Treball de Fi de Grau/Màster

Enginyeria en Tecnologies Industrials

Predicting The Severity Of Road Traffic Accidents In The City Of Barcelona

MEMÒRIA

Autor: Enric Reverter López
Director: Jordi Olivella Nadal

Convocatòria: Maig 2021

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Summary

The study of road traffic accidents and the adoption of measures to reduce them is one of the most important targets of the Sustainable Development Goals for 2030. In one of the major cities in EU, Barcelona, 205 people were badly injured with 30 of them having a fatal outcome and more than 10000 were mildly injured in 2015. In this project, accidents in Barcelona comprised between 2018 and 2019 are studied. It focuses on using machine learning techniques to analyze driver-injury severity. Predicting accident-related injury severity provides insights that benefit the prevention of casualties, which supposes a huge economic and social impact. That is, classifying existing casualties into different severity categories allows to explain the underlying patterns behind casualties. A random forest (RF) and classification tree (CART) approach is followed. The former is used to select the most important variables, which are then used as input for the latter; in this way, the variance is reduced and the classification rate is improved. Explanatory variables regarding driver, accident and vehicle characteristics, environmental status, and traffic density are used as input.

It is worth to mention the theoretical aspect of this project, the aim is to explain the causes of lower and higher injury severity in existing casualties rather than proposing a methodology to decrease the severity. Also, since multiple types of accidents and vehicles are taken into account the scope of the results is wider. Studies focused on single types and methodologies to prevent severe outcomes can be born from this project.

The whole code for this project can be accessed at:

eReverter/Predicting_The_Severity_Of_Road_Traffic_Accidents_In_The_City_Of_Barcelona: Final Degree Project at ETSEIB. (github.com).

Table of Contents

Table of Figures

Table of Tables

1. Preface

In this section, the origin of the project, the motivation and the requirements are presented.

1.1. Origin and Motivation

The origin of the project is a proposal published by Jordi Olivella, as well as the particular interest of the student in data analytics. To this day, road traffic accidents are a huge issue in our society and different studies have been published to prevent them and reduce its severity. Both traditional statistics and machine learning methods have been applied to provide insights in road traffic accidents, although the former have fallen behind. Open Data BCN has provided public access to the log of casualties maintained by the police, where each accident is described by its injury severity level. However, analysis from this data cannot be found, at least publicly. Thus, the opportunity to analyze it with non-traditional methods has been harnessed.

The interest in this topic was born from the contents learned in Informàtica (Computer science), *Estadística* (Statistics), *Optimització I Simulació* (Optimization and Simulation), and *Projecte* (Data Science with Python) during the degree. Another influential factor has been the curiosity to understand the world from data; that is, the ability to transform unreadable data to actual insights.

1.2. Requirements

In order to accomplish this project, some knowledge regarding statistics and programming are a must. The whole project has been coded with *R,* which altogether with *Python* is considered one of the best tools for data analysis. However, the most important requirement is the knowledge of how to develop a Data Science project. All the research involved, preprocess, explore and visualize the data, which methods are the most fit for the task, how to properly evaluate its performance, and how to extract the important insights from the results. Also, it comes without saying that a proficiency level of English is required since a huge amount of research is necessary.

2. Introduction

Road traffic accidents are one of the major causes of death over the world. More than half the deaths worldwide in 2016 were due to the top 10 causes, which include traffic ones. Road injuries alone killed 1.4 million people in 2016 (World Health Organization (WHO), 2016). In the Member States of the European Union about 25600 people died and more than 1.4 million got injured in the same year. (Europen Road Safety Observatory (ERSO), 2018). Also, the deaths from road injuries have not decreased below 1 million during the last decades (Institute for Health Metrics and Evaluation (IHME), 2017). The problem is serious enough that one of the targets which was established by all United Nations members' states back in 2015 is to reduce traffic injuries (United Nations (UN), 2015). In one of the major cities in EU, Barcelona, 205 people were badly injured with 30 of them having a fatal outcome and more than 10000 were mildly injured in 2015 (Dirección General de Tráfico (DGT), 2016). Causes of accidents and its victims injury severity are of special concern for researchers in the field of traffic safety, as having better insights on it would not only benefit the prevention of such, but also reduce its severity.

The city hall of Barcelona publishes information from public organisms, allowing anyone its access so it can be used by people and entities for the common good. In fact, information regarding all the road traffic accidents that happen in the city is updated every year. Five different datasets that describe accidents can be found, all of them related by a record code and managed by the local police. Each one of them contains different data from drivers' characteristics, vehicle characteristics, type of accident, cause of accident and injury severity. To this day, analysis of this particular data has not been found and so the objective of this project is achieving that.

There are two main type of analysis that can be performed on road traffic accidents; first, accident forecasting where the objective is predicting the number of future casualties, which can be useful to identify hot spots and asses' risks; and second, accident classification by severity, where the objective is classifying existing casualties to its associated known severity, which is used to assess the patterns of severity in accident outcomes. Given the available data, the second type is done in order to provide an insight into the problem that supposes road traffic accidents.

Reducing the severity of injuries from traffic accidents is one of the most effective means to improve highway safety. In order to reduce the number of people killed and/or injured in traffic accidents, many research studies have been conducted to identify the risk factors that can significantly influence the injury outcomes of traffic accidents (Chang & Wang, 2006). Predicting accidents injury severity provides insights that benefit the prevention of casualties, which supposes a huge economic and social impact. The worse the casualty the more expensive is to handle it. Also, fatal road traffic accidents concern job losses, loss of amenity and a critical impact on the functioning of the involved family.

The task of accurate injury severity prediction in road traffic accidents is difficult mainly due to the complex nature of traffic accidents. Not only are accidents rare events, but also very imbalanced regarding the level of severity. A great number of existing studies on this problem have used classical statistical methods, and often only considering descriptive measures, which leads to bad performance since the underlying relationships cannot be studied. A wide number of these published studies about traffic crash accidents have used parametric methods to get results. However, these methods limit the findings due to data-related issues such as: over-dispersion, under-dispersion, time-varying explanatory variables, low sample-means and size, crash-type correlation, under-reporting of crashes, omitted-variables bias, and issues related to functional form and fixed parameters (Lord & Mannering, 2010). In order to address those issues, non-parametric models have been gaining ground and being widely used: classification and regression trees (CART), support vector machines (SVM) and neural networks (NN) are some of the most common.

A summary of some methods can be depicted in Table 1. One of the simplest methods to address the classification problem is logistic regression. The predicted parameters give inference about the importance of each feature, so it can be used to get insights on the relationships between the features. The computational power required is very low due to its simplicity. Also, multi-class classification can be approached by multinomial models. However, on high dimensional datasets it can be easily over-fitted and non-linear problems cannot be solved since it has a linear decision surface. Overall, logistic regressions are used to compare the performance of other methods. Given the fact that logistic regression cannot handle nonlinear data, they are not used in this project.

METHODS	ADVANTAGES	DISADVANTAGES
Logistic Regression	- Simple algorithm - Inference about the importance of each feature - Can handle multi-class problems - Shows probabilistic explanation of the results	- Issues arise in high dimensional datasets - Nonlinear problems cannot be solved
Neural Network	- Capacity to work with missing information - Handles nonlinear data and multi-class problems - Can outperform almost every other machine learning method	- Black-box: cannot be understand how the algorithm got to the prediction - Requires huge amounts of data - More computationally expensive than other models
Support Vector Machine	- Powerful algorithm - Can be extended to multi-class classifications - Manage cases where the number of dimensions is higher than the number of variables	- Black-box - Computationally expensive - Probabilistic explanation of the classification is not provided
Classification Tree	- Easy to interpret - Does not require much preprocess of the data - Provides probabilistic explanation of the results	- Easily over-fitted - Suffers from high variance
Random Forest	- Robust algorithm with no over-fitting - Low variance	- High complexity - Results cannot be interpreted with ease

	- Manage cases where the number of dimensions is higher than the number of variables	

Table 1. Advantages and disadvantages of different machine learning algorithms.

Neural networks are like multiple little logistic regressions stacked one over another. One of the advantages of neural networks is the capacity of working with incomplete information. The performance will depend on the importance of the missing data, unlike other machine learning methods which perform poorly when data is missing. Also, they work perfectly with nonlinear data, as seen in image recognition. It has the capability to outperform every machine learning algorithm, but this comes at some expenses. First, it is a black-box, it is close to impossible to understand how the method came up with a given classification. Second, the amount of data required is huge, to the point that millions of labeled samples might be necessary. Third, they are much more computationally expensive compared to any other methods. Due to these disadvantages, neural networks are not used, as one of the objectives is understanding the patterns behind road traffic accidents.

Another powerful algorithm is known as support vector machine, which is also closely related to logistic regressions. Although it is intended for binary classes, it can be extended to multi-class problems with some extensions. Contrary to logistic regressions, their performance on high-dimensional and nonlinear data is not an issue. It can even manage cases where the number of dimensions is higher than the number of samples. However, it has similar disadvantages to neural networks, although its computational complexity is not as high. They are a black-box, it is not possible to understand how it came up with a certain classification. In fact, it does not even provide probabilistic explanation of the classification. Since support vector machine has a veiled performance, where the impacts of contributing factors cannot be accessed, it has not been used in this project. However, it is worth to mention that it is possible to perform a sensitivity analysis in order to illustrate variable influence on driver injury outcomes (without combinations on factors).

Tree-based methods are simple and useful for interpretation. Classification trees are simple methods which do not require normalization nor scaling of data. They can work with nonlinear data. Also, unlike other algorithms, the model is very intuitive and can be easily explained. The probabilistic explanation of the classifications is given. However, they suffer from a lot of variance; that is, a small change in the input data might result in a whole different model. Moreover, over-fitting a tree is really easy.

Random forests (RF) are a combination of many trees, in this way, the variance is reduced and overfitting is prevented. Same as in decision trees, feature scaling is not required and nonlinear data is not an issue. Similar to support vector machines, cases where the number of dimensions is higher than the number of features can be handled. Nonetheless, since the number of trees required is usually high, the complexity of the algorithm is large. Also, it is too complicated to understand the reason behind a classification, since the relationships between features cannot be easily displayed.

For these reasons, the followed approach to the problem involves RF and CART methods. Both methods have been widely used to predict accident injury severity. CART can be easily interpreted and thus the underlying relationships between explanatory variables can be extracted. However, as this algorithm suffers from a lot of variance, RF is previously applied in order to select the most important variables; in this way, the variance is reduced and the classification rate can be improved.

2.1. Objectives and Scope

The available data limits how far the problem can be analyzed. Thus, the main objective of this study is to classify the existing accidents between the different categories of injury severity with a decent performance. Also, the factors and underlying relationships that contribute to the increased severity of an injury in a road traffic accident can be analyzed as a result from the former. It is worth to mention that accidents are not being predicted but classified between different levels of severity; that is, the results describe the combinations of factors that influence the injury once the casualty has already happened. In order to do that, other objectives have to be met beforehand. First, the data has to be retrieved and cleaned until it can be properly used for the different analysis. Also, it is necessary to do a process of feature selection previous to the RF and CART. Finally, the models must be tuned to improve the performance before analyzing the results.

The study begins with the literature review of previous papers, where methodologies applied by different researchers is examined. This is followed by the data description, where multiple datasets are described and combined into one. Then, the combined dataset is briefly analyzed and feature selection is explained; that is, correlations between explanatory variables are analyzed in order to drop the redundant ones. Also, the contingency table between explanatory variables and the injury severity level is interpreted in order to observe which variables seem to be the most significant. This is later contrasted with the results from the models. Next, the RF and CART methodologies are explained in detail, as well as the issues with class imbalance, the best performance metrics, and the workflow used. This is followed by the variable importance selection resulted from the RF. These selected features are then used as input for the CART models. Then, the results from the classification trees are presented and analyzed before discussing its performance. These results are discussed and compared with the ones expected from the exploratory analysis. Finally, the conclusions and limitations are described.

3. Literature Review

Road traffic accidents have been widely investigated. This section shifts its focus on accident injury severity studies and the different methodologies applied. Non-parametric models have been increasingly used in recent years in order to overcome parametric models: Neural Networks (Delen, Sharda, & Bessonov, 2006), Tree methods ((Chang & Wang, 2006), (Chang & Chien, 2013)), SVM (Chen, Zhang, Qian, Rafiqul, & Tian, 2016),and combinations of multiple models to strengthen its weaknesses (Pillajo-Quijia, Arenas-Ramírez, González-Fernández, & Aparicio-Izquierdo, 2020), to name but a few. Even Bayesian networks have been applied (Oña, López, Mujalli, & J.Calvo, 2013). However, by no means are traditional approaches forgotten. Recent studies have used multinomial logit models ((Behnood & Mannering, 2017), (Li, et al., 2019)).

Behnood and Mannering (2017) used a random parameters logit model with heterogeneity in parameter means to study the effects of passengers on driver-injury severities. The data used is from police-reported crashes collected from Chicago over a nine-year period, from 2004 until 2012. A wide range of explanatory variables such as weather conditions, roadway, vehicle and driver characteristics were used. Three discrete injury severity levels are considered: no injury, minor injury and severe injury. Age and gender interactions were found significant.

Li et al. (2019) applied a mixed logit model and latent class model to examine rural single-vehicle crashes under raining conditions. A three-year crash dataset from South Central states is used, comprising 2012 to 2014. The final dataset contains information on contributing factors regarding characteristics of the crash, vehicle and driver. Injury severity is classified into five subtypes: fatal, incapacitating, visible, complaint and no apparent injury. Results show that factors such as curves, drugs and seat belt significantly increase the injury level.

Delen et al. (2006) used a series of artificial neural networks to analyze the causes of injury severity different levels. The data comprises observations over different areas of all United States within 1995 and 2005. The final dataset comprises information regarding accident, vehicle and driver characteristics. Five-scale injury levels are coded: no injury, possible injury, minor non-incapacitating injury, incapacitating injury and fatality. Results show that the combination of binary models outperform a model with a five-outcome category variable. Most results validated previous findings, although the authors are surprised to find that neither weather nor time of accident are significant.

Chang and Wang (2006) adjusted a CART model to establish the relationships between injury severity and crash factors. The 2001 accident data for Taipei is used. Explanatory variables regarding driver, vehicle, and accident characteristics are comprised in it. Three discrete severity levels are considered: fatality, injury and no-injury. The proposed model fails to identify fatality outcomes. Results show that pedestrians, motorcycle and bicycle riders are the most vulnerable. Collision type, circumstance and driver/vehicle interaction are also found significant.

Chang and Chien (2013) developed a non-parametric CART model to study truck-involved accidents. The data used comprises all the police-reported accident data in the Taiwan area between 2005 and 2006. Key variables about driver, vehicle and accident characteristics are provided in the dataset. Injury severity factor is split into three levels:

fatality, injury and no-injury. Results show that drinking-driving is the most important determinant for the injury severity of truck accidents on freeways.

Chen et al. (2016) used a CART model for variable selection and posterior support vector machine to examine driver injury severity in a rollover crash. A New Mexico two-year dataset, from 2010 to 2011, is used in the analysis. Explicit crash-level, vehicle and driver information are contained within the data. Injury severity is defined with three injury levels: property damage only, non-incapacitating injury, and incapacitating injury and fatality. Due to the SVM lack of capability to select significant variables, a CART model is first fit in order to make the selection. Results show that drinking-driving is the most significant cause of driver incapacitating and fatalities, as well as the seatbelt being the best protection.

Pillajo-Quijia et. Al (2020) selected the most important features using random forests and adjusted a CART model with them. A BLM and SVM are also modeled for comparison with the CART. Data over a period of nine years (2000-2008) from the DGT, in Spain, is used. Variables describe information regarding driver characteristics, vehicle characteristics, infrastructure, and environmental conditions. Injury severity is classified as either slightly injury or severely injured/killed. Results show the importance of: using the seatbelt, the psychophysical condition of the driver, and the injury location.

Oña et. Al (2013) used Bayesian networks to classify traffic accidents according to their injury severity. The dataset contains information regarding a three-year period, from 2003 to 2005. It is retrieved from the DGT, for rural highways in Granada, Spain. Features such as driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics, and atmospheric factors are described in said dataset. Injury severity is classified as either slightly injury or severely injured/killed. Based on the results, the worst-case scenario is probably due to a head-on or rollover accident, in a roadway without lightning, given the driver has between 18 and 25 years.

So far, only one study has been found with a similar content as this one. Albalate and Fernández-Villadangos (2010) studied motorcycle injury severity in Barcelona using a multinomial logistic regression. The used databased is retrieved from Barcelona's local police census (2002-2008). Information concerning demographic characteristics, environment and traffic conditions, primary causes of accidents, and regulatory measures are described in it. The dependent variable severity contains three increasing degrees of severity according to the database: non-severe, severe and fatal. Results show a negative relationship between traffic flow and injury severity on PTWs.

Ultimately, no studies have been found that consider the recent datasets given by the city hall of Barcelona. In this work, seven open access databases are retrieved to study injury severity inside the city. Random forests are used to select the most important features, followed by CART models to interpret the most significant factors (James, Witten, Haste, & Tibshirani, 2017). Model performance is evaluated for both methodologies. Information regarding driver, accident and vehicle characteristics, environmental status, and traffic density are taken into account to classify the injury severity.

4. Methodology

Once the background for this study has been analyzed, the methodology to follow can be briefly described. First, the data is retrieved from Open Data BCN, where information regarding public studies and administrations is published to be reused by anyone interested. Second, multiple datasets are merged in order to describe each accident by well-defined variables of interest. Then, the available features are described and analyzed. Such analysis has two objectives: understand the structure of the available variables and reduce the number of features. That is, drop the high correlated variables from the model in order to avoid redundant variables and improve the computational speed. Next, the methods used for sampling, modeling, and evaluating are discussed in detail. Afterward, different classifiers are trained depending on the number of classes to predict. RF methodology is used to select the best features for classifying between the given categories. Then, the most important features are used to model classification trees. Finally, the best CART models are examined in detail and its performance is evaluated to assess the applicability.

4.1. Data Description

This section is split into two sub-sections: retrieving and cleaning. In the former, a summary for each dataset can be depicted, where information regarding its contents and source is found. In the latter, the tweaks done in order to obtain the combined dataset are described. The structure of the final dataset can be observed at the start of the second sub-section, with all the transformations being justified after.

4.1.1. Retrieving the Data

Seven datasets from Open Data BCN have been accessed throughout the working, five of which are referred to the accident itself and interrelated by a record code. The remaining two datasets contain information regarding traffic density and meteorological status. It is worth to mention the accidents studied are all the ones comprised between 2018 and 2019 in the city of Barcelona, even though some of the datasets have records that go back to 2010. This is due to the fact that the traffic density started to be stored in October 2017. The summary of each accessed file can be depicted in the tables below.

Accidents

Dataset	All accident-related datasets
Source	Open Data BCN. Barcelona's City Hall Open Data Service. URL: https://opendata-ajuntament.barcelona.cat/data/es/organization/seguretat
Common attributes	● Record code. ● District and neighborhood. ● Location postal address and coordinates. ● Occurrence day and time. ● Kind of day (working or holiday). ● Local police working shift.
Time period	2010 onwards

Table 2. Common attributes of the accident-related datasets.

Dataset	**Elements**	**Specific attributes**

Accidents managed by the local police in the city of Barcelona	Accidents	● Number of dead, severely injured and mildly injured people. ● Number of vehicles involved.
Description of the accidents' handled by the police in the city of Barcelona causality.	Accidents	● Mediate cause of the accident (alcohol or no mediated cause).
Accidents managed by the Police in the city of Barcelona according to type	Accidents	● Type of accident (side-impact collision, head-on collision, head-on and side-impact collision, pedestrian hitting, animal hitting, fixed objects hitting, rollover, two-wheels vehicle fall, interior fall, stairs, route leaving and collision, route leaving and rollover, multiple, other, unknown).
People involved in accidents managed by the Police in the city of Barcelona	People that needed medical assistance	● When the accident is caused by a pedestrian behavior, the specific cause. ● Kind of vehicle involved. ● Age, gender and being driver, passenger or pedestrian. ● Kind of medical assistance needed.
Vehicles involved in accidents handled by the police in the city of Barcelona	Vehicles involved	● Type of vehicle. ● Make, model and color. ● Vehicle age.

Table 3. Specific attributes of the accident-related datasets.

Traffic density

Dataset	Traffic state information by sections of the city of Barcelona.
Source	Open Data BCN. Barcelona's City Hall Open Data Service. URL: https://opendata-ajuntament.barcelona.cat/data/es/dataset/trams
Elements	Observations (2/3 per hour)
Attributes	● Observation data and time. ● Street segment observed. ● State and previously forecasted state (0, non-available; 1, very fluid; 2, fluid; 3, dense; 4, very dense; 5, congested; 6, closed)
Time period	October 2017 and onwards.

Table 4. Attributes of the traffic density datasets.

Meteorological measures

Dataset	Measurements of the meteorological stations of the city of Barcelona. El Raval weather station.
Source	Open Data BCN. Barcelona's City Hall Open Data Service. URL: https://opendata-ajuntament.barcelona.cat/data/es/dataset/mesures-estacions-meteorologiques
Elements	Daily data

Attributes	● Maximum, minimum and average temperature. ● Average humidity. ● Accumulated precipitation. ● Average atmospheric pressure. ● Solar irradiation. ● Average wind speed and wind direction. ● Speed and direction of the maximum wind gusts.
Time period	1996 onwards

Table 5. Attributes of the meteorological datasets.

At this point, it is worth mentioning that the response variable, injury severity, is given by the kind of medical assist needed in the accident. It is an ordered factor with five levels: (1) “Ferit lleu: rebutja assistència sanitaria” [Slightly injured: refuses healthcare], (2) “Ferit lleu: amb assistència sanitaria en lloc d’accident” [Slightly injured: with medical assist in accident location], (3) “Ferit lleu: hospitalització fins a 24h” [Slightly injured: hospitalization up to 24h], (4) “Ferit greu: hospitalització superior a 24h” [Seriously injured: hospitalization for more than 24h], and (5) “Mort (dins 24h posteriors accident)” [Death (within 24h after the accident)]. In order to work with shorter descriptions, the degree of injury has been recoded respectively into: (1) No Injury, (2) Mild, (3) Moderate, (4) Severe, and (5) Fatal.

Following previous studies, in the case where there are multiple victims in a single accident, the injury severity of such casualty is determined according to the level of injury to the worst injury occupant (Chang & Wang, 2006). That is, if the output has multiple slightly injured victims, but one of them has a fatal outcome, the accident is immediately labeled as fatal. This mutation is described in the following section.

4.1.2. Cleaning the Data

In order to get a final merged database with which train and test the desired model, some tweaks are needed in each one of the datasets. It is worth to mention that some datasets are split by time: the accidents files are split by year and traffic density ones by month. The meteorological ones are split by stations, but each one contains all the historical data from past years until the present. This section is structured as follows: first, the features are described, whether they have been kept or not; second, the transformations computed to get them are explained. The variables descriptions can be depicted in the following tables:

KEPT FEATURES	DESCRIPTION
COD_EXPD	Code of record
NOM_DIST	Name of the district
NOM_DIA	Name of the day
DIA_TIPUS	Type of day
ANY	Year
NOM_MES	Name of the Month
TORN	Working shift by the police
INT_H	Interval of hours during a day
CAUSA_V	Cause relative to pedestrian
TIP_VEH	Type of vehicle

SEXE	Sex of the driver
TIP_ACC	Type of accident
CAUSA	Cause of the accident
DESC_VICT	Injury severity level
X	X coordinate as UTM
Y	Y coordinate as UTM
LONG	Longitude
LAT	Latitude
EDAT	Age of the driver
NUM_VICT	Number of victims
NUM_VEHS	Number of vehicles involved
AVG_ESTAT	Average traffic density
TM	Average daily temperature (ºC)
TX	Maximum daily temperature (ºC)
TN	Minimum daily temperature (ºC)
HRM	Average daily relative humidity (%)
PPT24H	Accumulated daily precipitation (mm)
HPA	Average daily atm (hPa)
RS24H	Daily global solar radiation (MJ / m2)
VVM10	Average daily wind speed (m/s)
DVM10	Average daily wind direction (º)
VVX10	Maximum daily wind gust (m/s)
DVX10	Direction of the maximum daily wind gust (º)
DATA	Short date and time

Table 6. Description of the final features in the combined dataset.

DROPPED FEATURES	**DESCRIPTION**
COD_DIST	Code of the district
COD_BARR	Code of the neighborhood
NOM_BARR	Name of the neighborhood
COD_CARR	Code of the street
NOM_CARR	Name of the street
NUM_POST	Postal code
DIA_SETM	Day of the week in number formatting
MES	Month in number formatting
DIA	Number of the day within the month
HORA	Hour of the day
MODEL_VEH	Vehicle model
MARCA_VEH	Vehicle brand
COL_VEH	Vehicle color
TIP_CARNET	Type of driving license
T_CARNET	Time of driving license
NUM_OBS	Number of observations of traffic density in an hour
ID_TRAM	Id of the section road being measured
ESTATACTUAL	Current state of traffic
ESTATPREVIST	Forecasted state of traffic

Table 7. Description of the dropped features from the datasets.

Accident-related Data

Each year has its own dataset, so, for the accident’s dataset, ten files are going to be used (five for 2018 and five for 2019).

To start with, the same group of datasets are to be merged to get a single file containing both years. However, the features between different years are not the same in some cases, or are labeled as different ones, which implies that a simple binding is not enough. In other words, the datasets have to be manually checked. For example, the dataset with the injury severity factor of 2019 has information regarding the motives of the victim's trip, but the 2018 one does not. Thus, this variable is deprecated. This is done for every file. It should be noted this issue is aggravated in past years.

Once the ten datasets are grouped by type in five different datasets, the common features (except for the record code) are stripped of to build a single dataset which has all these elements. Hence, there are six datasets to work on. The features in each dataset can be depicted in the following lists.

Common features

COD_EXPD	COD_DIST	NOM_DIST	COD_BARR	NOM_BARR
COD_CARR	NOM_CARR	NUM_POST	NOM_DIA	DIA_SETM
DIA_TIPUS	ANY	MES	NOM_MES	DIA
TORN	HORA	X	Y	LONG
LAT				

People involved in the accident

COD_EXPD	CAUSA_V	TIP_VEH	SEXE	EDAT
TIP_PERS	DESC_VICT			

Type of accident

COD_EXPD	TIP_ACC			

Vehicles involved in the accident

COD_EXPD	TIP_VEH	MODEL_VEH	MARCA_VEH	COL_VEH
TIP_CARNET	T_CARNET			

Cause of the accident

COD_EXPD	CAUSA			

Number of victims and vehicles involved

COD_EXPD	NUM_MS	NUM_LL	NUM_LG	NUM_VICT
NUM_VEHS				

As of now, some datasets have multiple observations for the same accident. For example, the involved vehicles dataset has one observation for each vehicle affected in the casualty. However, for this study, each accident should be characterized by a single one. Henceforth, the focus shifts to obtain a single observation for each accident. At present, only the datasets regarding common information, cause and counts have unique observations, so the others need a tweak to be grouped by the code of record.

People involved in the accident

As stated before, the injury severity is determined according to the level of injury to the worst victim. The main idea is to assign a single level of injury severity to each accident, in such a way that for multiple victims the worst injury becomes the defining one. This is done by looking at all the victims in the accident and linking the said record to the worst level of the ordered factor.

Also, this dataset has information of drivers, passengers and pedestrians, but only the ones labeled as drivers are important for the study. What is being studied is the severity of the accident, not the victims, so the important factors are the ones that define the accident. For example, the age and sex are relevant for the drivers, not the other victims. That is, all the observations labeled as passengers or pedestrians have been dropped. Also, some records contain data about more than one driver in separated rows (~1000), which can lead to duplicated observations (except for the driver features). In order to avoid this issue, which can later on influence the model, only the information of one driver is kept by accident. The driver to be kept is chosen randomly to avoid bias. The final number of observations is: 15124.

Type of accident

Each casualty is defined by all the different types of accident that originated it. For example, if three vehicles were involved in an accident, which started with a rear-end collision and ended with a rollover, both rear-end and rollover are accounted for the same accident in different observations. Thus, all the accidents labeled with different types, have been classified as “Multiple”. Therefore, each code is associated with a single type of accident. The final number of observations is: 19954.

Vehicles involved in the accident

The same issue arises with this dataset. However, it is impossible to use it as the information given about all the involved vehicles cannot be linked to the main drivers. There is no identifier between vehicles and drivers. For example, if two cars are involved, but only one of them is defined in the people dataset, trying to guess which one of the two is the one defined is futile. Consequently, this dataset is not used.

Henceforth, each dataset has unique observations for each causality and can be merged by its code. However, before merging, data dealing with traffic density and weather is going to be included at the common dataset.

Traffic Density Data

Regarding this database, each month has its own file, so, for the traffic density dataset, twenty-four files are going to be used (twelve for 2018 and twelve for 2019). The features are the same in all the files:

idTram	data	estatActual	estatPrevist	

Therefore, the files can be easily joined by binding all of them together. With all the files joined, there are 55.159.214 observations in total (81% from 2019 and 19% from 2018). From here, in order to get the average traffic density by hour the next transformations are computed: first, if the state is non-available or closed, the observation is dropped. The moving average is calculated for each hour grouped by idTram; thus, the variable is converted to continuous. Finally, as the link between the road section and accident coordinates would be complex, the moving average is computed by the whole city of Barcelona. Therefore, the final number of observations drops to 17.352 and the association between the accidents datasets and the traffic density can be done by the date and time of the accident. Note that two years have 17.520 hours, which means 168 hours do not have available data regarding traffic density. The final features of the traffic density dataset are:

DATA	AVG_ESTAT	NUM_OBS		

Meteorological Measures Data

There are four stations where measures are taken in a daily basis. Each station (Fabra, Raval, Universitària and Zoo) has its own records. The documentation started back in 1996, 2006, 2008 and 2006 respectively. The Zoo station has way less measures than the others, so it is not going to be used in the meteorological dataset. The observations are stored as per day, so the combined dataset of the three stations will have the mean as the final measure every day. That is, the max temperature of a random day is the mean of the max temperatures stored by each station. Unlike the traffic density dataset, there are complete cases of observations.

The features of the combined meteorological dataset are:

DATA	TM	TX	TN	HRM
PPT24H	HPA	RS24H	VVM10	DVM10
VVX10	DVX10			

Combined Dataset

Once all the observations are cleaned, the merging can take place. First, the average state of the road is added to the common features of the accidents' dataset by the date and hour of observations. Thus, each accident can now be described by the state of the traffic over the city. Then, the same procedure is used to merge the meteorological measures with the common dataset, but instead of doing it by date and hour, only date is available for the meteorological measures. It is worth to mention the fact that it is not known if it was raining where and when the casualty happened. Finally, all the accidents datasets, which only contain unique observations for each record, are merged by the expedient code. The final number of complete observations is: 14958.

Some features are added to the combined dataset. First, instead of having a factor with 24 levels depicting the hours, a new factor with 6 levels is created. This tweak is also done in existing literature (Pillajo-Quijia, Arenas-Ramírez, González-Fernández, & Aparicio-Izquierdo, 2020). Reducing the number of levels of a feature is key for the computational complexity. The hours are grouped by intervals of 4: [06:00,09:59], [10:00,13:59], [14:00,17:59], [18:00,21:59], [22:00,01:59] and [02:00,05:59]. The interval from 6h to 10h is considered the morning period which comprises the sunrise; from 10h to 14h the shift between morning and afternoon is included; the span between 14h and 18h is the afternoon; between 18h and 22h the evening is considered; from 22h to 2h midnight is enclosed; finally, between 2h and 6h represents the morning until dawn.

Also, the feature labeled as type of day contains only the category "Laboral", even if the accident took place in a holiday according to the calendar of Barcelona. Therefore, it is mutated given the next conditions: Weekday from 02:00 Monday until 21:59 Friday; Weekend from 22:00 Friday to 01:59 Monday; and Holiday according to the working calendar of Barcelona. The working calendar can be found in the city hall website.

Before starting the exploratory analysis, it has been noticed that some ages might have not been recorded correctly. Apparently, a great number of drivers with vehicles such as cars or vans have less than 18 years old, which is impossible. Thus, those observations have been imputed with the mean of the remaining observations.

Finally, as the different databases are written in Catalan, all of them are translated to English in the interest of the analysis and working. All the code regarding the commented procedures can be found in the annex.

4.2. Exploratory Analysis

This section comprises the first analysis done to the current data. The scope of it is to better understand the given data and, essentially, reduce the number of features that do not contribute to the outcome. That is, the feature selection step. First, the final features selected for the models are presented, both numerical and categorical. The former ones are summarized by basic statistical measures, such as the mean and median, being grouped by the level of injury severity. The latter ones are depicted in a contingency table. The contents of the depicted tables are also discussed in order to extract a first idea of what to expect from the results. Next, the correlation between all the features is computed. The tests used are: Pearson, Spearman, Chi-squared, Cramer's V and ANOVA. These methods justify the decision of dropping specific features that are already being described by the others. Thus, the final features can be depicted in the tables below.

NUMERICAL FEATURES	INJURY	OBS.	MEAN	SD	MEDIAN	MIN	MAX	RANGE
EDAT	No Injury	365	31.5	10.6	32.0	4.0	82.0	78.0
	Mild	3888	31.0	10.8	32.0	1.0	85.0	84.0
	Moderate	10350	30.8	10.9	32.0	1.0	86.0	85.0
	Severe	320	32.3	11.5	32.0	3.0	79.0	76.0
	Fatal	35	33.2	9.8	32.0	8.0	48.0	40.0
AVG_ESTAT	No Injury	365	2.0	0.4	2.1	1.1	2.6	1.5
	Mild	3888	2.0	0.3	2.1	1.1	2.6	1.6
	Moderate	10350	2.0	0.3	2.1	1.0	2.7	1.6
	Severe	320	1.9	0.4	2.0	1.1	2.6	1.5
	Fatal	35	1.6	0.5	1.4	1.1	2.4	1.3
TM	No Injury	365	16.8	6.2	15.2	4.9	30.7	25.8
	Mild	3888	16.7	6.1	15.4	1.6	31.3	29.7
	Moderate	10350	16.9	6.0	16.0	1.6	31.3	29.7
	Severe	320	17.5	6.2	17.4	5.7	30.7	25.0
	Fatal	35	18.4	6.3	19.8	7.2	30.7	23.5
HRM	No Injury	365	66.0	12.1	66.0	30.0	93.3	63.3
	Mild	3888	66.4	11.5	66.5	25.0	95.3	70.3
	Moderate	10350	66.4	11.4	66.3	25.0	95.3	70.3
	Severe	320	65.7	11.1	66.3	30.0	91.0	61.0
	Fatal	35	68.1	11.7	70.0	30.3	87.7	57.3
PPT24H	No Injury	365	2.7	8.1	0.0	0.0	43.9	43.9
	Mild	3888	2.2	8.8	0.0	0.0	99.1	99.1
	Moderate	10350	2.2	8.7	0.0	0.0	99.1	99.1
	Severe	320	1.8	7.7	0.0	0.0	64.0	64.0
	Fatal	35	1.7	4.7	0.0	0.0	22.6	22.6
HPA	No Injury	365	994.6	7.2	995.3	972.1	1,012.6	40.5
	Mild	3888	994.9	7.0	995.4	970.2	1,014.3	44.2
	Moderate	10350	995.2	6.9	995.5	970.2	1,014.3	44.2
	Severe	320	995.4	6.1	995.8	973.8	1,011.0	37.1
	Fatal	35	993.8	4.9	994.7	983.1	1,002.8	19.7
RS24H	No Injury	365	16.2	8.5	15.5	0.8	31.1	30.3

	Mild	3888	16.2	8.3	15.8	0.4	31.2	30.8
	Moderate	10350	16.2	8.3	15.8	0.4	31.2	30.8
	Severe	320	17.1	8.2	17.2	0.8	31.1	30.3
	Fatal	35	14.9	8.3	14.1	1.4	29.4	28.0
VVM10	No Injury	365	2.8	0.8	2.6	1.1	6.2	5.2
	Mild	3888	2.8	0.9	2.7	1.1	7.0	6.0
	Moderate	10350	2.8	0.9	2.7	1.1	7.0	6.0
	Severe	320	2.8	0.9	2.6	1.1	7.0	6.0
	Fatal	35	2.9	1.2	2.7	1.5	7.0	5.6
DVM10	No Injury	365	218.2	78.2	232.3	34.0	348.3	314.3
	Mild	3888	209.6	82.9	230.0	15.7	348.3	332.7
	Moderate	10350	208.5	83.5	229.0	15.7	348.3	332.7
	Severe	320	209.6	81.1	226.8	15.7	348.3	332.7
	Fatal	35	193.5	88.4	199.7	52.3	340.7	288.3

Table 8. Descriptive statistics from the numerical variables by injury severity level.

The age (EDAT) does not seem to impact the outcome of the casualty. It can be observed an increment in injury severity along with a decrease in traffic density (AVG_ESTAT). Regarding the average temperature (TM), the injury seems to be worse with a slight increase in it. The contrary can be said regarding the cumulative precipitation (PPT24H), the lower it is the worse the severity, which may indicate that when it rains drivers are more careful. Also, lower values of average daily radiation (RS24H) seem to be more frequent in fatal outcomes. The average wind speed (VVM10) seems to be trivial, although the direction (DVM10) slightly changes as the severity increases. Not much can be observed in relation to the average daily relative humidity (HRM) and pressure (HPA). Overall, nothing can be affirmed with robustness from looking at the descriptive table.

CATEGORICAL FEATURES	**NO INJURY (N=365)**	**MILD (N=3888)**	**MODERATE (N=10350)**	**SEVERE (N=320)**	**FATAL (N=35)**	**OVERALL (N=14958)**
CAUSA_V						
CROSS OUTSIDE PEDESTRIAN CROSSING	1 (0.8%)	29 (22.3%)	90 (69.2%)	7 (5.4%)	3 (2.3%)	130 (0.9%)
DISOBEY OTHER SIGNS	0 (0.0%)	1 (16.7%)	5 (83.3%)	0 (0.0%)	0 (0.0%)	6 (0.0%)
DISOBEY TRAFFIC LIGHT	2 (1.0%)	38 (19.0%)	143 (71.5%)	14 (7.0%)	3 (1.5%)	200 (1.3%)
NOT THE PEDESTRIAN FAULT	360 (2.5%)	3789 (26.1%)	10022 (69.1%)	295 (2.0%)	28 (0.2%)	14494 (96.9%)
OTHERS	2 (1.8%)	28 (25.5%)	78 (70.9%)	2 (1.8%)	0 (0.0%)	110 (0.7%)
WALK ALONG THE ROAD	0 (0.0%)	3 (16.7%)	12 (66.7%)	2 (11.1%)	1 (5.6%)	18 (0.1%)
TIP_VEH						
ARTICULATED BUS	0 (0.0%)	0 (0.0%)	2 (66.7%)	0 (0.0%)	1 (33.3%)	3 (0.0%)
BIKE	23 (2.2%)	260 (24.9%)	735 (70.5%)	24 (2.3%)	1 (0.1%)	1043 (7.0%)

BUS	1 (4.8%)	5 (23.8 %)	15 (71.4%)	0 (0.0%)	0 (0.0%)	21 (0.1%)
CAR	91 (4.5%)	565 (28.0 %)	1336 (66.3%)	19 (0.9%)	5 (0.2%)	2016 (13.5%)
CONSTRUCTION MACHINERY	0 (0.0%)	1 (33.3 %)	2 (66.7%)	0 (0.0%)	0 (0.0%)	3 (0.0%)
MOPED	22 (1.5%)	380 (26.4 %)	1005 (69.9%)	30 (2.1%)	0 (0.0%)	1437 (9.6%)
MOTORCYCLE	192 (2.0%)	2409 (25.5 %)	6594 (69.7%)	235 (2.5%)	25 (0.3%)	9455 (63.2%)
OFF-ROAD	1 (6.2%)	9 (56.2 %)	6 (37.5%)	0 (0.0%)	0 (0.0%)	16 (0.1%)
OTHER MOTOR VEHICLES	1 (6.7%)	3 (20.0 %)	10 (66.7%)	0 (0.0%)	1 (6.7%)	15 (0.1%)
OTHER NON-MOTOR VEHICLES	0 (0.0%)	0 (0.0%)	6 (100.0%)	0 (0.0%)	0 (0.0%)	6 (0.0%)
PERSONAL MOBILITY DEVICE (MOTOR)	9 (2.1%)	106 (24.5 %)	310 (71.8%)	7 (1.6%)	0 (0.0%)	432 (2.9%)
PERSONAL MOBILITY DEVICE (NO MOTOR)	1 (2.3%)	7 (16.3 %)	34 (79.1%)	1 (2.3%)	0 (0.0%)	43 (0.3%)
QUADRICYCLE < 75 CC	0 (0.0%)	2 (66.7 %)	1 (33.3%)	0 (0.0%)	0 (0.0%)	3 (0.0%)
QUADRICYCLE > 75 CC	0 (0.0%)	1 (100.0 %)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
RIGID TRUCK <= 3.5 TONS	1 (3.6%)	10 (35.7 %)	17 (60.7%)	0 (0.0%)	0 (0.0%)	28 (0.2%)
RIGID TRUCK > 3.5 TONS	0 (0.0%)	2 (25.0 %)	6 (75.0%)	0 (0.0%)	0 (0.0%)	8 (0.1%)
TAXI	9 (4.4%)	64 (31.5 %)	127 (62.6%)	2 (1.0%)	1 (0.5%)	203 (1.4%)
TRUCK TRACTOR	1 (14.3%)	2 (28.6 %)	4 (57.1%)	0 (0.0%)	0 (0.0%)	7 (0.0%)
UNKNOWN	0 (0.0%)	2 (66.7 %)	1 (33.3%)	0 (0.0%)	0 (0.0%)	3 (0.0%)
VAN	13 (6.0%)	60 (27.9 %)	139 (64.7%)	2 (0.9%)	1 (0.5%)	215 (1.4%)
SEXE						
FEMALE	114 (2.6%)	1094 (25.2 %)	3063 (70.7%)	60 (1.4%)	2 (0.0%)	4333 (29.0%)
MALE	251 (2.4%)	2794 (26.3 %)	7287 (68.6%)	260 (2.4%)	33 (0.3%)	10625 (71.0%)
NOM_DIST						
CIUTAT VELLA	14 (2.1%)	176 (26.5 %)	456 (68.8%)	14 (2.1%)	3 (0.5%)	663 (4.4%)
EIXAMPLE	102 (2.1%)	1346 (27.8 %)	3286 (67.9%)	103 (2.1%)	4 (0.1%)	4841 (32.4%)
GRÀCIA	27 (3.9%)	174 (25.2 %)	480 (69.5%)	8 (1.2%)	2 (0.3%)	691 (4.6%)

HORTA-GUINARDÓ	30 (3.1%)	206 (21.4%)	697 (72.5%)	26 (2.7%)	3 (0.3%)	962 (6.4%)
LES CORTS	30 (2.9%)	305 (29.1%)	689 (65.8%)	22 (2.1%)	1 (0.1%)	1047 (7.0%)
NOU BARRIS	14 (2.0%)	147 (21.5%)	511 (74.6%)	8 (1.2%)	5 (0.7%)	685 (4.6%)
SANT ANDREU	23 (2.5%)	206 (22.5%)	657 (71.7%)	26 (2.8%)	4 (0.4%)	916 (6.1%)
SANT MARTÍ	43 (2.3%)	432 (23.0%)	1344 (71.6%)	51 (2.7%)	6 (0.3%)	1876 (12.5%)
SANTS-MONTJUÏC	44 (2.7%)	446 (27.8%)	1083 (67.5%)	29 (1.8%)	2 (0.1%)	1604 (10.7%)
SARRIÀ-SANT GERVASI	38 (2.3%)	450 (26.9%)	1147 (68.6%)	33 (2.0%)	5 (0.3%)	1673 (11.2%)
DIA_TIPUS						
HOLIDAY	6 (2.2%)	62 (22.9%)	196 (72.3%)	6 (2.2%)	1 (0.4%)	271 (1.8%)
WEEKDAY	294 (2.5%)	3059 (25.8%)	8243 (69.6%)	234 (2.0%)	20 (0.2%)	11850 (79.2%)
WEEKEND	65 (2.3%)	767 (27.0%)	1911 (67.4%)	80 (2.8%)	14 (0.5%)	2837 (19.0%)
ANY						
2018	169 (2.3%)	1987 (26.8%)	5080 (68.5%)	170 (2.3%)	15 (0.2%)	7421 (49.6%)
2019	196 (2.6%)	1901 (25.2%)	5270 (69.9%)	150 (2.0%)	20 (0.3%)	7537 (50.4%)
NOM_MES						
APRIL	29 (2.3%)	347 (27.4%)	855 (67.6%)	30 (2.4%)	4 (0.3%)	1265 (8.5%)
AUGUST	22 (2.3%)	227 (23.3%)	688 (70.6%)	33 (3.4%)	4 (0.4%)	974 (6.5%)
DECEMBER	33 (2.8%)	283 (23.7%)	856 (71.8%)	18 (1.5%)	3 (0.3%)	1193 (8.0%)
FEBRUARY	22 (1.9%)	327 (28.2%)	783 (67.6%)	24 (2.1%)	2 (0.2%)	1158 (7.7%)
JANUARY	28 (2.2%)	351 (27.7%)	858 (67.6%)	29 (2.3%)	3 (0.2%)	1269 (8.5%)
JULY	39 (2.8%)	352 (25.7%)	943 (68.8%)	33 (2.4%)	3 (0.2%)	1370 (9.2%)
JUNE	22 (1.8%)	315 (25.9%)	852 (70.1%)	24 (2.0%)	2 (0.2%)	1215 (8.1%)
MARCH	31 (2.3%)	360 (27.1%)	914 (68.9%)	21 (1.6%)	1 (0.1%)	1327 (8.9%)
MAY	40 (3.0%)	362 (26.9%)	913 (67.9%)	28 (2.1%)	2 (0.1%)	1345 (9.0%)
NOVEMBER	41 (3.3%)	322 (26.2%)	833 (67.7%)	30 (2.4%)	4 (0.3%)	1230 (8.2%)

OCTOBER	31 (2.1%)	352 (24.1%)	1049 (71.8%)	27 (1.8%)	3 (0.2%)	1462 (9.8%)
SEPTEMBER	27 (2.3%)	290 (25.2%)	806 (70.1%)	23 (2.0%)	4 (0.3%)	1150 (7.7%)
CAUSA						
ALCOHOL	30 (6.1%)	132 (26.8%)	314 (63.7%)	16 (3.2%)	1 (0.2%)	493 (3.3%)
BAD SIGNALING	0 (0.0%)	0 (0.0%)	4 (100.0%)	0 (0.0%)	0 (0.0%)	4 (0.0%)
DRUGS	3 (6.1%)	13 (26.5%)	24 (49.0%)	8 (16.3%)	1 (2.0%)	49 (0.3%)
METEOROLOGICAL FACTORS	0 (0.0%)	5 (38.5%)	8 (61.5%)	0 (0.0%)	0 (0.0%)	13 (0.1%)
ROAD CONDITIONS	3 (2.3%)	44 (33.3%)	83 (62.9%)	2 (1.5%)	0 (0.0%)	132 (0.9%)
ROAD OBSTACLES	2 (7.7%)	6 (23.1%)	18 (69.2%)	0 (0.0%)	0 (0.0%)	26 (0.2%)
SPEEDING	2 (2.4%)	14 (16.5%)	51 (60.0%)	12 (14.1%)	6 (7.1%)	85 (0.6%)
UNKNOWN	325 (2.3%)	3674 (26.0%)	9848 (69.6%)	282 (2.0%)	27 (0.2%)	14156 (94.6%)
TIP_ACC						
ANIMAL COLLISION	1 (6.2%)	3 (18.8%)	11 (68.8%)	1 (6.2%)	0 (0.0%)	16 (0.1%)
CHAIN REACTION	24 (4.4%)	142 (26.2%)	376 (69.4%)	0 (0.0%)	0 (0.0%)	542 (3.6%)
DERAILMENT	0 (0.0%)	1 (12.5%)	4 (50.0%)	3 (37.5%)	0 (0.0%)	8 (0.1%)
DUMP (>2 WHEELS)	0 (0.0%)	3 (15.0%)	17 (85.0%)	0 (0.0%)	0 (0.0%)	20 (0.1%)
FALL (2 WHEELS)	47 (3.4%)	370 (26.9%)	954 (69.3%)	6 (0.4%)	0 (0.0%)	1377 (9.2%)
HEAD-ON	3 (1.2%)	50 (19.4%)	191 (74.0%)	13 (5.0%)	1 (0.4%)	258 (1.7%)
INSIDE VEHICLE FALL	0 (0.0%)	1 (5.0%)	6 (30.0%)	13 (65.0%)	0 (0.0%)	20 (0.1%)
MULTIPLE	20 (2.5%)	175 (21.6%)	540 (66.7%)	64 (7.9%)	10 (1.2%)	809 (5.4%)
OBSTACLE COLLISION	16 (5.2%)	75 (24.4%)	201 (65.5%)	8 (2.6%)	7 (2.3%)	307 (2.1%)
OTHERS	10 (3.8%)	86 (32.7%)	166 (63.1%)	1 (0.4%)	0 (0.0%)	263 (1.8%)
REAR-END	111 (3.2%)	934 (26.7%)	2427 (69.4%)	21 (0.6%)	2 (0.1%)	3495 (23.4%)
ROLLOVER	3 (0.7%)	82 (18.6%)	318 (72.1%)	32 (7.3%)	6 (1.4%)	441 (2.9%)
SIDE-IMPACT	43 (1.3%)	810 (24.9%)	2293 (70.5%)	100 (3.1%)	6 (0.2%)	3252 (21.7%)

SIDESWIPE	87 (2.1%)	1156 (27.9%)	2837 (68.5%)	58 (1.4%)	3 (0.1%)	4141 (27.7%)
UNKNOWN	0 (0.0%)	0 (0.0%)	9 (100.0%)	0 (0.0%)	0 (0.0%)	9 (0.1%)
INT_H						
02:00-05:59	8 (2.2%)	97 (26.3%)	242 (65.6%)	11 (3.0%)	11 (3.0%)	369 (2.5%)
06:00-09:59	83 (3.2%)	692 (26.3%)	1806 (68.7%)	40 (1.5%)	8 (0.3%)	2629 (17.6%)
10:00-13:59	94 (2.9%)	927 (29.0%)	2138 (66.8%)	40 (1.2%)	3 (0.1%)	3202 (21.4%)
14:00-17:59	81 (2.0%)	968 (23.9%)	2922 (72.1%)	79 (1.9%)	4 (0.1%)	4054 (27.1%)
18:00-21:59	66 (1.9%)	844 (24.5%)	2425 (70.4%)	104 (3.0%)	6 (0.2%)	3445 (23.0%)
22:00-01:59	33 (2.6%)	360 (28.6%)	817 (64.9%)	46 (3.7%)	3 (0.2%)	1259 (8.4%)
NUM_VICT						
1	315 (2.7%)	3107 (27.1%)	7807 (68.1%)	211 (1.8%)	20 (0.2%)	11460 (76.6%)
2-3	45 (1.4%)	730 (22.4%)	2368 (72.8%)	98 (3.0%)	12 (0.4%)	3253 (21.7%)
>3	5 (2.0%)	51 (20.8%)	175 (71.4%)	11 (4.5%)	3 (1.2%)	245 (1.6%)
NUM_VEHS						
1	43 (3.4%)	269 (21.2%)	882 (69.5%)	61 (4.8%)	14 (1.1%)	1269 (8.5%)
2	279 (2.3%)	3237 (26.8%)	8334 (69.0%)	218 (1.8%)	18 (0.1%)	12086 (80.8%)
3	35 (2.8%)	306 (24.4%)	884 (70.6%)	26 (2.1%)	1 (0.1%)	1252 (8.4%)
>3	8 (2.3%)	76 (21.7%)	250 (71.2%)	15 (4.3%)	2 (0.6%)	351 (2.3%)

Table 9. Contingency table of the categorical variables.

At first glance, it can be noticed a significant problem. The data is highly imbalanced, the observations with slightly injury (No Injury, Mild, Moderate) as an outcome account for approximately 98% of the observations. There are different possible approaches given data imbalance, which are discussed in the next section. Also, levels of some factors are imbalanced as well. For example, the pedestrian fault (CAUSA_V) is labeled as “Not the pedestrian fault” in most of the observations. This same issue is true for the type of vehicle (TIP_VEH) and type of day (DIA_TIPUS). There are models such as random forests which are able to perform fine with this kind of factors.

Regarding injury severity (DESC_VICT), when the pedestrian is one of the root causes in the casualty, it seems it tends to increase. Additionally, if the vehicle is two-wheeled (motorcycle, moped, bike) the observations contain more severe outcomes. The same can be said for male drivers, which suffer from more severe injuries in proportion to women. Interestingly, there are fewer accidents in August, but more of them are severe.

Also, some districts (NOM_DIST) have a higher concentration of severe outcomes: Horta-Guinardó, Sant Andreu and Sant Martí. When the cause (CAUSA) is either drugs or speeding, the severity increases a lot. The worst casualties seem to be linked with rollovers, derailments, inside vehicle fall and multiple accidents (TIP_ACC). The severity also increases as more people (NUM_VICT) and vehicles (NUM_VEHS) are involved. During the night more casualties seem to have a severe outcome. However, when only one vehicle is involved, more severe and fatal outcomes happen, this may be due to a relation with rollovers.

4.2.1. Correlation Between Explanatory Variables

As discussed, reducing the number of variables used in a model is crucial. Thus, correlations between all explanatory variables are calculated in order to drop the features that are highly correlated. The correlations have been computed between: numerical data, categorical data and both numerical and categorical data, although the last one cannot be used to decide which features are redundant.

Numerical variables

Two methods have been used for these correlations, Pearson and Spearman. The former evaluates the linear association between two interval or ratio-interval variables; the latter evaluates the monotonic relation, that is, it is based on the ranks of data points, rather than on their values (Boslaugh, 2013). A score of 0.7 has been used as a threshold to decide which variables are enough correlated that one can be dropped. Both correlation matrices for Pearson and Spearman's methods are depicted in the following figures.

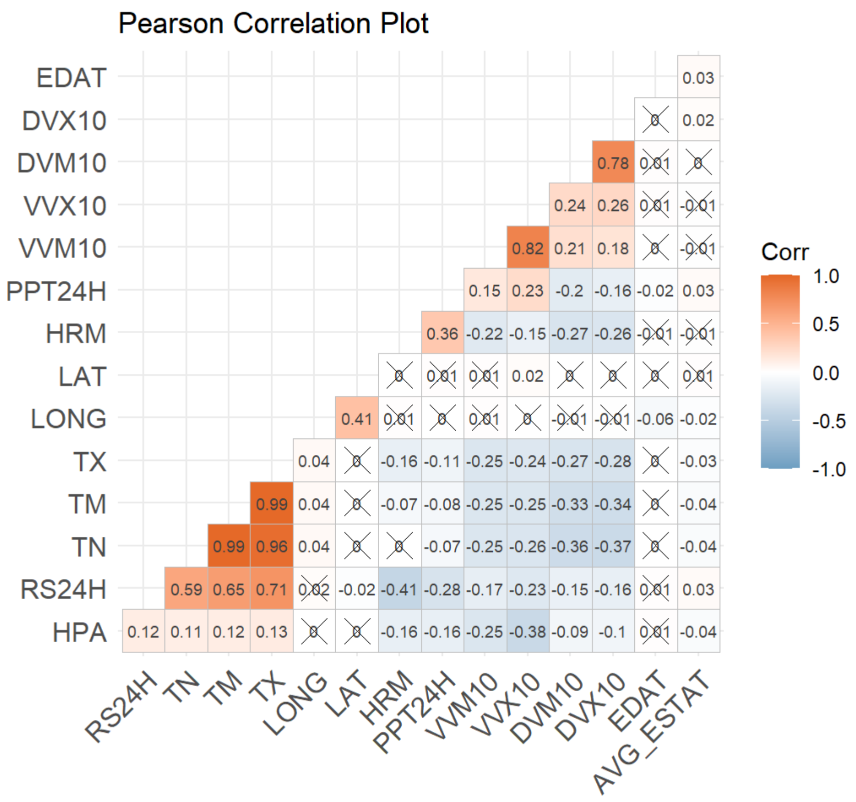


Figure 1. Pearson correlation plot. Correlations with a p-value higher than 0.05 are marked with a cross.

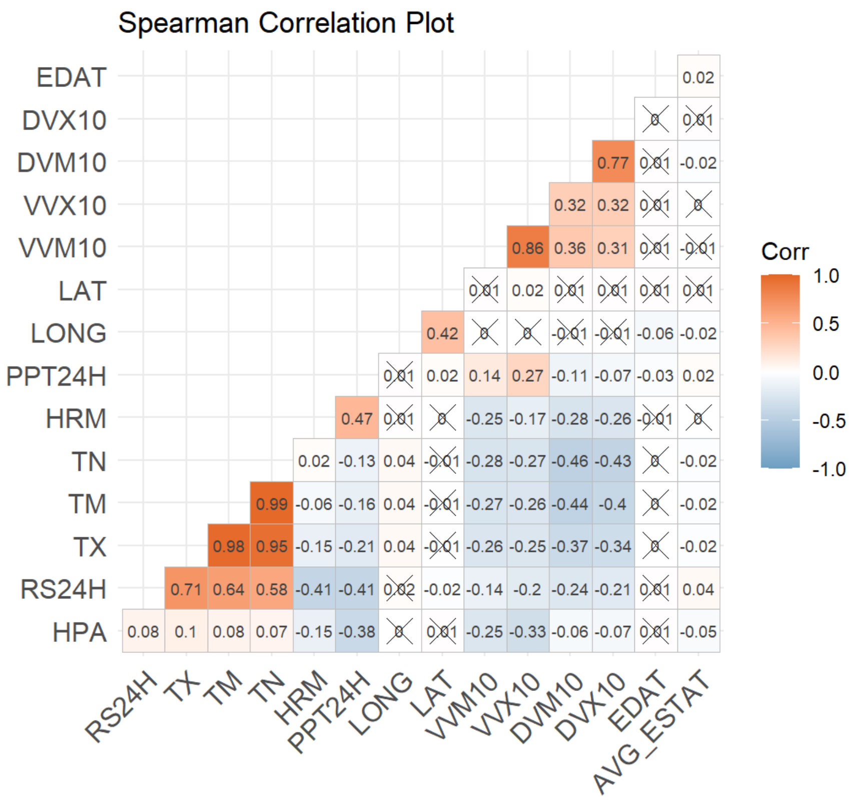


Figure 2. Spearman correlation plot. Correlations with a p-value higher than 0.05 are marked with a cross.

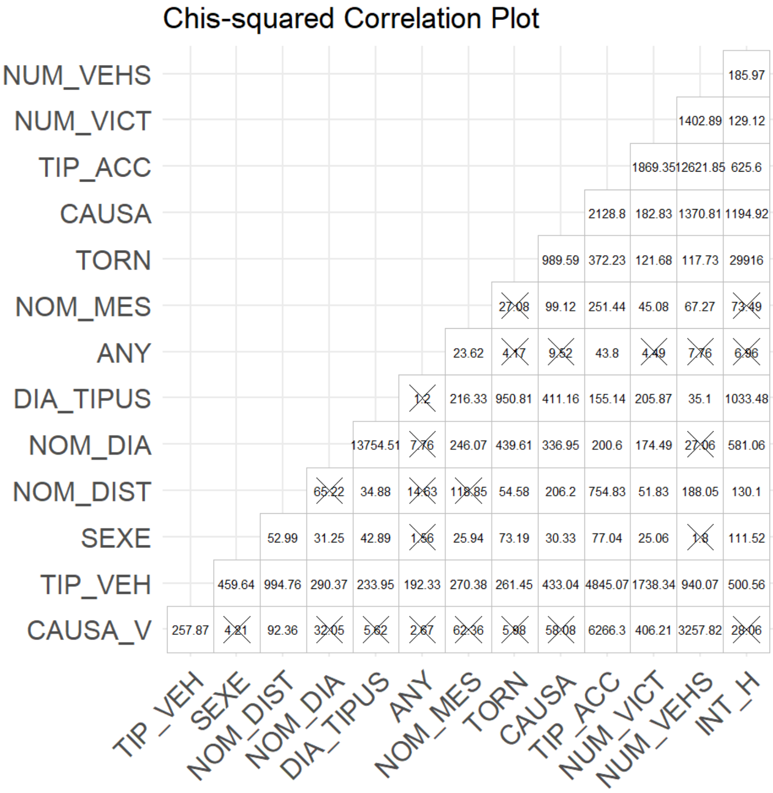
Both coefficients yield similar results. It can be noticed a high correlation between:

- Average daily temperature (TM) & Maximum daily temperature (TX)
- Average daily temperature (TM) & Minimum daily temperature (TN)
- Maximum daily temperature (TX) & Minimum daily temperature (TN)
- Daily global solar radiation (RS24H) & Maximum daily temperature (TX)
- Average daily wind speed (VVM10) & Maximum daily wind gust (VVX10)
- Wind average angle (DVM10) & Angle of the maximum wind gust (DVX10)

All high correlations scores by Pearson and Spearman coefficients are positive. However, slightly negative correlations can be observed between temperatures and wind, as well as between radiation and precipitations. Also, humidity and precipitation are slightly positive correlated. From the above results, average temperature (TM) is kept while maximum (TX) and minimum (TN) temperatures are dropped from the dataset; daily global radiation (RS24H) is not dropped; both average wind speed (VVM10) and wind average angle (DVM10) are kept while its maximums (VVX10 and DVX10) are dropped.

Categorical variables

Again, two methods have been used in order to build a correlation matrix for the categorical variables, Chi-squared and Cramer's V tests. The former is one of the most common ways to examine relationships between two or more categorical variables. It checks how away from being uniformly distributed are two variables. The latter computes a correlation measure following the first results in order to determine the strength of the relationship. Therefore, the matrix with the scores from the Chi-squared test is built; then, the Cramer's V test is used to check the association between the variables. Both matrices can be depicted in the figures below, although only the second is used to decide which features are redundant.



	TIP_VEH	SEXE	NOM_DIST	NOM_DIA	DIA_TIPUS	ANY	NOM_MES	TORN	CAUSA	TIP_ACC	NUM_VICT	NUM_VEHS	INT_H
NUM_VEHS													185.97
NUM_VICT												1402.89	129.12
TIP_ACC											1869.35	12621.85	625.6
CAUSA										2128.8	182.83	1370.81	1194.92
TORN									989.59	372.23	121.68	117.73	29916
NOM_MES								~~27.08~~	99.12	251.44	45.08	67.27	~~73.49~~
ANY							23.62	~~4.17~~	~~9.52~~	43.8	~~4.49~~	~~7.76~~	~~6.96~~
DIA_TIPUS						~~1.2~~	216.33	950.81	411.16	155.14	205.87	35.1	1033.48
NOM_DIA					13754.51	~~7.76~~	246.07	439.61	336.95	200.6	174.49	~~27.06~~	581.06
NOM_DIST				~~65.22~~	34.88	~~14.63~~	~~118.85~~	54.58	206.2	754.83	51.83	188.05	130.1
SEXE			52.99	31.25	42.89	~~1.56~~	25.94	73.19	30.33	77.04	25.06	~~1.8~~	111.52
TIP_VEH		459.64	994.76	290.37	233.95	192.33	270.38	261.45	433.04	4845.07	1738.34	940.07	500.56
CAUSA_V	257.87	~~4.21~~	92.36	~~32.05~~	~~5.62~~	~~2.67~~	~~62.36~~	~~5.98~~	~~58.08~~	6266.3	406.21	3257.82	~~28.06~~

Figure 3. Chi-squared correlation plot. Correlations with a p-value higher than 0.05 are marked with a cross.

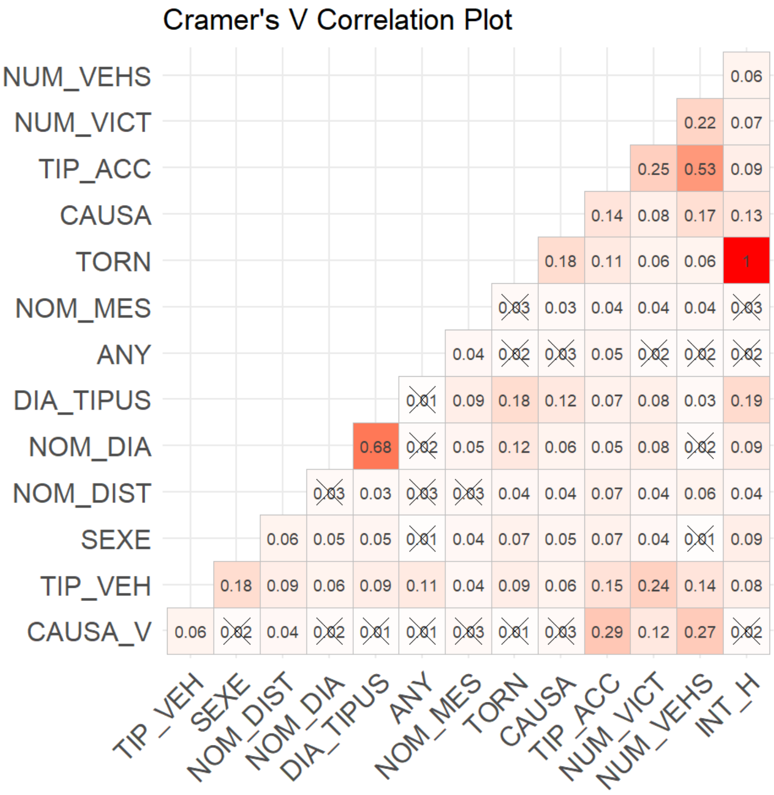


Figure 4. Cramer's V correlation plot. Correlations with a p-value higher than 0.05 are marked with a cross.

From the scores in both matrices, it can be observed that the next features are highly correlated:

- Working shift by the police (TORN) & interval of hours (INT_H)
- Name of the day (NOM_DIA) & type of day (DIA_TIPUS)

These results are expected, since for both correlations, one feature has been defined from the same source as the other. Working shift by the police (TORN) splits the day into three intervals, while interval of hours (INT_H) splits the day into six intervals. Name of the day (NOM_DIA) identifies each unique day of the week, while type of day (DIA_TIPUS) describes each day from a working calendar. Type of accident (TIP_ACC) and number of vehicles involved (NUM_VEH) are slightly correlated as accidents with multiple type of accidents involved need to have more than one vehicle involved most of the time. From the results, working shift by the police (TORN) and name of the day (NOM_DIA) are dropped from the dataset.

Numerical and categorical variables

Concerning the correlations between different types of variables, one testing method has been computed to build a correlation matrix, the ANOVA t-test. It is used to compare the mean values between independent groups. The results are merely used to grasp the underlying relationships as the strength of the associations has not been calculated. The matrix can be depicted in Figure 5.

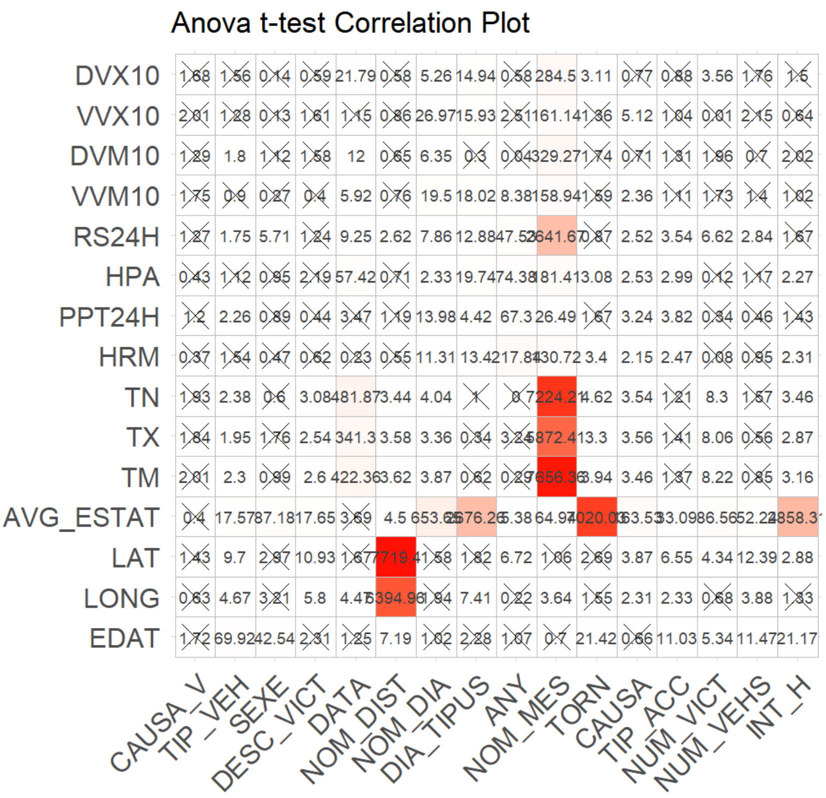


Figure 5. ANOVA t-test correlation plot. Correlations with a p-value higher than 0.05 are marked with a cross.

Regarding the correlations between categorical and numerical features, as the scores are not calculated, no variables are dropped from this test. However, it is interesting how the features are related. It can be observed that the association between coordinates and districts is significant, which is obvious. The association between meteorological measures, such as temperature and humidity, and month are also significant. The associations between traffic density and type of day as well as interval of hours are significant enough to be mentioned.

4.2.2. Redundant Features

From the exploratory analysis some of the features can be confidently dropped from the data that is used in the random forests and classification trees methods. The discarded features can be depicted the following Table 10. Hereby, the dataset is completely ready to be used as input for the supervised models.

DROPPED FEATURES	DESCRIPTION
TX	Maximum daily temperature
TN	Minimum daily temperature
DVX10	Maximum daily wind gust (m/s)
VVX10	Direction of the maximum daily wind gust (º)
TORN	Working shift by the police
NOM_DIA	Name of the day

Table 10. Dropped variables as a result of the exploratory analysis.

4.3. Modeling RF+CART

In this section, the methods adopted are discussed in detail. First, classification and regression trees are briefly described before random forests. Classification trees are used for its ease to logically interpret and visualize the underlying relationships between injury severity and explanatory variables, but they suffer from high variance. Random forests use sampling without replacement to randomly select each subset of data fitted to the trees. Also, in contrast to bagging, where all features are used in each tree, the algorithm randomly chooses m variables from the set of predictors available. Thus, the variance is highly reduced and the most used features can be selected as the most important. Hence, the input variables for the trees are selected according to the random forests results. Second, solutions to class imbalance and the used performance metrics are discussed. Third, given that injury severity is coded into five different levels in the dataset (No Injury, Mild, Moderate, Severe and Fatal) there are multiple options as for the predicted outcome. Models can be fit to classify between these five classes, or the levels can be collapsed to classify between less options. For example, a five-class classifier might misclassify more severe accident than a binary classifier that only classifies between severe and non-severe accidents, but might give insight on the milder severity casualties. Thus, different classifiers are fit in order to better answer different questions.

4.3.1. Classification and Regression Tree (CART)

Both regression and classification problems can be approached with decision trees. It is a supervised, nonparametric, binary segmentation learning technique. In this work, classification trees are adopted. In such a tree, class proportions are predicted for each observation, where the predicted class is given by the class with a higher probability. Recursive binary splitting is used to grow the tree, which is a top-down, greedy approach. Top-down as the split begins at the point where all observations belong to a single region; greedy because the best split is chosen at each step, rather than looking ahead to pick the split that will lead to better performance in the future. In contrast to regression trees, where the residual sum of squares (RSS) is used to select the best split, either Gini index or Entropy are preferred since classification error is not sensitive enough. In this analysis, Gini index is used, which is defined as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \log\left(\hat{p}_{mk}\right) \qquad [1]$$

Where $\hat{p}_{mk}$ represents the proportion of observations in the mth region that are from the kth class. The Gini index evaluates the quality of a particular split and it may be used for the node purity when pruning. While trees with a lot of nodes might overfit the data and consequently lead to poor performance; smaller trees with fewer splits might lead to better performance on the long run. The process with which a tree is reduced is known as pruning, where the cost complexity parameter is tuned. The cost complexity parameter is the price to pay for having a tree with many terminal nodes, defined as:

$$+\alpha|T| \qquad [2]$$

Where $|T|$ indicates the number of terminal nodes in a given tree. The tuning parameter is α. This is added to the function that describes the error. Note that when it is equal to 0, tree is not pruned. The optimal value for the cost complexity parameter is chosen using the cross-validated error. In this study, the smallest tree that is within one standard error of the best tree is selected. That is, if the hypothetical lowest cross-validated error = 0.6 and its standard error = 0.05, then the pruned tree is the one with the minimum number of splits that has its cross-validated error lower or equal to the addition of the both errors.

Classification trees are easy to interpret, as they can be displayed visually. However, trees fit with different subsets from the same data can yield completely different results. That is, they suffer from high variance and its predictions are not as accurate as other methods such as support vector machines.

4.3.2. Random Forest (RF)

Random forest is an improvement over bagging method. Bagging is the method with which multiple trees are fit with different bootstrapped samples from the same set and then averaged. Said trees are not pruned, they are deeply grown. Hence, each one has high variance and low bias. Thus, averaging the trees reduces the overall variance while keeping a low bias. Bagging is described as:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \qquad [3]$$

Where $\hat{f}^{*b}(x)$ represents each model fit with the bth bootstrapped subset. Thanks to bootstrapping, for each bagged tree (from 1 to n grown trees) the remaining observations not participating in the actual fit are used for predictions, which lead to the out of bag error (OOB). This error decreases as more trees are averaged until it converges, in other words, there is a point where growing more trees does not improve the accuracy. This estimate is practically equivalent to the leave-one-out cross-validation. However, for this study, cross-validation has been used as an evaluation method.

Random forest is similar to the described bagging method with a small tweak. Unlike bagging, where all the predictors are considered while growing each tree, only a random sample of m predictors is considered from the full set of p predictors, where usually $m = \sqrt{p}$. In this study, the parameter m is tuned. Thus, the resulting average of the trees is less variable and more reliable. This is known as decorrelating the trees.

As exposed, random forests usually result in a better performance than classification trees, yet it comes with an interpretation trade-off. It is impossible to represent the resulting model as attractive as a tree chart. Nonetheless, the importance of each predictor can be obtained through these models by adding up the Gini index, where the higher the impurity of the variable, the higher the importance (James, Witten, Haste, & Tibshirani, 2017). Detailed descriptions the models described can be looked at this book.

In the rear, classification trees are perfect for interpreting the model, but suffer from high variance and perform poorly compared to other models. Random forests results are much more reliable, but its interpretation is too complex. Therefore, both methods are used to analyze the given data. First, a random forest is fit in order to select the most important variables among all the predictors. A cut-off criterion of 75% is used to select them, where the 25% least important features are dropped from the model. Then, the chosen variables are used to fit a classification tree, so its performance is not affected by misleading explanatory variables. Finally, the tree is visualized in order to interpret the underlying relationships and the model performance is evaluated.

4.3.3. Class Imbalance and Model Performance

During the explanatory analysis, a high-class imbalance has been observed. The percentage of observations for each level of the response variable is depicted in the table below:

INJURY SEVERITY	PERCENTAGE OF OBSERVATIONS
No Injury	2,4 %
Mild	26,0 %
Moderate	69,2 %
Severe	2,1 %
Fatal	0,2 %

Table 11. Percentage of observations from the combined dataset.

Most machine learning classification methods perform poorly when the number of instances of the classes greatly differ from each other. In addition, if the metric to evaluate model performance is not adequate, the output is going to be terrible. For example, in this particular case, if accuracy is used as a performance metric, the chosen method might decide to classify all the observations as Moderate, since the accuracy will be high even if the other classes are ignored. In other words, the accuracy is a bad performance measure given the conditions. Another metric such as the area under the receiver operator characteristic (ROC) curve is much better as class distributions are taken into account.

Different approaches can be taken to deal with class imbalance. Cost function and sampling-based methods can be applied to mitigate this issue. The rationale behind cost sensitive learning is that a false negative count as n false negatives. Thus, a misclassified Fatal outcome is far worse than a misclassified Moderate outcome. Sampling methods can be sub-classified into three categories: undersampling (or downsampling), where observations are discarded until every class has the same number of instances as the minority class; oversampling (or upsampling), where observations are duplicated or generated until every class has the same number of instances as the majority class; and hybrids, where both methods are used. In this particular study, sampling methods are used. Two different sampling methods are used to adjust the imbalanced dataset. First, downsamplig is applied with an under-ratio of 1, which means that instances from each class are randomly deleted until its number of observations are all equal to the minority class. Second, upsampling with an over-ratio of 0.1, followed by downsampling with an under-ratio of 1 is applied, so observations are randomly duplicated until the threshold of 0.1 determined by the majority class is achieved, then the majority class is downsampled to the same level. Instead of upsampling, other algorithms such as smote can be used, where observations are generated according to the nearest neighbors rather than duplicated. The former method has the advantage that the computational time greatly decreases, but a lot of information regarding majority classes is lost. In the latter, information is not lost, but the fitted model may get overfitted by the duplicated observations and perform poorly on the test set.

Regarding the performance metrics, area under the ROC curve is used to tune the hyperparameters of the fitted models. Accuracy, sensitivity and specificity are also used to evaluate and compare the models. The relationship between sensitivity and specificity is displayed with the ROC curve. The optimal value should be determined when both are balanced. Sensitivity measures the proportion of true positives (TP) over all the positives observations (P), which are comprised within TP and false negatives (FN). Specificity measures the proportion of true negatives (TN) over the all the negative observations (N), which are comprised within TN and false positives (FP). Accuracy is simply the proportion of correct predictions over all the predictions. The formulas to calculate the above-mentioned metrics are depicted below.

$$Sensitivity = \frac{TP}{TP + FN} \qquad [4]$$

$$Specificity = \frac{TN}{TN + FP} \qquad [5]$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad [6]$$

4.3.4. Classifiers

Five classes are defined according to the injury severity level. Hence, different options to collapse the given classes are considered:

- Five-class classifier: No Injury > Mild > Moderate > Severe > Fatal
- Three-class classifier 1: No Injury + Mild > Moderate > Severe + Fatal
- Three-class classifier 2: No Injury + Mild + Moderate > Severe > Fatal
- Binary classifier 1: No Injury + Mild > Moderate + Severe + Fatal
- Binary classifier 2: No Injury + Mild + Moderate > Severe + Fatal

The idea behind the five types of models is that grouping the injury severity might help to improve the accuracy as well as answer different questions. The first (1) one is the classifier that does its thing with the classes already defined by the dataset. It might help to observe whether or not are clear differences between the injuries. The second (2) one tries to classify between low (No Injury, Mild), where no medical assist in hospital is needed; medium (Moderate), where the victim has to be carried to the hospital for less than 24h; and high (Severe, Fatal), where the victim has to stay into the hospital more than 24h or ends up passing away. The third (3) one classifies among non-severe (No Injury, Mild, Moderate), where the victim injuries are not severe; severe and fatal. The fourth (4) one classifies between injury (Moderate, Severe, Fatal), where the victim need some kind of medical assist in the hospital; and no injury (No Injury, Mild), where the victim can carry on without medical assist in the hospital. Finally, the fifth (5) one, which is the most common amongst the injury severity articles, classifies between slightly injured (No Injury, Mild, Moderate), where the victim injuries are not Severe; and severely injured (Severe, Fatal), where the victim is highly affected or passes away from the casualty.

3.3.5. Workflow

First and foremost, the injury severity is recoded according to the model that is being computed. That is, for each classifier the classes are grouped by its defined outcomes. For example, in the fourth (4) classifier the first two classes are labeled as “Mild-“ and the others as “Moderate+”. It is worth to mention that seeds are used for every pseudo-random process.

The data is split between training and testing sets (70% - 30%) with a stratified selection to have enough records of each class in both sets. The resampling technique is specified in the pre-processing of the workflow, in the recipe. Also, cross-validation with 10 folds is applied to the training set; this is crucial to avoid over-fitting, although random forests cannot be overfitted due to the bagging and random feature selection.

Hyperparameters are tuned using a big grid which ensures that most of the possibilities are accounted for. For random forests, the tuned parameters are the number of predictors used at each node and the minimum required number of data points in a node to split further from it. The number of trees to assemble is fixed at 1024. Hence, the best tunning parameters are chosen according to the best area under the ROC curve in order to take into account both specificity and sensitivity. Then, the model with better parameters is fit and stored for a later analysis. It is worth mentioning that only the final fit computes the variable importance as it takes more time and would substantially increase the computational time to do the hyperparameter tunning. Finally, the most important variables are selected from each model according to the Gini measure.

The most important features are selected as input for the classification trees, followed by the tuning of the minimum number of observations needed for a split. Next, the trees are fully grown before pruning. In order to prune, the tree with the minimum number of nodes and a cross-validated error lower or equal than the sum between the cross-validated error and the standard error of the best tree is chosen. Finally, the pruned trees are plotted for the analysis and the performance of both the random forests and classification trees is discussed. The whole workflows for each classifier can be observed at the annex.

5. Results

In this section, the results from the random forests are analyzed in order to select the best variables for the classification trees. Then, the trees are depicted and analyzed. Finally, the performance of the models is evaluated.

5.1. Variable Importance Selection

In regard to the key variables, the Gini index is used as a metric to choose the best. The most important variables are selected from the random forest models that have yield better classifications. That is, the classifiers with better area under the ROC curve. The performance of the binary classifier 1 (Moderate-, Severe+) can be depicted in Table 12.

SAMPLING METHOD	OOB ERROR	ACC	SENSITIVITY	SPECIFICITY	BALANCED ACCURACY	**ROC AUC**
DOWN	0.207	0.689	0.690	0.629	0.660	**0.725**
UP&DOWN	0.088	0.872	0.884	0.371	0.628	**0.699**

Table 12. Comparison between under and oversampling of the binary classifier 1.

It can be observed that in the model where the downsampling is applied more observations are misclassified. This is due to the fact that less observations are predicted within the majority class (Sens = 0.689, Spec = 0.690). However, the area under the ROC curve is better. Thus, the model is chosen to select the key features.

The variable importance plots for all the classifiers can be depicted in the figures below. Here, the most important variables for each classifier are ranked. It can be observed that some variables are important for all of them, such as traffic density, type of accident and HPA while others are equally unimportant such as year, sex and type of day.

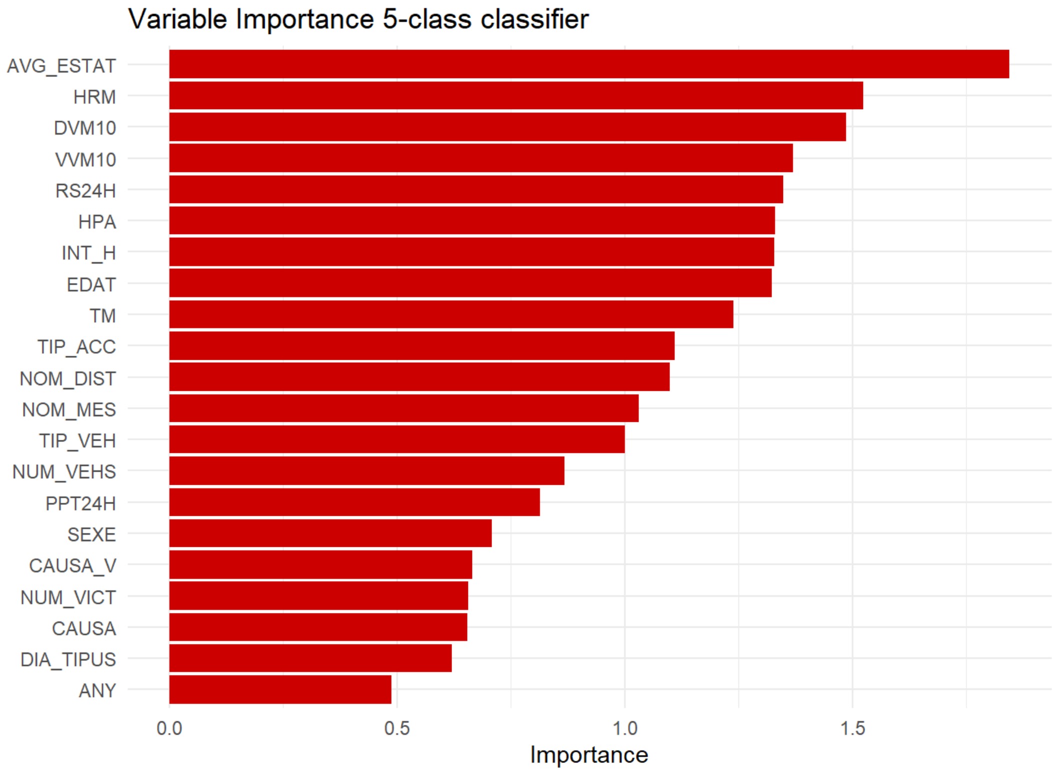


Figure 6. Variable Importance Plot for the 5-class classifier.

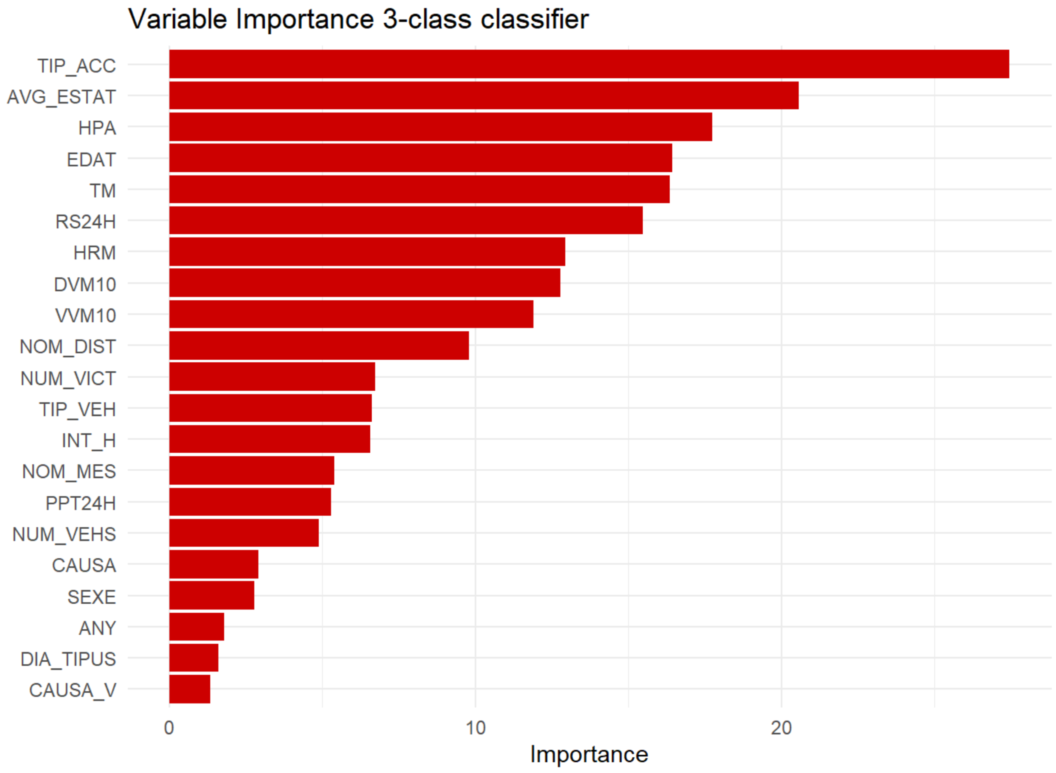


Figure 7. Variable Importance Plot for the 3-class classifier 1.

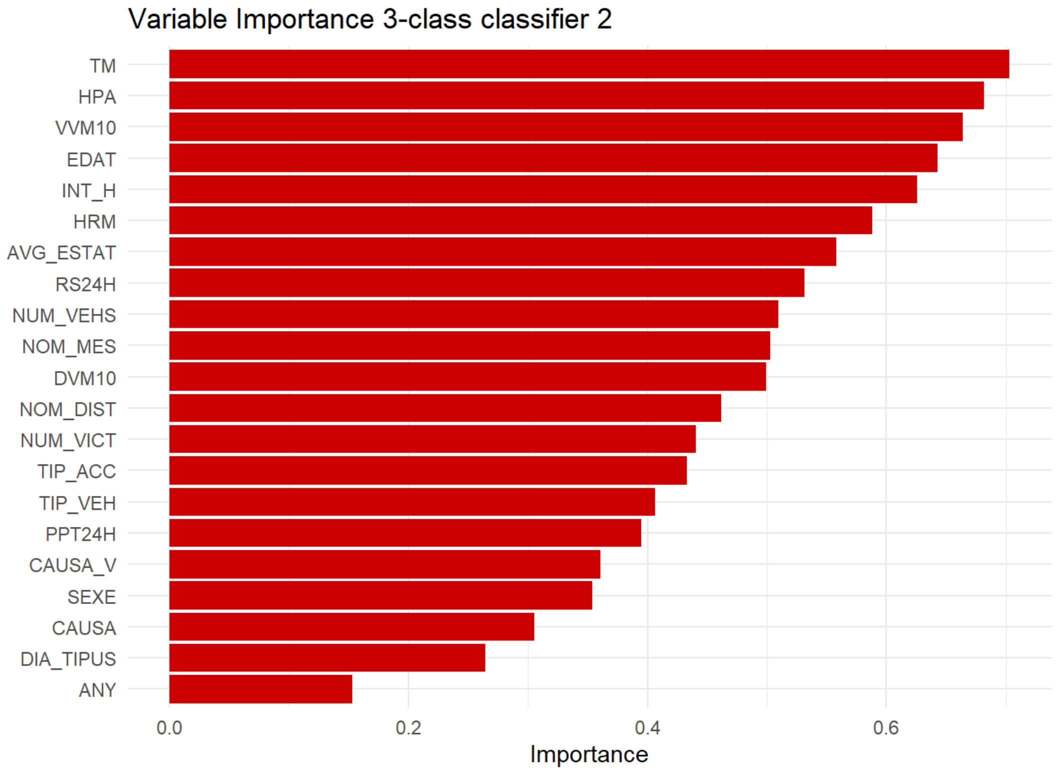


Figure 8. Variable Importance Plot for the 3-class classifier 2.

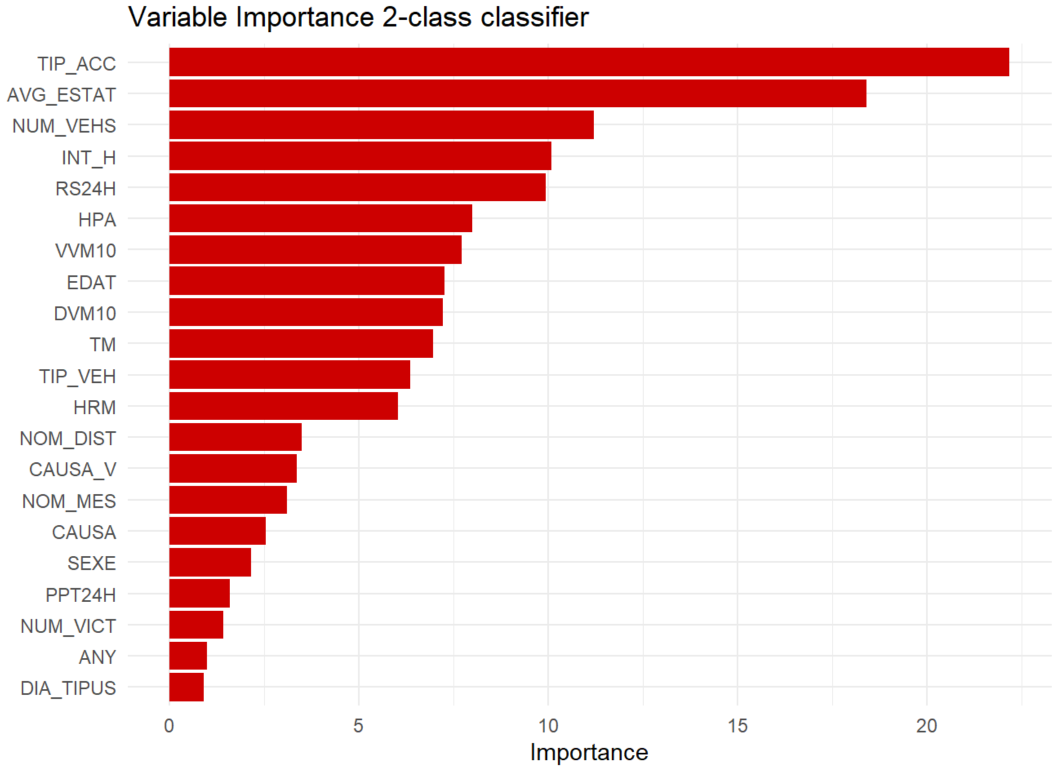


Figure 9. Variable Importance Plot for the binary classifier 1.

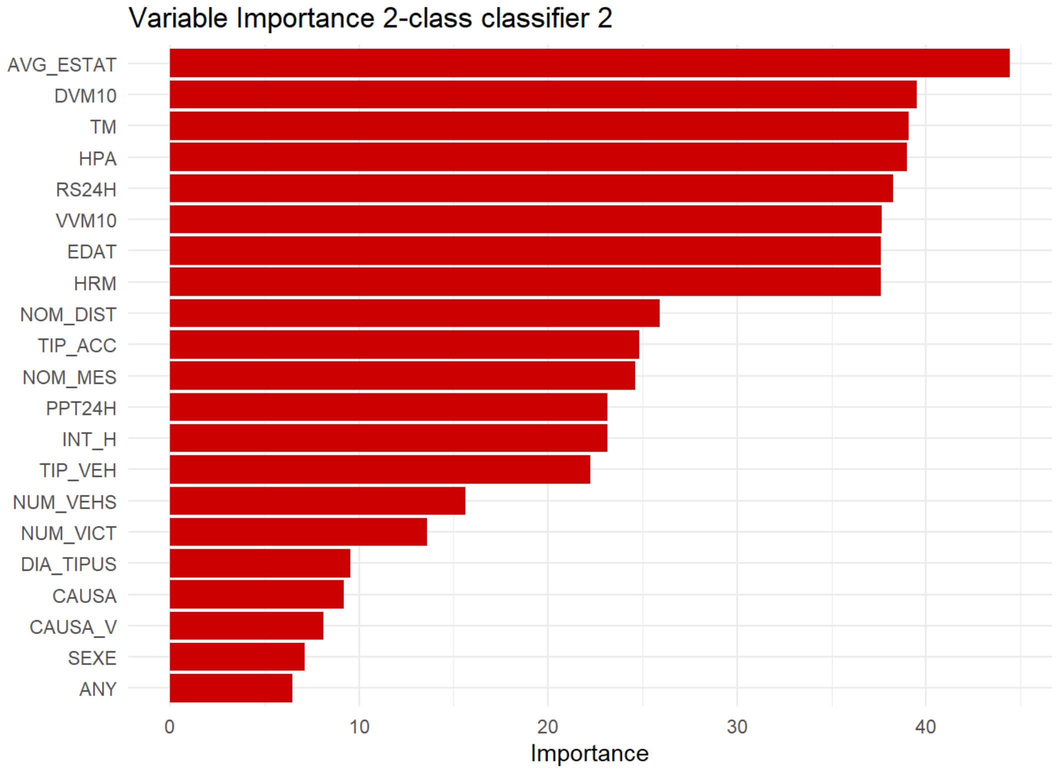


Figure 10. Variable Importance Plot for the binary classifier 2.

Henceforth, the cut-off criterion of 75% is applied to each model. The features dropped for each model can be depicted in Table 13.

CLASSIFIER	CUT-OFF VARIABLES
Five-Class Classifier	• CAUSA_V • NUM_VICT • CAUSA • DIA_TIPUS • ANY
Three-Class Classifier 1	• CAUSA • SEXE • ANY • DIA_TIPUS • CAUSA_V
Three-Class Classifier 2	• CAUSA_V • SEXE • CAUSA • DIA_TIPUS • ANY
Binary Classifier 1	• SEXE • PPT24H • NUM_VICT • ANY • DIA_TIPUS
Binary Classifier 2	• DIA_TIPUS • CAUSA • CAUSA_V • SEXE • ANY

Table 13. Dropped features as a result of the feature selection.

5.2. CART Models

The 16 significant variables of the random forests are used as input for the single classification tree models. Each classifier has as input the variables selected of its homolog random forest. That is, the five-class tree uses the best variables from the five-class random forest. There is one hyperparameter tuned previous to the pruning, the number of observations required at each split. Then, the tree is fully grown and pruned according to the optimal cross-validated error. The CP plot for the binary classifier 1 can be depicted in Figure 11, where it can be observed the corresponding pruned tree lies at CP=0.005. For the five-class classifier the corresponding pruned tree has CP = 0.01; for the 3-class classifier 1 it is CP = 0.013; for the three-class classifier 2 it is CP = 0.021; and for the binary classifier 2 it is CP = 0.015.

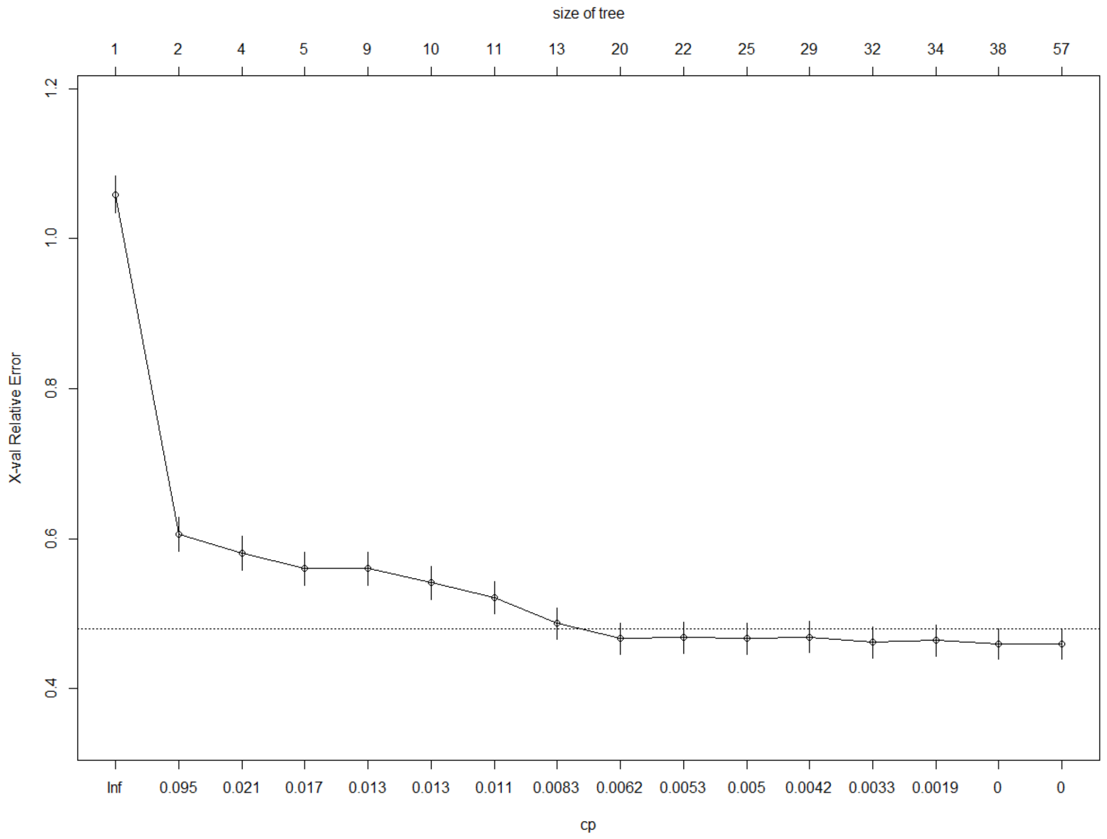


Figure 11. Size of tree versus the cross-validated error.

Five-Class Classifier

The area under the ROC curve is 0.655, which is good as is slightly above 0.5. Also, its misclassification is of 71.9 %. Figure 12 shows the results of the model, it has 9 splits and 10 terminal nodes. Each class is identified with a different color (No Injury: Red, Mild: Orange, Moderate: Grey, Severe: Blue, Fatal: Green). Also, the color intensity within a class may vary depending on its probabilities, that is, the darker the color the higher the probability of the predicted class.

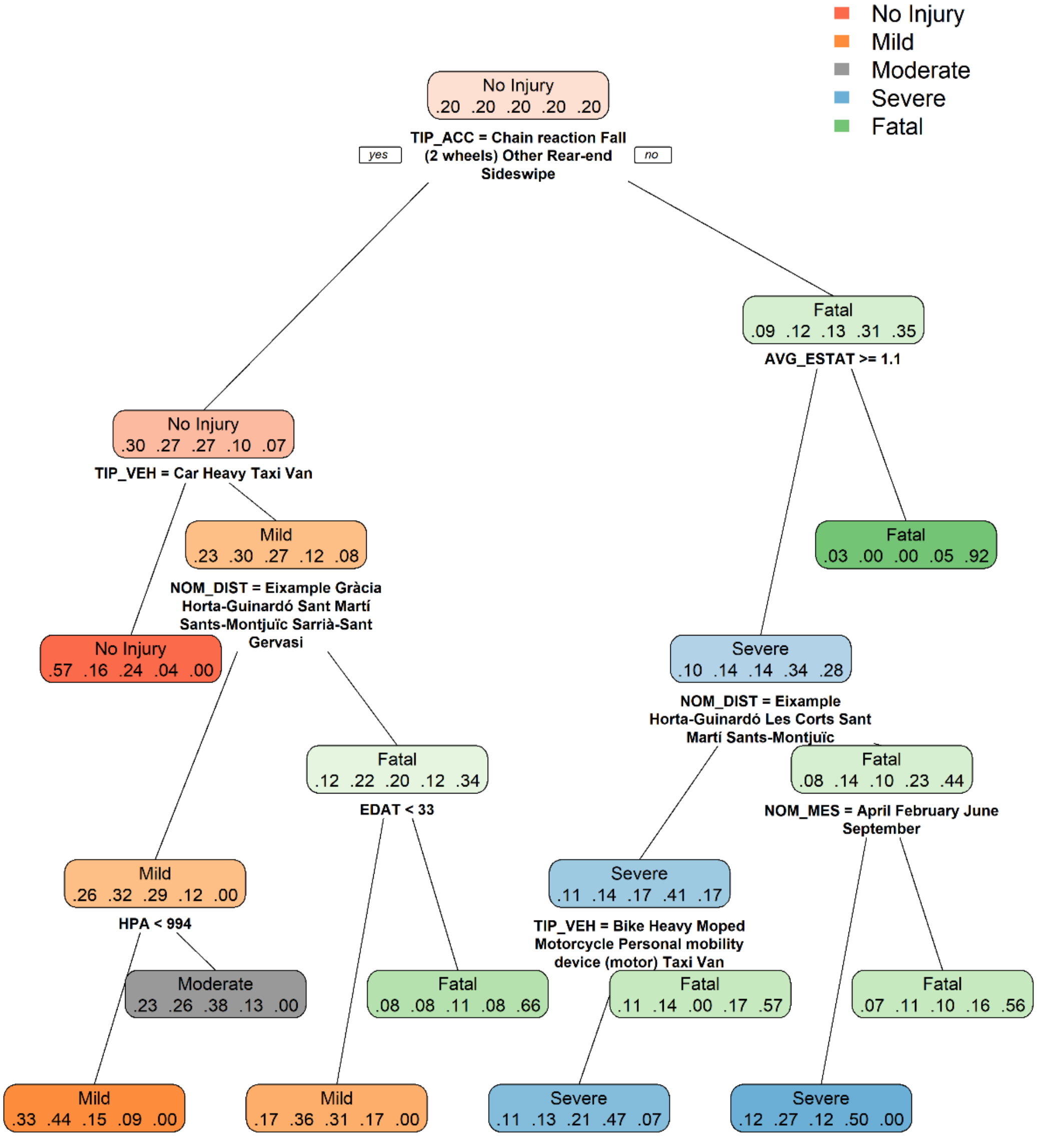

No Injury
Mild
Moderate
Severe
Fatal
No Injury .20 .20 .20 .20 .20
TIP_ACC = Chain reaction Fall (2 wheels) Other Rear-end Sideswipe
yes
no
Fatal .09 .12 .13 .31 .35
AVG_ESTAT >= 1.1
No Injury .30 .27 .27 .10 .07
TIP_VEH = Car Heavy Taxi Van
Mild .23 .30 .27 .12 .08
NOM_DIST = Eixample Gràcia Horta-Guinardó Sant Martí Sants-Montjuïc Sarrià-Sant Gervasi
Fatal .03 .00 .00 .05 .92
No Injury .57 .16 .24 .04 .00
Severe .10 .14 .14 .34 .28
NOM_DIST = Eixample Horta-Guinardó Les Corts Sant Martí Sants-Montjuïc
Fatal .12 .22 .20 .12 .34
EDAT < 33
Fatal .08 .14 .10 .23 .44
NOM_MES = April February June September
Mild .26 .32 .29 .12 .00
HPA < 994
Severe .11 .14 .17 .41 .17
TIP_VEH = Bike Heavy Moped Motorcycle Personal mobility device (motor) Taxi Van
Moderate .23 .26 .38 .13 .00
Fatal .08 .08 .11 .08 .66
Fatal .11 .14 .00 .17 .57
Fatal .07 .11 .10 .16 .56
Mild .33 .44 .15 .09 .00
Mild .17 .36 .31 .17 .00
Severe .11 .13 .21 .47 .07
Severe .12 .27 .12 .50 .00

Figure 12. 5-class classifier tree.

Analyzing the hierarchy of the tree, which classifies casualties between five different degrees of severity, the following ideas can be extracted:

Type of accident (TIP_ACC): when the type of accident encompasses: derailment, head-on collision, inside vehicle fall, multiple accidents, obstacle collision, rollover or side-impact the probability associated to it being a severe or fatal casualty increases to 31%

and 35%, respectively. Otherwise, for type of accidents such as chain reactions, fall (2 wheels), rear-end collisions and sideswipes, the probabilities of it being non-severe are equally split between no Injury, mild and moderate (90%). The first types of accidents are referred as the severe group from now onwards, the second types as the milder group.

Type of vehicle (TIP_VEH): the model identifies differences in severity according to two groups: those involved in the less severe type of accident and the ones involved in more severe type of accidents altogether with other factors. In the first group, the severity associated to cars, heavy vehicles, taxis and vans is lower than the associated to other vehicles (bike, moped, motorcycle, off-road, personal mobility devices and others). In its left node, which is the less severe one, almost all the accidents are comprised between no Injury and moderate (96%) while in its right node the severe and fatal accidents increase to 12% and 8%, respectively. In the second group of accidents, which is preceded by the combination of the more severe type of accidents, higher traffic density and the depicted districts, the severity associated to: cars, off-road vehicles and personal mobility devices (without motor) is higher, with the probability of a fatal outcome increasing from 7% to 57%.

Traffic Density (AVG_ESTAT): once one of the more severe accidents take place and the traffic is very fluid the probability of a fatal outcome greatly increases (92%), in fact, is the worst-case scenario for a casualty. On the contrary, if the traffic is not fluid at all, the probability of a fatal outcome decreases. This might be due to the fact that people drive faster in emptier roads.

District (NOM_DIST): there are two groups where the severity varies according to which district is located the accident. The first encompasses those injured by a milder type of accident and more dangerous vehicles, the second encompasses those injured by a more severe type of accident and higher traffic density. Both share some similitudes, as the severity tends to increase whenever one accident happens in one of the next districts: Ciutat Vella, Nou Barris and Sant Andreu. This is also true for Les Corts in the first combination and Gràcia and Sarrià-Sant Gervas in the second one.

Average daily pressure [atm] (HPA): this factor is important to identify whether in a non-severe injury scenario an ambulance is still needed. When the type of accident has been milder, the vehicle more dangerous and the district one of the safer ones, if the level of HPA is lower than 994 the prxºobabilities of requiring a trip to the hospital decreases, as the no injury and mild outcomes have a probability of 33% and 44%, respectively. On the contrary, for higher HPA levels the moderate outcome greatly increases (38%).

Age (EDAT): in the case where the type of accident has been one of the milder ones, the vehicle is more dangerous and the district is also more dangerous, the driver age is an influential factor. Drivers younger than 33 years old are less probable to suffer from a fatal outcome in comparison with older ones. An explanation could be that more experienced drivers are imbued with overconfidence and thus are less careful.

Month (NOM_MES): according to the model, the severity associated with the more severe type of accidents and districts increases if it does not happen in: February, April, June or September. It is high for all the months, but some of them just happen to contain more fatal outcomes.

Three-Class Classifier 1

The area under the ROC curve is 0.621. Also, its misclassification is of 64.7 %. Figure 13 shows the results of the model, it includes 7 splits and 8 terminal nodes. Each class is identified with a different color (Mild: Red, Moderate: Grey, Severe+: Green). Also, the color intensity within a class may vary depending on its probabilities, that is, the darker the color the higher the probability of the predicted class.

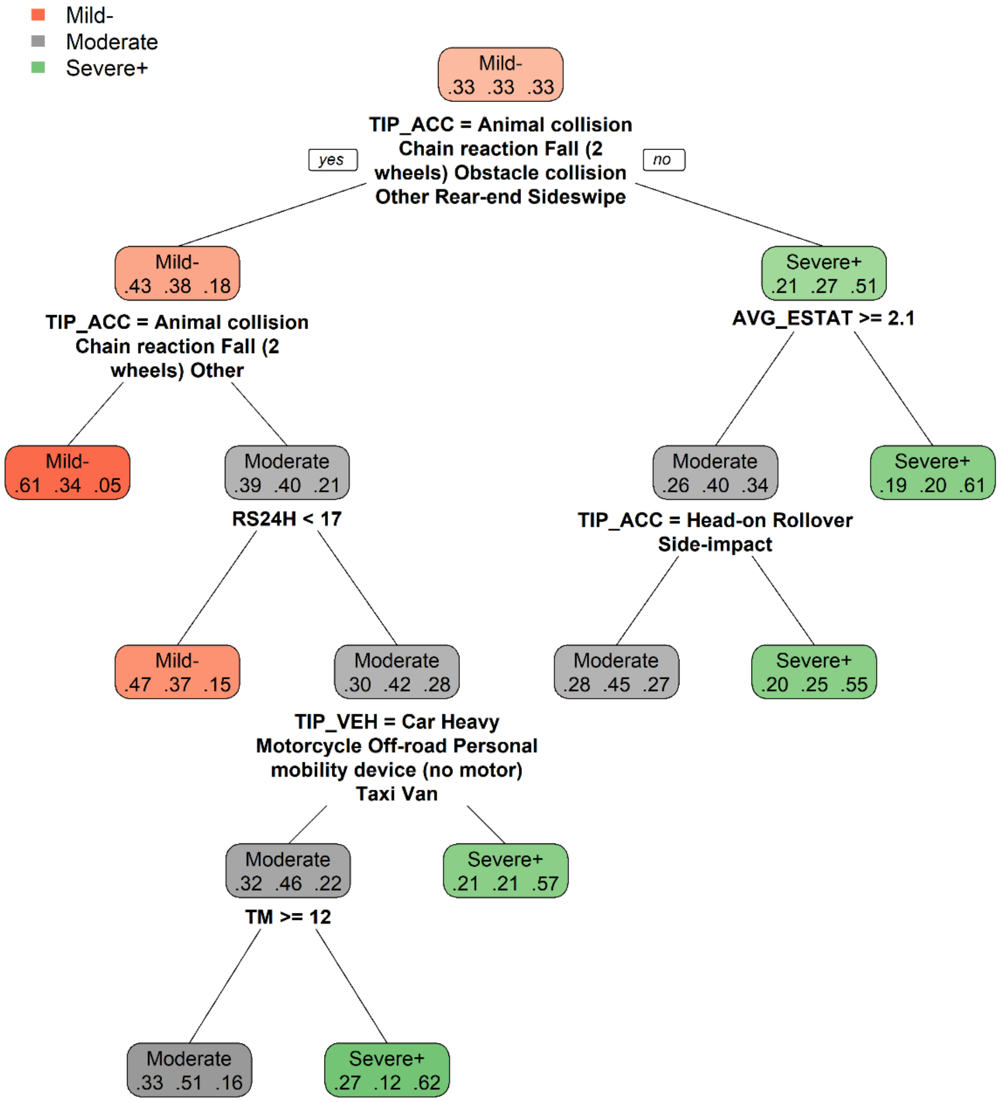


Figure 13. 3-class classifier 1 tree.

Type of accident (TIP_ACC): when the type of accident encompasses: derailment, head-on collision, inside vehicle fall, multiple accidents, rollover or side-impact the probability associated to it being a severe or fatal casualty increases to 27% and 51%, respectively. The model also differentiates severity according to two more groups: the first one is a direct split from the milder group of accidents type, the second one is a split from the more severe type of accidents altogether with a higher traffic density. In the former, when the type of accident is an animal collision, chain reaction, fall (2 wheels) or other less frequent type the related injury is milder, in fact, it is the milder case that the model identifies. In the latter, when the accident is a head-on collision, rollover or side-impact the related injury is also milder. The fact that a rollover has milder outcomes can be attributed to the bikes that the model has into consideration as well as the high traffic density.

Traffic density (AVG_ESTAT): once one of the more severe accidents takes place and the traffic is very fluid, the probability of a severe or even worse outcome increases up to a 61%. This is the worst combination of factors, the one with the higher level of severity associated. On the contrary, if the traffic is more congested the probability of a fatal outcome decreases (31%).

Average daily global solar radiation [MJ/m2] (RS24H): as reported by the model, when the type of accident is an obstacle collision, rear-end collision or sideswipe, higher levels of solar radiation increase the probabilities of a moderate outcome (42%). Lower levels are related to milder outcomes.

Type of vehicle (TIP_VEH): according to the model, in the above-mentioned case, where radiation is higher, accidents involving cars, heavy vehicles, motorcycles, off-road vehicles, personal mobility devices (without motor), taxis or vans are of a milder nature than those involving other type of vehicles.

Average daily temperature (TM): given the above-mentioned combination with the milder type of vehicles, the severity of the casualty increases for lower temperatures. This relationship is counterintuitive, as the Pearson correlation between these two variables is slightly positive. As these values are daily averages, one explanation could be that it rained in an unclouded day, which lowered the temperature average. However, it is not possible to associate the incidents to the rain as it is an unknown factor at the moment they happened.

Three-Class Classifier 2

The area under the ROC curve is 0.743. Also, its misclassification is of 47 %. Figure 14 shows the results of the model, it includes 6 splits and 7 terminal nodes. Each class is identified with a different color (Moderate-: Red, Severe: Grey, Fatal: Green). Also, the color intensity within a class may vary depending on its probabilities, that is, the darker the color the higher the probability of the predicted class.

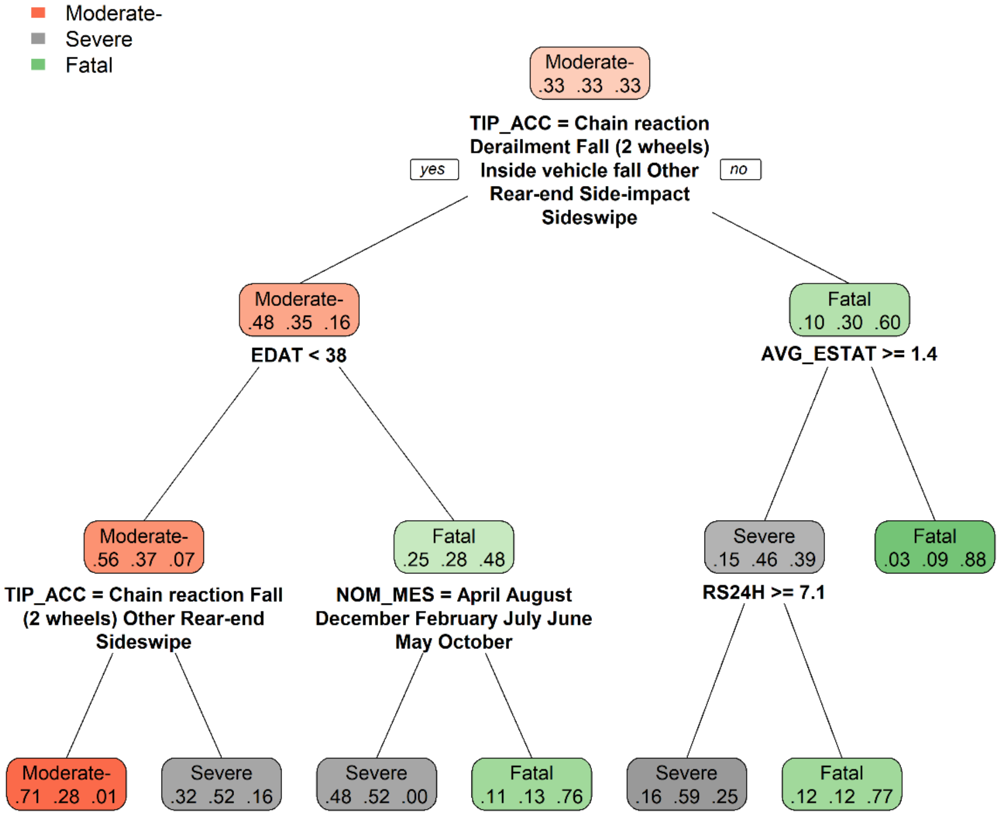


Figure 14. 3-class classifier 2 tree.

Type of accident (TIP_ACC): when the type of accident encompasses: head-on collision, multiple accidents, obstacle collision or rollover the probability of it being a severe or fatal accident increases to 30% and 60%, respectively. Also, within the milder type of accidents and the younger drivers, if the type is a chain-reaction, fall (2 wheels), rear-end, sideswipe or other less frequent casualties the associated severity decreases to moderate or less (71%).

Age (EDAT): according to the model, in the case where the type of accident is one of the milder kinds, drivers younger than 38 years old are less probable to suffer from a severe or fatal outcome in comparison with older ones, 7% for the younger and 48% for the older. An explanation could be that more experienced drivers are imbued with overconfidence and thus are less careful.

Traffic density (AVG_ESTAT): once one of the more severe type of accidents take place and the traffic is close to being perfectly fluid, the probability of a fatal outcome highly increases (88%). On the contrary, if the traffic is not fluid at all, the probability of a fatal outcome decreases to 39%.

Month (NOM_MES): according to the model, in the case where the driver is older and the type of accident one of the less severe, the severity decreases if the casualty has happened in: February, April, May, June, July, August, October or December. The severity on these months is still high, but it is not fatal.

Average daily global solar radiation [MJ/m2] (RS24H): when the type of accident is one of the more severe ones and the traffic density is high, lower levels of solar radiation increase the probabilities of a Fatal outcome (77%). Lower levels are related to Severe outcomes 59%.

Binary Classifier 1

The area under the ROC curve is 0.723. Also, its misclassification is of 25.2 %. Figure 15 shows the results of the model, it includes 21 splits and 22 terminal nodes. Each class is identified with a different color (Moderate: Blue, Severe+: Green). Also, the color intensity within a class may vary depending on its probabilities, that is, the darker the color the higher the probability of the predicted class.

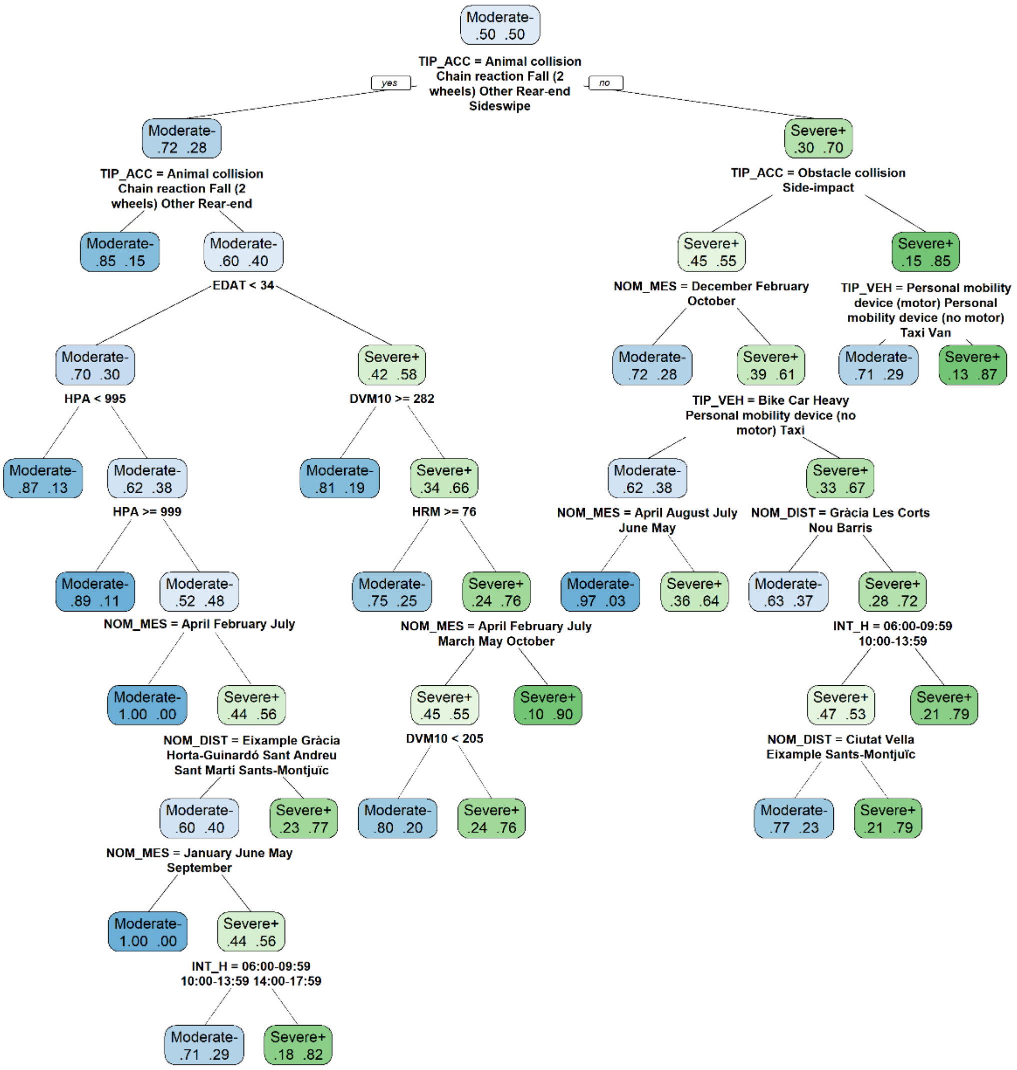

Moderate- .50 .50
TIP_ACC = Animal collision Chain reaction Fall (2 wheels) Other Rear-end Sideswipe
yes
no
Moderate- .72 .28
Severe+ .30 .70
TIP_ACC = Animal collision Chain reaction Fall (2 wheels) Other Rear-end
TIP_ACC = Obstacle collision Side-impact
Moderate- .85 .15
Moderate- .60 .40
Severe+ .45 .55
Severe+ .15 .85
EDAT < 34
NOM_MES = December February October
TIP_VEH = Personal mobility device (motor) Personal mobility device (no motor) Taxi Van
Moderate- .70 .30
Severe+ .42 .58
Moderate- .72 .28
Severe+ .39 .61
Moderate- .71 .29
Severe+ .13 .87
HPA < 995
DVM10 >= 282
TIP_VEH = Bike Car Heavy Personal mobility device (no motor) Taxi
Moderate- .87 .13
Moderate- .62 .38
Moderate- .81 .19
Severe+ .34 .66
Moderate- .62 .38
Severe+ .33 .67
HPA >= 999
HRM >= 76
NOM_MES = April August July June May
NOM_DIST = Gràcia Les Corts Nou Barris
Moderate- .89 .11
Moderate- .52 .48
Moderate- .75 .25
Severe+ .24 .76
Moderate- .97 .03
Severe+ .36 .64
Moderate- .63 .37
Severe+ .28 .72
NOM_MES = April February July
NOM_MES = April February July March May October
INT_H = 06:00-09:59 10:00-13:59
Moderate- 1.00 .00
Severe+ .44 .56
Severe+ .45 .55
Severe+ .10 .90
Severe+ .47 .53
Severe+ .21 .79
NOM_DIST = Eixample Gràcia Horta-Guinardó Sant Andreu Sant Marti Sants-Montjuïc
DVM10 < 205
NOM_DIST = Ciutat Vella Eixample Sants-Montjuïc
Moderate- .60 .40
Severe+ .23 .77
Moderate- .80 .20
Severe+ .24 .76
Moderate- .77 .23
Severe+ .21 .79
NOM_MES = January June May September
Moderate- 1.00 .00
Severe+ .44 .56
INT_H = 06:00-09:59 10:00-13:59 14:00-17:59
Moderate- .71 .29
Severe+ .18 .82

Figure 15. Binary classifier 1 tree.

Type of accident (TIP_ACC): when the type of accident encompasses: derailment, head-on collision, inside vehicle fall, multiple accidents, obstacle collision, rollover or side-impact the probability of it being a severe or fatal outcome increases to 70%. Otherwise, it is of 28%. The model also differentiates severity according to two more groups: the first one is a direct split from the milder group of type of accidents, while the second one is a direct split from the more severe type of accidents. In the former, when the type of accident is an animal collision, chain reaction, fall (2 wheels), rear-end or other less frequent types the associated injury is moderate or less (75%). In the latter, when the accident is a derailment, inside vehicle fall, multiple or a rollover the related injury is more severe (85%); otherwise, the probability decreases to 55%.

Age (EDAT): according to the tree, when the type of accident is a sideswipe, the age of the driver is influential to the outcome. For drivers older than 34 years old, the probabilities of the accident being severe or worse increase to 58%. On the contrary, the probability of it being severe for younger people is only of 30%. An explanation could be that more experienced drivers are imbued with overconfidence and thus are less careful.

Month (NOM_MES): multiple nodes are described by this variable. Three of them are under the milder group of type of accidents, the other two are under the more severe type of accidents. In the first case, when the combination is the above-mentioned with a younger driver and an average level of pressure, the severity is lower in February, April and July. Following the opposite node of this months, for the depicted districts the severity is higher in March, August, October, November and December. However, for older drivers and certain meteorological conditions the severity is higher in January, June, August, September, November and December. Regarding the two nodes that grow from the more severe accidents, whenever the type of accident is an obstacle collision or side-impact if the month is February, October or December the severity decreases down to 28%. Otherwise, if the vehicle involved is one of the depicted in its right node and it happens in Juanuary, March, September or November the severity increases to 64%.

Type of vehicle (TIP_VEH): the model identifies differences in severity according to two groups within the more severe type of accidents: those involved in the subgroup of more severe type of accidents (derailment, head-on collision, inside vehicle fall, multiple and rollover) and the ones involved in the subgroup of less severe type of accidents (obstacle collision and side-impact) and specific months. In the former, casualties involving bikes, mopeds, motorcycles, cars and heavy vehicles have a higher probability of being severe (87%), in fact, it is the worst-case scenario according to the model; otherwise, whenever the vehicle involved is a personal mobility device, taxi or van the severe outcome decreases to 28%. In the latter, casualties involving bikes, cars, heavy vehicles, personal mobility vehicles (without motor) and taxis tend to be less than moderate (62%); otherwise, for mopeds, motorcycles, personal mobility devices (with motor) and vans the moderate outcome decreases to 33%.

Average daily pressure [atm] (HPA): given the combination of factors where the driver is younger, the probabilities of a casualty being milder increase for more extreme values of pressure. The model may identify this relationship due to the fact that severe outcomes in the city rarely happen, so the majority of them occur during average days.

Average daily wind angle [m/s] (DVM10): similar to the pressure, when the combination of factors involves an older driver, casualties tend to be milder for certain values of wind direction. The model may identify this relationship due to the fact that severe outcomes in the city rarely happen, so the majority of them occur during average days.

Average daily relative humidity [%] (HRM): in the case where the accident is a side-swipe and the driver is older, lower levels of relative humidity are associated to higher severity if the wind speed is higher than average. Lower levels of relative humidity can aggravate people allergies, which may disturb the drivers and thus increase the risk of distraction while driving.

District (NOM_DIST): the model differentiates severity levels conforming to three groups: the first involves a sideswipe type of accident, younger drivers and average pressure. The second involves an obstacle collision or sideswipe, two-wheeled vehicles, cars or vans, and specific months. The third is a ramification from the second where the time of the casualty is also significant. Some districts are shared within the two first groups of, from which can be observed that accidents happening in Ciutat Vella and Sarrià-Sant Gervasi tend to be more severe. In the first group, Les Corts and Nou Barris are also associated to more severe outcomes. Eixample, Horta-Guinardó, Sant Andreu, Sant Martí and Sants-Montjuïc also have higher probabilities of being involved in worse outcomes. In the third group, if the time is comprised between the depicted hours and the district involved is Eixample, Ciutat Vella or Sants-Montjuïc the severity greatly decreases.

Hour (INT_H): the model separates severity levels according to two groups of hours, the first encompassing the milder group of accidents and the second the more severe group, although both precede from more complex combinations. The former involves sideswiped, younger drivers and average pressure as well as the depicted months. The latter involves obstacle collisions and side-impacts, vehicles such as two-wheeled and the depicted months and districts. They both share some similar times where the severity is increases, from 18:00 in the afternoon until to 6:00 in the morning. Also, the second group identifies that the range between 14:00 and 18:00 is more dangerous. Overall, the night shift is associated to a worse outcome.

Binary Classifier 2

The area under the ROC curve is 0.553. Also, its misclassification is of 49.3 %. Figure 16 shows the results of the model, it includes 6 splits and 7 terminal nodes. Each class is identified with a different color (Mild-: Blue, Moderate+: Green). Also, the color intensity within a class may vary depending on its probabilities, that is, the darker the color the higher the probability of the predicted class.

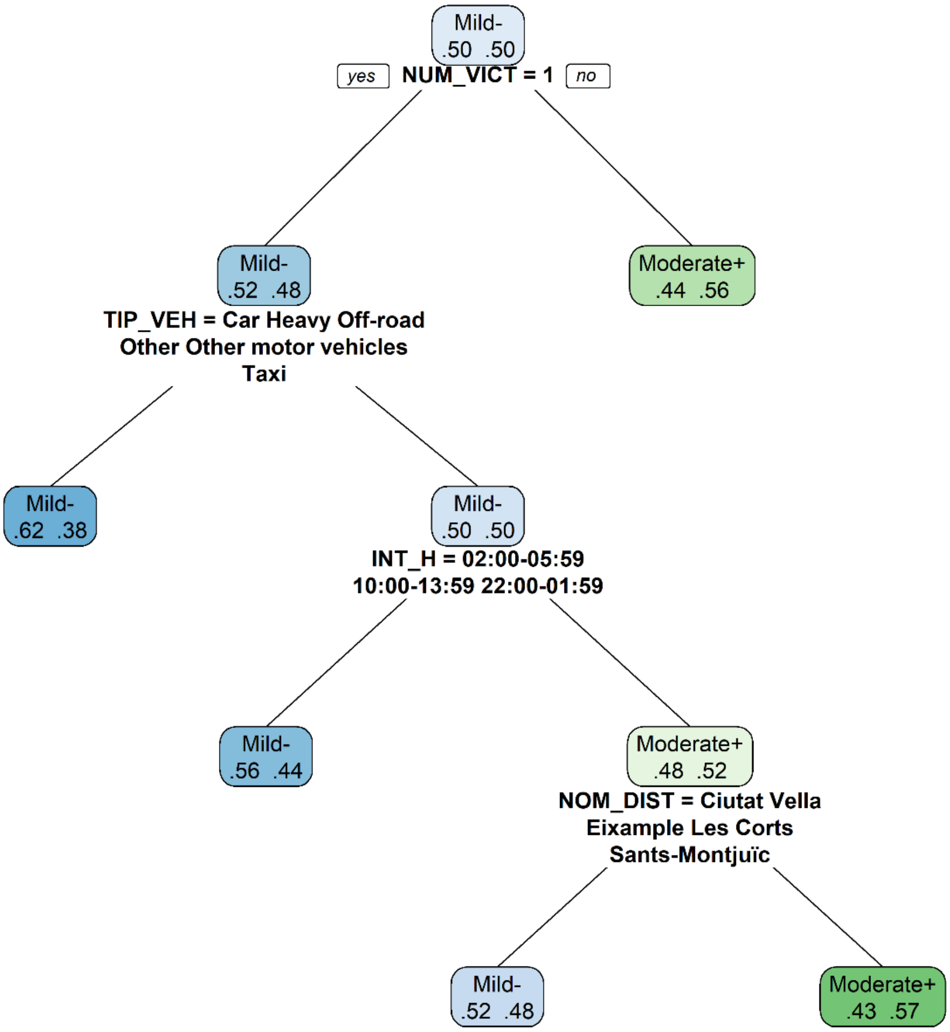


Figure 16. Binary classifier 2 tree.

Number of victims (NUM_VICT): the number of victims involved is the most important factor of influence in this classifier. If it is greater than 1 the probabilities of requiring transport to the hospital increase to 56%. Otherwise, the probability of being treated in the same place of the casualty is of 52%.

Type of vehicle (TIP_VEH): according to the model, when there is only one clear victim identified, some vehicles are safer than others when it comes down to suffering meaningful injuries: cars, heavy off-road vehicles, other motor vehicles and taxis. On the contrary, it is easier to suffer worse injuries in two-wheeled vehicles such as bikes, mopeds and motorcycles, and personal mobility devices.

Hour (INT_H): given the combination of one victim and the less safe vehicles, meaningful injuries (where transport to the hospital is required) are more probable during a casualty that happens between 6:00 and 10:00 in the morning or 14:00 and 22:00 in the afternoon.

Number of vehicles (NUM_VEHS): following the above-mentioned combination, if the time of the casualty is comprised during the night and midday, scenarios where either only one vehicle or more than three vehicles are involved is associated with a higher level of severity. Otherwise, for 2 and 3 vehicles the outcome is milder.

District (NOM_DIST): again, following the above-mentioned combination where the time of the accident encompasses more meaningful injuries, the district is also influential to the outcome. In fact, for Ciutat Vella, Eixample, Les Corts and Sants-Montjuïc the probability of requiring transport to the hospital decreases to 48%.

Month (NOM_MES): in the case where the district is one of the above-mentioned, the severity decreases if the casualty happens in: January, March, April, May, July, October, November or December, where the probabilities of it being milder increase to 54%. Otherwise, the probability of a notorious injury increases to 55%.

5.3. Model Performance

Table 14 shows the performance of the models where both RF and CART methods have been applied. Both methods have similar performance, although CART seems to perform slightly better in some cases, probably due to the fact that RF tuning time has been traded-off for computational speed. Looking at the accuracy, it is noted that most of the models misclassify the majority of observations, especially the ones which classify between more classes. Given the high imbalanced data, both sensitivity and specificity are depicted and comprised within the balanced accuracy and ROC curve. Balanced accuracy is higher than accuracy in most of the models, due to the fact that the minority classes have been prioritized. The ROC curve area shows that most of them perform much better than random models, except for the last binary classifier, which is closer to having the performance of a random classifier.

MODEL & METRICS	**RF**	**CART**
Five-Class Classifier		
Accuracy	0.315	0.281
Sensitivity	0.362	0.330
Specificity	0.804	0.803
Balanced Accuracy	0.583	0.567
ROC AUC	0.640	0.655
Three-Class Classifier 1		
Accuracy	0.344	0.353
Sensitivity	0.471	0.430
Specificity	0.671	0.667
Balanced Accuracy	0.571	0.549
ROC AUC	0.657	0.621
Three-Class Classifier 2		
Accuracy	0.511	0.531
Sensitivity	0.563	0.586

Specificity	0.725	0.794
Balanced Accuracy	0.644	0.690
ROC AUC	0.713	0.743
Binary Classifier 1		
Accuracy	0.689	0.748
Sensitivity	0.690	0.751
Specificity	0.629	0.623
Balanced Accuracy	0.660	0.687
ROC AUC	0.725	0.723
Binary Classifier 2		
Accuracy	0.530	0.507
Sensitivity	0.516	0.620
Specificity	0.536	0.463
Balanced Accuracy	0.526	0.541
ROC AUC	0.535	0.553

Table 14. Performance and comparison of both RF and CART methodologies.

Analyzing the performance by type of classifier and focusing on the are under the ROC curve, it can be seen that there is not a single factor that allows to classify between milder injuries, where no transport is needed to the hospital, and more moderate ones, where a trip to the hospital is needed. This idea is extracted from the last binary classifier, which performance resembles that of a random classifier. Also, even though the accuracy of the three-class classifier 1 (Mild-, Moderate and Severe+) is better than the five-class classifier (No Injury, Mild, Moderate, Severe, Fatal), its ROC AUC is similar but slightly worse. This may indicate that a factor able to classify between No Injury and Mild outcomes is missing; note that the decrease in the performance might be due to the fact that classifying between severe and fatal outcomes is easy, but as they are grouped the specificity slightly decreases and so does the ROC AUC. On the contrary, the performance of the three-class classifier 2 (Moderate-, Severe, Fatal) is much better than its homolog three-class classifier 1 (Mild-, Moderate and Severe+). This indicates that classifying between the milder group of casualties (No Injury, Mild and Moderate) needs more factors in order to be explained. Again, this model slightly outperforms the binary classifier 1 (Moderate-, Severe+) in the AUC ROC, although falls behind in accuracy, which may be explained due to the fact that the distinction between severe and fatal outcomes can be explained by some factors, so when grouped the specificity slightly decreases and so does the AUC ROC. Overall, the binary classifier 1 (Moderate-, Severe+) outperforms the rest of the models. The independent variables in this study are able to decently classify between non-severe, severe and fatal injuries.

6. Discussion

In this section, the results obtained are discussed. 16 variables are selected according to the importance ranking given by the RF method. A different ranking is yield depending on the number of injury severity categories. These variables are then used as input for the CART models; in this way, the classification rate can be improved. Also, classification trees allow a logical analysis given the ease with which the relationships can be visualized. The variable importance in RF slightly differs with the importance ranking of the CART, but most of the more relevant variables are coincident. The following variables are emphasized: type of accident, type of vehicle and traffic density.

Here, the results of the best models are compared with what was observed in the exploratory analysis. Unless the contrary is stipulated, the comparison is done with the binary classifier 1. A summary is depicted in Table 15, where the more severe factors selected by both analyses are shown. To start with, some accidents identified as the more severe are consistent with the frequencies from the exploratory analysis; that is, derailments, inside vehicle falls, multiple accidents and rollovers are also identified by the classifier as more severe. However, head-on collisions, obstacle collisions and side-impacts are also influential in more severe outcomes according to the model. On the contrary, animal collisions, falls (two-wheels), chain reactions, rear-end collisions and sideswipes are associated to lower degrees of severity. It is self-explanatory, but it can be noted in the second group the passengers do not suffer from a direct impact from an external force or it is absorbed by the vehicle itself, such as in the case of a rear-end collision.

Regarding the type of vehicle, in the previous analysis it was observed that two-wheeled vehicles were associated to more severe outcomes, but according to the models, this is true for a given combination of factors. In fact, when the type of accident is a derailment, head-on collision, inside vehicle fall, multiple accidents or rollover the severity greatly increases when a two-wheeled vehicle, car or van is involved. On the other hand, if the type of accident is an obstacle collision or side-impact the severity associated to mopeds, motorcycles, personal mobility devices (with motor) and vans seems to increase for some specific months such as August.

In the exploratory analysis is shown that for lower levels of traffic density the severity seems to increase. However, it has not been identified as a factor of influence in the best model, where the severity is classified either as moderate or severe, but in the models where there is a distinction between severe and fatal outcomes it is one of the most important ones. That is, when the casualty involves one of the most severe types of accident above-mentioned and the traffic density is fluid the severity greatly increases. This may be due to the fact that people drive faster in less congested roads.

In relation to the driver’s age, the descriptive statistics shown that age did not seem to impact the outcome of the casualty, however, in combination with other factors appear to be significant. Drivers older than ~34 years old are involved in worse casualties given a sideswipe type of accident, which may indicate that more experienced drivers are imbued with overconfidence compared to younger ones. Also, driver’s sex has not been influential in any case, although at first sight it seemed that males were associated to worse outcomes.

VARIABLE	EXPLORATORY ANALYSIS	CART
TYPE OF ACCIDENT	Derailments, inside vehicle falls, multiple accidents and rollovers	Derailments, inside vehicle falls, multiple accidents, rollovers, head-on collisions, obstacle collisions and side-impacts
TYPE OF VEHICLE	Two-wheeled vehicles	Two-wheeled vehicles, car and vans altogether with: derailments, head-on collisions, inside vehicle falls, multiple accidents or rollovers Mopeds, motorcycles, personal mobility devices (with motor) and vans in combination with: obstacle collisions or side-impacts
TRAFFIC DENSITY	Lower traffic density	Lower traffic density is influential to differentiate between severe and fatal outcomes whenever the type of accident is one of the above-mentioned
DRIVER AGE	-	Drivers older than 34 years old involved in sideswipes
DRIVER SEX	Male	-
METEOROLOGICAL FACTORS *All the meteorological factors influential in the models are so in very specific circumstances described in the previous section	Particular directions of wind Higher average temperature Lower cumulative precipitation Lower average daily radiation	Particular directions of wind Higher average temperature Lower average daily radiation Particular range of pressure
DISTRICTS	Horta-Guinardó, Sant Andreu and Sant Martí	Horta-Guinardó, Sant Andreu, Sant Martí, Nou Barris and Sants-Montjuïc are influential under very specific combinations of factors
MONTHS	August	All months are influential under very specific combinations of factors
HOUR	[22:00,06:00]	[22:00,06:00] for sideswipe accidents where young drivers are involved, among others combinations
CAUSE	Speeding and Drugs	-

Table 15. Variables with an associated higher severity according to the exploratory analysis and CART.

According to the models, some meteorological factors are significant but only in combination of many factors. For example, the best classifier has selected the average pressure, the average wind direction and the average relative humidity as influential factors under certain conditions which can be depicted in the previous section. In the exploratory analysis, nothing could be said about pressure and humidity, but the wind direction did seem to increase the severity for different values. Other meteorological variables such as average temperature and daily radiation are not influential in the binary classifier, but they appear to be in the three-class classifier 1, where the combination of higher radiation and lower temperatures altogether with other factors seem to increase the severity of the outcome. In the exploratory analysis, higher temperature and lower cumulative precipitation seemed to be influential as well as lower values of average daily radiation. Neither wind speed nor cumulative precipitation are significant in any case.

Districts, months and hours combined with other factors are also associated to different severity levels in the classifier. In the previous analysis, three districts were highlighted for having a higher frequency of severe outcomes: Horta-Guinardó, Sant Andreu and Sant Martí. This is true for some combinations of factors depicted in the previous sections, but other districts such as Nou Barris and Sants-Montjuïc, which did not have a high frequency of severe outcomes, have been selected as influential under some conditions. Regarding the month, only August was highlighted in the exploratory analysis due to the higher frequency of severe outcomes. However, it is only influential in specific combinations of other factors. The same can be said about the remaining months, which are only significant in particular cases. The time of the day is also found important in specific contexts, where the night shift seems to be associated to higher severity. This is consistent with what was observed in the exploratory analysis.

Other variables such as the accident cause, which seemed to be significant in the first analysis, have not been even considered by the classifiers. It was observed that whenever the cause of the casualty was speeding, drugs or the pedestrian fault the proportion of severe casualties increased, but it has not been selected by the CART nor RF as an important variable. Number of victims and vehicles involved are selected as important in the model that resembles a random classifier, so nothing can be said with robustness. The year has not been observed to be significant in any case.

The performance (ROC AUC) of the models that classify between non-severe and severe/fatal accidents is acceptable as they quite outperform random models. The metric can be depicted in Table 16. The performance of the models that also classify between non-severe categories such as no injury and mild have worse results, but still decent. However, the model that classifies between injuries that require a trip to the hospital and those that do not require the visit has a close performance to that of a random model. Other studies have shown that collapsing the categories down to two classes result in less variance and more robustness. Also, some methods such as support vector machines are known to be superior, but lack the explanatory power of CART.

CLASSIFIER (CART)	AREA UNDER THE ROC CURVE
5-Class Classifier	0.655
3-Class Classifier 1	0.621
3-Class Classifier 2	0.743
Binary Classifier 1	0.723
Binary Classifier 2	0.553

Table 16. Area under the ROC curve of the CART classifiers.

7. Project Cost

The project cost breakdown can be depicted in Table 17. Costs regarding materials used, software and work force are taken into account. First, the price of the equipment used is show. Then, the software used for coding the data analysis. Finally, the workforce is divided into the 2 main activities, research and data analysis. Both of them are crucial, but the analysis is the most important activity of the project.

Project Cost			
Tools	**Quantity (u)**	**Unit Cost (€/u)**	**Cost (u)**
Xiaomi Mi Air 13.3"	1	899	899
R Studio		0	0
		Tools Cost (€):	899
Workforce	**Time (h)**	**Cost (€/h)**	**Cost (€)**
Research	80	22	1760
Data Analysis	220	28	6160
		Workforce Cost (€):	7920
		Total Cost (€):	8819

Table 17. Project cost breakdown.

8. Economic, Social and Environmental Impact

In this section, the impact of the study regarding economic, social and environmental issues are discussed. First, in relation to the downsides, the cost of the project has been of 8819 €, which is mainly due to the workforce. About the social impact, no one's privacy has been compromised with the data used. Each accident is labeled with a record code which cannot be linked to the people involved. Regarding the environmental issues, the electric consumption has not been significantly higher than usual and thus the pollution is not an issue.

In relation to the upsides, the study could help to understand the underlying causes of the accidents that have happened in Barcelona; thus, help to reduce the injury severity of the casualties. According to an analysis from 30 countries in Europe (which includes Spain), the monetary valuation of preventing a fatality in a road crash varies from 0.7 to 3 million Euro. In fact, in Spain the estimate is approximately 1.6 million (Wim Wijnen, 2019).

Regarding social impact, fatal road traffic accidents concern job losses, loss of amenity and a critical impact on the functioning of the involved family. Also, fatal accidents require a bigger deployment of workforce and usually involve traffic jams, which are harmful since emissions start to go up when average speed decrease. Overall, reducing the severity of road traffic accidents would greatly reduce the economic, social and environmental impact.

Conclusions

Regarding the objectives, the main scope of this study was to predict the injury severity of the accidents that take place inside the city of Barcelona as well as identifying the patterns that contribute to the increased severity of an injury. Different classifiers have been modeled depending on the number of categories to classify. Information regarding driver, accident and vehicle characteristics, environmental status, and traffic density are used as input. The starting severity levels are: no injury, mild, moderate, severe and fatal. A RF+CART approach has been followed, where the former is used to select the most important variables to use as input for the latter; in this way, the variance is reduced and the classification rate can be improved. CART provides visual tools to easily analyze the relationships between factors. Both methodologies have selected similar ones as important, the next variables are highlighted: type of accident, type of vehicle and traffic density.

The type of accident is crucial for the severity. Derailments, head-on collisions, inside vehicle falls, multiple accidents, obstacle collisions, rollovers and side-impacts are associated to severe or fatal outcomes, whether animal collisions, chain reactions, falls (2 wheels), rear-end collisions, sideswipes and other less frequent type of accidents are associated to moderate or milder outcomes. These results are in line with the exploratory analysis. Regarding the type of vehicle, whenever a 2-wheeled vehicle, car or heavy vehicle is involved to a derailment, head-on collision, inside vehicle fall, multiple accident or rollover the associated severity is the highest. Also, traffic density has been identified as influential whenever the severity can be classifier between severe and fatal. In a case where the type of accident is one of the above-mentioned and the traffic is very fluid injury severity greatly increases. Other factors are important in more complex combinations. For example, in a sideswipe type of accident where the driver is younger than 34 years old, abnormal levels of pressure are associated to a higher severity. In this same case, whenever the level of pressure is higher and the accident has happened in a specific range of months and districts, accidents during the night are more severe.

The performance (area under the ROC curve) of the models that classify between non-severe and severe/fatal accidents is decent, as they quite outperform random models. Models that have had to classify between no injury, mild and moderate levels have not performed as well, which indicates that the factors studied are not enough to explain the differences between lower degrees of injury severity. Still, all the models have performed better than random classifiers.

There are some limitations that need to be discussed, which may affect the results obtained. First, this study is based on a two-year urban accidents database (ensembled with five datasets), where the number of observations involving severe or fatal outcomes is noticeably limited, resulting in a huge imbalance. In addition, each subset from the database has a different number of entries with the same casualty record and slightly differences in some features, which can lead to duplicated observations and greatly affect the results. In order to avoid duplicates, random entries (with the same casualty record) from a given subset have been dropped; thus, some information has been lost. Also, the scope of the analysis is really wide due to the fact that multiple types of accidents and vehicles are studied at the same time. For this reason, it is recommended to focus on specific types of accidents and vehicles to improve the performance.

Acknowledgements

To begin with, I would like to thank the community of *StackOverflow* for having an answer for almost every issue in relation to coding.

I would also like to thank my tutor, Jordi Olivella, for proposing this project and guiding me during it.

Above all, I must thank my family for supporting me through times that have been both good (usually) and not so good (tough months with health issues).

References

Behnood, A., & Mannering, F. (2017). The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach. *Analytic Methods in Accident Research, 14*, 41-53.

Boslaugh, S. (2013). *Statistics In A Nutshell.* O'Reilly.

Chang, L.-Y., & Chien, J.-T. (2013). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. *Safety Science, 51*(1), 17-22.

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.

Chen, C., Guohui, Z., Rafiqul, T., Jianming, M., Heng, W., & Hongzhi, G. (2015). A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. *Accident Analysis & Prevention, 80*, 76-88.

Chen, C., Zhang, G., Qian, Z., Rafiqul, A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention, 90*, 128-139.

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention, 38*, 434-444.

Dirección General de Tráfico (DGT). (2016). *Datos agregados de Barcelona 2015.* Retrieved from http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/informacion-municipal/agregados-provincias/barcelona.shtml

Europen Road Safety Observatory (ERSO). (2018). *Annual Accident Report.* Retrieved from https://ec.europa.eu/transport/road_safety/specialist/statistics_en#

Huelke, D. F., & Compton, C. P. (1983). Injury frequency and severity in rollover car crashes as related to occupant ejection, contacts and roof damage: —An analysis of national crash severity study data—. *Accident Analysis & Prevention, 41*(4), 395-401.

Institute for Health Metrics and Evaluation (IHME). (2017). *Global Burden of Disease Collaborative Network. Global Burden of Disease Study.* Retrieved from http://ghdx.healthdata.org/gbd-results-tool

James, G., Witten, D., Haste, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning.* Springer.

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian, Z., . . . T.Ma, D. (2019). Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. *Accident Analysis & Prevention, 124*, 219-229.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44*(5), 291-305.

Ng, K.-s., Hung, W.-t., & Wong, W.-g. (2002). An algorithm for assessing the risk of traffic accident. *Journal of Safety Research, 33*, 387-410.

Norman, G. R., & Streiner, D. L. (2003). *PDQ Statistics.* PMPH USA.

Oña, J. d., López, G., Mujalli, R., & J.Calvo, F. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention*, 1-10.

Pillajo-Quijia, G., Arenas-Ramírez, B., González-Fernández, C., & Aparicio-Izquierdo, F. (2020). Influential Factors on Injury Severity for Drivers of Light Trucks and Vans with Machine Learning Methods. *Sustainability*.

United Nations (UN). (2015). *Transforming our World: The 2030 Agenda for Sustainable Development.* Retrieved from Transforming ourWorld: The 2030 Agenda for Sustainable Development.

Wim Wijnen, W. W. (2019). An analysis of official road crash cost estimates in European countries. *Safety Science, 113*, 318-327.

World Health Organization (WHO). (2016). *Disease burden and mortality estimates.* Retrieved from https://www.who.int/healthinfo/global_burden_disease/estimates/en/

Yau, K. K. (2004). Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. *Accident Analysis & Prevention*, 333-340.