IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

Data Mining Approach to Analyze the Road Accidents in India

N. Usha Rani[1*], T. Vanaja[2]

[1,2]Dept. of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Sri Venkateswara University, Tirupati, India.

[*]*Corresponding Author: usha552@yahoo.com, Tel.: 9493564899*

Available online at: www.ijcseonline.org

Accepted: 18/Oct/2018, Published: 31/Oct/2018

***Abstract*—** The accident is an unplanned incident which leads to injury to people, damage to a plant, machinery or some other loss. The goal of this paper is an analysis of road accidents at country level and statewide of India. The analysis shows that accidental fatalities and injuries changes according to age, gender, month and time. Analysis of road accidents plays an important role in the transportation system. Road traffic injuries and fatalities are common in nature it would be impractical to predict one-to-one relationship among the safety measures in road accidents, injuries, and fatalities. Road safety is an important concern for both national and international level. Data mining tools and techniques are used to predict accident-prone locations. For every four minutes, one death is occurred due to road mishap in India. The crucial thing is an analysis of road accident data is its heterogeneousness. The relationship between road surface conditions, road type, severity, light conditions, etc. are investigated.

***Keywords*—** Road accidents, fatality, classification, clustering

I. INTRODUCTION

Road transportation is a dominant transport in India, in terms of traffic share and contribution to the national economy. Road accidents can be defined as "accident occurred on a way or street, results show that one or more people are killed or injured, one, one or more vehicles are involved [1].Thus crashes among vehicles; among vehicles and pedestrians; among vehicle and animals; among vehicle and geographical obstacles." The lack of road network leads to road accidents and road crash fatalities [2]. A lot of vehicles travelling on the roads every day and accidents may be happening at any time anywhere some accidents may lead to injuries some lead to deaths. Nowadays, road accident injuries are one of the most important causes of death, disabilities and hospitalization in India. Data mining apply different techniques and algorithms to determine the relationships among the attributes in the data set. The major problem is heterogeneity, thus heterogeneity must be considered during the analysis of the data otherwise, some relationship between the data may remain hidden [3].

The density of traffic accidents in India is the highest in the world. In this paper, consider states and the union territories of India and to know opposed causes and educational background of the driver in order to make possible road safety in the country [4]. Delhi and Chennai register that many numbers of accidents than other states in India. In Indian roads, the major accident-prone time is at the time of afternoon and evening. World Health Organization (WHO) spot that most of the traffic collisions occurred due to driver over speed, drunk and driving, drowsiness, and not wear helmets and seat belts.

Rest of the paper is organized as follows, Section II contains the related work which explains various data mining techniques and methods are used to analyzing road accidents. Section III contains datasets collection and flow of pre-processing is explains. Section IV details about methodology in which it has architecture and algorithm. Section V explains the results and discussion of the results. Section VI details about the conclusion and future scope.

II. RELATED WORK

Analyzing the road accidents in India by using data mining techniques fatality rate was reduced [5]. By using various road safety conditions fatality rate was reduced in both national and international level. In order to predict models for crashes on roads, a statistical approach was failed. The Poisson model and negative binomial model (regression model) are used to shaping data sets in order to give a concrete solution. Electing the independent variables for prediction model and crash analysis plays an important role. Fatality rate regarding accidents might be reduced by introducing emergency medical services (EMS). Classification and clustering techniques are used in order to predict the models for road safety.

In road networks, limited accident-prone areas are plotted by using statistical methods. Analysis of road accidents shows that deaths and injuries vary from age, gender, weather conditions, road surface conditions etc., although metropolitan cities face fewer accidents in India.

By using data mining techniques above-mentioned defects could be overcome. The main aim is to find the effective factor inflicting the various states of India based on the dataset and to accomplish all available factors that are directly or indirectly had a part in preparing the dataset. Clustering analysis could be used for finding the pedestrian [6] [7] deaths and decision trees could be used for predict the reasons for road accidents. The unstructured data is structured and prepared to fit for data mining techniques.

III. DATA SET COLLECTION AND PRE-PROCESSING

The dataset provides detailed information of the road accidents in India. The data is taken from the year 2013 to 2017 in many states and union territories of India for analyzing and evaluation of risks through the accidents, and various data mining techniques are applied on that data to generate concluded information. The data having states and union territories road accidents information, and it containing attributes like state, driver condition (alcohol consumption, over speed, dizziness), age, gender, road surface conditions, type of vehicle etc.

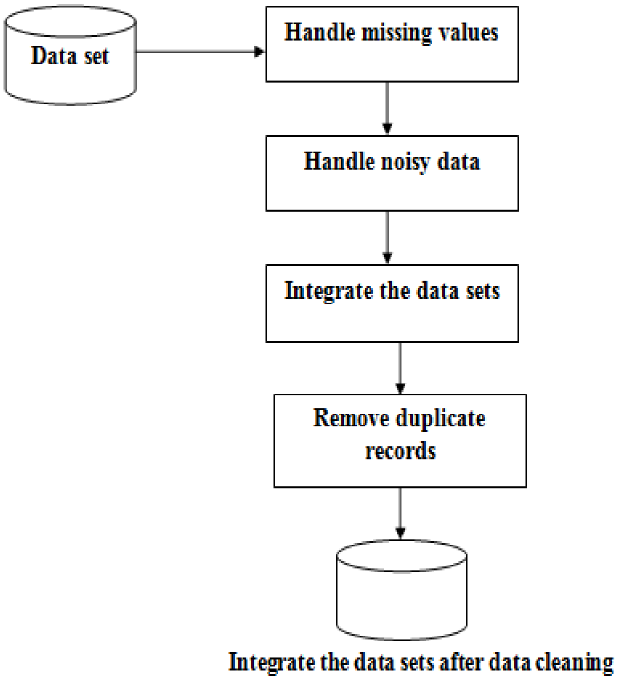

Figure 1: Flow of Data Pre-Processing Analysis

Pre-processing is an enormous task in any analysis of data. The data given to the system which is called as input is not in refined format and need to be prepared well before processing the actual data mining. The initial data set is in a raw state named as unstructured data, it is not analyzed directly it needs to be structured before analysis it is known as pre-processing. Structured data is quite easy to use in process improvements. Unstructured data is a generic term and is a mixture of textual and non-textual data. Unstructured non-textual data generally having images, video and audio files. Unstructured data is being used by social media companies because to understand their markets and customers in more depth.

Collect the datasets from the internet which is stored in the form of comma separated values (.CSV) format and retrieved in XL sheet. To do pre-processing of data few challenges are faced such as missing values handling (no data value stored for the variable) and noisy values handling (meaningless or corrupted data) is a part of data cleaning. Integrate the data sets from 3 different datasets by using vertical joins and removing the duplicate records for multiple index entries. It is necessary to eliminate missing values from the dataset as the values did not present which can affect the actual results in a negative way. So in dataset need to eliminate missing values. After removing missing values from the dataset, after that getting cleaning tasks. The next thing is removing the noisy data from the dataset. The noisy values can be anything from -1 to infinity values garbage integers in place of expected ranges or negative values. According to that handle the noisy values from some attributes by assigning the mean values of the attributes that belong to the same class as the attributes with the noisy values. This ensures there is no over fitting of the data. The next task in pre-processing is that integration. In this task regarding integration has an issue of vertical join and also of eliminating duplicate records due to multiple indices.

IV. METHODOLOGY

Accident dataset is collected from the internet. The collected data is not in a required format for analyzing it is in an unstructured format so it needs to be pre-processing the data.

Data pre-processing:
Data pre-processing is one of the predominant assignment in data mining. Pre-processing main work is to eliminate missing values and noisy data, separating unrelated data in order to prepare for analyzing the data. Here, the pre-processing goal is to make data structured data (which is understandable by everyone) which makes it ready for the analysis.

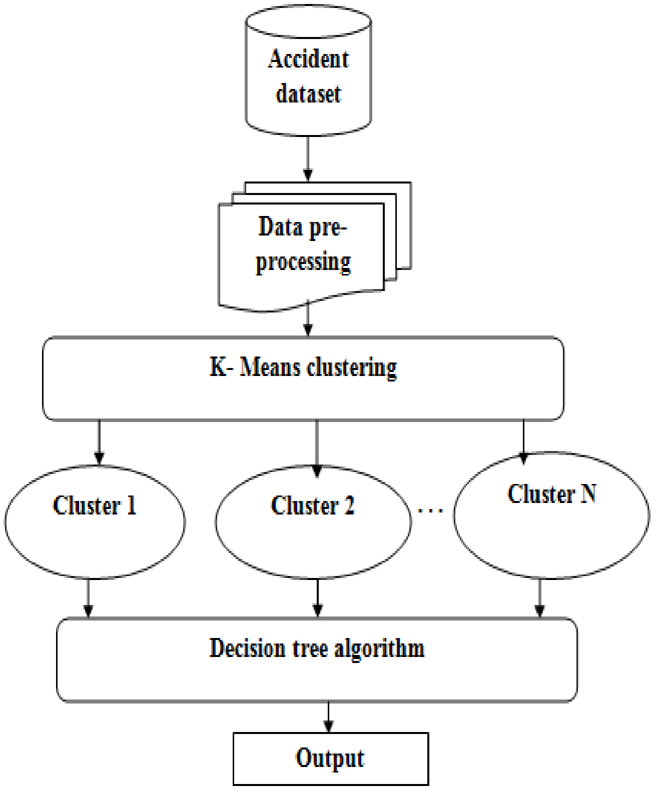

Figure 2: System Architecture

Clustering:
Clustering is nothing but splitting the data points into a number of groups which is defined as the data points in the same group are similar with each other comparing the data points in other groups dissimilar with each other [8]. The similar data points group would be considered as one cluster. Simply telling that the goal is to separate group with similar attributes and set them into clusters.

In this project considering some states and union territories of India for evaluating road accidents. Clustering does between the states regarding which state faced more number of accidents. Results show that Himachal Pradesh, Nagaland, and West Bengal register the highest number of accidents these states represent "high" risk factor for road accidents it can be considered as one cluster. Then Arunachal Pradesh, Chhattisgarh, and Sikkim register less number of road accidents compared with other states, it represents "low" risk factor for road accidents it can be considered as one cluster. Remaining states considered as another cluster.

Decision tree:
The Decision tree is like a tree structure and it having a root node, intermediate nodes, and leaf nodes. The decision depends on each node based on that decision, the tree progresses. It uses supervised learning methods. Tree structure methods allow predictive models with high reliability, stability. Compare with linear models, non-linear models map relationships quite well [9]. For solving any kind of problems decision trees are flexible (regression or classification). Identifying the main cause of road accidents by using classification methods.

Algorithm:

```
1.  Start
2.  Take the dataset i
3.  Create the root node 'n' for the tree
4.  Declare variables rt(road type), at(accident type),
    a(age), g(gender), dc(driver condition), s(severity).
5.  Read attributes 'rt', 'at', 'a', 'g', 'dc', 's'.
6.  Place 'rt' attribute at the root node 'n'.
7.  If 'rt' attribute is empty it will show null output.
8.  If (n==rt)
        {
        Display 'at', 'a', 'g', 'dc', 's'.
          }
          Else
          {
         It shows the null output
9.  Then we will enter another attribute name taken as
    'at'.
10. If (n==at)
            {
            Display 'rt', 'a', 'g', 'dc', 's'.
             }
            Else
            {
            It shows the null output.
            }
11. Repeat step 8 and step 9 with different attribute
    names on root node until you find leaf nodes in all
    the branches of the tree.
```

V. Results and Discussion

The simulation is done by using Eclipse IDE which is an Integrated Development Environment (IDE). Clustering (k-means) and classification (decision tree) techniques have been applied on different parameters of road accidental dataset to analyze and predict the accident prone locations which helps to reduce the frequency of accidental fatality and also finding the conditions of roads.

Statistics results:
The number of road accidents occurs in five years (2013-2017) shown in Fig 3. The more road accidents occurred highest in the year 2013 and least in 2016.

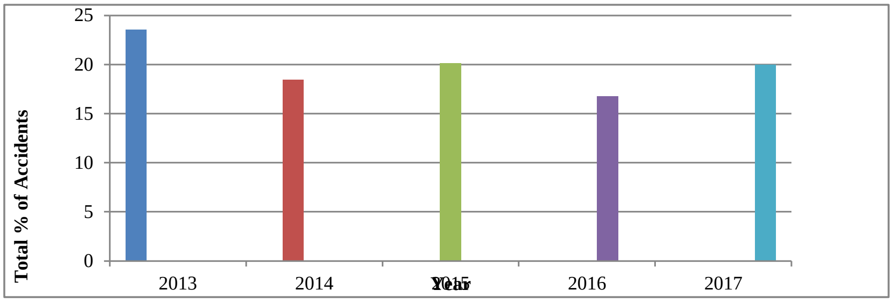

Figure 3: Number of Accidents per Year

Weather Conditions:

The percentage of accidents occurred in different weather conditions shows in figure 4, it compares with a percentage of accidents happened on Indian roads. More accidents occurred at “snow without high winds” weather, and less number of accidents occurred at “fine with high winds” weather. From this, it is clear that “snow without high winds” is the most wanted case of accidents, so people must be careful regarding this situation.

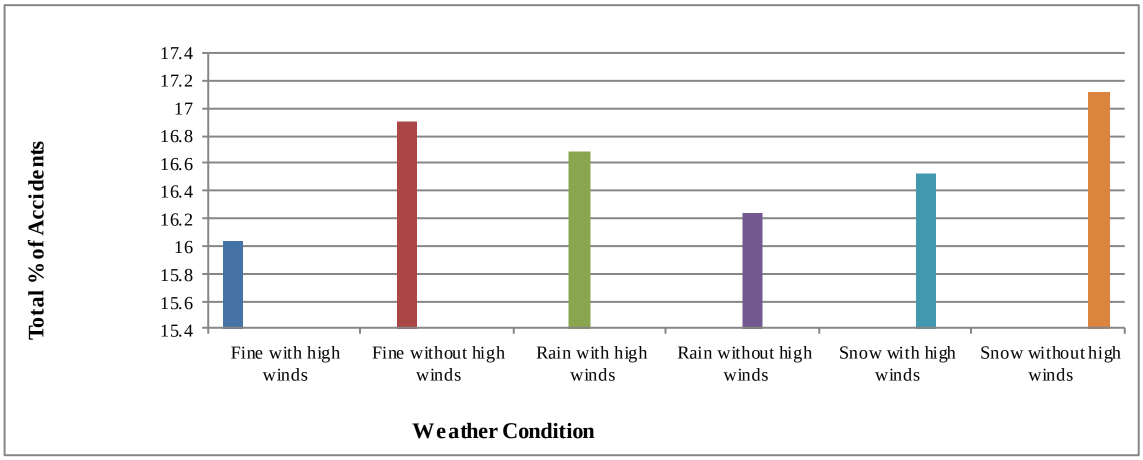

Figure 4: Accidents on Different Weather Conditions

Light conditions:

The percentage of road accidents occurred in various light conditions compares with a total number of road accidents involved is shown in figure 5. More accidents occurred at “daylight” condition why because traffic happens much more at daytime especially peak hours afternoon time compare to night, and less number of accidents occurred at “darkness without street light”.

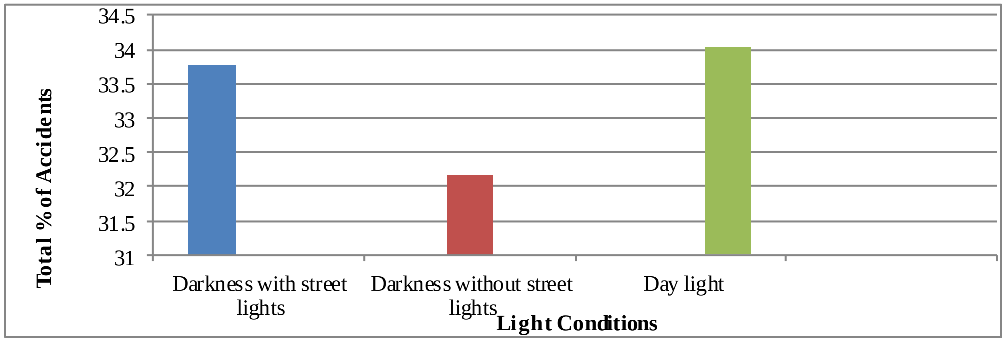

Figure 5: Accidents on Different Weather Conditions

Driver conditions:

The percentage of road accidents involved in different driver conditions compare with a total number of accidents happened shown in fig 6. More number of accidents happened due to drunk and drive condition because at that people lose their concentration on driving. It hampers vision due to dizziness. Compare to this less number of accidents happened due to sleeping mode.

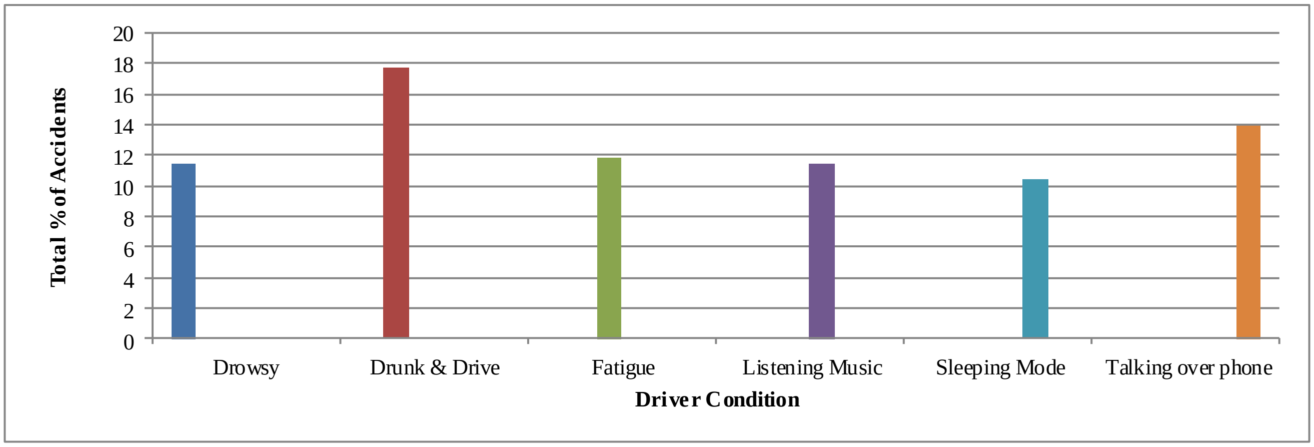

Figure 6: Accidents on Different Driver Conditions

VI. CONCLUSION AND FUTURE SCOPE

The highest number of accidents on roads creates a major problem worldwide. Finding the factors which cause the severity of accident consequences is an important thing of analysis. This study concentrates on analyzing accidents on roads in different states and union territories of India. Two data mining techniques were applied, classification (decision tree) technique predicts accident-prone locations and clustering (k-means) is used to determine pedestrian fatalities. The environmental factors like weather, road surface, light, driver conditions affect road accidents, while other human factors like accident type, drunk or not strongly affect the fatal rate.

In the future, intend to combine large data sources such as car insurance data, hospital records, and roadwork's data from different states and union territories of India transportation system. Along with this analyze the dataset in terms of time of occurrence of accidents can be done to determine most-danger time of the year and most-danger time of the day for road accidents and also death rate related with these accidents.

References

[1] Lilting Li, Shared Shrestha, Gongzhu Hu "Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques" 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp 363-3707, Sept. 2016.

[2] Prajakta S.Kabse, Apeksha Prajakta S. kasbe, Apeksha V. Sakhare "A review on road accident data analysis using data mining techniques" International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).March 2017.

[3] Gangapreet Kaur, Geetika Gandhi, Gobindgarh, India "A Framework for analyzing the Road accidents in Data Mining Using Rule Mining" International Journal of Innovative Research in Computer And Communication Engineering, pp.6931-6939, April 2017.

[4] Ayushi Jain1, Garima Ahuja2, Anuranjana3, Deepti Mehrotra" Data Mining Approach to Analyse the Road Accidents in India", pp.175-179, September 2016.

[5] K Jayasudha and C Chandrasekar. "An overview of data mining in road traffic and accident analysis". *Journal of Computer Applications*, pp.32–37, 2009.

[6] Carlo Giacomo Prato, Victoria Gitelman and Shlomo, "Mapping patterns of pedestrian fatal accidents in Israel", Accident Analysis and Prevention, pp.54–62, Jan 2012.

[7] Svetlana Bačkalić, Boško Matović and Dragan Jovanović, "Identification of hotspots road locations of traffic accidents with pedestrian in urban areas", International Co-operation on Theories and Concept in Traffic Safety, Dec 2014.

[8] Vikas Verma, Shaweta Bhardwaj and Harjit Singh,"A Hybrid K-Mean Clustering Algorithm for Prediction Analysis,"Indian Journal of Science and Technology, pp.0974-6846, (july2016).

[9] Frantisek Babi, Karin Zuskaova, Liangzheng Xia, "Descriptive and Predictive Mining on Road Accidents Data," IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 87-92, 2016.