Mining Road Traffic Accident Data to Improve Safety on Road-related Factors for Classification and Prediction of Accident Severity.

Jaideep Kashyap[1], Asst. Prof. Chandra Prakash Singh[2]

[1]*Jaideep Kashyap M.tech Student at CSE department, SRGI, Jhansi, Uttar Pradesh, India*
[2] *Asst. Prof. Chandra Prakash Singh, Department of CSE, SRGI, Jhansi, Uttar Pradesh, India*

---***---

Abstract - *Road Traffic Accident is very serious matter of life. The World Health Organization (WHO) reports that about 1.24 million people of the world die annually on the roads. The Institute for Health Metrics and Evaluation (IHME) estimated about 907,900, 1.3 million and 1.4 million deaths from road traffic injuries in 1990, 2010 and 2013, respectively. Uttar Pradesh in particular one of the state of India, experiences the highest rate of such accidents. Thus, methods to reduce accident severity are of great interest to traffic agencies and the public at large. In this paper, we applied data mining technologies to link recorded road characteristics to accident severity and developed a set of rules that could be used by the Indian Traffic Agency to improve safety and could help to save precious life.*

Key Words: Traffic Accident, Data Mining, Naïve Bayes, Classification, Prediction.

1. INTRODUCTION

Our life is priceless and we are not alone we are having family who depends on us. Road Traffic Accident is very serious matter of life and it should be controlled and reduced to very low or even negligible. Road Traffic is a crucial part to life, but the numerous road accidents carry serious bodily harm and loss of property. Each side of accidents contains a large amount of information, and data is the most common form of the most important information records. Data mining has been defined as the non-trivial extraction of previously unknown, implicit and potentially useful information from data via mining the data of Road Traffic Accident, we can analysis accident distinctiveness in multi-angles, multi-level and more comprehensive, and discover potential for reduction of accident. Data mining is the science of extracting useful information from large databases. It is one of the responsibilities in the process of knowledge discovery from the database [1]. There are two primary goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the data set to calculate unknown or future values of other variables of interest. On the other hand Description focuses on finding patterns describing the data that can be interpreted by humans.

The endeavor of this study is to investigate the performance of classification method using WEKA (Waikato Environment for Knowledge Learning) focuses on Traffic Accident Dataset. Choices of classifier used for this purpose is Naïve Bayes. Naïve Bayes classifier works best in two cases, when the features are completely independent and secondly when the features are functionally dependent. The worst performance is seen in between these two extremes [2]. It [3] has also shown that choosing structures by maximizing conditional likelihood while setting parameters by maximum likelihood also yields better results.

The popularity of naïve Bayes classifier has increased and is being adopted by many because of its simplicity, computational efficiency, and its good performances for real-world problems. The aim is to improve the general performance of the model through the best selection of predictive attribute data.

Thus using Naïve Bayes classifier to yield the probability of accident may occur or not on the bases of certain condition like Driver characteristics, Environment factors, Road orientation, Weather conditions etc. This help in control and reduction of accident on road and save the life which we got only once, by checking the probability on the data provided and thus accident could be control by taking of necessary steps as this informs before any accident may take place and thus one can try to remove that factor in which there is a chance of road accident may occur. In this problem we have used Naïve Bayes algorithm to build the model in which we used free data mining software available, WEKA under the GNU General Public License.

2. LITERATURE REVIEW

The costs of fatalities and injuries due to traffic accidents have a great impact on society. In recent years, researchers have paid a great attention at determining the factors that significantly affect driver injury severity in traffic accidents.

1. Tibebe et al. [4] focused on the contribution that various road-related factors have on the accident severity Ethiopian civilians. Modeling will be to combine road-related factors with driver information for better predictions, and to find interactions between the different attributes.

2. Jianfeng et al. [5] Adopting AHP classification

comparison to analyze accident influence factors condition which are driver factors, road, environmental factors, and state of the vehicle itself. And accident main influence factors are: driving experience, overload or not, road condition, weather conditions, etc.

3. Evgenikos, et al. [6] focused on exposure data related to the mobility of these vehicles (vehicle fleet, veh-kms, passenger-kms travelled), driving fatigue one of the most important accident factors related to the long distance lorry driving. Furthermore, the macroscopic analysis presented in this paper impact on HGVs and buses/coaches road safety and the underlining reasons behind their casualties.

4. Goel [7] focused over speeding/driver's fault (87 to 88%) should be checked by strict enforcement. It is observed that trucks/canter/buses are involved in maximum accidents (42%) followed by car/jeep (35%), 2-wheeler (13%) and others (10%). Buses account for 6% accidents. Enforcement measure should specially focus on the road truck/canter/bus as they are found in maximum number of accidents. More accidents take place during day time (61%) than in night time (39%). This may be attributed to less number of cars and 2-wheelers during night.

5. Batrakova et al. [8] synthesis of studies in the theory of the interaction of the driver with traffic environment and traffic safety on the roads, taking into account not only the technical capabilities of the car, but also psychological and physiological peculiarities of the perception of the driver of the road environment. Ensuring maximum reliability of the driver's activity and is most likely to hold the set speed is an effective tool for improving traffic safety at the design stage of roads.

6. Helen, et al. [9] focused on Naïve Bayes algorithms have good performance for classifying short, noisy snippets of text and are a simple, practical, easily implemented approach. This study shows that filtering with two Naïve Bayes models to selectively guide manual review successfully generated an unbiased estimate of the frequency of injury causation/events have observed that this simple approach is effective, but further improvement in overall performance may be attainable with other classifiers.

7. Taylor, et al. [10] showed that an important point is that to optimize the predictive models additional steps could be taken that would be likely to improve the performance of the model, such as increasing the sample size, trimming the word dropping common noise words to improve performance of the Naïve Bayes model.

8. Wong et al. [11] used a comparison of methodology approaches to identify causal factors of accident severity. Adopting the Taiwan single-auto-vehicle accident data set, the results indicated that accident fatality resulted from a combination of unfavorable factors, rather than from a single factor. Moreover, accidents related to rules with high or low support showed distinct features.

9. Chong, et al. [12] the critical factors influencing injury severity numerous data mining-related studies with results frequently varying depending on the socio-economic conditions and infrastructure of a given location.

10. Ali et al. [13] identifies most important factors which affect injury severity the crash data from the records of the Information and Technology Department of the Iranian Traffic Police from 2006 to 2008. The results indicated that seat belt is the most important factor associated with injury severity of traffic crashes.

11. Brijesh et al. [14] presents a single-vehicle crashes were extracted from the road traffic accident data between January 2004 and May 2008 in Beijing. The results shows that cause factors of Single vehicle crashes are lighting conditions, vehicle type, driving experience, whether wearing seat belt or not that affect the accident severity.

12. Ting, et al. [15] focused on Results showed that Naïve Bayes is the best classifiers against several common classifiers such as decision tree, neural network, and support vector machines in term of accuracy and computational efficiency. Even improve the time used to train and general the model, as proven in the experiment section.

3. PROPOSED WORK

In this work we have taken a Traffic Accident dataset contains thirteen columns (or attributes) i.e. Driver, Environment, Road, Vehicle, Location etc. Among these attributes, the automobiles plate number and name of drivers were not disclosed by the accident database for privacy purposes. These attributes were categorized with their data types displayed in table 1.

Table 1. Complete Description of the Attributes in the Dataset

S. No	Attribute Name	Contains	Description
1	Driver	Drunken, Etiquette, Sleepiness, Fatigue, FaultyPreparation, IgnoranceHighwayCode, TrafficOrders, Overspeeding, OverTaking	Detail of the driver causing the accident
2	DriverType	Experienced, learner	Whether driver have experienced about

			driving or not
3	Environment	Rain, Fog, Mist, Sunrays	Outlook of weather when accident occurred
4	Road	RDesign, PoliceChecks, Sandy, DrainageSystem, RQualityMaterial, RRatioMixtures	Condition of the road
5	LightCondition	DimLight, NoLight, Fluctuating	Street Light condition
6	Vehicle	VDesign, VBody, Breaksystem, Tyres, HeadLight, BackLight, Indicators, Engine	Vehicle condition caused accident
7	Automobile1	Truck, Car, 3Wheeler, Byke, Bicycle, Padestrian, Train	Accident between vehicles
8	Automobile2	Truck, Car, 3Wheeler, Byke, Bicycle, Padestrian, Train	Accident between vehicles
9	Animals	Cow, Dog, Pig, Donkey, Cat, Snake, Buffalo, Elephant	Accident may due to animal comes in front of vehicle
10	AccidentType	Fatal, Injury, Normal, Death, PhysicallyDisabled, PropertyLoss	Impact of accident
11	SeatBelt	Yes, No	Whether wearing seat belt or not
12	Location	City, Market, Highway	Location of accident
13	PredictAcc	Yes, No	Class attribute

Using WEKA a windows application tool that had executed Naïve Bayes algorithm for prediction and classification. The system is designed to read input (.arrf) file contains all the selected data by the user. It depends on user to apply cross validation method for classification or apply a test file for the prediction. The dataset which was input by the user, after successfully read, dataset was classified using k-fold cross-validation method for classification. Naïve Bayes algorithm splits dataset into training set and test set, this training set was used to calculate prior probability for each class. The conditional probabilities for each feature value is calculated using a single instance from test set. Using the training set, the prior probabilities of each class is calculated.

It takes less time in classifying the dataset. The workflow of the Naïve Bayes classifier have shown in the flowchart as follows (Figure 1). The conditional probabilities for each feature value is calculated using a single instance from the test set. Then for calculation of the posterior probabilities for each class these values are used. The class with the highest posterior probability is assigned as the class for that test instance. This process is done on each instance in the test set.

The number of correct classification is obtained by comparing these assigned class values to the actual class values of the test data, which is then used to calculate the accuracy of the classifier. For k-fold classification, the dataset is split into the number k entered by the user. The entire process from creating training set and test set to calculating the accuracy is performed k times using each set as the test set in each iteration. The training set is formed by merging the remaining k-1 sets. The accuracies obtained from all iterations are averaged to get the accuracy of the classifier.

Similarly when user want to predict the value in the test file by apply method to supply a test file for prediction, test file is also in .arrf file format which can be input from supply test file panel it will check whether file is in correct readable

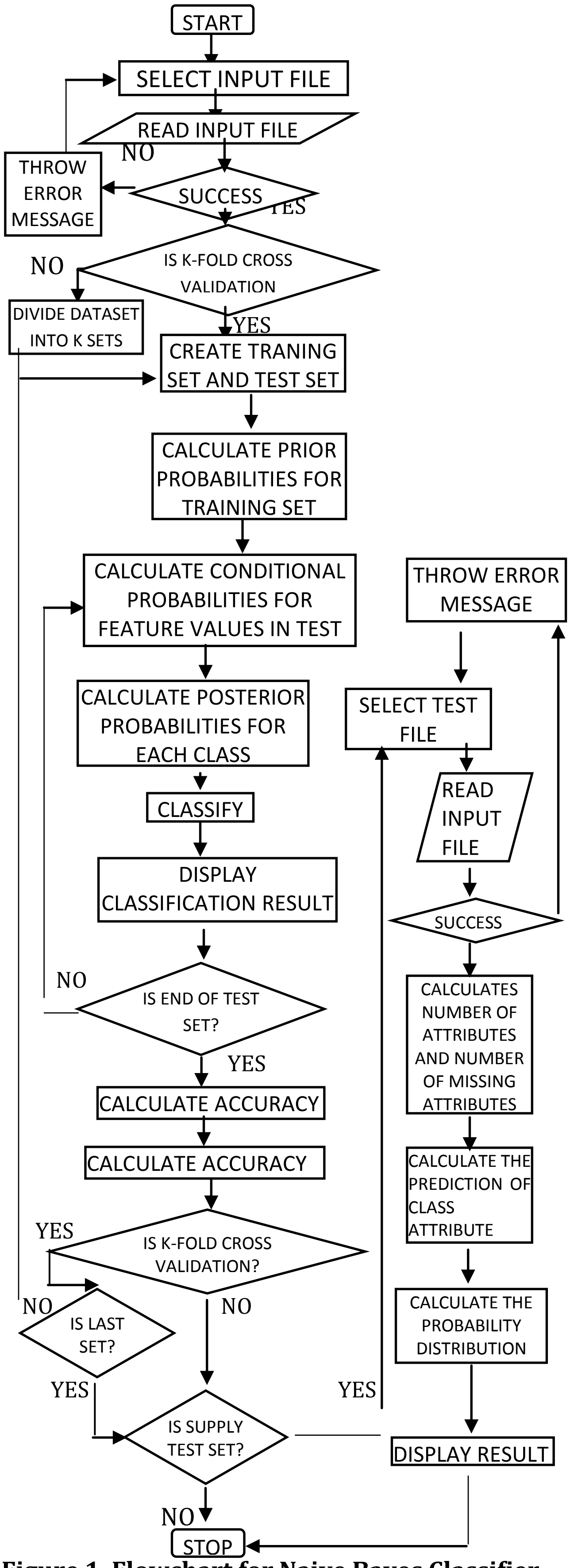

START
SELECT INPUT FILE
READ INPUT FILE
NO
THROW ERROR MESSAGE
SUCCESS
YES
NO
IS K-FOLD CROSS VALIDATION
DIVIDE DATASET INTO K SETS
YES
CREATE TRANING SET AND TEST SET
CALCULATE PRIOR PROBABILITIES FOR TRAINING SET
CALCULATE CONDITIONAL PROBABILITIES FOR FEATURE VALUES IN TEST
THROW ERROR MESSAGE
CALCULATE POSTERIOR PROBABILITIES FOR EACH CLASS
SELECT TEST FILE
CLASSIFY
READ INPUT FILE
DISPLAY CLASSIFICATION RESULT
SUCCESS
NO
IS END OF TEST SET?
CALCULATES NUMBER OF ATTRIBUTES AND NUMBER OF MISSING ATTRIBUTES
YES
CALCULATE ACCURACY
CALCULATE ACCURACY
CALCULATE THE PREDICTION OF CLASS ATTRIBUTE
YES
IS K-FOLD CROSS VALIDATION?
NO
IS LAST SET?
NO
CALCULATE THE PROBABILITY DISTRIBUTION
YES
YES
IS SUPPLY TEST SET?
DISPLAY RESULT
NO
STOP

Figure 1. Flowchart for Naive Bayes Classifier

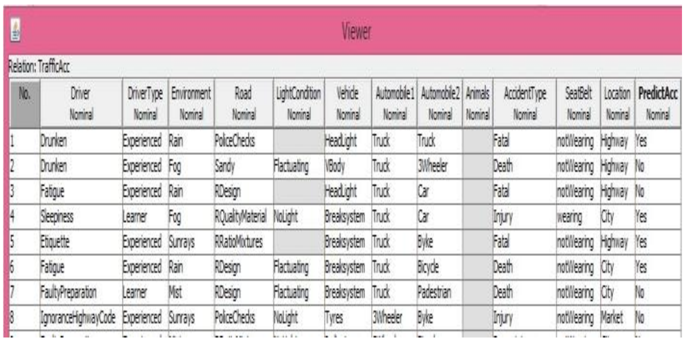
format if not it will throw an error file cannot be read, if file is read successfully then it proceed to further calculation for prediction, the outcome of the prediction will be calculate by Naïve Bayes algorithm using the dataset classified value of the class attribute. From these classified values it calculate predicted result as well as probability distribution value is also calculated and displayed the output on screen.

Thus for the prediction we have supplied a test set file having all attributes except class attributes which is going to predicted that whether in given conditions, what is the probability of an accident to be occurred.

4. EXPERIMENTATION

To predict accident severity, using classification algorithm Naive Bayes classifier. After assessing the data from dataset and selecting the predictive model to be used, a series of experiments were performed which yield different result and accuracy. Extensive data pre-processing resulted in a clean the dataset containing 1, 31,698 accidents instances with missing values. The class label ('PredictAcc') had two nominal values: 'Yes' and 'No'. During data exploration, different numbers of attributes were selected by different feature selection techniques which are provided in WEKA. Naïve Bayes classifier is selected in WEKA for classification and prediction that whether the accident may occur (Yes) or not (No) with 7 attributes, including 6 independent variables and one dependent variable (the class-label attribute 'PredictAcc'), were fed to explorer of WEKA. After that Naïve Bayes classifier was used, and an accuracy of 87.2527 % was achieved. In the second experiment, the number of attributes were increased to 8, including 7 independent variables and one dependent variable, an accuracy of 88.0613 % was achieved. In the third experiment, the number of attributes were increased to 13, including 12 independent variables and one dependent variable and accuracy of 89.4554 % was achieved.

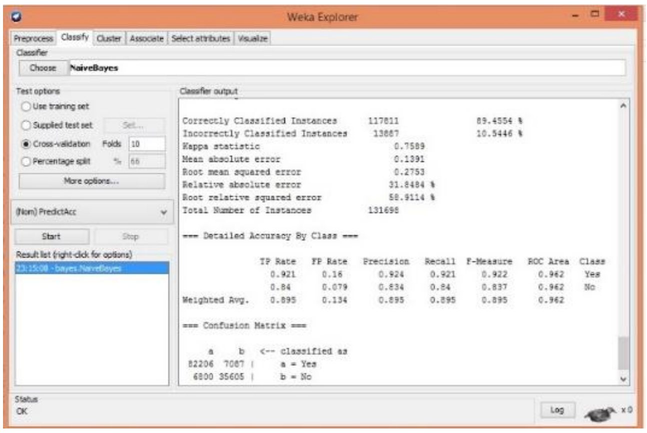
Experimental results of Naïve Bayes classifier are discussed in this section using the data mining tool WEKA. Traffic Accident data contains accident severity which represents the severity of the accident. The six kinds of accident severity are Injury, Fatal, Physical disabled, Property loss, Normal and Death. The data source for this research consumption is collected from various people that have been suffered from accident by answering the questionnaire, from municipal police department, towns and cities. The data stored in dataset in .arff file format. Dataset contains partial road accident records from year 2003 to 2015 that occurred in the city, markets Highway. To classify them correctly from the training data set the error rates and accuracy using classifiers are evaluated. In this study, all data is considered as instances and features in the data are known as attributes. For easier analysis and evaluation the simulation results are partitioned into several sub items. Different performance matrix i.e. TP rate, FP rate, and Precision, Recall, F-measure and ROC area were presented in numeric value during training and testing phase were displayed in the result. The summary of these results by running the algorithm in WEKA is reported in figure 3 also in table 2 and dataset shown in figure 2.



Relation: TrafficAcc

No.	Driver Nominal	DriverType Nominal	Environment Nominal	Road Nominal	LightCondition Nominal	Vehicle Nominal	Automobile1 Nominal	Automobile2 Nominal	Animals Nominal	AccidentType Nominal	SeatBelt Nominal	Location Nominal	PredictAcc Nominal
1	Drunken	Experienced	Rain	PoliceChecks		HeadLight	Truck	Truck		Fatal	notWearing	Highway	Yes
2	Drunken	Experienced	Fog	Sandy	Flactuating	VBody	Truck	3Wheeler		Death	notWearing	Highway	No
3	Fatigue	Experienced	Rain	RDesign		HeadLight	Truck	Car		Fatal	notWearing	Highway	No
4	Sleepiness	Learner	Fog	RQualityMaterial	NoLight	Breaksystem	Truck	Car		Injury	wearing	City	Yes
5	Etiquette	Experienced	Sunrays	RRatioMixtures		Breaksystem	Truck	Byke		Fatal	notWearing	Highway	Yes
6	Fatigue	Experienced	Rain	RDesign	Flactuating	Breaksystem	Truck	Bicycle		Death	notWearing	City	Yes
7	FaultyPreparation	Learner	Mist	RDesign	Flactuating	Breaksystem	Truck	Padestrian		Death	notWearing	City	No
8	IgnoranceHighwayCode	Experienced	Sunrays	PoliceChecks	NoLight	Tyres	3Wheeler	Byke		Injury	notWearing	Market	No

Figure 2: A snap of a part of dataset.



Classifier: NaiveBayes

Test options:
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) PredictAcc

Result list (right-click for options): 23:15:08 - bayes.NaiveBayes

Classifier output:

```
Correctly Classified Instances      117811              89.4554 %
Incorrectly Classified Instances     13887              10.5446 %
Kappa statistic                          0.7589
Mean absolute error                      0.1391
Root mean squared error                  0.2753
Relative absolute error                 31.8484 %
Root relative squared error             58.9114 %
Total Number of Instances           131698

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.921     0.16       0.924     0.921     0.922      0.962    Yes
                 0.84      0.079      0.834     0.84      0.837      0.962    No
Weighted Avg.    0.895     0.134      0.895     0.895     0.895      0.962

=== Confusion Matrix ===

     a     b   <-- classified as
 82206  7087 |     a = Yes
  6800 35605 |     b = No
```

Status: OK

Figure 3: Output of Naive Bayes Classifier

Summary of the output

Correctly Classified Instances	117811	89.4554 %
Incorrectly Classified Instances	13887	10.5446 %
Total Number of Instances	13198	

Table 2: Summary of the performance matrix of Naïve Bayes Algorithm.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Yes	0.921	0.16	0.924	0.921	0.922
No	0.84	0.079	0.834	0.84	0.837
Avg.	0.895	0.134	0.895	0.895	0.895

5. RESULT

Result in table 3 showed that increase in number of relevant attributes increase the accuracy of the Naive Bayes classifier, and probability prediction of accident by supplying test file with or without missing values to predict class attribute. These result can be very useful in controlling the road traffic accident by taking preventive measures and get aware before such accident might occur.

Table 3: Summary of the Experiments conducted.

S.no.	1	2	3
Classification Model	Naïve Bayes	Naïve Bayes	Naïve Bayes
Number of correctly classified instances	1,14,910	1,15,975	1,17,811
Accuracy in %	87.2527	88.0613	89.4554
Actual	YES	YES	YES
Predicted	YES	YES	YES

It can be easily seen that by collectively including more attributes i.e. collecting more important factors which are related to the traffic accident and apply them to the dataset yields better accuracy. In this work an accuracy of 89.4554 % was achieved which is more efficient regarding to control and minimize that occurrence of accident and also predict more accurate result for prediction that whether accident may occur or not on the applied conditions.

6. CONCLUSION

In this paper, we collected traffic accident data, and cleaned it, and attempted to construct novel attributes, and tested predictive model. The outputs of the model were presented for analysis to domain experts for feedback. The endeavor of this paper is to spot the causes of accidents how to reduce it. To accomplish these goals the WEKA data mining tools has been used to employ the Naive Bayes classifier. The assessment of the model using WEKA experimenter showed that Naive Bayes algorithm outperforms with an accuracy of 89.4554 %, In contrast with the previously published work of the authors, which focused on driver characteristics, here we focused on the contribution of various road-related factors role of environment, animals which suddenly come in front of running vehicles, weather conditions, condition of vehicles, hanging parts or material of vehicles, engine problems that have impact on the accident severity. The results of this study could be used by the respective stake holders to promote road safety. Thus this work could have tremendous impact on the well-being of Indian civilians as well as on other countries having the similar condition of road traffic accident. In the end, we have to conclude that by adding more attributes to the road traffic accident dataset we gets more accurate result and thus can be help in prevention of accident that occurred on road and safe ones precious life their physical condition as well as loss of properties. In future, we have to enlarge the classification accuracy of road traffic accidents types, data quality has to be improved. Another future effort is to test the applicability of other data mining techniques and clearing to get more accurate data.

REFERENCES

[1] Dipot, Olutayo. Et al," Using Data Mining Technique to Predict cause of accident and accident prone locations on highways" American Journal of Database Theory and Application 2012, 1(3): 26-38.

[2] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 22, pp. 41-46, 2001.

[3] D. Grossman and P. Domingos, "Learning Bayesian network classifiers by maximizing conditional likelihood,"(2004).

[4] Tibebe Shah, Shawndra Hill (2013)," Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia".

[5] Jianfeng Xi, Zhao et al "A Traffic Accident Causation Analysis Method Based on AHP Apriori 137 (2016) 680 – 687".

[6] Evgenikos, Yannis, et al "Characteristics and causes of heavy goods vehicles and buses accidents in Europe TRA 14 (2016) 2158 – 2167".

[7] Goel, Sachdeva "Analysis of Road Accidents on NH-1 between 98 Km to 148 Km [2016] pisc -256".

[8] Batrakova, Gredasova "Influence of Road Conditions on Traffic Safety 134 (2016)196–204".

[9] Helen, Wellman, et al "A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms (2015) 165–176".

[10] Taylor, Lacovara et al "Near-miss narratives from the fire service: A Bayesian analysis (2014) 119–129".

[11] Wong, J. and Y. Chung (2008). "Comparison of Methodology Approach to Identify Causal Factors of Accident Severity." Transportation Research Record 2083: 190-198.

[12] Chong, M., A. A., et al. (2005). "Traffic Accident Analysis Using Machine learning Paradigms." Informatica 29(1).

[13] Ali et al." A Data Mining Approach to Identify Key Factors Of Traffic Injury Severity" Traffic & Transportation, Vol. 23, 2011, No. 1, 11-17.

[14] Brijesh Kumar Baradwaj, Saurabh Pal," Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[15] Ting, Tsang et al "Naïve Bayes a Good Classifier for Document Classification [2011].