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

Damião Flávio dos Santos[1]*; Yuri Machado de Souza[2]

[1]Confederação Nacional do Transporte. Bacharel e Mestre em Estatística. QMSW 05, Lote 03, Bloco A; 70680-510 Sudoeste, DF, Brasil.
[2] Raízen. Mestre em Economia Aplicada. Avenida Brigadeiro Faria Lima, 4100, 11º andar – Itaim Bibi; 04538-132 São Paulo, SP, Brasil.
*autor correspondente: d.flaviostate@gmail.com

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

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. Para isso, 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 o 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.

Binary logistic regression model applied to accident data on federal highways in Brazil

Abstract

The 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. To this end, a binary logistic regression model was estimated. The event of interest is the circumstance of death in an accident with data from 2021. Following some variable selection procedures, the final model was obtained, and then a validation was made with data from 2022. The overall efficiency of the model for both 2021 and 2022 data was around 70%. Next, the odds ratio between some distinct categories was calculated, and how much of an increase in crash fatality it generates compared to the reference category. For example, in a crash, the pedestrian is 15.6 times more likely to be killed than the driver, and the bicyclist is 5.3 times more likely to die. Although most accidents have a human cause, some results show a need for public policy intervention that can help reduce these tragedies. Finally, a dashboard was developed to make the model's understanding concrete and dynamic. The user can 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

Introdução

Mundialmente, os acidentes de trânsito são uma das principais causas de morte, sendo a principal entre jovens na faixa etária de 15 a 29 anos, um fato que vai além do pesado

fardo que as lesões e óbitos ocorridos no trânsito representam para as economias nacionais e para as famílias (WHO, 2016).

No Brasil, o número elevado de acidentes ocorridos tanto no trânsito das cidades quanto em rodovias federais e estaduais tem causado bastante prejuízo e poderia ser evitado a partir de políticas de saúde pública. Todos os anos, milhares de pessoas têm suas vidas ceifadas em acidentes em rodovias federais causados por falha humana, falha na via, falha mecânica ou outros motivos.

De acordo com dados da Polícia Rodoviária Federal (PRF, 2021), foram registrados 64.515 acidentes em rodovias federais, 1,6% a mais do que no ano anterior. Além disso, os dados mostram que houve 5.395 óbitos, e a partir desses dados, a Confederação Nacional do Transporte (CNT, 2021) obteve a estimativa de que esses acontecimentos geraram um prejuízo em torno de 12,19 bilhões de reais para o país, sendo que cerca de 4,7 bilhões foram custos relacionados aos acidentes com vítimas fatais.

Destaca-se que a PRF, desde 2007, adaptou uma política de dados abertos e fornece as informações dos acidentes em rodovias federais. Por definição, de acordo com a PRF, acidente é um fato ocorrido em faixa de domínio de rodovia ou estrada federal que envolva veículo, que não seja premeditado e do qual resultem danos materiais em bens públicos ou particulares ou lesões em pessoas.

Ao considerar a relevância desses acidentes e principalmente os óbitos provocados, alguns autores, como Roquim et al. (2019), Miranda et al. (2021) e Junior et al. (2019), tem analisado e explorado esses dados, com a finalidade de identificar fatores que elevam a probabilidade de ocorrência desses eventos.

Ao considerar o desfecho do acidente como sendo uma variável resposta que pode ser “ocorrência de óbito” ou “não ocorrência de óbito” e ao incorporar outras variáveis no estudo, é possível, por meio de um modelo de regressão logística binária, explicar a probabilidade de ocorrência de ambos os desfechos.

Neste trabalho, será apresentada uma aplicação do modelo de regressão logística binária que consiga estimar a probabilidade de ocorrência de óbito em acidentes em rodovias federais no Brasil com dados de 2021 e que seja capaz de predizer a ocorrência com dados do primeiro trimestre de 2022. Após estimar o modelo, objetiva-se a construção de um *dashboard*, com o software Power BI, que estime a probabilidade de ocorrência de mortes no acidente. A classificação em óbito ou não óbito será realizada por dadas características, especificadas pelo usuário, e por um determinado valor de corte.

Para a aplicação de tais técnicas, será utilizado o software livre R versão 4.1.2. (CORE TEAM, 2021).

Material e Métodos

Modelo de regressão logística binária

De acordo com Fávero e Belfiore (2017), a regressão logística binária tem como objetivo principal estudar a probabilidade de ocorrência de um evento definido por Y que se apresenta na forma qualitativa dicotômica ($\mathrm{Y} = 1$ para descrever a ocorrência do evento de interesse e $\mathrm{Y} = 0$ para descrever a ocorrência do não evento), com base no comportamento de variáveis explicativas. Além disso, é definido um vetor de variáveis explicativas, com respectivos parâmetros estimados por meio da eq. (1):

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

onde, Z é conhecido por logito, α representa uma constante e $\beta_j\,(j = 1,2,\cdots,k)$ são os parâmetros estimados para cada variável explicativa.

Destaca-se que o objetivo da construção de Z não é representar a variável dependente, mas sim, a probabilidade de ocorrência do evento de interesse. Para tanto, é necessário considerar o conceito de chances de ocorrência de um evento de interesse ocorrer, como a razão entre a probabilidade de o evento ocorrer e a probabilidade de ele não ocorrer.

Conforme visto em Fávero e Belfiore (2017), na regressão logística binária define o logito Z como o logaritmo natural da chance, ou seja, da seguinte forma (eq. 2):

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

Como o objetivo é definir uma expressão para a probabilidade de ocorrência do evento de interesse em função do logito, ao isolar $\mathrm{p_i}$ da equação 2, após procedimentos algébricos, tem-se que a probabilidade do evento ocorrer é definido por $\mathrm{p_i} = \frac{e^{(Z_i)}}{1+e^{(Z_i)}}$ e por consequência a probabilidade do evento não ocorrer é definido por $1 - \mathrm{p_i} = 1 - \frac{e^{(Z_i)}}{1+e^{(Z_i)}}$.

Ao substituir $\mathrm{Z_i}$ pela expressão 1, se obtém que a probabilidade de ocorrência do evento é dada por (eq. 3):

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

e a probabilidade de ocorrência do não evento é dada por (eq. 4):

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

Estimadores dos parâmetros pelo método da máxima verossimilhança

Seja $Y\sim$ Bernoulli (p), sua função de distribuição de probabilidade é definida como (eq. 5):

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

onde, Y assume valores 0 (não ocorrência do evento) ou 1 (ocorrência do evento) e p_i é a probabilidade de ocorrência do evento de interesse.

Dessa forma, ao considerar os estimadores de máxima verossimilhança [EMV], tem-se que (eq. 6):

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

Por razões de otimização dos cálculos, é comum trabalhar com o logaritmo da função de verossimilhança. Sendo assim, ao aplicar o logaritmo em (5), tem-se (eq. 7):

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

Ao substituir o parâmetro p pela equação (2), tem-se a seguinte eq. (8):

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

onde, $\theta = (\alpha, \beta)$ e C é uma constante que não depende de θ.

Os estimadores EMV são os valores de θ que maximizam $L(\theta)$ ou, equivalentemente, o logaritmo de $L(\theta)$, que são estimados resolvendo-se o sistema de equações:

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

Para encontrar a solução deste sistema de equações para um conjunto de dados particular, é necessário utilizar um método numérico, para o qual, usualmente, utiliza-se o método de Newton-Raphson, algoritmo desenvolvido por McCullagh e Nelder (1989) que é implementado por um pacote estatístico.

Neste trabalho foi utilizado o software livre R na versão 4.1.2. (CORE TEAM, 2021), por meio da função glm() do pacote stats (R CORE TEAM, 2021).

Significância dos efeitos das variáveis

Obtidas as estimativas dos parâmetros do modelo de regressão logística binária, faz-se necessário avaliar a adequação do modelo ajustado. De acordo com Giolo (2017), o princípio em regressão logística é o mesmo usado em regressão linear, ou seja, comparar os valores observados da variável resposta com os valores preditos pelos modelos com e sem a variável sob investigação.

Teste χ^2 (qui-quadrado)

Segundo Fávero e Belfiore (2017), o teste qui-quadrado propicia condições à verificação da significância do modelo, uma vez que as hipóteses nula e alternativa para um modelo geral de regressão logística são, respectivamente:

$$\mathrm{H}_0\colon \beta_1 = \beta_2 = \cdots = \beta_\mathrm{k} = 0,$$

H_1: Existe pelo menos um $\beta_\mathrm{j} \neq 0$.

Sendo assim, o teste irá realizar uma verificação inicial sobre a existência do modelo que está sendo proposto, pois, caso todos os parâmetros estimados $\beta_j\ (j = 1{,}2, \cdots, k)$ forem estatisticamente iguais a 0, o comportamento de alteração de cada uma das variáveis explicativas não influencia em absolutamente nada a probabilidade de ocorrência do evento em estudo. A estatística χ^2 tem a seguinte expressão (eq. 9):

$$\chi^2 = -2(\mathrm{LL}_0 - \mathrm{LL}_{\mathrm{máx}}), \qquad (9)$$

em que LL_0 é o resultado da log verossimilhança do modelo nulo (sem a presença das variáveis explicativas) e $\mathrm{LL}_{\mathrm{máx}}$ é o resultado da log verossimilhança do modelo completo. O critério de decisão do teste qui-quadrado pode ser dado pelo $\mathrm{valor} - \mathrm{p}$ do teste, confrontado com o nível de significância $\alpha = 0{,}05$ estabelecido previamente. Caso o $\mathrm{valor} - \mathrm{p}$ seja menor do que 0,05, então existe pelo menos um $\beta_\mathrm{j} \neq 0$; caso contrário, todos os $\beta_\mathrm{j} = 0$, o que significa que nenhuma das variáveis explicativas são significativas para o modelo.

Teste Z de Wald

De acordo com Fávero e Belfiore (2017), o teste qui-quadrado avalia a significância conjunta das variáveis explicativas, não definindo qual ou quais dessas variáveis consideradas no modelo são estatisticamente significativas para influenciar a probabilidade de ocorrência do evento. Para isso, após a verificação de que pelo menos um $\beta_\mathrm{j} \neq 0$, o teste

conhecido por Z de Wald será aplicado (Wald, 1943). As expressões para o cálculo das estatísticas Z de Wald de cada parâmetro α e β_j são dadas, respectivamente, por (eq. 10):

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

onde, $s.e$ é o erro padrão da estimativa de cada parâmetro do modelo analisado.

De acordo com Carvalho (2011), sob a hipótese nula, tem-se que a estatística Z segue uma distribuição normal padrão. Então, para a conclusão do teste, o valor da estatística Z de Wald deverá ser confrontado com o valor crítico da tabela da distribuição normal, ao considerar um determinado nível de significância α estabelecido previamente. Além disso, assim como o teste qui-quadrado, a conclusão poderá ser dada pelo $\text{valor} - p$ do teste confrontado com o nível de significância. Assim, ao considerar $\alpha = 0{,}05$, o parâmetro estimado será significativo caso $Z < -1{,}96$ ou $Z > 1{,}96$, ou ainda caso $\text{valor} - p < 0{,}05$.

Razão de chances

De acordo com Giolo (2017), a chance de ocorrência de um evento de interesse é definida como a razão entre a probabilidade de o evento ocorrer e de o evento não ocorrer. Dessa forma, em estudos de coorte, indivíduos expostos e não expostos a um fator de interesse são acompanhados ao longo do tempo, a fim de se observar quantos deles desenvolvem a doença, por exemplo.

A razão de chances (“odds ratio”) fica definida, portanto, como a razão entre a chance de ocorrência da doença entre os expostos e a chance de ocorrência da doença entre os não expostos, da seguinte forma (eq. 11):

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

e substituindo as expressões do modelo de regressão logística binária, obtém-se a seguinte expressão (eq. 12):

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

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

Ainda segundo Giolo (2017), o intervalo de confiança para a OR, ao nível de $100(1-\alpha)\%$ de confiança, pode ser obtido por (eq. 13):

$$\mathrm{IC(OR)}_{100(1-\alpha)\%} = \mathrm{e}^{(\widehat{\beta_1} \pm \mathrm{z}_{\frac{\alpha}{2}} \times \mathrm{s.e}(\widehat{\beta_1}))}, \quad (13)$$

onde, $\mathrm{z}_{\alpha/2}$ denota o $100(1-\alpha)\%$ percentil da distribuição normal padrão e $\mathrm{s.e}$ é o erro padrão da estimativa de β_1.

Seleção de modelos

Teste de razão de verossimilhança

De acordo com Colosimo e Giolo (2006), o teste de razão de verossimilhança envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada sem restrição e sob H_0, ou seja, a comparação de $\log \mathrm{L}(\hat{\theta})$ (modelo completo) e $\log \mathrm{L}(\theta_0)$ (modelo reduzido). A estatística para o teste é dada por eq. (14):

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

em que sob $\mathrm{H}_0: \theta = \theta_0$ segue aproximadamente uma distribuição qui-quadrado com p graus de liberdade. Ainda de acordo com os autores, destaca-se que, para amostras grandes, H_0 é rejeitada a um nível de significância α, se $\mathrm{TRV} > \chi^2_{\mathrm{p},1-\alpha}$.

Critério de informação de Akaike [AIC]

O método proposto por Akaike (1974) é conhecido como critério de informação de Akaike [AIC]. De acordo com Santos (2017), sua ideia básica é selecionar um modelo que seja parcimonioso, ou seja, que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o logaritmo da função de verossimilhança cresce com o aumento do número de parâmetros do modelo, dessa forma, espera-se encontrar o modelo com menor valor para a eq. (15):

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

em que k é o número de parâmetros do modelo ajustado.

Critério de Informação de Akaike Corrigido [AICc]

Sugiura (1978) propôs uma correção do critério AIC, pois o AIC pode ter um desempenho ruim se existem muitos parâmetros em comparação com o tamanho da amostra. Desta forma, o AICc é apenas uma correção de segunda ordem do viés de AIC, dado pela seguinte expressão (eq. 16):

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

onde k é o número de parâmetros do modelo a serem estimados e n é o número de observação da amostra.

Critério de informação Bayesiano [BIC]

Proposto por Schwarz (1978), o critério de informação Bayesiano (BIC) é dado por eq. (17):

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

onde k é o número de parâmetros do modelo a serem estimados e n é o número de observações da amostra.

Análise de sensibilidade

No intuito de verificar a qualidade do ajuste de um modelo, algumas medidas podem ser avaliadas, como sensitividade e especificidade. De acordo com Giolo (2017), é necessário estabelecer uma probabilidade, denominada ponto de corte ("cutoff"), a partir da qual se estabeleça que a variável resposta receba o valor 1 para probabilidades preditas pelo modelo maiores ou iguais a esse ponto de corte e valor 0, caso contrário. A partir dessa definição, é possível construir uma tabela de dupla entrada entre os valores preditos e os valores reais, conhecida por matriz de confusão.

Tabela 1. Matriz de confusão para os valores observados e preditos sob um determinado ponto de corte

Resposta predita pelo modelo	Resposta observada		Totais
	$\mathrm{Y}=1$	$\mathrm{Y}=0$	
$\mathrm{Y}=1$	n_{11}	n_{12}	$\mathrm{n}_{1.}$
$\mathrm{Y}=0$	n_{21}	n_{22}	$\mathrm{n}_{2.}$
Totais	$\mathrm{n}_{.1}$	$\mathrm{n}_{.2}$	n

Fonte: Adaptado de Giolo (2017)

Desse modo, define-se três medidas muito importantes para o diagnóstico do modelo, sendo essas "Sensitividade", "Especificidade" e "Eficiência global do modelo".

Sensitividade

Conforme visto em Fávero e Belfiore (2017), a sensitividade diz respeito ao percentual de acerto (eq. 18), para um determinado "cutoff", considerando-se apenas observações que de fato são evento de interesse, ou seja:

$$\mathrm{sensitividade} = \frac{\mathrm{n}_{11}}{\mathrm{n}_{.1}} \times 100\%. \tag{18}$$

Especificidade

Segundo Fávero e Belfiore (2017), a especificidade, por outro lado, refere-se ao percentual de acerto (eq. 19), para um determinado ponto de corte ("cutoff"), considerando-se apenas as observações que não são evento de interesse, ou seja:

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

Eficiência global do modelo

Fávero e Belfiore (2017) definem a eficiência global do modelo como um percentual de acerto da classificação para um determinado ponto de corte (“cutoff”), que é dado pela soma da diagonal principal da matriz de confusão dividida pelo tamanho da amostra (eq. 20):

$$\text{EGM} = \frac{n_{11}+n_{22}}{n} \times 100\%. \quad (20)$$

Curva ROC

Como o valor do ponto de corte influencia no valor de sensitividade, especificidade e, consequentemente, na eficiência global do modelo, segundo Giolo (2017), é necessário identificar o valor de “cutoff” que produz o maior percentual de acertos. A curva ROC (em inglês, “Receiver Operating Characteristic”) é usualmente utilizada com essa finalidade.

Para a obtenção da curva ROC, pares de pontos (x,y) = (1-especificidade, sensitividade) são representados graficamente para vários pontos de corte. O modelo com discriminação perfeita corresponde àquele sem sensitividade e especificidade iguais a 1, o que implica que (x,y) = (0,1). Assim, pontos de corte localizados próximos ao canto superior esquerdo do gráfico (Figura 1) indicarão que o modelo ajustado produz o maior percentual de acertos, seja em termos de verdadeiros positivos ou de verdadeiros negativos. Ainda, quanto mais próxima de 1 for a área abaixo da curva ROC, melhor será o poder preditivo do modelo (Giolo, 2017).

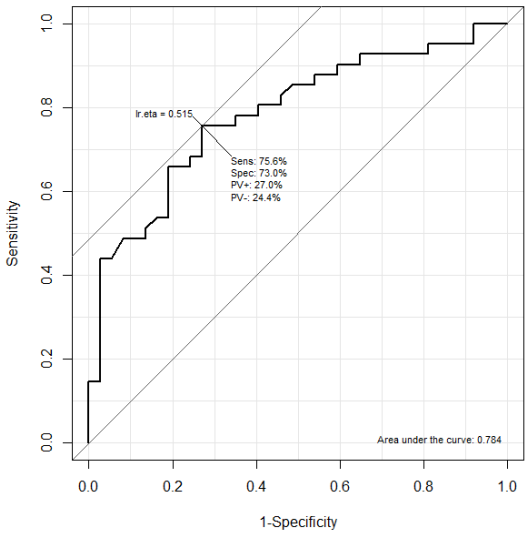


Figura 1: Exemplo de curva ROC associado ao modelo
Fonte: Adaptado de Giolo (2017)

Validação cruzada

Esta etapa consiste em dividir a base de dados em duas partes: treino e teste. Essa separação pode ser dada de forma aleatória, mas também podemos designar o tempo de observação para essa função. A validação cruzada representa uma das mais utilizadas em problema de "machine learning" e assim como muito utilizado na literatura, sugere-se a divisão entre 70% e 80% para a base de treino e de 30% a 20% para teste, conforme visto em Izbicki e Santos (2020).

Após essa divisão, há o desenvolvimento do modelo, como descrito anteriormente, mas utilizando apenas a base de treino. Em seguida, é avaliada a acurácia do modelo na base de teste. Neste trabalho, será desenvolvido o modelo com os dados de 2021 e validado com os dados de janeiro a março de 2022, como mostra a Figura 2.

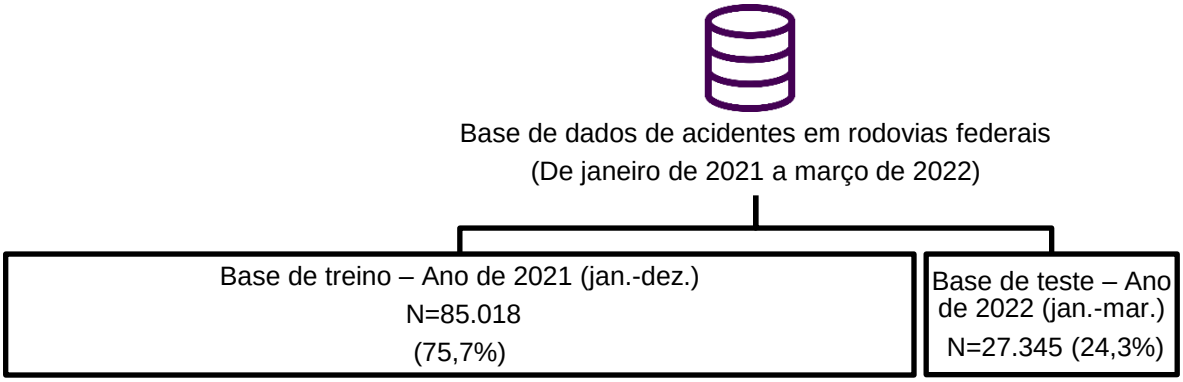


Figura 2. Fluxograma da segmentação da base de dados em treino e teste para validação do modelo
Fonte: Resultados originais da pesquisa

Banco de dados para aplicação

Os dados utilizados nesse trabalho são referentes aos acidentes em rodovias federais no Brasil ocorridos em 2021, que serão usados para estimar os parâmetros do modelo de regressão logística binária e, em seguida, testados utilizando os dados do primeiro trimestre de 2022. Esses fazem parte do banco de dados oficial da Polícia Rodoviária Federal, que contém dados a partir de 2007, como mostra a Figura 3.

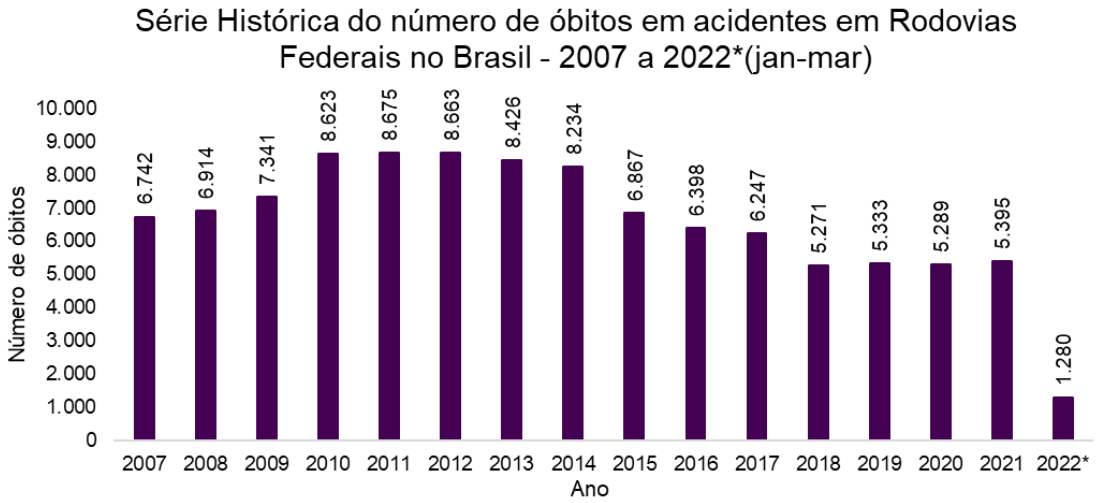


Figura 3. Série histórica do número de óbitos em acidentes em rodovias federais no Brasil, de 2007 a 2022
Fonte: Dados da Polícia Rodoviária Federal
Nota: *valor correspondente ao período de janeiro a março de 2022

Como se observa, há um número bem elevado de acidentes fatais todos os anos, e em 2021 não foi diferente. No total, em 2021 foram registradas 64.515 ocorrências de acidentes, sendo que houve 5.395 pessoas mortas nesses acidentes, o que significa que para cada 100 acidentes houve pelo menos 8 mortes. Ressalta-se que, nesses dados, há várias informações incompletas em relação tanto às variáveis explicativas quanto à variável resposta. Assim, foram considerados apenas as respostas válidas em todas as variáveis de interesse, totalizando 85.018 pessoas envolvidas em acidentes, sendo 81.974 pessoas que não vieram a óbito e 3.044 óbitos. Ou seja, o evento de interesse ocorreu em 3,6% das pessoas que se envolveram em acidentes em rodovias federais no Brasil.

Já de janeiro a março de 2022 ocorreram 14.958 acidentes, com 35.895 pessoas envolvidas e 1.280 óbitos. Assim como ocorreu em 2021, houve a presença de algumas variáveis com valores ausentes e, após sua exclusão, o banco de dados ficou com 27.345 observações, sendo 26.374 pessoas que não vieram a óbito e 971 óbitos, o que corresponde a 3,55% das pessoas que se envolveram nos acidentes, praticamente o mesmo percentual observado em 2021.

Variável resposta

A variável resposta para esse trabalho faz parte do banco de dados de acidentes e está indicada como o estado físico do envolvido. No banco original, essa variável está apresentada com 4 níveis de estado físico (Ileso, Lesões Leves, Lesões Graves e Óbito). Como o objetivo desse trabalho é estimar a probabilidade de ocorrência de óbito, então será realizada uma reclassificação da variável como (Figura 4):

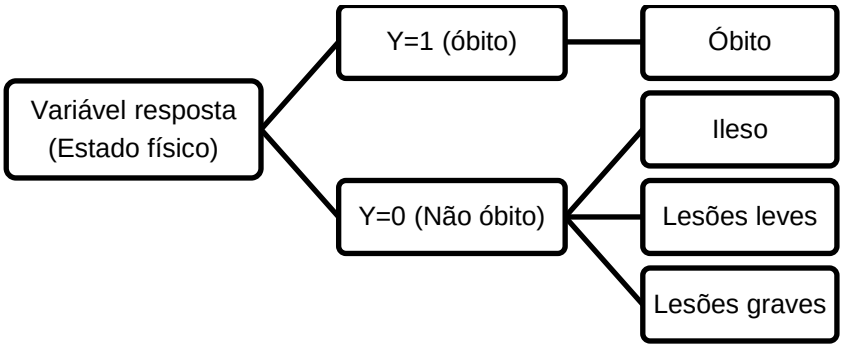


Figura 4. Fluxograma da reclassificação da variável resposta “Estado físico dos envolvidos”
Fonte: Resultados originais da pesquisa

Variáveis explicativas

As variáveis explicativas disponíveis no banco de dados de acidentes da PRF passaram por algumas transformações e ajustes devido à presença de variáveis com muitas categorias. A maioria delas são variáveis categóricas e será necessário a transformação em variáveis “dummies”. As variáveis trabalhadas como explicativas do modelo foram as que estão descritas na Figura 5 e suas categorias estão descritas na Tabela 3.

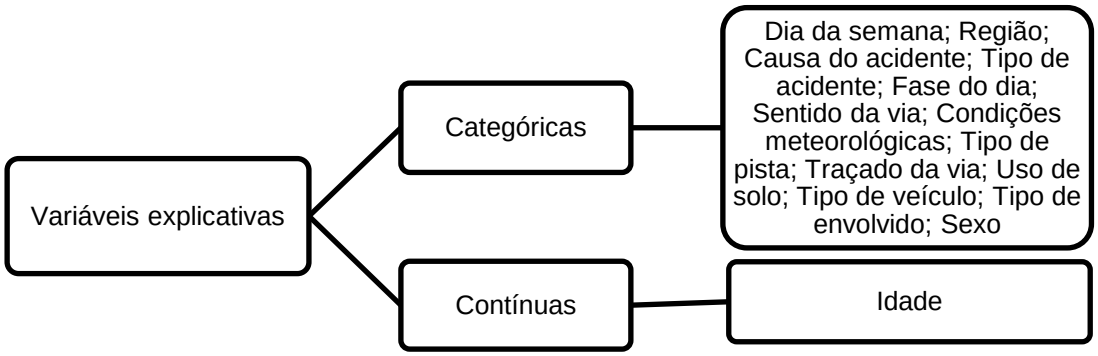


Figura 5. Fluxograma da composição das variáveis explicativas categóricas e contínuas
Fonte: Resultados originais da pesquisa

Resultados e Discussão

Nesta seção, serão apresentados os resultados e discussões do trabalho. Inicialmente, é realizada uma análise espacial dos óbitos ocorridos em acidentes em rodovias federais no Brasil. Pela densidade do mapa de calor, observa-se uma concentração mais elevada de mortes nas rodovias do Nordeste, Sudeste e Sul do país, além disso, observa-se uma maior concentração próximas das grandes cidades no país. Dessa forma, a localização do acidente pode ser um fator preponderante para a ocorrência de óbito.

Figura 6. Mapa de calor para a localização dos óbitos ocorridos em acidentes em rodovias federais no Brasil em 2021
Fonte: Dados da Polícia Rodoviária Federal

Em seguida, será apresentada uma análise descritiva dos dados, na qual, para a única variável numérica, calculou-se as principais medidas descritiva segundo o estado físico dos envolvidos em acidentes.

Conforme se observa na Tabela 2 e na Figura 7, a média e a mediana de idade entre as pessoas que foram a óbito é pelo menos 2 anos maior do que as pessoas que sobreviveram ao acidente. Essa diferença descritiva entre os dois grupos dá indícios de que a idade pode ser uma variável que consiga explicar a probabilidade de ocorrência de óbito

Tabela 2. Medidas descritivas para a idade dos envolvidos em acidentes em rodovias federais no Brasil em 2021, segundo seu estado físico após o envolvimento

Categorias	Estado físico	
	Não óbito	Óbito
Média	38,2	40,5
Moda	35,0	40,0
Mediana	37,0	39,0
Desvio-padrão	14,6	15,9
Mínimo	0	0
Máximo	104	87

Fonte: Dados da Polícia Rodoviária Federal

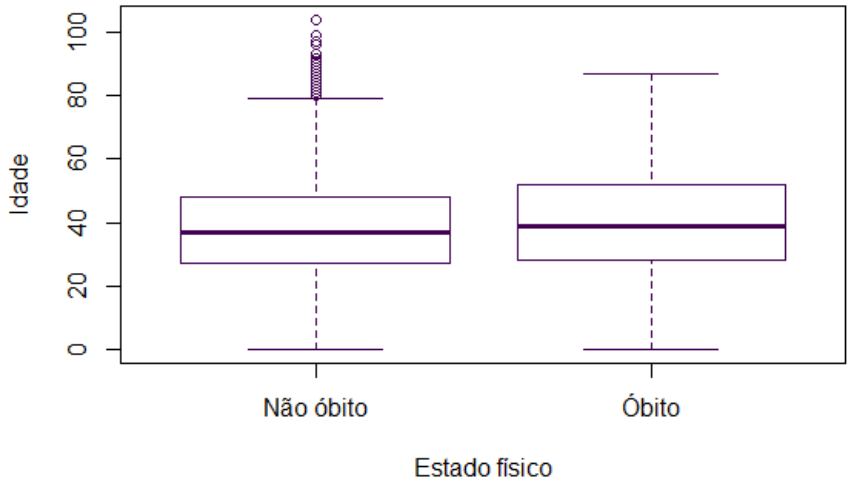


Figura 7. “Box plots” da idade, segundo o estado físico dos envolvidos após os acidentes em rodovias federais no Brasil em 2021
Fonte: Dados da Polícia Rodoviária Federal

A Tabela 3 apresenta uma análise bivariada por meio da distribuição conjunta entre as variáveis explicativas categóricas e a variável estado físico dos envolvidos. Destaca-se alguns pontos relevantes nessa análise descritiva, tais como: ocorrem mais mortes proporcionalmente durante o fim de semana, além de que, se considerando que o fim de semana foi composto por sábado e domingo, a média de mortes por dia foi de 608,5 óbitos; já durante a semana, foi de 365,4.

Os maiores percentuais de óbitos de forma individual são de pessoas do sexo masculino; das regiões Nordeste e Norte; tendo falha humana como causa do acidente; com atropelamento como tipo de acidente; durante o amanhecer; em neblina; em pista simples; em curva; bicicleta como veículo envolvido; e pedestre com elevado percentual de óbitos ao comparar com os outros tipos de envolvidos. Dessa forma, essas características podem ter efeito no modelo de regressão logística binária que irá calcular a probabilidade de ocorrência de óbito em rodovias federais no Brasil.

Tabela 3. Distribuição conjunta das frequências e proporções (em porcentagem), segundo as variáveis explicativas e o estado físico dos envolvidos nos acidentes

Variáveis	Categorias	Estado físico	
		Não óbito	Óbito
		N (%)	N (%)
Dia da semana	Fim de semana	27.440 (95,8%)	1.217 (4,2%)
	Semana	54.534 (96,8%)	1.827 (3,2%)
Sentido da via	Crescente	43.617 (96,3%)	1.677 (3,7%)
	Decrescente	38.357 (96,6%)	1.367 (3,4%)
Uso do solo	Urbano	37.669 (98,0%)	777 (2,0%)
	Rural	44.305 (95,1%)	2.267 (4,9%)
Sexo	Feminino	18.672 (97,4%)	489 (2,6%)
	Masculino	63.302 (96,1%)	2.555 (3,9%)
Região	Centro-Oeste	10.390 (96,4%)	390 (3,6%)
	Nordeste	16.992 (94,6%)	979 (5,4%)
	Norte	4.810 (95,7%)	214 (4,3%)
	Sudeste	25.096 (97,2%)	731 (2,8%)
	Sul	24.686 (97,1%)	730 (2,9%)
Causa do acidente	Falha humana	69.446 (96,3%)	2.691 (3,7%)
	Falha mecânica	4.601 (98,0%)	92 (2,0%)
	Falha na via	4.996 (96,6%)	178 (3,4%)
	Outras causas	2.931 (97,2%)	83 (2,8%)
Tipo de acidente	Atropelamento	4.028 (89,6%)	469 (10,4%)
	Capotamento/Tombamento	6.595 (96,9%)	210 (3,1%)
	Colisão	54.985 (96,7%)	1.884 (3,3%)
	Saída da pista	10.050 (96,1%)	403 (3,9%)
	Outros	6.316 (98,8%)	78 (1,2%)
Fase do dia	Amanhecer	3.351 (95,1%)	173 (4,9%)
	Pleno dia	45.998 (97,3%)	1.274 (2,7%)
	Anoitecer	5.100 (96,8%)	171 (3,2%)
	Plena noite	27.525 (97,3%)	1.426 (2,7%)
Condições meteorológicas	Chuva	7.216 (96,2%)	284 (3,8%)
	Neblina	695 (93,7%)	47 (6,3%)
	Nublado	12.861 (96,3%)	493 (3,7%)
	Sol	5.946 (97,5%)	154 (2,5%)
	Outros	55.256 (96,4%)	2.066 (3,6%)
Tipo de pista	Simples	41.134 (95,1%)	2.125 (4,9%)
	Dupla	33.506 (97,7%)	791 (2,3%)
	Múltipla	7.334 (98,3%)	128 (1,7%)
Traçado da via	Curva	12.907 (94,9%)	687 (5,1%)
	Reta	57.952 (96,5%)	2.104 (3,5%)
	Outros	11.115 (97,8%)	253 (2,2%)
Tipo de veículo	Automóvel	45.058 (97,2%)	1.300 (2,8%)
	Bicicleta	1.138 (90,5%)	119 (9,5%)
	Caminhão	13.021 (96,7%)	446 (3,3%)
	Moto	18.916 (94,9%)	1.008 (5,1%)
	Ônibus	2.343 (96,7%)	81 (3,3%)
	Outros	1.498 (94,3%)	90 (5,7%)
Tipo de envolvido	Condutor	60.260 (96,8%)	2.016 (3,2%)
	Passageiro	20.419 (97,1%)	601 (2,9%)
	Pedestre	1.274 (75,0%)	425 (25,0%)
	Outros	21 (91,3%)	2 (8,7%)

Fonte: Dados da Polícia Rodoviária Federal

Conforme descrito na equação (9), o teste qui-quadrado é realizado a partir do valor da log verossimilhança do modelo completo (com todas as variáveis explicativas) e utilizando o modelo apenas com o intercepto (α). O valor da log verossimilhança do modelo completo foi de $\mathrm{LL}_{\mathrm{máx}} = -11.461$. Já o modelo nulo foi de $\mathrm{LL}_0 = -13.124$. As hipóteses para o teste qui-quadrado são as seguintes:

$$\mathrm{H}_0\colon \beta_1 = \beta_2 = \cdots = \beta_{35} = 0,$$

$$\mathrm{H}_1\colon \text{Existe pelo menos um } \beta_j \neq 0.$$

Conforme descrito na Tabela 4, ao considerar um nível de significância α = 0,05, rejeita-se a hipótese nula (em que todos os β_j são estatisticamente iguais a 0), ou seja, há a existência de pelo menos um $\beta_j \neq 0$. Isso significa que, na prática, existe a possibilidade de um modelo de regressão logística binária com os dados analisados nesse estudo.

Tabela 4. Teste qui-quadrado para existência de modelo de regressão logística binária.

χ^2	g.l.	Valor-p
3.326,46	35	$< 2 \times 10^{-16}$

Fonte: Resultados originais da pesquisa

A Tabela 5 mostra as estimativas do modelo completo. Com o procedimento de "stepwise" foi possível estimar um modelo mais parcimonioso, ou seja, com menos variáveis explicativas sendo utilizadas para estimar a probabilidade da ocorrência de óbito em acidentes em rodovias federais no Brasil.

Com o teste Z de Wald, observa-se que o modelo completo apresenta algumas variáveis explicativas que não foram expressivas ao nível de 5% de significância, tais como: sentido da via (crescente); região (Norte); tipo de acidente (capotamento/tombamento); fase do dia (plena noite); condições meteorológicas (neblina); tipo de pista (múltipla); tipo de veículo (caminhão e ônibus); e tipo de envolvido (passageiro e outros). Assim, mesmo retirando essas variáveis explicativas, pelo procedimento "stepwise" é necessário verificar a eficiência desse modelo reduzido em comparação com o modelo completo, por meio de algumas medidas e teste de hipótese, conforme é apresentado na Tabela 6.

Tabela 5. Estimativas e teste de significância para o modelo completo e reduzido

Variáveis	Modelo completo			Modelo reduzido		
	Estimativa	Z	Valor-p	Estimativa	Z	Valor-p
α	-3,47582	-17,93	$< 2 \times 10^{-16}$	-3,52054	-27,468	$< 2 \times 10^{-16}$
Dia da semana (ref.: Fim de semana)						
Semana	-0,17376	-4,347	$1{,}4 \times 10^{-5}$	-0,17873	-4,499	$6{,}8 \times 10^{-6}$
Sentindo da via (ref.: Crescente)						
Decrescente	-0,06862	-1,777	0,07556	-	-	-
Uso do solo (ref.: Rural)						
Urbano	-1,04585	-21,741	$< 2 \times 10^{-16}$	-1,04626	-21,988	$< 2 \times 10^{-16}$
Sexo (ref.: Feminino)						
Masculino	0,400062	6,895	$5{,}4 \times 10^{-12}$	0,37018	7,113	$1{,}1 \times 10^{-12}$
Região (ref.: Centro-Oeste)						
Nordeste	0,28606	4,469	$7{,}9 \times 10^{-6}$	0,27062	4,896	$9{,}8 \times 10^{-7}$
Norte	0,03647	0,402	0,68795	-	-	-
Sudeste	-0,33690	-5,012	$5{,}4 \times 10^{-7}$	-0,35234	-5,959	$2{,}5 \times 10^{-9}$
Sul	-0,24868	-3,721	0,00020	-0,26061	-4,453	$8{,}5 \times 10^{-6}$
Causa do acidente (ref.: Falha humana)						
Falha mecânica	-0,42697	-3,808	0,00014	-0,43225	-3,881	0,00010
Falha na via	-0,34685	-4,178	$2{,}9 \times 10^{-5}$	-0,34781	-4,203	$2{,}6 \times 10^{-5}$
Outras causas	-0,55958	-4,401	$1{,}1 \times 10^{-5}$	-0,53574	-4,415	$1{,}0 \times 10^{-5}$
Tipo de acidente (ref.: Atropelamento)						
Capotamento/Tombamento	-0,02259	-0,17	0,864741	-	-	-
Colisão	0,23168	2,053	0,040067	0,25203	3,824	0,00013
Saída da pista	0,28328	2,303	0,021289	0,30857	3,860	0,00011
Outros	-0,71413	-4,478	$7{,}5 \times 10^{-6}$	-0,69773	-5,363	$8{,}2 \times 10^{-8}$
Fase do dia (ref.: Amanhecer)						
Pleno dia	-0,5355	-6,081	$1{,}2 \times 10^{-9}$	-0,50140	-12,067	$< 2 \times 10^{-16}$
Anoitecer	-0,45699	-3,956	$7{,}6 \times 10^{-5}$	-0,41844	-4,916	$8{,}8 \times 10^{-7}$
Plena noite	-0,04499	-0,517	0,60515	-	-	-
Condições meteorológicas (ref.: Chuva)						
Neblina	0,16492	0,957	0,33849	-	-	-
Nublado	-0,20371	-2,514	0,01193	-0,22824	-2,955	0,00313
Sol	-0,23816	-2,206	0,02736	-0,25854	-2,450	0,01428
Outros	-0,23663	-3,401	0,00067	-0,25966	-3,980	$6{,}9 \times 10^{-5}$
Tipo de pista (ref.: Dupla)						
Simples	0,56084	12,395	$< 2 \times 10^{-16}$	0,57184	13,201	$< 2 \times 10^{-16}$
Múltipla	-0,05264	-0,528	0,59765	-	-	-
Traçado da via (ref.: Curva)						
Outros	-0,77572	-9,856	$< 2 \times 10^{-16}$	-0,77069	-9,806	$< 2 \times 10^{-16}$
Reta	-0,42537	-8,575	$< 2 \times 10^{-16}$	-0,41846	-8,495	$< 2 \times 10^{-16}$
Tipo de veículo (ref.: Automóvel)						
Bicicleta	1,651616	15,452	$< 2 \times 10^{-16}$	1,66556	15,873	$< 2 \times 10^{-16}$
Caminhão	-0,09241	-1,549	0,12137	-	-	-
Moto	0,98353	20,619	$< 2 \times 10^{-16}$	1,00223	22,603	$< 2 \times 10^{-16}$
Ônibus	-0,04364	-0,36	0,71884	-	-	-
Outros	0,382344	3,111	0,00186	0,40558	3,333	0,00086
Tipo de envolvido (ref.: Condutor)						
Passageiro	0,040402	0,718	0,47255	-	-	-
Pedestre	2,744455	22,963	$< 2 \times 10^{-16}$	2,74812	32,144	$< 2 \times 10^{-16}$
Outros	1,116898	1,463	0,14352	-	-	-
Idade	0,010902	8,302	$< 2 \times 10^{-16}$	0,01072	8,238	$< 2 \times 10^{-16}$

Fonte: Resultados originais da pesquisa

De acordo com o descrito anteriormente, é necessário selecionar um modelo que seja parcimonioso, ou seja, que esteja bem ajustado e tenha um número reduzido de parâmetros

estimados. Para isso, utiliza-se os critérios de informação AIC, AICc e BIC, em que o modelo mais parcimonioso é o que tem menores valores para essas medidas, isto é, o modelo mais adequado para estimar a probabilidade de ocorrência do fenômeno estudado.

Assim, de acordo com a Tabela 6, verifica-se que o teste de razão de verossimilhança não rejeitou a hipótese de nulidade entre a diferença dos modelos ao considerar um nível de significância de 5%. No entanto, apesar desse resultado, não significa que qualquer um dos dois seja adequado, dado o princípio da parcimônia. Ressalta-se que a diferença da quantidade de parâmetros a serem estimados entre os modelos é igual a 10, e isso pode representar um ganho em termos de eficiência do modelo final. Como se observa na Tabela 6, tanto AIC quanto AICc e BIC do modelo reduzido foram menores do que do modelo completo. Dessa forma, optou-se pelo modelo reduzido como o modelo final.

Tabela 6. Medidas de critério de informação, log verossimilhança e teste de razão de verossimilhança para os dois modelos de regressão logística binária

Modelos	AIC	AICc	BIC	LL	TRV	Valor-p
Completo	22.994,3836	22.994,415	23.331,0059	-11.461,1918	9,6971	0,46746
Reduzido	22.984,0807	22.984,097	23.227,1968	-11.466,0403		

Fonte: Resultados originais da pesquisa

Após a escolha do modelo, buscou-se, para este estudo, identificar um valor de ponto de corte (“cutoff”) que seja capaz de captar o máximo de sensitividade e de especificidade. Ou seja, que esse valor de corte faça com que a taxa de acerto de previsão para os que serão evento seja igual à taxa de acerto para aqueles que não serão evento. Para isso, a curva de sensibilidade (Figura 8) mostra os valores possíveis de “cutoff” no eixo x e os valores de sensitividade e especificidade no eixo y.

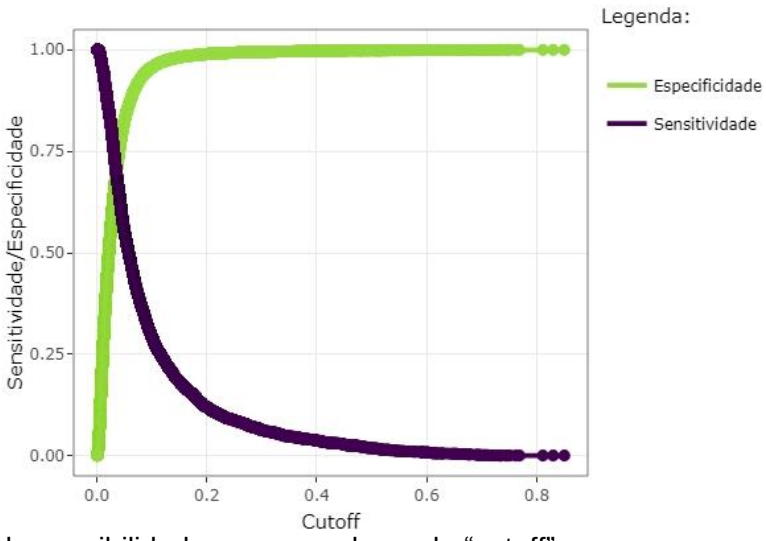


Figura 8. Curva de sensibilidade para os valores de “cutoff”
Fonte: Resultados originais da pesquisa

Ao observar a Figura 8, foi detectado o valor de ponto de corte ideal para a classificação dos eventos de interesse, de modo que esse consiga capturar a maior taxa de acerto em sensitividade, especificidade e eficiência global do modelo. O valor do "cutoff", portanto, foi de 0,036. Assim, teremos o seguinte critério de classificação:

- se $p_i \geq 0{,}036$, a observação i deverá ser classificada como óbito;
- se $p_i < 0{,}036$, a observação i deverá ser classificada como não óbito.

Assim, com esse ponto de corte, foi obtida a matriz de confusão e, a partir disso, uma eficiência global acima de 0,7, o que – para o objetivo do estudo – foi considerado bastante eficiente.

Tabela 7. Matriz de confusão para os valores observados e preditos sob o valor de *"*cutoff" = 0,036

Resposta predita pelo modelo	Resposta observada		Totais
	Óbito	Não óbito	
Óbito	2.116	24.395	26.511
Não óbito	928	57.579	58.507
Totais	3.044	81.974	85.018

Fonte: Resultados originais da pesquisa
Nota: Sensitividade = 0,6951; Especificidade = 0,7024; e Eficiência global do modelo = 0,7021

A área abaixo da curva igual a 0,774 e o coeficiente de Gini igual a 0,548 demonstram que o modelo foi bem ajustado e tem um poder preditivo bem significativo.

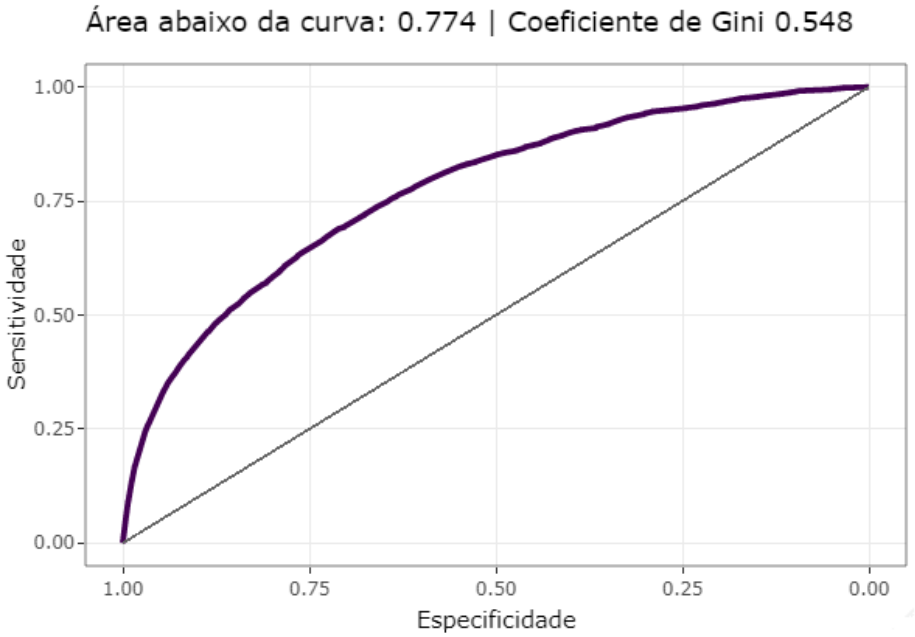


Figura 9. Curva ROC associada ao modelo de regressão logística binária escolhido
Fonte: Resultados originais da pesquisa

Após o ajuste do modelo, foram calculadas todas as combinações das categorias de todas as variáveis selecionadas para o modelo – ou seja, 3.456.000 combinações, a partir das

quais foi calculada a probabilidade de ocorrência de óbito. Ressalta-se que para esta análise, para a variável idade, como é contínua, considerou-se a média igual a 38 anos. Assim, como forma de visualizar as categorias que geram maiores e menores probabilidades de ocorrência de óbito, foram selecionadas as primeiras mil combinações e as últimas mil e, em seguida, geradas as duas nuvens de palavras, conforme é apresentado na Figura 10 e Figura 11.

Conforme já mencionado na análise descritiva, as falhas humanas; acidentes em pista simples; acidentes com bicicleta; com pedestres envolvidos; em curvas; com vítimas do sexo masculino; durante o fim de semana; no Nordeste em zona rural; faz com que a probabilidade de ocorrência de óbito aumente.

Em contrapartida, acidentes em área urbana; com pessoas do sexo feminino envolvidas; no Sudeste; em dia de semana; em pleno dia; em acidentes por outras causas; em outros traçados; e em pista dupla ou múltipla geram menor probabilidade de ocorrência de óbito das pessoas envolvidas.

Figura 10. Nuvem de palavras das categorias que geram maior probabilidade de ocorrência de óbito em rodovias federais no Brasil
Fonte: Resultados originais da pesquisa

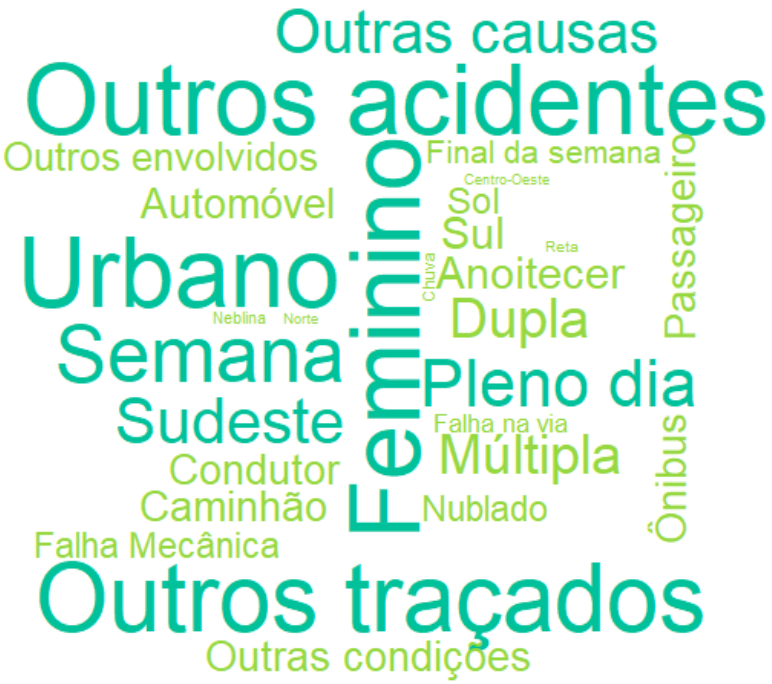


Figura 11. Nuvem de palavras das categorias que geram menor probabilidade de ocorrência de óbito em rodovias federais no Brasil
Fonte: Resultados originais da pesquisa

Outra abordagem para identificar o efeito de determinada variável no aumento da probabilidade de ocorrência de óbito nos acidentes ou o aumento na chance de letalidade é a utilização da razão de chances, conforme é apresentado na Tabela 8.

Assim como já observado nas análises anteriores, algumas características dos acidentes e das pessoas envolvidas fazem com que haja um aumento na probabilidade de óbito em acidentes em rodovias federais no Brasil. O resultado mais impactante nesta análise é o quanto de chance de letalidade o pedestre tem a mais do que o condutor do veículo – 15,6 vezes a mais, o que pode ser interpretado como um aumento de 1.461,3% na chance de letalidade do pedestre em relação ao condutor. Esse resultado corrobora com o estudo do Junior et. al (2019), que mostrou com um modelo de regressão logística binária, com dados de 2016, que a chance de letalidade do pedestre em relação ao ocupante de automóvel é 9,49 vezes maior.

De acordo com o DENATRAN, em abril de 2022 o número de condutores habilitados no Brasil foi 77.917.203, sendo que 64,8% são motoristas do sexo biológico masculino e 35,2% feminino. Apesar de muitas vezes serem discriminadas como sinônimo de motoristas ruins, as pessoas do sexo biológico feminino são as que menos morrem em acidentes e que menos sofrem acidentes, de acordo com os dados descritivos dos acidentes na Tabela 3. No modelo estimado, o sexo biológico masculino tem chance de morrer 1,44 vezes mais do que o feminino, o que equivale a um aumento de 44,7% na chance de letalidade.

Além disso, os ciclistas e os motociclistas são os perfis com maior chance de letalidade em relação aos motoristas de automóveis, conforme descrito na Tabela 8. O tipo de pista também é um fator que influencia a probabilidade de óbito, fazendo com que a pista simples aumente em 77,2% a chance de letalidade dos acidentes. Ressalta-se que de acordo com o Sistema Nacional de Viação (SNV, 2022) em 2022 a malha rodoviária do Brasil tem um total de 121.100,9 km de extensão, sendo que em sua maioria (57.309,5 km ou 47,3%) são pistas simples. Dessa forma, o investimento em infraestrutura, com a manutenção e construção de rodovias de pista dupla, pode ser visto como um fator primordial para a redução do número de mortes em acidentes no Brasil. É necessário que haja uma ação mais efetiva por parte das políticas públicas, com a finalidade de solucionar esse problema que tanto afeta socialmente e financeiramente tantas famílias vítimas desses eventos.

Os resultados negativos para o aumento na chance de letalidade, como os -64,9% para o uso do solo urbano, significa que para a categoria de referência (nesse caso o rural), há uma redução na chance de letalidade em 64,9%, caso o uso do solo durante o acidente seja urbano.

Já em relação a razão de chances da variável contínua, por exemplo, entre indivíduos com 65 anos e 18 anos de idade resulta em $\widehat{\mathrm{OR}} = \mathrm{e}^{(65-18)\times 0{,}01072} = 1{,}655$. Isso significa que indivíduos com 65 anos tem um aumento na chance de letalidade de 65,5% a mais do que indivíduos com 25 anos, além disso, que quanto mais idade o indivíduo tem, mais risco de óbito ele terá após o acidente.

Tabela 8. Razão de chances (“odds ratio”) associada ao modelo de regressão logística binária
(continua)

Variáveis	$\widehat{\beta_j}$	$OR = e^{\widehat{\beta_j}}$	$s.e(\widehat{\beta_j})$	$IC(OR)_{95\%}$	Aumento na chance de letalidade
Dia da semana (ref.: Fim de semana)					
Semana	-0,17873	0,836332	0,039722	(0,7737;0,9040)	-16,4%
Uso do solo (ref.: Rural)					
Urbano	-1,04626	0,351249	0,047582	(0,3200;0,3856)	-64,9%
Sexo (ref.: Feminino)					
Masculino	0,37018	1,447995	0,052045	(1,3076;1,6035)	44,8%
Região (ref.: Centro-Oeste)					
Nordeste	0,27062	1,310777	0,055274	(1,1762;1,4608)	31,1%
Sudeste	-0,35234	0,703041	0,059128	(0,6261;0,7894)	-29,7%
Sul	-0,26061	0,770581	0,05852	(0,6871;0,8642)	-22,9%
Causa do acidente (ref.: Falha humana)					
Falha Mecânica	-0,43225	0,649047	0,11139	(0,5217;0,8074)	-35,1%
Falha na via	-0,34781	0,706233	0,082759	(0,6005;0,8306)	-29,4%
Outras causas	-0,53574	0,585236	0,121332	(0,4614;0,7424)	-41,5%
Tipo de acidente (ref.: Atropelamento)					
Colisão	0,25203	1,286635	0,065916	(1,1307;1,4641)	28,7%
Saída da pista	0,30857	1,361477	0,079932	(1,1640;1,5924)	36,1%

Tabela 8. Razão de chances (“odds ratio”) associada ao modelo de regressão logística binária (conclusão)

Variáveis	$\widehat{\beta_J}$	$OR = e^{\widehat{\beta_J}}$	$s.e(\widehat{\beta_J})$	$IC(OR)_{95\%}$	Aumento na chance de letalidade
Outros	-0,69773	0,497714	0,130101	(0,3857;0,6423)	-50,2%
Fase do dia (ref.: Amanhecer)					
Pleno dia	-0,5014	0,605682	0,041553	(0,5583;0,6571)	-39,4%
Anoitecer	-0,41844	0,658073	0,08511	(0,5570;0,7775)	-34,2%
Condições meteorológicas (ref.: Chuva)					
Nublado	-0,22824	0,795933	0,07725	(0,6841;0,9260)	-20,4%
Sol	-0,25854	0,772178	0,105521	(0,6279;0,9496)	-22,8%
Outros	-0,25966	0,771314	0,06525	(0,6787;0,8765)	-22,9%
Tipo de pista (ref.: Dupla)					
Simples	0,57184	1,771524	0,043318	(1,6273;1,9285)	77,2%
Traçado da via (ref.: Curva)					
Outros	-0,77069	0,462694	0,078597	(0,3966;0,5398)	-53,7%
Reta	-0,41846	0,658059	0,049259	(0,5975;0,7248)	-34,2%
Tipo de veículo (ref.: Automóvel)					
Bicicleta	1,66556	5,288634	0,10493	(4,3055;6,4962)	428,9%
Moto	1,00223	2,72435	0,04434	(2,4976;2,9717)	172,4%
Outros	0,40558	1,500172	0,121695	(1,1818;1,9043)	50,0%
Tipo de envolvido (ref.: Condutor)					
Pedestre	2,74812	15,61325	0,085493	(13,2044;18,4615)	1.461,3%

Fonte: Resultados originais da pesquisa

Apesar das análises anteriores ter demonstrado que o modelo conseguiu ter um valor preditivo adequado para os objetivos do estudo, houve o interesse em utilizar os dados de acidentes em 2022 para validação do modelo, ou seja, esses dados não foram utilizados no momento de sua estimação. Assim, o modelo estimado será utilizado para estimar o evento de interesse, ao considerar as variáveis explicativas capturadas nos acidentes do primeiro trimestre de 2022 e em seguida confrontado com o verdadeiro valor do evento estudado.

Da mesma maneira como feito na validação do modelo estimado com os dados de 2021, será utilizado o valor de “cutoff” de 0,036, ou seja, teremos o seguinte critério de classificação:

- Se $p_i \geq 0{,}036$ a observação $\boldsymbol{i}$ deverá ser classificada como óbito.
- Se $p_i < 0{,}036$ a observação i deverá ser classificada como não óbito.

Dessa forma, foi obtido a matriz de confusão (Tabela 9) e a partir disso, uma eficiência global de 0,6982, valor muito próximo ao do modelo com os dados de 2021, demonstrando assim, que o modelo tem um poder preditivo bem significativo.

Tabela 9. Matriz de confusão para os valores observados em 2022 (jan.-mar.) e preditos sob o valor de "cutoff" = 0,036, com o modelo de regressão logística binária com dados de 2021

Resposta predita pelo modelo	Resposta Observada		Totais
	Óbito	Não óbito	
Óbito	660	7.940	8.600
Não óbito	311	18.434	18.745
Totais	971	26.374	27.345

Fonte: Resultados originais da pesquisa
Nota: Sensitividade = 0,6797; Especificidade = 0,6989; e Eficiência global do modelo = 0,6982

Além da eficiência do modelo, ressalta-se que alguns resultados obtidos nesse trabalho corroboram com outros trabalhos da literatura, como Junior et al. (2019), que, com um modelo de regressão logística binária com dados dos acidentes em rodovias federais no ano de 2016 identificou alguns fatores que aumentam a chance de letalidade, tais como a rodovia ser localizada na região Nordeste; a pista ser do tipo simples; o tipo de acidente ser por colisão; durante a madrugada; e no dia pertencente ao fim de semana.

No estudo de Roquim et al. (2019), com dados de 2018, com menos variáveis explicativas e com o uso do modelo de regressão logística binária, o resultado obtido foi similar no quesito causa do acidente por falha humana, que faz com que a probabilidade de óbito seja maior nessa característica de acidente.

Após todos esses resultados apresentados nesse trabalho, um dos objetivos foi elaborar um "dashboard" com os resultados descritivos dos dados de acidentes, para que o leitor se aprofunde mais sobre o tema e fazer simulações com o modelo estimado nesse trabalho. Assim, o usuário poderá selecionar as características dos acidentes e dos envolvidos nos acidentes e obter a probabilidade de óbito e a classificação se sobreviveria ou não ao acidente. O acesso ao *dashboard* poderá ser via link ou QR Code, conforme descrito na Figura 12.

Figura 12. QR Code do "dashboard" do modelo de regressão logística binária aplicada a dados de acidentes em rodovias federais no Brasil
Fonte: Resultados originais da pesquisa
Link:
https://app.powerbi.com/view?r=eyJrIjoiYzY0ODk4MTMtZmU2My00MDM2LTg3ZjktN2ZkM2U3YTY2ZTRhIiwidCI6IjM4YTZiNDFiLTJmOWEtNGFiMi1hYjJhLTYzOWE0M2M0ZDQ3YSJ9&pageName=ReportSectionf7a2416711706e78c82d

Considerações Finais

Ao utilizar o modelo de regressão logística binária no intuito de estimar a probabilidade de óbito em acidentes nas rodovias Federais no Brasil, foi possível identificar algumas características dos acidentes e dos envolvidos que elevam a sua chance de letalidade. Essas características podem ser utilizadas como objetos de estudos e de ações que ajudem a diminuir o número de fatalidades nas rodovias federais do nosso país, assim como o prejuízo financeiro causado. Uma das soluções possíveis é investir mais na qualidade das nossas rodovias federais, assim como na construção de mais rodovias com pistas duplas, dado que nessas localidades o risco de letalidade é menor. Além disso, como as maiores causas de acidentes são por falha humana, é necessário que haja mais políticas de educação no trânsito, principalmente em relação ao respeito pelos ciclistas, motociclistas e pedestres. Destaca-se que eficiência global do modelo estimado foi considerada bem significativa para o objeto de estudo e, por meio do “dashboard”, os usuários conseguirão de forma prática utilizar e entender como funciona os resultados de um modelo de regressão e que sirva para tomada de decisão.

Referências

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

Carvalho, M. S. et al. 2011. Análise de sobrevivência: teoria e aplicações em saúde. SciELO-Editora; FIOCRUZ. Rio de Janeiro, RJ, Brasil.

Colosimo, E. A.; Giolo, S. R. 2006. Análise de sobrevivência aplicada. São Paulo: Edgard Blucher.

Confederação Nacional do Transporte [CNT]. Painel CNT de Consultas Dinâmicas dos Acidentes Rodoviários – 2021. Disponível em: https://www.cnt.org.br/painel-acidente. Acesso em: 22 nov. 2021. 6 p.

Fávero, L. P.; Belfiore, P. 2017. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier, Rio de Janeiro, RJ, Brasil.

Giolo, S. R. 2017. Introdução à Análise de Dados Categóricos com Aplicações. Editora: Blucher - Projeto Fisher ABE.

Izbicki, R. e dos Santos, T. M. 2020. Aprendizado de máquina: uma abordagem estatística.

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

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

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 Plena 17(4). Disponível em: https://www.scientiaplena.org.br/sp/article/view/5897/2382. Acesso em: 31 out. 2021.

Polícia Rodoviária Federal [PRF]. 2021. Disponível em: https://arquivos.prf.gov.br/arquivos/index.php/s/n1T3lymvIdDOzzb. Acesso em: 22 de fevereiro de 2022.

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

R Core Team and contributors worldwide. stats: The R Stats Package. R package version 4.2.0. 2021. Disponível em: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html. Acesso em 30 out. 2021.

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, Alfenas 8(2): 19-28.

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

Schwarz, G. 1978. Estimating the dimensional of a model. Annals of Statistics, Hayward 6: 461-464.

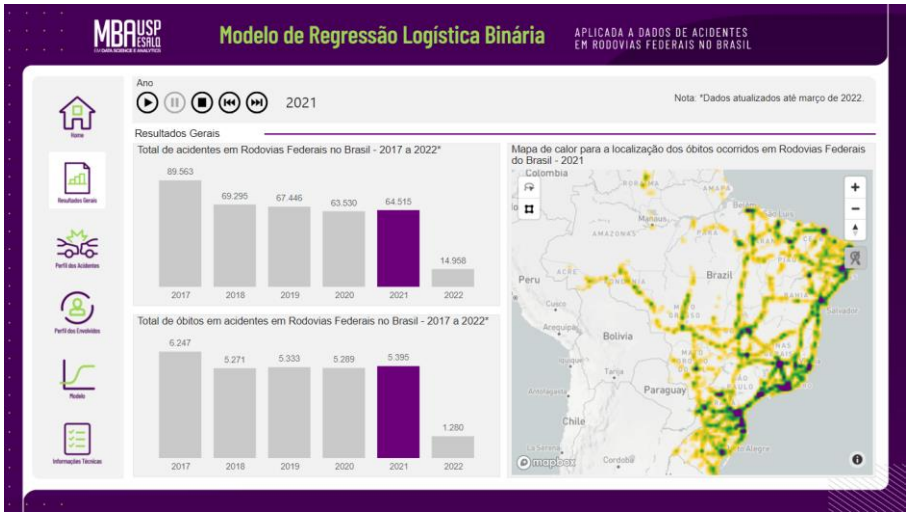
Sugiura, N. 1978. Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by Akaike's. Communications in Statistics – Theory and Methods 7(1): 13-26.

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

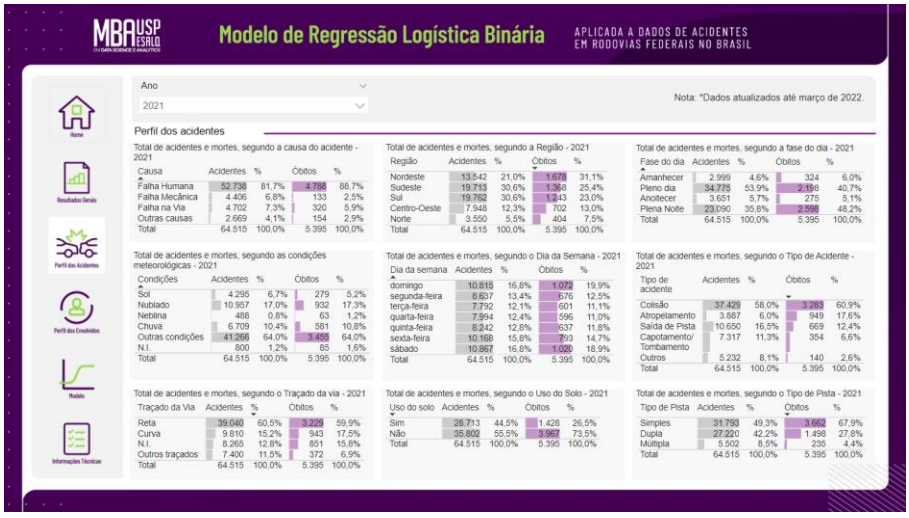
World Health Organization [WHO]. Global status report on road safety 2015. Geneva: WHO, 2015. Disponível em: https://shortest.link/whointviolenceinjuryprevention. Acesso em 30 out. 2021.

Apêndice

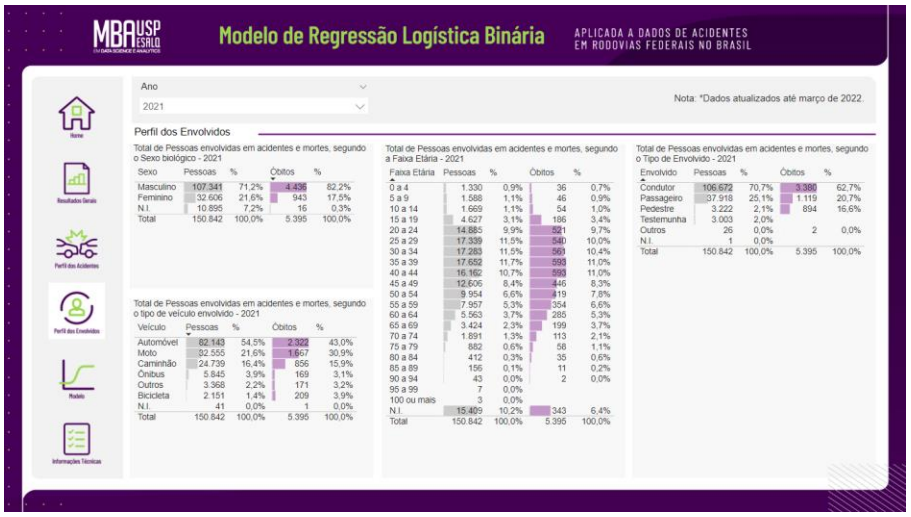
a)



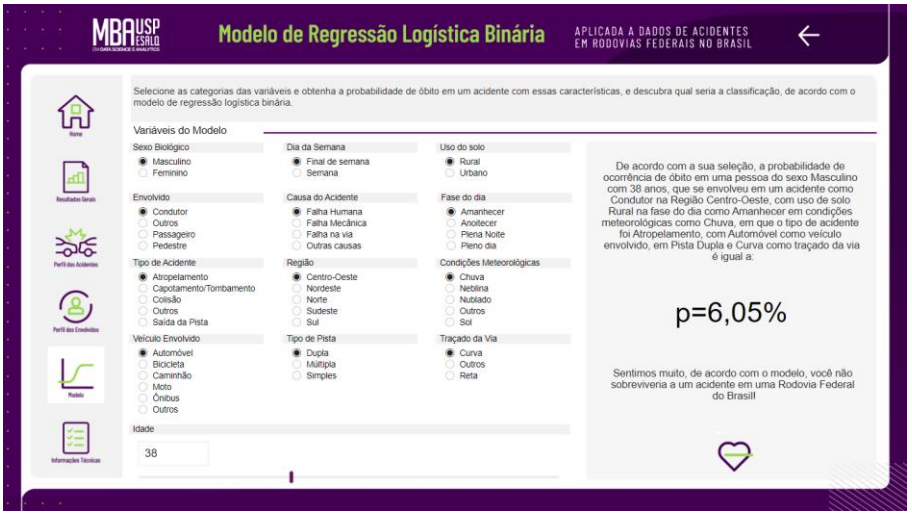
b)



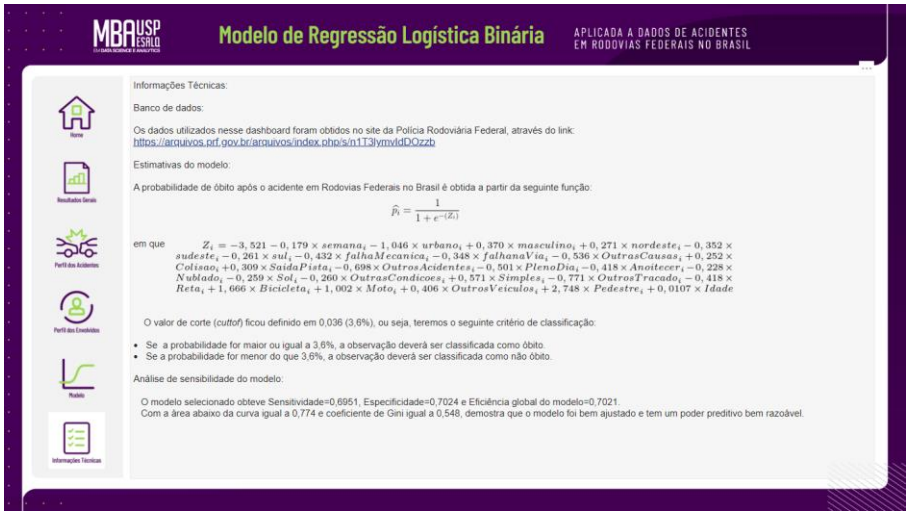
c)



d)



e)



f)

Figura 13. Dashboard desenvolvido. a) Capa; b) Resultados gerais: análise da série histórica do número de acidentes e sua respectiva localização georreferenciada; c) Perfil dos acidentes; d) Perfil dos envolvidos no acidente; e) Modelo: o usuário pode selecionar as variáveis e obter a probabilidade de óbito em decorrência de um acidente; f) Informações técnicas