ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter46892020.pdf
https://doi.org/10.30534/ijeter/2020/46892020

Predicting Factors of Vehicular Accidents using Machine Learning Algorithm

Aklilu Elias Kurika[1], Irfan Ahmad Ganie[2], Yuliyanti Kadir[3], Patrick D. Cerna[4], Frice L. Desei[5]
[1]Lecturer, Department of Information Technology, Wolaita Sodo University, Ethiopia, akliluelias123@gmail.com
[2] Irfan Ahmad Ganie, Department of Electrical Engineering, Indian Institute of Technology Jodhpur, India, ganieirfan27@gmail.com
[3]Yuliyanti Kadir, Assistant Professor, Department of Civil Engineering, Universitas Negeri Gorontalo, Indonesia
[4] Professor, Technology, Engineering and Research, PSHS-CRC, Philippines, pcerna@ieee.org
[5] Assistant Professor, Department of Civil Engineering, Universitas Negeri Gorontalo, Indonesia

ABSTRACT

Vehicle traffic accident is one of the major agenda for the government in which special attention has been given to continuously reduce its occurrence and related risks. Wolaita zone is one of the major areas in which increased vehicle traffic accident occurs. Government and concerned bodies have given special attention to reduce accident rate in the country. By having this point as the motivating factor for study, this work tried to predict factors of vehicle accidents by using machine learning algorithms. We used unbalanced datasets with 1611 instances, which was seven years data from year 2012-2019. In order to analyze data and evaluate patters of datasets, KDD process model was applied. The learning algorithms applied for experiments were J48 decision tree, Random forest tree, Rep tree, Naïve Bayes and Bayesian network classifiers. The experimental results, model evaluation and performance measurement shows that F-measure of J48 and Rep tree classifiers are comparatively similar i.e. 97.87% and 97.80% respectively and Random Forest tree performed less i.e. 90.9%. We identified the first experiment of J48 tree as the best model by performance and 23 best rules were generated from this experiment; best features were also identified. The most common victims, most commonly participated vehicles in accident and black spot areas for frequent accidents occurrences were identified. The findings of this study are significant for road and traffic authority and police commission for the revision and endorsement of the rules, regulations and standards related to traffic accidents; and therefore vehicle traffic accidents and related risks can be reduced generally in our country Ethiopia and specially at Wolaita Zone. We made accident data ready for further analysis in order to get most important patterns of datasets for any future researchers.

Key words : About four key words or phrases in alphabetical order, separated by commas.

1. INTRODUCTION

Road or vehicle traffic accident is a universal problem [1] and worldwide reports show that on average, more than four million peoples die because of many reasons in one year. Among this numbers, HIV AIDS and tuberculosis are the first and second cases for the deaths and vehicle traffic accident is the third known case for those dying on every day.

According to WHO and World Bank [2] in 2004, World Health Day, organized by the World Health Organization for the first time be devoted to Road Safety. Every year, according to the statistics, 1.2 million people are known to die in road accidents worldwide. The study conducted on Guardian [3] also shows that in the 2020, vehicle traffic accident will become the first factor that causes the death of human beings in the world. More than half the people killed in vehicle traffic crashes are young adults aged between 15 and 44 years often the breadwinners in a family. Furthermore, road traffic injuries cost low income and middle-income countries between 1% and 2% of their gross national product; more than the total development aid received by these countries WHO and World Bank [2]. A lot of researches were conducted on accidents from time to time in every parts of the world to reduce the accident rate and they used their own view on accident data according to their respective areas and country perspectives.

Even though plenty of researches were conducted, vehicle traffic accident increases rapidly and results in massive loss of humans' life, materials damage and other equivalent losses. WHO and World Bank [2] show that worldwide, an estimated 1.2 million people are killed in road crashes each year and as many as 50 million are injured. Projections indicate that these figures will increase by about 65% over the next 20 years unless there is new commitment to prevention. The increased loses and related injuries cause various problems to the economic development of respective countries. According to the perspectives of different countries, there are different kinds of attributes and contributing cases of the traffic accidents. The accident risk factors are more over determined in the developed countries and some preventive measures have been taken to reduce the risk. But traffic accident risks, related material damages and life lose increases from time to time in developing countries. In Ethiopia, some researches has been conducted but the risk factors cannot be reduced from time to time. In the case of

Wolaita Zone, the timely recorded data realities on ground show that traffic accident is the major issue that should be given special attention. The reason is that the risks of traffic accidents and related material and live loses show enormous increase from time to time. But the reasons for increased traffic accident factors are not well known. Additional deep analysis on accident data is indeed needed and this is also a motivating factor to conduct study by machine learning algorithms.

Generally the amount of data used by previous researchers is lesser; some others used secondary data, which is collected by questionnaire, as well as social media data for analysis. Using this kind of data for predicting factors of traffic accident is not feasible. Most of the studies that were conducted in the past literature are mainly focus on J48 decision tree algorithms. Other kinds of decision tree algorithms are not used for comparative analysis by most of the researchers. Thus, performance comparisons have not been made for more than two algorithms**.**

2. RELATED WORKS

Studies [5] and [4] are related to the locations of accident related factors; accordingly the road features are one of the contributing factors of traffic accidents. But the types of road features are not clearly specified in these studies.

Studies performed by authors [6] and [7] are comparative analysis in the performance measurement and accuracy of algorithms. The first author compared six algorithms (classification and regression tree, Random Forest, ID3, Functional trees, Naïve Bayes and J48) algorithms to determine the accidents severity level. It reveals that Naive Bayes value and J48 techniques value are approximately same in accuracy. The second one the comparative study on machine learning algorithms; the comparison has been made for decision tree and neural networks to determine factors of increased traffic injury. It comes up with that the decision trees are better than neural networks in performance.

Studies conducted by researchers explained in references [6], [7], [8] identified the factors of traffic accidents; their findings show that the causality factors are un- adopted speech, in-attention, behavior of passengers, roadway features, demographic features, environmental characters, technical characters, speed, age, gender, younger aged drivers, alcohol, less control, wrong over-taking and tire blow. These factors were identified in various areas as the contributing factors for the accidents. But it is impossible to blindly take control over all these characteristics to be considered in particular area. So accident factor analysis is needed to identify the most commonly contributing factors that hold a lion share of the commonly known determinant attributes. Some of the factors are common in one area and some other factors become common in other areas. While [9] used social media data as the primary data for predicting the causes of accident; secondary data is not suitable for analysis.

In Ethiopia, Wolaita zone is one of the most commonly known areas in which traffic accidents and related injuries take place. By analyzing the factors with learning algorithms, the most contributing factors will be determined from traffic accident data which is obtained from WZPC. Other contributing factors other than these might also be obtained for increased traffic accidents. The methodologies used by various researchers are of various types. Akinbola et al., [10] and [11] machine learning algorithms to predict the factors of traffic accidents. Both of these authors used only decision tree; and Tibebe et al., [12] is all about machine learning algorithm but it is not for determining the causes of traffic accidents and Gupta and Baluni [7] also used classification and machine learning algorithms to determine traffic injury occurrences.

3. MATERIALS AND METHODS

Classification algorithm has been identified as the best technique to attain our objectives in accordance with predetermined datasets we had. From various classification algorithms, decision tree classifiers (J48, Random Forest and Rep Tree) classifiers and from Bayesian classifiers (Naïve Bayes and Bayesian Network) classifiers were selected to conduct our experiments. We have computed 15 experiments, (three for each classifiers i.e. by 10 fold cross validation, by 66% split and by 90% split for each of them respectively.) We have identified 14 best features among 36 attributes with wrapper method.

Knowledge discovery in datasets (KDD) process modeling has been used as a study design based on Figure 1.

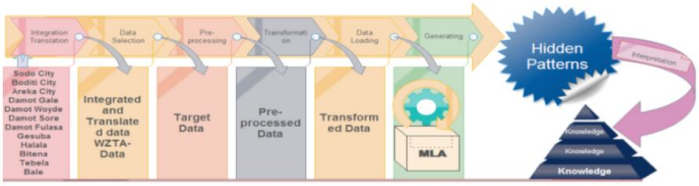


Figure 1: Knowledge discovery in datasets (KDD)

3.1 Data Integration:

To keep normal compliance of data, we integrated data to common format according our objectives and identified most important attributes to our study. Some of the attributes were ignored from the original data because they are less meaningful to our study. Accordingly, 36 important attributes were identified and 1611data was prepared for analysis, which is continuous 7 years data from 2005-2011 E.C. The amount of data was limited to this number; because five years (2000-2004) data was burned before it was being transformed to police commission from road and transport authority.

3.2 Data Selection

In order to get data for prediction, applicable data was selected from 12 districts and three city administrations of Wolaita Zone. The case study is limited to Wolaita zone only. This is because we wanted to define the scope of our study.

3.3 Data Preprocessing:

In this step the data cleaning, data reduction and data transformation has been made to prepare the best quality datasets for further analysis. The original data was obtained from Wolaita Zone police commission (PC) but, it has a lot of drawbacks such as spelling errors, unreadable data, misspelt attributes names, unknown values for some attributes and irrelevant personal representations of some terms. Some terms were inconsistent and considered to be outliers. We removed irrelevant attributes from the original Data. In this step we made the cleaning process of data before loading it to WEKA.

3.4 Data Transformation:

The original data was recorded in word processor while some data were in spreadsheet. The researcher transformed it to a .svc format which the weka workbench can read and supported.

Be aware of the different meanings of the homophones “affect” (usually a verb) and “effect” (usually a noun), “complement” and “compliment,” “discreet” and “discrete,” “principal” (e.g., “principal investigator”) and “principle” (e.g., “principle of measurement”). Do not confuse “imply” and “infer.”

Prefixes such as “non,” “sub,” “micro,” “multi,” and “"ultra” are not independent words; they should be joined to the words they modify, usually without a hyphen. There is no period after the “et” in the Latin abbreviation “*et al.*” (it is also italicized). The abbreviation “i.e.,” means “that is,” and the abbreviation “e.g.,” means “for example” (these abbreviations are not italicized).

An excellent style manual and source of information for science writers is [9].

4. EXPERIMENTATION

4.1 Most Prone Accident Vehicles

From the total 31 kinds of vehicles participated in accidents, we have identified 7 kinds vehicles as the most commonly participated. They account 75.34% and remaining 24 vehicles participation is only 24.66%. So we can conclude that if these vehicles were given separate road in cities specially Sodo-City (>25%) traffic accident can be possibly reduced.

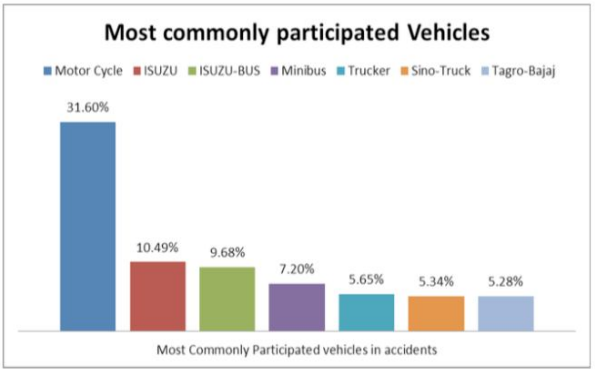


Figure 2 : Most Prone Accident Vehicles

4.2 Most Common Victims of Accidents

The above diagram shows that the most common victims of accidents are pedestrians (40.16%) and passengers (19.93%). Derives are less victims. So we can conclude that car traffic accident most commonly affects pedestrians and passengers in our case study. Males (53.8%) are most commonly affected by car traffic accidents compared to females (19.6%); which are opposite to study by [22] that revealed majority of participants as females in accidents. 18.75% of victims were aged between 1-18, 30.54% were aged between 19-30 and 18.56% were aged between 31-50.

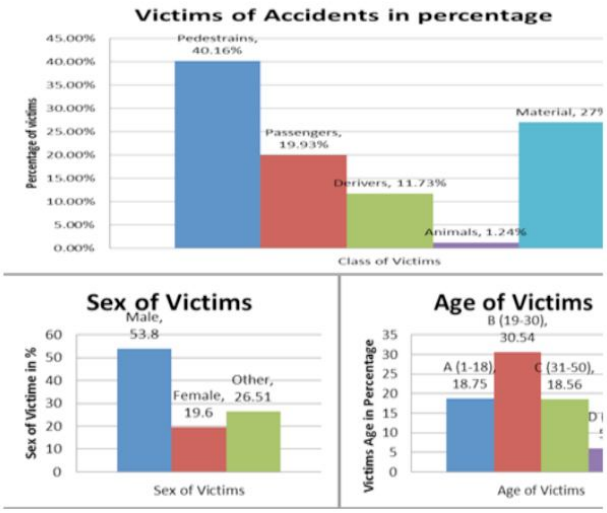


Figure 3: Most Common Vehicular Accident Victims

As it is known, the most productive human power is aged between 18 and 50. Therefore traffic accident affects the most productive classes of humans as we can conclude from the above result.

4.3 Most Common Black Spot Areas

We have selected 19 places with frequent accident occurrences from the above five Woredas. We selected areas with > = 15 accidents within 7 years. From the total accidents occurred, these places account 521 (32.34%) accidents. So concerned bodies has to give attention to these areas.

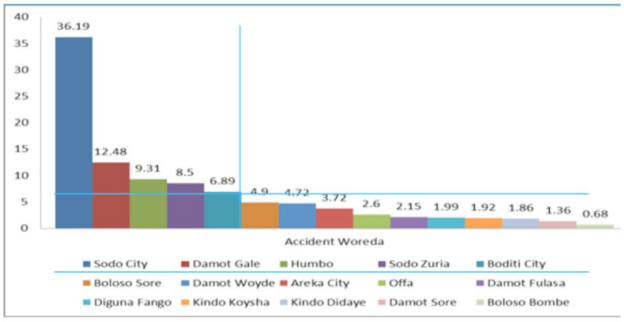


Figure 4: Most Common Black Spot Areas

From 15 different areas shown above, the first five (Sodo-city, Damot-Gale, Humbo, Sodo-Zuria and Boditi-City) account a lot accidents i.e. 73.37% of total accidents. The remaining 10 districts account only 26.63%. Each of them accounts > 5% accident occurrences from the total one, so we selected the black spot areas for frequent accidents occurrences from these five Woredas.

4.4 Determinant Cases of Accidents

The Most Determinant Cases and causality condition of Accidents are: Lack of attention (65.49%), over speed (10.62%), Prohibiting Priority (10.37%), lack of experience (6.33%) and technic failure (3.54%). The causality condition of accidents is mostly crossing the road (32.96%) straight crash (28.80%), roll down (16.70%), side to side crash (8.57%) and walking on the road (5.90%).

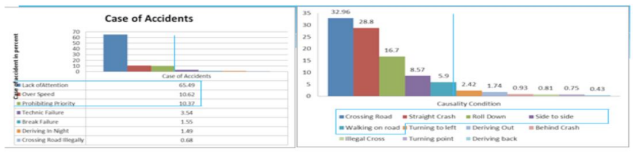


Figure 5: Determinant Cases of Accidents

Table 1: Summary of Experimental Results

Exp	Models	NL	ST	TP Rate	FP Rate	Precision	Accuracy
Exp.1	Trees.J48 -C 0.5-M 4 Testmode=10-fold Datasets=Unbalanced Attributes=All	141	145	0.984	0.030	0.984	98.45%
Exp.2	Trees.J48 -C 0.5-M 4 Testmode=Split=66% Datasets=Unbalanced Attributes=All	4	5	0.989	0.015	0.989	98.90 %
Exp.3	Trees.J48 -C 0.5-M 4 Testmode=Split=90% Datasets=Unbalanced Attributes=All	4	5	0.989	0.0014	0.989	98.91%
Exp.4	RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 Testmode=10-fold Datasets=Unbalanced Attributes=All	-	-	0.921	0.237	.926	92.12 %
Exp.5	RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 Testmode=Split=66% Datasets=Unbalanced Attributes=All	-	-	0.905	0.325	0.914	90.51 %
Exp.6	RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 Testmode=Split=90% Datasets=Unbalanced Attributes=All	-	-	0.901	0.301	0.912	90.06 %
Exp.7	trees.REPTree-M 2-V 0.001-N 3-S 1-L-1-I 0.0 Testmode=10-Fold Attributes=All Datasets=Unbalanced	4	5	0.984	0.026	.984	98.386 %
Exp.8	trees.REPTree-M 2-V 0.001-N 3-S 1-L-1-I 0.0 Testmode=Split=66% Attributes=All Datasets=Unbalanced	4	5	0.989	0.015	0.989	98.905 %
Exp.9	trees.REPTree-M 2-V 0.001-N 3-S 1-L-1-I 0.0 Testmode=Split=80% Attributes=All Datasets=Unbalanced	4	5	0.991	0.003	0.991	99.07 %
Exp.10	Bayes.NaïveBayes-output-debug-info Testmode=Split=90% Attribute= All Dataset=Unbalanced	-	-	0.946	0.068	0.948	94.60 %
Exp.11	Bayes.NaïveBayes-output-debug-info Testmode=split=66% Attribute= All Dataset=Unbalanced	-	-	0.954	0.066	0.956	95.438
Exp.12	Bayes.NaïveBayes-output-debug-info Testmode=split=90% Attribute= All Dataset=Unbalanced	-	-	0.969	0.027	0.971	96.894 %
Exp.13	Weka.Classifiers.bayes.net Testmode=10-Fold Attribute= All Dataset=Unbalanced	-	-	0.942	0.061	0.946	94.165 %
Exp.14	Weka.Classifiers.bayes.net Testmode=split=66% Attribute= All Dataset=Unbalanced	-	-	0.954	0.048	0.957	95.44 %
Exp.15	Weka.Classifiers.bayes.net Testmode=split=90% Attribute= All Dataset=Unbalanced	-	-	0.969	0.010	0.972	96.894 %

As we can see from the above experimental results and below diagram, J48 and Rep tree classifiers are comparatively similar by their accuracy. We computed average Precision and Recall of J48 and Rep tree and selected the J48 decision tree algorithm as a better than Rep tree. 1st Expt J48 tree Precision = 98% and Recall = 97.75%, (FM= 97.87%) 1st Expt. Rep tree Precision = 97.70% and Recall = 97.90%, (FM= 97.80%). The first experimental results of J48 decision tree, includes more features than exp.2 and 3 even though the number of leaves and size of tree generated are more. So we selected it as a working model and generated 23 best rules from this particular experiment.

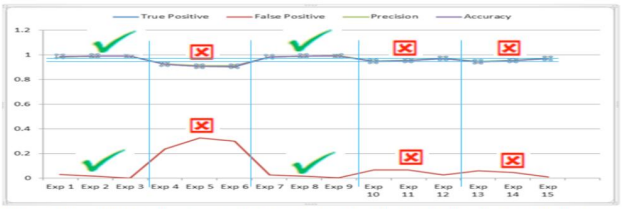


Figure 6: Diagrammatical representations of selected experiments

Below are some of the best rules generated:

1. **If** Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident □= Morning/Evening **Then Fatal in Accident: Yes.** □
2. If Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident □= Night and Number of Victims > **2: Then Fatal in Accident: Yes.** □
3. **If** Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident □= Afternoon and Type of Crashes = Vehicle With Pedestrian: **Then Fatal in Accident: No.** □
4. **If** Severity of Accident = Slight and Edu/n Level = Primary and Settlement of Road = Upward and Type of Causality Vehicle = Motor Cycle, ISUZU, ISUZU-Autobus, Minibus **Then Fatal in □Accident: Yes.** □
5. If Severity of Accident = Slight and Edu/n Level = Primary and Settlement of Road = Upward and □Type of Crashed Vehicle!= Motor Cycle **Then Fatal in Accident: No.** □

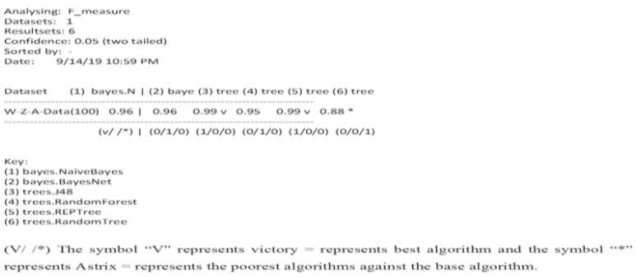
4.5 Performance Measurement of Learning Algorithms

In the experiment evaluation part, we have identified that J48 and Rep tree are comparatively similar and better that the remaining three classifiers. So we have used selected the first and third experiments for each classifiers and measured performance of their classifiers accuracy as follows.

Algorithms	Actual	Predicted		Recall	
J48 Tree		Non-fatal accidents	Fatal accidents		Accuracy
Exp.1	Non-fatal accidents	1213	11	99.10%	98.45%
	Fatal accidents	14	373	96.40%	
	Precision	98.90%	97.10%		
Exp.3	Non-fatal accidents	1217	7	99.40%	98.76%
	Fatal accidents	13	374	96.60%	
	Precision	98.90%	98.20%		
Rep Tree					
Exp.7	Non-fatal accidents	1211	13	98.90%	98.39%
	Fatal accidents	12	375	96.90%	
	Precision	99%	96.40%		
Exp.9	Non-fatal accidents	1203	21	98.30%	98.76%
	Fatal accidents	0	387	100%	
	Precision	100%	95.20%		

Figure 7: Confusion Matrix

Since the dataset we have was unbalanced, taking accuracy of the model to decide one model as best model is misleading. In such cases, it is advisable to take precision and recall for deciding whether one model is better than the other or not. In our cases, four of the experiments listed above have comparatively similar precision and recall values. But the 1st and 7th experiments were computed by 10 fold cross validation and the rest were computed by 90% split value for training and testing the model. So model with good predictive accuracy can be obtained by experiments performed with 10 fold cross validation tests according to expert judgments. Then we ignored the rest experiments with 90% split tests and accepted experiments with cross validation tests. Experiment 1st (98%) average precision and (97.75%) average recall for two class labels and 7th experiment (97.70%) average precision and (97.90%) average recall were selected to determine the best model with good predictive accuracy for fatal and non-fatal accident occurrences.


Analysing: F_measure
Datasets: 1
Resultsets: 6
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 9/14/19 10:59 PM

Dataset (1) bayes.N | (2) baye (3) tree (4) tree (5) tree (6) tree
W-Z-A-Data(100) 0.96 | 0.96 0.99 v 0.95 0.99 v 0.88 *
(v/ /*) | (0/1/0) (1/0/0) (0/1/0) (1/0/0) (0/0/1)

Key:
(1) bayes.NaiveBayes
(2) bayes.BayesNet
(3) trees.J48
(4) trees.RandomForest
(5) trees.REPTree
(6) trees.RandomTree

(V/ /*) The symbol "V" represents victory = represents best algorithm and the symbol "*" represents Astrix = represents the poorest algorithms against the base algorithm.

Figure 8: Model Evaluations

The above result shows that J48 Tree and Rep tree are significantly best by performance than all other classifiers with the given dataset. Naïve Bayes and Bayesian network classifiers are significantly good by their performance and the rest two algorithms (Random forest and Random tree) classifiers are poor by performance when compared to other classifiers with the given dataset.

5. CONCLUSION

In this study, machine Learning approaches have been applied for data analysis and prediction of car traffic accident datasets to explore important features and pattern relationships to car traffic accident occurrences. We addressed various statements of problems and objectives to determine determinant factors of car traffic accidents. We identified 7 most commonly participating vehicles, 20 areas for frequent accident occurrences, pedestrians and passengers as the most common victims and J48 and Rep tree as best algorithms by performance and model accuracy. 23 best rules were generated from the selected model for accident occurrences, results have been discussed and finally some points have been recommended for the future researchers

Based on the outcomes of this study, the following points were recommended for the future researchers. Comparatively better results might be obtained if they try accident predictions with techniques like support vector machine, multilayer perceptron and artificial neural networks. Add some unconsidered attributes to datasets and relate cases to behavior of derivers like amount of alcohol taken and mental normality of derivers to get better results. Try with deep learning with large amount of instances to get better result and integrate it with knowledge base to know cases for accident occurrences to use is as an expert system.

REFERENCES

[1] Micheale Kihishen Gebru, "Road traffic accident: Human security perspective," International Journal of Peace and Development Studies, vol. 8, no. ISSN 2141–6621, p. 16, March 2017.
[2] WHO and World Bank, "World Report on Traffic Injury Preventions," New York, 2013.
[3] Guardian. Traffic Accident Predictions. [Online]. http://politics.guardian.co.uk/homeaffairs/story/0 ,11026,1187637,00.html. 2012
[4] David Ian White, An Inverstigation of Factors Associated with Traffic Accidents and Causality Risk in Scotland. Scotland: Napier University, October 2002. [5] Durga Toshniwal2 Sachin Kumar1, A data mining approach to characterize road accident locations.: Published Online: Springerlink.com, 2016.
[6] Armit Kaur Maninder Singh, "A Review on Road accidents in Traffic system Using Data Mining Techniques," International Journal of Science and Research, p. 6, 2014.
[7] Mrs.Bhumika Gupta Pragya Baluni, "A comparative study of various Algorithms to explore factors for vehicle collision," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2012.
[8] Sani Salisu, Atomsa Yakubu, Yusuf Musa Malgwi, Elrufai Tijjani Abdullahi, I. A. Mohammed and Nuhu Abdul'alim Muhammad L. J. Muhammad, "Using Decision Tree Data Mining Algorithm to Predict Causes of Road Traffic Accidents, its Prone Locations and Time along Kano –Wudil Highway," International Journal of Database Theory and Applications, 2017.
[9] Claus Pastor, Manfred Pfeiffer, Jochen Schmidt Heinz Hautzinger, "Analysys for Accident and Injury Risk studies.," Heilbronn University, November 2007.
[10] Akinbola Olutayo2 Dipo T. Akomolafe1, "Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways," American Journal of Database Theory and Application, pp. 1-13, 2012.
[11] S. Vasavi, "Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques," in Information and Communication Technology Proceedings, Kanuru, AP, India, 2018, p. 11.
[12] Dejene Ejigu, Pavel Kromer, Vaclav Snasel, Jan Platos and Ajith Abraham Tibebe Beshah, "Mining Traffic Accident Features by Evolutionary Fuzzy Rules," IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems, 2013