A comparative study of various approaches to explore factors for vehicle collision

Pragya Baluni¹, Mrs. Bhumiaka Gupta²

¹² Dept of computer science & engineering, G.B Pant Engineering College, Pauri, Uttarakhand, India

Abstract: Data mining an non-trivial extraction of novel, implicit, and actionable knowledge from large data sets is an evolving technology which is a direct result of the increasing use of computer databases in order to store and retrieve information effectively. It is also known as Knowledge Discovery in Databases (KDD) and enables data exploration, data analysis, and data visualization of huge databases at a high level of abstraction, without a specific hypothesis in mind. Understanding the circumstances under which drivers and passengers are more likely to be killed or more severely injured in an automobile accident can help improve the overall driving safety situation. Factors that affect the risk of increased injury of occupants in the event of an automotive accident include demographic or behavioral characteristics of the person, environmental factors and roadway conditions at the time of the accident occurrence, technical characteristics of the vehicle itself, among others. Classifying and mining data can help in improving safety services and increasing the number of emergency services at the various locations. It can help in providing better guidelines for drivers to avoid collisions. This paper presents a comparative study of different approaches to explore various factors for vehicle collision. The results, mostly validated by the findings of previous studies, provide insight into the changing importance of crash factors with the changing injury severity levels. Experimental results revealed that decision trees outperformed neural networks.

Keywords: Data mining, KDD

1. INTRODUCTION

The ever increasing tremendous amount of data, collected and stored in large and numerous data bases, has far exceeded human ability for comprehension without the use of powerful tools. Consequently, important decisions are often made based not on the information rich data stored in databases but rather on a decision maker’s intuitions due to the lack of tools to extract the valuable knowledge embedded in the vast amounts of data. This is why data mining has received great attention in recent years. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, and pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis [1].

1.2 Functionalities

1.2.1. Characterization

Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

1.2.2. Discrimination

Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class.

1.2.3. Association analysis

Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability that an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.

1.2.4. Classification

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

1.2.5. Prediction

There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data and is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is referred to the forecast of missing numerical values, or increase/decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.
1.2.6. Clustering
It is the organization of data in classes and is similar to classification. In clustering class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification as the classification is not dictated by given class labels. There are many clustering approaches based on the principle of maximizing the similarity between objects in same class called intra-class and minimizing the similarity between objects of different classes called inter-class similarity.

1.2.7. Outlier analysis
Outliers are data elements that cannot be grouped in a given class or cluster. They are also known as exceptions or surprises and are often very important to identify. Outliers can reveal important knowledge in other domains, can be very significant and their analysis valuable.

1.2.8. Evolution and deviation analysis:
This pertains to the study of time related data that change with time. Evolution analysis model’s evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values [2].

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

2) Relational Databases: Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples.

3) Data Warehouses: A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof.

4) Transaction Databases: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items.

5) Multimedia Databases: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system.

6) Spatial Databases: Spatial databases are databases that store geographical information like maps, and global or regional positioning in addition to usual data.

7) World Wide Web: The World Wide Web is the most heterogeneous and dynamic repository available. Data in the World Wide Web is organized in inter-connected documents [2].

1.3 Road accidents and data mining:
Road safety experts and researchers deal with large volumes of quantitative information and collected statistics, in order to understand and estimate the social and economic cost of the accidents and to be able to introduce safety plans in order to prevent or reduce occurrences of accidents. The road traffic and accident statistics must be presented in such a way to make it easier to be both recognized and interpreted by a human operator. Previous works on accident analysis included statistical methods and formal techniques. Statistics tables and ordinary charting techniques are not sufficient for present day requirements and this causes difficulties in the effective visualization of results and patterns. Another disadvantage is that ordinary methods limit human involvement in the exploration tasks[3].

Traffic control system is one of the various areas, where critical data about the well-being of the society is recorded and kept. Various aspects of a traffic system like vehicle accidents, traffic volumes and concentration are recorded at different levels. In connection to this, injury severities resulted from road traffic accident are one of the areas of concern. The costs of fatalities and injuries due to traffic accidents have a great impact on the society. In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents. There are several approaches that researchers have employed to study this problem. These include neural network, nesting
logic formulation, log-linear model, fuzzy ART maps and so on[4].

2. RELATED WORK

An algorithm was proposed to study the application of BNN models for predicting motor vehicle crashes. To accomplish this objective, a series of models was estimated using data collected on rural frontage roads in Texas. Three types of models were compared: BPNN, BNN and the Negative Binomial (NB) regression models. The results of this study presented that in general both types of neural network models perform better than the NB regression model in terms of data prediction. Although the BPNN model can occasionally provide better or approximately equivalent prediction performance compared to the BNN model, in most cases its prediction performance is worse than the BNN model. In addition, the data fitting performance of the BNN model is consistently worse than the BNN model, which suggests that the BNN model has better generalization abilities than the BPNN model and can effectively alleviate the over-fitting problem without significantly compromising the nonlinear approximation ability. The results also proposed that BNNs could be used for other useful analyses in highway safety, including the development of accident modification factors and for improving the prediction[5]. The paper first described the fundamental principles of NB regression models commonly used in highway safety and the characteristics of neural network models. The review has shown that although neural network models have excellent function approximation abilities and do not require specifying a functional form linking the dependent variable to the explanatory variables, the over-fitting problem has significantly limited their application in highway safety[5]. Another paper presented the analysis of powered two-wheeler (PTW) crashes in Italy in order to detect interdependence as well as dissimilarities among crash characteristics and provide insights for the development of safety improvement strategies. Data mining techniques were used to analyze the data relative to the 254,575 crashes involving PTWs occurred in Italy in the period 2006–2008. Classification trees analysis and rules discovery were performed. Tree-based methods are nonlinear and non-parametric data mining tools for supervised classification and regression problems. They do not require a priori probabilistic knowledge about the phenomena under studying and consider conditional interactions among input data. Rules discovery is the identification of sets of items that occur together in a given event more often than they would if they were independent of each other. Both the classification trees and the rules discovery were effective in providing meaningful insights about PTW crash characteristics and their interdependencies. Even though in several cases different crash characteristics were highlighted[6]. Durmus Delen, Ramesh Sharda, Max Bessonov[7] proposed factors that affect the risk of increased injury of occupants in the event of an automotive accident include demographic or behavioral characteristics of the person, environmental factors and roadway conditions at the time of the accident occurrence, technical characteristics of the vehicle itself, among others. This study used a series of artificial neural networks to model the potentially nonlinear relationships between the injury severity levels and crash-related factors. It then conducted sensitivity analysis on the trained neural network models to identify the prioritized importance of crash-related factors as they apply to different injury severity levels. In the process, the problem of five-class prediction is decomposed into a set of binary prediction models (using a nationally representative sample of 30,358 police-recorded crash reports) in order to obtain the granularity of information needed to identify the “true” cause and effect relationships between the crash-related factors and different levels of injury severity. Eight binary MLP neural network models were developed with different levels of injury severity as the dependent variable. These eight models presented different levels of injury severity varying from the no-injury to fatality and from fatality to no-injury. All models were found to have better predictive power as compared to a model with a five-category outcome variable. In addition, this structure helped to identify the important explanatory variables at each level of distinction between the injury severities. Three another paper that presented The study intended to provide insight into pedestrian accidents by uncovering their patterns in order to design preventive measures and to allocate resources for identified problems. Kohonen neural networks are applied to a database of pedestrian fatal accidents occurred during the four-year period between 2003 and 2006. Results showed the existence of five pedestrian accident patterns: (i) elderly pedestrians crossing on crosswalks mostly far from intersections in metropolitan areas; (ii) pedestrians crossing suddenly or from hidden places and colliding with two-wheel vehicles on urban road sections; (iii) male pedestrians crossing at night and being hit by four-wheel vehicles on rural road sections; (iv) young male pedestrians crossing at night wide road sections in both urban and rural areas; (v) children and teenagers crossing road sections in small rural communities. From the perspective of preventive measures, results suggested the necessity of designing education and information campaigns for road users as well as allocating resources for infrastructural interventions and law enforcement in order to address the identified major problems[8]. DTs allow accident classification based on crash severity. They provide an alternative to parametric models due to their ability to identify patterns based on data, without the need to establish a functional relationship between variables. Moreover, such classification models can be used to determine interactions between variables that would be impossible to establish directly, using ordinary statistical modelling techniques[9].
Miao M. Chong, Ajith Abraham, Marcin Paprzycki[10] presented in a paper the severity of injury resulting from traffic accidents using artificial neural networks and decision trees. They applied them to an actual data set obtained from the National Automotive Sampling System (NASS) General Estimates System (GES). The experiments also showed that the model for fatal and non-fatal injury performed better than other classes. The ability of predicting fatal and non-fatal injury is very important since drivers’ fatality has the highest cost to society economically and socially. It has to be stressed again that it is a well-known fact that one of the most important factor differentiating injury level is the actual speed that the vehicle was going when the accident happened. Our dataset doesn’t provide enough information on the actual speed, since speed for 67.68% of the data records was unknown. If the speed was available, it might have helped to improve the performance of the two considered models. Kim[11] 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 driver behaviors of alcohol or drug use and lack of seat belt use greatly increase the odds of more severe crashes and injuries. Shankar [12] applied a nested logic formulation for estimating accident severity likelihood conditioned on the occurrence of an accident. The study 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. Dia used real-world data for developing a multilayered NN freeway incident detection model [13]. They compared the performance of the neural network model and the incident detection model in operation on freeways. The use of data mining to improve road safety can be categorised into two major approaches. The first approach concentrates on mining crash data, which includes various attributes relating to both driver and vehicle at the time of the crash [14,15,16]. The focus is on analyzing the data for the purpose of discovering useful, and potentially actionable, information. In [16] crash data was mined to identify the driver and vehicle attributes which are the main causes for road accidents. Principal Component Analysis was used to emphasize the relationships between characteristics such as age, gender and vehicle type, to the crash variables. Bedard [17] applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers’ fatality risk. It was found that increasing seatbelt use, reducing speed, and reducing the number and severity of driverside impacts might prevent fatalities. Sohn [18] 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 accidents. The individual classifiers used were neural network and decision trees. 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.

3. ARCHITECTURAL DESIGN

3.1 The Architecture of Data Mining:

The architecture of a typical data mining system has the following major components:

3.1.1. Database, data warehouse, or other information repository:

This component is one or a set of databases, data warehouses, spread sheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

3.1.2. Database or data warehouse server:

The component is responsible for fetching the relevant data, based on the data mining request of the user.

3.1.3. Knowledge base:

This is the domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. It includes concept hierarchies that are used to organize attributes or attribute values into different levels of abstraction.

3.1.4. Data mining engine:

This is an essential component of the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association analysis, classification, evolution and deviation analysis.

3.1.5. Pattern evaluation module:

This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns. It can access interestingness thresholds stored in the knowledge base. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. Efficient data mining is possible by pushing the evaluation of pattern interestingness deeply into the mining process so as to connect the search to only the interesting patterns.

3.1.6. Graphical user interface:

This module communicates between users and the data mining system, and allows the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. This component also allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

3.2 The Process of Data Mining

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections
to some form of new knowledge. The iterative process consists of the following steps:

3.2.1. Data cleaning:
It can also be termed as data cleansing. It is a phase wherein noise data and irrelevant data are removed from the collection.

3.2.2. Data integration:
In this stage multiple data sources that are heterogeneous can be combined in a common source.

3.2.3. Data selection:
In this phase the data relevant to the analysis is decided on and retrieved from the data collection.

3.2.4. Data transformation:
It can also be known as data consolidation and it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

3.2.5. Data mining:
This is a crucial phase wherein skillful techniques are applied to extract patterns potentially useful.

3.2.6. Pattern evaluation:
In this current phase strictly interesting patterns representing knowledge are identified with respect to the given measures.

3.2.7. Knowledge representation:
This is the final phase in which the discovered knowledge is visually represented to the user. This is an essential step which uses visualization techniques to help users understand and interpret the data mining results. The computed values for proportions and chi square are then compared with results obtained manually.

4. METHODOLOGY

4.1 Decision Trees
A DT is a predictive model which can be used to represent both classifiers and regression models. DTs are popular due to their simplicity and transparency; moreover, they are usually presented graphically as hierarchical structures, which make them easy to interpret. A DT is a simple structure that can be used as a classifier. Within a DT, each node represents an attribute variable X and each branch represents one of the states of this variable. Normally, a terminal node, or leaf, specifies the expected value of the class variable or variable in study C, depending on the information contained in the training data set, i.e. the set used to build the model. The set of data used to check the model is called test set. When we obtain a new instance or case of the test data set, we can make a decision or prediction about the state of the variable class following the path to the tree from the root node to a leaf node, using the sample values and the tree structure. Subsequently, the model obtained can be used to classify new examples (cases whose classes are not known a priori), to detect patterns, or simply to gain a better understanding of the phenomenon being analyzed. DTs are built recursively, following a descending strategy, starting with the full data set (made by the root node). Using specific split criteria, the full set of data is then split into even smaller subsets. Each subset is split recursively until all of them are pure (when the cases in each subset are all of the same class) or their “purity” cannot be increased. That is how the tree’s terminal nodes are called feature or predictor variable, are formed, which are obtained according to the answer values of the class variable. The main difference between DTs building procedures lies in the splitting criteria.

Methods for building decision trees:

4.1.1. CART:
Depending on the nature of the dependent variable, a classification tree (case discrete) or a regression tree (case continuous) will be built. The CART model generates binary trees by using impurity as a measure to split the Gini Index of diversity (which is a measure of the diversity of classes in a tree node being used). For variable C it is defined as:

$$g_{in}(C) = 1 - \sum_j p_j^2 (C = c_j)$$

The best split is the one that minimizes GIx(C,X). With this procedure, the maximal tree that overfits the data is created. To decrease its complexity, the tree is pruned using a cost-complexity measure that combines the precision criteria as opposed to complexity in the number
of nodes and processing speed, searching for the tree that obtains the lowest value for this parameter.

4.1.2. ID3:

Builds a tree in a manner similar to the CART method but without the binary restriction. It can only be used with discrete variables, does not allow pruning and the function used to measure impurity is the Shannon’s entropy (Shannon, 1948), which is an information-based uncertainty measure. The ID3 algorithm uses the Information Gain criterion to choose which attribute goes into a decision node. Information Gain could be defined as a difference of entropies in the current node, considering the information that an attribute variable gives us about the class variable. This split criterion can therefore be defined on an attribute variable X, given the class variable C, as follow:

$$\text{IG}(C, X) = \text{IG}(C, X) = \text{H}(C) - \sum_{j} p(C = c_j) \text{H}(C|X)$$

where H(C) is the entropy of C, \( \text{H}(C) = \sum_j p(C = c_j) \text{log}_2 p(C = c_j) \), with \( p(C = c_j) = p(C = c_j) \), the probability of each value of the variable class estimated in the training data set. In the same way, \( \text{H}(C|X) = -\sum_x \sum_{t,j} p(c_j|x_t) \text{log}_2 p(c_j|x_t) \), where \( x_t, t = 1, \ldots, |X| \), is each possible state of X and \( c_j, j = 1, \ldots, k \), each possible state of C.

4.1.3. C4.5:

In order to improve the ID3 algorithm, Quinlan (1993) introduces the C4.5 algorithm, where the Information Gain split criterion is replaced by an Information Gain Ratio criterion which penalizes variables with many states. Moreover, this model makes it possible to deal with continuous attributes and missing values, and to carry out a post-pruning process. The algorithm incorporates classification tree pruning once a tree has been induced, by applying a hypothesis test on whether or not to expand a branch. The Information Gain Ratio of an attribute variable X on a variable class C can be expressed as:

$$\text{IGR}(C, X) = \frac{\text{IG}(C, X)}{\text{H}(X)}$$

Results and Discussion:

The first step was to build DTs using the three algorithms (CART, C4.5 and ID3) with the aim of classification using 10 x10-fold CV procedure. In order to compare the results, corrected paired t-tests were conducted. The results of the tests, comparing the methods to each other on the indicators accuracy, sensitivity and specificity. C4.5 and CART show similar values for accuracy. ID3 shows significantly worse results than the other two algorithms. The accuracy values are within the range of values obtained in other studies in which classification methods with similar objectives were applied.

Table 1: Comparison of the parameters produced by the various algorithms.

<table>
<thead>
<tr>
<th></th>
<th>CART</th>
<th>C4.5</th>
<th>ID3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>55.87</td>
<td>54.16</td>
<td>52.72*</td>
</tr>
<tr>
<td>SENSITIVITY</td>
<td>54.00</td>
<td>55.00</td>
<td>53.00</td>
</tr>
<tr>
<td>SPECIFICITY</td>
<td>58.00</td>
<td>54.00</td>
<td>52.00</td>
</tr>
<tr>
<td>ROC AREA</td>
<td>57.00</td>
<td>54.00</td>
<td>53.00*</td>
</tr>
</tbody>
</table>

* The results worsen significantly.

The C4.5 algorithm gives a higher value than CART (55% vs. 54%) in the sensitivity parameter analysis. The improvement is not significant, however. CART gives a higher value than C4.5 for the specificity parameter, although the improvement is not significant either. For ID3, both sensitivity and specificity are poorer, in comparison to the values of the other two algorithms. A global measure given by the ROC area indicator shows that C4.5 gives the best results (57%) whereas ID3 obtains the lowest values again (53%). The computational time it took each algorithm to build the DT was another indicator analysed. It was obtained that the C4.5 algorithm requires the most time to build a tree, being 55 times slower than for the C4.5 algorithm and 42 times slower than ID3. C4.5 is the algorithm that takes the less time, needing only 0.03 s to build a DT with 19 variables and 1801 data. This result is logical because the CART algorithm is more complex, and in turn, C4.5 is more complex than ID3, since it has more optimization parameters in order to improve the results. The implementation of the C4.5 algorithm is optimized in Weka, and therefore the computational time is lower than for ID3. Taking the above results in consideration, it can be seen that the ID3 algorithm is the method that gives the worst results. The difference in improvement using CART and C4.5 is not significant, however. Although CART obtains slightly higher values in the precision and specificity parameters analysed, the improvement is not significant, and therefore, we cannot assert a priori that
One method is better than the other. It would be worthwhile to analyse the decision rules obtained with the algorithms that attained the best results: C4.5 and CART.

### TABLE 2: Description of the rules according to the CART.

<table>
<thead>
<tr>
<th>Node/rule</th>
<th>Rules</th>
<th>Then</th>
<th>S(%)</th>
<th>Po(%)</th>
<th>P(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>IF (SEX = M) AND (ACT = ROR OR ACT = CP) AND (DAY = WD OR DAY = BPH) AND (TIM = [6–12]) AND (PAS = N)</td>
<td>KSSI</td>
<td>1.59</td>
<td>2.38</td>
<td>66.6</td>
</tr>
<tr>
<td>5</td>
<td>IF (SEX / M) AND (LIG = IL OR LIG = WL)</td>
<td>KSI</td>
<td>2.22</td>
<td>3.65</td>
<td>60.8</td>
</tr>
<tr>
<td>15</td>
<td>IF (SEX = M) AND (ACT = ROR OR (ACT = CP) AND (ATF = LR) AND (DAY = WD OR DAY = BPH) AND (TIM = [6–12]) AND (PAS = N))</td>
<td>SI</td>
<td>4.60</td>
<td>6.51</td>
<td>70.7</td>
</tr>
<tr>
<td>6</td>
<td>IF (SEX ≠ M) AND (LIG ≠ IL OR LIG ≠ WL)</td>
<td>SI</td>
<td>8.02</td>
<td>11.67</td>
<td>68.7</td>
</tr>
<tr>
<td>4</td>
<td>IF (SEX = M) AND (ACT = RO OR ACT = CO OR ACT = OT)</td>
<td>SI</td>
<td>5.08</td>
<td>7.94</td>
<td>64.0</td>
</tr>
<tr>
<td>8</td>
<td>IF (SEX = M) AND (ACT = ROR OR ACT = CP) AND (ATF = LR)</td>
<td>SI</td>
<td>4.05</td>
<td>6.43</td>
<td>62.9</td>
</tr>
</tbody>
</table>

Fig. 3 shows the DT built using the CART method with 70% of the data for training and the remaining data (30%) for testing. The CART method creates a tree with 19 nodes and 10 terminal nodes. Table 1 shows a description of the six rules identified in the DT that verify the minimum values of the parameters S, Po and P in the training and in the test sets. Support varies from 1.6% (rule 16) to 8.0% (rule 6). All the rules include at least 1% of the population, and probability values are higher than 60.9%, with 70.7% being the highest value (rule 15).

With regards to the binomial test that was performed, all the rules obtained from the training set with the minimum threshold have a grade of different than 1. Hence the antecedent and consequent are independent. These results were not included in the paper because they are not important for our aims. The binomial test showed that all the rules given in Table 3 have no significant differences (at 0.05 level), based on support when they are applied on the test set. Only the rule 5 (see Table 3) has a high level of support in the test set compared to the support in the training set. This difference is significant at 0.05 level of significance. The root variable that generates the tree is SEX (see Fig. 3) which splits into two branches (nodes 1 and 2). For female drivers, and depending on LIG, nodes 5 and 6 are obtained, with different degrees of severity (see Fig. 3): accidents are KSI if LIG is insufficient or without lighting, with a probability of 61% (rule 5); while if LIG is sufficient, dusk or day light the severity is SI, with a probability of 69% (rule 6). This result shows a direct relationship between KSI accidents and female drivers on rural highways with insufficient or without lighting. The rest of the rules are attributable to male drivers (node 1). This result is coherent with the study data, given that in 84.5% of the accidents analysed the drivers were men. After this node, the tree splits according to ACT. The accident type has been identified in several previous as one of the key variables in analyses of accident severity. This study shows that if the accident type is rollover, collision with obstacles or other accidents types the probability of SI is 64.0% (rule 4 in Table 2). However, in the case of run off road or collision with pedestrian the probability of KSI is higher than the probability of SI (node 3 in Fig. 3). So, in this kind of facilities road safety managers should pay attention to this type of accidents (run off road and collision with pedestrian). Node 3 (Fig. 3) splits by the variable ATF: if ATF is light rain the accident is SI, with a probability of 63% (rule 8 in Table 2). This result proves that drivers try to be very careful under bad atmospheric conditions. In other cases, the tree continues to grow according to DAY. If DAY is on a weekend or public holiday (PH) or a working day after the weekend or public holiday (APH) the accident is KSI, with a probability of 65% (node 10 in Fig. 3). This result is coherent with the trend observed in Spain, where most of fatalities in road accidents occur on weekends (31.4% of the car accidents in 2009 occurred at the weekends, in which 818 deaths were recorded, that is 38.4% of the total number of fatalities in the year 2009). When DAY is a working day before the weekend or public holiday (BPH) or a regular working day (WD) the tree is divided according to TIM. From this point of DT’s structure, the rule interpretation is difficult because many variables are involved in the accident. However, the following results are highlighted: almost 71% whereas when it is not paved the severity is KSI (rule 16); and from [12–18] h, tree is divided by MON and LIG (see Fig. 3), however neither of the obtained nodes are rules because they do not meet the threshold limits for S, Po or P. Following Eq. (8), it is possible to obtain the importance of the variables in the model. Table 4 shows the normalized importance of these variables. 12 variables were detected as having the greatest influence on accident severity, with percent which varying from 100% to 9.9%. LIG is the most important variable, coinciding with previous studies. TIM has 77.1% importance in the model which is coherent because there is already a degree of relationship between the time and lighting variables.

### TABLE 3: Importance of the variables with CART.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Importance normalized(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The rules are generated from the training set with the minimum threshold have a severity of SI with a probability of almost 69%. When PAW is >7 m, the tree splits according to NOI, and when it is higher than 1, accidents are SI (seriously injured) in 64.4% of cases (rule 48); but when NOI is 1 and PAS is non-existent or impassable (rule 50) accidents are also SI in 70.8% of cases. These last three rules (rules 46, 48 and 50) are less useful to policy makers because they imply a combination of many more variables than in the preceding rules (6, 6 and 7 variables respectively), which makes it difficult to interpret the results and impossible to take direct preventive measures. Fourteen variables were detected as having the greatest influence on accident severity, with percent which varying from 100% to 11.2%. ACT is the most important variable in the C4.5 model, followed by CAU. The CART algorithm identified eleven of the previous fourteen variables. Moreover C4.5 identified VEH, PAW, and NOI.

Table 4: Description of the rules according to the C4.5.

<table>
<thead>
<tr>
<th>NODE/ RULE</th>
<th>RULE C4.5 IF.</th>
<th>THEN</th>
<th>S%</th>
<th>Po%</th>
<th>P%</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>IF (SEX = M) AND (ACT = CP) AND (PAS = Y)</td>
<td>KSI</td>
<td>4.68</td>
<td>77.97</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = MOT) AND (PAS = NE)</td>
<td>KSI</td>
<td>3.33</td>
<td>76.19</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = TRU)</td>
<td>KSI</td>
<td>3.17</td>
<td>67.50</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = CAR) AND (PAW = WID)</td>
<td>KSI</td>
<td>8.02</td>
<td>64.36</td>
<td></td>
</tr>
</tbody>
</table>
4.2. NEURAL NETWORKS:

ANN method is generally used for modeling the variables which are in nonlinear relation and good results are obtained [20]. Since the variables used in this study are in nonlinear relation, ANN method is preferred. In the application of ANN, a program code was written by using Matlab software. In the program, a software was developed being able to do cycling among transferring functions (tansigpurelinlogsig), training functions (trainbr-trainlm) and in the number of neuron (1,2,3,4,…….) in hidden layer. Other features of this developed software are that the networks’s taking the Mean Squared Error (MSE) as a criterion of determination performance, it is being able to change iteration number dependent on desires and the network’s putting an end to its training as it is desired. In the structures of alternative network input layer and neuron number in hidden layer, output layer, training function, transferring functions among the layers, R2, MSE and Akaiake Information Criteria (AIC) values were provided with its being able to be read in Microsoft Excel. Therefore, the most suitable network structure can be easily determined.

4.2.1. ANN:

ANN can be defined as a system designed to model the method to fulfill as a function of brain. An ANN is formed by being connected in various shapes of artificial neural cells with each other. ANN has an ability of gathering information, saving and generalization this information with connection weights between the cells after passing learning algorithm and learning process [19]. Generally, an artificial cell model consists of five components. These are; input, weights, transfer function, activation function and output (Figure 1). Features and superior sides of ANN are being nonlinear, parallelism, easiness of being reality, processing of local information, error tolerance, generalization, adaptation, hardware speed, learning, analysis and easiness of . Outputs of network are calculated by spending from an activation function of input data (Equation 1).

\[ y = \sum \alpha_{Wi} \times \tau \]

Where,
\[ w \]: Weight matrices of cell,
\[ x \]: Input vector cell,
\[ v \]: Net input cell,
\[ y \]: Output cell.

According to the features of transfer functions, data used in ANN must be calculated to a defined scale. Therefore, there must be minimum and maximum values existing in data set and must be dependent on according to the scales given below (Equation 2).

\[ Y_{new} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \]

Where,
\[ Y \]: original series
\[ Y_{min} \]: The minimum value of the original series
\[ Y_{max} \]: The maximum value of the original series

Learning algorithm in ANN

Learning algorithm consists of the main frame of an ANN. In the many learning algorithm, the most common ones are trainlm and trainbr.These two algorithms are called as Levenberg – Marguardt. Trainlm, which has an algorithm that reduce memory requirements, is used when a set of learning is very large. This algorithm is the fastest of modern learning algorithm. However, trainbr is a form of enhanced Tainlm algorithm. Trainbr algorithm is a learning method regulating bayesian. This method is a learning algorithm developed to improve the generalization skill of Levenberg – Marguardt. Trainbr algorithm partly reduces the problem of how much the structure of optimum network must be [21].
RESULTS AND DISCUSSION:
In this study, the prediction of the accidents occurring in coming months with ANN method was aimed by using accident data in recent months. When compared 6 models for ANN analysis, the best model is 2 neurons in input layer model which has minimum MSE value, maximum R2 value and AIC. In this study, the neuron number 6 in the hidden layer and the activation function trainbr and network structure for MSE’s (0.004) was accepted as the best network. At the end of the application, ANN -2 (two cells input model) model was determined as the most significant model as statistical. With monthly accident data occurring between the years 1994 to 2007, seasonal fluctuation in the months to come and the number of accidents were predicted. Therefore, an attention is attracted to accidents problems, more reliable analyses about the number of traffic accidents are made and it is emphasized that precautions are increased to prevent the accidents.

Table 6: ANN application results.

<table>
<thead>
<tr>
<th>numbers</th>
<th>AIC</th>
<th>MSR</th>
<th>R^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_1, Q_2</td>
<td>-2.31</td>
<td>-0.512</td>
<td>0.96</td>
</tr>
<tr>
<td>Q_1, Q_2, Q_3</td>
<td>-8.81</td>
<td>-0.004</td>
<td>0.98</td>
</tr>
<tr>
<td>Q_1, Q_2, Q_3, Q_4</td>
<td>-2.86</td>
<td>-0.049</td>
<td>0.97</td>
</tr>
<tr>
<td>Q_1, Q_2, Q_3, Q_4, Q_5</td>
<td>-3.12</td>
<td>-0.014</td>
<td>0.99</td>
</tr>
<tr>
<td>Q_1, Q_2, Q_3, Q_4, Q_5, Q_6</td>
<td>-3.30</td>
<td>-0.043</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figure 5: Network Architecture for ANN.

Figure 6: Network architecture sample used in this study.

Figure 7: One cell input model results (ANN-1).
5. CONCLUSION

In this paper we have analyzed the road accident patterns on the basis of data mining techniques. From above survey and studies it has been observed that DTs allow accident classification based on crash severity. They provide an alternative to parametric models due to their ability to identify patterns based on data, without the need...
to establish a functional relationship between variables. Moreover, such classification models can be used to determine interactions between variables that would be impossible to establish directly, using ordinary statistical modelling techniques. DTs permit certain potentially useful rules to be determined that can be used by road safety analysts and managers. Initially, they should focus on severe crashes and subsequently intervene in minor accidents.

Another approach of artificial neural networks was successfully implemented. Artificial Neural Network (ANN) was successfully performed on monthly traffic accidents in terms of number of accidents, injuries and deaths. Consequently, although the accident characteristic with months is quite complex with months is quite complex, the proposed ANN approach has been able to model them. It was observed that decision trees outperformed neural network. Hence further modifications can be done in order to reduce the complexities of neural network.

REFERENCES