta2x2+ \theta nxn…..$

$h_\theta(x)= \theta^T x$

Where θ0, θ1, θ2, …, θn are regression parameters.

LINEAR DISCRIMINANT ANALYSIS

The linear Discriminant analysis estimates the probability that a new set of inputs belongs to every class. The output class is the one that has the highest probability. That is how the LDA makes its prediction. In our analysis, we are only concerned with the case of two groups, x and y where x=MN($\mu1, \epsilon1$) and y=MN($\mu2, \epsilon2$) are two multivariate normal populations.

Here I have distinguished between X and Y based upon the values of our random variables $X^T=[X_1,X_2,……,X_p]$, where each group's values for each variable differ to some degree. Each group has a population consisting of the values of its variables defined by a probability density function $f_1(x) or\ f_1(x)$. The above mentioned guidelines are developed via a training sample. Two regions are formed, R_1 and R_2. The training sample splitted into majority of the original sample into two known or correctly classified regions and then each region R1 and R2 is associated with the group, x and y respectively. The remaining sample, n minus the size of training sample, is called the test sample. This I have used to test the validity of the classification rule formed by the training sample.

SYSTEM DESIGN

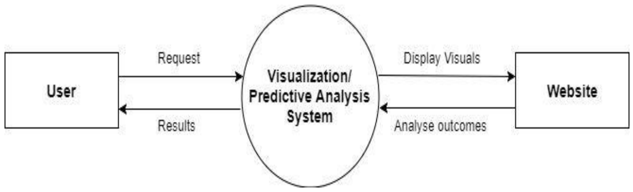


Fig 2. System Design

VI. RESULTS

Result of this research will be to present a system which will analyse, correlate and predict the crimes from huge data available. Results will be in the form of correlation between various Women Crime and location of crime i.e. state/city. Women Crime can also be correlated on the basis of age group, location of crime & type of Women Crime. Prediction of the crime will be presented using various techniques and Algorithm.

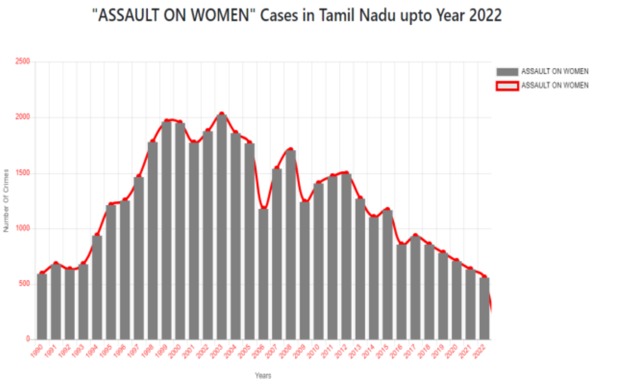


Fig.3. Assault of Women

The above Fig.3. presents the Assault of Women upto 2022 in Tamil Nadu . This is a visualization graph which represents the accuracy of crime

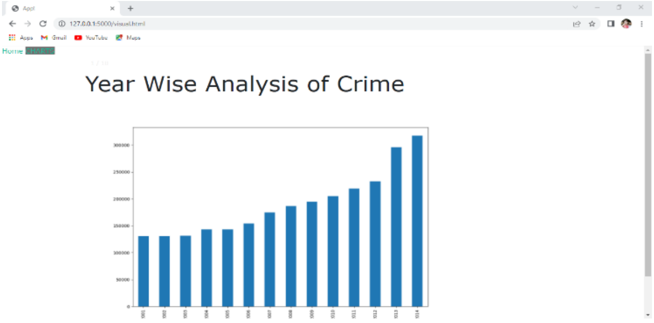
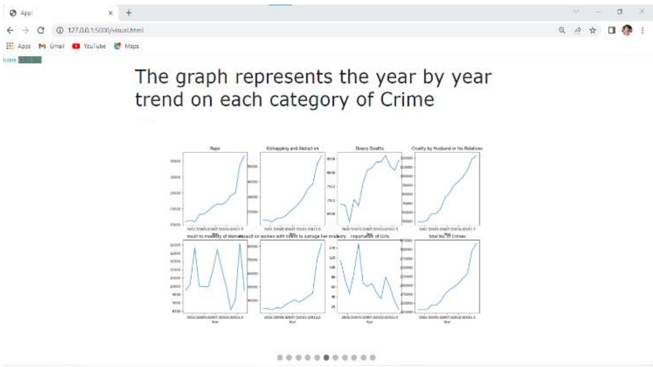


Fig.4. Crime Analysis Year wise

DOI : 10.35291/2454-9150.2022.0298

The above Fig.4. represents the year wise analysis of all listed crime in the form of bar chart.


App!
127.0.0.1:5000/visual.html
Apps
Gmail
YouTube
Maps
Rape
Kidnapping and Abduction
Dowry Deaths
Cruelty by Husband or his Relatives
Insult to modesty of Women
Assault on women with intent to outrage her modesty
Importation of Girls
Total No. of Crimes
Year

The graph represents the year by year trend on each category of Crime

Fig.5. Category of Crime

The above Fig.5.represents the year by year trend on each category of crime.

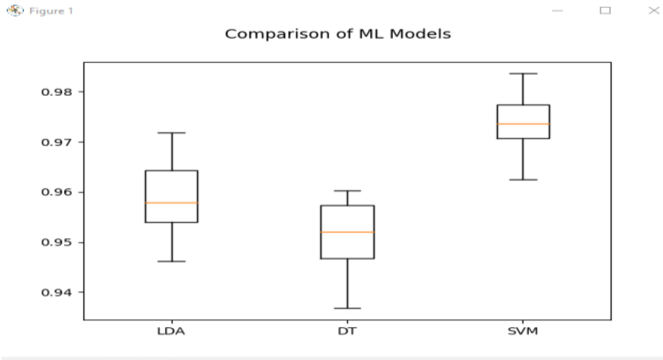


Fig.6. Comparison of Machine Learning Algorithms

The above Fig.6. shows the comparison of machine learning algorithms in which accuracy is increased in Support Vector Machine(SVM).

Another point noted is that the Women Crime rate is increasing and crime prevention has become an upheaval task. The legal force departments around the world are required to remain ahead in the eternal race between lawbreakers and law enforcers. So we are presenting the system which maintains, predict and visualize the crime records and the people who are involved whether the Government official or the Police. In the existing system the accuracy rate is higher in regression algorithm. But after applying some advanced techniques the accuracy rate is increased in SVM (Support Vector Machine). In [10], SVM are used to achieve higher accuracy.

VII. Conclusion

In this work, the analysis of crime against women using Machine Learning Algorithms was discussed. The efficiency of different Machine Learning algorithms is evaluated. The Crime type predictions were performed using the data and these predictions are displayed using simple visualization charts. It was observed that the efficiency of Support Vector Machine was superior when compared to other algorithms. This work will also help to analysis the crime rate against women and prevention measures can be used to reduce the crime. In future, more secure access to the website can be done by linking aadhar number.

References

[1] Puvi Prasad, Amirta Nair, Dr. S. Godfrey Winster Crime Against Women: Analysis and Prediction International Journal Engineering Research and Technology(IJERT) ,ISSN: 2278-0181,Vol. 10, Issue 05, May-2021.

[2] S. Lavanyaa, D. Akila Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019.

[3] Prasad D, Rachit Sharma , Dr.V.Anbarasu Analysis and Prediction of Crime against Woman Using Machine Learning Techniques Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021.

[4] R. Devakunchari, Bhowmick S, Bhutada S P, Shishodia Y Crimes Against Women in India using Regression International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 6S4, April 2019.

[5] P.Tamilarasi, Dr.R.Uma Rani Analysis of Crime against Women Using Simple Linear Regression and Ordinary Least Square Methods International Journal Of Science Technology and Management ISSN 2394-1537 Volume No.10, Issue No 04, April 2021.

[6] Vinay Narayan Bhat1, V Santhosh Kumar , Prof. Saravanan Analysis and Prediction of Crime against women in India using machine learning algorithms Journal of Emerging Technologies and Innovative Research (JETIR),ISSN-2349-5162, Volume 8, Issue 6, June 2021.

[7] P. Tamilarasi, Dr.R.Uma Rani Diagnosis of Crime Rate against Women using kfold Cross Validation through Machine Learning Algorithms, Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020) IEEE Xplore , ISBN:978-1-7281-4889-2.

[8] www.ncrb.gov.in

[9] www.crime-records-search.com.

[10] Harsh Namdev Bhor, & Dr. Mukesh Kalla. (2022). Analysis Of Performance Comparison Of Intrusion Detection System Between Svm, Naïve Bayes Model, Random Forest, K-Nearest Neighbor Algorithm. 17(05), 128–144. https://doi.org/10.5281/zenodo.6601187

[11] Anuradha K, Uma KP, Implementation of Fuzzy Cognitive Map and Support Vector Machine for Classification of Oral Cancers, EAI Endorsed Transactions on Energy Web and Information Technologies, 5(20), 1-5, 2018.

DOI : 10.35291/2454-9150.2022.0298