International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 05 | May -2020 ISSN: 2582-3930

ACCIDENT PREDICTION USING MACHINE LEARNING

Prof. Aemi kalaria[1] ,Abdul Shukoor[2] , Harshal[3], Allu Yagnesh[4]

Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India. 1
Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India. 2,3,4

Abstract- The paper introduces the use of accident prediction modelsfor the identification of hazardous road locations. The application of this method is presentedwith the machine learning method for prediction.The accident prediction model is represented by a linear model which, on the basisof the data available, determines the expected number of accidents for individual types of roads. A critical road segment is defined as a segment where the reported numberof accidents significantly exceeds the number of expected accidents on roads with similargeometric and traffic characteristics. This method can be used as an effective tool for roadnetwork safety management.

KEY WORDS: road safety inspection, hazardous road location, accident prediction model.

I. INTRODUCTION

Road traffic accidents are the major problem in the world.Road accidents are unquestionably the most frequent cause of loss. Throughout the world, cars, buses, trucks, motorcycles, pedestrians, animals, taxis and other categories of travelers, share the roadways, contributing to economic and social development in many countries. The *Global status report on road safety 2018*, launched by WHO in December 2018, highlights that the number of annual road traffic deaths has reached 1.35 million. Road traffic injuries are now the leading killer of people aged 5-29 years.Currently, road accidents are ranked as the ninth most serious cause of death in the world, and without new initiatives to enhance road safety, fatal accidents will likely rise to the third place by the year 2020 . In developed countries, road traffic accidents rates have decreased since the 1960s because of successful interventions such as seat belt safety laws, enforcement of speed limits, warnings about the dangers of mixing alcohol consumption with driving, and safer design and use of roads and vehicles. For example, road traffic fatalities have declined by about 25.0 percent in the United States from 2005 to 2014 and the number of people injured has decreased 13.0 percent from 2005 to 2014 [3].Speeding endangers everyone on the road: In 2018, speeding killed 9,378 people. We all know the frustrations of modern life and juggling a busy schedule, but speed limits are put in place to protect all road users. Learn about the dangers of speeding and why faster doesn't mean safer[3]. However, traffic accidents have increased in developing countries from 1990 to 2014; for example, number of road traffic accidents increased by 44% in Malaysia and by about 43% in China. The costs and consequences of these losses are significant. Three-quarters of every poor family who lost a part in a traffic accident reported a decrease in their standard of living, and around 61% revealed borrowing cash to cover costs following their misfortune. Although transportation agencies often try to identify the most hazardous road sites, and put great efforts into preventive measures, such as illumination and policy enforcement, the annual number of traffic accidents has not yet significantly decreased. For instance, 35092 traffic fatalities were recorded in United States of America during 2015; an increase of 7.2% compared to the previous year [13]. The fatality rate per 100 million vehicle miles traveled (VMT) increased by 3.7% between 2014 and 2015. Thirty-Five States had more motor vehicle fatalities in 2015 than in 2014 Except November, every month witnessed increases in fatalities from 2014 to 2015; the highest increases occurred in July and September.

II. COLLECTION OF DATA

Data on road traffic accidents was obtained from the sources of the Indian Police.For the purpose of this study road traffic accidents which occurred on rural

secondary roadsin the South region in 2018 to 2019 were used. Furthermore, the accidentswhich occurred at the non-junctions are considered and that of junctions are removed. The accidents at junction are monitored which are local.To create a successful machine learning model, it is imperative that an organization has the ability to train, test, and validate them prior to deploying into production.

The datasets were downloaded from the websites which have the data of accidents that has been occurred. The majority of accidentsare the result of a vehiclecollision, the crash of a vehicle with a solid obstacle,with the street divider, signboards and a crashwith another vehicle (84 % of cases). Collisions with animals represent 11 % of casesand collisions with pedestrians and other collisions represent 3 % or 2 % of casesrespectively.

Table 1: Description statistics of road accident from 2018-2019.

Variable	*Frequency*
ACCIDENT LOCALITY	
urban areas	13
rural areas	1395
TYPE OF ACCIDENT	
crash with non-rail vehicle	562
crash with solid obstacle	294
crash with pedestrian	39
crash with animal	158
crash with train	2
crash with other objects	326
other type of accident	27
ACCIDENT CHARACTER	
accident with injury	776
accident with property damage only	632
ACCIDENT LIABILITY	
motor vehicle driver	1173
non- motor vehicle driver	38
pedestrian	15
forest animals, domestic animals	158
road fault 3	
vehicle technical breakdown	12
other liability9	
ROAD CLASSIFICATION	
two-lane	1390
other	18
ROAD ALIGNMENT	
straight segment	638
straight segment after curve	307
curve	418
junction	45

III. WORKING

Data Cleaning and Data Transformation

After we have selected the dataset. The first step after collection of data is to clean the datasets and transform it into the desired format as it is possible the dataset we use can be of different format. It is also possible that there are multiple datasets sources from different sources which may be in different file formats. So to use them we need to convert them into the format we want to or the type that type prediction system supports. The reason behind this step is that it is possible that the data set contains the constraints which are not needed by the prediction system and including them makes the system complicated and may extend the processing time. Another reason behind data cleaning is the dataset may contain null value and garbage values too. So the solution to this issue is when the data is transformed the garbage values are replaced. There are many methods to perform that.

Data Processing and Algorithm Implementation

After the data is been cleaned and transformed it's ready to process further. After the data has been cleaned and we have taken the required constraints. We divide the whole dataset int o the two parts that can be either 70-30 or 80-20. The larger portion of the data is for the processing. The algorithm is applied on that part of data. Which helps the algorithm to learn on its own and make prediction for the future data or the unknown data. The algorithm is executed in which we take only the required constraints from the cleaned data. The output of the algorithm is in 'yes' and 'no'. It gives the error rate and the success rate.

Output and User Side Experience

After the prediction system is ready to use. The Website is developed for the user. The user just has to fill a form which consists of different options they need to select. They are like the type of climate, the type of vehicle and so on. Once the user submits the form the algorithm is triggered and the input given by the user is passed to the prediction system. The user is given how accident prone the road can be in percentage.

IV. METHODLOGY

Our model consists of Road accident data that upon data pre-processing will form a dataset. Data pre- processing is one of the most important task in data mining. It deals with handling missing values or removing attributes and makes it a structured form of data in order to perform analysis on it.

The dataset obtained will now be subjected to various data mining techniques. Clustering will be performed on the given dataset. The main aim of performing clustering is to divide the data into different clusters or groups such that the objects within a group are similar to each other whereas objects in other clusters are different from each other. There are several clustering algorithms available: Hierarchical clustering technique like Ward method, single linkage, complete linkage etc, K means and latest class clustering (LCC).Other clustering algorithms like K-modes clustering is an enhanced version of K means clustering.

The clusters are then subjected to other algorithms like Association rule mining and trend analysis. Association rule mining is a very popular data mining technique that extracts interesting and hidden relations between various attributes in a large dataset. Association rule mining produces a set of rules that define underlying patterns in the dataset. The associativity is known by the frequency of their occurrence together in the dataset.

The trend analysis is performed to determine the upcoming trends based on the total count of accidents for each cluster. The trends analysis can also be applied on the entire dataset as well as the clusters. The trends can show a positive or negative trend for the future based on the current and past trends. There can be difference in trends for various clusters as there might be different factors dominant in causing accidents for that particular cluster. This trend will help us analyse the extent to which the measures taken across the years to reduce accidents has contributed in reducing the accident rate.

The simulation is performed by using R tools. Various data mining techniques and exploratory visualization techniques is applied on the accident dataset to get interpreted results. The R tools help to develop and interactive user interface. Thus we can analyse the various factors contributing for the accidents by plotting various graphs, charts and other statistical and graphical representations.

Risk estimation of various factors causing road accidents helps to determine the risk related with various factors that contribute to the causing of accidents. Based on the accident dataset analysis we calculate the individual probabilities of the factors and its attributes that are most likely to cause accidents among a given set of factors. This risk analysis is carried out by using a technique called Naive Bayesian technique that determines the individual probability associated with various factors and its attributes. The probabilities are stored in a separate table which will be accessed to calculate the accident risk value. Based on the user input of the present factors we estimate the risk when the factors are considered together. For the input from the user, the output is the estimated risk in terms of Low risk, Medium risk and High risk.

These analysis and research helps in providing solutions in order to reduce the accident rate and

decrease the fatality in the number of deaths occurring due to these accidents as this analysis will help in understanding the overall causes of accidents, the degree to which they play a role in the accidents and how they can be reduced.

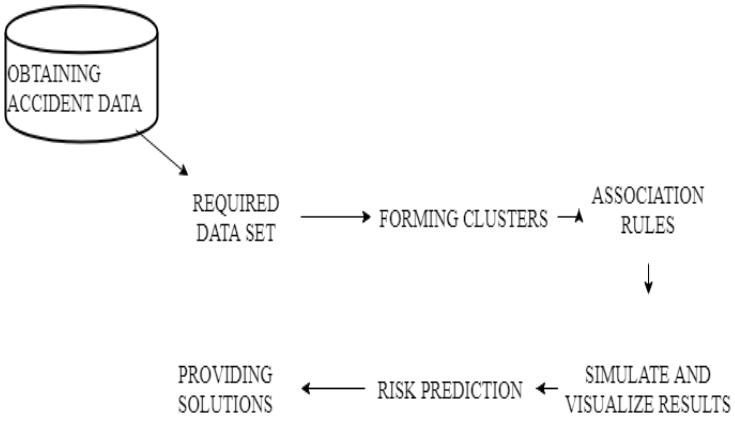

Fig: overview of proposed model

V.SYSTEM ARCHITECTURE

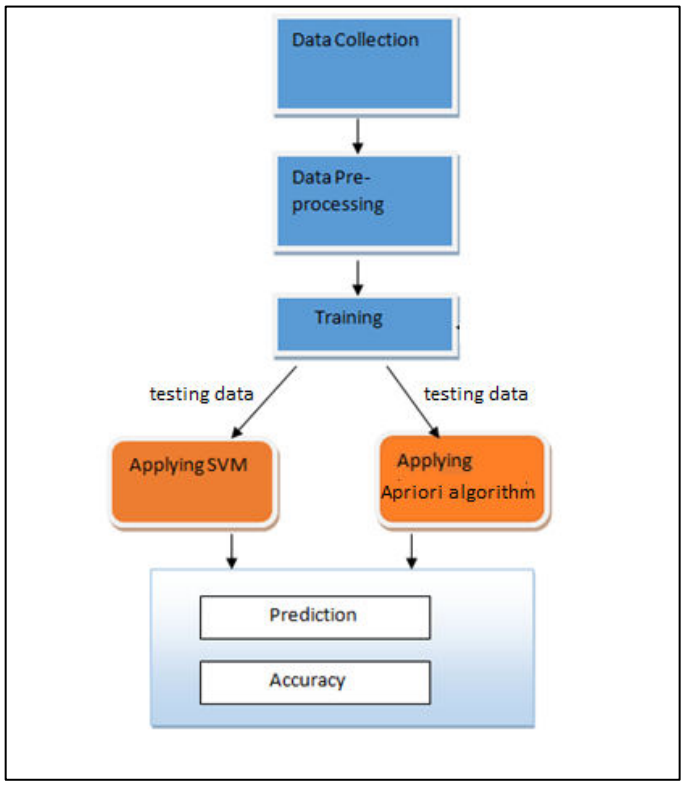

Fig. System architecture

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

- Importing the required Libraries
- Importing the Dataset
- Handling the Missing Data
- Encoding categorical data
- Splitting the Data set into Training set and Test Set
- Feature Scaling

SVM Algorithm:

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

Apriori Algorithm:

Aprioriis an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

$$
\begin{aligned}
&\text{Apriori}(T, \epsilon) \\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
&\quad k \leftarrow 2 \\
&\quad \textbf{while } L_{k-1} \neq \emptyset \\
&\qquad C_k \leftarrow \{c = a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a, \{s \subseteq c \mid |s| = k-1\} \subseteq L_{k-1}\} \\
&\qquad \textbf{for transactions } t \in T \\
&\qquad\quad D_t \leftarrow \{c \in C_k \mid c \subseteq t\} \\
&\qquad\quad \textbf{for candidates } c \in D_t \\
&\qquad\qquad count[c] \leftarrow count[c] + 1 \\
&\qquad L_k \leftarrow \{c \in C_k \mid count[c] \geq \epsilon\} \\
&\qquad k \leftarrow k + 1 \\
&\quad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

CONCLUSIONS

Road accidents prediction models are extremelyimportant tools in road safety programs used bytransportation agencies, police, health departments,education institutions that oversee road safety,vehicles, and the driver's education. They can be usedto predict both the frequency of accidents occurrenceand the contributing factors that could then beaddressed by transportation policies. According to theworld health organization (WHO), road crashes areranked as the ninth most serious cause of death in theworld, and present the world's leading cause of deathfor individuals between the ages of one and twenty-nine. The classification algorithm of the entire dataset. In the Road Accident prediction final result is to find the percentage of accident in particular area. Having lower number of features helps the algorithm to converge faster and increases accuracy. In the Road Accident prediction final result is to find the percentage of accident in particular area. Then we apply logistic regression on these features and obtain the least error.

REFERENCES

[1] .Leden, L. (2016). Pedestrian Risk decreases with pedestrianflow. A case study based on data from signalized intersectionsin Hamilton, Ontario – Accident Analysis and Prevention, Vol.34, p. 457-464.

[2] World Health Organization (2018) Global Status Report onRoad Safety 2018.
http://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.

[3] NHTSA—National Center for Statistics and Analysis (NCSA)(2016) NHTSA Studies Vehicle Safety and Driving Behaviorto Reduce Vehicle Crashes.
http://www.nhtsa.gov/NCSA

[4] Beirness, D.J. and Beasley, E. (2017) A Comparison of DrugandAlcohol-Involved Motor Vehicle Driver Fatalities.Canadian Centre on Substance Abuse, Ottawa.

[5] World Bank (2018) The World Bank-Transport for Development.

[6] NHTSA—National Center for Statistics and Analysis (NCSA)
(2016) NHTSA.

[7] Glenberg, A. (1996) Learning from Data: An Introduction toStatistical Reasoning. 2nd Edition, Lawrence ErlbaumAssociates, Mahwah.

[8] Gelman, A. and Hill, J. (2007) Data Analysis UsingRegression and Multilevel Hierarchical Models. CambridgeUniversity Press, London.

[9] Kim, D.G., Lee, Y., Washington, S. and Choi, K. (2007)Modeling Crash Outcome Probabilities at Rural Intersections:Application of Hierarchical Binomial Logistic Models.Accident Analysis and Prevention, 39, 125-134.

[10] Abdulhafedh, A. (2016) Crash Frequency Analysis. Journal ofTransportation Technologies, 6, 169-180.

[11] Blincoe, J., Miller, R., Zaloshnja, E. and Lawrence, A. (2015)The Economic and Societal Impact of Motor Vehicle Crashes,2010. National Highway Traffic, Washington DC.

[12] Park, S. and Lord, D. (2007) Multivariate Poisson-LognormalModels for Jointly Modeling Crash Frequency by Severity.Transportation Research Record, 2019, 1-6.

[13]Azad AbdulhafedRoad Crash Prediction Models: Different Statistical Modeling Approaches

AUTHORS

First Author – Prof. Aemikalaria, Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India.

Second Author -Abdul shukoor, B.E, Atria Institute of Technology and

Third Author –Harshal, B.E, Atria Institute of Technology and

Fourth Author – Allu Yagnesh, , B.E, Atria Institute of Technology