Extracting decision rules from police accident reports through decision trees

By: Juan de Oña, Griselda López and Joaquín Abellán

Juan de Oña, Griselda López and Joaquín Abellán (2013) *Extracting decision rules from police accident reports through decision trees*. **Accident Analysis and Prevention, 50, 1151– 1160.**

Direct access to the published version: http://dx.doi.org/10.1016/j.aap.2012.09.006

Extracting decision rules from police accident reports through decision trees

Juan de Oña[a,]*, Griselda López[a] and Joaquín Abellán[b]

[a] TRYSE Research Group. Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain)

[b] Department of Computer Science & Artificial Intelligence, ETSI Informática, c/Periodista Daniel Saucedo Aranda, s/n, 18071 Granada (Spain)

* Corresponding author. Phone: +34 958 24 99 79, jdona@ugr.es

Extracting decision rules from police accident reports through decision trees

ABSTRACT

Given the current number of road accidents, the aim of many road safety analysts is to identify the main factors that contribute to crash severity. To pinpoint those factors, this paper shows an application that applies some of the methods most commonly used to build Decision Trees (DTs), which have not been applied to the Road Safety field before. An analysis of accidents on rural highways in the province of Granada (Spain) between 2003 and 2009 (both inclusive) showed that the methods used to build DTs serve our purpose and may even be complementary. Applying these methods has enabled potentially useful decision rules to be extracted that could be used by road safety analysts. For instance, some of the rules may indicate that women, contrary to men, increase their risk of severity under bad lighting conditions. The rules could be used in road safety campaigns to mitigate specific problems. This would enable managers to implement priority actions based on a classification of accidents by types (depending on their severity). However, the primary importance of this proposal is that other databases not used here (i.e. other infrastructure, roads and countries) could be used to identify unconventional problems in a manner easy for road safety managers to understand, as decision rules.

Keywords: traffic accident; severity; decision trees; CART; C4.5; decision rules

1. Introduction

Traffic accidents are considered a major public health problem worldwide, claiming 1.27 million annual deaths and between 20 and 50 million injuries (WHO, 2009). Therefore, the aim of many studies to date has been to understand and identify the main factors that have an impact on road accident severity. Regression-type generalized linear models, Logit models and Probit models have been the techniques most commonly used to conduct such analyses (Kashani and Mohaymany, 2011; Savolainen et al., 2011; Mujalli and de Oña, in press). However most of them have their own model assumptions and pre-defined underlying relationships between dependent and independent variables (Chang and Wang, 2006).

Recently, data mining (DM) techniques have been used to study crash-injury severities by different researchers (Kuhnert et al. 2000; Sohn and Shin, 2001; Chang and Wang, 2006; Kashani and Mohaymany, 2011; Kashani et al., 2011; Pakgohar et al., 2010). The term Decision Trees (DTs) encompasses a series of techniques for extracting processable knowledge, implicit in databases, which is based on artificial intelligence and statistical analysis. One part of the DM could be defined as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data (Fayyad et al., 1996). These techniques are aimed at extracting knowledge from large amounts of data previously unknown and indistinguishable. DT techniques are particularly appropriate for studying crashes because they are non-parametric techniques that do not require prior probabilistic knowledge on the study phenomena. Furthermore, they consider conditional interactions among input data (Montella et al., 2011). Other advantages of DTs compared to other methods with similar aims include the extraction of decision rules of the "if-then" type (Kashani et al., 2011), and that they can be used to discover behaviours that occur within a specified set of data. Moreover, conclusions on behaviour can be drawn from the structure of DTs to understand the events leading up to a crash and identify the variables that determine how serious an accident will be.

There are many algorithms that can be used to build DTs, but CART (Classification and Regression Trees) developed by Breiman et al. in 1984 is the once most commonly used to

analyse crash severity. Authors such as Kuhnert et al. (2000) compared the results obtained with CART, multivariate adaptive regression splines (MARS) and logistic regression in the analysis of an epidemiological case-control study of injuries resulting from motor vehicle accidents. The findings indicated that non-parametric techniques such as CART and MARS can provide more informative and attractive models whose individual components can be displayed graphically. Chang and Wang (2006) studied the relationships between crash severity with characteristics related to drivers and vehicles, as well as variables related to roads, road accidents and the environment characteristics. Pakgohar et al. (2010) used CART and Multinomial Logistic Regression to study the role played by drivers' characteristics in the resulting crash severity. They found that the CART method provided more precise results, which were also simpler and easier to interpret. Kashani et al. (2011) studied the key factors that affect the injury severity of drivers involved in crashes on two-lane two-way rural roads. Subsequently, Kashani and Mohaymany (2011) used CART to identify the main factors that affect the injury severity of vehicle occupants involved in crashes on those roads.

However, CART always yields binary trees, which sometimes cannot be summarized as efficiently for interpretation and/or presentation (Breiman et al., 1984). In the case of road accidents, they may not be very practical when it comes to analysing the impact of a specific category of variable in crash severity. Liu (2009) mentions the existence of other popular algorithms for building DTs, such as C4.5. He does not apply it, however, because a binary DT is sufficient to develop his work.

Other simple algorithms, such as ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993), have been widely used in the literature of DM for building DTs, and do not involve the binary restriction. Therefore, this study proposes to make a comparison between the various methods, and to use CART and other methods to identify the main factors that affect crash severity. Then we extract certain decision or association rules (Agrawal et al., 1993) from the methods that give the best results. We show that the methods used complement our objective. Finally, our results could be used for the predictive purposes pursued by road safety analysts.

This paper is organized as follows: Section 2 gives an introduction to procedures for building DTs, focusing on the ones used in this study. It also describes the parameters used to assess the various methods, the procedure for extracting rules and the main features of the study data. Section 3 presents the results and a discussion on them. Finally, the last section presents the conclusions.

2. Materials and methods

2.1. 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[1] 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

[1] Also called *feature* or *predictor variable*

(cases whose classes are not known a priori), to detect patterns, or simply to gain a better understanding of the phenomenon being analysed.

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 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. The most commonly applied splitting criteria in simple algorithms are the Gini Index (which measures the degree of purity), used in the CART system (Breiman et al., 1984); Information Gain, used in the ID3 algorithm (Quinlan, 1986); and the Information Gain Ratio, used in the C4.5 algorithm (Quinlan, 1993). ID3 and C4.5 are based on the entropy, which measures the degree of confusion (the greater the confusion, less information). The procedures also differ in the strategies they use after building a tree, in the process known as pruning. This is when the model obtained is simplified and adjusted more closely to the data set used to build it.

2.2. Methods for building Decision Trees

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 a variable C, it is defined as:

$$gini(C) = 1 - \sum_j p^2(C = c_j)\,. \tag{1}$$

In this way, we can define the split criterion based on the Gini Index as:

$$\mathrm{GI}x(C,X) = gini(C|X) - gini(X)\,, \tag{2}$$

where $gini\,(C|X) = \sum_t p(x_t) gini\,(C|X = x_t)$ and X another known variable.

Thus, the best split is the one that minimizes $\mathrm{GI}x(C,X)$. 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. A more detailed description of the CART method can be found in Breiman et al. (1984).

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:

$$Information\,Gain\,(C,X) = IG(C,X) = H(C) - H(C|X) \tag{3}$$

Where $H(C)$ is the entropy of C, $H(C) = -\sum_j p(c_j) log p(c_j)$, with $p(c_j) = p(C = c_j)$, the probability of each value of the variable class estimated in the training data set. In the same way, $H(C|X) = -\sum_t \sum_j p(c_j|x_t) log p(c_j|x_t)$, where x_t, t=1,…,|X|, is each possible state of X and c_j, j=1,…,k each possible state of C.

Notice that the Information Gain criterion has implicit preference for splitting nominal attributes with lots of values. Therefore, it produces trees that discard the remaining attributes prematurely because they soon come to branches that have only a few cases. A more detailed description of the ID3 algorithm can be found in Quinlan (1986).

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:

$$IGR(C, X) = \frac{IG(C,X)}{H(X)} \tag{4}$$

2.3. Method assessment

Taking into consideration the indicators used to evaluate the goodness of a classification method in de Oña et al. (2011) and Mujalli and de Oña (2011), and that the variable class used shows 2 possible response categories (state A and state B), the parameters that can be defined are described below:

- *Accuracy* - The method's precision, defined as the percentage of cases correctly classified by the classifier.
- *Sensitivity* - The proportion of cases correctly classified as state A among all the observed as state A.
- *Specificity* - The proportion of cases correctly classified as state B among all the observed as state B.
- *Receiver Operating Characteristic Curve (ROC) Area* – This indicator represents the curve of positive cases correctly classified (sensitivity), as opposed to the cases of false positives (1-specificity), in such a way that a value 1 describes a perfect adjustment.

If the variable class is accident severity and its potential states are accidents with slightly injured -SI- (state A) and accidents with killed or seriously injured -KSI- (state B), the equations that define these indicators are:

$$Accuracy = \frac{TSI+TKSI}{TSI+TKSI+FSI+FKSI} 100\% \tag{5}$$

$$Sensitivity = \frac{TSI}{TSI+FKSI} 100\% \tag{6}$$

$$Specificity = \frac{TKSI}{TKSI+FSI} 100\% \tag{7}$$

Where, TSI - Number of cases of SI; $TKSI$- Number of cases of KSI; FSI- Number of false cases of SI (i.e. incorrectly classified as SI); $FKSI$- Number of false cases of KSI (i.e. incorrectly classified as KSI).

The software used to build the DTs was Weka (Witten and Frank, 2005), which is an open source freeware, available at: http://www.cs.waikato.ac.nz/ml/weka/

Moreover, in order to obtain a more reliable result for each method (CART, ID3 and C4.5) in classification, a repeated Cross Validation procedure (CV) was used. In our case, we use a 10x10-fold CV. In general, a k-fold CV uses the whole data set, and randomly divides the sample used in the training phase into k sets: Sequentially, each subset is kept to be used as a testing set against the tree model generated by the remaining k-1 subsets. Thus, different k models are obtained, in which the accuracy of the classifications in the training set (k-1) and in the testing subsets (k) can be evaluated and the optimal tree can be selected.

Finally, a corrected paired t-test implemented in Weka, which is a corrected version of the standard paired t-test, was used to compare the results of the trees generated with the different algorithms. This test checks whether a method is better or worse than another, on average, in all the training and testing data sets based on an initial data set. In our case, we used the classification results from the 100 test set for this test, i.e. the sets obtained from a 10x10-fold CV procedure. The level of significance used for this paired t-test was 0.1.

It should be pointed out that the ID3 algorithm implemented in Weka allows instances without classification. To compare results, we implemented a similar procedure but classified all the instances of the test set as in Abellán and Masegosa (2010). In the case of no classification, we took into account the decision in the parent node. For sake of simplicity, we call this procedure ID3 too.

2.4. Rules extraction and validation

The DT's structure was transformed into rules in order to extract its potentially useful information. The rules make a logic conditional structure of the type "X→Y", where in our case, X is a set of statuses of several attribute variables; and Y is only one state of the class variable:
IF (a set of statuses of several attribute variables) - THEN (status of the class variable).

For example:
IF (**accident type**=*rollover* & **atmospheric condition**=*light rain*) THEN (**severity**=*slightly injured accident*).

The part X of the rule is called the antecedent and the part Y is called the consequent.

In a DT, rules are configured from the root node, which is where the conditioned structure (IF) begins. Each variable that intervenes in tree division makes an IF of the rule, which ends in child nodes with a value of THEN, which is associated with the state resulting from the child node. The resulting state is the status of the class variable that shows the highest number of cases in the child node analysed.

A priori, as same number of rules can be identified as the number of terminal nodes on the tree. However, 3 parameters were used on each possible rule "X→Y", in order to extract significant rules that could provide useful information for the implementation of road safety strategies in the future.

It is known as support of X, as the percentage of the data set where X appears. In the same vein, we can talk about the support of the entire rule, as the percentage of the data set where X & Y appear. For each rule, the 3 parameters that we use are the following: *Support* (S), which will be the support of the entire rule; *Population* (Po), which is the support of the antecedent of the rule; and *Probability* (P), which is the percentage of cases in which the rule is accurate (i.e. P=S/Po expressed as percentage).

The concepts of support (S) and probability (P) are central to association rules and have been used by several authors (Agrawal et al., 1993; Pande and Abdel-Aty, 2009; Montella et al. 2011). Population (Po) is deduced from S and P (Po=S/P). Support is a measure of how frequently any given combination of antecedent and consequent occurs in a database. Probability[2] is defined by the percentage of cases in which a consequent appears, given that the antecedent has occurred. It essentially measures the strength of an association rule. For further clarification of these parameters see Pande and Abdel-Aty (2009).

Association rule discovery is the process of finding strong associations with a minimum support and probability. It is desirable for the rules to have a large probability factor and a high level of support. However, since some events of interest in traffic safety analysis are very rare (e.g., "crashes with fatal injury"), the support for some rules of interest could be quite low. The threshold values for these parameters depend on the nature of the data (balanced or unbalanced), significant interest in fatal crashes (rare events) and sample size (small or large databases). Pande and Abdel-Aty (2009) set 0.90% and 10% as threshold values for support and probability respectively. It means that no rules with support <0.90% and/or probability <10% would be considered. Montella et al. (2011) used lower thresholds for their analysis (0.10% and 1.00% for support and probability respectively). In this paper, as the sample size is not very large and the sample is balanced, the threshold values used are 0.60% for support and 60% for probability. With these thresholds the minimum population (Po) will be 1%. It is worth highlighting that if other lower threshold values were established, more rules could be obtained.

In order to test that spurious rules, and due to the large number of patterns considered, DTs could suffer from an extreme risk of type-1 error, that is, of finding patterns that appear due to chance alone to satisfy constraints on the sample data (Webb, 2007). To reduce this error and following other authors (Montella et al., 2011; Kashani and Mohaymany, 2011), the dataset was split randomly in two parts: a training set (70% of the data) and a testing set (remaining 30%).

The training set was used to build a DT and obtain the significant rules that satisfied the three parameters defined (S, Po and P). Next, the rules were validated in the testing set to prevent spurious rules (checking that they still met minimum values S, Po and P).

We also used a binomial test to check if the rule support measure deviates significantly (at 0.05 level) from the theoretically expected value (values from the training set) when the antecedent and the consequent items are independent.

2.5. Importance of the variables

The importance of the variables that intervene in the model is defined for a variable X with possible states $\{x_1...x\}$ by the following equation:

$$VIM\ X = \sum_{i=1}^{h} \frac{nx_i}{n} (I(C/X = x_i) - I(C)) \quad (8)$$

Where C is class variable (severity), n_{xi} the number of cases that $X=x_i$, n the number of total cases. I is Gini Index in CART, Information Gain in ID3 and Information Gain Ratio in C4.5

2.6. Description of the data

Accident data were obtained from the Spanish General Traffic Accident Directorate (DGT) for two-lane rural highways in the province of Granada (South of Spain) over a period of 7 years (2003-2009). In this study, rural highways with only two lanes (one for each direction) were used. The horizontal curves radius of these roads ranged from 16 m to 2,824 m. And the AADT

[2] Pande and Abdel-Aty (2009) and Montella et al. (2011) call this parameter confidence.

ranged from 210 to 8,681 veh/day. The accidents analysed involved 1 vehicle and they did not occur on intersections. The total number of 1,801 accidents met these conditions

In the period of study, the severity distribution for two-lane rural highways was: 6.1% fatal, 35.6% severe injury and 58.3% slight injury. For the same period, the severity distribution of all accidents (including accidents in freeways, multilane highways, two-lane highways, intersections, etc.) in the province of Granada was: 8.3% fatal, 40.1% severe injury and 51.6% slight injury. This study uses the DGT definition for injuries: severe injury is any person injured in a traffic accident and whose condition requires hospitalization for more than twenty-four hours; slight injury is any person that does not meet the severe injury definition; and fatal injury is any person that dies on the spot or within the subsequent 30 days as a result of a traffic accident.

Following previous studies (Chang and Wang, 2006; de Oña et al, 2011; Kashani and Mohaymany, 2011), severity of accident was defined according to the worst injured occupant, and two level of severity were identified: accident with slightly injured (SI) and accidents with killed or seriously injured (KSI).

To identify the main factors that affect accident severity, 19 variables were analysed (see Table 1). The variables chosen were based on:

- Variables available in the original dataset (from DGT).
- Variables selected in others studies with similar objectives (Chang and Wang, 2006; de Oña et al., 2011; Kashani and Mohaymany, 2011; Pakgohar et al., 2010).

The variables describe characteristics related to the driver (age and gender); accident (month, time, day, number of injuries, occupants involved, accident type and cause); road (safety barriers, pavement width, lane width, shoulder type, paved shoulder, pavements markings and sight distance[3]); vehicle (vehicle type); and context (atmospherics factors and lighting). Some variables were re-coded in a reduced number of categories to be able to work with them. For instance, in the original dataset MON had 12 categories (12 months), and it was recoded into four periods (see Table 1). Other variables, such as CAU, DAY, LAW, LIG, PAS, PAW, ROM, SEX, SHT, SID, were used as they were in the original dataset. Table 1 gives a description of the variables used for the analysis, together with the frequency distribution.

(insert Table 1 here)

3. 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 10x10-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, specificity and ROC_Area* are shown in Table 2.

(insert Table 2 here)

C4.5 and CART show similar values for accuracy. ID3 shows significantly worse values 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: Abdel Wahab and Abdel-Aty (2001) obtained 61% accuracy when they applied Bayesian networks and 58.1%

[3] The sight distance refers only to the horizontal visibility limitation at the site of the accident (i.e. the 'without restrictions' category means that there were no visibility limitations at the point of the accident; the 'building' category means that the visibility limitation at the point of the accident was a building; the same applies for topography, vegetation, atmospheric factors and others.

accuracy on neural networks. De Oña et al. (2011) obtained 58%, 59% and 61% accuracy applying Bayesian networks with different algorithms (AIC, MDL and BDeu, respectively).

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 CART 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 CART method 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 seconds to build a DT with 19 variables and 1,801 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 analyze the decision rules obtained with the algorithms that attained the best results: C4.5 and CART.

3.1. CART

Figure 1 shows the DT built using the CART method with 70% of the data for training and the remaining data (30%) for testing, as used by Montella et al. (2011). The CART method creates a tree with 19 nodes and 10 terminal nodes.

(insert Figure 1 here)

Table 3 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 lift (see Montella et al., 2011) 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.

(insert Table 3 here)

The root variable that generates the tree is SEX (see Figure 1) 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 Figure 1): 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 analyzed the drivers were men (see Table 1). After this node, the tree splits according to ACT. The accident type has been identified in several previous studies (Al-Ghamdi, 2002; de Oña et al., 2011; Kashani and Mohyamany, 2011) 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 3). 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 Figure 1). 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 (Figure 1) splits by the variable ATF: if ATF is light rain the accident is SI, with a probability of 63% (rule 8 in Table 3). 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 Figure 1). 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: from [6-12) h, accidents with SI are obtained when PAS is paved or non-existent (rule 15, which is the one that represents the highest probability: 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 Figure 1), 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%.

(insert Table 4 here)

LIG is the most important variable, coinciding with previous studies. Gray et al. (2008) identified that more severe injuries are predicted during darkness. Abel-Aty (2003) and Helai et al. (2008) found the same results. Pande and Abel-Aty (2009) concluded that there is a significant correlation between lack of illumination and high severity of crashes. De Oña et al. (2011) also pointed that KSI accidents are associated with roadways without lighting.

ATF is the second variable with 83.6% importance in the model. This result matches with other previous studies, such as Xie et al. (2009) and Mujalli and de Oña (2011). 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. Next comes ACT with 76.0%. Kcoleman and Kweon, (2002) and de Oña et al. (2011) also found this variable as one of the most important in the study of severity. SEX represented 72% of the variables' importance. The other variables (see Table 4) in the model are less important, with percentages between 55.9% and 9.9%.

3.2. C4.5

Figure 2 represents a DT built using the C4.5 algorithm based on the training set. It shows 52 nodes, with 39 terminal nodes. The increase in the number of nodes is justified be the fact that this algorithm creates a branch for each category of variable used in the analysis. In this case, however, only 9 rules that meet the minimal values for S, Po and P were obtained (see Table 5).

(insert Figure 2 here)

Since the tree generated with C4.5 is larger, only the rules extracted in Table 5 are used to describe the following tree structure. In this case, the rules in Table 5 also verify the threshold values for S, Po and P in both training and testing sets. For C4.5, the binomial test showed that all the rules obtained from the training set with the minimum threshold have a grade of lift (see Montella et al., 2011) different than 1. And none of the rules given in Table 5 have significant differences (at 0.05 level) on support when they are applied to the test set.

(insert Table 5 here)

As in CART, the root variable is the variable SEX. For female drivers when LIG is daylight, the rule with the highest population (9.9%) and support (6.8%) gives a severity result of SI (rule 8 in Table 5). This result agrees with the previous CART´s results: female drivers seem to be highly affected by lighting conditions.

Most of the tree is generated by male drivers (see Figure 2) and according to ACT, the same as CART. Figure 2 and Table 5 show the following patterns: if ACT is rollover the severity is SI, with a probability of 61% (rule 12); whereas, if ACT is collision with pedestrian, it depends on PAS. This result is very important because if PAS is paved the severity is KSI and we obtained the rule with the highest probability (78%) (rule 16). Thus, from the perspective of road safety, precautions against accidents could be taken by placing safety barriers on stretches of road where pedestrians walk on the shoulder (roads that link two towns that are close to each other).

The rest of the rules are obtained for run off road accidents (ROR represents of 82.9% of the accident analyzed) and depending on CAU. When CAU is a combination of factors, SID is without restriction and MON is spring the severity of accident is SI with almost 74% of the probability (rule 33). For CAU attributable to driver and depending on VEH the following patterns are shown: when VEH is a truck the accident is KSI (67.5%), rule 29; when is a motorbike or motorcycle and PAS is non-existent or impassable, the same severity (KSI) is obtained (rule 40); and for car two more rules are obtained depending on PAW. This result indicates the need to raise male drivers' awareness of vehicles of this type.

When PAW is between [6-7] meters and driver´s age is 28-60 (rule 46) the severity is SI with a probability of almost 69%. When PAW is > 7 meters, the tree splits according to NOI, and when it is higher than 1, accidents are SI 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. That is why Pande and Abdel-Aty (2009) restricted the number of variables in the antecedent to three.

Following Eq. 8, it is possible to obtain the importance of the variables in the C4.5 model (see Table 6).

(insert Table 6 here)

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. These results are in accordance with Al-Ghamdi (2002) and Kashani and Mohyamany (2011), who situate crash cause among the top variables influencing severity. The CART algorithm identified eleven of the previous fourteen variables. Moreover C4.5 identified VEH, PAW, and NOI.

4. Conclusions

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.

The main conclusions regarding the methods used in this paper to build DTs are the following:

- CART builds binary DTs and therefore certain categories of splitting variables are grouped in some branches, increasing node support, but making it impossible to analyze the influence of a specific category on severity. C4.5 creates a branch for each category, thereby permitting an analysis of the influence of all the categories of variables used to build the DT. Consequently, it could be said that the rules obtained with CART are less informative.
- C4.5 generates DTs with more branches than CART, and therefore it produces more rules. However, not all the rules meet the established minimal number of support, population and probability parameters, and therefore the rules may not be very useful for implementing future road safety strategies.
- The importance of the variables in the model can be obtained using either algorithm.
- The two algorithms have certain similarities with regards to the structure of the tree generated. For example, the root variable for both is SEX and tree density is obtained by the branch male drivers, and the value that continues to split the tree is ACT.

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. The approach proposed in this paper within each group will enable actions to be prioritized on the basis of support, population and probability. It is worth highlighting certain overall conclusions from a road safety perspective.

The rules drawn from the two methods are coincidental in that:

- Male drivers are the main causes of KSI crashes.
- The probability of KSI increases if pedestrians are involved (node 3 Figure 1 and rule 16 in Table 5).
- When women drivers are involved in an accident, both methods predict SI when lighting exists (full daylight, sufficient lighting and dusk) (rule 6 in Table 3; rule 8 in Table 5, and nodes 4 and 5 in Figure 2). However, both methods predict KSI when the lighting is non-existent or insufficient (rule 5 in Table 3 and nodes 6 and 7 in Figure 2). These rules are not observed for men and may indicate that women increase their risk of severity under conditions of less lighting on the road.

From a road safety point of view, most of the rules extracted coincide with the conventional problems found on rural highways in developed countries, as most previous studies point out. This validates the method proposed in this paper, and therefore it is positive. However, the primary importance of this proposal is that other data bases not used here (i.e. other infrastructure, roads and countries) could be used to identify unconventional problems in a manner easy for road safety managers to understand, as decision rules.

However, using these two types of DTs permitted the identification of a specific problem worthy of further study: Although less women than men are involved in accidents (15.3% vs. 84.5%, see Table 1), and accident severity is SI in 62.2% of cases, the two methods indicate that

women increase their risk of severity under conditions of non-existent or insufficient lighting. The efforts of multidisciplinary teams with experts on psychology, physiology, road safety and illumination should focus on a search of the reason why women, contrary to men, present higher risk of severity under conditions of less lighting on the road.

Finally, it should be stressed that each method has advantages and drawbacks, and reveals different information. Therefore, the two methods complement each other and the recommendation is to use both of them for a full analysis.

5. Future work

When we use a DT to obtain decision rules, such rules are highly dependent on the variable entered in the root node, which permits knowledge to be extracted only in the sense dictated by said root variable. For future research, it is worth studying the possibility of generating DTs by varying the root node and analysing all the rules that may be obtained from a single set of data.

For the same purpose, we would like to apply new split criteria based on new mathematical models for representing information, as well as the new procedures used in classification to date. These criteria and procedures can be seen in Abellán and Masegosa (2010) and Abellán et al. (2011).

Acknowledgements
The authors express their gratitude to the Spanish General Directorate of Traffic (DGT) for supporting this research and offering all the resources that are available to them. Griselda López wishes to express her acknowledgement to the regional ministry of Economy, Innovation and Science of the regional government of Andalusia (Spain) for their scholarship to train teachers and researchers in Deficit Areas, which has made this work possible.

The authors appreciate the reviewer´s comments and effort in order to improve the paper.

REFERENCES

Abdel Wahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6–13.
Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34, 597–603.
Abellán, J., Baker, R.M., Coolen, F.P.A., 2011. Maximising entropy on the nonparametric predictive inference model for multinomial data. European Journal of Operational Research 212(1), 112-122.
Abellán, J., Masegosa, A., 2010. An ensemble method using credal decision trees. European Journal of Operational Research, 205(1), 218-226.
Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD 1993), 207–216.
Al-Ghamdi, A., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34 (6), 729–741.
Breiman, L., Friedman, J.,Olshen, R., and Stone, C., 1984. Classification and Regression Trees. Belmont, CA: Chapman & Hall.
Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.
De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention 43, 402-411.
Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, 1–34.

Gray, R.C., Quddus, M.A., Evans, A., 2008. Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research 39, 483-495.
Helai, H., Chor, C.H., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40, 45–54.
Kashani, A., Mohaymany, A., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science 49, 1314-1320.
Kashani, A., Mohaymany, A., Ranjbari, A., 2011. A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. Promet-Traffic & Transportation, 23 (1), 11-17.
Kuhnert, P.M., Do, K.A., McClure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis, Vol 34(3), 371-386.
Liu, P., 2009. A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes. Neurocomputing 72, 2902-2908.
Montella, A., Aria, M., D´Ambrosio, A.,Mauriello, F., 2011. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, in press.
Mujalli, R.O., de Oña, J. (in press). Injury Severity Models for Motorized Vehicle Accidents: A review, Proceedings of the Institution of Civil Engineering - Transport. In Press.
Mujalli, R.O., de Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research, 42, 317–326
Pakgohar, A., Tabrizi, R.S., Khalilli, M., Esmaeili, A., 2010. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science 3, 764-769.
Pande, A., Abdel-Aty, M., 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. Safety Science 47, 145–154.
Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California.
Quinlan, J.R., 1986. Induction of decision trees. Machine Learning, 1(1), 81-106.
Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention. In press
Shannon, C., Weaver, W., 1964. The Mathematical Theory of Communication. The University of Illinois.
Sohn, S.Y., Shin, H.W., 2001. Data mining for road traffic accident type classification. Ergonomics 44, 107–117.
WHO, World Health Organisation, 2009. Informe Global sobre el estado de la Seguridad Vial: Tiempo para la Acción. Available at:
www.who.int/violence_injury_prevention/road_safety_status/2009
Webb, G.I., 2007. Discovering significant patterns. Machine Learning 68, 1–33.
Xie, Y., Zhang, Y., Liang, F., 2009. Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. Journal of Transportation Engineering ASCE, 135(1), 18–25.

List of figures:

List of tables:

Figure(s)

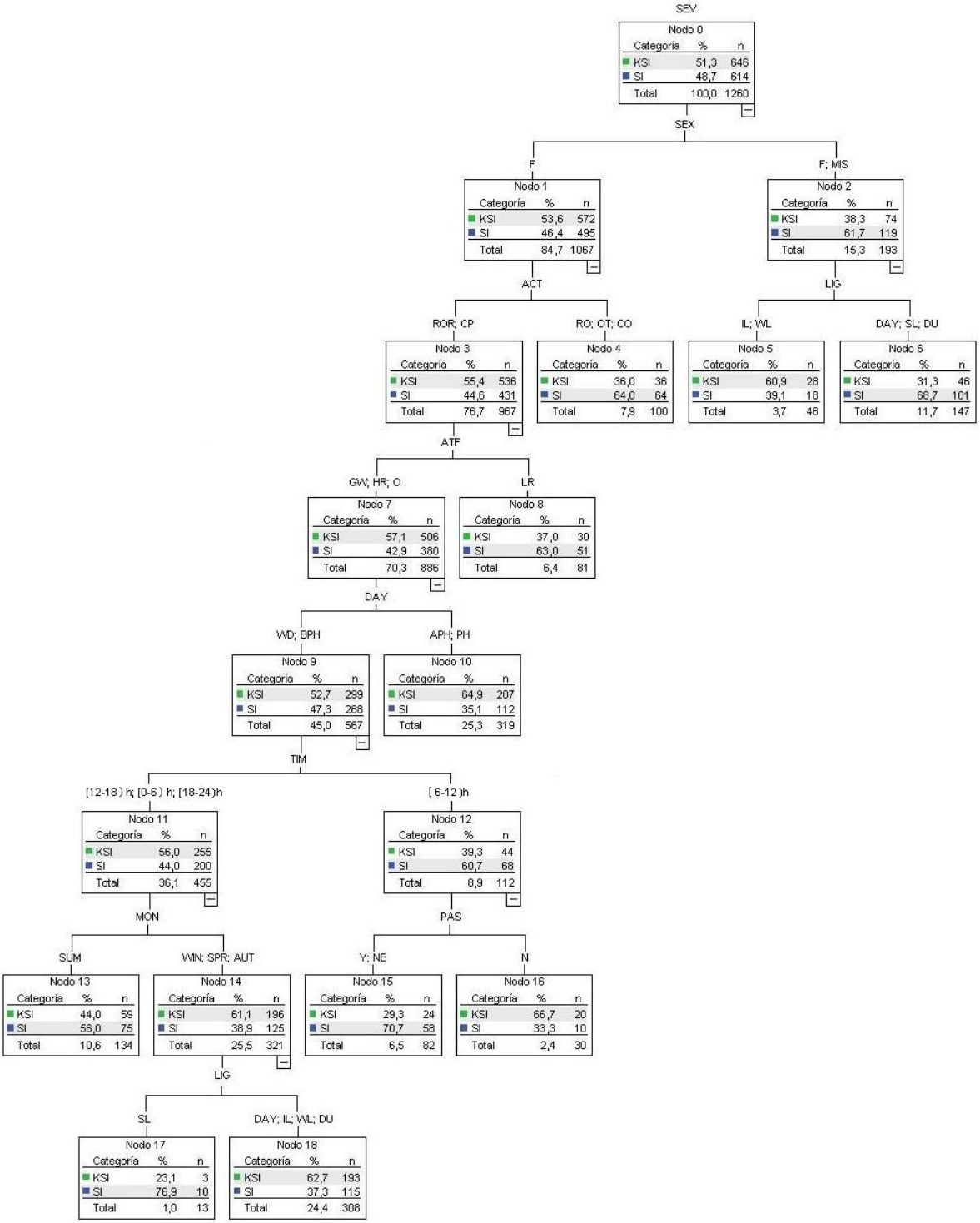

Figure 1. Decision tree built with CART.

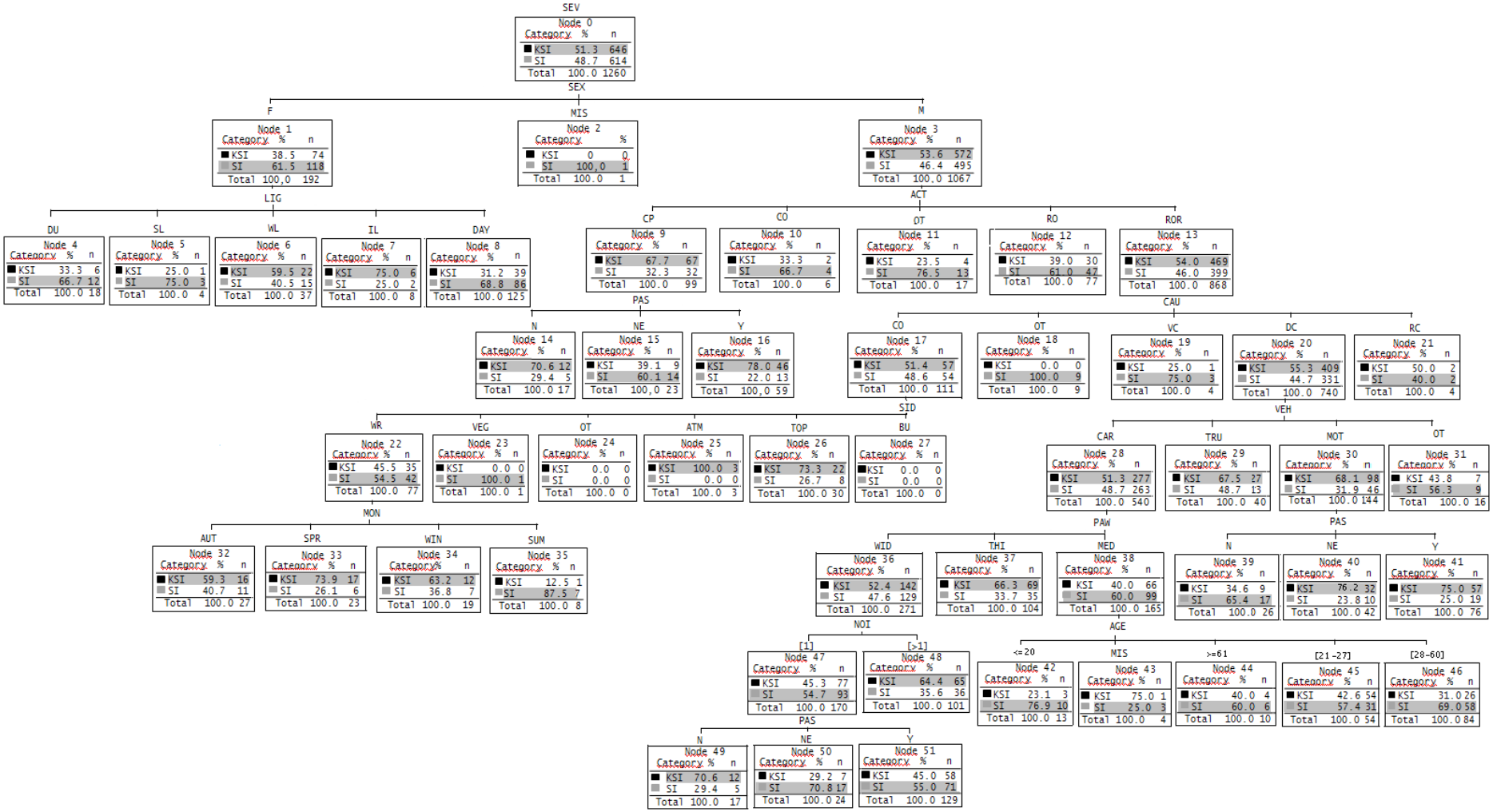

Figure 2. Decision tree built with C4.5.

Table(s)

NUM	VARIABLES			%TOTAL	SEVERITY	
	DESCRIPTION	CODE	VALUES		%SI	%KSI
1	ACT: Accident type	CO	Fixed objects collision	0.90	76.47	23.53
		CP	Collision with pedestrian	7.70	33.33	66.67
		OT	Other (collision with animals, etc.)	1.90	68.57	31.43
		RO	Rollover (in carriage without any collision)	6.60	61.86	38.14
		ROR	Run off road (with or without collision)	82.90	51.77	48.23
2	AGE: Age	≤ 20	≤ 20	12.22	52.73	47.27
		[21-27]	[21-27]	25.65	50.00	50.00
		[28-60]	[28-60]	53.64	51.76	48.24
		≥ 61	≥ 61	6.89	59.68	40.32
		UN	Unknown	1.61	27.59	72.41
3	ATF: Atmospheric factors	GW	Good weather	86.40	50.58	49.42
		HR	Heavy rain	2.10	63.16	36.84
		LR	Light rain	8.90	58.75	41.25
		O	Other	2.60	51.06	48.94
4	BAR: Safety barriers	N	No	96.90	48.30	54.70
		Y	Yes	3.10	53.60	46.40
5	CAU: Cause	DC	Driver characteristics	82.70	48.99	51.01
		CO	Combination of factors	13.40	61.16	38.84
		OT	Other	1.20	72.73	27.27
		RC	Road characteristics	1.40	84.00	16.00
		VC	Vehicle characteristics	1.20	63.64	36.36
6	DAY: Day	APH	Working day after the weekend or public holiday (Monday or day after public holiday)	8.40	57.62	42.38
		BPH	Working day before the weekend or public holiday (Friday or day before public holiday)	15.90	52.26	47.74
		PH	On a weekend (Saturday or Sunday) or public holiday	30.60	50.36	49.64
		WD	Regular working day (Tuesday, Wednesday or Thursday nor before neither after public holiday)	45.00	51.05	48.95
7	LAW: Lane width	THI	< 3,25 m	27.50	46.87	53.13
		MED	[3,25-3,75] m	70.20	53.20	46.80
		WID	> 3,75 m	2.30	58.54	41.46
8	LIG: Lighting	DAY	Daylight	53.10	55.49	44.51
		DU	Dusk	5.80	54.29	45.71
		IL	Insufficient (night-time)	7.30	51.15	48.85
		SL	Sufficient (night-time)	40.00	59.72	48.28
		WL	Without lighting (night-time)	29.80	43.10	56.90
9	MON: Month	AUT	Autumn	23.50	53.07	46.93
		SPR	Spring	25.20	53.64	46.36
		SUM	Summer	27.30	51.63	48.37
		WIN	Winter	24.00	47.92	52.08
10	NOI: Number of injuries	[1]	1 injury	69.60	53.43	46.57
		[>1]	> 1 injury	30.40	47.35	52.65
11	OI: Occupants involved	[1]	1 occupant	64.70	51.20	48.80
		[2]	2 occupants	22.50	51.48	48.52
		[>2]	> 2 occupants	12.70	53.71	46.29
12	PAS: Paved shoulder	N	No	17.10	49.35	50.65
		NE	Non existent or impassable	31.30	50.89	49.11
		Y	Yes	51.60	52.74	47.26
13	PAW: Pavement width	MED	[6-7] m	30.50	53.19	46.81
		THI	< 6 m	14.40	45.56	54.44
		WID	> 7 m	55.10	52.27	47.73
14	ROM: Pavement markings	DME	Does not exist or was deleted	9.40	52.35	47.65
		DMR	Separate margins of roadway	9.90	48.31	51.69
		SLD	Separate lanes and define road margins	75.80	52.23	47.77
		SLO	Separate lanes only	4.90	46.59	53.41
15	SEX: Gender	F	Female	15.30	62.18	37.82
		M	Male	84.50	49.61	50.39
		UN	Unknown	0.20	75.00	25.00
16	SHT: Shoulder type	THI	< 1,5 m	40.40	52.54	47.46
		MED	[1,5-2,5] m	10.50	50.28	49.72
		NE	Non existent or impassable	49.10	50.57	49.43
17	SID: Sight distance	ATM	Atmospheric	2.20	67.50	32.50
		BU	Building	0.60	36.36	63.64
		OT	Other	0.70	50.00	50.00
		TOP	Topography	22.70	49.39	50.61
		VEG	Vegetation	0.70	50.00	50.00
		WR	Without restriction	73.10	51.94	48.06
18	TIM: Time	[0-6)	[00:00-05:59]	20.00	48.06	51.94
		[6-12)	[06:00-11:59]	21.00	58.73	41.27
		[12-18)	[12:00-17:59]	32.10	52.77	47.23
		[18-24)	[18:00-23:59]	26.90	47.22	52.78
19	VEH: Vehicle type	CAR	Cars	70.90	47.10	52.90
		TRU	Trucks	4.90	53.80	46.20
		MOT	Mortorbikes and motorcycles	21.70	35.60	64.40
		OT	Other	2.50	50.60	49.40

Table 1. Variables used from the police accident reports.

	CART	C4.5	ID3
accuracy	55.87	54.16	52.72*
sensitivity	54.00	55.00	53.00
specificity	58.00	54.00	52.00
ROC Area	57.00	54.00	53.00*

** The results worsen significantly.*

Table 2. Comparison of the parameters produced by the various algorithms.

NODE /RULE	RULES CART: IF…	THEN	S (%)	Po (%)	P (%)
16	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)	KSI	1.59	2.38	66.67
5	IF (SEX≠M) AND (LIG= IL OR LIG=WL)	KSI	2.22	3.65	60.87
15	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)	SI	4.60	6.51	70.73
6	IF (SEX≠M) AND (LIG≠ IL OR LIG≠WL)	SI	8.02	11.67	68.71
4	IF (SEX=M) AND (ACT=RO OR ACT=CO OR ACT=OT)	SI	5.08	7.94	64.00
8	IF (SEX=M) AND (ACT=ROR OR ACT=CP) AND (ATF=LR)	SI	4.05	6.43	62.96

Table 3. Description of the rules according to the CART.

VARIABLES	IMPORTANCE NORMALIZED
LIG	100%
ATF	83.6%
TIM	77.1%
ACT	76.0%
SEX	72.0%
PAS	55.9%
DAY	54.9%
MON	49.9%
CAU	32.8%
AGE	30.6%
SID	28.4%
LAW	9.9%

Table 4. Importance of the variables with CART.

NODE /RULE	RULES C4.5: IF…	THEN	S(%)	Po(%)	P(%)
16	IF (SEX=M) AND (ACT=CP) AND (PAS=Y)	KSI	3.65	4.68	77.97
40	IF (SEX=M) AND (ACT=ROR) AND (CAU=DC) AND (VEH=MOT) AND (PAS=NE)	KSI	2.54	3.33	76.19
29	IF (SEX=M) AND (ACT=ROR) AND (CAU=DC) AND (VEH=TRU)	KSI	2.14	3.17	67.50
48	IF (SEX=M) AND (ACT=ROR) AND (CAU=DC) AND (VEH=CAR) AND (PAW=WID) AND (NOI=[>1])	KSI	5.16	8.02	64.36
33	IF (SEX=M) AND (ACT=ROR) AND (CAU=CO) AND (SID=WR) AND (MON=SPR)	SI	1.35	1.83	73.91
50	IF (SEX=M) AND (ACT=ROR) AND (CAU=DC) AND (VEH=CAR) AND (PAW=WID) AND (NOI=[1]) AND (PAS=NE)	SI	1.35	1.90	70.83
46	IF (SEX=M) AND (ACT=ROR) AND (CAU=DC) AND (VEH=CAR) AND (PAW=MED) AND (AGE=[28-60])	SI	4.60	6.67	69.05
8	IF (SEX=F) AND (LIG=DAY)	SI	6.83	9.92	68.80
12	IF (SEX=M) AND (ACT=RO)	SI	3.73	6.11	61.04

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

VARIABLES	IMPORTANCE NORMALIZED
ACT	100.0%
CAU	80.4%
SEX	69.1%
LIG	67.5%
VEH	65.7%
ATF	59.8%
PAW	42.8%
AGE	41.2%
TIM	39.7%
SID	36.3%
NOI	32.1%
DAY	25.7%
LAW	20.2%
MON	11.2%

Table 6. Importance of the variables with C4.5.