World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(CASE REPORT)

Check for updates

Discovering associated factors behind road accidents using association rule mining: A case study from Gujarat, Pakistan

M. Tariq [1,*], N. Q. Mehmood [1] and S. Z. Mahfooz [2]

[1] *Department of Computer Science & Information Technology, University of Lahore, Gujarat, Pakistan.*
[2] *Department of Computer Science and Engineering, University of Hafr Al-Batin, Saudi Arabia.*

World Journal of Advanced Research and Reviews, 2022, 15(03), 001–011

Publication history: Received on 28 July 2022; revised on 30 August 2022; accepted on 01 September 2022

Article DOI: https://doi.org/10.30574/wjarr.2022.15.3.0885

Abstract

Traffic and road safety is an international issue that is among the major concerns of the respective governing bodies around the world. Regulations and restrictions surrounding this issue can be legislated and applied only after identifying and understanding the numerous factors and conditions that increase the likelihood of traffic accidents. In this research work, we study and analyze the road accidents data and figure out the major factors that contribute to these incidents. The data was collected from Gujrat rescue office and outstanding results are achieved by applying association rules mining using Apriori algorithm. Our research and analysis found that some factors are greater contributors than others and need careful enforced legislation. Other researchers have also used the Apriori algorithm to analyze traffic accidents data from their regions and like our experience this algorithm has produced accurate and effective results. We also try to provide additional insights through visualization that may serve as guidelines to regulate better traffic control systems

Keywords: Data mining; Apriori algorithm; Association rule mining; Traffic accidents analysis

1. Introduction

In a global status report on road safety, World Health Organization (WHO) declared that around 1.35 million people die because of road traffic accidents each year and an additional 20-50 million people are left injured or disabled [1]. The speed at which a vehicle travels directly influences the risk of a crash, severity in injuries, and the likelihood of death occurring because of the crash [2]. Although the road traffic fatalities have become less constant since the year 2007, however; in developing countries the rate of fatalities is increasing [2]. The road traffic accidents are disproportionately higher in low and middle-income countries, and fatalities resulting from road traffic accidents are at least twice as likely in these countries [3]. This is alarming that the road traffic injuries are the eighth leading cause of death for all age groups and the leading cause of death amongst young adults and children between the age of 5-29 years, globally [1]. In 2016 alone, an estimated 181,384 accident casualties occurred in Britain with 1,792 of these accidents being reported as fatal [4]. For the possible remedy of traffic accidents, it is important that we understand the various conditions and factors affecting these incidents and try to determine preventable causes. The road accidents' data, that we want to analyze, in this paper was retrieved from the traffic police rescue office in Gujrat (Pakistan).

The data in its original form was not well-managed and it required few phases of data cleaning before we apply the necessary algorithm that helped us to achieve accurate and reliable results. We followed data analysis and visualization techniques to access, assess, and investigate this data. The break-down of data focused on factors such as injuries, location of the accident, day of the week, month of the year, time of day, intensity of light (day light versus dim light), weather conditions and other common causes contributing most to these accidents. To achieve effective solutions, we

* Corresponding author: : M. Tariq
Department of Computer Science and Information Technology, University of Lahore, Gujrat, Pakistan.

narrowed down our data to the most relevant and applicable factors and then applied the mining association rules using Apriori algorithm.

The nature of road traffic is different in different areas and cities as well as the rush hour timings. This majorly depends on the size and structures of the roads, number of vehicles, papulation size, socio-economic activities, and day timing etc. [5-6]. To confirm generalization of our approach, we compare our method and results with other available scholarly work. The methodology is not bound to any area, city, or country, but with the availability of the traffic data, this framework provides a base to analysis and interpret results from any road traffic data having the relevant information. It can identify and clearly distinguish the most common factors contributing to traffic accidents out of the given data. We provide statistics to suggest that which contributors should be considered while creating provincial & federal legislation as well as other infrastructure development measurements. These measures will eventually prevent and decrease the amount of traffic accidents.

In Section II, we present some literature overview. Section III explains our methodology and experimental results are described in Section IV. Finally, in Section V, we conclude this research work.

2. Literature Review

Traffic accidents are a global issue, and no country or part of the world is immune to the devastating effects and dangers resulting from road incidents. Researchers are focusing more of their studies around association rules that can help determine the relationship between road and accident severity using accident data. In [6], the researchers apply association rule mining to find the frequent patterns that coexist during accidents on the roads. In the advancement of their approach, the same authors apply association rules upon the data set after first finding different clusters within the existing data using different clustering techniques, especially K-means and Hierarchical clustering [7].

The severity of road accidents in Dubai is analyzed by using Apriori and Predictive Apriori algorithm [8-9]. The results show that Apriori algorithm is more effective in generating rules and exploring the severity level of accidents. Analytic Hierarchy Process (AHP)-Apriori is used to sequence the causation factors contributing to traffic accident and to determine the most relevant ones [10]. It then uses the Apriori algorithm to analyze the degree of the accident and its influence. The work in [11] also supports the importance of association rules in determining the relationship between roads and accidents, as it reflects that association rule mining is a key technique used in identifying the correlation in different parameters of road accidents. The number of traffic accidents in young vs. older age groups are analyzed and association rule was found as the most relevant factors during the studies [11]. The researchers in [12], uses association rule mining to find the basic summary statistics and rule descriptions at a limited scale as they focus only on time and vehicle relevant attributes. Similarly, in [13], the authors use the same technique to find spatial features and infrastructure play a key role in the road crashes. By using the association rule, they were able to reach accurate and reliable results in the correlation between age and traffic accidents, like what we have accomplished in our study. Utilizing the same mining technique as the last study, the authors in [14] identify the road accidents that occurred on Sunday have more fatal consequences. Using same approach [15], the effects of road accidents are identified, such as driving skills, road type and road visibility situations. They say that the age of vehicle and weather conditions have no significant effects.

3. Material and methods

After collection of data, there are always involved some steps of data cleaning, data profiling and data preprocessing. Some modern libraries and packages available for different programming languages and analysis tools reduce the overhead associated with the research work. They help researchers to focus on their objectives while performing all such activities conveniently. For performing machine learning related tasks, we used a popular tool, called Waikato Environment for Knowledge Analysis (Weka) [16-17]. Weka is a free software developed at the University of Waikato, New Zealand and it provides a platform to performs machine learning tasks with the help of a collection of machine learning algorithms. Our research methodology is depicted through Fig. 1, and the details of its major implementations are as follows:

3.1 Data Collection

Usually, it is hard to get data from any government organization in Pakistan. Also, the data available is mostly ill-managed. We were able to collect year wise data from 2018 to 2020. The data was first sorted and consolidated into a single worksheet. A basic analysis helped us to consider factors that can determine and identify the categories contributing to the most significant and thorough relations to the desired results.

3.2 Data cleaning and pre-processing

Data was cleaned by removing redundant entries, misspelled words, missing values, and corrupt data. Pre-processing helps to remove noise and irrelevant attributes to make data ready for analysis. In this step, our aim is to preprocess the road accident records to make our data appropriate for the analysis.

The following Fig.1 is the model process of our efforts to generate rules and find hidden patterns or association between different attributes. The data is transformed into CVS (Comma Separated Values) and to ARFF (Attribute Relation File Format). Eventually, input the data from the ARFF into the Weka. After patterns are found, further association rules were discovered and applied by using the Apriori algorithm and finally some useful results were generated. Association rule mining is one of the techniques in Data Mining methods that explores the relationships between different attributes of data [18]. Apriori is a basic algorithm that finds frequent item-sets based on candidate generation. It helped to recognize the frequent item-sets in our data to discover associate rules and generate results, reflecting specifically on the most common and frequent contributors to traffic accidents in the area.

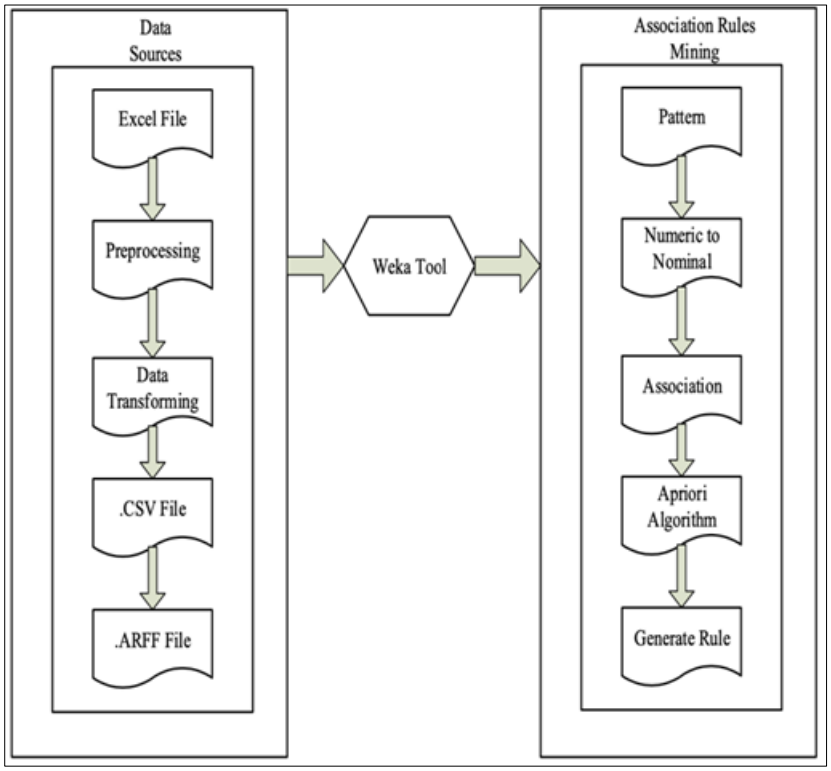

Figure 1 Methodology

4. Results

The results are generated through data mining by using the Weka Tool Kit and by applying the Apriori algorithm that highlights the most relevant and common contributors to traffic accidents. The factors used to generate these contributors are location of the accident, injuries, day of the week, month of the year, time of day, light intensity (day light versus dim light), weather conditions, and overall common causes contributing most significantly to the accidents.

4.1 Location wise statistics

The data includes various locations that fall into various categories. Most common categories may include G.T. road, filling-station, hotel, chowk (Intersections), and bridge. Amongst the top locations analyzed (Gt Road, Filling Station, Hotel, Chowk, and Bridge), it can be seen in Fig. 2. that the greatest number of traffic accidents occurred on chowk with almost nine hundred occurrences; followed by G.T. road, filling Station, hotel, and bridge, respectively.

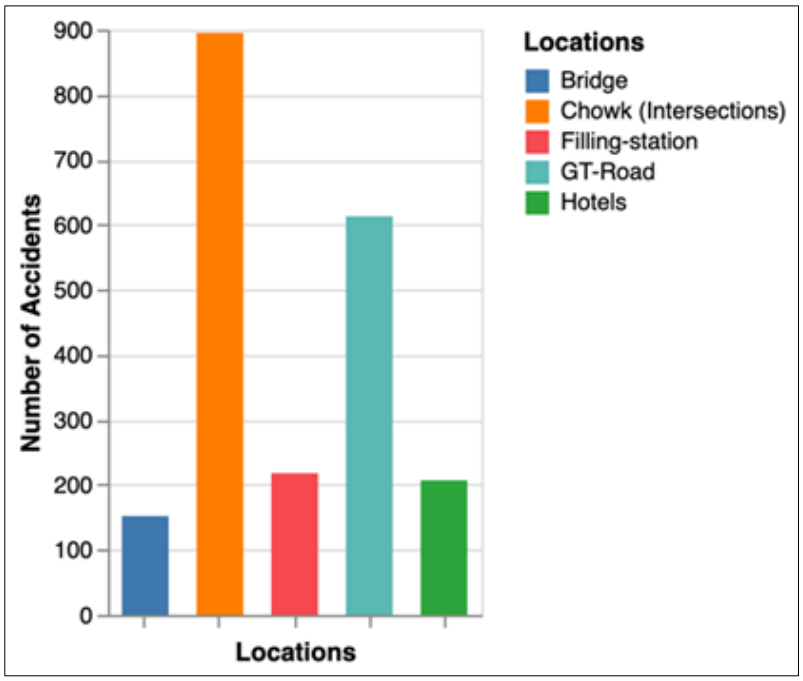

Figure 2 Locations and occurrences of accidents

4.2 Survey of injury

The next factor considered is the severity of injuries resulting from these accidents, ranging from slight to severe. As can be seen in Fig. 3, the generated results reflect that most injuries (almost 1600) were in the moderate range and slightly over two hundred injuries ranked in the slight and severe injury category.

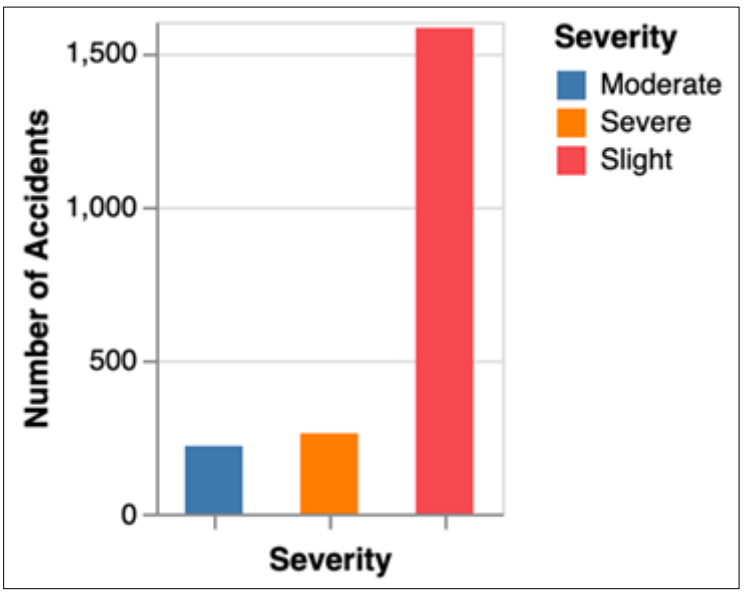

Figure 3 Severity of Injuries

4.3 Days of week

In Fig. 4, the days of the week are columned into three separate categories: number 1 represents Saturday and Sunday, number 2 represents Monday, Tuesday and Wednesday, and number 3 denotes Thursday and Friday. The greatest number of accidents occurred in column 2, meaning on Mondays, Tuesdays, and Wednesdays. That is probably because of congested traffic on these working days. They are the busiest work and school days of the week. The roads and highways are not as busy on Thursdays, Fridays, Saturdays, and Sundays.

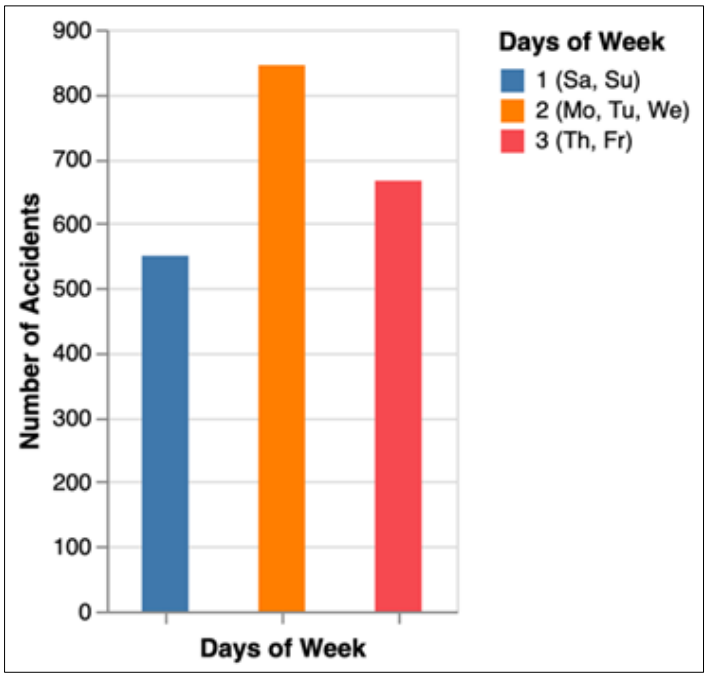

Figure 4 Accidents in different days of week

4.4 Time of the day

The time of the day these incidents occurred is also a relevant factor to be considered. As seen in Fig. 5, traffic accidents are most likely to occur during the afternoon. This can logically be attributed to rush hours, thus, resulting in more traffic than usual and increasing the likelihood of accidents occurring.

Fig. 6 reflects that traffic incidents are more likely to occur during day light versus dim light, which can be linked to the simple fact that more people are on the road during the day.

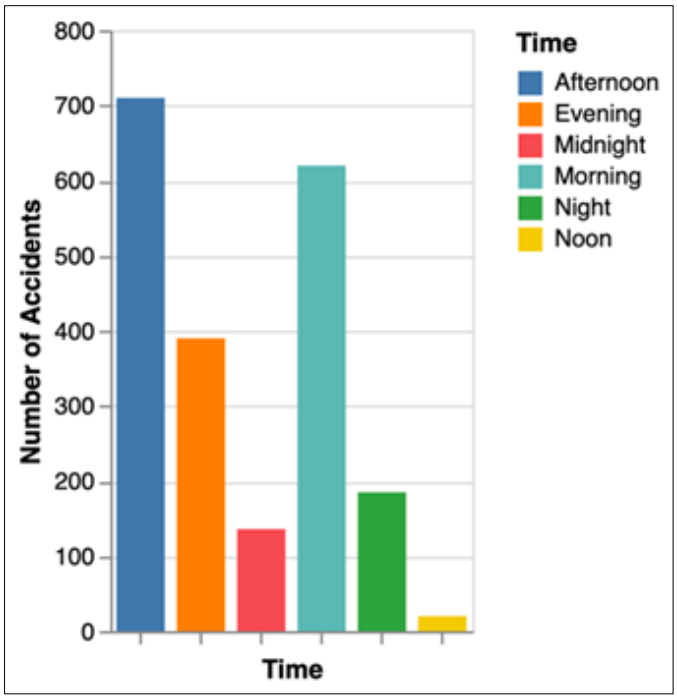

Figure 5 Time of day

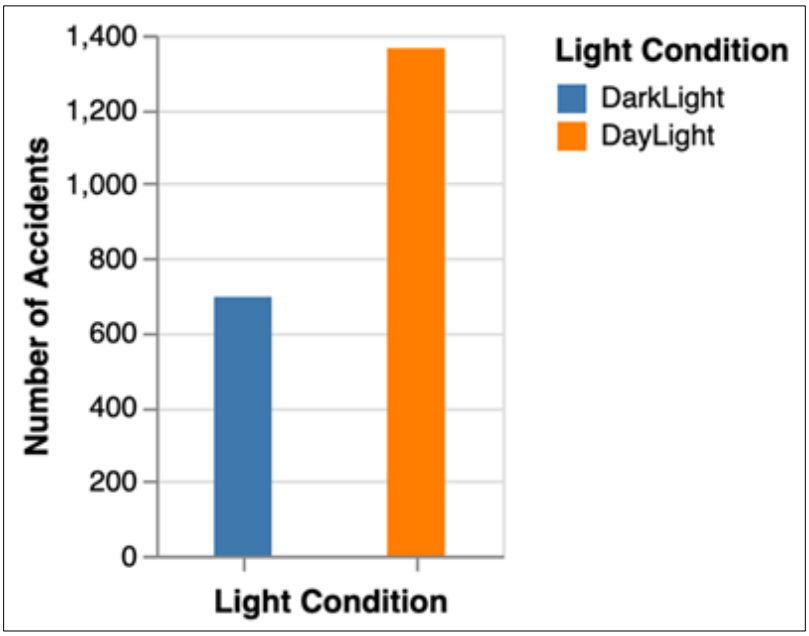

Figure 6 Accidents in different light conditions

4.5 Months

Furthermore, the months of the year are separated into four quarters. quarter one represents January, February, and March; quarter two represents April, May, and June; quarter three consists of July August and September; and the last quarter four includes the months of October November and December. The number of accidents in these quarters are displayed in Fig. 7. The greatest number of accidents occurred in the fourth quarter, standing for the last three months of the year: October, November, and December.

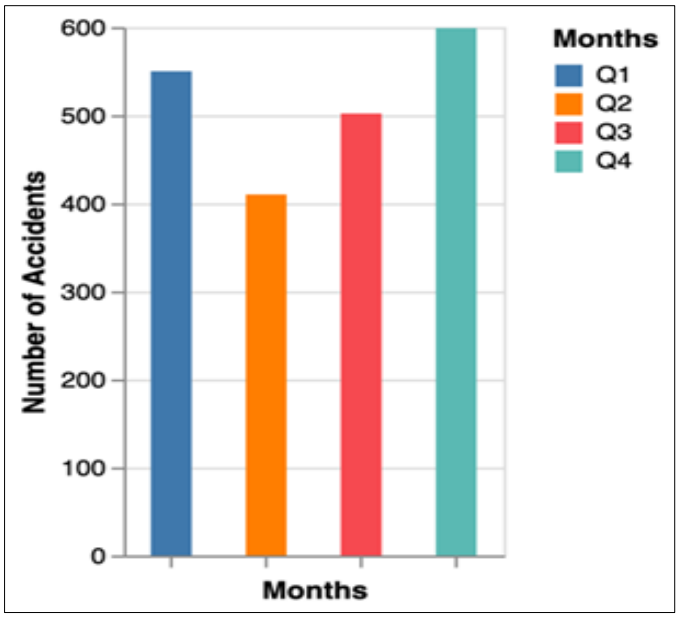

Figure 7 Number of accidents around different quarters of the year

4.6 Weather condition

However, upon generating results for weather conditions, the weather in these last three months bears no link to the increased number of accidents; since Fig. 8 reflects that the greatest number of accidents have in fact occurred in clear weather.

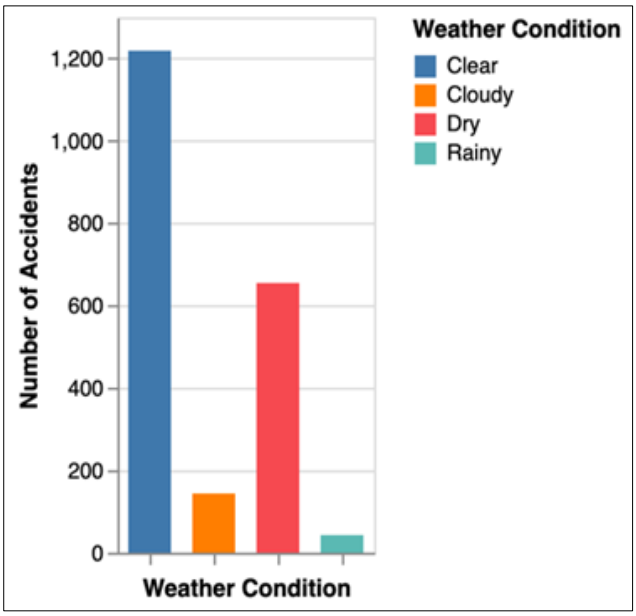

Figure 8 No. of accidents by weather conditions

4.7 Common causes of road accidents

Lastly, we focus on other big factors contributing to the traffic accidents, such as over-speeding, careless driving, driver effect, vehicle condition and pedestrian crossing. These results are reflected in Fig. 9, and as expected, over-speeding has the largest effect, among others. With over eight hundred of these incidents are because of over-speeding.

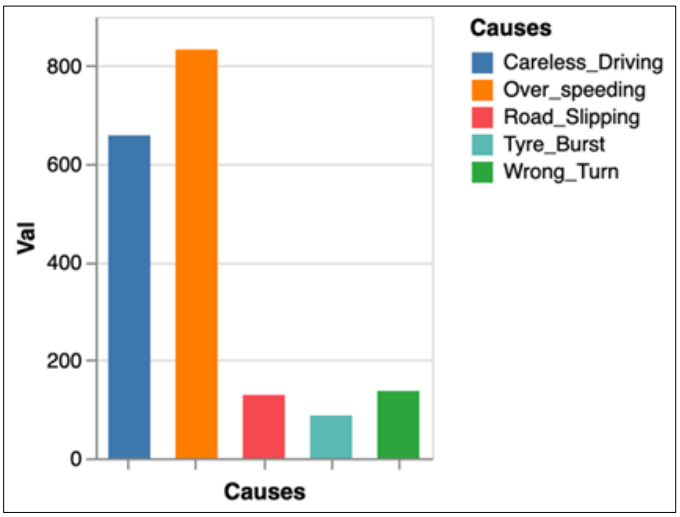

Figure 9 No. of accidents and their major causes

4.7.1 Association Rule Mining

After the preprocessing and transforming of data based on the selected attributes, the association rule mining is applied for generating the best rules through Apriori algorithm. This allows us to see the hidden patterns and relationship of different attributes between each other. In association rule mining support & confidence are two exciting measures which are used for evaluating the quality of rules.

Support (Sp)

Support represents the probability of the occurrence of event A and B simultaneously i.e. A → B. Support is also known as frequency constraint. The frequent item sets are used to generate some association rules. To generate the support value the following equation can be used.

$$S_p = \frac{P(A \cap B)}{N} \text{ ..... (1)}$$

(Where, N is total number of accidents)

Confidence (C_f)

Confidence is the probability of occurrence of A and B together to the occurrence of A. When the confidence value of A → B is increased, the probability of occurrence of B also increases with occurrence of A. The confidence values are assessed through the following equation.

$$C_f = \frac{P(A \cap B)}{P(A)} \text{ ..... (2)}$$

Lift (L_t)

Lift is confidence divided by the quantity of all instances that are covered by the consequence. The predicting values can be calculated using the following equation.

$$L_t = \frac{P(A \cap B)}{P(A)P(B)} \text{ ..... (3)}$$

The value of lift is the probability ratio of the occurrence of confidence and expected confidence of rule. Expected confidence is that the occurrence of A and B together with existence of B. The lift value greater than one gives positive association between A and B. The lift value equals to one show that there is no association between them and when the lift value is less than 1, it reflects a negative association of event A and B. Lift values which are greater than one is the most beneficial for us and are used in our calculations given in the Rules Table [2].

For a better understanding of support, confidence, and lift, consider the following short example in Table. 1. From road accident data in which the set of items is I= {Location Chowk, Overspeed, Clear Weather, Time Afternoon, Severe Accident}. In Table. 1, the value 1 shows the presence of the item and a zero indicates the absence of the item. Rule example is {Chowk, Overspeed} → {Severe Accident} specifies that a severe accident is possible when over-speeding in a chowk.

Table 1 Example dataset

Chowk	Overspeed	Clear	Afternoon	Severe Accident
1	1	1	0	1
1	1	1	1	0
0	0	1	1	0
0	0	0	0	0
1	1	1	1	1

Support value is provided to select the interesting rules. Support and confidence value of above item sets is computed below by using Eq. (1) and (2).

$$S_p = \frac{P(Chowk \cap \text{Overspeed} \cap \text{SeverAccident})}{N}$$

$$S_p = \frac{2}{5} = 0.4$$

A support value of 0.4 indicates that in 40 % of accident records, chowk, over-speed, and severe accident occur together.

$$C_f = \frac{P(Chowk \cap \text{Overspeed} \cap \text{SeverAccident})}{P(Chowk \cap \text{Overspeed})}$$

$$C_f = \frac{2/5}{3/5} = \frac{0.4}{0.6} = 0.66$$

The confidence value of 0.66 means that in 66 % of accident cases when cause is overspeed and location is chowk occur together, then Severe Accident also occurs. This value indicates that there are 66% chances of Severe accident if the location is chowk and vehicle, is over-speeding. The lift value calculated by using Eq. (3).

$$L_t = \frac{P(Chowk \cap \text{Overspeed} \cap \text{SevereAcc})}{P(Chowk) \cap \text{P(Overspeed)} \cap \text{P(SevereAcc)}}$$

$$L_t = \frac{0.4}{0.6x0.6x0.4} = 2.77$$

The lift result shows that chowk location, over-speeding and severe accidents are strongly correlated with each other. This rule is useful to predict future accident with good accuracy.

5. Discussion

As evident in the Table 2, we identified that the maximum accidents are between a car and a bike driver; especially in the afternoon when there are no light issues. In many cases it is evident that, this happens because of violating the over speeding rule from a bike driver. Such accidents especially happened at the chowk location with moderate injuries. The possibility to occur such an accident increases during the middle of the week.

Table 2 Apriori Algorithm Extracted Rules

Rule	Confidence	Lift
Time_of_Accident=After_Noon Vehicle_Opponent=Car → Vehicle_Type =Bike	0.91	1.4
Vehicle_Opponent=Car →Vehicle_Type=Bike	0.86	1.33
Days=3 →Injuries=Moderate	0.8	1.03
LocationsRoundOf=Chowk Vehicle_Type=Bike →Injuries=Moderate	0.8	1.03
LocationsRoundOf=Chowk→Injuries=Moderate	0.79	1.02
Weather_Condition=Clear Vehicle_Type=Bike →Injuries=Moderate	0.79	1.02
Causes=Over_speeding →Injuries=Moderate	0.78	1.1
Vehicle_Type=Bike→Injuries=Moderate	0.78	1.02
Causes=Over_speeding Time_of_Accident=After_Noon →Vehicle_Type=Bike	0.74	1.14
Time_of_Accident=After_Noon → Vehicle_Type=Bike	0.72	1.1
Time_of_Accident=After_Noon → Vehicle_Type=Bike	0.72	1.1
Time_of_Accident=Evening →Vehicle_Type=Bike	0.71	1.09
Months_Quarter=Q3 → Weather_Condition=Clear	0.69	1.15
Time_of_Accident=After_Noon Locations Round Of=Chowk Weather_Condition →Vehicle_Type=Bike	0.69	1.07
Causes=Careless_Driving → Vehicle_Type=Bike	0.69	1.06
Causes=Over_speeding → Vehicle_Type=Bike	0.68	1.05
Causes=Over_speeding Weather_Condition=Clear →Vehicle_Type=Bike	0.68	1.05
Days=1 → Vehicle_Type=Bike	0.66	1.02
Time_of_Accident=Morning →Vehicle_Type=Bike	0.65	1.01

6. Conclusion

Traffic accidents is a global issue effecting every country around the world. The situation is even more critical in developing and middle-to-low-income countries such as Pakistan. The main purpose of this research is to analyze road accidents data and to identify factors that majorly contribute to such events. Association Apriori algorithm is used on road accidents data collected from Traffic Police Rescue office located at Gujrat city, Pakistan. Through association rule mining using Apriori algorithm, we were able to find hidden patterns and by extracting rules, we were able to find the relationship of attributes.

After analysis of this data, we identified that bikes committing over-speeding frequently face accidents. The situation is severe on the chowks during the middle of the week especially in afternoon and in many cases, the opponent is a car.

The results obtained from our analysis correspond and align with other available research work, reflecting that some situations are more likely to experience traffic accidents than others. By following our suggested framework and model, local authorities can analyze core causes of road accidents and implement laws that will contribute to decrease such incidents. In future, we planned to identify different clusters with respect to the three sizes of the vehicles, such as small, medium, and large to find causes of accidents against each category. We are also planned to differentiate the geographical locations into different small regions and to target them individually.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare no conflict of interest.

References

[1] World Health Organization(WHO), Global Status Report on Road Safety [Internet], 2018. https://www.who.int/publications/i/item/9789241565684

[2] Tsala, Simon Armand Zogo, Merlin Zacharie Ayissi, Gerald Azeh, Pierre Anicet Noah, Fabien Betene Ebanda, and Louis Max Ayina Ohandja. "An in-depth analysis of the causes of road accidents in developing countries: case study of Douala-Dschang Highway in Cameroon." Journal of transportation technologies, 2021, 11, no. 3, 455-470.

[3] Abegaz, Teferi, and Samson Gebremedhin. "Magnitude of road traffic accident related injuries and fatalities in Ethiopia." 2019, PloS one 14, no. 1: e0202240.

[4] Jackson, Lydia, and Richard Cracknell. "Road accident casualties in Britain and the world, House of Commons Library Briefing Paper CBP-7615." (2018).

[5] Lu, Juan, Bin Li, He Li, and Abdo Al-Barakani. "Expansion of city scale, traffic modes, traffic congestion, and air pollution." 2021, Cities 108: 102974.

[6] Kumar, Sachin, Durga Toshniwal, and Manoranjan Parida. "A comparative analysis of heterogeneity in road accident data using data mining techniques." 2017, Evolving systems 8, no. 2: 147-155.

[7] Kumar, Sachin, and Durga Toshniwal. "A data mining framework to analyze road accident data." 2015, Journal of Big Data 2.1: 1-18.

[8] Li, Liling, Sharad Shrestha, and Gongzhu Hu. "Analysis of road traffic fatal accidents using data mining techniques." In 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), pp. 363-370. IEEE, 2017.

[9] A. A. El Tayeb, V. Pareek, and A. Araar, "Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai," Int. J. Soft Comput. 2015, Eng., no. 4, pp. 2231–2307.

[10] J. Xi, Z. Zhao, W. Li, and Q. Wang, "A Traffic Accident Causation Analysis Method Based on AHP-Apriori," 2016, Procedia Eng., vol. 137, pp. 680–687.

[11] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," J. Mod. Transp., 2016, vol. 24, no. 1, pp. 62–72.

[12] Priya, S., and R. Agalya. "Association rule mining approach to analyze road accident data." In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-5. IEEE, 2018.

[13] Moradkhani, Farzaneh, Somayya Ebrahimkhani, and B. Sadeghi Begham. "Road accident data analysis: A data mining approach." Indian J. Sci. Res 2014, 3: 437-443.

[14] Lian, Yanqi, et al. "Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles." Accident Analysis & Prevention 2020, 146: 105711.

[15] Castro, Yuri, and Young Jin Kim. "Data mining on road safety: factor assessment on vehicle accidents using classification models." International journal of crashworthiness, 2016, 21.2: 104-111.

[16] Frank, Eibe, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H. Witten, and Len Trigg. "Weka-a machine learning workbench for data mining." In Data mining and knowledge discovery handbook, pp. 1269-1277. Springer, Boston, MA, 2009.

[17] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 2009, Volume 11, Issue 1.

[18] Shahin, Mahtab, Sijo Arakkal Peious, Rahul Sharma, Minakshi Kaushik, Sadok Ben Yahia, Syed Attique Shah, and Dirk Draheim. "Big data analytics in association rule mining: A systematic literature review." In 2021 the 3rd International Conference on Big Data Engineering and Technology (BDET), 2021, pp. 40-49.