Aprendizado de máquina, Internet das Coisas e a modernização do gerenciamento de trânsito no Brasil

Claudio Viana de Sousa [1]*; Mateus Modesto[2]

[1] Empresa de Tecnologia e Informações da Previdência. Analista de Processamento. R. Cosme Velho, 6 – Cosme Velho; 22241-900 Rio de Janeiro, RJ, Brasil
[2] Pecege. Msc. Eng. de Produção e de Manufatura. Parque Tecnológico - R. Cezira Giovanoni Moretti, 600 - Santa Rosa, Piracicaba, SP, Brasil
*autor correspondente: claudio.vianas@gmail.com

Aprendizado de máquina, Internet das Coisas e a modernização do gerenciamento de trânsito no Brasil

Resumo

A expansão acelerada dos sistemas de transporte colaborou para um aumento significativo no número de acidentes no ecossistema do trânsito. Nas últimas décadas houve uma aceleração no grau desses acidentes, sendo estes motivados por diversos fatores e acarretando um crescimento de vítimas. Considerando este cenário dinâmico e de constante transformação, este estudo apresenta como proposta a análise dos fatores críticos envolvidos, buscando orientar e dar suporte à tomada de decisões. Deste modo, permitindo o desenvolvimento de propostas, controles e a aplicação mais adequada de recursos e ferramentas tecnológicas baseadas em Internet das Coisas. Além disso, almeja permitir a observação dos impactos com as ações de tomada de decisões, de forma que os acidentes automobilísticos possam ser reduzidos ou evitados. Nesse sentido, para tais análises, foram utilizados os dados de acidentes rodoviários registados em território brasileiro no período entre 2017 e 2021 e disponibilizados pela Polícia Rodoviária Federal. Quanto aos métodos de análise, foram utilizadas as técnicas supervisionadas de aprendizado de máquina de regressão logística multinomial, árvore de decisão, redes neurais e “Gradient Boosting”. Por fim, durante as análises, foram estimados parâmetros e correlações entre os diversos fatores referentes às ocorrências, tais como as causas e tipos de acidentes.
Palavras-chave: acidentes; decisão; regressão; neurais; boosting.

Machine learning, Internet of things and the modernization of traffic management in Brazil

Abstract

The accelerated expansion of transport systems has contributed to a significant increase in the number of accidents in the traffic ecosystem. In recent decades there has been an acceleration in the degree of these accidents, which are motivated by several factors and causing an increase in victims. Considering this dynamic and constantly changing scenario, this study proposes the analysis of the critical factors involved, seeking to guide and support decision-making. Thus, allowing the development of proposals, controls and the most appropriate application of resources and technological tools based on the Internet of Things. In addition, it aims to allow the observation of impacts with decision-making actions, so that car accidents can be reduced or avoided. In this sense, for such analyses, data from road accidents recorded in Brazilian territory in the period between 2017 and 2021 and made available by the Polícia Rodoviária Federal were used. As for the analysis methods, supervised machine learning techniques of multinomial logistic regression, decision tree, neural networks and Gradient Boosting were used. Finally, during the analyses, parameters and correlations were estimated between the various factors referring to the occurrences, such as the causes and types of accidents.
Keywords: accidents; decision; regression; neural; boosting.

Introdução

No cenário de crescimento e disseminação das novas tendências tecnológicas impulsionadas pela evolução das redes móveis, como por exemplo a rede 5G[1], a Internet das

[1] 5G - Consiste na evolução da tecnologia de banda larga sem fio, sendo o passo seguinte da geração do 4G.

Coisas [IoT] avança para se consolidar dentre as tecnologias denominadas disruptivas. Nesta conjuntura, com um amplo potencial de transformação e valor, emergem as condições para incrementar melhorias de estrutura, integração e performance. Permitindo, assim, que todo o sistema de mobilidade urbana se torne cada vez mais eficaz e seguro à sociedade.

No ano de 2020 o Laboratório Hitachi-UTokio [H-UTOKYO LAB], no Japão, publicou a obra “Society 5.0: A People-centric Super-smart Society”, onde mencionou que o IoT está presente através de objetos físicos interligados e integrados por sensores que coletam informações sobre inúmeras atividades nas redes. Tais sensores visam analisar as atividades mais próximas de sua ocorrência, trazendo mais agilidade nas tomadas de decisões baseadas em informações; podendo estas serem referentes a determinada localidade, área ou região. Ainda, neste sentido, determinada parcela dos veículos das novas gerações possuem uma considerável quantidade de dispositivos tecnológicos embarcados cuja finalidade é medir, integrar, monitorar e analisar; visando garantir um melhor auxílio aos seus condutores.

Ainda na obra “Society 5.0: A People-centric Super-smart Society”, o H-UTOKYO LAB revela que tecnologias como IoT, em conjunto com “machine learning”, permitiram à cidade de Barcelona, na Espanha, otimizar o controle do tráfego de veículos por meio de ajustes dos semáforos de trânsito. Tal medida permitiu uma melhoria no fluxo viário dentro do perímetro urbano. Logo, se verifica que através da utilização de dados e técnicas de aprendizado de máquina, é possível se atingir um significativo progresso no comando e fiscalização, conforme proposto nos modelos de cidades inteligentes.

De acordo com Maschietto et al. (2021), bem como para Pinheiro e Crivelaro (2020), em escala global, a Internet das Coisas possui cada vez mais influência nos grandes centros urbanos, vindo a ser um fator de relevância para que as cidades tenham um melhor aproveitamento de seus recursos, possibilitando uma expressiva e contínua evolução às suas populações. Em consequência disso, se observa que inúmeras ações e meios podem ser empregados a fim de evitar as ocorrências de acidentes. Por exemplo, o aumento na capacidade de vigilância e melhoria no monitoramento climático em determinadas áreas das cidades. Nesse sentido, no ano de 2015, o Instituto de Pesquisa Econômica Aplicada [IPEA] estimou que uma das principais causas de mortes no Brasil está relacionada aos acidentes de trânsito.

Considerando todo esse cenário de transformação, este estudo tem como propósito explorar e analisar os dados dos acidentes de trânsito no Brasil. E, a partir disso, observar como o tratamento e modelagem desses dados, por meio de técnicas de aprendizado de máquina, podem orientar para uma melhor aplicação dos recursos de Internet das Coisas na administração do trânsito. Portanto, visando fornecer informações que permitam aprimorar a segurança na mobilidade e melhorar o gerenciamento do tráfego nas rodovias brasileiras.

Material e Métodos

Para a realização do estudo foi utilizada a base de dados de acidentes rodoviários registrados em território brasileiro, no período entre 2017 e 2021, disponibilizada pela Polícia Rodoviária Federal [PRF] do Brasil.

Na preparação e tratamento dos conjuntos de dados foram utilizados procedimentos de “data wrangling”, objetivando a formatação, estruturação, limpeza, normalização e padronização.

Além disso, para atendimento das propostas do estudo, definiu-se como a variável de resposta a classificação do acidente. Por vez, como categoria de referência, foi selecionada a classe com vítimas fatais.

Em relação às variáveis explicativas, foram utilizadas aquelas referentes às causas, tipos, condições meteorológicas, classificação do acidente, data, além de outras previamente selecionadas.

Regressão logística

De acordo com Fávero e Belfiore (2017) os modelos de regressão são técnicas supervisionadas de aprendizado de máquina. Estes modelos estão entre as melhores formas de análise de previsão de comportamentos. Para a elaboração de um modelo preditivo eficaz, foram realizados procedimentos e técnicas baseadas no modelo de regressão logística. Tal método integra a categoria dos modelos lineares que utilizam uma função linear na classificação dos registros. Este modelo de classificação estimará os valores discretos e valores binários que preveem a probabilidade de ocorrência de um evento, através do ajuste de uma função “logit”. Com isso, por meio dos resultados das previsões, compreender através da variável preditora quais acidentes com vítimas fatais mais tendem a ocorrer no trânsito. Abaixo é apresentada a eq. (1) referente a regressão logística multinomial:

$$\ln\left(\frac{P_{im}}{1- P_{im}}\right) = \propto m + \beta 1m.X1i + \beta 2m.X2i + \cdots + \beta km.Xki \qquad (1)$$

onde, P_m: significa a probabilidade de acontecimentos de cada uma das m categorias da variável Y dependente; $\propto$ representa a constante; β são os parâmetros importantes para cada variável X que são as variáveis explicativas; i corresponde a cada observação da amostra.

Por fim, para a criação do algoritmo de regressão logística multinomial, foi utilizado o pacote “sklearn” da linguagem Python, no qual foram efetuados ajustes dos hiperparâmetros padronizados para o referido modelo de regressão.

Árvore de decisão

A análise de classificação através do modelo de árvore de decisão foi realizada com o objetivo de verificar uma melhor acurácia e precisão. Desta maneira, foi possível identificar as características que mais resultaram em acidentes com vítimas fatais, visando propor ações necessárias para a redução destes números.

As árvores de decisão são algoritmos utilizados para regressão e classificação que se baseiam em estrutura de árvore, podendo serem associadas às inferências do tipo “se/então”. Este modelo possui como objetivo a construção de uma árvore que represente, explicitamente, a estrutura do conjunto de dados. Uma árvore de decisão fornece, através da disposição de nós, uma interpretação intuitiva do conjunto de variáveis explicativas para o critério de classificação. Também, há o índice de Gini (eq. 2)que é um critério baseado em impurezas que mede as divergências entre as distribuições de probabilidade do alvo e dos valores do atributo (Rokach e Maimon, 2014).

$$\text{Gini}(D) = \sum_{i=1}^{m} Pi^2 \qquad (2)$$

onde, Pi: significa a probabilidade das instâncias em D que se referem a classe i.

Para a elaboração do modelo de árvore de decisão foram executados ajustes, tais como a profundidade da árvore e o índice de Gini, respectivamente, a fim de evitar a ocorrência de sobreajuste, além de medir a qualidade das divisões.

Na Figura 1 é apresentado o pseudocódigo do algoritmo de uma árvore de decisão “Classification and Regression Trees” [CART]. Na linha 1, se tem a decisão do critério de parada da construção da árvore, no qual se D pertence a uma única classe, se retorna uma folha. Por vez, na linha 7 ocorre o critério de escolha do melhor atributo através do índice de Gini. Já na linha 11 é verificado se D está vazio. Assim, caso inexista valor, será adicionado um nó à folha, na classe mais comum de Dv. Por fim, na linha 14, ocorre o retorno do subconjunto Dv do algoritmo CART, terminando com a adição da árvore. Assim, o algoritmo escolhe as melhores divisões e repete esse processo, recursivamente, até que o conjunto ideal seja encontrado.

```
CART(D)
Entrada: Um valor de atributo do dataset D
01:   Se D ∈ a uma única classe então
02:   retorna um nó folha.
03:   Senão
04:   Se D estiver vazio então
05:   retorna um nó folha com a classe mais comum de D.
06:   Senão
07:   Escolhe o melhor atributo F e cria dois nós.
08:   Para cada possível valor vi de F
09:   Seja Dv subconjunto que tenha valor vi para F
10:   Adicione as arestas a partir dos nós com o valor vi.
11:   Se Dv estiver vazio então
12:   Adicione um nó folha ligado a aresta com a classe v mais comum de Dv.
13:   Senão
14:   Árvore = CART(Dv)
15:   Retorne a Árvore
```

Figura 1. Pseudocódigo do modelo de árvore decisão
Fonte: Okada e Neves (2019)

Redes neurais

Também foram empregados algoritmos de redes neurais artificiais [RNA], no qual foi elaborado um modelo preditivo por classificação, através de redes neurais “Multilayer Perceptron” [MLP], a fim de utilizá-lo na variável “target” de classificação dos acidentes com vítimas fatais. De modo que, através dos resultados obtidos, seja possível identificar quais os acidentes de trânsito tendem a ocorrer com maior frequência. E assim, permitir extrair uma melhor predição e, consequentemente, melhores ações na prevenção destes acidentes.

As RNA são inspiradas e têm como objetivos simular as redes neurais biológicas. Um “perceptron”, por exemplo, é a forma mais simples de um modelo de rede neural. Possuindo apenas um neurônio de entradas binárias e assumindo valores de 0 ou 1. Assim, se a soma das saídas ponderadas for maior que 0, a saída do “perceptron” será 1; sendo que uma classe é reconhecida. No entanto, caso contrário, o “perceptron” não reconhece uma classe (Grus, 2021). A seguir, observa-se a Figura 2.

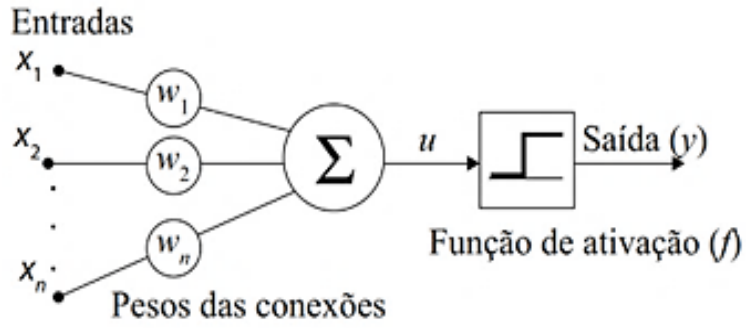


Figura 2. Interconexão entre os neurônios artificiais e seus valores de entrada e saída
Fonte: Silva et al. (2019)

As redes neurais “Multilayer Perceptron” são estruturadas com mais de uma camada de neurônio. Possuem uma camada de entrada e uma camada intermediária oculta, com um fluxo de combinação linear que passa por uma função de ativação, na qual se define o sinal que será propagado pela rede ou não, sendo uma camada de saída. Desse modo, as camadas são ligadas entre si através de pesos ajustáveis e o seu aprendizado, geralmente, é realizado através do algoritmo de retropropagação do erro (Silva et al., 2019).

No âmbito deste estudo foi utilizado o algoritmo “Multilayer Perceptron” com um neurônio de saída (Y) para a variável dependente com vítimas fatais e (X) para os valores das variáveis explicativas na entrada, além de uma camada oculta com três neurônios. Sendo utilizado a função de ativação “Rectified Linear Unit” [ReLu] para calcular os valores dos neurônios na camada oculta e na camada de saída, além do otimizador estocástico de Adam para correção dos pesos e “random_state” no valor de três, para geração de números aleatórios para inicialização de pesos e viés, ambos do pacote sklearn.

“Gradient Boosting”

Com o objetivo de ampliar as análises e obter resultados complementares, também foi utilizado o método “Gradient Boosting”. Este é um método de “ensemble” que atua combinando modelos e cuja finalidade é melhorar o desempenho. Nesse sentido, os métodos mais conhecidos são “Bagging”, “Boosting” e “Stacking”. O método de “Gradient Boosting” dispõe de elevado fator de potência, uma vez que ele ajusta o novo previsor aos erros residuais ocasionados pelo previsor anterior, assim melhorando a performance do modelo (Géron, 2019).

Para o desenvolvimento do algoritmo de “Gradient Boosting” foram realizados os ajustes dos hiperparâmetros, tais como a quantidade de árvores, limitação da profundidade e taxa de aprendizado. Na Figura 3 é mostrado a construção do modelo de “Gradient Boosting” através do pseudocódigo.

Initialize $f_0(x)$ to be a constant, $f_0(x) = argmin_\rho \sum_{i=1}^{N} L(y_i\ ,\ \rho)$.

For $m = 1\ to\ M$ do

For $i = 1$ to n do

Compute the negative gradient

$$z_{im} = -\left[\frac{\partial L\left(y_i\ , f(x_i)\right)}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

End;

Fit a regression tree $g_m(x)$ to predict the targets z_{im} from covariates x_i for all training data.

Compute a gradient descent step size as

$$\rho_m = argmin_\rho \sum_{i=1}^{n} L(y_i\ , f_{m-1}(x_i) + \rho g_m(x_i)\ )$$

Update the model as

$$f_m(x) = f_{m-1}(x) + \rho_m g_m(x)$$

End;

Output the final model $f_M(x)$

Figura 3. Pseudocódigo do modelo de “Gradient Boosting”
Fonte: Zhang e Haghani (2015)

Métricas de avaliação

Ao longo do estudo foram utilizadas algumas métricas para avaliar a qualidade dos modelos e suas performances. De tal modo, foram considerados os problemas propostos na pesquisa, visto que existem métricas específicas para cada tipo de problema, sejam eles de classificação ou regressão. Assim, neste estudo, foram utilizadas métricas relacionadas a classificação na qual, por exemplo, um acidente com vítimas fatais é considerado de classe positiva.

Deste modo, se tem a matriz de confusão que é um recurso gráfico importante que visa facilitar a observação dos registros realizados pelo modelo em análise. Em uma de suas diagonais se encontram os registros corretamente classificados. Por vez, na diagonal oposta, são aplicados os registros classificados como incorretos. Abaixo, na Figura 4, se observa a demonstração da estrutura de uma matriz de confusão.

		Valores Reais	
		Positivos	**Negativos**
Valores Preditos	**Positivos**	Verdadeiros Positivos (VP)	Falso Positivos (FP)
	Negativos	Falso Negativos (FN)	Verdadeiros Negativos (VN)

Figura 4. Representação da matriz de confusão
Fonte: Dados originais da pesquisa

Uma das métricas mais utilizadas é a acurácia, pois ela informa o quanto o modelo gerado acertou em suas previsões. Sendo, a soma dos verdadeiros positivos e verdadeiros negativos, dividida pela soma total das previsões, cuja eq. (3) segue abaixo:

$$\text{Acurácia} = VP + VN \,/\, VP + VN + FP + FN \quad (3)$$

onde, respectivamente, VP são os valores verdadeiros positivos; VN são os valores verdadeiros negativos; FP sendo os valores falsos positivos; e FN sendo os valores falsos negativos.

De acordo com Rokach e Maimon (2014), existem outras métricas, tal como a precisão, que verifica o número de registros que o modelo previu como positivos o quanto realmente o são. Nisso, o principal objetivo é limitar os falsos positivos. Por vez, o “recall” é uma métrica que verifica todos os registros que realmente são positivos, quanto o modelo previu corretamente como positivos. Já o “F1-score” é a média harmônica entre as métricas de precisão e o “recall”. Por fim, a curva das características operacionais do receptor [ROC] que é uma métrica que combina a sensibilidade e a especificidade. Além disso, a partir da ROC também é possível medir a área abaixo da curva, conhecida por AUC.

Validação cruzada

Conforme dispõe Géron (2019), a validação cruzada de “K-fold” melhora o modelo através da validação dos dados. Com isso, K é a quantidade de separações de “folds” da base de dados, a qual é dividida em duas partes, sendo elas de treino e teste. Ou seja, este método divide o conjunto de dados em subconjuntos com um número “folds” e repete o método de validação cruzada K em determinado número de vezes até que a validação esteja completa. Desta maneira, a validação cruzada melhora a performance do modelo, uma vez que toda a base de treino e teste é analisada de forma mais eficiente. A seguir, a Figura 5 representa a estrutura de funcionamento da validação cruzada “K-fold”.

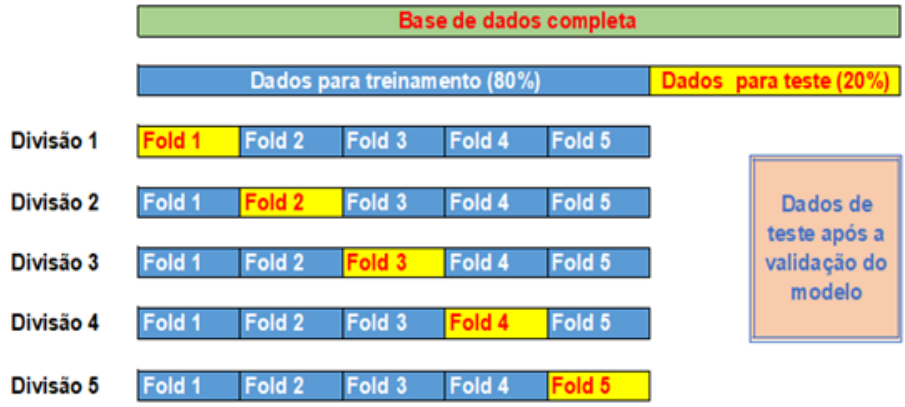


Figura 5. Representação da validação cruzada “K-fold”
Fonte: Dados originais da pesquisa

Em relação ao banco de dados utilizado no estudo foi verificada a ocorrência de desbalanceamento entre os registros de acidentes com vítimas fatais e os demais acidentes com vítimas feridas. Por tal motivo, optou-se por utilizar uma técnica apropriada a fim de efetuar o balanceamento da referida base de dados. Para isso, empregou-se a técnica de “Smote Synthetic Minority Oversampling” que é um procedimento de sobreamostragem dos dados de forma sintética. E assim, estabelecer um ajuste para a classificação desbalanceada, tendo em vista que a assimetria da base de dados pode influenciar durante a criação do modelo preditivo (Molin, 2019).

Em conclusão, para a execução dos procedimentos e técnicas, bem como para a análise e construção dos modelos e resultados, foi utilizada a plataforma Google Colab, baseada em linguagem Python. Além disso, para a execução das técnicas, manipulações e procedimentos relativos aos conjuntos de dados, foi usada a linguagem R em ambiente RStudio. Ainda, utilizou-se recursos do Microsoft Excel.

Resultados e Discussão

Para a concepção e elaboração do estudo foi realizado o carregamento dos registros de dados de acidentes através do portal da Polícia Rodoviária Federal. Esses dados contêm informações sobre vítimas, veículos, causas, tipos de acidentes, entre outras classes relevantes para o desenvolvimento deste estudo.

A Polícia Rodoviária Federal atende cerca de 70 mil quilômetros de rodovias e estradas federais em todo território brasileiro, sendo uma de suas demandas públicas o atendimento a acidentes rodoviários.

Para o estudo foram utilizados os dados coletados entre os anos de 2017 e 2021. Esses dados estavam disponibilizados, para cada ano, em formato de arquivo “comma-separated values” [csv]. Foram observadas diversas variáveis de relevância para o âmbito da

pesquisa, tais como: classificação do acidente, tipo do acidente, causa do acidente, clima, pessoas, veículos, além de variáveis de via (sentido da via, tipo da pista e traçado), variáveis de localização (UF, BR, Km e município), variáveis de data (data, ano, dia semana, horário e fase do dia). Desse modo, essas classes foram inicialmente estabelecidas como as mais adequadas para o início da análise exploratória.

Através dos históricos de acidentes e por meio dos algoritmos dos modelos supervisionados, foi possível identificar e classificar as características que mais ocasionam acidentes com vítimas fatais e advertir sobre ações preventivas para redução desses números. Além disso, também se almejou verificar como a aplicação das tecnologias relacionadas à Internet das Coisas podem elevar a eficiência e otimizar a segurança no cenário do trânsito. Pretendeu-se, também, obter um melhor entendimento sobre os eventuais impactos positivos que podem ser acarretados com a utilização de IoT no contexto da mobilidade urbana, assim como no auxílio e desenvolvimento de gestão das cidades.

De início, durante as análises descritivas, foi efetuada a junção dos arquivos dos períodos de 2017 a 2021, de forma que apresentassem o mesmo formato. Ainda, nesse sentido, se realizou o tratamento de valores faltantes, padronizações e a exclusão de variáveis irrelevantes ao contexto do estudo. Realizadas as ações iniciais, se verificou a quantidade de 353.625 mil observações e 23 variáveis.

Desta forma, conforme disposto na Tabela 1, foi realizada a análise exploratória a fim de examinar a classificação dos acidentes. Logo, observou-se que a quantidade de ocorrências com vítimas fatais foi de 6% e a quantidade de eventos com vítimas feridas correspondeu a 70% do total de acidentes. Essa desproporção entre o número de vítimas fatais e feridas aponta para um desequilíbrio na base de dados, indicando um desbalanceamento entre as classes, o que tende a acarretar inconsistências à análise preditiva do modelo.

Tabela 1. Quantidade de acidentes por classificação

Classificação do Acidente	Quantidade	Porcentagem
Com vítimas fatais	23.448	6,63%
Com vítimas feridas	249.615	70,59%
Sem vítimas	80.562	22,78%
Total	353.625	100,00%

Fonte: Polícia Rodoviária Federal

A seguir, na Tabela 2, é apresentado o quantitativo dos três tipos de classificação, considerando-se os dez estados com mais registros no período abarcado pelo estudo.

Tabela 2. Dez estados por classificação de acidentes

UF	Com vítimas fatais	Com vítimas feridas	Sem vítimas
MG	2.882	34.245	9.691
PR	2.294	28.294	10.012
BA	2.017	12.329	3.973
SC	1.634	31.763	9.104
PE	1.400	9.015	3.491
RJ	1.346	17.017	5.383
RS	1.301	16.249	6.429
GO	1.177	12.196	4.107
SP	1.033	16.257	5.789
MA	934	3.708	1.528

Fonte: Polícia Rodoviária Federal

Em complemento, na Tabela 3, foi exposta a análise das principais causas de acidentes fatais, onde foi constatado que a falta de atenção à condução esteve em primeiro lugar com 28%, sendo seguida por velocidade incompatível com 17% e a falta de atenção do pedestre com 13% do total de acidentes.

Tabela 3. Principais causas de acidentes fatais

Causa	Quantidade	Porcentagem
Falta de atenção à condução	4.956	28%
Velocidade incompatível	3.078	17%
Falta de atenção do pedestre	2.392	13%
Desobediência às normas de trânsito	2.376	13%
Ingestão de álcool	1.255	7%
Ultrapassagem indevida	1.243	7%
Condutor dormindo	1.129	6%
Defeito na via/pista escorregadia	604	3%
Transitar na contramão	515	3%
Mal súbito	421	2%
Total	17.969	100%

Fonte: Polícia Rodoviária Federal

Abaixo, na Tabela 4, foi apresentada a análise dos principais acidentes com vítimas fatais, considerando os tipos das ocorrências. Assim sendo, foi observado que as colisões frontais corresponderam a 27% dos casos. Em seguida estão os atropelamentos de pedestres, respondendo por 19% dos registros.

Tabela 4. Principais tipos de acidentes fatais

(continua)

Tipo	Quantidade	Porcentagem
Colisão frontal	6.257	27%
Atropelamento de pedestre	4.507	19%
Saída de leito carroçável	2.806	12%
Colisão traseira	2.586	11%
Colisão transversal	1.980	8%
Colisão lateral	1.147	5%

Tabela 4. Principais tipos de acidentes fatais

(conclusão)

Tipo	Quantidade	Porcentagem
Colisão com objeto estático	986	4%
Tombamento	975	4%
Capotamento	569	2%
Queda de ocupante de veículo	472	2%
Atropelamento de animal	357	2%
Colisão com objeto	258	1%
Colisão lateral mesmo sentido	169	1%
Colisão lateral sentido oposto	139	1%
Engavetamento	89	0%
Colisão com objeto em movimento	83	0%
Derramamento de carga	21	0%
Danos eventuais	20	0%
Eventos atípicos	19	0%
Incêndio	8	0%
Total	23.448	100%

Fonte: Polícia Rodoviária Federal

Em sequência das análises, na Figura 6, é mostrado um gráfico de barras com os cinco principais tipos de acidentes com vítimas fatais, respectivamente, em ordem crescente: colisão transversal, com 1.980 casos; colisão traseira, 2.586 casos; saída de leito carroçável, com 2.806; atropelamento de pedestre, com 4.507 casos; e colisão frontal, com 6.257 casos.

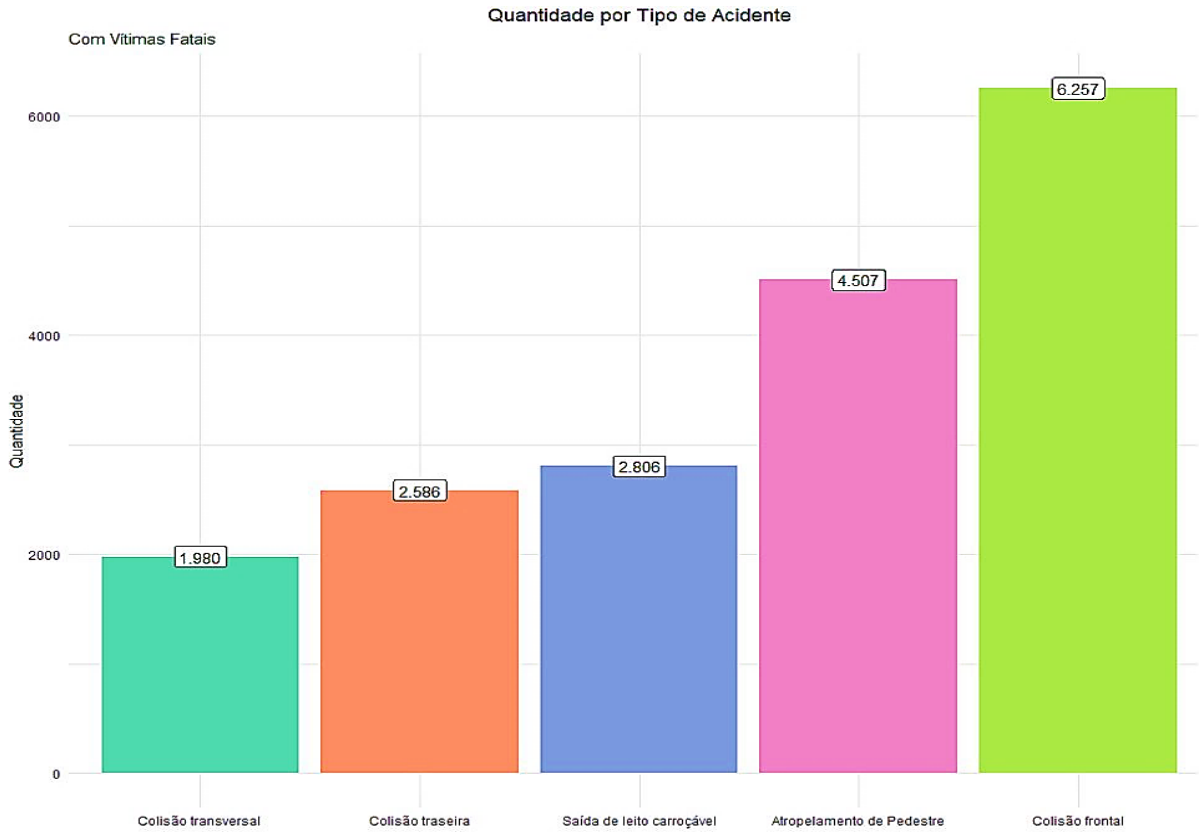


Figura 6. Quantidade de acidentes com vítimas fatais por tipo
Fonte: Polícia Rodoviária Federal

Abaixo, na Figura 7, se tem a análise exploratória que identificou a existência de relação entre os dias dos acidentes e a incidência de vítimas fatais. Nesse sentido, identificou-se que o número de vítimas tende a aumentar conforme a proximidade do período compreendido entre às sextas-feiras e domingos.

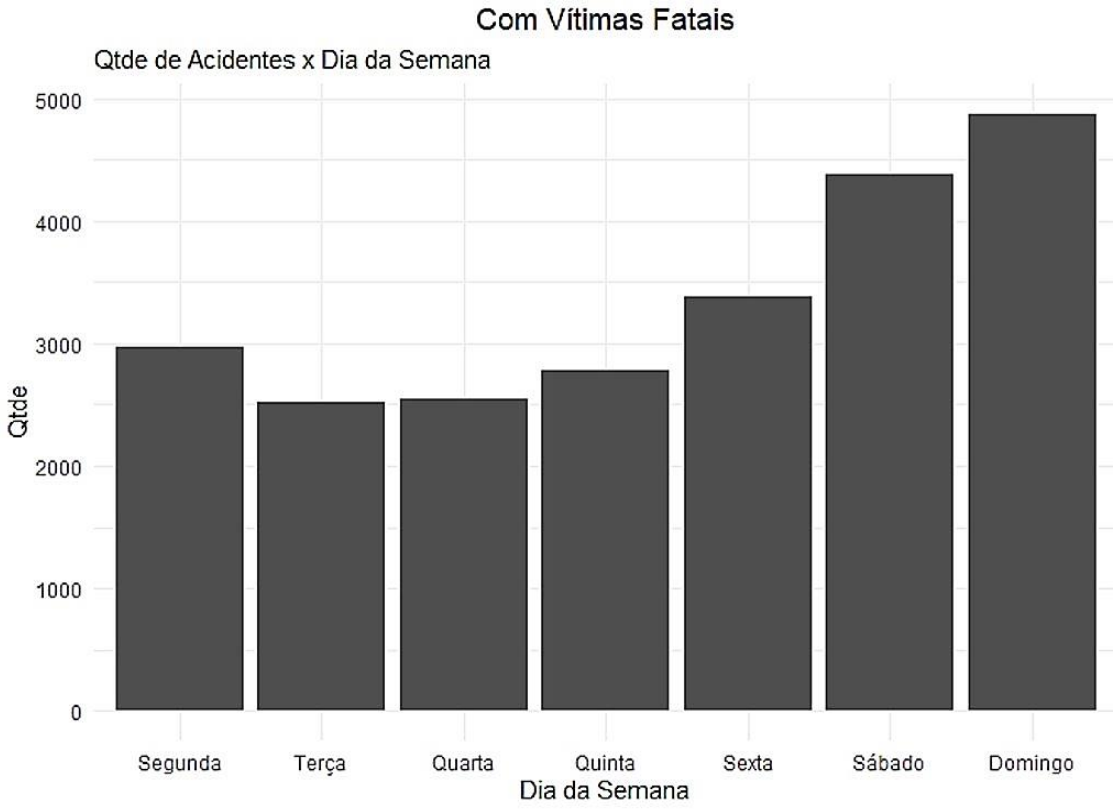


Figura 7. Quantidade de acidentes com vítimas fatais por dia da semana
Fonte: Resultados originais da pesquisa

Por vez, na Tabela 5, são apresentados os quantitativos de pessoas envolvidas em acidentes, considerando-se os dias da semana.

Tabela 5. Quantidade de pessoas envolvidas em acidentes por dia da semana

Dia	Pessoas envolvidas
Domingo	58.509
Sábado	58.937
Sexta-feira	54.666
Segunda-feira	47.784
Quinta-feira	46.141
Quarta-feira	44.282
Terça-feira	43.306
Total	353.625

Fonte: Polícia Rodoviária Federal

Observa-se, na Figura 8, que em relação as fases do dia, o período noturno é responsável por 50% dos acidentes com vítimas fatais. Além disso, também foi verificado que o período diurno é responsável por 57% dos acidentes com vítimas feridas.

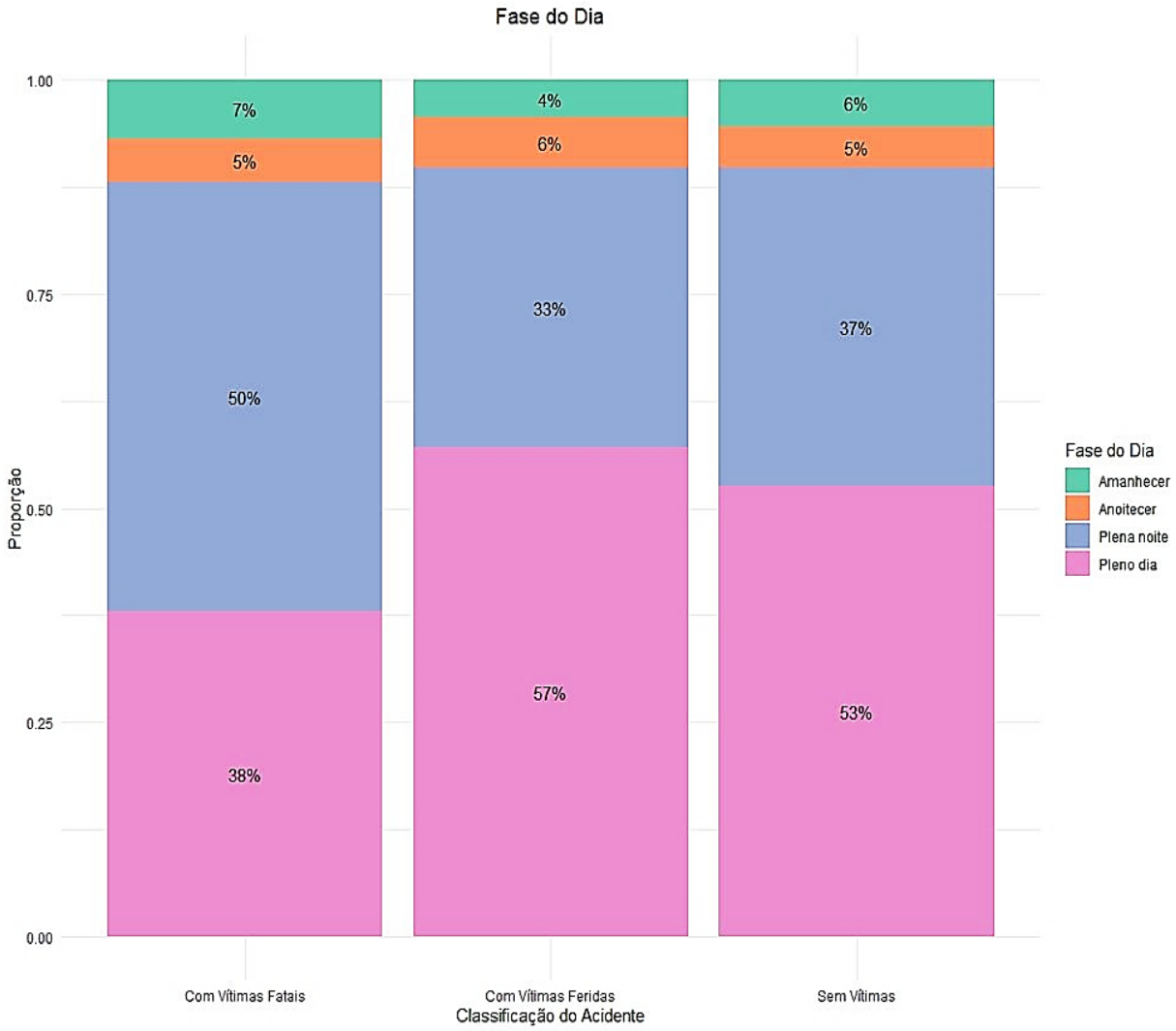


Figura 8. Proporção de acidentes por fase do dia
Fonte: Resultados originais da pesquisa

A seguir, na Figura 9, é apresentado um gráfico com os níveis de correlação e força entre as variáveis. Assim, expondo o nível de intensidade com que as variáveis estão se relacionando entre si. Por exemplo, se observa que as variáveis pessoas e veículos possuem uma correlação moderada entre si. Nesse sentido, também é observado que há considerável relação entre o tipo de acidente e veículos, o que pode demonstrar tendência e potencial de ocorrência entre esses fatores.

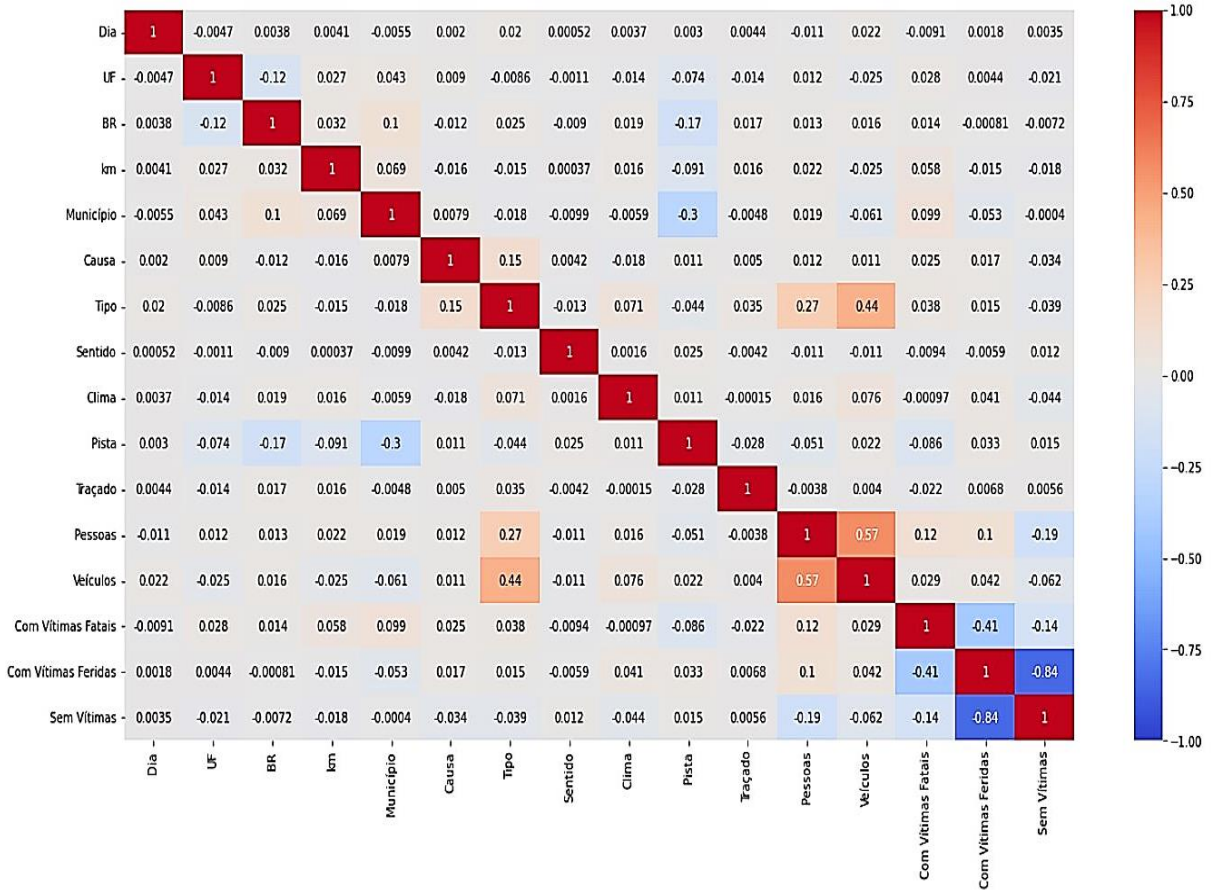


Figura 9. Gráfico de correlações
Fonte: Resultados originais da pesquisa

Para melhorar o aprendizado dos modelos construídos foi executado o processamento das variáveis e ajustes de transformações como, por exemplo, o realizado na coluna de classificação de acidentes, na qual se utilizou a forma de transformação “one-hot-encoding”. Isso, a fim de separar em novas variáveis com vítimas fatais, com vítimas feridas e sem vítimas os dados de cada categoria distinta, na forma binária de 0 ou 1, conhecidas como “dummies”. Neste caso, a variável com vítimas fatais foi utilizada como a variável dependente e as demais como variáveis explicativas.

Nesse contexto, também foi verificado que as variáveis pessoas e veículos apresentaram a incidência de “outliers”, os quais são pontos de dados discrepantes que se afastam da média e da variância, conforme exposto nos gráficos de “Boxplot”, na Figura 10.

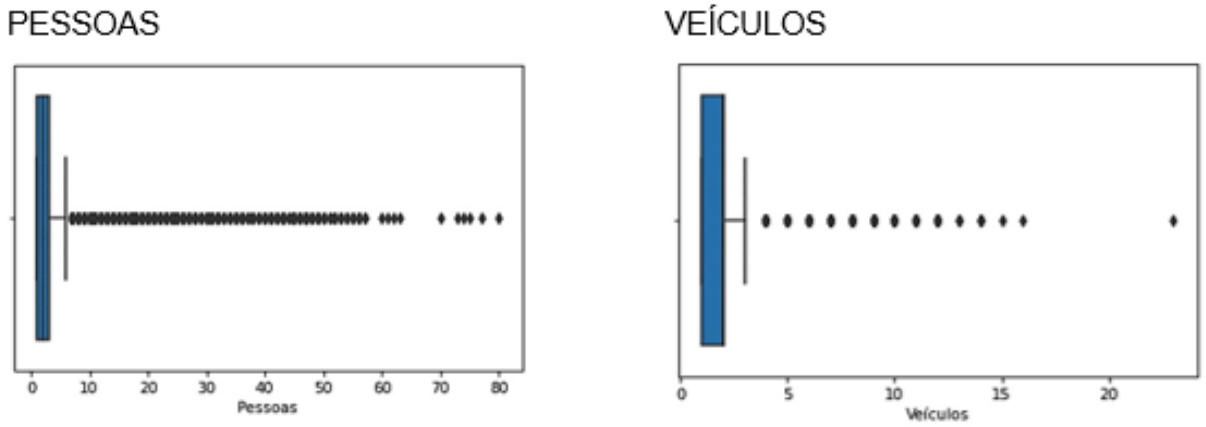


Figura 10. Gráficos de “Boxplot” das variáveis com “outliers”
Fonte: Resultados originais da pesquisa

Ressalta-se que uma vez identificados os problemas com “outliers”, é recomendável que eles sejam tratados, tendo em vista que podem interferir e gerar prejuízos às execuções e análises dos modelos. Assim sendo, foi realizada a remoção das observações divergentes da amostra, de modo a permitir uma melhoria nos resultados, conforme visto na Figura 11.

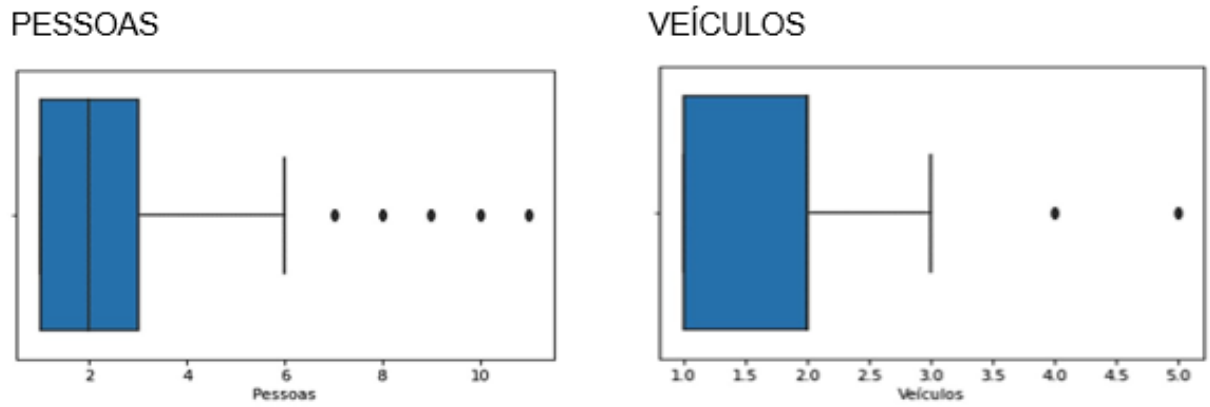


Figura 11. Gráficos de “Boxplot” com correção de “outliers”
Fonte: Resultados originais da pesquisa

Na sequência das ações se verificou que após realizar a criação das variáveis “dummies”, houve a ocorrência de elevada correlação entre as variáveis. Indicando, assim, a ocorrência de multicolinearidade**,** sendo isso confirmado com uso da métrica conhecida como “Variance Inflation Factor” [VIF]. Uma vez efetuada a análise foi verificado o valor VIF>5, que indica alta multicolinearidade. Após essa observação, removeu-se a variável com alta correlação e realizou-se um novo teste onde foi encontrado um valor VIF<5, indicando que a multicolinearidade foi solucionada (Fávero e Belfiore, 2017).

Durante a realização do pré-processamento foram utilizados três modelos de algoritmos com características singulares e de comparação, considerando-se os resultados obtidos para predição dos possíveis acidentes, sendo eles: regressão logística, redes neurais

e árvore de decisão. Além disso, se utilizou os métodos “ensemble”, a fim de melhorar o potencial destas previsões baseados nos modelos “Boosting”.

Nessa conjuntura, por meio da análise exploratória, foi detectado que a variável dependente com vítimas fatais, a ser utilizada como “target”, estava apresentando desbalanceamento, conforme disposto na Figura 12.

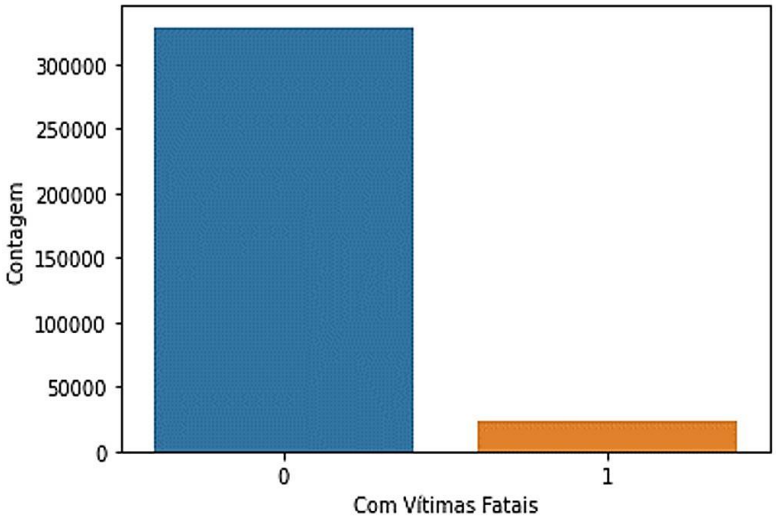


Figura 12. Gráfico de desbalanceamento da variável dependente
Fonte: Resultados originais da pesquisa

Em princípio, para a verificação das métricas, foram efetuados os treinamentos dos modelos sem balanceamento. Em seguida, avaliou-se os mesmos modelos, porém com balanceamento. Além disso, foi realizada a divisão dos dados, respectivamente, em 20% para teste e 80% para treinamento. Após a execução dos modelos se verificou algumas métricas de avaliação, tal como a acurácia.

Assim sendo, após as análises dos modelos sem balanceamento, foram extraídos os seguintes resultados, conforme a Tabela 6.

Tabela 6. Métricas dos modelos sem balanceamento

Modelo	Acurácia	Precisão	Recall	F1 score
Regressão logística	93,90%	95%	99%	97%
Árvore de decisão	95,75%	96%	99%	98%
Redes neurais	94,03%	96%	98%	97%
Gradient Boosting	96,05%	97%	99%	98%

Fonte: Resultados originais da pesquisa

Ressalta-se que durante a execução do modelo de rede neural foi utilizado o “Multilayer Perceptron” com três camadas ocultas e função de ativação ReLu.

Após o exame dos resultados preliminares obtidos a partir dos modelos desbalanceados, foi efetuado o balanceamento através do método de “Smote”. Essa é uma técnica “oversampling”, na qual se realiza o ajuste do conjunto de dados. Nisso, são criados dados sintéticos de modo que haja uma igualdade entre as classes positivas e negativas, evitando-se a ocorrência de “overfitting”, ou seja, o sobreajuste do modelo. O resultado do procedimento de balanceamento pode ser observado na Figura 13.

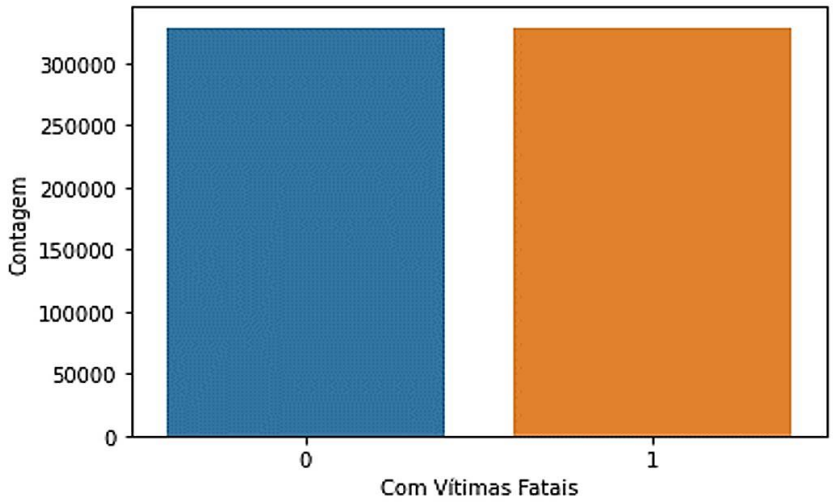


Figura 13. Gráfico de balanceamento da variável dependente pelo método “Smote”
Fonte: Resultados originais da pesquisa

Posteriormente, já com a base balanceada, se realizou um novo treinamento dos modelos. Nesse sentido, visando melhorar a análise e performance, também foi incluído nesta nova execução um modelo que utiliza o método de “ensemble” do tipo “Boosting”, conhecido como “Gradient Boosting”. Este modelo utiliza os erros residuais do modelo anterior para corrigir o modelo seguinte. Após a finalização dos treinamentos foram obtidas as métricas com os melhores ajustes, conforme verificados na Tabela 7.

Tabela 7. Métricas dos modelos após o balanceamento

Modelo	Acurácia	Precisão	Recall	F1 score
Regressão logística	89,35%	96%	82%	89%
Árvore de decisão	88,45%	91%	85%	88%
Redes neurais	89,36%	97%	81%	88%
Gradient Boosting	91,04%	96%	86%	91%

Fonte: Resultados originais da pesquisa

Ainda, com o objetivo de aplicar um melhor ajuste e performance, foi realizada a técnica de validação cruzada “K-fold”. Esta técnica consiste em uma divisão dos dados de treino numa combinação escolhida em partes iguais, onde essa parte é treinada com as combinações e validada com o restante. Após a execução da técnica foi verificada a média

da acurácia e seus intervalos. Nisso, se observa que o modelo de “Gradient Boosting” possui a melhor acurácia dentre os modelos testados, conforme visto na Tabela 8.

Tabela 8. Métricas dos modelos após a validação cruzada

Modelo	Acurácia Média	Intervalo de Acurácia
Regressão logística	93,85%	93,65% ~ 94,05%
Árvore de decisão	94,90%	93,25% ~ 96,56%
Redes neurais	94,24%	93,94% ~ 94,54%
Gradient Boosting	96,03%	95,79% ~ 96,28%

Fonte: Resultados originais da pesquisa

Em relação aos procedimentos de validação cruzada, também foram construídas as matrizes de confusão, possibilitando verificar o comportamento do algoritmo, conforme demonstrado na Figura 14.

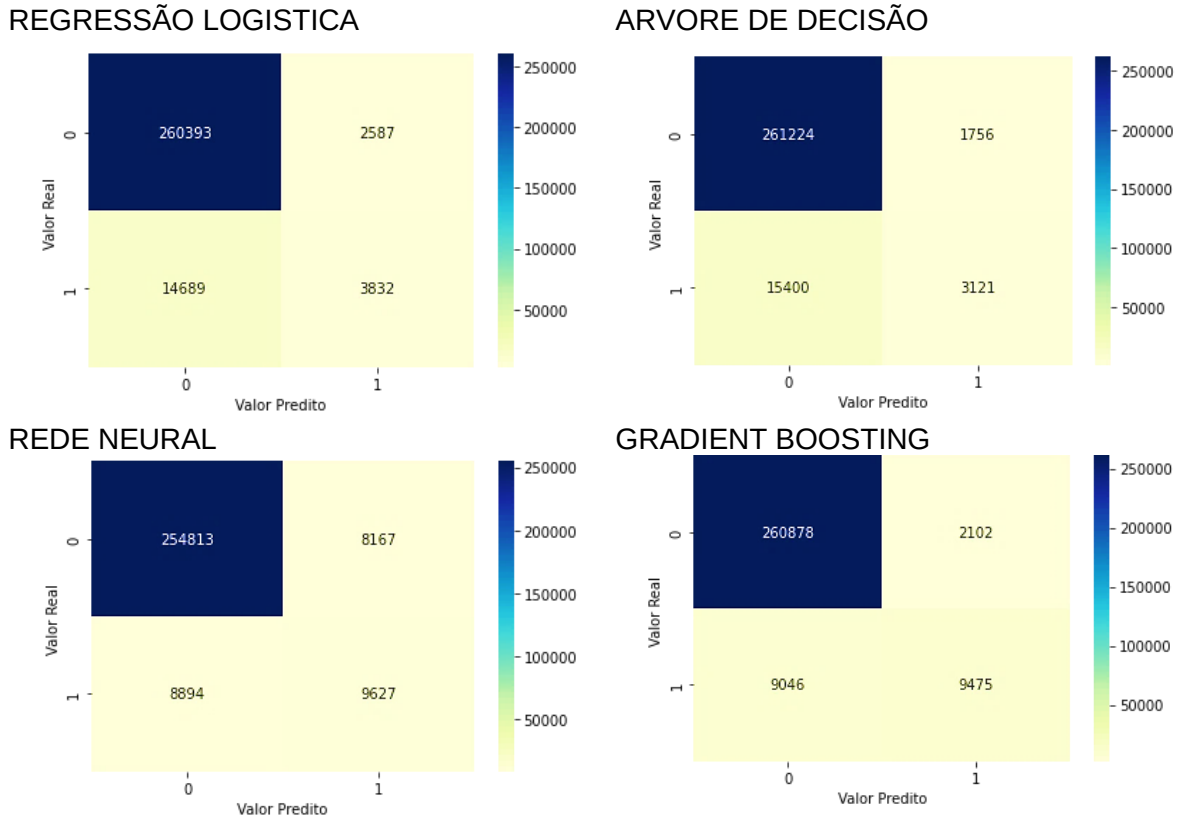
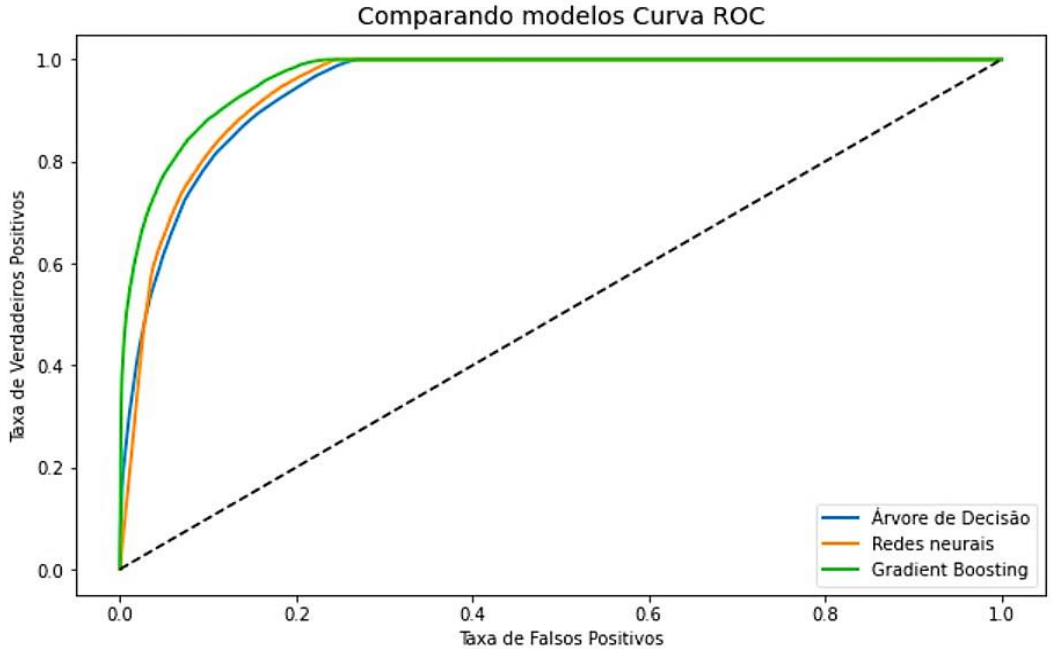


Figura 14. Matrizes de confusão dos modelos de validação cruzada
Fonte: Resultados originais da pesquisa

Uma outra métrica importante na comparação entre os modelos é a curva ROC. Nesse método, quanto mais próximas as linhas dos gráficos estiverem do valor de 1.0, melhor é o ajuste. Assim, quanto maior a curva ROC, melhor será o modelo.

Deste modo, após a comparação dos três modelos com melhor performance nas métricas anteriores, se verificou que o modelo de “Gradient Boosting” apresentou um melhor ajuste, quando em comparação aos demais modelos, conforme verificado na Figura 15.




```
Área sob a curva ROC - Modelo Árvore de Decisão: 94.39%
Área sob a curva ROC - Modelo Redes neurais: 94.64%
Área sob a curva ROC - Modelo Gradient Boosting: 94.64%
```

Figura 15. Gráfico de comparação dos modelos em Curva ROC
Fonte: Dados originais da pesquisa

Considerações Finais

O propósito deste estudo consistiu em identificar e analisar os fatores que influenciaram para as ocorrências de acidentes de trânsito nas rodovias e estradas federais brasileiras, entre os anos de 2017 e 2021.

De igual modo, procurou compreender melhor como as ferramentas e técnicas de aprendizado de máquina supervisionados podem contribuir e acelerar o emprego das tecnologias de Internet das Coisas no gerenciamento do trânsito no Brasil.

Para atingir tais objetivos foram realizadas análises de dados, por meio de técnicas e métodos, baseados em modelos de regressão logística multinomial, árvore de decisão, rede neural e “Gradient Boosting”.

Em relação ao desempenho dos modelos, o “Gradient Boosting” se apresentou como o mais eficaz, obtendo uma acurácia de 96% na classificação dos acidentes com vítimas fatais.

No contexto geral, entende-se que o estudo atingiu os propósitos inicialmente previstos, sendo considerados como satisfatórios os resultados e informações coletadas a partir das pesquisas e análises desenvolvidas, sobretudo em relação as técnicas de aprendizado de máquina que foram exploradas.

Ao longo do estudo foram extraídas diversas informações relevantes. Nisso, se observou que alguns estados apresentam rodovias mais perigosas que os demais. Também foi verificado que a falta de atenção à condução é o fator mais proeminente para o quantitativo de acidentes com vítimas fatais. Em sequência está o desrespeito aos limites de velocidade. Observou-se, ainda, que as colisões frontais e os atropelamentos de pedestres apresentam expressiva contribuição para a quantidade de vítimas fatais. Além disso, notou-se que há uma tendência no aumento do número de ocorrências aos finais de semana. Por vez, o período noturno é a fase do dia no qual ocorrem a maioria dos acidentes com vítimas fatais.

Diante disso, embora os resultados positivos obtidos pelo estudo, compreende-se que novas pesquisas se fazem necessárias. Assim, em virtude da importância do assunto, sugere-se que os estudos futuros explorem outras bases de dados. Também é aconselhável o experimento com os demais métodos de aprendizado de máquina, tais como as técnicas de modelos não supervisionados.

Em conclusão, diante dessas considerações, espera-se que o presente estudo possa contribuir para a projeção e desenvolvimento de recursos analíticos e tecnológicos relacionados à gestão do trânsito. Estima-se que o emprego de novas ferramentas de tecnologia, tal como a Internet das Coisas, ocasionará um avanço na administração da mobilidade. Permitindo, assim, uma maior agilidade e eficiência na tomada de decisões, visando reduzir o número de acidentes e aumentar a segurança viária no Brasil.

Agradecimento

Primeiramente, agradeço a Deus por esta oportunidade. Também, a minha esposa Ana Paula, minha filha Lívia, minha mãe Jô e meu irmão Rafael, por todo o apoio. Do mesmo modo, ao professor Mateus Modesto pela orientação e cooperação neste projeto.

Referências

Fávero, L.P.; Belfiore, P. 2017. Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. 1ed. Elsevier Editora, Rio de Janeiro, RJ, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788595155602>. Acesso em: 22 nov. 2021.

Géron, A. 2019. Mãos à Obra Aprendizado Máquina com Scikit-Learn e TensorFlow. 1ed. Alta Editora, Rio de Janeiro, RJ, Brasil.

Grus, J. 2021. Data Science do Zero. 2ed. ALTA BOOKS, Rio de Janeiro, RJ, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788550816463>. Acesso em: 22 nov. 2021.

Hitachi-UTokyo Laboratory [H-UTOKYO LAB]. 2020. Society 5.0: A people-centric super-smart society. 1ed. SpringerOpen, Bunkyo-ku, Tokyo, Japan.

Instituto de Pesquisa Econômica Aplicada [IPEA]. 2015. Relatório de pesquisa, estimativa dos custos dos acidentes de trânsito no Brasil com base na atualização simplificada das pesquisas anteriores do Ipea. Disponível em: <https://www.ipea.gov.br/portal/images/stories/PDFs/relatoriopesquisa/160516_relatorio_estimativas.pdf>. Acesso em: 20 nov. 2021.

Maschietto, L.G.; Vieira, A.L.N.; Torres, F. 2021. Arquitetura e infraestrutura de IoT. 1ed. SAGAH, Porto Alegre, RS, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9786556901947>. Acesso em: 22 nov. 2021.

Molin, S. 2019. Hands-On Data Analysis with Pandas. 1ed. Packt Publishing, Birmingham, UK, United Kingdom.

Okada, H.; Neves, A. 2019. Análise de algoritmos de indução de árvores de decisão. Escola Superior Batista da Amazônia, Manaus, AM, Brasil. Disponível em < https://doi.org/10.33448/rsd-v8i11.1473>. Acesso em: 18 ago. 2022.

Pinheiro, A.C.B.; Crivelaro, M. 2020. Edificações Inteligentes: Smart Buildings para Smart Cities.1ed. Editora Saraiva, São Paulo, SP, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788536532677>.Acesso em: 22 nov. 2021.

Polícia Rodoviária Federal. 2022. Acessos a informação de dados abertos de acidentes. Disponível em: < https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-acidentes>. Acesso em: 24 jan. 2022.

Rokach, L.; Maimon, O. 2014. Data mining with decision trees: theory and applications. 2ed. World Scientific, Danvers, MA, United States.

Silva, F.M.D.; Lenz, M.L.; Freitas, P.H.C.; Santos, S.C.B.D. 2019. Inteligência artificial. 1ed. SAGAH, Porto Alegre, RS, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788595029392>. Acesso em: 16 nov. 2021.

Zhang, Y.; Haghani, A. 2015. A gradient boosting method to improve travel time prediction. University of Maryland, DC, United States. Disponível em < https://doi.org/10.1016/j.trc.2015.02.019>. Acesso em: 17 ago. 2022.