Research, Society and Development, v. 11, n. 15, e120111536833, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i15.36833

# Binary logistic regression model applied to data on accidents occurred on federal highways in Brazil

## Modelo de Regressão logística binária aplicada a dados de acidentes em rodovias federais no Brasil

## Modelo de Regresión logística binaria aplicado a datos de accidentes en carreteras federales en Brasil

Received: 10/21/2022 | Revised: 11/02/2022 | Accepted: 11/05/2022 | Published: 11/12/2022

**Damião Flávio dos Santos**
ORCID: https://orcid.org/0000-0001-9411-1403
Universidade de Brasília, Brasil
E-mail: d.flaviostate@gmail.com
**Yuri Machado de Souza**
ORCID: https://orcid.org/0000-0001-5569-8547
Universidade de São Paulo, Brasil
E-mail: yurimachadodesouza95@gmail.com

**Abstract**
Accidents on federal highways in Brazil lead to social and economic impacts on the country. Data from the Federal Highway Police reveal that thousands of people lose their lives in these accidents year after year. This paper aims to examine the factors that influence the probability of death based on the occurrence of the accident. The estimation of a binary logistic regression model took place, in which the event of interest is the circumstance of death in an accident with data from 2021. Following variable selection procedures, it was possible to obtain the final model, which was later validated with data from 2022. The accuracy of the model for both 2021 and 2022 data was around 70%. Then, the odds ratio was calculated between some distinct categories, and how much of an increase in accident lethality it generates compared to the reference category. For example, in a crash, a pedestrian is 15.6 times more likely to die when compared to the driver, while a cyclist is 5.3 times more likely to die. Although most accidents have a human cause, some results show the need of public policies that can help reduce these tragedies. To explain the model, a dashboard was created in a way that the user is able to obtain the probability of death by selecting specific accident characteristics and those involved.
**Keywords:** Supervised analysis; Machine learning; Odds ratio; Lethality of accidents; Highway accidents.

**Resumo**
Os acidentes ocorridos em rodovias federais no Brasil geram impactos sociais e econômicos para o país. Dados da Polícia Rodoviária Federal revelam que, ano após ano, milhares de pessoas perdem suas vidas nesses acidentes. Este trabalho objetiva explorar os fatores que influenciam a probabilidade de óbito a partir da ocorrência do acidente. Foi estimado um modelo de regressão logística binária, em que o evento de interesse é a circunstância de óbito em um acidente com dados de 2021. Atendendo alguns procedimentos de seleção de variáveis, foi obtido a modelo final e, em seguida, feita uma validação com dados de 2022. A eficiência global do modelo, tanto nos dados de 2021 quanto em 2022, ficou em torno de 70%. Em seguida, foi calculada a razão de chances entre algumas categorias distintas e o quanto gera de aumento na letalidade do acidente em relação à categoria de referência – como o pedestre, que tem 15,6 vezes mais chance de letalidade do que o condutor em um acidente, assim como o uso de bicicleta, que tem 5,3 vezes mais chances do que o automóvel. Apesar de a maioria dos acidentes ter causa humana, alguns resultados demonstram que existe a necessidade de intervenção por parte de políticas públicas que podem ajudar na redução dessas tragédias. Para tornar mais concreto e dinâmico o entendimento do modelo, foi elaborado um *dashboard* para que o usuário obtenha a probabilidade de óbito por meio da seleção de determinadas características do acidente e dos envolvidos.
**Palavras-chave:** Análise supervisionada; Aprendizado de máquina; Razão de chances; Letalidade dos acidentes; Acidentes rodoviários.

**Resumen**
Los accidentes en las carreteras federales de Brasil generan impactos sociales y económicos para el país. Datos de la Policía Federal de Caminos revelan que, año tras año, miles de personas pierden la vida en estos accidentes. Este trabajo tiene como objetivo explorar los factores que influyen en la probabilidad de muerte por la ocurrencia del accidente. Se estimó un modelo de regresión logística binaria, en el que el evento de interés es la circunstancia de muerte en accidente con datos de 2021. Dados algunos procedimientos de selección de variables, se obtuvo el modelo

final y luego se validó con datos de 2022. La eficiencia global del modelo, tanto en datos de 2021 como de 2022, rondaba el 70%. Luego, se calculó la razón de posibilidades entre algunas categorías diferentes y cuánto genera un aumento en la letalidad de accidentes en relación con la categoría de referencia, como el peatón, que tiene 15,6 veces más posibilidades de letalidad que el conductor en un accidente, así como como el uso de una bicicleta, que es 5,3 veces más probable que un coche. Aunque la mayoría de los accidentes son causados por el hombre, algunos resultados muestran que existe la necesidad de una intervención de políticas públicas que ayuden a reducir estas tragedias. Para hacer más concreta y dinámica la comprensión del modelo, se creó un tablero para que el usuario obtenga la probabilidad de muerte seleccionando ciertas características del accidente y los involucrados.
**Palabras clave:** Análisis supervisado; Aprendizaje automático; Razón de probabilidades; Letalidad de los accidentes; Accidentes de carretera.

## 1. Introduction

Worldwide, traffic accidents are one of the main causes of death, being the main cause among young people between the ages of 15 and 29, a fact that goes beyond the burden that traffic injuries and deaths represent for national economies and for families (WHO, 2016).

In Brazil, the high number of accidents that occur both in traffic in cities and federal and state highways has caused considerable damage, and public health policies could help avoid these tragedies. Every year, accidents on federal highways are the cause of death of thousands of people, often caused by human error, road failure, and mechanical failure, among other reasons.

According to the data from the Federal Highway Police - PRF (2021), 64,515 accidents occurred on federal highways, 1.6% more than the previous year. In addition, the data show that there were 5,395 deaths, and from these data, the National Transport Confederation - CNT (2021) estimates that these events impacted in a loss of around 12.19 billion reais for the country, of which about 4.7 billion were costs related to accidents with fatal victims.

It is important to emphasize that since 2007, PRF has adopted an open data policy and provides information on accidents on federal highways. By definition, and according to information from the PRF website (2021), an accident is an event that occurs on a highway or federal road domain that involves a vehicle, it is not premeditated and results in material damage to public or private property or injuries to people.

When considering the relevance of these accidents and especially the deaths caused, some authors, such as Roquim et al. (2019), Miranda et al. (2021) and Junior et al. (2019), analyzed and explored these data aiming to identify factors that increase the probability of occurrence of these events.

By considering the accident outcome as a response variable that can be “death occurrence” or “non-death occurrence” and by incorporating other variables in the study, it is possible, through a binary logistic regression model, to explain the probability occurrence of both outcomes.

This article presents an application of the binary logistic regression model that can estimate the probability of death in accidents on federal highways in Brazil with data from 2021 and that is able to predict the occurrence with data from the first quarter of 2022. After estimating the model, a dashboard was built using Power BI, allowing the estimation of the probability of deaths in accidents. The classification of death or non-death considers certain characteristics, specified by the user and by a cut-off value. For the application of such techniques, the free software R version 4.1.2. comes in use (Core Team, 2021).

## 2. Methodology

### 2.1 Binary logistic regression model

According to Fávero and Belfiore (2017), a binary logistic regression aims to study the probability of occurrence of an event defined by Y that is presented in a qualitative dichotomous form (Y=1 to describe the occurrence of the event of interest

and Y =0 to describe the occurrence of the non-event), based on the behavior of explanatory variables. In addition, a vector of explanatory variables is defined, with respective parameters estimated using:

$$Z_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \cdots \beta_k \cdot X_{ki}, \quad (1)$$

Where $Z$ is known as logit, $\alpha$ represents a constant, and $\beta_j$ $(j = 1,2,\cdots,k)$ are the estimated parameters for each explanatory variable.

It is important to highlight that the purpose of the construction of Z is not to represent the dependent variable, but the probability of occurrence of the event of interest. Therefore, it is necessary to consider the concept of chances of occurrence of an event to occur, as the ratio between the probability that the event will occur and the probability that it will not occur.

As seen in Fávero and Belfiore (2017), the binary logistic regression defines the logit Z as the natural logarithm of chance, that is, as follows:

$$\ln\left(\frac{p_i}{1-p_i}\right) = Z_i. \quad (2)$$

As the objective is to define an expression for the probability of occurrence of the event of interest as a function of the logit, by isolating $p_i$ from equation 2, after algebraic procedures, we have that the probability of the event occurring is defined by $p_i = \frac{e^{(Z_i)}}{1+e^{(Z_i)}}$ and therefore the probability of the event not occurring is defined by $1 - p_i = 1 - \frac{e^{(Z_i)}}{1+e^{(Z_i)}}$.

By replacing $Z_i$ for expression. 1, it is possible to obtain that the probability of occurrence of the event is given by:

$$p_i = \frac{e^{(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}{1+e^{(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}} = \frac{1}{1+e^{-(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}, \quad (3)$$

and the probability of occurrence of the non-event is given by:

$$1 - p_i = 1 - \frac{e^{(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}{1+e^{(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}} = \frac{1}{1+e^{(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}. \quad (4)$$

**2.2 Parameter estimators by the maximum likelihood estimation**

If $Y \sim$ Bernoulli $(p)$, its probability distribution function is defined as:

$$P(Y_i) = p_i^{Y_i}(1-p_i)^{1-Y_i}, \quad (5)$$

in which $Y$ assumes values 0 (nonoccurrence of the event) or 1 (occurrence of the event) and $p_i$ is the probability of occurrence of the event of interest.

Thus, when considering the maximum likelihood estimators (MLE), we have:

$$L(Y_i,p) \propto \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i}. \quad (6)$$

When calculating optimization, it is common to work with the logarithm of the likelihood function. Therefore, when applying the logarithm in (5), we have:

$$\log\left(L(Y_i,p)\right) = \sum_{i=1}^{n} y_i \log(p) + \left(n - \sum_{i=1}^{n} y_i\right)\log(1-p). \quad (7)$$

By replacing the parameter p with equation (2), we have the following equation:

$$\log(L(Y_i,\theta)) = \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}\right) + \left(n - \sum_{i=1}^{n} y_i\right) \log\left(\frac{1}{1+e^{(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}\right) + C \quad (8)$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}\right) + (1-y_i) \log\left(\frac{1}{1+e^{(\alpha+\beta_1 \cdot X_{1i}+\beta_2 \cdot X_{2i}+\cdots\beta_k \cdot X_{ki})}}\right)\right] + C.$$

in which $\theta = (\alpha, \beta)$ and $C$ is a constant that does not depend on $\theta$.

The estimators MLE are the values of $\theta$ that maximize $L(\theta)$ or, equivalently, the logarithm of $L(\theta)$, which are estimated by solving the system of equations:

$$U(\theta) = \frac{\partial log L(\theta)}{\partial \theta} = 0.$$

To find the solution of this system of equations for a specific dataset, it is necessary to use a numerical method, usually the Newton-Raphson method, an algorithm developed by McCullagh and Nelder (1989) that, using a statistical package, is possible to be applied.

For the development of this research, the free software R in version 4.1.2. (Core Team, 2021) came in use, through the *glm()* function of the stats package (Core Team, 2021).

### 2.3 Significance of variable effects

Once the estimates of the parameters of the binary logistic regression model were obtained, it was necessary to evaluate the adequacy of the adjusted model. According to Giolo (2017), the principle in logistic regression is the same used in linear regression, that is, comparing the observed values of the response variable with the values predicted by the models with and without the variable under investigation.

#### 2.3.1 Test $\chi^2$ (chi-squared test)

According to Fávero and Belfiore (2017), the chi-square test provides conditions to verify the significance of the model, since the null and alternative hypotheses for a general logistic regression model are, respectively:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

$H_1$: There is at least one $\beta_j \neq 0$.

Therefore, the test performs an initial check on the existence of the proposed model, because, if all the estimated parameters $\beta_j$ $(j = 1,2,\cdots,k)$ are statistically equal to 0, the behavior of alteration of each of the explanatory variables does not influence at all the probability of occurrence of the event of interest. The statistic $\chi^2$ presents the following expression:

$$\chi^2 = -2(LL_0 - LL_{max}), \quad (9)$$

In which $LL_0$ is the result of the log likelihood of the null model (without the presence of explanatory variables) and $LL_{max}$ is the result of the log likelihood of the complete model. By comparing the level of significance $\alpha = 0.05$ previously stablished,

the $p-value$ of the test defines the decision criterion for the chi-square. If $p-value$ is less than 0.05, then there is at least one $\beta_j \neq 0$; otherwise, all $\beta_j = 0$, which means that none of the explanatory variables are significant for the model.

### 2.3.2 Wald Z test

According to Fávero and Belfiore (2017), the chi-square test assesses the joint significance of the explanatory variables, without defining which of these variables considered in the model are statistically significant to influence the probability of occurrence of the event. Thus, after verifying that at least one $\beta_j \neq 0$, the application of the test known as Wald was possible (Wald, 1943). The expressions for calculating Wald Z statistics for each parameter $\alpha$ and $\beta_j$ are, respectively:

$$Z_\alpha = \frac{\alpha}{s.e(\alpha)} \quad \text{and} \quad Z_{\beta_j} = \frac{\beta_j}{s.e(\beta_j)}, \tag{10}$$

in which $s.e$ is the standard error of the estimate of each parameter of the analyzed model.

According to Carvalho (2011), under the null hypothesis, we have that the Z statistic follows a standard normal distribution. Then, for the conclusion of the test, the value of Wald $Z$ statistic was compared with the critical value of the normal distribution table, when considering a certain level of significance $\alpha$ previously established. Furthermore, as with the chi-square test, $p-value$ offers a conclusion, by comparing the level of significance. Thus, when considering $\alpha = 0.05$, the estimated parameter will be significant if $Z < -1.96$ ou $Z > 1.96$, or even if $p-value < 0.05$.

### 2.3.3 Odds ratio

According to Giolo (2017), the chance of an event of interest occurring is the ratio between the probability of the event occurring and the event not occurring. Thus, in cohort studies, individuals exposed and not exposed to a factor of interest are monitored over time, to observe how many of them develop the disease, for example.

Therefore, it is possible to define the odds ratio as the ratio between the chance of the disease occurring among those exposed and the chance of the disease occurring among the unexposed, as follows:

$$OR = \frac{p(x=1)/[1-p(x=1)]}{p(x=0)/[1-p(x=0)]}, \tag{11}$$

and substituting the expressions of the binary logistic regression model, it was possible to obtain the following expression:

$$OR = \frac{\left(\frac{e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})}}{1+e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})}}\right)\Big/\left(\frac{1}{1+e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})}}\right)}{\left(\frac{e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})}}{1+e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})}}\right)\Big/\left(\frac{1}{1+e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})}}\right)} \tag{12}$$

$$= \frac{e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})}}{e^{\hat{\alpha}}} = e^{(\hat{\alpha}+\hat{\beta_1}\cdot x_{1i}+\hat{\beta_2}\cdot x_{2i}+\cdots\hat{\beta_k}\cdot x_{ki})-\hat{\alpha}} = e^{\hat{\beta_1}}.$$

Also, according to Giolo (2017), it is possible to obtain the confidence interval for the OR, at the confidence level $100(1-\alpha)\%$:

$$IC(OR)_{100(1-\alpha)\%} = e^{\left(\hat{\beta_1} \pm z_\alpha \times s.e(\hat{\beta_1})\right)}, \tag{13}$$

in which $z_{\alpha/2}$ shows the $100(1-\alpha)\%$ percentile of the standard normal distribution and $s.e$ is the standard error of the estimate of $\beta_1$.

**2.4 Model selection**

**2.4.1 Likelihood-ratio test**

According to Colosimo and Giolo (2006), the likelihood ratio test involves comparing the values of the logarithm of the maximized likelihood function without restriction and under $H_0$, that is, the comparison of $log\ L(\hat{\theta})$ (complete model) and $log\ L(\theta_0)$ (reduced model). The statistic for the test is:

$$LRT = -2log\left[\frac{L(\theta_0)}{L(\hat{\theta})}\right] = 2[log\ L(\hat{\theta}) - log\ L(\theta_0)] \quad (14)$$

in which sob $H_0{:}\theta = \theta_0$ follows approximately a chi-square distribution with $p$ degrees of freedom. Also, according to the authors, we emphasize that, for large samples, $H_0$ is rejected at a significance level $\alpha$, if $LRT > \chi^2_{p,1-\alpha}$.

**2.4.2 Akaike information criterion (AIC)**

The method proposed by Akaike (1974) is known as the Akaike information criterion (AIC). According to Santos (2017), its basic idea is to select a model that is parsimonious, that is, that is well adjusted and has a reduced number of parameters. As the logarithm of the likelihood function grows with the increase in the number of model parameters, in this way, we expect to find the model with the lowest value for the function:

$$AIC = -2\ log\ L(\hat{\theta}) + 2k, \quad (15)$$

in which $k$ is the number of parameters of the adjusted model.

**2.4.3 Akaike Information Criterion Fixed (AICc)**

Sugiura (1978) proposed a correction of the AIC criterion, as the AIC may perform poorly if there are too many parameters compared to the sample size. Thus, the AICc is just a second-order correction of the AIC bias, given by the following expression:

$$AICc = -2\ log\ L(\hat{\theta}) + 2k + 2\frac{k(k+1)}{n-k-1}, \quad (16)$$

in which $k$ is the number of model parameters to be estimated and $n$ is the sample observation number.

**2.4.4 Bayesian information criterion (BIC)**

Proposed by Schwarz (1978), the Bayesian information criterion (BIC) is given by:

$$BIC = -2\ log\ L(\hat{\theta}) + klog(n), \quad (17)$$

in which $k$ is the number of model parameters to be estimated and n is the number of observations in the sample.

### 2.5 Sensitivity analysis

To verify the quality of an adjusted model, it is possible to evaluate some measures, such as sensitivity and specificity. According to Giolo (2017), it is necessary to establish a probability, called cutoff point, from which it is established that the response variable receives value 1 for probabilities predicted by the model greater than or equal to this cutoff point and value 0, otherwise. From this definition, it is possible to build a double-entry table between the predicted values and the actual values, known as a confusion matrix, as shown in Table 1:

**Table 1 -** Confusion matrix for observed and predicted values on a given cut-off point.

| Response predicted by the model | Observed Response | | Total |
|---|---|---|---|
| | $Y = 1$ | $Y = 0$ | |
| $Y = 1$ | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| $Y = 0$ | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| Total | $n_{.1}$ | $n_{.2}$ | $n$ |

Sour**e:** Adapted from Giolo (2017)

Thus, three very important measures are defined for the diagnosis of the model, these being "Sensitivity", "Specificity" and "Accuracy".

#### 2.5.1 Sensitivity

As seen in Fávero and Belfiore (2017), sensitivity concerns the success rate for a given cutoff, only considering observations that are an event of interest, therefore:

$$sensitivity = \frac{n_{11}}{n_{.1}} \times 100\%. \quad (18)$$

#### 2.5.2 Specificity

According to Fávero and Belfiore (2017), specificity, on the other hand, refers to the success rate, for a given cutoff point, only considering observations that are not an event of interest, therefore:

$$specificity = \frac{n_{22}}{n_{.2}} \times 100\%. \quad (19)$$

#### 2.5.3 Accuracy

Fávero and Belfiore (2017) define the accuracy of the model as a percentage of correct classification for a given cutoff point, which is given by the sum of the main diagonal of the confusion matrix divided by the sample size, therefore:

$$Overall\ model\ efficiency = \frac{n_{11}+n_{22}}{n} \times 100\%. \quad (20)$$

### 2.6 ROC curve

Since the cutoff value influences the sensitivity value, specificity and, consequently, the accuracy, according to Giolo (2017) it is necessary to identify the cutoff value that produces the highest percentage of correct answers. The Receiver Operating Characteristic – ROC curve, better known by its acronym, is usually used for this purpose.

To obtain the ROC curve, pairs of points (x,y) = (1-specificity, sensitivity) are plotted for various cut-off points. The model with perfect discrimination corresponds to the one without sensitivity and specificity equal to 1, which implies that

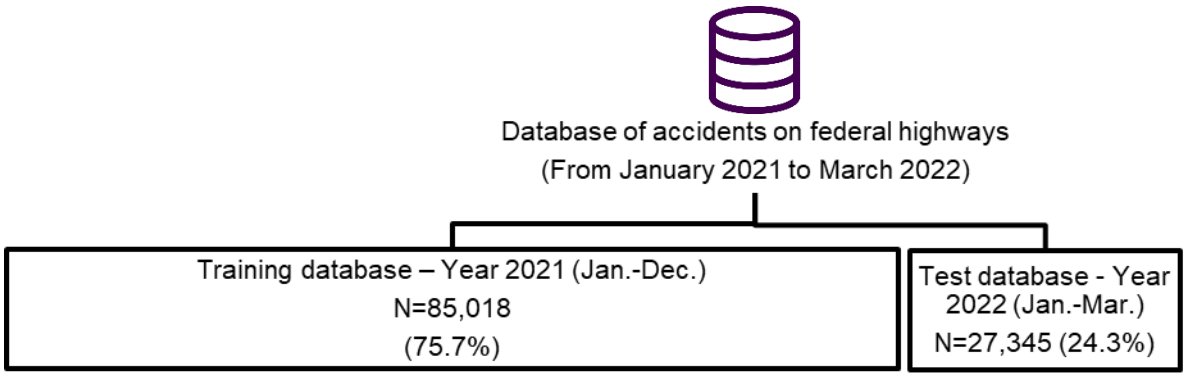
(x,y) = (0,1). Thus, cut-off points located near the upper left corner of the graph will indicate that the fitted model produces the highest success rate, whether in terms of true positives or true negatives. Also, the closer to 1 the area under the ROC curve is, the better the predictive power of the model (Giolo, 2017).

**2.7 Cross validation**

This step consists of dividing the database into two parts: training and testing. This separation can be random, but it is also an option to assign the observation time to this function. Cross-validation represents one of the most used in "machine learning" problems and, as it is widely used in the literature, we suggest the division between 70% and 80% for the training base and 30% to 20% for the test, as seen in Izbicki and Santos (2020).

After this division, the development of the model is initiated, as described above, but only using the training database. Then, the model's accuracy on the test base is evaluated. In this article, the model was developed with data from 2021 and validated it with data from January to March 2022, as shown in Figure 1.

**Figure 1 -** Database segmentation flowchart in training and testing for model validation.




Source: Prepared by the authors based on research data (2021)

**2.8 Database for application**

The data used in this article comes from the official Federal Highway Police database (available since 2007), referring to accidents on federal highways in Brazil that occurred in the year 2021. This is the dataset used to estimate the parameters of the binary logistic regression model, and then tested using data from the first quarter of 2022.

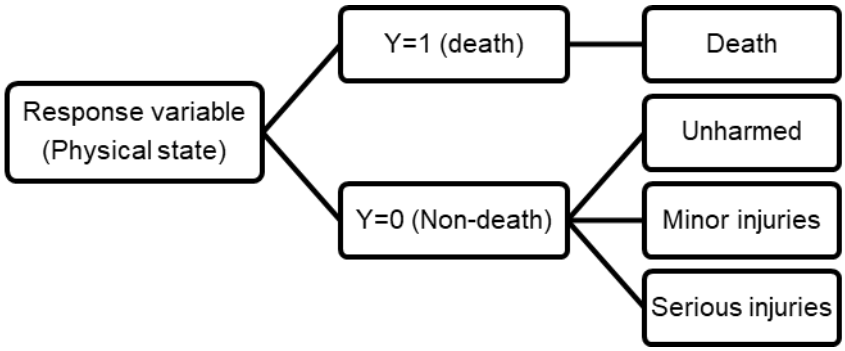
There is a high number of fatal accidents every year, and 2021 was no different. In total, in 2021, 64,515 accidents were recorded, with 5,395 people killed in these accidents, which means that for every 100 accidents there were at least 8 deaths. It is noteworthy that, in these data, there is a lot of incomplete information regarding both the explanatory variables and the response variable. Thus, only valid responses in all variables of interest were considered, adding to a total of 85,018 people involved in accidents, with 81,974 people who did not die and 3,044 deaths. That is, the event of interest occurred in 3.6% of people who were involved in accidents on federal highways in Brazil.

From January to March 2022, there were 14,958 accidents, with 35,895 people involved and 1,280 deaths. As in 2021, there were some variables with missing values and, after their exclusion, the database had 27,345 observations, with 26,374 people who did not die and 971 deaths, which corresponds to 3.55% of people involved in accidents, close to the same percentage observed in 2021.

### 2.8.1 Response variable

The response variable for this work is part of the accident database and is indicated as the physical state of the person involved. In the original database, this variable is presented with 4 levels of physical status (Unharmed, Minor Injuries, Serious Injuries and Death). As the objective of this work is to estimate the probability of death, it was important to perform a reclassification of the variable, as shown in Figure 2:

**Figure 2 -** Flowchart of the reclassification of the response variable "Physical state of those involved".

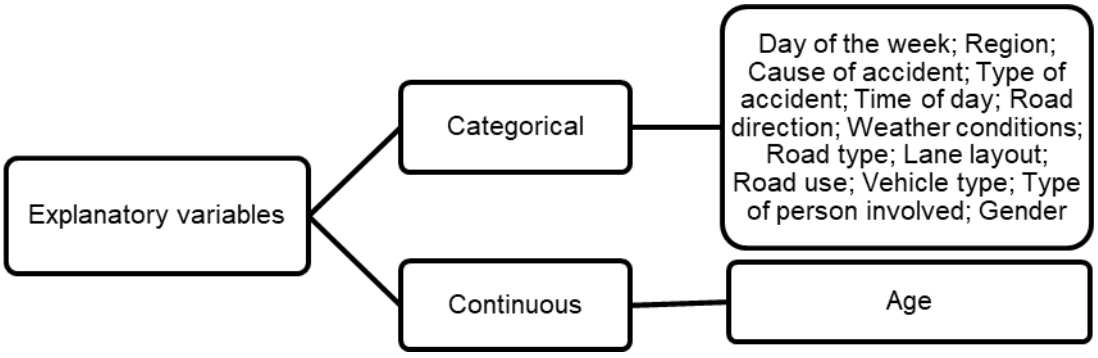



Source: Prepared by the authors based on research data (2021).

### 2.8.2 Explanatory variables

The explanatory variables available in the PRF accident database underwent transformations and adjustments due to the presence of variables with many categories. Most of them are categorical variables and will need to be transformed into dummy variables. The variables used to explain the model were those described in Figure 3 and their categories are described in Table 3.

**Figure 3**: Flowchart of the composition of categorical and continuous explanatory variables.

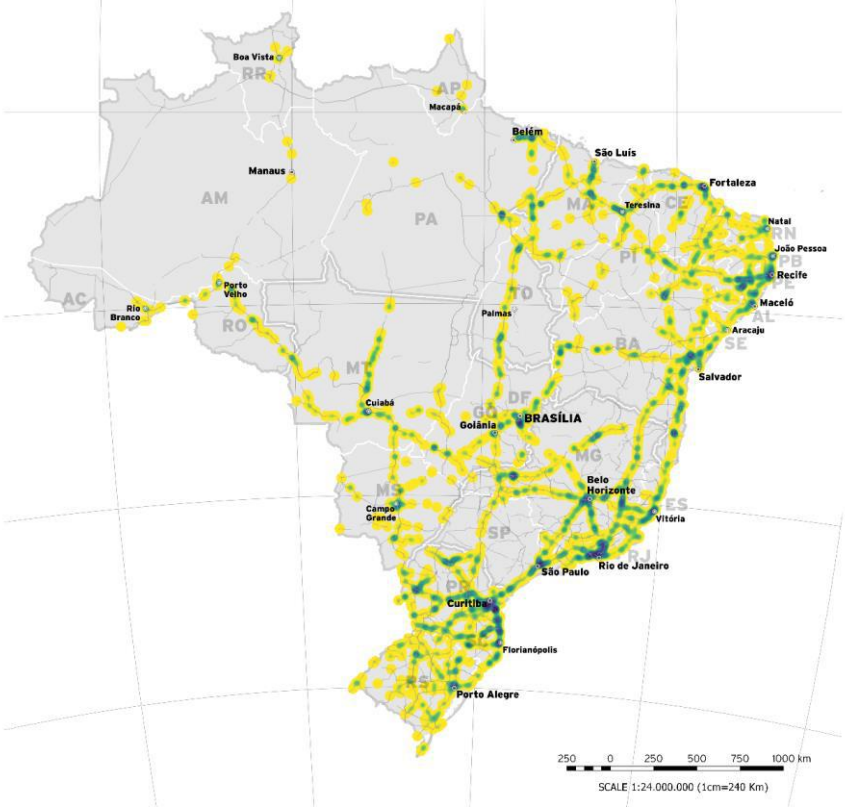



Source: Prepared by the authors based on research data (2021).

## 3. Results and Discussion

Initially, it was essential to conduct a spatial analysis of deaths occurred in accidents on federal highways in Brazil. By the density of the heat map, a higher concentration of deaths on the highways of the Northeast, Southeast and South of the country is clear, in addition, it was possible to observe a greater concentration near the large cities in the country. Thus, the location of the accident can be a preponderant factor for the occurrence of death.

**Figure 4 -** Heat map for the location of deaths in accidents on federal highways in Brazil in 2021.




Source: Data from the Federal Highway Police (2021).

According to the descriptive analysis of the data, for the only numerical variable, the main descriptive measures were calculated according to the physical state of those involved in accidents.

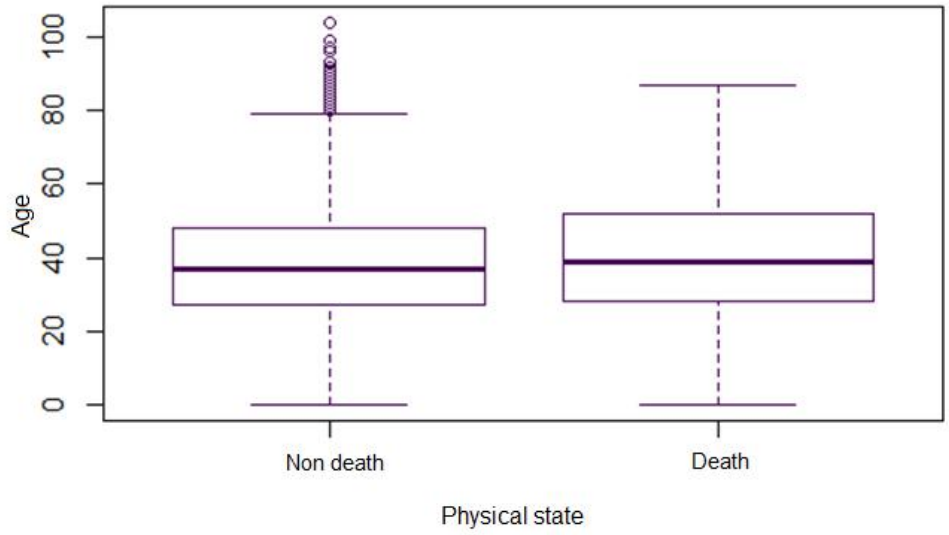
As seen in Table 2 and Figure 5, the mean and median age among people who died is at least 2 years older than the people who survived the accident. This descriptive difference between the two groups gives evidence that age may be a variable that can explain the probability of death.

**Table 2 -** Descriptive measures for the age of those involved in accidents on federal highways in Brazil in 2021, according to their physical state after an accident.

| Categories | Physical state | |
|---|---|---|
| | Non-death | Death |
| Mean | 38.2 | 40.5 |
| Mode | 35.0 | 40.0 |
| Median | 37.0 | 39.0 |
| Standard deviation | 14.6 | 15.9 |
| Minimum | 0 | 0 |
| Maximum | 104 | 87 |

Source: Federal Highway Police data (2021).

**Figure 5 -** Box plots of age, according to the physical state of those involved in accidents on federal highways in Brazil in 2021.




Source: Data from the Federal Highway Police (2021).

Table 3 presents a bivariate analysis through the joint distribution between the categorical explanatory variables and the variable "physical state" of those involved. It is relevant to highlight important relevant points in this descriptive analysis, such as: proportionally more deaths occur during the weekend, in addition to that, considering that the weekend consisted of Saturday and Sunday, the average number of deaths per day was 608,5 deaths; during the week, it was 365,4.

The highest percentages of individual deaths are among men; from the Northeast and North regions; having human error as the cause of the accident; with being run over as the type of accident; during the morning; when the weather presented fog; in single lane; in a curve; having a bicycle as a vehicle involved; and pedestrian with a high percentage of deaths when compared with the other types of people involved. Thus, these characteristics may have an effect on the binary logistic regression model that will calculate the probability of death on federal highways in Brazil.

**Table 3 -** Joint distribution of frequencies and proportions (in percentage), according to the explanatory variables and the physical state of those involved in accidents.

| Variables | Categories | Physical state | |
|---|---|---|---|
| | | Non-death | Death |
| | | N (%) | N (%) |
| Day of the week | Weekend | 27,440 (95.8%) | 1,217 (4.2%) |
| | Weekday | 54,534 (96.8%) | 1,827 (3.2%) |
| Road direction | Ascending | 43,617 (96.3%) | 1,677 (3.7%) |
| | Descending | 38,357 (96.6%) | 1,367 (3.4%) |
| Road use | Urban | 37,669 (98.0%) | 777 (2.0%) |
| | Rural | 44,305 (95.1%) | 2,267 (4.9%) |
| Sex | Female | 18,672 (97.4%) | 489 (2.6%) |
| | Male | 63,302 (96.1%) | 2,555 (3.9%) |
| Region | Midwest | 10,390 (96.4%) | 390 (3.6%) |
| | Northeast | 16,992 (94.6%) | 979 (5.4%) |
| | North | 4,810 (95.7%) | 214 (4.3%) |
| | Southeast | 25,096 (97.2%) | 731 (2.8%) |
| | South | 24,686 (97.1%) | 730 (2.9%) |

| | | | |
|---|---|---|---|
| Cause of accident | Human error | 69,446 (96.3%) | 2,691 (3.7%) |
| | Mechanical failure | 4,601 (98.0%) | 92 (2.0%) |
| | Road failure | 4,996 (96.6%) | 178 (3.4%) |
| | Other causes | 2,931 (97.2%) | 83 (2.8%) |
| Type of accident | Run over | 4,028 (89.6%) | 469 (10.4%) |
| | Overturning | 6,595 (96.9%) | 210 (3.1%) |
| | Collision | 54,985 (96.7%) | 1,884 (3.3%) |
| | Runway exit | 10,050 (96.1%) | 403 (3.9%) |
| | Others | 6,316 (98.8%) | 78 (1.2%) |
| Time of day | Morning | 3,351 (95.1%) | 173 (4.9%) |
| | Day | 45,998 (97.3%) | 1,274 (2.7%) |
| | Evening | 5,100 (96.8%) | 171 (3.2%) |
| | Night | 27,525 (97.3%) | 1,426 (2.7%) |
| Weather conditions | Rain | 7,216 (96.2%) | 284 (3.8%) |
| | Fog | 695 (93.7%) | 47 (6.3%) |
| | Cloudy | 12,861 (96.3%) | 493 (3.7%) |
| | Sun | 5,946 (97.5%) | 154 (2.5%) |
| | Others | 55,256 (96.4%) | 2,066 (3.6%) |
| Lane type | Simple | 41,134 (95.1%) | 2,125 (4.9%) |
| | Two-way | 33,506 (97.7%) | 791 (2.3%) |
| | Multiple | 7,334 (98.3%) | 128 (1.7%) |
| Lane layout | Curve | 12,907 (94.9%) | 687 (5.1%) |
| | Straight | 57,952 (96.5%) | 2,104 (3.5%) |
| | Others | 11,115 (97.8%) | 253 (2.2%) |
| Vehicle type | Car | 45,058 (97.2%) | 1,300 (2.8%) |
| | Bicycle | 1,138 (90.5%) | 119 (9.5%) |
| | Truck | 13,021 (96.7%) | 446 (3.3%) |
| | Motorcycle | 18,916 (94.9%) | 1,008 (5.1%) |
| | Bus | 2,343 (96.7%) | 81 (3.3%) |
| | Others | 1,498 (94.3%) | 90 (5.7%) |
| Type of person involved | Driver | 60,260 (96.8%) | 2,016 (3.2%) |
| | Passenger | 20,419 (97.1%) | 601 (2.9%) |
| | Pedestrian | 1,274 (75.0%) | 425 (25.0%) |
| | Others | 21 (91.3%) | 2 (8.7%) |

Source: Data from the Federal Highway Police (2021).

As described in equation (9), the chi-square test is performed from the value of the log likelihood of the complete model (with all explanatory variables) and using the model only with the intercept (α). The value of the log likelihood of the complete model was $LL_{max} = -11{,}461$. The null model was $LL_0 = -13{,}124$. The hypotheses for the chi-square test are as follows:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{35} = 0,$$

$H_1$: There is at least one $\beta_j \neq 0$.

According to the description in Table 4, when considering a significance level α = 0.05, the null hypothesis is rejected (in which all $\beta_j$ statistically equal to 0), that is, there is at least one $\beta_j \neq 0$. This means that a binary logistic regression model with the data analyzed in this study is possible.

**Table 4 -** Chi-square test for the existence of a binary logistic regression model.

| $\chi^2$ | g.l. | P-value |
|---|---|---|
| 3,326.46 | 35 | $< 2 \times 10^{-16}$ |

Source: Prepared by the authors according to research data (2021)

Table 5 shows the estimates for the full model. With the stepwise procedure, it was possible to estimate a more parsimonious model, that is, with fewer explanatory variables used to estimate the probability of death in accidents on federal highways in Brazil.

With Wald Z test, it was possible to observe that the complete model presents some explanatory variables that were not significant at the 5% significance level, such as: direction of the road (ascending); region (North); type of accident (overturning); time of day (night); weather conditions (fog); road type (multiple); type of vehicle (truck and bus); and type of person involved (passenger and others). Thus, even removing these explanatory variables, using the stepwise procedure, it is necessary to verify the efficiency of this reduced model compared to the full model, through some measures and hypothesis testing, as shown in Table 5.

**Table 5 -** Estimates and significance test for the full and reduced model.

| Variables | Complete model | | | Reduced model | | |
|---|---|---|---|---|---|---|
| | Estimate | Z | P-value | Estimate | Z | P-value |
| A | -3.47582 | -17.93 | $< 2 \times 10^{-16}$ | -3.52054 | -27.468 | $< 2 \times 10^{-16}$ |
| Day of the week (ref.: Weekend) | | | | | | |
| Weekday | -0.17376 | -4.347 | $1.4 \times 10^{-5}$ | -0.17873 | -4.499 | $6.8 \times 10^{-6}$ |
| Road direction (ref.: Ascending) | | | | | | |
| Descending | -0.06862 | -1.777 | 0.07556 | - | - | - |
| Road use (ref.: Rural) | | | | | | |
| Urban | -1.04585 | -21.741 | $< 2 \times 10^{-16}$ | -1.04626 | -21.988 | $< 2 \times 10^{-16}$ |
| Sex (ref.: Female) | | | | | | |
| Male | 0.400062 | 6.895 | $5.4 \times 10^{-12}$ | 0.37018 | 7.113 | $1.1 \times 10^{-12}$ |
| Region (ref.: Midwest) | | | | | | |
| Northeast | 0.28606 | 4.469 | $7.9 \times 10^{-6}$ | 0.27062 | 4.896 | $9.8 \times 10^{-7}$ |
| North | 0.03647 | 0.402 | 0.68795 | - | - | - |
| Southeast | -0.33690 | -5.012 | $5.4 \times 10^{-7}$ | -0.35234 | -5.959 | $2.5 \times 10^{-9}$ |
| South | -0.24868 | -3.721 | 0.00020 | -0.26061 | -4.453 | $8.5 \times 10^{-6}$ |
| Cause of accident (ref.: Human error) | | | | | | |
| Mechanical failure | -0.42697 | -3.808 | 0.00014 | -0.43225 | -3.881 | 0.00010 |
| Road failure | -0.34685 | -4.178 | $2.9 \times 10^{-5}$ | -0.34781 | -4.203 | $2.6 \times 10^{-5}$ |
| Other causes | -0.55958 | -4.401 | $1.1 \times 10^{-5}$ | -0.53574 | -4.415 | $1.0 \times 10^{-5}$ |
| Type of accident (ref.: Run Over) | | | | | | |
| Overturning | -0.02259 | -0.17 | 0.864741 | - | - | - |
| Collision | 0.23168 | 2.053 | 0.040067 | 0.25203 | 3.824 | 0.00013 |
| Runway exit | 0.28328 | 2.303 | 0.021289 | 0.30857 | 3.860 | 0.00011 |
| Others | -0.71413 | -4.478 | $7.5 \times 10^{-6}$ | -0.69773 | -5.363 | $8.2 \times 10^{-8}$ |
| Time of day (ref.: Morning) | | | | | | |
| Day | -0.5355 | -6.081 | $1.2 \times 10^{-9}$ | -0.50140 | -12.067 | $< 2 \times 10^{-16}$ |
| Evening | -0.45699 | -3.956 | $7.6 \times 10^{-5}$ | -0.41844 | -4.916 | $8.8 \times 10^{-7}$ |
| Night | -0.04499 | -0.517 | 0.60515 | - | - | - |
| Weather conditions (ref.: Rain) | | | | | | |
| Fog | 0.16492 | 0.957 | 0.33849 | - | - | - |
| Cloudy | -0.20371 | -2.514 | 0.01193 | -0.22824 | -2.955 | 0.00313 |
| Sun | -0.23816 | -2.206 | 0.02736 | -0.25854 | -2.450 | 0.01428 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Others | -0.23663 | -3.401 | 0.00067 | -0.25966 | -3.980 | $6.9 \times 10^{-5}$ |
| Lane type (ref.: Double) | | | | | | |
| Simple | 0.56084 | 12.395 | $< 2 \times 10^{-16}$ | 0.57184 | 1.201 | $< 2 \times 10^{-16}$ |
| Multiple | -0.05264 | -0.528 | 0.59765 | - | - | - |
| Lane layout (ref.: Curve) | | | | | | |
| Others | -0.77572 | -9.856 | $< 2 \times 10^{-16}$ | -0.77069 | -9.806 | $< 2 \times 10^{-16}$ |
| Straight | -0.42537 | -8.575 | $< 2 \times 10^{-16}$ | -0.41846 | -8.495 | $< 2 \times 10^{-16}$ |
| Vehicle type (ref.: Car) | | | | | | |
| Bicycle | 1.651616 | 15.452 | $< 2 \times 10^{-16}$ | 1.66556 | 15.873 | $< 2 \times 10^{-16}$ |
| Truck | -0.09241 | -1.549 | 0.12137 | - | - | - |
| Motorcycle | 0.98353 | 20.619 | $< 2 \times 10^{-16}$ | 1.00223 | 22.603 | $< 2 \times 10^{-16}$ |
| Bus | -0.04364 | -0.36 | 0.71884 | - | - | - |
| Others | 0.382344 | 3.111 | 0.00186 | 0.40558 | 3.333 | 0.00086 |
| Type of person involved (ref.: Driver) | | | | | | |
| Passenger | 0.040402 | 0.718 | 0.47255 | - | - | - |
| Pedestrian | 2.744455 | 22.963 | $< 2 \times 10^{-16}$ | 2.74812 | 32.144 | $< 2 \times 10^{-16}$ |
| Others | 1.116898 | 1.463 | 0.14352 | - | - | - |
| Age | 0.010902 | 8.302 | $< 2 \times 10^{-16}$ | 0.01072 | 8.238 | $< 2 \times 10^{-16}$ |

Source: Prepared by the authors based on research data (2021).

According to the data described above, it is necessary to select a model that is parsimonious, that is, that is well adjusted and has a reduced number of estimated parameters. For this, the information criteria AIC, AICc and BIC came in use, in which the most parsimonious model is the one with the lowest values for these measures, that is, the most adequate model to estimate the probability of occurrence of the phenomenon studied.

Thus, according to Table 6, it was possible to verify that the likelihood ratio test did not reject the null hypothesis between the difference of the models when considering a significance level of 5%. However, despite this result, it does not mean that either one is suitable, given the principle of parsimony. It is important to emphasize that the difference in the number of parameters to be estimated between the models is equal to 10, and this can represent a gain in terms of efficiency of the final model. As seen in Table 6, both AIC and AICc and BIC of the reduced model were smaller than those of the complete model. Thus, the reduced model was chosen as the final model.

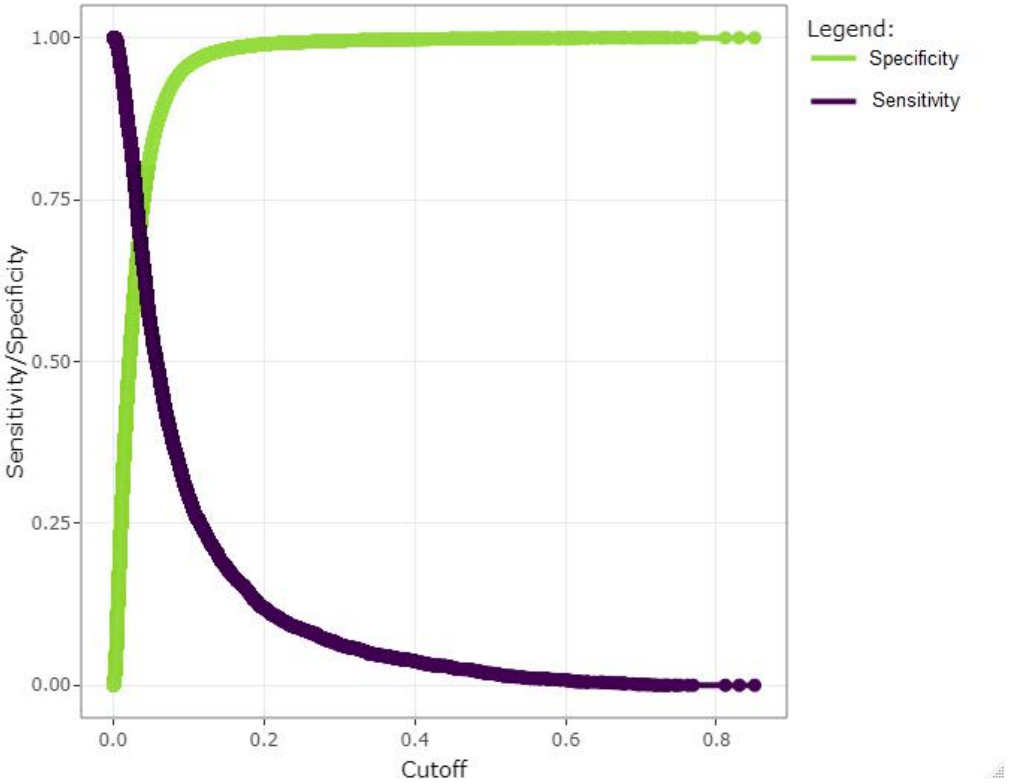
**Table 6 -** Information criterion measures, log likelihood and likelihood ratio test for the two binary logistic regression models.

| Models | AIC | AICc | BIC | LL | TRV | P-value |
|---|---|---|---|---|---|---|
| Complete | 22,994.3836 | 22,994.415 | 23,331.0059 | -11,461.1918 | 9.6971 | 0.46746 |
| Reduced | 22,984.0807 | 22,984.097 | 23,227.1968 | -11,466.0403 | | |

Source: Prepared by the authors based on research data (2021).

After choosing the model, it was important to identify a cutoff value capable of capturing the maximum sensitivity and specificity. That is, that this cutoff value causes the prediction success rate for those that will be an event equal to the success rate for those that will not be an event. For this purpose, the sensitivity curve shows the possible cutoff values on the x-axis and the sensitivity and specificity values on the y-axis.

**Figure 6** - Sensitivity curve for cutoff values




Source: Prepared by the authors based on research data (2021).

When looking at Figure 6, the ideal cut-off value was detected for the classification of events of interest, so that it can capture the highest success rate in sensitivity, specificity, and accuracy. The cutoff value, therefore, was 0.036. Thus, the following classification criteria is suitable:

- If $p_i \geq 0.036$, observation $i$ should classify as death.
- If $p_i < 0.036$, observation $i$ should classify as non-death.

With this cut-off point, the confusion matrix was obtained and, from that, an accuracy above 0.7, which – for the purpose of the study – was considered efficient.

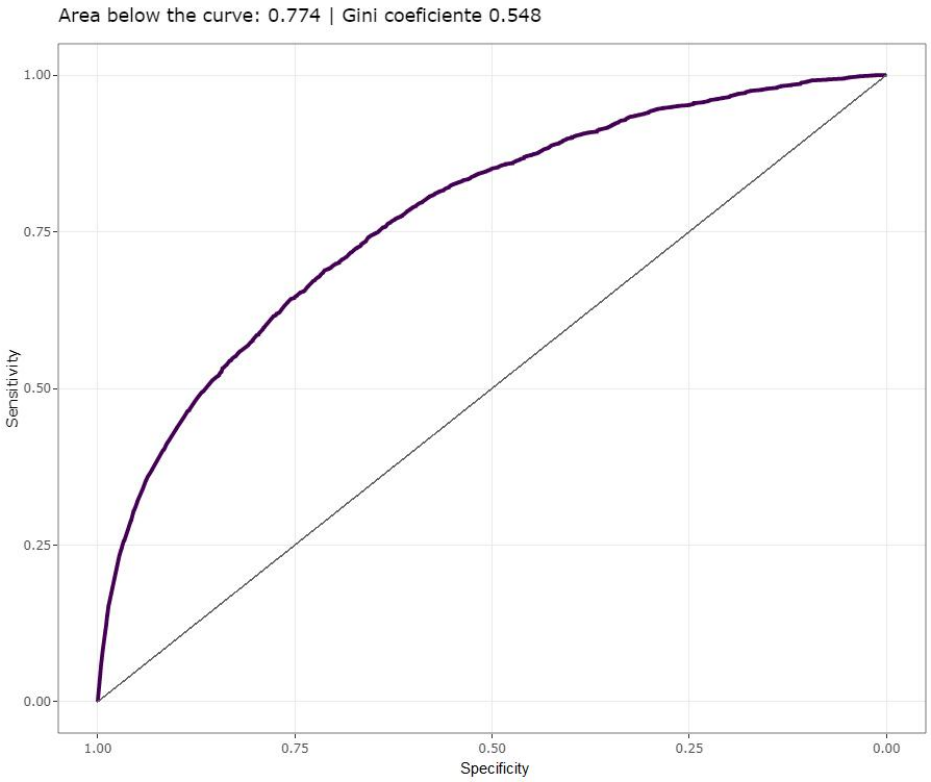
**Table 7 -** Confusion matrix for observed and predicted values under the cutoff value = 0.036.

| Answer predicted by the model | Observed Response | | Total |
|---|---|---|---|
| | Death | Non-death | |
| Death | 2,116 | 24,395 | 26,511 |
| Non-death | 928 | 57,579 | 58,507 |
| Total | 3,044 | 81,974 | 85,018 |

Note: Sensitivity = 0.6951; Specificity = 0.7024; and Accuracy = 0.7021. Source: Prepared by the authors based on research data (2021)

The area under the curve that equals to 0.774 and the Gini coefficient that equals to 0.548 demonstrate that the model was well adjusted and has a very significant predictive power.

**Figure 7** - ROC curve associated with the chosen binary logistic regression model.




Source: Prepared by the authors based on research data (2021)

After adjusting the model, all combinations of the categories of all the variables selected for the model were calculated – that is, 3,456,000 combinations, from which the probability of death was calculated. It is relevant to highlight that for this analysis, for the age variable, as it is continuous, we considered the mean to be 38 years old. Thus, as a way of visualizing the categories that generate higher and lower probabilities of death, the first thousand combinations and the last thousand were selected and then two word clouds were generated, as shown in Figure 8 and Figure 9.

As already mentioned in the descriptive analysis, human errors; single lane accidents; bicycle accidents; with pedestrians involved; in curves; with male victims; during the weekend; in the Northeast in rural areas; increases the probability of death.

On the other hand, accidents in urban areas; with females involved; in the Southeast; on weekdays; during the day; in accidents due to other causes; in other lane layouts; and in double or multiple lanes, have a lower probability of death of the people involved.

**Figure 8 -** Word cloud of the categories that generate the highest probability of death on federal highways in Brazil.


Source**:** Prepared by the authors based on research data (2021).

**Figure 9 -** Word cloud of categories that generate less probability of death on federal highways in Brazil


Source**:** Prepared by the authors based on research data (2021).

Another approach to identify the effect of a given variable on the increase in the probability of death in accidents or the increase in the chance of lethality is the use of the odds ratio, as shown in Table 8.

As already observed in previous analyses, some characteristics of the accidents and the people involved cause an increase in the probability of death in accidents on federal highways in Brazil. The most striking result in this analysis is how much more chance of fatality the pedestrian experiences rather than the driver of the vehicle – 15.6 times more, which points to a 1,461.3% increase in the chance of fatality of the pedestrian in relation to the driver. This result corroborates the study by Junior et al. (2019), who showed with a binary logistic regression model, with data from 2016, that the chance of pedestrian lethality in relation to car occupants is 9.49 times greater.

According to the National Traffic Department-DENATRAN, in April of 2022 the number of licensed drivers in Brazil was 77,917,203, with 64.8% being male drivers and 35.2% female. Although often discriminated against as synonymous of bad drivers, biologically female people are the ones who die less in accidents and suffer fewer accidents, according to the descriptive data of accidents in Table 3. In the estimated model, males are 1.44 times more likely to die than females, which is equivalent to a 44.7% increase in the chance of lethality.

In addition, cyclists and motorcyclists are the profiles with the highest chance of fatality in relation to car drivers, as described in Table 8. The type of lane is also a factor that influences the probability of death, with the single lane increasing the chance of accident lethality by 77.2%. It is relevant to highlight that according to the National Road Traffic System - SNV (2022) in 2022 the road network in Brazil will have a total length of 121,100.9 km, most of which (57,309.5 km or 47.3%) are simple lanes. Therefore, investment in infrastructure, with the maintenance and construction of dual-lane highways, could be a key factor in reducing the number of deaths in accidents in Brazil. It is necessary to have more effective public policies in place, in order to solve this problem that affects socially and financially so many families that are victims of these events.

The negative results for the increase in the chance of lethality, such as -64.9% for urban land use, mean that for the reference category (in this case, rural), there is a reduction in the chance of lethality by 64.9 % if the land use during the accident is considered urban.

In relation to the odds ratio of the continuous variable, for example, between individuals aged 65 years and 18 years, it results in $\widehat{OR} = e^{(65-18)\times 0.01072} = 1.655$. This means that individuals in the age of 65 have a 65.5% higher chance of death than 25 year-olds, in addition, the older the individual is, greater is the risk of death after the accident.

**Table 8 -** Odds ratio associated with the binary logistic regression model.

| Variables | $\widehat{\beta_j}$ | $OR = e^{\widehat{\beta_j}}$ | $s.e(\widehat{\beta_j})$ | $IC(OR)_{95\%}$ | Increased chance of lethality |
|---|---|---|---|---|---|
| Day of the week (ref.: Weekend) | | | | | |
| Weekday | -0.17873 | 0.836332 | 0.039722 | (0.7737;0.9040) | -16.4% |
| Road use (ref.: Rural) | | | | | |
| Urban | -1.04626 | 0.351249 | 0.047582 | (0.3200;0.3856) | -64.9% |
| Sex (ref.: Female) | | | | | |
| Male | 0.37018 | 1.447995 | 0.052045 | (1.3076;1.6035) | 44.8% |
| Region (ref.: Midwest) | | | | | |
| Northeast | 0.27062 | 1.310777 | 0.055274 | (1.1762;1.4608) | 31.1% |
| Southeast | -0.35234 | 0.703041 | 0.059128 | (0.6261;0.7894) | -29.7% |
| South | -0.26061 | 0.770581 | 0.05852 | (0.6871;0.8642) | -22.9% |
| Cause of accident (ref.: Human error) | | | | | |
| Mechanical Failure | -0.43225 | 0.649047 | 0.11139 | (0.5217;0.8074) | -35.1% |
| Road failure | -0.34781 | 0.706233 | 0.082759 | (0.6005;0.8306) | -29.4% |
| Other causes | -0.53574 | 0.585236 | 0.121332 | (0.4614;0.7424) | -41.5% |
| Type of accident (ref.: Run over) | | | | | |
| Collision | 0.25203 | 1.286635 | 0.065916 | (1.1307;1.4641) | 28.7% |
| Runway exit | 0.30857 | 1.361477 | 0.079932 | (1.1640;1.5924) | 36.1% |
| Others | -0.69773 | 0.497714 | 0.130101 | (0.3857;0.6423) | -50.2% |
| Time of day (ref.: Morning) | | | | | |
| Day | -0.5014 | 0.605682 | 0.041553 | (0.5583;0.6571) | -39.4% |
| Evening | -0.41844 | 0.658073 | 0.08511 | (0.5570;0.7775) | -34.2% |

| | | | | | |
|---|---|---|---|---|---|
| Weather conditions (ref.: Rain) | | | | | |
| Cloudy | -0.22824 | 0.795933 | 0.07725 | (0.6841;0.9260) | -20.4% |
| Sun | -0.25854 | 0.772178 | 0.105521 | (0.6279;0.9496) | -22.8% |
| Others | -0.25966 | 0.771314 | 0.06525 | (0.6787;0.8765) | -22.9% |
| Lane type (ref.: Double) | | | | | |
| Simple | 0.57184 | 1.771524 | 0.043318 | (1.6273;1.9285) | 77.2% |
| Lane layout (ref.: Curve) | | | | | |
| Others | -0.77069 | 0.462694 | 0.078597 | (0.3966;0.5398) | -53.7% |
| Straight | -0.41846 | 0.658059 | 0.049259 | (0.5975;0.7248) | -34.2% |
| Vehicle type (ref.: Car) | | | | | |
| Bicycle | 1.66556 | 5.288634 | 0.10493 | (4.3055;6.4962) | 428.9% |
| Motorcycle | 1.00223 | 2.72435 | 0.04434 | (2.4976;2.9717) | 172.4% |
| Others | 0.40558 | 1.500172 | 0.121695 | (1.1818;1.9043) | 50.0% |
| Type of person involved (ref.: Driver) | | | | | |
| Pedestrian | 2.74812 | 15.61325 | 0.085493 | (13.2044;18.4615) | 1,461.3% |

Source: Prepared by the authors based on research data (2021).

Although previous analyzes have shown that the model was able to have an adequate predictive value for the purposes of the study, there was an interest in using accident data in 2022 to validate the model, that is, with data not previously used at the time of its estimation. Thus, the estimated model will be used to estimate the event of interest, when considering the explanatory variables captured in the accidents of the first quarter of 2022 and then compared to the true value of the event studied.

The cutoff value of 0.036 comes in use, the same way as done in the validation of the estimated model with the 2021 data. This way, the classification criterion was:

- If $p_i \geq 0.036$ observation $i$ should classify as death.
- If $p_i < 0.036$ observation $i$ should classify as non-death.

In this way, it was possible to obtain the confusion matrix and from that, an accuracy of 0.6982, a value very close to that of the model with the 2021 data, thus demonstrating that the model has a very significant predictive power.

**Table 9 -** Confusion matrix for the values observed in 2022 (Jan.-Mar) and predicted under the cutoff value = 0.036, with the binary logistic regression model with data from 2021.

| Response predicted by the model | Observed Response | | Total |
|---|---|---|---|
| | Death | Non-death | |
| Death | 660 | 7,940 | 8,600 |
| Non-death | 311 | 18,434 | 18,745 |
| Total | 971 | 26,374 | 27,345 |

Note: Sensitivity = 0.6797; Specificity = 0.6989; and Accuracy = 0.6982. Source: Prepared by the authors based on research data (2021)

In addition to the efficiency of the model, it is important to emphasize that some results obtained in this work corroborate other works in the literature, such as Junior et al. (2019), who, using a binary logistic regression model with data from accidents on federal highways in 2016, identified some factors that increase the chance of lethality, such as the highway being located in the Northeast region; single lane; the type of accident being by collision; during nighttime; and on weekends.

In the study conducted by Roquim et al. (2019), with data from 2018, with fewer explanatory variables and using the binary logistic regression model, the result obtained was similar when pointing to a greater probability of death for accidents caused due to human error.

After all the results presented in this work, one of the objectives was to develop a dashboard[1] with the descriptive results of the data, so that the reader can have a deeper understanding of the topic and perform simulations with the model estimated in this work. Thus, the user will be able to select the characteristics of the accidents and those involved in the accidents and obtain the probability of death and the classification of whether the individual would survive the accident or not.

## 4. Conclusion

By using the binary logistic regression model to estimate the probability of death in accidents on Federal highways in Brazil, it was possible to identify some characteristics of accidents and those involved that increase their chance of lethality. These characteristics can be objects of studies and actions that help to reduce the number of fatalities on federal highways in the country, as well as the financial loss caused. One of the possible solutions is to invest more in the quality of federal highways, as well as in the construction of more dual-lane highways, given that the risk of lethality is lower in these locations. In addition, as the biggest causes of accidents are due to human error, there needs to be more traffic education policies, especially regarding respect for cyclists, motorcyclists, and pedestrians. The accuracy of the estimated model was considered very significant for the object of study, and, through the dashboard, users will be able to understand how the results of a regression model work and use it for decision making.

Furthermore, for future studies, the insertion of new variables related to the quality of the road is suggested. Additionally, it can be used the multinomial logistic regression model, when considering the original target variable before applying the transformation, i.e. with four categories. Another approach is to obtain data on road accidents in other countries and conduct the same study, with a comparative focus on Brazil.

## Acknowledgments

To the Institute for Research and Continued Education in Business Economics and Management (Instituto de Pesquisas e Educação Continuada em Economia e Gestão de Empresas – PECEGE), through the MBA course in Data Science and Analytics – USP/ESALQ.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6), 716-723.

Carvalho, M. S. et al. (2011). *Análise de sobrevivência: teoria e aplicações em saúde*. FIOCRUZ.

Colosimo, E. A. & Giolo, S.R. (2006). *Análise de sobrevivência aplicada*. Edgard Blucher.

Core Team. (2021). A language and environment for statistical computing. *Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org.

Core Team. (2021). Core Team and contributors worldwide stats: The R Stats Package. *R package version 4.2.0*. 2021.. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html.

CNT. Confederação Nacional do Transporte (2021). *Painel CNT de Consultas Dinâmicas dos Acidentes Rodoviários.* https://www.cnt.org.br/painel-acidente.

Fávero, L. P. & Belfiore, P. (2017). M*anual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. Elsevier.

Giolo, S. R. (2017). *Introdução à Análise de Dados Categóricos com Aplicações*. Projeto Fisher ABE.

[1] Access to the dashboard can be through the link: https://app.powerbi.com/view?r=eyJrIjoiMTU1M2I1YzUtNzQ2YS00MjE5LTkyODYtYjU0ODU5YmZkNWQxIiwidCI6IjM4YTZiNDFiLTJmOWEtNGFiMi1hYjJhLTYzOWE0M2M0ZDQ3YSJ9

Izbicki, R. & dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki,.

Junior, G. T. B., Bertho, A. C. S. & Veiga, A. C. (2019). A letalidade dos acidentes de trânsito nas rodovias federais brasileiras. *Revista Brasileira de Estudos de População*, 36, 1-22.

Miranda, R., Silva, W. P. & Dutt-Ross, S. (2021). Identificação de fatores determinantes da severidade das lesões sofridas por pedestres nas rodovias federais brasileiras entre 2017 e 2019: Análise via regressão logística multinomial. *Scientia Plen*a, 17 (4).

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London – New York. Second edition, Chapman and Hall, 1989.

PRF. Polícia Rodoviária Federal. (2021). https://arquivos.prf.gov.br/arquivos/index.php/s/n1T3lymvIdDOzzb.

Roquim, F. V., Nakamura, L. R., Ramires, T. G. & Lima, R. R. (2019). Regressão logística: o que leva um acidente rodoviário a ser uma tragédia? *Sigmae*, 8 (2), 19-28.

Santos, D. F. (2017). *Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência*. (Dissertação de Mestrado) . Universidade de Brasília, Brasília, Brasil.

Schwarz, G. (1978). Estimating the dimensional of a model. *Annals of Statistics*, 6, 461-464.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by Akaike's. *Communications in Statistics – Theory and Methods*, 7 (1), 13-26.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc*, 462-482.

WHO. World Health Organization. (2015). *Global status report on road safety 2015*. https://shortest.link/whointviolenceinjuryprevention.