International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Traffic Accidents Classification and Injury Severity Prediction

Shyam Sunder Pabboju, P.Satya Shekar Varma, Surya Prakash Jella

Abstract: Traffic accidents are one of the most life-threatening dangers to human being. Deaths and injuries due to traffic accidents have a great impact on society. Traffic accidents information and data provided by public can be useful to classify these accidents according to their type and severity, and consequently try to build predictive model. Detecting and identifying injury severity in traffic accidents in real time is primordial for speeding post-accidents protocols as well as developing general road safety policies. In this project we are using Logistic Regression algorithm to classify accident data. The data to be analysed is collected from various sources, is both structured and unstructured and has several attributes. In this project we are going to detect and analyse data together to generate decision trees that give insights on previous accidents.

Keywords: Traffic accidents; Decision tree; Logistic Regression; Injury Severity Prediction;

I. INTRODUCTION

Traffic accidents are unavoidable. The cost of deaths and injuries due to traffic accidents have a great impact on society. Traffic accidents and their severity are the results by several factors ranging from driver behaviour, roads characteristics, vehicles types, to weather conditions, to cite few of them. At present road accidents are one of the most life-threatening dangers to everyone [2]. Many strategies can be deployed to reduce deaths during traffic accidents, and one of them, consists in speeding postaccidents attention [1]. In this sense, predicting accidents severity could be a key for quick response to such accidents.

Road accidents are unpredictable incidents and their analysis requires a good knowledge of the factors affecting them.The major problem in the analysis of accident data is its Heterogenous nature [3].Thus, heterogeneity should be considered during the analysis of the data, otherwise some of the relationship between the data may remain hidden.

The severity of the road accidents is more in densely populated cities. Every year nearly 1.3 million people die [3] and 20-50 million are injured in road crashes, on average 3,287 deaths a day [3]. More than half of all deaths due to traffic accidents occur among young adults ages 15-44.

Revised Manuscript Received on April 30, 2020.
* Correspondence Author
Shyam Sunder Pabboju*, Assistant Professor, CSE, JNTUH, MGIT, Hyderabad, India. shyampabboju@gmail.com
P.Satya Shekar Varma, Assistant Professor CSE, JNTUH, MGIT, Hyderabad, India. satyashekarvarma@gmail.com
Surya Prakash Jella, student, CSE, JNTUH, MGIT, Hyderabad. suryaprakashjella@gmail.com

Over 90 percent of all road fatalities occur in low income countries [4], which have less than half of the world's vehicles. Road crashes cost USD 518 billion globally, costing individual countries from 1-2 percent of their annual GDP [7]. Unless action is taken, road accidents are predicted to become the fifth leading cause of death by 2030.

II. LITERATURE SURVEY

Harnen (2003) developed a generalized linear model for predicting motorcycle accidents at three-legged major minor priority junctions in Malaysia and found that motorcycle accidents were proportional to the power of traffic flows. Increases in non-motorcycle and motorcycle flows entering junctions were associated with an increase in motorcycle accidents [5].Table I gives information on the literature survey of traffic accidents classification and injury severity prediction.

Table I: Literature survey of traffic accidents classification and injury severity prediction.

Author	Objective	Data Mining Techniques	Accuracy
Chaozhonget.al (2009)	To identify the factors significantly influencing single vehicle crash severity.	Random Forest, Rough set theory	60.73%
Ali et.al (2010)	To identify Most important factors which affect injury severity	Classification & Regression tree	72.49%
Liping et.al (2010)	To predict Traffic accident duration of incident and driver information system	Artificial neural Networks	85.35%
DipoT.Akomolafe, Akinbola Olutayo (2012)	To predict causes of accidents and accident prone locations.	Decision tree: Id3, Functional tree	70.27%
Tibebe et.al (2013)	To Explore the possible application of data mining technology for developing a classification model	Classification & Regression tree	87.47%

Chang and Yeh (2006) developed two logistic regression models,one for non-motorcycle drivers and the other for motorcyclists. It was found that on an average, motorcyclists had approximately three times higher fatality risk than other drivers [5]. In-order to reduce the high risk of fatality sustained by these two classes of drivers, enhancing the driver's seatbeltuse rate, the rider's risk perceptions, and the enhancement of road quality were recommended.

Zambon and Hasselberg (2014) through a stepwise logistic regression procedure identified factors that affect the severity of motorcycle injuries by considering variables related to the individual, the environment, the vehicle and the crash in Sweden [8]. They found that suspicion of alcohol consumption emerged as the strongest determinant of a severe outcome. They also suggest that alcohol related crashes can be prevented through law enforcement and a multiplicity of policies at local and national levels.

III. ANALYSIS

Analysis of Dataset :

To get good prediction we should consider a large data set. In this project UK road accident data of size 18mb which has 140057 records is taken.Figure 1 displays the screenshot of the actual dataset used for this model in csv file.The feature extraction is done and as a result, few of the features were removed in order to get better result [3].The description of the columns of the dataset is below:

Accident_Index: The first column i.eAccident_Index acts as the unique key in the data set and helped us to map every column to its respective key.

Location_Country,Location_State, Longitude, Latitude: The second, third, fourth and the fith column of the dataset i.e. Location_Country, Location_State, Longitude and Latitude gives data about the accident location. These location related data helped to plot the results on google maps.

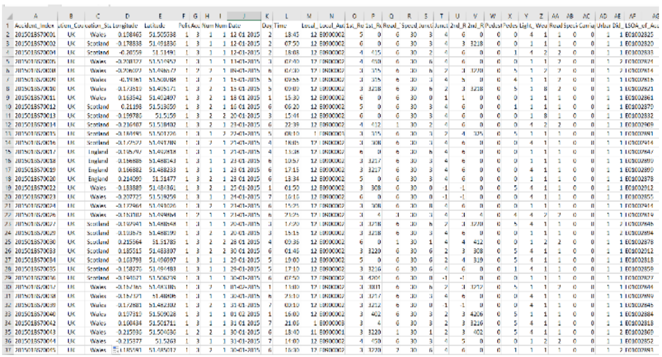

Fig. 1. Screenshot of dataset in excel sheet.

Police_Force: The sixth column is Police_Force that provides data related to police force at the accident lo-cation.

Accident_Severity: The seventh column is Accident_Severity. This column indicates the severity of the accident based on numbers. A number 3 indicates a more severe accident whereas number 1 indicates a less severe accident. This column is acting as a class label for the classification Machine Learning algorithm.

Number_of_Vehicles and Number_of_Casualties: *The* eighth and ninth column explains the number of vehicles involved in accident and the number of casualties occurred in the accident.

Date,Day_of_Week and Time: The tenth, eleventh and twelfth column provides data about date, day and time of the accident occurring. Using these we can analysis that on which day of the week or time most accidents occur.

Local_Authority_(District) and Local_Authority_Area: The thirteenth and four-teenth column gives us the district and highway numbers. Doing analysis of these two columns we can visualise which highways and districts are most prone to severe accidents.

1st_Road_Class, 1st_Road_Number, Road_Type, 2nd_Road_Class and 2nd_Road_Number:The fifteenth, sixteenth, seventeenth, twenty- first and twenty-second column is the road class which gives us the classification of the road.

Speed_limit: The eighteenth column is the speed limit of the road.

Junction_Detail and Junction_Control: The nineteenth and twentieth column gives us the junction and junction control details.

Pedestrian_Crossing-Human_Control and Pedestrian_Crossing-Physical_Facilities: The twenty-third and twenty-fourth column provides data about Pedestrian Crossing.

Light_Conditions: The twenty-fifth column indicates the light conditions near the accident on a scale of 1-7.

Weather_Conditions: The twenty-sixth column is the weather conditions on a scale of 1-9.

Road_Surface_Conditions: The twenty-seventh column is the road-surface-condition which is of the range 1-5

Special_Conditions_at_Site: The twenty-eighth column indicates if there were any other conditions in the accident sight which might have caused the accident and it ranges from 0-7.

Carriageway_Hazards: The twenty-ninth column is the carriageway-hazard which indicates any barriers in the road. The number ranges from 0-7.

Urban_or_Rural_Area: The thirtieth column determines if the road belongs to a rural or an urban area.

Did_Police_Officer_Attend_Scene_of_Accident: The thirty-first column gives us information whether police came to the accident spot or not.

Accident_Location_Zone: The thirty-second column is the zone of accident location which is the geo-graphical location surrogate for latitude and longitude.

Cleaning of Dataset:

The dataset has been cleaned in order to get better result and accuracy. The null values have been removed and also, a check is in place in order to look for duplicate records. We have removed few features as a part of cleaning and ex-traction data [8]. The features such as accident ID, location specific features, information related to post accident de-tails, authority etc. have been removed. During the run of Machine learning algorithm[6], the information related to specific location details, column with unique values hinders to generalise the information; it memorizes the data and tends to learn the specific data [4]. As a result, it will not generalise the result when the code is run using another dataset. The location related data is helpful to visualize the results and plot the locations on google maps [9]. Figure 2 shows step by step procedure of data pre-processing technique.

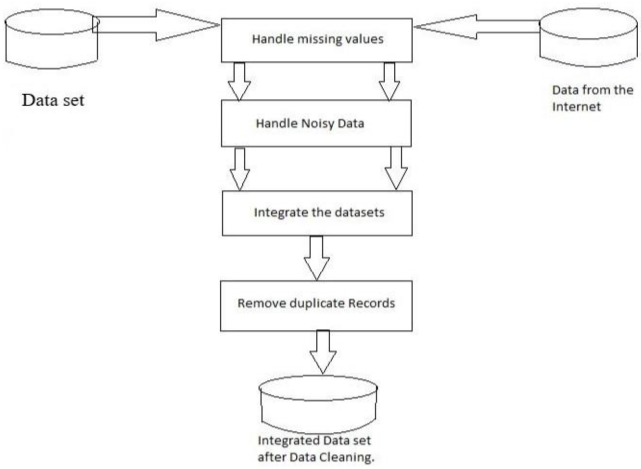

Fig. 2. Data pre-processing technique.

As shown in Figure 2 initially data is filtered by removing noisy data and duplicate records. After cleaning the data is integrated[6].

The feature normalisation is done for feature Time so as to normalize the range of values. Normalisation is important to get better result and include Time as a feature in the algorithm [2]. The results provide insight to accidents occurring are more at what time in the day.

IV. SYSTEM ARCHITECTURE

The system basically uses databricks notebook which sets the path to read the data set. The data set is revived from the data base and send for data preprocessing[5]. In the preprocessing the data is extracted,selected values are converted to continuous values then the data isused to train model[3]. After training evaluate new data with the model to get predicted values. Figure 3 shows the system architecture of the proposed system.

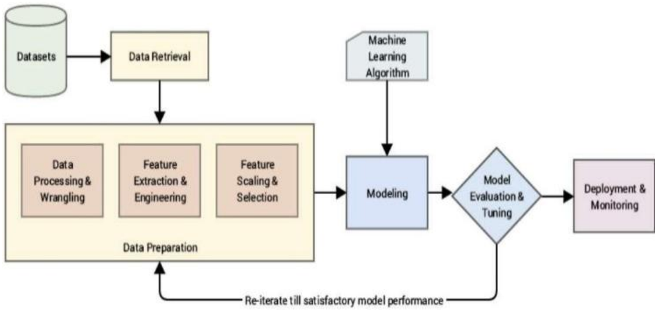

Fig. 3. System Architecture

A Data Flow Diagram is a pictorial representation of the flow of the data [1]. A Data Flow Diagram is used to create an overview of the system without going into detail [6]. Figure 4 shows what kind of information will be input to and output from the system, and where the data will be stored [3].

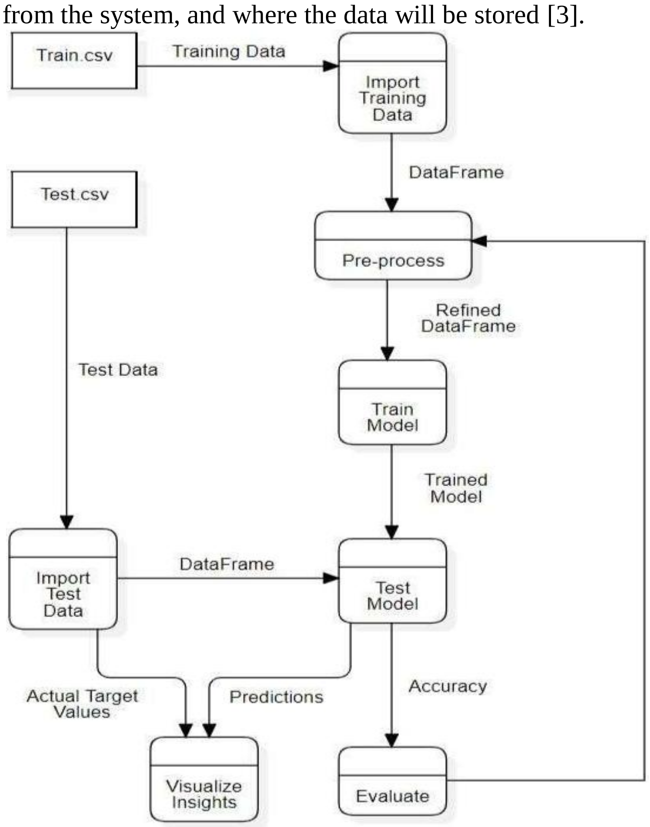

Fig. 4. Data Flow Diagram.

Figure 4 shows data flow diagram of the proposed model and gives overview of the system without going into detail.

V. METHODOLOGY

Approach for problem statement:

The problem is treated as a Classification problem of machine learning where a dataset and corresponding class label is provided. This kind of data can be used for supervised learning [4]. Thus, we will build a model using classification or supervised learning algorithm based on logistic [9] regression.

The model is then trained on training set and results are tested on test set. We have implemented Decision tree and Logistic Regression for the problem. The dataset has been divided into 80 - 20 in order to train the model on 70 percent and to test on rest of the 30 percent [8].

Brief overview of the Algorithms used:

Decision tree:Decision tree learning is the construction of a decision tree from classlabelled training tuples. A decision tree is similar to a flowchart, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label [5]. The topmost node in a decision tree is the root node. Decision tree goal is to generate a model that predicts the value of the target variable based on several input variables.

The dataset is split in the ratio of 80:20 [3]. The 80 percent of the data (training data) fed to the decision tree classifier and a tree is returned [2]. When this tree is fed with the test data for the prediction, it results in an accuracy of eighty five percent and fifteen percent error approximately.

Logistic Regression: The logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary or multinomial variable and one or more metric independent variables [1]. When the dependent variable is dichotomous then Logistic regression is the appropriate regression analysis. For logistic regression we will train the model using training set and test the model created by logistic regression algorithm using test set [4]. In this project we are using multinomial logistic regression for the prediction which results 85 percent accuracy and error of 15 percent approximately.The outcome of logistic regression [7] model lies between 0 and 1 whichworks on the below formulae.

$$\ln[p/(1-p)] = \beta_0 + \beta_1 X$$

whereβ_0 is constant and p is the probability that the event Y occurs which has range of 0 to 1.

p/(1-p) is the "odds ratio".

The odds ratio has a range of 0 to ∞ with values greater than 1 associated with an event being more likely to occur than to not occur and values less than 1 are associated with an event [9] that is less likely to occur than not occur. Figure 5 shows comparison of logistic regression model and linear probability model.

The estimated probability is:$p = \frac{e^{(\beta_0 + \beta_1 X)}}{1+e^{(\beta_0 + \beta_1 X)}}$

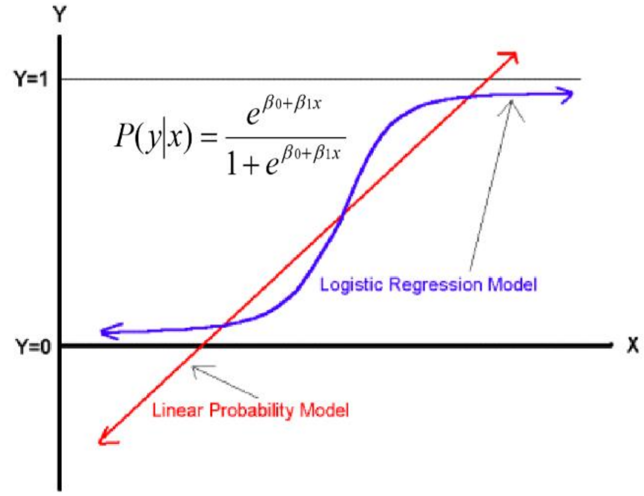

Fig. 5. Comparison of the logistic regression model and linear probability model.

if $\beta_0 + \beta_1 X=0$, then p = 0.5.
if$\beta_0 + \beta_1 X$ is big, then p approaches 1.
if $\beta_0 + \beta_1 X$ is small, then p approaches 0.

VI RESULTS AND PREDICTION

We have used Databricks cloud for implementation of project. The community edition is used that comes free of cost. Databricks provides a cloud platform on top of Apache Spark [10].Databricks empowers anyone to build and deploy advanced analytical solutions. Apache Spark is thecommoncluster computing system [10] which is very fast and provides high level APIs in Java and Python.

A. Accuracy:

Accuracy is a metric for the performance analysis of the any classification model. We can calculate accuracy by dividing the number of correct predictions with the total number of predictions [9]. The below Table II shows the parameters for prediction.

Table II: parameters for prediction

	Actual Positives	Actual Negatives
Positive Predictions	True Positives (TP)	False Positives (FP)
Negative Predictions	False Negatives (FN)	True Negatives (TN)

$$\text{Accuracy}=\frac{\text{number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Accuracy} =\frac{TP + TN}{TP + FP + FN + TN}$$

B. Precision:

Precision is the fraction of correctly predicted positive instances to the total number of positive instances.

$$\text{Precision}=\frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

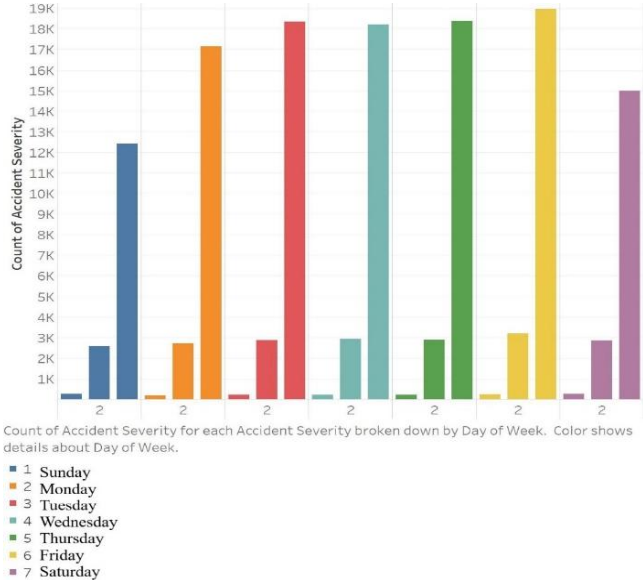

Fig.6. Number of accidents occurred in a week.

The Figure 6 shows that the number of accidents occurred in a week (1-Sunday through 7-Saturday). We can observe that maximum accidents occurred on friday. The Figure 7 shows that the number of accidents occurring in each month with severity. The data is startling as the number of severe accidents are quite high as compared to low severity accidents with July being the highest and February being the lowest.

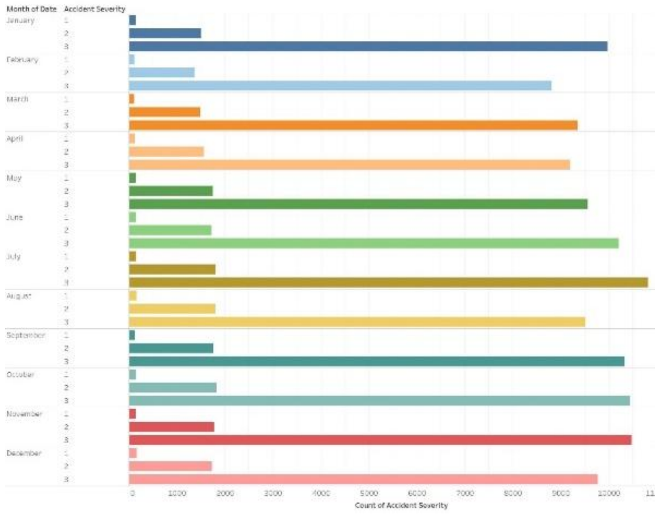

Fig.7. Accident severity in all months of the year.

The below figure 8shows the Share of different vehicle types in road accidents on a pie chart.

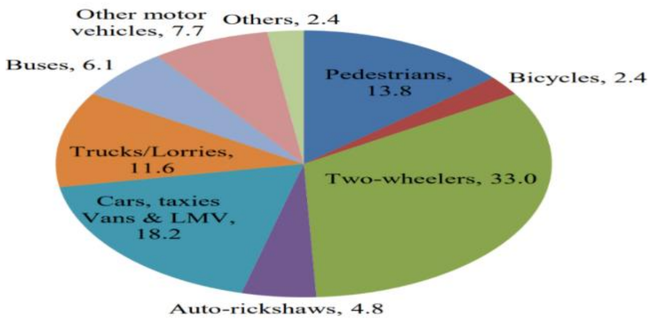

Fig.8.Share of different vehicle types in road accidents

The Figure 9 shows the most accidentprone areas on google map.

Instructions on how run the code using Databricks notebook.

1) Go to https://community.cloud.databricks.com and sign up.
2) Select the Community Edition.
3) Sign up for Databricks Community Edition by entering personal details.
4) Login with your email and password.
5) Select Workspace from the left-hand column, and under workspace, choose shared and select import under the drop-down.
6) Create Clusters by choosing the clusters in the left-hand column, click on the create cluster tab and enter the cluster name and use Apache Spark Version Spark 2.0 and above.

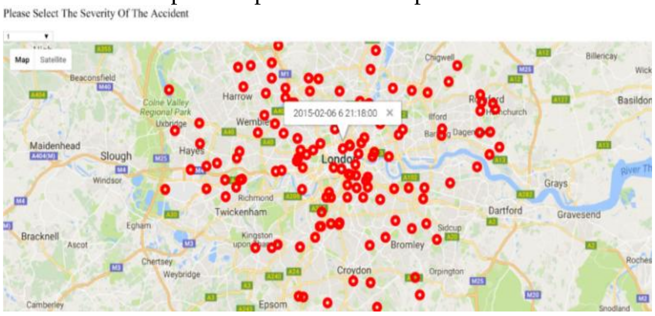

Fig.9.Google map displaying accident prone areas.

C. Code:

The source code with the dataset developed for Traffic accidents classification and injury severity prediction is available at:

https://drive.google.com/file/d/1YG6FcFOM13Hy0iENoYB4rMKOWONKxKFO/view

VII. CONCLUSION

Increasing road accidents gives a wake-up call to everyone. It has been found that most of the road accidents are caused by various factors such as types of vehicles, age of the driver, weather condition, age of the vehicle, road structure and so on. Thus, we build a model that gives efficient prediction of road accidents which results in 85 percent accuracy and error of 15 percent approximately.It has been clearly demonstrated that logistic regression analysis has been successfully applied to predict road accidents. There is a lot of scope for Traffic accidents classification and injury severity prediction.

It is suggested to further improve the model reported in this study using more variables (e.g., Drugs, Animal Crossings, Traffic volume etc.) to get a more realistic picture in predicting or forecasting accidents.Accidents are uncertain so we can not exactly predict future trends by using any model or theory but this model is very useful to take mitigation measures in advance by studying future trends to minimize the accident rate to certain extent and to take other safety measures.

REFERENCES

1. V. Gopinath, K Purna Prakash, Challa Yallamandha, G Krishna Veni, S Krishna Rao, "Traffic accidents analysis with respect to road users using data mining techniques", International Journal of Emerging Trends & Technology in Computer Science, vol.6, no.3, 2017.
2. Alkheder, S.; Taamneh, M.; Taamneh, S. Severity prediction of the traffic accident using an artificial neural network. J. Forecast. 2017, 36, 100–108.
3. Turaiki, Maryam, "Modeling traffic accidents in Saudi Arabia using classification techniques", 4th Saudi International Conference on Information Technology IEEE transaction, 2016.
4. J. Kashyap, Chandra Prakash Singh, "Mining road traffic accident data to improve safety on road-related factors for classification and prediction of accident severity", International Research Journal of Engineering and Technology, vol. 03, no. 10, 2016.
5. Olutayo, A.Eludire, "Traffic accident analysis using decision trees and neural networks", I.J. Information Technology and Computer Science, vol. 6, no. 2, pp. 22-28, 2014.
6. Tadesse Kebede Bahiru, " Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity", International Journal, Part Number: CFP18BAC-ART, ISBN:978-1-5386-1974-2, IEEE 2018. [Online]. Available:https://ieeexplore.ieee.org/document/8473265
7. Maze, Agarwai& Burchett. (2006). "Whether weather matters to traffic demand, traffic safety, and traffic operations and flow".
8. Tesema, T. B., Abraham, A., & Grosan, C. (2005). "Rule mining and the classification of road traffic accidents using adaptive regression trees". International Journal of Simulation, 6(10), 80-94.
9. Peden, M. (2004) "World report on road the traffic injury prevention". Geneva: World Health Organization.
10. Ming Zheng, "Traffic Accident's Severity Prediction: A DeepLearning ApproachBased on CNN ", IEEE Xplore Journal, vol. 7, March 2019. [Online]. Available:https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8661485

AUTHORS PROFILE

Shyam Sunder Pabboju, working as Assistant Professor, CSE , MGIT. He is currently pursuing Ph.D. in Osmania University, Hyderabad. He is life time member of CSI .He published several international and national journals and attended several conferences. He has 14 years of teaching experience and his areas of interest are cloud computing, machine learning, deep learning.

Satya Shekar Varma Poranki , working as Assistant Professor, CSE MGIT. He is life time member of ISTE. He published several international and national journals and attended several conferences. He has 13 years of teaching experience and his areas of interest are Big data, image processing.

Surya Prakash Jella pursuing final year B.Tech, CSE in MGIT.