












Article

Factors That Influence the Type of Road Traffic Accidents: A Case Study in a District of Portugal

Paulo Infante ^{1,2,*}, Gonçalo Jacinto ^{1,2,*}, Anabela Afonso ^{1,2}, Leonor Rego ², Pedro Nogueira ^{3,4}, Marcelo Silva ^{3,4}, Vitor Nogueira ^{5,6}, José Saias ^{5,6}, Paulo Quaresma ^{5,6}, Daniel Santos ⁶, Patrícia Góis ⁷ and Paulo Rebelo Manuel ¹

¹ CIMA, IIFA, University of Évora, 7000-671 Évora, Portugal

² Department of Mathematics, ECT, University of Évora, 7000-671 Évora, Portugal

³ ICT, IIFA, University of Évora, 7000-671 Évora, Portugal

⁴ Department of Geosciences, University of Évora, 7000-671 Évora, Portugal

⁵ Algoritmi Research Centre, University of Évora, 7000-671 Évora, Portugal

⁶ Department of Informatics, ECT, University of Évora, 7000-671 Évora, Portugal

⁷ Department of Visual Arts and Design, EA, University of Évora, 7000-208 Évora, Portugal

* Correspondence: pinfante@uevora.pt (P.I.); gjcj@uevora.pt (G.J.); Tel.: +351-266-745-370 (P.I. & G.J.)

Abstract: Road traffic accidents (RTAs) are a problem with repercussions in several dimensions: social, economic, health, justice, and security. Data science plays an important role in its explanation and prediction. One of the main objectives of RTA data analysis is to identify the main factors associated with a RTA. The present study aims to contribute to the identification of the determinants for the type of RTA: collision, crash, or pedestrian running-over. These factors are essential for identifying specific countermeasures because there is a relevant relationship between the type of RTA and its severity. Daily RTA data from 2016 to 2019 in a district of Portugal were analyzed. A statistical multinomial logit model was fitted. The identified determinants for the type of RTA were geographical (municipality, location, and parking areas), meteorological (air temperature and weather), time of the day (hour, day of the week, and month), driver's characteristics (gender and age), vehicle's features (type and age) and road characteristics (road layout and type). The multinomial model results were compared with several machine learning algorithms, since the original data of the type of RTA are severely imbalanced. All models showed poor performance. However, when combining these models with ROSE for class balancing, their performance improved considerably, with the random forest algorithm showing the best performance.

Keywords: imbalance data; machine learning algorithms; multinomial logit model; ROSE technique; type of road traffic accident



Citation: Infante, P.; Jacinto, G.; Afonso, A.; Rego, L.; Nogueira, P.; Silva, M.; Nogueira, V.; Saias, J.; Quaresma, P.; Santos, D.; Góis, P.; Manuel, P.R. Factors That Influence the Type of Road Traffic Accidents: A Case Study in a District of Portugal. *Sustainability* **2023**, *15*, 2352. <https://doi.org/10.3390/su15032352>

Academic Editors: Griselda López and Randa Oqab Mujalli

Received: 10 December 2022

Revised: 13 January 2023

Accepted: 19 January 2023

Published: 27 January 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Road traffic accidents (RTA) are a reality with great expression and relevant social consequences, and can include at least three types of simultaneous problems: *public security*—for the social, economic, and political consequences; *justice*—due to the high number of crimes of homicide and potential offenses to physical integrity; *health*—due to the impact it can have on the lives and health of those involved and their families.

In 2016, RTA injuries were the eighth world-leading cause of death and are predicted to become the seventh leading cause of death by 2030, with the RTA cost being 1–3% of the gross domestic product (GDP) worldwide [1]. In 2019, Portugal recorded the sixth-highest rate of road fatalities among the 27 members of the European Union (EU), with more than 16 fatalities per million inhabitants than the EU as a whole [2]. In Portugal, the impact caused by RTAs has an economic and social impact equivalent to 1.2% of GDP [3], in addition to the social impact caused by the fatalities and severe injuries from RTAs.

The district of Setúbal, with a special focus on the roads under operation area of the Territorial Command of Setúbal of the National Republican Guard (Comando Territorial de Setúbal da Guarda Nacional Republicana—TC-GNR), in 2018 recorded the highest number of fatalities due to RTAs among all districts in the country. However, the district is not with the highest number of RTAs.

Different types of RTAs may occur under substantially different circumstances and may be associated with predictor variables in different ways. However, crash prediction models that look into the type of RTA have not been developed, possibly due to the difficulty in collecting the necessary data. Kim et al. [4] argued that crash type models are useful for at least three reasons: (a) need to identify locations that are high risk for specific crash types, but that may not be revealed through total crash modeling; (b) countermeasures are likely to affect only a subset of all crashes; and (c) different crash types are usually associated with road geometry and geography, the environment, and traffic variables in different ways.

A better understanding of factors that influence the type of RTA, which have a close relation to their severity, is fundamental to implementing appropriate strategies to improve road safety. The objective of this paper is to identify those factors when the type of RTA is classified as a collision, a crash, or a pedestrian being run over. There is a lack of studies about these types of RTA.

In recent years, several methodological approaches (both statistical approaches and machine learning (ML) algorithms) have been used to analyze RTA data, but in the majority of them, the RTA is seen as a binary variable (see, for example, [5–8], and a systematic literature review can be found in [9]).

Despite the existence of many studies about RTA severity, there is a lack of studies about the types of those accidents, which are closely related to their severity. Knowledge of factors that enhance the existence of pedestrians being run over, a collision or a crash, can be very useful in preventing road accidents.

The logistic regression model was used to determine the possible factors contributing to pedestrian hit-and-run accidents [10]. Road and environmental factors that were found to contribute to the number of pedestrian hit-and-run accidents incidents included lighting and weather condition, road description, median separation, road surface condition and repair, location type of the accident and traffic conditions.

Multinomial logit models were used to predict different types of RTA outcomes [11–13]. These studies focused on different categories of collisions, and did not consider pedestrian running over. Intini et al. [14] used grouped random parameter multinomial logit structure for predicting crash types, at segments and intersections, which allowed the capturing of some specific unobserved characteristics of segments and intersections. Iranitalab and Khattak [15] compared the performance of four statistical and ML methods, including multinomial logit, nearest neighbour classification, support vector machines and random forests, in predicting traffic crash severity classified in four levels, based on the most serious injury outcome in each crash. Christoforou et al. [16] used multivariate probit models to find the effects of various traffic parameters on the type of RTA, classified as a single-vehicle crash, multiple collisions, a rear-end crash with two vehicles and with more than two vehicles and a sideswipe crash.

In Boo and Choi [17], a prediction of the RTA severity was studied using different ML models combined with different over and under-sampling techniques. In Guo et al. [18], a traffic crash risk prediction model was established using logistic regression with the synthetic minority oversampling technique (SMOTE). In Ding et al. [19], a different method to generate synthetic crash data, a deep generative approach, was used for a crash frequency model with heterogeneous imbalanced data. Yu et al. [20] studied the RTA occurrence using a convolutional neural network modeling approach with refined loss functions, to explore the multidimensional, temporal–spatial, correlated crash data and to overcome the low classification accuracy issue brought by the imbalanced data. Rella Riccardi et al. [21] modeled the unmeasured heterogeneity considering a random parameter multinomial logit, and the imbalance was treated by introducing weights, which forced the estimator

to learn on the basis of the importance that was given to a particular severity level. In Vilaça et al. [22] to identify the risk factors that affect the severity of a vulnerable road users injured when involved in a motor vehicle crash, different resampling techniques were used, in particular the random over-sampling examples (ROSE) approach, to deal with the imbalance data problem. Rella Riccardi et al. [23] used a mixed logit model to identify each pattern on the pedestrian crash occurrence. Their results show that driver behavior and psychophysical status are highly related to an overrepresentation of pedestrian collisions.

The main purpose of this paper is to identify some determinants for the types of RTA: collision, crash, or pedestrian running-over. We analyze the daily data about RTAs that occurred between 2016 and 2019, in the district of Setúbal (Portugal). A statistical multinomial logit model and several ML algorithms are used. However, the imbalanced nature of the data can compromise the performance of these methods. Therefore, another aim of this study was to combine the previous models with methods for data balancing and compare the results. To deal with imbalanced data, we use an adaptation of ROSE, a bootstrap-based technique that aids the task of binary classification in the presence of rare classes.

This paper is organized as follows. Section 2 presents the study area, data description and the methodology used in the paper. Section 3 presents the results of the statistical multinomial logistic regression and ML methods for the type of RTA with and without imbalanced data. Section 4 presents the final remarks.

2. Materials and Methods

2.1. Study Area

Setúbal is the eighth-largest district in Portugal with a land area of 5064 km² divided into 13 municipalities and six protected natural areas. The northern part of this district is characterized by high population density and high traffic flow during the week, mainly toward Lisbon, the capital of Portugal. The remaining areas are sparsely populated and have little traffic during the week. The district is crossed by important access roads to Lisbon and to important tourist spots, which increase the traffic flow during the summer holidays and weekends.

This district contains approximately 293 km of National Road (EN—Estrada Nacional), 219 km of Highway (AE—Autoestrada), 19 km of Principal Itinerary (IP—Itinerário Principal), 90 km of Complementary Itinerary (IC—Itinerário Complementar) and the bridges Vasco da Gama and 25 de Abril that cross the Tagus River in Lisbon. The TC-GNR of Setúbal has a jurisdiction area of approximately 96% of this territory, including responsibility for the Vasco da Gama Bridge. A map of the geographical setting of the study area and its municipalities is presented in Figure 1.

Between 2016 and 2019, the district of Setúbal was fifth out of 18 with the highest number of RTA in Portugal, and fifth in the number of serious injuries and deaths as a result of the RTA. There were reported 50,726 vehicles involved in a total of 28,103 RTA, of which 407 were RTAs resulting in serious injury or death (183 casualties died) and 5436 RTAs were with minor injuries.

2.2. Data

This work analyzes historical RTAs that occurred in the district of Setúbal, between 1 January 2016 and 31 December 2019, and were reported in the Statistical Bulletin of Road Accidents (BEAV [24]). The pandemic due to COVID-19 dramatically changed the number of accidents, and therefore we chose to exclude data already available for the year 2020. This bulletin has information to characterize the circumstances in which the RTA occurred, as well as the people and vehicles involved in the accident [24]. The data were collected and validated by the TC-GNR of Setúbal. Information about 30-day victims was given by National Road Safety Authority (Autoridade Nacional de Segurança Rodoviária—ANSR). Meteorological information at the time and place of the accident was provided by Portuguese Institute for Sea and Atmosphere (Instituto Português do Mar e da Atmosfera—

IPMA) at the meteorological station closest to the accident. Some information about the road characteristics was provided by Portugal Infrastructures (Infraestruturas de Portugal).

Due to some missing observations and incorrect registration of some data, only 28,002 RTA were considered. The accidents were classified by their type, that is, the situation that led to the occurrence of the accident [24]:

- Crash ($n = 5226$)—an RTA in which the driver loses control of the vehicle, being able to deviate or leave the traffic lane or carriageway in which it circulates and/or collide with other users of the public road or obstacles outside the carriageway (includes pavement, vehicle stops, poles, vertical and light signs and other road equipment or trees, rocks, etc.).
- Collision ($n = 21,800$)—crash resulting from a conflict situation between a moving vehicle and other vehicle(s) (moving, stopped or parked) or obstacles in the carriageway (includes eyelets, dividers, fences, safety guards, center plates and other road equipment or potholes, rocks, etc.).
- Pedestrian running over ($n = 976$)—accident resulting from a conflict situation between a moving vehicle and a pedestrian or animal. It does not include situations in which the pedestrian or animal contributed to the occurrence of the accident, but was not hit by the vehicle (there was no collision).

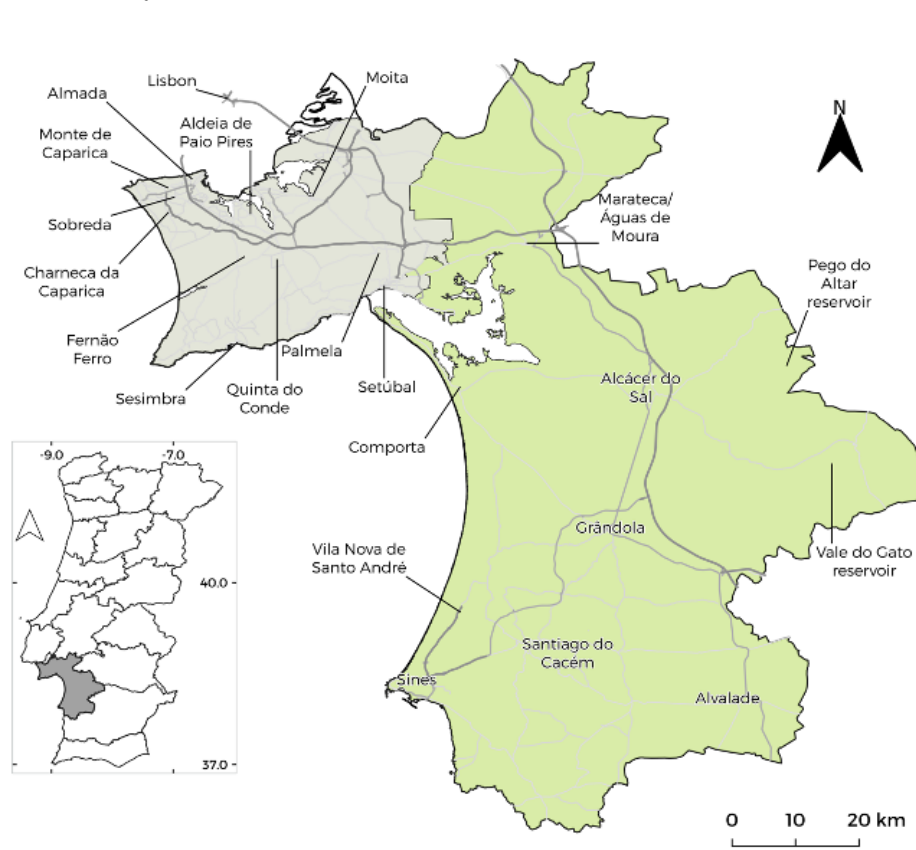


Figure 1. Geographical setting of the study area, the district of Setúbal, its municipalities, and relevant areas. The grey area is referred to as the northwestern region, whereas the green area is referred to as the eastern region. Major roads are the grey lines, with their relative importance displayed in different thicknesses.

The several databases with different structures were merged into a single dataset. In this analysis, the following variables were used:

- Accidents and geographical variables: municipality, location, type, causes, occurring in parking.

- Road variables: type of road, type of roadside, road layout, type of lane, road conservation state, the existence of works on the road, the existence of light signals and the existence of pavement marks.
- Vehicles variables: age and type of the vehicle(s) involved in the RTA.
- Drivers variables: gender, age, the driver ran away from the RTA scene.
- Victims variables: types of victims and injury severity of the victim within 30 days.
- Meteorological variables: precipitation, temperature, wind speed, and if the weather is considered good, that is, there was no fog, no rain, no strong wind, no snow, no smoke cloud, no hail.
- Time variables: date and hour.

The categories and the number of observations of each variable used to describe the RTAs are presented in Table 1.

In the data processing stage, the data were checked for incorrect registration, cleaned and prepared so that they could be properly used. Such a stage also involved handling missing data, encoding values and assigning types to variables among other standard techniques.

Table 1. Categories (or mean \pm standard deviation (SD) for continuous variables) and frequency (n) of each variable used to describe the RTA (categorized as they were considered in the models).

Variable	Categories (or Mean \pm SD)	n
Accident and geographical variables		
Municipality	Setúbal/Alcochete/Seixal/Montijo	7865
	Alcácer do Sal	1184
	Sesimbra/Almada	7760
	Moita/sines	2876
	Palmela/Barreiro	5565
	Grândola	1195
	Santiago do Cacém	1557
Accident location	Inside urban area	19,321
	Outside urban area	8681
Causes	Distraction	3452
	Irregular manoeuvre	868
	Lack of dexterity	807
	Disregard of vertical signs	803
	Disregard of safety distances	588
	Excessive speed	555
	Alcohol influence	343
	Other reasons	882
Occurring in a parking	Yes	846
	No	27156
Road variables		
Type of road	Highway/Bridge	2411
	EN	5308
	IC/IP	1776
	Other type	18,507
	Type of roadside	Paved
Road layout	Unpaved or non-existent	4289
	Curve	4302
Type of lane	Straight	23,650
	With central separation	4400
Road conservation state	Without central separation	23,601
	Good	4613
Existence of works on the road	Regular/bad	3562
	Yes	190

Table 1. Cont.

Variable	Categories (or Mean±SD)	n
Existence of light signals	No	7389
	Yes	456
Existence of pavement marks	No	7070
	Yes	4854
	No	2912
Vehicle variables		
Median vehicle age	11.72 ± 6.04	27,624
Type of the vehicle(s) involved	Light passenger	23,272
	Motorcycles but not heavy vehicles	2655
	Heavy	1964
Driver variables		
Age of the oldest driver (years)	≤20	616
	(20, 25]	1311
	(25, 30]	1501
	(30, 40]	4415
	(40, 55]	8854
	(55, 82]	9465
≥50% male drivers	>82	126
	Yes	21,753
Driver ran away from the RTA scene	No	4867
	Yes	3987
	No	24,015
Victims variables		
Severity	Fatal	163
	Serious injuries	404
	Minor injuries	5404
	Property damages	22,033
Drivers injured or dead	No	1291
	Yes	4678
Number of passengers	0	4288
	≥1	1651
Number of pedestrian	0	5345
	≥1	624
Meteorological variables		
Temperature (°C)	<17 or [20, 28)	21,403
	[17, 20)	4965
	[28, 50)	1634
Precipitation (mm/min)	0.06 ± 0.73	28,002
Wind velocity (m/s)	1.16 ± 0.98	28,002
Good weather	Yes	24,327
	No	3627
Time variables		
Month	Feb./Mar./May/Jul./Sep.	11,546
	Oct. to Jan.	9264
	Apr./Jun./Aug.	7192
Day of the week	Thursday/Friday	8454
	Weekend	7446
Hour of the day	Monday to Wednesday	12,102
	6–10 p.m.	6961
	0–1 a.m./ 2 a.m.	12,436
	2 a.m.	3395
	4 a.m./ 5–7 a.m./ 11 a.m.–1 p.m./ 4–5 p.m./ 11 p.m.	4781
	2–3 p.m.	254
	8–10 a.m.	175

2.3. Methodology

2.3.1. Multinomial Logistic Model

To identify the influential factors, such as road environments, vehicles, and atmospheric conditions, that influence the type of RTA (pedestrian running over, collision, and crashes), a statistical multinomial logit model was used.

To obtain a parsimonious multivariable model, we followed the Hosmer–Lemeshow methodology [25]. The level of significance used in the univariable analysis was 0.05. In this phase, the time factors month, day of the week and hour were categorized by their coefficients and using a likelihood ratio test. To select the interactions, $\alpha = 0.001$ was considered to obtain a simpler model, easier to understand and interpret. The analysis of the coefficients of the multivariable model and the likelihood ratio test were used to collapse categories of a variable. In the final model, the assumption of the linearity with the logit was tested for the continuous variables, and the ones where this assumption fails were categorized since an intuitive transformation of such variables was not possible, despite some discriminative ability being lost. To evaluate the goodness of fit of the multivariable model, three logit models were fitted with the same variables selected in the multivariable model. For each logit model, the functional form of continuous variables was checked using the Lowess method, the GAM method and fractional polynomials. Multicollinearity was checked using the VIF values, and residual analysis was performed to check the existence of influential observations and outliers. Model validation was performed using a bootstrap technique. The goodness of fit was assessed using the Hosmer–Lemeshow and the Cessie–van Houwenlingen tests. The discrimination ability was obtained using sensitivity, sensibility, and the ROC curve.

2.3.2. Machine Learning

ML techniques were used to compare their classification performance with that obtained with the statistical multinomial logit model. The following supervised ML algorithms were used: random forest (RF), naive Bayes, support vector machine (SVM), K-nearest neighbors (KNNs) and decision trees with the C5.0 algorithm. RF is one of the most-used classification methods, building decision trees on different samples and providing the final prediction by the majority voting. Naive Bayes is a simple classification algorithm based on Bayes' theorem and assumes that the predictors are independent. SVM finds a hyperplane with dimensions equal to the number of features that separate the classes of data points with the maximum distance between points. KNN classifies a new case based on the similarity between it and the available categories. C5.0 is a decision tree algorithm that uses information entropy to determine the best rule to split the data. A detailed description of the ML algorithms can be found in [9].

For the ML methods, data were pre-processed to be used in the ML algorithms being deleted some observations due to missing values and no missing values' imputation was made. The dataset was also transformed using a design matrix, by expanding factors to a set of dummy variables.

2.3.3. Oversampling Approach

Since the type of RTA has strongly imbalanced categories, with a much lower number of cases of pedestrian running over, we compared the performance of both statistical and ML methods when the classes are heavily imbalanced and when an approach used to balance classes is used: the random over-sampling examples approach [26]. ROSE is based on a smoothed bootstrap form of re-sampling from data to deal with binary classification problems in the presence of imbalanced classes. It handles both continuous and categorical data by generating synthetic examples from a conditional density estimate of the two classes.

As the minority class (pedestrian running over) had 21 times fewer observations than the majority class (collisions) and since ROSE only allows generating synthetic data for dichotomous variables, the following approach was taken: (1) Random sampling of half of

the collision accidents (the majority class). (2) The creation of a dichotomous variable joining the classes collision and pedestrian running over. The crash RTAs are now oversampled using the ROSE technique, leaving this category with twice the number of its initial cases. (3) Create another dichotomous variable joining the classes collisions and crashes and, using ROSE technique, the number of pedestrians being run over is oversampled, leaving this category with 5 times more cases. We chose to use this adaptation for a multiclass variable of the ROSE technique since we do not intend to create too many synthetic observations for the minority class, nor sample too many observations of the majority class.

With this approach, we balanced the number of cases in each category, without an extreme over or under-sampling of the classes. When no correction was made to the imbalanced data, we had the following distribution for the type of RTA: pedestrian running over (976 cases, 3.49%), collisions (21,800 cases, 77.85%) and crashes (5226 cases, 18.66%). After applying the ROSE approach, the distribution of RTA by type is pedestrian running over (4874 cases, 18.53%), collisions (10,933 cases, 41.57%) and crashes (10,493 cases, 39.90%). Therefore, we obtained a similar number of cases of collisions and crashes, and the number of accidents by pedestrians running over with around half of the frequency of the collisions and crashes.

2.3.4. Performance Metrics

The performance of each model was evaluated by its discrimination ability: accuracy, sensitivity, and specificity for each RTA type outcome. For the multinomial classification, an RTA is considered as a true positive for a given type of RTA if the model predicts correctly that RTA type. It is considered a true negative case for a given RTA type if a given accident is not of that type and the model correctly predicts that the accident is not of that type. Using this terminology, sensitivity (also called recall) is the ability of the model to detect a true positive case, and specificity is the ability of the model to detect a true negative case. The accuracy is then the proportion of model correct predictions. The balanced accuracy is obtained by averaging the sensitivity values of each class. We also obtain the balanced accuracy weighted, making use of the balanced accuracy by multiplying each sensitivity by the weight of each class.

According to He and Garcia [27], for imbalanced data, the sensitivity is more interesting than the specificity. Another popular classification metric for imbalanced data is the F-score or the F-measure, which combines, into a single measure, the balance between positive predictive values and sensitivity. For a multcategory classification problem, the calculation of the F-score, usually, is done by averaging methods. We consider the Macro F-score, which is the arithmetic mean of the F-score per class. Matthew's correlation coefficient (MCC) measures the correlation between the true class and the predicted class and is one of the best measures to use when the data is imbalanced. An extension for multcategory classification issues has been proposed in Gorodkin [28]. Cohen's Kappa measures the concordance between the true class and the predicted class and, similarly to the MCC, can be extended to multcategory classification issues.

To correctly compare both statistical and ML methods, we obtained the performance measures for the same testing data. For the statistical multinomial logit model, it is also possible to obtain the significant variables and measure their effect on the response variable using the odds ratio.

All statistical analyses were conducted using R version 4.0.4 [29]. The main packages used for both the statistical and ML models were the `mlogit` package for the statistical multinomial logistic model [30], the `caret` package for the ML models [31], the `MLmetrics` package for the performance evaluation metrics [32] and the `ROSE` package for the ROSE approach [33].

3. Results

With the univariable statistical multinomial logistic models, the following variables were selected to be considered in the multivariable models:

- Geographic factors: municipality, RTA in a parking area, and RTA located inside/outside an urban area;
- Time factors: month, day of the week and hour of the RTA;
- Weather factors: temperature, and weather conditions (good or rain/other conditions);
- Road characteristics factors: road layout, and type of road;
- Driver's characteristics factors: % of male drivers, and age of the oldest driver;
- Vehicle's characteristics factors: type of vehicle, and median vehicle's age.

In the following sections, we present each of the methods and their performance metrics and, at the end of the section, we perform a comparison between methods.

3.1. Statistical Multinomial Logit Model

Using the methodology described in Section 2.3, the statistical multinomial logit model was fitted using the pedestrian running over as the reference class. The adjusted model is presented in Table 2. This model adequately describes the data and presents a good discriminative ability: Hosmer–Lemeshow test p -value = 0.22, Nagelkerke R^2 = 0.33, AUC = 0.811 (OR95% = (0.805; 0.818)). Positive coefficients are associated with variables/categories with higher odds of a collision/crash RTA occurring, while negative coefficients are associated with higher odds of a pedestrian running over.

The model coefficients were obtained using all 28,002 RTA. We compared the coefficients of the adjusted model with the ones obtained when using only the 80% of training observations used for the ML models, with and without the ROSE approach. We verified that the coefficients are very close to each other, asserting the robustness of the model.

In general, we can conclude that the probability of an RTA being a crash (relative to a pedestrian running over) increases when the RTA occurs in the following circumstances: (1) in the municipalities of Alcochete, Montijo, Setúbal and Seixal; (2) outside localities; (3) with higher temperature; (4) not good weather, i.e., the occurrence of atmospheric factors that could cause a reduction in visibility and loss of vehicle adherence to the road (5) on the weekend; (6) in April, June, or August; (7) outside the period from 6 pm to 11 pm; (8) in a road with curves; (9) with the oldest driver younger than 20 years old (when compared with older drivers); (10) involving motorcycles; and (11) involving older vehicles (RTAs increase with the age of the vehicles involved, Figure 2).

We can also conclude, in general, that the probability of a RTA being a collision (relative to a pedestrian running over) increases when the RTA occurs in the following circumstances: (1) in the municipalities of Alcochete, Montijo, Setúbal and Seixal; (2) in a parking area; (3) with higher temperature; (4) on the weekend; (5) in April, June, or August; (6) outside the period from 6 p.m. to 11 p.m., in particular between 2 p.m. and 4 p.m.; (7) on a curve; (8) on a national road, IP/IC or other roads; (10) with the oldest driver up to 20 years old and not greater than 82 years old (when compared with younger drivers); and (11) not involving light vehicles.



Figure 2. Evolution of the odds of a crash RTA occurring compared to a pedestrian running over as a function of the difference between the median age of the vehicles involved in the accidents.

Table 2. Statistical multinomial logistic fitted model for the type of RTA. In the table, we present the significant variables of the final multivariable regression model, with a complete description of their categories. The coefficients of the model, the standard errors and the *p*-values obtained from Wald statistic are presented. Reference category: pedestrian running over.

Variable	Collision			Crash		
	Coef.	St. Err.	OR	Coef.	St. Err.	OR
Intercept	1.28 ***	0.33	3.36	2.22 ***	0.34	8.13
Municipality: ref. Setúbal/Alcochete/Seixal/Montijo						
Alcácer do Sal	−0.68 ***	0.21	0.50	0.80 ***	0.21	2.19
Sesimbra/Almada	−0.25 *	0.11	0.78	−0.51 ***	0.12	0.60
Palmela/Barreiro	−0.39 ***	0.11	0.62	−0.10	0.12	0.84
Grândola	−0.93 ***	0.18	0.36	−0.03	0.19	0.94
Santiago do Cacém	−1.39 ***	0.13	0.25	−0.89 ***	0.14	0.42
Accident location: ref. Inside urban area						
Outside urban area	0.10	0.12	1.18	0.98 ***	0.12	2.80
Occurred in a parking: ref. No						
Yes	0.89 **	0.30	2.51	−0.35	0.34	0.69
Temperature °C: ref. < 17 or [20,28)						
[17,20)	−0.28 ***	0.09	0.70	−0.23*	0.09	0.73
[28,50)	1.00 ***	0.27	2.81	0.88**	0.28	2.34
Good weather: ref. Yes						
No	−0.01	0.11	0.93	0.92 ***	0.11	2.32
Day of the week: ref. Thursday/Friday						
Weekend	0.31 **	0.10	1.40	0.62 ***	0.10	1.97
Monday to Wednesday	0.19 *	0.08	1.20	0.23 **	0.09	1.29
Month: ref. Feb./Mar./May/Jul./Sep.						
Oct. to Jan.	−0.30 ***	0.08	0.76	−0.36 ***	0.09	0.72
Apr./Jun./Aug.	0.23 *	0.10	1.26	0.28 **	0.10	1.34
Hour of the day: ref. 6–10 p.m.						
0–1 a.m./ 2 a.m.	0.21 *	0.08	1.14	0.58 ***	0.09	1.81
2 a.m.	−1.09 ***	0.26	0.33	0.22	0.27	0.99
4 a.m./ 5–7 a.m./ 11 a.m.–1 p.m./ 4–5 p.m./ 11 p.m.	−0.77	0.41	0.83	1.08 **	0.41	5.50
2–3 p.m.	0.82 ***	0.15	2.53	1.04 ***	0.16	3.42
8–10 a.m.	0.22 *	0.10	1.17	0.24 *	0.12	1.29
Road layout: ref. Curve						
Straight	−0.37 **	0.13	0.75	−1.46 ***	0.13	0.25
Type of road: ref. Highway/bridge						
EN	0.57 ***	0.15	1.75	0.003	0.15	0.98
IC/IP	0.57 **	0.18	1.82	−0.12	0.19	0.92
Other type	0.71 ***	0.16	2.14	0.17	0.16	1.24
≥ 50% male drivers: ref. No						
Yes	0.61 ***	0.08	1.89	−0.07	0.09	0.97
Oldest driver: ref. < 20 years old						
(20,25]	0.19	0.26	1.28	−0.46	0.26	0.64
(25,30]	0.57 *	0.26	1.96	−0.73 **	0.26	0.5
(30,40]	0.69 **	0.24	2.11	−1.25 ***	0.24	0.30
(40,55]	1.08 ***	0.23	3.01	−1.37 ***	0.23	0.24
(55,82]	1.49 ***	0.23	4.91	−1.34 ***	0.23	0.27
> 82	0.58	0.48	1.82	−1.30 *	0.51	0.26
Type of vehicle: ref. Light passenger vehicle						
Motorcycles but not heavy vehicles	0.64 ***	0.17	1.78	1.53 ***	0.18	4.49
Heavy vehicles	1.02 ***	0.21	2.73	0.52 *	0.22	1.87
Median vehicle's age	0.01	0.01	1.00	0.06 ***	0.01	1.06

p-value: * < 0.05, ** < 0.01, *** < 0.001.

3.2. Machine Learning Algorithms

The variables used in the ML models were precisely the ones used in the statistical multinomial logistic final model. To train the ML models, a random sample of 80% of observations was selected, and 20% of the remaining data was used for validation. As explained, the ML models were fitted first without any correction for the imbalanced data and then refitted after the implementation of the ROSE approach to balance the classes of the type of RTA.

For the model with imbalance data, the following distribution of the training data of the RTA by their type was pedestrian running over (696 cases), collision (16,288 cases) and crashes (3799 cases). The remaining 20% of cases used for model validation have the following distribution: pedestrian running over (179 cases), collision (4031 cases) and crashes (959 cases). After proceeding for a correction on the data imbalance using the ROSE approach, we obtained the following distribution of the training data by type of the accident: pedestrian being run over (3849 cases), collision (8807 cases) and crashes (8395 cases). Additionally, 20% of cases used for validation have the following distribution: pedestrian running over (1025 cases), collision (2126 cases) and crashes (2098 cases).

Since the ML models assume that all the data are numeric, factors need to be converted into dummy variables, resulting in a total of 34 predictors. The fitted statistical multinomial logit model uses the same validation data as the ML models to evaluate the models' performance.

Cross-validation (10-fold) was performed with three repetitions for all models. For the RF model, the number of variables randomly collected to be sampled at each split time was 36. The results for the C5.0 algorithm were obtained by using a rules model with 18 trials. For the KNN, the final model used the 17 closest observations. The SVM used a linear kernel with an upper bound for the constraint of the minimization problem equal to 1. The naive Bayes classifier used a constant Laplace smoother and a Gaussian density.

The detailed results of the performance measures for all models without the use of the ROSE approach are presented in Table 3. The class with the lowest number of cases, the pedestrian running over, was not correctly classified either in the statistical multinomial logistic model or in all ML algorithms. Only the class with the higher number of cases (collisions) was correctly classified by the models. All the performance measures that combine other metrics (balanced accuracy weighed, Macro F1, MCC and Cohen Kappa) have tiny values. The imbalance of the data implies very low correct prediction, in all models, for the minority class.

Table 3. Performance measures for the statistical multinomial logit regression model and the ML models for the type of RTA, without using the ROSE approach.

Measure	Mult.	ML Algorithms				
		RF	SVM	Naive-Bayes	C5.0	KNN
Accuracy	0.818	0.797	0.788	0.744	0.811	0.802
Sensitivity (run. over)	0.000	0.006	0.000	0.118	0.006	0.011
Sensitivity (collision)	0.964	0.931	0.958	0.842	0.964	0.978
Sensitivity (crash)	0.369	0.382	0.219	0.450	0.318	0.208
Specificity (run. over)	1.000	0.994	1.000	0.947	0.999	0.999
Specificity (collision)	0.324	0.353	0.204	0.480	0.292	0.192
Specificity (crash)	0.961	0.932	0.955	0.889	0.959	0.975
Balanced Acc. Weigh.	0.818	0.797	0.788	0.744	0.811	0.801
Macro F1	-	0.457	-	0.469	0.519	0.511
MCC	0.391	0.336	0.240	0.298	0.351	0.286
Cohen Kappa	0.354	0.321	0.203	0.398	0.312	0.223

Table 4 presents the detailed results of the performance measures for all models using the ROSE approach. It is clear that the ROSE approach allows a considerable improvement in all performance measures of the models. The statistical multinomial logistic model,

the SVM and naive Bayes methods, despite improving the classification of the minority class, still do not predict accurately this class and present small values of the combined metrics. On the other hand, the ML algorithms that use decision trees, the RF and C5.0, achieve good performance, with each class correctly classified with a precision higher than 80%. The combined metrics also present very interesting values, with an MCC around 0.80 and a Macro F1 over 0.85.

Table 4. Performance measures for the statistical multinomial logit regression model and the ML models for the type of RTA, using the ROSE approach.

Measure	Mult.	ML Algorithms				
		RF	SVM	Naive-Bayes	C5.0	KNN
Accuracy	0.578	0.881	0.570	0.523	0.858	0.666
Sensitivity (run. over)	0.140	0.982	0.120	0.420	0.970	0.863
Sensitivity (collision)	0.715	0.806	0.727	0.681	0.787	0.548
Sensitivity (crash)	0.651	0.907	0.630	0.413	0.876	0.688
Specificity (run. over)	0.960	0.978	0.960	0.806	0.973	0.840
Specificity (collision)	0.586	0.940	0.558	0.591	0.919	0.850
Specificity (crash)	0.759	0.891	0.776	0.871	0.880	0.807
Balanced Acc. Weigh.	0.577	0.881	0.570	0.523	0.858	0.666
Macro F1	0.524	0.893	0.512	0.512	0.871	0.680
MCC	0.316	0.816	0.305	0.270	0.780	0.501
Cohen Kappa	0.306	0.814	0.293	0.262	0.779	0.493

Figure 3 presents the importance of the explanatory variables only for the RF model, which was the one that produced the best results. For each tree, the prediction accuracy on the out-of-bag portion of the data was recorded. Then the same was done after permuting each predictor variable. The difference between the two accuracies was then averaged over all trees and normalized by the standard error. For pedestrian running over, the variables median age of the vehicle and accident’s location had the most influence on the classification process; in collisions, were the variables age of the oldest driver and the median age of the vehicles; and in crashes, were the median variables of the age of the vehicles, the type of road and the location of the accident.

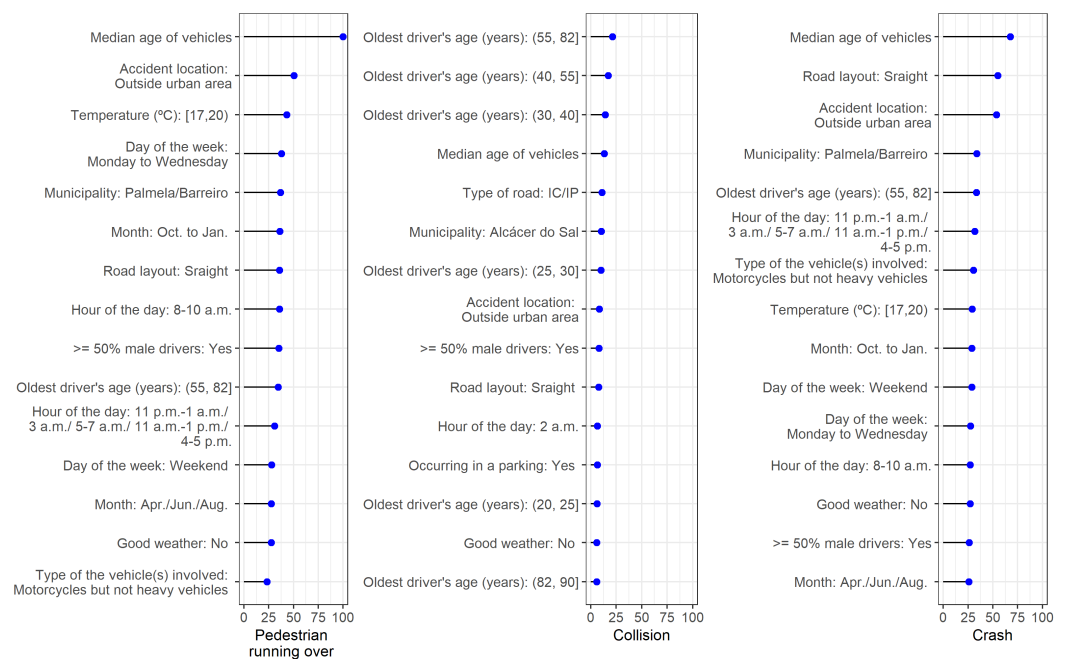


Figure 3. RF variable importance for the type of road accidents.

3.3. Comparison of Models

When comparing the results obtained in previous sections, the best approach to correctly classify a multiclass problem for the type of RTA with severe imbalanced data was the RF algorithm using the ROSE approach. When the data were imbalanced for this multiclass problem, none of the methods was able to correctly predict the minority classes of the type of the RTA. The accuracy and the balanced weighted accuracy are the only metrics that have reasonable values. Despite the pedestrian running over class being highly misclassified, their contribution to the total number of observations is so low that it does not have an impact on the value of these two metrics. Correcting the imbalance in the data is notorious, with a considerable improvement in all methods and all metrics evaluated. This result allows us to conclude that in this kind of problem, the need to correct the imbalance in the data is essential for a good classification of a multiclass issue.

Both statistical and ML approaches had poor performance when the multiclass data presented a considerable imbalance. When the imbalance is corrected by the ROSE approach, the statistical approach as well as some ML algorithms, improve slightly but still cannot correctly classify the minority class. However, the ML algorithms that use decision trees, the RF and the C5.0 algorithms, were able to classify all the categories with a high percentage of correct predictions. The combined metrics, such as the Macro F1, the MCC and the balanced accuracy weighted, all have values higher than 0.80, particularly for the RF model.

4. Final Remarks

In this paper, we analyzed data from RTA in a district of Portugal. The initial challenge was to create a unique dataset joining data from several sources with different structures, with several distinct information: data about the vehicles, the drivers, the victims, the road characteristics, the location, the time of the day and the weather conditions of a given RTA.

The main objective of the paper was to obtain the main factors that influence the type of RTA (collision, crash, or pedestrian running over). For that, we compared the performance of a statistical multinomial logit model with some ML algorithms that can be used in a multiclass problem.

The multinomial logit model allowed us to conclude that, in general, the probability of a crash (relative to a pedestrian running over) increases in the following cases: (a) in the municipalities of Alcochete, Montijo, Setúbal and Seixal; (b) outside localities; (c) with higher temperature; (d) with bad weather; (e) on the weekend; (f) in April, June, or August; (g) outside the period from 6 p.m. to 11 p.m.; (h) on a curve; (i) with the oldest driver younger than 20 years old (when compared with older drivers); (j) it involves motorcycles; and (k) with the age of vehicles involved (increases with the median age of the vehicles). We also can conclude that the probability of a collision occurring (relative to a pedestrian running over) increases in the following cases: (a) in the municipalities of Alcochete, Montijo, Setúbal and Seixal; (b) in a car park; (c) with higher air temperature; (d) at the weekend; (e) in April, June, or August; (f) outside the period from 6 p.m. to 11 p.m., in particular between 2 p.m. and 4 p.m.; (g) on a curve; (h) on an EN, IP/IC or other types of secondary roads; (i) with the oldest driver up to 20 years old and not greater than 82 years old (when compared with younger drivers); and (j) not involving light vehicles.

Moreover, this multiclass problem has the characteristic of having data imbalance, with a given class with a small percentage of cases. We concluded that neither the statistical model nor the ML algorithms were able to correctly predict the classes with the small number of cases.

Therefore, we adapted a technique that is commonly used for binary classification to balance classes, the ROSE technique. This adapted approach for the multinomial response variable allowed us to synthesize new observations of the minority classes and obtain a more balanced dataset. We then evaluated the performance of all models with this new balanced dataset and concluded that all models had a considerable improvement in their

performance. Only the decision tree algorithms were able to achieve a good prediction of all three classes of the type of RTA, and the RF model was the best one.

Obtaining the main factors that influence the type of RTA is essential on its own, but also could have implications in the definition of policies for prevention, whether at the level of the security forces, the road, or awareness campaigns in terms of pedestrians being run over, among others.

In future research, we should analyze if the need to correct imbalanced data is more necessary in a multiclass problem than in a binary problem, and if it is always the decision tree algorithms that present a better performance in this kind of multicategory issue.

Author Contributions: Conceptualization all authors; methodology, P.I., G.J., A.A., P.N., V.N., P.Q. and J.S.; software, P.I., G.J., A.A., L.R., P.N., M.S., V.N., J.S., P.Q. and D.S.; validation, P.I., G.J., A.A., L.R. and P.R.M.; formal analysis, P.I., G.J., A.A. and L.R.; investigation, all authors; resources, all authors; data curation, A.A., V.N., P.Q., J.S. and D.S.; writing—original draft preparation, P.I., G.J., A.A. and L.R.; writing—review and editing, all authors; visualization, all authors; supervision, P.I.; project administration, P.I. and V.N.; funding acquisition, P.I., G.J., A.A., P.Q., P.N., V.N., J.S. and P.R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, grant number FCT DSAIPA/DS/0090/2018, “MOPREVIS—Modelação e Predição de Acidentes de Viação no Distrito de Setúbal”, within the scope of the National Initiative on Digital Skills e.2030, Portugal INCoDe.2030.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from the Portuguese GNR in the context of the MOPREVIS project.

Acknowledgments: The authors are grateful for the data support given by ANSR, Infraestruturas de Portugal and IPMA.

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

Abbreviations

The following abbreviations are used in this manuscript:

AE	Highway
ANSR	Autoridade Nacional de Segurança Rodoviária (National Road Safety Authority)
BEAV	Statistical Bulletin of Road Traffic Accidents
EN	National Road
EU	European Union
GDP	Gross Domestic Product
GNR	Guarda Nacional Republicana (National Republican Guard)
IC	Complementary Itinerary
IP	Principal Itinerary
IPMA	Instituto Português do Mar e da Atmosfera (Portuguese Institute for Sea and Atmosphere)
KNN	K-Nearest Neighbor
MCC	Matthew’s Correlation Coefficient
ML	Machine Learning
MOPREVIS	Modeling and Prediction of Road Traffic Accidents in the District of Setúbal
RF	Random Forest
7 ROSE	Random Over-Sampling Examples
RTA	Road Traffic Accident
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TC-GNR	Territorial Command of the GNR

References

1. World Health Organization. Projections of Mortality and Causes of Death, 2015 and 2030. 2013. Available online: https://www.who.int/healthinfo/global_burden_disease/projections2015_2030/en/ (accessed on 25 January 2022).
2. Eurostat. Road Accidents: Number of Fatalities Continues Falling, 2021. Available online: <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/ddn-20210624-1> (accessed on 25 January 2022).
3. Lusa. Sinistralidade Rodoviária Tem Impacto Económico e Social Negativo de 1,2% do PIB—Governo. 2018. Available online: https://www.rtp.pt/noticias/pais/sinistralidade-rodoviaria-tem-impacto-economico-e-social-negativo-de-12-do-pib-governo_n1112193 (accessed on 25 January 2022).
4. Kim, D.G.; Washington, S.; Oh, J. Modeling crash types: New insights into the effects of covariates on crashes at rural intersections. *J. Transp. Eng.* **2006**, *132*, 282–292. [[CrossRef](#)]
5. Infante, P.; Jacinto, G.; Afonso, A.; Rego, L.; Nogueira, V.; Quaresma, P.; Saias, J.; Santos, D.; Nogueira, P.; Silva, M.; et al. Comparison of statistical and machine-learning models on road traffic accident severity classification. *Computers* **2022**, *11*, 80. [[CrossRef](#)]
6. Zhang, J.; Li, Z.; Pu, Z.; Xu, C. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access* **2018**, *6*, 60079–60087. [[CrossRef](#)]
7. Rezapour, M.; Moomen, M.; Ksaibati, K. Ordered logistic models of influencing factors on crash injury severity of single and multiple-vehicle downgrade crashes: A case study in Wyoming. *J. Saf. Res.* **2019**, *68*, 107–118. [[CrossRef](#)] [[PubMed](#)]
8. Fiorentini, N.; Losa, M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures* **2020**, *5*, 61. [[CrossRef](#)]
9. Silva, P.B.; Andrade, M.; Ferreira, S. Machine learning applied to road safety modeling: A systematic literature review. *J. Traffic Transp. Eng.* **2020**, *7*, 775–790. [[CrossRef](#)]
10. Aidoo, E.N.; Amoh-Gyimah, R.; Ackaah, W. The effect of road and environmental characteristics on pedestrian hit-and-run accidents in Ghana. *Accid. Anal. Prev.* **2013**, *53*, 23–27. [[CrossRef](#)]
11. Geedipally, S.R.; Patil, S.; Lord, D. Examination of methods to estimate crash counts by collision type. *Transp. Res. Rec.* **2010**, *2165*, 12–20. [[CrossRef](#)]
12. Bham, G.H.; Javvadi, B.S.; Manepalli, U.R. Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban US highways in Arkansas. *J. Transp. Eng.* **2012**, *138*, 786–797. [[CrossRef](#)]
13. Chen, Y.; Wang, K.; King, M.; He, J.; Ding, J.; Shi, Q.; Wang, C.; Li, P. Differences in factors affecting various crash types with high numbers of fatalities and injuries in China. *PLoS ONE* **2016**, *11*, e0158559. [[CrossRef](#)]
14. Intini, P.; Berloco, N.; Fonzone, A.; Fountas, G.; Ranieri, V. The influence of traffic, geometric and context variables on urban crash types: A grouped random parameter multinomial logit approach. *Anal. Methods Accid. Res.* **2020**, *28*, 100141. [[CrossRef](#)]
15. Iranitalab, A.; Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accid. Anal. Prev.* **2017**, *108*, 27–36. [[CrossRef](#)] [[PubMed](#)]
16. Christoforou, Z.; Cohen, S.; Karlaftis, M.G. Identifying crash type propensity using real-time traffic data on freeways. *J. Saf. Res.* **2011**, *42*, 43–50. [[CrossRef](#)]
17. Boo, Y.; Choi, Y. Comparison of Prediction Models for Mortality Related to Injuries from Road Traffic Accidents after Correcting for Undersampling. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5604. [[CrossRef](#)]
18. Guo, M.; Zhao, X.; Yao, Y.; Yan, P.; Su, Y.; Bi, C.; Wu, D. A study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data. *Accid. Anal. Prev.* **2021**, *160*, 106328. [[CrossRef](#)] [[PubMed](#)]
19. Ding, H.; Lu, Y.; Sze, N.; Chen, T.; Guo, Y.; Lin, Q. A deep generative approach for crash frequency model with heterogeneous imbalanced data. *Anal. Methods Accid. Res.* **2022**, *34*, 100212. [[CrossRef](#)]
20. Yu, R.; Wang, Y.; Zou, Z.; Wang, L. Convolutional neural networks with refined loss functions for the real-time crash risk analysis. *Transp. Res. Part C Emerg. Technol.* **2020**, *119*, 102740. [[CrossRef](#)]
21. Rella Riccardi, M.; Mauriello, F.; Sarkar, S.; Galante, F.; Scarano, A.; Montella, A. Parametric and Non-Parametric Analyses for Pedestrian Crash Severity Prediction in Great Britain. *Sustainability* **2022**, *14*, 3188. [[CrossRef](#)]
22. Vilaça, M.; Macedo, E.; Coelho, M.C. A Rare Event Modelling Approach to Assess Injury Severity Risk of Vulnerable Road Users. *Safety* **2019**, *5*, 29. [[CrossRef](#)]
23. Rella Riccardi, M.; Galante, F.; Scarano, A.; Montella, A. Econometric and Machine Learning Methods to Identify Pedestrian Crash Patterns. *Sustainability* **2022**, *14*, 15471. [[CrossRef](#)]
24. ANSR. Manual de Prenchimento. Boletim Estatístico de Acidente de Viação. 2013. Available online: <http://www.ansr.pt/Estatisticas/BEAV/Documents/MANUALPREENCHIMENTOBEAV.pdf> (accessed on 25 January 2022).
25. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
26. Menardi, G.; Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl.* **2014**, *28*, 92–122. [[CrossRef](#)]
27. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [[CrossRef](#)]
28. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **2004**, *28*, 367–374. [[CrossRef](#)] [[PubMed](#)]

29. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
30. Croissant, Y. Estimation of Random Utility Models in R: The mlogit Package. *J. Stat. Softw.* **2020**, *95*, 1–41. [[CrossRef](#)]
31. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [[CrossRef](#)]
32. Yan, Y. MLmetrics: Machine Learning Evaluation Metrics. R Package Version 1.1.1. 2016. Available online: <https://CRAN.R-project.org/package=MLmetrics> (accessed on 1 December 2022).
33. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *R J.* **2014**, *6*, 82–92. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.