International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1039-1045 www.ijaem.net ISSN: 2395-5252

A Machine Learning Based Approach for Road Traffic Accidents Prediction

S Revathi[1], S Deeksha[2], M Dinesh[3]
[1]*Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology,*
[2,3]*Student, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India*

Submitted: 10-03-2021 Revised: 30-03-2021 Accepted: 01-04-2021

ABSTRACT: Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In order to give safe driving suggestions, careful analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents.

Globalization has affected many countries. There has been a drastic increase in the economic activities and consumption level, leading to expansion of travel and transportation. The increase in the vehicles, traffic lead to road accidents. Considering the importance of the road safety, government is trying to identify the causes of road accidents to reduce the accidents level. The exponential increase in the accidents data is making it difficult to analyse the constraints causing the road accidents. We find associations among road accidents and predict the type of accidents for existing as well as for new roads. We make use of cluster algorithm rules to discover the patterns between road accidents and as well as predict road accidents.

In this paper we apply statistics analysis and data mining algorithms on the FARS Fatal Accident datas et as an attempt to address this problem. The relationship between fatal rate and other attributes including collision manner, weather, surface condition, light condition, and drunk driver were investigated.

Certain safety driving suggestions were made based on statistics and cluster obtained.

I. INTRODUCTION :

There are a lot of vehicles driving on the roadway every day, and traffic accidents could happen at any time any where. Some accident involves fatality, means people die in that accident. As human being, we all want to avoid accident and stay safe. To find out how to drive safer, data mining technique could be applied on the traffic accident data set to find out some valuable information, thus give driving suggestion.

Data mining uses many different techniques and algorithms to discover the relationship in large amount of data. It is considered one of the most important tool in information technology in the previous decades.Association rule mining algorithm is a popular methodology to identify the significant relations between the data stored in large database and also plays a very important role in frequent item set mining.

India has second largest road network in the world. Road accidents happen quite frequently and they claim too many lives every year. It is necessary to find the root cause for road accidents in order to avoid them. Suitable data mining approach has to be applied on collected datasets representing occurred road accidents to identify possible hidden relationships and connections between various factors affecting road accidents with fatal consequences.

The results obtained from data mining approach can help understand the most significant factors or often repeating patterns. The generated pattern identifies the most dangerous roads in terms of road accidents and necessary measures can be taken to avoid accidents in those roads.

Due to increase in the population rate, there has been a major growth in number of vehicles used. Though public is requested to use the common public ransport, this has been a major increase in the vehicles being used. In proportion to this, road accidents have also being increased due to lack of road safety precaution and other related factors like climatic conditions and surface conditions, drunken drivers and also condition of vehicles.

The data mining methodology classification method primarily seeks at building a model (classifier) from a training data set that can be used to classify documents of unknown class labels. The Naïve Bayes method is one of the very fundamental probability-based classification techniques based on the hypothesis of the Bayes

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1039-1045 www.ijaem.net ISSN: 2395-5252

with the presumption that each pair of variables is independent. The algorithmic chase program is used to determine the moving cluster position.

In our day-to-day life multiple number of accidents occurs due to various reasons and the accident count increases and the transportation system and the road safety is the important factor to reduce the accident count. In accidents, many people get injured and also some may die. Accidents can occur in different ways depending upon different situations. There are many factors and reasons for road accidents which may include road conditions, lighting conditions, whether conditions, weight of the vehicle, no. of people travelling in the vehicle, speed of the vehicle at the time of accident, etc.,

Information Mining is a computational framework to oversee generous and complex informational collection and these informational collections can be of standard, apparent and mixed. It is extremely easy to use in variety of room have a place with science and organization; moreover, it could be used as a piece of deception recognizing evidence and various more intelligent cases and furthermore in setback reality issue.

Damage like property, people in view of street mishap are annoying. Normally, it happened that street incident scenes are more commonplace at particular places that can help in perceiving factors behind them. Connection run mining is a procedure that perceives the relationship in different parameter of street incidents.

II. LITERATURE SURVEY

Review of literature is important in any research work. Many researchers have carried out research work in the area of road accidents. Some of them have analyzed accident data in different ways. Some of them Identification of Black spot zone. Some of them have developed accident models for forecasting future accident trends. They have also proposed strategies for road safety.

In the present chapter literature review is carried out covering the different issues related to road accident and road safety.

Yannis T.H. (2014) was presented A Review of The Effect of Traffic and Weather Characteristics on Road Safety. Despite the existence of generally mixed evidence on the effect of traffic parameters, a few patterns can be observed. For instance, traffic flow seems to have a nonlinear relationship with accident rates, even though some studies suggest linear relationship with accidents. Regarding weather effects, the effect of precipitation is quite consistent and leads generally to increased accident frequency but does not seem to have a consistent effect on severity.

K. Meshram and H.S. Goliya (2013) were presented an analysis of accidents on small portion NH-3 Indore to Dhamnod. The data for analysis is collected for the period of 2009 to September 2011. More accidents occurred in Manpur region by faulty road geometry. The trend of accidents occurring in urban portion (Indore) is more than 35 % to rate of total accidents in each year.This may due to high speeds and more vehicular traffic. In the present study area the frequency of fatal accidents are 2 in a week and 6 for minor accidents in a week. More number of accident observed in 6 p.m. to 8 p.m. duration because in that time more buses are travels between villages and city.

One fatal and five casualties are occurring per km per year in the study area. The volume of the trucks passing through study corridor is increasing by year. At Rajendra Nagar from 2000 onwards the traffic is reduced due to the construction of by passes in that area.

E.S.Park (2012) studies the safety effect of wider edge lines was examined by analyzing crash frequency data for road segments with and without wider edge lines. The data from three states, Kansas, Michigan, and Illinois, have been analyzed. Because of different nature of data from each state, a different statistical analysis approach was employed for each state: an empirical Bays, before-after analysis of Kansas data, an interrupted time series design and generalized linear segmented regression analysis of Michigan data, and a cross sectional analysis of Illinois data.

Although it is well-known that causation is hard to establish based on observational studies, the results from three extensive statistical analyses all point to the same findings.

Data Preparation diagram

Data preparation was performed before each model construction. All records with missing value (usually represented by 99 in the dataset) in the chosen attributes were removed. All numerical values were converted to nominal value according to the data dictionary in attached user guide. Fatal rate were calculated and binned to two categories: high and low.

Several variables are calculated from other independent variables.

Here are two examples:

FATAL RATE:

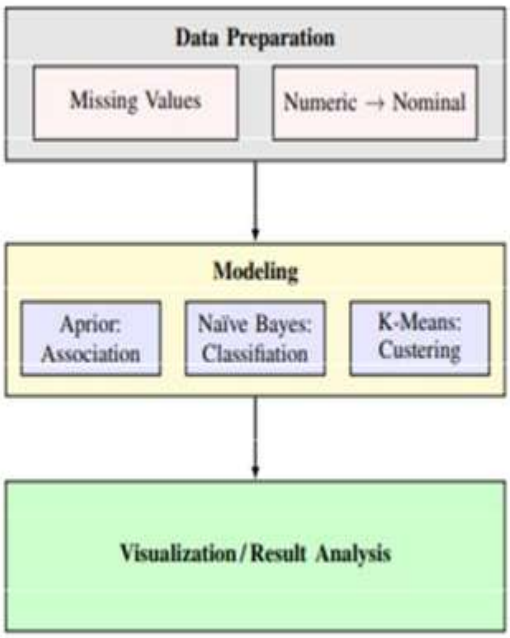
This variable denotes the percentage of fatality in a fatal accident computed as F AT AL RAT E = F AT ALS/P ERSONS, where F AT ALS

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1039-1045 www.ijaem.net ISSN: 2395-5252

is the number of fatalities and P ERSONS is the number of persons involved in the accident. It is also referred as "rate" in the analysis.

ARRIVAL TIME:

This variable is the arrival time of emergency staff in minutes, calculated as ARRIV AL TIME = 60 × (ARR HOUR − HOUR) +ARR MIN −MINUTE. All records with missing values on these time-related attributes are removed, and the early morning hours after 12:00 midnight are added by 24 to make it computationally easier.



a) Modeling

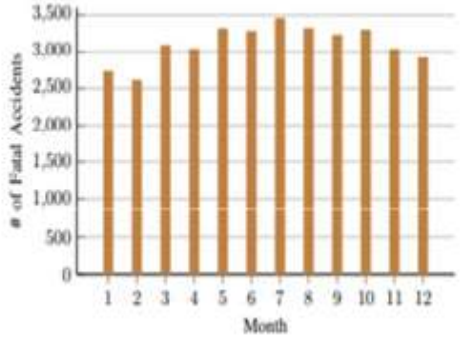
We first calculated several statistics from the dataset to show the basic characteristics of the fatal accidents. We then applied association rule mining, clustering, and Naïve Bayes classification to find relationships among the attributes and the patterns.

Result Analysis

The results of our analysis include association rules among the variables, clustering of states in the USA on their populations and number of fatal accidents, and classification of the regions as being high or low risk of fatal accident. We used a data analytic tool Weka to perform these analyses.

b) Statistics results

The number of fatal accidents in each months are shown. The most fatal accidents happened in July and the least in February.



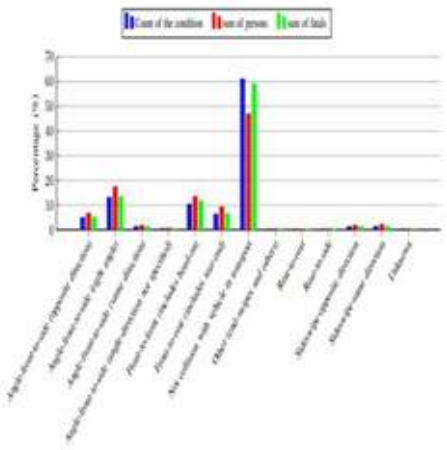
shows the percentage of fatal accidents in four other variables: SP LIMIT (speed limit), LGT COND (light condition), WEATHER (weather condition), and SUR COND (road surface condition).

1) Collision Type:

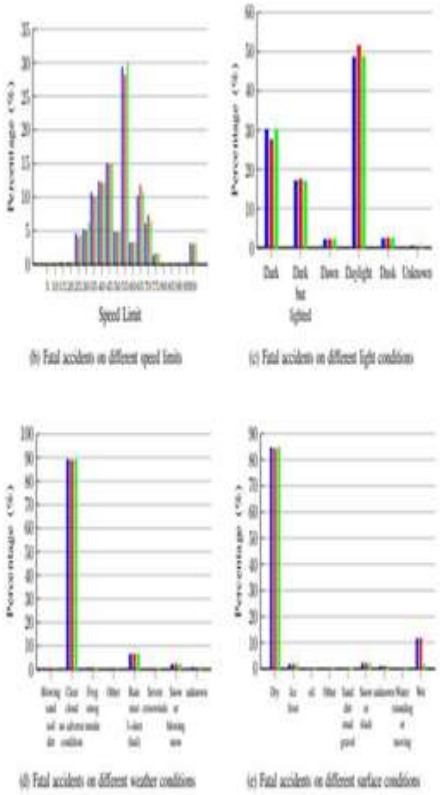
The percentage of fatal accidents happened on different collision types in comparison of people and fatal involved are shown in Fig 3(a). Surprisingly, the most fatal accidents are not in collision with motor vehicle in transportation. In Front-to-Front (Head-on Collision), the percentage of people and fatal involved are much higher than the percentage of accident number, which reveals that head-on collision has higher fatal rate in a fatal accident.

2) Speed Limit:

The percentage of fatal accidents happened at different speed limit in comparison of people involved and fatal involved are shown in Fig 3(b). Most of fatal accidents happened at speed limit 55 mph. The value "99" infers the missing value on attribute SP LIMIT.



International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1039-1045 www.ijaem.net ISSN: 2395-5252



3.Light Condition:

The percentage of fatal accidents happened on different light condition in comparison of people and fatals involved are shown in Fig 3(c). Unsurprisingly, most fatal accidents happen in day light condition because much more roadway traffic happens in day time other than at night.

4.Weather Condition:

The percentage of fatal accident happened on different weather is shown in comparison with percentage of people and fatals involved. Most fatal accidents happened at clear/cloud weather. This is understandable because clear/cloud is the most usual case of weather condition.

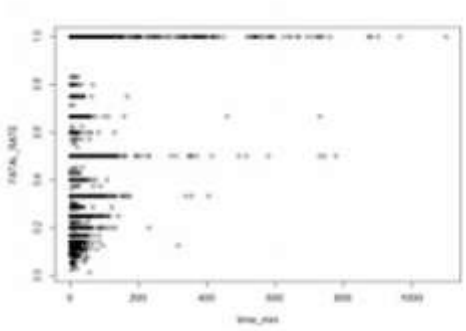
5. Roadway Surface Condition:

The percentage of fatal accident happened on different roadway surface condition is shown in Fig 3(e). Most fatal accidents happened on dry surface. This is understandable because the most usual case of road condition is that the road surface is dry. Evanco pointed out that there are "Golden Ten Minutes" for trauma in traffic accident.

To find out the possible relation between emergence service (EMS) arrival time (in minute) and fatal rate, the correlation is performed by R, as shown in Fig 5. Most fatals happened in short time and there is no significant relation between the EMS arrival time and fatal rate (correlation=0.1132231).

The minimum arrival time is 0 minute, which either because the EMS was at scene or data entry error. The average time take for EMS arrival is 18.27 minutes, the median is 10 minutes, and the maximum time is over 18 hours. We also need to mention that all the data is about fatal accident, so no matter how long would it take for EMS to arrive, there would always be fatals.

Also, there is no variable recording at what time the death happened, and a lot of records are missing value at time, so very limit information could be inferred from the time relevant attributes.



and the results are not significant so are not included here. After performing basic statistics and related work research, five attributes (collision type, weather, surface condition, light condition, drunk driver) are considered and selected as affecting fatal rate.

A. Association Rule Mining

Before applying the algorithms, the tuples with missing value in chosen attributes were removed, the numerical values were converted to nominal values according to data dictionary in the user guide [11]. The clean data was stored in CSV format and ready to be analyzed by the data analyzing tool Weka. The clean data for association rule mining and classification contains 36,789 tuples, 5 condition attributes, and 1 decision attribute. A small partial sample of the dataset is shown in Table I.

All values were converted to nominal values. After applying Apriori algorithm with minimum support = 0.4 and minimum confidence = 0.6 in Weka, association rules with fatal rate at the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1039-1045 www.ijaem.net ISSN: 2395-5252

right side as decision were generated. The best 13 rules are shown in Table II.

We could see that fatal accidents involving drunk driver have higher fatal rate, which means drunk drivers are much more dangerous than others. Also the clear/cloud weather condition with day light has high fatal rate, this reveals that not only the accident percentage is higher, as shown in basic statistics, but also the fatal rate are high (with confidence level = 0.65).

B. Classification

Naïve Bayes classifier was built on the cleaned data. Of the total 36,789 records, 24,994 were correctly classified giving a 67.95% accuracy rate. The various evaluation measures are given in Table III. The Naive Bayes Classifier shows that the fatal rate does not strongly depend on the given attributes, although they are considered feature in comparison to other attributes in the dataset.

C. Clustering of States

To find out which states are similar to each other considering fatal rate, and which states are safer or more risky to drive, clustering algorithm was performed on the fatal accident dataset. To perform the clustering, total number of fatality per state was calculated. Also the population data for each states in 2007 was obtained from U.S. Census Bureau.

With the fatal accident and the population dataset, fatalities per million people in the state was calculated. This allowed us to compare relative fatal rate in a state regardless of population of state. The simple K-means algorithm with Euclidean distance as the dissimilarity measure was applied to the data of 48 states (without Wisconsin, Wyoming, and District of Columbia, for some reason) with two variables: population (in 100,000) and number of fatal accidents. The states were grouped into 3 clusters .

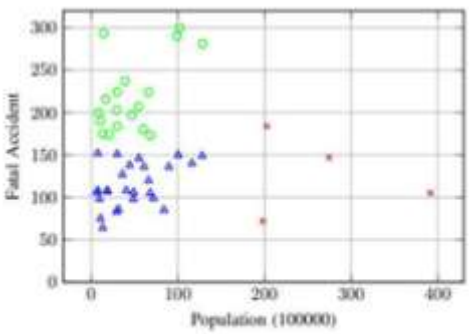


Fig. 6. Clusters of states by fatality of per million people in states

The three clusters are:

Cluster A (blue):

26 states. Those states in cluster A represents the safe state with relatively lower fatal rate per million people.

Cluster B (green):

18 states were clustered to cluster B which had relatively higher fatal rate.

Cluster C (red):

4 states, California, Texas, New York and Florida, formed cluster C.

These states have relatively large population and lower fatal rate. They are considered safe driving region and also outliers. After careful observation it was found that none of the states from mid-west or north-east region lied on cluster A, and almost all the states from south were in cluster B. Only two states from the south were located in region of cluster A, which is considered to be safe. Virginia and Washington DC, those too were in the southern region bordering north east. Georgia had highest fatal rate per million people, whereas Mississippi had highest number of per million people involved in fatal accidents. But Montana from west also had as much people involved in fatal accident as Mississippi. All four regions, North East, Midwest, West, and South, were also compared to each another to find out what's the difference between them.

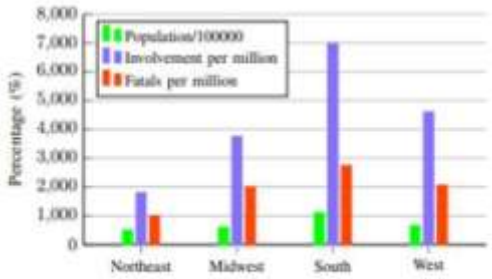


Fig. 7. Fatal accident in different regions

The southern region seemed to have 350% more people involved in an accident and almost 300% higher fatal rate compared to north east. This means that south is much more risky compared to rest regions. North east is the safest region and followings are mid-west and west.

Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1039-1045 www.ijaem.net ISSN: 2395-5252

Heroku is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud.

In this project we have created a web of the fatal accidents prediction using Streamlit library and deployed it on heroku platform.

The App will be able to predict whether a person has met with an accident or not depending on the factors given.

III. CONCLUSION

As seen in statistics, association rule mining, and the classification, the environmental factors like roadway surface, weather, and light condition do not strongly affect the fatal rate, while the human factors like being drunk or not, and the collision type, have stronger affect on the fatal rate. From the clustering result we could see that some states/regions have higher fatal rate, while some others lower. We may pay more attention when driving within those risky states/regions. Through the task performed, we realized that data seems never to be enough to make a strong decision. If more data, like non-fatal accident data, weather data, mileage data, and so on, are available, more test could be performed thus more suggestion could be made from the data.

The tree generated is pruned to large extent due to memory restrictions and varied type of data. Further room for improvement exists by adding more clusters to the distributed processing module & using more user friendly visualizations. The analysis can be used to develop preventive measures using Image Processing techniques for the vehicles violating traffic rules or for the vehicles that match many attributes in this project.Preventive measure can be developed in the locations which are more prone to accidents found from the analysis.

REFERENCE

[1]. Sachin Kumar , Durga Toshniwal ,"Analyzing Road Accident Data Using Association Rule Mining, International Conference on Computing, Communication and Security (ICCCS)", IEEE 2015.

[2]. An Shi,Zhang Tao, Zhang Xinming, Wang Jian, "Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining", Fifth International Conference on Intelligent Systems Design and Engineering Applications,2014.

[3]. Eyad Abdullah, Ahmed Emam, "Traffic Accidents Analyzer Using Big Data", International Conference On Computational Science and Computational Intelligence, 2015.
[4]. Seoung-hunPark ,Young-guk Ha, "Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction", Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing,2014.
[5]. Lokesh Hebbani, "Road Safety Scenario in India Problems & Solutions", 5th Foundation Day Lecture CiSTUP, IISC January 10, 2014. 6. Costabilea. J., Walla, J., Vecovskia, V &Baileya, "The rapid deployment of an effective road safety counter measure through a smart phone application- The story of Speed Adviser", Proceedings of the Australasian Road Safety Research, Policing & Education Conference November,2014.
[6]. Divya Bansal and Lekha Bhambhu. Execution of Apriori algorithm of data mining directed towards tumultuous crimes concerningwomen. International Journal of Advanced Research in Computer Science andSoftware Engineering, 3(9), September 2013.
[7]. Amira A El Tayeb, Vikas Pareek, and Abdelaziz Araar. Applying association rules mining algorithms for traffic accidents in dubai. International Journal of Soft Computing and Engineering, September 2015.
[8]. William M Evanco. The potential impact of rural mayday systems on vehicular crash fatalities. Accident Analysis & Prevention, 31(5):455–462, September 1999.
[9]. K Jayasudha and C Chandrasekar. An overview of data mining in road traffic and accident analysis. Journal of Computer Applications, 2(4):32–37, 2009.
[10]. S. Krishnaveni and M. Hemalatha. A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7):40–48, June 2011.
[11]. Liling Li, Sharad Shrestha and Gongzhu Hu , Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques, Department of Computer Science Central Michigan University, USA