Highway to... Determining Fatal Outcomes in Traffic Accidents Based on Police Reports

Arthur M. P. Gabardo*, Guilherme A. A. Schünemann, Pablo A. Jaskowiak, Benjamin G. Moreira, and Ricardo J. Pfitscher

Interdisciplinary Data and Artificial Intelligence Laboratory (IDA Lab)
Federal University of Santa Catarina (UFSC) - Joinville, SC

Abstract. Brazil faces significant traffic safety challenges with its vast territory and one of the world's largest road networks. Road traffic accidents, particularly on federal highways, remain a leading cause of death in the country, with serious economic and social consequences. This work presents a case study of three machine learning methods — Random Forest (RF), k-Nearest Neighbors (kNN), and Multilayer Perceptron (MLP) — for classifying the severity of traffic accidents in the Brazilian southern region. Using an open dataset from the Brazilian Federal Highway Police (PRF) covering the years 2021 to 2024, extensive preprocessing was carried out, including categorical variable encoding, feature selection, and the application of the SMOTE technique to address class imbalance. Model performance was assessed through statistical metrics such as specificity, $F1$-score, and AUC-ROC. The results show that RF and kNN (with SMOTE) achieved the best performance in predicting fatal accidents, both with AUC-ROC of 0.99. In addition to an in-depth model evaluation, this study presents a post-hoc analysis of feature importance and contributions through Shapley Additive Explanations (SHAP) for the best performing model, in order to support knowledge discovery and highlight the most influential factors associated with fatalities.

Keywords: Traffic accident · Road safety · Machine learning · Explainable AI.

1 Introduction

Brazil is a continental-sized country, with a territory exceeding 8.5 million square kilometers. It has the fourth-largest road network in the world, spanning 1.7 million kilometers. This vast network — composed of federal, state, and municipal roads, both public and concession-operated — is fundamental to the country's economic development, playing a crucial role in regional integration by enabling the transportation of people and goods. Such characteristics, associated with poor road conditions, contribute to a high incidence of traffic accidents [3].

According to the Brazilian National Confederation of Transport [5], traffic accidents in Brazil led to an economic cost of R$14 billion in 2024, with over 67,000

* corresponding author: `arthur.miguel@posgrad.ufsc.br`

incidents reported. The southern region stands out with particularly alarming figures, with the highest rates of both accidents and deaths per inhabitants in the entire country. In 2024 alone, the states in the southern region recorded over 21,000 accidents, 6.4% of which were fatal, also resulting in an economic impact exceeding R$4.3 billion. This represents 63.5 accidents (4.4 deaths) per 100,000 inhabitants, as depicted in Fig. 1, a rate far higher than the center-western region (CW) in accidents. Another alarming trend is observed when analyzing the evolution of traffic incidents in recent years. Between 2021 and 2024, Brazil has experienced a steady increase in both the total number of road accidents and the proportion involving fatalities. This trend is also observed in the southern region, with an increase in the number of reported incidents in the past years.

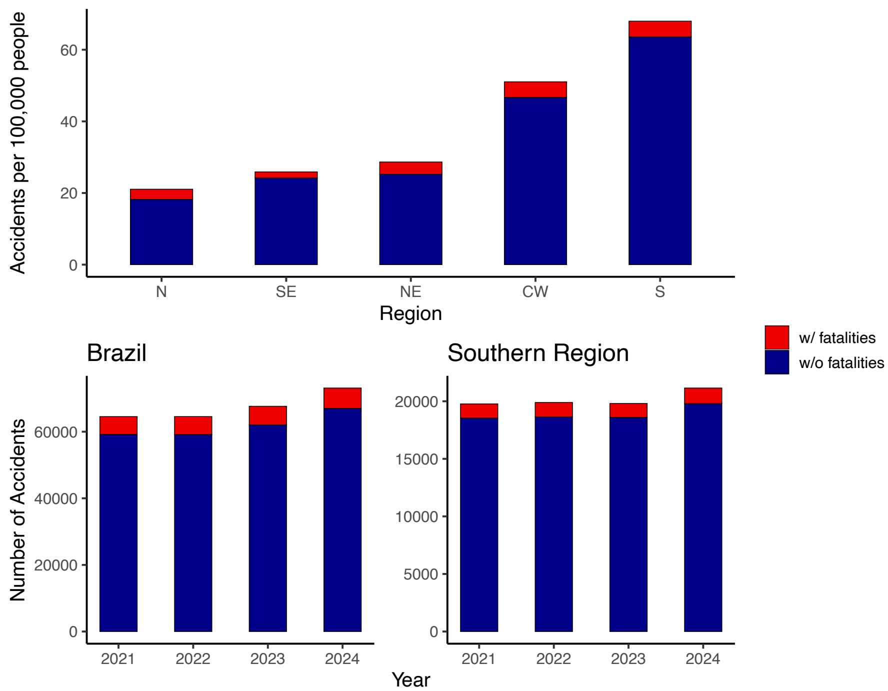

Fig. 1: Rate of accidents and deaths in Brazilian federal roads.

These numbers underscore the severity of the issue and the urgent need for public policies to improve road safety and consequently reduce fatalities. Given that the primary causes of accidents include human error, inadequate infrastructure, and ineffective enforcement [14], detailed analysis of traffic accident data is essential to an informed decision-making process regarding infrastructure improvements and to support targeted traffic safety campaigns. Machine learning and data mining techniques have gained prominence in this context, as they identify complex patterns and trends that may not be easily captured through traditional analytical approaches [2].

This study presents a comparative evaluation of different classification models to predict the severity of traffic accidents on federal highways in the Brazilian southern region, using an open-access dataset provided by the Brazilian Federal Highway Police [4]. The comparison comprises three algorithms — k-Nearest Neighbors (kNN), Random Forest (RF), and Multilayer Perceptron (MLP) — assessed through statistical performance metrics such as $F1$-score and AUC-ROC. The analysis also takes into account a grid search for hyperparameter tuning, and oversampling techniques (i.e., SMOTE) to address the class imbalance in the target variable. A post-hoc feature importance analysis is also conducted on the best-performing models to support knowledge discovery and interpretability, highlighting the most influential variables associated with accident severity.

2 Related Works

Given the scale and multivariate nature of road traffic accident data, traditional statistical methods often struggle in capturing complex patterns and nonlinear interactions among various features. Conversely, machine learning models have proven effective not only in improving predictive performance but also in supporting the extraction of meaningful knowledge from large-scale datasets. Kumeda et al. [13] demonstrated that classification algorithms, such as RF and MLP, can successfully model accident severity by leveraging road, meteorological, vehicles, and involved persons information, achieving accuracies as high as 85%. Similarly, Gan et al. [9] also argued that the severity of traffic accidents is rarely determined by linear associations alone and, therefore employed methods such as Random Forests and kNN to uncover complex data relationships. The authors reported accuracies close to 93%. Infante et al. [10], in a study focused on Portuguese traffic data, compared statistical approaches to ML classifiers, revealing that while ML often outperforms statistical models in balanced datasets, their advantage becomes limited in highly imbalanced data, outlining a key characteristic in accident severity classification problems.

Indeed, class imbalance remains one of the main challenges in this domain, especially when severe or fatal accident cases are underrepresented in historical records. Various studies have sought to mitigate this issue using resampling strategies. Jeong et al. [11] compared under-sampling and over-sampling techniques in combination with ensemble methods such as bagging and majority voting, and found out that decision trees, paired with over-sampling achieved the best performing results. Cuenca et al. [7] also addressed the disproportion between non-severe and severe accidents using under-sampling to balance the training data, subsequently evaluating classification performance using Naive Bayes, Deep Neural Networks (DNN), and Gradient Boosting Trees in the context of decision-making for Spanish traffic agencies, achieving a $F1$-score of 87% with DNN as the best performing model. More related to our work, Mokoatle et al. [16] employed the Synthetic Minority Over-sampling Technique (SMOTE) to improve the representation of minority classes in South African accident data. Their results confirmed the effectiveness of SMOTE in enhancing the performance of XGBoost over logistic regression, especially when dealing with sparse and noisy data from police reports. These findings reinforce the importance of handling class imbalance to prevent biased predictions and ensure better recall of rare but critical cases such as fatalities.

In addition to predictive accuracy, recent studies have stressed the importance of interpretability and feature importance analysis as part of knowledge discovery in intelligent accident analysis systems. Transparent models and post-hoc explanation methods enable practitioners to understand not only what the model predicts, but why. Dong et al. [8] applied Shapley Additive Explanations (SHAP) to interpret boosting-based ensemble models, revealing that month of the year, accident cause, driver age, and collision type were significant determinants of injury outcomes in highway crashes in Pakistan. Likewise, Ahmed et al. [1] employed SHAP to a RF model trained on New Zealand crash data, revealing that road type and vehicle count were key predictors of injury severity.

3 Materials and Methods

The main goal of this work is to classify the severity of traffic accidents — distinguishing between fatal and non-fatal cases — through a comprehensive analysis of accident data collected by the Brazilian Federal Highway Police (PRF) from 2021 to 2024 in the southern region of Brazil. This section describes the procedures and decisions involved in the data collection, preprocessing, feature engineering, evaluation of models, and feature importance assessment.

3.1 Data Collection

The dataset used in this study comprises official traffic accident records obtained from the PRF's publicly available databases. These records include structured reports of accidents that occurred from 2021 to 2024 in all Brazilian federal roads. The choice of this time frame is justified by a methodological change in accident reporting adopted nationwide in 2021, which standardized the annotation of key variables and improved data reliability across years. Although the dataset covers the entire country, the study focused on the southern region of Brazil because of two main factors: (i) the South consistently records the highest rates of traffic accidents and fatalities per capita among all Brazilian regions; and (ii) the country's road infrastructure varies significantly across regions — with notable differences in road quality, traffic density, and urban-rural distribution — making region-specific analyses essential to accurately capture such heterogeneity.

3.2 Data Preprocessing

Initial preprocessing involved merging the datasets from each year into a single consolidated collection. Records with inconsistent or implausible values were removed, such as vehicle manufacturing years marked as zero or negative, and ages of involved individuals below zero or above 130. Furthermore, entries containing missing or unreported values in any of the essential features were excluded to ensure the integrity of subsequent analyses.

Each accident record contains various features, including road and weather conditions, time of occurrence, types of vehicles involved, accident typology, and demographic data of those affected. Table 1 describes all features in the dataset. From this initial set of features, categorical features were standardized to reduce variability caused by spelling inconsistencies or semantically equivalent labels. For instance, the number of categories in the `type_vehicle` feature was reduced from 15 to 6, and in the `type_accident` feature from 17 to 9. Along the same line, the features `id_road` and `km_road` were concatenated into a single locality identifier named `br_km`.

Table 1: Features of occurrences present in the dataset.

Feature	Type	Description
`severity`	target	Fatal or non-fatal accident
`type_accident`	categorical	Type of accident (e.g., frontal collision, rollover)
`type_involved`	categorical	Whether involved was a passenger or driver
`age`	numerical	Age of the person involved
`gender`	categorical	Gender of the person: male, female, or undefined
`type_vehicle`	categorical	Type of vehicle (e.g., car, bus, motorcycle)
`year_vehicle`	numerical	Year the vehicle was manufactured
`day_week`	categorical	Day of the week when the accident occurred
`day_period`	categorical	Period of the day (e.g., morning, night)
`met_cond`	categorical	Weather conditions (e.g., sunny, raining, windy)
`type_road`	categorical	Type of road (e.g., single, double, or multilane)
`out_road`	categorical	Whether the vehicle went off the road
`id_road`	categorical	Road identifier
`km_road`	numerical	Road segment (kilometer marker)
`direction_road`	categorical	Traffic direction of the road

Due to the high cardinality of the categorical features, we applied a k-fold target encoding approach using five distinct folds to encode each category with the average value of the target variable in the training set (i.e., the remaining four folds), thus preserving meaningful patterns and avoiding data leakage and overfitting [17]. Subsequently, we centered and scaled all features to ensure a balanced contribution to model training. Finally, due to the highly imbalanced nature of the reports of fatal accidents compared to overall accidents, the Synthetic Minority Oversampling Technique (SMOTE), proposed by Chawla et al. [6], was employed to synthetically increase the representation of fatal accidents, preserving the distribution of objects features. It is important to highlight that SMOTE was applied exclusively in the training sets.

3.3 Model Selection and Validation

This study evaluates three distinct models to predict accident severity: k-nearest Neighbors (kNN), Random Forest (RF), and Multilayer Perceptron (MLP). We performed a hyperparameter grid search for each model to identify the optimal configuration (Table 2).The full dataset was initially partitioned into 80% for training and 20% for final testing. To assess model performance and mitigate overfitting, we also evaluated the training set through Monte Carlo cross-validation with five random resampling iterations, each using 80% of the training data for model fitting and 20% for internal validation [12]. This resampling strategy corresponds to the Leave-Group-Out Cross-Validation (LGOCV) method implemented in the caret package[1], without specifying grouping factors.

Table 2: Hyperparameters used for grid search tuning.

Model	**Hyperparameter**	**Interval**
kNN	k	$[3, 5, 7, ..., 15]$
RF	mtry; split rule; min. node size	$[1, 2, ..., 12]$; $\{gini,\ extratrees\}$; $[1, 5, 10]$
MLP	size; decay	$[11, 19, 27, ..., 115]$; $10^{[-4,-3,...,-1]}$

A comprehensive set of classification metrics was applied to evaluate each model's performance, particularly those that reflect the model's ability to correctly identify fatal accidents — the positive class in this study. Given the highly imbalanced nature of the dataset, relying solely on overall accuracy would provide a misleading assessment of model effectiveness. Therefore, additional metrics were considered to capture the model's discriminatory power across both classes. Sensitivity (or recall) allows us to measure the proportion of correctly identified fatal accidents, while specificity quantifies the model's ability to classify non-fatal cases correctly. The $F1$-score is the harmonic mean of precision and recall, providing a balanced view of the model's performance concerning the positive class. Finally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) allows for the assessment of the model's ability to distinguish between classes across varying decision thresholds.

3.4 Feature Importance

This study includes a feature importance analysis on the best-performing model through the Shapley Additive Explanations (SHAP) method. SHAP is a model-agnostic interpretability technique grounded in cooperative game theory, introduced by Shapley [18] and later adapted to machine learning contexts by Lundberg and Lee [15]. The SHAP value of a given feature represents its marginal contribution to the prediction outcome, calculated by comparing the model's output with and without the feature across all possible combinations of feature subsets (coalitions). This approach provides a consistent and locally accurate attribution of importance to each feature, enabling a more transparent understanding of the model's decision-making process.

[1] caret package – `https://cran.r-project.org/web/packages/caret/caret.pdf`

4 Results and Discussion

This section presents the classification results, including the effects of data oversampling, the consistency of feature distributions after preprocessing, the evaluated models' comparative performance, and the best-performing model's interpretability through Shapley value analysis. The analysis includes a particular emphasis on understanding the role of SMOTE in enhancing sensitivity towards the positive class (fatal accidents), as well as on identifying the most influential factors associated with fatal outcomes in road traffic accidents.

4.1 Oversampling

Applying the Synthetic Minority Oversampling Technique effectively addresses the class imbalance in the dataset, increasing the proportion of fatal to non-fatal accidents from 7.1% to 49.8%. As shown in Fig. 2, where features are encoded, centered, and scaled, the distribution of key features remained consistent before and after oversampling. Although the figure only depicts a subset of features (due to space constraints), the other variables show similar patterns. This consistency is justified by SMOTE's method to interpolate new samples based on existing minority class instances rather than introducing random, artificial distortions.

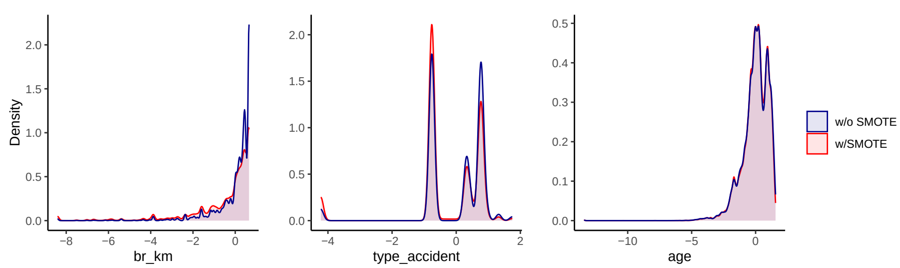

Fig. 2: Feature distributions before and after oversampling.

4.2 Model Performance

Table 3 summarizes the performance of the three evaluated models using the test set. Without oversampling, kNN and MLP exhibited limited ability to detect fatal accidents, with sensitivity values of only 19% and 14%, respectively. Such a result confirms the challenge posed by the imbalanced dataset, in which models tend to favor the majority class (non-fatal accidents), achieving deceptively high accuracy and specificity at the expense of recall. After applying SMOTE, both kNN and MLP showed significant improvements in sensitivity — reaching 100% and 71%, respectively — though with a corresponding drop in specificity and overall accuracy, particularly for MLP.

Table 3: Performance metrics of all models in test dataset.

Model	Accuracy	Sensitivity	Specificity	AUC-ROC	F1-Score
w/o SMOTE					
kNN	0.94	0.19	0.99	0.92	0.30
RF	0.99	0.99	1.00	0.99	0.99
MLP	0.93	0.14	0.99	0.82	0.23
w/ SMOTE					
kNN	0.86	1.00	0.85	0.99	0.50
RF	0.99	0.99	0.99	0.99	0.99
MLP	0.77	0.71	0.77	0.82	0.29

In contrast, the Random Forest model demonstrated exceptional performance across all evaluation scenarios. Even without oversampling, RF achieved near-perfect scores in accuracy, sensitivity, specificity, $F1$-score, and AUC-ROC. Its performance remained consistently high with SMOTE, confirming the model's robustness and capability of generalization in learning complex patterns associated with accident severity.

4.3 Interpretability

In classification problems, SHAP values provide a consistent and locally accurate method for attributing the predicted probability of a class to each input feature. For a given instance, positive SHAP values indicate features that push the prediction toward a specific class (in this case, a fatal outcome), while negative values pull it in the opposite direction. When aggregated across the dataset, SHAP values reveal the global importance of features and allow the identification of dominant patterns that influence the model's decision-making process. In this work, we used SHAP values to quantify each variable's contribution to the classification of accident severity and explore how specific feature categories influence the likelihood of fatal outcomes.

The SHAP summary plots presented in Fig. 3 provide insights into the inner workings of the RF classifier. The absolute mean SHAP value plot demonstrates that features related to the location and nature of accidents (`br_km`, `accident_type`, and `road_type`) are the most influential in determining the likelihood of a fatal outcome. Such a result suggests that road conditions and contextual factors have significant impacts on accident severity, which might be a consequence of infrastructure and traffic policies of a given location (e.g., speed limits, police surveillance, and road maintenance), as well as how some types of accidents are naturally more prone to result in fatalities. In contrast, demographic characteristics such as the age and gender of the individuals involved exhibit low average SHAP values, indicating a marginal influence on the model's predictions and suggesting that these attributes may not be strong determinants of fatality within the dataset analyzed.

The beeswarm plot (Fig 3 bottom) shows how individual feature values influence the predicted outcome. Among the top six features with the highest mean

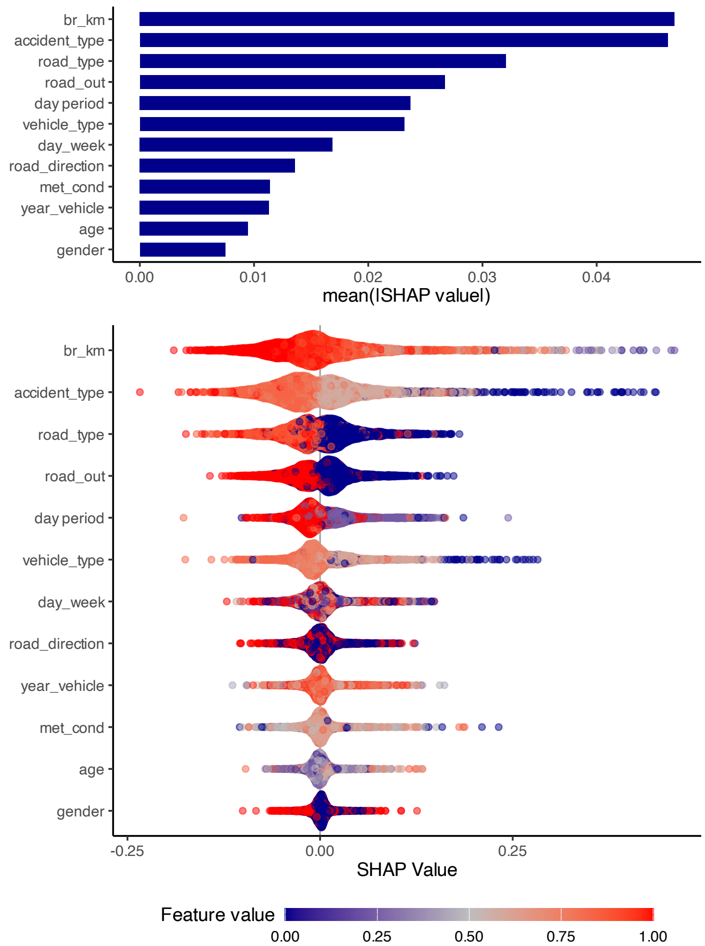

Fig. 3: Shapley values for RF features in test set.

SHAP values (`br_km`, `accident_type`, `road_type`, `road_out`, `day_period`, and `vehicle_type`) there is a notable trend: higher encoded feature values tend to produce lower SHAP values, and vice versa. This pattern indicates a potential inverse relationship between these categorical encodings and the model's predicted probability of fatality. Additionally, several features display diffuse and spread SHAP value distributions, a characteristic often associated with interactions between features. This spread, along with good performance metrics, suggests that the RF model can capture complex relationships that may involve dependencies among variables.

To better understand the importance of the variable `br_km`, we leverage a spatial analysis of fatal accidents, as illustrated in Fig. 4. In the map, each point

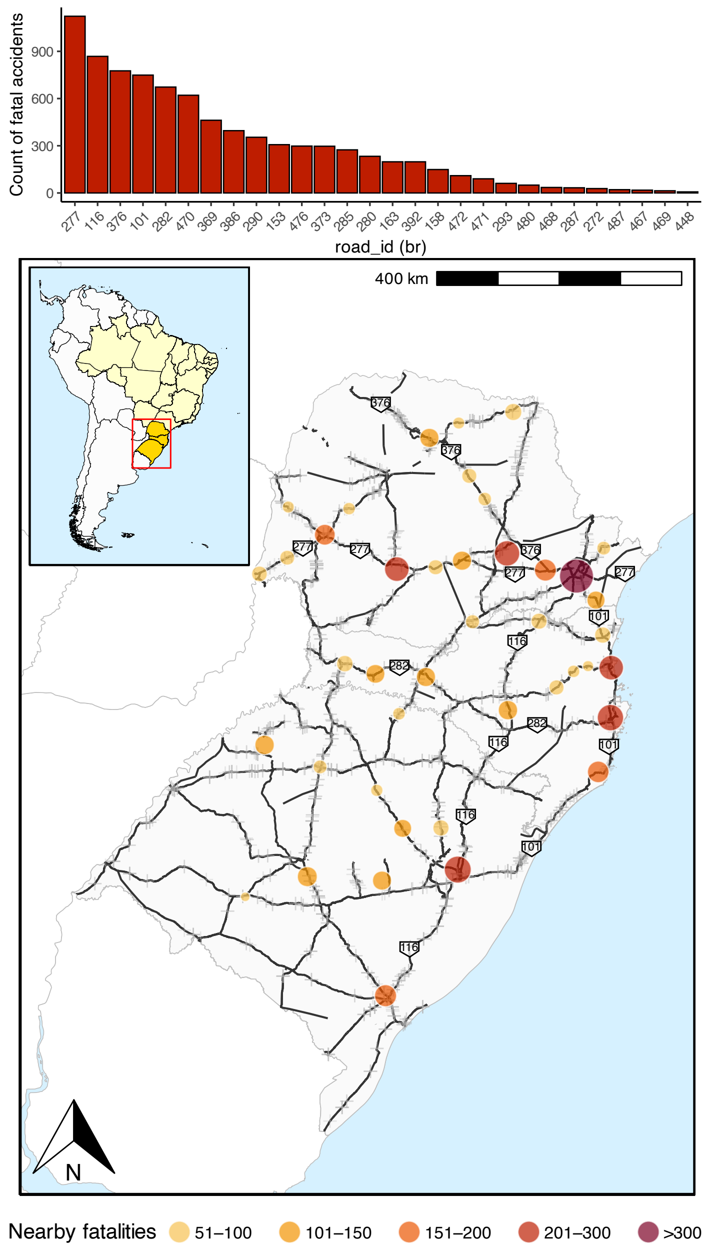

Fig. 4: Road sections (`br_km`) with the most number of fatal accidents.

represents a geographic region where numerous fatal accidents occurred, with the size and color of the points indicating the concentration of fatalities in the nearby areas. Larger and darker points represent higher death counts, and the gray cross markers denote regions of low fatal accident frequencies. The bar plot shows the total number of accidents per federal highway (road_id) within the southern region of Brazil. The five most critical roads regarding fatal accident counts — BR-277, BR-116, BR-376, BR-101, and BR-282 — are highlighted on the map for spatial context.

These five highlighted highways exhibit structural and operational characteristics that may contribute to this elevated rate of fatal accidents. BR-277 and BR-282 are important in connecting the eastern and western regions of Santa Catarina and Paraná, the two states with the highest accident rates in the southern region. These roads often cover rural areas with long sections of single lanes, few overtaking zones, and poor road maintenance, factors that might contribute to a higher incidence of severe crashes, such as front collisions. In contrast, BR-116 and BR-101 are long, heavily trafficked longitudinal highways that pass through densely populated urban areas and industrial hubs. Their frequent intersections and proximity to residential zones increase the likelihood of fatal incidents. Finally, BR-376 connects the Southern region in an important logistics route with high volumes of cargo and heavy vehicles along with sinuous roads on hill and valley areas with sharp curves and challenging driving conditions, where braking distance and driver reaction time are often compromised.

The influence of the remaining most impactful categorical features on model predictions is investigated through violin and boxplots of SHAP values constructed for each category of the accident_type, road_type, road_out, day_period, and vehicle_type variables (Fig. 5). These plots illustrate the distribution and magnitude of each feature category's contribution to the predicted probability of fatal outcomes. Higher SHAP values indicate a more substantial influence toward the fatal class, while values closer to zero suggest a more neutral or non-discriminative role.

The accident_type feature presents distinct patterns among its categories. Run-over, front collision, and overturn accidents have overall positive SHAP values, indicating that they substantially increase the likelihood of fatal outcomes. In contrast, rear-end, side collisions, atypical events, and cases involving vehicles leaving the road tend to have more neutral effects on the model's predictions. Interestingly, although rollovers and fires are typically associated with severe outcomes, they display lower SHAP values in this dataset, likely due to their relative rarity. Notably, run-over accidents are particularly high-risk events, consistent with the lack of protection for pedestrians involved in such incidents.

Regarding road infrastructure, the road_type feature reveals that single-lane roads are strongly associated with fatal outcomes, showing higher SHAP values than multiple-lane or double-lane roads. Such a result is likely due to the increased risk of head-on collisions during overtaking maneuvers, a mechanism also reflected in the high fatality rates associated with front collisions in accident_type. The road_out variable supports this interpretation, indicating that accidents involving vehicles staying in the roadway exhibit elevated SHAP values, reinforcing the link between risky maneuvers and accident severity. The key distinction of off-road category within accident_type from the

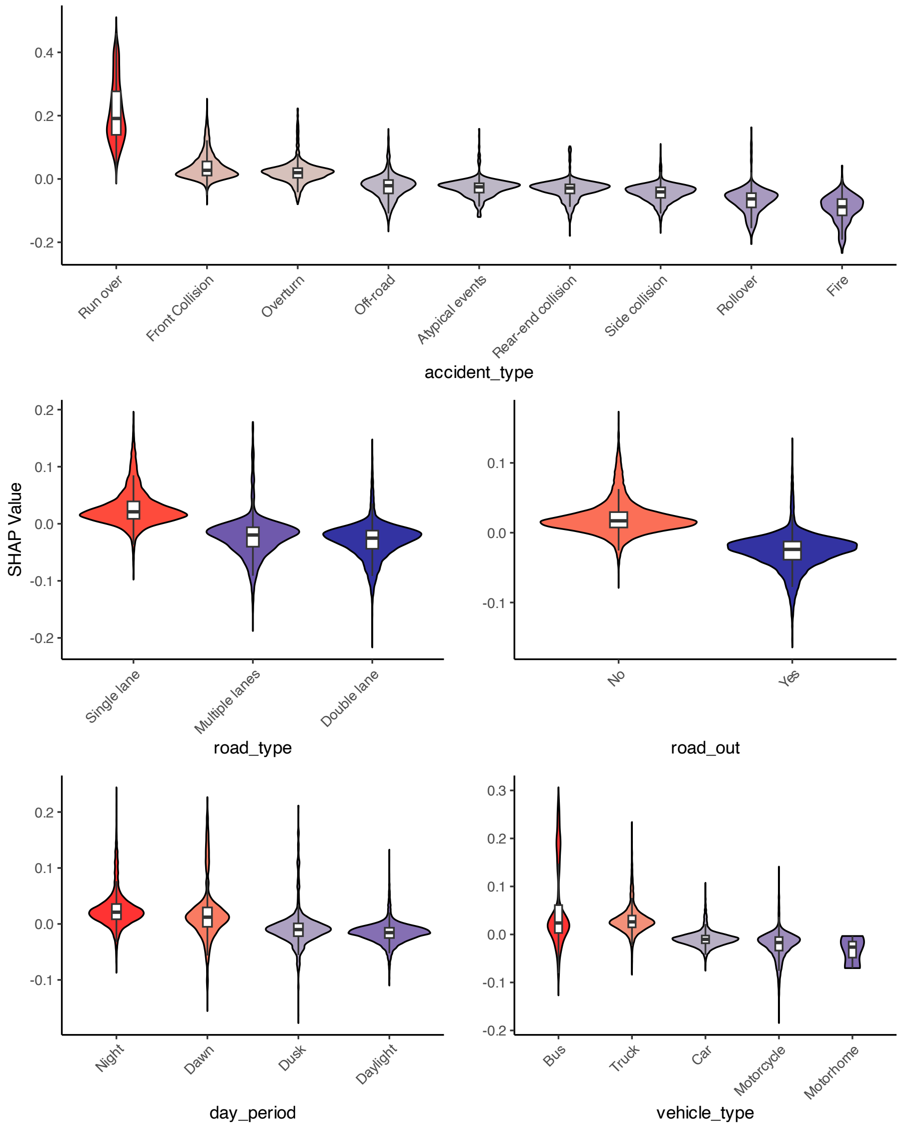

Fig. 5: Shapley values for each category in most influential features.

broader `road_out` feature is that while off-road specifically describes that the vehicle exits the roadway as the primary event, the `road_out` feature is a more general indicator that flags whether any of the vehicles involved in the accident ultimately left the roadway, regardless of the accident's initial type.

Lighting conditions, represented by the `day_period` feature, are also relevant. Accidents occurring at night and dawn are associated with significantly higher SHAP values than those at dusk or in daylight, showing the impact of reduced visibility and possibly driver fatigue or riskier behavior in low-light conditions. Lastly, the `vehicle_type` plot reveals that incidents involving heavier vehicles, such as trucks and buses, result in higher SHAP values, suggesting an increased risk of fatalities, likely due to the greater forces involved in collisions. Cars and motorcycles display a more neutral contribution. Motorhomes, in contrast, show lower SHAP values, which may be attributed to their very low representation in the dataset during the study period, with no recorded fatal outcomes.

5 Conclusions

This paper presented a comparative analysis of classification models applied to traffic accident data from the southern region of Brazil to predict accident severity. The motivation for this study is the alarming number of fatal accidents reported in this region and their substantial economic impact. The dataset used in this analysis was obtained from an open-access source provided by the Federal Highway Police and includes information on accidents recorded between 2021 and 2024. This data includes both fatal and non-fatal incidents, with significant class imbalance (only 7.1% of cases are fatal). To address class imbalance, SMOTE was applied, increasing the proportion of fatal accidents to nearly 50% while preserving the original feature distributions. Three classification methods were evaluated through a comprehensive set of performance metrics. The results suggest that kNN and MLP were not effective to detect fatal outcomes without oversampling, whereas RF achieved near-perfect performance across all metrics. Although SMOTE improved overall sensitivity for kNN and MLP models, it hindered specificity and accuracy. RF achieved satisfactory results across all evaluation scenarios, confirming its robustness and generalization capability.

The study also explored SHAP values for model interpretability, which provided insights into feature contributions to the classification outcomes. Road conditions and characteristics were the most influential in fatal outcomes, whereas demographic features had the least impact. A spatial analysis revealed that certain federal highways concentrate a high count of fatalities, notably: BR-277, BR-116, BR-376, BR-101, and BR-282. Indeed, BR-277 is frequently referred to as the “Death Highway” (*Rodovia da Morte*), a grim title earned due to its high incidence of fatal accidents and hazardous road conditions. The aforementioned roads traverse both rural and densely populated regions, leading to challenging structural and operational conditions. Moreover, factors such as weather conditions and the type of vehicles involved are highly relevant in determining the severity of accidents. Low-light conditions and heavy vehicles are frequently associated with more severe outcomes, including runovers and frontal collisions.

Overall, this work succeeded in developing and evaluating predictive models to determine fatal outcomes using data from federal highways in the southern region of Brazil. By addressing class imbalance and utilizing model interpretability techniques, efficient models were developed, capable of accurately identifying fatal accidents and extracting valuable insights into the underlying contributing factors. These insights are of practical significance for shaping public policies, planning road maintenance, and designing traffic safety awareness campaigns.

As future work, we plan to conduct a comprehensive analysis covering all regions of Brazil to identify regional variations in the factors contributing to fatal accidents, while also considering additional methods for improved predictions and insights. Such comparative studies enhance the understanding of nationwide traffic dynamics and support more targeted and effective intervention strategies across diverse geographic and socio-economic contexts. Finally, we also plan to analyze the impact of infrastructure improvements on reducing fatal accidents. This could lead to the development of a comprehensive methodology for analyzing data over time, enabling the detection of shifting patterns in accident hotspots. Such an approach would not only identify areas where accidents emerge and fade, but also provide insights into the factors that influence these temporal changes, allowing for more dynamic and proactive traffic safety measures.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). The authors also thank the financial support provided by the Fundação Stemmer para Pesquisa, Desenvolvimento e Inovação (FEESC) through the Conexões para Inovar program.

Data Availability

Code and data used in this work is available at:
`https://github.com/arthur-miguel/Highway-to-Classifying-Accidents`

References

[1] Ahmed, S., Hossain, M.A., Ray, S.K., Bhuiyan, M.M.I., Sabuj, S.R.: A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. Transportation research interdisciplinary perspectives **19**, 100814 (2023)
[2] Amorim, B.d.S.P.: Uso de aprendizado de máquina para classificação de risco de acidentes em rodovias. Mestrado em Ciência da Computação, Universidade Federal de Campina Grande (UFCG), Paraíba (2019)
[3] Barros, P.H., Baggio, I.: Uma análise espacial da malha rodoviária brasileira: relações com o desenvolvimento econômico regional. Crescimento e desenvolvimento numa perspectiva interdisciplinar: ensaios sobre o crescimento econômico brasileiro **1**, 360–372 (2022)
[4] Brazilian Federal Highway Police: Dados abertos (2024), disponível em: `https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos`

[5] Brazilian National Confederation of Transport: Painel CNT: Acidentes Rodoviários (2024), disponível em: https://www.cnt.org.br/pesquisas
[6] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
[7] Cuenca, L.G., Puertas, E., Aliane, N., Andres, J.F.: Traffic accidents classification and injury severity prediction. In: 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE), pp. 52–57 (2018)
[8] Dong, S., Khattak, A., Ullah, I., Zhou, J., Hussain, A.: Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with shapley additive explanations. International Journal of Environmental Research and Public Health **19**(5) (2022)
[9] Gan, J., Li, L., Zhang, D., Yi, Z., Xiang, Q.: An alternative method for traffic accident severity prediction: Using deep forests algorithm. Journal of Advanced Transportation **2020**(1), 1257627 (2020)
[10] Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, V., Quaresma, P., Saias, J., Santos, D., Nogueira, P., Silva, M., et al.: Comparison of statistical and machine-learning models on road traffic accident severity classification. Computers **11**(5), 80 (2022)
[11] Jeong, H., Jang, Y., Bowman, P.J., Masoud, N.: Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. Accident Analysis & Prevention **120**, 250–261 (2018)
[12] Kuhn, M., Johnson, K.: Applied predictive modeling, vol. 26. Springer (2013)
[13] Kumeda, B., Zhang, F., Zhou, F., Hussain, S., Almasri, A., Assefa, M.: Classification of road traffic accident data using machine learning algorithms. In: 2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN), pp. 682–687 (2019)
[14] Lobato, C.R.V.: Análise dos acidentes nas rodovias federais concedidas no Brasil. Mestrado em Geotecnia E Transportes, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte (2018)
[15] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
[16] Mokoatle, M., Vukosi Marivate, D., Michael Esiefarienrhe Bukohwo, P.: Predicting road traffic accident severity using accident report data in south africa. In: Proceedings of the 20th Annual International Conference on Digital Government Research, p. 11–17, Association for Computing Machinery, New York, NY, USA (2019)
[17] Pargent, F., Pfisterer, F., Thomas, J., Bischl, B.: Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics **37**(5), 2671–2692 (2022)
[18] Shapley, L.S.: A Value for N-Person Games. RAND Corporation, Santa Monica, CA (1952)