ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

Design a Machine Learning Model to Identify and Predict the Cause of Vehicle Accidents in East Gojjam, Amhara, Ethiopia

Lijalem Megibaru [1], Baye Atnafu [2]

Department of Software Engineering, School of Computing, Debere Markos Institute of Technology, Debre Markos, Ethiopia[1]

Department of Software Engineering, School of Computing, Debere Markos Institute of Technology, Debre Markos, Ethiopia[2]

ABSTRACT: Road accidents are the main cause of death, grievous injury as well as minor injuries in the world. According to world health organisation (WHO) report approximately 1.25 million people are death annually and between twenty and fifty million peoples suffer non- fatal injuries. Especially in the developing countries where the rates at which road traffic accident occur is more than the critical limit. So, knowing the cause of vehicle accidents are play the vital role to reduce the accident that occurred frequently. Currently, accident data generating time to time and frequently from various sources which rich huge raw data. However, analysing and exploring useful pattern from this huge raw data is a big challenge. So, in order to easily exploring the useful pattern/ information from huge data which will help the decision makers making the decision effectively and properly. The main aim of this study is to design a machine learning model to analysis, identify and predict the cause of vehicle accidents in order to reduce the vehicle accident in East Gojjam. Amhara, Ethiopia. There are various parameters which are useful to investigate the cause of vehicle accident. From those parameters, this study selects useful parameters such as, types of severity, nature of crush, crush hour, driver relation with vehicle, experience of the drivers, road type, road geometry, intersection type, type of road pavement, road condition, lighting condition, weather condition and cause of accident using feature selection algorithms. In the proposed model, a random tree, J48, and Naïve Bayes, machine learning algorithms will apply on the road traffic accident data. Then this study will compare the performance of each algorithm using the collected vehicle accident data and finally the prediction model is designing using algorithms which are performing better than other three algorithms. After detail comparison and validation of different machine learning tools, the proposed study will use WEKA data mining tool for analysis. The data set for analysis is obtained from East Gojjam Police Authority.

KEYWORDS: Machine Learning; random tree classifier; cause of the accident; accident severity

I. INTRODUCTION

The identification and prediction of the cause of road accident is an important area of research as it can help in reducing the number of road traffic accidents. According to World Health Organization (WHO) reports approximately 1.25 million people die and 20-50 million people suffer non-fatal injuries each year as a result of road traffic crashes [1]. According to the road traffic accident data provided by states, Amhara records increased rapidly compares to the previous years.. However, this trend can change in future as it is hard to predict the rate at which road traffic accidents occur as it can occur in any situation. Therefore, we need to investigate the cause of road traffic accidents and its contributory factors using Machine learning paradigms.. There are a number of classification algorithms like Random forest, Random tree, j48, Naïve Bayes, REP Tree, SVM, and alike that can be used to construct models to predict the future data. In this study authors select Random Forest, J48, Random tree and Naive Bayes machine learning algorithm to design a model that predict the cause of road traffic accidents based on 145 road traffic accident records obtained

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

from East Gojjam police Authority which is recorded from 2007-2010 E.C. For each road accident record, 13 diverse attributes were collected at the time of the accident by East Gojjam police Authority. Those attributes are types of severity, nature of crush, crush hour, driver relation with vehicle, experience of the drivers, road type, road geometry, and intersection type, type of road pavement, road condition, lighting condition, weather condition and cause of accident. This dataset is pre-processed for removal of outliers, redundant and missing values. After pre-processing, the dataset contains 109 records with twelve independent variables and one dependent variable. Then the data set is decomposed into training set of 98 records (90%) and testing set of 11 records (10%) using random sampling technique without replacement. In this study, the data set is applied for each classification algorithm which is mentioned above, then we choose random forest machine learning algorithm in our study, since we get better result from this algorithm.

II. RELATED WORK

Yina Wu, Mohamed Abdel-Aty & Jaeyoung Lee [4] used logistic regression model to recognize the various factors contributing to increased vehicle crash risk during fog condition and investigate the situations in which crash is more likely to occur. This study also investigates the change of traffic characteristics and crash risks during fog conditions using actual traffic flow and weather condition data. The analysis results show that drivers would be more careful when fog is present and the chances of crash would be more near ramp areas.

Bhaven Naik, Li-Wei Tung, Shanshan Zhao & Aemal J. Khattak [11] combined two diverse sources of data and applied random parameters (mixed) ordered logit model to consider the distinct heterogeneity in the data. The results depicted that rain, air temperature, humidity, wind speed and rain are the main factors for serious injury crashes. Out of these factors, warmer air temperature and rain is associated with high severe injuries while less severe injuries are associated with higher levels of humidity.

Hasan Mehdi Naqvi & Geetam Tiwari [6] used a binomial logistic regression model to identify the various causes that influence the motorcycle crashes. The results of the study show that for variable “collision type”, chances of occurring rear-end, sideswipe and head -on collision are 42 times, 35 times and 25 times more than hit pedestrian respectively; for variable “number of vehicle”, chances are thrice as “single vehicle” than two or more vehicles”; and, for variable “number of lane”, the probability is more on two-lane NH than four-lane NH.

So Young Sohn and Hyungwon Shin [7] used Artificial Neural Network, Liner Regression and Decision Tree algorithm to select a set of influential factors and build up classification models for accident severity. The result of three algorithms is then compared in terms of classification performance. The analysis result depicted that accuracy does not differ significantly for each classifier while the defensive device is the powerful factor in the accident severity variation.

Velivela Gopinath et al [8] in this paper, they have been applied the Support Vector Machine (SVM), Naïve Bayes and decision tree classification algorithms on the data set and also they have been implemented Self Organized Maps (SOM), K-modes clustering algorithms. The result of the analysis shows better accuracy were achieve on the basis of casualty class and they investigate who is involved more in an accident between the driver, passenger or pedestrian. The cluster analysis result depicts SOM clustering provided better results as compared to K-modes clustering.

Sachin Kumar & Durga Toshniwal [9] used k-means clustering algorithm to investigate the high and low-frequency accident locations. They also used association rule mining algorithm to recognize the association between the various factors related to road traffic accidents at various places with changeable accident occurrences. The result shows that more accidents occur on highways, foot-travelers are more vulnerable to road accidents at roads that have intersections, curves on roads bordered by agriculture land are risky for multi-vehicle crashes and intersections on roads which fall upon marketplaces are more vulnerable to severe accidents.

Wang, Yubian and Wei Zhang [10] used logistic regression model to identify and analyse the impact of various roadways and environmental factors on the traffic crash severities and predict the accident severity levels. Authors investigated that factors like crash location, road function class, road alignment, light condition, road surface condition and speed limit have significant impact on crash severity. The results show that higher crash severity is linked with

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

rural roadways, major arterials, locations without intersection, locations with curves, night-time driving, dry roadway conditions and high-speed limits.

Dursun Delen, Leman Tomak, Kazim Topuz & Enes Eryarsoy [11] focused on recognizing the person, vehicle and accident-related risk factors that are significant in building a variance in automobile crashes. The authors used a number of predictive analytics algorithms to find the composite associations between various stages of injury severity and the risk causes related to crash. The authors also found the importance of crash-related risk factors after applying a systematic series of information fusion-based sensitivity analysis on the trained predictive models. Sensitivity analysis results prove that use of a preventive system (e.g. seatbelt), nature of crash and drug usages are the main predictors of the injury severity.

III. METHDOLOGY

In the proposed study, WEKA 3.8 data mining tool is used to analyze and predict cause of road traffic accidents. There are a number of classification algorithms like Random forest, Random tree, j48, Naïve Bayes, REP Tree, SVM, and alike that can be used to construct models to predict the future data. In this study authors select Random Forest, J48, Random tree and Naive Bayes machine learning algorithm to design a model that predict the cause of road traffic accidents based on 145 road traffic accident records obtained from East Gojjam police Authority which cover accident data from 2007-2010 E.C. For each road accident record, 13 diverse attributes were collected at the time of the accident by East Gojjam police Authority. The various steps followed to build the prediction model are:

Step1: Road traffic accident data is collected from east Gojjam Police Authority.

Step2: Data pre-processing and it includes tasks like data cleaning, data integrating, data reduction and data transformation.

Step 3: Split cleaned data set into a training set and testing set.

Step 4: Select the appropriate data mining technique according to the nature of problem and target variables.

Step 5: Apply the desired algorithm to the cleaned data set.

Step 6: Compare the performance of each classification algorithm and build the prediction model using the algorithm that outperforms other classifiers.

Step 7: Evaluate the prediction model using testing set data and other evaluation techniques.

Step 8: Result interpretation and conclusion

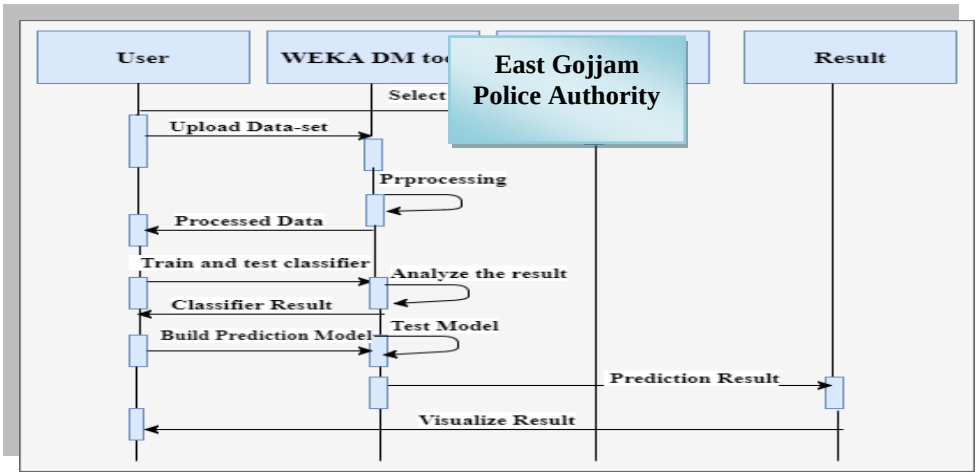

Figure 1. Sequence diagram of proposed study

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

Attribute	Distinct values	Attribute value	Code
Type of severity	5	fatal	1
		Serious injury	2
		Slight injury	3
		Non-injury	4
		Property damage	5
	4	Over turning	1
		Vehicle to pedestrian	2
		Vehicle to animal	3
		Vehicle to vehicle	4
Crush hour	6	1:00-4:59	T1
		5:00-8:59	T2
		9:00-12:59	T3
		13:00-16:59	T4
		17:00-20:59	T5
		21:00-24:59	T6
Day of the week	7	Monday	-
		Tuesday	-
		Wednesday	-
		Thursday	-
		Friday	-
		Saturday	-
		Sunday	-
Driver R/S with vehicle	3	Employee	1
		Private	2
		rent	3
Road type		One way	1
		Two way	2
Road geometry	4	steep	1
		Straight	2
		upward-gradient 3	
		downward-gradient	4
Road Intersection	2	No intersection	1
		Curve	2
Type of road pavement	2	Gravel	1
		Asphalt	2
Road condition	2	Dry	1
		Wet	2
Lighting Condition	2	Day	1
		Night	2
Weather condition	2	Fine	1
		Fog	2
Cause of accident	4	Overturning	1
		Over-speed	2
		Overloading	3
		Driving without license	4

Table I. DESCRIPTION OF DATASET AND ITS ATTRIBUTES WIT

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

ANALYSIS

A. IMPACT OF 'ROAD GEOMETRY' ON CAUSE OF ACCIDENTS

The results of road geometry on the causes of accidents (predicted class) are shown in Fig. 4. It is observed that accidents that occurred due to over speeding and overloading is higher when the road geometry is straight and upward gradient compared to other road geometry.

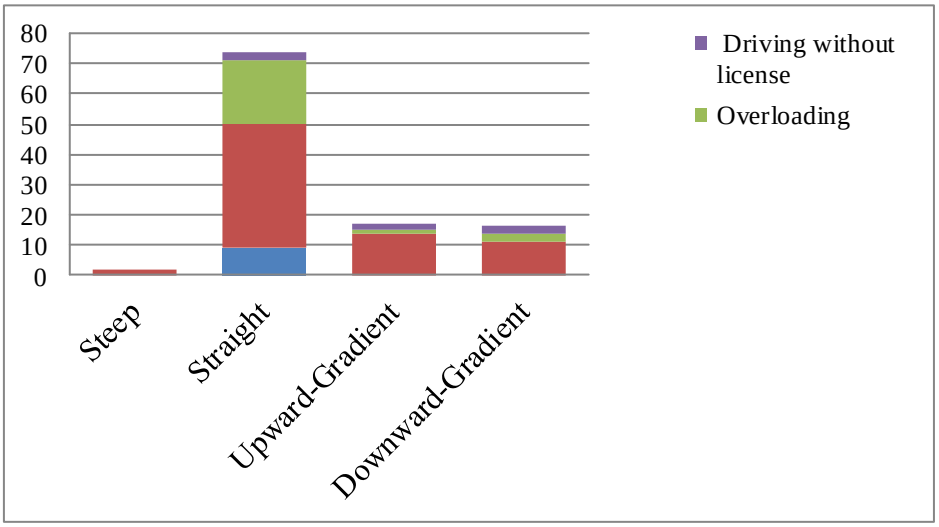

Figure 4. Impact of road geometry on causes of accidents (predicted classes)

B. IMPACT OF DRIVER R/S WITH VEHICLE' ON CAUSES OF ACCIDENTS

The results of driver relationship with vehicle on the causes of accidents (predicted class) are shown in Fig. 4. It is observed that accidents that occurred due to over speeding and overloading is higher when the driver is an employed compared to privet and rent drivers.

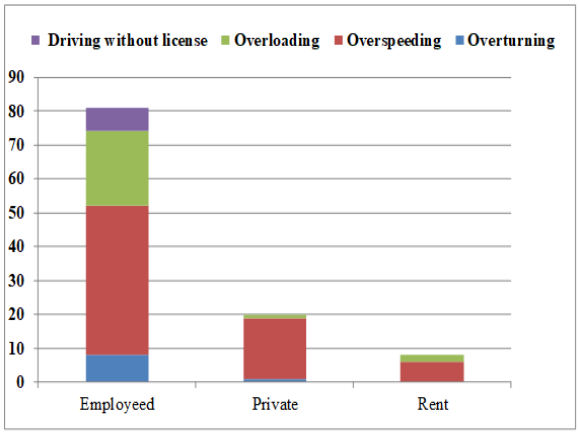

Figure 5. Impact of driver r/s with vehicle on cause of accidents

C. IMPACT OF 'ACCIDENT SEVERITY' ON CAUSES OF ACCIDENTS

The results of on the causes of accidents (predicted class) are shown in Fig. 4. It is observed that accidents that occurred due to over speeding and overloading cause higher serious injury and property damage compare to fatal

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

severity.

.

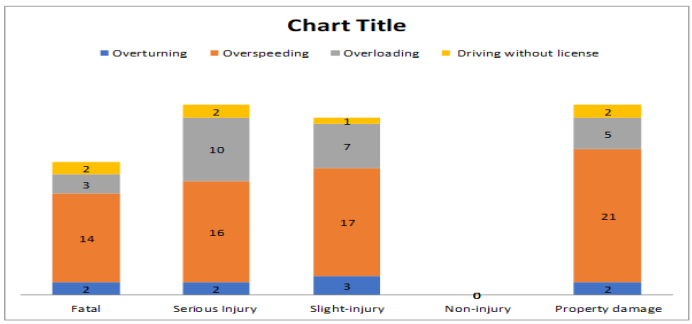

Figure 6. Association of Accident severity and cause of accidents

D. IMPACT OF 'TIME' ON CAUSES OF ACCIDENTS

The results of the impact of time on the causes of accidents are shown in Figure 7. It is observed that accident occurring rate is different in different time intervals. For variable "Time", it is observed that the accident occurred due to over-speeding and overturning causes accident with time interval in descending order of: **13:00-16:59, 9:00-12:59**, 17:00-20:59, 5:00-8:59. It can be concluded that the time intervals T4and T3 have a higher impact on the predicted class as compared to T5 and T2 time intervals. But T1 and T6 have no accident record.

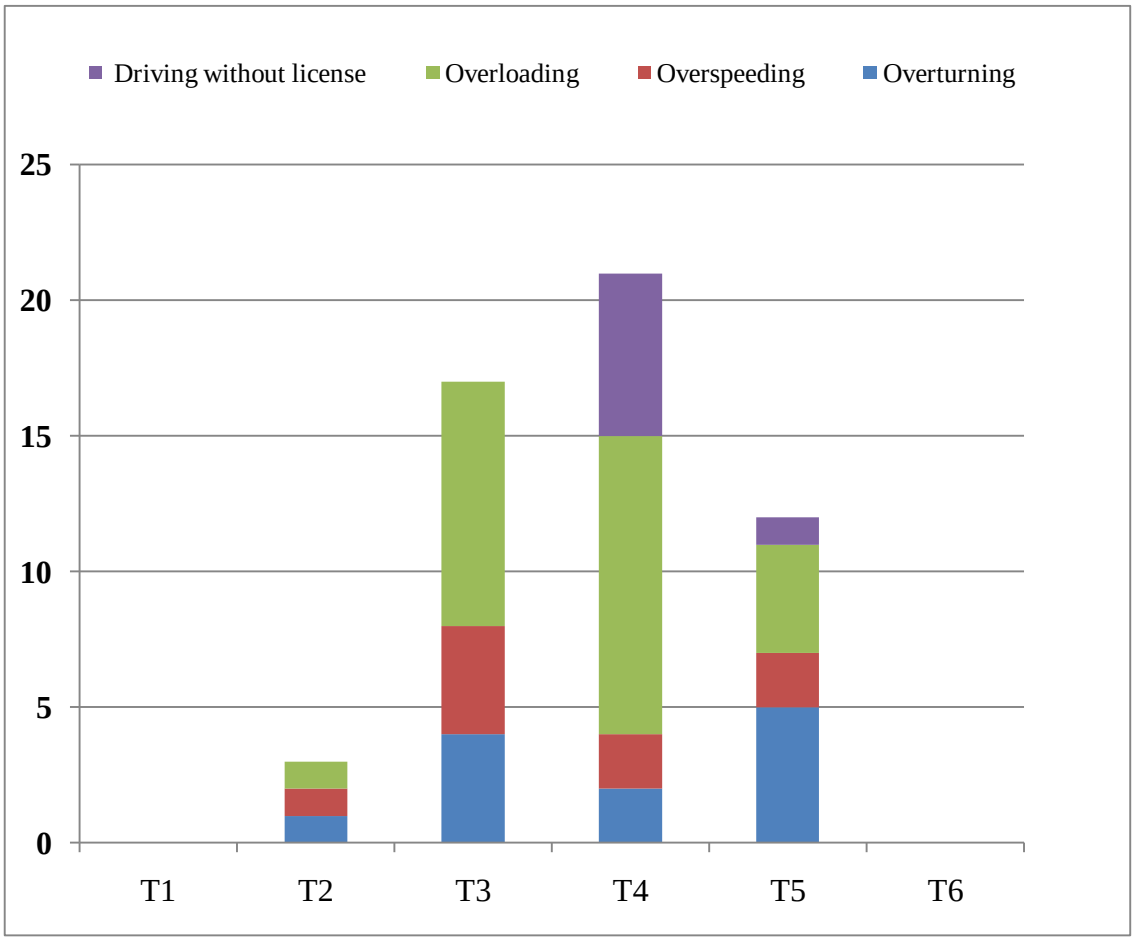

Figure 7. Impact of time on causes of accidents (predicted classes)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

IV. CLASSIFCATION RESULT

Random tree classifier is built using the data obtained after pre-processing i.e. 109 records with twelve independent variables and one dependent variable. The data set is then decomposed into a training set of 98 records (90%) and testing set of 10 records (10%). As shown in Table II out of 98 instances, all 98 instances are correctly classified giving 100% accuracy rate. Some of the various evaluation measures and their values for random tree classifier are:

A. TP RATE = RECALL= SENSETIVITY

It is the fraction of the relevant instance or documents that are successfully classified. It can be calculated as:
TP rate/Recall= TP/P or TP/TP+FN

B. FP RATE

FP rate is the proportion of all negatives that still yield positive test outputs. It can be calculated as:
FP rate= FP/N or FP/FP+TN

C. ACCURACY

Accuracy means the exactness of a predicted value to a known value. It can be calculated as:
Accuracy= (TP+TN)/(P+N)=(TP+TN)/(TP+TN+FP+FN)

D. PRECISION

Precision means positive predicted values, it can be calculated as:
Precision =TP/TP+FP

E. F-MEASURE

It measures the similarity of the known value and Prediction value distributions.
F-Measure =2(Precision*Recall)/ (Precision +Recall)

Table II. EVALUATION MEASURES FOR RANDOM TREE CLASSIFIER

```
=== Summary ===

Correctly Classified Instances          98               100      %
Incorrectly Classified Instances         0                 0      %
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               98

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 1.000    0.000    1.000      1.000    1.000      1
                 1.000    0.000    1.000      1.000    1.000      2
                 1.000    0.000    1.000      1.000    1.000      3
                 1.000    0.000    1.000      1.000    1.000      4
Weighted Avg.    1.000    0.000    1.000      1.000    1.000

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  9  0  0  0 |  a = 1
  0 60  0  0 |  b = 2
  0  0 19  0 |  c = 3
  0  0  0 10 |  d = 4
```

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 12, December 2019

V. CONCLUSION

In this study, prediction of the cause of road traffic accidents is done by considering various parameters like types of severity, nature of crush, crush hour, driver relation with vehicle, experience of the drivers, road type, road geometry, intersection type, type of road pavement, road condition, lighting condition, weather condition and cause of accident using random tree classifier. In this study, we have built the prediction model to identify the causes of accidents that can help a lot of drivers, traffic police, society, government and others alike. Through the work performed, we understood that data never seems to be enough to make a strong decision. If more data like driver related data, weather data, road condition, vehicle related data, accident severity level data, mileage data with dialed attribute values are available then more tests could be performed and more recommendations could be made from the data. Also, it is possible to improve the prediction model performance of random tree classifier by using ensemble and hybrid methods.

REFERENCES

[1] WHO. Road traffic safety. http://www.who.int/ media Centre/ factsheets/fs358/en, 2017.
[2] 400 road deaths per day in India. Up 5% to 1.46 lakh in 2015. http://timesofindia.indiatimes.com/lifestyle/health-fitness/health-news/400-road deaths-per-day-in-India-up-5-to-1-46-lakh-in 2015/articles how/51920988.cms, 2017.
[3] Yina Wu, Mohamed Abdel-Aty and Jaeyoung Lee. 2017. Crash risk analysis during fog conditions using real-time traffic data. Accident Analysis & Prevention.
[4] Bhaven Naik, Li-Wei Tung, Shanshan Zhao and Aemal J. Khattak. 2016. Weather impacts on single-vehicle truck crash injury severity. Journal of Safety Research.
[5] Hasan Mehdi Naqvi and Geetam Tiwari. 2017. Factors Contributing to Motorcycle Fatal Crashes on National Highways in India. Transportation Research Procedia, 25, 2089-2102.
[6] So Young Sohn and Hyungwon Shin. 2001. Pattern recognition for road traffic accident severity in Korea. Ergonomics, 44:1, 107-117.
[7] Velivela Gopinath et al. 2017. Traffic Accidents Analysis with respect to Road Users using Data mining Techniques. International Journal of Emerging Trends & Technology in Computer Science, 6(3), 2278-6856.
[8] Sachin Kumar and Durga Toshniwal. 2016. A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24(1): 62-72.
[9] Wang, Yubian, and Wei Zhang. 2017. Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities. Transportation Research Procedia, 25, 2124-2130.
[10] Dursun Delen, Leman Tomak, Kazim Topuz & Enes Eryarsoy. 2017. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. Journal of Transport & Health, 4, 118-131.
[11] Classification. Prediction and clustering. http://shareengineer. blogspot.in /2012/09 /classification-and-clustering.html, 2017.
[12] Sewaiwar, Purva, and Kamal Kant Verma. Comparative Study of Various Decision Tree Classification Algorithm Using WEKA. International Journal of Emerging Research in Management &Technology. 4 (2015): 2278-9359.
[13] Rapidminer.Random tree. https://docs. Rapidminer.com/studio/ operators/modeling/ predictive/trees/random_tree.html accessed, 2017.
[14] Zhao, Yongheng, and Yanxia Zhang. 2008. Comparison of decision tree methods for finding active objects. Advances in Space Research. 41.12, 1955-1959.
[15] Kenneth Lai and Narciso Cerpa. 2003. Support vs. confidence in association rule algorithms. In Proceedings of the OPTIMA Conference, Curicó.
[16] Aher, S. B., & Lobo, L. M. R. J. 2012. A comparative study of association rule algorithms for course recommender system in e-learning. International Journal of Computer Applications, 39(1): 48-52.
[17] L. Zhichun and Y. Fengxin. 2008. An improved frequent pattern tree growth algorithm. Applied Science and Technology, vol. 35, no. 6, pp. 47–51.
[18] C. Jun and G. Li. 2013. An improved FP-growth algorithm based on item head table node. Information Technology, vol. 12, pp. 34–3.
[19] B. Zheng and J. Li. 2008. An improved algorithm based on FP-growth. Journal of Pingdingshan Institute of Technology, vol. 17, no. 4, pp. 9–12.