Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia

Social and Behavioral Sciences

Procedia - Social and Behavioral Sciences 53 (2012) 106 – 114

SIIV - 5th International Congress - Sustainability of Road Infrastructures

Using Decision Trees to extract Decision Rules from Police Reports on Road Accidents

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

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

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

[c] Ph. D. Department of Computer Science & Artificial Intelligence, University of Granada, ETSI Informática, c/Periodista Daniel Saucedo Aranda s/n, 18071, Granada, Spain

Abstract

The World Health Organization (WHO) considers that traffic accidents are major public health problem worldwide, for this reason safety managers try to identify the main factors affecting the severity as consequence of road accidents. In order to identify these factors, in this paper, Data Mining (DM) techniques such as Decision Trees (DTs), have been used. A dataset of traffic accidents on rural roads in the province of Granada (Spain) have been analyzed.

DTs allow certain decision rules to be extracted. These rules could be used in future road safety campaigns and would enable managers to implement certain priority actions.

Keywords: Type your keywords here, separated by semicolons ;

1. Introduction

The World Health Organization (WHO) considers that traffic accidents are major public health problem worldwide that every year claiming 1.27 million annual deaths and between 20 and 50 million injuries [1]. For this reason, many safety researchers have attempted to identify affecting the severity as consequence of road accidents. Different techniques, such as: Regression-type generalized linear models, Logit/Probit models, ordered Logit/Probit models have been used to achieve these objectives [2, 3, 4]. However most of these models have their own model assumptions and pre-defined underlying relationships between dependent and independent

* Corresponding author. Tel.: +34-958-249-450; fax: +34-958-246-138.
E-mail address: griselda@ugr.es

doi:10.1016/j.sbspro.2012.09.864

variables [5]; if these assumptions are violated, the model could lead to erroneous estimations, for example of the likelihood of severity accident.

To solve these limitations other method such as, Classification and Regression Trees (CART), have been used in the field of Road Safety. Kuhnert et al. [6] compared the results obtained with logistic regression, CART, multivariate adaptive regression splines (MARS) in the analysis of study of injuries resulting from motor vehicle accidents. The findings indicated that non-parametric techniques (CART and MARS) could provide more informative and attractive models whose individual components can be displayed graphically. Chang and Wang [5] employed CART to study 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. They obtained that vehicle type was the most important affecting the severity of the accident. Recently, Pakgohar et al. [7] 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 are also simpler and easier to interpret. Kashani et al. [8] studied the most important factors that affect the injury severity of drivers involved in crashes on two-lane two-way rural roads. Subsequently, Kashani and Mohaymany [2] used CART to identify the main factors that affect the injury severity of vehicle occupants involved in crashes on those roads. And the results indicated that improper overtaking and not using a seatbelt was the most important factors associated with crash severity.

CART is particularly appropriate for studying traffic accident because is non-parametric techniques that do not require a priori probabilistic knowledge about the phenomena under studying and consider conditional interactions among input data [9].

Moreover, CART method allow certain decision rules of the "if-then" type to be extracted [8], and these rules can be used to discover behaviours that occur within a particular set of data. So, the aim of this work is to use CART method to identify the main factors that affect of the traffic injury severity and to extract certain decision rules which could be used in future road safety campaigns.

The paper is organized in four mayor sections. Section 2 presents an introduction to the main concepts of CART method, Decision Rules and the database used in the analysis. Section 3 presents the results and discussion. And, finally, section 4 presents the main conclusions of the study.

2. Materials and Methods

2.1. CART

A decision tree (DT) could be defined as a predictive model which can be used to represent both classifiers and regression models (depending on the nature of the variable class). When the value of the target variable is discrete, a regression trees is developed, whereas a regression trees is developed for the continuous target variable. CART method is a particularly type of DTs which allow developed either type of tree. In this wok a classification tree is developed because target variable (injurity severity) is discrete (slight injured -SI; killed or seriously injured –KSI).

A DT is a simple structure formed by number finite of “nodes” (which represent an attribute variable) connected by “branches” (which represents one of the states of the one variable) and finally, “terminal nodes or leafs” which specify the expected value of the variable class or target variable. The principle behind tree growing is to recursively partition the target variable to maximize “purity” in the child node. DTs are built recursively, following a descending strategy. The root node (which contained all of the data), is divide by two branchs (because the CART model generates binary trees) on the basis of an independent variable (splitter) that creates the best homogeneity. Each branch connected with a child node, the data in each child node are more homogenous than those in the upper parent node. Then, each child node is split recursively until all of them are pure (when all the cases are 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 variable class.

There are different splitting criteria, however in the CART system the most commonly applied splitting criteria is the Gini index (GI); it could be defined for node c, as:

$$\text{gini}(c) = 1 - \sum_j p^2(j|c) \tag{1}$$

With: $p(j|c) = {p(j,c)}/{p(c)}$, $p(j,c) = {\pi(j)N_j(c)}/{N_j}$ and $p(c) = \sum_j p(j,c)$. Where: j – number of target variable or classes; $\pi(j)$ - prior probability for class j; $p(j|c)$ - conditional probability of a case being in class j provided that is in node m, $N_j(c)$ - number of cases of class j of node m, N_j - number of cases of class j in the roof node.

GI is one measure the degree of purity of the node, so when GI is equal to cero, the node is pure (all the cases in the node have the same class). When CART is development the aim is to achieve the maximum purity in the nodes, so the best split is the one that minimizes GI. Following 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. At great length description of the CART method could be found in Breiman [10].

Following de Oña et al., [11], the goodness of a classification method is evaluated by accuracy. Accuracy is the percentage of cases correctly classified by the classifier of the method, and it is defined by following equation:

$$\text{accuracy} = \frac{TSI + TKSI}{TSI + TKSI + FSI + FKSI}.100\% \tag{2}$$

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).

On the other hand, one of the most valuable outcome provided by CART analysis is the value of the importance of independent variables that intervene in the model, which shows the impact of such predictor variables on the model.

2.2. Decision Rules

Decision Rules (DRs) could be obtained from the DT's structure. DRs are important because could be used to extract the potentially useful information from the data. The rules have the form of logic conditional: if "A" then "B", where "A" is the antecedent (a state or a set of statuses of one or several variables) and "B" is the consequent (one status of the variable class).

So, the conditioned structure (IF) of DR, begins in root node. 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 class resulting (the status of the variable class that shows the highest number of cases in the terminal node) from the child node. A priori, as same number of rules can be identified as the number of terminal nodes on the tree.

However, 2 parameters (population -Po; class probability -P) were used in order to extract important rules that could provide useful information for the implementation of road safety strategies in the future. The parameters that have been used could be defined as: population (Po), is the percentage of cases of a node in relation to the total number of cases analysed; and class probability (P), is the percentage of cases for the resulting class. The minimum values used so the selected rules will be representative are: Po≥1% and P≥60%.

2.3. Data

In this work, traffic accident data for rural highways for the province of Granada (South of Spain) have been used. These data have been obtained from Spanish General Traffic Accident Directorate (DGT). The period of the study is 5 years (2004-2008), and only data for 1 vehicle involved were used for this analysis. The total number of accident's records used is 1,801.

Considering that the main objective of this study is to identify the principal factors that affect the severity of traffic accidents, 17 explanatory variables were used based on De Oña et al. [11], and as a class variable, the injury severity level was considered with two classes (SI or KSI).

The data included variables describing the conditions that contributed to the accident and injury severity (see Table 1): characteristics of the accidents (month, time, day type, number of injuries, number of occupants, accident type and cause); weather information (atmospherics factors and lighting); driver characteristics (age and gender); and road characteristics (pavement width, lane width, shoulder width, paved shoulder, road markings and sight distance).

Table 1. Explanatory variables description

VARIABLE (CODE)	DESCRIPTION (CODE)	KSI	SI cases
Accident type (ACT)	Fixed objets collision (CO)	4	13
	Collision with pedrastian (CP)	92	46
	Other (OT)	11	24
	Rollover (RO)	45	73
	Run off road (ROR)	720	773
Age (AGE)	≤ 20	104	116
	(20-27]	231	231
	(27-60]	466	500
	>60	50	74
	Missing (MIS)	21	8
Atmospheric factors (ATF)	Good weather (GW)	769	787
	Heavy rain (HR)	66	94
	Light rain (LR)	14	24
	Other (O)	23	24
Cause (CAU)	Driver characteristics (DC)	760	730
	Combination of factors (COF)	94	148
	Other (OT)	6	16
	Road characteristics (RC)	4	21
	Vehicle characteristics (VC)	8	14
Day (DAY)	After holiday (AH)	137	150
	Before holiday (BH)	64	87
	Holiday (H)	274	278
	Working day (W)	397	414
Lane width (LAW)	< 3,25 m (THI)	263	232
	[3,25-3,75] m (MED)	592	673
	> 3, 75 m (WID)	17	24
Lighting (LIG)	Daylight (DAY)	426	531
	Dusk (DU)	48	57
	Insufficient (IL)	64	67
	Sufficient (SL)	29	43
	Without lighting (WL)	305	231
Month (MON)	Autumn (AUT)	199	225
	Spring (SPR)	210	243
	Summer (SUM)	238	254
	Winter (WIN)	225	207
Number of injuries	1 injury	584	670

(NOI)	> 1 injury	288	259
Occupants involved (OI)	1 occupant	569	597
	2 occupants	197	209
	> 2 occupants	106	123
Paved shoulder (PAS)	No (N)	156	152
	Non existent or impassable (NE)	277	287
	Yes (Y)	439	490
Pavement width (PAW)	[7-6] m (MED)	257	292
	< 6 m (THI)	141	118
	> 7 m (WID)	474	519
Road markings (ROM)	Does not exist or was deleted (DME)	81	89
	Separate margins of roadway (DMR)	92	86
	Separate lanes and define road margins	652	713
	Separate lanes only (SLO)	47	41
Gender (SEX)	Female (F)	104	171
	Male (M)	767	755
	Missing (MIS)	1	3
Shoulder type (SHT)	< 1,5 m (THI)	345	382
	[1,5-2,5] m (MED)	90	100
	Non existent or impassable (NE)	437	447
Sight distance (SID)	Atmosferic (ATM)	13	27
	Building (BU)	7	4
	Other (OT)	6	6
	Topological (TOP)	207	202
	Vegetation (VEG)	6	6
	Without restriction (WR)	633	684
Time (TIM)	[0-6] h	187	173
	(6-12] h	156	222
	(12-18] h	256	229
	(18-24] h	273	305

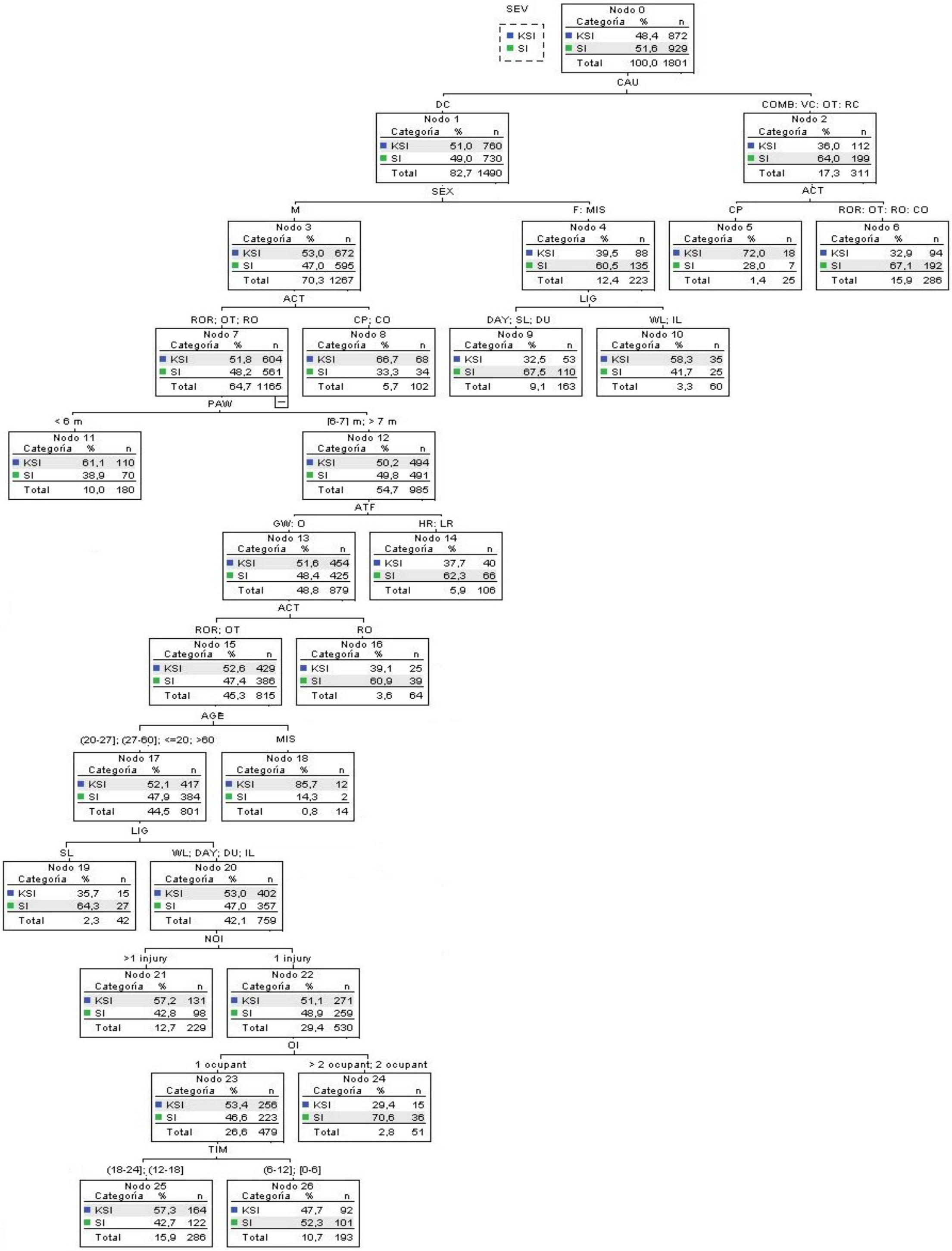
3. Results

The accuracy obtained for CART method was 54.43%. This value is within the range of values obtained in other studies in which classification methods with similar objectives [12; 11].

DT obtained contains 27 nodes (14 of them are terminal nodes). The identifier number, total number of accidents present in that node, and node classification based on the 2 categories (SI and KSI) are indicated for each node. Figure 1 shows the DT built and the interpretation is given below.

A root node is variable CAU, which is divided into two child nodes (node 1 and 2, see Figure 1). Node 2 shows accidents which not due to the driver, and depending on type, nodes 5 and 6 are obtained, with varying degrees of severity in collision with pedestrians the resulting severity is KSI with a probability of 72%, while for other accident types the severity is SI with a probability of 67.1%. Node 1 shows data related to accidents which are due to driver. This node is divided by gender variable. So, if the driver is a woman, nodes 10 and 9 are obtained, depending on road lighting: if the lighting is insufficient or without lighting, the accident is KSI with a probability of 58.33%; whereas if the road is sufficiently lit, it is broad daylight or dusk, the severity is SI in 67.48%.

However, most of the tree is generated by male driver (node 3). This node is splitted according to the ACT variable. For accidents involving pedestrians (with or without obstacles), accident severity is KSI. Whereas, for all other accident types, the tree splits by the variable lane width. In lanes narrower than 6m (narrow lanes), accidents are KSI in 61.11% of cases. In lanes wider than 6 m, severity with light or heavy rain is SI with a probability of 62.26%, depending on ATF variable.



Figue 1. Classification tree.

When atmospheric factors of good weather and others, the tree continues to grow according to ACT variable. When ACT is rollover, the accident is SI with a probability of 60.94%, and if it is run off the road or another type, the tree is divided according to the age of the driver involved in the accident. For most of the age groups (4 of the 5 analysed), the variable lighting causes the tree to grow, so accidents are SI if the road lighting is sufficient, whereas severity in the other cases is related to the number of injured.

If there is more than one injured person involved, the accident has a 57% probability of being KSI (node 21), while if there is only one injured person, the severity will also depend on the number of occupants, so that if the number is equal or more than two, the accident will be SI. If there is only one occupant, depending on the time of day, the 2 last nodes of the decision tree are obtained: from and from 12-24 h accident severity is KSI (node 25) and from 0-12 h accidents with SI (node 26).

3.1. Variable importance

The CART modelling process has an important phase in which the variables that are of key importance in the prediction of the dependent variable are identified. This is achieved by using the importance index [2].

Using this index, thirteen variables were detected as having the greatest influence on accident severity (see Table 2). Accident type is the most important variable, coinciding with previous studies [13, 11, 9]. The next important variable has been causes of the accident with a 57,6% importance, result that is coherent with other studies [14, 2], who situate crash cause among the top variables influencing severity. The variable lighting has 42.3% importance in the model. Lighting conditions were also highlighted in studies by Abel-Aty [15], Gray et al. [16], Heali et al [17], De Oña et al. [11] and Montella et al [18]. And gender variable has 34.9%. The other variables in the model are less important, with percentages of 26.9% to 4.5%.

Table 2. Importance of the variables.

ACT	CAU	LIG	SEX	OI	TIM	ATF	PAW	AGE
100%	57.6%	42.3%	34.9%	26.9%	23.5%	20.5%	18.2%	17.8%

3.2. Decision Rules

DT obtained has 14 terminals nodes, 9 of them has been identified as DRs. Table 3 shows a description of the DRs obtained which have been ordered by number of the node . As it could be observed, most of the rules are SI rules (6 of 9), however 3 important KSI rules have been identified.

About parameter analyzed, all the rules include at least 1% of the population, having rule 6 a percent of 16% of the population. Probability parameter, it could de remark that probability values are higher than 60%, with 70.59% being the highest value (rule 24).

About the length of one rule it could be said that less numbers of variables involved in the rule to imply the higher its predictive capacity of the rule. In DRs analyzed, rule length varies from 2 variables (as in node 5) to a maximum of 11 variables (as in node 26), so DRs obtained are enough informative

Seeing Table 3, it could be remark that:

- Two of three KSI rules have a male drivers involved.
- When there are pedestrians involved in accident, the probability of KSI increases: two out of three accidents involving pedestrians and male drivers will be KSI (rule 8).
- In general, accidents due to causes not attributable to the driver tend to have minor consequences, SI accident (rule 6). With a probability of 67.1% , this rule representing almost 16% of the total population.
- Also, a higher probability of KSI for accidents that were not collisions caused by male drivers on roads with a pavement width of less than 6m are identified.
- About DRs obtained with variable LIG, when women drivers cause a crash, CART methods predict SI accident if lighting exists (full daylight, sufficient lighting and dusk) (node 9). However, when the lighting is

non-existent or insufficient (node 10) the rule obtained is KSI accident. This rule is not observed for men and may indicate that women increase their risk of severity under conditions of less lighting on the road.

Table 3. DRs obtained.

NODE	VARIABLES OF THE RULES: [IF (AND … AND)]	THEN	Po (%)	P (%)
5	IF [(CAU≠DC) AND (ACT=CP)]	KSI	1.39	72.00
6	IF (CAU ≠ DC) AND (ACT=ROR OR ACT=OT OR ACT = ROOR ACT = CO).	SI	15.88	67.13
8	IF (CAU=DC) AND (SEX=M) AND (ACT=CP OR ACT = CO).	KSI	5.66	66.67
9	IF (CAU =DC) AND (SEX ≠ M) AND (LIG ≠ WL AND LIG ≠ IL).	SI	9.05	67.48
11	IF (CAU =DC) AND (SEX = M) AND (ACT ≠ CP AND ACT ≠ CO) AND (PAW = THI).	KSI	9.99	61.11
14	IF (CAU =DC) AND (SEX = M) AND (ACT ≠ CP AND ACT≠ CO) AND (PAW ≠ THI) AND (ATF = LR OR ATF = HR).	SI	5.89	62.26
16	IF (CAU=DC) AND (SEX=M) AND (ACT≠CP AND ACT≠CO) AND (PAW≠THI) AND (ACT=RO).	SI	3.55	60.94
19	IF (CAU=DC) AND (SEX=M) AND (ACT≠CP AND ACT≠CO) AND (PAW ≠ THI) AND (ATF ≠ LR AND ATF ≠ HR) AND (ACT =ROR OR ACT=OT) AND (AGE ≠ UN) AND (LIG = SL).	SI	2.33	64.29
24	IF (CAU =DC) AND (SEX = M) AND (ACT ≠ CP AND ACT ≠ CO) AND (PAW ≠ THI) AND (ATF ≠ LR AND ATF ≠ HR) AND (ACT =ROR OR ACT=OT) AND (AGE ≠ UN) AND (LIG ≠ SL) AND (NOI ≠ [>1]) AND (OI = [>2] OR OI = [2]).	SI	2.83	70.59

4. Conclusions

CART method allows classification based on crash severity and provides an alternative to parametric models because of their ability to identify patterns based on data, without the need to establish a functional relationship between variables. In fact, CART analysis does not need to specify a functional form as ordinary statistical modelling techniques, such as regression models. In regression analysis if the model is misspecified, the estimated relationship between dependent variable and independent variables as well as model predictions will be erroneous. So, CART model has a number of benefits compared to other widely used parametric models.

One of the most important advantages of the CART model is that the outcomes of the analysis are easy to understand and perform due to the graphical nature of its results. Also, the CART analysis allows a great many explanatory variables and it can easily find the important variables of the model.

Moreover, CART has permitted certain potentially useful rules to be determined that can be used by road safety analysts and managers. DRs obtained have been classified based on their severity, so, firstly the safety analysts should focus on severe or mortal crashes and subsequently intervene in accidents whose results are slight injuries. The approach proposed in this work within each group will enable the actions to give priority on the basis of population and probability.

Analyzing DRs, certain overall conclusions from a road safety perspective could be remarked:

- Due to length of DRs obtained, it could be said that they are enough informative.
- The CART method enables to obtain the importance of the variables in the model. In this case, the most important variables are: accident type, cause of the accident and lighting.
- The structure of the tree is generated by variable "cause of accident".
- Male drivers are the main cause of KSI crashes.
- Women drivers have more probability than men drivers from suffering KSI accident when the lighting is non-existent or insufficient.

However, DTs models are often unstable. They could suffer variations if different strategies such as stratified random sampling (with injury severity as the stratification variable) are applied for creating learning and testing datasets [5].

Finally, the main problem observed with CART is that only binary trees can be built. For this reason certain categories of splitting variables are grouped in some branches, increasing node support, but making impossible to analyse the influence of a specific category on crash severity. For this reason it could be suitable to use other methods to built DTs which allow trees without binary restriction in the branches.

Acknowledgements

The authors are grateful to the Spanish General Directorate of Traffic (DGT) for providing the data necessary for this research. 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.

References

[1] WHO, World Health Organization, (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

[2] Kashani, A. and Mohaymany, A. (2011). "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models". Safety Science 49, pp.1314-1320.

[3] 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

[4] 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.

[5] Chang, L.Y. and Wang, H.W. (2006). "Analysis of traffic injury severity: an application of non-parametric classification tree techniques". Accident Analysis and Prevention 38, pp. 1019–1027.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Breiman, L., Friedman, J.,Olshen, R., and Stone, C., (1984). "Classification and Regression Trees". Belmont, CA: Chapman & Hall.

[11] 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, pp. 402–411.

[12] 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.

[13] Kockelman, K.M., Kweon, Y.J., (2002). Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34, 313–321.

[14] 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.

[15] Abdel-Aty, M., (2003). "Analysis of driver injury severity levels at multiple locations using ordered probit models". Journal of Safety Research 34, 597–603.

[16] 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.

[17] 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.

[18] Montella, A., Aria, M., D´Ambrosio, A., Mauriello, F., (2011). Data Mining Techniques for Exploratory Analysis of Pedestrian Crashes. Transportation Research Record: Journal of Transportation Research Board, No. 2237, Transportation Research Board of the National Academies, Washington, D.C., 107-116.