



## A ROAD ACCIDENT PREDICTION MODEL USING DATA MINING TECHNIQUES

**Mr. N.Eleswara Rao, G.PUJITHA, K.V.V.KARTHIKEYA & V.AKANKSHA**

<sup>1</sup>Assistant Professor, Department of Information Technology, CMR College of Engineering & Technology

<sup>2,3,4</sup>B-Tech, Department of Information Technology, CMR College of Engineering & Technology

### **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.

## OBJECTIVE:

- To study the causes of the accident by features extraction from the images given as the input.
- To understand the severity of the accident based on these features.
- To classify it as fatal, grievous, simple injury or motor collision.
- To carry out the above mentioned algorithms to predict the performance and accuracy of each.
- To conclude the fastest algorithm.
- To understand the effect of each feature on the accident and conclude how much is it responsible for the accident

## IMPLEMENTATION

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.

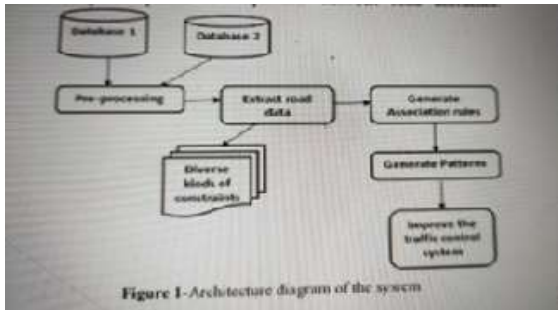
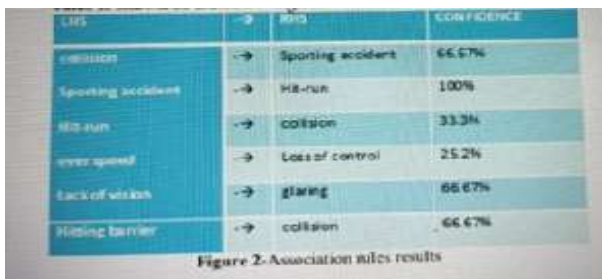


Fig 1 Architecture diagram of system



LHS	RHS	Support/Confidence
collision	→ Sporting accident	66.67%
Sporting accident	→ Hit-run	100%
Hit-run	→ collision	33.33%
over speed	→ Loss of control	25.2%
Lack of wipers	→ glaring	66.67%
Flipping barrier	→ collision	66.67%

Fig 2 Association rules results

## PROPOSED SYSTEM:

### Objective of Proposed Model:

we have 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

### Algorithms Used for Proposed Model

#### Support Vector Machines

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms,

which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

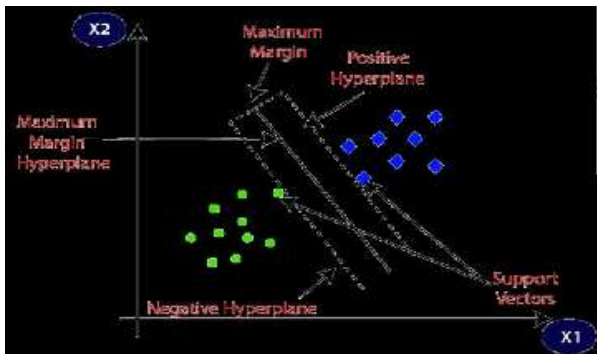


Fig 3 support vector machine

### Advantages of SVM Algorithm:

- Effective in high-dimensional cases.
- Its memory is efficient as it uses a subset of training points in the decision function called support vectors.
- Different kernel functions can be specified for the decision functions and its possible to specify custom kernels.

### Disadvantages of SVM Algorithm:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and

regularization term is crucial.

- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

### KNN-K Nearest Neighbour:-

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems



## RESULTS AND DISCUSSION

### Comparison of Existing Solutions

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. 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.

### Data Collection and Performance Metrics

The user interface of the model based application outputs a graphical visualization of the factors that have been responsible for causing accidents relative to a specified area in the past. Based on this, a categorical prediction as high or low risk relative to accident occurrences is made for an area chosen by the user. The overall model has helped to give an understanding of the combinations of factors that have proven fatal in accident scenarios. A provision to further improve the dataset for future use has also been made in the form of an option to enter details of new accident cases.

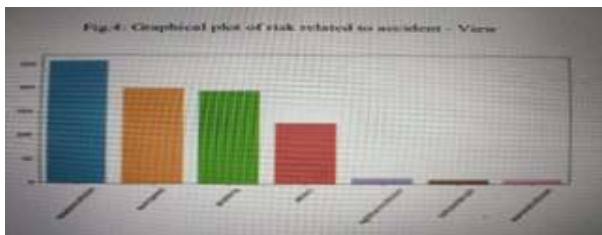


fig 5 graphical plot of risk related to accident

### CONCL

### USION:

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 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.

## 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 expectation maximization 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](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] Soujanya, K. (2018). Ontology based variability management for dynamic reconfiguration of software product lines. *Journal of Advanced Research in Dynamical and Control Systems*, 9(18), 2361-2375.
- [12] Gurumoorthi, E., & Ayyasamy, A. (2020). Cache agent based location aided routing using distance and direction for performance enhancement in VANET. *Telecommunication Systems*, 73(3), 419-432.
- [13] Niranjana, G., Poongodai, A., Soujanya, K.L.S., 2022, Biological inspired self-organized secure autonomous routing protocol and secured data assured routing in WSN: Hybrid EHO and MBO approach,



International Journal of Communication  
Systems, 10.1002/dac.5044

[14] Rajalingam, S., Karuppiah, N.,  
Muthubalaji, S., Shanmugapriyan, J., 2022,  
Power quality improvement in the distribution  
system by interconnecting PV using hybrid  
DSTATCOM, International Journal of  
Advanced Technology and Engineering  
Exploration, 10.19101/IJATEE.2021.875154

[15] Ahmed, M., Laskar, R.H., 2022, Eye  
center localization using gradient and intensity  
information under uncontrolled environment,  
Multimedia Tools and Applications,  
10.1007/s11042-021-11805-z