IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

Three Class Classification Technique To Predict Road Accident Severity

Ramesh M Chakrasali[1*], Naganandini G[2], Ancy Thomas[3]

[1] Department of Computer Science, Acharya Institute Of Technology, Bengaluru, India
[2] Department of Computer Science, Acharya Institute Of Technology, Bengaluru, India
[3]Department of Computer Science, Acharya Institute Of Technology, Bengaluru, India

[*]*Corresponding Author: rameshmc1998@gmail.com, Tel.: 9620932672*

DOI: https://doi.org/10.26438/ijcse/v7si14.380385 | Available online at: www.ijcseonline.org

***Abstract*—** In recent years, road accidents are becoming more and more due to the larger growth in population. The growth of population and the increase in number of vehicles has led to a traffic congestion and sometimes may results in accidents. There are many factors that may lead to the road accidents and those maybe the driver's carelessness, drunk and drive, road conditions etc. Using the technology, necessary measures can be taken in order to predict the accidents at prior and to prevent the occurrence of accidents. In this research paper we use Gretl tool to identify the factors that are significantly contributing to the accidents, applied the logistic regression classification technique to build the machine learning model in order to predict the accident severity using the predictors like number of vehicles involved, road conditions, weather conditions, light conditions etc. Here we consider the accident severity as a dependent variable which is of three classes that is slight, serious and fatal. The main objective of this paper is that the accident has already occurred, in which we are predicting the severity of that accident**.**

***Keywords*—**AccidentSeverity,Predictions,LogisticRegression,Gretl,Tableau

I. INTRODUCTION

The road traffic accident (RTA) is considered to be more injurious when the mishap occurs. The accident predictions would go upto maximum by 2020. The RTA accidents are eportedly high when compared to 2013. Road traffic accidents are the leading cause of death among young people.

India is again no exception and road accident data shows that more than 1.3 lakh people die through RTA on Indian roads, making the India to the top in the global list of fatalities from road crashes. Rapid urbanization, pressure on the roads due to increase in number of vehicles, lack of appropriate road engineering, poor awareness levels, non-existent injury prevention programmes, and poor enforcement of traffic laws has made the situation critical. The human agitating factor has contributed to the increase in road accidents. Drunken driving, over speeding, refusal to follow traffic rules, and reckless driving are maybe one of the main reasons for road accidents.

Drunken driving is one of the major causes of road traffic accidents and a separate laws has to be put up in order to reduce the accidents due to drunken driving. Data shows drunken driving to be responsible for 70% of road fatalities. The risk of road accidents due to drunk and drive increases significantly above a blood alcohol concentration (BAC) of 0.04 g/dl. So, laws that establish blood alcohol concentration are maybe more effective at reducing number of alcohol related crashes. Over speeding increases the probability of road accidents. Setting and enforcing mandatory helmet for riders may decrease the severity of the road accident and sometimes may reduce many deaths.

Carelessness of driver like use of mobile phones during driving, non-use of helmets, non-use of seat-belts are the significant contributing factors for road traffic accidents and must be avoided. Sometimes driver fatigue and sleepiness can also contribute to the crashes. Improper designing of roads and lack of pedestrian pavement are other contributing factors. Only 28 countries have comprehensive road safety laws on major key risk factors like drunken driving, speeding, and failing to use helmets, seat-belts, and child restraints. This is a major cause of concern and both society and government should work together to reduce this preventable cause of death.

II. RELATED WORK

In [1] research on "The traffic accident hotspot prediction based on the Logistic Regression method" quotes about accidents that occurs . The results shows location of car in road transects, the road safety grade, the road surface condition, the visual condition, the vehicle condition and the driver state are the most prominent factors which leads to traffic accident The relationship between the identified factors were analysed and the predictions were made on the accuracy of the accidents which was found to be 86.6%.

In [2] research on "Predicting Traffic Accident Severity Using Classification Techniques",they extracted the feature information on traffic accidents and stored them into an accidental feature database. Features that were extracted were split into two sets which were disjoint, then applied classification methods such as decision trees ,neural network and Bayesian network to generate the classification models in order to improve the prediction accuracy of traffic accidents severity.

In [3] research on "Data Mining Methods for Traffic Accident Severity Prediction". In this research, Three classification techniques (decision trees, ANN and SVM) were applied to identify features that were influencing the road accidents. Further a prediction model is built and tested using real dataset obtained from department of transport of UK observed an accuracy by 61% using ANN and then 54.8% using SVM.

In [4] research on "Traffic Accident Severity Prediction using Artificial Neural Network". In this research ,a data mining software called WEKA is used in order to build ANN classifier and using the same to predict the injury severity of the traffic accidents based on 5973 traffic accident records occurred in Abu Dhabi over 6 year period. The overall prediction performance for training and testing data were 81.6% and 74.6% respectively. An ordered probit model was used as comparative benchmark with ANN in order to validate the performance of ANN. Later it is revealed that accuracy of 59.5% obtained from the ordered probit model was clearly less than ANN accuracy value of 74.6%.

III. METHODOLOGY

The main objective of the proposed methodology is to build the prediction model using the Logistic regression. This section explains the proposed research methodology to check the performance of the Logistic regression.

A. Overall Research Design

Figure (1) shows the different phases of our implementation used in this research. In the first phase we have collected the dataset and pre-processing is done by oversampling of data when there is a less minority class, and transformed the categorical data to numerical data and then finally feature scaling is done using normalization technique. Then using the pre-processed dataset we used the logistic regression classification technique to classify the data into three classes that is whether the accident is slight, serious or fatal. The results of the three classifiers are evaluated based on four evaluation measures that is Accuracy, Precision, Recall and F1-Measure.And then later in the next phase we used the same classification model to predict the severity of the particular accident.And later in our research we have used another business intelligence tool called tableau to visualize the accident data in order to provide the analysis of the road accidents.

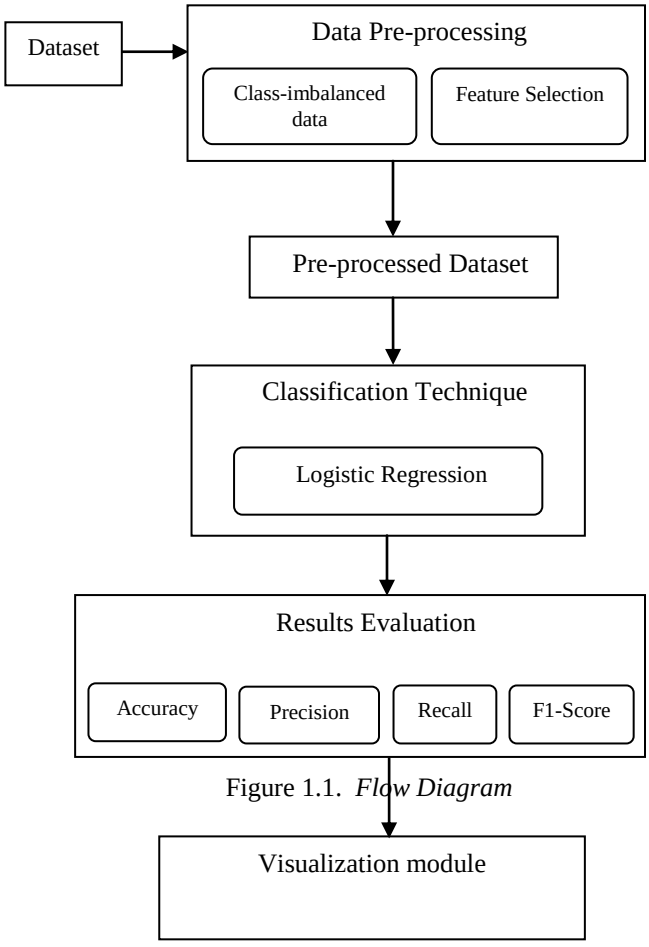

Figure 1.1. *Flow Diagram*

B. Research Phases

The Dataset

In our research we used road accident data that is available at (https://data.gov.uk/dataset/road-accidents-safety-data) which was published by the Department for Transport of the United Kingdom in the year 2015.

Dataset descritption

Table 1.1 Dataset description

Feature Name	Feature description and values
Speed Limit	The Speed limitation of the road where the accident occurred.
Number of Vehicles	Number of Vehicles involved in an accident.
Number of Casualities	Number of Casualities involved in an accident.

Road Type	1. Roundabout 2. One-way street 3. Dual carriageway 4. Single carriageway 5. Slip road
Junction Detail	0. Not at junction or within 20 metres 1. Roundabout 2. Mini-roundabout 3. T or Staggered junction 5. Slip road 6. Crossroads 7. More the 4 arms 8. Private drive or entrance 9. Other junction
Light conditions	1. Daylight 2. Darkness- lights lit 3. Darkness- lights unlit 4. Darkness- no lighting 5. Darkness- lighting unknown
Road Surface Conditions	1. Dry 2. Wet or Damp 3. Snow 4. Frost or ice 5. Flood over 3cm deep 6. Oil or diesel 7. Mud 8. Data missing or out of range
Weather Conditions	1. Fine no high winds 2. Raining no high winds 3. Snowing no high winds 4. Fine and high winds 5. Raining and high winds 6. Snowing and high winds 7. Fog or mist 8. Other 9. Unknown
Day of the week	Day of the week when the accident occurred

Table 1.2 Dependent variable description

Accident Severity	Code	Number of instances	Percentage
Fatal	1	4500	1%
Serious	2	48747	13%
Slight	3	302884	85%

Dataset Splitting

After pre-processing the dataset we split the dataset into training and testing data as 80% and 20% respectively.

Data Pre-Processing

Data pre-processing is an important stage for handling the data before applying a classification technique. This process involves various steps including cleaning, normalization, feature selection, transformation and coming up with a solution for Class-Imbalanced data by applying the oversampling technique to minority cases

C. Class-Imbalanced Data

Class imbalance problem is a major issue in the classification in which the solution for this issue can be achieved by under-sampling the majority class, over-sampling the minority class, or a hybrid of over and under sampling approaches. . Resampling techniques can be achieved by Our dataset is imbalanced because the major samples are for the Slight class while the minority samples are for the classes Fatal and Serious. In such situation, most of the classifiers are biased towards the major classes and hence provide poor classification rates on minor classes. In addition, it is also possible that classifiers predict everything as a major class and ignore the minor class, such in our case where the logistic regression classifiers predicted all classes as Slight classes and misclassified the Fatal and Serious classes and treated them as slight as well. To handle this issue and to improve the classification accuracy of class-imbalanced data, we used re-sampling techniques, namely Under-sampling, Oversampling, and hybrid sampling.

D. Feature Selection

Feature selection, also known as attribute selection or variable selection, is a process of selecting a subset of relevant features for using in model construction. The used dataset contains 31 features, in addition to 1 for the dependant variable. We used threshold p-value of 5%(0.05) to check whether the particular feature is significantly contributing to the accident or not. If the particular feature has a threshold p-value less than 5%(0.05) then it is inferred that feature is significantly contributing to the accidents. We used Gretl software to check whether the particular feature is significantly contributing or not. Then the used algorithm is applied to the dataset with these selected features, and the accuracies of them were compared and repeated this process with the multi thresholds to obtain the highest accuracies.

Tool used Gretl

Gretl, which has the acronym of Gnu Regression, Econometrics and Time-series library, is the open source statistical package, mainly for Econometrics. The Gretl software is written in C and uses Cross-platform operating system. Here we used Gretl to build model initially to check

whether the factors are significantly contributing to the accidents are not.

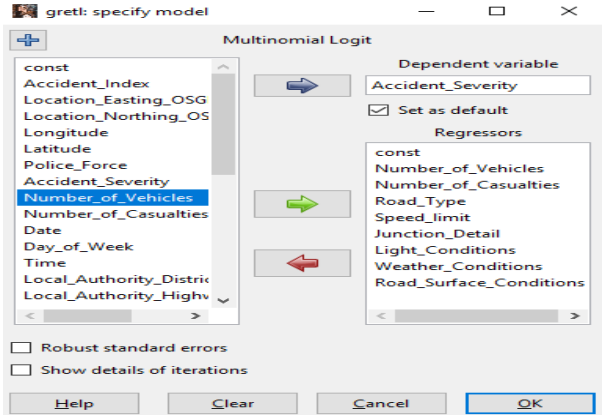

Figure 1.2 *Variables input to build a model*

In our Gretl software, we have input both dependent and independent variables. Here we considered Accident Severity as the dependent variable and others factors like Number of Vehicles, Number of Casualities, Road type, Lightning conditions and weather conditions as an independent variables.

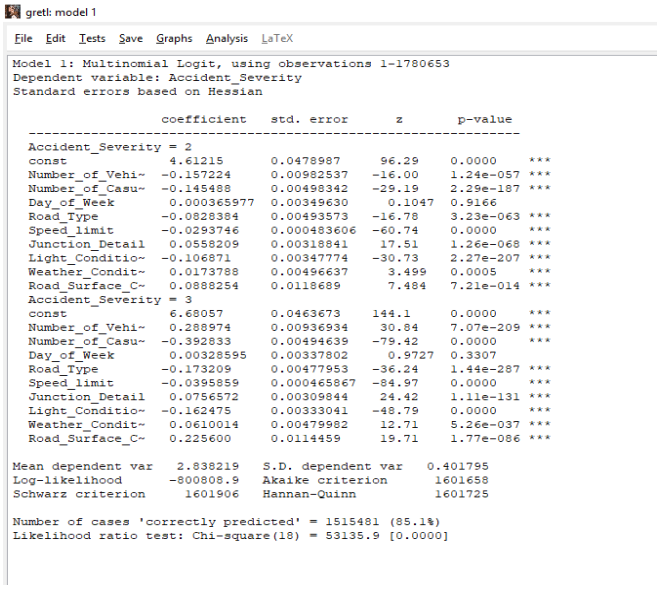

Figure 1.3 *Model Estimation in Gretl*

Classification technique

Considering that there are n factors x1,x2,…xn affecting the occurrence of road accidents,the multinomial logit model is used to classify those factors with respect to the dependent variable.

$$\Pr(Y_i = 1) = \frac{e^{\beta_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\beta_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

$$\Pr(Y_i = K-1) = \frac{e^{\beta_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}$$

IV. RESULTS AND DISCUSSION

This section presents and discusses about the experiments and outcomes of different classifiers that is logistic regression and KNN classifier. Accuracy, precision and recall were used in this comparisons.

Results

When the logistic regression classifier is applied to our dataset, we get the result as follows:

	precision	recall	f1-score	support
1	0.04	0.00	0.00	4500
2	0.41	0.00	0.00	48747
3	0.85	1.00	0.92	302884
micro avg	0.85	0.85	0.85	356131
macro avg	0.43	0.33	0.31	356131
weighted avg	0.78	0.85	0.78	356131

0.8504089787185053

Figure 1.4 *Performance measurement*

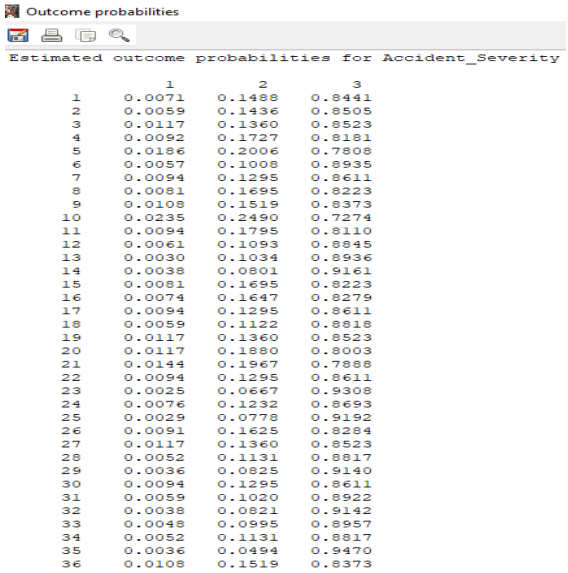

Figure 1.5 *Estimated Outcome Probabilities*

gretl: display data

Model estimation range: 1 - 1780653

	Accident_Severity	fitted	residual
1	2.00	3.00	1.00
2	3.00	3.00	0.00
3	3.00	3.00	0.00
4	3.00	3.00	0.00
5	3.00	3.00	0.00
6	3.00	3.00	0.00
7	3.00	3.00	0.00
8	3.00	3.00	0.00
9	3.00	3.00	0.00
10	3.00	3.00	0.00
11	3.00	3.00	0.00
12	3.00	3.00	0.00
13	3.00	3.00	0.00
14	3.00	3.00	0.00
15	3.00	3.00	0.00
16	3.00	3.00	0.00
17	2.00	3.00	1.00
18	3.00	3.00	0.00
19	3.00	3.00	0.00
20	2.00	3.00	1.00
21	3.00	3.00	0.00
22	3.00	3.00	0.00
23	3.00	3.00	0.00
24	3.00	3.00	0.00
25	3.00	3.00	0.00
26	3.00	3.00	0.00
27	3.00	3.00	0.00
28	3.00	3.00	0.00
29	3.00	3.00	0.00
30	3.00	3.00	0.00
31	2.00	3.00	1.00
32	3.00	3.00	0.00
33	3.00	3.00	0.00
34	3.00	3.00	0.00
35	3.00	3.00	0.00
36	3.00	3.00	0.00
37	3.00	3.00	0.00

Figure 1.6 *Model Estimation- Fitted and Residual*

And then later, the graph is plotted for independent variables against the dependent variables. We plotted the actual vs fitted curve for different factors like Road conditions, Weather conditions, Number of vehicles, Lightning conditions etc with respect to the dependent variable that is Accident Severity having classes (Serious,Slight,Fatal).

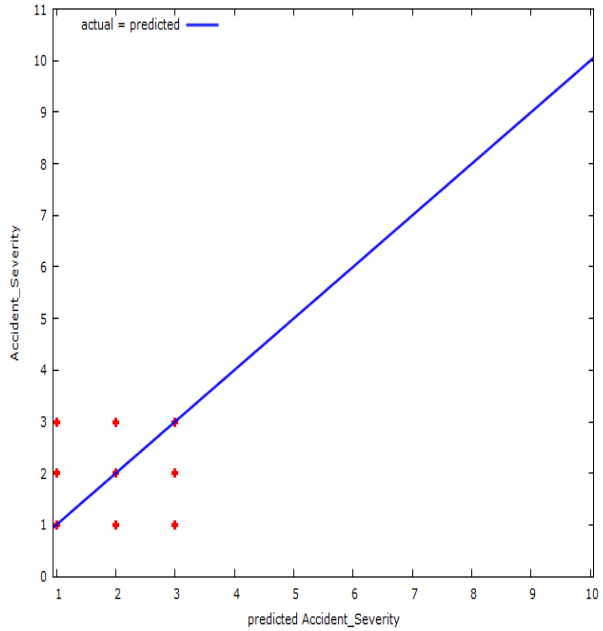

Figure 1.7 *Actual vs Predicted Accident Severity*

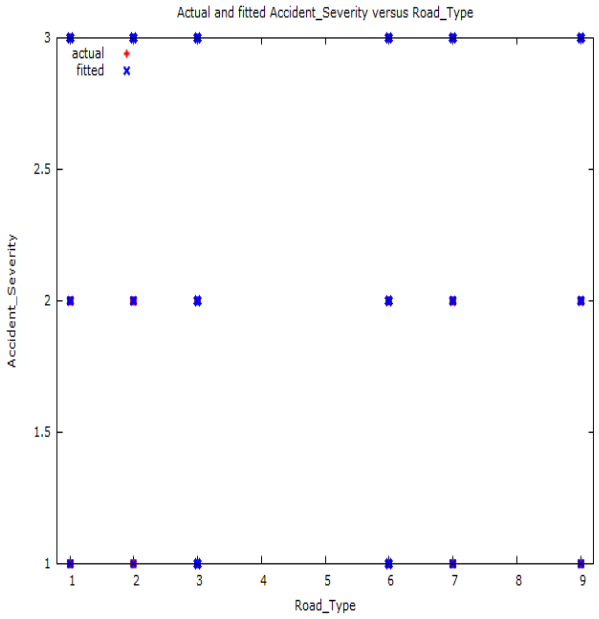

Figure 1.8 *Severity vs Road_Type*

And similarly when we apply the K-nearest neighbor classifier we get the accuracy of 82% which is lesser than logistic regression.

Data Visualization

For data visualization we used the tool called Tableau.

Tableau: It is the business intelligence tool used for data visualization. Here, in our research we used this tool in order to visualize the entire dataset of road accidents and to provide an inference from those.

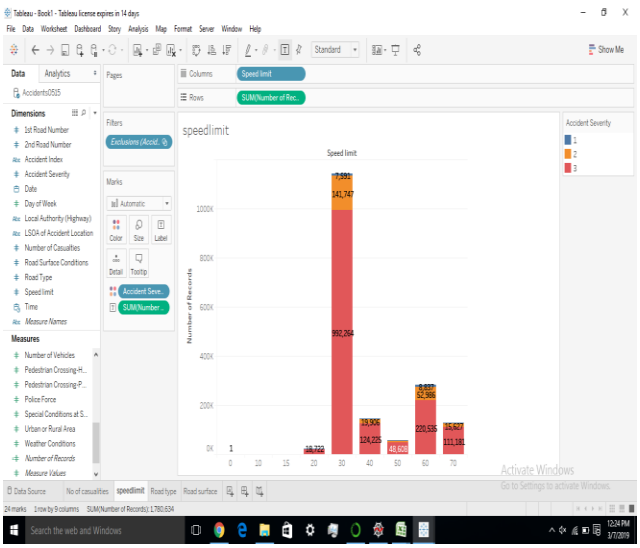

Figure 1.9 *Accident Severity vs Speed Limit*

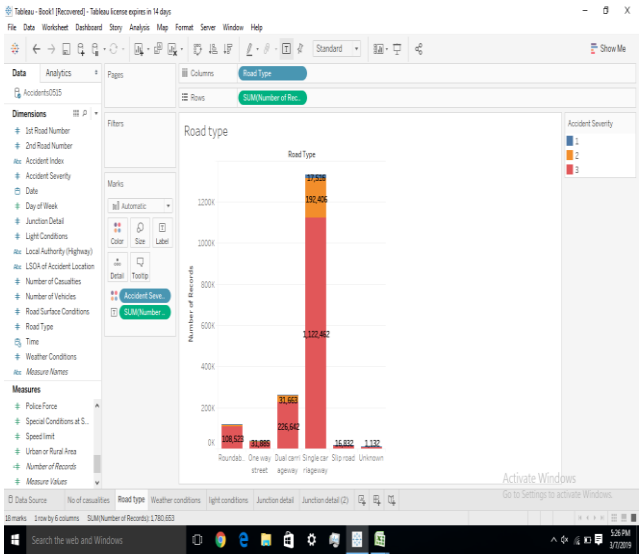

Figure 1.10 *Accident Severity vs Road Type*

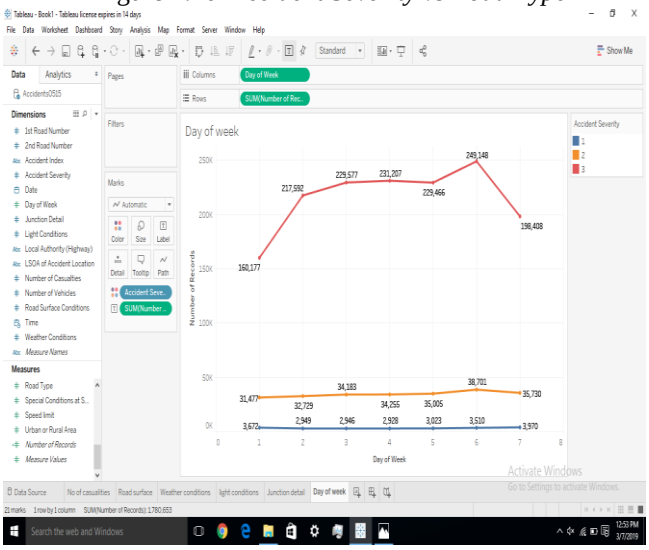

Figure 1.11 *Accident Severity vs Day of Week*

Similarly,we have visualised for the other factors with respect to the Accident Severity.

V. CONCLUSION AND FUTURE SCOPE

Classification of Road accident severity using Logistic Regression, we got the prediction with the accuracy of upto 85%.

From the data visualization in our research we can infer:

1. Speed limit: Accidents are more in the roads having the speed limit restricted to 30km.
2. Roadtype: Accidents are more in the road of type single carriageway compared to dual carriageway and Roundabout.
3. Day of Week: Accidents are more in the 6th day of the week.

VI. Preventive Measures

1.Speed monitoring cameras and radars and speed-limiting governors in vehicles are useful devices in enforcing the speed limit.

2. Laws that establish blood alcohol concentration of 0.05g/dl or below are effective at reducing no. of alcohol-related crashes.

3. Banning drivers from using hand-held mobile phones.

4. New gadgets are to be developed for collision prevention and should be fitted on all vehicles.

5. Gadgets can be developed to automatically slow down the vehicle,if safe distance commensurate with the speed of the vehicle in front is not maintained.

REFERENCES

[1] Tao Lu, Yan Lixin, Zhu Dunyao, Zhang pan "The traffic accident hotspot prediction: Based on the Logistic Regression method" The 3rd International Conference on Transportation Information and Safety, June 25 – June 28, 2015, Wuhan, P. R. China.

[2] Maher Al-Zuhairi, Biswajeet Pradhan "Severity Prediction of Traffic Accidents with Recurrent Neural Networks" Article in Applied sciences

[3] SharafAlkheder, Madhar M. Taamneh, Salah Taamneh "Traffic Accident Severity Prediction Using Artificial Neural Network" Journal of Forecasting,J.Forecast(2016) Published in Wiley Online Library.

[4] Rui Garrido, Ana Bastas, Ana de Almeida, Jose Paulo Elvas" Prediction of Road Accident Severity using the Ordered ProbitModel" ElseVier publication.

Authors Profile

Mr. Ramesh M Chakrasali pursuing Bachelor of Engineering in Department of Computer Science in Acharya Institute Of Technology, Bengaluru, India.He is a member of Computer Society of India since 2018.He has got selected in project exhibition of Old Dominion University,USA.And his main interest is on Machine Learning,Data Mining,Data Analytics.

Mrs Naganandini.G pursued Bachelor of Engineering Master of Engineering from U V C E, India in the year 2010. Pursuing Ph.D. from REVA University and currently working as Assistant Professor in Department of Computer Science , Acharya Institute of Technology. She has published more than 5 research papers. Her main research work focuses on Data Mining ,Bioinformatics and Machine Learning. She has 11 years of teaching experience .

Mrs Ancy Thomas pursued Bachelor of Engineering Master of Engineering from R V C E, India in the year 2009. Pursuing Ph.D. from VTU and currently working as Assistant Professor in Department of Computer Science , Acharya Institute of Technology. She has published more than 4 research papers. Her main research work focuses on Data Mining and Machine Learning. She has 14 years of teaching experience .

