Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.6274

Artigo de Pesquisa/Research Paper

Técnicas de Aprendizado de Máquina para Predição de Gravidade de Acidentes em Rodovias do Estado do Rio de Janeiro

Machine Learning Techniques for Predicting the Severity of Accidents on Highways in the State of Rio de Janeiro

Evandro V. Mafort [Centro Federal de Educação Tecnológica Celso Suckow da Fonseca | *evandro.mafort@aluno.cefet-rj.br*]
Marco A. A. Kappel [Centro Federal de Educação Tecnológica Celso Suckow da Fonseca |*marco.kappel@cefet-rj.br*]

Centro Federal de Educação Celso Suckow da Fonseca, Av. Gov. Roberto Silveira, 1900, Jd. Ouro Preto, Nova Friburgo, RJ, 28635-000, Brasil.

Resumo. Devido à sua dependência do transporte rodoviário, a segurança viária é um desafio significativo no Brasil. O estado do Rio de Janeiro apresenta padrões recorrentes de acidentes graves, demandando a identificação de seus principais fatores de risco. Este estudo aplicou técnicas de aprendizado de máquina para desenvolver um modelo preditivo para a gravidade de acidentes ocorridos nas rodovias federais do estado entre 2020 e 2023, utilizando dados da Polícia Rodoviária Federal (PRF). Foram comparados sete algoritmos de classificação supervisionada, com hiperparâmetros otimizados e avaliação por validação cruzada. O modelo de Regressão Logística apresentou o desempenho mais robusto, alcançando 73,85% de acurácia. A análise de interpretabilidade com a ferramenta LIME indicou que o tipo de acidente, a fase do dia, as condições meteorológicas, a localização geográfica e a densidade de acidentes por quilômetro foram os preditores mais influentes. Os resultados podem subsidiar a formulação de políticas públicas de prevenção mais eficazes.

Abstract. Due to its reliance on road transportation, road safety is a significant challenge in Brazil. The state of Rio de Janeiro exhibits recurring patterns of severe accidents, demanding the identification of its main risk factors. This study applied machine learning techniques to develop a predictive model for accident severity on the state's federal highways between 2020 and 2023, using data from the Federal Highway Police (PRF). Seven supervised classification algorithms were compared, with optimized hyperparameters and evaluation using cross-validation. The Logistic Regression model showed the most robust performance, achieving 73.85% accuracy. Interpretability analysis with the LIME tool indicated that the accident type, time of day, weather conditions, geographical location, and accident density per kilometer were the most influential predictors. The results can support the formulation of more effective public prevention policies.

Palavras-chave: Aprendizado de Máquina, Gravidade de Acidentes, Segurança Viária

Keywords: Machine Learning, Accident Severity, Road Safety

Recebido/Received: 29 July 2025 • **Aceito/Accepted:** 12 October 2025 • **Publicado/Published:** 31 October 2025

1 Introdução

A República Federativa do Brasil é o maior país da América do Sul em extensão territorial, ocupando 47,3% do continente e totalizando 8,5 milhões de km² [Brasil, 2022]. Desse total, cerca de 1,8 milhão de quilômetros são cobertos por estradas, sendo aproximadamente 75.800 km sob jurisdição federal, dos quais 65.400 km são pavimentados [Brasil, 2023]. No estado do Rio de Janeiro, cerca de 2.546 km de rodovias federais são geridos por órgãos federais ou parcerias público-privadas [DER-RJ, 2024].

Entre os principais desafios do sistema rodoviário está a alta incidência de acidentes de trânsito. Entre 2007 e 2024, foram registrados mais de 2,1 milhões de acidentes em rodovias federais, com 120.853 mortes [CNT, 2024]. No Rio de Janeiro, foram 179.392 ocorrências e 7.412 óbitos no mesmo período. Além do impacto humano, estima-se que os custos anuais com acidentes alcancem R$ 50 bilhões, abrangendo atendimento médico, reparos e outros prejuízos à infraestrutura [IPEA, 2020].

A redução desses números é dificultada pela complexidade dos fatores envolvidos, como clima, comportamento humano e condições estruturais das vias. Nesse contexto, técnicas de Aprendizado de Máquina vêm sendo adotadas para identificar padrões e realizar previsões a partir de grandes volumes de dados, com aplicações que incluem predição da gravidade, localização e frequência dos acidentes [Martins and De Andrade, 2021; Kraut and Sapia, 2022; Costa *et al.*, 2021].

Apesar do crescente interesse no tema, a literatura existente apresenta lacunas importantes que este trabalho busca endereçar. Primeiramente, observa-se uma carência de estudos focados especificamente nas rodovias federais do estado do Rio de Janeiro, cujas características de relevo, clima e fluxo de veículos diferem de outras regiões do país. Adicionalmente, muitos estudos nacionais, embora robustos, frequentemente se concentram em cenários mais amplos, como toda a Região Sudeste ou o território nacional, ou excluem variáveis contextuais relevantes, como a densidade de acidentes por trecho de via. Portanto, há uma oportunidade clara para desenvolver um modelo preditivo ajustado às particularidades locais, contribuindo com insights direcionados para a gestão da segurança

viária fluminense.

Este trabalho propõe o uso de algoritmos de Aprendizado de Máquina aplicados a dados públicos da Polícia Rodoviária Federal para identificar padrões associados a acidentes ocorridos nas rodovias federais do estado do Rio de Janeiro. São analisadas variáveis como densidade de acidentes por quilômetro, localização, sentido da via, clima, horário, tipo e causa do acidente. O objetivo é contribuir para estratégias de prevenção, políticas públicas baseadas em evidências e estudos futuros sobre predição da severidade de acidentes em diferentes regiões do Brasil.

A Seção 2 apresenta a revisão da literatura. A Seção 3 detalha os dados, o pré-processamento e os algoritmos utilizados. A Seção 4 discute os resultados obtidos, e a Seção 5 apresenta as implicações práticas e sugestões para pesquisas futuras.

2 Trabalhos relacionados

Apesar da disponibilidade de bases de dados abertas da PRF e de órgãos municipais e estaduais, ainda há escassez de estudos focados especificamente na predição da severidade dos acidentes em rodovias nacionais. No entanto, pesquisas correlatas, como predição da frequência de acidentes, identificação de padrões por regras de associação e análise de locais críticos, fornecem bases teóricas e metodológicas relevantes para esse tipo de investigação.

Entre os estudos nacionais, destaca-se [Amorim, 2019], que utilizou algoritmos supervisionados para prever a severidade de acidentes em rodovias federais da Região Sudeste, com dados da PRF entre 2007 e 2017. Foram testados quatro cenários com dados balanceados e desbalanceados, com e sem atributo que demarca a frequência de acidentes por quilômetro de estrada. O melhor desempenho foi alcançado pela Rede Neural MLP (85% de acurácia) e por combinações como *Random Forest* com *BernoulliNB*, embora não haja detalhamento sobre as variáveis mais influentes. Em linha semelhante, [Scholz and Pinheiro, 2023] analisaram acidentes em todo o território nacional entre 2017 e 2022, incorporando variáveis criadas a partir de dados da Fundação Instituto de Pesquisas Econômicas (FIPE) sobre o valor dos veículos envolvidos. O cenário mais eficaz envolveu dados sobre veículos e ambiente, com *Random Forest* atingindo 86% de acurácia e F1-score de 59%, sendo fatores como tipo e valor do veículo, ano de fabricação, marca e horário os mais relevantes para a predição, porém excluindo fatores ligados a outras características dos acidentes.

Outros estudos empregaram regras de associação para identificar padrões recorrentes em acidentes. [Kraut and Sapia, 2022] analisaram dados da cidade de São Paulo (2019–2021), obtendo um modelo preditivo com precisão de 96,04% na identificação de locais com alta concentração de acidentes. Já [Malaquias *et al*., 2021] focaram na previsão das causas dos acidentes em rodovias federais entre 2017 e 2019, atingindo 69% de acurácia e F1-score em seu melhor cenário, após a aplicação de algoritmos supervisionados em cinco diferentes configurações de dados. [Costa *et al*., 2014] também contribuíram ao aplicar técnicas de mineração de dados e aprendizado de máquina aos registros da PRF de 2012. Utilizando a ferramenta *Weka*, geraram 38 regras de associação com confiança superior a 0,8 e valores de AUC acima de 0,5 para todas as classes analisadas. Além do uso de técnicas computacionais, metodologias estatísticas também foram exploradas. [De Almeida *et al*., 2013], por exemplo, analisaram mais de 118 mil acidentes em Fortaleza entre 2004 e 2008, identificando maior letalidade em atropelamentos, acidentes ocorridos em rodovias federais e entre motoristas com pouca experiência de habilitação.

No cenário internacional, diversos estudos também abordaram a predição da severidade de acidentes. [Atwah and Al-Mousa, 2021] utilizaram *Random Forest* e *SVM* para analisar dados britânicos entre 2005 e 2014, atingindo mais de 80% de acurácia. [Balfaqih *et al*., 2021] investigaram o uso de sensores veiculares com modelos como *GMM*, *Naive Bayes*, *Decision Tree* e *CART*. [Hadjidimitriou *et al*., 2020] focaram em acidentes com motocicletas nos Estados Unidos (2010–2015), enquanto [Santos, 2021] aplicaram técnicas de classificação e agrupamento em acidentes ocorridos em Setúbal, Portugal, entre 2016 e 2019. Por fim, [Iranitalab and Khattak, 2017] desenvolveram modelos baseados em dados do estado de Nebraska, também nos EUA.

3 Metodologia

A metodologia adotada no presente estudo foi estruturada para garantir a organização, o processamento e a análise eficiente dos dados sobre acidentes de trânsito nas rodovias federais brasileiras. Inicialmente, foram selecionadas e preparadas as bases de dados da Polícia Rodoviária Federal (PRF), abrangendo informações detalhadas sobre os acidentes registrados. Em seguida, realizou-se o pré-processamento dos dados, que incluiu a limpeza, a normalização e a correção de possíveis desbalanceamentos, garantindo a integridade e a qualidade dos dados utilizados. A partir desse conjunto tratado, foram aplicados algoritmos de aprendizado de máquina, otimizados por meio de ajuste de hiperparâmetros e validados por técnicas de validação cruzada. Por fim, os resultados foram avaliados e comparados com estudos anteriores para uma análise quantitativa das variáveis. Na Figura 1 é apresentado o fluxograma detalhado da metodologia empregada, destacando as principais etapas do estudo.

Desde 2007, a PRF mantém uma base de dados com registros de acidentes em cerca de 70.000 km de rodovias federais sob sua jurisdição, abrangendo dados até janeiro de 2025. Até 2016, os dados eram coletados via sistema BR-Brasil, substituído a partir de 2017 pelo Sistema BAT (Boletim de Acidentes de Trânsito), que passou a organizar as informações de forma mais estruturada, favorecendo a engenharia de variáveis. As bases de dados são divididas em duas categorias: por ocorrência e por pessoa. Neste estudo, optou-se pelo uso da base por ocorrência, por fornecer uma visão completa do acidente sem necessidade de agregação de dados.

Dado o grande volume de registros históricos, optou-se por trabalhar com o recorte de 2020 a 2023, totalizando 260.524 registros. As bases anuais foram concatenadas e passaram por um tratamento inicial para a remoção de nulos. Para garantir uma metodologia robusta, o conjunto foi dividido em 75% para treino e 25% para teste, sendo este último isolado. A otimização de hiperparâmetros foi realizada no

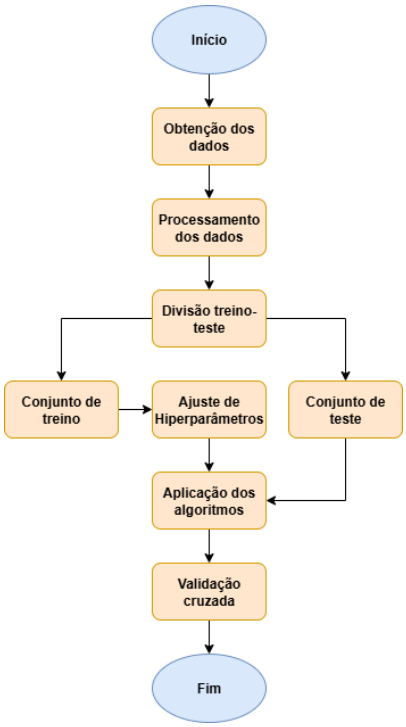


Figura 1. Fluxograma da metodologia utilizada no estudo.

conjunto de treino por meio de validação cruzada, onde um pipeline de pré-processamento, envolvendo codificação de variáveis categóricas, normalização e subamostragem aleatória, foi aplicado internamente em cada partição. A avaliação de desempenho final foi conduzida em duas frentes: Uma avaliação da capacidade de generalização no conjunto de teste isolado e uma segunda avaliação de estabilidade, por meio de uma nova rotina de validação cruzada aplicada apenas ao conjunto de treino. Por fim, foi feita a análise das variáveis mais importantes e as suas importâncias para a predição das ocorrências de trânsito.

Todo o pipeline foi implementado em Python (versão 3.11.9), utilizando o Jupyter Notebook (v7.0.6) e bibliotecas como NumPy[1], Pandas[2], Matplotlib[3], Seaborn[4] que foram utilizadas para tratamento e visualização dos dados. Para as tarefas de aprendizado de máquina, a biblioteca scikit-learn[5]. Já a busca por hiperparâmetros foi conduzida com Skopt[6] e Optuna[7]. A ferramenta de explicabilidade de modelos utilizada no estudo foi a LIME[8]

3.1 Obtenção e preparação dos dados

A base de dados foi obtida na seção da Polícia Rodoviária Federal, sediada no site do Governo Brasileiro[9]. As bases de dados encontram-se no formato .csv (Comma Separated Values, ou Valores Separados por Vírgula) e são identificadas por 'datatran_ANO', onde, no caso, 'ANO' vai de 2020 a 2023 e corresponde a registros ocorridos entre 1 de janeiro a 31 de fevereiro de cada ano. Cada um dos arquivos csv tem em média 18 megabytes de tamanho e contém informações sobre acidentes de trânsito em rodovias federais por todo o país. Junto com o arquivo csv, vem um arquivo pdf que serve como 'dicionário' informando o nome e a descrição da variável. A estrutura dos arquivos selecionados para o experimento é a mesma, contendo cerca de 30 colunas que são descritas na Tabela 1.

Tabela 1. Descrição das variáveis do conjunto de dados.

Nome	Descrição
Id	Identificador único do acidente.
data_inversa	Data da ocorrência.
dia_semana	Dia da semana do acidente.
horario	Horário da ocorrência.
uf	Unidade federativa.
br	Número da rodovia federal.
km	Quilômetro onde ocorreu o acidente.
municipio	Município da ocorrência.
causa_acidente	Causa principal do acidente.
tipo_acidente	Tipo de acidente.
classificacao_acidente	Gravidade do acidente.
fase_dia	Fase do dia.
sentido_via	Sentido da via no local do acidente.
condicao_meteorologica	Clima no momento do acidente.
tipo_pista	Tipo de pista.
tracado_via	Formato da via.
uso_solo	Zona urbana ou rural.
pessoas	Pessoas envolvidas no acidente.
mortos	Pessoas mortas na ocorrência.
feridos_leves	Feridos com lesões leves.
feridos_graves	Feridos com lesões graves.
ilesos	Pessoas sem ferimentos.
ignorados	Pessoas com dados ausentes.
feridos	Soma de feridos leves e graves.
veiculos	Veículos envolvidos.
latitude	Latitude do local.
longitude	Longitude do local.
regional	Superintendência da PRF responsável.
delegacia	Delegacia da PRF responsável.
uop	Unidade operacional da PRF.

As bases de dados, divididas por ano, foram concatenadas verticalmente em uma única tabela para centralizar a aplicação dos códigos. Em seguida, foram extraídos os dados de acidentes do Estado do Rio de Janeiro, reduzindo a base para 19.090 linhas (cerca de 7% do total original). Também foram removidas 266 entradas sem informações relevantes, como registros com condição meteorológica 'ignorado' ou BR igual a 0. Por fim, colunas não essenciais, como *delegacia*, *uop* e *regional*, foram eliminadas. A variável *causa_acidente* também foi removida, pois sua determinação ocorre apenas após o acidente, não representando uma característica disponível previamente para previsão. Seu uso poderia levar

[1] https://numpy.org
[2] https://pandas.pydata.org
[3] https://matplotlib.org
[4] https://seaborn.pydata.org
[5] scikit-learn.org
[6] https://pypi.org/project/scikit-learn-intelex
[7] https://optuna.org
[8] https://christophm.github.io/interpretable-ml-book/lime.html
[9] https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf

a um problema de vazamento de dados, comprometendo a confiabilidade do modelo preditivo [Amorim, 2019; Scholz and Pinheiro, 2023]. Com o intuito de tornar mais evidente a distinção entre as variáveis originais e aquelas derivadas por meio de transformações ou combinações, optou-se por adotar a nomenclatura no padrão *Camel Case* exclusivamente para as variáveis derivadas.

A variável escolhida como alvo (ou target), denominada *gravidade*, foi criada a partir das colunas *ilesos*, *feridos_leves*, *feridos_graves* e *mortos*. Seus valores foram definidos como 'grave' e 'não-grave'. Como os dados brutos apenas informam a quantidade de ilesos, feridos e mortos, mas não expressam diretamente a severidade do acidente, foi necessário elaborar a variável *gravidade* [Amorim, 2019]. Um acidente é classificado como 'grave' se houver pelo menos um morto ou um ferido grave; caso contrário, ao apresentar apenas feridos leves ou ilesos, será considerado 'não-grave'. Ao final do cálculo, essas colunas foram removidas, pois poderiam causar *target leakage* (ou vazamento para o alvo), já que fornecem informações sobre o estado futuro do acidente antes mesmo da previsão [Larsen and Becker, 2021].

Outro fator considerado foi a frequência de acidentes por quilômetro em uma estrada. Para isso, foi criada a variável *frequenciaAcidente*, que indica a quantidade média de acidentes por quilômetro em uma determinada rodovia [Amorim, 2019]. Essa métrica busca identificar padrões de concentração de acidentes em determinados trechos, permitindo ao modelo capturar contextos de maior risco e, assim, melhorar a previsão da gravidade dos acidentes. Além disso, foi criada a variável *ehFeriado*, contendo os valores 'sim' ou 'não', para mostrar se um acidente ocorreu em um fim de semana. A coluna foi criada com base nas tabelas de datas fornecidas pela AMBIMA (Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais) que demarcam feriados bancários de 2000 a 2099. Paralelamente, foi criado um atributo derivado de *dia_semana*, chamado *fimDeSemana*, utilizando valores binários para sinalizar se o dia faz parte de um fim de semana ou não, sendo 0 para não e 1 para sim.

Inicialmente, as informações relativas a data e horário dos acidentes estavam no formato *string*, exigindo a decomposição das variáveis *data_inversa* e *horario* em *dia*, *mes*, *ano* e *hora*. Além disso, como os algoritmos de previsão não reconhecem a natureza cíclica dessas variáveis, poderiam interpretá-las de forma equivocada. Para corrigir isso, aplicou-se a transformação seno-cosseno, que converte os valores em representações numéricas contínuas, preservando a periodicidade dos dados e permitindo que o modelo capture melhor os padrões temporais [Scholz and Pinheiro, 2023]. Por fim, a coluna *data_inversa* foi removida do conjunto de dados por atingir o seu objetivo.

Seguindo com a etapa de tratamento dos dados, foi necessário padronizar os valores da coluna *tracado_via*. Essa variável contém informações sobre o local do acidente, porém de maneira que não permite compreender a importância de um fator sobre o outro. Alguns dos 174 rótulos presentes na variável incluem: 'interseção de vias; retorno regulamentado', 'viaduto; interseção de vias; curva; aclive' e 'rotatória; reta'. Para extrair o máximo de informação, os rótulos foram desmembrados e reagrupados, resultando na criação de cinco novos atributos derivados. Os atributos derivados criados e os seus respectivos rótulos encontram-se na Tabela 2. O desmembramento das *strings* contidas na variável original em novas variáveis pode ajudar no entendimento de informações sobre o local do acidente.

Tabela 2. Atributos derivados criados e seus respectivos rótulos

Atributo derivado	Rótulos do atributo
condicaoPista	'Em obras', 'Normal'
tipoInclinacao	'Plano', 'Declive', 'Aclive'
tipoSuperficie	'Reta', 'Curva'
tipoManobra	'Nenhum', 'Interseção', 'Rotatória', 'Retorno Regulamentado'
tipoEstrutura	'Nenhum', 'Viaduto', 'Ponte', 'Túnel'

Após o tratamento dos dados, a base final passou a contar com 18.104 linhas e 25 colunas, representando uma redução de aproximadamente 6% no número de linhas e 16% no número de colunas em relação ao início do processo. Por fim, a variável alvo foi isolada do restante da base para preparar os dados para modelagem, assegurando um treinamento adequado do modelo e prevenindo o vazamento de informações durante o aprendizado.

3.2 Pre-processamento dos dados

Nesta etapa, os dados foram tratados de forma que os algoritmos obtivessem o melhor resultado possível. Para isso, primeiramente dividimos o conjunto de dados em dados de treino e teste e, a partir disso, três técnicas foram utilizadas: A transformação de valores não numéricos para numéricos, a padronização dos dados e a subamostragem aleatória.

Codificação de variáveis categóricas. A conversão de dados categóricos em numéricos é fundamental, pois muitos modelos de aprendizado de máquina trabalham com operações matemáticas, como cálculo de probabilidades e distâncias, o que impede o uso direto de variáveis qualitativas. Para isso, cada rótulo categórico foi substituído pelo valor médio da variável alvo correspondente, a qual, para permitir este cálculo, foi representada numericamente ('grave' como 1 e 'não grave' como 0). Essa abordagem preserva a relação entre os atributos categóricos e o resultado esperado, ao mesmo tempo que evita a introdução de uma ordem artificial entre as categorias, que poderia gerar vieses no modelo. [Scholz and Pinheiro, 2023].

Padronização dos dados. Ao lidar com múltiplas variáveis que representam diferentes características, é comum que estejam em escalas distintas, o que pode prejudicar algoritmos de aprendizado de máquina baseados em distância. Para resolver isso, aplicou-se uma padronização dos dados, de forma que todas as variáveis passassem a ter média zero e desvio padrão igual a um. A Equação 1 expressa esse processo, no qual uma variável x é transformada em z por meio da subtração da média $\mu(X)$ e divisão pelo desvio padrão $S(X)$. Essa normalização torna as variáveis comparáveis entre si e melhora o desempenho de modelos sensíveis à escala dos dados.

$$z = \frac{x - \mu(X)}{S(X)} \tag{1}$$

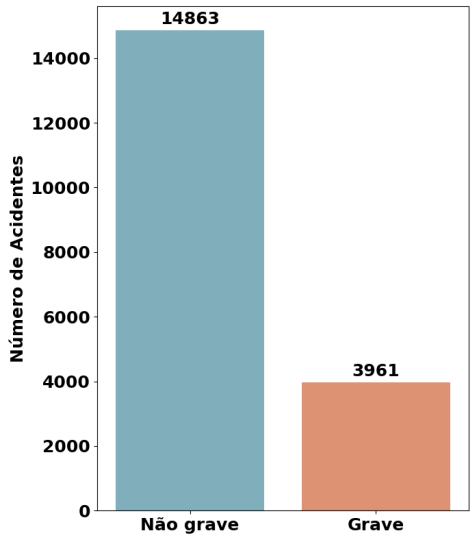


Figura 2. Proporção das classes 'grave' e 'não-grave' antes da subamostragem.

Transformação de coordenadas geográficas. As variáveis de localização (*latitude* e *longitude*) foram inicialmente padronizadas e, em seguida, reduzidas utilizando a Análise de Componentes Principais (PCA). Esse procedimento permitiu capturar a variabilidade espacial em duas dimensões principais, gerando a nova variável *pcLatLon*, que substituiu as coordenadas originais conseguindo reter 99,5% das informações contidas nas duas colunas. Essa transformação auxiliou na preservação das informações espaciais relevantes, ao mesmo tempo que simplificou a representação geográfica dos dados [Demšar *et al*., 2013].

Subamostragem aleatória. A subamostragem é uma estratégia utilizada para corrigir o desequilíbrio entre classes em conjuntos de dados, reduzindo aleatoriamente o número de amostras da classe mais representada até igualá-lo ao da classe menos frequente. Esse procedimento é essencial em tarefas de Aprendizado de Máquina, pois o excesso de amostras em uma classe pode induzir o modelo a negligenciar a classe minoritária. Neste estudo, observou-se que 79% dos registros referem-se a acidentes não graves, o que motivou a aplicação da subamostragem para nivelar a distribuição das classes, como mostra a Figura 3. Como resultado, ambas as classes passaram a contar com 3.791 exemplos.

3.3 Técnicas de Aprendizado de Máquina

Neste trabalho, foram avaliadas sete técnicas de Aprendizado de Máquina amplamente utilizadas em tarefas de classificação binária, com o objetivo de identificar os fatores mais relevantes para determinar se um acidente é grave ou não. Para isso, diferentes combinações de hiperparâmetros foram testadas para cada modelo, a fim de mitigar o sobreajuste e melhorar o desempenho.

Regressão Logística. Apesar do nome, trata-se de um algoritmo de classificação binária que estima a probabilidade de um evento ocorrer por meio da função sigmoide, resultando em valores entre 0 e 1. É simples, eficiente e robusto mesmo com variáveis que não seguem distribuição normal [Géron, 2021; Minussi *et al*., 2002].

Naive Bayes. Baseado no teorema de Bayes, é um classificador probabilístico que assume independência entre as variáveis preditoras. Apesar dessa suposição raramente se confirmar na prática, o algoritmo costuma apresentar bom desempenho mesmo com poucos dados [Harrison, 2019; Kamel *et al*., 2019].

K-Vizinhos Próximos (kNN). Classifica uma nova amostra com base na "distância" para seus vizinhos mais próximos. É um método de aprendizado por instância e não paramétrico, simples de implementar, mas sensível ao número de variáveis devido à maldição da dimensionalidade [Harrison, 2019; Grus, 2021].

Árvore de Decisão. Constrói modelos em formato de árvore, dividindo os dados com base em critérios como entropia ou índice de Gini. É interpretável, lida bem com variáveis categóricas e numéricas, e exige pouco pré-processamento [Grus, 2021; Géron, 2021].

Floresta Aleatória. Também conhecido como *Random Forest*, é um conjunto de Árvores de Decisão combinadas por meio do método de *bagging*, o que reduz o risco de sobreajuste. Oferece maior robustez, mantendo a interpretabilidade e adaptabilidade a diferentes tipos de dados [Harrison, 2019; Grus, 2021].

Máquina de Vetores de Suporte (SVM). Busca um hiperplano que melhor separa as classes no espaço de características, maximizando a margem entre elas. Embora tenha custo computacional elevado, apresenta excelente desempenho em diversos problemas de classificação [Magalhães *et al*., 2022].

Perceptron Multicamadas (MLP). Por fim, o Perceptron Multicamadas (MLP) é um tipo de rede neural artificial que emprega múltiplas camadas, incluindo uma de entrada, camadas ocultas e uma de saída. O algoritmo utiliza funções de ativação e ajusta seus pesos por meio do algoritmo de retro-propagação [Grus, 2021].

3.4 Ajuste de Hiperparâmetros.

A otimização bayesiana foi empregada no ajuste de hiperparâmetros. O algoritmo opera de forma iterativa, utilizando dois componentes principais para guiar a busca: um modelo substituto probabilístico e uma função de aquisição. O modelo substituto, tipicamente um Processo Gaussiano, é continuamente ajustado com as observações de desempenho de cada configuração testada. Ele estima a superfície de desempenho dos hiperparâmetros, prevendo o resultado de combinações ainda não exploradas. A função de aquisição, por sua vez, utiliza essas previsões para encontrar equilíbrio enquanto investiga áreas de alta incerteza e também foca em áreas com alta probabilidade de sucesso. Este processo de aprendizado contínuo permite que a busca seja direcionada de forma mais inteligente, convergindo para uma solução ótima com um número menor de iterações [Bischl *et al*., 2023; Cardoso, 2023]. Para o presente estudo, foram definidas 50 iterações para a busca bayesiana, visando equilibrar velocidade de execução e abrangência na exploração do espaço de parâmetros. Os únicos hiperparâmetros não submetidos à otimização foram os de estado aleatório, para garantir a reprodutibilidade, e, para o SVM, a configuração que permite a predição de proba-

bilidades, essencial para o cálculo da métrica AUC-ROC.

3.5 Validação Cruzada

A fim de garantir a confiabilidade dos resultados, a validação cruzada foi um componente central da metodologia, utilizada em duas etapas distintas. A primeira foi na otimização de hiperparâmetros, onde se aplicou a busca bayesiana combinada com validação cruzada aninhada (com 3 *folds*) sobre o conjunto de treino. Essa abordagem robusta minimiza a influência de variações aleatórias na escolha da melhor configuração para os modelos.

A segunda aplicação da validação cruzada foi como uma medida de estabilidade do desempenho. Após a definição dos hiperparâmetros, uma nova rotina de validação cruzada, com 5 *folds*, foi executada, também restrita ao conjunto de treino, para gerar uma estimativa da performance média e sua variabilidade. Em ambas as aplicações, a técnica baseou-se no método *k-Fold* estratificado, e o pré-processamento e o balanceamento foram aplicados isoladamente aos subconjuntos de treinamento em cada iteração, garantindo uma estimativa não viciada das métricas.

3.6 Análise da importância das variáveis

Além da avaliação de desempenho, buscou-se compreender os fatores que influenciam as previsões do modelo. Para essa finalidade, foi empregada a técnica LIME (Local Interpretable Model-agnostic Explanations), um método que explica predições individuais de qualquer classificador de forma agnóstica. O LIME opera gerando perturbações em torno de uma instância específica e treinando um modelo local, mais simples e interpretável, para aproximar o comportamento do modelo complexo naquela vizinhança. Os pesos desse modelo local indicam quais variáveis mais contribuíram para a predição original [Ribeiro *et al*., 2016].

Neste estudo, o LIME foi aplicado para analisar as previsões do modelo com melhor desempenho, permitindo a identificação dos principais preditores em nível de instância. A abordagem em questão faz com que se tenha, ao invés de apenas as métricas puras de performance, uma visão geral dos elementos que mais agregam a predição para ambas as classes, permitindo se ter uma visão prática mais clara para o entendimento do problema do presente estudo.

4 Resultados e Discussão

Esta seção avalia o desempenho de diferentes algoritmos de classificação na predição da gravidade de acidentes de trânsito, identificando o modelo mais eficaz e suas características mais influentes.

Na Tabela 3 são exibidas as configurações finais dos hiperparâmetros selecionados, revelando estratégias de otimização específicas para cada modelo. Por exemplo, a Regressão Logística adotou regularização L1 com um valor baixo de $C \approx 0{,}01$ e tolerância estrita. Embora pertençam à mesma família, Árvore de Decisão e Floresta Aleatória utilizaram critérios de divisão distintos, Gini e entropia, respectivamente, e diferiram em estrutura: a primeira com profundidade mais rasa e limiares mais altos para folhas.

O parâmetro *var_smoothing* do Naive Bayes, moderadamente baixo, reflete novamente sua sensibilidade a dados padronizados. O KNN selecionou 18 vizinhos com ponderação por distância e utilizou o algoritmo *ball_tree* com $p = 1$, adaptando seu desempenho à distribuição dos dados. Já o MLP exigiu mais de 5.000 iterações e empregou duas camadas ocultas com 100 neurônios cada e ativação logística.

A Figura 3 resume as principais métricas de classificação, incluindo acurácia, AUC-ROC, precisão, *recall* e F1-Score, reportadas separadamente para cada classe. As barras de erro indicam a variabilidade, evidenciando os trade-offs entre estabilidade e desempenho dos modelos. Embora nenhum classificador tenha superado os demais em todas as métricas, a Regressão Logística, o Perceptron Multicamadas e o Floresta Aleatória apresentaram, de forma consistente, os maiores valores de AUC-ROC (69,77%, 69,68% e 69,05%, respectivamente), conforme ilustrado na Figura 5. Apesar de uma AUC-ROC inferior, o *SVM* demonstrou o melhor desempenho em termos de *recall* para a classe grave (70,36%), mas ao custo da menor precisão (27,27%) e da mais alta variabilidade. Destaca-se ainda que o *Naïve Bayes* obteve alto *recall* para a classe não grave (77,99%), mas com desempenho insatisfatório na classe grave (*recall* = 45,42%). A regressão logística destacou-se por equilibrar a maior acurácia (73,85%), a melhor AUC-ROC (69,77%) e o maior F1-Macro (62,75%). Além disso, seu baixo desvio padrão em todas as métricas e a estrutura interpretável posicionam esse modelo como o "vencedor comparativo" do estudo.

Uma vez que a explicabilidade do LIME se restringe a instâncias individuais, torna-se necessário analisar múltiplas previsões para se obter uma visão geral de seu comportamento. Das 100 predições aleatórias do conjunto de teste explicadas pelo LIME, muitas tiveram uma probabilidade acima de 80% para acidentes não graves. Assim, para uma análise mais detalhada dos fatores de decisão do modelo, escolheu-se a predição com as probabilidades mais equilibradas, apresentada na Figura 4: 40% de chance de ser um acidente não grave e 60% de ser um acidente grave.

A Figura 4(a) detalha as evidências que levaram a essa decisão, mostrando o peso de cada variável na predição. As características mais relevantes para prever um desfecho grave foram, em ordem de importância, o *tipo_acidente*, que apresentou um peso de 0.25, a *fase_do_dia* e a *condicao_metereologica*. Para justificar essa influência, a Figura 4(b) exibe os valores reais e padronizados desta instância, destacando o valor extremamente alto de 2.42 para a *condicao_metereologica* e o valor positivo de 1.09 para a *fase_do_dia*. Por outro lado, a análise na Figura 4(a) revela que, para este caso específico, o modelo não identificou nenhuma variável com forte influência contrária que favorecesse um desfecho não grave, o que, por sua ausência de contrapontos, reforça a classificação final para a classe de maior risco.

5 Conclusão

O presente trabalho teve como objetivo comparar diferentes técnicas de aprendizado de máquina, analisando seus pontos fortes e limitações, com vistas ao desenvolvimento de um modelo preditivo capaz de estimar a gravidade de acidentes de trânsito nas rodovias federais do Estado do Rio de Janeiro. Independentemente do grau de severidade, acidentes de trânsito representam um problema relevante, com impactos na

Tabela 3. Hiperparâmetros finais encontrados para os algoritmos utilizados no estudo.

Algoritmo	Hiperparâmetros selecionados
Regressão Logística	$C = 0{,}0101$, penalty = 'l1', solver = 'saga', fit_intercept = True, max_iter = 50, tol = 1e-6
Árvore de Decisão	criterion = 'gini', max_depth = 86, min_samples_split = 30, min_samples_leaf = 50, ccp_alpha = 1e-6
Floresta Aleatória	n_estimators = 49, criterion = 'entropy', max_depth = 43, min_samples_split = 4, min_samples_leaf = 27, max_features = 0.5, bootstrap = False, ccp_alpha = 5.72e-5
Naïve Bayes	var_smoothing = 0.00156
K-Vizinhos Próximos	n_neighbors = 18, algorithm = 'ball_tree', metric = 'minkowski', $p = 1$, weights = 'distance', leaf_size = 74
Máquina de Vetores de Suporte	$C = 0{,}58$, kernel = 'rbf', gamma = 'auto', max_iter = 2000, tol = 7e-5
Perceptron Multicamadas	activation = 'logistic', solver = 'adam', alpha = 0.00352, learning_rate = 'constant', learning_rate_init = 0.0091, max_iter = 5075, tol = 0.0037, hidden_layer_sizes = (100, 100)

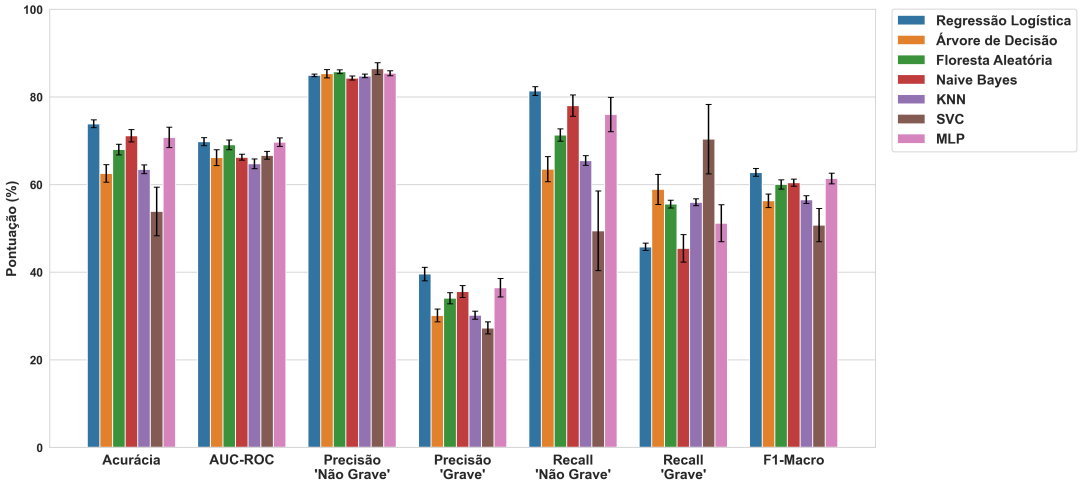
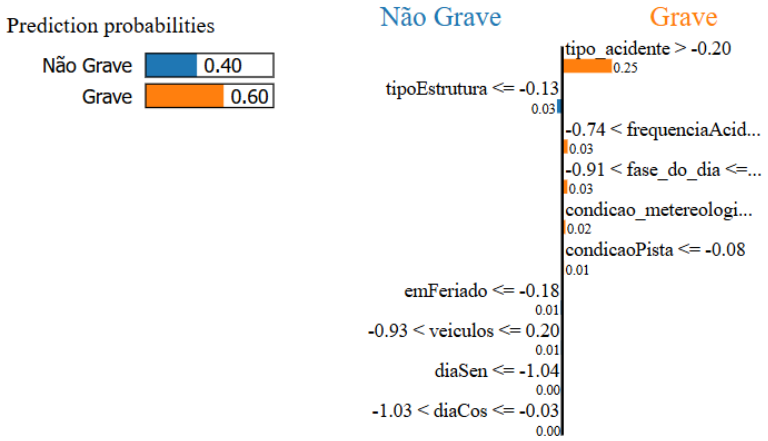


Figura 3. Resultados finais obtidos pela validação cruzada de cada algoritmo com 5 folds



(a) Probabilidades de predição e regras locais.

Feature	Value
tipo_acidente	0.46
tipoEstrutura	-0.13
frequenciaAcidente	-0.55
fase_do_dia	1.09
condicao_metereologica	2.42
condicaoPista	-0.08
emFeriado	-0.18
veiculos	0.20
diaSen	-1.38
diaCos	-0.31

(b) Impacto das features na predição da instância em questão.

Figura 4. Explicabilidade da predição para uma instância específica via LIME.

saúde pública, nos custos governamentais e nas estruturas viárias, além de consequências sociais profundas.

Após etapas de filtragem, enriquecimento de variáveis e pré- processamento, os dados foram balanceados e os mo-

delos treinados com validação cruzada. Os hiperparâmetros foram otimizados via busca bayesiana. Diferentes classificadores foram avaliados quanto ao desempenho, e a importância das variáveis foi analisada com base no modelo mais eficaz. Dentre os sete algoritmos avaliados, o modelo de Regressão Logística foi identificado como o de melhor desempenho comparativo, alcançando uma acurácia de 73,85% e AUC-ROC de 69,77%. A performance do modelo no contexto foi notada não apenas pelas métricas de classificação, mas também por sua estabilidade, evidenciada pelo baixo desvio padrão durante a validação cruzada.

A principal contribuição deste estudo reside na resposta a lacunas previamente identificadas na literatura. Ao focar especificamente no estado do Rio de Janeiro, foi possível construir um modelo preditivo ajustado às particularidades locais, algo pouco explorado em trabalhos anteriores. A análise de interpretabilidade, realizada com a ferramenta LIME, indicou que variáveis como o tipo de acidente, a fase do dia em que ocorre e as condições meteorológicas são preditores cruciais da gravidade. Adicionalmente, a relevância de fatores contextuais, como a localização geográfica e a densidade de acidentes por quilômetro, foi confirmada, preenchendo uma segunda lacuna de estudos que frequentemente desconsideram tais atributos.

Espera-se que estes achados possam subsidiar a formulação de políticas públicas de segurança viária mais eficazes. A identificação de fatores de risco específicos permite o aprimoramento de estratégias de prevenção, como a intensificação da fiscalização em horários ou trechos de maior perigo e a implementação de melhorias na infraestrutura. Embora o foco tenha sido o Rio de Janeiro, os resultados podem servir de referência para investigações em outras regiões do país.

Para pesquisas futuras, recomenda-se a incorporação de dados sobre os envolvidos nos acidentes, bem como informações mais detalhadas sobre as condições das rodovias. A análise conjunta de acidentes em rodovias federais e estaduais pode, igualmente, ampliar as perspectivas e a capacidade de generalização dos modelos.

Declarações complementares

Contribuições dos autores

EVM contribuiu para a concepção, execução do estudo, análise dos resultados e redação do manuscrito. MAAK atuou como orientador do trabalho, auxiliando na definição da metodologia, discussão dos resultados e revisão crítica do texto. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Visando a transparência e a reprodutibilidade dos resultados, o código-fonte desenvolvido para este trabalho encontra-se publicamente disponível em https://github.com/evandrovieiramafort/artigoRioML

Referências

Amorim, B. d. S. P. (2019). Uso de aprendizado de máquina para classificação de risco de acidentes em rodovias. Master's thesis, Universidade Federal de Campina Grande, Campina Grande. Dissertação de Mestrado. Disponível em: `http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/7340`.

Atwah, A. and Al-Mousa, A. (2021). Car accident severity classification using machine learning. In *International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 186–192, online. DOI: 10.1109/3ICT53449.2021.9581646.

Balfaqih, M. *et al*. (2021). An accident detection and classification system using internet of things and machine learning towards smart city. *Sustainability*, 14(1):1–13. DOI: 10.3390/su14010210.

Bischl, B. *et al*. (2023). Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2):1–43. DOI: 10.1002/widm.1484.

Brasil (2022). Geografia. disponível em: `https://www.gov.br/mre/pt-br/embaixada-bogota/o-brasil/geografia`. Acesso em: 20 Janeiro 2025.

Brasil (2023). Rodovias federais. disponível em: `https://www.gov.br/transportes/pt-br/assuntos/conteudo/rodovias-brasileiras`. Acesso em: 20 Janeiro 2025.

Bruce, P. and Bruce, A. (2017). *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly, Sebastopol.

Cardoso, S. P. C. (2023). Optimizing process mining algorithms: A hyperparameter tuning approach. Master's thesis, Universidade do Porto, Porto.

CNT (2024). Painel cnt de acidentes rodoviários. disponível em: `https://www.cnt.org.br/painel-acidente`. Acesso em: 20 Janeiro 2025.

Costa, A. D. M., De Freitas, A. G. O., and Pinheiro, R. P. (2021). Mineração de dados na construção de modelo de predição de acidentes com vítimas em recife. *Revista de Engenharia e Pesquisa Aplicada*, 6(3):70–80. DOI: 10.25286/repa.v6i3.1707.

Costa, J. D. J., Bernardini, F. C., and Viterbo Filho, J. (2014). A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. *AtoZ: novas práticas em informação e conhecimento*, pages 139–157. DOI: 10.5380/atoz.v3i2.41346.

De Almeida, R. L. F. *et al*. (2013). Via, homem e veículo: fatores de risco associados à gravidade dos acidentes de trânsito. *Revista Saúde Pública*, 47(4):718–732. DOI: 10.1590/S0034-8910.2013047003657.

Demšar, U., Harris, P., Brunsdon, S., Fotheringham, A. S., and McLoone, S. (2013). Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers*, 103(1):106–128. DOI: 10.1080/00045608.2012.689236.

DER-RJ (2024). Mapa rodoviário. Secretaria de Estado de Infraestrutura e Obras Públicas, Departamento de Estradas de Rodagem. Escala 1:450.000. `https://www.der.rj.gov.br/documentos/mapas/Mapa%20do%20Rio%20de%20Janeiro.pdf`.

Grus, J. (2021). *Data Science do Zero - Noções Fundamentais com Python*. Alta Books, Rio de Janeiro, 2ª edition.

Géron, A. (2021). *Mãos à Obra - Aprendizado de Máquina com Scikit-Learn, Keras e TensorFlow*. Alta Books, Rio de Janeiro, 2ª edition.

Hadjidimitriou, N. S. *et al*. (2020). Machine learning for severity classification of accidents involving powered two wheelers. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4308–4317. DOI: 10.1109/TITS.2019.2939624.

Harrison, M. (2019). *Machine Learning - Guia de Referência Rápida: Trabalhando com dados estruturados em Python*, volume 1. Novatec, Rio de Janeiro.

IPEA (2020). Custos dos acidentes de trânsito no brasil: Estimativa simplificada com base na atualização das pesquisas do ipea sobre custos de acidentes nos aglomerados urbanos e rodovias. Relatório Técnico 1415-4765, IPEA, Brasília. `http://repositorio.ipea.gov.br/handle/11058/10075`.

Iranitalab, A. and Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, 108:27–36. DOI: 10.1016/j.aap.2017.08.008.

Kamel, H., Abdulah, D., and M. Al-Tuwaijari, J. (2019). Cancer classification using gaussian naive bayes. In *International Engineering Conference (IEC)*, pages 165–170, Erbil. DOI: 10.1109/IEC47844.2019.8950650.

Kraut, C. and Sapia, H. M. (2022). Aprendizado de máquina utilizando agrupamento e regressão na previsão de locais de acidentes de trânsito em zonas urbanas. *Colloquium Exactarum*, 14:1–11. DOI: 10.5747/ce.2022.v14.n1.e380.

Larsen, K. R. and Becker, D. S. (2021). *Automated Machine Learning for Business*. Oxford University Press, Oxford, 1ª edition.

Li, D. and Kanoulas, E. (2018). Bayesian optimization for optimizing retrieval systems. In *Web Search and Data Mining Conference*, volume 11, pages 360–368, Marina Del Rey. DOI: 10.1145/3159652.3159665.

Magalhães, D., Pozo, A., and Machado, S. (2022). Técnicas de aprendizado de máquinas aplicadas à classificação de decisões judiciais. *Revista de Estudos Empíricos em Direito*, 9. DOI: 10.19092/reed.v9.573.

Malaquias, E. O. *et al*. (2021). Acidentes em rodovias brasileiras: Um estudo com técnicas de machine learning para classificar a causa das ocorrências. In *Congresso de Pesquisa e Ensino em Transporte da ANPET*, number 35, pages 2322–2334, On-line. `https://www.anpet.org.br/anais/documentos/2019/Tr%C3%A1fego%20Urbano%20e%20Rodovi%C3%A1rio/Seguran%C3%A7a%20Vi%C3%A1ria%20II/5_686_AC.pdf`.

Martins, I. E. S. and De Andrade, M. H. S. (2021). Aplicação de técnicas de aprendizado de máquina na análise de ocorrências de trânsito de belo horizonte - mg. *Journal of Innovation and Science: Research and Application*, 1(1):67–74. DOI: 10.56509/joins.2021.v1.101.

Minussi, J. A., Damacena, C., and Ness Jr., W. L. (2002). Um modelo de previsão de solvência utilizando regressão logística. *Journal of Contemporary Administration*, 6(3):109–128. DOI: 10.1590/S1415-65552002000300007.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. ACM. DOI: 10.1145/2939672.2939778.

Santos, D. e. o. (2021). Machine learning approaches to traffic accident analysis and hotspot prediction. *Computers*, 10(12):1–15. DOI: 10.3390/computers10120157.

Scholz, J. C. W. and Pinheiro, Y. P. (2023). Prevendo a gravidade de acidentes rodoviários no brasil: A influência do ambiente e características do veículo. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação). Disponível em: `http://repositorio.utfpr.edu.br/jspui/handle/1/32667`.

Wu, J. *et al*. (2019). Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*. DOI: 10.11989/JEST.1674-862X.80904120.