# BAYESIAN NETWORKS MODELING IN TRAFFIC ACCIDENTS: CASE STUDY FOR BRAZIL

**Maria Lígia Chuerubim**
Faculty of Civil Engineering
Federal University of Uberlândia
**Alan Valejo**
Institute of Mathematical and Computer Sciences
School of Engineering of São Carlos
University of São Paulo
**Irineu da Silva**
Department of Transport Engineering
School of Engineering of São Carlos
University of São Paulo

**ABSTRACT**
The construction of networks from databases of road accidents is considered a challenging issue in road safety. The accident databases are naturally unbalanced, which requires the use of models that allow the simultaneous analysis of multiple variables, without a priori prerequisites to be established. Bayesian Networks (BNs) have presented promising results in relation to traditional methods. Based on this motivation, this paper analyses the effectiveness of BNs in the classification of the severity of accidents, together with factors related to the causality, the road infrastructure and the environmental conditions. The results indicate that, when used in conjunction with balancing techniques, the BNs perform better in the classification of the minority classes, with an average accuracy of 84.0%, while the classification of the unbalanced data indicates an average pseudo-accuracy of 56.3%, diagnosed by the low proportion of instances correctly predicted for the minority class.

**RESUMO**
A construção de redes a partir de bases de dados de acidentes rodoviários é considerada um tema desafiador em segurança viária. Os bancos de dados de acidentes são naturalmente desbalanceados, o que requer o uso de modelos que permitam a análise simultânea de múltiplas variáveis, sem que pré-requisitos a priori sejam estabelecidos. As Redes Bayesianas (RBs) têm apresentado resultados promissores em relação aos métodos tradicionais. Com base nesta motivação, este trabalho analisa a efetividade das RBs na classificação da gravidade de acidentes, em conjunto com fatores relacionados a causalidade, a infraestrutura viária e as condições ambientais. Os resultados indicam que, quando utilizado em conjunto com técnicas de balanceamento, as RBs apresentam um melhor desempenho na classificação das classes minoritárias, com acurácia média de 84,0%, enquanto que a classificação dos dados desbalanceados indica uma pseudo-acurácia média de 56,3%, diagnosticada pela baixa proporção de instâncias corretamente previstas para a classe minoritária.



## 1. INTRODUCTION

In environments of a complex nature, such as highways, the methods based on regression models and extraction of decision rules if they are inadequate for the treatment of multicausal problems, by the limited interpretation and representation of the relations between the predictive variables and the target variable. In this environment, a likely cause of an accident, for example, may be related to the driver's behavior, vehicle type, road characteristics, and environmental conditions.

In recent years, different approaches have been proposed with the purpose for modeling the nebulous relationship between the factors that contribute to the occurrence of accidents. In this perspective, Bayesian probability inference has been widely used in the process of modeling complex, multidimensional real-world problems in the form of network structures (Khoo and Ahmed, 2018; Fu *et al.*, 2018; Zhu *et al.*, 2017). In a Bayesian Network (BN), the nodes represent the observed data and their respective attributes, regardless of their distribution, whereas which the arcs express the probabilistic relationships between them (Zhu

*et al.*, 2017; Wach, 2016).

The BNs are applied by means of data mining techniques, in order to represent the non-linearity between the predictive variables and the target variable, as well as the uncertainties present in the modeling development. The modeling based in BNs, permit calculate the posterior distribution of the parameters related to the model obtained, from the prior distribution of the instantiations of the variables, by means of the analysis of the frequencies obtained by multivariate regression and by the probability distribution of the data (Khoo and Ahmed, 2018; Schlüter *et al.*, 2018). When there is no the knowledge a priori of the distribution of the database, it is assumed that all variables are equally distributed (Soro and Wayoro, 2017, Spiegelman *et al.*, 2010).

The BNs have been showing promising results in relation to traditional methods in the classification of traffic accidents according to the severity of the injury (De-Ona *et al.*, 2011; Lord and Mannering, 2010), and to assess the risk of accidents (Deublein *et al.*, 2013, Zhao and Deng, 2015). This modeling not require the prior knowledge of the data distribution and have the capacity to assimilate the uncertainty, that is, the evidence of an accident based on probability theory (Mannering *et al.*, 2016, Zong *et al.*, 2013).

However, the BNs modeling has been limited by the amount of data and the quality of records of the occurrence of accidents, especially in developing countries such as Brazil, which makes it difficult to apply them in classification studies in the context road safety. This drawback contributes to the generation of unbalanced databases, in which the number of instances in each category of the target variable is not equally distributed (Crone and Finlay, 2012). The imbalance between class instances leads to incorrect classifications of the minority classes, such as accidents that result in fatal and non-fatal victims (Japkowicz, 2000).

The objective of this study is to apply the modeling by BNs in the assessment of the degree of severity of traffic injury, considering the multiple factors related to the occurrence of incidents, probable cause, the geometric of the segments of the highway, the climatic conditions and the particularities of the individual record of each accident. Based on the model is intended to understand, the main factors that influence the level of safety in an ideal urban stretch of a highway. The experimental results, obtained by the use of the SMOTE (Synthetic Minority Over-sampling TEchnique) balancing technique (Li *et al.*, 2018) associated with the algorithm of classification of Naive Bayes networks, provided a performance of the classification of the severity of the accidents compatible with the values found in the literature. In addition, they allowed identifying the main factors that conditioned the eventuality of accidents on the highway.

## 2. RELATED WORK

Several authors applied a NMB in the context of Road Safety. This study had as its main theme the influence of accidents, environmental conditions and mitigation of the severity of road accidents (Chen *et al.*, 2015, Liang *et al.*, 2017).

For the knowledge of the orientations used, we directed the reader to Zhu *et al.* (2017) based on the BNM as relations between the position positions collected with the Global Navigation Satellite System, with a collision frequency and an individual exposure to the risk, in order to evaluate the behavior of the driver behind the wheel. The authors concluded that drivers

driving at speeds above the established threshold and accelerating or rapidly decelerating on expressways are more likely to be involved in collision events. In addition, younger female drivers, who make more trips and travel longer stretches of the highway, represent the driver profile with the highest risk of exposure to accidents.

In another study of Mujalli *et al.* (2016) applied different data balancing techniques and Bayes classifiers to identify factors that affect the severity of an accident. To do so, they used a database of unbalanced traffic accidents observed over the three-year period in Jordan (2009-2011). The results indicated that the use of balanced databases, especially those obtained with oversampling techniques, when used in conjunction with BN, improves the performance of the classification of the severity of a road accident. Consequently, reduces the error in the classification of occurrences with serious or fatal injuries (minority class) in relation to occurrences with mild or moderate injuries (major class). Based on this model, the authors also identified the factors that contributed to the occurrence of accidents with serious or fatal injuries, such as the number of vehicles involved, type of accident, lighting, runway condition and speed limit.

De Oña *et al.* (2011) also applied the modeling by BNs in the study of accident database for a highway in Spain, aiming to classify the accidents in different levels of moderate, severe and fatal injury. For this, he selected 18 variables intrinsic to the driver, the highway, the vehicle and the accident. In addition, it applied different models of networks to reduce the number of variables and simplify the study problem, which contributed to the identification of statistically significant variables such as type of accident, driver profile (age and gender), highway lighting, number of injured persons and occupants in each vehicle. The author concluded that it is feasible to reduce the number of variables applying BNs without modeling accuracy is degraded.

Zou *et al.* (2017) discussed the ability of BNs to store modeling uncertainty, resulting in superior accuracy when compared to traditional statistical techniques, such as regression models, in predicting the severity of traffic accidents. Similar observations were made by Deublein *et al.* (2014), which obtained in its studies an accuracy level of 86.0% with a tolerance of 25% in the process of modeling the severity of traffic accidents.

The motivation of the paper is based on the fact that others researches in context of the road safety no do not address the fundamentals and relevance of the process of the data balancing process in Bayesian networks.

## 3. MATERIALS AND METHOD

### *3.1 Road accident database*

The data of road accidents in Brazil refer to the Dom Pedro I highway (SP-065), located in the urban perimeter of the city of Campinas, State of São Paulo, for a period of 8 years (2009-2016). The total number of accidents observed in this period was 4,259. As the objective of this work is to classify the degree of severity of the accidents and to identify the main factors that conditioned the traffic occurrences, the observations recorded during the construction period in the stretch between km 125 to km 145.5. Due to the difficulty of observational records in this situation, of incomplete records of occurrences or errors in relation to the positional location of accidents. In addition, with the exclusion of the data construction

period, it is avoided to generate underestimated indicators of the highway safety levels.

In this study, twelve variables were used to identify the main factors that conditioned the variation in the degree of injury from traffic accidents. The selected variables are based on observational records collected by the Rota das Bandeiras Concessionaire and empirical data collected in the field. The data include discretized variables in different values and associated to the accident type (rear-end collision, head-on collision, sideswipe collision, transverse collision, pile-up, rollover, overturning, pedestrian collision, crash with fixed or mobile object and fall of motorbikes and motorcycles), the accident cause (driver, vehicle, road/environment and others), the period (morning, afternoon and night), the milestone (km), the visibility condition (normal, partial and adverse), the weather condition (good, rain, cloudy (good, partial and poor). In addition, the pavement condition (dry, wet and oily), the geometry road (straight, smooth curve and sharp curve), the profile road (slope, slope and level), the horizontal signalling (exists and does not exist), the vertical signalling (there is and does not exist), and intervention scenario (before and after). The dependent variable (target) was traffic accidents without victims and with victims (fatal and non-fatal).

*3.2 Bayesian network modeling*

The methodology used in this experiment consists in the preliminary application of the SMOTE database balancing technique (Li *et al.*, 2018). In order, to balance the original unbalanced data between the majority (no injury - NI) and minority (with injury - WI) classes present in the base of highway accident records SP-065, as well as improving the performance of the classification process (Thammasiri *et al.*, 2014).

The SMOTE technique resulted in 353 records, according to the phylogenetic of the minority class by interpolation (Weiss, 2004; Sain and Purnami, 2015), of which 61.19% correspond to NI and 38.81% to WI accidents. It used because considered one of the most efficient sampling techniques since it defines the closest neighbours for each minority class (Sain and Purnami, 2015). After rescaling or balancing the database, the accident severity modeling applied using the BNs approach to construct the model based on the data and using the Naive Bayes algorithm.

In the process of learning the BN at each iteration, the database was divided from the cross-validation into a set of test and training with, respectively, 70.0% and 30.0% of the data. Figure 1 schematically shows the previous balance of the database by the SMOTE technique and the modeling of the severity of the accidents in the highway SP-065 using the modeling with BNs and the Naive Bayes classifier.

The Weka software adopted to aid the Bayesian modeling process. It is a free data mining software, developed by the University of Waikato, New Zealand, in Java language and available to users under the GNU (GPL) license.

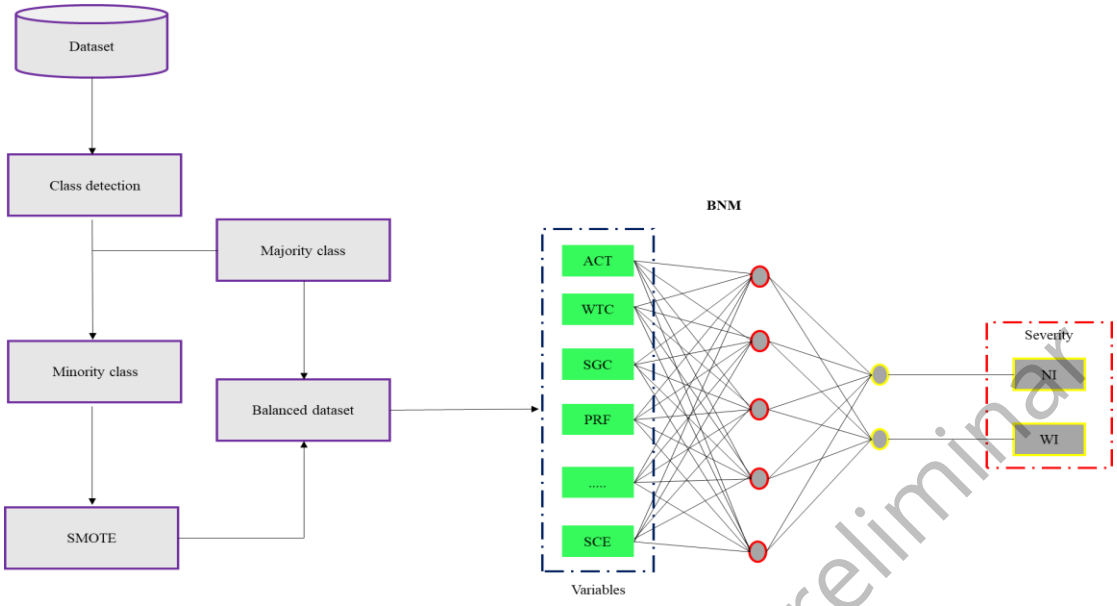


**Figure 1:** Balancing methodology and BN modeling

## 4. EXPERIMENTS AND ANALYSIS OF THE RESULTS OBTAINED

The general accuracy of the classification of the severity of the accidents with a modeling by Bayesian networks was 84.5%. Out of 353 records, 262 instances (74.2%) were classified class and 91 instances (25.8%) were classified incorrectly. A table of classification error (Table 1) allows analysing the results obtained from the perspective of road safety (Mujalli *et al.*, 2016).

The results provided a considerable gain in the accuracy of the classification by Naive Bayes in networks linking its use to the database balanced by the SMOTE technique. The classification of the original database presents a degraded accuracy of 44.0%, in which 3,185 instances were correctly classified (74.8%) and 1,074 erroneously classified (25.2%).

Table 1 shows the general performance statistics of the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) ratings, considering the classification error obtained with the original database and the database (Chen *et al.*, 2015). The TP corresponds to the number of real positives correctly classified, that is, the number of NI and WI accidents. The TN defines the number of negative real duly classified, that is, the number of accidents that culminated in fatal and non-fatal injuries. The FP expresses the percentage of negative instances incorrectly classified as positive, while the FN indicates the percentage of positive instances erroneously classified as negative.

The class accuracy obtained in the BN modeling process using the Naive Bayes classifier based on the unbalanced (Table 2) and balanced (Table 3) database. The AUC accuracy of the modeling using unbalanced data was of the order of 56.3% and based on the proposed model of 84%.

The rate of change in TP indicates the proportion of predicted instances correctly as positive for all positive real cases, while the rate of change in FP expresses the proportion of instances predicted and incorrectly classified as positive for all negative real cases (Chen *et al.*, 2015).

In this experiment, considering the proposed model, TP rates ranged from 0.690 to 0.825, respectively, for the NI and WI accident classes, which indicates that the Naive Bayes classifier performs well for the classes under analysis, demonstrating the capacity to classify 82.5% of the instances corresponding to the WI accidents and 69.0% of the instances corresponding to the NI accidents. Considering that the rates of variation TP and FP are not performance metrics, one must evaluate the accuracy, F-Measure and AUC (Table 3).

**Table 1:** Matrix confusion

| Original Dataset | | | Balanced Dataset | | |
|---|---|---|---|---|---|
| Severity | NI | WI | Severity | NI | WI |
| NI | 3,181 (TP) | 3 (FN) | NI | 149 (TP) | 67 (FN) |
| WI | 1,071 (FP) | 4 (TN) | WI | 24 (FP) | 113 (TN) |

**Table 2:** Accuracy of the Naive Bayes classifier in traffic accident modeling using the original database

| Severity | TP Rate | FP Rate | Precision | F-Measure | AUC |
|---|---|---|---|---|---|
| NI | 0.999 | 0.996 | 0.748 | 0.856 | 0.563 |
| WI | 0.004 | 0.001 | 0.571 | 0.007 | 0.563 |
| Weighted average | 0.748 | 0.745 | 0.704 | 0.641 | 0.563 |

**Table 3:** Accuracy of the Naive Bayes classifier in traffic accident modeling using the balanced database

| Severity | TP Rate | FP Rate | Precision | F-Measure | AUC |
|---|---|---|---|---|---|
| NI | 0.690 | 0.175 | 0.861 | 0.766 | 0.840 |
| WI | 0.825 | 0.310 | 0.628 | 0.713 | 0.840 |
| Weighted average | 0.742 | 0.228 | 0.771 | 0.745 | 0.840 |

Precision matches the proportion of instances that are properly sorted across all instances and provides general information about classifier performance. However, this metric is not efficient in evaluating unbalanced databases (Mujalli *et al.*, 2016). In these cases, a pseudo-precision of the classification process is obtained using the balanced database, since the classifier is sensitive to class distribution, presenting a high hit rate for the majority class (accidents without victims) and detriment rate for the minority class (casualty accidents).

In this study, an accuracy of 86% was obtained for the classification of the NI instances and 62% for the WI instances for the balanced database. These values are consistent since, for unbalanced databases, close to 90.0% indicate an exaggerated adjustment of the data to the model, that is, in these cases, the precision of the model is based on the classification of the majority class of the data (De Oña *et al.*, 2013; Mujalli *et al.*, 2016).

Sensitivity and specificity are generally adopted to monitor the performance of classification in two classes separately. Sensitivity represents the proportion of accurately predicted accidents as NI among all observed as NI. In this study, considering the proposed methodology, it was obtained a sensitivity of 86.1% for the NI class and 62.8% for the NI, for WI, which indicates that the classifier was more sensitive to the instances corresponding to NI accidents (Mujalli *et al.*, 2016).

The F-Measure is obtained by the harmonic mean considering the precision and the sensitivity of the modeling, is indicated to the study of unbalanced databases (Mujalli *et al.*, 2016; Wang *et al.*, 2015). The values obtained for the F statistic for the NI and WI classes were, respectively, 76.6% and 71.3%, can be considered statistically compatible considering the balanced database (Mujalli *et al.*, 2016).

The AUC/ROC (Area Under the curve of the Receiver Operating Characteristic) represents the relationship between sensitivity and specificity (Mujalli *et al.*, 2016). It is useful in describing the general performance of the classification, in which values close to 1.0 indicate a perfect classification and values close to 0.50 are considered negligible (Mujalli *et al.*, 2016).

Figure 2 and Figure 3 present, respectively, the variation of the AUC considering the original database and, respectively, the NI and WI instances. Figure 4 and Figure 5, respectively, describe, respectively, the variation of the AUC considering the balanced database, respectively, for the NI and WI instances. The results obtained indicated an excellent fit of the data to the model (0.84), that is, of 84%, which is compatible with the values found in the literature (Mujalli *et al.*, 2016).

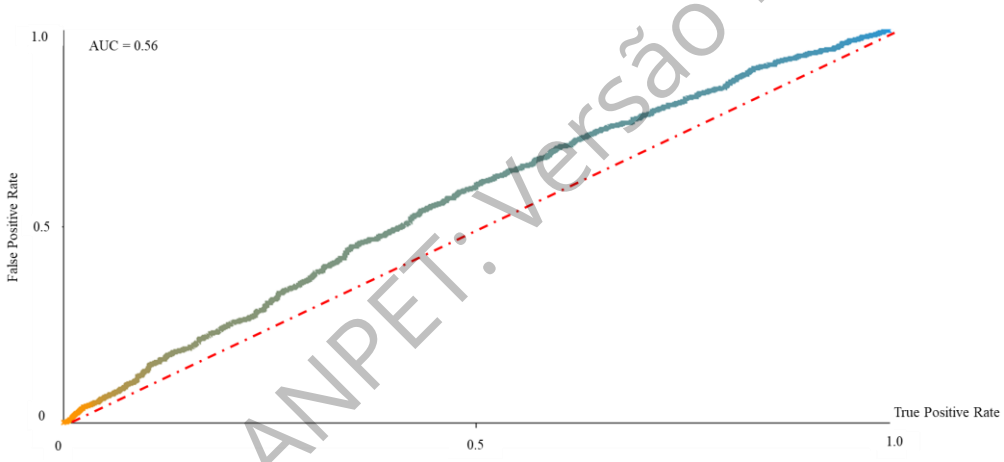


**Figure 2:** AUC of category NI with the original database

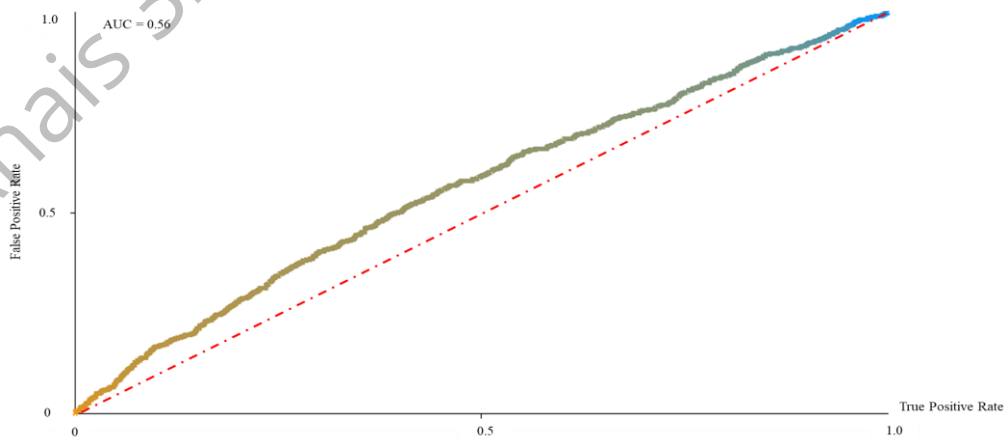


**Figure 3:** AUC of category WI with the original database

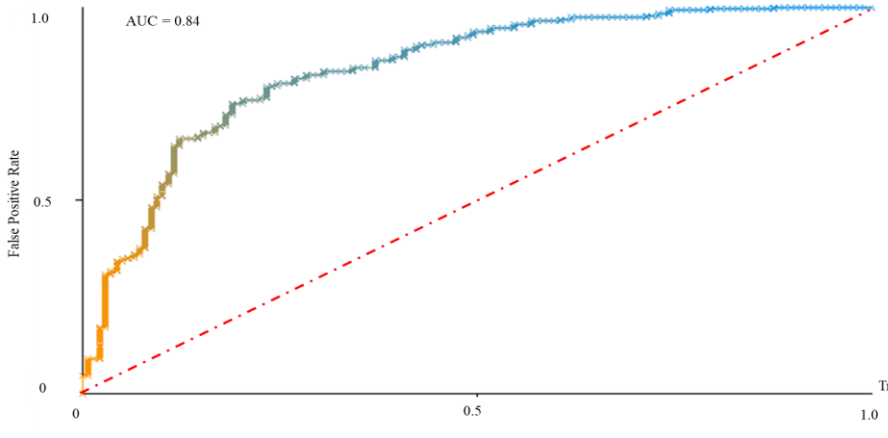


**Figure 4:** AUC of category NI with the balanced database

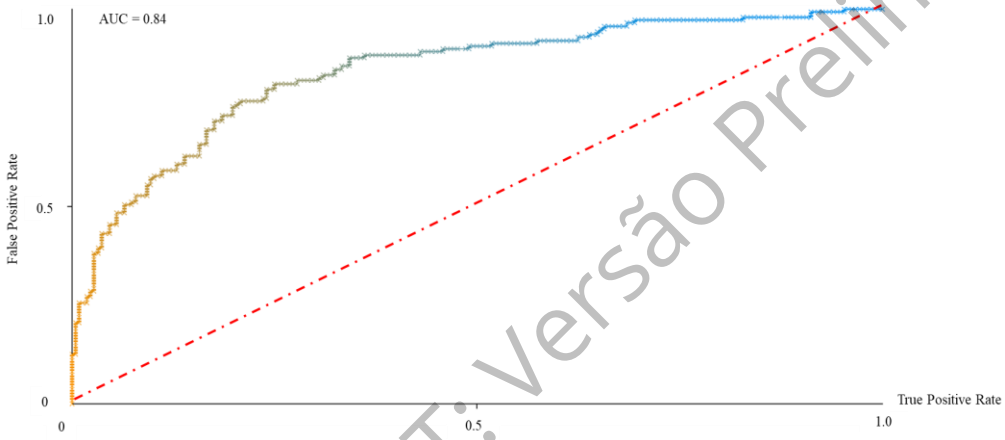


**Figure 5:** AUC of category WI with the balanced database

Table 4 presents the results obtained with BNM with the means and standard deviations for each variable under analysis. It is verified that the least significant variables, highlighted in Table 4, were vertical signalling and intervention scenario. In this way, the analyses will be extended only to the more explanatory variables. The most explanatory variables in this case study for severity modeling were type of accident and probable cause.

**Table 4:** Modeling results with Bayesian networks

| | Original Dataset | | | | Balanced Dataset (SMOTE) | | | |
|---|---|---|---|---|---|---|---|---|
| | NI | | WI | | NI | | WI | |
| Variables | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Accident type | 4.9 | 3.3 | 5.5 | 3.4 | 5.3 | 3.2 | 6.1 | 1.6 |
| Accident cause | 5.5 | 1.4 | 5.6 | 1.2 | 5.6 | 1.2 | 6.0 | 1.0 |
| Period | 1.8 | 0.7 | 1.8 | 0.8 | 1.8 | 0.7 | 2.2 | 0.8 |
| Milestone | 137.8 | 5.6 | 137.7 | 5.8 | 137.7 | 6.0 | 136.6 | 5.5 |
| Visibility Condition | 1.4 | 0.6 | 1.4 | 0.6 | 1.4 | 0.6 | 1.7 | 0.5 |
| Weather Condition | 1.4 | 1.0 | 1.2 | 0.8 | 1.6 | 0.9 | 1.3 | 0.2 |
| Pavement condition | 1.2 | 0.6 | 1.2 | 0.5 | 1.5 | 0.4 | 1.5 | 0.3 |
| Geometry road | 1.2 | 0.6 | 1.2 | 0.6 | 1.2 | 0.6 | 1.3 | 0.7 |
| Profile road | 2.0 | 0.8 | 2.0 | 0.8 | 2.0 | 0.7 | 2.0 | 0.8 |
| Horizontal signalling | 1.1 | 0.3 | 1.0 | 0.3 | 1.0 | 0.3 | 1.1 | 0.3 |
| Vertical signalling | 1.0 | 0.3 | 1.0 | 0.3 | 0.1 | 0.5 | 0.0 | 0.3 |
| Intervention scenario | 0.3 | 0.5 | 0.3 | 0.5 | 0.3 | 0.5 | 0.2 | 0.4 |

The classification of NI accidents (Figure 6) resulted in an average score of 0.82 and a minimum and maximum value of, respectively, 0.51 and 1.00. While the classification of WI accidents (Figure 7) resulted in an average score of 0.76 and a minimum and a maximum score of, respectively, 0.51 and 0.97.

Considering the instances classified with scores greater than or equal to 0.7 (Hosmer and Lemeshow, 2000), the key factors that contributed to the occasionally of NI and WI accidents were evidenced.

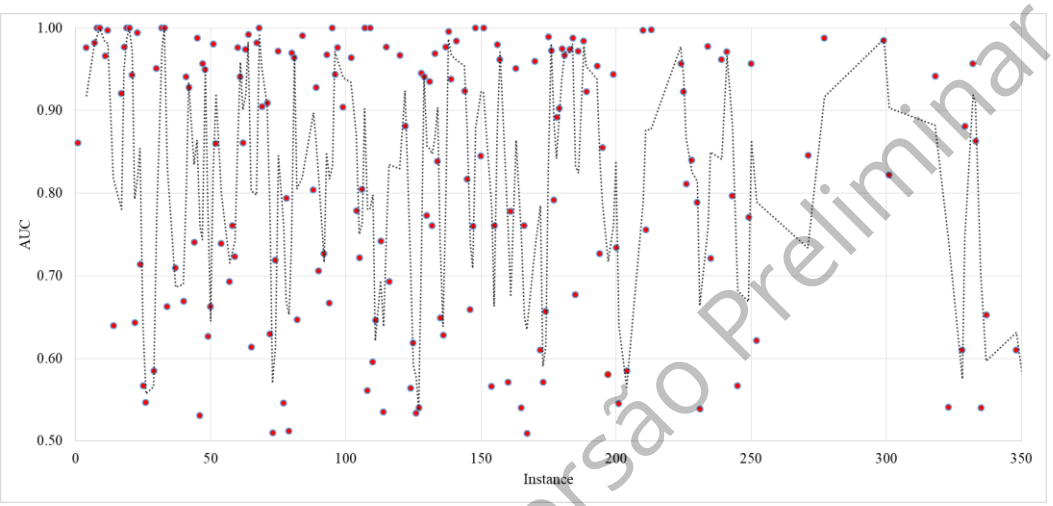


**Figure 6:** Scores classification of NI classes

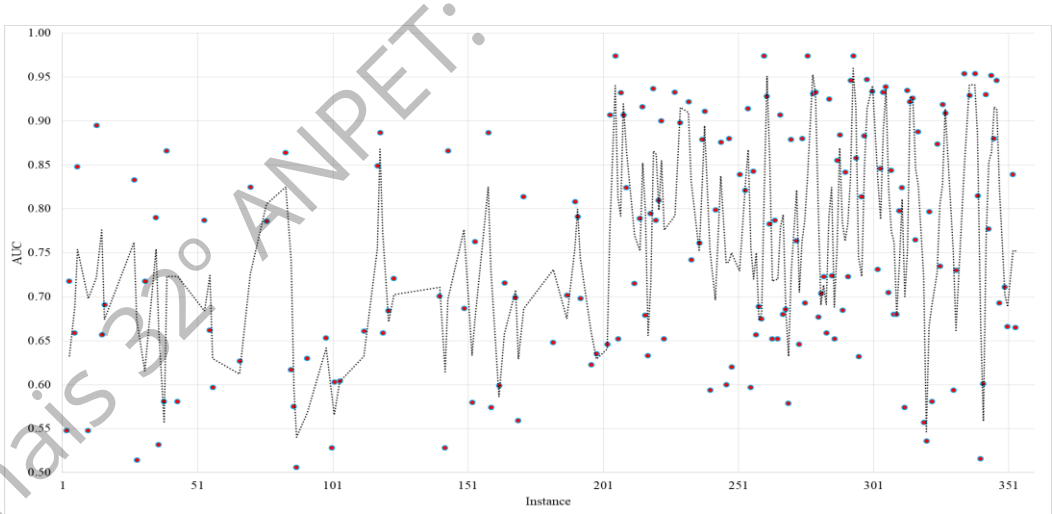


**Figure 7:** Scores classification of WI classes

For the stretch between km 125 and km 145.5 of the SP-065 highway in the period 2009-2016, it was verified that the accidents that are most likely to cause NI accidents were rear-end collision (100%), head-on collision (100%), crash with fixed or mobile object and fall with motorbikes and motorcycles (84.6%). The accidents of the type transverse collision, sideswipe collision, and pile-up have equivalent probability rates of causing NI and WI accidents. These accidents are most likely to occur in weather condition good (45.2%), rain (100%) e cloudy (100%). They are most likely observed in normal (70.2%) and adverse

(80.0%) visibility condition. It is probably in straight (54.8%) and smooth curve (41.2%) track layout.

The NI accidents were observed period occur in dry (45.5%) and wet (84.4%) track condition. These accidents are probably in the morning (69.2%) and afternoon (49.6%) day period. The probable cause of NI accidents are probably related to a cyclist on track (100%), traffic jam (100%), road and environment (100%), vehicle (70%), driver (57.3%) and other factors (51.5%). This type accident is frequently in all highway study stretch.

The WI accidents most likely they are of the type rollover (92.9%), pedestrian collision (93.3%) and overturning (83.3%). These accidents occur predominantly in good weather condition (54.8%) and partial visibility condition (72.5%). They are probably in sharp curve track layout, but also occur in stretches of the highway with straight (45.2%) or smooth curve (58.8%) track layout. These accidents were observed dry (54.5%) and wet (15.6%) track condition. The WI accidents are probably in afternoon (50.4%) and night (70.9%) day period. The probable cause of WI accidents is pedestrian on track (97.8%), self-destruction (100%), driver (42.7%) and other factors (48.5%). These accidents are more likely to occur at km 128, km 130, km 132 at km 137 and at km 142.

The NI and WI accidents occur with all probability equivalents in all track profiles. These accidents occur with equivalent probabilities in all trajectory profiles and in locations where there is the presence of horizontal and there is no vertical signalling. NI accidents occur predominantly in the scenario after the intervention of the concessionaire in the section under study due to the implementation of accident countermeasures. WI accidents, in turn, are more likely to occur in the scenario that precedes intervention in the stretch.

## 5. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

This paper used BN modeling to analyse traffic accident data with the purpose of evaluating the performance of this methodology in the classification of events in an urban environment of a highway in Brazil. We used Naive Bayes Classifier and different scores metrics with precision, F-Measure, and AUC.

The results obtained for these metrics vary significantly between the original database unbalanced and the database balanced by the SMOTE sampling technique. The accuracy of the classification in the first case was 44.0% and in the second case 84.5%. These values are confirmed in the literature and indicate that the use of the classification algorithm based on neural networks when used in conjunction with sampling techniques results in better classification performance. In this way, the rarer events in the database that correspond to the minor classes (casualties with fatal and non-fatal victims) are more adequately classified.

Considering the balanced database, it was verified that the variables most significant to the classification process of NI and WI accidents were the type of accident, probable cause, mileage, and track profile. The WI accidents were still influenced by the time of day.

The results suggest that NBM is adequate for predicting the severity of traffic accidents when compared to traditional techniques such as regression since it allows simultaneous analysis of the influence of all the variables involved in the database without previous restrictions being established.

However, the proposed approach is limited to the available data, since in developing countries such as Brazil there is still no standardization for the collection of this information. This particularity leads to the obtaining of incomplete and incoherent records, which implies that pre-processing techniques such as balancing are applied in the correction of these data.

To improve the performance of the proposed approach there is a need to complement the data, with the insertion of more instances and variables associated with the road characteristics of the stretch and the profile of the driver. In addition, it is recommended to be the performance of different Bayes classifiers and several networked approaches such as homogeneous and bipartite complex networks.

**REFERENCES**

Chen, C.; G. Zhang; H. Wang; J. Yang; P. J. Jin and C. M. Walton (2015) Bayesian network-based formulation and analysis for toll road utilization supported by traffic information provision. *Transportation Research Part C*, v. 60, p. 339–359.

Crone, S. F. and S. Finlay (2012) Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, v. 28, p. 224-239.

De Oña, J.; R. O. Mujalli and F. J. Calvo (2011) Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis and Prevention*, v. 43, n. 1, p. 402-411.

De Oña, J.; G. López; R. Mujalli and Calvo, F. J. (2013) Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis and Prevention*, v. 51, p. 1-10.

Deublein, M.; M. Schubert; B. T. Adey; J. Köhler and M. H. Faber (2013) Prediction of road accidents: A Bayesian hierarchical approach. *Accident Analysis and Prevention*, v. 51, p. 274-291.

Deublein, M.; M. Schubert and B. T. Adey (2014) Prediction of road accidents: comparison of two Bayesian methods. *Structure and Infrastructure Engineering Maintenance*, v. 10, n. 11, p. 1394-1416.

Fu, Y.; C. Li; T. H. Luan; Y. Zhang and G. Mao (2018) Infrastructure-cooperative algorithm for effective intersection collision avoidance. *Transportation Research Part C*, v. 89, p. 188-204.

Hosmer, D. W. and S. Lemeshow (2000). *Applied logistic regression* (2ª ed.). John Wiley and Sons, New York.

Japkowicz, N. (2000) The Class Imbalance Problem: Significance and Strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence*, Nevada, p. 111-117.

Khoo, H. L. and M. Ahmed (2018) Modeling of passengers' safety perception for buses on mountainous roads *Accident Analysis and Prevention*, v. 113, n. 1, p. 106-116.

Liang, C.; M. Ghazel and O. Cazier (2017) Risk analysis on level crossings using causal Bayesian network based approach based approach. *Transportation Research Procedia*, v. 25, p. 2167-2181.

Li, J.; S. Fong; R. K. Wong and V. W. Chu (2018) Adaptive multi-objective swarm fusion for imbalanced data classification. *Information Fusion*, v. 39, p. 1-24.

Lord, D. and F. Mannering (2010) The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A*, v. 44, n. 5, p. 291-305.

Mannering, F. L.; V. Shankar and C. R. Bhat (2016) Unobserved heterogeneity and the statistical analysis of highway accident data *Analytic Methods in Accident Research*, v. 11, p. 1-16.

Mujalli, R. O.; G. López and L. Garach (2016) Bayes classifiers for imbalanced traffic accidents datasets. *Accident Analysis and Prevention*, v. 88, p. 37-51.

Sain, H. and S. W. Purnami (2015) Combine sampling support vector machine for imbalanced data classification. *The Third Information Systems International Conference. Procedia Computer Science*, v. 72, p. 59 - 66.

Schlüter, P. J.; J. J. Deely and A. J. Nicholson (2018) Ranking and Selecting Motor Vehicle Accident Sites by Using a Hierarchical Bayesian Model. *Journal of the Royal Statistical Society*, v. 46, n. 3, p. 293-316.

Soro, W. L. and D. Wayoro (2017) A Bayesian analysis of the impact of post-crash care on road mortality in Sub-Saharan African countries. *IATSS Research*, v. 41, n. 3, p. 140–146.

Spiegelman, C.; E. S. Park and L. R. Rilett (2010) *Transportation Statistics and Microsimulation*. CRC Press.

Thammasiri, D.; D. Delen; P. Meesad and N. Kasap (2014) A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student. *Expert Systems with Applications*, v. 41, p. 321-330.

Wach, W. (2016) Calculation reliability in vehicle accident reconstruction. *Forensic Science International*, v. 263, p. 27-38.

Wang, J.; Y. Zheng; X. Li; C. Y; K. Kodaka and K. Li (2015) Driving risk assessment using near-crash database through data mining of tree-based model. *Accident Analysis and Prevention*, v. 84, p. 54-64.

Weiss, G. M. (2004) Mining with Rarity: A Unifying Framework. *SIGKDD Explorations*, v. 6, p. 7-19.

Zhao, J. and W. Deng (2015) The use of Bayesian network in analysis of urban intersection crashes in China. Transport, v. 30, n. 4, p. 411-420.

Zhu, X.; Y. Yuan; X. Hu; Y. Chiu and Y. Ma (2017) Bayesian Network model for contextual versus non-contextual driving behavior assessment. *Transportation Research Part C*, v. 81, p. 172-187.

Zong, F.; H. Xu and H. Zhang (2013) Prediction for traffic accident severity: comparing the Bayesian Network and Regression Models. *Mathematical Problems in Engineering*, v. 2013, p. 1-9.

Zou, Y.; X. Zhong; J. Ash; Z. Zeng; Y. Wang; Y. Hao and Y. Peng (2017) Developing a Clustering-Based Empirical Bayes Analysis Method for Hotspot Identification. *Journal of Advanced Transportation*, v. 2017, p. 1-9.