Women Safety Prediction using Logistic Regression Model

V. Sushma Swaraj, L. Bhavya, G. Pooja, R. DevaRevathi

***Abstract:** Safety of Women has become a major issue in India. Especially at night women think a lot before coming out of their homes. We daily come up with news of how women are subjected to a lot of violence and harassment or get molested in public areas. This paper focuses on the issue of helping Women that they don't ever never feel alone in the middle of any situations. The project idea is to predict whether the given place at any time is safe for a women to go or not. There are many pre-existing applications that are useful at the time of crisis situations. At some situations when a women is in trouble, she is not able to use those applications. And there are also so many rehabilitation centres which are used after the situation has happened. But our proposed model will help women to take precautions so that they never ever get that situation. For this idea we used Machine Learning. Machine learning is used to train the data and make quality predictions by recognizing the patterns in data. We applied different algorithms like Naïve Bayes, K-Nearest Neighbours, Logistic Regression models. Logistic regression is the best fit among other machine learning algorithms and it is more effective than others. In this paper, we used Logistic regression algorithm of Sklearn machine learning library to classify the dataset. Information about some set of areas in Tamilnadu are collected and was used in our project. When a women alone want to go out for any personal work or any financial work without knowing any safety details about the place she wants to go our application helps more better.*

***Keywords:** K-Nearest Neighbours, Logistic Regression, Machine Learning, Naïve Bayes, Sklearn, Women safety*

I. INTRODUCTION

We are living in 21st century. Day by day science and technology is advancing. Today we are developed scientifically and technologically in many areas such as having medicine for almost all the disease, super fast transportation, etc. But across the world we are still far behind when we are talking about "Women's Safety". There is no safety for women as per seeing the last few crimes in India such as acid attacks, rape cases, harassments etc.

According to researchers in every 15 to 20 minutes, a woman gets tortured, whether it is a crowded area or an uncrowded area. Time has changed but attitude towards women have never been changed yet. Though there is a lot of development in various areas, problems to women also rising constantly.

World health organization has stated in a report that a huge proportion of women been suffered from sexual violence. Crimes against women in the form of murders, rapes, dowry threats, have been on the up in the past decade[4].

Women safety has become a serious issue in modern days. Especially it is more common in populated countries. Today some of the women has been coming out and give competition to men in all areas including employment, education, politics. But however the problems to women also increasing significantly. Many of the women harassed news scared the women. Especially the harassment against women during their journey to office or from higher officials has a terrible impact on women. There are also situations where women have been restricted to home from their parents in fear of unsafe society.

Although precautions and rules have been introduced by the government, a new problem against women safety also enhancing. There are many technical solutions for women safety but they are in vain as the days are passing. There are so many applications which helps user can easily trigger the calling function by shaking the phone or explicitly by simply pressing panic button on the screen[3]. As there is a large development in technology, we developed a solution by using machine learning for women safety. Our idea helps women to take decisions about whether to go to a particular place alone or not. For this decision, our project gives the level of safety to that particular place using machine learning.

Machine learning is one of the emerging technology that has various applications in business, education, science and day-to-day life. In simple words, machine learning means training a computer with previous data and recognize patterns and give predictions for given input with accuracy. Instead of programming the computer every time step-by-step by a programmer, this approach gives instructions to the computer that allows it to learn from data. This means we can perform new, complicated tasks that could not be done manually.

In this paper , we first issued the cases of women safety. Next we provide the existing solutions on women safety using machine learning. Then we proposed our idea and its outcomes along with methodology. Then we describe how our project significance towards women safety by means of visualizing. Then we provide the result analysis of our idea with conclusion at last.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author

V. Sushma Swaraj*, Student, Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: Sushma.swaraj15@gmail.com

L. Bhavya, Asst. Professor (Adhoc), Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: bhavyalevadala@gmail.com

G. Pooja , Student, Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: gajaramnagarajareddy@gmail.com

R. DevaRevathi , Student, Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: devarevathi033@gmail.com

II. RELATED WORK

Machine learning consists of three different types of algorithms. They are Supervised, Unsupervised and Reinforcement. From a given set of predictors or independent variables, supervised algorithm consists of dependent variable are predicted. In Supervised learning there are many classification and regression algorithms like Linear regression, Support Vector Machine, Naïve Bayes , KNN, Logistic Regression, etc.
large datasets which we can easily build with complicated parameters. Direct acyclic graphs with one parent and several children together makes this model.

B. Logistic Regression

Logistic Regression is one of the powerful numerical classification algorithms. This statistical method is used for producing a binary outcome by analyzing the dataset. It is mainly used for binary classification problem and easy to implement. This is used to predict the binary variable by recognizing the relation between one dependent binary output and independent features. There are three types of logistic regression. Depending upon the output variables, logistic regression is classified into binomial, multinomial and ordinal.

C. K-Nearest Neighbour

KNN is simple classification algorithm. This algorithm assumes the similarity between the new data and available dataset and put the new data into the category that is most similar to the available categories. K-NN does not make any assumption on underlying data which means it is a non-parametric algorithm. It is based upon the similarity among variables. By using K-NN algorithm it can be easily classified in to a well suite category when a new data appears.

III. PROBLEM STATEMENT AND SOLUTION

Our problem statement is to make women feel safe after stepping out of their homes. In the past years, rehabilitation centres, many NGO'S and helpline numbers have been made operational. But they are not the preventions that we need, they are just all cures to the harassment that has already happened.
Women could walk around freely, without the fear of being attacked at anytime, anywhere by our predictions of which place is safest at a particular time. To solve this problem we have chosen machine learning technology. In machine learning we have different types of algorithms that are suitable for our problem. We have applied some supervised algorithms like Logistic Regression, K-NN and Naïve Bayes.
Our data set consists of attributes that include time to travel, source, destination, time, people frequency, police station availability, presence of bars, tier, residence level and class which is the output variable.
For example, if women want to travel from a particular source to destination which takes more time like 6 to 7 hours. Some places are safer at day times and unsafe at night times. Likewise if she started her journey alone in the evening and reaches by night, it means that place will be unsafe for her. In those cases, we predict the place safe or not according to place and time earlier which can avoid crisis situations and can take precautions.

A. Naïve Bayesian

Naïve Bayes is one of the powerful classification algorithms. Naïve bayes uses Bayesian theorem that performs the classification depends upon the probabilities arrived with the belief that all the variables are independent to each other. This classification algorithm takes in to account of the Mean and Variance of variables. Particularly it is useful for

The main attributes that effect the targeted variable are people frequency, bar presence, police station availability. TheSource and destination attributes have nearly 27 areas covered in tamilnadu. The bar availability has yes or no values. The police station attribute has yes or no values. The tier says whether the place is inner, outer or middle of the city. The time attribute has the values morning, afternoon, evening, night. The residence level attribute has values low, medium and high. The people frequency attribute has values low, medium, high. The targeted attribute class has values Safe and Unsafe. All these values which are in categorical nature are converted into numerical values for easy visualization.

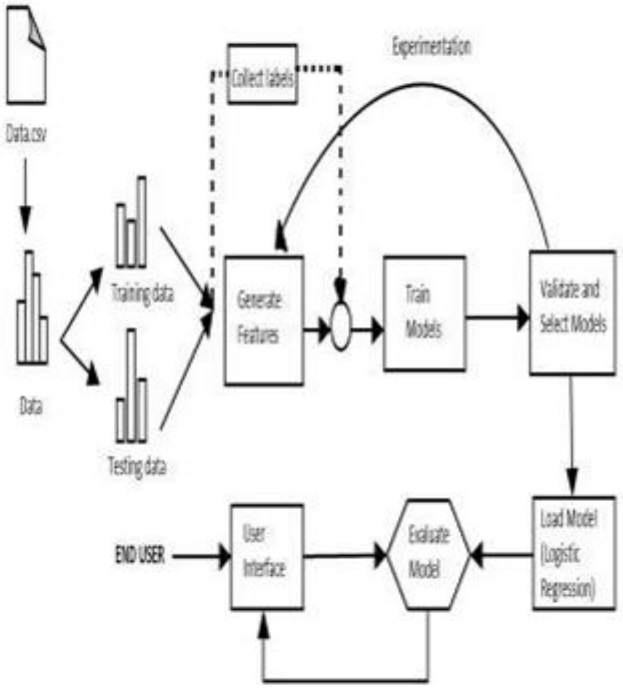

Fig 1: Architecture of the women safety prediction model

The architecture is as follows. Here first we divide the 75 percent of data into training and remaining to testing. Training data is used by model to recognize pattern by using existing algorithms. Then we used testing data to evaluate the model by comparing the predicted outcomes with the actual outcomes. Before training and testing we have to preprocess the data. In preprocessing we remove the missing data i.e., data in which some attributes are missing. Then we convert the data to numerical data because some of our machine learning algorithms works better on numerical data.

IV. METHODOLOGY

A machine learning project can be splitted up into three important steps. Data collection, Data modeling and deployment. All influence one another. Data is collected

Libraries:

The various libraries essential for the project are numpy, scikit-learn, pandas, matplotlib.

Numpy: Numpy package has a more significance in data science for multi-dimensional array applications. Numpy is a also one of the popular python library. It uses Fourier transforms, linear algebra and random number capabilities I its implementation.

Pandas: For data analysis, Pandas is also mostly used python library. In our project we used pandas for reading csv file. Data extraction and preparation was developed scientifically in case of pandas. For data analysis, it presents wide variety tools and high level data structures. It represents many inbuilt methods such as combining, filtering data and groping.

Matplotlib: For data visualization, Matplotlib is often used in python library. For visualizing the patterns in the data by a programmer, it particularly comes in handy. For creating 2D graphs and plots we use 2D plotting library. Using module name pyplot makes it simple for a programmer as it provides features to control line styles, formatting axes, font properties, etc. In our project we used matplotlib for easy visualization by means of plotting histogram.

Scikit-learn: Scikit-learn is the heart of machine learning that consist of powerful algorithms. It builts with top two basic python libraries, viz., scipy and numpy. It can also be used for data mining and data analysis. When anyone starts with machine learning it will be a great tool.

Data Visualization:

We will read the dataset by using the pandas library. And then summarization of data is done by finding the dimensions of the data, their statistical summary of all attributes and class distribution of the data. Visualizing the dataset involves boxplots, histograms and countplots.

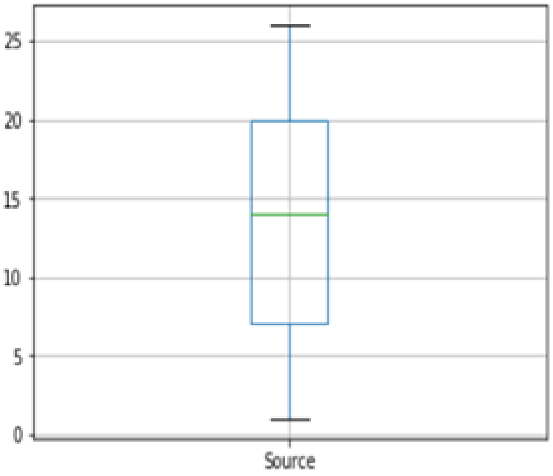

Fig 2: Boxplot of the attribute ‘Source’

Boxplots means how well distributed the data in a dataset can be measured. It divides the dataset into three quarters. This graph in the dataset represents the minimum, median, maximum, first quartile and third quartile. In the box the line that divides into two parts shows the median of the

and converted it into a structured data in a spreadsheet. Our dataset consists of 2580 rows and 10 columns. We converted our dataset into numerical values for easy understanding purpose and analyzation.

data. In a box the end represents the lower and upper quartiles. The extreme line shows the highest and lowest value excluding
outliers. In fig 2, the boxplot doesn’t have any outlier s.

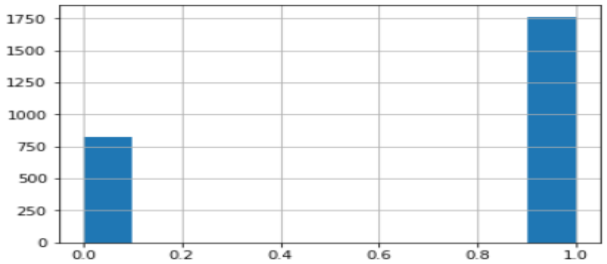

Fig 3: Histogram of the attribute ”Class”

A histogram represents distribution of numerical data with an accurate graphical representation. It takes only one numerical variable as input. The variables are separated into several bins and the highest of the bar is shown by the number of observation per bin.

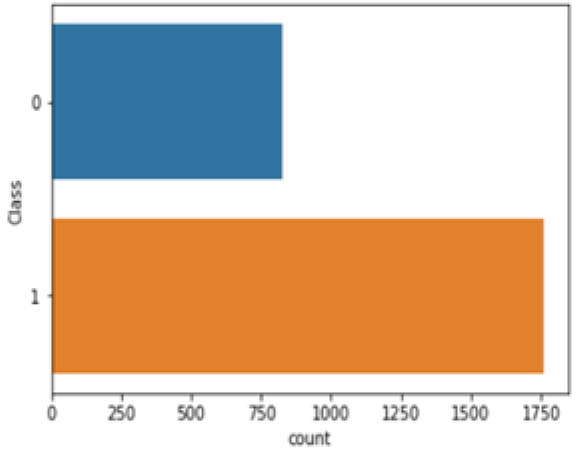

Fig 4: Countplot of attribute”Class”

A countplot is kind of like a histogram or a bar graph for some categorical area. Based on a certain type of category, it simply shows the number of occurrences of an item.

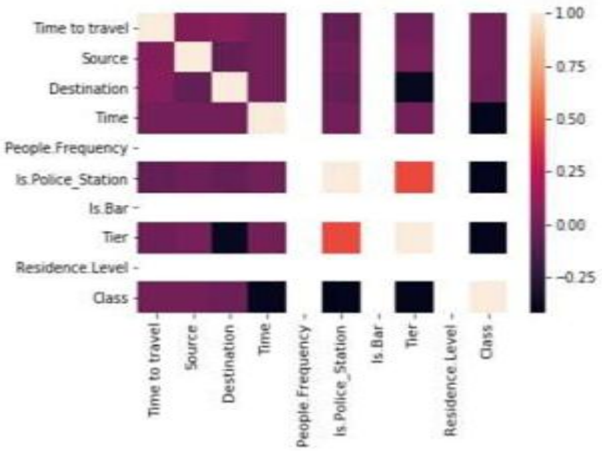

Fig 5: Heatmap of the dataset

Heatmap shows graphical representation of data. In a matrix which contains colors are represented as individual values. It is really useful for displaying a general form of numerical data. But not used to extract specific data point.

Training & Evaluating Models:

To evaluate the models firstly the dataset should be trained. For this we divided the dataset into training annd testing datasets with different percentages.We divided dataset with 60, 70, 80 percentages as training datasets and 40, 30, 20 percentages as testing datasets.

We approached different classification methods for comparison which helps to find out the efficient method.We compared accuracies of different algorithms like Naïve Bayes, Logistic Regression, K-Nearest Neighours.We fit the trained data in the classification models and predict the accuracies.In a classification problem, summary results are predicted by a confusion matrix. The principal diagonal values in confusion matrix justify the model for predicting happened values. And the remaining describes the model for negative decisions. It may misleads while doing predictions. It gives not only errors but also type of errors being made by a classifier.

For finding the precision we divided the correctly predicted cases to the total number of predicted cases just as given below

Precision=TP/TP+FP

The precision value we got from the experimetal results is 88 percetage.

For finding the recall i.e., the proportion of incorrectly identified we divide the non-principal values in confusion matrix to the to the sum of values in confusion matrix.

Recall=TP/TP+FN

The recall value we got from the experimetal results is 91 percetage.

ROC curve is a used as performance measurement in classification problem at different thresholds settings . It is a probability curve. It defines about distinguishing between classes, how much the model is capable. As AUC being higher, it shows the better model at predicting true as true and false as false

Taking FPR on X-axis and TPR on Y-axis ,we drawn a ROC Curve as shown below. Here then precision is calculated by getting the correctly identified observations and total observations using (TP / (TP+FN)).. Also recall is calculated by getting incorrectly predicted observations using (FP / (TN+FP)). By this we come to know about the truthfulness and falsity of our trained model.

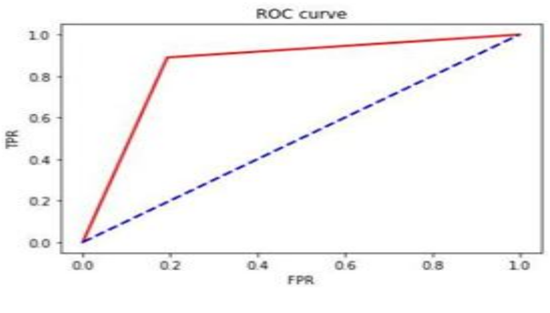

Fig 6: ROC Curve

In the figure 6, we can see the ROC curve for our trained regression Model, The red line shows the ROC curve and blue dotted lines shows the AUC (Area under curve).

V. RESULT ANALYSIS

The machine learning model is trained by varying the training and testing split percentage. From our observations, the trained model gives more accuracy when we give 70 percent of data to training purpose and remaining data for testing.

Table-I: Comparisons of Algorithms

Metrics	Naive Bayes	Logistic Regression	KNN
Accuracy Testing-40% Training-60%	73%	85%	83%
Accuracy Testing-20% Training-80%	71%	86%	74%
Accuracy Testing-30% Training-70%	72%	86%	84%

As compared logistic regression with naïve bayes and knn, KNN is inefficient since the entire training data is processed for every prediction. Naïve bayes calculates the prior probability . An error rate is shown by classification decision. The input data is very sensitive to form by using sample attribute independence, when we use sample attributes are dependent then the effect is not good.

From our research and observations, Logistic Regression is a best fit model for project and it works well with less training data . Naïve bayes model performance is not as good as Logistic Regression because of overfitting. KNN is a distance based technique gives medium results but less compared to naïve bayes.

VI. CONCLUSION

Thus, the purpose of this project i.e., to develop an authoritative model and a predictive model by using logistic regression is successfully achieved, without the fear of being attacked at anytime, anywhere, women could roam around safely. It seems women not to face any helpless situations. Sometimes women wont know about a particular places. With this project a women can know whether a place is safe or unsafe to go alone at a particular time. Accurately we predicted 85 percent about safety of women. Further we can add live location tracking system and we can also add a system in which it can send a message automatically to some saved contacts about reaching her destination.Using logistic regression model we performed processing on classification of data.

This research can be further used for implementations on other domains of predictions like location tracking, sending messages automatically etc. By changing the independent variables the results will be differ in nature.

REFERENCE

1. Avantika Bhate, Parveen Sultana HSmart, Wrist Band for Women Security using Logistic Regression Technique.
2. Sunpreet Kaur, Sonalika Jindal, A Survey on Machine Learning Algorithms.
3. Dhruv Chand, Shivani parikh, Sunil Nayak, Amita ajith Kamath, Karthik bhat, Yuvraj singh, A Mobile Application for Women's Safety.
4. R. Devakunchari, Bhowmick S, Bhutada S P, Shishodia Y, Analysis of Crimes Against Women in India using Regression.

a. Deepak Kumar, Shivani Agarwal, Analysis of Women Safety in Indian Cities using Machine Learning on Tweets, 2019 Amity International Conference on Artificial Intelligence(AICAI).

5. Dantu Sai Prashanth, Gowtam Patel, Dr. B. Bharathi, Research and Development of a mobile based Women Safety Application with real-time database and data stream network[ICCPCT].
6. National Crime Records Bureau, Crime against Women, 2014.

a. M. Maria Dominic, R. Shinoj Robert, A Literature Review on Machine Learning.

7. http://www.towardsdatascience.com
8. S. Celine, M. Maria Dominic, M. Savitha Devi, Logistic Regression for Employability Prediction, International Journal of Innovative Technology and Exploring engineering(IJITEE),Volume-9 Issue-3, January 2020.
9. B. Sathyasri, U.Jaishree Vidhya, G.V.K. Jothi Sree, T. Pratheeba, K. Ragapriya, Design and Implementation of Women Safety System Based on Iot Technology.

AUTHOR PROFILE

V. Sushma Swaraj, Student, Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: Sushma.swaraj15@gmail.com

L. Bhavya, Asst. Professor (Adhoc), Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: bhavyalevadala@gmail.com

G. Pooja , Student, Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: gajaramnagarajareddy@gmail.com

R. DevaRevathi , Student, Dept of CSE, JNTUACEP, Pulivendula – 516390, YSR (Dist), A.P, Email: devarevathi033@gmail.com