ACCIDENTS ON BRAZILIAN HIGHWAYS: IDENTIFICATION AND CLASSIFICATION OF THEIR CAUSES FROM MACHINE LEARNING

Eveline Oliveira Malaquias[1], Marielce de Cássia Ribeiro Tosta[1], Gisele de Lorena Diniz Chaves[2], Glaydston Mattos Ribeiro[3]

[1] Universidade Federal do Espírito Santo

[2] Universidade Federal de Santa Catarina

[2] Universidade Federal do Rio de Janeiro

Abstract

Traffic accidents or crashes are a worldwide problem with social and economic consequences, beyond incalculable losses. These accidents result from multiple factors: human, road, vehicle, and environmental causes. Despite the complexity of the associated factors, some accidents deny the event's cause identification, whether due to the lack of witnesses, death, or divergent information between those involved. In this way, the cause identification of traffic accidents requires in-depth studies to assist security officers in recording occurrences and contributing to a reliable database. In this sense, the objective of this study is to propose a model for identifying and classifying the causes of traffic accidents that occur on Brazilian federal highways. From it, an analysis was carried out to find the best predictive model for the available data and their performance indicators were compared. To this end, exploratory data analysis and Machine Learning models were employed to process Brazilian accident data in the years 2017, 2018 and 2019. The results point to trends in the pattern and major causes of accidents: drivers, vehicle, road and external factors. The Random Forest algorithm presented better predictive precision of 69%. Considering the diversity and variability of causes, this performance value is acceptable. The model was able to learn the behavior of the data and generalize new occurrences. The main contribution involves the identification of variables that influence road accidents in Brazil, as well as the main risk factors, which can assist public policies for their prevention. Furthermore, the classification of these causes can assist security officers in analyzing the occurrences.

Keywords: traffic accidents; road crash; forecast model

1. INTRODUCTION

Traffic Accidents (TA) represent the eighth cause of death worldwide for people of any age. This category of accidents is the only cause of death among the ten most recurrent in the world that is not directly associated with any kind of disease (OMS, 2018). If no action is taken to reverse the conditions of transport and mobility, it is estimated that in 2030 the TA will reach the fifth leading cause of death in the world (Pradhan e Sameen, 2020). The number of deaths is three times higher on highways or roads in developing countries (Samson e Adewale, 2020). Surpassing the average mortality rate of 15.6 deaths per 100,000 inhabitants of the American continent, Brazil has an estimated mortality rate of 19.7 (Paho, 2019). The country has expressive TA numbers, occupying fifth place in the world ranking, behind only India, China, USA, and Russia (OMS, 2018).

After the enactment of stricter traffic safety laws, a reduction in deaths on Brazilian federal highways was noticed in 2017. However, despite the drop in the number of deaths, accidents continue to impact the health system inducing economic and social losses (Pradhan e Sameen, 2020). Road traffic injuries generate costs with treatments, rehabilitation, and investigation, as well as reduced or lost productivity. Hospitalizations related to traffic injuries increased by 33% from 2009 to 2018 (CFM, 2019). The cost of TA represents about 3% of the Gross Domestic Product (GDP) of countries (PAHO,

2019), and in Brazil, it reached 3.7% of GDP in 2015. In 2020, total expenses with accidents on Brazilian highways were estimated at around BRL 40 billion (IPEA e ANTP, 2020).

Identifying the main causes of accidents (or chashes) makes it possible to design the best way to act on critical points (Zhang et al., 2020). This mapping requires analyzing data to observe patterns and develop efficient preventive measures (Luoma and Sivak, 2007). Furthermore, the complexity of TA associated with multiple factors (Chen, 2017; Rios et al., 2020) compromises data consistency. However, factors available in the accident occurrence contain valuable information for their understanding (Pradhan e Sameen, 2020; Ghandour et al., 2020; Zhang et al., 2020).

In this sense, technological tools can be used to identify and classify the causes of accidents (Chen, 2017; Zheng e Huang, 2020). The use of computational methods emerges as an alternative to improve traffic safety control and they are increasingly used to predict accident-related factors (Azimi et al., 2020). With several possibilities of predictive modeling, Machine Learning techniques have been used in this analysis (Dadashova et al., 2020). However, Brazil does not have enough tools allocation nor significant numbers of scientific publications with TA prediction models (Zou et al., 2020).

Since the success of safety programs depends on the reliability of accident data, this study aimed to obtain a treated database to be used in a Machine Learning model to evaluate the prediction results of different algorithms used in the literature. In this paper we identified and classified the causes of the accidents on Brazilian highways using machine learning. From this, it was possible to analyze the importance of the data pre-processing steps in the results obtained. An analysis was also carried out to find the best predictive model for the available data to compare the performance measures indicated for classification models. The results can support the Federal Highway Police (PRF) of Brazil in fulfilling and improving the Traffic Accident Bulletin as well as support more precise changes and adjustments in safety and prevention policies.

2. TRAFFIC ACCIDENTS: STUDIES WITH MACHINE LEARNING MODELS

A traffic accident is an interaction between roads, vehicles, drivers, and environment (Mohamed, 2014). The literature studies the main aspects that cause accidents to propose interventions that reduce incidents (Ghandour et al., 2020). In the literature, 33 main features used as input data in ML models were identified. They can be categorized by driver, road condition, environment, temporal and travel-related factors, or accident (Table 1).

Table 1 - Variables used in studies of traffic accidents

Group	Variable
Driver	Age, Gender, Time of driving experience, Nationality, Occupation/Position, Family income, Frequency of accidents, History of infractions, Psychological and socioeconomic factors
Road conditions	Number of lanes, Terrain conditions, Road quality, Land use, Road surface, Traffic flow, Lane width
Temporal	Time, Date, Day of the Week
Environment	Weather conditions, Light conditions, Average temperature, Location
Travel	Average speed, Distance traveled, Vehicle type, Vehicle age
Accident	The severity of the accident, Cause of the accident, Mechanical failures, Use of alcohol, Improper overtaking, Violation of speed, Number of vehicles involved

Source: adapted from Malaquias et al. (2021)

With the advancement of mathematical and computational models, models such as Machine Learning - ML can be useful to find combinations of factors that lead to accidents, assist traffic authorities in decision making and design effective preventive countermeasures (Rios et al., 2020). The use of this tool rose in road safety with accident prediction models, which support the

identification of critical points and propositions of improvements (Yannis et al., 2017). Hegde and Rokseth (2020) highlight the ten most used algorithms in order of the largest number of applications: Artificial Neural Networks, Support Vector Machine, Decision Trees, Random Forests, Classification and Regression Tree, Naive Bayes, K-Means, K-Nearest Neighbors Algorithm, Logistic Regression, and Boosted Regression Trees. Table 2 summarizes the different objectives and algorithms used in ML models applied to TA, demonstrating its great versatility.

Table 2 - Algorithms used in predictive studies of traffic accidents

Algorithm	Objective	Studies
Random Forests	Predict frequency of collisions; Predict severity of injuries; Identify variables related to accidents/injuries/frequency	Hassan & Abdel-Aty (2013); Ghandour *et al.* (2020); Zhang *et al.* (2020); Dadashova *et al.* (2020).
Support Vector Machine	Predict frequency of collisions; Predict severity of injuries; Identify variables related to accidents/injuries/frequency; Predict the cause of traffic accidents.	Chong *et al.* (2005); Li *et al.* (2008); Li *et al*. (2012); Yu & Abdel-Aty (2013); Mohamed (2014); Sun & Sun (2016).
K-means	Predict frequency of collisions; Predict severity of injuries	Sohn & Lee (2003); Sun & Sun (2016).
Artificial Neural Networks	Predict frequency of collisions; Identify variables related to accidents/injuries/frequency; Predict severity of injuries	Chong *et al.* (2005); Akgüngör & Doğan (2009); Huilin &Yucai (2011); Cigdem & Ozden (2018); Ghandour *et al.* (2020).
Decision Trees	Predict severity of injuries	Chong *et al.* (2005); Martín *et al.* (2014).
Classification and Regression Tree	Predict frequency of collisions	Chang & Chen (2005); Moradkhani *et al* (2014).
Logistic Regression	Identify variables related to accidents/injuries/frequency; Predict severity of injuries; Determine the impact of drinking and driving on the severity of injuries	Moradkhani *et al.* (2014); Ghandour *et al.* (2020); Dadashova *et al.* (2020).
K-Nearest Neighbors Algorithm	Identify variables related to accidents/injuries/frequency	Lv *et al.* (2009).
Naive Bayes	Identify variables related to accidents/injuries/frequency; predict injury severity	Shanthi & Ramani (2012); Ghandour *et al.* (2020).

Source: prepared by the authors

Despite the advance in the use of computational methods, ML studies on this topic are still incipient (Zheng e Huang, 2020). Some studies focus on discussing the relationship between accidents and the factors that influence them, but there is still little progress in terms of predictive models (Zhang et al., 2020; Ghandour et al., 2020).

3. METHODOLOGY

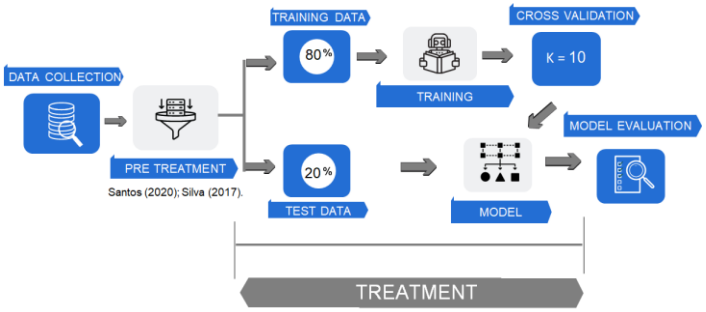
The study used data from the Brazilian Federal Highway Police Department (DPRF, 2022) considering the occurrences of accidents in the years 2017, 2018 and 2019. It was decided to exclude the data after the beginning of the pandemic, as there was a considerable change in behavior and causes in this period, as pointed out by the International Transport Forum (2021). The Python programming language was used to organize the data in the temporal order of occurrences and organize the database for analysis. In total, the database resulted in 531,470 valid records and 0.6% missing one or more explanatory variables. Classes of accident causes categorized into four main groups were adopted, according to Chen (2017):

- Driver: driver sleeping, sudden illness, ingestion of alcohol, ingestion of psychoactive substances, not keeping a safe distance, incompatible speed, improper overtaking, lack of attention while driving, driver disobedience to traffic regulations;

- Vehicle: excessive load and, or poorly packed, damages and/or excessive wear on the tire, mechanical defect, deficiency or non-activation of the lighting/signaling system;
- Road: animals on the road, road defects, slippery road, insufficient or inadequate road signage, static objects on the road;
- External factors: natural phenomena, external aggression, disobedience to traffic rules by the pedestrian, ingestion of alcohol and/or psychoactive substances by the pedestrian, restriction of visibility, and lack of attention by the pedestrian.

The construction of a machine learning model involves the following macro steps: collecting input data, pre-processing, processing, output data, and its evaluation (Canhoto e Clear, 2020). In this study, incremental steps of data pre-processing were performed in five models, to measure the impact of each incremental treatment for data adjustment, as proposed by Han et al. (2011) and Mohamed (2014). The steps adopted for the construction of the five models (named A, B, C, D and E) of Machine Learning in this study are shown in Figure 1, as well as the synthesis of the different criteria and pre-processes adopted.

Figure 1 - Pre-processing steps for each Machine Learning model



Pre-processing steps	A	B	C	D	E
Manual selection of variables		✓	✓	✓	✓
Deleting all outliers and missing values			✓		
Selection of variables by mathematical methods			✓	✓	✓
Exclusion of outliers and missing values of predictor variables only				✓	✓
Balance between categories					✓

Model A was built without any pre-processing methods. All variables were considered, as well as all occurrences with outliers and missing data, totaling 531,470 records in the database. In Model B the pre-processing of data involved the selection of variables manually. Of the variables available in the database, only those cited in the literature were selected (Table 2). The variables "accident classification", "type of accident", "uninjured", "slightly injured", "seriously injured", "dead", and "physical state" were excluded because they refer to a consequence of the accident. The variables “time” and “date” were also excluded due to the high number of classes (1,434 and 1,095, respectively), which impairs the performance of the models. In Model C, the column “date”, originally informed in the format day/month/year (Brazilian standard), was replaced by the temporal variable “month”. With the variable “year of vehicle manufacture”, a new one entitled “vehicle age” was obtained, calculated by subtracting the year of the accident from the year of vehicle manufacture. Another adapted variable was the “hour”, extracted from the HH:MM:SS format. Then, a ranking of features was enumerated to ensure that the model was trained only with the most relevant attributes. The Information Gain Ratio method evaluates the value of a variable and helps the model to have greater generalization capacity (Gong et al., 2020) and was also used by Mohamed (2014) in a model for classifying accident causes. The Gini Index was also used, an indicator that quantifies the degree of impurity and information gain, used by Zhang et al. (2020) and Dadashova et al. (2020). Input variables that did not score in either method was excluded. In this modeling, all occurrences that contained missing data in some variable, as well as occurrences with outliers for dimensionality reduction, were excluded. Therefore, the dataset used after pre-processing had 415,744 occurrences.

For a more refined pre-processing, in Model D, data filled with "not informed", "invalid" or wrongly informed values were discarded. In total, 62,378 instances were excluded by this criterion, totaling 469,092 occurrences for Model D. Model E involved adjustments in the dimensioning of

accident causes to minimize data imbalance between the considered categories. The dataset was randomly resampled, with the subsampling technique, indicated to treat unbalanced data (Ganganwar, 2012). The distribution of the categories used is shown in Table 3. It is observed that the predominant driver category has more balanced participation in Model E.

To start processing, data were randomly divided into 80% for training and 20% for testing, also used by Ghandour et al. (2020). In all cases, 80% of the data is used to train the algorithms. For this, cross-validation with 10 parts was used. Cross-validation is paramount to provide the model's ability to generalize to independent data and a subset of 10 parts (k = 10) is generally used in the literature (Zhang et al., 2020). Among the ten most frequently used algorithms (Goodfellow; Bengio; Courville, 2016; Hegde and Rokseth, 2020), the models considered in this article were built from the following algorithms: Random Forest, Naive Bayes, Logistic Regression and Support Vector Machine. The evaluation of the models used the performance measures suggested by Mohamed (2014) e Panicker e Ramadurai (2022). These metrics are suitable for multi- category classification models (Menon et al., 2020). Accuracy was not used since it is recommended for balanced databases, that is, with the same proportion of data for each class (Jiang et al., 2020).

Table 3 - Distribution of target variable categories with the inclusion of Model E

Category	Models A and B		Model C		Model D		Model E	
	Number of events	Proportion (%)	Number of events	Proporção dados (%)	Number of events	Proporção dados (%)	Number of events	Proporção dados (%)
Driver	428,718	80.67	337,924	81.28	380,466	81.11	40,000	31.09
External factors	26,773	5.04	17,540	4.22	23,135	4.93	23,135	17.99
Road	41,830	7.88	34,366	8.27	37,257	7.94	37,257	28.97
Vehicle	34,106	6.41	25,914	6.23	28,234	6.02	28,234	21.95

Recall is the ability of a classification model to identify all relevant cases, where the number of true positives is divided by the sum of true positives and false negatives. (Elassad et al., 2020). Precision refers to the classification model's ability to identify only relevant data. It is defined by the number of true positives divided by the number of true positives plus the number of false positives. (Adekitan et al., 2019). There is a trade-off between recall and precision: as you choose to maximize one, you decrease the other. Then, a combination of recall and precision is given by the harmonic mean between the two, known as the F1 score. A model that has a high F1 is a model balanced between recall and precision. It can vary between 0 and 1, and the closer to 1, the better the performance of the model (Jiang et al., 2020; Panicker and Ramadurai, 2022). The F1 score is a measure that is suitable for models with unbalanced classes and is more suitable for performance evaluation than the precision alone (Elassad et al., 2020). To identify the best performance, a confusion matrix was considered for each possibility. The Confusion Matrix is the result of a visual agreement analysis between prediction and reality. The matrix diagonal represents the ideal case where the instance was correctly classified, while all off-diagonal cells represent misclassified instances. Thus, it is possible to calculate how many data were correctly classified and, if not, with which category they were confused (Marom et al., 2010).

4. RESULTS

Model A did not receive any treatment and took all the noise and errors from the database, in addition to its high dimensionality. For this reason, the Logistic Regression, Random Forest, and Support Vector Machine algorithms presented errors due to the huge amount of data and could not be processed due to the incapacity of computational memory. The Naive Bayes algorithm resulted in an F1-score of 0.776 in the training and 0.774 in the test. The precision value in the test indicates that,

of all the predictions made by the model in each category of variables, on average 79.6% of the predictions were correct. The recall points out that about 75.9% of all cases (without subdivision by category) had assertive classification. Measures of precision, recall and F1 score are an aggregate of the result for each category (driver, road, vehicle, and external factors). The individual results of the categories, however, indicated the disproportionate concentration of the values of each measure in the Model A Confusion Matrix. The "driver" category got 70,530 occurrences right out of a total of 85,655 real cases and the model performed well as it managed to get most of the data right, achieving a precision of 0.89 and a recall of 0.83. However, the performance for the other categories was not satisfactory with few hits (true positives) and low values in precision and recall. Of the total of 106,294 data tested, Model A missed 35,764 classifications. Therefore, the biased model for accidents caused by drivers was not able to generalize the classifications.

The reduction of the data dimension, in Model B, allowed the algorithms Logistic Regression, Random Forest and Support Vector Machine to be processed (Table 4). For Random Forest, the F1 score with a value above 0.8 in the test for the aggregated categories provides a satisfactory result. Precision and recall averages maintained a high and balanced pattern with each other. Despite this, there was no expressive number of correct answers (true positives) for the minority categories. When comparing the results of the training and testing stages, Naive Bayes and Logistic Regression were the only algorithms that showed a tendency to overfit, represented by the decrease in the performance of the F1 score from the training to the test stage.

Table 4 - Model B performance evaluation

Model	**Training**			**Test**		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Logistic Regression	0,721	0,651	0,807	0,719	0,649	0,806
Naive Bayes	0,736	0,729	0,801	0,735	0,719	0,800
Random Forest	0,803	0,840	0,843	0,806	0,844	0,845
Support Vector Machine	0,712	0,672	0,771	0,714	0,665	0,789

The confusion matrix presented in Figure 1 shows that the number of false positives is significant, especially in the driver category. Colors indicate the concentration of data: blue color indicates true positives, pink color indicates wrong predictions, and white cells mean little or no occurrence. In blue and pink colors, darker tones indicate a greater number of occurrences. Comparison between actual and predicted events show that the model learning problems, missing many classifications in the test phase. In all algorithms, there was a strong tendency to classify the instances for the majority class (driver), evidencing the consequences of designing models with unbalanced categories. The lack of variables representing the phenomenon and the presence of noise in the data are also factors that can negatively affect learning. Despite this, the reduction in the size of the database enabled the processing of all the proposed algorithms.

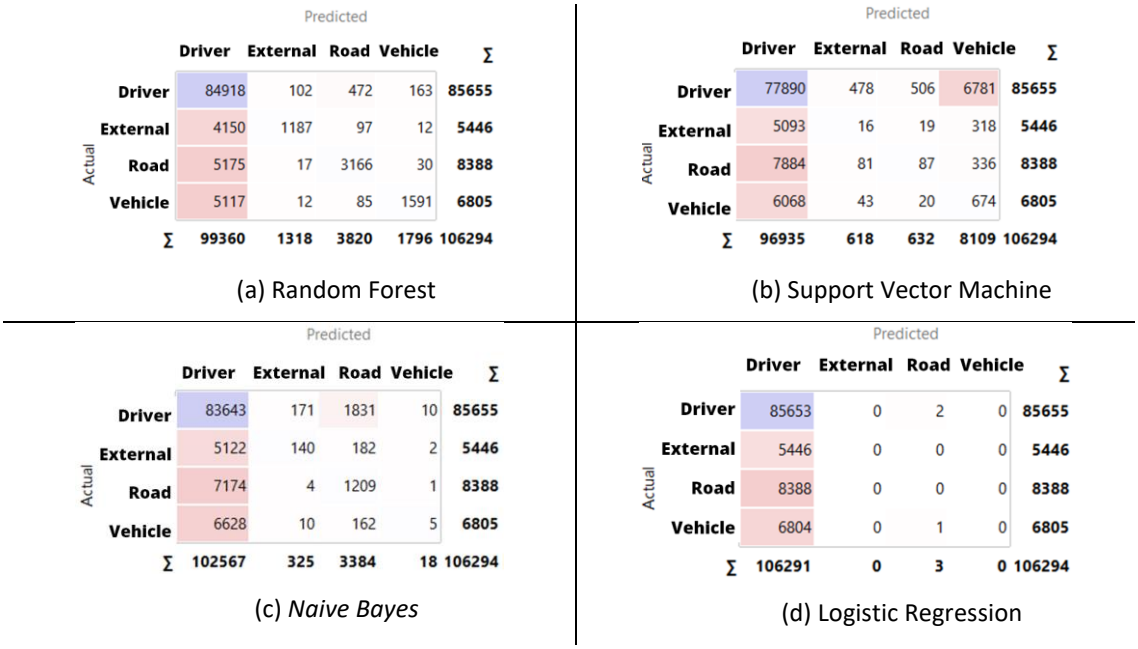


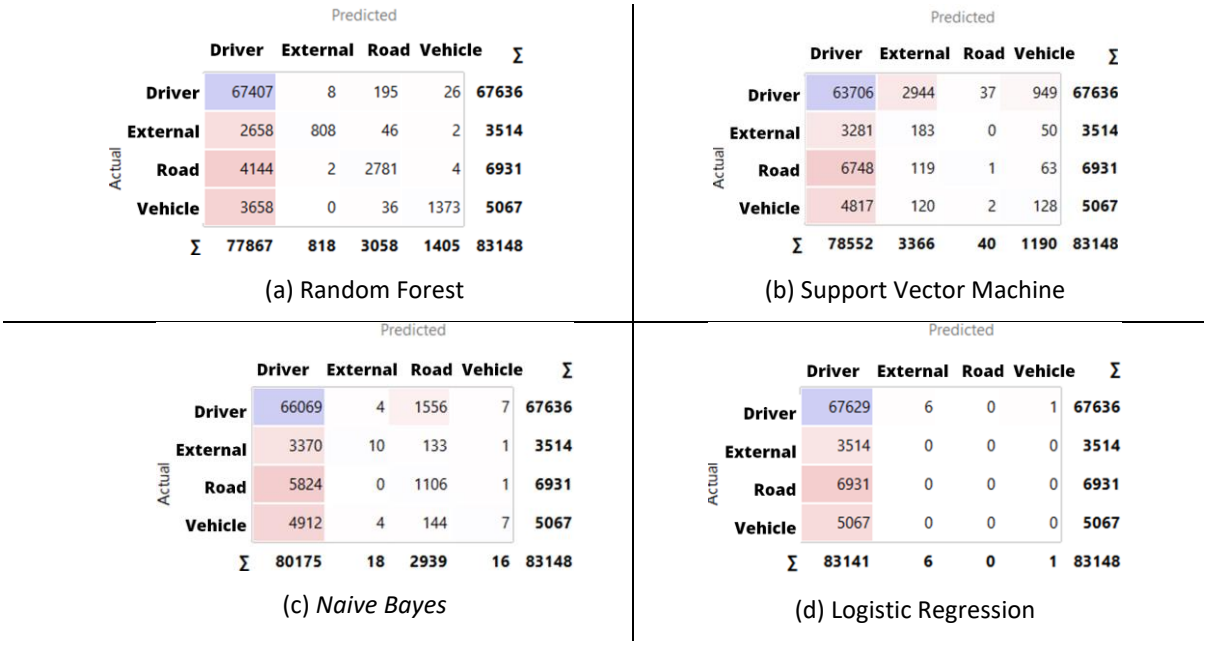
Figure 1 - Confusion matrix for the algorithms in Model B

When processing Model C, it was found that the manipulated variables (vehicle age and period) were more relevant to characterize the data than the vehicle year and time (HH:MM:SS format). Table 5 shows that the aggregate performance of all algorithms improved, indicated by the F1 score. Unlike Model B, in which the test performance had been lower than in the training for Logistic Regression and Naive Bayes, in Model C this did not occur in any case, whereas there was an improvement for the four algorithms.

Table 5 - Model C performance evaluation

Model	Training			Test		
	F1	Precision	Recall	F1	Precision	Recall
Logistic Regression	0,729	0,663	0,813	0,730	0,662	0,813
Naive Bayes	0,744	0,741	0,807	0,746	0,752	0,808
Random Forest	0,834	0,877	0,865	0,842	0,881	0,870
Support Vector Machine	0,698	0,679	0,727	0,715	0,671	0,770

The Confusion Matrix in Figure 2 shows that the Random Forest had better predictions, but many instances were misclassified in the minority categories. As the aggregate result of this algorithm presented high values, it can be concluded that these were, in the case of the recall, influenced by the individual performance of the driver category, as in Models A and B. The predictions by the Random Forest and Naive Bayes algorithms do not show a significant difference in the trend between the categories to Model B. In these two algorithms, the number of forecast errors remained high for all minority classes, as well as the real data was poorly identified. Many erroneous predictions were classified as driver-caused, indicating that the algorithms remained biased in this category. For the Support Vector Machine, despite having classified a lot of data as “driver”, it also got more instances right in the “external factors” class than in the previous model and worsened the prediction for the “road” and “vehicle” categories. The selected variables and the exclusion of spurious data may have negatively affected the model with the lack of representative data for road and vehicle categories in this algorithm. The Logistic Regression model remained with a low predictive capacity in the minority classes and, despite a satisfactory performance for the "driver" category, the model was biased, and the result does not reflect machine learning. The treatments performed were not enough to make the models satisfactory.



(a) Random Forest

(b) Support Vector Machine

(c) *Naive Bayes*

(d) Logistic Regression

Figure 2 - Confusion matrix for Model C algorithms

In Model D, all occurrences of accidents that had missing data or outliers in one or more variables were excluded. Table 6 shows that there was a small decrease in the values of the metrics, in all algorithms, indicating a possible drop in the performance of the model compared to Model C, but superior performance to Model B.

Table 6 - Model D performance evaluation

Model	Training			Test		
	F1	Precision	Recall	F1	Precision	Recall
Logistic Regression	0,727	0,661	0,811	0,725	0,656	0,810
Naive Bayes	0,743	0,731	0,805	0,741	0,738	0,804
Random Forest	0,835	0,878	0,866	0,840	0,882	0,869
Support Vector Machine	0,695	0,677	0,719	0,704	0,663	0,754

The confusion matrix (Figure 3) shows that although the Logistic Regression does not present a low F1 score, the ability to generalize the data is not perceived. In the matrices of Models B, C and D there were no prediction hits for the minority categories with this algorithm. As with previous models, the aggregate F1 score results were influenced by the high incidence of true positives in the "driver" category. Thus, accuracy and recall metrics remain high due to data imbalance. In none of the four models was an algorithm able to achieve satisfactory rates of correct classifications in the other categories. Although the total number of data has decreased in this model, the proportion of records by categories has not changed significantly. The driver category concentrated more than 80% of the instances in all models, which explains the bias in the results presented so far. Therefore, Model D did not present any improvements compared to the previous models.

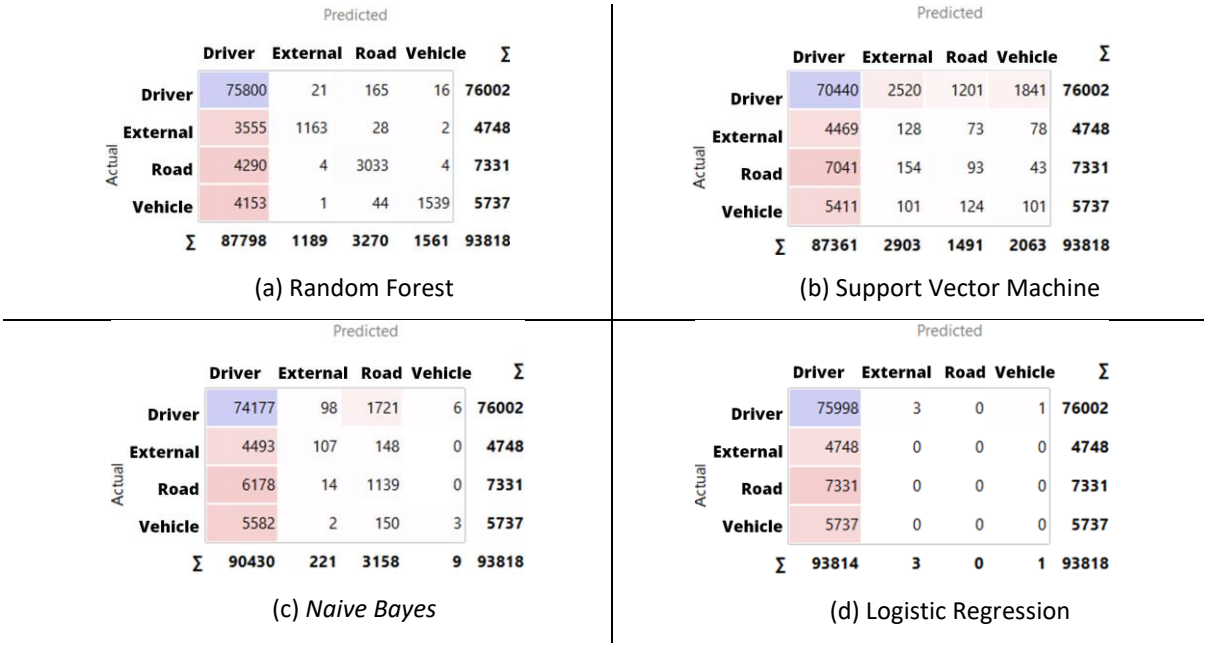


Figure 3 - Confusion matrix for Model D algorithms

Finally, in Model E, the four classes were balanced to investigate whether this treatment results in a decrease in the model's bias. Table 7 shows that, despite this last model receiving more pre-processing steps, its aggregate indicators had reduced performance. However, this result must be analyzed together with the confusion matrix, shown in Figure 4.

Table 7 - Model E performance evaluation

Model	Training			Test		
	F1	Precision	Recall	F1	Precision	Recall
Logistic Regression	0,466	0,474	0,479	0,471	0,479	0,484
Naive Bayes	0,455	0,455	0,463	0,459	0,459	0,467
Random Forest	0,675	0,679	0,676	0,689	0,692	0,689
Support Vector Machine	0,281	0,344	0,322	0,253	0,305	0,304

This model resulted in the first matrix with a balanced distribution between classes, that is, fewer accidents were labeled as caused by the driver. Despite the smaller amount of data for training and testing, Model E resulted in higher numbers of correct predictions for the minority categories, as demonstrated in Figure 4. Algorithms were forced to learn the behavior of the data to get the classification right and not just label lots of data in the majority class for good accuracy and recall. Although the algorithms still present true negative, false positive and false positive classification errors, there was an improvement in the rates of true positives and the tendency to classify the instances as a driver in all algorithms was reduced. Even with a reduction in the performance of the evaluation metrics, the results of this model were satisfactory since the predictions occurred in a balanced way between the categories. The Logistic Regression algorithm had more correct predictions, evidencing that the balancing of the classes results in a greater ability to generalize the Machine Learning models. The Random Forest was the model that obtained the best results, with a proportion of true positives greater than errors due to false positives and false negatives in all categories, which had not yet happened in any previous model. For this algorithm, the precision for each category was: driver - 62%, external - 77%, road - 72%, and vehicle - 70%. The recall obtained in each category was: driver - 65%, external - 67%, road - 77%, and vehicle - 65%.

(a) Random Forest

(b) Support Vector Machine

(c) *Naive Bayes*

(d) Logistic Regression

Figure 4 - Confusion matrix for Model E algorithms

5. CONCLUSION

The analysis of the confusion matrices of the five proposed models showed that the incremental steps of data pre-processing present important contributions to the ability to classify new accident occurrences. Model E presented the best predictive capacity with the Random Forest algorithm and was still the only model to correctly classify data from minority classes with the Logistic Regression algorithm. The results of this Model E allow us to conclude that by carrying out all the data pre-processing steps indicated by the literature, it is possible to obtain Machine Learning models capable of dealing with new data, which is the main objective of a prediction or classification model. Based on historical data, it is expected that after training a model with treated data it is possible to obtain a design that can deal with new data and then without human intervention obtain correct results. We emphasize the importance of the data balancing step in a multi-category classification model, which is the decisive treatment for the generalization capacity of Model E.

The models did not show significant improvements before balancing the target variable data, due to the strong influence of the majority class on the models' performance. An important contribution of the comparison between the models is the use of the confusion matrix for the interpretation of the prediction results since the quantitative results are highly influenced by an unbalanced model and do not point out the deficiencies of the model.

Also noteworthy is the variable selection step, which directly contributed to the reduction of the size of the database and allowed the processing of all algorithms from Model B. Aiming to contribute to the reduction of data, the mathematical methods that punctuate the importance of the variables can help the qualitative understanding of traffic accidents, as they point out the characteristics that are associated with a risk event. In addition, the replacement of the time, date and year of vehicle manufacturing variables with other variables with similar information, but with a smaller number of classes, also allowed for an improvement in the data processing. These adjusted variables show that the researcher's work influences the improvement of Machine Learning models.

The F1 score, accuracy and recall measures for the designed models were not close to the maximum value of 1, indicating that improvements from Model E can still be made. However, it should

be noted that the steps taken were essential to creating the proposed model, in which it was possible to remove the bias from the analysis and start discussions about learning the model for the available data. Thus, it is still possible to conclude that the data provided by the Department of Highway Police in Brazil provides the variables that have the potential to explain the phenomenon studied.

The results show that the Random Forest classifier has the best performance for the modeled data, while the Vector Support Machine had the worst performance. The performance obtained in the Random Forest (69%) is acceptable as a classifier but could be improved. Suggestions for improvements to the classification model are in the adjustment of hyperparameters as regulators of the algorithms, to obtain a more robust model and, consequently, better prediction results. It is recommended to test other algorithms such as Artificial Neural Networks and Decision Trees. A combination of methods can also be designed to improve model accuracy. Furthermore, additional data pre-processing steps can be added following the steps here provided.

This study provides initial notes on the applicability of machine learning to help predict accidents on Brazilian highways, pointing out the best algorithms and data pre-processing steps necessary to improve results. Another contribution involves the identification of variables that influence road accidents in Brazil, as well as the main risk factors, which can assist public policies for their prevention. Furthermore, the classification of these causes can assist security officers in analyzing the occurrences. Knowledge of the risks of traffic accidents associated with different factors constitutes relevant information for the definition of priorities in public health and safety policy to mitigate the seriousness and adversities arising from them.

REFERENCES

Adekitan, A. I., Abolade, J., & Shobayo, O. 2019. Data mining approach for predicting the daily Internet data traffic of a smart university. Journal of Big Data, 6(1), 1-23.

Akgüngör, A. P., & Doğan, E. 2009. An artificial intelligent approach to traffic accident estimation: Model development and application. Transport, 24(2), 135-142.

Azimi, G., Rahimi, A., Asgari, H., & Jin, X. 2020. Severity analysis for large truck rollover crashes using a random parameter ordered logit model. Accident Analysis & Prevention, 135, 105355.

Canhoto, A. I., & Clear, F. 2020. Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. Business Horizons, 63(2), 183-193.

Conselho Federal de Medicina (CFM). 2019 Em dez anos, acidentes de trânsito consomem quase R$ 3 bilhões do SUS. Brasília, Brasil.

Chang, L. Y., & Chen, W. C. 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research, 36(4), 365-375.

Chen, C. 2017. Analysis and forecast of traffic accident big data. In ITM Web of Conferences (Vol. 12, p. 04029). EDP Sciences.

Chong, M., Abraham, A., & Paprzycki, M. 2005. Traffic accident analysis using machine learning paradigms. Informatica, 29(1).

Cigdem, A., & Ozden, C. 2018. Predicting the severity of motor vehicle accident injuries in Adana-Turkey using machine learning methods and detailed meteorological data. International Journal of Intelligent Systems and Applications in Engineering, 6(1), 72-79.

Dadashova, B., Arenas-Ramires, B., Mira-McWillaims, J., Dixon, K., & Lord, D. 2020. Analysis of crash injury severity on two trans-European transport network corridors in Spain using discrete-choice models and random forests. Traffic injury prevention, 21(3), 228-233.

DPRF - Departamento de Polícia Rodoviária Federal. Dados Abertos – Acidentes. 2022. Disponível em: <https://portal.prf.gov.br/dados-abertos-acidentes>. Access in 20 jan. 2022.

Ganganwar, V. 2012. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4), 42-47.

Ghandour, A. J., Hammoud, H., & Al-Hajj, S. 2020. Analyzing factors associated with fatal road crashes: a machine learning approach. International journal of environmental research and public health, 17(11), 4111.

Gong, F., Jiang, L., Zhang, H., Wang, D., & Guo, X. 2020. Gain ratio weighted inverted specific-class distance measure for nominal attributes. International Journal of Machine Learning and Cybernetics, 11(10), 2237-2246.

Hassan, H. M., & Abdel-Aty, M. A. 2013. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. Journal of safety research, 45, 29-36.

Hegde, J., & Rokseth, B. 2020. Applications of machine learning methods for engineering risk assessment–A review. Safety science, 122, 104492.

Huang, F., Zhang, J., Zhou, C., Wang, Y., Huang, J., & Zhu, L. 2020. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. Landslides, 17(1), 217-229.

Huilin, F.; Yucai, Z. 2011. The Traffic Accident Prediction Based on Neural Network. In: 2nd International Conference on Digital Manufacturing & Automation.

International Transport Forum ITF (2021), Road Safety Annual Report 2021: The Impact of Covid-19, OECD Publishing, Paris.

Jiang, L., Xie, Y., Wen, X., & Ren, T. 2020. Modeling highly imbalanced crash severity data by ensemble methods and global sensitivity analysis. Journal of Transportation Safety & Security, 1-23.

Li, X., Lord, D., Zhang, Y., & Xie, Y. 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis & Prevention, 40(4), 1611-1618.

Li, Z., Liu, P., Wang, W., & Xu, C. 2012. Using support vector machine models for crash injury severity analysis. Accident Analysis & Prevention, 45, 478-486.

Lv, Y., Tang, S., & Zhao, H. 2009, April. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In 2009 international conference on measuring technology and mechatronics automation (Vol. 3, pp. 547-550). IEEE.

Malaquias, E. O. ; Tosta, M. C. R. ; Chaves, G. L. D. ; Ribeiro, G. M. . Acidentes em rodovias brasileiras: um estudo com técnicas de machine learning para classificar a causa das ocorrências. In: . Anais do 35 Congresso de Ensino e Pesquisa em Transportes, 2021, online. Rio de Janeiro: ANPET, 2021. v. 1. p. 1

Marom, N. D., Rokach, L., & Shmilovici, A. 2010, November. Using the confusion matrix for improving ensemble classifiers. In 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel.

Martín, L., Baena, L., Garach, L., López, G., & De Oña, J. 2014. Using data mining techniques to road safety improvement in Spanish roads. Procedia-social and behavioral sciences, 160, 607-614.

Menon, A. P., Varghese, A., Joseph, J. P., Sajan, J., & Francis, N. 2020. Performance Analysis of different Classifiers for Earthquake prediction: PACE.

Mohamed, E. A. 2014. Predicting causes of traffic road accidents using multi-class support vector machines. Journal of Communication and Computer, 11(5), 441-447.

Moradkhani, F., Ebrahimkhani, S., & Sadeghi Begham, B. 2014. Road accident data analysis: A data mining approach. Indian J. Sci. Res, 3, 437-443.

OMS - Organização Mundial da Saúde. Global Status Report in Road Safety. 2018. Supporting a decade of action. Genebra, Suíça, 2018.

PAHO. Pan American Health Organization. Status of Road Safety in the Region of the Americas. Washington, D.C., 2019.

Panicker, A. K., & Ramadurai, G. (2022). Injury severity prediction model for two-wheeler crashes at mid-block road sections. International journal of crashworthiness, 27(2), 328-336.

Pradhan, B., & Sameen, M. I. 2020. Modeling traffic accident severity using neural networks and support vector machines. In Laser Scanning Systems in Highway and Safety Assessment (pp. 111-117). Springer, Cham.

Samson, A. O., & Adewale, L. A. 2020. Traffic crashes prediction of states in Nigeria using time series analysis. Global Journal of Engineering and Technology Advances, 3(1), 015-026.

Shanthi, S., & Ramani, R. G. 2012, October. Feature relevance analysis and classification of road traffic accident data through data mining techniques. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 24-26). sn.

Sohn, S. Y., & Lee, S. H. 2003. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. Safety Science, 41(1), 1-14.

Sun, J., & Sun, J. 2016. Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model. IET intelligent transport systems, 10(5), 331-337.

Yannis, G., Dragomanovits, A., Laiou, A., La Torre, F., Domenichini, L., Richter, T., ... & Karathodorou, N. 2017, October. Road traffic accident prediction modelling: a literature review. In Proceedings of the institution of civil engineers-transport (Vol. 170, No. 5, pp. 245-254). Thomas Telford Ltd.

Yu, R., & Abdel-Aty, M. 2013. Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis & Prevention, 51, 252-259.

Zhang, X., Waller, S. T., & Jiang, P. 2020. An ensemble machine learning-based modeling framework for analysis of traffic crash frequency. Computer-Aided Civil and Infrastructure Engineering, 35(3), 258-276.

Zhang, Z., Yang, W., & Wushour, S. 2020. Traffic accident prediction based on LSTM-GBRT model. Journal of Control Science and Engineering, 2020.

Zheng, J., & Huang, M. 2020. Traffic flow forecast through time series analysis based on deep learning. IEEE Access, 8, 82562-82570.

Zou, X., Vu, H. L., & Huang, H. 2020. Fifty years of Accident Analysis & Prevention: a bibliometric and scientometric overview. Accident Analysis & Prevention, 144, 105568.