informatics

Article

The Prediction of Road-Accident Risk through Data Mining: A Case Study from Setubal, Portugal

David Dias [1,2], José Silvestre Silva [1,3,4,*] and Alexandre Bernardino [2,5]

[1] Portuguese Military Academy, Rua Gomes Freire, 1169-203 Lisbon, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal
[3] Military Academy Research Center (CINAMIL), Rua Gomes Freire, 1169-203 Lisbon, Portugal
[4] Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys-UC), 3000-370 Coimbra, Portugal
[5] Institute for Systems and Robotics (ISR/IST), 1049-001 Lisbon, Portugal
[*] Correspondence: jose.silva@academiamilitar.pt

Abstract: This work proposes a tool to predict the risk of road accidents. The developed system consists of three steps: data selection and collection, preprocessing, and the use of mining algorithms. The data were imported from the Portuguese National Guard database, and they related to accidents that occurred from 2019 to 2021. The results allowed us to conclude that the highest concentration of accidents occurs during the time interval from 17:00 to 20:00, and that rain is the meteorological factor with the greatest effect on the probability of an accident occurring. Additionally, we concluded that Friday is the day of the week on which more accidents occur than on other days. These results are of importance to the decision makers responsible for planning the most effective allocation of resources for traffic surveillance.

Keywords: risk prediction; road accidents; supervised classification; classical methods; deep neural networks

check for updates

Citation: Dias, D.; Silva, J.S.; Bernardino, A. The Prediction of Road-Accident Risk through Data Mining: A Case Study from Setubal, Portugal. *Informatics* **2023**, *10*, 17. https://doi.org/10.3390/informatics10010017

Academic Editor: Olga Kurasova

Received: 31 December 2022
Revised: 22 January 2023
Accepted: 25 January 2023
Published: 30 January 2023

1. Introduction

Road accidents cause multiple deaths each year and result in economic and physical damage to their victims; additionally, they incur the loss of public resources. Preventive action by the security forces has focused on what is known as Information-Guided Policing [1]. Since accident-related data are stored in the National Guard database, it is possible to discover patterns correlated with the occurrence of accidents and to create knowledge that is useful in decision-making. Data-mining techniques have evolved significantly in recent decades and are being widely applied to several real-world problems. Current data-mining methods can be used on a database to rapidly extract knowledge that can help to guide policing methods and thus improve accident-prevention techniques and awareness campaigns produced by the security forces.

This work aims to develop a tool to aid Information-Guided Policing in traffic management. Several data mining algorithms were applied to different types of datasets, including the National Guard database, which contains multiple accident reports. To complement the data provided by the National Guard, other publicly available databases were explored, such as meteorological data sources and the annual calendar.

This work is one of the limited number of research projects carried out by Portuguese researchers using data from the Portuguese National Guard to analyze and predict road accidents. One of the objectives of this work is to provide statistical and predictive information on traffic accidents for the National Guard and other researchers.

This investigation is original because, unlike other works that use categorical variables to identify the variables that most influence the severity of accidents, it sets out to predict the number of accidents likely to occur in a future time frame. One of the main objectives of this work is to make possible the prediction of accidents using categorical variables, combining a number of factors from past events with anticipated future data (e.g., meteorological conditions) to forecast the places where there will be a higher risk of traffic accidents occurring. A further objective of this work is to compare road accidents occurring prior to the COVID-19 pandemic with those occurring during the pandemic.

This work is divided into five sections. The first section sets out an introduction to the theme explored in this work and additionally relevant topics are also described. The second section examines the state of the art and is divided into two parts: the first part analyzes classical classification methods, and the second part analyzes deep neural network methods. The third section presents and develops concepts including the discovery of knowledge in databases and the respective steps used in data filtering to select the data relevant to preprocessing and to prepare the data for data mining algorithms, mainly classifiers and performance-evaluation metrics. In the fourth section the results are presented and analyzed; they are compared in order to identify the most effective algorithm for the intended task. The fifth section sets out our conclusions, where the main information extracted during this work is summarized.

2. Related Work

The related works mentioned in this section are divided into classical methods and deep learning methods.

2.1. The Classical Approach

In 2016, Castro et al. [2] selected 81,690 records from a large database of 451,462 UK road accidents occurring from 2010 to 2012. The WEKA tool was used, and seven input variables were considered: the type of road, the lighting conditions, the weather conditions, the road surface conditions, the vehicle's maneuvers, the vehicle's fuel type, the age of the vehicle, and the severity of the accident. Three mining algorithms were used: the BayesNet implementation of the Bayesian network available in the WEKA software, the decision tree algorithm, and a neural network algorithm. The problem was framed as a classification task, where classes were labels that represented the severity of the accident (fatal, serious, or normal). The accuracy of the prediction of the severity of the accident by the three mining algorithms was very similar, with a value of approximately 72%.

In the same year, Keshyap et al. [3] sought a link between road conditions and accident severity. Instead of decision trees, the Bayesian network algorithm was used through the WEKA software. In that work, several attributes were included: the driver's condition, the driver's experience, the weather conditions, the type of road, the lighting conditions, the condition of the vehicle, the type of vehicle, the severity of the accident, the use of seat belt, and the location. A total of 31,698 accidents between 2003 and 2015 were analyzed using information from the surveys conducted among those people who had suffered the accidents. The authors attempted to use feature selection; however, the best result showed an accuracy of 89% without feature selection.

In 2019, Hussain et al. [4] carried out an evaluation of different classical data-mining methods in road accidents. The best-performing algorithms were Multilayer Perceptron, the Decision Tree, and Naive Bayes. In the same year, Kumeda et al. [5] applied six classic classification algorithms, including naive Bayes, multilayer perceptron, and random forest, to find the factors that are most influential in road accidents.

Although the works mentioned above deal with classification rather than regression problems, they are important in highlighting the type of variables used. Most of the literature that uses classical methods does not aim to predict numbers of accidents, but to predict classes of accident-related factors, e.g., accident severity. Furthermore, all the works mentioned above use small-scale datasets (on one road or on a small number of roads) with a limited number of features.

2.2. The Deep Learning Approach

Some recent works have tried to create predictive models of road accidents with Deep Learning Models. Chen et al. [6] used data from approximately 1.6 million GPS records and a history of accident records to build a model that relates human mobility to the risk of accidents. That model evaluates the risk of accidents in real time as a classification problem for each zone of the map. Data such as the geolocation of the accident and the levels of human mobility in real time proved to be essential. The author also points out that there are many factors that lead to traffic accidents, including driver behavior, the weather, and road conditions.

In 2018, Yuan et al. [7] developed methods of predicting the risk of an accident according to time, place, and day. A deep learning approach was used, based on the spatial and temporal heterogeneity of the data, which is an important characteristic of road accidents. The study was carried out with data from the state of Iowa, USA, and the sample contained 375,690 accidents occurring from 2006 to 2014. This study was supplemented by information from external databases and included data on traffic volume, road conditions, precipitation, and ambient temperature; the information was drawn from four different databases spanning eight years. The algorithm used was a variant of the long and short memory convolutional neural network, and it was developed using software that creates a predictive model for each region of the country.

2.3. Other Relevant Works

Krukowicz et al. [8] analyzed problems associated with animal-related vehicle accidents in Poland and concluded that there is no relationship between the abundance of a particular animal species and the number of road crashes, but that there is a correlation between the number of crashes and the overall length of the road network.

Billah et al. [9] used data collected over 10 years by the Texas Crash Record and Information System database to investigate how some of the most prominent driving behaviors leading to crashes and severe injuries vary by gender in San Antonio, Texas. They adopted bivariate analysis and logistic regression modeling that facilitated the identification of the effect of different variables on crash occurrence and severity by gender. It was concluded that male drivers were more likely to be involved in a crash related to speeding, DUI, or lane departure, with subsequent severe injuries, while female drivers were slightly more associated with distracted-driving crashes and subsequent injuries.

Saveliev et al. [10] proposed a fully automatic methodology for the reconstruction of an accident scenario with a highly accurate in measuring distances from the relative location of objects. After the three-dimensional scene of an accident was built, objects of interest were segmented using a deep learning model, SWideRNet with Axial Attention; there followed a two-dimensional reconstruction of the road accident based on marked-up data and the use of the image transformation method. The results achieved by the intersection over union (IoU) metric were 0.771 on average.

Tajnik et al. [11] analyzed the significance of variables influencing road crashes on rural roads to estimate crash frequencies during different conditions. They used a holistic approach and analyzed a wide range of driver/vehicle/road/environment variables. The results showed that the crash frequencies and driving speeds have strong daily and weekly seasonality: the average hourly crash frequencies per kilometer driven during the week varied between 0.2 and 2.2 crashes per million kilometers, and the major cause was speeding, which contributed to almost 32% of fatal crashes.

Bokaba et al. [12] aimed to assess prediction-model designs for road traffic accidents (RTAs) to help transport authorities and policy makers. They used Naive Bayes, logistic regression, k-Nearest Neighbor, AdaBoost, a Support Vector Machine, and Random Forest; the input data were taken from a real-life RTA dataset from Gauteng, South Africa. The results allowed them to conclude that random forest performed marginally better across the experiments in terms of accuracy, precision, recall, and ROC (AUC) when compared with the other classifiers.

Islam et al. [13] evaluated the factors that influence the frequency and severity of Road Traffic Crashes (RTCs) involving adolescent road users aged 15 to 44 in fatal and significant-injury RTCs in Al-Ahsa, Saudi Arabia. The prediction models used a logistic regression and CART (Classification and Regression Tree) to study the RTC characteristics affecting the target age group's involvement in RTCs. The results of logistic regression and CART models confirm that victims in the target age group were involved in serious traffic accidents with comparatively higher numbers of injuries and fatalities. The CART model also showed that overturn RTCs that occur due to driver distraction, speeding, failure to give way, or sudden turning, are more likely to involve victims from the target age group.

Islam et al. [14] also explored crash-severity prediction using tree-based ensemble models (gradient boosting and random forest) and a logistic regression model. The results were compared to prediction the Road Traffic Crash (RTC) severity. The random forest method outperformed other models in terms of injury severity, individual class accuracy, and collective prediction accuracy when using k-fold (k = 10) based on various performance metrics.

Mesquitela et al. [15] proposed a data fusion process from different information sources such as road accident, weather conditions, local authority reporting tools, traffic, fire brigade, which allows the creation of knowledge for local municipalities using local data. Using ArcGIS Pro, the authors applied kernel density and hotspot analysis (Getis-Ord Gi) tools to identify the existence of blackspots in terms of location and context conditions, and they evaluated the possible human, environmental, and circumstantial factors that may influence the severity of accidents.

Guido et al. [16] used two machine learning algorithms, including the data handling group method (GMDH) type neural network and a combination of a Support Vector Machine (SVM) and the grasshopper optimization algorithm (GOA). The seven factors that affect transport safety, including daylight (DL), weekday (W), type of accident (TA), location (L), speed limit (SL), average speed (AS), and annual average daily traffic (AADT) of rural roads in Cosenza, southern Italy were used as input. The results showed that, in the investigated rural area, the type of accident has the greatest importance whereas location has the lowest importance, and that the GOA-SVM model achieved a better degree of accuracy and robustness than the GMDH model.

Kim et al. [17] proposed a model for estimating run-off road crash (RORC) severity based on fixed objects, roadway geometry, traffic conditions, and the road traffic environment. This model included a learning method with tree-augmented naive Bayes, and the input data related to a section of highway in South Korea. The results allowed them to conclude that the factors affecting RORC severity were the density of fixed objects, the horizontal distance between the roadway and the fixed object in the crash, the vertical slope, and the pier, when these factors either exceeded or fell short of the set threshold values. Among all types of fixed objects, piers had the greatest impact.

Rodionova et al. [18] investigated factors that explain road crash severity levels in Saint Petersburg, Russia. The research takes into account factors such as lighting conditions, weather conditions, infrastructure factors, human factors, accident types, vehicle category and color to assess their influence on the severity of the crash. The ordered probit regression method was selected as the tool for their analysis. Their work allowed them to conclude that missing road illumination had the highest impact on crash severity; precipitations were the main factor negatively influencing crash severity, and other important factors included the absence of road barriers, the absence of restraint systems for pedestrians at appropriate locations, defective traffic lights, and problems affecting horizontal road markings.

Infante et al. [19] analyzed the determinants that contribute to road traffic accidents involving victims, as well as the determinants for fatalities and/or serious injuries in accidents involving victims. They used a logistic regression model, and the results were compared with machine-learning models (random forest, naive Bayes, SVM and kNN). They conclude that machine-learning models generally do not perform better than statistical models; however, they perform similarly when the sample is large and has a small imbalance.

3. Theoretical Framework

Current technology allows the storage of large and multiple databases. The analysis of these data is often useful; however, it is impractical without the aid of computational tools. The knowledge discovery in databases (KDD) process uses computational tools to identify valid and potentially useful patterns in the data and to generate knowledge [20–24]. Typically, this process includes the following steps:

Data selection/Problem definition: the domain of available data is defined, as are the information and data that are relevant and the knowledge-discovery objectives.

Preprocessing: this aims to prepare the data for the algorithms of the next stage. This involves performing data cleaning, data integration, data reduction, and data transformation/normalization.

Data Mining: the algorithms are applied to the data in search of knowledge and in order to extract patterns from the data. The choice of algorithm to be applied depends on the type of task to be performed.

Evaluation and representation of results: the models produced are interpreted, and evaluation metrics are used to estimate the quality of the results. Tools are used to visualize the data produced as output.

We aim to solve a regression problem in which the target variable is the number of accidents that occur on each road in a range of time periods. The learning is supervised once we already have the annotated data related to accidents, in order to train the model. The input data is categorical and the target variable is numeric.

Supervised learning occurs when data already have an associated output. As is the case with our data, we will only implement algorithms that fit this profile. For example, if the objective of a data mining problem is to predict male or female gender from the image of a face, it is necessary to have a set of faces with the gender already correctly identified. It is important to distinguish regression problems, where the data for which we want to predict the value are numerical values, from classification problems, where the data are categorical values [23,25–28].

Different techniques were analyzed in [26] and it was concluded that decision trees, naive Bayes, and support vector machines are the most frequently used techniques. Other frequently used supervised learning algorithms are k-nearest neighbors (kNN) [25–27,29] and the artificial neural network (ANN) [25,27,30]. Based on this information, these algorithms were implemented.

The most important attributes for road traffic accidents [5,31–37] were divided into three groups and listed in Table 1.

Table 1. Attributes considered important for traffic accidents that were found in the literature.

Weather Conditions:	**Precipitation, Temperature; Wind Force**
Human behavior:	Seat belt use, cell phone use, alcohol consumption calendar
Road conditions:	Road networks, luminosity, road identification, traffic volume

For the selection of attributes, it is important to analyze the correlation between the different variables and the target variable. The Pearson correlation coefficient is often used to compute the linear correlation between continuous numeric variables. However, we must use a different metric to compute the correlation between categorical variables, as is the case with our dataset. The Cramer V correlation is used to compute the correlation between nominal categorical variables with more than two (non-binary) values [38].

The Cramer's V correlation is defined as [39]:

$$\varnothing_c = \sqrt{\frac{X^2}{N(k-1)}} \tag{1}$$

where $\varnothing_c$ is the value of V of Cramer, X^2 is the value of chi-squared, N is the number of samples, and k is the number of categories of the variable with the smallest number of categories. The chi-square value is defined as:

$$X^2 = \sum \frac{(o_{ij} - e_{ij}\ )^2}{e_{ij}} \tag{2}$$

where e_{ij} is the expected frequency value and o_{ij} is the observed frequency value of a combination of two values, one of variable i, the other of variable j. The expected frequency value can be computed as

$$e_{ij}\ = \frac{o_i \,.\, o_j}{N} \tag{3}$$

and represents the expected frequency of a combination of two values (one of i, the other of j). In the previous formula, o_i is the marginal frequency of one of the values of the variable i, o_j is the marginal frequency of one of the values of j, and N is the total number of samples.

The interpretation of the strength of the correlation between two nominal categorical variables as a function of Cramer's V is given in Table 2 [36].

Table 2. Interpretation of Cramer's V coefficient.

Values of Cramer V Coefficient, $\varnothing_c$	Interpretation
[0.25; 1.00]	Very Strong
[0.15; 0.25]	Strong
[0.10; 0.15]	Moderated
[0.05; 0.10]	Weak
[0; 0.05]	Very Weak

To achieve a universal standard for deleting attributes with low correlation values, it is important that all calculated correlations be comparable. The Kruskal-Wallis is equivalent to the chi-square also used in Cramer's V, so the values achieved can be compared in the two measures. The expression for the Kruskal-Wallis test [29,40,41] is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\overline{r}_i - \overline{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \overline{r})^2} \tag{4}$$

where N is the total number of samples across all groups, g is the number of groups, n_i is the number of samples in group i, r_{ij} is the rank value of sample j that belongs to group i, $\overline{r_i} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the mean value of the rank of all observations j in group i; and $\overline{r} = \frac{r}{2}(N+1)$ is the average value of the sum of all classifications r_{ij}, i.e., the expected value for the average of all groups.

Relief-based feature selection (RBA) and sequential backward selection (SBS) were used for the selection of features [42–45]. Starting from an empty set of features, the SBS gradually adds features selected by a performance measure, which measures the extent to which each feature improves or worsens a mining method. At each iteration, the feature to be included in the feature set is selected from those available in the feature set.

To evaluate the different mining algorithms, we use the mean absolute error (*MAE*), which is an error measurement that sums the absolute error between the observations and the value obtained by the model. The mean squared error was not used, because the number of accidents has many outliers that significantly bias this metric. The *MAE* is given by the following equation:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|\overline{y_j} - y_j\right| \quad (5)$$

As the purpose of this work is to present the risk of accidents rather than to predict the exact value of accidents, the predicted values and the actual values are grouped into three risk groups: low, medium, and high. After making this grouping we can compute the classification accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

The classification accuracy measures the ratio of correct predictions to the total number of instances evaluated, where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives [29].

4. Results and Discussion

In this section, the results produced by the methodology are presented and analyzed. When more than one technique is presented, these are compared in order to assess which technique best suits the task in question.

The implemented methodology was developed using a Lenovo IdeaPad 3 computer with an AMD Ryzen 5 5500 U processor and AMD Radeon Graphics processor. The Python language was used through Jupyter in Anaconda. The Python libraries used were Keras, Numpy, Scikit-learn, Matplotlib, and Pandas.

4.1. Dataset

The data provided by the Portuguese National Guard correspond to the years 2019 to 2021 in the district of Setubal, a peripheral city of the capital of Portugal, Lisbon. This information includes road accidents and also data on administrative offenses that contain the number of inspections carried out, the number of drivers who had consumed excessive alcohol, the number of drivers who were speeding, and other administrative offences. In the present work, only information relating to traffic accidents was considered relevant. Regards data selection, in Table 3, all the attributes selected from the accident reports are presented.

Table 3. Selected attributes from the National Guard Database.

Attribute	Type/Format of Data
Identification of accident	Serial number
Date	dd/mm/yyyy
Time	{Morning, morning work, morning rush hours, lunch break, afternoon work, afternoon rush hours, night}
Type of local	{Motorway, itineraries or national roads, village roads}
Localization	{Urban location, non-urban location}
Type of accident	{Damage only, with injured}
Day of the week	{Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}
Holiday	Boolean
Alcohol	Numerical with 2 decimal digits (g/L)
Administrative offenses	Numerical with 2 decimal digits
Weather conditions	{Good weather, fog, rain, strong wind, hail, smoke cloud}

The data distributed among several time intervals is presented in Figure 1.

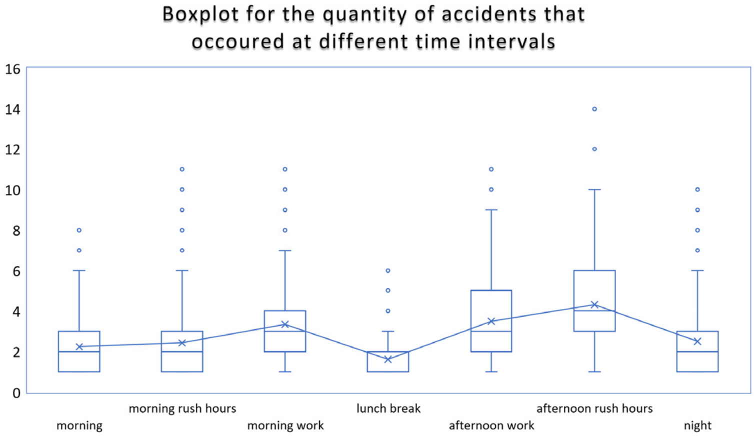


Figure 1. Box plots of the frequency of accidents that occurred at the different time intervals (according to the TIME field of Table 3).

With regard to meteorological factors, it was possible to categorize the accidents according to the different weather conditions in which they occurred. Taking into account the probability of rain as $P(C)$ and the probability of having an accident as $P(A)$, the graph in Figure 2 allows us to extract the probability of rain being related to an accident, i.e., $P(C \mid A)$. This is intended to facilitate a comparison between the probability of having an accident in rainy weather conditions, $P(A \mid C)$, and the probability of having an accident in fine weather conditions, $P(A \mid B)$. To make this comparison, the month of December was used as an example. Thus, $P(C \mid A)$ for the month of December is given by:

$$P(C|A) = \frac{144 + 126 + 132}{404 + 300 + 346} = 0.38 \tag{7}$$

where the numbers on the numerator are the number of accidents on rainy days in the months of December in 2019, 2020, and 2021, and the numbers in the denominator are the total number of accidents in the same month.

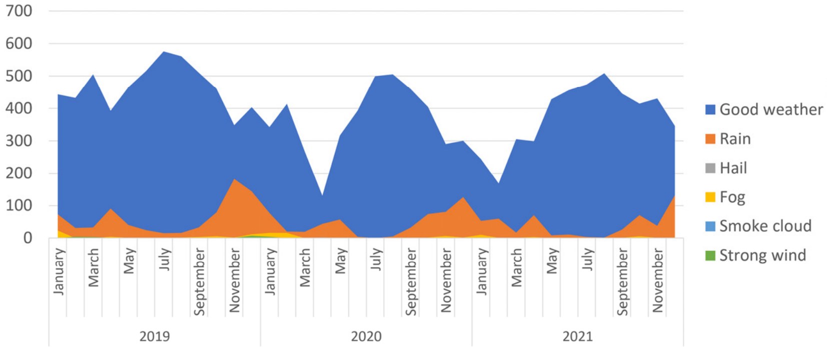


Figure 2. Number of accidents grouped by month, year, and type of weather condition.

As the average number of rainy days for the month of December in the Setubal district is 8.5 (information extracted from the Weather Spark website (https://pt.weatherspark.com/y/32195/Clima-caracter%C3%ADstico-em-Set%C3%BAbal-Portugal-durante-o-ano#Sections-Precipitation, (accessed on 1 November 2022))), we have:

$$P(C) = \frac{8.5}{31} = 0.27 \tag{8}$$

Using Bayes' theorem, we can compute:

$$P(A|C) = \frac{P(C|A) \times P(A)}{P(C)} = 1.4 \times P(A) \tag{9}$$

Using the same procedure for good weather, it can be concluded that the probability of an accident when it is raining is greater than the probability of an accident when the weather is good:

$$P(A|B) = 0.85 \times P(A) < 1.4 \times P(A) = P(A|C) \tag{10}$$

The location type of the accident (i.e., inside or outside an urban region) was grouped by month and year.

Figure 3 represents the number of traffic accidents grouped by day of the week and by year.

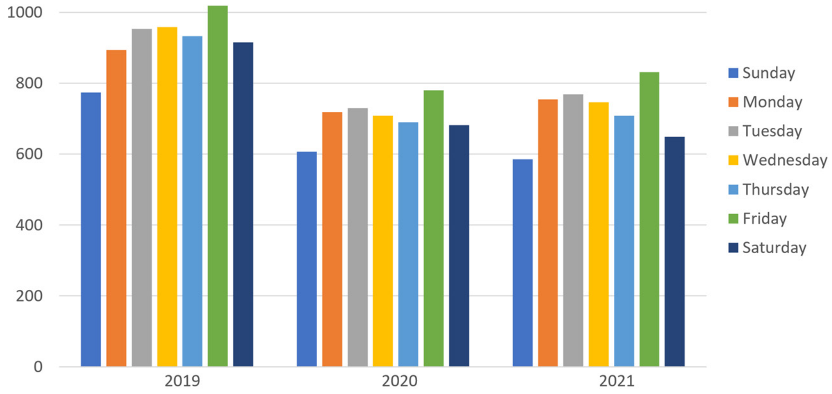


Figure 3. Grouping of the number of accidents by day of the week and year.

Figure 3 shows that the day of the week with more accidents than any other is Friday (that is the day of the week when most traffic congestion occurs [46]), and the days with the fewest accidents are Saturday and Sunday.

Figure 4 presents data on road accidents grouped by month to facilitate the comparison of accidents between different years but in the same month.

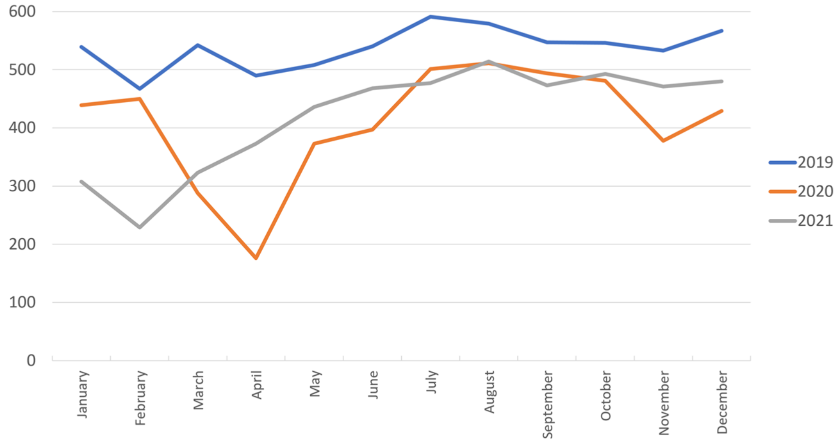


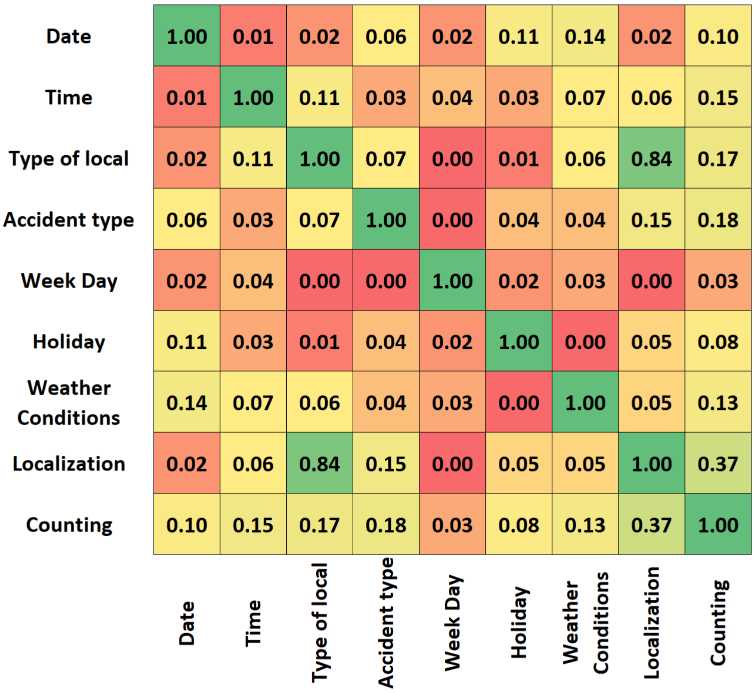
Figure 4. Number of monthly accidents before COVID-19 (2019) and during the COVID-19 pandemic (2020 and 2021).

The information presented in Figure 4 shows an approximate average value of 550 accidents per month in 2019. In early 2020, COVID-19 expanded worldwide, leading to a pandemic being declared in March 2020, and many countries declared a lockdown that affected almost all activities. As a result, the number of accidents reached a minimum in April 2020, with fewer than 200 accidents. The monthly number of accidents gradually increased, with slight reductions in November 2020 and February 2021 due to government measures to encourage remote working, owing to concern about the peaks in the prevalence of the disease in Europe.

4.2. Selection of Attributes

It was possible to achieve the different correlation values for pairs of nominal and numerical categorical variables (using the Kruskal-Wallis test) and for pairs of nominal categorical variables with nominal categorical variables (using Cramer's V).

In Figure 5 we can see that, in the variable "counting", which represents the accident count, the variables with the highest correlation are the time of day, the type of place, the location, and the meteorological factors. The type of accident, which here represents its severity, was considered only to be verified if there was a correlation between the severity of the accident and the number of accidents that occurred; this is confirmed, since there is a low correlation of 0.18 for this pair of variables.



	Date	Time	Type of local	Accident type	Week Day	Holiday	Weather Conditions	Localization	Counting
Date	1.00	0.01	0.02	0.06	0.02	0.11	0.14	0.02	0.10
Time	0.01	1.00	0.11	0.03	0.04	0.03	0.07	0.06	0.15
Type of local	0.02	0.11	1.00	0.07	0.00	0.01	0.06	0.84	0.17
Accident type	0.06	0.03	0.07	1.00	0.00	0.04	0.04	0.15	0.18
Week Day	0.02	0.04	0.00	0.00	1.00	0.02	0.03	0.00	0.03
Holiday	0.11	0.03	0.01	0.04	0.02	1.00	0.00	0.05	0.08
Weather Conditions	0.14	0.07	0.06	0.04	0.03	0.00	1.00	0.05	0.13
Localization	0.02	0.06	0.84	0.15	0.00	0.05	0.05	1.00	0.37
Counting	0.10	0.15	0.17	0.18	0.03	0.08	0.13	0.37	1.00

Figure 5. Correlations of Cramer V and Kruskal-Wallis test for the most relevant pairs of variables.

The RBA and SBS were used for feature selection processing data only from motorways, since it was concluded that only for motorways is it possible to obtain a credible model for accident prediction. Despite the results depending on the classification algorithm used, there were several attributes where both algorithms agreed (see Table 4).

Table 4. Relevance of features for the creation of predictive models obtained with RBA and SBS for incidents that occur on motorways.

Motorways	Considered Relevant by Both Algorithms	Considered Irrelevant by Both Algorithms
RBA & SBS	Rain, morning work, afternoon rush hours, Friday, Saturday, August, February	Sunday

4.3. Data Mining

Owing to the importance of the location of accidents, it was decided to group the data by their location: motorways; national roads or itineraries; and village roads. To achieve an accident-risk evaluation, we decided to divide the risk into classes, as shown in Table 5.

Table 5. Intervals of number of accidents that correspond to the different classes of risk.

Classes	Range of the number of accidents
Low Risk	<1.5
Medium Risk	≥1.5 & <2.5
High Risk	≥2.5

The data were grouped according to accidents on motorways, on village roads, on itineraries, on national roads, and in municipalities. It was decided to divide the classes using the intervals defined in Table 5. The purpose of keeping the classification range was to facilitate an understanding of the behavior of each individual model for each type of road.

Accidents on motorways represented 9.3% of all accidents; accidents on itineraries or national roads represented 30% of all accidents; and accidents outside the previous two catgories, including those in village streets, represented 60.7% of all accidents.

Starting with the motorway dataset, the best models produced for each algorithm according to the metrics used are represented in Tables 6–8.

Table 6. Results for motorways: 9.3% of total accidents.

Algorithm	*MAE* (Distance)	*Accuracy* (%)
kNN	0.74	56%
Linear Regression	0.63	57%
Lasso Regression	0.60	54%
Ridge Regression	0.61	52%
Decision Tree	0.69	56%
Neural Network	0.57	89%

Table 7. Results for itineraries or national roads: 30.0% of total accidents.

Algorithm	*MAE* (Distance)	*Accuracy* (%)
kNN	0.30	81%
Linear Regression	0.27	86%
Lasso Regression	0.28	86%
Ridge Regression	0.28	80%
Decision Tree	0.31	76%
Neural Network	0.55	87%

Table 8. Results for village roads: 60.7% of total accidents.

Algorithm	*MAE* (Distance)	*Accuracy* (%)
kNN	0.93	48%
Linear Regression	0.85	50%
Lasso Regression	0.80	51%
Ridge Regression	0.79	50%
Decision Tree	0.91	55%
Neural Network	0.52	88%

This option was chosen based on information set out in the box diagrams shown in Figure 6. As can be seen, the variance of values is greater for the dataset relating to motorways and village roads.

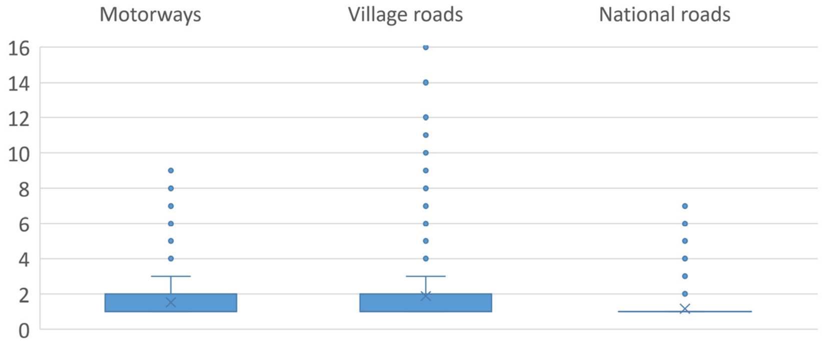


Figure 6. Box plots for the values of the frequency of accidents on the different type of roads.

In the dataset of itineraries or national roads, there is, in most cases, only one accident in each of the time intervals; therefore, a higher frequency of accidents would be necessary for the model to be of use.

From Table 9 we see that the motorway, despite being the location with the lowest number of accidents (for the district of Setubal), is the location with the highest concentration of accidents per area when compared with villages, and with the highest concentration of accidents per individual motorway when compared with the concentration of accidents per individual national road. The motorway is the type of road where there are more injuries and deaths per accident; this can be seen in Table 10.

Table 9. Summary of the best results from Tables 6–8.

Algorithm (Regression with Neural Network)	*MAE* (Distance)	*Accuracy* (%)
General model	0.49	88%
Motorways (9.3% of total accidents)	0.57	89%
Itineraries or national roads (30% of total accidents)	0.55	87%
Village roads (60.7% of total accidents)	0.52	88%

Table 10. Information relating to the number of injuries and deaths per accident.

Type of Road	Percentage of Accidents involving Injuries or Deaths	N° of Injured/Dead per Accident
Motorway	25.1%	1.7
Village roads	17.3%	1.2
Itineraries or national roads	29.8%	1.42

Additionally, the motorway is the location where it is possible for the National Guard to carry out more effective surveillance; village roads, by contrast, are extremely numerous, and there is a large area where accidents can occur.

5. Conclusions

In this work, data-mining methods for the prediction of the risk of road accidents were analyzed. Data on accident reports were made available by the National Guard and related to accidents that occurred in the Setubal region from 2019 to 2021. We describe the process followed to develop accident-prediction methods. This process consists of three modules: (i) data selection and collection, (ii) pre-processing, and (iii) the use of mining algorithms.

Through a preliminary data analysis, it was concluded that the highest concentration of accidents is seen between 17 h and 20 h. It was also possible to conclude that rain is the meteorological factor with the highest probability of increasing the risk of an accident. A further conclusion is that the day of the week on which more accidents occur than any other is Friday. These conclusions are consistent with the literature [47].

Through an analysis of the correlation between the different variables, it was possible to conclude that location is the variable that most influences the frequency of accidents. Following on from this conclusion, the information characterizing the accidents was grouped according to the type of road where the accidents occurred. For this reason, it was necessary to create different models for each set. In addition to the location, the correlation between variables also highlighted other factors that influenced the frequency of accidents, such as the time of day, the meteorological conditions, and whether the accident occurred in a village or elsewhere. After dividing the data set into the three types of location (motorways, national roads or itineraries, and villages), it was possible, using the feature-selection algorithms, to understand which features most influence each type of accident location.

The data-mining problem was approached as a regression problem, since the target variable was the frequency of accidents in the defined time range. The mining algorithms tested were kNN, simple linear regression, Lasso and Ridge, the Decision Tree for regression, and the traditional neural network, both for the initial dataset and for the datasets divided by location in the following sets: motorways, national roads or itineraries, and villages. The best result was achieved through the neural network. However, for each set, different models were produced, with different architectures (number of nodes, training periods, etc.). The best result occurred for the motorway dataset. The motorway, despite being the location with the lowest number of accidents, is the one with the highest density of accidents per area when compared with villages; it also features the highest density of accidents per road, when compared with the concentration of accidents on national routes or roads. In addition, the motorway is the location where there are more injuries and deaths per accident. The motorway is also the location where it is possible for the National Guard to carry out more effective surveillance, since in the villages there are a large number of roads, and consequently there is a vast area where accidents can occur; however, the density of accidents on village roads is low.

This work is of value owing to the fact that it was possible to obtain good results for the prediction of the risk of accidents on motorways, but with variables that can be predicted in a future time frame. For example, it is possible today to make a weather forecast for the next week; we can distinguish the different days of the week in the future; we know which days will be holidays, etc. By using input data that relating only to future events, we are able to obtain an accident-risk result for a day in the future and thus enable the police to improve their forward planning.

In future work, the first step would be to improve data collection to ensure that the geolocation of accidents was acquired, making it possible to opt for more complex approaches. Another important variable to obtain would be the level of human mobility; it would be possible to acquire this by using applications such as Google Maps or Waze, or simply by recording the speed at which Uber taxis or other companies' vehicles travel.

Author Contributions: J.S.S. and A.B. proposed the idea and concept; D.D. developed the software under the supervision of J.S.S. and A.B.; all authors revised and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Military Academy Research Center (CINAMIL) and by FCT through the projects UID/FIS/04559/2019, HAVATAR (PTDC/EEI-ROB/1155/2020) and LARSyS (UIDB/50009/2020).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable (This study does not involve humans).

Data Availability Statement: The data used in this work was imported into the huge private National Guard database. Other researchers who intend to use this data in the future must formalize in writing a request for access to the National Guard's private database, which will decide on a case-by-case basis.

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

References

1. Hengst, M.D.; Mors, J.T. Community of Intelligence: The Secret Behind Intelligence-Led Policing. In Proceedings of the 2012 European Intelligence and Security Informatics Conference, Odense, Denmark, 22–24 August 2012; pp. 22–29.
2. Castro, Y.; Kim, Y.J. Data mining on road safety: Factor assessment on vehicle accidents using classification models. *Int. J. Crashworthiness* **2016**, *21*, 104–111. [CrossRef]
3. Kashyap, J.; Chandra, A.; Singh, P. Mining Road Traffic Accident Data to Improve Safety on Road-related Factors for Classification and Prediction of Accident Severity. *Int. Res. J. Eng. Technol.* **2016**, *10*, 2395–2456.
4. Hussain, S.; Muhammad, L.J.; Ishaq, F.S.; Yakubu, A.; Mohammed, I.A. Performance evaluation of various data mining algorithms on road traffic accident dataset. *Smart Innov. Syst. Technol.* **2019**, *106*, 67–78. [CrossRef]
5. Kumeda, B.; Zhang, F.; Zhou, F.; Hussain, S.; Almasri, A.; Assefa, M. Classification of road traffic accident data using machine learning Algorithms. In Proceedings of the 2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN), Chongqing, China, 12–15 June 2019; pp. 682–687. [CrossRef]
6. Chen, Q.; Song, X.; Yamada, H.; Shibasaki, R. Learning deep representation from big and heterogeneous data for traffic accident inference. In Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016), Phoenix, AZ, USA, 21 February 2016; pp. 338–344.
7. Yuan, Z.; Zhou, X.; Yang, T. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, United Kingdom, 19 August 2018; Volume 18, pp. 984–992. [CrossRef]
8. Krukowicz, T.; Firląg, K.; Chrobot, P. Spatiotemporal analysis of road crashes with animals in Poland. *Sustainability* **2022**, *14*, 1253. [CrossRef]
9. Billah, K.; Sharif, H.O.; Dessouky, S. How Gender Affects Motor Vehicle Crashes: A Case Study from San Antonio, Texas. *Sustainability* **2022**, *14*, 7023. [CrossRef]
10. Saveliev, A.; Lebedeva, V.; Lebedev, I.; Uzdiaev, M. An approach to the automatic construction of a road accident scheme using UAV and deep learning methods. *Sensors* **2022**, *22*, 4728. [CrossRef]
11. Tajnik, S.; Luin, B. Impact of Driver, Vehicle, and Environment on Rural Road Crash Rate. *Sustainability* **2022**, *14*, 15744. [CrossRef]
12. Bokaba, T.; Doorsamy, W.; Paul, B.S. Comparative study of machine learning classifiers for modelling road traffic accidents. *Appl. Sci.* **2022**, *12*, 828. [CrossRef]
13. Islam, M.K.; Gazder, U.; Akter, R.; Arifuzzaman, M. Involvement of Road Users from the Productive Age Group in Traffic Crashes in Saudi Arabia: An Investigative Study Using Statistical and Machine Learning Techniques. *Appl. Sci.* **2022**, *12*, 6368. [CrossRef]
14. Islam, M.K.; Reza, I.; Gazder, U.; Akter, R.; Arifuzzaman, M.; Rahman, M.M. Predicting Road Crash Severity Using Classifier Models and Crash Hotspots. *Appl. Sci.* **2022**, *12*, 11354. [CrossRef]
15. Mesquitela, J.; Elvas, L.B.; Ferreira, J.C.; Nunes, L. Data Analytics Process over Road Accidents Data—A Case Study of Lisbon City. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 143. [CrossRef]
16. Guido, G.; Shaffiee Haghshenas, S.; Shaffiee Haghshenas, S.; Vitale, A.; Astarita, V.; Park, Y.; Geem, Z.W. Evaluation of Contributing Factors Affecting Number of Vehicles Involved in Crashes Using Machine Learning Techniques in Rural Roads of Cosenza, Italy. *Safety* **2022**, *8*, 28. [CrossRef]
17. Kim, H.; Kim, J.-T.; Shin, S.; Lee, H.; Lim, J. Prediction of Run-Off Road Crash Severity in South Korea's Highway through Tree Augmented Naïve Bayes Learning. *Appl. Sci.* **2022**, *12*, 1120. [CrossRef]
18. Rodionova, M.; Skhvediani, A.; Kudryavtseva, T. Prediction of crash severity as a way of road safety improvement: The case of Saint Petersburg, Russia. *Sustainability* **2022**, *14*, 9840. [CrossRef]
19. Infante, P.; Jacinto, G.; Afonso, A.; Rego, L.; Nogueira, V.; Quaresma, P.; Saias, J.; Santos, D.; Nogueira, P.; Silva, M. Comparison of statistical and machine-learning models on road traffic accident severity classification. *Computers* **2022**, *11*, 80. [CrossRef]
20. Goldschmidt, R.; Passos, E.; Bezerra, E. *Data Mining, Conceitos Técnicas, Algoritmos, Orientações e Aplicações*; Elsevier: Rio de Janeiro, Brasil, 2015.
21. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 82–88.
22. Hendrickx, T.; Cule, B.; Meysman, P.; Naulaerts, S.; Laukens, K.; Goethals, B. Mining association rules in graphs based on frequent cohesive itemsets. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Ho Chi Minh City, Vietnam, 19–22 May 2015; pp. 637–648.
23. Agarwal, S. Data mining: Data mining concepts and techniques. In Proceedings of the 2013 International Conference on Machine Intelligence and Research Advancement, Katra, India, 21 December 2013; pp. 203–207.

24. Zhang, S.; Zhang, C.; Yang, Q. Data preparation for data mining. *Appl. Artif. Intell.* **2003**, *17*, 375–381. [CrossRef]
25. Mueller, J.P.; Massaron, L. *Deep Learning for Dummies*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
26. Berry, M.W.; Mohamed, A.; Yap, B.W. *Supervised and Unsupervised Learning for Data Science*; Springer: Cham, Switzerland, 2020.
27. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
28. Sen, P.C.; Hajra, M.; Ghosh, M. Emerging Technology in modelling and graphics. *Singap. Springer Singap.* **2020**, *937*, 99.
29. Belanche, L.A.; González, F.F. Review and evaluation of feature selection algorithms in synthetic problems. *arXiv* **2011**, arXiv:1101.2320.
30. Indrakumari, R.; Poongodi, T.; Singh, K. Introduction to Deep Learning. In *Advanced Deep Learning for Engineers and Scientists*; Springer: Cham, Switzerland, 2021; pp. 1–22.
31. Eisenberg, D. The mixed effects of precipitation on traffic crashes. *Accid. Anal. Prev.* **2004**, *36*, 637–647. [CrossRef]
32. Hayat, R.B.; Debbarh, M.; Antoniou, C.; Hayat, R.B.; Debbarh, M.; Antoniou, C.; Yannis, G. Explaining the road accident risk: Weather effects. *Accid. Anal. Prev.* **2013**, *1*, 456–465. [CrossRef]
33. Tamerius, J.D.; Zhou, X.; Mantilla, R.; Greenfield-Huitt, T. Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions. *Weather. Clim. Soc.* **2016**, *8*, 399–407. [CrossRef]
34. Febres, J.D.; Garca-Herrero, S.; Herrera, S.; Gutirrez, J.M.; Lpez-Garca, J.R.; Mariscal, M.A. Influence of seat-belt use on the severity of injury in traffic accidents. *Eur. Transp. Res. Rev.* **2020**, *12*, 1–12. [CrossRef]
35. Musile, G.; Pigaiani, N.; Sorio, D.; Colombari, M.; Bortolotti, F.; Tagliaro, F. Alcohol-associated traffic injuries in Verona territory: A nine-year survey. *Med. Sci. Law* **2021**, *61*, 7–13. [CrossRef] [PubMed]
36. Song, Y.; Kou, S.; Wang, C. Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach. *J. Saf. Res.* **2021**, *76*, 64–72. [CrossRef]
37. Martn-delosReyes, L.M.; Martnez-Ruiz, V.; Rivera-Izquierdo, M.; Jimnez-Mejas, E.; Lardelli-Claret, P. Is driving without a valid license associated with an increased risk of causing a road crash? *Accid. Anal. Prev.* **2021**, *149*, 1–7. [CrossRef]
38. Zhang, Z.; McDonnell, K.T.; Zadok, E.; Mueller, K. Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE Trans. Vis. Comput. Graph.* **2015**, *21*, 289–303. [CrossRef]
39. Bhattacharya, A.; Dunson, D.B. Simplex factor models for multivariate unordered categorical data. *J. Am. Stat. Assoc.* **2012**, *107*, 362–377. [CrossRef]
40. Leon, A.C. Descriptive and Inferential Statistics. *Compr. Clin. Psychol.* **1998**, *3*, 243–285. [CrossRef]
41. Sun, J. *The Microbiome in Health and Disease Preface*; Academic Press: Cambridge, MA, USA, 2020; Volume 171, pp. XV–XVI.
42. Urbanowicz, R.J.; Meeker, M.; La Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* **2018**, *85*, 189–203. [CrossRef]
43. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
44. Marcano-Cedeño, A.; Quintanilla-Domínguez, J.; Cortina-Januchs, M.; Andina, D. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In Proceedings of the IECON 2010—36th Annual Conference on IEEE Industrial Electronics Society, Glendale, AZ, USA, 7 November 2010; pp. 2845–2850.
45. Molina, L.C.; Belanche, L.; Nebot, À. Feature selection algorithms: A survey and experimental evaluation. In Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, 9 December 2002; pp. 306–313.
46. SeguroPorDias. O Congestionamento nas Estradas da Cidade do Porto (Congestion on the Roads of the City of Porto). Available online: https://seguropordias.pt/blog/tr%C3%A2nsito-porto-portugal (accessed on 29 December 2022).
47. Ren, H.; Song, Y.; Wang, J.; Hu, Y.; Lei, J. A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3346–3351. [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.