Acta logistica - International Scientific Journal about Logistics
Volume: 10 2023 Issue: 3 Pages: 413-422 ISSN 1339-5629

Predicting crash injury severity in road freight flows with association rules algorithms
Luis David Berrones-Sanz, Estefania Perez-Diaz, Dulce Maria Monroy Becerril, Esteban Martinez Diaz

https://doi.org/10.22306/al.v10i3.410 *Received: 25 Mar. 2023; Revised: 13 July 2023; Accepted: 11 Aug. 2023*

Predicting crash injury severity in road freight flows with association rules algorithms

Luis David Berrones-Sanz
Instituto Politécnico Nacional, Av. Plan de Agua Prieta 66, Plutarco Elías Calles, Miguel Hidalgo, 11350 Ciudad de México, Mexico, lberroness@ipn.mx (corresponding author)
Estefania Perez-Diaz
Instituto Politécnico Nacional, Manuel Carpio 471, Plutarco Elías Calles, Miguel Hidalgo, 11350 Ciudad de México, Mexico, fanny.pd1201@gmail.com
Dulce Maria Monroy Becerril
Instituto Politécnico Nacional, Manuel Carpio 471, Plutarco Elías Calles, Miguel Hidalgo, 11350 Ciudad de México, Mexico, dmonroy@ipn.mx
Esteban Martinez Diaz
Instituto Politécnico Nacional, Manuel Carpio 471, Plutarco Elías Calles, Miguel Hidalgo, 11350 Ciudad de México, Mexico, emartinez@ipn.mx

Keywords: Association rule mining, Apriori algorithm, crash risk prediction, road freight transport.
Abstract: The purpose of this study is to evaluate the use of the Apriori association rule mining algorithm to classify and predict the severity of the 718,565 accidents involving freight transport vehicles in Mexico, which occurred between 2009 and 2018. The accidents were classified into those in which there was only material damage or injured people {Severity=0} and in those in which people died {Severity=1}. 115 association rules were obtained, 79 corresponding to non-fatal accidents, and 36 to fatal ones. The main factors associated with the severity of the accident belong to male subjects, involved in accidents that occur on weekends and in suburban areas, and where the probability of the accident being fatal is 1.69 times greater. Thus, the results of using the association rules to relate demographic and circumstantial characteristics of the accident with the severity of the injuries show an accuracy of just over 65%. Therefore, despite the limitations that may occur due to the omission of relevant variables, and the fact that the results show little precision, the feasibility of using machine learning techniques and, specifically, the association rules as promising tools to help analyze accidents and help launch road safety interventions more effectively is manifested.

1 Introduction

In Mexico, each year over sixteen thousand people die as a consequence of a road accident [1], this is equal to 132 deaths per million inhabitants, which is why this cause represents a severe problem of public healthcare. Regarding, this country, diverse studies have been realized in order to analyse the trend of deaths and injuries of motorbikes, bicycles, drivers and pedestrians, as well as the prevalence in the use of helmet, safety belt or alcohol consumption, among other subjects about risk factors [2-4]. As far as accidents in road freight transport, in logistics activities, over 2500 deaths per year are acknowledged, and the accidents are recorded by the National Institute of Statistics and Geography [5] and the Mexican Institute of Transportation [6].

The records include general characteristics of the accidents, in which the types of vehicles are included, as well as the behavioral and socioeconomic data from the injured people. However, despite the great deal of information, the data analysis reduces to the descriptive statistics, and only in some case of correlational manner [7]. Likewise, though in international literature the risk factors of road accidents have been widely [8], the research about road accidents of road freight transportation in Mexico are very limited and it hasn't been deeply investigated about the groups and characteristics of drivers in road freight flows which have particularly high risks of accident [9]. In this sense, the data have been wasted and, as a result, their potential has been limited in the elaboration of public policies or in business programs for accident prevention.

The great amount of available information requires applying interdisciplinary methods of data analysis and realizing studies of an explanatory-casual kind, generalizing behaviors or making inferences in such way that the patterns of relationships between the variables could be acknowledged and extract a better understanding of the high volume of data. Under this context, this research started with the purpose of identifying explanatory or predictive variables of the fatalities of the road freight transport , in which the flow of goods takes place, and as a case study of the statistics of Land Traffic Accidents in Urban and Suburban Zones (ATUS) of the INEGI. Multivariate analysis techniques were attempted; however, the assumptions were not met, or the data did not fit the models. It is how it was thought to use machine learning techniques and, in order to predict the severity of traffic accidents in freight transport in Mexico, the supervised technique of association rules with discrete labels was used

to relate the variables of accidents with the severity of the injuries, that is to say, whether or not it was fatal.

2 Literature review

Technological advances in sensors, positioning systems, information collection devices and, in general, smart devices connected to communication networks provide a large amount of data in all areas and activities on the planet. In this sense, transport is one of the activities with the greatest number of applications; this is demonstrated by Tang et al. [10] who conducted a review of the Big Data literature in forecast research and found that, in addition to a rapid growth in the number of investigations, about 16.55% of the so-called hotspots correspond to transportation issues, and 11.38 % to traffic forecast models.

Therefore, it is not surprising that there are more and more records and information on traffic accidents, transport systems and logistics activities, which generate more research on the application of Data Science to predict, classify and identify the causes of incidents with new models and techniques that improve prediction and information on future prospects in complex situations, without linearity or seasonality.

In this sense, to try to explain the behavior of accidents, all kinds of machine learning techniques have been used. For example, Iranitalab and Khattak [11] used Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVMs) and Random Forests (RF) to predict the severity of traffic accidents; and they conclude that NNC had higher prediction accuracy in general and in the most severe crashes, the RF and SVMs methods followed with sufficient performance, while the MNL method was the weakest.

On their behalf, Arhin and Gatiba [12] used Gaussian naïve Bayes classifiers and support vector machines (SVMs) algorithms to predict the severity of injuries caused by accidents at intersections without traffic lights. Although the first technique had very low performance with an accuracy of 48.5%; with the second, SVMs, they achieved an accuracy of 83.2%, so they conclude that these models can be applied by transportation officials to recognize and perform necessary mitigation actions at collision-prone intersections.

Das et al. [13] also made predictions for injury severity in pedestrian-involved crashes in two cities in the United States and evaluated three algorithms: support vector machines (SVMs), random forests (RFs) and Extreme Gradient Boosting (XGBoost). In this study, different levels of precision were found for each of the evaluated cities and the different algorithms; for SVMs 0.6250 and 0.5849, for RF 0.6250 and 0.5769, and for XGBoost 0.7059 and 0.66. Thus, the best performance was for the XGBoost algorithm.

These studies show that there is no clear hierarchy of methods in terms of their performance in predicting the severity of injuries in traffic accidents. In addition, there are several studies that use techniques such as association rule mining (ARM) to identify sets of attributes and subgroups with a higher probability of severity in traffic accidents [14-16]. Yao et al. [17] found with ARM that the behavior of road users, vehicle factors, geometric characteristics of the road and environmental conditions are among the factors associated with severe traffic accidents. Likewise, Feng et al. [18] found a strong correlation between accidents and environmental factors, speed, and location.

Nonetheless, the studies that use association rules to identify sets of factors that determine the severity of accidents show the importance of carrying out analyzes in geographical locations, with different cultural factors, and for each of the types of transport. Thus, for example, another study analyzing motorcycle accidents with ARM in Australia [19] shows that among the factors with a higher probability of severe injuries are the collision with a truck or a static object, the accident while trying to pass a car or with a vehicle from the opposite direction, and during certain periods of the day, either during early morning or late at night. While Das et al. [20] found that lighting conditions, vehicle turns, and the age range of pedestrians 45 years and older, are frequently present in the factors determined by the association rules.

However, even though these methods have been in use for several decades, there are still relatively few studies on the severity of road accidents. In this way, the diversity of factors determined by the region and culture make it significant to investigate the elements that influence the severity of traffic accidents, the different road users, and the various types of transport; in order to provide perspectives that help understand the causes of the severity of injuries, and develop initiatives and effective public policies that help reduce the severity and deaths caused by traffic accidents.

3 Method

Open data corresponding to the period between 2009 and 2018 of the statistics of Land Traffic Accidents in Urban and Suburban Zones (ATUS) of the National Institute of Statistics and Geography [5] were used . Ten annual files were collected, for which a total of 3,889,989 records of traffic accidents were accumulated. All claims involving a cargo transport vehicle were selected, and participate in the flow of goods, giving a total of 891,172 records corresponding to 22.9% of the total.

Specifically, for the selection of these records, the variable type of vehicle was used, which INEGI [21] classifies into thirteen categories, and among which there are three types of vehicles dedicated to the goods flow; cargo vans, with a capacity of up to one ton; cargo trucks, which carry between one and five tons; and tractors with or without a trailer, which transport more than five tons of goods. After discarding the null records, the database on cargo accidents for the study decade consisted of 718,565

records, 7,464 classified as fatal accidents {Severity=1} and 711,101 as non-fatal {Severity=0}.

Subsequently, with multivariate analysis techniques, an attempt was made to identify explanatory or predictive variables of fatalities in freight transport; however, the assumptions of normality, homoscedasticity, multicollinearity, or the nature of the variables were not met; even some methods of interdependence froze the computer. Given the binary nature of the main variable, which refers to whether there were fatalities in the freight transport accident, it was considered to use binary logistic regression. However, even though the global precision was 76.5%, the sensitivity had a very low value (1.8%), and the total goodness of fit (P <0.001) showed that the data do not fit the model.

This is how it was thought to use machine learning techniques and, with the objective to predict the severity of traffic accidents in freight transportation in Mexico, the supervised technique of association rules was used to investigate among the variables related to the fatalities in traffic accidents of vehicles that execute the flows of goods by road.

4 Association rule mining

The Arules package, created by Hahsler et al. [22] was used for the programming environment, R. Arules provides a basic data analysis infrastructure that results in a set of elements and association rules [23] with a interface capable of processing the Apriori algorithm proposed by [24].

Essentially, the algorithm consists of analyzing a subset of elements contained in the attributes, which give rise to a rule defined as an implication of the form if X ⇒ Y; where the patterns obtained in the sets of elements X, which are called antecedents or Left-Hand-Side (LHS), can be used to predict the class of unclassified records Y, called consequential or Right-Hand-Side (RHS). In addition, the Arules package outputs a diversity of rules, with their respective confidence (1), support (2) and lift (3) values, to quantify the statistical strength of the different patterns. The software internally calculates the parameters with formulas such as the following:

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X} \quad (1)$$

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions} \quad (2)$$

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y} \quad (3)$$

For the case study of this research, the Apriori algorithm was executed with variables from the statistics of Land Traffic Accidents in Urban and Suburban Zones (ATUS), which includes aspects related to people: sex and age; behavioral variables: alcoholic breath and use of seat belt; temporality variables: related to the month of the year, the day of the week, and the time of day; and the variables related to the road, which include the region of the country, and whether or not it is an urban area. The variables that were used, and their discretized values are shown in the following (Table 1).

Table 1 Definition and categorization of variables

Variable	Description	Values
Sex	Sex of the driver	1 Male
		2 Woman
Alcoholic	Evidence of alcoholic breath	1 Yes
		2 No
		3 It is ignored
Belt	Usage of safety belt	0 No
		1 Yes
		3 It is ignored
Region	Region of the country where the accident happened	1 Northwest Zone
		2 Northeast Zone
		3 Occident Zone
		4 Center Zone
		5 Southeast Zone
Season	Season of the year in which the accident happened	1 Spring
		2 Summer
		3 Fall
		4 Winter
Time	Time of day when the accident happened	1 From 00:01 to 05:59
		2 From 06:00 to 11:59
		3 From 12:00 to 17:59
		4 From 18:00 to 23:59

Weekend	Day of the week in which the accident happened	1	From Monday to Friday
		2	Saturday and Sunday
Urban	If the accident happened in Urban Zone	0	Suburban Zone
		1	Urban Zone
Age	Age Group	1	From 1 to 20 years old
		2	From 20 to 30 years old
		3	From 31 to 40 years old
		4	From 41 to 50 years old
		5	51 or older
Severity	Severity of the accident	0	Non-fatal
		1	Fatal

Thus, considering these variables, it is sought to explain the severity of the accidents. For this, the records were classified in which there were only material damages or injured persons {Severity=0} and in which there were deceased persons {Severity=1}. In this sense, for the prediction of fatalities in traffic accidents in freight transport in Mexico, the consequences that refer to accidents in which there were deceased persons were filtered {Severity=1} and their statistical factors were analyzed , to determine the antecedents with the patterns with the most significant associations.

Previously, the variables were decodified and established discretely and encode as a factor. To avoid overfitting and rebalancing the classes between fatal {Severity=1} and non-fatal {Severity=0} accidents, just over one hundredth of the non-fatal records (n=7,463) was randomly selected. These were combined with fatal accidents (FatalAcc; n=7,464) to form a database on rebalanced load accidents (Accdf_Rebalanced).

To monitor the algorithm, the rebalanced data set (n=14,927) was separated into two groups: the first (Accdf_train), denominated as the training set and made up of 70% of the records, and; the second (Accdf_test) that contains the rest of the data (30%) and, was named the test set. Therefore, using the training data, the Apriori algorithm was executed and the rules and associated parameters were obtained. However, given that what was sought was to identify regularities between the different variables and the accidents where there were deceased persons {Severity=1} or not {Severity=0}, a filter was applied on the consequent, to only consider those related to the accident fatality variable (Severity).

Subsequently, using the test data and, to determine the number of rules that each of the records meet, two functions were used: one that fragments the rules and serves to identify the values of the different attributes that integrate each of the rules (function ruleSeparation from the code in the supplemental material); and another that compares and quantifies the rules with the values of the cases, in such a way that it is possible to know which cases meet one or more of the rules (function RuleVsCases).

The number of rules per case was considered as the classification threshold, which is why the sensitivity, specificity and precision of the model were calculated considering from the fulfillment of one and up to ten of the rules in each record. Once the number of rules that presented the best indicators for classification had been established, the prediction of the group belonging to the initial set that contains all the records in which a cargo vehicle was involved was made and; based on this set of rules, the variables related to the level of severity of the accident were established. The R code and data used, including all commands and functions to generate all results, graphs and figures shown in this document, are included in the supplementary material.

5 Results

By applying the Apriori algorithm on the training data (with parameters support = .1, confidence = .5), 785 association rules were obtained. After applying the filter on the RHS to consider only accidents according to the category of whether there were {Severity=1} or not {Severity=0} deceased persons, a total of 115 rules were obtained, 79 associated with accidents where there were not deceased persons, and 36 where there were. Figure 1 shows a graph that considers the support (in the size of the spheres) and the lift (according to the intensity of the color) and it can be seen that, despite not having cases where both values are high, there are combinations where fatalities in cargo transport {Severity=1} show high lift or high support, so it can be inferred that there is a set of rules where fatalities are an important consequent.

Predicting crash injury severity in road freight flows with association rules algorithms
Luis David Berrones-Sanz, Estefania Perez-Diaz, Dulce Maria Monroy Becerril, Esteban Martinez Diaz

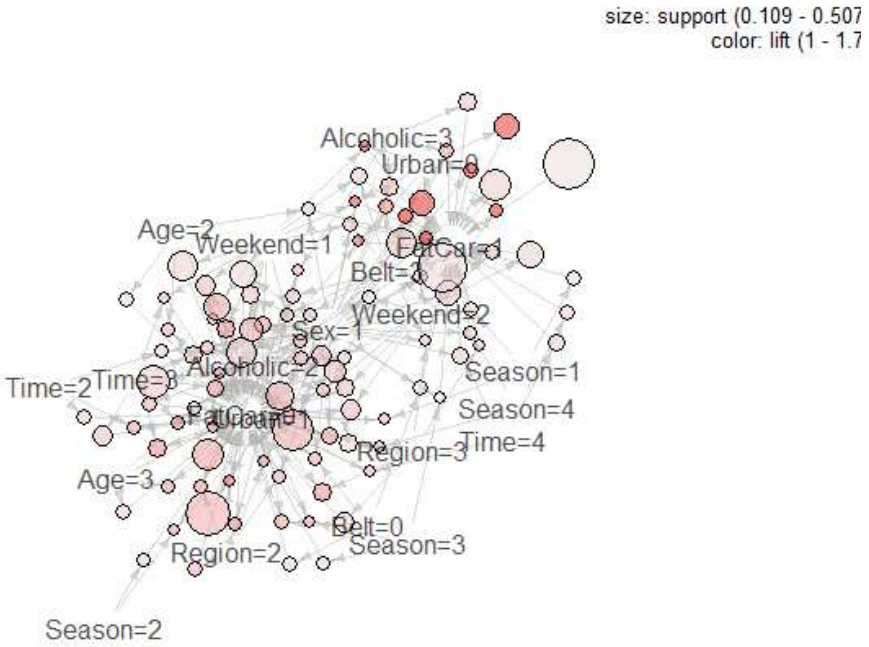

Figure 1 Association rules graphic

Likewise, while inspecting the rules it was found that, for example, when the driver is male, the accident occurs on the weekend and in a suburban area, {Sex = 1, Weekend = 2, Urban = 0} the probability that the accident is fatal is 1.69 times greater. The antecedents and consequences of other rules, as well as their associated parameters can be seen in (Table 2).

Table 2 Some association rules and their associated parameters

LHS		RHS	Support	Confidence	Coverage	Lift	Count
{Weekend=2,Urban=0}	=>	{FatCar=1}	0.127459	0.862887	0.147712	1.700279	1309
{Sex=1,Weekend=2,Urban=0}	=>	{FatCar=1}	0.123661	0.862186	0.143427	1.698897	1270
{Sex=1,Urban=0}	=>	{FatCar=1}	0.234177	0.847129	0.276436	1.669228	2405
{Sex=1,Weekend=1,Urban=0}	=>	{FatCar=1}	0.110516	0.830893	0.133009	1.637236	1135
{Weekend=1,Urban=0}	=>	{FatCar=1}	0.114703	0.825508	0.138948	1.626625	1178
{Sex=1,Alcoholic=2,Urban=0}	=>	{FatCar=1}	0.106134	0.794461	0.133593	1.565447	1090
{Alcoholic=2,Urban=0}	=>	{FatCar=1}	0.109542	0.791696	0.138364	1.560000	1125

The occurrence in suburban areas, gender, the use of the seat belt and the season of the year in which the accident occurs are attributes that are repeated several times in the different rules; This can also be seen in Figure *2*, where a graph of support bars by frequencies is shown with the ten elements that are repeated the most.

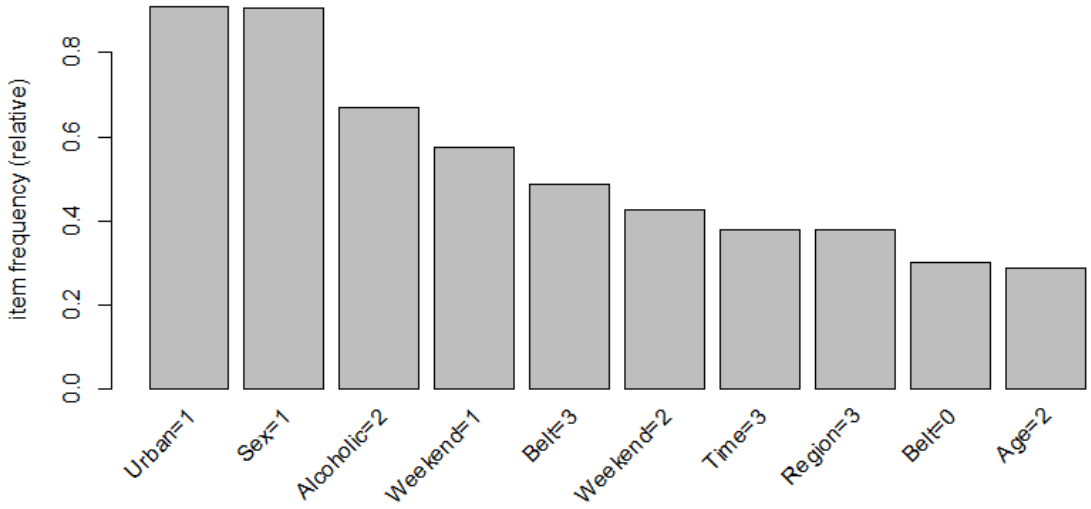

Figure 2 Frequency bar plot

Nonetheless, it is known beforehand that these variables represent risk factors and that, therefore, they increase the probability that the accident is [8]. Given it is sought to identify the elements that are related to fatalities in freight transport accidents, the function to separate rules (ruleSeparation) and another to compare with the cases (RuleVsCases) was used. With these functions, it was possible to identify whether the variables, for each of the cases, obtained the values indicated by the association rules, in addition to quantifying the number of rules that are fulfilled per case. The histogram of Figure 3 shows the frequency of the rules associated with accidents with fatalities and the frequency presented in the test data, it can be seen that, for eight rules per case, the highest frequency of fatalities in vehicle accidents is presented freight involved.

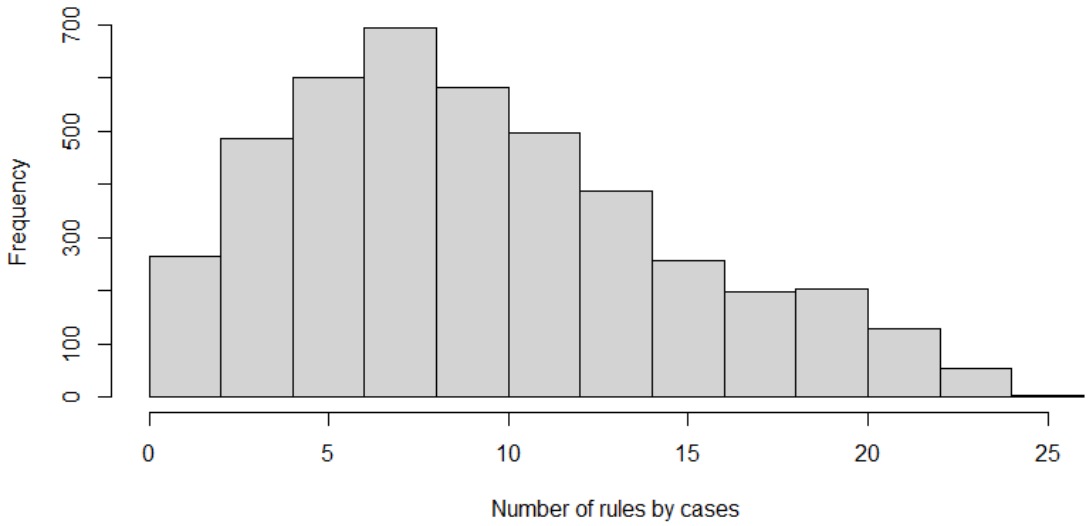

Figure 3 Histogram of the number of rules per case

In this way, eight or more rules are proposed per case to determine the discrimination threshold and, therefore, if the forecast of the cases will be established as a fatal load accident {Severity=1} or as an accident without fatalities {Severity=0}. Thus, by applying the rules to the test data and classifying the cases that are within the threshold, the confusion matrix can be established (Figure 4).

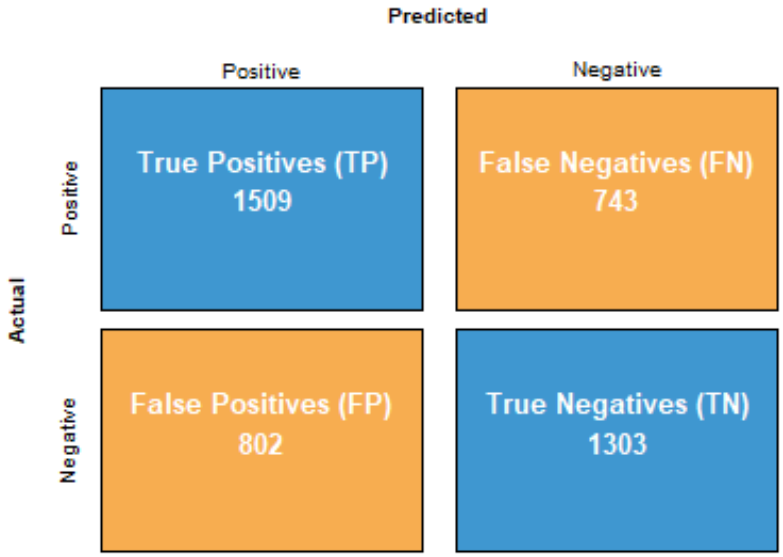

Figure 4 Confusion Matrix for the threshold of 8 or more rules

In addition, if the sensitivity, specificity and precision of the predictions are graphed with different thresholds according to the number of rules, Figure 5 is obtained. In this, it can be seen that near the eight rules per case is where the balance between the value of the sensitivity and that of the specificity is obtained; and values of 0.67 are obtained for the sensitivity, 0.62 for the specificity, and 0.65 for the precision.

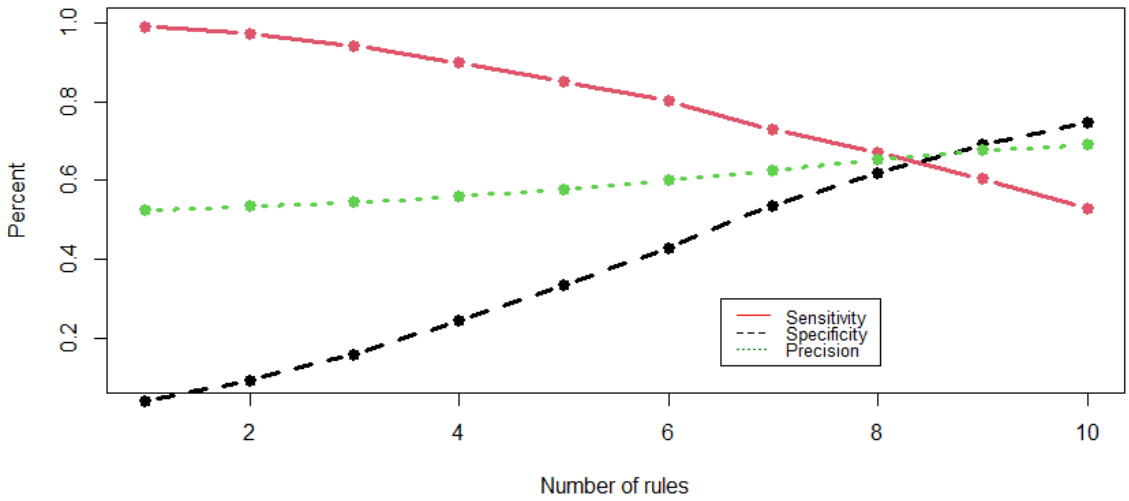

Figure 5 Sensitivity, specificity and precision of the model by number of rules per case

6 Discussion and conclusions

Every day the use of technologies that record information is being integrated and, therefore, a large amount of data on transport and the supply chain activities is being accumulated. Most of the databases are not analyzed in their entirety, they have non-linear relationships, the models are of the non-deterministic polynomial-time hardness (NP-hardness) class or the meta-analyzes are short-range.

The analysis of traffic accidents is an essential task for freight transport companies, and avoid breaks in the flow of goods. However, in Mexico, there is a gap in accident investigation methodology, as most investigations are based on traditional statistical methods.

Thus, it is highlighted the importance and growth that data mining and different machine learning techniques have had to discover patterns, generalize behaviors and make inferences that help to understand traffic accidents better; among which there are, for example, neural networks [25], regression trees [26] or text mining Techniques [27].

In addition, the use of machine learning techniques has become cheap and easy to apply due to free software, such as R or Python, which include libraries and functions that form a very flexible and extensible work environment, and that make it possible to systematize in a constant basis the analysis of information.

In this study, an extensive set of data on traffic accidents in Mexico is used and the Apriori algorithm of association rules is applied to predict the severity of accidents in freight transport. The results of using the association rules to relate the demographic and circumstantial characteristics of the accident with the severity of the injuries show a precision of just over 65%.

It can be inferred that the level of imprecision may be due to the nature of the variables; given that the records are retrospective and were not specifically designed to classify their severity. For example, in the international literature there are well-known risk factors regarding the increase in the severity of the accidents; These include the speed at which the vehicles circulate, weather conditions such as rain, snow or fog, ergonomic conditions of the vehicle, psychosocial factors and the drivers' days, among other working conditions that are not included in the database and which may be omissions of relevant [8,9].

The use of association rules is an alternative that, in addition to indicating the possibility of an accident due to the exposure to some risk such as the odds ratios of traditional epidemiological methods, can be used to classify individuals with a set of characteristics or who are found under specific circumstances or patterns. Therefore, despite the limitations and the fact that the results show insufficient precision, the association rules are a promising tool in the analysis of traffic accidents and, therefore, to provide elements that contribute to launching interventions of road safety more effectively, and that at the same time allows the correct flow of merchandise in logistics activities.

It is proposed to use machine learning algorithms in companies dedicated to the transport of goods with data collected according to a theoretical framework on risk factors in freight transport. In addition, the group of drivers who carry out their journeys without mishaps must be included to establish the variables and characteristics of the protective factors. In this way, characteristics of the drivers, shipments and vehicles of each of the trips that the company makes will be obtained, under the premise that better performance will be obtained, and a more practical application, when applying the algorithms to identify the variables related to accidents and, therefore, in the prediction of the conditions and the drivers that represent less risk.

In conclusion, these types of prediction models help to identify the particularities of those injured and killed by traffic accidents. The determination is relevant in the sense that the characteristics must be used for the elaboration of government actions that promote protective factors and, therefore, reduce the number of accidents and ruptures in the flow of goods that are transported by road. One strategy could be the development of road safety campaigns in which the factors and attitudes that drivers carry out and that lead to accidents are disclosed. Kolter [28] indicates that a campaign to change social behavior must contain some guidelines such as identifying the objective, the factors that prevent the negative attitude of the people to whom the campaign is directed, validation and feedback, execution, monitoring, and campaign evaluation. In general, in Mexico, there are very few efforts and there are very few campaigns dedicated to the self-transportation of cargo, rather they are dedicated to cars to avoid driving under the influence of alcohol, fatigue, or for the use of seat belts.

Thus, with the information resulting from these analyses, and if this type of model is used, it will be possible to establish categories that integrate common arguments and that contemplate both physical, chemical, biological, and mechanical risks; demands derived from the organization and technical division of labor; and elements related to psychosocial and behavioral factors. With this information, alternatives should be sought in the media that are accessible to professional drivers, among these could be social networks, the radio, training courses, the facilities where they load and unload, or where they process their driver's licenses. It is also important to use the mass media and place billboards on-road sections as a graphic means of advertising information.

Acknowledgement
This work has been sponsored by the Universidad Autónoma de la Ciudad de México Project UACM CCYT-2023-IMP-14 and National Polytechnic Institute (IPN), number 20232252

References

[1] STCONAPRA, Report on the situation of road safety: Mexico 2017, México: Secretaría de Salud, Secretariado Técnico del Consejo Nacional para la Prevención de Accidentes (STCONAPRA), [Online], Available: https://www.gob.mx/salud/acciones-y-programas/acerca-del-stconapra [05 Mar 2023], 2018. (Original in Spanish)

[2] MURO-BÁEZ, V.A., MENDOZA-GARCÍA, M.E., VERA-LÓPEZ, J.D., PÉREZ-NÚNEZ, R.: Analysis of road traffic injuries in Mexican cyclists, *Gaceta Médica de México*, Vol. 153, No. 6, pp. 653-661, 2017. https://doi.org/10.24875/GMM.17002632 (Original in Spanish)

[3] SÁNCHEZ, H., CHIAS, L., RESÉNDIZ, H.: Evolution of Traffic Accidents in the Urban and Suburban Areas in Mexico during 1997-2016: Higher Risk Exposure and Lower Lethality, *Revista Gerencia y Políticas de Salud*, Vol. 18, No. 37, pp. 1-24, 2019. (Original in Spanish)
https://doi.org/10.11144/Javeriana.rgps18-37.eatz

[4] BERRONES-SANZ, L.D.: Analysis of accidents and injuries on motorcycles in Mexico, *Gaceta Médica de México*, Vol. 153, No. 6, pp. 662-671, 2017. https://doi.org/10.24875/GMM.017002812 (Original in Spanish)

[5] INEGI, *Road traffic accidents in urban and suburban areas*, [Online], Available: https://www.inegi.org.mx/programas/accidentes [05 Mar 2023], 2023. (Original in Spanish)

[6] IMT, *Statistical yearbook of accidents on federal highways (2018)*, Sanfandila: Instituto Mexicano del Transporte (IMT), [Online], Available: https://www.gob.mx/imt [05 Mar 2023], 2019. (Original in Spanish)

[7] GUTIÉRREZ, J., CUEVAS, A., SORIA, V., VILLEGAS, N.: Correlation between Vehicle Composition and Accidents in the Federal Highway Network, during the Period 2006-2016, Phase I. Sanfandila: Instituto Mexicano del Transporte, *Publicacion Tecnica*, Vol. 2018, No. 529, pp. 1-67, 2018. (Original in Spanish)

[8] ELVIK, R., HOYE, A., TRULS, V., SORENSEN, M.: *The handbook of road safety measures*, 2nd ed., Bingley UK, Emerald, 2009.

[9] BERRONES-SANZ, L.D., CANO, P., SÁNCHEZ, D., MARTÍNEZ, J.L.: Injuries, diseases, and occupational accidents of cargo drivers in Mexico, *Acta Universitaria*, Vol. 28, No. 3, pp. 47-55, 2018. https://doi.org/10.15174/au.2018.1946 (Original in Spanish)

[10] TANG, L., LI, J., DU, H., LI, L., WU, J., WANG, S.: Big Data in Forecasting Research: A Literature Review, *Big Data Research*, Vol. 27, No. February, 2022. https://doi.org/10.1016/j.bdr.2021.100289

[11] IRANITALAB, A., KHATTAK, A.: Comparison of four statistical and machine learning methods for crash severity prediction, *Accident Analysis & Prevention*, Vol. 108, pp. 27-36, 2017. https://doi.org/10.1016/j.aap.2017.08.008

[12] ARHIN, S.A., GATIBA, A.: Predicting crash injury severity at unsignalized intersections using support vector machines and naïve Bayes classifiers, *Transportation Safety and Environment*, Vol. 2, No. 2, pp. 120-132, 2020. https://doi.org/10.1093/tse/tdaa012

[13] DAS, S., LE, M., DAI, B.: Application of machine learning tools in classifying pedestrian crash types: A case study, *Transportation Safety and Environment*, Vol. 2, No. 2, pp. 106-119, 2020. https://doi.org/10.1093/tse/tdaa010

[14] PRATI, G., ANGELIS, M.DE, PUCHADES VM, FRABONI F, PIETRANTONI L. Characteristics of cyclist crashes in Italy using latent class analysis and association rule mining, *PLoS ONE*, Vol. 12, No. 2, pp. 1-28, 2017. https://doi.org/10.1371/journal.pone.0171484

[15] SHIN, D.-P., PARK, Y.-J., SEO, J., LEE, D.-E.: Association Rules Mined from Construction Accident Data, *KSCE Journal of Civil Engineering*, Vol. 22, No. 4, pp. 1027-1039, 2018. https://doi.org/10.1007/s12205-017-0537-6

[16] XU, C., BAO, J., WANG, C., LIU, P.: Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China, *Journal of Safety Research*, Vol. 67, pp. 65-75, 2018. https://doi.org/10.1016/j.jsr.2018.09.013

[17] YAO, Z., DENG, W., WU, D.: *Association Rule Analysis of Contributory Factors to Severe Traffic Accidents*, CICTP 2018: Intelligence, Connectivity, and Mobility - Proceedings of the 18th COTA International Conference of Transportation Professionals, 2018. https://doi.org/10.1061/9780784481523.186

[18] FENG, M., ZHENG, J., REN, J., XI, Y.: *Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK*, In: Ren, J., et al. Advances in Brain Inspired Cognitive Systems, BICS 2019, Lecture Notes in Computer Science, Vol. 11691, Springer, Cham., 2020. https://doi.org/10.1007/978-3-030-39431-8_50

[19] JIANG, F., YUEN, K., LEE, E.: Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology, *Journal of Safety Research*, Vol. 75, pp. 292-309, 2020. https://doi.org/10.1016/j.jsr.2020.09.004

[20] DAS, S., TAMAKLOE, R., ZUBAIDI, H., OBAID, I., ALNEDAWI, A.: Fatal pedestrian crashes at intersections: Trend mining using association rules, *Accident Analysis & Prevention*, Vol. 160, No. September, 2021. https://doi.org/10.1016/j.aap.2021.106306

[21] INEGI, Methodological synthesis of the statistics of land traffic accidents in urban and suburban areas 2016, Instituto Nacional de Estadística y Geografía, México: Instituto Nacional de Estadística y Geografía (INEGI), [Online], Available: https://www.inegi.org.mx/programas/accidentes [05 Mar 2023] 2016. (Original in Spanish)

[22] HAHSLER, M., GRÜN, B., HORNIK, K.: Introduction to arules — Mining Association Rules and Frequent Item Sets, *ResearchGate*, Vol. 2006, No. August, pp. 1-28, 2006.

[23] HAHSLER, M., GRÜN, B., HORNIK, K.: Arules - A Computational Environment for Mining Association Rules and Frequent Item Sets, *Journal of Statistical Software*, Vol. 14, No. 15, pp. 1-25, 2005. https://doi.org/10.18637/jss.v014.i15

[24] AGRAWAL, R., SRIKANT, R.: *Fast Algorithms for Mining Association Rules in Large Databases*, In: Bocca J, Jarke M, Zaniolo C, editors, Proceedings of the 20th International Conference on Very Large Data

Bases. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, 1994, pp. 487-499, 1994.

[25] PAN, G., FU, L., THAKALI, L.: Development of a global road safety performance function using deep neural networks, *International Journal of Transportation Science and Technology*, Vol. 6, No. 3, pp. 159-173, 2017. https://doi.org/10.1016/j.ijtst.2017.07.004

[26] WAHAB, L., JIANG, H.: Severity prediction of motorcycle crashes with machine learning methods, *International Journal of Crashworthiness*, Vol. 25, No. 5, pp. 485-492, 2020. https://doi.org/10.1080/13588265.2019.1616885

[27] ZHANG, X., GREEN, E., CHEN, M.: Souleyrette RR, Identifying secondary crashes using text mining techniques, *Journal of Transportation Safety & Security*, Vol. 12, No. 10, pp. 1338-1358, 2020. https://doi.org/10.1080/19439962.2019.1597795

[28] KOLTER, P., ROBERTO, E.: *Mercadotecnia social: Estrategias para cambiar la conducta pública*, España: Diaz de Santos, 1989. (Original in Spanish)

Review process
Single-blind peer review process.