Predicting Highway Accident Severity in Brazil: Environmental Factors and Vehicle Features

Júlio C. W. Scholz
Informatics
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
julioscholz@alunos.utfpr.edu.br

Yan P. Pinheiro
Informatics
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
yanp@alunos.utfpr.edu.br

Thiago H. Silva
Informatics
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
thiagoh@utfpr.edu.br

***Abstract*—This work supports the problematization of the high rate of traffic accidents due to the increased use of automobiles for utility or transport in Brazil. Every year, the lives of approximately 1.3 million people are interrupted due to traffic accidents worldwide. In this regard, this study proposes the use of data mining as an approach to analyze and explore datasets of traffic accidents that occurred on Brazilian highways between the years 2017 and 2022, as provided by the Federal Highway Police. Additionally, vehicle price data were included, allowing for a more comprehensive analysis that also considers the financial value of the vehicles. The goal is to assess the predictive capability of classification models regarding the severity of accidents, focusing on vehicle characteristics and environmental factors. By applying classification algorithms and machine learning explainability techniques, we acquired relevant knowledge regarding the studied data, contributing to understanding and preventing accidents. As a result, the attributes related to vehicle characteristics had a more positive impact on the predictive capability of the models when compared to the attributes describing the environment and other variables.**

***Index Terms*—road accidents, data mining, classification, explicability, Shapley values**

I. Introduction

The adoption of automobiles as technical means and modes of transportation significantly impacts the organization of society, as automobiles play a significant role in shaping the structure of urban environments [1]. A substantial part of industrial development and urban planning has been directed towards establishing the automotive system. Thus, having become a social necessity, automobiles constitute a foundational part of contemporary life, entailing various consequences.

The intensive incorporation of automobiles into daily life brings significant benefits such as agility in human and cargo transportation, comfort, and a reduction in travel time, for example. However, there is also a harmful side to this reality. Regarding the drawbacks, it is possible to highlight noise, environmental pollution, and a more severe problem: deaths caused by traffic accidents.

It is public knowledge that traffic accidents are a frequent problem in society. Every year, the lives of approximately 1.3 million people are terminated abruptly as a result of a traffic accident. Between 20 and 50 million people suffer non-fatal injuries, with many incurring disabilities as a result of their injuries [2]. The World Health Organization classifies traffic accidents as the leading cause of death for children and young people between 5 and 29 years old. According to the National Registry of Traffic Accidents and Statistics (RENAEST) [3] 2022, there were approximately 1 million accidents in Brazil, involving just over 1.5 million people, resulting in 20,856 deaths.

Being a global problem, the United Nations General Assembly has determined to achieve a 50% reduction in the worldwide number of deaths and injuries due to traffic accidents [2]. Brazil is one of the signatories of this global effort and has proposed adopting a series of policies and measures, such as preventive campaigns, reduction of speed limits, infrastructure improvements, etc. [4]. However, few efforts have been made to reach this goal. According to the Institute of Applied Economic Research[1] (IPEA), traffic accidents in the country are estimated to cost an average of R$ 130 billion, thus revealing that the high number of traffic accidents remains a reality.

Therefore, initiatives to reduce and prevent traffic accidents are more necessary than ever. In this sense, Urban Computing [5] could play an important role, as suggested by a vast collection of studies addressing analysis and data mining of accidents in recent years [6]–[10]. This type of study is essential to help understand the risk factors and predict characteristics of serious accidents, which could enable the development and implementation of highly effective preventive measures.

In Brazil, the Federal Highway Police (PRF) has been collecting accident data since 2007 and makes this data available through the open data section on the government's website regarding accidents that occurred on federal highways in all states. This dataset includes occurrences recorded from 2007 to the present day. During this period, it is possible to observe that there was no significant reduction in the number of deaths and an insufficient reduction in the number of injured, even with a decrease in the number of people involved in accidents.

This study enhances the PRF dataset by adding attributes related to vehicle specifications. Then, data classification algorithms are applied to measure the impact of external factors, the environment, and attributes related to the vehicle on the

[1] https://repositorio.ipea.gov.br/handle/11058/10611

severity of accidents. Evidence was found, for instance, that vehicle attributes should not be neglected in models predicting accident severity. The findings of this study also offer valuable insights into the causes of accidents so that public and private entities can use this information to help reduce the severity and number of accidents on Brazilian highways.

The rest of the study is organized as follows. Section II presents the related work. Section III introduces data and methods explored in this work. Section IV presents the results followed by the conclusions (Section V).

II. Related Work

Urban computing is an interdisciplinary area that studies urban issues using state-of-the-art technologies and digital data [5], including, for instance, social media data [11]–[13] and open data [14]–[16]. Specifically related to traffic issues, applying data mining techniques to complex urban traffic accident data facilitates the discovery of non-trivial patterns and relationships [6]. Thus, urban computing in traffic safety research is helping to generate new ideas and hypotheses [17].

In this direction, Yap et al. [6] aimed to identify and categorize aspects of traffic accidents based on the characteristics of risk factors and classify them depending on the level of injury severity from an accident. The database explored consists of accident records from a state in the United States from 2004 to 2018. The authors explored a Decision Tree classifier method to predict the injury severity level (i.e., no injuries, minor injuries, serious injuries, and death). The primary variable that predicts injury severity in an accident is whether motorcycle-type vehicles were involved, resulting in the possibility of accidents with more serious injuries. It was also discovered that the involvement of a pedestrian in an accident significantly increases the probability of injury. However, geographic and environmental factors were shown to be less significant in the model's prediction. The model obtained an accuracy of 65.78%.

Labib et al. [7] classify the severity degree of the nearly 43,000 traffic accidents in Bangladesh from 2001 to 2015. First, they explored different classifiers to predict the severity of accidents (i.e., Fatal, Serious, Minor Injury, and Motor Collision). Naive Bayes classifier and adaptive stimulus algorithms showed the highest accuracy, about 80% of accuracy. In a second experiment, only the Fatal and Serious accident classes were used for classification. The vehicle type variables and the time of accidents were the most important in the predictive capacity of the models.

Kwon et al. [8] propose using a Naive Bayes classifier and a Decision Tree to identify the relative importance of accident risk factors concerning the injury severity level. The database used consists of accident data collected by the California Highway Patrol, focusing on accident reports that occurred on California highways during the period from 2004 to 2010, as only in 2004 did they begin to record attributes related to the characteristics of the vehicles involved, the type of highway, the date and time of the accident, weather conditions, and the type of accident. As a result, it was discovered that in the dataset under study, when dependency is taken into account, the most important factors are: type of collision, fault, local population, state highway, and movement before the collision.

In the study by Zhang et al. [9], four different machine learning models and two statistical models were compared in correctly categorizing the severity level of injuries in accidents. They explored a dataset composed of accidents that occurred on divergent highways in Florida, USA. They found that the model generated by the Random Forest method achieved the highest overall prediction accuracy in the test set (53.9%), and all other machine learning models were more accurate than the statistical methods (Ordered Probit Model and Multinomial Logit Model). The models have a lower capacity to accurately predict more severe accidents.

Ahmed et al. [10] evaluated different machine learning models to predict the severity of road accidents based on a dataset of road accidents in New Zealand from 2016 to 2020. Furthermore, the predicted results were analyzed, and an explainable machine learning (XML) technique was applied to assess the importance of factors contributing to accidents. To predict road accidents with different injury severities, different algorithms were considered, such as Random Forest (RF), Decision Jungle (DJ), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and Categorical Boosting (CatBoost). The comparison results showed that the RF model achieved the best classification, with 81.45% accuracy and 81.04% F1-Score. The results showed that the road category and the number of vehicles involved in an accident significantly impact the severity of injuries. The characteristics identified as most relevant through SHAP analysis were used to retrain the ML models and measure their performance. The results showed an increase of 6%, 5%, and 8%, respectively, in the performance of the DJ, AdaBoost, and CatBoost models.

Our research distinguishes itself from prior studies by adopting an incremental approach divided into three scenarios. In each scenario, new characteristics are incorporated, and the performance of the models is evaluated and compared. This strategy allows us to monitor the evolution and impact of each attribute in detail across the scenarios. In this way, we enrich the model with a broader range of variables related to vehicles, thereby providing a deeper insight into how these specific characteristics can influence the outcomes of traffic accident severity predictions.

To better understand the impact of each attribute on the final scenario, we applied the Shapley values technique. By integrating Shapley values into our analysis, we better understood how different attributes weigh on the prediction outcomes. This reinforces the comprehensiveness of our study and enhances the precision of our insights into the dynamics influencing traffic accident severity predictions.

III. Data and Methods

A. Data studied

The dataset utilized in this paper comprises records of road accidents reported by the police on federal highways in Brazil

from 2017 to 2022. We call this dataset PRF dataset. These data are publicly available through the Federal Highway Police portal[2]. This dataset includes a wide range of information regarding each accident, such as the date, time, location, severity, and contributing factors. Additionally, the dataset contains detailed information such as the type of road, road layout, whether the accident occurred in an urban or rural environment, weather conditions, and the type of individual involved in the accident (driver, passenger, or pedestrian). It also includes details about the brand, model, and year of manufacture of the vehicles involved and the type of vehicle (motorcycle, truck, car, etc.).

In order to enhance the analysis of vehicle attributes and provide a more comprehensive understanding of accident characteristics, we sought to acquire additional data from the FIPE Table[3]. The FIPE Table, an acronym for *Fundação Instituto de Pesquisas Econômicas* (Foundation Institute of Economic Research), is a crucial resource for evaluating average vehicle prices in the Brazilian automotive market. By incorporating this supplementary data in the PRF dataset, we aim to enrich our analysis and gain deeper insights into the relationships between vehicle characteristics and accident outcomes.

B. Preprocessing

Initially, the PRF dataset comprised 985,220 records. A data preprocessing stage was initiated to conduct the desired analysis, focusing on removing irrelevant records. Exclusions were made for entries lacking information on the victim's physical state and those categorizing individuals involved in the accident as "Witness" or "Horse Rider" (individuals using animals for transport). The discovery of null values in attributes such as age, gender, brand, and model of vehicles led to the removal of these data instances, culminating in a refined dataset of 862,465 records.

The preprocessing revealed that the brand and model variables frequently suffered from typing errors and incomplete information, primarily because these details were merged into a single "brand/model" column in the PRF dataset. The absence or incorrect placement of separator characters was a common issue, though many records could be corrected. For extreme cases, the solution was to categorize the brand and model as "Others." Typographical errors were another prevalent problem, with common mistakes being addressed for ease of identification.

Following the brand and model data cleanup, the next phase involved integrating vehicle price data from the FIPE table. This effort enabled assigning prices to 557,429 records, leveraging the brand, model, and year of manufacture for each entry. The remaining 305,036 records, lacking direct price information, had their values estimated by grouping them by vehicle type and year of manufacture to calculate an average price, which was then used to infer the missing prices.

Data encoding emerged as a critical step in the preprocessing of categorical data, essential for subsequent analysis and machine learning modeling. Label Encoding, which assigns sequential integer numbers to each category, introduces a risk of artificial ordering and potential bias, as models might misconstrue the categories as having inherent order or magnitude. To counteract this, creating dummy variables transforms each category into a binary variable, representing the presence or absence of categories and thus mitigating artificial ordering issues at the cost of increased data dimensionality. For cyclical variables like time of day, sine-cosine transformations are applied to preserve their cyclical nature continuously, enabling more accurate modeling of cyclical patterns without distortion.

The study established three distinct scenarios for training and testing to compare factors influencing the severity of accidents. This approach allows for exploring various variables and characteristics pertinent to accidents, facilitating a thorough analysis and the identification of potential patterns or correlations. This multifaceted methodology provides a comprehensive understanding of the factors contributing to accident severity.

The first scenario, *Scenario 1 - Base*, focuses on variables unrelated to vehicles or the environment, providing a general overview of highway accidents in Brazil. The insights gained from models trained on these characteristics serve as a benchmark for evaluating the impact of additional variables introduced in subsequent scenarios.

The second scenario, *Scenario 2 - Environment*, expands upon the first by incorporating environmental variables related to the accident site, such as geographical data, track features, and weather conditions. This allows for an assessment of how these additional factors influence model predictability.

The third scenario, *Scenario 3 - Vehicles*, builds on the second by including variables related to the vehicles involved in the accidents. This comprehensive analysis considers both environmental characteristics and specific vehicle attributes, offering deeper insights into the myriad factors affecting accident severity and their interrelations.

C. Models and Performance Evaluation

In this study, we utilized a diverse set of five classification algorithms:

- Tree-based models: Decision Tree (DT) and Random Forest (RF)
- A proximity-based algorithm: K-Nearest Neighbors (KNN).
- A Statistics-based approach: Naive Bayes (NB)
- And a Neural network architecture: Multilayer Perceptron (MLP)

In each scenario, the five machine learning algorithms mentioned were evaluated based on their ability to accurately predict whether an accident would be severe or not. As the dataset under analysis is imbalanced – meaning some classes have a significantly larger number of examples than others, in this case, 85.68% of the records are of minor accidents, and only 14.31% are of severe accidents – relying solely on accuracy as an evaluation metric is inadequate. This is because machine learning models may become biased toward

[2]https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos.

[3]https://veiculos.fipe.org.br.

the dominant class, resulting in a decreased ability to correctly predict the minority class.

To assess and compare the classification models, the following metrics were used:

- Accuracy: Measures the proportion of examples correctly classified relative to the total number of examples. It represents the overall precision of the model in correctly classifying the data.
- Average F1-Score: Computes the arithmetic mean of the F1-score values for each class individually. The F1-score combines precision and recall, providing a balanced measure of the model's performance for all classes.
- Geometric Mean (G-mean): Captures the model's overall performance by considering the true positive rate of all classes in a balanced manner. This metric provides a more balanced evaluation of the model, especially in imbalanced datasets.

We used the k-fold technique to train and test the algorithms with k = 5. This approach is used to evaluate the performance of machine learning models. This technique divides the dataset into k equal parts, called folds. The model is trained k times, where in each iteration, one of the folds is used as the test set, and the remaining folds are used as the training set. At the end, we obtain k performance measures, usually the mean, which can be used to evaluate the model's performance. The k-fold technique allows for a more robust and realistic evaluation of the model, as it uses all the data for training and testing, helping to reduce the variance of the results and providing a more accurate estimate of the model's performance on unseen data.

The Randomized Search in Hyperparameters technique selected the ideal parameters for each classifier. This is an efficient and automated approach for finding the ideal hyperparameters in machine learning models. Unlike an exhaustive search that examines all possible combinations of hyperparameters, Randomized SearchCV performs a random search within a predefined space of hyperparameters. The Randomized SearchCV technique helps to find a suitable set of hyperparameters to maximize the model's performance, resulting in better prediction and fitting results.

IV. Results

A. Scenario 1 - Base

Table I displays the performance of the classification models in Scenario 1 - Base. It is observed that the data is imbalanced, as the average accuracy is high (84.6%), while the average F1-score and geometric mean are low in comparison (51.6%). Tree-based models showed better performance according to the F1-score and G-mean. Other algorithms yielded similar results among themselves.

The importance of permutation variables for Scenario 1 is represented in Figure 1. It can be observed that the time of day was the most important variable for predicting whether an accident would be severe or not. It is important to note that the sine-cosine transformation applied to the time variable is not ideal for tree-based models because they make splits based on one attribute at a time, and sine/cosine attributes should be considered simultaneously to correctly identify points in time within a period. Since one piece of information is represented in two features, mathematically, more weight will be attributed to it from the algorithm's point of view.

TABLE I
Performance of classification models in Scenario 1-Base

Models	Accuracy	F1-score	G-mean
DT	82%	52%	31%
RF	83%	52%	30%
NB	86%	52%	25%
MLP	86%	51%	24%
KNN	86%	51%	23%

However, the importance of the time variable is much higher than all other variables, indicating that there may be a relationship between accident severity and the time it occurs. The second most important variable is the age of the injured person, and the variable that had the most negligible impact on the model's performance indicates whether the accident occurred on a holiday.

B. Scenario 2 - Environment

When analyzing the performance metrics contained in Table II, which refer to Scenario2-Environment, it is noticeable that the Decision Tree differs from the other models. Specifically, its accuracy experienced a reduction of approximately 7.3% compared to Scenario 1; however, metrics considering the imbalance of the target variables increased, mainly the geometric mean. This suggests a slight improvement in the decision tree's ability to identify severe accidents compared to the same model in the previous scenario.

TABLE II
Performance of classification models in Scenario 2 - Environment

Models	Accuracy	F1-score	G-mean
DT	76%	54%	44%
RF	85%	54%	32%
NB	86%	52%	26%
MLP	86%	51%	23%
KNN	85%	52%	26%

On the other hand, the other models did not achieve significant results, especially considering the average F1-score, which only increased by 2% for the Random Forest and Decision Tree models, and for the NB, MLP, and KNN classification algorithms, there was no noticeable difference, indicating a difficulty for these models in capturing possible relationships in the dataset.

Figure 2 shows the same pattern presented in Scenario 1-Base, where the variables "time," "age," and "day of the week" occupy the top positions in importance, was repeated in this second scenario; however, the six new features related

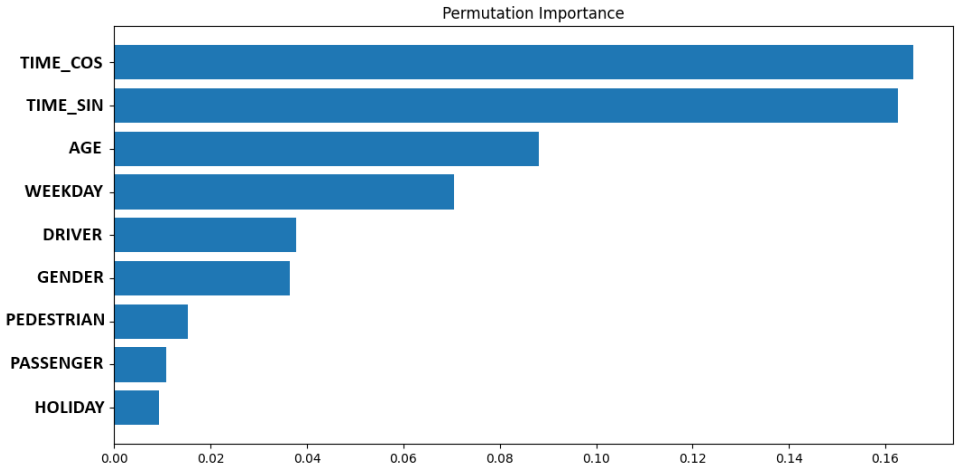

Fig. 1. Permutation Importance for the Decision Tree in Scenario 1-Base

to the environment occupied the next six most important positions. Especially the region and road type were the ones that contributed the most to the predictive capacity of the model; in contrast, the variable added in this scenario, which obtained the lowest importance value, was the ”time of day.”

C. Scenario 3 - Vehicles

Table III presents the results of the classification models for this scenario, demonstrating consistent improvements in the average F1-Score and Geometric Mean across all models. The Random Forest and Decision Tree models showed a considerable increase in predictive capacity, while regarding the average F1-score, the Naive Bayes model yielded the best result: 61%, indicating a percentage increase (relative variation) of 17.31% compared to scenarios 1 and 2. Figure 3 suggests that this increase may be attributed, in part, to the influence of attributes related to vehicles, with vehicle type being the most influential.

TABLE III
PERFORMANCE OF CLASSIFICATION MODELS IN SCENARIO 3-VEHICLES

Models	Accuracy	F1-score	G-mean
DT	79%	59%	52%
RF	86%	59%	41%
NB	84%	61%	49%
MLP	86%	55%	31%
KNN	85%	57%	37%

Broeck et al. [18] stated that calculating Shapley values for probabilistic models like Naive Bayes is impractical due to its complexity. Shapley values assess individual contributions of features in cooperative game theory scenarios. Since Naive Bayes calculates probabilities independently for each feature, the concept of Shapley values does not directly apply to it. Therefore, the focus here is to analyze the Decision Tree, the second-best model regarding the F1-score and G-mean metrics.

In Figure 4, it is possible to verify the distribution of importance based on SHAP values for the Decision Tree. The higher the SHAP value (further to the right on the X-axis), the more impactful it is in deciding whether an accident will be severe, and conversely, negative values indicate if an accident will be mild. The order of features on the Y-axis corresponds to each variable’s average absolute value of SHAP values. This order represents the average impact of each variable in deciding the severity of an accident. In this case, the ”vehicle type” variable was the most impactful on the model’s predictive capacity, while the variable indicating whether an accident occurred on a holiday had the least impact.

Upon analyzing the variable ”type of involved party in the accident,” it is observed that the model exhibits a high tendency to classify the accident as severe when it involves a pedestrian. Since the number of occurrences involving pedestrians accounts for only 1.57% of the total number of those involved in accidents, this explains why this variable did not assume a position of higher importance in the analysis of SHAP values.

Price and year of manufacture are the next variables in order of importance and are related to vehicles. In the scatter plot represented in Figure 11, it is possible to visualize the effect of the ”year of manufacture” variable on the predictions made by the model. It is common knowledge that newer vehicles usually have a higher monetary value, and this trend is evidenced in Figure 5. It is also noted that as the year of manufacture of vehicles decreases, there is a subtle reduction in SHAP values, suggesting that newer vehicles have a lesser contribution to predicting severe accidents and a more significant contribution to mild accidents.

In Figure 5, it is also possible to observe a concentration of data instances with a lower year of manufacture of the vehicle, presenting SHAP values above 0.2. This suggests that older vehicles play a significant role in predicting severe accidents. On the other hand, recently manufactured vehicles are more distributed in the SHAP value range between -0.2 and 0.2.

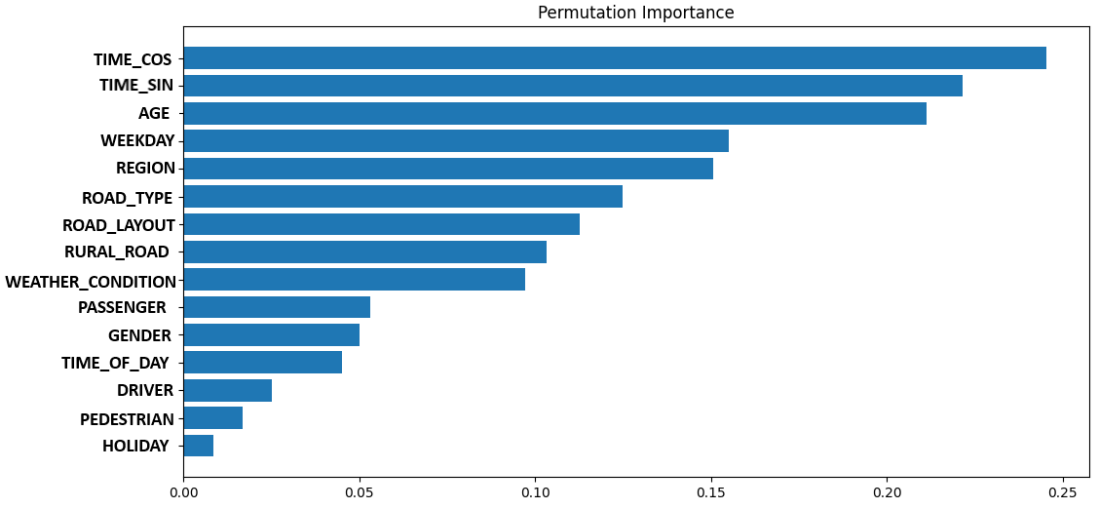

Fig. 2. Permutation Importance for the Decision Tree in Scenario 2-Environment

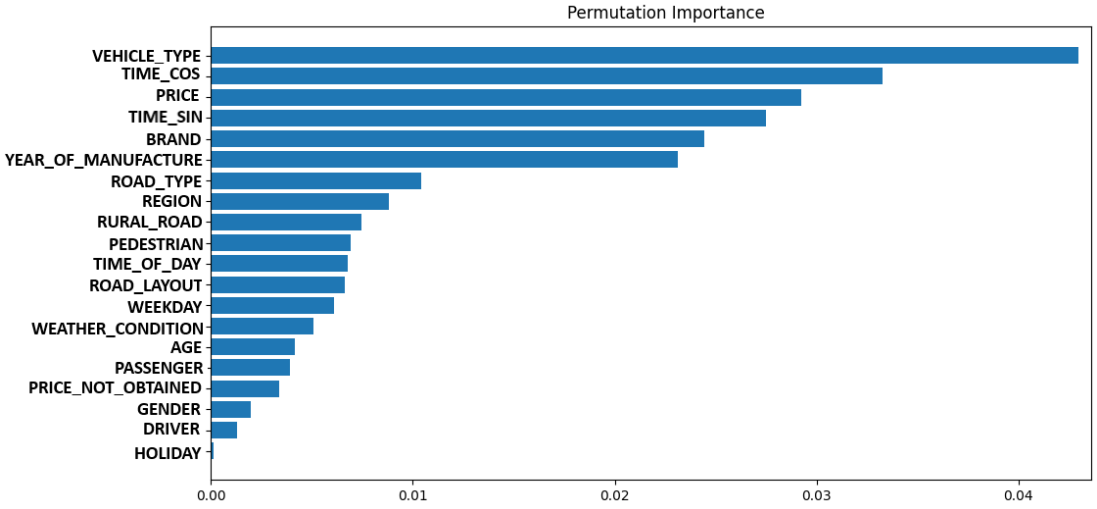

Fig. 3. Permutation Importance for the Decision Tree in Scenario 3-Vehicles

V. Final Discussions, Conclusions and Future Work

The incremental approach of variables by different scenarios evaluated in this study offered a deeper understanding of the impact of added variables on the models, allowing the identification of their relevance in prediction. Although this strategy may limit the analysis of interactions between variables, the scenarios were built progressively, without excluding variables.

Data imbalance was observed in all three evaluated scenarios, with an average accuracy of 84.07%, while the average F1-score and geometric mean were lower considering the imbalance. Decision Trees and Random Forests showed better performance in terms of F1-score and geometric mean. At the same time, the other algorithms had similar results, except for Naive Bayes, which stood out in Scenario 3-Vehicles with an average F1-score of 61%. All models were evaluated using k-fold and demonstrated consistency and stability across different folds, with a standard deviation (SD) of the mean F1-score of less than 0.5% in all scenarios, as indicated in Table IV.

TABLE IV
Standard deviation (SD) of F1-score across all scenarios

Scenarios	Models				
	DT	RF	NB	MLP	KNN
Scenario 1 F1-Score SD	0.19%	0.13%	0.45%	0.40%	0.39%
Scenario 2 F1-Score SD	0.056%	0.11%	0.48%	0.73%	0.21%
Scenario 3 F1-Score SD	0.06%	0.45%	0.28%	0.45%	0.22%

An analysis by scenarios conducted in this study, see a summarization in Table V, indicates that variables related to the environment (Scenario 2-Environment) contribute little to the

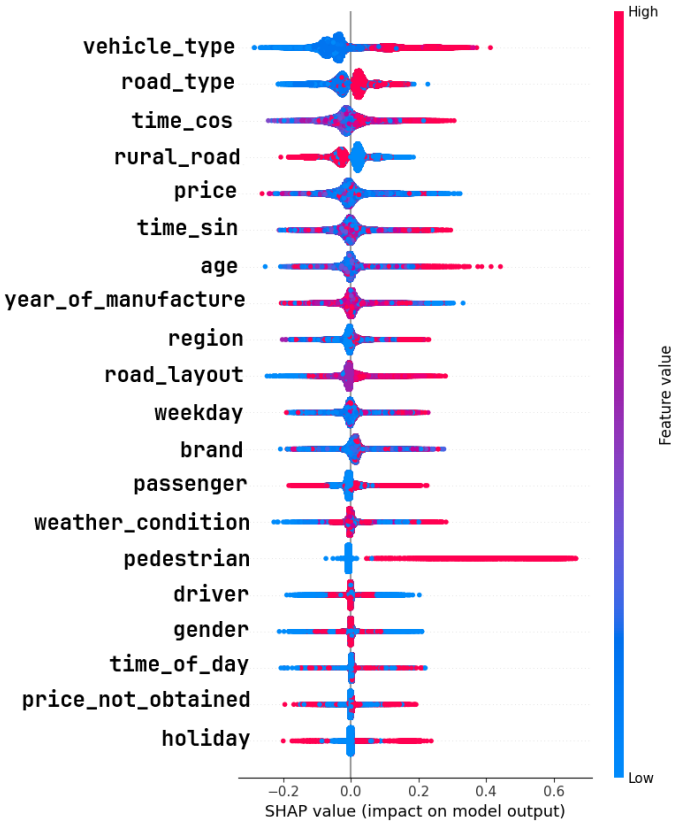

Fig. 4. Shapley Values for Decision Tree in Scenario 3-Vehicles

TABLE V
F1-SCORE OF THE CLASSIFICATION MODELS ACROSS ALL SCENARIOS

Models	F1-score		
	Scenario 1	Scenario 2	Scenario 3
DT	52%	54%	59%
RF	52%	54%	59%
NB	52%	52%	61%
MLP	51%	51%	55%
KNN	51%	52%	57%

predictive capacity when compared to vehicle characteristics added in Scenario 3-Vehicles. In this third scenario, all models showed better performance, indicating a relationship between these attributes and the severity of accidents. The Naive Bayes classifier achieved a significant increase in the F1-score, with a percentage increment of 17.31% compared to Scenarios 1 and 2. At the same time, the Random Forest and Decision Tree recorded a relative increase in F1-score of 13.46% compared to Scenario 1 and 9.26% compared to Scenario 2.

Table V also demonstrates that the K-Nearest Neighbors and Multilayer Perceptron classifiers failed to capture any importance from the environment variables added in Scenario 2, as their predictive capacity remained basically unchanged compared to the first scenario. Only with the addition of vehicle factors was the improvement in classification results achieved; however, KNN and MLP did not perform better than tree-based models (DT and RF) and the Naive Bayes probabilistic classifier.

By analyzing and comparing the results of the different scenarios, it is possible to identify which factors are most relevant for the prediction task. The ”vehicle type” variable was identified as the primary one for classifying the severity of an accident. The most important vehicle type for the predictive capacity of the DT is ”Motorcycle,” a result not exclusive to the dataset under study, as this pattern was also observed in the analysis conducted in the work of Yap et al. [6], which analyzed a dataset of accidents in the United States from 2004 to 2018.

It is evident from Figure 4 that the variables ”road type” and ”rural road” play a vital role in the classification capacity of the Decision Tree in Scenario 3-Vehicles. However, analyzing the metrics of Scenario 2-Environment and the importance graph, they are not as impactful, and this increase occurred only when vehicle-related attributes were added. Thus, such attributes should not be neglected, as the predictive capacity of the models increased considerably.

As evidenced earlier, for the ”rural road” factor, there is a predictive trend to classify severe accidents if they occur in a non-urban environment. This trend was also observed in a study analyzing a dataset of accidents that occurred in New Zealand [10]. In Brazil, considering that highways crossing urban areas are subjected to more safety restrictions than non-urban highways [19], such as speed limits and more frequent signage, this may be one of the reasons explaining the trend observed in the Decision Tree.

Initially, the improvement in metrics observed across scenarios may seem insignificant. However, it is important to emphasize that this improvement occurred precisely in metrics that take into account the imbalance present in the dataset. In other words, as new variables were added, the models were able to capture relevant patterns that contributed to the increase in predicting severe accidents, which represent the minority and most critical class. Although preventing accidents as a whole is an important goal, directing resources and efforts towards predicting and preventing severe accidents can bring substantial benefits to society, including reducing deaths, reducing both public and private costs, and overall improving road safety.

The methodology used demonstrated that the analysis divided into scenarios contributed to understanding each factor’s impact on accidents. Creating new features was fundamental in expanding the existing explanations of accident causes. Thus, future work can continue the analysis of occurrences on Brazilian highways, since there are numerous attributes related to these occurrences that were not present in the PRF dataset and can be collected and analyzed in the future. Considering that the geographical location where the accident occurred has some importance, an interesting possibility would be to conduct a more specific analysis of the regions of Brazil and use the Latitude and Longitude of the accident point, which are attributes present in the PRF database. Another possibility is to analyze the distribution of traffic signs and speed radars

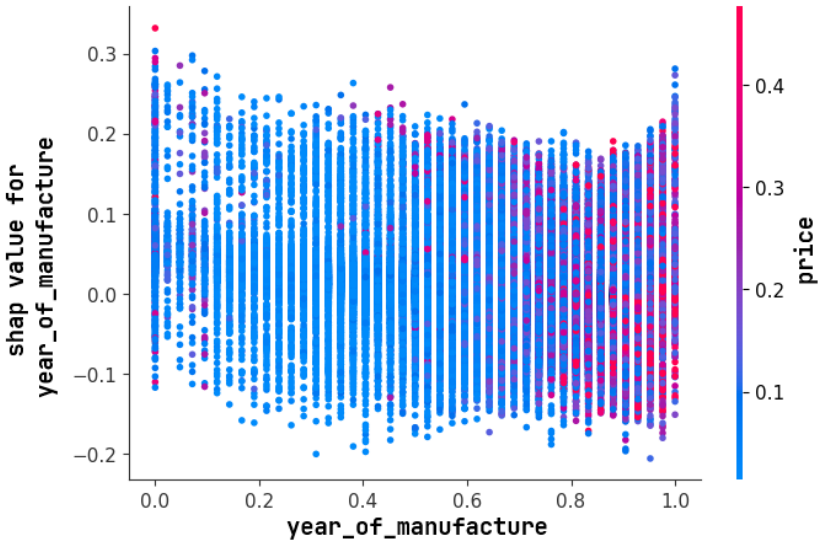

Fig. 5. Dependency analysis on the year of manufacture of the vehicle using SHAP

to determine if there's a correlation with accident sites, which could offer valuable insights.

The SHAP values explainability technique is precious and offers a visual explanation of the relationships in the model's predictive capacity. However, in this work, we restricted its use only to the Decision Tree due to the high computational cost and significant time required to perform the calculations. Future work may expand the use of this technique to explain other machine-learning models. Thus, it would be possible to obtain comprehensive and interpretable insights into the contribution of each variable in different prediction algorithms. This would allow for a deeper and more robust understanding of the factors influencing the predictive capacity of these models. More advanced data mining techniques, such as Deep Neural Networks, could be an interesting approach to analyzing highway accidents.

Acknowledgments

This research was partly supported by the project SocialNet (grant 2023/00148-0 from Sao Paulo Research Foundation - FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (process 314603/2023-9 and 441444/2023-7).

References

[1] T. Schor, "O automóvel e o desgaste social," *São Paulo em perspectiva*, vol. 13, pp. 107–116, 1999.
[2] W. H. O. WHO, "Road traffic injuries," Jun 2021.
[3] RENAEST, "Registro nacional de acidentes e estatísticas de trânsito," Apr 2022.
[4] I. d. P. E. A. IPEA, "Por uma agência nacional de prevenção e investigação de acidentes de transportes," 2021.
[5] T. H. Silva, A. C. Viana, F. Benevenuto, L. Villas, J. Salles, A. Loureiro, and D. Quercia, "Urban computing leveraging location-based social network data: A survey," *ACM Comput. Surv.*, vol. 52, pp. 17:1–17:39, Feb. 2019.
[6] L. Yap, H. N. Chua, Y. C. Low, M. Akmar, and A. Lee, "A data mining approach to analyse crash injury severity level," *Journal of Engineering Science and Technology*, pp. 1–14, 02 2022.
[7] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das, and F. Nawrine, "Road accident analysis and prediction of accident severity by using machine learning in bangladesh," in *2019 7th international conference on smart computing & communications (ICSCC)*, pp. 1–5, IEEE, 2019.
[8] O. H. Kwon, W. Rhee, and Y. Yoon, "Application of classification algorithms for analysis of road safety risk factor dependencies," *Accident Analysis & Prevention*, vol. 75, pp. 1–15, 2015.
[9] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing prediction performance for crash injury severity among various machine learning and statistical methods," *IEEE Access*, vol. 6, pp. 60079–60087, 2018.
[10] S. Ahmed, M. A. Hossain, S. K. Ray, M. M. I. Bhuiyan, and S. R. Sabuj, "A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance," *Transportation Research Interdisciplinary Perspectives*, vol. 19, p. 100814, 2023.
[11] S. A. de Brito, A. L. Baldykowski, S. A. Miczevski, and T. H. Silva, "Cheers to untappd! preferences for beer reflect cultural differences around the world," in *Proc. of Americas Conf. on Information Systems (AMCIS'18)*, (New Orleans, USA), 2018.
[12] A. P. G. Ferreira, T. H. Silva, and A. A. F. Loureiro, "Profiling the mobility of tourists exploring social sensing," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 522–529, 2019.
[13] F. A. Santos, T. H. Silva, A. A. F. Loureiro, , and L. A. Villas, "Automatic extraction of urban outdoor perception from geolocated free-texts," *Social Network Analysis and Mining*, vol. x, no. x, 2020.
[14] F. A. Santos, D. O. Rodrigues, T. H. Silva, A. A. F. Loureiro, R. W. Pazzi, and L. A. Villas, "Context-aware vehicle route recommendation platform: Exploring open and crowdsourced data," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–7, 2018.
[15] F. R. Gubert, P. Santin, M. Fonseca, A. Munaretto, and T. H. Silva, "On strategies to help reduce contamination on public transit: a multilayer network approach," *Applied Network Science*, vol. 8, no. 1, pp. 1–22, 2023.
[16] D. Silver and T. H. Silva, "A markov model of urban evolution: Neighbourhood change as a complex process," *PLOS ONE*, vol. 16, pp. 1–29, 01 2021.
[17] M. A. Raihan, M. Hossain, and T. Hasan, "Data mining in road crash analysis: the context of developing countries," *International journal of injury control and safety promotion*, vol. 25, no. 1, pp. 41–52, 2018.
[18] G. V. d. Broeck, A. Lykov, M. Schleich, and D. Suciu, "On the tractability of shap explanations," *Journal of Artificial Intelligence Research*, May 2021.
[19] C. L. d. Carmo and A. A. Raia Junior, "Segurança em rodovias inseridas em áreas urbanas na região sul do Brasil," *urbe. Revista Brasileira de Gestão Urbana*, vol. 11, p. e20170182, 2019.