Road Accident Analysis Using Data Mining Techniques

A Project Report submitted in partial fulfilment of the requirement for the award of the degree of

Master of Technology

in

Computer Science & Engineering

Under the Supervision of

Dr. Pradeep Kumar Singh

(Supervisor)

By

Sachin Verma
Enrollment No: 142203

Jaypee University of Information TechnologyWaknaghat, Solan – 173234, Himachal Pradesh

Dedicated to

my mother, Mrs. Meena Verma, who has always emphasized the
importance of education, discipline, integrity and has been a constant source of
inspiration for me, my entire life

and

my father, Dr. Deen Dyal Verma, who has always been my role model
for hard work, persistence, patience and always supported me open heartedly in
all my endeavors.

Certificate

I hereby declare that the work presented in this report entitled **"Road Accident Analysis Using Data Mining Techniques"** in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2015 to December 2015 under the supervision of **Dr. Pradeep Kumar Singh** Assistant Professor (Senior Grade) department of Computer Science & Engineering and Information Technology**.**

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
Sachin Verma, 142203

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)
Dr. Pradeep Kumar Singh
Assistant Professor (Senior Grade)
Department of Computer Science & Engineering and Information Technology

Acknowledgements

I express my deepest gratitude towards my guide **Dr. Pradeep Kumar Singh** for his constant help and encouragement throughout the project work. I have been fortunate to have a guide who gave me the freedom to explore on my own and at the same time helped me plan the project with timely reviews and constructive comments, suggestions wherever required. A big thanks to him for having faith in me throughout the project and helping me walk through the new avenues of research papers and publications.

I also take this opportunity to thanks all those teachers, staff and colleagues who have constantly helped me grow, learn and mature both personally and professionally throughout the process. A BIG thanks goes to my dearest friends who have always supported, guided and even criticized me, always for the right reasons and have helped me stay sane throughout this and every other chapter of my life. I greatly value their friendship and deeply appreciate their belief in me. Special thanks to all the new friends from M.Tech. I have made without whom the journey wouldn't have been so interesting and memorable!

Most importantly, none of this would have happened without the love and patience of my family my parents, to whom this dissertation is dedicated. I would like to express my heart-felt gratitude to my family.

Sachin Verma (142203)
JUIT, Wakanaghat.
Solan.
June 2016.

Table of Contents

List Of Figures

List of Tables

Chapter 1.
INTRODUCTION

1.1 INTRODUCTION

A traffic collision, also known as a motor vehicle collision (MVC) among others, occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree or utility pole. Traffic collisions may result in injury, death and property damage.

A number of factors contribute to the risk of collision, including vehicle design, speed of operation, road design, road environment, and driver skill, impairment due to alcohol or drugs, and behavior, notably speeding and street racing. Worldwide, motor vehicle collisions lead to death and disability as well as financial costs to both society and the individuals involved. Road injuries occurred in about 54 million people in 2013 [1]. This resulted in 1.4 million deaths in 2013, up from 1.1 million deaths in 1990 [2]. About 68,000 of these occurred in children less than five years old [2].Almost all high-income countries have decreasing death rates, while the majority of low-income countries have increasing death rates due to traffic collisions. Middle-income countries have the highest rate with 20 deaths per 100,000 inhabitants, 80% of all road fatalities by only 52% of all vehicles. While the death rate in Africa is the highest (24.1 per 100,000 inhabitants), the lowest rate is to be found in Europe [3]. (March 2006) Road traffic accidents—the leading cause of death by injury and the tenth-leading cause of all deaths globally—now make up a surprisingly significant portion of the worldwide burden of ill-health. An estimated 1.2 million people are killed in road crashes each year, and as many as 50 million are injured, occupying 30 percent to 70 percent of orthopedic beds in developing countries hospitals, and if present trends continue, road traffic injuries are predicted to be the third-leading contributor to the global burden of disease and injury by 2020. One percent of world’s population belongs to Iran whereas country has 1 out of 40 of deaths from road accidents in the world. Hence, attempts to improve travels’ safety and to reduce hazards of road accidents through development and application of traffic safety programs are essential tasks. What is still considered by the traffic experts is to identify the factors affecting the incidence or severity of an occurred crash. The traffic safety is subjected to vast and complex dimensions in which to be interacted together and consequently demands various knowledge and experiences. These knowledge and experiences are applied in three categories of factors including: Human factors, Road, and Vehicles [4]. Previous studies indicated that the role of the human factor influence on road safety and accidents close to 90 percent [4]. Accident Analysis & Prevention provides wide coverage of the general areas relating to accidental injury and damage, including the pre-injury and immediate post-injury phases.

1.2 MACHINE LEARNING

Machine Learning controls how computers can learn and boost their performance based on data. Main aim of a research aspect since a long time is to make computers such intelligent that they can make independent intelligent decisions in any situation without human intervention. They should implicitly learn to observe, identify complex patterns and make intelligent decisions based on the data of their own.

Types of Machine Learning:

- **Supervised Learning**: based upon the learning from available data and constructs a model, which classifies the new tuples to a particular class. Some examples of supervised learning techniques are SVM, decision tree [4] etc.
- **Unsupervised Learning:** implies clustering because initially no predefined classes are there in the data set. Clusters are built from the tuples which holds some similarity and after that user can map these clusters to a particular class. Commonly we use clustering to identify the classes within the data. Unsupervised model built cannot tell us about the semantic meaning of the clusters identified, because training data is unlabeled.eg. KNN.

1.3 DATA MINING

Data Mining is relatively young and interdisciplinary field in computer science which deals with analyzing and discovering interesting and useful patterns from large data sets. This field involves various tasks for analyzing data out of which the most important task relevant in the context of our work is: - Classification. Classification task involves generalizing a known structure or pattern among the available data already assigned some special class or label. This generalized pat- tern can then be used to predict the class of a new and unknown data. On the contrary, Clustering is also a data mining task of discovering groups and structures of data which are in some way similar to each other and differ in similar way from other groups, without any prior knowledge of the structure of the data.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays.

1.3.1 Data, Information, and Knowledge

1.3.1.1 Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro-economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

1.3.1.2 Information

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

1.3.1.3 Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

1.3.2 Data Warehouses

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely.

1.3.4 How does data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

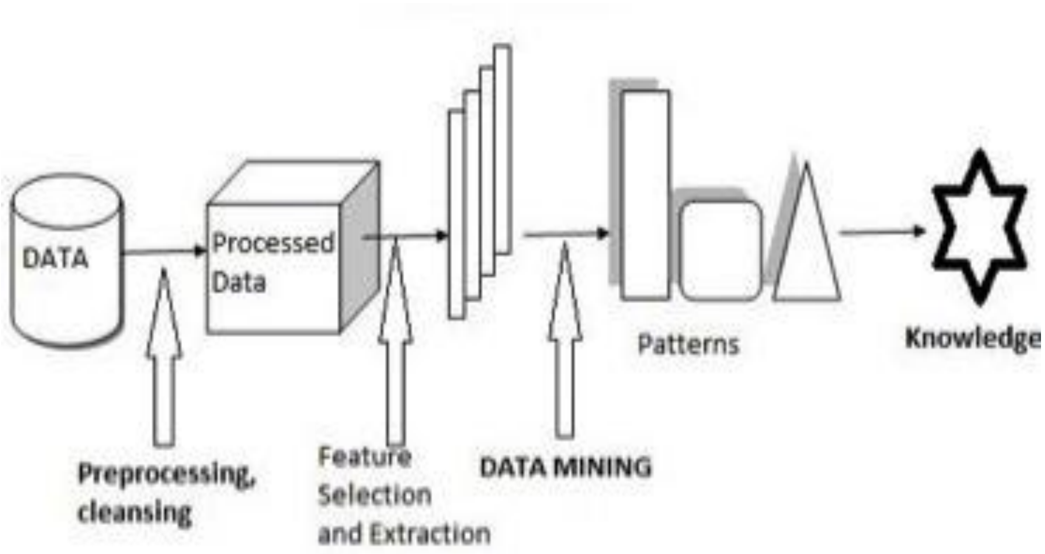


Figure 1: Steps followed in Knowledge Retrieval.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

1.3.5 What technological infrastructure is required?

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. NCR has the capacity to deliver applications exceeding 100 terabytes. There are two critical technological drivers:

- **Size of the database**: the more data being processed and maintained, the more powerful the system required.

- **Query complexity**: the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

1.4 Outline of Thesis

Brief introduction about the thesis topic, road traffic, data mining have been discussed in chapter 1. In chapter 2 survey of literature in the context of road accident analysis has been done and all techniques have been discussed conceptually. Chapter 3 describes the problem statement and objectives of the work done in this thesis. Chapter 4 details the solution proposed to address the problems stated in chapter 3. Chapter 5 explains the methodology adopted to deal with the problem and various performance metrics, which have been observed to determine the performance of the proposed solution. Finally implementation and results analysis have been done in chapter 6. Chapter 7 holds the conclusive remarks for the work done in thesis and future work.

Chapter 2.

LITERATURE REVIEW

2.1 Introduction

All the data mining techniques which are used by various authors around the world in context to weather forecasting have been discussed with respect to the problems they have addressed, as well as conceptually also. Obviously not a single technique can be used for all purposes. A technique giving good performance in one area doesn’t imply that it will surely give the better results in different aspect. Every technique is associated or influenced with the application of domain in which they are used, upon the dimensionality of the data as well as the aim of the work.

2.2 Classification

Following are the examples of cases where the data analysis task is Classification −

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.

- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. How well the predictions are done is measured in percentage of predictions hit against the total number of predictions. Optimal solution is a rule with 100% prediction hit rate, which is impossible to achieve. Thus approximation algorithms example-statistical algorithms ID3 can only solve classification except for few examples, C4.5, CART. The ID3 algorithm induces classification models, or DECISION TREE, from data. As it is supervised algorithm, previous years or day’s data is used for training the classifier and then used for prediction purpose of unlabeled data. After being trained, the algorithm should be able to predict the class of a new item. ID3 identifies attributes that differentiate one class from another. All attributes must be known in advance, and must also be either continuous or selected from a set of known values. For instance, temperature (continuous), and country of citizenship (set of known values) are valid attributes. To determine which attributes are the most important, ID3 uses the statistical property of entropy. Entropy measures the amount of information in an attribute. This is how the decision tree, which will be used in testing future cases, is built. One of the limitations of ID3 is that it is very sensitive to attributes with a large number of values (e.g. social security numbers). The entropy of such attributes is very low, and they don’t help you in performing any type of prediction. The C4.5 algorithm overcomes this problem by using another statistical property known as information gain. Information gain measures how well a given attribute separates the training sets into the output classes. Therefore, the C4.5 algorithm extends the ID3 algorithm through the use of information gain to reduce the problem of artificially low entropy values for attributes such as social security numbers. J48 in WEKA implements C4.5. A supervised classification algorithm [5] which have different varieties in internal node splitting algorithm like ID3, CART, C4.5, Gini Index have been implemented by authors in predicting temperature, rainfall, evaporation, wind speed and weather events [5].

2.3 Review Analysis

In 2009 the accuracy of data mining techniques viz. discriminant analysis, logistic regression, Bayes classifier, nearest neighbor, artificial neural networks, and classification trees has been investigated in analyzing customers’ default credit payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods [6]. The results reveal that artificial neural network is the only one that can accurately estimate the real probability of default credit payments.

CART model has been modeled to find the relationship between injury severity and driver/vehicle characteristics, highway/environment variables, and accident variables in Taiwan accident data from the year 2001 [7].

Logistic regression model and classification tree method have been compared in determining social-demographic risk factors which have affected depression status of women in separate postpartum periods [8]. They projected that Classification tree method gives more information than logistic regression model with details on diagnosis by evaluating a lot of risk factors. A comparative study has been conducted between data mining and statistical techniques by varying the number of independent variables, the types of independent variables, the number of classes of the independent variables, and the sample size [9]. The results have shown that the artificial neural network performance improved faster than that of the other methods as the number of classes of categorical variables increased.

FCBF algorithm [10] is designed for high dimensional data and has been shown effective in removing both irrelevant features and redundant features. The limitation of Mutual Information Feature Selector (MIFS) has been analyzed in [11] and proposed a method to overcome this limitation. The basics and implementation of various feature selection algorithms have been discussed in [12].

The combination of the AdaBoost and random forests algorithms were used for constructing a breast cancer survivability prediction model [13]. It was proposed to use random forests as a weak learner of AdaBoost for selecting the high weight instances during the boosting process to improve accuracy, stability and to reduce over-fitting problems [13].

Several voting algorithms, including Bagging, AdaBoost, and Arc-x4, have been studied using decision tree and Naïve Bayes to understand why and when these algorithms affect classification error [14].

The accuracies of simple classification algorithms such as C4.5, C-RT, CS-MC4, Decision List, ID3, Naive Bayes and Random Tree have been evaluated using the accuracy measures such as Precision, Recall and ROC curve [15]. The accuracy of classifiers using feature selection algorithms has been compared and the results have shown that Random Tree using Feature Ranking algorithm better performs other algorithms in modeling the vehicle collision patterns in road accident data [16].

Abdelwahab et al. studied the 1997 accident data for the Central Florida area [17]. The analysis focused on vehicle accidents that occurred at signalized intersections. The injury severity was divided into three classes: no injury, possible injury and disabling injury. They compared the performance of Multi-layered Perceptron (MLP) and Fuzzy ARTMAP, and found that MLP classification accuracy is higher than Fuzzy ARTMAP. Levenberg- Marquardt algorithm was used for MLP training and achieved 65.6 and 60.4 percent classification accuracy for the training and testing phases, respectively. Fuzzy ARTMAP achieved a classification accuracy of 56.1 percent. Yang et al. used neural network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs [18]. They performed the Cramer's V Coefficient test [19] to identify significant variables that cause injury to reduce the dimensions of the data. Then, they applied data transformation method with a frequency-based scheme to transform categorical codes into numerical values. They used the Critical Analysis Reporting Environment (CARE) system, which was developed at the University of Alabama, using a Back propagation (BP) neural network. They used the 1997 Alabama interstate alcohol-related data, and further studied the weights on the trained network to obtain a set of controllable cause variables that are likely causing the injury during a crash. The target variable in their study had two classes: injury and non-injury, in which injury class included fatalities. They found that by controlling a single variable (such as the driving speed, or the light conditions) they could reduce fatalities and injuries by up to 40%.

Sohn et al. applied data fusion, ensemble and clustering to improve the accuracy of individual classifiers for two categories of severity (bodily injury and property damage) of road traffic accident [20]. The individual classifiers used were neural network and decision tree. They applied a clustering algorithm to the dataset to divide it into subsets, and then used each subset of data to train the classifiers. They found that classification based on clustering works better if the variation in observations is relatively large as in Korean road traffic accident data.

Mussone et al. used neural networks to analyze vehicle accident that occurred at intersections in Milan, Italy [21]. They chose feed-forward MLP using BP learning. The model had 10 input nodes for eight variables (day or night, traffic flows circulating in the intersection, number of virtual conflict points, and number of real conflict points, type of intersection, accident type, road surface condition, and weather conditions). The output node was called accident index, which was calculated as the ratio between the number of accidents for a given intersection and the number of accidents at the most dangerous intersection. Results showed that the highest accident index for running over of pedestrian occurs at non-signalized intersections at night time.

Dia et al. used real-world data for developing a multilayered MLP neural network freeway incident detection model [22]. They compared the performance of the neural network model and the incident detection model in operation on Melbourne’s freeways. Results showed that neural network model could provide faster and more reliable incident detection over the model that was in operation on Melbourne’s freeways. They also found that failure to provide speed data at a station could significantly deteriorate model performance within that section of the freeway.

Shankar et al. applied a nested logic formulation for estimating accident severity likelihood conditioned on the occurrence of an accident [23]. They found that there is a greater probability of evident injury or disabling injury/fatality relative to no evident injury if at least one driver did not use a restraint system at the time of the accident.

Kim et al. developed a log-linear model to clarify the role of driver characteristics and behaviors in the causal sequence leading to more severe injuries. They found that alcohol or drug use and lack of seat belt use greatly increase the odds of more severe crashes and injuries [24].

Abdel-Aty et al. used the Fatality Analysis Reporting System (FARS) crash databases covering the period of 1975-2000 to analyze the effect of the increasing number of Light Truck Vehicle (LTV) registrations on fatal angle collision trends in the US [25]. They investigated the number of annual fatalities that resulted from angle collisions as well as collision configuration (car-car, car-LTV, LTV-car, and LTV-LTV). Time series modeling results showed that fatalities in angle collisions will increase in the next 10 years, and that they are affected by the expected increase in the percentage of LTVs in traffic.

Bedard et al. applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers’ fatality risk [26]. They found that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities. Evanco conducted a multivariate population-based statistical analysis to determine the relationship between fatalities and accident notification times [27]. The analysis demonstrated that accident notification time is an important determinant of the number of fatalities for accidents on rural roadways.

Ossiander et al. used Poisson regression to analyze the association between the fatal crash rate (fatal crashes per vehicle mile traveled) and the speed limit increase [28]. They found that the speed limit increase was associated with a higher fatal crash rate and more deaths on freeways in Washington State.

Furthermore, some researchers studied the relationship between drivers' age, gender, vehicle mass, impact speed or driving speed measure with fatalities [29, 30, 31, 32, 33].

2.4 CONCLUSION

We have studied the accident data and investigated the performance of neural network, decision tree, support vector machine and a hybrid decision tree- neural network for predicting driver's injury severity. As some techniques are independent of the interdependent attributes and some are inefficient in giving better results when multidimensional data is used. Thus by using decision tree we found out the effectiveness of the computational abilities of the training data set.

Chapter 3.

PROBLEM STATEMENT

3.1 INTRODUCTION

The costs of fatalities and injuries due to road traffic accidents (RTAs) have a tremendous impact on societal well-being and socioeconomic development. RTAs are among the leading causes of death and injury worldwide, causing an estimated 1.2 million deaths and 50 million injuries each year (World Health Organization, 2004).Therefore, finding the road accident patterns from a data set is essential for better and effective prediction so that proper measures can be used and better and enhanced technology can be used.

3.2 PROBLEM STATEMENT

Finding accurate and reliable technique or method through which extraction of data and information can be analyzed effectively and efficiently through the provided data set by concerned authorities. In this research work we focused on finding accident patterns in road accidents based on causes using various classification algorithms.

3.3 OBJECTIVES

- To find the road accident patterns using various classification algorithms.
- To study Data Mining Techniques, and how information can be extracted from the raw data.
- To find a robust, reliable and accurate technique or classifier for better understanding accident patterns and predicting them.

3.4 Profile of the Problem

In developed countries, road accident rates have decreased since the 1960s because of successful interventions and approaches such as seat belt safety laws, enforcement of speed limits, warnings about the dangers of mixing alcohol consumption with driving, and safer design and use of roads and vehicles.

3.4 CONCLUSION

A reliable and effective data mining technique is needed to address the problem of inefficiency of generating the data and extraction of data collected from reliable source and use that data set and narrow down the algorithms so that better effectiveness can be achieved.

CHAPTER 4.
PROPOSED SOLUTION

4.1 PROPOSED SOLUTION

To predict accident rate, various classification models were built by using different set of predictive classifiers. Decision trees properties are easy to analyze and to understand the build, that can manage both continuous and categorical variables, and can perform classification as well as regression analysis. They automatically handle interactions between variables and identify important variables.

The general objective of the research is to investigate the role of road-related factors in accident, using road accident data on highways from available data set and predictive models. Our three specific objectives include:

- Exploring the underlying road related variables that impact vehicle accident severity,
- Predicting accident time and causes by using different data mining techniques, and
- Comparing standard classification models for this task.

4.2 Genesis of Problem

The basis of this research is that accidents are not randomly occurs along the road network, and that drivers are not involved in accidents at random. There are complex relationships between several varying characteristics (driver, road, car, etc.) and the accidents occurring simultaneously. As such, one cannot improve safety without successfully relating accident frequency and severity to the causative variables. We will attempt to extend the previous work in this area by generating additional attributes and focusing on the contribution of road-related factors to accident severity. This will help to identify the parts of a road that are risky, thus supporting traffic accident data analysis in decision-making processes.

4.3 CONCLUSION

Road accidents are lethal and requires proper attention, for this we need relevant data set so that proper decision making and predictive models can be used for better results and prediction. In this research analysis we will use decision trees because they are easy to build and understand and can easily identify the interactions between the variables.

CHAPTER 5.
METHODOLOGY

5.1 INTRODUCTION

The systematic, theoretical analysis of the methods and parameters observed to resolve our problem of hailing prediction and verifying the performance of classifier as well as sensitivity. A number of parameters have been observed for all the kernel functions of SVM.

- Confusion Matrix
- Accuracy
- Precision
- Recall
- Receiver Operating Characteristics (ROC) Curve area
- Sensitivity
- 10 fold cross validation
- Cost/Benefit Analysis

5.1.1 Confusion Matrix

Also known as error matrix, is a tabular representation of the performance of an algorithm or classifier. It holds the information about the predicted values with respect to the actual value. It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

Table 1: Confusion Matrix

Positive	**Negative**	**Class**
True +ve	False -ve	**Positive**
False +ve	True -ve	**Negative**

5.1.2 Accuracy

- True results among the total cases examined.
- TP=True Positive
- FP=False Positive
- FN=False Negative
- TN=True Negative

Accuracy=TP+TN /(TP+FP+TN+FN)

5.1.3 Precision & Recall

In pattern recognition and information retrieval with binary classification, **precision** (also called positive predictive value) is the fraction of retrieved instances that are relevant, while **recall** (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

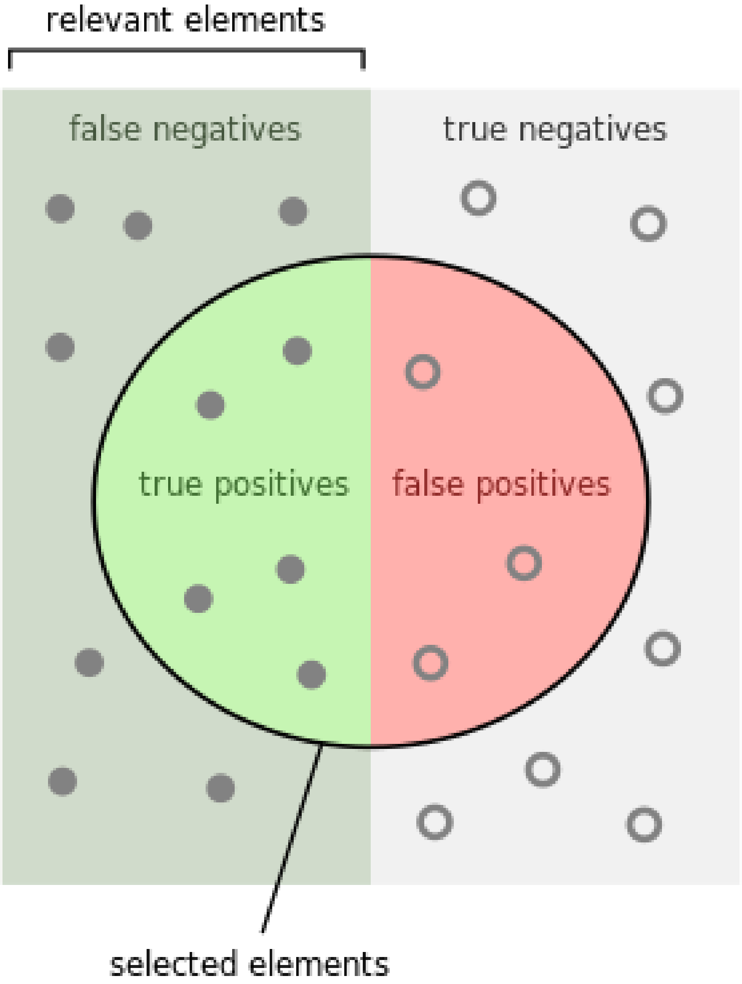


Figure 2: Precision & Recall

Precision=How many selected items are relevant. i.e. = TP/TP+FP

Recall=How many relevant items are selected. i.e. = TP/TP+FN

5.1.4 ROC Curve

In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as (1 - specificity). The ROC curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability in x-axis.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

The relationship between True Positive Rate (TPR) i.e. Sensitivity and True Negative Rate (TNR) i.e. Specificity as well the performance of the classifier can be visualized by the ROC curve. Area under the TPR on one axis and TNR on other axis curve gives the metric to determine the better classifier. More the area will be better performance of the classifier will be.

5.1.5 Sensitivity

How the model or classifier behavior changes when the system is perturbed, describes the reliability a very important measure to assess the classifier performance. As in our aspect sensitivity can be checked by training the classifier with different size of data sets and observe the variation in the performance of the classifier.

5.1.6 10 Fold Cross Validation

It is away to check the performance on trained training samples. Training set is divided in 10 parts and 9 is used for training purpose and all performance metrics are observed over the remaining 1 part and this whole procedure is repeated for 10 times by using each partition as a test set. Suppose we have a model with one or more unknown parameters, and a data set to which the model can be fit (the training data set). The fitting process optimizes the model parameters to make the model fit the training data as well as possible. If we then take an independent sample of validation data from the same population as the training data, it will generally turn out that the model does not fit the validation data as well as it fits the training data. This is called over fitting, and is particularly likely to happen when the size of the training data set is small, or when the number of parameters in the model is large. Cross-validation is a way to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available.

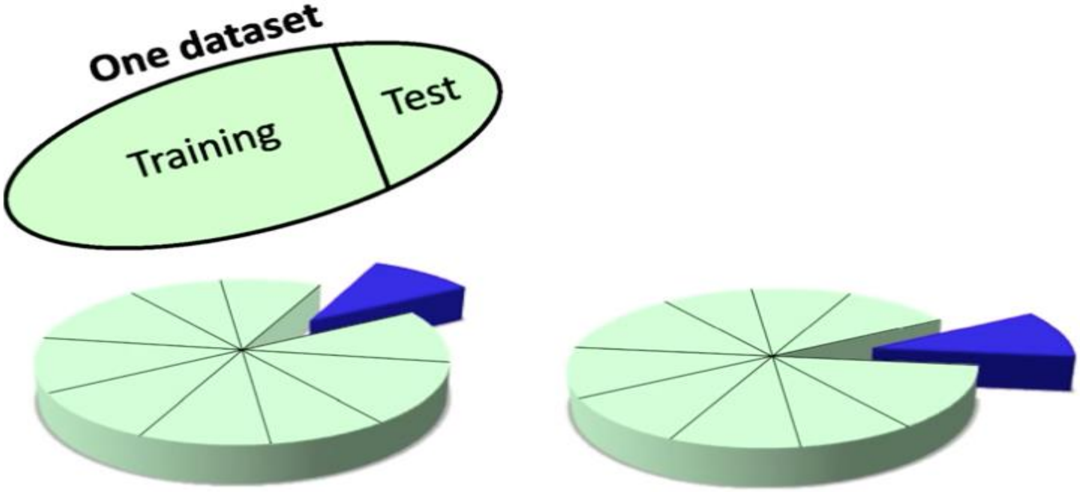


Figure 8: 10 Fold cross Validation

In summary, cross-validation combines (averages) measures of fit (prediction error) to correct for the optimistic nature of training error and derive a more accurate estimate of model prediction performance.

5.1.7 Cost/Benefit Analysis

Cost–benefit analysis (CBA), is a systematic process evaluation for estimating the strengths and weaknesses of alternatives that satisfy transactions, activities or functional requirements. It is a technique that is used to determine options that provide the best approach for the adoption and practice in terms of benefits in labor, time and cost savings etc. The CBA is also defined as a systematic process for calculating and comparing benefits and costs of a project, decision or a specified policy.

Broadly, CBA has two purposes:

1. To determine if it is a sound investment/decision (justification/feasibility),
2. To provide a basis for comparing projects. It involves comparing the total expected cost of each option against the total expected benefits, to see whether the benefits outweigh the costs, and by how much.

CBA is related to, but distinct from cost-effectiveness analysis. In CBA, benefits and costs are expressed in monetary terms, and are adjusted for the time value of money, so that all flows of benefits and flows of project costs over time (which tend to occur at different points in time) are expressed on a common basis in terms of their net present value.

5.2 Data Set Attributes and Description

Table 2. Attributes and Description

Attributes	**Description**
Time Shift	Morning, Afternoon, Evening, Night, Late-night.
Nature of Accident	Overturning, Head on collision, Rear end collision, Collision brush/Side Wipe, Right turn collision, Skidding, Right turn collision, Others.
Classification Of Accident	Fatal, Grievous Injury, Minor Injured, Non-Injury.
Causes	Drunken, Over speeding, Vehicle out of control, Fault of driver of motor vehicle/driver of other vehicle/cyclist/pedestrian/passenger, Defect in mechanical condition of motor vehicle/road.
Road Feature	Straight road, Slight Curve, Sharp Curve, Flat Road, Gentle incline, Steep incline, Hump, Dip.
Road Condition	Straight road, Slight Curve, Sharp Curve, Flat Road, Gentle incline, Steep incline, Hump, Dip.
Intersection Type	T Junction, Y Junction, Four arm junction, Staggered junction, unction with more than 4 arms, Roundabout junction, Manned Rail crossing, Unmanned Rail crossing.
Weather Condition	Fine, Mist/Fog, Cloud, light rain, Heavy rain, Hail/sleet, Snow, Strong Wind, Dust Storm, Very Hot, Very Cold, Other extra ordinary weather condition.
Vehicle Responsible	Truck, Bus, Car etc.
Help by Ambulance/Patrol	Help Provided Through Ambulance and highway patrol.

5.2.3 Training and deploying of a classifier

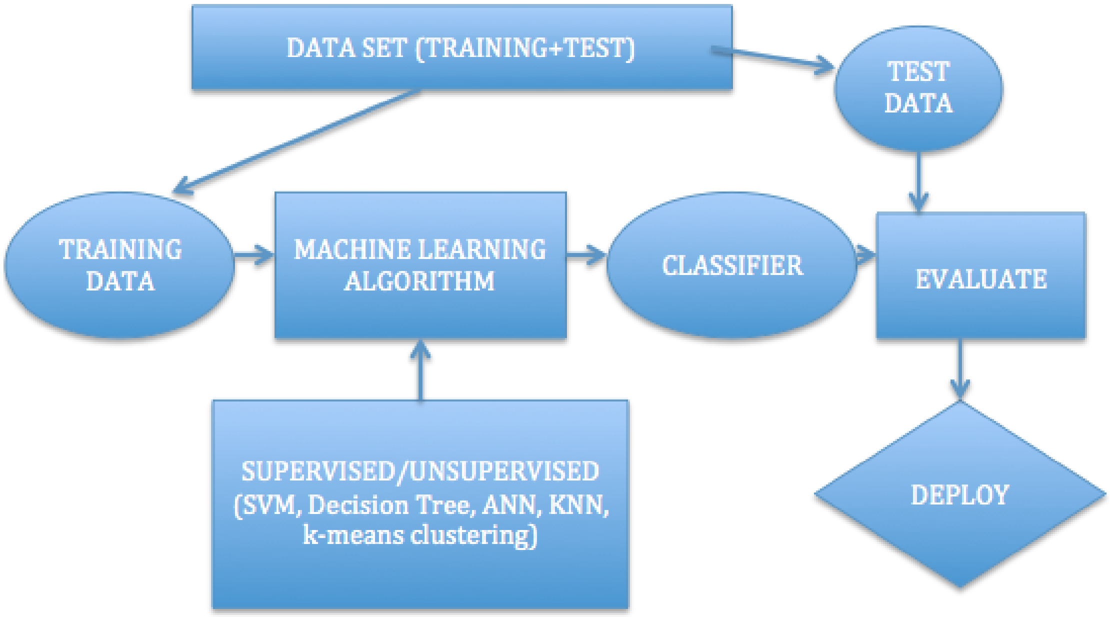

DATA SET (TRAINING+TEST)
TEST DATA
TRAINING DATA
MACHINE LEARNING ALGORITHM
CLASSIFIER
EVALUATE
SUPERVISED/UNSUPERVISED (SVM, Decision Tree, ANN, KNN, k-means clustering)
DEPLOY

Figure 3: Machine learning algorithms significance in building a classifier

5.3 CONCLUSION

By using WEKA (Waikato Environment of Knowledge Analysis) an open source data mining techniques implementation framework, by training the classifier with different size of training set and testing on Test data set.

CHAPTER 6.

IMPLEMENTATON AND RESULT ANALYSIS

WEKA tool is a workbench that contains a collection of visualization techniques and algorithms for data analysis and modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of WEKA was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Make file-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains.

But the more recent fully Java-based version (WEKA 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of WEKA include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

6.1 PERFORMANCE EVALUATION AND ANALYSIS

Data set is divided in 2 years 2014 and 2015. Data set is of the highway of Ambala highway, with 869 instances and 11 attributes in 2014 and in 2015, 11 attributes and 639 instances.

6.1.1 CONFUSION MATRIX for CART (2014)

1. Confusion matrix generated is in terms of causes is given below.

Table 3: Confusion Matrix (Causes)

a	b	c	d	e	f	<-- classified as
680	0	0	0	0	0	**a = O**
71	0	0	0	0	0	**b = DI**
3	0	0	0	0	0	**c = OTH**
27	0	0	0	0	0	**d = D**
77	1	0	0	0	0	**e = FOD**
10	0	0	0	0	0	**f = VOC**

Accuracy of correct instances computed: 78.2509 %

2. Confusion Matrix generated in terms of Timeshift.

Table 4: Confusion Matrix (Timeshift)

a	b	c	d	e	<-- classified as
331	0	0	0	37	**a = Morning**
132	0	0	0	14	**b = LateNight**
80	0	0	0	12	**c = Afternoon**
37	0	0	0	8	**d = Night**
186	0	0	0	32	**e = Evening**

Above result represents that in 2014, maximum accidents happen due to drivers fault because of over speeding in the morning time.

6.1.2 CONFUSION MATRIX for CART (2015)

1. Confusion matrix generated is in terms of causes is given below.

Table 5: Confusion Matrix (Causes)

a	**b**	**c**	**d**	**e**	**<-- classified as**
72	104	0	2	0	**a = O**
64	**188**	0	5	0	**b = FOD**
13	32	0	0	0	**c = VOC**
28	77	0	0	0	**d = DI**
23	30	0	1	0	**e = D**

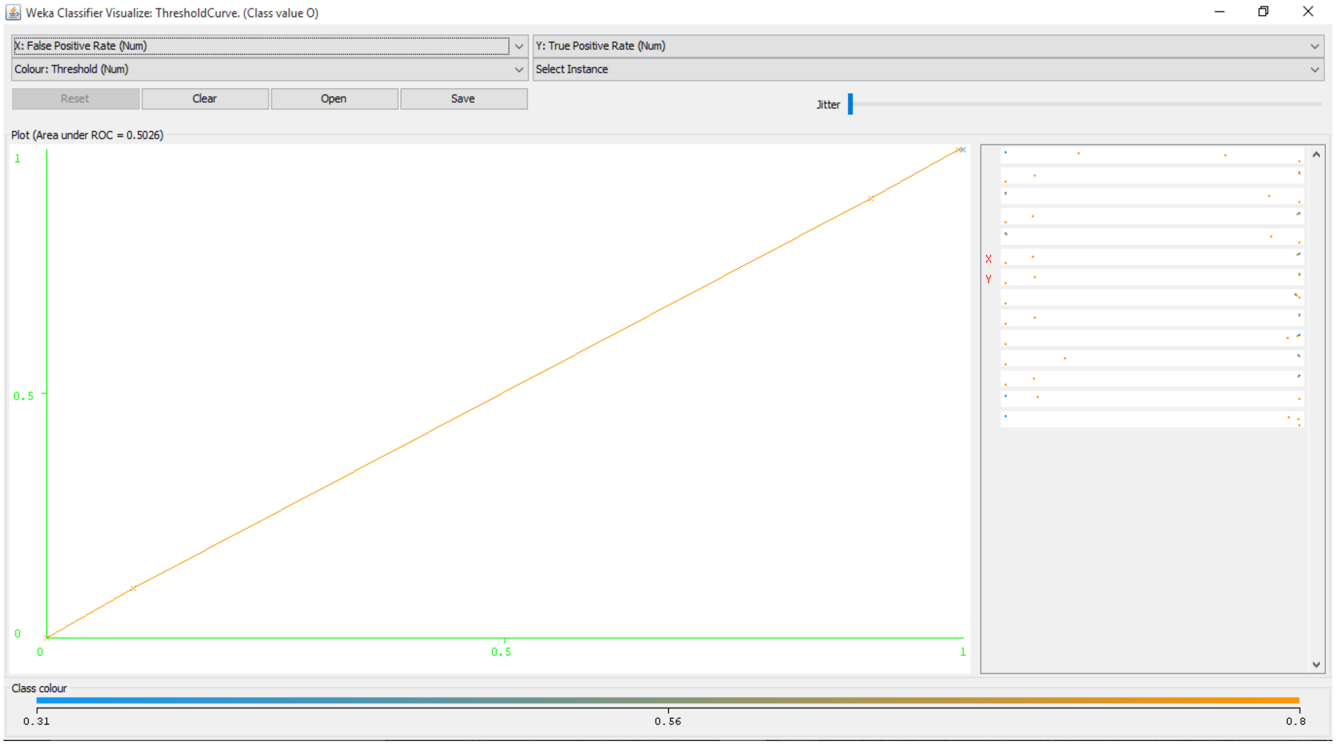
2. Confusion Matrix generated in terms of Time shift

Table 6: Confusion Matrix (Time shift)

a	**b**	**c**	**d**	**e**	**<-- classified as**
63	0	88	7	0	**a = EVENING**
9	0	32	1	0	**b= Night**
56	0	**177**	8	1	**c = Morning**
31	0	90	13	0	**d = late night**
18	0	42	0	1	**e= Afternoon**

Result of given data indicates that in 2015, maximum accidents occurs in morning time due to the fault of the driver.

6.4 ROC CURVES FOR CART



Weka Classifier Visualize: ThresholdCurve. (Class value 0)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.5026)

Figure 4: ROC curve for CART (causes) 2014

ANALYSIS: In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. In the above plotted ROC curve the area generated is 0.5026. ROC analysis helps tools to select optimal solutions and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC curve generally compares two operating characteristics i.e. TFR and FPR, in this we will compare the remaining values of area under the curve so as to select the better area under the curve for our predictive model and come up with a better and improve solution.

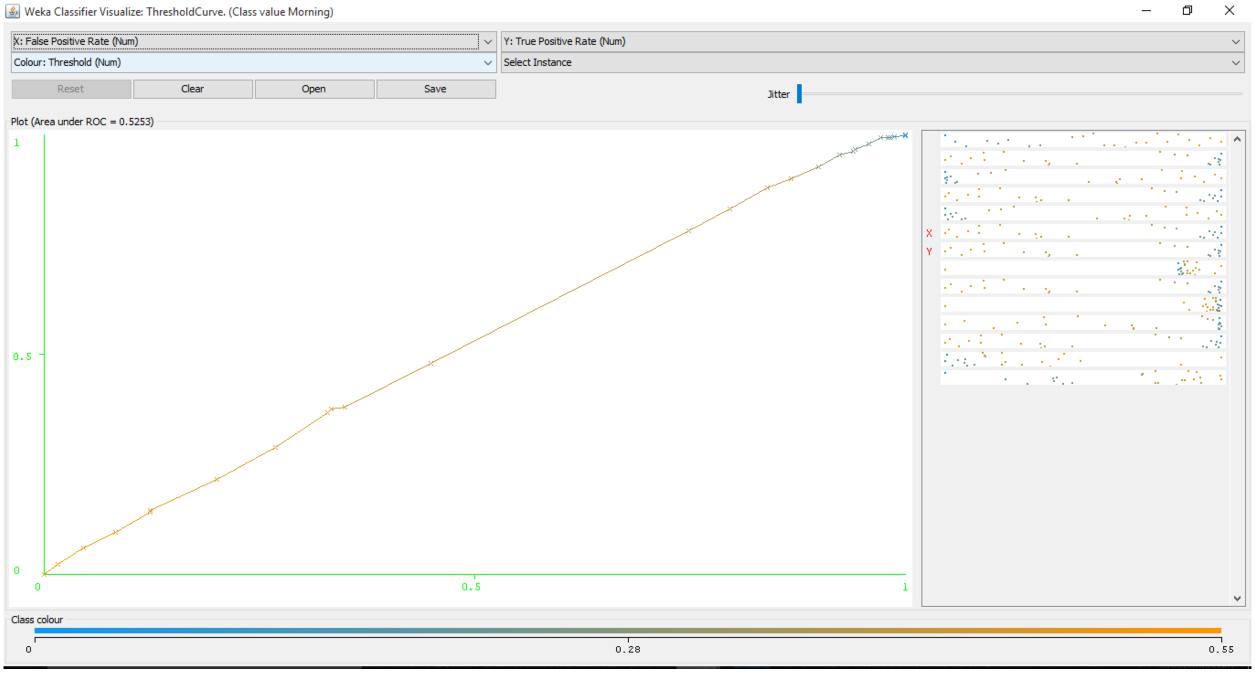

Weka Classifier Visualize: ThresholdCurve. (Class value Morning)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.5253)
1
0.5
0
0
0.5
1
X
Y
Class colour
0
0.28
0.55

Figure 5: ROC curve for CART (Time Shift) 2014

ANALYSIS: In this ROC curve area plotted is in numerical terms having different cut-off points varies in the function of false positive rate and true positive rate. The area under the curve generated in CART classifier is given in numerical terms which is equal to 0.5253 which shows the decent sensitivity analyzed by the graph generated by the classifier. Now as we have checked the area under the curve generated by the classifier, now we have to compare these values to find an optimal solution and discard the sub optimal solution through the analysis of the classifier.

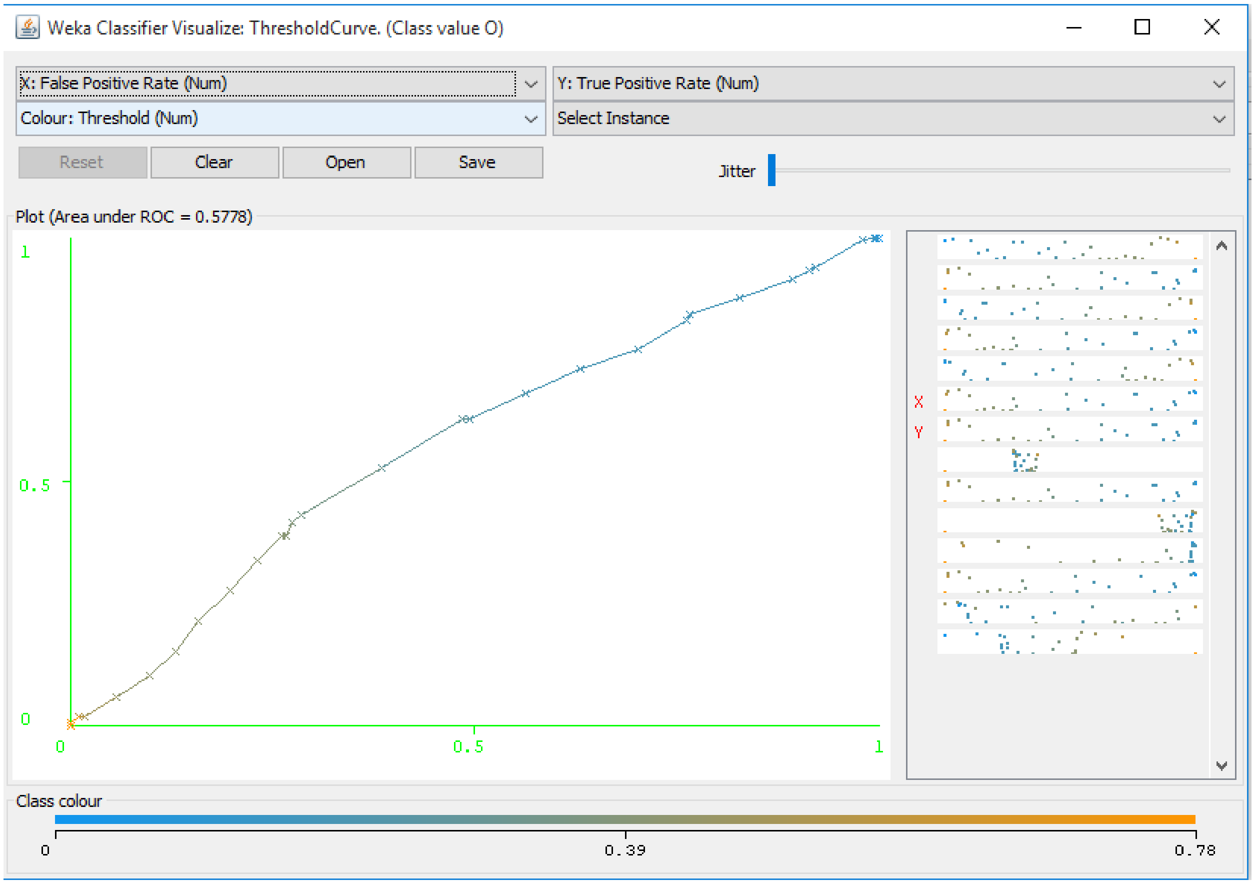
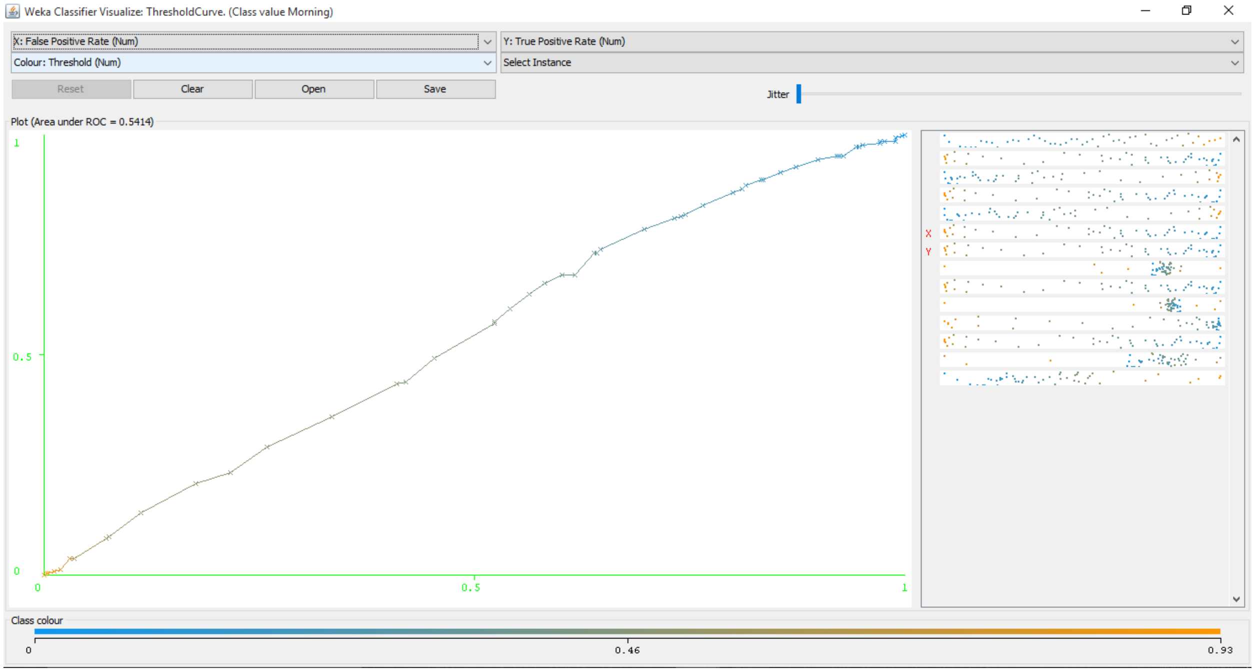

Weka Classifier Visualize: ThresholdCurve. (Class value O)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.5778)
1
0.5
0
0
0.5
1
X
Y
Class colour
0
0.39
0.78

Figure 6: ROC curve for CART (causes) 2015

ANALYSIS: Above analysis is for the causes of the accidents and is analyzed by the CART classifier of 2015 data set. The area under the curve shows the value of 0.5778 which is an optimal solution and a decent sensitivity of the area under the curve. In this two operating characteristics (TPR and FPR) showing the optimal solution of the curve generated by CART classifier in 2015 data set through diagnostic test evaluation.



Weka Classifier Visualize: ThresholdCurve. (Class value Morning)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.5414)

Figure 7: ROC curve for CART (TimeShift) 2015

ANALYSIS: This analysis is done with respect to the attribute Timeshift and area under the ROC curve is computed in numeric value of 0.5414 which is an optimal solution for comparing the values of the ID3 classifier. . In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Now as we have checked the area under the curve generated by the classifier, now we have to compare these values to find an optimal solution and discard the sub optimal solution through the analysis of the classifier.

6.5 CONFUSION MATRIX FOR ID3 (2014)

1. Confusion matrix generated is in terms of causes is given below.

Table 7: Confusion Matrix (Causes)

a	**b**	**c**	**d**	**e**	**f**	**<-- classified as**
569	24	2	9	19	0	**a = O**
45	12	0	0	5	0	**b = DI**
0	0	2	1	0	0	**c = OTH**
19	1	1	0	3	0	**d = D**
55	5	0	3	4	0	**e = FOD**
7	1	0	0	0	0	**f = VOC**

2. Confusion Matrix generated in terms of Time shift

Table 8: Confusion Matrix (Time shift)

a	**b**	**c**	**d**	**e**	**<-- classified as**
229	38	24	6	38	**a = Morning**
85	22	5	4	15	**b = Late Night**
47	6	6	3	21	**c = Afternoon**
21	3	2	4	11	**d = Night**
104	11	20	9	49	**e = Evening**

Result of given data indicates that in 2014, maximum accidents occurs in morning time due to the fault of the driver.

6.6 CONFUSION MATRIX FOR ID3 (2015)

1. Confusion matrix generated is in terms of causes is given below.

Table 9: Confusion Matrix (Causes)

a	**b**	**c**	**d**	**e**	**<-- classified as**
67	46	5	14	11	**a = O**
50	**134**	5	17	10	**b = FOD**
14	16	3	4	2	**c = VOC**
26	39	4	20	3	**d = DI**
21	14	2	8	2	**e = D**

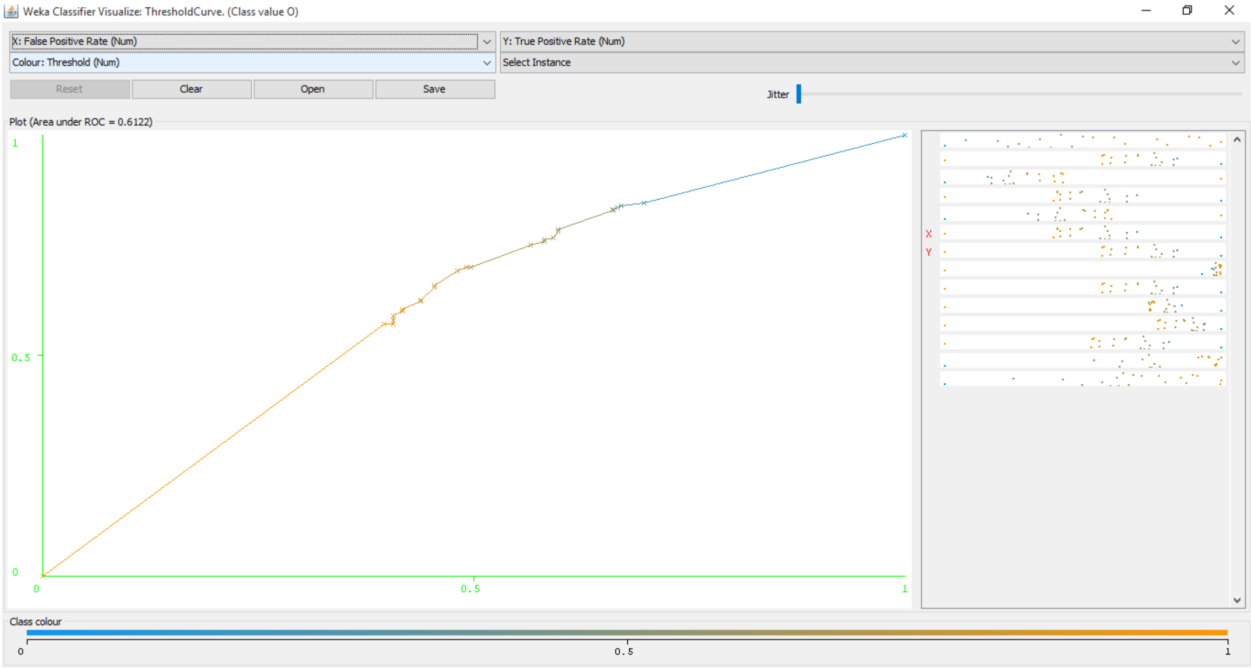
2. Confusion Matrix generated in terms of Timeshift

Table 10: Confusion Matrix (Timeshift)

a	**b**	**c**	**d**	**e**	**<-- classified as**
48	6	48	17	8	**a = EVENING**
6	2	19	5	3	**b = Night**
48	8	116	21	11	**c = Morning**
31	5	51	21	4	**d = latenight**
16	4	13	3	10	**e= Afternoon**

Result of given data indicates that in 2015, maximum accidents occurs in morning time due to the fault of the driver.

6.7 ROC CURVES FOR ID3



Weka Classifier Visualize: ThresholdCurve. (Class value 0)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

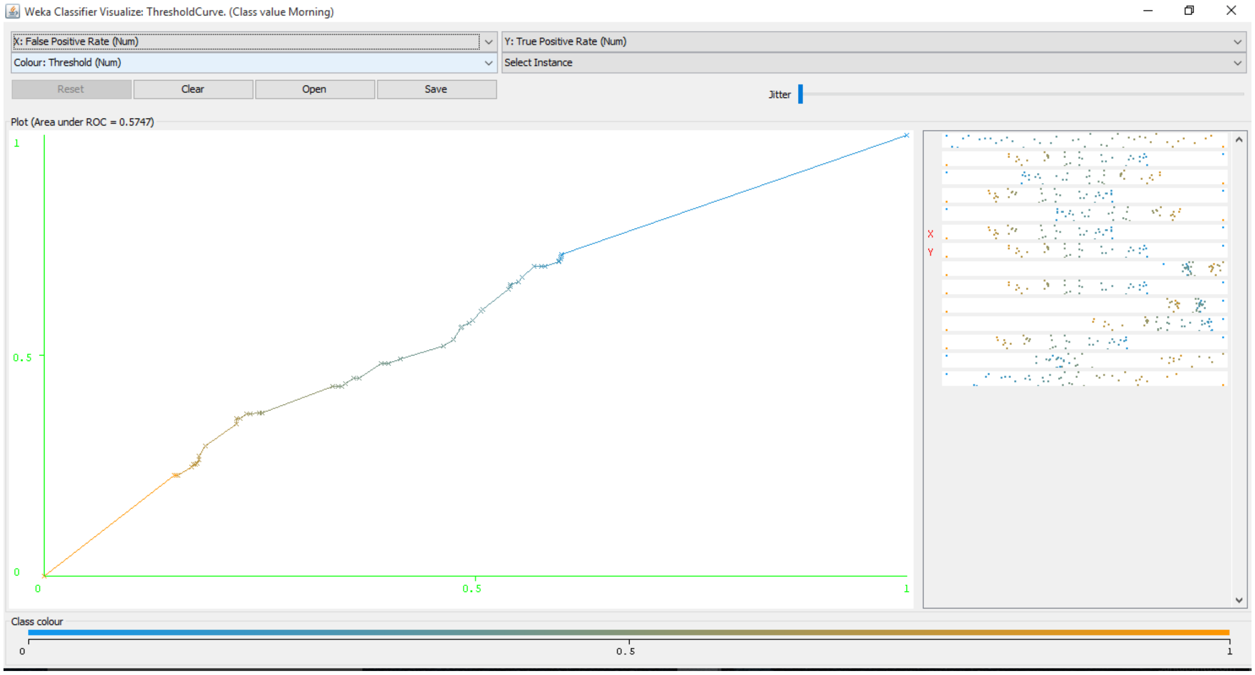
Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.6122)

Figure 8: ROC curve for ID3 (causes) 2014

ANALYSIS: In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. In the above plotted ROC curve the area generated is 0.6122. This curve provides tools to select optimal solutions and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC curve generally compares two operating characteristics i.e. TFR and FPR, in this we will compare the remaining values of area under the curve so as to select the better area under the curve for our predictive model and come up with a better and improve solution.



Weka Classifier Visualize: ThresholdCurve. (Class value Morning)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

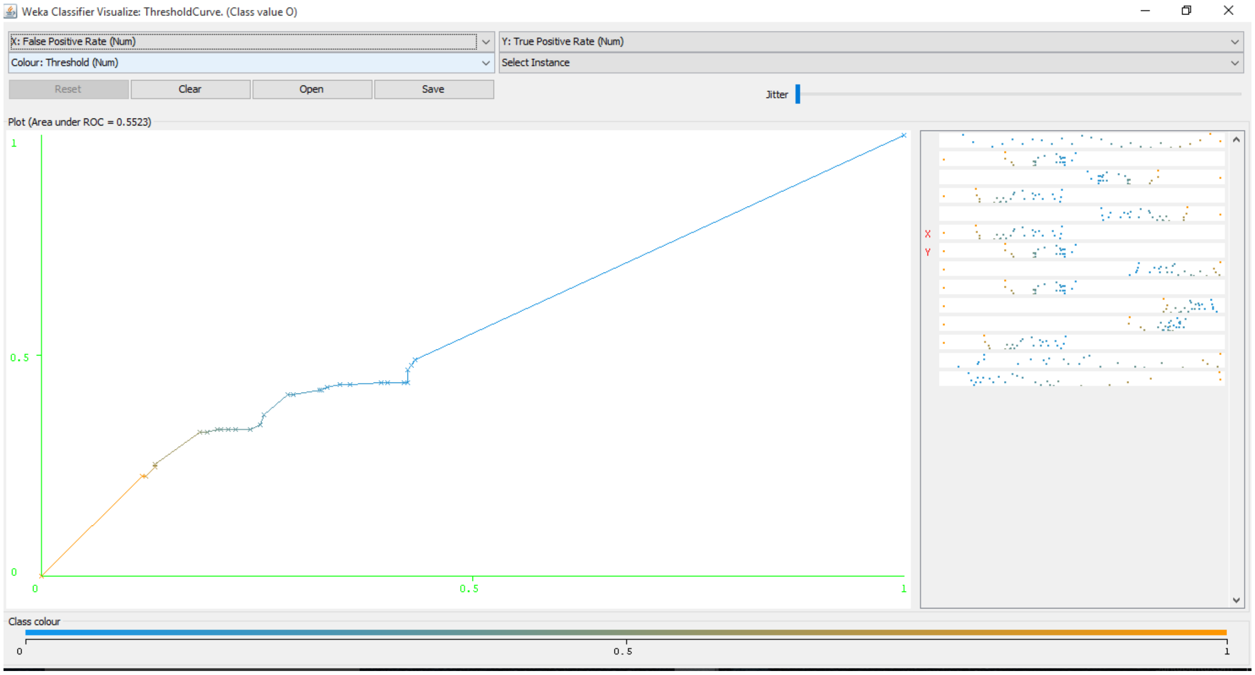
Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.5747)

Figure 9: ROC curve for ID3 (TimeShift) 2014

ANALYSIS: In this ROC curve area plotted is in numerical terms having different cut-off points varies in the function of false positive rate and true positive rate. The area under the curve generated in ID3 classifier is given in numerical terms which is equal to 0.5747 which shows the decent sensitivity analyzed by the graph generated by the classifier. Now as we have checked the area under the curve generated by the classifier, now we have to compare these values to find an optimal solution and discard the sub optimal solution through the analysis of the classifier.



Weka Classifier Visualize: ThresholdCurve. (Class value 0)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

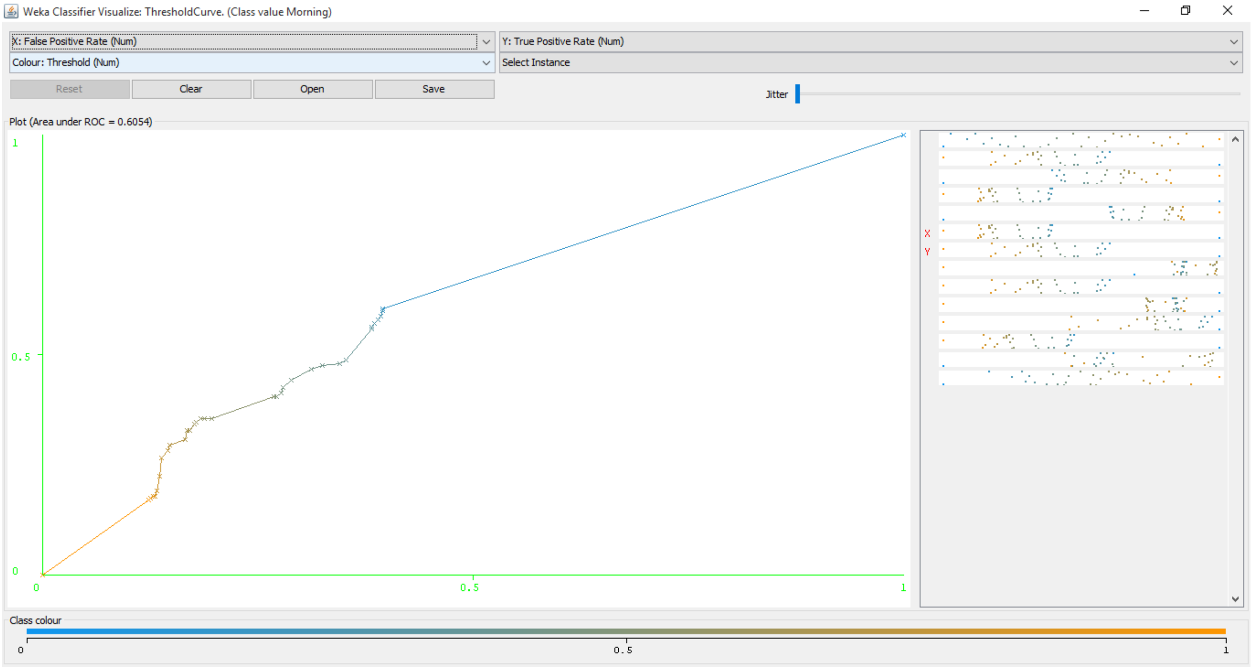
Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.5523)

Figure 10: ROC curve for ID3 (causes) 2015

ANALYSIS: Above analysis is for the causes of the accidents and is analyzed by the ID3 classifier of 2015 data set. The area under the curve shows the value of 0.5523 which is an optimal solution and a decent sensitivity of the area under the curve. In this two operating characteristics (TPR and FPR) showing the optimal solution of the curve generated by ID3 classifier in 2015 data set through diagnostic test evaluation.


Weka Classifier Visualize: ThresholdCurve. (Class value Morning)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot (Area under ROC = 0.6054)

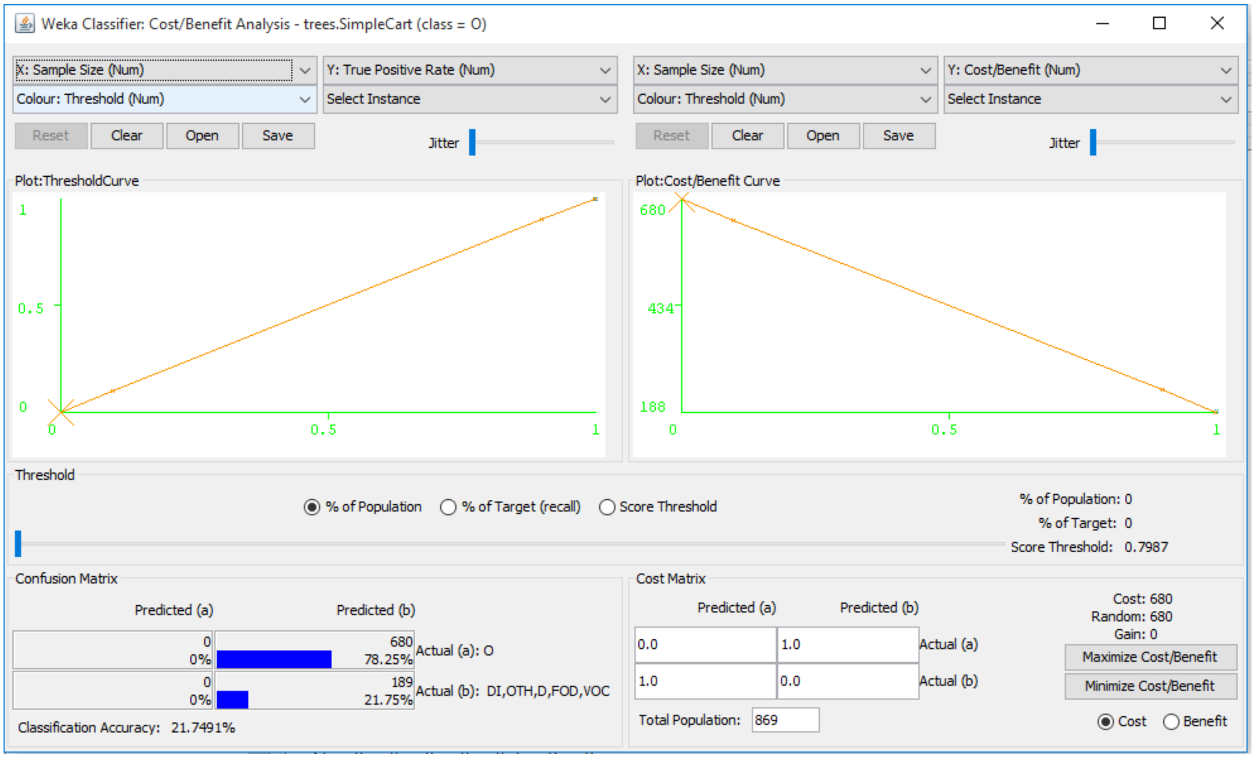
Figure 11: ROC curve for ID3 (TimeShift) 2015

ANALYSIS: This analysis is done with respect to the attribute Timeshift and area under the ROC curve is computed in numeric value of 0.5414 which is an optimal solution for comparing the values of the ID3 classifier. . In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Now as we have checked the area under the curve generated by the classifier, now we have to compare these values to find an optimal solution and discard the sub optimal solution through the analysis of the classifier.

6.8 Analysis of ROC Curve of both CART and ID3

As the ROC curves plots the graph with respect to true positive rate and false positive rate which decides the performance of the classifier. As in above ROC Curves the area given by Decision Tree (ID3) is more than that given by CART, hence it is concluded that ID3 can be used in real scenarios of road accident analysis.

6.8 COST/BENEFIT ANALYSIS FOR OVERSPEEDING



Weka Classifier: Cost/Benefit Analysis - trees.SimpleCart (class = O)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Threshold (Num) | Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.7987

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	680 78.25%	Actual (a): O
0 0%	189 21.75%	Actual (b): DI,OTH,D,FOD,VOC

Classification Accuracy: 21.7491%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

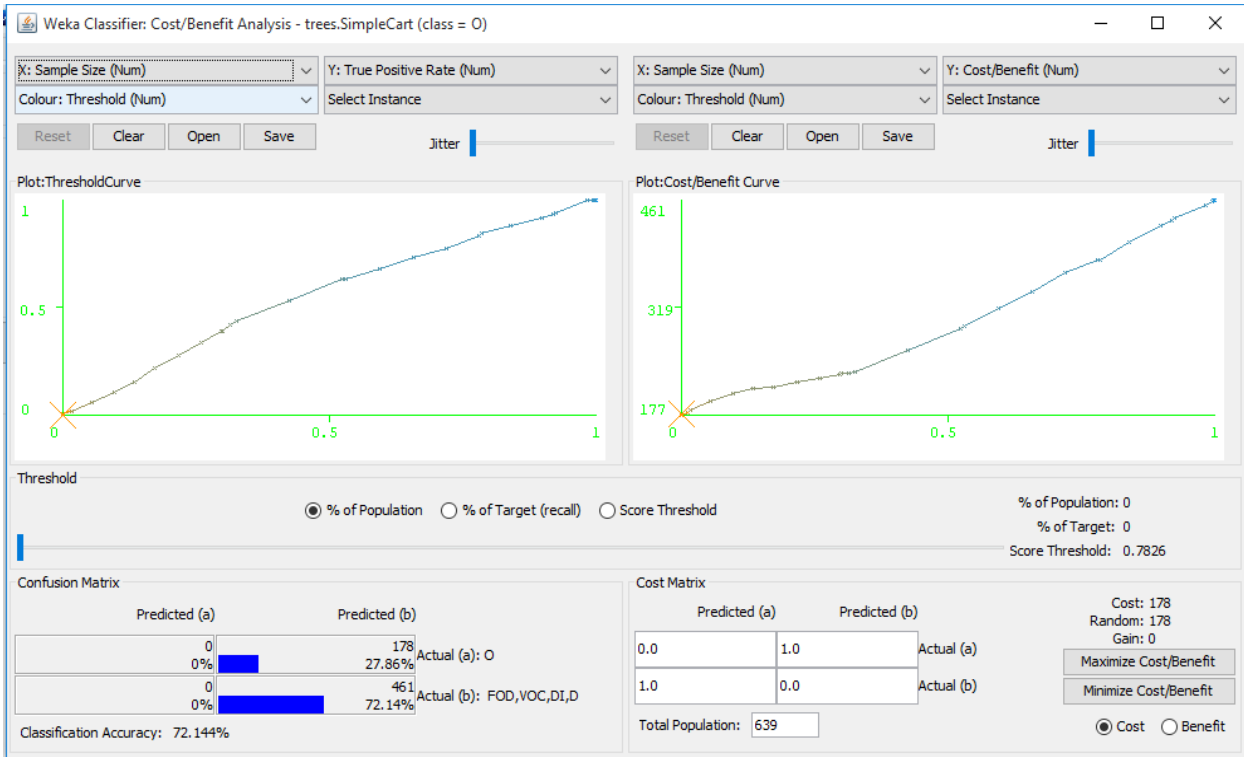
Total Population: 869

Cost: 680
Random: 680
Gain: 0

● Cost ○ Benefit

Figure 12: COST/BENEFIT ANALYSIS (CART) 2014

ANALYSIS: Its systematic architecture is used for calculating and comparing benefits and costs of underlying analysis, in above cost benefit analysis 680 instances are generated having accuracy of 78.25% which indicates the total cost of the damages occurs with other causes in data set of 2014 data set through ID3 classifier.



Weka Classifier: Cost/Benefit Analysis - trees.SimpleCart (class = O)

X: Sample Size (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance

X: Sample Size (Num)
Y: Cost/Benefit (Num)
Colour: Threshold (Num)
Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.7826

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	178 27.86%	Actual (a): O
0 0%	461 72.14%	Actual (b): FOD,VOC,DI,D

Classification Accuracy: 72.144%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 639

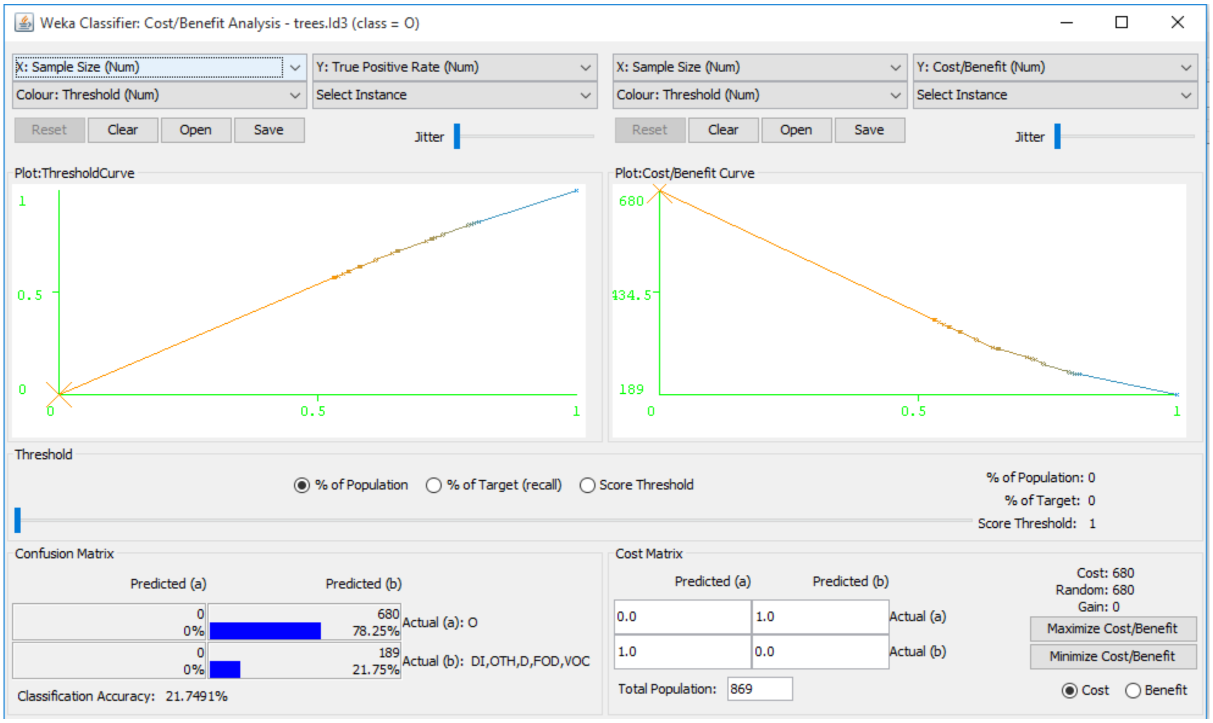
Cost: 178
Random: 178
Gain: 0

● Cost ○ Benefit

Figure 13: COST/BENEFIT ANALYSIS (CART) 2015

ANALYSIS: Its systematic architecture is used for calculating and comparing benefits and costs of underlying analysis, in above cost benefit analysis 680 instances are generated having accuracy of 27.86% which indicates the total cost of the damages occurs with other causes in data set of 2014 data set through CART classifier.

6.9 COST/BENEFIT ANALYSIS FOR OVERSPEEDING



Weka Classifier: Cost/Benefit Analysis - trees.Id3 (class = O)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Threshold (Num) | Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	680 78.25%	Actual (a): O
0 0%	189 21.75%	Actual (b): DI,OTH,D,FOD,VOC

Classification Accuracy: 21.7491%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

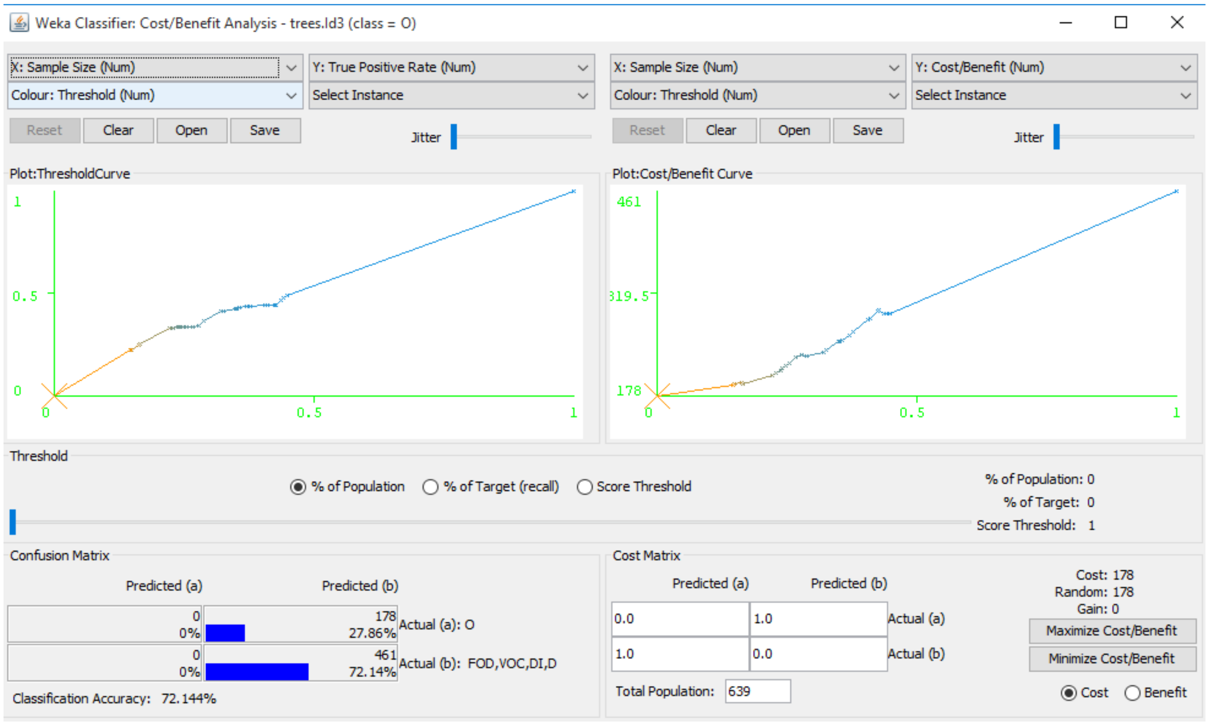
Total Population: 869

Cost: 680
Random: 680
Gain: 0

● Cost ○ Benefit

Figure 14: COST/BENEFIT ANALYSIS (ID3) 2014

ANALYSIS: In data set of 2014 the ID3 classifier is used which classifies 680 instances having 78.25% accuracy which concludes that the overspeeding in highways is not effective against cost issues as compare to the other causes of road accident analysis. For providing a basis for comparing analysis, it involves comparing the total expected cost of each option against the total expected loss.



Weka Classifier: Cost/Benefit Analysis - trees.Id3 (class = O)

X: Sample Size (Num) | Y: True Positive Rate (Num)
Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num)
Colour: Threshold (Num) | Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	178 27.86%	Actual (a): O
0 0%	461 72.14%	Actual (b): FOD,VOC,DI,D

Classification Accuracy: 72.144%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 639

Cost: 178
Random: 178
Gain: 0

● Cost ○ Benefit

Figure 15: COST/BENEFIT ANALYSIS (ID3) 2015

ANALYSIS: Its systematic architecture is used for calculating and comparing benefits and costs of underlying analysis, in above cost benefit analysis 680 instances are generated having accuracy of 27.86% which indicates the total cost of the damages occurs with other causes in data set of 2014 data set through ID3 classifier.

6.10 Cost/Benefit Analysis

Above analysis describes the actual rate due to overspeeding which is 27.86% for 178 instances in 2015 and 78.25% with 680 instances in 2014 with classifier ID3, similarly in 2014 CART predicts the actual rate 27.86% in 2015 for 178 instances and 78.25% in 2014 for 680 instances.

CHAPTER 7
CONCLUSION AND FUTURE WORK

A literature review has given a general idea in published studies on the relationship between road characteristics. In this analysis, we collected and cleaned traffic accident data, attempted to construct attributes, and tested predictive models. Finally, knowledge was presented in the form of data stimulation in WEKA, which helps in pin pointing the accidents pattern and causes, this results may also be found in different highways of India. Through this simulation we can take proper safety measures and precautions so that any type of fatal injury can be reduced for a better safety of individuals.

After observing the simulated results and various performance metrics it is clear that variation in the size of input data affects the outputs and performance of classifier significantly. In small data size simulation ID3 comes out to be best among others having large areas under ROC curve and all other parameters and when size of input data size gets large, CART performance decreases while others gives better results. Thus ID3 with appropriate selection of training data size which gives better results with all its different test functions and can be deployed in real scenario.

We extended the research to possible injury, causes and timeshift of an accident. Our experiments showed that the model for causes and timeshift performed better than other classes. The ability of predicting causes and timeshift is very important since drivers' fatality has the highest cost to society economically and socially.

In future work, we proposed using the IoT (Internet of Things) devices to collect data by using different type of sensors and use that data for better prediction using data mining and machine learning algorithms for better and frequent prediction so that proper safety measures can be taken.

REFERENCES

1. Global Burden of Disease Study 2013, Collaborators (22 August 2015). "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.". Lancet (London, England) **386** (9995): 743–800.
2. GBD 2013 Mortality and Causes of Death, Collaborators (17 December 2014). "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.". Lancet 385: 117–71.
3. "Global status report on road safety 2013: Supporting a decade of action" (PDF) (in English and Russian). Geneva, Switzerland: world health organization WHO. 2013. Retrieved 3 October 2014.
4. L. Rokach and O. Maimon, "Data mining with decision trees: theory and applications" World scientific, 2014.
5. N. Prasad,P. Kumar and M. Naidu,"An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree",4th International Conference on Intelligent Systems Modelling & Simulation (ISMS), IEEE ,january 2013 pp. 56-60.
6. I-Cheng Yeh, Che-hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", Expert Systems with Applications, vol.36, pp.2473-2480, 2009.
7. Chang L. and H. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques Accident analysis and prevention", Accident analysis and prevention, vol. 38(5), pp.1019-1027, 2006
8. Handan Ankarali Camdeviren, Ayse Canan Yazici, Zeki Akkus, Resul Bugdayci, Mehmet Ali Sungur, "A Comparison of logistic regression model and classification tree: An application to postpartum depression data", Expert Systems with Applications, vol.
   32, pp. 987–994, 2007.

9. Yong Soo Kim, “Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size”, Expert Systems with Applications, vol. 34, pp. 1227–1234 2008.

10. Andreas G.K. Janecek, Wilfried N. Gansterer, Michael A. Demel Michael, Gerhard F. Ecker, “On the Relationship Between Feature Selection and Classification Accuracy”, JMLR Workshop and Conference Proceedings, pp.90-105, 2008.

11. Nojun Kwak and Chong-Ho Choi, “Input Feature Selection for Classification Problems”, IEEE Transactions on Neural Networks, vol. 13, No. 1, 2002.

12. Isabelle Guyon, Andr´e Elisseeff, “An Introduction to variable and Feature Selection”, Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

13. Jaree Thongkam, Guandong Xu and Yanchun Zhang, “AdaBoost algorithm with random forests for predicting breast cancer survivability”, International Joint Conference on Neural Networks,2008.

14. 14. Eric Bauer, Ron Kohavi, “An Empirical Comparison of Voting Classification Algorithms: Bagging Boosting, and Variants”, Machine Learning, Vol.36, pp. 105-139, 1999.

15. S. Shanthi, Dr. R. Geetha Ramani, “Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms”, International Journal of Computer Applications, Vol.35, No.12, pp.30-37, 2011.

16. S. Shanthi, Dr. R. Geetha Ramani, “Classification of Seating Position Specific Patterns in Road Traffic Accident Data through Data Mining Techniques”, Proceedings of Second International Conference on Computer Applications, ICCA 2012, Vol.5, pp. 98-104, January, 2012.

17. Abdelwahab, H. T. and Abdel-Aty, M. A., Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. Transportation Research Record 1746, Paper No. 01-2234.

18. Yang, W.T., Chen, H. C., & Brown, D. B., Detecting Safer Driving Patterns By A Neural Network Approach. ANNIE ’99 for the Proceedings of Smart Engineering System Design Neural Network, Evolutionary Programming, Complex Systems and Data Mining, Vol. 9, pp 839-844, Nov. 1999.

19. Zembowicz, R. and Zytkow, J. M., 1996. From Contingency Tables to Various Forms of Knowledge in Database. Advances in knowledge Discovery and Data Mining, editors, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. AAAI Press/The MIT Press, pp.329-349.

20. Sohn, S. Y., & Lee, S. H., Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Traffic Accidents in Korea. Safety Science, Vol. 4, issue1, February 2003, pp. 1-14.

21. Mussone, L., Ferrari, A., & Oneta, M., An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention, Vol. 31, 1999, pp. 705-718.

22. Dia, H., & Rose, G., Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. Transportation Research C, Vol. 5, No. 5, 1997, pp. 313-331.

23. Shankar, V., Mannering, F., & Barfield, W., Statistical Analysis of Accident Severity on RuralFreeways. Accident Analysis and Prevention, Vol. 28, No. 3, 1996, pp.391-401.

24. Kim, K., Nitz, L., Richardson, J., & Li, L., Personal and Behavioral Predictors of Automobile Crash and Injury Severity. Accident Analysis and Prevention, Vol. 27, No. 4, 1995, pp. 469-481.

25. Abdel-Aty, M., and Abdelwahab, H., Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility. Accident Analysis and Prevention, 2003.

26. Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P., The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. Accident analysis and Prevention, Vol. 34, pp. 717-727, 2002

27. Evanco, W. M., The Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities. Accident Analysis and Prevention, Vol. 31, 1999, pp. 455-462

28. Ossiander, E. M., & Cummings, P., Freeway speed limits and Traffic Fatalities in Washington State. Accident Analysis and Prevention, Vol. 34, 2002, pp. 13-18.

29. Buzeman, D. G., Viano, D. C., & Lovsund, P., Car Occupant Safety in Frontal Crashes: A Parameter Study of Vehicle Mass, Impact Speed, and Inherent Vehicle Protection. Accident Analysis and Prevention, Vol. 30, No. 6, pp. 713-722, 1998.

30. Kweon, Y. J., & Kockelman, D. M., Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models. Accident Analysis and Prevention, Vol. 35, 2003, pp. 441-450.
31. Martin, P. G., Crandall, J. R., & Pilkey, W. D., Injury Trends of Passenger Car Drivers In the USA. Accident Analysis and Prevention, Vol. 32, 2000, pp. 541-557.
32. Mayhew, D. R., Ferguson, S. A., Desmond, K. J., & Simpson, G. M., Trends In Fatal Crashes Involving Female Drivers, 1975-1998. Accident Analysis and Prevention, Vol. 35, 2003, pp. 407-415.
33. Tavris, D. R., Kuhn, E. M, & Layde, P. M., Age and Gender Patterns In Motor Vehicle Crash injuries: Importance of Type of Crash and Occupant Role. Accident Analysis and Prevention, Vol. 33, 2001, pp. 167-172.