Analysis of the Severity of Road Accidents Using Combined Data Mining Techniques

César Corrales *, Juan Carlos Rubio-Romero, María del Carmen Pardo-Ferreira

Posted Date: 15 May 2026

doi: 10.20944/preprints202605.1045.v1

Keywords: road safety; data mining; decision trees; association rules; accident severity; heavy vehicles

 doi:10.20944/preprints202605.1045.v1

Article

Analysis of the Severity of Road Accidents Using Combined Data Mining Techniques

César Corrales [1,*], Juan Carlos Rubio-Romero [2] and María del Carmen Pardo-Ferreira [2]

[1] Department of Engineering, Pontifical Catholic University of Peru, Lima 15088, Peru

[2] Department of Economics and Business Administration, University of Málaga, Málaga 29071, Spain

* Correspondence: ccorral@pucp.edu.pe; Tel.: +51 999092797

Abstract

Road traffic accidents represent a critical road safety problem whose severity depends on the complex interaction of multiple factors. The study of these factors and their interrelationship has therefore long been a focus of scientific literature. The objective of this study is to analyze the factors that determine the severity of road accidents, identifying the most important ones and their correlations. An accident dataset incorporating variables related to infrastructure, location, time, and vehicle type was used to predict the Injury Severity Index (ISI), applying Association Rules to identify latent correlations and an Optimized Decision Tree (CART) model for hierarchical risk classification. The results reveal that the Type of Collision is the primary predictor of severity; collisions with objects or pedestrians showed a 100% confidence in resulting in low severity, while maximum severity is associated with heavy traffic and head-on or side-impact collisions. Critical scenarios were also identified during the early morning hours and in rural areas, primarily linked to trucks. The combined use of both tools provides a solid scientific basis for designing interventions on highly vulnerable road segments and during high-risk time periods.

Keywords: road safety; data mining; decision trees; association rules; accident severity; heavy vehicles

1. Introduction

The World Road Safety Status Report 2023 indicates that the annual number of traffic-related deaths has decreased to 1.19 million from 1.25 million in 2010. However, traffic injuries remain the leading cause of death among people aged 5 to 29, and 9 out of 10 deaths occur in low- and middle-income countries [1]. Traffic crashes are projected to be the seventh leading cause of death globally by 2030 [2]. Furthermore, the direct economic cost is very high; the National Highway Traffic Safety Administration (NHTSA) estimated that traffic accidents cost the U.S. **$871 billion** annually [3], and the International Road Assessment Programme (iRAP) estimates that road deaths and injuries worldwide cost $3.6 trillion annually, equivalent to more than 3% of global GDP [4].

While it is true that road safety has improved in many countries, the opposite is true in many developing countries, where fatalities have increased significantly, as is the case in Bangladesh, where the number of people killed in traffic accidents rose from 1,483 in 1993 to 4,046 in 2000—an increase of more than 200% [5]. This aligns with data presented by the World Health Organization, which highlights that low-income countries have a traffic accident mortality rate of 24.1 per 100,000 inhabitants, compared to 9.2 per 100,000 inhabitants in high-income countries [6]. In Peru, this rate is above 10 per 100,000, with 3,316 deaths recorded in 2023 [7], and the National Road Network accounts for 60% of all traffic accident fatalities [8]. In the United States, however, when considering non-urban areas—that is, rural roads and highways—41% of traffic fatalities occur on these roads [9].

It is very important to identify the factors that affect transportation safety, particularly on rural roads and highways, in order to have a positive impact on reducing fatalities and on the economy in the long term. Both globally and locally, there is a clear need to reduce traffic-related deaths and injuries

 doi:10.20944/preprints202605.1045.v1

[10]. In this regard, little is known about the determinants of the severity of road accidents, particularly involving buses, in developing countries [11]. Several factors are attributed to fatalities, including driver behavior, vehicle characteristics, road infrastructure, and system characteristics [12]. Driver drowsiness contributes to a substantial proportion of all traffic accidents [13]. On the other hand, the probability of traffic accidents is expected to increase on sections of road with a high volume of heavy vehicles [14]. Other risk factors for road accidents include time of day, lighting conditions, weather, driver characteristics and behavior, infrastructure characteristics, road geometry, and type of collision, among others [15,16].

In all cases, there are many causes involved in accidents, so it is important to establish the relationship between these causes and the significance of each of them in order to seek to reduce road accidents. While numerous studies have investigated the factors influencing accident risks, most of them have focused on common traffic accidents. Consequently, the factors contributing to road accidents involving buses—which typically result in a high number of casualties—and, above all, the interdependence among them, have not been studied with the depth required. As a result, measures taken to reduce the risk of these accidents may not be as effective. Furthermore, due to differences in the characteristics and mechanisms of accidents with serious casualties and ordinary accidents, traditional countermeasures for ordinary accidents may not effectively reduce the risks of accidents with serious casualties. In this regard, the use of data mining through statistical and computational methods to search for behavioral patterns in the data can provide us with hidden insights into these processes. The two main objectives of data mining tend to be prediction (using a dataset to predict unknown or future values) and description (finding patterns that describe the data) [17].

The main tasks of data mining include selecting the observed data (time series, cross-sectional, panel), the sample or population period, the frequency (annual, quarterly, etc.) and its measurement, selecting predictors (searching for a specific correlation structure), model specification: changing the model's assumptions, diagnostic tests, and data exploration [18]. Data mining extracts useful and interesting information from very large datasets and has become increasingly common in many fields, such as banking, insurance, medicine, retail, biology, and agriculture [19], and may involve the use of artificial intelligence, expert systems, machine learning, pattern recognition, statistics, intelligent databases, knowledge acquisition, data visualization, and other fields. Its task is to create models from data, and its development will largely determine the direction of data development [20]. Taking all this into account, road accident data was analyzed using association rule mining and association tree techniques, which aim to extract knowledge from observational data, considering that road accidents—which are typically monitored by SUTRAN—are generally caused by complex interactions among various contributing factors. The use of these techniques could reveal valuable relationships that have not been identified in existing studies. By identifying the contributing factors and their interdependencies, important information can be obtained to understand the reasons behind the occurrence of accidents and to develop effective policies and countermeasures to improve safety.

The main objective of this study is to demonstrate the importance of using combined data mining tools to determine the main factors contributing to accidents with serious injuries and their interdependencies, laying the groundwork for providing information and developing safety improvement policies and strategies to reduce these accidents.

2. Literature Review

Negative binomial models have been used to explore the impact of various accident factors on their frequency [21–23]. The use of Poisson regression models is also widespread in road safety management and analysis, particularly for determining the importance of factors in traffic accidents [24,25]. In recent years, other types of models have emerged: artificial neural networks, Bayesian networks, decision trees, and genetic programming [26].

In recent years, the use of data mining tools has become particularly important. Data mining involves extracting implicit, previously unknown, and potentially useful information from data. The

 doi:10.20944/preprints202605.1045.v1

idea is to create computer programs that automatically scan databases for regularities or patterns [27]. Data mining allows for filtering out noise in the data, predicting patterns, forecasting outcomes, and generating information to support informed and agile decision-making. It is closely related to statistics, artificial intelligence, and machine learning. In predictive modeling, the goal is to estimate the value of a target attribute based on training data, using tools such as Association Rules and classification and regression trees [28].

Association rules, also known as basket analysis, are one of the most popular approaches for discovering patterns in databases [29]. Algorithms for the discovery of association rules attempt to identify products that tend to be sold together [30]. The association rule method in data mining has been successfully used to discover patterns or hidden rules in a variety of fields, including shopping basket analysis, product recommendation, and medical record analysis [31], and allows for the identification of potential cause-and-effect relationships among the many factors that play a role, for example, in workplace accidents in the construction industry [32].

The goal of association rule mining is to identify any real associations in the data without specifically designating any variable as a dependent or independent variable. An association rule in the context of accident data indicates that the presence of a certain characteristic in an accident implies the presence of another characteristic in that accident. These rules can be searched for in the database using the a priori algorithm [30]. Improvements in the quantity and quality of data have sparked interest in new ways of analyzing and interpreting data. In particular, various authors have applied association rules to search for hidden patterns in accident databases in very general or very specific terms [33–35]

The Apriori algorithm has been used to extract strong association rules among the values of accident attributes in China, such as vehicle type, weather, and time of day, among others [36]. Combined with the WEKA platform to identify hidden factors in road accident records, with cross-validation of the method across different regions and cities [37],or combined with complex graph structures to reflect interactions between variables such as improper operations, overloading, mountainous terrain, and running off the road [38], or combined with Complex Network Analysis to examine causal mechanisms of serious accidents by identifying human factors (speeding, fatigue, insufficient distance), vehicle factors (trucks), and environmental factors (national highways, nighttime hours) as the main causes [39]. Association rules can also be used with more complex tools such as Random Forest + SHAP, which combine a powerful machine learning model with a robust interpretability framework. The combination of RF-SHAP (to determine the individual importance of factors) with the Apriori algorithm (to explore interactions among multiple factors simultaneously) allows us to overcome the limitations of each method separately: RF-SHAP identifies which factors matter individually, while Apriori reveals how they interact with one another to cause serious accidents [40].

Classification trees, meanwhile, are a specific case of partitioning-based testing strategies [41]. Classification is a unique example of predictive modeling in which a dataset is already segmented into pre-specified groups, and patterns in the data are identified to distinguish those groups. The patterns explored can then be used to categorize another dataset where the appropriate group description for the target attribute is unknown. Regression analysis is also an example of predictive modeling with a numerical target attribute, and the goal is to predict that value for new data [28]. The method involves identifying factors relevant to the test, and based on each of them, an exclusive classification is made, which in turn can be further reclassified into subcategories, represented graphically in the form of a tree. Test cases will be generated by combining the different elements from the various classifications performed. One of the advantages of this method is that it allows all the information to be managed in a structured manner in small groups or parts, making it easier to understand and document [41].

Studies have been found that apply CART and Random Forest (RF) to heavy vehicle accident data in Malaysia [42]and others that combine decision trees with statistical analysis in SPSS to analyze accidents in tunnels on mountainous highways in China [43].

 doi:10.20944/preprints202605.1045.v1

Association Rules have been applied after segmenting data into homogeneous clusters, integrating results with GIS to identify black spots, determining that serious accidents occur between trucks and motorcycles in low-density areas, with the main causes being excessive speed and improper lane changes [44]. Decision trees (DT) have also been used with Python to identify severity factors related to driver behavior and socioeconomic characteristics [45].

Moving toward more complex uses of these tools, there are studies that integrate various methods to overcome the individual limitations of each. Among the use of combined methods is the application of three classification algorithms—Decision Tree, LightGBM, and XGBoost—to UK accident data (2020), with hyperparameter tuning. It identifies that most accidents occur in daylight conditions on dry surfaces, and that speed limits of 30 mph (urban roads) see a higher concentration of accidents, although highways are more lethal [46]. There is a comparative study between the use of Logistic Regression (LR), CART, and Random Forest (RF) to predict severity on roads in Taiwan [47]. There is also the use of Ordered Probit, Association Rules, and CART to identify severity factors for vulnerable users at road-rail crossings, finding that speed is the most important factor [48].

Unlike traditional literature, which has relied primarily on Poisson and negative binomial regression models to study traffic accidents, this study aims to enhance the analysis by using data mining tools, specifically Association Rules and Classification and Regression Trees (CART). These tools were chosen because the factors causing traffic accidents are complex and often interrelated, requiring tools capable of extracting relevant information for sound decision-making. Association Rules facilitate the discovery of cause-and-effect relationships among variables or factors that interact simultaneously, such as driver behavior, the environment, and the vehicle. The integration of the CART model allows this information to be handled in a structured manner by partitioning the dataset into specific groups. This dual approach overcomes the individual limitations of each method, making this methodological combination a significant advance over studies that use isolated tools or basic descriptive statistics. Despite the rise of highly complex supervised learning algorithms, such as Random Forest, LightGBM, and XGBoost, there remains a critical need for models that not only predict severity but also reveal the architecture of interactions between factors, thereby offering more robust and accurate insights for road safety management and the reduction of road accidents.

3. Materials and Methods

The Data mining techniques widely used in the study of complex phenomena with multiple interrelated variables were employed in this study. First, association rules were applied to identify frequent patterns and significant relationships among categorical variables within large datasets; second, regression and classification trees (CART) were used to facilitate both severity prediction and the identification of factors with the greatest discriminatory power. These techniques and the methodology followed in this process are described below.

3.1. Association Rules

Rules take the form $A \rightarrow B$, where A is the antecedent and B is the consequent. In association rules, rules can be expressed in terms of support, confidence, and lift. Support is the percentage of a rule that exists in the entire dataset. Confidence is the proportion of consequents among the antecedents. Lift is a mathematical measure to quantify the statistical dependence of a rule. The three indices can be calculated as follows:

$$\text{Support}(A \rightarrow B) = \frac{\#(A \cap B)}{N} \quad (1)$$

$$\text{Confidence} = \frac{(\text{Support}(A \rightarrow B)}{\text{Support}\ (A)} \quad (2)$$

 doi:10.20944/preprints202605.1045.v1

$$\text{Lift} = \frac{\text{Support}(A \rightarrow B)}{\text{Support}\ (A) \times\ \text{Support}\ (B)} \tag{3}$$

where N is the number of observations and #(A∩B) is the number of observations in which conditions A (antecedent) and B (consequent) are met. The lift of the rule indicates the ratio of the actual co-occurrences of the antecedent and consequent to the expected co-occurrences under the assumption that the antecedent and consequent are independent. A value less than 1 indicates a negative dependence between the antecedent and the consequent. A value equal to 1 indicates independence, and a value greater than 1 indicates a positive dependence. The higher the lift, the stronger the association rule [35]. It is desirable for rules to have a high level of support, high confidence, and a lift value considerably greater than one. To identify strong associations, threshold values for support (S), confidence (C), and lift (L) were set as follows: S ≥ 4%, C ≥ 20%, and L ≥ 2. For rules with lift values greater than 10, the support threshold was set at 1% [31,33]

For example, in the rule "reckless driving → alcohol (support = 1%, confidence = 50%, lift = 5)", support indicates that the proportion of observations that include both reckless driving errors and alcohol is 1% across the entire dataset. ; confidence indicates that the proportion of observations that include both reckless driving errors and alcohol is 50% in the dataset that includes alcohol; and lift indicates that reckless driving errors are positively associated with alcohol [35].

3.2. Classification and Regression Tree

A decision tree model consists of root, internal, and leaf nodes. The metrics used in a decision tree model for branch selection are information gain, entropy, and Gini impurity. Figure 1 shows a schematic diagram of a decision tree model. It uses a top-down recursive method at each node in the sample set to select the branch attribute according to the given criteria, from a root node to a leaf node. A leaf node represents the value of an objective function, determined by the input variables along the path from the root node to the leaf node [49]

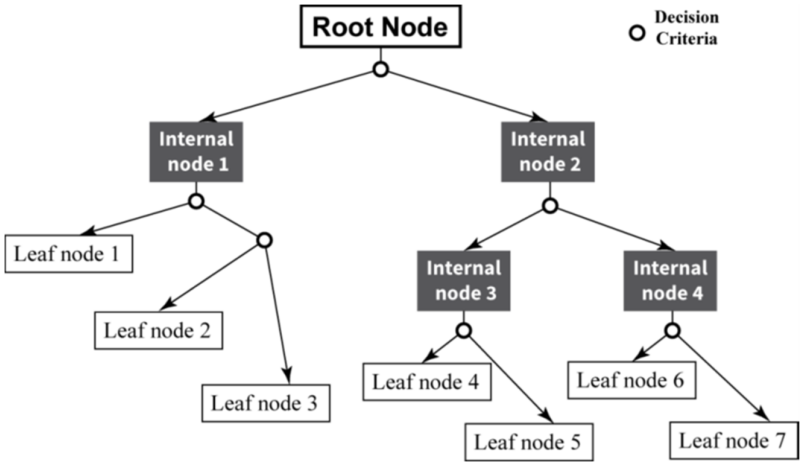

Figure 1. Schematic diagram of a decision tree model.

The CART algorithm is a decision tree-specific algorithm widely used today. For classification problems, Gini impurity is used as the criterion for branch selection, whereas for regression problems, the mean squared error (MSE) is used. Gini impurity represents the probability that a randomly selected sample will be misclassified in the sample set. Gini impurity can be calculated as follows:

$$\text{Gini}(A) = 1 - \sum_{k=1}^{n} P_k^{\,2} \tag{4}$$

 doi:10.20944/preprints202605.1045.v1

where A represents node A; n is the number of elements in a dataset, i = 1, 2, 3, ..., n; and Pk is the probability that a set with elements ordered in D belongs to class Ci. For regression problems, the objective of CART is to minimize the MSE. The MSE represents the difference between the predicted value and the actual value at a leaf node, and the calculation formula is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (5)$$

where n is the number of elements in a dataset, i = 1, 2, 3, ..., n; y_i is the actual value; and $\hat{y}_i$ is the predicted value [49,50].

The application of the non-parametric Classification and Regression Tree (CART) model does not require prior probabilistic knowledge of the phenomenon under study or the fulfillment of strict assumptions, either regarding the type of relationship or the distribution of the dependent variable. These aspects represent the main advantages over parametric techniques. Each node in the tree indicates the predicted value, the number of experimental units contained in the node, and its descriptive percentage. CART offers both theoretical and practical advantages over parametric models. In fact, from a theoretical perspective, the advantage of the CART method is that it does not require prior specification of the model's functional form or the assumption of an additive relationship between the dependent and independent variables. Another advantage is that CART analysis can effectively handle collinearity issues [51].

3.3. Methodology

A procedure for analyzing accident rates is proposed that consists of identifying critical factors in traffic accidents and their interrelationships by combining the use of Association Rules and CART, which can be fully generalized to any road safety study. This procedure consists of the phases shown in Figure 2.

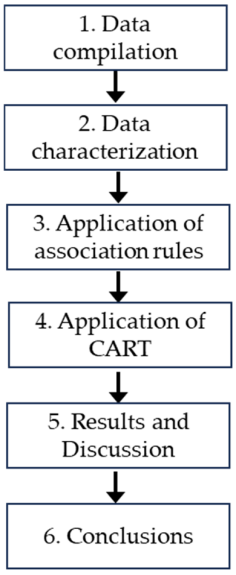

Figure 2. Stages of the methodology to be employed.

3.3.1. Data Collection

In this stage, accident records from the National Police are identified and obtained. The time period and geographic scope of the study are defined. The geographic scope is based on previous studies of accident rates on Peruvian roads.

 doi:10.20944/preprints202605.1045.v1

3.3.2. Data Characterization

The collected data is transformed into a structured set of variables. Each recorded attribute (type of accident, environmental conditions, human factor, etc.) is defined as a variable, specifying its nature (nominal, ordinal, numerical). The different categories for each variable are defined, and in the case of continuous numerical variables, they are discretized into intervals for analysis. A special coding system is used to identify the variables and their categories. The result is a data matrix characterized by clearly defined variables and categories.

3.3.3. Application of Association Rules

In this stage, association rules were extracted from the dataset using the R programming language and the specialized arules library (which contains the Apriori algorithm). Minimum support and confidence thresholds are defined, generating a set of frequent itemsets and, from these, association rules of the form {antecedent} → {consequent}. Each rule is evaluated using the metrics of support, confidence, and lift. The most relevant rules are filtered and sorted to identify co-occurrence patterns among the conditions present in the accidents.

3.3.4. Application of CART

In this stage, the R language and the rpart package are used to implement the Classification and Regression Trees (CART) algorithm, using accident severity as the target variable. This algorithm allowed the data space to be segmented using recursive binary partitions, optimizing the homogeneity of the terminal nodes by reducing variance to identify the variables with the greatest discriminatory power.

3.3.5. Discussion of Results

The findings from the techniques applied in steps 3 and 4 are discussed individually and then compared and integrated. Association rules provide patterns of co-occurrence among factors, while the CART tree establishes hierarchies of predictive importance. The agreement and complementarity between the results of both approaches are analyzed. The identified patterns are interpreted in light of the theoretical framework of the domain and compared with the existing literature.

3.3.6. Conclusions

In this stage, practical implications (preventive measures, intervention policies) are derived, and the study's limitations are noted, suggesting lines of future research.

4. Results

4.1. Data Collection

The road accident data was obtained from the Superintendency of Land Transport of Passengers, Cargo, and Goods (SUTRAN), a public institution whose functions include the regulation and oversight of all types of land transport activities, where data recorded by the Peruvian National Police is processed. The data for each accident includes the time, date, kilometer marker, vehicles involved, and driver information, among other details. Accidents recorded between 2014 and 2020 on the Pan-American Highway North—the road with the highest accident rate in Peru in recent years—were considered [52]. In addition to the data obtained, based on the kilometer where the accident occurred, three additional important factors were identified: road geometry, proximity to a town, and the intersection of a main and an auxiliary road.

4.2. Data Characterization

This dataset was structured so that each record represents a traffic accident, characterized by different variables such as the kilometer where it occurred, the date, the day, the time, the number of

 doi:10.20944/preprints202605.1045.v1

vehicles involved, the type of vehicle, among others. Based on this characterization, three main categories of variables or factors were defined: road factors (location, time), vehicle factors (type and number), and other factors (type of collision and severity). To facilitate rule analysis, continuous variables were transformed into discrete categories (bins). For example, distance was divided into 10-km ranges (DIS1, DIS2, etc.), and time into hourly blocks (TIM1 to TIM5). Finally, abbreviated labels were assigned to optimize computational processing (e.g., VEC1 for BUS, TYP6 for collision with an object). It is important to note that to characterize the severity of accidents, the IPA term was used, an adaptation of the IPA proposed by Peru's MTC [53]. Thus, IPA = 4NM + NL, where NM = number of fatalities and NL = number of injuries, creating 5 IPA categories, from least to most severe. The results can be seen in Table 1.

4.3. Application of Association Rules

To perform an Association Rule analysis, as previously indicated, the Apriori algorithm was used under the R language interface, utilizing the arules package. This technique is applied to discover frequent relationships between accident factors and their outcomes.

Table 1. Characterization of the variables to be considered in the study.

Factor Category		Abbreviation			Definitions
Road factors	City proximity	City_prox	N/F		Near/Far
Road factors	Place in which crash ocurred	Kilometer	DIS1/DIS2/DIS3/DIS4/DIS5/DIS6/DIS7/DIS8/DIS9/DIS10	1/2/3/4/5/6/7/8/9/10	KM 0-10/10-20/20-40/40-60/60-80/80-100/ 100-120/120-140/140-160/160-180/180-200
Time factors	Month of year	Month	MON1/MON2/MON3/MON4/MON5/MON6/MON7/MON8/MON9	1/2/3/4/5/6/7/8/9	January/February/March/April,May,June/July/August/September, October/November/December
Time factors	Day of week	Day	DAY1/DAY2/DAY3/DAY4/DAY5/DAY6/DAY7	1/2/3/4/5/6/7	Sunday/Monday/Tuesday/Wednesday/Thursday/Friday/Saturday
Time factors	Time of day	Time	TIM1/TIM2/TIM3/TIM4/TIM5	1/2/3/4/5	07:00-11:00/11:00-14:00/14:00-18:00/18:00-24:00/00:00-07:00
Vehicle factors	Vehicle type involved	Vec_type	VEC1/VEC2/VEC3/VEC4/VEC5	1/2/3/4/5	BUS/Truck/Private car/BUS-TRUCK/Others
Vehicle factors	Number of vehicles involved	Vec_num	NUM1/NUM2/NUM3/NUM4/NUM5	1/2/3/4/5	1/2/3/4/Multiple
Other factors	Type of crash	Crash_type	TYP1/TYP2/TYP3/TYP4/TYP5/TYP6/TYP7	1/2/3/4/5/6/7	Run off the road/Head-on collision/Rear-end collision/Sideswipe collision/Turnover/Hit object or pedestrian/ Others
Other factors	Crash severity	IPA	IPA1/IPA2/IPA3/IPA4	1/2/3/4	IPA 1-5/IPA5-10/IPA10-20/IPA>20

Combinations of "antecedents" (causes/conditions) leading to a "consequent" (severity outcome or type of collision) were sought. The data were filtered using the Support, Lift, and Confidence metrics. The rules with the highest confidence were considered for the analysis, since high confidence guarantees that the rule has a high success rate and a reduction of false positives. Table 2 shows the rules with the highest confidence.

The rules with the highest confidence (1.00 or 100%) indicate an absolute relationship between the type of collision with objects or pedestrians and an IPA severity between 1 and 5. Rule 87 indicates that, if this type of accident occurs, it always results in an IPA1 between 1 and 5. Also, based on Rules

 doi:10.20944/preprints202605.1045.v1

491 and 890, when this type of collision occurs, it almost always (Confidence > 0.90) involves a single vehicle. Based on Rules 820 and 921, there is a strong association between the involvement of a truck in a two-vehicle collision and the generation of a collision classified as "other." This frequently occurs in the 80–100 km range (DIS5) (Rule 763), suggesting a stretch of road where trucks have conflicts with other vehicles. Rule 820 adds the TimeBin_TIM4 factor (6:00 PM to midnight). This indicates a specific pattern of two-vehicle collisions during that time block. From Rule 452, it follows that if there is a single vehicle and it is a truck, there is a 91% probability of resulting in an IPA severity of 1 to 5. Analyzing the Lift, Rule 890 (Lift: 3.02), with the highest Lift, indicates that if a bus is involved in a collision with objects or pedestrians, the probability of it being an IPA severity accident between 1 and 5 is three times higher than normal. Rule 1022 (Lift: 2.91) associates run-off-road incidents with trucks and locations far from cities to predict severity and single-vehicle involvement. Trucks appear in multiple high-confidence rules during the early morning hours (TIM5: 00:00–07:00), for example, in rules 855 and 701. The combination of fatigue and driving heavy vehicles during these hours almost always results in accidents of a certain severity. Finally, truck accidents far from the city, particularly in the months of April/May/June, have a 93.7% confidence level of being severe. In general terms, the rules with the highest confidence and "Lift" involve buses and trucks.

Table 2. Results of Applying the Association Rule.

Top Association Rules for CrashType/Severity

	Antecedents	Consequents	Support	Confidence	Lift
87	(CrashType_TYP6)	(SeverityBin_IPA1)	0.130346	1.000000	1.316354
218	(Proximity_F, CrashType_TYP6)	(SeverityBin_IPA1)	0.063136	1.000000	1.316354
287	(Proximity_N, CrashType_TYP6)	(SeverityBin_IPA1	0.067210	1.000000	1.316354
439	(VehType_VEC1, CrashType_TYP6)	(SeverityBin_IPA1)	0.089613	1.000000	1.316354
488	(NumVeh_NUM1, CrashType_TYP6)	(SeverityBin_IPA1)	0.118126	1.000000	1.316354
684	(NumVeh_NUM1, Proximity_F, CrashType_TYP6)	(SeverityBin_IPA1)	0.057026	1.000000	1.316354
742	(Proximity_N, NumVeh_NUM1, CrashType_TYP6)	(SeverityBin_IPA1)	0.061100	1.000000	1.316354
886	(VehType_VEC1, NumVeh_NUM1, CrashType_TYP6)	(SeverityBin_IPA1)	0.085540	1.000000	1.316354
724	(Proximity_N, NumVeh_NUM1, VehType_VEC2)	(SeverityBin_IPA1	0.059063	0.966667	1.272475
349	(Weekday_DAY4, NumVeh_NUM1)	(SeverityBin_IPA1)	0.057026	0.965517	1.270962
820	(VehType_VEC2, TimeBin_TIM4, NumVeh_NUM2)	(CrashType_TYP7)	0.050916	0.961538	2.000489
890	(VehType_VEC1, CrashType_TYP6)	(SeverityBin_IPA1, NumVeh_NUM1)	0.085540	0.954545	3.023754
371	(TimeBin_TIM3, CrashType_TYP7)	(SeverityBin_IPA1)	0.067210	0.942857	1.241134
855	(NumVeh_NUM1, TimeBin_TIM5, VehType_VEC2)	(SeverityBin_IPA1	0.063136	0.939394	1.236575
510	(Proximity_F, VehType_VEC2, MonthBin_MON4)	(SeverityBin_IPA1)	0.061100	0.937500	1.234082
319	(VehType_VEC2, MonthBin_MON4)	(SeverityBin_IPA1)	0.083503	0.931818	1.226602
279	(Proximity_N, NumVeh_NUM1)	(SeverityBin_IPA1)	0.138493	0.931507	1.226193
549	(TimeBin_TIM4, Proximity_F, VehType_VEC2)	(SeverityBin_IPA1)	0.052953	0.928571	1.222329
580	(TimeBin_TIM5, Proximity_F, VehType_VEC2)	(SeverityBin_IPA1)	0.052953	0.928571	1.222329
701	(Proximity_N, NumVeh_NUM1, TimeBin_TIM5)	(SeverityBin_IPA1)	0.050916	0.925926	1.218846

 doi:10.20944/preprints202605.1045.v1

776	(NumVeh_NUM1, VehType_VEC2, MonthBin_MON4)	(SeverityBin_IPA1)	0.050916	0.925926	1.218846
912	(CrashType_TYP1, VehType_VEC2)	(SeverityBin_IPA1, NumVeh_NUM1)	0.097760	0.923077	2.924069
461	(CrashType_TYP1, VehType_VEC2)	(SeverityBin_IPA1)	0.097760	0.923077	1.215096
908	(CrashType_TYP1, NumVeh_NUM1, VehType_VEC2)	(SeverityBin_IPA1)	0.097760	0.923077	1.215096
1022	(CrashType_TYP1, Proximity_F, VehType_VEC2)	(SeverityBin_IPA1, NumVeh_NUM1)	0.069246	0.918919	2.910898
631	(CrashType_TYP1, Proximity_F, VehType_VEC2)	(SeverityBin_IPA1)	0.069246	0.918919	1.209622
1017	(CrashType_TYP1, NumVeh_NUM1, Proximity_F, VehType_VEC2)	(SeverityBin_IPA1)	0.069246	0.918919	1.209622
402	(TimeBin_TIM5, VehType_VEC2)	(SeverityBin_IPA1)	0.091650	0.918367	1.208896
1047	(VehType_VEC2, SeverityBin_IPA1, Proximity_F, NumVeh_NUM2)	(CrashType_TYP7)	0.067210	0.916667	1.907133
1044	(VehType_VEC2, Proximity_F, CrashType_TYP7, NumVeh_NUM2)	(SeverityBin_IPA1)	0.067210	0.916667	1.206658
141	(MonthBin_MON5, Proximity_F)	(SeverityBin_IPA1)	0.065173	0.914286	1.203524
452	(NumVeh_NUM1, VehType_VEC2)	(SeverityBin_IPA1)	0.171079	0.913043	1.201888
746	(Proximity_N, CrashType_TYP6)	(SeverityBin_IPA1, NumVeh_NUM1)	0.061100	0.909091	2.879765
491	(CrashType_TYP6)	(SeverityBin_IPA1, NumVeh_NUM1)	0.118126	0.906250	2.870766
949	(SeverityBin_IPA1, Proximity_F, NumVeh_NUM2, MonthBin_MON4)	(CrashType_TYP7)	0.059063	0.906250	1.885461

4.4. Application of Classification and Regression Trees (CART)

In the case of the Classification and Regression Tree (CART), an algorithm was used within the R language interface to hierarchically identify which variables best discriminate severity (IPA). The five IPA categories were reduced to three to facilitate the analysis. The model selects the root node (the most important variable) and then continues its division to create specific risk profiles until reaching the terminal nodes. To apply CART, the same variables considered in the association rules were used, but the Accident Severity variable was modified as follows:

IPA 0-1 ➔ IPA 1
IPA 2-4 ➔ IPA 2
IPA 5-88 ➔ IPA3

This resulted in the tree shown in Figure 3.

The tree identifies that the most important variable for separating the data is the type of collision. If the accident involves a collision with an object or a pedestrian, the model automatically classifies it as having a higher probability of IPA2 (low-to-moderate severity, range 5–10). This branch represents 14% of the total sample. The other types of collisions account for 86% of the remaining cases and require more variables to determine severity.

The model also identifies a specific scenario in which severity is most likely to be IPA1 (the lowest level on the scale). The path would be: if the accident is not a collision with an object or pedestrian➔ Traffic Level = 1 (low)➔ Collision type is run-off-road, rollover, or other. If the above conditions are met and the accident occurs on the sections corresponding to kilometers 0–10, 180–200, 20–40, 40–60, and 80–100, the model predicts IPA1. This node represents 32% of the data, making it the largest and most distinct group for IPA1 severity.

 doi:10.20944/preprints202605.1045.v1

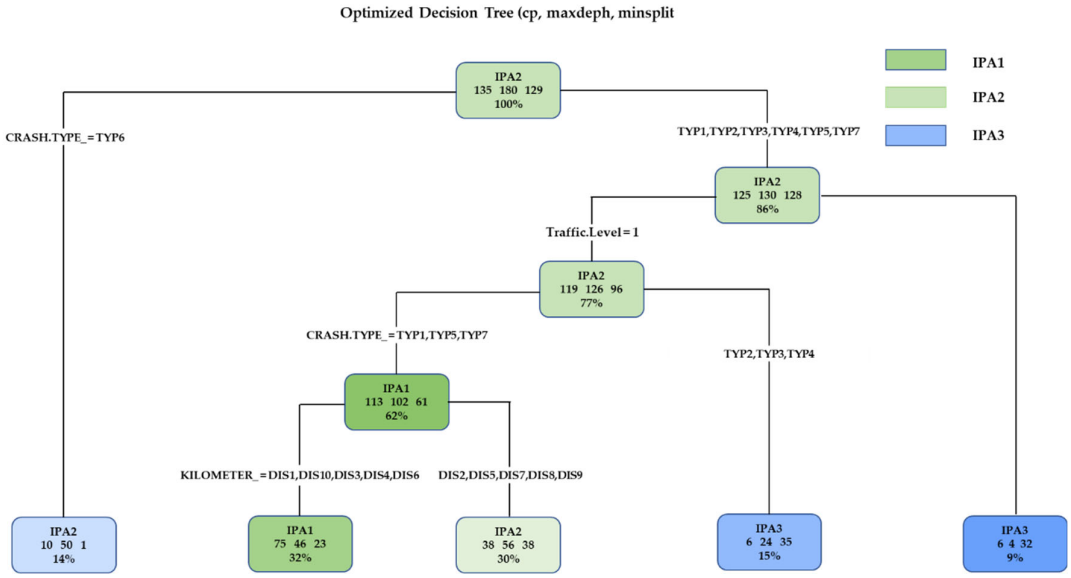

Figure 3. Application of regression and classification trees (CART).

The tree identifies two clear paths leading to the highest-risk category visible in the graph (IPA3). The first indicates that if the crash type is not TYP6 and the Traffic Level is 2 or 3, the prediction falls directly into IPA3 (9% of cases), suggesting that as congestion increases, severity tends to increase; the second indicates that if the traffic level is low (Level 1), but the crash type is TYP2 (Head-on Collision), TYP3 (Rear-End Collision), or TYP4 (Side Collision), the model predicts IPA3 (15% of cases).

The tree identifies that the most important variable for separating the data is the type of collision. If the accident involves a collision with an object or a pedestrian, the model automatically classifies it as having a higher probability of IPA2 (low-to-moderate severity, range 5–10). This branch represents 14% of the total sample. The other types of collisions account for 86% of the remaining cases and require more variables to determine severity.

The model also identifies a specific scenario in which severity is most likely to be IPA1 (the lowest level on the scale). The path would be: if the accident is not a collision with an object or pedestrian➔ Traffic Level = 1 (low)➔ Collision type is run-off-road, rollover, or other. If the above conditions are met and the accident occurs on the sections corresponding to kilometers 0–10, 180–200, 20–40, 40–60, and 80–100, the model predicts IPA1. This node represents 32% of the data, making it the largest and most distinct group for IPA1 severity.

The tree identifies two clear paths leading to the highest-risk category visible in the graph (IPA3). The first indicates that if the crash type is not TYP6 and the Traffic Level is 2 or 3, the prediction falls directly into IPA3 (9% of cases), suggesting that as congestion increases, severity tends to increase; the second indicates that if the traffic level is low (Level 1), but the crash type is TYP2 (Head-on Collision), TYP3 (Rear-End Collision), or TYP4 (Side Collision), the model predicts IPA3 (15% of cases).

Upon performing the combined analysis, it was found that while the association rules showed that TYP6 is strongly linked to IPA1 with high confidence, the CART tree refined this view by showing that, depending on the traffic context and location, this or other types of collisions can escalate to IPA2 or IPA3 severity levels. Furthermore, it was identified that heavy vehicles (VEC1-BUS and VEC2-Truck) and frontal/side collisions in dense traffic are the scenarios with the highest severity.

 doi:10.20944/preprints202605.1045.v1

5. Discussion

Joint analysis using association rules and decision trees (CART) reveals critical patterns linking the nature of the accident, the type of vehicle, and the operating environment to the severity of the event. Furthermore, takes advantage of the benefit of the CART method, which lies in the fact that it does not require prior specification of the model's functional form or the assumption of an additive relationship between the dependent and independent variables, as proposed by Pagliara [51]. The results demonstrate that there is a direct relationship between the type of crash (Crash_type) and the severity of injuries, an approach similar to that of Beshah and Hill [53]. They also show an absolute correlation between crash types involving buses (VEC1) and trucks (VEC2) with consistent severity levels. This aligns with the findings of Samerei [54], who indicate that the presence of heavy vehicles and multi-vehicle collisions are additional factors contributing to increased fatalities in accidents involving buses, and with Kashani [55], who noted that heavy vehicle transport carries a high probability of accidents. The association rules indicate that collision type TYP6 (Impact with object or pedestrian) involving a bus (VEC1) invariably results in severity IPA1 (confidence 1.00), a finding further supported by Samerei [54], who argues that direct collisions between buses and pedestrians on roads greatly increase the probability of fatalities.

This suggests that the inertial mass of heavy vehicles eliminates variability in the impact outcome, turning any collision with fixed objects into an event of guaranteed severity. The CART decision tree shows that severity depends not only on the impact but also on traffic flow conditions. It is observed that under Traffic Level 2 and 3 conditions, the probability of reaching a higher severity (IPA3) increases to 9%, regardless of whether the collision is a side-impact or a rear-end collision. Conversely, on specific road sections such as KM 80–100 (DIS5), trucks tend to be involved in TYP7-type crashes (Other/Rollovers), which is associated with driver fatigue due to these areas being far from the city (Proximity_F). It is also worth noting that road sections influence the frequency and severity of accidents, which can be linked to the road's geometry—a finding also documented in the specialized literature [56,57]. It is important to highlight the danger posed by nighttime and early morning time blocks. The combination of TIM5 (00:00–07:00) with vehicle type VEC2 (Truck) has a confidence level exceeding 92% for causing IPA1 severity. These findings suggest that reduced visibility and fatigue in areas far from cities (Proximity_F) increase the severity of run-off-road accidents (TYP1), confirming that the early morning hours are the most dangerous due to reduced driver alertness [58].

Another interesting finding is that the CART model reveals that as congestion increases, accident severity tends to rise. This is consistent with the findings of Kashani [57]. The incidence of accidents and the severity of injuries depend on traffic volume on this highway. It is significant to note that while the CART model identifies the Type of Collision as the primary data separator, the association rules add the Month of the Year (MON4, MON5) layer as an aggravating factor in remote areas. This indicates that prevention campaigns should be seasonal and geographically targeted at the identified kilometer markers (e.g., DIS5 and DIS10).

The academic literature supports the combined use of data mining tools; while the CART algorithm defines the hierarchical structure of accident severity, the Association Rules allow for the capture of complex and specific scenarios (such as the interaction between the TIM5 block, the Kilometer, and the Vehicle Type) that traditional linear models often omit, hence the advantages of using them in combination.

6. Conclusions

The joint use of CART and Association Rules demonstrated that the combined use of data mining tools yields more comprehensive results than conventional linear models for this type of phenomenon, as it enables the capture of both the hierarchy of factors causing accidents and the interrelationships among them. Future studies on road safety should adopt this combined approach as the standard for analysis.

 doi:10.20944/preprints202605.1045.v1

The identified patterns reveal that a uniform road safety policy is insufficient: distant road segments (Proximity_F), night/early morning time blocks (TIM5), and the operation of heavy vehicles constitute specific risk scenarios. This calls for specific control measures, focused on different combinations of factors such as fatigue controls in remote areas, nighttime speed limits for trucks, and infrastructure improvements at identified critical kilometers (DIS5, DIS10).

The emergence of months MON4 and MON5 as aggravating factors in remote areas, combined with the influence of the road segment on the frequency and severity of accidents, indicates that the risk of accidents on the road is not uniform and depends on the time of year and geographic location. Thus, control measures may be seasonal in nature and take geographical location into account, for example, by evaluating road geometry and signage on segments with the highest accident rates and by increasing police presence during certain times of the year.

The fact that heavy vehicles (buses and trucks) are significant factors in accidents and generate consistent and predictable levels of severity, reaching a 100% confidence level in scenarios involving collisions with pedestrians or objects, suggests that stricter temporal and spatial restrictions for heavy transport should be evaluated on the identified critical road segments, as well as strengthening the technical and human monitoring systems associated with these two types of vehicles.

Author Contributions: Conceptualization, C.C. and J.R.; methodology, C.C. and C.P.; software, C.C.; validation, C.P. and J.R.; formal analysis, C.C.; investigation, C.C.; resources, CC.; writing—original draft preparation, C.C.; writing—review and editing, C.C. C.P. and J.R.; visualization, C.C.; supervision, J.R.

Funding: This research received no external funding.

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: This study did not involve humans.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to legal reasons.

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

References

1. World Health Organization. *Global Status Report on Road Safety 2023*; Geneva, 2023. https://repository.gheli.harvard.edu/repository/12838/ (accessed 2026-04-06).
2. Ahmed, S. K.; Mohammed, M. G.; Abdulqadir, S. O.; El-Kader, R. G. A.; El-Shall, N. A.; Chandran, D.; Rehman, M. E. U.; Dhama, K. Road Traffic Accidental Injuries and Deaths: A Neglected Global Health Issue. *Health Sci. Rep.* **2023**, *6* (5). [CrossRef]
3. Blincoe, L.; Miller, T. R.; Wang, J.-S.; Swedler, D.; Coughlin, T.; Lawrence, B.; Guo, F.; Klauer, S.; Dingus, T. *February 2023 6. Performing Organization Code 7. Authors 13. Type of Report and Period Covered NHTSA Technical Report 14. Sponsoring Agency Code Unclassified*; 2023. https://rosap.ntl.bts.gov.
4. International Road Assessment Programme (iRAP). *2024 Annual Report: Celebrating Partners' Global Impact*; Bracknell, Reino Unido, 2025. https://irap.org/2025/05/2024-annual-report/ (accessed 2026-04-06).
5. Barua, U.; Tay, R. Severity of Urban Transit Bus Crashes in Bangladesh. *J. Adv. Transp.* **2010**, *44* (June 2010), 36–41. [CrossRef]
6. OMS. La Seguridad Vial 2013. *Informe Sobre La Situación Mundial De La Seguridad Vial 2013* **2013**, 12.
7. Observatorio Nacional de Seguridad Vial (ONSV). *Estadísticas de Siniestros de Tránsito 2023*; Lima, 2025. http://www.onsv.gob.pe/ (accessed 2026-04-06).
8. Ministerio de Transportes y Comunicaciones. *MTC Impulsa Proyecto de Recolección de Información Sobre Accidentes En Vías Concesionadas Para Reforzar La Seguridad*; Lima, Perú, 2025. https://www.gob.pe/institucion/mtc/noticias/1092860-mtc-impulsa-proyecto-de-recoleccion-de-informacion-sobre-accidentes-en-vias-concesionadas-para-reforzar-la-seguridad (accessed 2026-04-06).
9. International Transport Forum (ITF). *Road Safety Annual Report 2024*; Paris, 2024. https://www.itf-oecd.org/road-safety-annual-report-2024 (accessed 2026-04-06).

 doi:10.20944/preprints202605.1045.v1

10. Ahmed, M.; Patnaik, J. L.; Whitestone, N.; Hossain, M. A.; Alauddin, M.; Husain, L.; Hossain, M. P.; Islam, M. S.; Hossain, M. I.; Imdad, K.; Cherwek, D. H.; Congdon, N. Visual Impairment and Risk of Self-Reported Road Traffic Crashes Among Bus Drivers in Bangladesh. *Asia. Pac. J. Ophthalmol. (Phila).* **2022**, *11* (1), 72–78. [CrossRef]
11. Nguyen, T. C.; Nguyen, M. H.; Armoogum, J.; Ha, T. T. Bus Crash Severity in Hanoi, Vietnam. *Safety* **2021**, *7* (3), 1–14. https://doi.org/10.3390/safety7030065. [CrossRef]
12. Verma, A.; Sasidharan, S.; Bhalla, K.; Allirani, H. Fatality Risk Analysis of Vulnerable Road Users from an Indian City. *Case Stud. Transp. Policy* **2022**, *10* (1), 269–277. [CrossRef]
13. Miller, K. A.; Filtness, A. J.; Anund, A.; Maynard, S. E.; Pilkington-Cheney, F. Contributory Factors to Sleepiness amongst London Bus Drivers. *Transp. Res. Part F Traffic Psychol. Behav.* **2020**, *73*, 415–424. [CrossRef]
14. Law, T. H.; Daud, M. S.; Hamid, H.; Haron, N. A. Development of Safety Performance Index for Intercity Buses: An Exploratory Factor Analysis Approach. *Transp. Policy (Oxf).* **2017**, *58* (August 2016), 46–52. https://doi.org/10.1016/j.tranpol.2017.05.003. [CrossRef]
15. Zegeer, C. V.; Huang, H. F.; Stutts, J. C.; Rodgman, E.; Hummer, J. E. Commercial Bus Accident Characteristics and Roadway Treatments. *Transp. Res. Rec.* **1994**, No. 1467, 14–22.
16. Kaplan, S.; Prato, C. G. Risk Factors Associated with Bus Accident Severity in the United States: A Generalized Ordered Logit Model. *J. Safety Res.* **2012**, *43* (3), 171–180. [CrossRef]
17. Polo, J. R.; Riveros, C. C.; Diaz, W. A.; Cansaya, A. C.; Anticona, M. R. Caracterización Del Nivel de Estrés de Alumnos de Ingeniería Mediante Herramientas de Data Mining. *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology* **2021**, *2021-July* (January). https://doi.org/10.18687/LACCEI2021.1.1.489. [CrossRef]
18. Spanos, A. Revisiting Data Mining: 'Hunting' with or without a License. *Journal of Economic Methodology* **2000**, *7* (2), 231–264. [CrossRef]
19. Xianfang, T.; Yachao, J.; Ru, Z. The Infiltration of Mathematical Modeling Thoughts in College Mathematics Teaching. *J. Phys. Conf. Ser.* **2019**, *1168* (5), 1–5. https://doi.org/10.1088/1742-6596/1168/5/052018. [CrossRef]
20. Wu, Y. The Modes of Data Development in the Internet Age. *Data Sci. J.* **2007**, *6* (SUPPL.), 962–967. [CrossRef]
21. Chand, S.; Li, Z.; Alsultan, A.; Dixit, V. v. Comparing and Contrasting the Impacts of Macro-Level Factors on Crash Duration and Frequency. *Int. J. Environ. Res. Public Health* **2022**, *19* (9). [CrossRef]
22. Li, F.; Jiang, K. Application of Random-Parameter Negative Binomial Model to Examine the Relationship between the Severity of Traffic Accident. *2020 IEEE 5th International Conference on Intelligent Transportation Engineering, ICITE 2020* **2020**, 351–354. [CrossRef]
23. Mahmud, A.; Gayah, V. v. Estimation of Crash Type Frequencies on Individual Collector Roadway Segments. *Accid. Anal. Prev.* **2021**, *161* (August), 106345. [CrossRef]
24. Ghadban, N. R.; Abdella, G. M.; Alhajyaseen, W.; Al-Khalifa, K. N. Analyzing the Impact of Human Characteristics on the Comprehensibility of Road Traffic Signs. *Proceedings of the International Conference on Industrial Engineering and Operations Management, Bandung, Indonesia*. **2018**, 2210-2219.
25. Kraidi, R.; Evdorides, H. Pedestrian Safety Models for Urban Environments with High Roadside Activities. Saf. Sci. 2020, 130. [CrossRef]
26. Mujalli, R. O.; de Ona, J. Injury Severity Models for Motor Vehicle Accidents: A Review. *Proceedings of the Institution of Civil Engineers: Transport*. 2013. [CrossRef]
27. Witten, I.; Frank, E.; Hall, M. *Data Mining*, Third.; Morgan Kaufmann Publishers: Burlington, 2011.
28. Dandge, S. S.; Chakraborty, S. A Data Mining Approach for Analysis of a Wire Electrical Discharge Machining Process. *Management and Production Engineering Review* **2021**, *12* (3), 116–128. [CrossRef]
29. Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. **1993**, No. January 1993, 207–216. [CrossRef]
30. Pande, A.; Abdel-Aty, M. Discovering Indirect Associations in Crash Data through Probe Attributes. *Transp. Res. Rec.* **2008**, No. 2083, 170–179. [CrossRef]

 doi:10.20944/preprints202605.1045.v1

31. 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* (4), 1451–1463. [CrossRef]
32. Cheng, C. W.; Lin, C. C.; Leu, S. Sen. Use of Association Rules to Explore Cause-Effect Relationships in Occupational Accidents in the Taiwan Construction Industry. *Saf. Sci.* **2010**, *48* (4), 436–444. [CrossRef]
33. Montella, A.; Aria, M.; D'Ambrosio, A.; Mauriello, F. Analysis of Powered Two-Wheeler Crashes in Italy by Classification Trees and Rules Discovery. *Accid. Anal. Prev.* **2012**, *49*, 58–72. [CrossRef]
34. Daher, J. R.; Chilkaka, S.; Younes, A.; Shaban, K. Association Rule Mining on Five Years of Motor Vehicle Crashes. *MATEC Web of Conferences* **2016**, *81* (2016). [CrossRef]
35. Wang, K.; Qin, X. Exploring Driver Error at Intersections: Key Contributors and Solutions. *Transp. Res. Rec.* **2015**, *2514*, 1–9. [CrossRef]
36. Liu, S.; Kang, L.; Sun, H.; Wu, J.; Amihere, S. Exploring the Factors of Major Road Traffic Accidents: A Case Study of China. *Frontiers of Engineering Management* **2025**, *12* (2), 414–424. [CrossRef]
37. M. Tariq; N. Q. Mehmood; S. Z. Mahfooz. Discovering Associated Factors behind Road Accidents Using Association Rule Mining: A Case Study from Gujarat, Pakistan. *World Journal of Advanced Research and Reviews* **2022**, *15* (3), 001–011. [CrossRef]
38. Gu, C.; Xu, J.; Gao, C.; Mu, M.; Guangxun, E.; Ma, Y. Multivariate Analysis of Roadway Multi-Fatality Crashes Using Association Rules Mining and Rules Graph Structures: A Case Study in China. *PLoS One* **2022**, *17* (10 October). [CrossRef]
39. Huang, S.; Jin, C.; Chen, T.; Wang, Z. W.; Wang, J. Analysis of Major Road Traffic Accident Causes Using a Combined Method of Association Rule and Complex Network. *J. Adv. Transp.* **2025**, *2025* (1). [CrossRef]
40. Wang, J.; Ma, S.; Jiao, P.; Ji, L.; Sun, X.; Lu, H. Analyzing the Risk Factors of Traffic Accident Severity Using a Combination of Random Forest and Association Rules. *Applied Sciences (Switzerland)* **2023**, *13* (14). [CrossRef]
41. Grochtmann, M.; Grimm, K. Classification Trees for Partition Testing. *Softw. Test. Verif. Reliab.* **1993**, *3*, 63-82. [CrossRef]
42. Azhar, A.; Ariff, N. M.; Bakar, M. A. A.; Roslan, A. Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest. *Sustainability (Switzerland)* **2022**, *14* (7). [CrossRef]
43. Le, K. G.; Tran, Q. H.; Do, V. M. Urban Traffic Accident Features Investigation to Improve Urban Transportation Infrastructure Sustainability by Integrating GIS and Data Mining Techniques. *Sustainability (Switzerland)* **2024**, *16* (1). [CrossRef]
44. Abdullah, P.; Sipos, T. Drivers' Behavior and Traffic Accident Analysis Using Decision Tree Method. *Sustainability (Switzerland)* **2022**, *14* (18). [CrossRef]
45. Megnidio-Tchoukouegno, M.; Adedeji, J. A. Machine Learning for Road Traffic Accident Improvement and Environmental Resource Management in the Transportation Sector. *Sustainability (Switzerland)* **2023**, *15* (3). [CrossRef]
46. Wang, H.; Liang, G. Association Rules Between Urban Road Traffic Accidents and Violations Considering Temporal and Spatial Constraints: A Case Study of Beijing. *Sustainability (Switzerland)* **2025**, *17* (4). [CrossRef]
47. Chen, M. M.; Chen, M. C. Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest. *Information (Switzerland)* **2020**, *11* (5). [CrossRef]
48. Ghomi, H.; Bagheri, M.; Fu, L.; Miranda-Moreno, L. F. Analyzing Injury Severity Factors at Highway Railway Grade Crossing Accidents Involving Vulnerable Road Users: A Comparative Study. *Traffic Inj. Prev.* **2016**, *17* (8), 833–841. [CrossRef]
49. Yang, X.; Ji, Y.; Gu, J.; Niu, M. An Electricity Consumption Disaggregation Method for HVAC Terminal Units in Sub-Metered Buildings Based on CART Algorithm. *Buildings* **2023**, *13* (4). [CrossRef]
50. Kim, H.; Kim, W.; Kim, J.; Lee, S. J.; Yoon, D.; Jo, J. A Study on Re-engagement and Stabilization Time on Take-over Transition in a Highly Automated Driving System. *Electronics (Switzerland)* **2021**, *10* (3), 1–13. [CrossRef]

 doi:10.20944/preprints202605.1045.v1

51. Pagliara, F.; Mauriello, F.; Ping, Y. Analyzing the Impact of High-Speed Rail on Tourism with Parametric and Non-Parametric Methods: The Case Study of China. *Sustainability (Switzerland)* **2021**, *13* (6). [CrossRef]
52. SUTRAN. *Reporte Estadístico de Siniestros Viales 2022*; Lima, 2023. https://www.gob.pe/institucion/sutran/informes-publicaciones/4171345-reporte-estadistico-de-siniestros-viales-2022 (accessed 2026-04-22).
53. Beshah, T.; Hill, S. Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. *AAAI Spring Symposium: Artificial Intelligence for Development* **2010**
54. Samerei, S. A.; Aghabayk, K.; Mohammadi, A.; Shiwakoti, N. Data Mining Approach to Model Bus Crash Severity in Australia. *J. Safety Res.* **2021**, *76*, 73–82. [CrossRef]
55. Kashani, A. T.; Zandi, K.; Okabe, A. Investigation of Factors Associated with Heavy Vehicle Crashes in Iran (Tehran–Qazvin Freeway). *Sustainability (Switzerland)* **2023**, *15* (13). [CrossRef]
56. Costa, J. O. D.; Freitas, E. F.; Jacques, M. A. P.; Pereira, P. A. A. Collision Prediction Models with Longitudinal Data: An Analysis of Contributing Factors in Collision Frequency in Road Segments in Portugal. *17th International Conference Road Safety on Five Continents (RS5C 2016), Rio de Janeiro, Brazil* **2016,** 1-12.
57. Besharati, M. M.; Tavakoli Kashani, A. Factors Contributing to Intercity Commercial Bus Drivers' Crash Involvement Risk. *Arch. Environ. Occup. Health* **2018**, *73* (4), 243–250. [CrossRef]
58. Pakgohar, A.; Tabrizi, R. S.; Khalili, M.; Esmaeili, A. The Role of Human Factor in Incidence and Severity of Road Crashes Based on the CART and LR Regression: A Data Mining Approach. In *Procedia Computer Science*; 2011; Vol. 3, pp 764–769. [CrossRef]