safety

Article

A Rare Event Modelling Approach to Assess Injury Severity Risk of Vulnerable Road Users

Mariana Vilaça *, Eloísa Macedo and Margarida C. Coelho

Centre of Mechanical Technology and Automation, Department of Mechanical Engineering,
University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal; macedo@ua.pt (E.M.);
margarida.coelho@ua.pt (M.C.C.)

* Correspondence: mvilaca@ua.pt; Tel.: +351-234-378-172

Received: 12 March 2019; Accepted: 5 May 2019; Published: 7 May 2019

Abstract: Vulnerable road users (VRUs) represent a large portion of fatalities and injuries occurring on European Union roads. It is therefore important to address the safety of VRUs, particularly in urban areas, by identifying which factors may affect the injury severity level that can be used to develop countermeasures. This paper aims to identify the risk factors that affect the severity of a VRU injured when involved in a motor vehicle crash. For that purpose, a comparative evaluation of two machine learning classifiers—decision tree and logistic regression—considering three different resampling techniques (under-, over- and synthetic oversampling) is presented, comparing both imbalanced and balanced datasets. Crash data records were analyzed involving VRUs from three different cities in Portugal and six years (2012–2017). The main conclusion that can be drawn from this study is that oversampling techniques improve the ability of the classifiers to identify risk factors. On the one hand, this analysis revealed that road markings, road conditions and luminosity affect the injury severity of a pedestrian. On the other hand, age group and temporal variables (month, weekday and time period) showed to be relevant to predict the severity of a cyclist injury when involved in a crash.

Keywords: road crashes; vulnerable road users; imbalanced data; injury severity; logistic regression; decision tree; machine learning

1. Introduction

Road crashes are among the leading causes of death, disability, property loss and yield costs to society, representing 1–3% of GDP worldwide [1]. More than one million people lose their lives every year in road crashes and 20 to 50 million people are injured [1]. Pedestrians and cyclists are vulnerable road users (VRUs) since they are unprotected and they represent the majority of people killed and injured on the European Union (EU) roads [2]. Despite long-term trends in reducing death and injury rates, in 2017, 21% of fatalities on European Union roads were pedestrians and 8% were cyclists, decreasing at a lower rate than other fatalities [3]. For Portugal, in 2017, the percentage of VRU fatalities were 25% of the total (21% being pedestrians and 4% being cyclists) [4].

The transportation systems are becoming more sophisticated and confront more risks. This situation increases the difficulty of regulators to ensure safety [5]. There are many factors related to road crash risks, such as human factors, environmental conditions, roadway infrastructure, traffic characteristics and vehicle conditions. However, the identification of risk factors that can contribute to the injury severity of a specific type of road user may be different [6,7]. For that purpose, it is particularly important to give special attention to VRUs' safety, by providing a better understanding of the factors affecting the outcome in terms of injury severity [8].

Predicting road crashes and finding patterns from crash registrations is an important step to develop safety measures. Among a lot of research into the evaluation of crash likelihood and frequency, several studies have focused on factors affecting the injury severity [9–18]. Notwithstanding the personal injury, the economic and societal cost of crashes varies substantially based on the severity level of injury. Some of these studies have focused their investigation on identifying which variables can increase the injury severity of a crash involving pedestrians [19–24], cyclists [25–30] or provide a joint analysis of pedestrian and cyclist risk factors [31–40]. From these studies, several risk factors have been reported to affect the injury severity of VRUs. For instance, pedestrians' and cyclists' age and gender and specific hours (specially related with night-time) were generally identified as risk factors [19,20,23,25,26,34,38]. It has also been recognized that the road environment, in particular vehicle speed, mixed land use and intersections, can affect the level of injury severity of these road users [31,33,41]. Few studies have recognized weather conditions (namely fog and rain) as risk factors considering pedestrians' injury severity [22,24]. Moreover, the month was considered an important temporal variable to predict the severity of cyclist injury [29].

There exists a variety of methodologies that have been used to investigate injury severity data. The logistic regression and decision trees are two predictive/classification techniques that have been widely explored. In particular, logistic regression is often used for binary response variables and decision trees is a nonparametric method that does not require prior probabilistic knowledge on the phenomena under study [42–44]. To predict crash severity, logistic regression is among the most popular techniques [11,19,32,34,37]. This fact can be explained since multinomial logistic regression does not require premises and these model are easily interpreted by their coefficients, giving the influence of the particular input fields and predefined relationships between dependent and independent variables [45,46]. Decision trees have also been a useful methodology for analysing traffic crashes based on the level of injury severity [9,29,47].

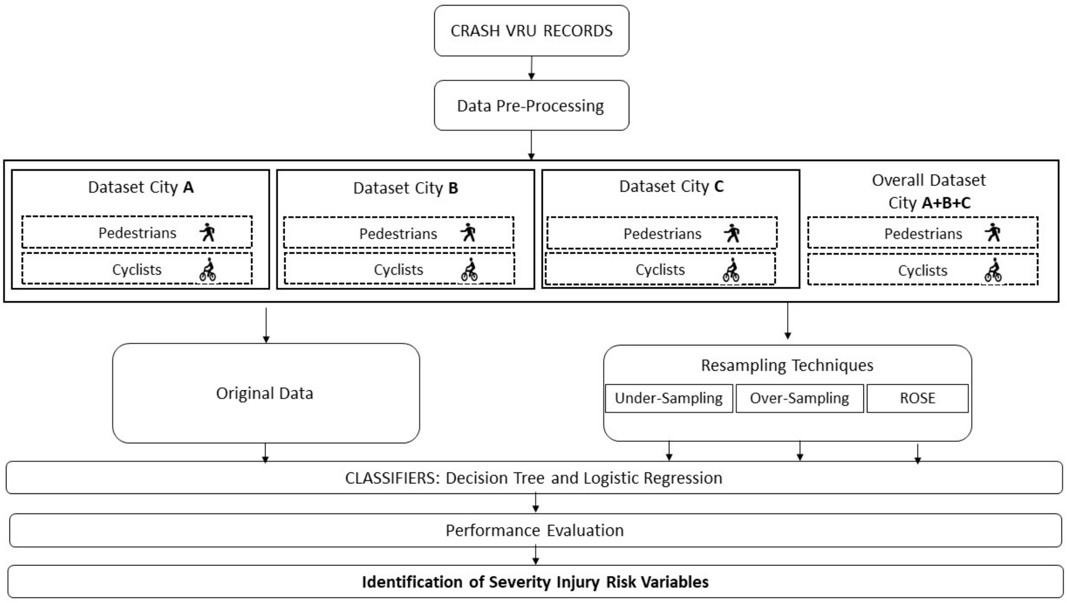
Class imbalance problem is one of the most fundamental challenges considering the learning process and has been found in different domains. There are several examples where resampling techniques may improve the performance of a classifier [48–52]. Hence, many resampling techniques can be applied to balance a dataset; the best approach is hardly dependent on the type of dataset and the technique applied [53,54]. A two-class dataset is said to be imbalanced when one of the minority class is underrepresented regarding the majority class. If it is not suitably addressed, the prediction results would be asymmetrically biased by the weight of the majority class and might produce misleading conclusions [54]. In general, road crash databases are limited and imbalanced in terms of injury severity level since the majority of records are non-severe. Due to the inherent complex characteristics of imbalanced data, some studies have been developed considering resampling techniques applied to the prediction of crash severity risk factors. Pei et al. [55] founded a more accurate and reliable prediction model to determine crash risk factors and crash severity factors applying a bootstrap resampling approach to Poisson regression. Mujalli et al. [56] compared three different data balancing techniques—random undersampling, synthetic minority oversampling (SMOTE) and a combination of both sampling techniques—with three different Bayes classifiers. The results show that the classification of a crash according to its severity improved with the use of oversampling techniques [56]. Delen et al. [13] developed a fusion-based sensitivity analysis on several predictive algorithms to identify risk factors that contribute to the injury severity in motor vehicle crashes using an undersampling method for balancing data. On the other hand, Mussone et al. [57] applied an oversampling technique in order to develop a back propagation neural network model and a generalised linear mixed model to identify factors affecting crash severity level. Madhar et al. [58] show that the prediction accuracy of different data-mining techniques improves when using a resampling dataset to bias the accident severity distribution. More recently, Al-Radaideh and Daoud [47] applied down-sample, up-sample and SMOTE functions to a road traffic accidents dataset to perform three classification techniques—decision trees, Artificial Neural Network (ANN) and Support Vector Machine (SVM)—revealing the best performance for decision trees using a hybrid sampling dataset.

Jiang et al. [59] introduced three methods to model unbalanced data—random forest, AdaBoost and Gradient Boost—showing that the latter two generate more balanced prediction accuracies.

The novelty of this study is the suggested methodological approach, which applies different resampling techniques to imbalanced pedestrian and cyclist motor vehicle crash datasets to perform a comparative evaluation of two commonly used but different classifiers: decision tree and logistic regression. For that purpose, road crash records involving a motor vehicle and pedestrians/cyclists from six years (2012–2017) and three different cities were used. This database was organized in injury severity level, which is classified into severe (which includes serious injuries and fatalities) and non-severe (light injuries). Since the proportion of the minority class is significantly lower than the majority class, the dataset is imbalanced. Thus, different resampling techniques (under-, over- and synthetic oversampling) were applied. The best resampling method is selected based on the classifier performance through receiver operating characteristics (ROC) curves. The developed models allow the identification of risk factors that can affect pedestrian or cyclist injury severity when involved in a motor vehicle crash. These findings can be used as a tool for local authorities to develop road safety strategies. From the variables under evaluation, we should highlight the road conditions and markings sometimes neglected in the literature.

2. Methodology

This section describes the techniques used for data resampling, followed by a short description of the classifier methods: decision tree and logistic regression. The classifiers will be applied to identify the variables which are statistically significant in predicting the injury severity of a VRU involved in a crash. Lastly, data characteristics and pre-processing as well as case studies are described. The conceptual framework designed for this study is presented in Figure 1. This process was applied for each city individually and a global dataset including all crash and injury information.



Legend: VRU—Vulnerable Road User ROSE: Bootstrap Random Examples technique.

Figure 1. Overview of the methodology process.

In order to perform the proposed methodology, an open-source software for statistical computing, R software [60], is employed to handle the crash dataset using specific packages for resampling imbalanced data and for applying the two classifiers.

2.1. Resampling Techniques

Crash records can be considered an imbalanced dataset, since the target variable (injury severity) is predominantly imbalanced, with the majority of instances belonging to the non-severe class and only a small percentage of the instances in the severe class. Resampling is a commonly used data level approach to deal with class imbalance [53,54]. Resampling methodologies are processes of continuously drawing samples from a dataset and refitting a given model. These methodologies can be divided into undersampling and oversampling [54].

Three resampling techniques are proposed for application to such a dataset in order to construct a more balanced one. In this study, the widely used random undersampling and random oversampling methods as well as a synthetic oversampling method were analysed. Since we are interested in using open-source software to perform our analysis, the ROSE R package [61] was explored, since it has many built-in resampling techniques. The best resampling method was then selected based on the classifier performance (see Section 2.2) through ROC curves.

2.1.1. Undersampling

The undersampling method consists of constructing a balanced dataset by randomly removing instances from the majority class until the desired ratio has been reached in order to adjust a class distribution of a dataset [41]. The main advantage of an undersampling method is the reduction of a training data size when the original data is large. On the other hand, removing instances may imply a loss of valuable information of the majority class [52].

2.1.2. Oversampling

In the oversampling method, the balanced dataset is constructed by randomly duplicating instances from the minority class until the desired ratio has been reached [41]. The advantage of an oversampling technique is that it leads to no information loss [52]. Therefore, although it is widely used, oversampling might be ineffective at improving recognition of the minority class and may lead to overfitting [54].

2.1.3. ROSE

ROSE (bootstrap random oversampling examples technique) is a method that generates a synthetic sample from the feature space around the minority class according to a smoothed-bootstrapping approach. According to this, ROSE combines oversampling and undersampling by generating an augmented sample of the data (mainly belonging to the minority class). Three steps are involved in the development of ROSE methodology: (1) Resampling data of the majority class using a bootstrap resampling technique to remove instances of the majority class considering a ratio of 50%—undersampling; (2) Repeat the same process for the minority class—oversampling; (3) Generate a new synthetic data in its neighbourhood, where the shape is determined by a function provided by the ROSE R package. A new synthetic training sample of approximately equal size to the original dataset is generated [62].

Studies have been showing that generating new synthetic data to balance a skewed dataset is an alternative to the above resampling techniques, and is being associated to a reduction of the risk of overfitting and an improvement of the ability of generalisation compromised by the oversampling methods [63].

2.2. Supervised Learning Classifiers

In order to identify risk factors significantly affecting the VRU injury severity, two classification techniques were explored and the results compared. Classifiers can be trained using historical data with the known outcome to predict an associated class. A trained model aims to be able to classify unseen new data correctly.

In this study, the dependent variable (injury severity) presents a binary classification with two possible outcomes (non-severe or severe). Two widely used supervised classification techniques will be explored, namely decision tree and logistic regression.

A stratified holdout procedure was applied to split data into training and testing sets, which ensures each class is represented in both sets. The training set (70% of instances) is used to build the models, which are then tested over the testing set (the remaining 30% of instances) to evaluate its predictive accuracy [64]. The classifier performance is evaluated through receiver operating characteristics (ROC) curves [65,66]. The ROC curve represents the relationship between both sensitivity and specificity in a graphical representation of the true positive (i.e., a severe injury correctly classified) rate against the false positive rate. The overall performance can be given with the area under the ROC curve (AUC), a summary measure that allows one to quantify how accurately a model can discriminate. In particular, an AUC under 0.50 reflects a poor model. Hence, a higher AUC score represents a better classifier. Moreover, when comparing models, the best model is the one that yields the dominant ROC curve (most significant AUC).

The following sections briefly describe the two classifiers used to develop the models.

2.2.1. Decision Tree

Decision tree methodology is a commonly used nonparametric data-mining method. It classifies instances by sorting them based on attribute values. Each node in a decision tree represents a feature in an instance to be classified and is a test on an attribute. Each branch represents a value that the node can assume and is an outcome of the test. Finally, a leaf node represents a class label. The feature that best divides the training data would be the root node of the tree. At each node, one attribute is chosen to split training examples into disjoint classes as much as possible. This procedure is repeated on each partition of the divided data, resulting in subtrees until the training data is divided into subsets of the same class. Thus, a decision tree classifies an instance as belonging to a specific class by following a suitable path from the root to a leaf node, which represents a classification rule [67]. The advantages of using a decision tree model are threefold: it requires minimal knowledge of the underlying data relationships, provides useful information regarding the most important variables in the dataset that are placed as top nodes and is less sensitive to missing data and outliers [68].

2.2.2. Logistic Regression

Logistic regression is a linear, parametric method for binary classification. The logistic regression method is used to explain the relationship between the dependent variable and the independent variables and has been the most commonly used statistic method for studying injury severity risk factors [13]. The outcomes of the regression equation can vary without limit, but constrain the predictions of the dependent variable to values between 0 and 1. The multiple binary logistic regression model expression [69] is given by:

$$\pi(x) = \left(\exp\left(x^T\beta\right)\right)/\left(1 + \exp\left(x^T\beta\right)\right), \tag{1}$$

where x is the vector of the explanatory variables, β is the vector of the coefficients of the model and $\pi(x)$ is the probability of a severe VRU injury. The logistic regression model can be used for continuous and/or categorical explanatory variables as well as interaction terms to investigate potential combined effects of the explanatory variables. In fitting the data, logistic regression fits a straight line to divide the space into two. A single linear boundary can sometimes be limiting for logistic regression.

2.3. Data Description and Case Studies

A crash dataset involving pedestrians and cyclists from three different cities of Portugal was originally acquired from the National Authority of Road Safety (ANSR). Crash registrations correspond to six years (from 2012 to 2017), which gave a total of 6876 observations. These crashes yielded 7155 injured VRUs, 86% corresponding to injured pedestrians and 14% to cyclists. The original dataset contains specific information about the number and severity of injuries, gender and age of the injured, temporal information (year, month, day and hour), location/position by address and geocode, road characteristics, weather conditions, and luminosity information.

The dataset covers three different cities located in the north, centre, and centre-south of Portugal, namely Aveiro, Porto and Lisbon. These case studies were chosen based on their differences in terms of land use, transport demand and demographic contexts, and also due to their relatively high share of walking and cycling modes, which vary between 19% and 22% for pedestrians and 0.2% and 3% for cyclists [70].

The previously described methodology was applied to three different datasets considering each city. Afterward, the same process was applied to a third dataset considering all recorded samples of each city, which yielded an overall perspective.

2.4. Pre-Processing Data

The analysis focused on the injury severity level, which is subdivided into two classes: non-severe (light injuries) and severe (including serious injuries and fatalities). In order to have a representative sample with common characteristics, records with missing information or uninjured VRUs were removed from the dataset. This preliminary step eliminated 1.5% of the records. Hence, the dataset used in this study contains a total of 7048 injured VRUs, 6% being categorised into severe injuries or fatalities and 94% into the non-severe injury class. The crash dataset represents an imbalance of a 1/16 ratio considering non-severe and severe injuries.

In order to identify the main factors that can significantly affect the injury severity of a VRU involved in a crash, 10 independent variables were selected. Table 1 describes all the variables analysed.

Table 1. Description of the variables and number of reported injuries classified by severity.

Variable	Code	Description	Aveiro		Porto		Lisbon	
			NSI	SI	NSI	SI	NSI	SI
Gender	0	Male	245	25	942	44	2072	178
	1	Female	218	19	1062	27	2097	119
Age	1	≤11 years old	22	4	89	3	213	11
	2	12–17 years old	29	2	146	4	291	14
	3	18–24 years old	75	4	235	7	544	28
	4	25–49 years old	140	15	572	11	1298	87
	5	50–65 years old	102	8	472	14	820	51
	6	>65 years old	95	11	490	32	1003	106
Month	1	January	38	4	159	6	340	30
	2	February	33	3	150	8	327	28
	3	March	36	3	172	4	325	25
	4	April	35	3	133	7	328	10
	5	May	35	6	162	4	391	30
	6	June	38	3	165	4	326	20
	7	July	37	2	187	3	344	24
	8	August	39	5	117	3	255	24
	9	September	37	5	201	9	387	21
	10	October	43	2	194	10	389	27
	11	November	55	3	182	5	387	29
	12	December	37	5	182	8	370	29
Weekday	1	Sunday	42	3	150	6	342	29
	2	Monday	66	8	334	10	619	38
	3	Tuesday	80	3	297	12	671	44
	4	Wednesday	70	5	341	11	682	46
	5	Thursday	88	9	328	16	715	50
	6	Friday	65	10	342	9	723	56
	7	Saturday	52	6	212	7	417	34

Table 1. *Cont.*

Variable	Code	Description	Aveiro		Porto		Lisbon	
			NSI	SI	NSI	SI	NSI	SI
Time	1	00:00–06:00 h	22	4	67	4	222	22
	2	07:00–10:00 h	112	9	418	18	947	38
	3	11:00–15:00 h	123	6	658	25	1256	91
	4	16:00–19:00 h	163	18	637	14	1302	82
	5	20:00–23:00 h	43	7	224	10	442	64
Weather	0	Bad	71	5	315	14	472	35
	1	Good	392	39	1689	57	3697	262
Luminosity	1	Daylight	344	29	1519	50	3074	182
	2	Sun glare	1	1	8	1	20	3
	3	Dawn or dusk	10	0	29	0	131	13
	4	Night with road lights	91	10	419	19	925	92
	5	Night without road lights	17	4	29	1	19	7
Road Conditions	1	Good	239	19	1306	34	1998	143
	2	Regular	219	22	691	37	2080	151
	3	Bad	5	3	7	0	91	3
Road Markings	1	Without	183	16	382	7	1736	71
	2	Separating directions	121	12	384	7	468	41
	3	Separating directions and lanes	159	16	1238	57	1965	185

NSI: non-severe injuries; SI: severe injuries.

The main attributes considered in the forthcoming analyses are:

- VRU profile: gender and age group;
- Temporal variables: month, weekday and time period;
- Weather conditions: subdivided into good or bad (including any adverse situation, e.g., rain, fog, snow, strong winds);
- Luminosity: subdivided based on the national authority classification as daylight, sun glare, dawn or dusk, night with road lights or night without road lights;
- Road characteristics: describing the conservation conditions of the pavement (road conditions) and the presence of road surface markings for separating directions or directions and lanes (road markings).

This exploratory step is often crucial for obtaining a good fit of the model and better predictive ability.

3. Results

The results are presented considering two different aims:

1. To evaluate the most efficient prediction model based on three resampling techniques (undersampling, oversampling and ROSE);
2. To explore and compare the results of two supervised classification techniques in order to identify which variables can significantly affect pedestrian and cyclist injury severity when involved in a motor vehicle crash.

The three resampling techniques were applied to the datasets, resulting in six different datasets for each city and the overall perspective. Table 2 shows an overview of the dataset modifications and their distribution amongst the severity classes.

The two classifiers described in Section 2.2 were used to develop injury severity prediction models. Therefore, two models were developed for each dataset considering the different case studies (Aveiro, Porto, Lisbon and overall). For the developed models, the datasets were divided into training and test sets considering a 2/3 ratio, as described in the methodology. A total of 64 models were developed. The performance of the models was examined based on the area under the ROC curve (AUC). Table 3 shows the AUC results for the developed models.

Table 2. Number of injuries and severity distribution for the different datasets.

Datasets	**Total**		**NSI**		**SI**	
Aveiro	🚶	🚴	🚶	🚴	🚶	🚴
Original	249	258	222	241	27	17
Undersampling	54	34	27	17	27	17
Oversampling	444	482	222	241	222	241
ROSE	249	258	126	131	123	127
Porto						
Original	1849	226	1780	224	69	2
Undersampling	138	4	69	2	69	2
Oversampling	3560	448	1780	224	1780	224
ROSE	1849	226	959	110	890	116
Lisbon						
Original	3990	476	3713	456	277	20
Undersampling	554	40	277	20	277	20
Oversampling	7426	912	3713	456	3713	456
ROSE	3990	476	2060	257	1930	219
Overall						
Original	6088	960	5715	921	373	39
Undersampling	746	78	373	39	373	39
Oversampling	11430	1842	5715	921	5715	921
ROSE	6088	960	3085	497	3003	463

Key: 🚶 represents pedestrians; 🚴 represents cyclists.

Table 3. AUC values for comparison of models' performance. Bold formatting depicts the best-performing technique of a dataset.

Decision Tree	**VRU**	**Original**	**Undersampling**	**Oversampling**	**ROSE**
Aveiro	🚶	0.576	0.611	**0.773**	0.536
	🚴	0.505	0.667	**0.839**	0.635
Porto	🚶	0.538	0.524	**0.796**	0.656
	🚴	0.547	0.500 [1]	0.962	**0.974**
Lisbon	🚶	0.558	**0.671**	0.660	0.636
	🚴	0.574	0.571	**0.931**	0.547
Overall	🚶	0.500 [1]	0.584	**0.624**	0.615
	🚴	0.614	0.552	**0.894**	0.672
Logistic Regression	**VRU**	**Original**	**Undersampling**	**Oversampling**	**ROSE**
Aveiro	🚶	0.547	0.580	**0.652**	0.540
	🚴	0.561	0.556	**0.772**	0.738
Porto	🚶	0.652	0.512	0.684	**0.692**
	🚴	0.680	- [2]	**0.861**	0.814
Lisbon	🚶	0.653	0.694	**0.697**	0.679
	🚴	0.578	0.510	**0.692**	0.623
Overall	🚶	0.630	0.649	**0.662**	0.619
	🚴	0.539	**0.667**	0.631	0.584

[1] Difficulty in the learning process: the resulting model suggests no discrimination ability to classify an instance as severe or non-severe; [2] The sample size was too small to perform the analysis; VRU—Vulnerable Road User.

In general, results showed that applying resampling methods in a class-imbalanced dataset tends to improve the classification power of the classifiers to discriminate between severe and non-severe VRU injuries when involved in a motor vehicle crash. Improvement in the classification power can be verified for oversampling techniques for both classifier models; however, this is not always the case regarding the undersampling technique and ROSE.

The best results (highlighted in Table 3) revealed that oversampling is the best resampling technique for Aveiro, independent of the classifier used and the VRU under study. Regarding Porto, for the pedestrians database, the oversampling technique was revealed to improve the classifier power of a decision tree and ROSE yielded the best performance for logistic regression. On the other hand, the cyclist database of Porto revealed ROSE as the best technique when a decision tree is applied and oversampling when the logistic regression is applied. Regarding the Lisbon and overall cases, considering the pedestrian databases, oversampling is the best technique, except in the case of the Lisbon pedestrian database, where the decision tree classifier is applied. Considering the cyclist database for Lisbon, oversampling is the best resampling technique for both classifiers. Besides, this technique also presents the best performance when considering the overall case for the cyclist database, apart from logistic regression, where undersampling is the best approach.

From Table 3, it can be seen that the decision tree performed more accurately for predicting VRU injury severity level for Aveiro and Porto, showing an appropriate predictive ability of 77% to 80% for pedestrians and 84% to 97% for cyclists. The same can be verified for Lisbon and the overall perspective regarding the cyclist database when the decision tree presents a predictive ability between 89% and 93%. Considering the pedestrian database, logistic regression presents a better predictive ability of 66% to 70%.

The comparison between the two different classifiers—decision tree and logistic regression—shows that the decision tree presents the best performance results.

The risk variables that can affect VRU injury severity were identified. To accomplish this goal, decision tree and logistic regression models were developed. In particular, a decision tree has a subprocess with an attribute weighting scheme; this weight (from 0 to 100) provides attribute importance information considering the occurrence of severe injury. Table 4 presents the results of the decision tree model, considering the three variables with higher importance scores (numbers shown in brackets) for each case study and database approach (original, undersampling, oversampling and ROSE). Some models present only one significant variable since the weight of this variable is 100.

The significant variables presented take into account mainly the results of the best classifier's performance. Regarding Aveiro, the significant variables that present the highest weight considering oversampling (the best classifier performance) are road conditions and road markings. The profile of cyclists (especially age) and temporal variables (such as month, weekday and time) present an important role in identifying severe injuries of cyclists. Regarding the Porto case study, road markings, age and month are the most significant variables considering pedestrians' severe injury. On the other hand, the variables most significant for cyclists are age, gender and month. For the Lisbon case study, road markings are considered the most important variable concerning pedestrians' severe injury, and beyond that, age and month also present significant importance. Regarding cyclists, temporal variables (month, weekday and time) present an important role too. Lastly, the overall dataset results clearly show that road markings, gender and age are the main risk factors affecting the injury severity of a pedestrian involved in a motor vehicle crash, while for cyclists, age, road conditions and luminosity are considered the important variables to predict injury severity.

Concerning the possible risk factors identified using the logistic regression, Tables 5 and 6 present the details of each model. Based on the p-value considered, we can conclude whether or not a variable included in the model is significantly contributing to the model's ability to predict the injury severity level of a VRU. The coefficients are given, followed by the p-value range. Variables with p-values (p) < 0.1 are considered significant.

Table 4. Weight-based contribution of studied variables for the resampling techniques and cities under study (attribute importance information from 0 to 100).

Resampling Techniques	VRU	Aveiro	Porto	Lisbon	Overall
Original	🚶	Age (28) Month (20) Luminosity (14)	Road Markings (100)	Road Markings (93) Age (4) Road Conditions (3)	
	🚴	Age (31) Gender (28) Month (18)	Age (100)	Age (42) Weekday (36) Month (19)	Weekday (47) Age (36) Time (16)
Undersampling	🚶	Month (51) Luminosity (15) Time (10)	Age (36) Month (24) Road Markings (17)	Road Markings (25) Age (24) Month (14)	Month (33) Luminosity (19) Road Markings (19)
	🚴	Month (46) Age (12) Time (12) Luminosity (12) Road Conditions (12)		Road Markings (77) Gender (8) Age (8) Month (8)	Road Conditions (30) Month (22) Age (17)
Oversampling	🚶	Road Conditions (19) Month (17) Age (14)	Month (18) Road Markings (14) Age (14)	Road Markings (42) Time (21) Age (19)	Age (43) Gender (30) Road Markings (20)
	🚴	Weekday (18) Month (17) Age (14) Time (14)	Month (61) Time (17) Luminosity (9)	Weekday (22) Time (22) Month (15)	Age (21) Month (20) Weekday (19)
ROSE	🚶	Gender (41) Road Markings (16) Time (11)	Road Markings (23) Gender (20) Age (18)	Road Markings (42) Luminosity (36) Age (11)	Gender (28) Road Markings (25) Age (20)
	🚴	Gender (31) Age (17) Road Conditions (16)	Month (30) Gender (18) Time (17)	Luminosity (24) Weather (19) Weekday (13) Time (13)	Road Conditions (19) Age (18) Luminosity (16)

Table 5. Statistically significant variables for pedestrians considering the logistic regression model.

Dataset	Resampling Technique	Gender	Age	Month	Weekday	Time	Weather	Luminosity	Road Conditions	Road Markings
Aveiro	**Original**							0.4115 *		
	Undersampling						2.0020 *			
	Oversampling						0.9262 **	0.2262 **	0.4465 **	
	ROSE	−0.8828 ***						0.2323 *		0.4263 **
Porto	**Original**	−0.8858 ***								1.0265 ***
	Undersampling		0.3958 **						0.8862 *	
	Oversampling	−0.9121 ***	0.2355 ***		−0.0515 **	−0.2357 ***	−0.2142 *	0.1561 ***	0.7761 ***	0.6581 ***
	ROSE	−0.5685 ***	0.1181 ***			−0.1227 **		0.1065 ***	0.4855 ***	0.4051 ***
Lisbon	**Original**	−0.3684 **	0.2090 ***				0.4627 *	0.1456 **		0.5239 ***
	Undersampling	−0.3795 *	0.2079 **							0.2833 **
	Oversampling	−0.4822 ***	0.2601 ***	−0.0256 ***		0.1009 ***	0.1849 **	0.1653 ***		0.3945 ***
	ROSE	−0.4406 ***	0.1959 ***			0.1158 ***	0.2369 **	0.1072 ***		0.3403 ***
Overall	**Original**	−0.5991 ***	0.1812 ***					0.16196 ***	0.3441 ***	0.4130 ***
	Undersampling							0.2438 ***	0.2932 *	0.3413 ***
	Oversampling	−0.5041 ***	0.1970 ***	−0.0166 **		0.0609 ***	0.2412 ***	0.1573 ***	0.2865 ***	0.3219 ***
	ROSE	−0.3931 ***	0.1325 ***	−0.020 **		0.0655 **		0.0804 ***	0.1874 ***	0.2808 ***

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Considering the analysis for pedestrians (Table 5), for Aveiro, Lisbon and the overall dataset, oversampling is the best resampling technique applied for the logistic regression model. Only Porto presented ROSE as the best resampling approach when applied to this classifier. Based on this, we can conclude that weather, luminosity and road conditions are statistically significant variables to predict severe injuries of pedestrians for Aveiro. In the Porto case study, gender and age as well as time, luminosity, road conditions and road markings are the risk factors to predict the severity of these VRU injuries. For Lisbon, considering the oversampling approach, we can conclude that VRU profile (gender and age), month, time, weather, luminosity and road markings are the risk factors that contribute to predicting the injury severity of a pedestrian. For the overall case, considering oversampling as the best approach, only weekdays seems not to be considered as a risk factor in predicting pedestrian injury severity.

Table 6. Statistically significant variables for cyclists considering the logistic regression model.

Dataset	Resampling Technique	Gender	Age	Month	Weekday	Time	Weather	Luminosity	Road Conditions	Road Markings
Aveiro	Original		1.0863 **							
	Undersampling									
	Oversampling	−1.0377 ***	1.0251 ***		0.5456 ***	0.7132 ***		−0.4234 ***	1.3226 ***	0.3846 **
	ROSE	−0.9348 ***			0.1494 *				1.2172 ***	
Porto	Original									
	Undersampling									
	Oversampling			−0.9513 ***	−0.6974 ***					−2.0614 ***
	ROSE			−0.9218 ***			5.7643 **			−0.3771 *
Lisbon	Original									
	Undersampling			0.4457 *						
	Oversampling			0.0626 **			0.0342 **	0.1109 *		
	ROSE			0.0748 **		0.1847 **				
Overall	Original									
	Undersampling									
	Oversampling	−0.3509 **	0.2066 ***		0.0669 **	0.2386 ***			0.6649 ***	
	ROSE					0.1407 **				

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Similarly, concerning cyclists, oversampling presented the best approach when a logistic regression classifier is applied to the database, except for the overall perspective, when undersampling was shown to be the best approach. Based on this, for the Aveiro case study, only month and weather are variables that are not statistically significant in predicting injury severity of cyclists. Considering the Porto case study, month, weekday and road markings are the risk factors to predict the injury severity for these VRUs. Regarding Lisbon, the variables that were shown to be risk factors are month, weather and luminosity. For the overall case, considering that undersampling is the best approach, we conclude that any variable seems to be a risk factor that contributes to predicting the injury severity of a cyclist; see Table 6.

Comparing the results obtained from the two models—decision tree and logistic regression—variables such as gender, luminosity and road conditions are considered statistically significant for more datasets when logistic regression is applied to the pedestrian datasets. On the other hand, gender and month are more representative variables considering the decision tree models when applied to the cyclist datasets.

Results highlight that an overall perspective is not always the best approach, considering two main reasons: First, the seriousness of the road crash can be affected by the specificities of each city; secondly, the overall perspective results can be biased from the most prominent database (in this case, the Lisbon database).

4. Discussion

In this paper, the performance of two classifiers (decision tree and logistic regression) to predict risk factors that can affect the severity of a VRU's injury were investigated. To deal with the imbalanced data problem, three resampling techniques were applied: undersamping, oversampling, and ROSE methods. The effectiveness of each resampling method applied to each of two classifiers was evaluated based on AUC as a performance metric.

The results showed that the performance of the classifier can be improved by processing the data with one of those resampling techniques. However, the small samples of VRU crash data may explain why the application of an undersampling technique does not improve the explanatory power of the studied classifiers for almost all the datasets, due to the loss of information related with this resampling technique. Emphasis is given for the Porto case study, where the proportion of severe injuries presented the lowest values (4% for pedestrians and 1% for cyclists).

Regarding the overall performance of the two classifiers, the decision tree slightly outperformed the logistic regression. This can be explained by the fact that the coefficient correlations were not considered and all the variables were analysed. Nevertheless, it is well known that decision tree models are robust to identify outliers, so when a domain problem is given, the decision tree technique naturally captures the relationship between variables, leading to higher classification performance.

Considering the identification of risk factors that can affect the injury severity level of a VRU, having a joint overview of the two classifiers enables finding out some main results. Considering the pedestrian injury severity risk, we can conclude that:

- Gender and age factors seem to play an important role in this type of VRU;
- Road markings are a risk factor considering pedestrian injury severity, especially for bigger cities;
- The luminosity of the road seems to be more important than weather conditions.

On the other hand, considering cyclist databases, the main results allow us to conclude that cyclist age group and month are the main identified risk factors in predicting the injury severity of a cyclist. These results can be related to exposure values, namely the most people of active cycling age and the fluctuation between the number of people cycling in summer and winter. Road conditions seem not to affect the severity of both pedestrian and cyclist injuries in Lisbon and cyclist injuries in Porto.

Although these results are based on a crash database of three cities, the methodology and results can be generalized to small and medium-sized cities, since our results are similar to those reported in the literature review. For instance, the importance of age to most severe outcomes involving pedestrians [19] and the importance of environmental factors, such as time and environmental conditions, are relevant categories to consider in motor vehicle–bicycle collisions [29,40]. Road conditions and surface markings, variables which are sometimes neglected in the literature, are essential factors to be taken into account.

Limitations and Future Research

Due to the complexity of reporting a road crash, our study presents some limitations: (1) more detailed information about vehicle characteristics and drivers are missing in our database; (2) unobserved heterogeneity of data was not considered in the analyses. Future research will try to address these issues. Also, it would be interesting to explore other methods to handle the class imbalance problem, to extend to a national database and to focus our research on some unobserved variables that may affect the prediction models.

5. Conclusions

In this paper, an approach to reveal the most significant risk factors that can possibly affect VRU injury severity when involved in a motor vehicle crash is presented. Since prediction model performance can be biased when imbalanced data is used, three well-known resampling techniques were examined in an attempt to improve the model's classification performance. Two widely used supervised methods were applied: decision tree and logistic regression.

The machine learning classifiers were able to correctly classify both the majority and the minority classes with relatively high accuracy. It is known that the performance of a resampling method depends on the classifier used, and no method would always outperform the other. Nevertheless, the decision tree model revealed to be a more accurate model considering the crash severity data under evaluation.

Results showed that the oversampling technique (used to balance the dataset) always improves the effectiveness of both classifiers (decision tree and logistic regression) to identify risk factors.

The classifiers were applied considering the original and developed dataset based on three different resampling techniques. Based on an attribute weighting scheme presented by decision tree models and p-values < 0.1 considering the logistic regression model, the risk variables that can significantly affect pedestrians and cyclists injury severity were identified. A joint analysis of the obtained results allows us to conclude that road markings, road conditions and luminosity significantly affect the severity of a pedestrian's injury when involved in a crash. On the other hand, age group and temporal variables (month, weekday and time period) are the risk factors that were revealed to be the most significant to predict the severity of a cyclist's injury when involved in a motor vehicle crash.

Furthermore, it should be emphasised that the identification of risk factors is relevant to the development of road safety measures that aim to reduce the injury severity of crashes between VRUs and motor vehicles, which is crucial information to help decision-makers in the definition of road safety policies and strategies.

Author Contributions: Conceptualization, M.V., E.M., and M.C.C.; Resources, M.C.C.; Investigation, M.V., E.M., and M.C.C.; Methodology, E.M., M.V., and M.C.C.; Data curation, M.V. and E.M.; Software, E.M.; Formal analysis, M.V. and E.M; Writing—original draft preparation, M.V. and E.M.; Writing—review and editing, E.M. and M.C.C.; Project administration, M.C.C.; Funding acquisition, M.C.C.

Funding: This research was funded by TEMA (CENTRO-01-0145-FEDER-022083), Strategical Project (UID/EMS/00481/2013—FCT), @CRUiSE project (PTDC/EMS-TRA/0383/2014, funded within Project 9471—Reforçar a Investigação, o Desenvolvimento Tecnológico e a Inovação and supported by the European Community Fund FEDER), MobiWise (P2020 SAICTPAC/0011/2015, cofunded by COMPETE2020, Portugal2020—Operational Program for Competitiveness and Internationalization, European Union's ERDF and FCT), and CISMOB (PGI01611, funded by the Interreg Europe Programme). This work was also financially supported by the project POCI-01-0145-FEDER-029463 (DICA-VE) and POCI-01-0145-FEDER-029679 (InFLOWence), funded by FEDER through COMPETE2020—Programa Operacional Competitividade e Internacionalização (POCI), and by national funds (OE) through FCT/MCTES.

Acknowledgments: The authors acknowledge the Portuguese National Authority of Road Safety (ANSR) for providing crash data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. *Global Status Report on Road Safety 2018*; License: CC BY-NC-SA 3.0 IGO; World Health Organization: Geneva, Switzerland, 2018.
2. European Commission. *Pedestrians and Cyclists*; Directorate General for Transport, February 2018; European Commission: Brussels, Belgium, 2018.
3. European Commission. *2017 Road Safety Statistics: What Is behind the Figures?* European Commission: Brussels, Belgium, 2018; pp. 1–4.
4. ANSR. *Annual Report 2017 (30 Days Victims)—National Authority of Road Safety*; Portuguese National Authority for Road Safety: Lisbon, Portugal, 2017. (In Portuguese)
5. National Academies of Sciences, Engineering, and Medicine. *Critical Issues in Transportation 2019*; National Academies Press: Washington, DC, USA, 2019; ISBN 978-0-309-48676-7.
6. Monsere, C.; Wang, H.; Wang, Y.; Chen, C. *Risk Factors for Pedestrians and Bicycle Crashes—Final Report—SPR 779*; Oregon Department of Transportation: Salem, OR, USA, 2017.
7. Liu, G.; Chen, S.; Zeng, Z.; Cui, H.; Fang, Y.; Gu, D.; Yin, Z.; Wang, Z. Risk factors for extremely serious road accidents: Results from national Road Accident Statistical Annual Report of China. *PLoS ONE* **2018**, *13*, e0201587. [CrossRef] [PubMed]
8. SafetyNet Pedestrians & Cyclists. Available online: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/specialist/knowledge/pdf/pedestrians.pdf (accessed on 17 June 2018).
9. Abellán, J.; López, G.; De Oña, J. Analysis of traffic accident severity using Decision Rules via decision trees. *Expert Syst. Appl.* **2013**, *40*, 6047–6054. [CrossRef]
10. Ma, X.; Chen, S.; Chen, F. Multivariate space-time modeling of crash frequencies by injury severity levels. *Anal. Methods Accid. Res.* **2017**, *15*, 29–40. [CrossRef]
11. Wang, Z.; Yue, Y.; Li, Q.; Nie, K.; Tu, W.; Liang, S.; Wang, Z.; Yue, Y.; Li, Q.; Nie, K.; et al. Analyzing Risk Factors for Fatality in Urban Traffic Crashes: A Case Study of Wuhan, China. *Sustainability* **2017**, *9*, 897. [CrossRef]
12. Ferreira, S.; Amorim, M.; Couto, A. Risk factors affecting injury severity determined by the MAIS score. *Traffic Inj. Prev.* **2017**, *18*, 515–520. [CrossRef] [PubMed]
13. Delen, D.; Tomak, L.; Topuz, K.; Eryarsoy, E. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *J. Transp. Health* **2017**, *4*, 118–131. [CrossRef]

14. Fountas, G.; Anastasopoulos, P.C.; Abdel-Aty, M. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. *Anal. Methods Accid. Res.* **2018**, *18*, 57–68. [CrossRef]
15. Fountas, G.; Anastasopoulos, P.C.; Mannering, F.L. Analysis of vehicle accident-injury severities: A comparison of segment- versus accident-based latent class ordered probit models with class-probability functions. *Anal. Methods Accid. Res.* **2018**, *18*, 15–32. [CrossRef]
16. Fountas, G.; Anastasopoulos, P.C. Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances. *Anal. Methods Accid. Res.* **2018**, *20*, 30–45. [CrossRef]
17. Ramachandiran, V.M.; Babu, P.N.K.; Manikandan, R. Prediction of Road Accidents Severity using various algorithms. *Int. J. Pure Appl. Math.* **2018**, *119*, 16663–16669.
18. Duddu, V.R.; Kukkapalli, V.M.; Pulugurtha, S.S. Crash risk factors associated with injury severity of teen drivers. *IATSS Res.* **2018**, *43*, 37–43. [CrossRef]
19. Senserrick, T.; Boufous, S.; de Rome, L.; Ivers, R.; Stevenson, M. Detailed Analysis of Pedestrian Casualty Collisions in Victoria, Australia. *Traffic Inj. Prev.* **2014**, *15*, 197–205. [CrossRef] [PubMed]
20. Pour-Rouholamin, M.; Zhou, H. Investigating the risk factors associated with pedestrian injury severity in Illinois. *J. Saf. Res.* **2016**, *57*, 9–17. [CrossRef]
21. Xin, C.; Guo, R.; Wang, Z.; Lu, Q.; Lin, P.S. The effects of neighborhood characteristics and the built environment on pedestrian injury severity: A random parameters generalized ordered probability model with heterogeneity in means and variances. *Anal. Methods Accid. Res.* **2017**, *16*, 117–132. [CrossRef]
22. Kim, M.; Kho, S.Y.; Kim, D.K. Hierarchical ordered model for injury severity of pedestrian crashes in South Korea. *J. Saf. Res.* **2017**, *61*, 33–40. [CrossRef]
23. Uddin, M.; Ahmed, F. Pedestrian Injury Severity Analysis in Motor Vehicle Crashes in Ohio. *Safety* **2018**, *4*, 20. [CrossRef]
24. Zhai, X.; Huang, H.; Sze, N.N.; Song, Z.; Hon, K.K. Diagnostic analysis of the effects of weather condition on pedestrian crash severity. *Accid. Anal. Prev.* **2019**, *122*, 318–324. [CrossRef]
25. Kaplan, S.; Vavatsoulas, K.; Prato, C.G. Aggravating and mitigating factors associated with cyclist injury severity in Denmark. *J. Saf. Res.* **2014**, *50*, 75–82. [CrossRef]
26. Chen, P.; Shen, Q. Built environment effects on cyclist injury severity in automobile-involved bicycle crashes. *Accid. Anal. Prev.* **2016**, *86*, 239–246. [CrossRef]
27. Wall, S.P.; Lee, D.C.; Frangos, S.G.; Sethi, M.; Heyer, J.H.; Ayoung-chee, P.; DiMaggio, C.J. The Effect of Sharrows, Painted Bicycle Lanes and Physically Protected Paths on the Severity of Bicycle Injuries Caused by Motor Vehicles. *Safety* **2016**, *2*, 26. [CrossRef]
28. Behnood, A.; Mannering, F. Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances. *Anal. Methods Accid. Res.* **2017**, *16*, 35–47. [CrossRef]
29. Prati, G.; Pietrantoni, L.; Fraboni, F. Using data mining techniques to predict the severity of bicycle crashes. *Accid. Anal. Prev.* **2017**, *101*, 44–54. [CrossRef]
30. Useche, S.; Montoro, L.; Alonso, F.; Oviedo-Trespalacios, O. Infrastructural and Human Factors Affecting Safety Outcomes of Cyclists. *Sustainability* **2018**, *10*, 299. [CrossRef]
31. Zahabi, S.; Strauss, J.; Manaugh, K.; Miranda-Moreno, L. Estimating Potential Effect of Speed Limits, Built Environment, and Other Factors on Severity of Pedestrian and Cyclist Injuries in Crashes. *Transp. Res. Rec. J. Transp. Res. Board* **2011**, *2247*, 81–90. [CrossRef]
32. Torrão, G.; Coelho, M.; Rouphail, N. Modeling the impact of subject and opponent vehicles on crash severity in two-vehicle collisions. *Transp. Res. Rec.* **2014**, *2432*, 53–64. [CrossRef]
33. Amoh-Gyimah, R.; Saberi, M.; Sarvi, M. Macroscopic modeling of pedestrian and bicycle crashes: A cross-comparison of estimation methods. *Accid. Anal. Prev.* **2016**, *93*, 147–159. [CrossRef]
34. Yuan, Q.; Chen, H. Factor comparison of passenger-vehicle to vulnerable road user crashes in Beijing, China. *Int. J. Crashworthiness* **2017**, *22*, 260–270. [CrossRef]
35. Heydari, S.; Fu, L.; Miranda-Moreno, L.F.; Jopseph, L. Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. *Anal. Methods Accid. Res.* **2017**, *13*, 16–27. [CrossRef]
36. Vilaça, M.; Silva, N.; Coelho, M.C. Statistical Analysis of the Occurrence and Severity of Crashes Involving Vulnerable Road Users. *Transp. Res. Procedia* **2017**, *27*, 1113–1120. [CrossRef]

37. Vilaça, M.; Macedo, E.; Tafidis, P.; Coelho, M. Frequency and severity of crashes involving vulnerable road users—An integrated spatial and temporal analysis. In Proceedings of the Annual Meeting Transportation Research Board, Washignton, DC, USA, 12–16 January 2018; pp. 1–17.
38. Salon, D.; McIntyre, A. Determinants of pedestrian and bicyclist crash severity by party at fault in San Francisco, CA. *Accid. Anal. Prev.* **2018**, *110*, 149–160. [CrossRef]
39. Ouni, F.; Belloumi, M. Spatio-temporal pattern of vulnerable road user's collisions hot spots and related risk factors for injury severity in Tunisia. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *56*, 477–495. [CrossRef]
40. Weast, R. Temporal factors in motor-vehicle crash deaths: Ten years later. *J. Saf. Res.* **2018**, *65*, 125–131. [CrossRef]
41. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
42. Rokach, L.; Maimon, O. *Data Mining with Decision Trees: Theory and Applications*; Series in Machine Perception and Artificial Intelligence; World Scientific: Singapore, 2014; Volume 81, ISBN 978-981-4590-07-5.
43. Dovom, H.Z.; Saffarzadeh, M.; Dovom, M.Z.; Nadimi, N. An Analysis of Pedestrian Fatal Accident Severity Using a Binary logistic regression Model. *ITE* **2012**, *82*, 38.
44. Moudon, A.V.; Lin, L.; Jiao, J.; Hurvitz, P.; Reeves, P. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. *Accid. Anal. Prev.* **2011**, *43*, 11–24. [CrossRef]
45. Harrell, F.E. *Regression Modeling Strategies—With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*; Springer Science + Business Media: New York, NY, USA, 2015; ISBN 978-3319194240.
46. Abdulhafedh, A. Incorporating the Multinomial logistic regression in Vehicle Crash Severity Modeling: A Detailed Overview. *J. Transp. Technol.* **2017**, *7*, 279–303. [CrossRef]
47. Al-Radaideh, Q.A.; Daoud, E.J. Data mining methods for traffic accident severity prediction. *Int. J. Neural Netw. Adv. Appl.* **2018**, *5*, 1–12.
48. Japkowicz, N. Assessment Metrics for Imbalanced Learning. In *Imbalanced Learning*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2013; pp. 187–206.
49. López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **2013**, *250*, 113–141. [CrossRef]
50. More, A. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv* **2016**, arXiv:1608.06048.
51. Burnaev, E.; Erofeev, P.; Papanov, A. Influence of Resampling on Accuracy of Imbalanced Classification. In Proceedings of the Eighth International Conference on Machine Vision (ICMV 2015), Barcelona, Spain, 19–20 November 2015.
52. Tantithamthavorn, C.; Hassan, A.E.; Matsumoto, K. The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models. *IEEE Trans. Softw. Eng.* **2018**, 1–20. [CrossRef]
53. Crone, S.F.; Finlay, S. Instance sampling in credit scoring: An empirical study of sample size and balancing. *Int. J. Forecast.* **2012**, *28*, 224–238. [CrossRef]
54. He, H.; Garcia, E.A. Learning from Imbalanced Data. *Knowl. Data Eng. IEE Trans.* **2009**, *21*, 1263–1284.
55. Pei, X.; Sze, N.N.; Wong, S.C.; Yao, D. Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong. *Accid. Anal. Prev.* **2016**, *95*, 512–520. [CrossRef]
56. Mujalli, R.O.; López, G.; Garach, L. Bayes classifiers for imbalanced traffic accidents datasets. *Accid. Anal. Prev.* **2016**, *88*, 37–51. [CrossRef] [PubMed]
57. Mussone, L.; Bassani, M.; Masci, P. Analysis of factors affecting the severity of crashes in urban road intersections. *Accid. Anal. Prev.* **2017**, *103*, 112–122. [CrossRef] [PubMed]
58. Taamneh, M.; Alkheder, S.; Taamneh, S. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *J. Transp. Saf. Secur.* **2017**, *9*, 146–166. [CrossRef]
59. Jiang, L.; Xie, Y. Modelling highly unbalanced crash injury severity data by ensemble methods and global sensitivity analysis. In Proceedings of the Transportation Research Board 98th Annual Meeting, Washington, DC, USA, 13–17 January 2019.
60. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Development Core Team: Vienna, Austria, 2008.

61. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: Random Oversampling Examples. ROSE-Package, Version 0.0-3, License GPL-2, CRAN. 2014. Available online: https://rdrr.io/cran/ROSE/man/ROSE-package.html (accessed on 1 March 2019).
62. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *R J.* **2014**, *6*, 79–89. [CrossRef]
63. Menardi, G.; Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Discov.* **2014**, *28*, 92–122. [CrossRef]
64. Liu, H.; Gegov, A.; Cocea, M. Unified Framework for Control of Machine Learning Tasks towards Effective and Efficient Processing of Big Data. In *Data Science and Big Data: An Environment of Computational Intelligence*; Springer: Cham, Switzerland, 2017; pp. 123–140. ISBN 978-3-319-53473-2.
65. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
66. Akobeng, A.K. Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatr.* **2007**, *96*, 644–647. [CrossRef]
67. Williams, G. *Data Mining with Rattle and R*; Springer: New York, NY, USA, 2011; ISBN 978-1-4419-9889-7.
68. Ryza, S.; Laserson, U.; Owen, S.; Wills, J. *Advanced Analytics with Spark: Patterns from Learning from Data at Scale*; O'Reilly Media, Inc.: Sevastopol, CA, USA, 2017; ISBN 9781491972946.
69. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 3rd ed.; A Wiley-Interscience Publication: Hoboken, NJ, USA, 2013; ISBN 0-471-35632-8.
70. INE. *Censos 2011 Resultados Definitivos—Portugal*; Portuguese Statistics Institute: Lisbon, Portugal, 2011. (In Portuguese)