



A ROAD ACCIDENT PREDICTION MODEL USING DATA MINING TECHNIQUES

**K. VENKATESWARLU¹, P BHAVANI², V VARUN REDDY³, N SATHISH⁴,
GOLUSULA RAMA DEVI⁵**

¹Assistant Professor, Department of CSE, Malla Reddy College of Engineering
Hyderabad, TS, India.

^{2,3,4,5}UG students, Department of ITE, Malla Reddy College of Engineering
Hyderabad, TS, India.

ABSTRACT

Due to the exponentially increasing number of vehicles on the road, the number of accidents occurring on a daily basis is also increasing at an alarming rate. With the high number of traffic incidents and deaths these days, the ability to forecast the number of traffic accidents over a given time is important for the transportation department to make scientific decisions. In this scenario, it will be good to analyze the occurrence of accidents so that this can be further used to help us in coming up with techniques to reduce them. Even though uncertainty is a characteristic trait of majority of the accidents, over a period of time, there is a level of regularity that is perceived on observing the accidents occurring in a particular area. This regularity can be made use of in making well informed predictions on accident occurrences in an area and developing accident prediction models. In this paper, we have studied the inter relationships between road accidents,

condition of a road and the role of environmental factors in the occurrence of an accident. We have made use of data mining techniques in developing an accident prediction model using Apriori algorithm and Support Vector Machines. Bangalore road accident datasets for the years 2014 to 2017 available in the internet have been made use for this study. The results from this study can be advantageously used by several stakeholders including and not limited to the government public work departments, contractors and other automobile industries in better designing roads and vehicles based on the estimates obtained.

INTRODUCTION

The alarming rate of increase of accidents in India is now a cause for serious concern. According to some recent statistics [1], India accounts for roughly six percent of global road accidents while owning only one percent of the global vehicle population. There



are a lot of accident cases reported due to the negligence of two-wheelers, whereas over-speeding is also another contributing factor. Accidents caused while under the influence of alcohol or during general traffic violations are also common. In spite of having set regulations and the highway codes, the negligence of people towards the speed of the vehicle, the vehicle condition and their own negligence of not wearing helmets has caused a lot of accidents. While the major cause of road accidents is attributed to the increasing number of vehicles, the role played by the condition of the roads and other environmental factors cannot be overlooked.

The number of deaths due to road accidents in India is indeed a cause for worry. The scenario is very dismal with more than 137,000 people succumbing to injuries from road accidents. This figure is more than four times the annual death toll from terrorism. Accidents involving heavy goods vehicles like trucks and even those involving commercial vehicles used for public transportation like buses are some of the most fatal kind of accidents that occur, claiming the lives of innocent people. Weather conditions

like rain, fog, etc., also play a role in catalysing the risk of accidents. Thus, having a proper estimation of accidents and knowledge of accident hotspots and causing factors will help in taking steps to reduce them. This requires a keen study on accidents and development of accident prediction models.

To implement a well designed road framework management system for looking into road security aspects, it is often desired to have an optimized accident prediction model which can analyze potential issues arising due to infrastructure fallbacks and to estimate the effect of existing models in reducing the occurrence of accidents. The main challenge involved in the creation of such a model include the evaluation of the weight that can be attributed to the impact of each variable in contributing to the accident and assessing how the model can be best designed to incorporate the effect of all such variables. Data mining techniques and models have in the past been found useful for the purpose of data interpretation in a variety of domains including but not limited to credit risk management, fraud detection, healthcare informatics, recommendation systems and so on. Approaches involving artificial intelligence and machine



learning have further helped to augment these studies. For this paper, we have investigated the inter-relationship between the occurrences of road accidents and the roles played by the underlying road conditions and environmental factors in contributing to the same. Since such a study requires us to cover several aspects affecting accidents, we can make use of data mining techniques to analyze this data to extract relevant details from them, as these huge volumes of data would otherwise be meaningless without the right interpretation applied to them.

In this paper, we are discussing the effects of such an accident prediction model in identifying the risks involved in road accident scenarios. The next section discusses the prior works done with respect to analyzing the different accidents that have taken place over the years. This is followed by a summarized description of the methodology used in this work. Further, the different components of implementation including the system architecture, software and languages used, simulation, user interface and screenshots of the developed application are discussed. Finally, the discussion and conclusions derived from the

present study and the future scopes are outlined in the last two sections. The results from this study have been used to propose a model that can be used as a tool to estimate the possibility of road accidents in a particular area chosen by the user.

EXISTING SYSTEM

Williams et al. [5] have found through their studies that the age and experience of a driver also play a major role in the occurrence of accidents. Suganya, E. and S. Vijayarani [6] in their paper have analysed the road accidents in India and compared the performance of different classification algorithms such as linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm using accuracy, error rate and execution time as a measure of performance. They have found the performance of KNN to be better than that of the others.

Sarkar et al. [7] have done a comparative study on the type of roads that are prominent in accidents. While exploring the other components associated with accidents, they have found that the occurrence of accidents in highways is more common than in a normal road similar to [4]. Stewart et al. [8] have utilized original data in



building a neural network model to predict accidents. They found that this model was able to give quicker results than those being used in the models built on Indian roads.

Zheng et al. [9] have studied the range of injuries that come forth in a motor vehicle accident and have also analyzed the emotions of the drivers involved in the accidents that could have been a causal factor. Arun Prasath N and Muthusamy.

Punithavalli [10] have conducted an extensive survey on the different techniques used in road accident detection over the years, the approaches implemented in them and discusses their merits and de-merits.

George Yannis et al. [11], in their paper, have discussed about the current practices used in the development of accident prediction models on an international level. Detailed information on various models have been collected with the help of questionnaires and they have made use of this data to identify which could be the most useful model that can be applied for accident prediction.

Anand, J. V [12] has developed a method to determine the effect of different variables in the detection and

prediction of atmospheric deterioration all over the world. Fuzzy C means clustering, R-studio, and the ARIMA frame work have been made use of in creating this method. A similar approach can also be tried in evaluating the impact of various factors on road accidents. Analyzing the original cause of accidents is important because this will tell us the impact factor and contribution of each attribute towards road accidents. Tiwari et al. [13] have made use of self-organizing maps, K-mode clustering techniques, Support Vector Machines, Naïve Bayes and Decision tree to classify the data from road accidents based on the type of road users.

Disadvantages

- 1) The system doesn't have facility to train and test on large number of numbers.
- 2) The system doesn't measure an accurate road accident due to poor classification models.

PROPOSED SYSTEM

In the proposed system, the system has built an application that is capable of predicting the possibility of occurrence of accidents based on available road accident data. Data pre-processing is done on this road accident data to obtain a dataset. The data preprocessing step includes cleaning to remove the null and



garbage values, and normalization of the data, followed by feature selection, where only relevant features from the original dataset are selected to be included in the final dataset. The dataset is then subjected to different data mining techniques. Clustering is performed on this dataset. The clusters are then subjected to other algorithms like Support Vector Machines (SVM) and Apriori. Since the data being used for the study has an unknown distribution and we need to sort out the frequent and infrequent items in the dataset, the former (SVM) is used to predict the probable risk of accidents while the latter (Apriori) is applied to perform rule mining, that is, to generate a frequent item set based on given support and confidence values.

Rules have been set considering different combinations of factors which have caused accidents of varying nature and severity in different road types and weather conditions. For the frequently occurring item sets, the chosen support and confidence values imply the higher probability of the particular combination of attributes in leading to an accident. For example, based on the rule mining done, the probable risk of an accident occurring even during fine weather in a junction on account of over-speeding is

high and could prove to be fatal based on the training dataset. SVM classification has been used to characterize each accident event into a high or a low risk category. Various data mining techniques and exploratory visualization techniques are applied on the accident dataset to get the interpreted results..

Advantages

- 1) These optimized models can be efficiently utilized by the government to reduce road accidents and to implement policies for road safety.
- 2) The overall model has helped to give an understanding of the combinations of factors that have proven fatal in accident scenarios.

IMPLEMENTATION

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Browse DataSets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View All Road Accident Prediction, Find Road Accident Prediction Type Ratio, View Road Accident Ratio Results, Download Predicted Data Sets, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like PREDICT ROAD ACCIDENT STATUS, VIEW YOUR PROFILE.

Operation:

Raw datasets were terribly filthy, not in a format that computer machines could understand, and provided partial data to use as is. The effectiveness of the crash severity prediction model will be reduced if such datasets are used. As a result, irrelevant datasets should be eliminated in order to generate maximum data. Before designing the model, the researchers was using an expensive data preparation technique to obtain relevant and determinant potential risks, such as cleaning the data, missing value handling, outlier management, dealing with absolute value encoding, and standardization.



Fig.1. Home page.

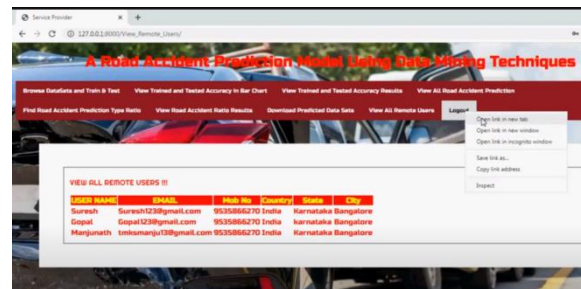


Fig.2. Login details.

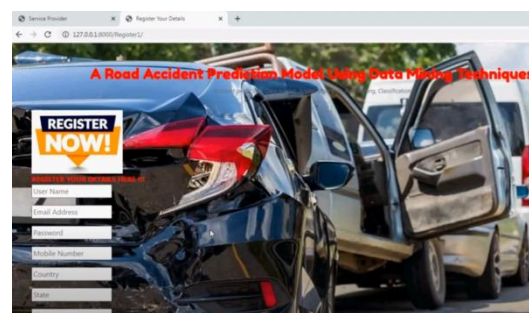


Fig.3. Registration page.

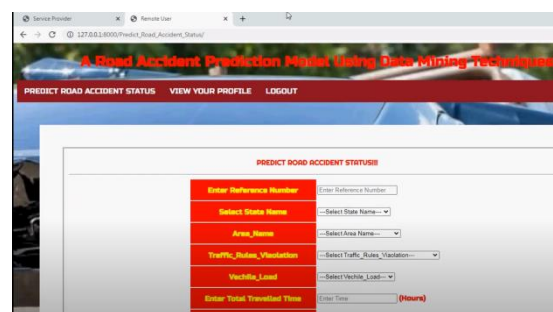


Fig.4. Prediction of dataset.



CONCLUSION

An accident can change the lives of many people. It is up to each of us to bring down this increasing number. This can be made possible by adopting safe driving measures to an extent. Since all instances of accidents cannot be attributed to the same cause, proper precautionary measures will also need to be exercised by the road development authorities in designing the structure of roads as well as by the automobile industries in creating better fatality reducing vehicle models. One thing within our capability is to predict the possibility of an accident based on previous data and observations that can aid such authorities and industries. This project was successful in creating such an application that can help in efficient prediction of road accidents based on factors such as types of vehicles, age of the driver, age of the vehicle, weather condition and road structure, This model was implemented by making use of several data mining and machine learning algorithms applied over a dataset for Bangalore and has been successfully used to predict the risk probability of accidents over different areas with high accuracy. The model can be further optimized in future to include several constraints that have been left

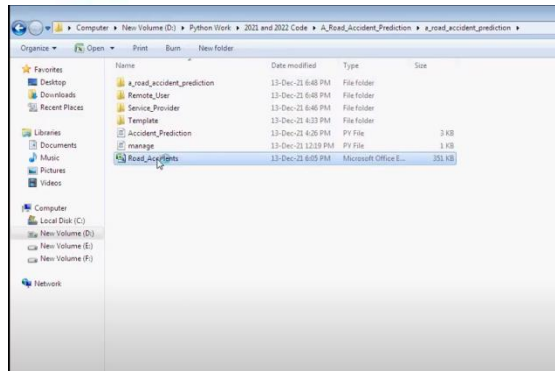
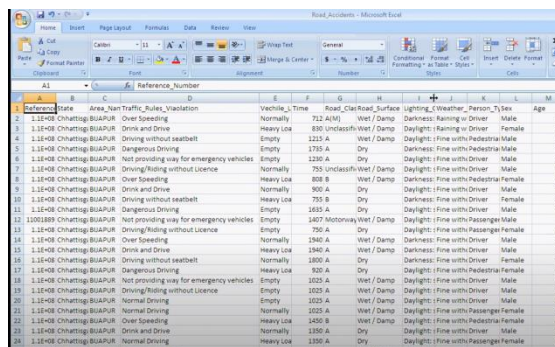


Fig.5. Dataset details.



Reference_Number	Area	Area Name	Traffic_Signs_Violation	Vehicle_Age	Road_CapRoad_Surface	Lighting_Weather	Person_Type	Age	
1	1.1E+08	Chhattisgarh	Over Speeding	Normally	712 A(M)	Wet / Damp	Darkness:Raining w/ Driver	Male	43
2	1.1E+08	Chhattisgarh	Drms and Drive	Heavy Loa	890 Unclassif	Wet / Damp	Daylight: F Rain w/ Driver	Female	35
3	1.1E+08	Chhattisgarh	Driving without seatbelt	Empty	1225 A	Wet / Damp	Daylight: Fine with Pedestria	Male	46
4	1.1E+08	Chhattisgarh	Dangerous Driving	Empty	1795 A	Dry	Darkness: Fine with Pedestria	Male	35
5	1.1E+08	Chhattisgarh	Not providing way for emergency vehicles	Empty	1290 A	Dry	Daylight: Fine with Driver	Male	29
6	1.1E+08	Chhattisgarh	Driving/Riding without licence	Normally	755 Unclassif	Wet / Damp	Daylight: Fine with Driver	Male	25
7	1.1E+08	Chhattisgarh	Over Speeding	Heavy Loa	808 B	Wet / Damp	Darkness: Fine with Pedestria	Female	13
8	1.1E+08	Chhattisgarh	Drms and Drive	Normally	900 A	Dry	Daylight: Fine with Driver	Male	45
9	1.1E+08	Chhattisgarh	Driving without seatbelt	Heavy Loa	755 B	Dry	Darkness: Fine with Driver	Female	32
10	1.1E+08	Chhattisgarh	Dangerous Driving	Empty	1885 A	Dry	Daylight: Fine with Driver	Male	35
11	1.00E+09	Chhattisgarh	Not providing way for emergency vehicles	Empty	1407 Moonrwy	Wet / Damp	Daylight: Fine with Passenger	Male	1
12	1.1E+08	Chhattisgarh	Driving/Riding without licence	Empty	750 A	Dry	Daylight: Fine with Passenger	Female	23
13	1.1E+08	Chhattisgarh	Over Speeding	Normally	2940 A	Wet / Damp	Darkness: Fine with Driver	Male	26
14	1.1E+08	Chhattisgarh	Drms and Drive	Heavy Loa	1940 A	Wet / Damp	Darkness: Fine with Driver	Male	32
15	1.1E+08	Chhattisgarh	Driving without seatbelt	Normally	1800 A	Dry	Darkness: Fine with Driver	Female	22
16	1.1E+08	Chhattisgarh	Dangerous Driving	Heavy Loa	920 A	Dry	Daylight: Fine with Pedestria	Female	11
17	1.1E+08	Chhattisgarh	Not providing way for emergency vehicles	Empty	1025 A	Wet / Damp	Daylight: Fine with Driver	Male	46
18	1.1E+08	Chhattisgarh	Driving/Riding without licence	Empty	1025 A	Wet / Damp	Daylight: Fine with Driver	Male	33
19	1.1E+08	Chhattisgarh	Normal Driving	Normally	1025 A	Wet / Damp	Daylight: Fine with Driver	Male	46
20	1.1E+08	Chhattisgarh	Normal Driving	Normally	1025 A	Wet / Damp	Daylight: Fine with Passenger	Female	36
21	1.1E+08	Chhattisgarh	Over Speeding	Heavy Loa	2450 B	Wet / Damp	Daylight: Fine with Pedestria	Female	41
22	1.1E+08	Chhattisgarh	Drms and Drive	Normally	1195 A	Dry	Daylight: Fine with Driver	Male	45
23	1.1E+08	Chhattisgarh	Normal Driving	Heavy Loa	1195 A	Dry	Daylight: Fine with Passenger	Female	40

Fig.6. Dataset with rows and columns.

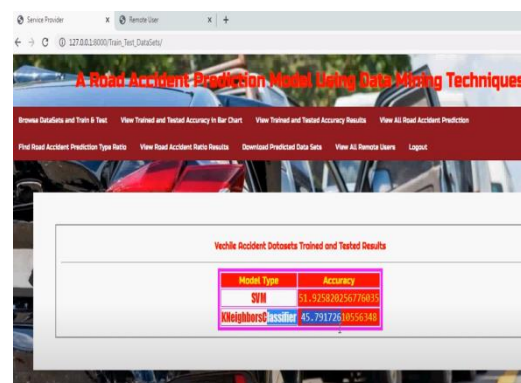


Fig.7. SVM accuracy.



Fig.8. Accuracy of output.



out in the current study. These optimized models can be efficiently utilized by the government to reduce road accidents and to implement policies for road safety. Another scope of this work would be to develop a mobile app that will help the drivers in choosing a route for a ride. A call out to the driver through the maps service can also be implemented that would also announce the risk probability in a chosen route along with the directions. This can then be implemented by service provider companies such as Uber, Ola and so on in future. This will also be useful in having a better surveillance of accident prone areas and providing emergency services in the event of an accident. Better road safety instructions can also be installed along the highways taking into account the risks obtained from this model.

REFERENCES

- [1] <https://www.statista.com/topics/5982/road-accidents-in-india/>
- [2] Srivastava AN, Zane-Ulman B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In Aerospace Conference, IEEE. IEEE 3853-3862.
- [3] Ghazizadeh M, McDonald AD, Lee JD. (2014). Text mining to decipher free-response consumer complaints: Insights from the nhtsa vehicle owner's complaint database. *Human Factors* 56(6): 1189-1203. <http://dx.doi.org/10.1504/IJFCM.2017.089439>.
- [4] Chen ZY, Chen CC. (2015). Identifying the stances of topic persons using a model-based expectationmaximization method. *J. Inf. Sci. Eng* 31(2): 573-595. <http://dx.doi.org/10.1504/IJASM.2015.068609>
- [5] Williams T, Betak J, Findley B. (2016). Text mining analysis of railroad accident investigation reports. In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009- V001T06A009. <http://dx.doi.org/10.14299/ijser.2013.01>.
- [6] Suganya, E. and S. Vijayarani. "Analysis of road accidents in India using data mining classification algorithms." 2017 International Conference on Inventive Computing and Informatics (ICICI) (2017):1122-1126.
- [7] Sarkar S, Pateshwari V, Maiti J. (2017). Predictive model for incident occurrences in steel plant in India. In ICCCNT 2017, IEEE, pp. 1-5. <http://dx.doi.org/10.14299/ijser.2013.01>.
- [8] Stewart M, Liu W, Cardell-Oliver R, Griffin M. (2017). An interactive web-



based toolset for knowledge discovery from short text log data. In International Conference on Advanced Data Mining and Applications. Springer, pp. 853-858. http://dx.doi.org/10.1007/978-3-319-69179-4_61.

[9] Zheng CT, Liu C, Wong HS. (2018). Corpus based topic diffusion for short text clustering. *Neurocomputing* 275: 2444-2458. <http://dx.doi.org/10.1504/IJIT.2018.090859>.

[10] ArunPrasath, N and Muthusamy Punithavalli. "A review on road accident detection using data mining techniques." *International Journal of Advanced Research in Computer Science* 9 (2018): 881-885.

[11] George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Thomas Richter, Stephan Ruhl, Francesca La Torre, Lorenzo Domenichini, Daniel Graham, Niovi Karathodorou, Haojie Li (2016). "Use of accident prediction models in road safety management – an international inquiry". *Transportation Research Procedia* 14, pp. 4257 – 4266.