sustainability

MDPI

Article

Applying Data Mining Approaches for Analyzing Hazardous Materials Transportation Accidents on Different Types of Roads

Shanshan Wei [1], Xiaoyan Shen [2], Minhua Shao [1,*] and Lijun Sun [1]

[1] College of Transportation Engineering, Tongji University, Shanghai 201804, China; weishan1995@foxmail.com (S.W.); 2018222024@chd.edu.cn (L.S.)
[2] School of Automotive, Chang'an University, Xi'an 710064, China; sxy719@163.com
* Correspondence: shaominhua@tongji.edu.cn; Tel.: +86-135-6409-9112

Abstract: With the increase in the demand for and transportation of hazardous materials (Hazmat), frequent Hazmat road transport accidents, high death tolls and property damage have caused widespread societal concern. Therefore, it is necessary to carry out risk factor analysis of Hazmat transportation; predict the severity of accidents; and develop targeted, extensive and refined preventive measures to guarantee the safety of Hazmat road transportation. Based on the philosophy of graded risk management, this study used a priori algorithms in association rule mining (ARM) technology to analyze Hazmat transport accidents, using road types as classification criteria to find rules that had strong associations with property-damage-only (PDO) accidents and casualty (CAS) accidents under different road types. The results indicated that accidents involving PDO had a strong association with weather (WEA), traffic signals (TS), surface conditions (SC), fatigue (FAT) and vehicle safety status (VSS), and that accidents involving CAS had a strong association with VSS, equipment safety status (ESS), time of day (TOD) and WEA when urban roads were used for Hazmat transportation. Among Hazmat transport incidents on rural roads, the incidence of PDO accidents was associated with intersections (IN), SC, WEA, vehicle type (VT), and segment type (ST), while the occurrence of CAS accidents was associated with qualification (QUA), ESS, TS, VSS, SC, WEA, TOD, and month (MON). Strong associations between the occurrence of PDO accidents and related items, such as IN, SC, WEA and FAT, and the occurrence of CAS accidents and related items, such as ESS, TOD, VSS, WEA and SC, were identified for Hazmat road transport accidents on highways. The accident characteristics exemplified by strongly correlated rules were used as the input to the prediction model. Considering the scarcity of these events, four prediction models were selected to predict the severity of Hazmat accidents on each road type employing four analyses, and the most suitable prediction model was determined based on the evaluation criteria. The results showed that extreme gradient boosting (XGBoost) is preferable for predicting the severity of Hazmat accidents occurring on urban roads and highways, while nearest neighbor classification (NNC) is more suitable for predicting the severity of Hazmat accidents occurring on rural roads.

Keywords: hazardous materials; association rules mining; accident prevention; different road types

check for updates

Citation: Wei, S.; Shen, X.; Shao, M.; Sun, L. Applying Data Mining Approaches for Analyzing Hazardous Materials Transportation Accidents on Different Types of Roads. *Sustainability* **2021**, *13*, 12773. https://doi.org/10.3390/su132212773

Academic Editors: Changxi Ma and Xuecai Xu

Received: 9 October 2021
Accepted: 15 November 2021
Published: 18 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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

China has become the world's largest producer and seller of chemicals, and the accompanying logistics have also increased rapidly with the booming development of production, sales and related activities. Due to the uneven geographical distribution of product supply and product demand in China's industries, approximately 95% of hazardous materials (Hazmat) in China must be transported off-site [1]. Due to policy constraints, geographical differences, and nonuniform technical conditions, information systems are not interoperable, and railroads, waterways and other modes of transport are not fully utilized. As a result, most Hazmat must be transported by road. In 2020, China's total shipments of Hazmat reached 1.7 billion tons, of which approximately 1.2 billion tons,

or 69% of the total transport of Hazmat, was moved by road, accounting for 3.5% of the total road transport of goods. Nearly 95,000 heavy-duty Hazmat vehicles carry 2.2 million tons of Hazmat on roads every day [2].

The continuous increase in the frequency of transport makes transport accidents increasingly frequent as well. The substances that characterize Hazmat are flammable, explosive, toxic and corrosive, and they have other dangerous characteristics that often cause major accidents and result in casualties and property damage. In addition, Hazmat leakage can damage ecological safety barriers and reduce the ability to provide sustainable ecological services for human survival and development [3]. Between 2006 and 2017, 5203 traffic participants died in 3974 events involving the transportation of Hazmat in China. These data demonstrate that, each day, more than one person dies in China as a result of Hazmat accidents [4]. Li Wei et al. [5] pointed out that the statistics of accidents involving Hazmat that occurred in China from 2010–2017 show that 278 accidents occurred in the transportation segment, resulting in 306 deaths—a fatal accident rate second only to that of the production process.

Since China places great emphasis on curbing serious accidents, implementing safety grading control and hidden danger investigation and management, the 14th Five-Year Plan for China's national economic and social development has also put forward new requirements for road safety. Therefore, it is beneficial to build a safe and sustainable national transportation system by using road type as the grading standard, conducting a comprehensive safety risk assessment of Hazmat road transportation routes, exploring and analyzing the causes of Hazmat road transportation accidents, and conducting risk grading and control. Risk grading and control can achieve targeted, comprehensive coverage and support refinement to prevent accidents, and the primary aspect of accident prevention is to perform an in-depth investigation and data analysis.

The elements of a transport system interact with each other, and changes in the behavior or properties of any one of them impact the functioning of the entire system. Accidents in the road transport of Hazmat are a direct consequence of the dysfunction, loss of control or failure of one or more parts of the transport system. The research methods used in previous literature are mainly statistical methods that require predefined relationships between dependent and independent variables, have a sound theoretical basis and clear calculation structures, and effectively reveal the characteristics of Hazmat road transport accidents [6–11]. Multiple studies have found that factors, such as people, vehicles, equipment, Hazmat, roads, environment and management, all have a relationship with the occurrence of accidents. Between 1986 and 1987, Andersson [12] employed statistical approaches to evaluate 570 Hazmat accidents. He determined that the kind of Hazmat, road, vehicle, and location all impacted the severity of the incidents. According to Yang et al. [13], during 2000–2008, 46.6% of Hazmat road transport incidents were caused by bad road conditions, 13.7% by driver mistake, and 9% by mismanagement. Xing et al. [14] built a random parameter ordered probit model to investigate the effect of contributing variables on the severity of accidents. The findings suggested that a greater degree of injury may be associated with Hazmat type, mishandling, driver tiredness, speeding, tunnels, hills, county roads, dry roads, winter, night, more than two cars, rear-end collisions, and explosions. This research by Azimi et al. [15] used a random parameter logit model to examine the severity of heavy truck rollover collisions in Florida. They found that crashes are more severe when Hazmat spills are present. Ma et al. [16] used an ordered logit model to predict the risk of several Hazmat incidents. A study of the Hazmat accident severity factors using elasticity theory. In addition, the severity of road Hazmat accidents was shown to be influenced by illegal activities and dangerous driving conduct.

Unfortunately, the correlation between crash risk factors as independent variables hurt the statistical analysis has been reported by some literature that argues that, once the assumptions of the generalized linear model (GLM) are violated, it could introduce biased inferences about the influence of the factors of interest [17]. Machine learning approaches are adaptable to processing outliers, missing data, and noisy data and are

versatile, requiring no or few previous assumptions about input variables [18–26]. These methods can effectively solve the problems associated with the above statistical methods and achieve more accurate predictions of accident severity [17]. Huting et al. [27] used the random forest model to identify factors that affected the probability of a responsible bus accident in the Minneapolis–Saint Paul, Minnesota, metropolitan area. They found that bus drivers are at greater risk toward the middle of their shift, especially when in dense traffic. Yassin et al. [28] used a hybrid k-means and random forest algorithm approach to road accident prediction and model interpretation. They found that driver experience and day, light condition, driver age, and service year of the vehicle were the decisive contributing factors for serious injury, light injury, and fatal severity, respectively. Harb et al. [29] investigated the features of drivers, vehicles, and settings associated with accident avoidance strategies. Additionally, the random forests approach was used to prioritize the drivers, vehicles, and environmental variables of accident avoidance operations. They discovered that obstructions to drivers' sight, physical disability, and attention were all connected with collision avoidance actions during incidents. Additionally, the speed limit was connected with avoidance movements for rear-end crashes, and vehicle type was associated with avoidance efforts for head-on and angle collisions. Lv et al. [30] investigated how to identify the traffic accident potential by using the k-nearest neighbor method with real-time traffic data and found that the k-nearest neighbor method outperformed the conventional c-means clustering method. An investigation by Ma et al. [31] of the 3146 traffic deaths in Los Angeles between 2010 and 2012, using a methodological framework of XGBoost and grid analysis, revealed the eight most essential elements that contributed to the fatalities. Drunk driving, partying, rear-end crashes, poor illumination, pedestrian contact, motorcycle contact, the day of the week, and the hour of the day were the most significant influences, in that order. Soleimani et al. [32] utilized XGBoost to determine the relative importance of crossing closure criteria using accidents data from 18,485 road-rail grade crossings in the United States. The model's accuracy was 0.991, which was higher than that of decision trees and random forests. Parsa et al. [33] applied XGBoost and Shapley Additive exPlanations (SHAP) for real-time accident detection and characterization. The findings indicated that XGBoost could reliably detect accidents with a 99% detection rate, 79% accuracy rate, and a 0.16% false alarm rate. Additionally, it was suggested that speed, population, network, land use, and weather conditions all substantially affected the likelihood of accidents.

However, since machine learning methods are 'black box' approaches, the analysis and prediction of severity classification often lack a direct and clear interpretation of accident severity and related variables [34]. In contrast, the association rule mining algorithm, as an unsupervised algorithm that does not rely on any assumptions or a priori knowledge to discover hidden but meaningful connections in a dataset, can discover the associations between different accident characteristics, including their severity [35–37]. This data mining methodology has been identified as a potential decision support tool for traffic safety engineers [38–41]. Montella et al. [42] investigated the contributory crash factors in 15 urban roundabouts located in Italy and to study the interdependences between these factors. They identified numerous contributory factors related to the road and environment deficiencies but unrelated to the road user or the vehicle. Das et al. [43] adopted an association rules mining method to investigate driver lane-keeping ability in foggy weather conditions. Their study indicated that affected visibility, male drivers, a higher number of lanes, the presence of horizontal curves, was associated with poor lane-keeping performance in several rules. Langford et al. [44] utilized an unsupervised association mining approach to uncover trends in a database of vehicle-pedestrian collisions. They discovered that highlighting traffic illumination helped to mitigate the severity of pedestrian accidents. According to Xu et al. [45], the association rule mining approach was used to find sets of accident contributing elements that were often found together in significant casualty collisions. According to researchers, there is a complicated connection between road user behavior, vehicle parameters, road geometry qualities, and environmental elements that

lead to significant casualty collisions. Yu et al. [46] used an a priori approach to find significant correlations between crash severity and crash-related parameters. The created rules showed that male drivers aged 29 are more likely to be engaged in fatal incidents on non-separable roads, while property damage crashes are more likely to occur in towns.

Furthermore, despite this discovery, there is still a lack of study that uses data mining technologies to uncover the hidden correlations in Hazmat road transport accident-related datasets. A primary objective of this study is to apply the association rule mining (ARM) approach to extensively explore the characteristics and contributing factors of Hazmat road transport accidents that occur on different kinds of roads in light of this understanding. At the same time, multiple prediction models are evaluated to determine the best severity prediction model for accidents occurring on different road types. The findings of this research will aid in the complete understanding of basic patterns of Hazmat road transport accidents on various road types to target and guide policy and decision-making initiatives to enhance the safety of Hazmat road transport.

2. Methods

2.1. Association Rule Mining

ARM is a typical unsupervised learning technique that uses data mining ideas to uncover hidden correlations between variables in a database [36]. Its functions include discovering frequent itemsets and discovering association rules, and its process is composed of the following two steps:

(1) The frequent itemset mining method is used to find all the frequent itemsets.
(2) Strong association rules are produced according to the obtained frequent itemsets.

2.1.1. Apriori Algorithm

The Apriori algorithm is a classic data mining algorithm that follows the a priori principle; that is, if an itemset is an infrequent itemset, then all its supersets are also infrequent itemsets, and if a rule does not have a strong association relationship, then all the subsets of the rule also do not have a strong association relationship. This approach can avoid the calculations caused by infrequent candidate itemsets. After several passes over the dataset, multiple robust candidate itemsets and multiple strongly correlated rules can be generated [37].

The process of determining the set of frequent items by the Apriori algorithm is shown in Figure 1.

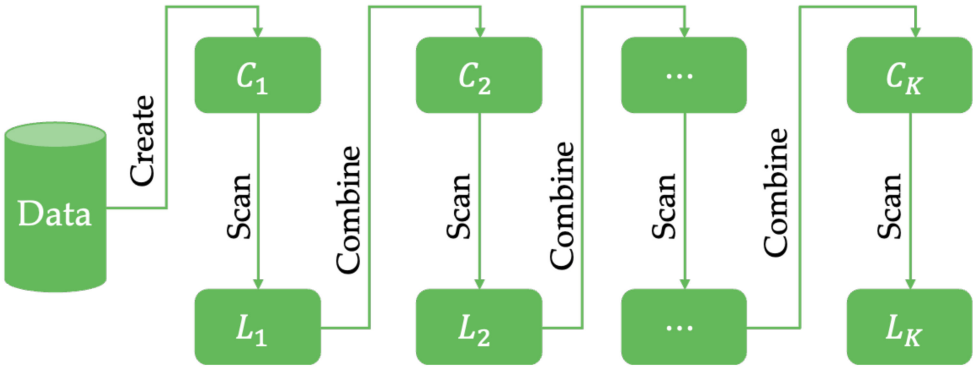


Figure 1. The process of determining a set of frequent items.

$C_1, C_2, \cdots, C_k \cdots, C_K$ denote 1-item sets, 2-item sets..., k-item sets, respectively. $L_1, L_2, \cdots, L_k \cdots, L_K$ denote the frequent itemsets with k items. Scan represents the dataset scanning function, which filters the itemsets by the set minimum support and discards those that do not meet the minimum support. The remaining itemsets that meet the requirements constitute the set L_k. The different frequent k itemsets are combined into the candidate $K + 1$ itemsets.

After determining the frequent itemsets, the association rule mining criteria are used to find strong association relationships. The process is as follows. First, we start with a frequent itemset, create a list of rules with only one element on the right-hand side, and then calculate those rules' confidence and lift values. Next, the remaining rules are merged to create a new list of rules with two elements on the right-hand side of the rule, and the confidence and lift values of those rules are calculated. This step is repeated by adding elements to the rule's right-hand side, iterating through all the rules, and finally selecting the rules that satisfy the threshold.

2.1.2. Association Rule Assessment Criteria

Support, confidence and lift values are often used assessment metrics for frequent itemsets and strong association rules. An implication is defined in the Hazmat road transport accident dataset D for two sets of itemsets X (the antecedent) and Y (the consequent) of the form $X \to Y$ that satisfy the requirements $X, Y \subseteq I$ and $X \cap Y = \{\varnothing\}$.

The support of the rule is the probability that X and Y hold together among all the possible presented cases. Support can be mathematically defined, as shown in Equation (1) below.

$$\text{Support}(X \to Y) = \text{P}(X \cap Y) = \frac{|X \cup Y|}{|D|}, \tag{1}$$

where $|X \cup Y|$ is the number of times both itemsets X and Y occur together and $|D|$ is the number of items in the accident database.

The confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X, as defined as Equation (2).

$$\text{Confidence}(X \to Y) = \text{P}(Y|X) = \frac{\text{P}(XY)}{\text{P}(X)} = \frac{|X \cup Y|}{|X|}, \tag{2}$$

where $|X|$ denotes the number of occurrences of itemset X, and $|X \cup Y|$ denotes the number of occurrences of both X and Y itemsets.

The lift takes into account how much the likelihood of occurrence of Y varies as a result of X. Equation (3) below may be used to compute the lift value mathematically.

$$\text{Lift}(X \to Y) = \frac{\text{Confidence}(X \to Y)}{\text{Support}(Y)} = \frac{\text{Support}(X \cup Y)}{\text{Support}(X) \cdot \text{Support}(Y)}. \tag{3}$$

Lift = 1 indicates no correlation between the antecedent and consequent, Lift > 1 indicates a positive correlation between the antecedent and consequent, and Lift < 1 indicates a negative correlation between the antecedent and consequent.

2.2. Prediction Models

2.2.1. Ordinal Logit (OL)

Ordered logit models are derived from econometric models and are one of the common models used to perform ordered discrete data analysis and forecasting [16]. These models map the latent, difficult-to-observe, continuous variable y_i^* into an observable ordered variable y to represent the severity propensity, and y_i^* and y_i are related by Equation (4).

$$y_i = j, if \gamma_{j-1} < y_i^* \le \gamma_j, \tag{4}$$

where $\tau = (\gamma_0, \gamma_1, \cdots \gamma_j, \cdots \gamma_J)$ denotes the set of accident severity grading points.

Accident severity is represented by the ordered variable y, and the various characteristics affecting accident severity are represented by X. The general form of the model is $y_i^* = \beta X_i^T + \varepsilon_i$.

Where $X_i^T = x_{i1}, x_{i2}, \ldots, x_{ik} \ldots x_{iK}; n = 1, \ldots, N; k = 1, \ldots, K$ is the vector of accident severity influencing factors; $\beta = (\beta_1, \beta_2, \ldots \beta_k, \beta_K)$ is the parameter corresponding to an influencing factor, where x_{ik} is the observed value of the kth influencing factor of the ith

accident; N is the total number of accident samples; K is the number of influencing factors for each accident; and ε_i is the random error term, which is the sum of other factors that are difficult to observe but have an impact on the severity of the accident.

In the ordered logit model, ε_i obeys the Gumbel distribution, its probability density function is $f(\varepsilon_i)$, and its cumulative distribution function is $F(\varepsilon_i)$, $E(\varepsilon_i) = 0$.

From Equations (1) and (2), it can be derived that the probability of the ith accident being of severity j is

$$\mathrm{P}(y_i = \mathrm{j}|\, X_i, \beta, \tau) = P\left(\gamma_{j-1} - \beta X_i^T < y_i^* \le \gamma_j - \beta X_i^T\right) = F\left(\gamma_j - \beta X_i^T\right) - F\left(\gamma_{j-1} - \beta X_i^T\right),$$

where the ith accident occurrence ratio (odds) is $\frac{P(y_i \le j|X_i,\beta,\tau)}{1-P(y_i \le j|X_i,\beta,\tau)} = \frac{P(y_i \le j|X_i,\beta,\tau)}{P(y_i > j|X_i,\beta,\tau)} = \exp\left(\gamma_j - X_i^T\beta\right)$.

2.2.2. Nearest Neighbor Classification (NNC)

NNC, sometimes referred to as the k nearest neighbors method, classifies an observation of interest by examining the closest k observations, and if the majority of these k instances belong to a specific class, then the new data belongs to that class. Its essential elements are the k value [47], the distance between two instances in the feature space [48], and the classification decision rule. The choice of k value starts from k = 1 and gradually increases, and the k value is determined according to the classification effect. The choice of the distance calculation method is decided according to the scenario of application and the characteristics of the data itself, which are generally Euclidean distance and Manhattan Distance [49]. The classification decision rule is generally a majority voting rule (majority voting rule), that is, the majority of the k neighboring categories are used as the categories of the test samples.

2.2.3. Random Forests (RF)

The core of the RF algorithm is to construct multiple mutually independent evaluators and then to average or majority vote principle on their predictions to decide the results of the evaluators. The primary computational process includes sample set selection, construction of decision tree, and combination in three parts [50].

(1) Sample set selection.

In an original training set containing n samples, K rounds of data extraction are performed; in each round of data extraction, random sampling is performed, one sample is sampled each time, and the sample is put back into the original training set before the following sample is taken, so that n times are collected. Finally, the K datasets are as large as the original training set is obtained. Since it is random sampling, the other sampled sets are also different each time the dataset is different from the original dataset.

(2) Decision tree construction.

The core problem of decision tree is to find out the right features to make judgments, that is, how to branch. When each sample has M attributes, and each node of the decision tree requires splitting, m attributes are randomly chosen from these M attributes that fulfill the criterion $m \ll M$. Then, using some approach (Gini coefficient or Information Gain), one of these m properties is chosen as the node's splitting attribute. It continues until no more splitting is possible.

(3) Decision tree combination.

A decision tree's importance is equated to the significance of the outcomes since each decision tree in this research is autonomous. In the RF combination phase, the weight of each decision tree is equal. All of the decision trees weigh in on the final categorization outcomes.

2.2.4. Extreme Gradient Boosting (XGBoost)

The objective function of XGBoost [51] is expressed as Equation (5).

$$Obj = \sum_{i=1}^{m} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad (5)$$

where i is the i th sample in the dataset, m is the total amount of data imported into the kth tree, and K is all trees created. When creating t trees solely, the equation should be $\sum_{k=1}^{t} \Omega(f_k)$. y_i is the actual label, $\hat{y}_i$ is the predicted value, and Ω is an equation that determines the tree model's complexity based on the tree's structure.

When t trees are created, the predicted value $\hat{y}_i$ in the traditional loss function is expressed as Equation (6).

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t-1} f_k(x_i) + f_t(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \quad (6)$$

As a result, the classic loss function is connected to all well-established trees. $\hat{y}_i$ stores the outcomes of all tree iterations, making a direct connection between the tree's structure and the model effect. The objective function is expressed as Equation (7).

$$Obj = \sum_{i=1}^{m} l\left(y_i^{(t)}, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{k=1}^{t-1} \Omega(f_k) + f_t. \quad (7)$$

Using Taylor's formula as a guide, the objective function may be expressed as shown in Equation (8) after expansion.

$$Obj = \sum_{i=1}^{m} \left[l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right) + f_t(x_i) g_i + \frac{1}{2} (f_t(x_i))^2 h_i \right] + \sum_{k=1}^{t-1} \Omega(f_k) + \Omega(f_t), \quad (8)$$

where $g_i = \frac{\partial l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}}$ and $h_i = \frac{\partial^2 l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right)}{\partial^2 \left(\hat{y}_i^{(t-1)}\right)}$ are the first- and second-order derivatives of the loss function $l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right)$ over $\hat{y}_i^{(t-1)}$, respectively.

The constant term is irrelevant to the result of the tth iteration, so the constant terms $l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right)$ and $\sum_{k=1}^{t-1} \Omega(f_k)$ are removed from the objective function. The objective function is expressed as Equation (9).

$$Obj = \sum_{i=1}^{m} \left[f_t(x_i) g_i + \frac{1}{2} (f_t(x_i))^2 h_i \right] + \Omega(f_t). \quad (9)$$

The structure of the tree is redefined according to Equation (10).

$$f_t(x_i) = w_{q(x_i)}, \quad (10)$$

where $q(x_i)$ is the leaf node where sample x_i is located. $w_{q(x_i)}$ is the score obtained by this sample falling in the $q(x_i)$ leaf node of the tth tree.

If a tree has a total of T leaf nodes, each with an index of j, the weight of the samples in the leaf nodes is w_j. Equation (11) describes the complexity of the model $\Omega(f)$.

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} {w_j}^2. \quad (11)$$

The objective function may be turned into Equation (12) by including the tree's structure into the loss function and specifying the set of samples stored on a leaf with index j as I_j.

$$Obj = \sum_{j=1}^{T} \left[w_j \sum_{i \in I_j} g_i + \frac{1}{2} {w_j}^2 \left(\sum_{i \in I_j} h_i + \lambda \right) \right] + \gamma T. \quad (12)$$

2.2.5. Predictive Performance Evaluation Indexes

The confusion matrix is a special kind of table that is used to visualize an algorithm's performance. Table 1 illustrates the confusion matrix for a two-class classifier, where TN represents the number of correct predictions that an instance is negative, FP represents the number of incorrect predictions that an instance is positive, FN represents the number of incorrect predictions that an instance is negative, and TP represents the number of correct predictions that an instance is positive. While the optimal outcome is to achieve a high overall model prediction accuracy, greater preference is given to the prediction of CAS accidents; that is, it is more desirable to capture the occurrence of a few categories of accidents. Additionally, the influence of the imbalance of sample categories on the index results in the actual accident data should be eliminated. Therefore, the evaluation index for the overall effectiveness of the model, accuracy; the evaluation index that can capture the particular category, recall; and the index that can equalize the impact of the sample imbalance on the index results, the area under the receiver operating characteristics (ROC) curve (AUC), were chosen [52].

Table 1. Confusion matrix.

Confusion Matrix		**Predicted Condition**	
		Positive	**Negative**
True condition	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Accuracy is the proportion of all correctly judged results, as shown in Equation (13). Recall is the probability of being predicted as a positive sample out of an actual positive sample, as shown in Equation (14). FPR is the proportion of false positive prediction values within the sum of true negative and false positive values, as shown in Equation (15). When the distribution of positive and negative samples in the test set changes, the ROC curve with the TPR as the *y*-axis and the FPR as the *x*-axis can be kept constant; the higher the TPR (Recall) and the smaller the FPR, the more efficient the model and algorithm. From a geometric point of view, the larger the AUC is, the better the model, so the AUC can be used as a metric measuring the reliability of the algorithm and the model.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{13}$$

$$\text{Recall(TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}. \tag{15}$$

3. Data Sources

In this paper, we selected 900 accidents resulting from the transportation of Hazmat by road between 2016 and 2020, and these data were obtained from the Hazardous Chemicals Registration Center of the Ministry of Emergency Management of China. After screening and integration, the final data used for analysis included 862 accident cases, mainly involving attributes such as accident casualties, driver attributes, vehicle attributes, road attributes, environmental attributes, and Hazmat types. According to the road types where the accidents occurred, they were divided into three road types (rural road, urban road and highway) with large differences and analyzed separately, accounting for 11.14%, 23.43%, and 65.43% of the total number of accidents, respectively. Depending on the casualties of the accidents, the accident severities were divided into property-damage-only (PDO) and casualty (CAS) categories, accounting for 43.97% and 56.03% of the total number of accidents, respectively. To facilitate the modeling and analysis of the data, the accident

characteristics need to be coded. The statistical results after feature coding are shown in Table 2.

Table 2. Coding and descriptive statistics of features.

Feature	Code and Description	Count	Feature	Code and Description	Count
Hazardous Materials: HM	Gases: 2	191	Road Alignment: RA	Straight: 1	530
	Flammable liquids: 3	487		Ramps: 2	82
	Flammable solids: 4	11		Curved ramp: 3	7
	Oxidizers and organic peroxides: 5	16		Curve: 4	243
	Poisonous and infectious substances: 6	15	Vehicle Type: VT	Tank: 1	745
		141		Cargo-truck: 2	96
	Corrosives: 8	1		Other: 3	21
Season: SEA	Spring: 1	221	Surface Condition: SC	Dry: 1	726
	Summer: 2	248		Wet: 2	83
	Autumn: 3	208		Ice: 3	28
	Winter: 4	185		Waterlogged: 4	25
Month: MON	January: 1	60	Segment Type: ST	Ordinary segment: 1	671
	February: 2	41		Tunnel: 2	40
	March: 3	78		Bridge: 3	32
	April: 4	75		Entrance and exit: 4	26
	May: 5	68		Station: 5	74
	June: 6	62		Risky segment: 6	19
	July: 7	102	Intersection: INT	Yes: 1	128
	August: 8	84		No: 0	734
	September: 9	74	Traffic Signal: TS	Yes: 1	837
	October: 10	74		No: 0	25
	November: 11	60	Fatigue: FAT	Yes: 1	175
	December: 12	84		No: 0	687
Time of Day: TOD	[1–3]: 1	162	Moving Status: MS	Go straight: 1	487
	[4–6]: 2	122		Stop: 2	63
	[7–9]: 3	150		Turn: 3	252
	[10–12]: 4	92		Downhill: 4	11
	[13–15]: 5	167		Avoid: 5	49
	[16–18]: 6	31	Weather: WEA	Sunny: 1	778
	[19–21]: 7	51		Rain: 2	46
	[22–24]: 8	87		Snow: 3	22
Equipment Safety Status: ESS	Safety: 1	723		Fog: 4	16
	Malfunction: 0	139	Qualification: QUA	Yes: 1	808
Vehicle Safety Status: VSS	Safety: 1	779		No: 0	54
	Malfunction: 0	83			

4. Results and Discussion

4.1. Association Rule Mining

To arrive at significant results, it is critical to calibrate the minimal support and confidence levels. Defining proper cutoff points will result in the discovery of novel rules. A trial-and-error approach using iterative support and confidence combinations was utilized to develop a fair set of thresholds for investigations, including different levels of road. Then, using the lift values, itemsets with a high association to accident severity were retrieved. Increased lift values suggest higher links between the rule's or right-side item's (RSI or Y) consequence and the rule's or left-side item's antecedent (LSI or X).

4.1.1. Urban Roads

The minimum support, confidence and lift thresholds were defined as 0.3, 0.9, and 1.1, respectively. A total of 50 rules were generated using accident severity as a consequence. The top ten rules in descending order of lift values for different severity levels were selected and are presented in Table 3. Figure 2 shows the relationship between each antecedent and consequent.

Table 3. Top 10 rules ranked by the lift value of each severity (urban roads).

No.	Association Rules	Support	Confidence	Lift
1	{WEA-1, TS-1, SC-1}→{Severity-PDO}	0.255	0.902	2.059
2	{SC-1}→{Severity-PDO}	0.280	0.951	2.050
3	{WEA-1, SC-1}→{Severity-PDO}	0.280	0.951	2.050
4	{TS-1, SC-1}→{Severity-PDO}	0.275	0.941	2.045
5	{FAT-0}→{Severity-PDO}	0.275	0.941	2.045
6	{WEA-1}→{Severity-PDO}	0.295	0.980	2.026
7	{WEA-1, TS-1}→{Severity-PDO}	0.290	0.971	2.021
8	{TS-1}→{Severity-PDO}	0.295	0.990	1.995
9	{VSS-1}→{Severity-PDO}	0.255	0.902	1.954
10	{TS-1, VSS-1}→{Severity-PDO}	0.255	0.902	1.954
1	{VSS-1, ESS-1, TOD-1}→{Severity-CAS}	0.275	0.960	2.276
2	{VSS-1, ESS-1}→{Severity-CAS}	0.275	0.960	2.276
3	{TOD-1, ESS-1}→{Severity-CAS}	0.280	0.970	2.210
4	{ESS-1}→{Severity-CAS}	0.280	0.970	2.202
5	{WEA-1, TOD-1, ESS-1}→{Severity-CAS}	0.246	0.900	2.188
6	{WEA-1, ESS-1}→{Severity-CAS}	0.246	0.900	2.181
7	{WEA-1, VSS-1, ESS-1}→{Severity-CAS}	0.246	0.900	2.089
8	{TOD-1, VSS-1, WEA-1}→{Severity-CAS}	0.246	0.900	2.089
9	{VSS-1}→{Severity-CAS}	0.250	0.910	2.081
10	{VSS-1, TOD-1}→{Severity-CAS}	0.250	0.910	2.081

(1) PDO Accidents.

As shown in Table 3, the occurrence of PDO accidents had a strong association with WEA, TS, SC, FAT and VSS. The highest lift value is 2.059 for the LSI term X {WEA-1, TS-1, SC-1}, which indicates that the probability of PDO accidents occurring under clear weather, dry road surface and up to standard road traffic signs is 2.059 times that of the average occurrence of PDO accidents on urban roads. This means that clear weather, a good road surface environment and standard sign markings in the city have certain helpful effects on reducing the severity of accidents. These benefits may exist because clear weather provides drivers with a clear view and a better grasp of the surrounding environment [43]; the dry road surface ensures that there is enough friction between the vehicle and the road surface, which can balance with the large inertia force of the heavy-duty Hazmat transport vehicle and allows the driver to control the vehicle better when danger occurs; and the presence of sign markings regulates the behavior of road users, controls the speed of motor vehicles [29], and effectively separates pedestrians, nonmotorized vehicles and motor

vehicles, reducing the possibility of other road participants being involved in accidents and increasing the possibility of escape from Hazmat subaccidents.

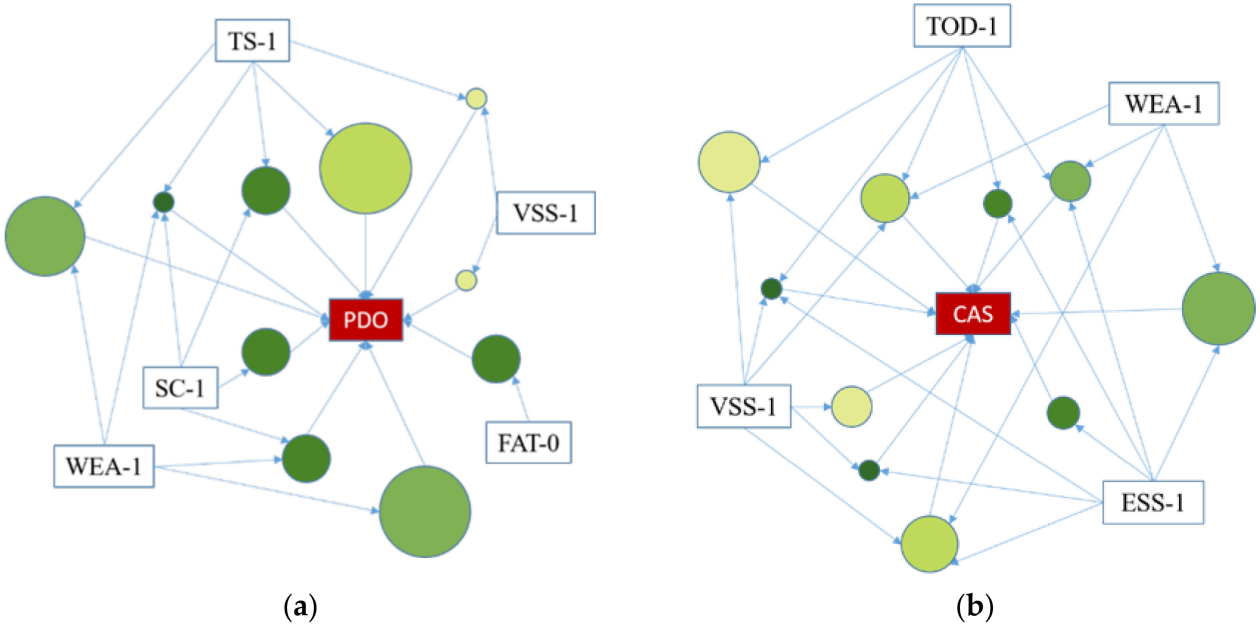


Figure 2. Graph-based visualization of association rules (urban roads). (**a**) Related to the PDO accidents; (**b**) Related to the CAS accidents.

(2) CAS Accidents.

As shown in Table 3, the occurrence of CAS accidents showed a higher propensity to be linked to VSS, ESS, TOD, WEA and QUA. The highest lift value is found to be 2.276 with rule {VSS-1, ESS-1, TOD-1} → {Severity-CAS}. This finding indicates that the probability of CAS accidents occurring at 1–3 a.m. under transport vehicles with good loading equipment and vehicle technology is 2.276 times that of the average occurrence of CAS accidents on urban roads. This means that although the vehicles entering the city and with their loading equipment are in great technical condition, the probability of causing casualties in accidents that occur in the early morning hours is also high. The reason for this may be that the urban transport management of Hazmat transport vehicles has access to strict standards, so access to the technical condition of the vehicle is relatively good [15]. Meanwhile, the urban area has strict requirements on the access time and roadway of Hazmat transport vehicles, the more concentrated access time is 23:00–5:00. According to human physiological characteristics, in the early morning hours, individuals are prone to fatigue and sleepiness, and the ability to accurately evaluate the driving environment and the correct handling of risk are reduced [14]. In addition, because there are fewer road users and law enforcement officers during the night, drivers may engage in illegal driving, hit-and-run and other dangerous behaviors.

(3) Proposals to Improve Safety in Hazmat Transport on Urban Roads.

To improve the safety of Hazmat road transport on urban roads, the following approaches should be taken into consideration. Law enforcement departments should increase supervision, enforcement, and accident tracking while increasing the cost of violations to eliminate unsafe driver behaviors. Road units should be used with increased investment in science, technology and personnel to provide timely detection and effective handling of dangerous road surface environments according to three aspects: initial forecasts (weather forecasts, event monitoring and regular analysis), timely warnings (information dissemination, extensive channels and directed push), and active interventions (road control, variable information and on-site command). It is also important to set standardized signs and markings [46]. Transportation companies should conduct psychological

tests for drivers to avoid hiring aggressive and dangerous drivers. Specialized departments and transport companies should also conduct regular emergency rescue training and drills for Hazmat transport accidents.

4.1.2. Rural Roads

Minimum support, confidence, and lift levels of 0.2, 0.80, and 1.1, respectively, were specified. For accident severity, a total of 67 rules were produced. Among these, the best ten rules ranked by lift values for various severity levels were chosen and are given in Table 4. The link between each antecedent and consequent is shown in Figure 3.

Table 4. Top 10 rules ranked by the lift value of each severity (rural roads).

No.	Association Rules	Support	Confidence	Lift
1	{IN-0, SC-1, WEA-1 }→{Severity-PDO}	0.201	0.901	1.943
2	{IN-0, SC-1}→{Severity-PDO}	0.201	0.887	1.924
3	{IN-0, WEA-1}→{Severity-PDO}	0.204	0.887	1.883
4	{VT-1, WEA-1}→{Severity-PDO}	0.213	0.877	1.872
5	{ST-1, WEA-1}→{Severity-PDO}	0.214	0.875	1.867
6	{IN-0}→{Severity-PDO}	0.205	0.871	1.865
7	{VT-1, SC-1}→{Severity-PDO}	0.231	0.868	1.861
8	{VT-1, SC-1, WEA-1}→{Severity-PDO}	0.217	0.863	1.857
9	{VT-1}→{Severity-PDO}	0.225	0.851	1.844
10	{WEA-1}→{Severity-PDO}	0.213	0.847	1.742
1	{QUA-0, TOD-1, VSS-1} →{Severity-CAS}	0.202	0.906	2.432
2	{MON-10, WEA-1, SC-1}→{Severity-CAS}	0.202	0.906	2.432
3	{QUA-1, ESS-1, TS-1}→{Severity-CAS}	0.213	0.895	2.413
4	{QUA-1, ESS-1, TS-1, VSS-1}→{Severity-CAS}	0.213	0.895	2.413
5	{ESS-1, TS-1}→{Severity-CAS}	0.248	0.865	2.222
6	{VSS-1, ESS-1}→{Severity-CAS}	0.244	0.850	2.160
7	{ESS-1}→{Severity-CAS}	0.220	0.837	2.131
8	{SC-1, ESS-1, WEA-1}→{Severity-CAS}	0.221	0.825	2.117
9	{TS-1}→{Severity-CAS}	0.235	0.820	2.099
10	{VSS-1, TS-1}→{Severity-CAS}	0.239	0.823	2.096

(1) PDO Accidents.

As shown in Table 4, the features with strong association rules with the occurrence of PDO accidents were IN, SC, WEA, VT, and ST. The highest lift value is found to be 1.943 with rule {IN-0, SC-1, WEA-1} → {Severity-PDO}. This rule signifies that the probability of PDO accidents occurring at nonintersections with clear weather and dry road surface environments is 1.943 times that of the average occurrence of PDO accidents on rural roads. This implies that the probability of a serious accident at an intersection is higher in clear weather and under good road surface conditions. The reasons for this phenomenon include the following: (1) Hazmat transport vehicles are mostly heavy semitrailers, with a higher center of gravity, in the process of turning, the centrifugal force of the curve and the lateral force of the vehicle rotation on the tires increase the lateral slip force, making the vehicle susceptible to rolling over. Moreover, large body, long wheelbase and the high driver position increase the vehicle blind spots and the area of the inner wheel difference [12]; (2) Road junctions are not equipped with signal lights or other traffic signs and markings; motor vehicles, nonmotorized vehicles and pedestrians are mixed; and personnel are more concentrated; (3) The supervision of road transportation of Hazmat in rural areas is low, and there are many driving violations, such as running red lights and speeding at intersections.

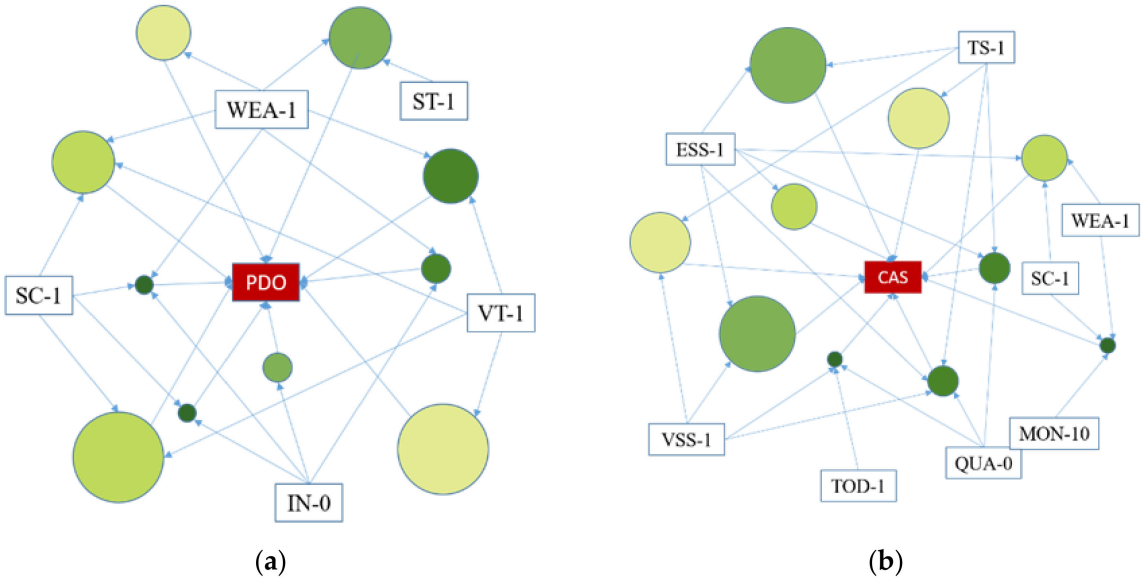


Figure 3. Graph-based visualization of association rules (rural roads). (**a**) Related to the PDO accidents; (**b**) Related to the CAS accidents.

(2) CAS Accidents.

We identified a strong association between the occurrence of PDO accidents and related items such as QUA, ESS, TS, VSS, SC, WEA, TOD and MON, as shown in Table 4. The rule with the highest lift value of 2.432 is {QUA-0, TOD-1, VSS-1} → {Severity-CAS}. This rule demonstrates that the probability of CAS accidents occurring in the early morning hours when drivers who are not qualified to drive tankers transporting Hazmat is 2.432 times that of the average occurrence of CAS accidents on rural roads. This means that driver qualification, accident time, and vehicle type significantly influence whether the accident will cause casualties. Possible reasons are mainly that rural areas have inadequate supervision over front-line transportation and Hazmat transportation enterprises and the lack of long-term management mechanisms. Some enterprises that have not obtained Hazmat transport qualifications attempt to avoid supervision by choosing rural roads. Drivers who are not qualified for transportation have insufficient knowledge of the physical and chemical characteristics of Hazmat, transportation requirements, precautions, rescue measures, and so forth. Moreover, their awareness of safety and legal systems is weak. Vehicles without transport qualifications do not meet the requirements for vehicle stability, braking, tank pressure resistance and impact, making them susceptible to leakage, fire or explosions. Road lights in rural areas are not well configured and have poor driving visibility in the early morning [28], and drivers are prone to fatigue, leading to a decrease in the perception of the surrounding environment and the ability to perform driving operations. The physical and chemical properties of different Hazmat differ greatly from each other, and the consequences of an accident are diverse and complex. Rescue work is highly professional and difficult to perform, requiring coordination with relevant departments to scientifically configure emergency rescue resources and equipment. However, a lack of resources for emergency treatment exists in rural areas, often resulting in missing the best time for disposal due to the lengthy delivery time. In addition, limited medical care in rural areas makes emergency medical assistance difficult and may miss the best time to treat the injury and cause it to worsen.

The lift value of rule {MON-10, WEA-1, SC-1} → {Severity-CAS} is also 2.432, which is interpreted as the probability of CAS accidents in October, when the weather is sunny and the road surface is dry, is 2.432 times higher than the average rate of CAS accidents on rural roads. This means that month, road surface conditions and weather conditions are strongly correlated with the occurrence of casualties in rural road accidents [43]. The possible reason for the above phenomenon is that October is the autumn harvest season, and roads with

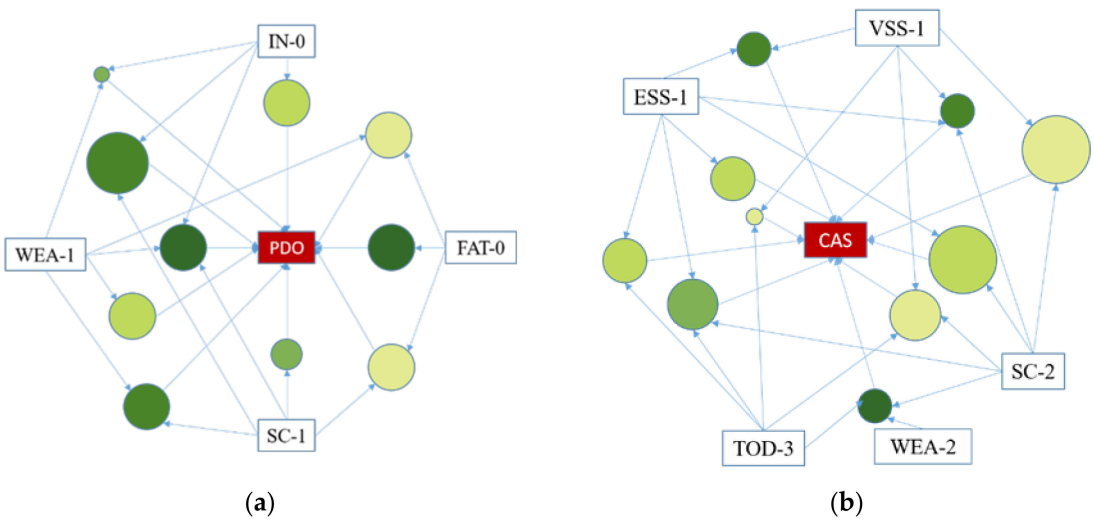
good road surface conditions are illegally occupied by farmers for grain drying in sunny weather. At this time, the flying chaff seriously affects driver and pedestrian vision; the surface of the rounded grain and smooth straw reduces the stability of the vehicle; and the contact of straw with the vehicle is likely to induce mechanical failure of the vehicle and can even ignite Hazmat in the process of friction, causing a fire or explosion and seriously affecting the safety of road traffic.

(3) Proposals to Improve Safety in Hazmat Transport on Rural Roads.

Additional mobile inspection stations for Hazmat should be set up at appropriate locations on rural roads to increase on-site supervision of Hazmat transport in rural areas. Led by the government, the joint management of several departments should crack down on the unlicensed transport of Hazmat, strengthen the source of management, and establish a long-term management mechanism. It is crucial to increase the number of streetlights and optimize traffic signal devices at intersections to improve the technical conditions of rural roads [53]. Observation windows should be fitted into the copilot doors of heavy vehicles, and these vehicles should be equipped with other side assistance systems, such as blind spot cameras and radar, to reduce the impact of visual blind spots on transport safety. By linking transport enterprises and regulatory units, the whole process of the transport supervision system can be established. Digital registration, intelligent query and route management of Hazmat, drivers and vehicle information can provide the behavior of drivers and escort personnel, the state of Hazmat, and the supervision and analysis of the state of vehicles and loading equipment to ensure the safety of the whole process of transportation. Access standards for Hazmat transportation drivers can be improved, including driving skills, risk avoidance skills, and risk awareness in the audit criteria, and driver education should be ongoing throughout drivers' professional careers. Finally, it is important to preset or optimize emergency rescue sites for Hazmat road transport accidents in rural areas and strengthen the linkage with the local public security traffic police, emergency fire, medical and health departments.

4.1.3. Highways

The minimum support, confidence and lift thresholds were defined as 0.2, 0.77, and 1.5, respectively. A total of 77 rules were generated using accident severity as a consequence (RSI). The top ten rules in descending order of lift values for different severity levels were selected and are presented in Table 5. Figure 4 shows the relationship between each antecedent and consequent.



(**a**) (**b**)

Figure 4. Graph-based visualization of association rules (highways). (**a**) Related to the PDO accidents; (**b**) Related to the CAS accidents.

Table 5. Top 10 rules ranked by the lift value of each severity (highways).

No.	Association Rules	Support	Confidence	Lift
1	{IN-0, SC-1, WEA-1}→{Severity-PDO}	0.210	0.966	2.044
2	{FAT-0}→{Severity-PDO}	0.210	0.889	1.961
3	{IN-0, SC-1}→{Severity-PDO}	0.235	0.877	1.883
4	{SC-1, WEA-1}→{Severity-PDO}	0.225	0.862	1.855
5	{SC-1}→{Severity-PDO}	0.213	0.843	1.745
6	{IN-0, WEA-1}→{Severity-PDO}	0.203	0.825	1.676
7	{WEA-1}→{Severity-PDO}	0.223	0.782	1.664
8	{IN-0}→{Severity-PDO}	0.224	0.773	1.656
9	{WEA-1, FAT-0}→{Severity-PDO}	0.223	0.772	1.645
10	{FAT-0, SC-1}→{Severity-PDO}	0.220	0.766	1.631
1	{SC-2, WEA-2, TOD-3}→{Severity-CAS}	0.213	0.903	2.482
2	{ESS-1, VSS-1, SC-2}→{Severity-CAS}	0.216	0.902	2.339
3	{ESS-1, VSS-1}→{Severity-CAS}	0.213	0.895	2.237
4	{ESS-1, TOD-3, SC-2}→{Severity-CAS}	0.224	0.887	2.203
5	{TOD-3, ESS-1}→{Severity-CAS}	0.220	0.873	2.151
6	{SC-2, ESS-1}→{Severity-CAS}	0.230	0.879	2.148
7	{ESS-1}→{Severity-CAS}	0.220	0.872	2.146
8	{TOD-3, VSS-1, SC-2}→{Severity-CAS}	0.224	0.874	2.070
9	{TOD-3, VSS-1}→{Severity-CAS}	0.204	0.861	2.067
10	{SC-2, VSS-1}→{Severity-CAS}	0.230	0.840	2.064

(1) PDO Accidents.

The severity of PDO accidents showed a higher propensity to be linked to IN, SC, WEA, and FAT, as shown in Table 5. The highest lift value is 2.044 for LSI {IN-0, SC-1, WEA-1}. The results reveal that the probability of PDO accidents occurring at non-intersections in clear weather with dry road surfaces is 2.044 times that of the average occurrence of PDO accidents on highways. In clear weather and good road surface conditions, the probability of serious accidents is higher at highway entrances and exits, especially at exits [5]. This is mainly because, in the exit diversion area, the speed difference between vehicles moving straight and vehicles turning becomes greater than the speed difference between vehicles on the general roadway. In particular, when the distance of the road sign at the front of the exit is not set reasonably (the sign is too close to the diversion nose), the driver needs to brake sharply and turn sharply before driving off ramp. However, compared to ordinary vehicles, Hazmat transport vehicles are heavier and have greater inertia, which makes it challenging to drive smoothly into the exit in a short time, thus causing traffic accident.

(2) CAS Accidents.

We identified strong associations between Hazmat road transport accidents involving casualties and related items, such as ESS, TOD, VSS, WEA and SC, as shown in Table 5. The highest lift value is found to be 2.482 with rule {SC-2, WEA-2, TOD-3} → {Severity-CAS}, which is interpreted as the probability of CAS accidents occurring at 7:00–9:00 a.m. on wet road surfaces being 2.482 times greater than that of the average occurrence of CAS accidents on highways. This rule signifies that there is a strong association among weather, road surface conditions, time of day and the occurrence of CAS accidents. This is mainly because of the fast travel speed and large traffic flow on the highway; at this time, any changes in the driving environment may bring safety hazards. For example, rainfall will reduce the visibility of the road, affecting the driver's ability to judge visually, and rain will also reduce the friction between the wheels and the ground, affecting the braking performance of the vehicle [13]. The fourth category of Hazmat regarding being in contact with water or moisture indicates that a violent chemical reaction will occur, releasing a large amount of flammable gas and heat, and in conditions that do not require an open flame, Hazmat may also burn or explode. Fog will cause diffusion and absorption of light and, coupled with small droplets of water in the air, it will result in objects on

the road becoming blurred, seriously hindering the driver's sight and easily causing rear-end accidents and other accidents [43]. The impact of snow and ice on transport safety is mainly in reduced visibility and the road friction coefficient. At the same time, according to the physical and chemical properties of Hazmat, certain types of Hazmat will change state under high temperature or cold conditions and influence the safety of load-bearing equipment. Additionally, adverse weather conditions can also have a negative impact on the rescue work of Hazmat transport accidents. Furthermore, although nighttime (23:00–06:00) prohibitions have been developed and implemented for Hazmat road transport vehicles, transport companies are driven by would-be interests to keep drivers in transport, which will lead to driver fatigue in the early morning and loss of accurate perception of the road environment and the ability to deal with emergencies. At the same time, because of the inherent physical and chemical characteristics of Hazmat, after an accident occurs, leakage, fire and explosion can easily occur; in the case of a concentration of a large number of vehicles, mass death and injury can easily occur.

(3) Proposals to Improve Safety in Hazmat Transport on Highways.

Modifications to accident-prone exits, such as installing speed feedback devices, appropriately increasing the distance between exit signs and ramps, placing crash barrels in exit triangles, and establishing emergency rescue facilities and equipment storage stations for Hazmat in service areas near entrances and exits, are suggestions for improving safety [14]. According to regional, seasonal, and other characteristics, regular Hazard surveys and updates of the permitted hours for road transport in Hazmat should be conducted. In addition, the following recommendations warrant further consideration: strengthening the inspection of fatigue driving at night, establishing joint liability and several liabilities between enterprises and drivers for fatigue driving, increasing the cost of noncompliance, and forcing enterprises to take primary responsibility for traffic safety. Road operators are able to deploy real-time weather monitoring systems and establish variable speed limit signs and treble horns to set reasonable speed limits and provide drivers with real-time information on the weather and road environment based on weather conditions.

4.2. Performance of the Prediction Models

The features that strongly correlate with accident severity under different road types are used as the input of each prediction model; the output results are also evaluated based on the evaluation indexes, and the evaluation results are shown in Table 6. From this analysis, it can be seen that XGBoost is more suitable for predicting the severity of road transport accidents involving Hazmat that occur on urban roads and highways, and NNC is more suitable for predicting the severity of accidents that occur on rural roads.

Table 6. Model assessment results.

Models		Urban Roads			Rural Roads			Highways		
		Accuracy	Recall	AUC	Accuracy	Recall	AUC	Accuracy	Recall	AUC
OL	PDO	0.516	0.376	0.503	0.603	0.389	0.506	0.612	0.472	0.517
	CAS		0.435			0.448	0.506		0.457	
NNC	PDO	0.801	0.772	0.870	0.815	0.828	0.915	0.794	0.800	0.860
	CAS		0.876			0.920	0.915		0.774	
RF	PDO	0.776	0.676	0.801	0.764	0.640	0.817	0.787	0.717	0.831
	CAS		0.966			0.940	0.817		0.903	
XGBoost	PDO	0.872	0.873	0.943	0.832	0.763	0.889	0.854	0.819	0.921
	CAS		0.951			0.890	0.889		0.973	

5. Conclusions

Safety accidents involving Hazmat during road transport occur occasionally, often causing high casualties, property damage and environmental damage, and the safety management of Hazmat transportation has gained widespread concern in society. Exploring the

leading causes and predicting the severity of Hazmat road transport accidents on different road types using road types as grading criteria is meaningful for building a community with traffic safety as a priority.

The main contributions of the paper are summarized below:

(1) The use of ARM can both compensate for the negative impact of correlation between risk factors as independent variables in accident severity analysis and fill the shortcoming in which machine learning cannot provide a reasonable explanation for the antecedents and consequences of accident occurrences. This approach also provides meaningful relationship maps for factors that are strongly associated with the occurrence of accidents of different severities under different road types.

The contributory factors for accidents of different severity on different road types explored using the Apriori algorithm are shown below:

(a) The features that had a strong association with the occurrence of PDO accidents during the transportation of Hazmat on urban roads were WEA, TS, SC, FAT and VSS, and the rule with the highest lift value was {WEA-1, TS-1, SC-1} → {Severity-PDO}. The features that had a strong association with the occurrence of accidents involving human casualties were VSS, ESS, TS, WEA and QUA, and the rule with the highest lift value was {VSS-1, ESS-1, TOD-1} → {Severity-CAS}.

(b) In accidents involving the transport of Hazmat occurring on rural roads, IN, SC, WEA, VT and ST were strongly associated with the occurrence of PDO accidents, and the highest lift value was found for the association rule {IN-0, SC-1, WEA-1} → {Severity-PDO}. The occurrence of CAS accidents had a strong association with QUA, ESS, TS, VSS, SC, WEA, TON and MON, and the highest lift values of the association rules were {QUA-0, TOD-1, VSS-1} → {Severity-CAS} and {MON-10, WEA-1, SC-1} → {Severity-CAS}.

(c) The occurrence of PDO accidents on highways had a strong association with IN, SC, WEA, and FAT. {IN-0, SC-1, WEA-1} → {Severity-PDO} was the rule with the highest lift value. Casualties on highways were more likely to be associated with ESS, TOD, VSS, WEA, and SC, and {SC-2, WEA-2, TOD-3} → {Severity-CAS} was the rule with the most significant lift value.

Based on the results of the study, possible preventive measures provided for the safety of road transport of Hazmat on different road types are as follows:

(a) To improve the safety of road transportation of Hazmat in urban areas, the road administration unit needs to continuously ensure good road surface conditions. The transportation management department should improve access standards and monitoring of Hazmat transport vehicles entering urban areas. Law enforcement departments need to increase the frequency of supervision, prosecution and punishment of Hazmat transport violations at night to eliminate dangerous driver behaviors. However, the main consideration is to avoid the routing of Hazmat transport vehicles through densely populated urban areas;

(b) Strengthening the monitoring and punishment of the illegal transport of Hazmat; improving the basic knowledge of traffic safety, safety and risk awareness of participants in traffic travel; optimizing the traffic infrastructure; and setting up more Hazmat rescue stations and equipping them with special materials for Hazmat accident rescue can reduce the incidence and severity of Hazmat road transport accidents in rural areas;

(c) The safety of highway transportation can be improved by establishing a whole-process supervision system for the transportation of Hazmat with the help of fifth-generation (5G) networks, big data, the Internet of Things, biotechnology and other technologies. The supervisory system can maintain continuous attention to driver fatigue, the state of Hazmat, the driving speed of the vehicle, and the driving environment of the highway and make appropriate interventions according to the actual situation in a timely manner.

(2) Selecting multiple prediction models, the features that exhibit strong correlation rules with accident severity are used as inputs to the prediction models, allowing the best prediction model to be determined for each road type for accident severity prediction in the transportation of Hazmat. The risk features discovered by the Apriori algorithm on different road types that lead to accidents of different severity were input into different prediction models for case studies and it was found that, when predicting the severity of Hazmat road transport accidents, XGBoost should be chosen for urban roads and highways, and NNC should be chosen for rural roads.

(3) Limitations and future research.

   (a) In this paper, when classifying the severity of Hazmat road transport accidents, only human casualty determinants are considered, and the salient features of environmental damage caused by Hazmat transport accidents are not reflected. In future research, it will be necessary to quantify the data on damage to the environment to achieve a more comprehensive analysis of the severity of accidents;

   (b) In this paper, when analyzing the factors influencing accident severity, objective factors such as roads, vehicles and the external environment are considered to influence accident severity, but the subjective aspects of drivers' psychological and physiological states are not analyzed. In future research, we need to obtain more information about the subjective state of drivers through questionnaires, surveillance videos and physiological state testing instruments to analyze the influence of drivers on the occurrence of accidents.

Author Contributions: Conceptualization, S.W. and M.S.; methodology, S.W.; software (python3.8, software manufacturer: Python Software Foundation, the Netherlands), S.W.; validation, S.W., M.S. and X.S.; formal analysis, S.W.; investigation, S.W.; resources, X.S.; data curation, S.W.; writing—original draft preparation, S.W.; writing—review and editing, S.W. and M.S.; visualization, S.W.; supervision, L.S.; project administration, L.S.; funding acquisition, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key Research and Development Project (2019YFE0112100), the Scientific Research Program Project of Shanghai Science and Technology Commission (16DZ1203602) and the National Natural Science Foundation of China (51208379).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. National Bureau of Statistics of the People's Republic of China. *China Statistical Yearbook*; National Bureau of Statistics of China: Beijing, China, 2020.
2. Ministry of Transport of the People's Republic of China. Statistical Data. Available online: http://www.mot.gov.cn/shuju/ (accessed on 4 September 2021).
3. Shen, X.; Wei, S. Severity analysis of road transport accidents of hazardous materials with machine learning. *Traffic Inj. Prev.* **2021**, *22*, 324–329. [CrossRef]
4. Zhao, L.; Qian, Y.; Hu, Q.-M.; Jiang, R.; Li, M.; Wang, X. An Analysis of Hazardous Chemical Accidents in China between 2006 and 2017. *Sustainability* **2018**, *10*, 2935. [CrossRef]
5. Wei, L.; Minghu, W.; Nan, Y. Statistical analysis of hazardous chemical accidents in a province from 2010 to 2017. *Ind. Saf. Environ. Prot.* **2018**, *44*, 54–57.
6. Oggero, A.; Darbra, R.M.; Munoz, M.; Planas, E.; Casal, J. A survey of accidents occurring during the transport of hazardous substances by road and rail. *J. Hazard. Mater.* **2006**, *133*, 1–7. [CrossRef]
7. Vlakveld, W.P.; Twisk, D.; Christoph, M.; Boele, M.; Sikkema, R.; Remy, R.; Schwab, A.L. Speed choice and mental workload of elderly cyclists on e-bikes in simple and complex traffic situations: A field experiment. *Accid. Anal. Prev.* **2015**, *74*, 97–106. [CrossRef]

8. Yuan, Q.; Yang, H.; Huang, J.; Kou, S.; Li, Y.; Theofilatos, A. What factors impact injury severity of vehicle to electric bike crashes in China? *Adv. Mech. Eng.* **2017**, *9*. [CrossRef]
9. Chen, F.; Chen, S. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. *Accid. Anal. Prev.* **2011**, *43*, 1677–1688. [CrossRef] [PubMed]
10. Zhang, H.-D.; Zheng, X.-P. Characteristics of hazardous chemical accidents in China: A statistical investigation. *J. Loss Prev. Process. Ind.* **2012**, *25*, 686–693. [CrossRef]
11. Wang, B.; Wu, C.; Reniers, G.; Huang, L.; Kang, L.; Zhang, L. The future of hazardous chemical safety in China: Opportunities, problems, challenges and tasks. *Sci. Total Environ.* **2018**, *643*, 1–11. [CrossRef] [PubMed]
12. Andersson, S.-E. Safe Transport of Dangerous Goods: Road, Rail or Sea? A Screening of Technical and Administrative Factors. *Eur. J. Oper. Res.* **1994**, *75*, 499–507. [CrossRef]
13. Yang, J.; Li, F.; Zhou, J.; Zhang, L.; Huang, L.; Bi, J. A survey on hazardous materials accidents during road transport in China from 2000 to 2008. *J. Hazard. Mater.* **2010**, *184*, 647–653. [CrossRef] [PubMed]
14. Xing, Y.; Chen, S.; Zhu, S.; Zhang, Y.; Lu, J. Exploring Risk Factors Contributing to the Severity of Hazardous Material Transportation Accidents in China. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1344. [CrossRef]
15. Azimi, G.; Rahimi, A.; Asgari, H.; Jin, X. Severity analysis for large truck rollover crashes using a random parameter ordered logit model. *Accid. Anal. Prev.* **2019**, *135*, 105355. [CrossRef]
16. Ma, C.; Zhou, J.-B.; Yang, D. Causation Analysis of Hazardous Material Road Transportation Accidents Based on the Ordered Logit Regression Model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1259. [CrossRef]
17. Iranitalab, A.; Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accid. Anal. Prev.* **2017**, *108*, 27–36. [CrossRef]
18. Chen, C.; Zhang, G.; Qian, Z.; Tarefder, R.A.; Tian, Z. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accid. Anal. Prev.* **2016**, *90*, 128–139. [CrossRef] [PubMed]
19. de Oña, J.; López, G.; Abellán, J. Extracting decision rules from police accident reports through decision trees. *Accid. Anal. Prev.* **2013**, *50*, 1151–1160. [CrossRef]
20. Abellán, J.; López, G.; de Oña, J. Analysis of traffic accident severity using Decision Rules via Decision Trees. *Expert Syst. Appl.* **2013**, *40*, 6047–6054. [CrossRef]
21. Zeng, Q.; Huang, H. A stable and optimized neural network model for crash injury severity prediction. *Accid. Anal. Prev.* **2014**, *73*, 351–358. [CrossRef] [PubMed]
22. Kashani, A.T.; Rabieyan, R.; Besharati, M.M. A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. *J. Saf. Res.* **2014**, *51*, 93–98. [CrossRef]
23. Liu, X. Risk Analysis of Transporting Crude Oil by Rail: Methodology and Decision Support System. *Transp. Res. Rec.* **2016**, *2547*, 57–65. [CrossRef]
24. Cui, Y.; He, Q.; Khani, A. Travel Behavior Classification: An Approach with Social Network and Deep Learning. *Transp. Res. Rec.* **2018**, *2672*, 68–80. [CrossRef]
25. Mafi, S.; AbdelRazig, Y.; Doczy, R. Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. *Transp. Res. Rec. J. Transp. Res. Board* **2018**, *2672*, 171–183. [CrossRef]
26. Trepanier, M.; Leroux, M.-H.; De Marcellis-Warin, N. Cross-analysis of hazmat road accidents using multiple databases. *Accid. Anal. Prev.* **2008**, *41*, 1192–1198. [CrossRef] [PubMed]
27. Huting, J.; Reid, J.; Nwoke, U.; Bacarella, E.; Ky, K.E. Identifying Factors That Increase Bus Accident Risk by Using Random Forests and Trip-Level Data. *Transp. Res. Rec.* **2016**, *2539*, 149–158. [CrossRef]
28. Yassin, S.S. Pooja Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Appl. Sci.* **2020**, *2*, 1–13. [CrossRef]
29. Harb, R.; Yan, X.; Radwan, E.; Su, X. Exploring precrash maneuvers using classification trees and random forests. *Accid. Anal. Prev.* **2009**, *41*, 98–107. [CrossRef]
30. Lv, Y.; Tang, S.; Zhao, H. Real-Time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method. In Proceedings of the 2009 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA, Zhangjiajie, China, 11–12 April 2009; Volume 3, pp. 547–550.
31. Ma, J.; Ding, Y.; Cheng, J.C.P.; Tan, Y.; Gan, V.J.L.; Zhang, J. Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective. *IEEE Access* **2019**, *7*, 148059–148072. [CrossRef]
32. Soleimani, S.; Mousa, S.R.; Codjoe, J.; Leitner, M. A Comprehensive Railroad-Highway Grade Crossing Consolidation Model: A Machine Learning Approach. *Accid. Anal. Prev.* **2019**, *128*, 65–77. [CrossRef]
33. Parsa, A.B.; Movahedi, A.; Taghipour, H.; Derrible, S.; Mohammadian, A.K. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid. Anal. Prev.* **2020**, *136*, 105405. [CrossRef]
34. Tang, J.; Liang, J.; Han, C.; Li, Z.; Huang, H. Crash injury severity analysis using a two-layer Stacking framework. *Accid. Anal. Prev.* **2018**, *122*, 226–238. [CrossRef]
35. Zhao, Q.; Bhowmick, S.S. *Association Rule Mining: A Survey*; Nanyang Technological University: Singapore, 2003; Volume 135.
36. Geng, X.; Liang, Y.; Jiao, L. ARC-SL: Association rule-based classification with soft labels. *Know.-Based Syst.* **2021**, *225*, 107116. [CrossRef]

37. Le, B.; Le, D.P.; Tran, M.-T. Hiding sensitive association rules using the optimal electromagnetic optimization method and a dynamic bit vector data structure. *Expert Syst. Appl.* **2021**, *176*, 114879. [CrossRef]
38. Hong, J.; Tamakloe, R.; Park, D. Application of association rules mining algorithm for hazardous materials transportation crashes on expressway. *Accid. Anal. Prev.* **2020**, *142*, 105497. [CrossRef]
39. Du, W.; Yang, J.; Powis, B.; Zheng, X.; Ozanne-Smith, J.; Bilston, L.; Wu, M. Understanding on-road practices of electric bike riders: An observational study in a developed city of China. *Accid. Anal. Prev.* **2013**, *59*, 319–326. [CrossRef]
40. Das, S.; Dutta, A.; Jalayer, M.; Bibeka, A.; Wu, L. Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration using 'Eclat' association rules to promote safety. *Int. J. Transp. Sci. Technol.* **2018**, *7*, 114–123. [CrossRef]
41. Weng, J.; Zhu, J.-Z.; Yan, X.; Liu, Z. Investigation of work zone crash casualty patterns using association rules. *Accid. Anal. Prev.* **2016**, *92*, 43–52. [CrossRef] [PubMed]
42. Montella, A. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. *Accid. Anal. Prev.* **2011**, *43*, 1451–1463. [CrossRef] [PubMed]
43. Das, A.; Ahmed, M.M.; Ghasemzadeh, A. Using trajectory-level SHRP2 naturalistic driving data for investigating driver lane-keeping ability in fog: An association rules mining approach. *Accid. Anal. Prev.* **2019**, *129*, 250–262. [CrossRef]
44. Langford, B.C.; Chen, J.; Cherry, C.R. Risky riding: Naturalistic methods comparing safety behavior from conventional bicycle riders and electric bike riders. *Accid. Anal. Prev.* **2015**, *82*, 220–226. [CrossRef]
45. Xu, C.; Bao, J.; Wang, C.; Liu, P. Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. *J. Saf. Res.* **2018**, *67*, 65–75. [CrossRef] [PubMed]
46. Yu, S.; Jia, Y.; Sun, D. Identifying Factors that Influence the Patterns of Road Crashes Using Association Rules: A case Study from Wisconsin, United States. *Sustainability* **2019**, *11*, 1925. [CrossRef]
47. Cover, T.M.; Hart, P.E. *Approximate Formulas for the Information Transmitted by a Discrete Communication Channel*; IEEE: Manhattan, NY, USA, 1952; Volume 24.
48. Grahne, G. *Encyclopedia of Database Systems*; Liu, L., Özsu, M., Eds.; Springer: New York, NY, USA, 2016. [CrossRef]
49. Lê, S.; Josse, J.; Rennes, A.; Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **2008**, *25*, 1–18. [CrossRef]
50. Scornet, E.; Biau, G.; Vert, J.-P. Consistency of random forests. *Ann. Stat.* **2015**, *43*, 1716–1741. [CrossRef]
51. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
52. Savolainen, P.T.; Mannering, F.L.; Lord, D.; Quddus, M.A. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accid. Anal. Prev.* **2011**, *43*, 1666–1676. [CrossRef]
53. Severino, A.; Pappalardo, G.; Curto, S.; Trubia, S.; Olayode, I.O. Safety Evaluation of Flower Roundabout Considering Autonomous Vehicles Operation. *Sustainability* **2021**, *13*, 10120. [CrossRef]