Mining of the Correlations for Fatal Road Accident using Graph-based Fuzzified FP-Growth Algorithm

Soniya Mudgal, Mahesh Parmar

***Abstract**: Rapid population growth and economic activity have caused a continuous growth of motor vehicles and the increase in population and vehicle traffic injuries is increasing each day. Injury and death traffic accidents lead to not only significant economic losses however too severe mental & physical illness. Social issues have been created by the increasing road accident as a result of death and suffering and death. FP Growth Algorithm, Support Vector Machine (SVM) Cluster classification models and simple C-means clustering Algorithm formed Association laws. Some suggestions for safety driving were made based on data, association guidelines, classification model and obtained clusters. In this paper, we will attempt to address this problem by applying statistical study and FARS fatal accident DM algorithms. The findings suggest that the algorithm proposed is more efficient and faster than the algorithm of the previous research.*

***Keywords:** Spatial data mining, air pollution, association rule mining, Fuzzification, graphical representation.*

I. INTRODUCTION

Traffic accident affect developing countries significantly more than in developed countries. Transport network has been expanded at a high rate, and vehicle safety is a problem for all due to reports of loss of human life and property and fatal accidents and regular traffic blockage. Efficient mobility and accessibility roles are provided by national highways. Social problems have been exacerbated by rising accidents on the roads because of fatalities and human suffering [1].

The interactions between cars, road users and road conditions are the primary causes of road accidents. Each of these fundamental elements includes some components such as pavements, geometrical characteristics, traffic characteristics, behavior, vehicle design, driver characteristics, and environmental factors. Accident cause can be well understood by analyzing accident statistics, which can give insights into many factors that lead to road accidents.

Revised Manuscript Received on May 15, 2020.
* Correspondence Author

Soniya Mudgal*, Computer Science and Engineering, Madhav Institute of Technology and Science (MITS), Gwalior, India. Email: mudgalsoniya23@gmail.com

Mahesh Parmar, Computer Science and Engineering, Madhav Institute of Technology and Science (MITS), Gwalior, India. Email: maheshparmarcse@gmail.com

Many studies have focused on road accidents and study innovative work on analyzing road accidents. In several India, studies have been carried out for forecasting road accidents using population and the vehicle population in several cities such as Bangalore, Kolkata, Hyderabad, Ahmadabad, Delhi, and Chennai. [2] In most Indian cities, urban transport facilities inadequately and over the years have deteriorated. In terms of quality as well as quantity, the development of the system of public transport has remained unchanged. Travel and traffic risk grow much faster as the growth of registered vehicles increases constantly with population growth.

India has now been provided a dubious distinction by traffic accidents, which have overtaken China to the highest level of road fatalities with nearly 140 000 deaths a year. India is the world's only country facing more than 15 fatalities and 53 injuries every hour in the immediate aftermath of road crashes. Because the situation is generally improving in many developed and developing countries like China, India faces a worsening situation. The principal objective of this study is a national, state, and urban analysis of traffic accidents in India. Identifying key issues in road safety and addressing counter-measures that might address the specific road safety issues [3]. Road accidents (RAs) are currently one of India's leading causes of death, disability & high socio-economic hospitalization. Some road accidents have put a significant social and economic burden on accident victims.

The main goal of this study is the analysis of road accidents at the national, state and metropolitan levels in India. The goal would be to recognize important issues in road safety and to discuss countermeasures that could fix special problems in road safety.

II. REVIEW OF LITERATURE

[4]Survey showed that, as regards the accident death rate per thousand vehicles, small states in India have a doubtful record. In Arunachal Pradesh, it was 5.7% higher and in Sikkim 3.6% higher. The highest accident rate in Nagaland too was 92.1%, followed by 89.7% in Mizoram, compared to 28.4% in the region. In all Indian cities, approximately 21.5% of the total incidents during 1977, which slightly decreased by 5% to 16.9% in 2001 and was reportable in 7 metropolitan cities, Chennai, Hyderabad, Delhi, namely Ahmadabad, Kolkata, Mumbai, and Bangalore. [5] The purpose of this study was to identify the monthly and annual accident rate variability, the impact on the traffic accident rate and the development model with the use of AADT and road conditions. Accidents / Km / year = C0 + C1

(AADT) + C2 (Road condition Rank) Predicts equation:. Airport accident prediction is equivalent to the number of accidents with AADT increases per km and decreases as the conditions of the roadside shoulder improve.

[6] A case study of RAs analysis in Patna found that congestion & prejudice were major reasons for road accidents.

[7] Models of road accidents developed for India's major metropolitan cities. The main purpose of the project was to establish models through the review of road accident data both for large cities and all of India. To measure the type and depth of the event by using Smeed's concept and Andreassen's equation, the data for the 25 years from 1977 to 2001 were analyzed. The basic conclusion drawn from this research was that major policies could be modified to minimize the number of people using public transport vehicles, to reduce the growth of personalized vehicles.

[8] A research on urban roads in Malaysia was carried out to develop models to predict motorcycle accidents at signaled crossings. Research concluded that motorcycle crashes are directly proportional to the number of motorcycles entering the traffic signals.

[9] Considered trying to detect the impact on driving performance of using cell phones. Specific researchers have shown that drivers who spoke on their cell phones are less likely to stop completely at the stop sign.

[10] ANN-oriented method was demonstrated to estimate deaths in motor vehicle accidents and the results showed that the ANN model is an approach that predicts fatalities in traffic accidents.

[11] Research on accident severity prediction using artificial neuron network techniques was presented. In this study, ANN models can be used for accident frequency estimations and crash-related significant factors.

[12] A research on traffic accident modeling in Turkey was shown. The conclusion was that a lack of standards is the main driver of safety on highways.

III. METHODOLOGY

Problem Statement:

A. Data Preparation

Factors behind Traffic Mortality on Our Roads — FARS is a national census that gives annual data about fatal injuries suffered in motor vehicle crashes from NHTSA, Congress, & US government. https://data.world/nhtsa/fars-data. To identify problems, the program collects data for the analyzes of road safety crashes and assesses counter-measures to reduce injury and property damage arising out of crashes. Every fatal crash is defined in the FARS dataset in the standard format. The data elements that describe the accident, the vehicles and the persons involved are more than 100 coded each crash. There are also motor vehicle accident details the information sort that FARS-big application-processes.

B. Designing

1) **Rules of Association:** Rule Mining Association [28] is a popular technology for DM and extracts interesting and hidden links between different attributes of a large dataset. The association for rule mining establishes a set of rules that describes the trends underlying the data package. The data set specifies the average frequency of incidence of the two incident features. A rule A→ B shows that if A occurs then B also occurs B.

There are two important steps to identify the co-located patterns

a) Conversion of the spatial data

- The spatial data information.
- Input representation.

b) Co-location pattern mining

- FP tree construction.
- Mining from FP-Tree.

When developing the FP tree [8] is used to use frequent patterns from a compact tree arrangement FP tree mining process. Divide-and-conquer functions of FP-growth. The first database scan extracts a list of frequently selected items by order of the frequency. In the frequency descending list which collects information on the association of objects, the data basis is packed into a pattern tree or FP tree.

C. Graph-Based Fuzzified FP-Tree

In regular itemset mining, time complexity is one of the major issues and our proposed new solution is to overcome it. Interestingly, this Graph-Based Approach searches the entire database once and is the most sought-after technique in the area of regular sets defining large amounts of candidates. Before scanning, a graph is created, which is a directed graph. The graph size is stored in the main memory and is shown as an adjacent matrix. In this case, the data set elements are represented at the top of the directed diagram, and the vertex weight indicates the support count of one element instead of several sets. The vertices with weights are connected by a surface. For mining large k-itemsets, relationships must be established (k>=3). Some transactions in a database will include the same set of objects, but two operations are fundamentally distinct from each other and two transactions may include the same itemsets such that their sub-sets are similar.

Algorithm for Graph-based approach

Input: Transaction Set D and the complete number and occurrence of itemsets.

Output: Frequent itemsets: in so several steps our algo functions:

a) Scan: scan database & start it in form of adjacency matrix.

b) Identify: update & identify values of every element of matrix Aij & Aij.count

c) Construct: Delete the corresponding row and column of element Aij.count= 0 for i = j only, create the reduced adjacency matrix. The recurrent 1 collection shows the patterns sometimes mined.

d) Mine: At this step, more level from each row is extracted using the logical AND operator from row elements.

D. SVM Algorithm

SVM (Support Vector Machine) is supervised machine learning (ML) algo, and can be utilized to classification problems. This is therefore seen primarily in problems of classification.

In SVM algo, any data element is plotted as a space in n dimensions (where n is several features that you have) and each feature's value is the value of a given coordinate.

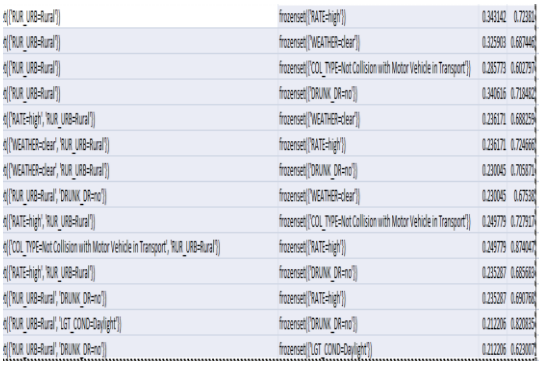
Naive Bayes comes under the section of generative classification structures. This determines the inverse chance from the class conditions. The output is thus the probability of belonging to a class.

SVM conversely is depended on the discriminant function given by y = w.x+b. The bias parameter b and weights w are derived from the training results. It seeks to find a hyperplane optimizing the margin, and in this way, there is optimization function.

Performance-wise SVMs can do best as they manage data non-linearity.

E. Proposed Algorithm

```
Step 1  Start
Step 2  Load the FARS 2015 dataset
Step 3  Preprocess the dataset by Removing the null
        values from the dataset
Step 4  Convert dataset from numeric to nominal
Step 5  Apply FP Growth Association rule mining
        algorithm
Step 6  Train the dataset using Naïve Bayes
Step 7  Cluster the dataset using C-means clustering
Step 8  $\sum_{i=1}^{N}\sum_{j=1}^{c} \imath$ , 1≤m<∞
Step 9  Clustering will demonstrate the number of states
        with fatal accidents
Step 10 Terminate
Step 11 End
```



t({'RUR_URB=Rural'})	frozenset({'RATE=high'})	0.343142	0.72381
t({'RUR_URB=Rural'})	frozenset({'WEATHER=clear'})	0.325903	0.687446
t({'RUR_URB=Rural'})	frozenset({'COL_TYPE=Not Collision with Motor Vehicle in Transport'})	0.285773	0.602797
t({'RUR_URB=Rural'})	frozenset({'DRUNK_DR=no'})	0.340616	0.718482
t({'RATE=high', 'RUR_URB=Rural'})	frozenset({'WEATHER=clear'})	0.236171	0.688259
t({'WEATHER=clear', 'RUR_URB=Rural'})	frozenset({'RATE=high'})	0.236171	0.724666
t({'WEATHER=clear', 'RUR_URB=Rural'})	frozenset({'DRUNK_DR=no'})	0.230045	0.705871
t({'RUR_URB=Rural', 'DRUNK_DR=no'})	frozenset({'WEATHER=clear'})	0.230045	0.67538
t({'RATE=high', 'RUR_URB=Rural'})	frozenset({'COL_TYPE=Not Collision with Motor Vehicle in Transport'})	0.249779	0.727917
t({'COL_TYPE=Not Collision with Motor Vehicle in Transport', 'RUR_URB=Rural'})	frozenset({'RATE=high'})	0.249779	0.874047
t({'RATE=high', 'RUR_URB=Rural'})	frozenset({'DRUNK_DR=no'})	0.235287	0.685683
t({'RUR_URB=Rural', 'DRUNK_DR=no'})	frozenset({'RATE=high'})	0.235287	0.690768
t({'RUR_URB=Rural', 'LGT_COND=Daylight'})	frozenset({'DRUNK_DR=no'})	0.212206	0.820835
t({'RUR_URB=Rural', 'DRUNK_DR=no'})	frozenset({'LGT_COND=Daylight'})	0.212206	0.623007

Fig. 1.Number of rules showing higher fatality rate cause.

FCM (Fuzzy c-means) is a clustering method that allows for the use of two or more clusters for one piece of data. When it is used as a process for model recognition (developed by Dunn in 1973 and enhanced by Bezdek in 1981). This is based on the following objective function being minimized:

$$J_m \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m}||x_i - c_j||^2 \text{ ,} 1 \leq \qquad (1)$$

When m is a valid number greater than 1, uij is xi of cluster j, d-dimensional calculations are xi, cj is the d-dimensional cluster center and ||*|| is normal for the similarity among calculated and center data.

IV. RESULTS AND DISCUSSIONS

Fig. 2 shows no. of fatal accidents per month. During August and February, the most fatal accidents were the least.

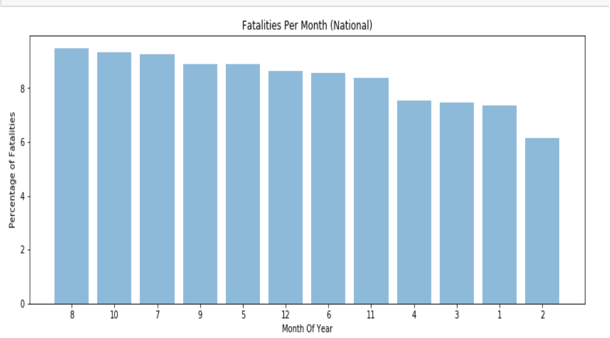


Fig. 2.Number of fatal accidents per month.

Fig. 2 illustrates no. of fatal accidents every hour of a day. In August and February, the most fatal accidents happened.

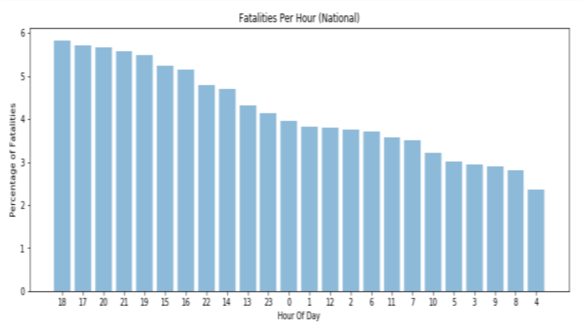


Fig. 3.Number of fatal accidents per hour.

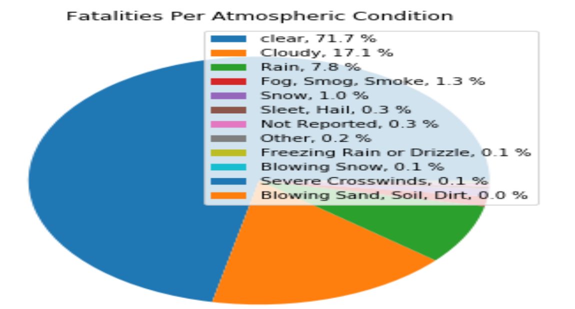


Fig. 4.Number of fatal accidents on different atmospheric condition.

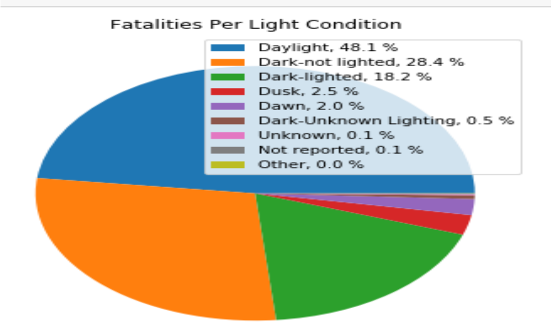


Fig. 5.Amount of fatal accidents in various lighting conditions.

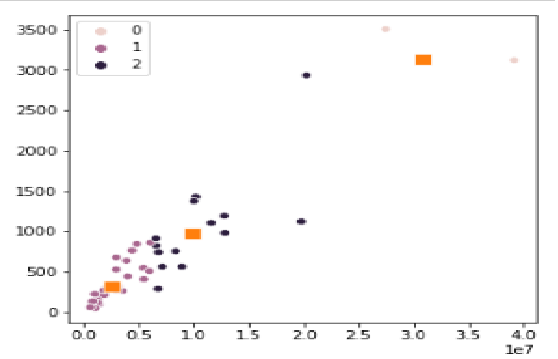
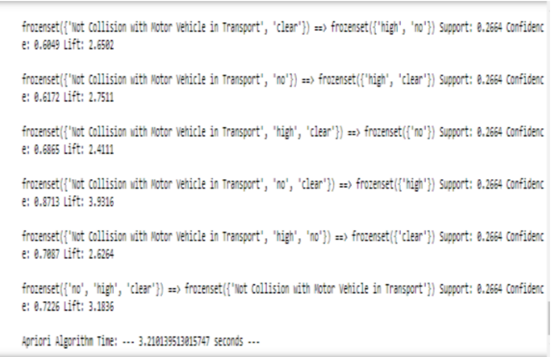


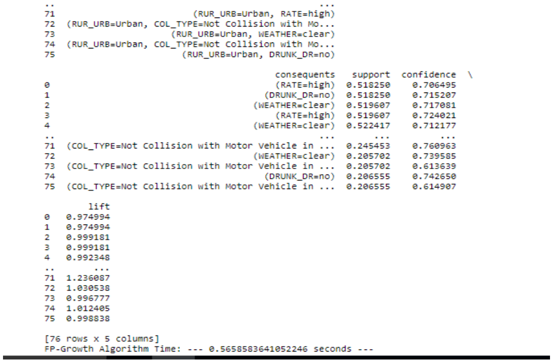
Fig. 6.States clusters per million people per million in states.

Figure 6 illustrates number, in the form of low, higher fatality rates, of States, affected by fatal rates and shows that cluster 0 has a relatively high population and fatality rate in California and Texas. Cluster 1 consists of 29 states and is fatal less, compared to Cluster 2 has 15 states.



```
frozenset({'Not Collision with Motor Vehicle in Transport', 'clear'}) ==> frozenset({'high', 'no'}) Support: 0.2664 Confidenc
e: 0.6049 Lift: 2.6502

frozenset({'Not Collision with Motor Vehicle in Transport', 'no'}) ==> frozenset({'high', 'clear'}) Support: 0.2664 Confidenc
e: 0.6172 Lift: 2.7511

frozenset({'Not Collision with Motor Vehicle in Transport', 'high', 'clear'}) ==> frozenset({'no'}) Support: 0.2664 Confidenc
e: 0.6865 Lift: 2.4111

frozenset({'Not Collision with Motor Vehicle in Transport', 'no', 'clear'}) ==> frozenset({'high'}) Support: 0.2664 Confidenc
e: 0.8713 Lift: 3.9316

frozenset({'Not Collision with Motor Vehicle in Transport', 'high', 'no'}) ==> frozenset({'clear'}) Support: 0.2664 Confidenc
e: 0.7087 Lift: 2.6264

frozenset({'no', 'high', 'clear'}) ==> frozenset({'Not Collision with Motor Vehicle in Transport'}) Support: 0.2664 Confidenc
e: 0.7226 Lift: 3.1836

Apriori Algorithm Time: --- 3.210139513015747 seconds ---
```

Fig. 7.Time taken by Apriori Algorithm for mining rules.

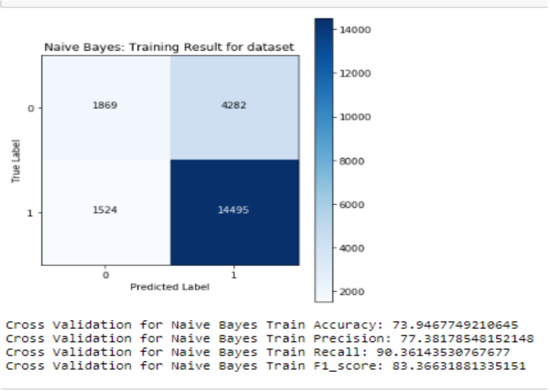


```
..                                               ...
71                          (RUR_URB=Urban, RATE=high)
72  (RUR_URB=Urban, COL_TYPE=Not Collision with Mo...
73                       (RUR_URB=Urban, WEATHER=clear)
74  (RUR_URB=Urban, COL_TYPE=Not Collision with Mo...
75                        (RUR_URB=Urban, DRUNK_DR=no)

                                     consequents   support  confidence  \
0                                    (RATE=high)  0.518250    0.706495
1                                  (DRUNK_DR=no)  0.518250    0.715207
2                                (WEATHER=clear)  0.519607    0.717081
3                                    (RATE=high)  0.519607    0.724021
4                                (WEATHER=clear)  0.522417    0.712177
..                                           ...       ...         ...
71  (COL_TYPE=Not Collision with Motor Vehicle in ...  0.245453    0.760963
72                               (WEATHER=clear)  0.205702    0.739585
73  (COL_TYPE=Not Collision with Motor Vehicle in ...  0.205702    0.613639
74                                 (DRUNK_DR=no)  0.206555    0.742650
75  (COL_TYPE=Not Collision with Motor Vehicle in ...  0.206555    0.614907

        lift
0   0.974994
1   0.974994
2   0.999181
3   0.999181
4   0.992348
..       ...
71  1.236087
72  1.030538
73  0.996777
74  1.012405
75  0.998838

[76 rows x 5 columns]
FP-Growth Algorithm Time: --- 0.5658583641052246 seconds ---
```

Fig. 8.Time taken by FP Algorithm for mining rules.

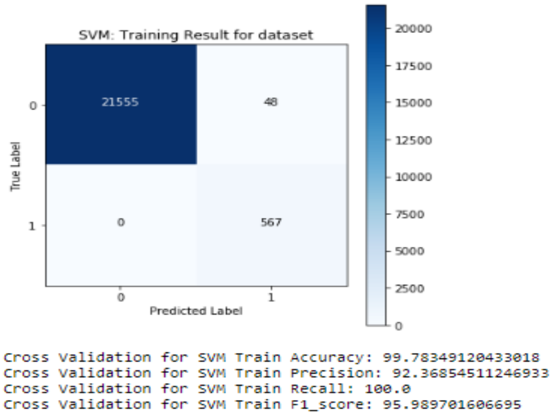
Figures 7 and 8 show the time taken by the association rule mining algorithms for mining the most efficient rules. FP-Growth algorithm took 0.5658 seconds whereas the Apriori algorithm took 3.2101 seconds for mining rules. Therefore FP Growth is faster and more accurate than the Apriori algorithm.

Classification Algorithm SVM shows 99.78 precision, while considered as a limitation compared to other data set attributes, the fatal rate does not depend heavily on the attributes given.



```
Cross Validation for Naive Bayes Train Accuracy: 73.9467749210645
Cross Validation for Naive Bayes Train Precision: 77.38178548152148
Cross Validation for Naive Bayes Train Recall: 90.36143530767677
Cross Validation for Naive Bayes Train F1_score: 83.36631881335151
```

Fig. 9.Accuracy measurements of trained Naïve Bayes.



```
Cross Validation for SVM Train Accuracy: 99.78349120433018
Cross Validation for SVM Train Precision: 92.36854511246933
Cross Validation for SVM Train Recall: 100.0
Cross Validation for SVM Train F1_score: 95.989701606695
```

Fig. 10. Accuracy measurements of trained SVM.

Table- I: Results of the SVM

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	99.78	92.36	100	95.98
Naïve Bayes	73.94	77.38	90.36	83.36

V. CONCLUSION

The country road accident scenario is of great concern. Every year the number of accident deaths rises because of the congestion of commercial vehicles over speed on all categories of roads. To determine the cause/characteristics of accidents the FARS data set was analyzed. We could see that some states and regions have a higher rate of fatality from the clustering results while others are lower. Association rule mining algorithm determines the major reasons behind the fatal accidents happening every day. From the classification and ARM, we found that accidents do not fully depend on any particular attribute but human factors also. In the future, we will try to work on other attributes also and will see the result, how the other attributes can affect the dataset and the obtained results.

REFERENCES

1. Census Bureau of India. Population Census Data. Available at Ministry of Home Affairs, Government of India website: http://www.censusindia.net (retrieved 18 October 2003).
2. UNICEF. The state of the world's children. New York: UNICEF; 1999.

3. Society of Indian Automobile manufacturers. Available at SIAM, New Delhi website: http://siamindia.com/General/ domestic-sales-trend.aspx (retrieved 4 February 2004).
4. Peden M, McGee K, Sharma G. The injury chart book: A graphical overview of the global burden of injuries. Geneva: WHO; 2002.
5. WHO. Making a difference: The World health report 1999. Geneva: WHO; 1999.
6. Murray CJ, Lopez AD. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. Lancet 1997;349:1436–42.
7. Gore G. Searching the medical literature. Inj Prev 2003;9:103–4.
8. Sathiyasekran BWC. Study of injured and injury pattern in a road traffic accident. Ind J Forensic Sciences 1991;5:63–8.
9. Mishra BK, Mohan D. Two-wheeler injuries in Delhi, India: A study of crash victims hospitalized in the neuro-surgery ward. Accid Anal Prev 1984;16:407–16.
10. Srivastava AK, Gupta RK. A study of fatal road accidents in Kanpur. J Indian Acad Forensic Med 1989;11:24–8.
11. Sathiyasekaran BWC. A population-based cohort study of injuries. Injury 1996;27:695–8.
12. Banerjee KK, Agarwal BB, Kohli A, Aggarwal NK. Study of head injury victims in fatal road traffic accidents in Delhi. Indian J Med Sci 1998;52:395–8.

AUTHORS PROFILE

Soniya Mudgal pursed Master of Technology from Madhav Institute of Technology and Science, Gwalior and Bachelor of Engineering from Shri Ram College of Engineering and Management (SRCEM) in year 2016. He is currently pursuing M. Tech in Cyber Securities. In bachelor Degree she did a live project on dot net technology by developing a web application.

Mr. Mahesh parmar received M.E. degree in Computer Engineering from SGSITS Indore. He is currently working as Assistant Professor in CSE/IT Department in MITS Gwalior and having 10 years of Academic and Professional experience. since 10 years. He has guided several students at Master and Under Graduate level. His areas of current research include Data mining and Image Processing. He has published more than 30 research papers in the journals and conferences of international repute. He has also published 02 book chapters. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, and IET etc.