Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 5183 - 5188
Received 25 April 2021; Accepted 08 May 2021.

Analysis and Prediction of Crime against Woman Using Machine Learning Techniques

Prasad DS[1],Rachit Sharma[2],Dr. V. Anbarasu[3]

[1,2]UG Student, Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur.

[3]Associate Professor, Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur.
prasaddsprasadds@gmail.com[1],rachitsharma48@gmail.com[2], anbarasv2@srmist.edu.in[3]

ABSTRACT

Even after living in the country where most of the population pray women/girls, there has been a tremendous increase in the crimes involved against the same. Here in India Hypocrisy can be seen at peak where people pray women on one hand and treat them like nothing on another hand. So, after talking the help of big data which has been created over the years about the crimes involved against the women, we will be looking at the patterns and see if any correlation is there in between those factors. Other than just analysing we'll also be visualising the required parameters. Crime prediction will assist law enforcement in developing policies to deter crime against women and taking proactive measures to reduce crime.

KEYWORDS: Crimes, Machine Learning, Data Analysis, Visualization, Crime detection.

I. INTRODUCTION

Crimes in India is increasing at a very tremendous rate. Over the past years, the has seen plenty of growth and has moved on the way of success and India being one in every of them. India being one of the countries which has tried to balance between the advancement and their culture. We Indians pray woman one hand then attempt to supress their voice on the opposite hand. the rise within the number of crimes against woman within the past few decades indicates the statement.

In a country where the economy is blooming and is growing in each and every particular state and sector. In spite of all this, there has been a huge increase in the number of crimes against woman.

According to the report of WHO on crime against woman published on 29th Nov 2017, one out of each 3 women across the globe faces some crime a minimum of once in their life. So, if we glance at the statistics, around 35 you look after the girl bear this and are mostly done by their partners or knowns.

II. REVIEW OF RELATED STUDIES:

In [1], Author analyses and see the prime factors that have an effect on the amount of crimes in bound regions of the country. They have used graph-based clustering to see in which particular areas the crime is taking place the most and crime is taking place in specific. Along with this, authors have also implemented Community detection Algorithm. When we look out on a bigger perspective, we come to know that this particular model doesn't works well with incremental Dataset i.e. it isn't effective.

In [2], Author takes a lot of factors in consideration. In this, the authors did visualization and analysis both in which the focus was mainly on the Visualization part. They used matplotlib in most of the course and they did the analysis part using the Linear Regression which is not a recommendable choice as it is a primitive algorithm.

In [3], Authors has done the whole project using visualization and they have done the analysis part with the help of that. They have visualized each and every crime that has happened in the state or district of National Capital Territory (NCT) and explained the conclusion in a very comprehensive manner taking every possible into consideration.

In [4], Authors have done the diagnosis of crime rates against women using k-fold cross validation. They just analyzed the crime rate on a bigger scale rather than analyzing the features. They used feature scaling technique along with k fold validation to see how the crime rates are affecting or to check if it follows any pattern.

In [5], the authors have presented the paper that presents both the analysis & prediction of occurrence of crimes and the crime rate employing systematized method that segregates and inspects the occurring crime patterns. An optimized K means algorithm is implemented instead of traditional K means clustering algorithm to achieve reduced time complexity

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 5183 - 5188
Received 25 April 2021; Accepted 08 May 2021.

and improvised efficiency can be seen in the results. However, scalability is one of the limitations. Determination of optimum separation point appears to be biased while actually implementing the optimized K means algorithm.

In [6], the authors propose various approaches and techniques such as K-Means clustering algorithm, co-web clustering algorithm, density-based clustering algorithm, and other techniques based on filtered clustering to be implemented and utilized for interpreting the collected data and analyzing various ways to reduce the crime rates. They also employ Bayesian neural networks for crime prediction. But, density-based clustering is an unsupervised learning approach in machine learning. Hence, less precision & accuracy is expected in this algorithm due to unknown & unlabeled data.

In [7], the authors propose framework that aims to forecast the probability of a crime that may occur in a city by analyzing the crime dataset and visualizing the findings. The system determines the low areas of crime occurrence, medium areas of crime occurrence and high areas of crime occurrence based on the available dataset K-means processes. Unfortunately, K-means algorithm is incapable of handling any noisy data and unfit to identify non-convex shaped clusters.

In [8], the authors propose a technique to predict crimes that is based on hybrid approach of combining 2-Dimensional Hotspot analysis(which uses clustering) along with machine learning. The model designed and built for crime prediction is tested on the San Francisco dataset. However, given the spatiotemporal dataset, the model needs to be evaluated for performance & accuracy.

In [9], the authors propose employing an extension of K-means algorithm called the expectation-maximization algorithm. Here, the data is being categorized and partitioned based on its features and parameters. The data mining framework being proposed targets and deals with the occurring crime's geospatial plot. However, K means algorithm implemented often terminates at a local optimum which serves as limitation. Another limitation is that it is incapable of handling any noisy data and potential outliers.

In [10] the authors propose paper which outlines the concept of combining approaches of machine learning and data mining. Here, the combined approach can be employed fetching and assessing the complex criminal patterns and underlying criminal behaviours. They use the dataset available containing details of crime occurring in India that contains records of various crimes committed in all states is tested & they implement k-means clustering to find generic & recurring patterns. One of the few limitations is that K-means clustering algorithm can't handle categorical variables properly[13].

In [11] the authors propose implementation of K-means clustering analysis on collected dataset having sample of 190 countries. This using 7 dimensions of cybercrime such as fraud, spam, malware, piracy,GDP, and at last the internet use. And henceforth, classify nations into 4 categories based on cyber crime activity. Drawback of this work was the measure of malware data was sparse and not indicative enough.

In [12] the authors propose work in which K-means clustering algorithm is used for clustering various types of crimes by accessing the stored data. Also helps in identifying relevant and complex crime patterns, associated link prediction, underlying hidden links, and the statistical analysis approach towards the collected crime data. However, crime patterns are dynamic as they change over a period of time. Though can extract new crime factors by moving through the crime data, expected or even decent accuracy is not achieved[14]. The authors should have included more crime causing attributes instead of having fixed number of them unchanged over time.

III. **METHODOLOGY**:

The following modules are there in the proposed work:

1. Data Requirement Selection
2. Data Collection
3. Data Preprocessing
4. Data Analysis
   4.1 Prediction model
5. Data Interpretation
   5.1 Data Visualization

Algorithms used: Logistic Regression, Decision Tree, Naïve Bayes and Random Forest.

Description:

1) Data Requirement Selection:

The dataset required for this major project needs to be of a time period of around 10 years. The dataset should specify the type of crime committed against a particular victim. The dataset should geographically cover an entire region or a country

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 5183 - 5188
Received 25 April 2021; Accepted 08 May 2021.

(say India). Preferably state-wise or pin code-wise data. The dataset should contain row entries around 8000 to 10000 to have accurate and efficient clustering. We are using the dataset which has got entries from all over the country and has records of all the crimes that has happened.

2) Data Collection:

The data being used in this project is collected from https://data.gov.in. The dataset contains data on numerous crimes committed against women in different states and cities between the year 2001 to 2012. The data includes over 8 different types of crimes that has happened across the country. The dataset has 10 columns and 9018 row entries.

3) Data Preprocessing:

The data pre-processing is a technique that does all the transforming work to convert the vague data input to some understandable data. The dataset we collected is of real-world data which may often be inconsistent or incomplete. May also lack certain trends or behaviors. Hence, we clean the data and make it ready for our algorithms so that it becomes easy to make predictions. Also eliminate occurrence of any errors.

4) Data Analysis:

The data analysis step is the process of analyzing data to extract insights that support decision-making. Here, we employ certain methods and techniques in order to gain useful and significant insights from our dataset. Here we have worked on some different prediction algos to get some predicted values out of the same

4.1) Prediction Model

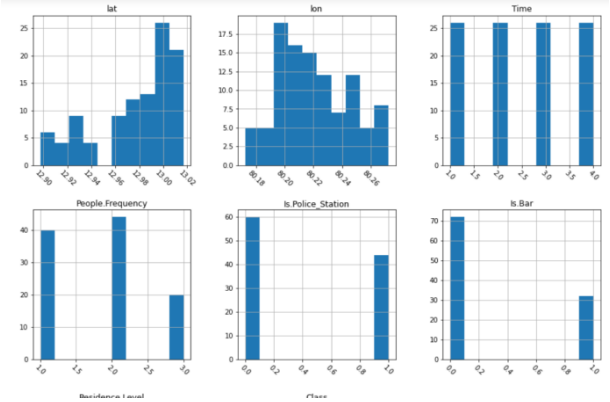
Prediction Model building step involves working on different statistical method, training those models with the help of test and then use the same to predict values out of it to get some the useful information which can help the end user in some manner, let it be predicting crimes or detect a safe zone.

5) Data Interpretation:

Data interpretation is the implementation of process of by which some useful insights or some pattern can be detected by the end user and help them to come at a conclusion. It comprises of the following things: Working on the output of data analysis and then seeing the things which we can infer from the same, and finally, thinking of the ways where we can apply it i.e. concluding. Here, along using DATA VISUALIZATION with DATA INTERPRETATION, we'll get to know about the different crimes happening in different states and cities.

5.1) Data Visualization:

Data visualization is the technique of converting the information which was provided to us in the form of textual data into some great visuals, like Bar Plot, Scatter Plot, Pie Chart to make it easier for the end user to understand and to work upon. Here, we plot graphs for various crimes committed against women such that yearly distribution and state-wise distribution of crime is visually represented.
This helps us in better understanding and accurate interpretations of useful insights from the raw dataset with the help of data analysis and data visualization.



Now talking about the Algorithms, let's go through them one by one.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 5183 - 5188
Received 25 April 2021; Accepted 08 May 2021.

1. **Logistic Regression**: Logistic Regression classifier is an statistical method which is used for predictive analysis. It is somewhat similar to linear regression as even this method predicts the value of an dependent variable based upon the independent one. The difference being that the predicted value in case of Logistics Regression is between 0 and 1. The value is given by the following formula:

$$S(x) = \frac{1}{1 + e^{-x}}$$

2. **Decision Tree**: Just like the regression, Decision tree is also a supervised machine learning Algorithm used to predict value out by learning decision rules from features. This machine learning algorithm is a classification algorithm which is extremely intuitive, easy to interpret and understand. they're recursively constructed ranging from the foundation node till the kid nodes. The attributes of the info and also the model hyperparameters together assists within the node construction.

3. **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 algo 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 build the model and rest is used to predict the accuracy.

4. **Naïve Bayes Classifier**: Naïve Bayes classifier works on the probability of the events which has happened in the past, using that along with Bayes theorem, it helps us to find the probability of the event that is going to happen in the future. This algorithm works completely fine with the events which are linearly separable but it can also works well with the ones which are not linearly separable.

IV. IMPLEMENTATION AND RESULT:

- First, we are going to split our data. Y contains column 'Class' and its corresponding row entries. X contains all other columns and their corresponding row entries.
  Then, we are going to divide train size and test size in the ratio of 3:1. That is test size is going to be 0.25(total dataset size)

PREDICTIVE MODELLING:

Now, we are going to apply various machine learning approaches.

We are going to implement certain machine learning prediction models relevant and compatible to our dataset.
1) Random Forest
2) Naive Bayes
3) Decision Tree
4) Logistic Regression

I. We implemented Random Forest Classifier and trained it using our train data. Train data was 75% of the total dataset. Then , we tested the model using the test data which consists of 25% of the total dataset.
We, then tested accuracy score of the predictive model and we got decent accuracy score of **0.769230 7692307693.** This is 76.923% in terms of accuracy percentage.

II. We train our model using Naive Bayes algorithm using MultinomialNB() method and trained it using our train data. Train data was 75% of the total dataset. Then , we tested the model using the test data which consists of 25% of the total dataset. We, then tested accuracy score of the predictive model and we got good accuracy score of **0.8076923076923077**. This is 80.769% in terms of accuracy percentage.

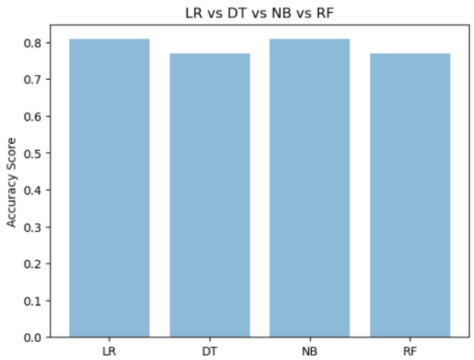
III. we implement our model employing Decision Tree algorithm. Train data was 75% of the total dataset. Then , we tested the model using the test data which consists of 25% of the total dataset.
We, then tested accuracy score of the predictive model and we got decent accuracy score of **0.7692307692307693**. This is 76.923% in terms of accuracy percentage.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 5183 - 5188
Received 25 April 2021; Accepted 08 May 2021.

IV. We applied Logistic Regression and trained it using our train data. Train data was 75% of the total dataset. Then , we tested the model using the test data which consists of 25% of the total dataset. We, then tested accuracy score of the predictive model and we got good accuracy score of **0.8076923076923077**. This is 80.769% in terms of accuracy percentage.

V. CONCLUSION:

Finally, we do side-by-side comparison of these 4 implemented predictive models comparing them in terms of accuracy. On a scale of 0 to 1, we get 2 algorithms with accuracy score of 0.7692 and other 2 algorithms with accuracy score 0.80769.



REFERENCES:

[1] 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)

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

[3] 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)

[4] 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)

[5] S.G Krishnendu, P.P Lakshmi, L Nitha, "Crime Analysis and Prediction using Optimized K-Means Algorithm", (ICCMC) 2020 (IEEE Xplore : April 2020)

[6] ShraddhaRamdasBandekar, C.Vijayalakshmi, "Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India", The 9th World Engineering Education Forum (WEEF-2019) (Procedia Computer Science : January 2020)

[7] Wasim A. Ali, HusamAlalloush, Manasa K.N, "CRIME ANALYSIS AND PREDICTION USING K-MEANS CLUSTERING TECHNIQUE", EPRA International Journal of Economic and Business Review (August 2020) (EPRA IJRD Volume: 5 | Issue: 7 | July 2020)

[8] Gaurav Hajela, Dr. Meenu Chawla, Dr. Akhtar Rasool, "A Clustering Based Hotspot Identification Approach For Crime Prediction", (ICCIDS 2019) (Procedia Computer Science : 2020)

[9] Vineet Jain, Yogesh Sharma, Ayush Bhatia, Vaibhav Arora, "Crime Prediction using K-means Algorithm", GRD Journals - Global Research and Development Journal for Engineering (Volume 2 | Issue 5 | April 2017)

[10] GouriJha, Laxmi Ahuja, Ajay Rana, "Criminal Behaviour Analysis and Segmentation using K-Means Clustering", 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (June 2020)

[11] Alex Kigerl, "Cyber Crime Nation Typologies: K-Means Clustering of Countries Based on Cyber Crime Rates", International Journal of Cyber Criminology (Vol 10 Issue 2 July – December 2016)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 5183 - 5188
Received 25 April 2021; Accepted 08 May 2021.

[12] SnehalDhaktode, MiralDoshi, NeerajVernekar, Ditixa Vyas, "Crime Rate Prediction Using K-Means", IOSR Journal of Engineering (IOSR JEN) (2019).

[13] M .Baskar, J. Ramkumar, V.Venkateswara Reddy, G.Naveen Reddy, “Cricket Match Outcome Prediction using Machine Learning Techniques”, International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1863-1871, ISSN: 2005-4238, April 2020.

[14] M .Baskar, J. Ramkumar,Ritik Rathore, Raghav Kabra, “A Deep Learning Based Approach for Automatic Detection of Bike Riders with No Helmet and Number Plate Recognition”, International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1844-1854, ISSN: 2005-4238, April 2020.