ISSN No: - 2456 – 2165

A Radical Approach to Forecast the Road Accident Using Data Mining Technique

Anupama Makkar[1], Harpreet Singh Gill[2]
[1]M.Tech Scholar, Dept. of Computer Science& Engineering, SKIT, Jaipur, 302017 (India)
[2]Assistant Professor, Dept. of Computer Science& Engineering, SKIT, Jaipur, 302017 (India)
anupama.makkar@gmail.com, harpreet@skit.ac.in

Abstract:- Modern Era inventions have resulted in a better life span and a safer and better world. In spite the fact that there is a remarkable development in health status and standard across the globe. One area that still haunts human life the most is the road accidents, resulting in colossal loss of immensely precious human life every year. In this work, we avail data mining techniques to analyze the record of accident dataset in recent years. We proposed approach by using combinations of machine learning algorithms i.e. Bayes Net, j48 decision tree, j48 graft and analyze their performance in forecasting the road accidents. The experimental results reveal that the proposed approach is more suitable than the single algorithm to predict occurrence of road accidents in future. This work provides an insight to reduce the road accidents to certain extent.

Keywords:-Road Accident, Data Mining, Weka.

I. INTRODUCTION

In the recent years, the increased number of transportation has given rise to greater number of road accidents. Transport accidental costs; in terms of casualties and injuries have profound implication on the survivors. Apparently survivors, their families both have to sustain the sufferings, caused by the accident [1].

Demographic data of low income countries elaborates that they bear a large share of the burden, accounting for 85 percent of annual deaths and 90 percent of the disability-adjusted life years lost because of road traffic injury [6]. In low and middle income nations, people from lower socio economic strata are more prone to be involved in a road traffic accident [12]. Though, a lot has been done in this field, still it leaves much to be desired, to prevent these mishaps [16].

Databases are increasing in size to a level where traditional techniques of analysis and visualization of the data are cracking down [9]. Wide availability of large volume of data can be turned into useful information through data mining. The basic assumption of data mining is to analyze the data from different aspects [10].The study highlights analyzing the accident data in past records to reduce the accident rate and death rate.
Section II includes a literature review on work done by recent researchers. Section III describes the accident dataset including description of attributes. Section IV describes the methodology employed in this work by using machine learning algorithms to evaluate the performance of different algorithms. Section V describes performances of algorithms with accuracy. Section VI draws the conclusion and includes some possible areas for further work.

II. LITERATURE SURVEY

This section describes the review of published studies in recent years. Authors paid a great interest in recognizing the accident data set by using different data mining approaches to determine the cause of road accidents that significantly affects the severity of accidents.

Ayushi Jain et. al [1] focused to create a model that expelled the discrepancy of the data by clustering to group similar objects together to locate the accident prone areas and these formed clusters are labeled for using decision tree. The author attempts to identify the accident prone areas in states and territories of India against the causes that contribute to road accident and the experiences of drivers by using K-means clustering, decision tree algorithms to analyze the accident data set to improve road safety management.

Isra et.al [12] determines the causes of road accidents by using various data mining techniques to figure out the factors leads to car accidents severity in Riyad. The result of the models has comparable accuracy values. The intent of this paper lies on analyzing the road accidents records by using three classification techniques i.e.; CHAID, j48 and Naïve Bayes are used to predict the factors responsible for car accidents and also highlight crashes involving distracted drivers that can assist the authorities to implement effective measures.

Sachin Kumar et.al [4] focused on analyzing the accident data by using association rule mining algorithms and clustering algorithms to identify various factors associated with road accidents. Synthesis of k mode clustering along with association rule mining is very impressive as it provides vital data that would not be possible if no segmentation has been done before generating association rules. Then trend analysis is made for each cluster and entire data set of the road accident, finds different trends in the different cluster.

Ramya et. al [2] analyses the accident patterns from many different dimensions or angles to predict the probability of

accident to occur and also determines the death rates due to road accidents. The combination of approaches CRISP-DM(Cross-Industry Standard Process for Data Mining) and SEMMA(sample, explore, modify, model, assess) and the algorithms such as: random forest, J48, Naïve Bayes, lib SVM, Regression Analysis are used to identify the accident zones to improve road safety and draws attention in the view of road accidents statistics.

Frantisek Babic et al. [3] describes the way to apply the compiled data to extract frequent patterns and vital factors causing different accidents by using Decision Tree algorithm, apriority algorithm. Analyzing the real sample data using two alternatives i.e.; predictive mining through decision tree algorithms, and using descriptive mining by the apriority algorithm.

An Shi et. al [20] describes the traffic flow on highways, when crashes took place. The author proposed a time series model by using temporal data mining method and Clustering analysis was done to study the transport dynamics. Result of the proposed method was carried out by numerical experimentation. The newly developed method highlights the increase of traffic flow and the resultant crashes on highways.

Maninder et. al [23] analyze the accident patterns of collision occurred during road accidents that involves head on collisions between vehicles and animals, vehicles and pedestrians and vehicles and fixed obstructions due to the high speed of road traffic. To predict accident severity by using different data mining techniques that are classification and regression trees, random forest, ID3, functional tree, naïve Bayes and j48 to prevent the road accidents.

Dipo T. Akolmalage et. al [7] made attempts to analyze the cause and occurrence of crashes on highways using various data mining techniques to overcome road mishaps. They used different techniques to analyses the reasons of collision along this route and the effects of accidents. These mishaps were also explored by using decision tree: ID3, functional tree. To find the cause of road collision and accident prone locations proposed using Lagos –Ibadan Highway.

Hiwot Teshome et. al [14] determines the factors that cause road accidents to reduce the accident rate attempted the possible application of data mining to predict the relationship between driver's experience including different attributes and road traffic accident. They followed the CRISP-DM model, and machine learning tool Weka is used to implement the Naïve Bayes, J48, and PART algorithms to find the factors associated with the road accident.

Kwon OH et. al [18] Analyzing the risk factor of road accident and the result of the work was evaluated by comparing the classification methods: Naïve Bayes Classifier, decision tree classifier and Binary Logistic Regression Model.

M. Gupta et. al [15] proposed approach in which association rule mining is used to find out the severity of road accident data obtained from police records in Mujjafarnagar district, Uttar Pradesh, India. To identify the accident severity based on fatal, major injury and minor/no injury occurred during accidents. The association rules exposed different factors that are associated with road accidents in each of the categories. Author describes the way to find some hidden factors that can be used to understand the factors that cause road accident.

Dheeraj Khera et.al [8] predicts the severity of road accidents techniques by using classification and regression tress, random tree, ID3, functional tree, J48, naïve Bayes and PART over different data sets to yields the better results with accuracy. Using the different methods of data mining, the author concluded that J48 performed with better accurate results as compared to other techniques used to improve the transportation system.

Bahram Sadeghi et. al [5] contributes to determining the various factors leads to road accidents by using data mining algorithms including Market basket analysis, association rule, pattern mining that would help transportation system to overcome the accidents and analyses accident patterns to predict future behaviors.

Hence in previous work various different data mining techniques were used to analyze the record of accident dataset to improve the traffic system and to reduce factors leading to road accidents. The following table summarizes the above described work done to prevent road accidents.

Author	**Objectives**	**Data Mining Techniques**
Ayushi jain, garima Ahuja (2016)	To find the accident prone areas with respect to different accident – factors.	k-means clustering, decision tree
Isra Al-Turaiki, Maryam Aloumi, Nour Aloumi, Khulood Alghamdi (2016)	Understanding the various factors that cause car accidents by comparing the performance of classification techniques to compute the accurate results.	CHAID, J48 and Naive Bayes

ISSN No: - 2456 – 2165

Sachin kumar, Durga Toshiwal (2015)	Analysis of data to identify the factors associated with road accidents	Association rule mining algorithms and K-mode clustering algorithms
Ramya (2016)	Analysis of road accidents to predict the probability of the accident to occur.	Random forest, J48, Naïve Bayes, Lib SVM, Regression Analysis
Frantisek Babic, Karin Zuskacona (2016)	To use the collected data about road accidents to mine frequent patterns and important factors causing different types of accident	Decision tree algorithm, Apriori algorithm
An Shi, Zhang Tao, Zhang Xinming, Wang Jian(2014)	This paper describes traffic flow evolution by creating a Time series model using temporal data mining. The newly developed method highlights the evolution of traffic flow on highways	Clustering analysis, temporal data mining
Maninder Singh, Amrit Kaur (2016)	To determine the factors causes road accident that provides the methods for traffic accident prevention to predict the rate of accident severity	Classification and regression tree, ID3, random forest, functional tree, J48, naïve bayes.
Dipo T. Akolmalage, Akinbola Olutayo (2012)	Analyze the cause and occurrence of accidents prone locations	Decision tree: Id3, Functional tree
Hiwot teshome, Tibebe beshah	To determine the factors leads to road accidents in relation with training of the driver.	J48, PART, Naïve Bayes.
Kwon oH, Rheew, Yoon Y	Analyzing the risk factor of road accident by comparing the classification methods	Naïve Bayes Classifier, Decision tree classifier, Binary Logistic Regression Model.
Meenu Gupta, Vijender Kumar Solanki, Vijay Kumar Jain (2016)	To identify the traffic accidents severity by understanding the hidden factors behind road accidents	Association rule Mining
Dheeraj Khera, Williamjeet Singh (2014)	Determine the degree of injury severity in road traffic accident by evaluating the variables that contribute in severity of injury occurred during accident.	classification and regression tress, random tree, ID3, functional tree, J48, naïve bayes and PART
Bahram sadeghi Bigham (2014)	To identify the factor behind road accidents analyzing the accident patterns to predict future behaviors.	Market basket analysis, association rule, Pattern mining.

Table 2.1: Summary of Literature Review

ISSN No: - 2456 – 2165

III. DATA SET COLLECTED

The first step in data mining process is to extract the relevant information from the available database. Accident data set have been produced by data repository. To understand the accident dataset in perspective of machine learning [1]. Convert the format of selected accident data set to the required ARFF (attribute relation) format file that is machine readable format. This dataset contains 7 attributes related to road accident data i.e. Fatal, Commercial Vehicle, Alcohol, Year, gender, driver city, and accident. Dataset type can be either nominal or numeric form. The table given below describes the detailed description of attributes.

The detailed description of attributes can be viewed by the table.

S.no	Attributes	Type	Description
1	Fatal	Nominal	Describes accident that causes someone to die or not.
2	Commercial Vehicle	Nominal	Vehicle is commercial or not
3	Alcohol	Nominal	Describes that the driver has drunken or not.
4	Year	Numeric	As specified in data set i.e., Real value
5	Gender	Nominal	It includes the gender of the driver.
6	Driver City	Nominal	Describes that city in which in accident occurred
7	Accident	Nominal	Predict that accident will occur or not on the bases of certain conditions

Table 3.1: Detailed Description of Attributes.

The above described attributes are related to road accidents, each attribute contains the different data as described. So, it's required to convert the format of data into the machine understandable format (.arff) format file that is automatically converted by loading the selected data set in weka.

ARFF format includes three parts: relation, attribute and data. Attributes information in machine readable format is described below:

- @relation accident data
- @attribute Fatal {yes, no}
- @attribute Commercial Vehicle {yes, no}
- @attribute Alcohol {yes, no}
- @attribute Year {2014, 2015, 2016}
- @attribute Gender {male, female}
- @attribute Driver city {Annapolis, Rockville, Baltimore, Cumberland}
- @attribute accident {yes, no}

IV. METHODOLOGY

Data repositories are growing in size to a level where traditional techniques have become incompetent to handle huge data sets [1].

Data Mining is a cross-disciplinary area that intend to extract useful patterns of data from enormous amount of data and there are various approaches to discovering features of data sets [7]. Machine Learning is a sub-field of data science that provides an exciting set of technologies that includes practical tools for analyzing data and making predictions [8].

A. Bayesian Network Classifier

Bayesian networks are graphical models, which mean that they contain a part that can be depicted as a graph. They are directional, thus being capable of representing cause-effect relationships.

A Bayes net model can be used to calculate the probability of any event involving variables in the domain, conditional on any other event.

A Bayesian network is a directed acyclic graph in which a set of random variables makes up the nodes in the network that represent a set of random variables, $Y = Y_1, Y_i \ldots Y_n$, from the domain and a set of directed arrows connects pairs of nodes $Y_i \rightarrow Y_j$, representing the direct dependencies between variables. Where each node has a conditional probability table that quantifies the effects the parents have on the node [22].

$$P(Y_1 \ldots Y_n) = \Pi_i (P(Y_i \mid Pa(Y_i))) \quad \ldots\ldots\ldots (1)$$

Algorithm for Bayesian Network:
Algorithm: Algo_Bayes net
Input: Training Data set
Output: Structure of Bayesian network

Procedure

a). Phase I: (Drafting)

- Initiate a graph G (V, E) where V= {all the nodes of a data set}, E= { }. Initiate two empty ordered set S, R.
- For each pair of nodes (Yi, Yj) where Yi, Yj ε Y, estimate mutual information I (Yi, Yj) using equation (1). For the pairs of nodes that have mutual information greater than a certain small value ε, sort them by their mutual information from large to small and put them into an ordered set S.
- Get the first two pairs of nodes in S and remove them from S. Add the corresponding arcs to E. (the direction of the arcs in this algorithm is determined by the previously available nodes ordering.)
- Get the first pair of nodes remained in S and removes it from S. If there is no open path (open path means every node in the path is active) between the two nodes (these two nodes are d-separated given empty set), add the corresponding arc to E; Otherwise, add the pair of nodes to the end of an ordered set R.
- Repeat step 4 until S is empty.

b). Phase II: (Thickening)

- Get the first pair of nodes in R and remove it from R.
- Find a block set that blocks each open path between these two nodes by a set of minimum number of nodes. Conduct a Conditional independency test. If these two nodes are still dependent on each other given the block set, connect them by an arc.
- Go to step 6 until R is empty.

c). Phase III: (Thinning)

- For each arc in E, if there are open paths between the two nodes besides this arc remove this arc from E temporarily and call procedure find_block_set (current graph, node1, node2). Conduct a CI test on the condition of the block set. If the two nodes are dependent, add this arc back to E; otherwise remove the arc permanently.

B. J48 Decision Tree

A decision tree is a decision support tool that decides on the basis of certain conditions, a graphical model in which an internal node represents a test on an attribute, each branch denotes the outcome of the test and leaf node represents a class label distribution.J48 Decision tree is the implementation of ID3. J48 is an open source Java implementation of the C4.5 algorithm and by applying J48 decision tree on available dataset would help to predict the target variable of a new dataset record [11]. It follows the following algorithm:

Algorithm for J48 Decision Tree:
Algorithm: J48_ALGO
Input: Training dataset
Output: Decision tree

Procedure:

- In this step, on the basis of attributes value of available training data set, identify the value of the attributes and distinguish the various instances clearly.
- Selecting recursively, classify the "best attribute" to have the highest information gain.
- If there is any value for which there is no ambiguity, then we stop and assign the target value as decision attribute for the node.
- In other cases, then investigate for other attribute that have the highest information gain. Hence, we proceed in this way until, we either obtain a clear decision of the combination of attributes that denotes a target value, or run out of attributes.
- In case, we run out of attributes or if unable to get the as it is value for the target variable from the available data, we assign this branch a target value that most of the items in this branch possess.
- In this decision tree, follow the order of attribute selection as we obtained for tree. Verify all the attributes, their values with the values in the decision tree and then assign the target value of this new instance.

C. J48 Graft

Grafted Decision tree implemented in J48 Graft component of the weka machine learning workbench, grafting would allow adding nodes to the tree to increase the performance of decision tree as good as possible.

ISSN No: - 2456 – 2165

Grafting and Pruning are corresponding terms that helps in decision making process. Pruning provides more clarity and Grafting add nodes to the decision tree to get better predictive accuracy.

The Purpose of Grafting to correctly classify the instances that falls apart from the areas covered in training data and also find out the areas, with no occupancy and enhances the poor class assignments which increase the accuracy and decreases the prediction error.

Grafting is also used as a post-process when a decision tree has been built, it increase the complexity of the tree, but most significantly reduces the error of the tree [21].
The algorithm for the modified decision tree is described in simple steps:

Algorithm: J48 Graft_ALGO:
Input: Accident data set
Output: Modified decision tree

Procedure:

- Let each leaf L in the tree; explored the region represents the leaf L.
- Consider each attribute a, and find out the highest information gain from splitting on attribute a.
- Let b best be the attribute that has the highest information gain.
- Create a decision node that splits on a best.
- Recourse on the sub lists created by splitting on b best, adds those nodes as children of the node.

Adding nodes to the existing decision tree would improve the predictive accuracy, and also reduces the prediction error. Grafting at each leaf decreases the number of faults; the first level of the tree is a single header node, it is a pointer node to its sub node as shown in above structure of modified decision tree.

D. Proposed Approach

The dataset was split into training set and test set by the cross-validation technique and percentage split technique. 10-fold cross validation and Percentage split 66% is used to evaluate the results.

The workflow diagram below should give us an intuitive understanding of how they are connected. We extracted certain features from our raw data; we would now randomly split our dataset into training and a test dataset.
The training dataset will be used to train the model, and the purpose of the test dataset is to evaluate the performance of the final model at the very end as shown in flow chart.

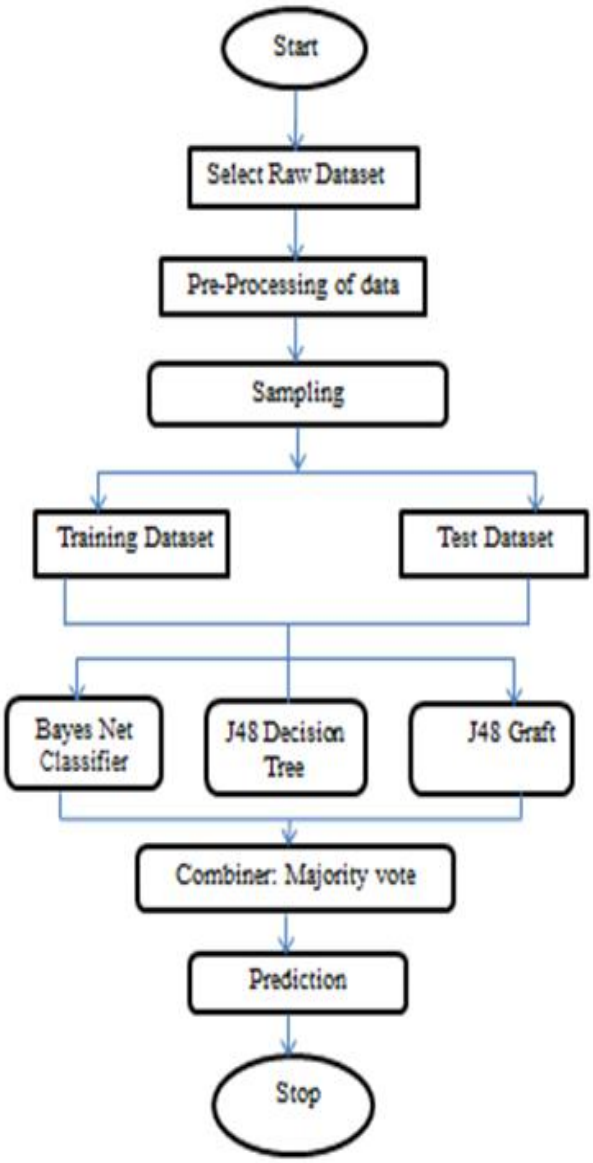

Fig 4.1: Flow Chart of Proposed Work.

V. PERFORMANCE ANALYSIS

Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if you have more than two classes in your dataset. However, this type of model takes it one step further, and it is common practice to take an entire training set and divide it into two parts: take about 66 percent of the data and put it into our training set, which we will use to create the model; then take the remaining data and put it into a test set, which we'll use immediately after creating the model to test the accuracy of our model. A major focus of this paper predicts the results of the accident that includes two nominal values yes and no. Machine learning algorithms are applied to the road accident dataset includes two aspects of training and testing. Results of each classifier

are obtained by 10-fold cross validation and 66% of split validation with accuracy.

Techniques	TP Rate	FP Rate	Precision	Accuracy
Bayesian Network Classifier	0.777	0.222	0.781	77.7%
J48 Decision Tree	0.779	0.213	0.803	77.9%
J48 Graft	0.779	0.213	0.803	77.9%
Proposed Approach	0.781	0.211	0.807	78.1%

Table 5.1: Table Illustrates the Accuracy of Data Mining Techniques.

The above table illustrates the accuracy of data mining techniques for road accident data analysis. Performance of the algorithms Bayes Net, J48, J48 Graft includes TP rate, FP rate, Precision and accuracy. The accuracy value obtained by different algorithms is in table ayes net 77.7%, J48 decision tree 77.9%, J48 Graft 77.9% and Proposed Approach 78.1%.

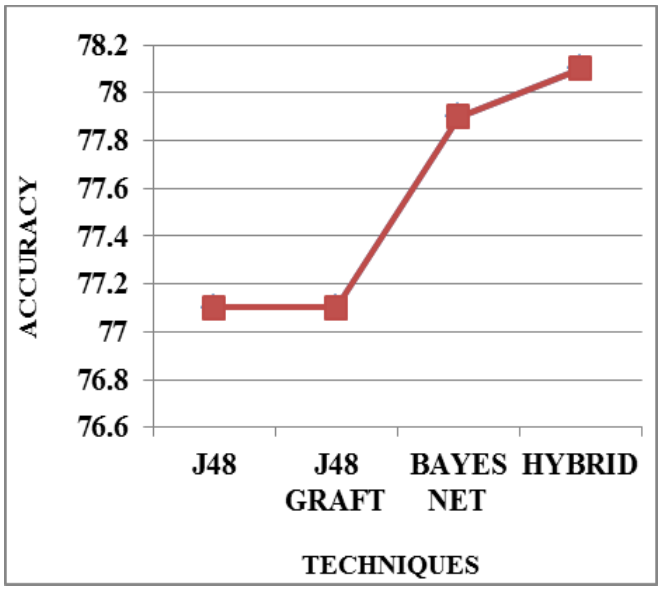

Fig 5.2: Graphical Representation of Accuracy.

The above graph represents the accuracy of each technique used in this work. Finally, the results concluded that proposed approach performed with better accuracy rate as compare to other.

The results obtained in this work would be helpful in preventing accident rate. This work gives an insight of how the analysis of traffic accident data can be done to reduce the accident rate to certain extent.

VI. CONCLUSION & FUTURE WORK

This paper endeavors to explore the application of data mining techniques on road accident by the means of using machine learning algorithms to predict accident rate in future to reduce crash deaths and injuries. A review of the literature revealed several published studies on various aspects of road crash accident data analysis implementing data mining techniques. The accident dataset for the study comprises of traffic accident reports of different cities analyzed by machine learning algorithms to predict the accident rate.
The result of this paper reveals that proposed approach performed with higher accuracy rate as compared to other techniques used. Results obtained from the work will be helpful for the prevention and control of road traffic accidents. In future, we have to increase the classification accuracy of road traffic accidents types; data quality has to be enhanced. Another future effort is to investigate the viability of other data mining techniques.

VII. ACKNOWLEDGEMENT

I owe my gratitude to all those people who have made this work possible and because of whom my experience has been one that I will cherish forever.

REFRENCES

[1]. Ayushi Jain,Garima Ahuja, Anuranjana, Deepti Mehrotra. "data mining approach to analyze the road accidents in India," International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), JULY 2016.
[2]. RamyaV, "Analysis and Prediction of Bangalore Traffic South Road Accidents," International Journal on Recent and Innovation Trends in Computing and Communication, July 2016.
[3]. Frantisek Babic, Karin Zuskacova "Descriptive and Predictive data mining on road accidents data," IEEE 14th International Symposium on Applied Machine Intelligence and Informatics, March 2016.
[4]. Sachin Kumar, Durga Toshniwal, "Analyzing road accident data using association rule mining," International conference on computing, communication and security, January 2016.
[5]. Bahram Sadeghi Bigham, "Road Accident data analysis: A Data Mining Approach," Indian Journal of Scientific Research, May 2014.
[6]. Tibebe Shah, Shawndra Hill, "Mining Road Traffic Accident Data to Improve Safety: Role of Road- related Factors on Accident Severity in Ethiopia.," Conference: Artificial Intelligence for Development, Papers from the

2010 AAAI Spring Symposium, Technical Report SS-10-01, Stanford, California, USA, January 2010.

[7]. DipoT. Akomolafe, Akinbola Olutayo, "Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways," American Journal of Database Theory and Application, 2012.

[8]. Dheeraj Khera, Williamjeet Singh, "A Review on Injury Severity in Traffic System using Various Data Mining Techniques," International Journal of Computer Applications, August 2014.

[9]. M. Sowmya, Dr.P.Ponmuthuramalingam, "Analyzing the Road Traffic and Accidents with Classification Techniques," International Journal of Computer Trends and Technology (IJCTT), november 2013.

[10]. PragyaBaluni Y.1 P. RaiwaniV, "Extraction of Road accident patterns in Uttarakhand using the neural network," International Journal of Computer Engineering and Technology, August 2014.

[11]. S.Vigneswaran, A.Arun Joseph, E.Rajamanickam, "Efficient Analysis of Traffic Accident Using Mining Techniques," International Journal of Software and Hardware Research in Engineering, March 2014.

[12]. Isra Al-Turaiki, Maryam Aloumi, Nour Aloumi, Khulood Alghamdi, "Modeling traffic accidents in Saudi Arabia using classification techniques," Saudi International Conference on Information Technology (Big Data Analysis) (KACSTIT), December 2016.

[13]. MS.R. Saravanya,Ms.Mangayarkarasi, "A Study and Analysis of Road Accident in Tamilnadu using Data mining Technique," International Journal of Recent Engineering Science (IJRES), August 2015.

[14]. Hiwot Teshome, TibebeBeshah, "Traffic Accident Analysis from Drivers Training Perspective Using Predictive Approach," HiLCoE Journal of Computer Science and Technology, December 2013.

[15]. Meenu Gupta, Vijender Kumar Solanki, Vijay Kumar Singh, "A NovelFramework to Use Association Rule Mining for classification of traffic accident severity," Ingeniería Solidaria, January 2017.

[16]. Farzaneh Moradkhani, SomayyaEbrahimkhani, BahramSadeghiBegham, "Road Accident Data Analysis: A Data Mining Approach," Indian Journal of Scientific Research, May 2014.

[17]. Jaideep Kashyap1, Chandra Prakash Singh, "Mining Road TrafficAccident Data to Improve Safety on Road-related Factors for Classification and Prediction of Accident Severity," International Research Journal of Engineering and Technology (IRJET), October 2016.

[18]. Kwon OH,Rhee W, Yoon Y "Application of classification algorithms for analysis of road safety risk factor dependencies," US National Library of Medicine National Institutes of Health, November 2016.

[19]. Hamzah Al Najada, Imad Mahgoub, "Big vehicular traffic Data mining:Towards accident and congestion prevention," International Wireless Communications and Mobile Computing Conference (IWCMC), September 2016.

[20]. An Shi, Zhang Tao, Zhang Xinming, and Wang Jian, "Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining," International Conference on Intelligent Systems Design and Engineering Applications, December 2014.

[21]. Emil Brissman, Kajsa Eriksson, “Classification: Grafted Decision Trees,” Linköping University Campus Norrköping, December 2011.

[22]. Jie Cheng, David A. Bell, Weiru Liu, "An Algorithm for Bayesian Belief Network Construction from Data," University of Ulster, U.K., April 2010.

[23]. Maninder Singh, Amrit Kaur, "A Review on Road Accident in Traffic System using Data Mining Techniques, " International Journal of Science and Research (IJSR), January 2016.