# Analysis of pedestrian injury severity using Data Mining techniques

Joana Alexandre de Sousa

Instituto Superior Técnico - Universidade de Lisboa

June, 2017

**Abstract**

*Road accidents are a major problem worldwide, particularly those involving pedestrians. Portugal is one of the Member States of the European Union with the highest rates of pedestrian fatalities per million inhabitants. Despite all the strategies that have been applied the severity of pedestrian accidents in Portugal is a real problem that requires the development and implementation of specific measures of road safety. Therefore, using data mining techniques, the present study aims to find and understand the factors that contribute to increase the severity of pedestrian injuries, in order to minimize them. Bayesian networks, decisions trees and ordered logistic regression models were used to analyse a road accidents database of the Portugal national road safety authority (ANSR). The choice of the best model was based on performance measures such as, accuracy, sensitivity, specificity, F-score and AUC. The several models analysed performed well, however, the bayesian networks showed to be one step ahead, in particular, regarding the classification of the classes of serious injuries and fatal injuries of pedestrians. Throughout this analysis, were found causes for the severity of pedestrian injuries common to all models. These include accidents occurring outside urban areas, heavy vehicles, the Lisboa region and highway, principal itinerary or complementary itinerary. All these causes were pointed out by the three models as crucial to increase the severity of pedestrian's injuries.*

**Keywords**: bayesian network, data mining, decision tree, ordered logistic regression, pedestrian injuries

## Introduction

Road accident statistics in Europe show a need for more systematic mechanisms for the analysis and prediction of accidents. According to the World Health Organization (WHO), it is estimated that road traffic injuries become the seventh leading cause of death by 2030 (WHO - Global status report on road safety [15]). In 2012, Portugal National Road Safety Authority registered 29867 accidents with victims resulting in 38105 injured people. Regarding accidents involving pedestrians, during the same period, 5245 people suffered injuries, of which 159 resulted in fatalities. The pedestrian death toll equals to 22% of all Portugal road deaths in 2012 [1]. All EU-15 Member States show a declining trend in the rate of pedestrian fatalities per million inhabitants. However, Portugal remains one of the countries with the highest rates. It follows that, despite all the strategies applied, the severity of pedestrian's accidents in Portugal is a real problem that requires the development and implementation of specific measures of road safety.

Several studies have been used to analyse the causes of the severity of accidents, such as, driver related factors, vehicle characteristics and road conditions, mainly by using logistic regression models (Al-Ghamdi (2002), Bédard et al. (2002), Yau et al. (2006) and Milton et al. (2008)) or ordered probit models (Kockelman and Kweon (2002), Zajac and Ivan (2003) and Clifton et al. (2009)). Logistic and probit regression models belong to the family of generalized linear models, which have their own assumptions and establish relationships between explanatory and response variables, as verified by Chang and Wang (2006) and Õna et al. (2011). In order to avoid incorrect estimates or difficulties in dealing with these pre-defined assumptions, some researchers have recourse to other types of techniques, such as decision trees and more rarely bayesian networks or neural networks.

Decision trees models were adopted, for instance, in Chang and Wang (2006), due to having no relation between independent and dependent variables. In this research the goal was to establish a relationship between the severity of the injury and the variables of the accident. Chang and Wang reported the vehicle category as the most important variable associated with the severity of the accident. Furthermore, they also verified that pedestrians, riders and bicycle riders are those who are at greater risk of injury. On the other hand, the bayesian network were applied by Õna et al. (2011) to classify traffic accidents according to their severity of injury. In this analysis three different bayesian networks were built, one for each severity class of injury (slight, severe or fatality). Type of accident, driver's age, lighting conditions and number of injuries were factors associated with fatalities or severe injuries.

Although each of these different methods (regressions, decision trees and bayesian networks) have their advantages described in several studies, there are few that compare the methods to each other for this type of dataset. Zong et al. (2013) was one of the few studies that presents a comparison between bayesian networks and regression models (logistic and probit). The comparison between the methods was based on the accuracy of each model.

The study reported that the bayesian network model was more suitable to predict the severity of the accident than the regression models, in terms of model accuracy.

With regard to the severity of the pedestrian's injury, this is a problem that has been much less studied and analysed than the severity of the accident as a whole. Researches that folded on this topic have usually resorted to logistic regression. Examples are: Sze and Wong (2007) and Kim et al. (2008).

The present study aims to find and understand the factors that contribute to increase the severity of pedestrian's injuries, focusing only on accidents involving injured pedestrians. For that, ANSR provided a database with detailed information on the accidents and the drivers and pedestrians involved. The goal will be to determine which factors actually have a real influence on the severity of pedestrian's injuries. For this analysis we will resort to several techniques of data mining. Furthermore, we will compare the results obtained with the different techniques in order to determine which of them gives us the best contributes to our problem.

# Variable Selection Methods

## Correlations Measures

Correlation measures are one of the most common methods for selecting variables. They measure the association strength between two variables. In cases where the variables are categorical, the usual correlation coefficient (Pearson) can not be used. In this case, we need to use other type of coefficients, measures or even statistical tests to analyse the relationship between two variables. Spearman coefficient or the Kruskal Wallis test are examples of possible alternatives.

## Mutual Information Criteria

Mutual Information (MI) measure the information shared by two variables, this can be observed when a variable decreases in uncertainty by assuming knowledge of another. For discrete random variables X and Y with (joint) Probability Mass Function (PMF) $\mathsf{P}_{X,Y}(x,y)$ and marginal PMF $\mathsf{P}_X(x)$ and $\mathsf{P}_Y(y)$, the MI criteria can be defined as:

$$MI(X,Y) = \sum \mathsf{P}_{X,Y}(x,y) \log\left(\frac{\mathsf{P}_{X,Y}(x,y)}{\mathsf{P}_X(x)\mathsf{P}_Y(y)}\right)$$

A potential problem with this type of selection is that two variables may be highly informative in relation to the response variable, but very similar among them so it would be reasonable to choose only one.

## Minimum Redundancy Maximum Relevance Criteria

Taking into account the problem of mutual information Peng et al. (2005) suggested the minimum Redundancy Maximum Relevance (mRMR) criteria. This criteria bases the selection of variables not only on the relevance towards response variable, but also on the similarity with the already chosen explanatory variables. Formally, this method attempts to maximize mutual information with the response variable, keeping the pairwise mutual information among the explanatory variables as low as possible. This is done by choosing in the $m$-th step the variable $X_k$ such that:

$$k : \arg \max_{t \notin S_{m-1}} \frac{MI(X_t, Y)}{\sum_{j \in S_{m-1}} MI(X_t, Y)}$$

where $S_m$ is the set of indexes of the first $m$ selected variables and $Y$ the response variable.

## Stepwise Regression

Stepwise regression is another method used in the exploratory analysis of the dataset to identify useful variables to explain the response variable. In each step the process systematically adds the most significant variable or removes the least significant, depending on the type of search: forward, backward or both. Forward selection starts with a model without variables and rehearses the addition of each variable using a comparison criteria (e.g. AIC, BIC, p-value) and repeats the process until there is no model improvement. Backward selection starts with a model with all variables and tests its removal using a comparison criteria. The process is repeated until no further improvement is possible. Selection in both directions is the combination of the two above, where variables are tested at each step in order to be included or excluded from the model.

## Imbalanced Dataset

Sampling techniques are the most traditional choice when dealing with the problem of imbalanced classes and can be considered two approaches: oversampling and undersampling. Oversampling replicate instances from the minor class and repeats them until the classes have equal frequency. Undersampling focuses on the majority class by discarding instances until its reaches the size of the minor ones. Another technique is SMOTE propose by Chawla et al. (2002), which is a combination of oversampling the minority class and undersampling the majority. According to Chawla et al. this method can achieve better performance of the classifier than just the undersampling approach.

# Statistical Models

## Bayesian Networks

A Bayesian Network (BN) is a graphical representation of a probability distribution over a set of $k$ random variables $\{X_1, \ldots, X_k\}$. Hence, a BN can be described as a pair $\mathcal{B} = (\boldsymbol{X}, \mathcal{G})$ composed of a random $k$-dimensional variable $\boldsymbol{X} = (X_1, \ldots, X_k)$ and a $\mathcal{G}$ Directed Acyclic Graph (DAG). This acyclic graph is composed by a set of $k$ vertices (represented each of the random variables $X_i$, $i = 1, \ldots, k$) and a set of arrows between these vertices, which represents direct dependences among the variables. For each variable $X_i$, with parents $Y_1, \ldots, Y_n$ there is attached the conditional probability table $\mathsf{P}(X_i|Y_1, \ldots, Y_n)$. If $X_i$ has no parents, then the table is reduced to the unconditional probability $\mathsf{P}(X_i)$. Denote as $\Pi_{X_i}$ the set of random variables that are the immediate predecessors (the parents) of the variable $X_i$. According to the DAG, the joint probability distribution of $\mathcal{B}$ is given by the following equation:

$$\mathsf{P}_{\mathcal{B}}(\boldsymbol{X}) = P(X_1 = x_1, \ldots, X_k = x_k) = \prod_{i=1}^{k} \mathsf{P}_{\mathcal{B}}(x_i|\Pi_{X_i})$$

where $x_i$ is the realization of $X_i$. In other words the joint PDF of the network can be factorized into smaller PDF, each involves the node and its parents.

In order to construct a BN (i.e, to establish the joint probability distribution of the random vector) it is necessary to specify the conditional dependence (or independence) relation between the variables and the conditional probability distributions. The structure of the network can be defined a priori or by means of an estimate made from the data (learning structure). The algorithms to learn the structure of a BN can be classified as: constraint-based algorithms; score-based algorithms; or hybrid algorithms. Constraint-based algorithms use a conditional independence test (such as, $\mathcal{X}^2$, $G$-test or mutual information) on the data to detect and search a network consistent with the observed.

Score-based algorithms rank network structures in relation to a goodness-of-fit score. First, it quantifies the fit of a BN considering a scoring function. Second, it defines a search engine to find a structure that maximizes the defined score. Both steps are applied iteratively until there is no modification that can improve the score, that is, no new structure is better than the previous one. Two different types of scoring functions can be considered: bayesian scoring functions such as K2, Bayesian Dirichlet (BD) and its variants (BDe and BDeu); and information theoretic scoring functions such as Log-Likelihood (LL), AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). BDe score is a particular case of the Bayesian Dirichlet score (BD) and can be defined as follows:

$$\mathsf{BDe}(\mathcal{B}, \mathcal{T}) = \mathsf{P}(\mathcal{B}) \times \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left( \frac{\Gamma(N'_{ij})}{\Gamma(N_{ij} + N'_{ij})} \times \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + N'_{ijk})}{\Gamma(N'_{ij})} \right)$$

Assuming that, $\mathcal{T}$ is the dataset, $\rho(\Theta_D|\mathcal{G})$ is Dirichlet with equivalent sample size $N'$ for some complete DAG $\mathcal{G}$ in $\mathcal{T}$, $r_i$ the number of states of $X_i$, $N_{ijk}$ the number of instances in $T$ where the variable $X_i$ takes it $k$-th value $x_{ik}$, $N'_{ijk} = N' \times \mathsf{P}(X_i = x_{ik}, \prod X_i = w_{ij}|\mathcal{G})$ and $q_i = \prod_{X_j \in \prod_{X_i}} r_j$ the number of possible configurations of the parent set $\prod_{X_i}$ of $X_i$.

Regarding information theoretic scoring functions, an example is the Akaike Information Criterion (AIC) which can be defined as:

$$AIC(\mathcal{B}|\mathcal{T}) = LL(\mathcal{B}|\mathcal{T}) - \left| \sum_{i=1}^{n} q_i(r_i - 1) \right|$$

where $LL(\mathcal{B}|\mathcal{T})$ is the log-likelihood score defined as:

$$LL(\mathcal{B}|\mathcal{T}) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \left( \frac{N_{ijk}}{N_{ij}} \right)$$

For the second step of a score-based algorithm it is necessary to define a search algorithm that goes through the network. For this, some classic search algorithms such as hill-climbing, tabu search or simulated annealing (also known as metropolis) can be applied.

## Decision Trees

There are different types of decision tree algorithms, one of the most used is the CART (Classification And Regression Tree). A CART tree is a binary tree that is constructed by splitting a node into two nodes repeatedly. First find the variable that is the best split (root) and then find the best division of this variable. If the stop rules are not satisfied, the variable is split using its best division found in the previous step. The paths from the root to the leaves represent the classification rules. For a given node $t$, the split $s$ is chosen in order to maximize a splitting criterion $\Delta G(s, t)$. The most used criterion is Gini Index, one of the splitting criterion that allows the use of categorical data. Given a node $t$ with points belonging to one of the $m$ classes, $c_1, c_2, \ldots, c_m$ its Gini Index is defined by:

$$G(t) = \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} \hat{\mathsf{P}}(c_i|t) \hat{\mathsf{P}}(c_j|t)$$

where $\hat{\mathsf{P}}(c_i|t)$ is the estimate probability of an element of $t$ belonging to class $i$.

For a given node $t$, a split $s$ and $p_l$ and $p_r$ the proportions of nodes $t_l$ (left branch) and $t_r$ (right branch) respectively, the Gini Gain (Gini splitting criterion) is the decrease of impurity defined as:

$$\Delta G(s, t) = G(t) - p_l G(t_l) - p_r G(t_r)$$

The greater the Gini gain, the greater the decrease in the impurity of the class after its partition, the better the variable that originated the split.

Conditional inference tree is an algorithm developed by Hothorn et al. (2006) for regression and classification that uses binary recursive partition, such as the CART algorithm. This new algorithm based the tree construction on statistical properties associate to the variables and can be divided into three steps:

1. Tests the null hypothesis of independence between each explanatory variable and the response variable. Stops if the hypothesis cannot be rejected, otherwise, select the variable with the highest $p$-value for the test of the partial null hypothesis;

2. Implements a binary split in the selected variable;

3. Iterates the Steps 1 and Step 2 recursively for both elements of the split obtained in Step 2, until the null hypothesis of independence cannot be reject.

## Ordered Logistic Regression

An ordered logistic model (proportional odds model) is a statistical technique used to analyse the relationship between a multilevel ordinal response variable and one or more explanatory variables. Let $Y$ be a random ordinal response variable with $m$ categories. Assuming that it is only possible to observe $Y$ at specific thresholds of $Y^*$, which can be seen as the regions of the real line corresponding to the different ordinal categories of $Y$. Formally, $Y_i^*$ is defined as a continuous random variable with $i = 1, \ldots, n$ and $\pi_{ic}$ as the probability of $i$ belonging to the $c$ category with $c = 1, \ldots, m$. Then, the cut-off points of the distribution of $Y_i^*$, $\gamma_{i,1}, \ldots, \gamma_{i,m-1}$ for each $i$ are defined by

$$\begin{aligned} P(Y_i^* \leq \gamma_{i1}) &= \pi_{i1} \\ P(Y_{i,c-1} < Y_i^* \leq \gamma_{ic}) &= \pi_{ic}, \quad c = 2, \ldots m-1 \\ P(Y_i^* > \gamma_{i,m-1}) &= \pi_{im} \end{aligned}$$

Consider the cumulative probabilities, $\pi_{ik}^* = \sum_{c=1}^{k} \pi_{ic}$ with $0 < \pi_{ic} < 1$, $k = 1, \ldots, m$ and $\pi_{ic}^* > \pi_{i,c-1}^*$ holds. The cumulative logit model with proportional odds is given by

$$\begin{aligned} \log\left(\frac{P(Y^* \leq \gamma_{ic})}{P(Y^* > \gamma_{ic})}\right) &= \log\left(\frac{\pi_{i1} + \cdots + \pi_{ic}}{\pi_{i,c+1} + \cdots + \pi_{im}}\right) \\ &= \beta_{0c} - \beta_{1c}x_{i1} - \cdots - \beta_{k-1,c}x_{i,k-1} \\ \text{or} \quad log\left(\frac{\pi_{ic}^*}{1-\pi_{ic}^*}\right) &= \beta_{0c} - \beta_{1c}x_{i1} - \cdots - \beta_{k-1,c}x_{i,k-1} \end{aligned}$$

where $x_1, \ldots, x_{k-1}$ are the $k-1$ explanatory variables and $\beta_{01} < \cdots < \beta_{0,m}$ to ensure that $\pi_{i,c}^* \geq \pi_{i,c-1}^*$. The linear predictor $\boldsymbol{x}^T\boldsymbol{\beta}_c$ is restricted so that the intercept $\beta_{0,c}$ may depend on $c$, but the effects of the other predictor variables be constant across the response categories (proportional odds).

# Model Selection

### Performance metrics

Different models can be fitted using the same data. Different algorithms can be applied and, even the same algorithm can have different results with changes of its parameters. In real problems it is not possible to find the perfect model, so it is necessary that we can determine the best model for each case. There are several measures that can be used to evaluate the performance of the models, usually based on the measures obtained with a confusion matrix. Confusion matrix is a special case of a contingency table, which allows the visualization of the performance of an algorithm. Table 1 represents a confusion matrix for a model whose response variable has three levels or classes.

**Table 1:** Example of a confusion matrix for a response variable with three classes.

| | | Reference | | | |
|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | |
| Model | Class 1 | $TP_1$ | $FN_{12}$ | $FN_{13}$ | |
| | Class 2 | $FN_{21}$ | $TP_2$ | $FN_{23}$ | |
| | Class 3 | $FN_{31}$ | $FN_{32}$ | $TP_3$ | |
| | | | | | Total |

In Table 1, $TP_i$ describes the true positives for the $i$-th class, which are the observations that the model predicts belong to class $i$ to which the observation actually belong. In contrast, $FN_{ij}$ describes the false negatives for the $i$-th class. In this case, the model predicts that observations belong to class $i$ when they actually belong to class $j$. Table 2 summarizes the main performance measures that will be estimated based on the true positives and false positives presented in a binary confusion matrix.

**Table 2:** Performance measures based on probabilities, where $\phi$ denotes the classification (on the model), $c^+$ denotes the data observations classified as class $i$ which are actually of class $i$ and $c^-$ denotes the data observations classified as class $i$ which in fact belongs to one of the remaining classes of the response variable $Y$.

| Accuracy | Precision |
|---|---|
| $p(\phi(\boldsymbol{X}) \neq C)$ | $p(C = c^+ \mid \phi(\boldsymbol{X}) = c^+)$ |
| **Sensitivity** | **Specificity** |
| $p(\phi(\boldsymbol{X}) = c^+ \mid C = c^+)$ | $p(\phi(\boldsymbol{X}) = c^- \mid C = c^-)$ |

Without loss of generality, it will be used the terms accuracy, precision, sensitivity and specificity as the estimates of the measures instead of the true probabilities of them.

Accuracy evaluated the global performance of the model by estimating the predicted data observations correctly/incorrectly classified. Precision represents the fraction of data observations classified as class $i$ (or $c^+$

by the definition in Table 2) which are actually belong to class $i$. Sensitivity measures the positive class observations correctly classified.In a problem with three levels there are three positive classes observations, one for each class. Conversely, specificity measures the fraction of the correctly identified negative class of observations. This means that for class 1 of the confusion matrix in Table 1, specificity measures the fraction of negative class of observations correctly identified. Note that, for example, when computing the sensitivity for class 1 of the confusion matrix of Table 1, $c+$ (from the definition of Table 2) represents class 1. On the other hand, when computing specificity for class 1 $c-$ represents the remaining classes (class 2 and 3). Another measure used is the harmonic mean of the precision and sensitivity, traditional called F-measure or F-score and can be given by:

$$F = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}$$

Another technique to analyse the performance of the model is the ROC (Receiver Operating Characteristic) curve. ROC curve is a graph that illustrates the model's performance representing the true positive rate (sensitivity) against the false positive rate (1-specificity). The area under the ROC curve is another widely used performance measure which allows to verify the ability of the model to identify different classes. In a multiclass problem, there are a sensitivity and a specificity value for each class of the response variable. Therefore, there will be a ROC curve and its AUC for each of these classes. To tacking this, Hand and Till 2001 suggested an extension of the AUC definition for cases that we have more than two classes, which is based on pairwise comparisons. Thus, the AUC for a multiclass problem can be given by:

$$AUC = \frac{2}{c(c-1)} \sum_{i<j} A(i,j)$$

where $c$ represents the number of classes of the response variable and $A(i,j)$ is the estimated probability of a randomly chosen member of class $j$ having a lower estimated probability of belonging to class $i$ than a member of class $i$ also chosen randomly, with $i$ and $j$ two different classes of the response variable.

# Dataset

The dataset refers to road accidents involving injured pedestrians occurred in Portugal in 2011-2012. This set was provided by Portugal national road safety authority and has detailed information on all accidents recorded with victims. Table 3 describe all variables of the dataset.

Each observation (dataset row) refers to an injured pedestrian and contains information that can be divided into: accident's information (including road conditions), driver's details and pedestrian's details. In this dataset, the ANSR uses the definition of fatalities adopted by the Vienna Convention of 1968 which considers fatalities all victims who passed way within 30 days of the accident. Pedestrians with serious injuries are victims of bodily injury requiring medical assistance for more than 24 hours and do not die within 30 days of the accident. Pedestrians with minor injuries are all other cases of injured pedestrians who do not die within 30 days of the accident. A more complete and detailed information can be found in the ANSR Annual Report ([1]).

**Table 3:** Description of each variable of the dataset.

| Variables | Description |
|---|---|
| `Month` | Month |
| `DaysWeek` | Days of the Week |
| `Hour` | Hour |
| `RoadType` | Road Type |
| `Region` | Administrative Region |
| `Location` | Location |
| `Weather` | Weather Conditions |
| `Light` | Lighting Conditions |
| `Grip` | Grip Conditions |
| `Direction` | Road Directions |
| `Intersection` | Road Intersection |
| `SegmentType` | Road Segment Type |
| `TrafficLanes` | Traffic Lanes |
| `DriverAge` | Driver's Age |
| `DriverGen` | Driver's Gender |
| `DriverInjuries` | Driver's Injuries |
| `DriverAction` | Driver's Actions |
| `License` | Driver's License |
| `VehicleCat` | Vehicle Category |
| `Alcohol` | Driver's Blood Alcohol Content |
| `PedestAge` | Pedestrian's Age |
| `PedestGen` | Pedestrian's Gender |
| `PedestAction` | Pedestrian's Actions |
| `PedestInjuries` | Pedestrian's Injuries |

The dataset has information on 8431 injured pedestrians, of whom 7479 were minor injured, 666 were seriously injured and 286 were fatalities. `PedestInjuries` is the response variable, which is divided into minor, serious and fatal injuries, being an ordinal variable. The analysis shows that the response variable is extremely unbalanced. As such, to enhance the results, we decided to balanced the dataset using the SMOTE method provided by the R *DMwr* package. Moreover, the initial dataset was split into two sets: the train set (75%) and the test set (25%) in order to evaluate the models through performance measures.

There are only three continuous variables in this dataset. Two of them are related to age (pedestrian and driver) and the other describes the driver's blood alcohol content. Detailed analysis of these variables indicated different levels of pedestrian injury severity, depending on the age group and the different levels of alcohol. Therefore, we decided to discretize these three variables and, in addition, regroup classes of several other categorical vari-

ables. The final categories of each variable are described in Table 13.

A selection of variables was performed in order to determine which variables should be included in the model. The selection was based on correlation measures, mutual information, mRMR and stepwise regression. The variable's selection was performed in the original dataset and in the discretized and regrouped dataset. After analysing the results of all the approaches, we defined some criteria to select the variables. First, we choose the variables selected by all techniques for both datasets. Thus, the first selected variables were: `RoadType` and `PedestAge`. The second step was to choose the variables selected at least in one dataset by each technique. In this step were selected: `Region`, `Location`, `Light`, `DriverAge`, `DriverInjuries`, VehicleCat, `Alcohol` and `PedestAction`. At the end, and tackling into account all the considerations, the selected variables that will be used to fit the models are describe in Table 4.

**Table 4:** Set of selected variables.

| Final Selection of Variables | | |
|---|---|---|
| `RoadType` | `Region` | `Location` |
| `Light` | `DriverInjuries` | `VehicleCat` |
| `Alcohol` | `PedestAge` | `PedestAction` |

# Results

### Bayesian Network

In the process of finding the best bayesian network, we resorted to the R package called *bnlearn*. We used two learning algorithms to build the networks: hill-climbing and tabu search. At the same time, all the score functions available for categorical variables (AIC, BIC, BDe, mBDe, K2 and Loglik) were also applied. The differences between the results of both algorithms were minimal so, only the results of the hill-climbing algorithm will be analysed in detail. The results of the performance measures allowed to conclude that the highest values of accuracy, sensitivity and specificity belong to the BDe and mBDe scores. However, the models learned from these two scores were not capable of conducting a proper identification of the causes of serious or fatal pedestrian's injuries. Thus, by weighing all the performance measures of the remaining scores we chose as the best model the one learned with loglik score. The choice fell on this score since, in this particular problem, we considered it more important that the model be able to identify cases of serious injury or even fatal injuries (true positive rate) than to identify all cases in general (accuracy).

In short, the chosen model was learned with the hill-climbing algorithm with the loglik score and the parameter estimation by the maximum likelihood method. The respective bayesian network is shown in Figure 1 and its results will be analysed below.

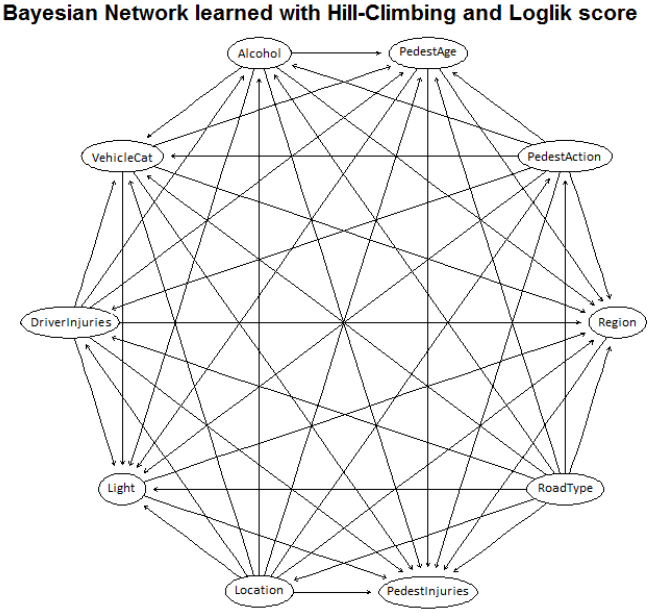


**Figure 1:** Bayesian network learned with hill-climbing algorithm and loglik score from the selected variables.

Once the links between the nodes of the network are specified, the next step is to quantify the relationships between those nodes. Which is done by specifying a conditional probability distribution for each node. In this case, all the variables are discrete and, so, the relationships are represented in a conditional probability table (CPT). Each node has a CPT and to analyse it is necessary to look at all possible combinations of values of its parent nodes. In this study, the goal is to find the causes of pedestrian injury severity therefore, the node of interest is `PedestInjuries` and its parents. In this particular network to analyse the node `PedestInjuries` is necessary taking into account all the other nodes since all of them are parents of this particular node. Thus, for `PedestInjuries` we have a CPT for each combination of all possible values of the parent nodes. This has resulted in hundreds of CPT, which have become unreasonable to analyse all them. Table 5 represents the estimated probability of pedestrian injuries by road type conditioned to: the crash location being in urban areas; the lighting conditions be daylight; the administrative region belongs to group 1; the driver be unharmed; the vehicle be a light vehicle; the driver's blood alcohol content be less than 0.2 g/L; the pedestrian be less than 20 years old; and the pedestrian's action be illegal. In order to simplify Table 5 will show the estimated probabilities for the street and other type of road levels of `RoadType` variable.

The analysis to Table 5 shows that in the streets the highest estimated probability is for minor injuries. While in other type of roads are the serious and fatal injuries that had the higher estimated probability.

Consider now Table 6, which represents either the estimated conditional probability for each level of pedestrian's injuries for the same conditions of Table 5. However, instead of the administrative region belongs to

**Table 5:** Estimated conditional probability for each `RoadType` class per pedestrian's injuries.

| | **RoadType** | |
|---|---|---|
| | Street | Other type of Roads |
| Minor Injury | 0.6667 | 0.0000 |
| Serious Injury | 0.3333 | 0.5714 |
| Fatal Injury | 0.0000 | 0.4286 |

group 1, in this case belongs to group 2 (Porto).

**Table 6:** Estimated conditional probability for each `RoadType` class per pedestrian's injuries.

| | **RoadType** | |
|---|---|---|
| | Street | Other type of Roads |
| Minor Injury | 0.7333 | 0.0000 |
| Serious Injury | 0.2667 | 0.9999 |
| Fatal Injury | 0.0000 | 0.0000 |

Comparing the results of both tables the significant differences are found in the class of other type of roads. The estimated probability of a fatality of a pedestrian decreases from 0.4286 to almost 0. However, the highest estimated probability still to belong to serious injuries. Regards to the streets there is almost no difference. The estimated probability of a pedestrian suffering minor injuries increases to 0.7333 (it was 0.6667), while the estimated probability of a pedestrian suffering serious injuries decreases to 0.2667 (was 0.333).

### Decision Tree

The second approach was to fit a decision tree to our dataset. In order to built it, two different procedures were used: decision trees (DT) and conditional trees (CT). Here DT stands for the CART algorithm from the R *rpart* package, while CT stands for the conditional inference tree built with the **ctree** function of the R *party* package.

Results of the performance measures of both algorithms shows very similar scores, with a slight advantage for CT. However, sensitivity, specificity and F-score of DT are more balanced for the three classes of the response variable. Therefore, given the size and consequent difficulty of visualizing the CT results, we chose analyse the DT results.

The split criterion used to build DT was the gini gain. The tree presents five primary splitters: crash location, vehicle category, pedestrian's age, administrative region and road type. Thus, these are the critical variables for analyse the pedestrian injury severity in road accidents.

The initial split at node 1 is based on the variable that represents the location of the accident. Which means that this is the single best variable to evaluate pedestrian injury severity. The tree indicated that accidents occurring in rural areas are more likely to result in serious injury (61%). While in urban areas the pedestrian is almost as likely to suffer fatal or serious injuries (23% and 24%, respectively). However, they are most likely to suffer minor injuries (53%).

In regards to the vehicle category the tree predicts that when heavy vehicles are involved 55% of pedestrians are more likely to be fatal injured against to 28% to be serious injured and only 17% to be minor injured.

Regarding the variable road type, conditioned to the pedestrian's age, vehicle category and crash location the tree splits its into two terminal nodes. In the left branch, forming terminal node 1 are the streets, while in the terminal node 2 are the remaining classes of road type. In the streets the tree estimates that pedestrians are 71% most likely to be minor injury. While in all the other road categories they are most likely to be serious injury (with an estimate probability of 53%, against 38% for minor injuries). On the other hand, when road type is conditioned by the administrative region, vehicle category and crash location the tree estimates that in accidents that occur in streets pedestrians are only 55% likely to suffer minor injuries compared to 25% serious injuries and 20% of fatal injuries. While in all other road types they are more likely to suffer serious injuries (52%) against 20% minor injuries and 29% fatal injuries.

The last factor is the administrative region whose tree split group 1 (Lisboa) and group 4 (Braga and Viana do Castelo) against the remain groups. Group 1 and group 4 represent the regions where a pedestrian is more likely to suffer a fatality (45%). However the estimate probability of a pedestrian being minor injured is 39%.

### Ordered Logistic Regression

A final approach for fit a model to the dataset was ordered logistic regression. In particular, an ordered multinominal logistic regression model since our response variable has three levels. Likelihood ratio test for overall significance of the 10 coefficients showed that at least one of them is significant to explain the severity of pedestrians injuries. Next, we analyse the particular results for each model variable. The following tables present the values of odds ratio, p-value, lower bound and upper bound of the ordered logistic regression model for some of the selected variables. The odds ratio (OR) values represented the increase (OR $> 1$) or decrease (OR $< 1$) in the odds between the classes of each variable. The Lower Bound (LB) and the Upper Bound (UB) are the limits of a confidence interval of 95% ($CI_{95\%}$) for rejecting the null hypothesis in the significance test (univariate Wald test). Note that the analysis of the odds ratio of each one of the variables assume fixed all the remaining variables.

**Table 7:** Estimates of OR for the variable: `Location`.

| `Location` | OR | $p$-value | LB | UP |
|---|---|---|---|---|
| Urban areas | - | - | - | - |
| Rural areas | 1.81 | 0.00 | 1.54 | 2.13 |

With regards to crash location variable, the results indicate that when there are accidents in rural areas, there

is a 81% increase in the odds of giving a response that indicates higher levels of severity of pedestrian injuries when compared to accidents occurring in urban areas.

**Table 8:** Estimates of OR for the variable: `RoadType`.

| `RoadType` | OR | $p$-value | LB | UP |
|---|---|---|---|---|
| Streets | - | - | - | - |
| National Road | 1.60 | 0.00 | 1.25 | 2.04 |
| Highway, principal or complementary itinerary | 2.98 | 0.00 | 2.31 | 3.85 |
| Other type of roads | 1.89 | 0.00 | 1.65 | 2.17 |

The results in Table 8 show that when we compare accidents on street with each of the other categories, there is an increase in the odds of all of them. In accidents on national roads the odds increase by 60% giving a response that indicates higher levels of pedestrian injury severity. Accidents on highway, principal itinerary or complementary itinerary have 198% increase in odds indicating higher levels of pedestrian injury severity. For other types of roads, these results show a 89% increase in the odds.

**Table 9:** Estimates of OR for the variable: `VehicleCat`.

| `VehicleCat` | OR | $p$-value | LB | UP |
|---|---|---|---|---|
| Light vehicle | - | - | - | - |
| Heavy vehicle | 4.88 | 0.00 | 4.16 | 5.73 |
| Two motorized wheels | 0.99 | 0.91 | 0.76 | 1.28 |
| Other type of vehicles | 0.57 | 0.00 | 0.41 | 0.78 |

Compared with light vehicles, results for vehicle category indicate a 388% increase in the odds of giving a response that indicates higher levels of pedestrian injury severity for heavy vehicles. While, for other type of vehicles, the results shows a 57% decrease in the odds of giving a response that indicated higher levels of pedestrian injury severity.

**Table 10:** Estimates of OR for the variable: `Alcohol`.

| `Alcohol` (g/L) | OR | $p$-value | LB | UP |
|---|---|---|---|---|
| $\leq 0.2$ | - | - | - | - |
| $]0.2, 0.5[$ | 1.55 | 0.06 | 0.98 | 2.42 |
| $[0.5, 1.0[$ | 1.32 | 0.01 | 1.06 | 1.64 |
| $\geq 1$ | 1.71 | 0.00 | 1.20 | 2.43 |

In regard to driver's blood alcohol content, when compared to a level below 0.2 g/L, a level between $]0.2, 0.5[$ g/L shows a 55% increase in odds to give a response that indicated higher levels of pedestrian injury severity. Surprisingly, a level between 0.5 g/L and 1.0 g/L shows an increase of just 32%, which are lower than the levels below of 0.5 g/L. As expected, the highest increase belongs to alcohol levels above 1.0 g/L, which shows a 71% increase in the odds of giving a response that indicated higher levels of pedestrian injury severity.

**Table 11:** Estimates of OR for the variable: `Light`.

| `Light` | OR | $p$-value | LB | UP |
|---|---|---|---|---|
| Daylight | - | - | - | - |
| Night with light | 1.58 | 0.00 | 1.41 | 1.77 |
| Night without light | 2.73 | 0.00 | 2.33 | 3.20 |
| Dawn and twilight | 1.03 | 0.82 | 0.77 | 1.38 |

In respect to lighting conditions, compared to daylight, the results show that there is a 58% increase in the odds of giving a response that indicated higher levels of pedestrian injury severity of the accident that occurs at night with illumination. As could be expected at night without illumination the odds increase further to 173% giving a response that indicated higher levels of pedestrian injury severity.

**Table 12:** Estimates of OR for the variable: `DriverInjury`.

| `DriverInjuries` | OR | $p$-value | LB | UP |
|---|---|---|---|---|
| Unharmed | - | - | - | - |
| Injured | 0.49 | 0.00 | 0.41 | 0.58 |

Considering the severity of the driver's injury the results indicate that when the driver is injured there is a 51% decrease in the odds of giving a response that indicated higher levels of pedestrian injury severity when compared to accidents where the driver is unharmed.

Regarding the variable of pedestrian's actions, the results show that when the action is legal there is a 56% decrease in the odds of giving a response that indicates higher levels of pedestrian injury severity when compared to accidents where the action of pedestrian is illegal.

Considering the variable administrative region, using Lisboa as the basis for comparison, all other groups of regions show a decrease in the odds of giving a response that indicates higher levels of pedestrian's injuries. However, it is Porto (group 2), which shows the largest decrease. Conversely it is group 4 that shows the lowest.

In respect to pedestrian's age, using pedestrians under 20 years old as a base of comparison, the results shows an increase in the odds of 77% for pedestrians with age between [20, 29] years and a huge increase of 209% for pedestrians with age between [40, 49] years giving a response that indicate more serious levels of pedestrian injuries. In otherwise, pedestrians over 50 years old present a decrease in the odds of giving a response that indicates higher levels of pedestrian injury severity.

# Discussion

Throughout this analysis, three different approaches have been presented to find a model that meets the aims of

this work. From now on, to simplify the analysis, we designate the best model of bayesian networks by BN, the best decision tree by DT and the ordered logistic regression model by LR.

Based on performance measures, it is unclear which model is the best choice for this dataset. However, some considerations can be taken into account. LR model obtained globally the highest values of all measurements except AUC. By contrast, were LR and DT models that had smallest value of F-score for serious and fatal injuries. In general was BN model that obtained balanced values for the three levels of the pedestrian's injuries variable. Moreover the accuracy and AUC for BN, in comparison with the remaining, are an acceptable result. Notice that, although this is not a classification problem, the model has to be capable to state the reasons behind the causes of the pedestrian injury severity, correctly. Which means, that in this case, the true positive rate and the sensitivity for each level has more importance than accuracy. As a result, we are able to say that BN is the approach that builds a model which better fits this dataset.

Decision tree model identifies the road type, the administrative region, the crash location, the vehicle category and the pedestrian's age as the most critical factors for the analysis of the severity of pedestrian injury. According to this model, pedestrians between 20 and 29 years old and between 40 and 49 years old have a higher risk of serious or fatal injury. With regards to the type of road, is on national roads, highway, principal or complementary itinerary and other types of roads that accidents cause more serious injuries to pedestrians. This model also predicts that heavy vehicles are strongly associated with serious injuries of pedestrians. In respect to administrative regions, are the group 1 (Lisboa) and group 4 (Braga and Vila Real) the regions that presents higher risk of a pedestrians suffer a serious or even a fatal injury. Finally, crash or collision location is another critical variable associated with severity in pedestrian injuries, since this model estimates a probability of 61% of serious injuries in accidents which occurs in rural areas.

Regarding the results of logistic regression estimation, the vehicle category was the variable that present the largest OR. Heavy vehicles have an increase of 388% in the odds of giving a response that indicated higher levels of pedestrian injury severity when compared with light vehicles. Moreover, all administrative regions groups, when compared to Lisboa (the nation's capital), have a decrease in the odds of a response that indicate higher levels of pedestrian injury severity. Some road-related factors were also consider interesting. For instance, this model predicts that accidents occurring in highway, principal or complementary itinerary have an increase of 198% in the odds of giving a response that indicated higher levels of pedestrian injury severity when compared with accidents occurring in streets. Which can be explained by the fact that the speed limits of these roads are much higher compared to streets. Indeed all the road categories has an increase in the odds when compared to accidents that occur in the streets. In other way, regarding lighting conditions, accidents occurring during night without illumination have an increase of 173% in the odds of giving a response that indicated higher levels of pedestrian injury severity when compared to accidents occurring in daylight. Crash location in rural areas is another factor that has an increase in the odds of giving a response that indicated higher levels of pedestrian injury severity. Lastly, the pedestrians between 40 and 49 years old show an huge increase of 209% in the odds of giving a response indicating higher levels of pedestrian injury severity when compared with younger than 19 years old. Regards pedestrian's action, when comparing legal actions to illegal we have a decrease in the odds giving a response that indicated higher levels of pedestrian injury severity.

Bayesian network results were not so easy to interpret since we have hundreds of CPT for all combinations of the parents of pedestrian's injuries node. However, the analysis of the network shows that crash location, lighting conditions, driver's injuries, vehicle category, driver's blood alcohol content, pedestrian's age, pedestrian's action, administrative region and road type influence the severity of pedestrian injury. Moreover, the conditional probability tables allow us to support all the results obtained by decision tree and logistic regression models.

# Conclusion

In short, throughout the analysis were found several factors considered relevant for the analysis of the causes of pedestrian injury severity, common to all the estimated models. In fact, vehicle category, administrative region, road type and crash location were indicated in all the three approaches as being strictly related to serious or fatal pedestrian's injuries. On the other hand, the driver's blood alcohol content and lighting conditions were considered relevant only by BN and LR models.

Some of the key conclusions of the study are considered critical and should be prioritized in action and prevention plans to promote pedestrian safety. The absence of illumination is associated with injuries of greater severity. Successful campaigns to increase the use of reflectors by pedestrians are likely to have significant benefits to their safety. Reflectors are not only useful for reducing the probability of an accident in the first place, but also to reduce its severity. An adequate separation between the pedestrian traffic and the heavy vehicle lane is of extreme importance given the significantly greater severity in accidents involving pedestrians and heavy vehicles. To tackling the problem of accidents with pedestrians in highway, principal itinerary or complementary itinerary should be action taken in order to prevent pedestrian to move around this type of roads or build appropriate accesses. Moreover, large road sides must be provide to ensure that pedestrians from other accidents or breakdowns are safe. Drunk drivers

are another important issue. Not only from an accident frequency standpoint but as shown here, drivers with blood alcohol are strongly associated with more serious pedestrian accidents. Rigorous fines for the driver are recommended in order to discourage them to drive with any level of alcohol in blood.

A more extensive dataset, with more refined measures of pedestrian exposure to danger, including time distance walked, the number of cross streets and the numbers of pedestrians and vehicles at intersections can be an interesting topic in future research. However obtaining this data for large area analysis is a challenge. Moreover, this type of study based on police records does not contain the physical causes for the injury severity (actual impact speed, vehicle mass and movement of the pedestrian). Detailed data of this type can only be collected through crash testing using dummies or by computer simulations. However, a dataset with this type of information allows to determine the real causes of accidents involving pedestrians. Other pedestrian's information that is not available in the provided dataset and in which some studies have shown some relevance was the pedestrian's blood alcohol content. Since it was the bayesian network that showed better results, as future work could be try to apply another type of network model, such as neural networks. On the other hand can simply try to apply another type of learning algorithms to learn the bayesian network. Namely resort to simulated annealing or a genetic algorithm.

# References


[1] Autoridade Nacional de Segurança Rodoviária - *Portugal National Road Safety Authority* (2015). Relatório anual - vítimas a 30 dias.

[2] Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(06):729–741.

[3] Bédard, M., Guyatt, G. H., Stones, M. J., and Hirdes, J. P. (2002). The independent contribution of driver, crash and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention*, 34(06):717–727.

[4] Chang, L.-Y. and Wang, H.-W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, 38(05):1019–1027.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[6] Clifton, K. J., Burnier, C. V., and Akar, G. (2009). Severity of injury resulting from pedestrian-vehicle crashes: What can we learn from examining the built environment? *Transportation Research Part D: Transport and Environment*, 14(06):425–436.

[7] Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186. doi:10.1023/A:1010920819831.

[8] Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

[9] Kim, J.-K., Ulfarsson, G. F., Shankar, V. N., and Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. *Accident Analysis & Prevention*, 40(5):1695–1702.

[10] Kockelman, K. M. and Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis & Prevention*, 34(03):313–321.

[11] Milton, J. C., Shankar, V. N., and Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis & Prevention*, 40(01):260–266.

[12] Õna, J., Mujalli, R. O., and Calvo, F. J. (2011). Analysis of traffic accident injury severity on spanish rural highways using bayesian networks. *Accident Analysis & Prevention*, 43(01):402–411.

[13] Peng, H., Long, F., and Ding, C. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 03(02):185–205.

[14] Sze, N. N. and Wong, S. C. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. *Accident Analysis & Prevention*, 39(6):1267–1278.

[15] World Health Organization (2015). Global status report on road safety 2015. `http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/`.

[16] Yau, K. K. W., Lo, H. P., and Fung, S. H. H. (2006). Multiple-vehicle traffic accidents in hong kong. *Accident Analysis & Prevention*, 38(06):1157–1161.

[17] Zajac, S. S. and Ivan, J. N. (2003). Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural connecticut. *Accident Analysis & Prevention*, 35(03):369–379.

[18] Zong, F., Xu, H., and Zhang, H. (2013). Prediction for traffic accident severity: Comparing the bayesian network and regression models. *Mathematical Probelms in Engineering*, 2013.

**Table 13:** Description and explanation of each class of each variable of the dataset after discretization and classes regrouped.

| Variables | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| `Month` | 1 | January | 2 | February | 3 | March | 4 | April |
| | 5 | May | 6 | June | 7 | July | 8 | August |
| | 9 | September | 10 | October | 11 | November | 12 | December |
| `DaysWeek` | 1 | Business day | 2 | Weekend | | | | |
| `Hour` | 1 | 01h - 06h | 2 | 07h - 09h | 3 | 10h - 15h | 4 | 16h - 18h |
| | 5 | 19h - 24h | | | | | | |
| `RoadType` | 1 | Street | 2 | National Road (EN) | 3 | Highway (A), Principal Itenerary (IP) | | |
| | 4 | Other type of roads | | | | or Complementary Itenerary (IC) | | |
| `Region` | 1 | Lisboa | 2 | Porto | 3 | Aveiro | | |
| | 4 | Braga and Viana do Castelo | 5 | Bragança, Coimbra, Guarda, Viseu and Vila Real | | | | |
| | 6 | Castelo Branco, Leiria, Setúbal and Santarém | | | 7 | Évora, Beja, Faro and Portalegre | | |
| `Location` | 1 | Urban area | 2 | Rural area | | | | |
| `Weather` | 1 | Good weather | 2 | Adverse weather conditions (rain, hail, amongst others) | | | | |
| `Light` | 1 | Daylight | 2 | Night with illumination | | | | |
| | 3 | Night without illumination | | | 4 | Dawn or twilight | | |
| `Grip` | 1 | Clean and dry | 2 | Damp or wet | | | | |
| `Direction` | 1 | Two ways road | 2 | Only one direction road | | | | |
| `Intersection` | 1 | Outside intersections | 2 | Roundabout | 3 | Entry or exit lanes, connecting | | |
| | 4 | Crossroad or intersection | | | | road, rail way crossing | | |
| `SegmentType` | 1 | Straight | 2 | Curved | | | | |
| `TrafficLanes` | 1 | Right | 2 | Left | 3 | Central | | |
| `DriverAge` | 1 | $\leq$ 19 years | 2 | [20 - 29] years | 3 | [30 - 39] years | 4 | [40 - 49] years |
| | 5 | [50 - 59] years | 6 | [60-69] years | 7 | $\geq$ 70 years | | |
| `DriverGen` | 1 | Male | 2 | Female | | | | |
| `DriverInjuries` | 1 | Unharmed | 2 | Injured | | | | |
| `DriverAction` | 1 | Direction change | 2 | Regular driving | 3 | Others situations | | |
| `License` | 1 | With driving license | 2 | Without driving license, amongst others particular cases | | | | |
| `VehicleCat` | 1 | Light vehicle | 2 | Heavy vehicle | 3 | Vehicle two motorized wheels | | |
| | 4 | Others type of vehicles | | | | | | |
| `Alcohol` | 1 | $\leq$ 0.2 g/L | 3 | ]0.2 - 0.5[ g/L | 3 | [0.5 - 1[ g/L | 4 | $\geq$ 1 g/L |
| `PedestAge` | 1 | $\leq$ 19 years | 2 | [20-29] years | 3 | [30-39] years | 4 | [40-49] years |
| | 5 | [50-59] years | 6 | [60-69] years | 7 | $\geq$ 70 years | | |
| `PedestGen` | 1 | Male | 2 | Female | | | | |
| `PedestAction` | 1 | Illegal | 2 | Legal | | | | |
| `PedestInjuries` | 1 | Minor injury | 2 | Serious injury | 3 | Fatal injury | | |