A REVIEW ON ROAD TRAFFIC ACCIDENT DATA ANALYSIS

[1]Sindhu Sumukha,[2]George Philip C
[1]Student (MTech),[2]Associate Professor,
[1,2]Department of Information Science and Engineering, Ramaiah Institute of Technology, Bangalore, India

Abstract: Vehicle crashes occur because of numerous factors. It leads to loss of lives and permanent incapacity. Due to vehicle crashes, the budgetary expenses for both individuals as well as for the nation are influenced from the vehicle crashes. According to Road accidents statistics a total of 4,64,910 road accidents were reported in India, claiming 1,47,913 lives and causing injuries to 4,70,975 persons, which translates into 405 deaths and 1,290 injuries each day from 1,274 accidents.
In recent years many researches have been carried out and many models like Sequential models, Complex Linear Models, Time Sequence Models etc. has been created for analysis and characterizing purpose. Several Data Mining Techniques are also available, and it deals with two important steps i.e. Prediction of severity and identifying important factors for the accident. Several Machine Learning algorithms are available among those Artificial Neural Networks (ANN), Linear Regression (LR), Random Forest, Decision Trees, SVM are widely used. These are the algorithms which are widely used for predicting severity of the accidents. Tools available are Weka, Tangara, Knime, Orange, R, NLTK. The widely used tools are R and Weka.
This review paper discusses about the algorithm that shows better performance in the earlier researches done, and the widely used tools. The models are implemented based on the performance of the model of the previous research in order to overcome the shortfalls of the earlier work and the dataset is taken from Kaggle.

Index Terms - *ANN, LR, Decision Trees, Random Forest, R, Weka, SVM*

I. INTRODUCTION

Vehicle crashes occur because of numerous factors like speed of vehicle, diversion of driver concentration, road surface condition, driver expertise and so forth. Because of vehicle crashes individuals lose their lives orthey will experience the ill effects of incapacity. In any case, the budgetary expenses for both individuals as well as for the nation are influenced from the vehicle crashes.

The statistics of Road accidents, injuries and fatalities were released by India in 2017. According to the statistics, a total of 4,64,910 road accidents were reported in the country, claiming 1,47,913 lives and causing injuries to 4,70,975 persons, which translates into 405 deaths and 1,290 injuries each day from 1,274 accidents. This also means that 16 people are killed and another 53 are injured every hour on Indian roads.

In recent years many researches have been carried out and many models has been created to analyze and characterize the traffic accidents. Some are as follows:

- Sequential models
- Complex Linear Models
- Time Sequence Models
- Epidemiological Models
- Process Models
- Systemic Models
- Non-Linear Models

For all the above models Machine Learning (ML) and Artificial Intelligence (AI) are the two emerging technologies used to build the models.

Many Researchers have considered the Data Mining Techniques for analyzing the traffic accidents. These Data Mining Techniques are used in other fields like Health Care, Text mining, Pattern Selection etc. The recent works shows that traffic accident data analysis can be based on two categories of Data Mining Techniques:

- Prediction of Severity of Traffic Accidents
- Identification of Important factors that are responsible for Traffic Accidents.

Here, this paper considers the supervised machine learning approach for carrying analysis on the dataset. This Paper deals with Prediction of Severity and Analysis of various factors responsible for Traffic Accidents.The dataset used is UK dataset which is sourced from Kaggle. The dataset contains three files namely accident, vehicles and causalities and all the files are in .csv format. The total fields in each file is 34, 28 and 16 respectively. The attributes that are considered is around 28 for the study.

The remainder of the paper is organized as follows. In Section 2, the literature survey on prediction of severity of accidents is presented. The section 3 deals with algorithms, section 4 deals with Tools, section 5 deals with Conclusion.

II. LITERATURE SURVEY

Halil et al. [14] studied the road accident information data from Traffic Insurance Information Center (TRAMER) of Istanbul which was founded in 2003. According to TRAMER there were 854 road accidents in the year 2013 in Istanbul. The algorithms which were used to analyze and predict the Severity are Decision Tree Classification Algorithms (CART, ID3), Statistical Classification algorithms (Naïve Bayes, OneR), Classification Algorithms for Distance (IbK). The raw data has been cleaned and the dataset contained 1000 records with only two types of Severity (i.e., fatal, non-fatal). The CART algorithm outperformed the other models in predicting the performance, and Kappa Statistics. The algorithms can be used to improve the accident estimation system.

Cigdem et al. [11] studied the traffic accident data of 10 years which contains fatal, non-fatal accidents along with meteorological records collected from Adana, Turkey (2005-15). This data was provided by General Directorate of Security-Traffic Services Department and Turkish State Meteorological Service. The Data set consisted of 25,015 records with 20 parameters was considered for study and this data set was unbalanced hence they have used cross validation method to balance the data. The algorithms used for the study are Naïve Bayes Classifier, K-Nearest Neighbor method, Decision tree, Multilayer Perceptron, Logistic Regression. To classify fatal accidents Decision tree and logistic regression gave better results, the accuracy of the prediction model was found to be high with Decision trees. The data sets used in the study lacks information about the driver which is the main disadvantage.

Ameera et al. [13] Studied the traffic accident data of the Island of Abu Dhabi for the period of 3 years(i.e.,2013-16). There were 257 accidents records in the datasets. Here they have considered four levels of severity i.e., Minor, Moderate, Severe and Fatal. The Ordered Logistic model was used to predict the severity of the Traffic accidents which yielded 84.8% of accuracy. Running through red light was identified as one of the main factors for severity of the accidents.

Laura et al. [4] studied the traffic accident data from Spanish Traffic Agency (DGT) statistical portal of the year 2011 in which injuries of the people in the traffic accidents were recorded for a period of 24 hours in 30 days. The dataset contained 34 attributes with 1018204 records. The dataset was unbalanced hence they have used random sampling to balance the data. They have used Naïve Bayes, Gradient Boosting and Deep Learning algorithms to predict the severity of the accident. The best result was obtained by using Deep Learning and it can be used for decision making in the account of controlling the accidents.

Miao Chong et al. [16] Studied the data from National Automotive Sampling System (NASS) General Estimate System (GES). The dataset contained records for the year (1995 -2000), the total number of records are 417670. GES data set contained the information about drivers, and it doesn't include information about the passengers. This data set mainly contains information about driver, vehicle and police jurisdiction who handled the case, Injury severity classes are NO Injury, Non-incapacitating Injury, Incapacitating Injury, Fatal Injury. Here the crash occurred in 7 categories i.e., not collision, rear-end head-on, rear-to-rear, angle, sideswipe direction and sideswipe opposite direction. In this paper they have used Hybrid Learning Artificial Neural Networks, Decision Trees, Support Vector Machines, Hybrid Decision Tree-ANN for building the model. The Performance of Neural network, decision tree, Support Vector Machine, Combination of Decision tree and ANN were evaluated and found that for non-incapacitating injury, incapacitating injury, fatal injury, no Injury, possible injury classes hybrid approach gave better results than other algorithms, for no injury class and possible injury decision tree performed better. The important factor of injury severity is speed of the vehicle.

Olutayo et al. [17] studied the dataset from Nigeria Road Safety Corps. The dataset contains data of 24 months from Jan-2002 to Dec-2003 on the first 40 Km from Ibadan to Lagos. It contains attributes like Vehicle Type, Time of Delay, Season and causes. Some of the unknown factors discovered are Law of enforcement agent problems, Drivers condition, Attitude of the road users, Inadequate traffic road signs etc. Artificial Neural Network, Decision Tree were used for building the model In case of ANN two types of algorithms were used Multilayer Perceptron (MLP), Radial Basis Function (RBF). For Radial Basis Function model training and testing performance of 54.73% and 40.56% respectively was achieved. RBF had an accuracy rate of 0.547. For the MLP model training and testing performance was 52.70% and 45.20% respectively and it attained the accuracy rate of 0.399. Decision Tree used ID3 algorithm where training and testing performance attained was 77.70% and 70.27% and accuracy rate attained was 0.703. From the above results it is evident that Decision tree outperformed the ANN methods.

The below Table 1 shows the comparison of different works along with the title of the paper, objective, performance, methods used, Result of the papers.

From the below table it is evident that the most used algorithms are ANN, Decision Trees, Naïve Bayes to predict and to analyze the severity of the accidents

Table 1: Comparison of Different Reviews

Authors	Title	Methods	Objective	Algorithm Performance	Result
Laura Garcia et al [4]	Traffic Accidents Classification and Injury Severity Prediction	Naïve Bayes, Gradient Boosting Trees, deep Learning	To compare different ML algorithms for developing a classification model that determines whether the crash is fatal / not	Accuracy: Naïve Bayes=76.89, Gradient Boosting=87.12, Deep Learning = 87.47	To analyze accident is fatal or not and to understand the cause of the accidents like driver behavior
Fabio Galatioto et al. [15]	Advanced accident Prediction models and Impact Assessment	Neural Networks and Ensemble Techniques	To develop and test methods and tools in order to improve the ways of predicting the collision and estimation of Collison impacts. Comparing the statistical methods with ML models for predicting the main causes of road collision such as collision frequency, number of causalities and severity	Accident Rate Model, Causality Model, Accident Severity Model	1. Development of MAIA toolkit a web-based simulation platform 2. Parametric approach outperformed the non-parametric approach 3. Causality modeling reveled that accuracy of model declined for the events with more severe and less frequent accidents
Roop Kumar et al. [8]	Data Analysis in Road Accidents Using ANN and Decision Trees	ANN, Decision Trees i.e., ID3, C50	To Identify the Key factors for accident severity	Accuracy for C50=79.8%, ID3=77.7% ANN=79%	The Key factors of road accident severity are Light Conditions, Road Surface Condition and Weather Condition
Qasem et al. [7]	Data Mining Methods for Traffic Accident Severity Prediction	Decision Trees (CART, Random Forest, J48, Random Tree), ANN, SVM	To build the prediction model based on the classification rules	Accuracy: Random Forest=80.65%, ANN=61.445%, SVM=54.843%	Various Classification Algorithms were used to detect the influential environmental parameters on Road traffic accidents and Weka tool was used to generate the

					rules in order to build the prediction model and R tool was used applying sampling techniques for unbalanced data
Tadesse Kebede Bahiru et al..[2]	Comparative study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity	J48, ID3, CART, Naïve Bayes	Identifying the accident factors in order to predict severity of the accident	Accuracy for J48=96.3, ID3=94.09, CART=96.22, Naïve Bayes=94.62	Speed Limit, weather condition, Road Condition, number of lanes, lightning condition and time of accident were found to be the main influential parameters for the severity of the accidents
S. Vasavi [3]	Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques		To generate association rules that will analyze to discover the hidden patterns which are the root causes for accidents among different combination of attributes of a larger dataset. To propose a framework based on cluster analysis using K-medoids and Expectation Maximization		From the analysis accident distribution is even in normal days and high during weekends, Accidents are more in cold nights than in hot and clear conditions, Fatal accidents are found to be more in old-aged group, non-fatal in young and middle-aged group, Accidents were high in august and low in June.
A. Priyanka et al. [10]	A Comparative Study of Classification Algorithms Using Accident Data	SMO, J48, IBK	To Explore the Data Mining Techniques for traffic accidents data	Accuracy: SMO = 94%, J48=92.8%, IBK=88.3%	Several Models were built using SMO for identifying and extracting the rules. The best performing

					model was chosen and predicted that the traffic accident possibility in Coimbatore city are high
Ms. Gagandeep et al. [1]	Prediction of the cause of accident- and Accident-Prone Location on Roads Using Data Mining Techniques	Correlation Analysis, Exploratory Analysis	To identify location such as State highway /Ordinary district road where more accidents occur		From correlation analysis: inverse relation between the progress of road and length of road. From Exploratory analysis: Frequency of accidents are maximum on state Highways and minimum on other types of roads
Maher Ibrahim Sameen et al. [12]	Severity Prediction of Traffic Accidents With RNN	RNN, MLP, BLR	To Develop RNN model that predicts the severity of the accidents based on the temporal structure of accident data	Accuracy: MLP=65.48%, BLR=58.30%, RNN=71.7%	Developed RNN model to predict the injury severity of traffic accidents and the model outperformed MLP and BLR. Prediction of injury severity for drivers in dry surface, dark lightning condition
. Baye Atnafu et al. [18]	Survey on Analysis and Prediction of Road Traffic Accident Severity Levels Using Data Mining Techniques in Maharashtra, India	Random Tree, J48, Naïve Bayes, Association rule mining algorithms	To compare different ML algorithms and different tools which are used to run the ML algorithms		The Road accidents severity is changing in nature. The changing and increasing road traffic accident severity leads to issues of not understanding the behavior of accident, factors influencing it

					and managing large volumes of data
Halil Ibrahim et al. [14]	Analysis for Status of The Road Accident Occurrence and Determination of The Risk Of Accident By Machine Learning In Istanbul	AdaBoost, CART, C4.5, Naïve Bayes, OneR, IbK	To Analyze different ML Algorithms and to find which is more acceptable algorithm for road accident cause and the status detection of roads	Performance %, Kappa Criteria, Precision, Recall of CART was high, ROC of Naïve Bayes was high when compared to other algorithms	From the study, CART gave better performance in most of the performance measures
Dheeraj Khera et al. [5]	Prediction and Analysis of Injury Severity in Traffic Systems Using Data Mining Techniques	Naïve Bayes, ID3, Random tree	To analyze the performance of different algorithms in two different tools like Weka and Tangara	Accuracy In Weka: Naïve Bayes=50.7, ID3=25.35, Random Tree=45.07. In Tangara: Naïve Bayes=67.6, Id3=57.74, Random Tree=92.25	Comparison of the results obtained from two different tools. Time taken for building the model is considered in the study. Random forest gives the best result in Tangara than weka tool.

III. ALGORITHMS

Several Algorithms are available for implementing the Traffic Accident Severity Prediction. Among them some are as follows.

- Artificial Neural Networks (ANN)

Artificial Neural Networks are computational algorithms which is inspired by the biological neural networks. It is mainly used for classification, clustering and pattern recognition problems [19].

ANN is a powerful modeling tool for classification and prediction. Several Techniques like back propagation [7] The Kohonen ANN, Counter-propagation ANN etc. Some of the advantages of using ANN are flexibility with missing and noisy data removal techniques, ANN has capability to deal with untrained and complex patterns.

- Linear Regression (LR)

Linear Regression is the most basic Machine Learning algorithm used for Predictive analysis. LR is used to examine two types of problems (a). will a set of predictor variables predict the dependent variable (i.e., outcome). (b) which predictor variables are most significant in predicting the outcome. Linear Regression is represented using equation given below

$$Y=mX+c$$

Where Y = dependent Variable

X = Independent variable

m = regression co-efficient

c = constant

Regression analysis is mainly used for predicting the strength of independent variable, for forecasting an effect and for trend forecasting. Several types of Linear Regression are available they are Simple Linear Regression, Multiple Linear Regression [20]. It is easy and simple to learn and, we can easily find the relationship between the variables. The main disadvantage of LR are it is applicable only for linear problems and it builds up the noise if we have a greater number of parameters [21].

- Support Vector Machine (SVM)

SVM is a supervised Machine Learning algorithm which can be used for classification or regression problems. IN SVM Kernel trick technique is used for transforming the data and to find the optimal boundary between the possible outcomes [22]. The main objective of using SVM is to find the hyper plane that distinctly classifies the data points in N dimensional phase. Hyper planes act as the decision boundaries to separate the two classes i.e., for classifying the data points. SVM can be used for text classification, Bio informatics, Generalized Predictive Control etc. [23].

The main advantage of using SVM are When we don't know what kind of data, we are dealing with we use SVM, it works well for unstructured and semi-structured data, it is helpful in solving complex problems using kernel functions.

The main disadvantage of SVM is choosing appropriate kernel function is difficult, while processing large data set the time it takes is more, the end model is difficult to understand [24].

- Decision Tree

Decision tree is a supervised machine learning algorithm that is used for classification problems for both continuous and categorical variables. Based on the type of variables used we have two types of Decision trees Continuous Variable Decision trees and Categorical Variable Decision Trees. Decision trees creates split on all the nodes and select the most appropriate one. There are four types in selecting the split by the decision tree they are Gini Index, Chi-Square, Information Gain, Reduction in variance. The overfitting of the data can be avoided by limiting the tree size and Pruning [25]. Decision trees can be used for Business Management, CRM, Fraudulent Statement Detection, Engineering, Energy Consumption etc. It is easy to interpret, use and understand as it will be in the form of diagram. One of the main disadvantages of Decision tree is over-fitting.

- Random Forest

Random forest algorithm is supervised machine learning approach which falls under ensemble technique this can be used for regression and classification problems. Random forest has the behavior of both Decision trees and bagging. By using Random forest algorithm, it is very easy to measure the importance of features, it also becomes easy to reduce the impurity in different levels of trees. Random forest algorithms are simple and easy to use, as it is used for both classification and regression problems, we can minimize the overfit problems that usually occur. One disadvantage of Random Forest Algorithm is large number of trees makes the processing slow and for real-time prediction it is not suitable [27]

IV. TOOLS

Several Open Source Tools are available for Processing the data, they are as follows

- Rapid Miner

Rapid Miner is a data science software platform while was called earlier as YALE (Yet Another Learning Environment). Rapid Miner is written in JAVA code. Rapid miner is integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. According to Bloor Research, the rapid miner provides 99% analytical solutions using template-based frameworks this will provide speed delivery and the need to write the code [28].

- R tool

R is a programming Language and free software environment used for statistical computing and graphics. R is greatly used by statisticians, data miners for data analysis. R is written using C, Fortran and R itself. R includes libraries for Linear, Non-Linear, Clustering, classification, Time Series analysis etc. R is compatible with JAVA, C++, .Net, Python Languages. R is basically Interpreted Language which uses Command Line Interpreter (CLI). R provides flexibility for the users to create their own packages and to use them. The current version of R is 3.5 which is widely being used [26].

- Weka

Waikato Environment for Knowledge Analysis (WEKA). Weka is written in Java Language. It contains visualization tools for data analysis and predictive modeling. The main advantages of using Weka tool are it is easy to use, portable, freely available, contains all models like data preprocessing models, modelling techniques etc. Using Weka, we can do clustering, classification, preprocessing, modelling, feature selection etc.

Some of the other tools used are Orange, Knime and NLTK.

V. CONCLUSION

This Paper, A review on road traffic accident data analysis on traffic accident data set of United Kingdom, discusses about the latest work that has been carried out in the field of traffic accident analysis. In the recent years, there is increase in volume of road traffic and so the factors responsible for accidents vary accordingly. It is important for the government to lay down the rules for minimizing the growth of traffic and lay down the regulations for road from previous incidents for avoiding the future accidents which may lead to injuries, death, loss of organs etc. There are several gaps identified by the researchers during the research some are factors affecting the severity of the accidents, reason for accidents. There are several challenges in analyzing the traffic accidents which include modeling the algorithms for finding which models best suits the traffic accidents data and to detect the

severity levels. In order to fill some of the gaps this study identifies the suitable algorithms and tools from the recent studies for severity prediction and analysis of the influential attributes of the traffic accidents.

REFERENCES

[1] Ms. Gagandeep Kaur, Er. Harpreet Kaur, "Prediction of the cause of Accident-prone location on road using data mining techniques," IEEE-40222 [8th ICCNT 2017 July 3-5,2017, IIT Delhi, India].

[2] Tadesse Kebede Bahiru, Prof. Dheeraj Kumar Singh, Engdaw Ayalew Tessfaw, "Comparative study on data mining classification algorithms for predicting road traffic accident severity," Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (2018).

[3] S. Vasavi, "Extracting hidden patterns within road accident data using machine learning techniques," Springer Nature Singapore Pte Ltd. 2018. [Information and Communication Technology Proceedings of ICICT 2016, Mishra D.K, Azar A.T, Joshi A(Eds.) 2018, XVIII, 340 p, illus, Softcover ISBN: 978-981-10-5507-2]

[4] Laura Garcia Cuenca, Enrique Puertas, Nourdine Aliane, Javier Fernandez Andres, "Traffic accidents classification and injury severity prediction," 2018, 3dr International Conference on Intelligent Transportation Engineering.

[5] Dheeraj Khera, Williamjeet Singh, "Prediction and Analysis of Injury Severity in Traffic System using Data Mining Techniques," International Journal of Computer Application (0975-8887) [National Conference on Advances in Computing Communication and Application (ACCA-2015).

[6] Seyed Hessam-Allah Hashmienejad, Seyed Mohammad Hossein Hashmienejad, "Traffic accident severity prediction using a novel multi-objective genetic algorithm," Taylor & Francis, International Journal of Crashworthiness,2017 vol 22, no 4, 425-440.

[7] Qasem A. Al-Radaideh and Esraa J. Daoud, "Data mining methods for traffic accident severity prediction," International Journal of Neural Networks and Advanced Applications, volume 5,2018.

[8] Roop Kumar R, Ramamurthy B, "Data analysis in road accidents using ANN and Decision tree," International Journal of Civil Engineering and Technology, volume 9, Issue 4, April 2018, PP. 214-221, Article ID: IJCIET_09_04_023, IAEME Publication.

[9] Fang Zong, Hongguo Xu, Huiyong Zhang, "Prediction for traffic accident severity: comparing the Bayesian Network and Regression models," Hindawi Publishing Corporation Mathematical Problems in Engineering, volume 2013, Article ID: 475194, 9 pages.

[10] A Priyanka, K Sathiyakumari, "A comparative study of classification algorithm using accident data," International Journal of Computer Science and Engineering Technology (IJCSET), vol 5 No 10 oct 2014.

[11] Cigdem ACI, Cevher OZDEN, "Predicting the severity of motor vehicle accident injuries in Adana-Turkey using machine learning methods and detailed meteorological data," International Intelligent Systems and Applications in Engineering (IJISAE), 2018,6(1),72-79.

[12] Maher Ibrahim Sameen, Biswajit Pradhan, "Severity prediction of traffic accidents with Recurrent Neural Networks," Applied Science,2017,7,476; doi:10 3390/app7060476.

[13] Ameera Al-Khalifa, Abdulla Galadari, "Identifying the risk factors affecting crash severity at intersections with considering crash characteristics and signal configuration using an Ordered Logistic model".

[14] Halil Ibrahim BULBUL, Tarik KAYA, Yusuf TULGAR, "Analysis for status of the road accident occurrence and determination of the risk of the accident by machine learning in Istanbul," 2016 15th IEEE International Conference on Machine Learning and Applications.

[15] Fabio Galatioto, Mario Catalano, Nabeel Shaikh, Ecaterina McCormick, Ryan Johnston, "Advanced accident prediction models and impact assessment," The Institute of Engineering and Technology 2018, IET Intelligent Transport Systems, 2018, Vol.12, Issue 9, pp. 1131-1141.

[16] Miao Chong, Ajith Abraham, Marcin Paprzycki, "Traffic accident data mining using Machine learning paradigms," ResearchGate.

[17] Olutayo V.A, Eludire A.A, "Traffic accident analysis using Decision Trees and Neural Networks," I.J Information Technology and Computer Science, 2014, 02, 22-28.

[18] Baye Atnafu, Gagandeep Kaur, "Survey on analysis and prediction of road traffic accident severity levels using data mining techniques in Maharashtra, India," International Journal of Current Engineering and Technology 2017, INPRESSCO, Vol 7, No 6.

[19] https://www.xenonstack.com/blog/artificial-neural-network-applications/.

[20] https://www.statisticssolutions.com/what-is-linear-regression/

[21] https://www.quora.com/What-are-the-advantages-and-disadvantages-of-linear-regression

[22] https://community.alteryx.com/t5/Data-Science-Blog/Why-use-SVM/ba-p/138440

[23] https://data-flair.training/blogs/applications-of-svm/

[24] https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/

[25] https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in- python/#one

[26] https://en.m.wikipedia.org/wiki/R_(programming_language).

[27] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[28] https://en.wikipedia.org/wiki/RapidMiner.