International Journal of Engineering & Technology, 7 (3.10) (2018) 40-44

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analysis of Road Accidents Using Data Mining Techniques

Ms. Nidhi. R[1], Ms. Kanchana V[2]

[1,2]*Department of Computer Science*
[1,2]*Amrita School Of Arts and Science,*
[1,2]*Amrita Vishwa Vidyapeetham,*
Mysuru, India
[*]*Corresponding author: E-mail: Nidhigowri23@gmail.com*[1]

Abstract

Road Accident is an all-inclusive disaster with consistently raising pattern. In India according to Indian road safety campaign every minute there is a road accident and almost 17 people die per hour in road accidents. There are different categories of vehicle accidents like rear end, head on and rollover accidents. The state recorded police reports or FIR's are the documents which contains the information about the accidents. The incident may be self-reported by the people or recorded by the state police. In this paper the frequent patterns of road accidents is been predicted using Apriori and Naïve Bayesian techniques. This pattern will help the government or NGOs to improve the safety and take preventive measures in the roads that have major accident zones.

Keywords— *Apriori, Naïve Bayes, Pattern Prediction, Road Accident***.**

1. Introduction

At present, the street movement security is a one of the genuine social issues inside the entire world. For each demise on streets there are an anticipated 4 for all time debilitating damages, for example, harm to the mind or spinal line, 10 genuine wounds and 60 minor wounds. These disturbing numbers have incited the Indian Commission to make a move at national level to diminish mishaps on the streets in the years 2011-2018. Important help in this circumstance states to a noticeable proof of the key elements causing street/car crashes. Use of practical information mining strategies on the gathered datasets exhibiting distinctive circumstances on the streets and happened mischances can help comprehend the most striking components. The accomplishment of such investigation depends clearly on the nature of the information accessible for the trials, e.g. not just information depicting the parameters of the mishap, however data details related to climate conditions or street qualities as well. Gotten brings about the type of prescient models or produced standards can help applicable chiefs to recognize the most risky places as far as street movement, to make and position vital activities to enhance the street security and to plan some broad street movement wellbeing approaches on nearby or national level. Breaking down, deciphering and making greatest utilization of the information is a troublesome and asset requesting undertaking because of the exponential development of numerous organizations, administrative and logical databases.

Proposed framework can be actualized as an ongoing application. This subject can be executed as government segment application. Current framework can be utilized by general society to know the mischances designs and the sort of the mishap in a specific territory or city.

Inspiration aimed at the introduced work was to recognize conceivable concealed relationships and associations between several components depicting happened street mischances with mortal significances. Comparable separated learning on account of direct collaboration with every significant partner as police, state and nearby administration can enhance the street security in India as well. Current framework is manual where administration division make utilization of record information and dissects the information physically, in view of the investigation they will take the precaution measures to lessen the quantity of mischances. We additionally get many devices and programming to keep up street mishaps, these devices just gathers the information stores in separate however no investigation is finished. Current framework is a manual procedure, tedious, costly, absence of information revelation and less productive. The incentive for the proposed work is to identify probable unseen relationships and linkages between various factors, describe the occurrence of traffic accidents and have fatal consequences. In the case of direct cooperation with all most important stakeholders including the police department, state government and central governments, similar knowledge extraction can help improve road safety. The paper consists of four main components: the beginning section explains the present situation and our inspiration for analyzing data on road traffic accidents; the second presents the survey data set and performs pre-processing operations to prepare mining data; applying selected methods and assessments.

Data mining derives up with a set of tools and methods which when applied to this managed data, make available knowledge to long distance riders or travelers for making appropriate decisions during travel.

2. Related Works

Using appropriate system methods to improve safety measures is an remarkable challenge, particularly when the sum of existing records often increases. From the present circumstance, you can choose some existing studies that mainly address local traffic accident data. Iraq is one of the nations with high traffic accident deaths and casualties. In the past 3 years, traffic accidents in Iraq resulted in an average of 24,000 deaths per day (three people per hour), and about 240,000 people were injured every year [1], [2]. This fact prompted a group of writers to classify the most noteworthy affecting the severity of driver injury in these road traffic accidents [2]. They used accident details from the accounts of the Ministry of Information and Technology of the Iraq Traffic Police Department from 2006 to 2008. The target is severity level three: No injuries, injuries and deaths; the record includes more than 169,000 drivers. The choice factor is based on the adaptable importance percentage (VIM) of individual of the outputs of the CART (Classification and Regression Tree) method. The outcomes show that the seat belt is the greatest significant factor related to the harshness of traffic mishap injury, and not using it will considerably increase the possibility of injury or death.

The authors Chang and Wang [3] used related methods to generate CART models to determine the association between injury severity and motorist/motor vehicle features, path/surrounding variables, and collision variables. Using the 2001 collision data from Taiwan (Taipei), they identified the kind of vehicle as the maximum important adaptable in the severity of the collision. In addition, statistics from central Taipei was used in the study by the author Yau-Ren Shiau et al. [4] First determine the best key factors affecting more than 2400 traffic road accidents in 2011, and then use the fuzzy robust principal module analysis, back propagation neural network and Logistic regression mining methods to create the predictable classification model. By combining the first two methods mentioned above, the best accuracy rate is 85.89%.

Nayak et al. [5] analyzed Queensland Department of Transportation and Australia's major road traffic accidents and road data, which contained more than 42,000 records from 2004 to 2007. The author uses a taxonomy that breaks it down into numerous stages of crash tendencies (certain streets, designs, or conditions, which have greater collision proportions than others) and inspires collective effort by one more writer [6]. For experiments, the authors used a chi-squared test tree, regression trees using f-tests, neural networks, logistic regression, and Bayesian models. Decision trees show better performance than other test simulations.

Alternative study is devoted to road traffic safety in UAE. In the UAE, about six hundred people are killed each year in car mishaps; road traffic accidents are the 2nd leading cause of death [7]. The author used additional than 1 800 000 records between 2008 and 2010, of which 19 attributes covered accidents, motorists, and path/means of transportation conditions. The target characteristic contains 5 values: disastrous, severe, modest, slight, and none. The ultimate analytical model was generated in the WEKA support tool and the subsequent algorithm: PART, Bayesian network, J48 decision tree, and multilayer perceptron. The most accurate model was created by a multi-layer perceptron (more than 99%), and the Bayesian network established the fastest model (0.17 seconds).

The authors of the Hong Kong government transportation department collectively applied similar algorithms [8]. This data set contains additional 34,000 records; genetic algorithms are implemented for feature selection; classification experiments are also conducted in the WEKA mine kit, and J48 provides a more precise classification model to estimate the severity of traffic accident injuries.

The basic values for refining road safety in the Andalucía region of Spain through data mining techniques are defined in [9]. The complete procedure is based on what is known as the susceptibility improvement element, which is defined as a section on the road that shows different road conditions than best road safety criterions (eg, layout, signals, crossings, tunnels, etc.). A selection of data mining methods such as decision trees, neural networks or association rules will be implemented to integrate datasets of 3 original datasets: element, roads, and collapses.

Lastly, the work of Flach et al. [10] was defined since the authors used the same data set as required by the Hampshire National Council in the UK and wanted to understand the profile of road safety development over the past 20 years. The study was conducted by seven methodical teams who applied a series of data mining techniques, including time series clustering, text mining , multi-relation data mining, subgroup discovery, and association rule learning. Some of the outcomes obtained can be used as an sample: Suppose that if the speed limit is 60 miles per hour, which occurs after 8 p.m. and the vehicle has 2 wheels, the accident will have fatal consequences. This is 70,000 records of 100,000 records. Confirmed. In addition, the investigation of the association between road accidents and the hour or day of the week resulted in the following: Most of the accidents occurred around 8 am and 4 pm to 5 pm.

All the proposed studies have the following commonalities: they use limited data samples for selected time periods or local areas - in view of the local features can be considered positive, but on the other hand some major universal issues can be left; In many cases, traditional data mining algorithms are apparently used as decision trees because they are easy to understand and interpret; almost all studies produce highly accurate models, but this value strongly depends on the feature range method used. The experiments described in this article involve large amounts of data that require new methodologies and supportive environments. As a result, the in-memory method [11] expressed in the platform H2O and R languages was chooses as a proper grouping for a particular goal.

3. Methodology

In order to predict the pattern of new road accident, an association and classification data mining technique are used that is, Apriori and Naïve Bayes classifier, which are highly scalable. Even if we are working on a data set with millions of records with some attributes, this classifier can yield best results. There are models that assign class labels to problem instances, which are represented as vectors of feature values, and the class labels are drawn from some finite set. The data is collected from police stations which are restricted to an area. The below figure represents the architecture diagram for predicting the road accidents where a data repository is created based on the data collected from different police stations. Based on this uploaded data, system predicts the patterns between road accidents.

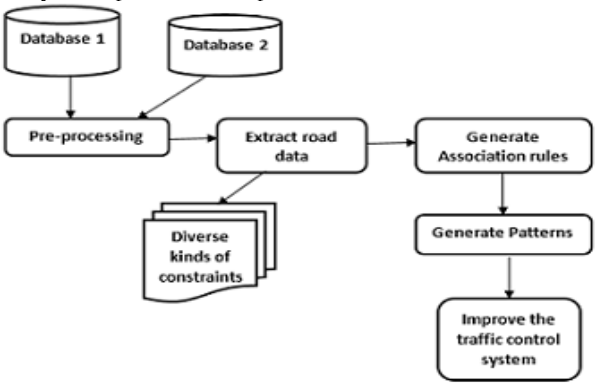

Figure 1-Architecture diagram of the system

The proposed methodology uses naïve Bayes classification technique to predict the type of road accident in a newly

constructed road. It contains 61 constraints, based on which the road accident prediction is done. They are Age, vehicle type, weather conditions, day or night time, road surface, male or female, speed, school zone, highway, Blood number of riders, light conditions, junction details, age of vehicle, driving experience, engine capacity, vehicle reference number are some of the constraints

Once the data warehouse is developed based on the constraints on road wise, city wise, date wise, month wise etc. are used to find the association rule using the Apriori algorithm. A sample of the rules is shown in the below figure.

LHS	->	RHS	CONFIDENCE
collision	->	Sporting accident	66.67%
Sporting accident	->	Hit-run	100%
Hit-run	->	collision	33.3%
over speed	->	Loss of control	25.2%
Lack of vision	->	glaring	66.67%
Hitting barrier	->	collision	66.67%

Figure 2-Association rules results

Based on the association rules a prediction model is developed to predict the frequently occurring accident types. Naiye Bayes algorithm is used where input is the association rules for a selected road and predicts the accident type for a new road.

4. Results and Discussion

Based on the data the following patterns are generated using Naiye Bayes algorithm.

1. Urban Versus Rural

Analysis of urban and rural road traffic accidents shows that rural areas are more prone to traffic accidents. Compared with the number of accidents in rural areas, the total number of urban road accidents is small. The table shows that rural roads require substantial investment and improvement to reduce accidents in rural areas.

Table 1 - Number of people killed/injured in urban and rural areas

Category	Accident	Killed	Injured
Rural	403598	121126	2589
Urban	15461	4091	890

The below figure shows the graphical representation for the above table.

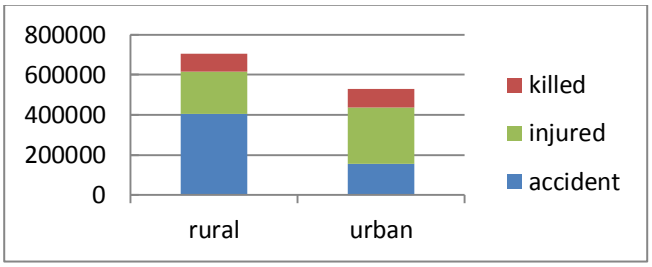

Figure 3 x- axis: accidents,y -axis:person killed on road type

2. Accidents at Traffic/ Police Controlled Areas

It can be seen that the largest number of accidents occurred in uncontrolled areas, causing 128,263 accidents in traffic control/police control areas, and the number of reported accidents was 1,66,158.

Table 2 - Accidents in controlled areas

	Accident	killed	Injured
Traffic signal	15125	4322	12995
Police controlled	11386	3076	11761
Stop sign	11221	3609	11002
Blinker	10009	3012	10138
Uncontrolled	128263	40010	124491

The below figure shows the pie chart of the accidents in the urban as well as rural area where the road is controlled by traffic or police.

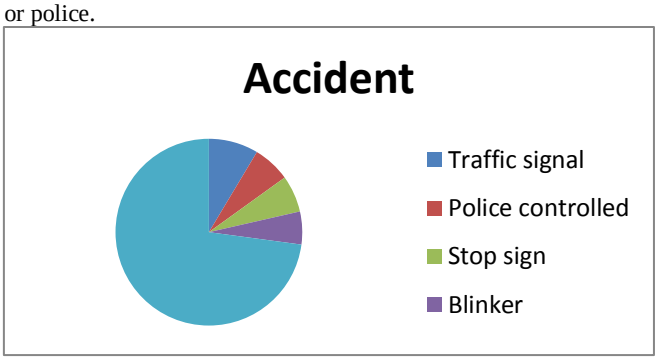

Figure 4: Accidents in controlled area

3 Accidents based on the Age of Vehicles

The below table and the corresponding figure shows the data regarding the age of vehicles involved in the accidents. During the year, fewer than five-year-old vehicles had the highest number of accidents in the country (3,94,198), of which 56,329 were dead and 2,030,042 were injured.

Table 3-Accident based on age of vehicles

Age of vehicle	Accidents	killed	injured
Less than 5 years	394198	56329	203042
5-10 years	157370	49536	160642
10-15 years	74149	23775	72982
15 years & above	45358	17073	47391
Age not known	10598	3921	10238

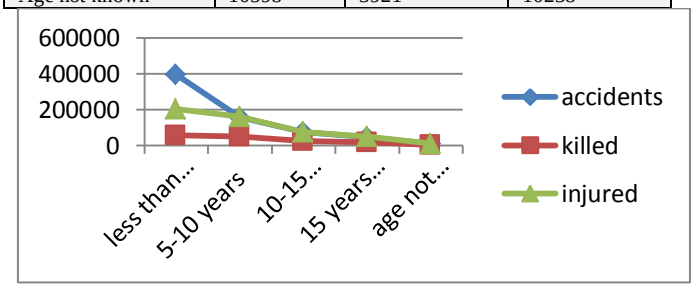

Figure 4 : Graphical representation of age of vehicles

4 Persons killed in Road Accidents in terms of Road User Categories

The road is used by two categories namely vulnerable road users who are largely unprotected like pedestrian bicycle riders and two wheeler riders. The second category are the drivers.

Table 4-Accident on based on type of vehicle

Road Users	No of persons killed
Pedestrian	15746
Bicycles	2585
Two wheelers	52500
Auto rickshaws	7150
Cars, taxis, vans, bus	26923

Figure 5: Causality in type of road users

The above table and the figure shows that the major causality is affected to the vulnerable road users since they are the one who are 0more exposed. They have the highest tendency towards accidents

5 Age of Persons Killed (Gender wise) in Road Accidents

In the deaths of road traffic accidents,the male and female genders clearly showed that the total number of men and women killed during the year was 1,27,435 and 23,332 respectively.

Table 5: Accident on person's age groups

Age Group	Male	Female
Less than 18	8347	2275
18-25	27417	4358
25-35	32609	5467
35-45	28564	4994
45-60	18592	3582
60 & above	6964	1850
Age not known	4960	806
total	127453	23332

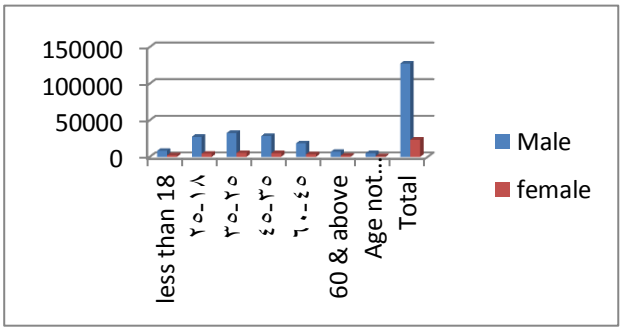

Figure 6 : Chart for the accident on person's age groups

The trend in accidents revel that young people in the age of 20 to 35 males disobey the traffic rules and hence prone to huge number of causalities in the road accidents.

5. Conclusion

From the statistical results, it can be seen that the rural mortality rate is higher, while the city is lower. Statistical analysis also includes other limiting factors such as the age of the vehicle, the type of vehicle, the age group of the person, and the category of road users. The predicted data results are displayed in a graphical representation. Graphical representations help the public understand accident metrics that help reduce mortality.

Acknowledgement

First and foremost, we feel deeply indebted to Her Holiness Most Revered Mata Amritanandamayi Devi(Amma) for her inspiration and guidance both in unseen and unconcealed ways.
Whole heartedly, we thank our college, Amrita School of Arts and Sciences, Mysuru campus, Karnataka, India, for providing the necessary environment, infrastructure, encouragement and for extending the support possible at each stage of project.
We express our sincere gratitude and indebtedness to our parents who have bestowed there great guidance at appropriate times by providing encouragement in planning and carrying out the project

References

[1] F.M.O.I. Forensic Medicine Organization of Iran; Statistical Data, Accidents, online avail-able on: http://www.lmo.ir/?siteid=1&pageid=1347
[2] A.T. Kashani et al., "A Data Mining Approach to Identify Key Factors of Traffic Injury Severity", PROMETTraffic& Transportation, 23(1), pp. 11-17, 2011.
[3] L.Y. Chang, H.W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques", Accident Analysis and Prevention, 38(5), pp. 1019-1027, 2006.
[4] S. Yau-Ren et al. "The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents", Mathematical Problems in Engineering, Volume 2015 (2015), pp. 1-8., 2015. F. Babi and K. Zuská ová • Descriptive and Predictive Mining on Road Accidents Data– 92
[5] R. Nayak et al., "Road Crash Proneness Prediction using Data Mining". Ailamaki, Anastasia & Amer-Yahia , Sihem (Eds.) Proceedings of the 14th International Conference on Extending Database Technology, Association for Computing Machinery (ACM), Upp-sala, Sweden, pp. 521-526, 2011.
[6] V. Shankar, J. Milton, F. Mannering, "Modeling accident frequencies as zero-altered probability processes: An empirical inquiry", Accident Analysis & Prevention, 29(6), pp. 829-837, 1997.
[7] A Araar et al., "Mining road traffic accident data to improve safety in Dubai", Journal of Theoretical and Applied Information Technology, 47(3), pp. 911-927, 2013.
[8] S. Vigneswaran et al., "Efficient Analysis of Traffic Accident Using Mining Techniques", International Journal of Software and Hardware Research in Engineering, Vol. 2, No. 3, 2014, pp. 110-118, 2014.
[9] L. Martin et al. "Using data mining techniques to road safety improvement in Spanish roads", XI Congreso de Ingeniería del Transporte (CIT 2014), Procedia - Social and Behavioral Sciences 160 (2014), pp. 607–614, 2014.
[10] P. Flach et al., "On the road to knowledge: Mining 21 years of UK traffic accident reports", Data Mining and Decision Support: Aspects of Integration and Collaboration, Springer, pp. 143-155, 2003.
[11] H. Zhang et al., "In-Memory Big Data Management and Processing: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 7, pp. 1920–1948, 2015.
[12] J. Hipp, U. Güntzer, G. Nakhaeizadeh, "Algorithms for Association Rule Mining &Mdash; a General Survey and Comparison", SIGKDD Explor Newsl 2, pp. 58–64, 2000.
[13] A.T. Kashani et al., "A Data Mining Approach to Identify Key Factors of Traffic Injury Severity", PROMETTraffic& Transportation, 23(1), pp. 11-17, 2011.
[14] P.J. Ossenbruggen, J. Pendharkar et al., "Roadway safety in rural and small urbanized areas", Accidents Analysis & Prevention, 33(4), pp. 485-498, 2001.
[15] R. Agrawal, T. Imieli ski, A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, pp. 207–216, 1993.
[16] [16]R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Data-bases", Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 487-499, 1994.
[17] L. Breiman, "Random Forests", Machine Learning, Vol. 45, pp. 5-32, 2001.

[18] T. Calders et al., "Machine Learning and Knowledge Discovery in Databases", Part II. Nancy, France, Springer, pp. 203-232, 2014.

[19] EC: Mobility and Transport, Road Safety, Statistics – accidents data, online available on: http://ec.europa.eu/transport/road_safety/specialist/statistics/index_en.htm

[20] J. H. Friedman, "Stochastic gradient boosting", Computational Statistics & Data Analysis - Nonlinear methods and data mining, 38(4), pp. 367-378, 2002

[21] Ahalya, C. S., Abin, K. O., & Kanchana, V. (2017, May). Building up an information archive for putting away pesticide data. In Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference on (pp. 2125-2128). IEEE.

[22] Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In Data Mining and Advanced Computing (SAPIENCE), International Conference on (pp. 91-95). IEEE.