ISSN: 2317-0840

# Classificação das Causas de Acidentes em Rodovias de Minas Gerais Utilizando Machine Learning

Lucas Ferreira Rosa[1†], Miguel Carvalho Nascimento[1], Wélson Antônio de Oliveira[1], Momate Emate Ossifo[2], Paulo Henrique Sales Guimarães[3]

[1]*Programa de pós-graduação em estatística e experimentação agropecuária, Instituto de ciências exatas e Tecnológicas, Universidade Federal de Lavras (UFLA).*
[2]*Escola Superior de Desenvolvimento Rural, Universidade Eduardo Mondlane, Moçambique.*
[3]*Departamento de Estatística, Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Lavras (UFLA).*

**Resumo:** *Os acidentes de transporte terrestre são uma das principais causas de mortalidade global, especialmente entre jovens e adultos, e causam impactos significativos em áreas essenciais como saúde e economia. Este estudo visa comparar e selecionar um modelo para classificar as causas de acidentes em rodovias federais de Minas Gerais, Brasil, utilizando métodos de tratamento de dados e técnicas de Machine Learning. Dados da Polícia Rodoviária Federal, de 01/01/2023 a 30/09/2023, foram usados para analisar os algoritmos Random Forest (RF), Support Vector Machine (SVM) e K-Nearest Neighbors (K-NN). O SVM apresentou a melhor acurácia e índice Kappa, enquanto o RF teve um desempenho razoável e o K-NN foi inferior e mais lento. A análise reforça a importância de escolher criteriosamente o modelo, considerando desempenho e eficiência computacional. O estudo visa apoiar as autoridades de segurança, como a Polícia Rodoviária Federal, na análise e registro das ocorrências, fortalecendo a construção de uma base de dados robusta e confiável para o preenchimento do Boletim de Acidente de Trânsito.*

**Palavras-chave:** *Aprendizado Supervisionado; K-Nearest Neighbor; Random Forest; Support Vector Machine.*

# Classification of Accident Causes on Highways in Minas Gerais Using Machine Learning

**Abstract:** *Road traffic accidents are one of the leading causes of global mortality, especially among young people and adults, and they have significant impacts on essential areas such as health and the economy. This study aims to compare and select a model to classify the causes of accidents on federal highways in Minas Gerais, Brazil, using data processing methods and Machine Learning techniques. Data from the Federal Highway Police, from January 1, 2023, to September 30, 2023, were used to analyze the Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (K-NN) algorithms. The SVM showed the best accuracy and Kappa index, while the RF had reasonable performance, and the K-NN was inferior and slower. The analysis reinforces the importance of carefully choosing the model, considering performance and computational efficiency. The study aims to support security authorities, such as the Federal Highway Police, in analyzing and recording occurrences, strengthening the construction of a robust and reliable database for filling out the Traffic Accident Report.*

**Keywords:** *Supervised Learning; K-Nearest Neighbor; Random Forest; Support Vector Machine.*

[†]Autor correspondente: lucasferreira_77@hotmail.com.br

Manuscrito recebido em: 24/07/2024
Manuscrito revisado em: 30/09/2024
Manuscrito aceito em: 03/10/2024

## Introdução

De acordo com o *Global Status Report on Road Safety* (GSRRS) de 2023 da *World Health Organization* (WHO), desde 2019, com mais de 2 mortes por minuto e mais de 3.200 por dia, os acidentes de trânsito continuam a ser a principal causa de morte de crianças e jovens com idades entre 5 e 29 anos e a 12ª principal causa de morte considerando todas as idades. Dois terços dessas mortes ocorrem entre pessoas em idade produtiva (18 a 59 anos), causando enormes prejuízos à saúde, além de impactos sociais e econômicos em toda a sociedade (WORLD HEALTH ORGANIZATION, 2023a).

Estima-se que ocorreram 1,19 milhão de mortes no trânsito em 2021, o que representa uma redução de 5% em relação às 1,25 milhão de mortes registradas em 2010. Esse declínio, embora modesto, reflete os esforços globais para melhorar a segurança nas estradas, resultando em uma significativa diminuição no número de óbitos, mesmo diante do substancial aumento da frota global de veículos, da expansão das redes viárias e do crescimento da população mundial em quase um bilhão de pessoas (WORLD HEALTH ORGANIZATION, 2023b).

O Brasil tem participado regularmente do GSRRS, fornecendo dados desde 2009. De acordo com o *Road Safety Brazil* 2023 da WHO, foram reportadas 31.468 fatalidades devido a acidentes de trânsito em 2021, embora estima-se que o número real seja aproximadamente 33.586. Esses dados evidenciam a persistente preocupação com a segurança viária no Brasil e refletem a necessidade contínua de políticas e intervenções eficazes para reduzir o impacto dos acidentes de trânsito na população (WORLD HEALTH ORGANIZATION, 2023c).

Conforme o Departamento de Trânsito de Minas Gerais (DETRAN-MG), o estado brasileiro que mais sofre com acidentes de trânsito é Minas Gerais, com registros de mais de 600 ocorrências diárias em rodovias federais (DETRAN-MG, 2022). Além disso, de acordo com a Secretaria de Estado de Saúde de Minas Gerais (2023), entre 2010 e 2023, o estado registrou 57.183 óbitos decorrentes de acidentes de trânsito, sendo 3.490 somente em 2023.

Identificar as causas principais desses acidentes é vital para orientar estratégias e ações preventivas nos locais críticos. No entanto, essa tarefa é desafiadora devido à complexidade resultante da grande quantidade de fatores envolvidos. Assim, o uso de ferramentas tecnológicas, como métodos de Aprendizado de Máquina (*Machine Learning* – ML) podem ser empregados para identificar e classificar essas causas (MALAQUIAS; TOSTA; CHAVES; RIBEIRO, 2021).

Os métodos ML têm sido amplamente utilizados em diversos campos de aplicação da sociedade e no campo acadêmico, bem como na engenharia de transportes. Eles são aplicados em áreas como a identificação de locais de estrada propensos a acidentes e a determinação da gravidade dos danos/lesões de acidentes, entre outros. O ML tem se mostrado eficaz na análise de grandes volumes de dados, permitindo a identificação de padrões e a previsão de ocorrências de acidentes, o que é crucial para a implementação de estratégias de segurança viária (MEGNIDIO-TCHOUKOUEGNO; ADEDEJI, 2023).

Diante desse cenário, este trabalho tem como objetivo principal explorar e comparar diferentes algoritmos de ML na classificação de acidentes de trânsito, enfatizando a importância das etapas de pré-processamento para os resultados preditivos. Almejamos, com os resultados obtidos neste estudo, fornecer dados robustos que possam auxiliar órgãos públicos, como a Polícia Rodoviária Federal, no preenchimento do Boletim de Acidente de Trânsito, contribuindo para o desenvolvimento de políticas de segurança e prevenção de acidentes no estado de Minas Gerais. Isso visa reduzir a incidência e, consequentemente, os impactos negativos dos acidentes de trânsito na população.

## Metodologia

O processo de modelagem consiste em quatro estágios principais, abrangendo a investigação e a preparação dos conjuntos de dados. Esses estágios incluem a compreensão dos dados,

o tratamento dos valores ausentes, o treinamento do modelo com um algoritmo de aprendizado de máquina e, por fim, a avaliação do desempenho do modelo utilizando métricas de classificação existentes.

Os dados utilizados para a modelagem foram obtidos por meio do Departamento da Polícia Rodoviária Federal, abrangendo acidentes ocorridos entre 01/01/2023 e 30/09/2023 em rodovias federais do estado de Minas Gerais, Brasil. Esses dados institucionais, classificados como Dados Abertos, são informações fornecidas em um formato legível por máquinas, sem restrições de licenças, patentes ou mecanismos de controle. Isso permite que qualquer pessoa os utilize, reutilize e redistribua livremente (POLÍCIA RODOVIÁRIA FEDERAL, 2023).

Inicialmente, o conjunto de dados era composto por 430.288 registros, contendo 37 variáveis e de todos os estados brasileiros. A descrição das variáveis encontra-se na Tabela 1 abaixo:

Como o objetivo é classificar as causas de acidentes no Estado de Minas Gerais, e nem todas as variáveis influenciam na variável resposta, a base de dados passou por um pré-processamento de filtro e seleção de variáveis. Portanto, os dados relativos aos demais estados da federação foram removidos do banco de dados, resultando na redução para 59.683 registros. As variáveis excluídas foram: "latitude", "longitude", "regional", "uf", "delegacia", "fase_dia", "uop", "id", "pesid", "causa_principal", "ordem_tipo_acidente", "uso_solo", "id_veiculo", "estado_fisico", "ilesos", "feridos_leves", "feridos_graves", "mortos", "marca", "classificacao_acidente". Isso ocorreu porque elas se referem a consequências do acidente ou não interferem no que diz respeito à causa do acidente.

Ainda nesta fase, foram identificados e retirados os registros preenchidos com "NA", totalizando 13.144 instâncias. Assim, o banco de dados utilizado no trabalho contém 46.539 registros válidos com informações de 17 variáveis, sendo uma delas a variável estudada.

Cada uma das variáveis passou por uma análise minuciosa para identificação de valores inconsistentes ou irreais. Para isso, as variáveis estudadas, "km", "data_inversa", "dia_semana", "horario", "br", "causa_acidente", "tipo_acidente", "sentido_via", "condicao_metereologica", "tipo_pista", "tracado_via", "tipo_veiculo", "ano_fabricacao_veiculo", "tipo_envolvido", "idade" e "sexo", passaram por tratamento de substituição dos outliers pela média da variável.

Além disso, foi necessário realizar a categorização de algumas variáveis na criação de classes para "horario", "data_inversa" e "causa_acidente". A variável "horario" foi categorizada como "madrugada" (de 0:00 às 5:59), "manhã" (de 6:00 às 11:59), "tarde" (de 12:00 às 17:59) e "noite" (de 18:00 às 23:59). A variável "data_inversa" foi transformada para representar o mês de ocorrência, correspondendo aos meses de "janeiro", "fevereiro", "março", "abril", "maio", "junho", "julho", "agosto" e "setembro".

A variável resposta "causa_acidente" inicialmente apresentava uma classificação em 72 categorias (Causas Correspondentes). Devido a essa grande dimensionalidade e visando evitar possíveis problemas de programação nos algoritmos, as causas de acidentes foram reorganizadas e realocadas nas categorias de Condutor, Pista, Veículo e Outros, conforme apresentado na Tabela 2.

A análise descritiva dos dados iniciou-se com a construção de gráficos de barras para melhor visualização do comportamento das variáveis em estudo. A análise inferencial do trabalho consiste no ajuste e avaliação de três modelos de ML supervisionados utilizados para a tarefa de classificação: *Random Forest* (RF), *Support Vector Machine* (SVM) e *K-Nearest Neighbors* (K-NN). Estes modelos foram implementados na linguagem R, utilizando as seguintes bibliotecas e versões: `randomForest` versão 4.7-1.1 (LIAW; WIENER, 2022), `e1071` versão 1.7-12 (MEYER *et al.*, 2023) e `class` versão 7.3-22 (Ripley; Venables, 2023). Para o ajuste, o conjunto de dados foi dividido em conjuntos de treino e teste, sendo 75% dos dados separados aleatoriamente para o treinamento do algoritmo, enquanto os 25% restantes foram reservados para avaliar o desempenho do mesmo.

Table 1: Description of accident variables.

| Variável | Descrição |
|---|---|
| id | Identificador único do acidente. |
| pesid | Identificador da pessoa envolvida no acidente. |
| data_inversa | Data do acidente, no formato reverso (YYYY-MM-DD). |
| dia_semana | Dia da semana em que o acidente ocorreu (ex.: segunda-feira, terça-feira). |
| horario | Horário do acidente. |
| uf | Unidade Federativa (estado) onde o acidente ocorreu. |
| br | Número da rodovia federal onde o acidente ocorreu. |
| km | Quilômetro da rodovia onde o acidente ocorreu. |
| municipio | Município onde o acidente ocorreu. |
| causa_principal | Causa principal do acidente. |
| causa_acidente | Descrição detalhada da causa do acidente. |
| ordem_tipo_acidente | Código que ordena o tipo de acidente. |
| tipo_acidente | Tipo de acidente (ex.: colisão, capotamento). |
| classificacao_acidente | Classificação do acidente em termos de gravidade (ex.: leve, grave). |
| fase_dia | Fase do dia em que o acidente ocorreu (ex.: manhã, tarde). |
| sentido_via | Sentido da via onde o acidente ocorreu (ex.: norte, sul). |
| condicao_meteorologica | Condições meteorológicas no momento do acidente (ex.: chuva, sol). |
| tipo_pista | Tipo de pista onde o acidente ocorreu (ex.: pista simples, pista dupla). |
| tracado_via | Traçado da via onde o acidente ocorreu (ex.: reta, curva). |
| uso_solo | Tipo de uso do solo nas imediações do acidente (ex.: urbano, rural). |
| id_veiculo | Identificador do veículo envolvido no acidente. |
| tipo_veiculo | Tipo de veículo envolvido (ex.: carro, caminhão). |
| marca | Marca do veículo. |
| ano_fabricacao_veiculo | Ano de fabricação do veículo. |
| tipo_envolvido | Tipo de pessoa envolvida (ex.: motorista, passageiro). |
| estado_fisico | Estado físico da pessoa envolvida (ex.: ileso, ferido). |
| idade | Idade da pessoa envolvida no acidente. |
| sexo | Sexo da pessoa envolvida no acidente. |
| ilesos | Número de pessoas ilesas no acidente. |
| feridos_leves | Número de feridos leves no acidente. |
| feridos_graves | Número de feridos graves no acidente. |
| mortos | Número de mortos no acidente. |
| latitude | Latitude do local do acidente. |
| longitude | Longitude do local do acidente. |
| regional | Regional da PRF responsável pelo local do acidente. |
| delegacia | Delegacia da PRF responsável pelo local do acidente. |
| uop | Unidade Operacional da PRF responsável pelo local do acidente. |

Source: from the authors (2023).

Os códigos para a reprodução das análises podem ser acessados no repositório do GitHub [ii]. O processamento e análise dos dados foi feito através de recursos e pacotes disponíveis no software R (R Core Team, 2023). Os componentes de hardware da máquina utilizada na execução dos algoritmos possuem as seguintes especificações: Processador: Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz, 8,00 GB (utilizável: 7,85 GB), Sistema operacional: Windows 11 Home Single Language, RStudio Team (2023) versão 4.2.2.

[ii]Disponível em: `https://github.com/lucasrosa3/Classifica-o_Acidentes_Rodovias_MG_ML`.

Table 2: Causes of Accidents Grouped by Category.

| Categoria | Causas Correspondentes |
|---|---|
| Condutor | Acessar a via sem observar a presença dos outros veículos, Ausência de reação do condutor, Condutor usando celular, Conversão proibida, Desrespeitar a preferência no cruzamento, Estacionar ou parar em local proibido, Frear bruscamente, Ingestão de álcool pelo condutor, Ingestão de substâncias psicoativas pelo condutor, Mal súbito do condutor, Manobra de mudança de faixa, Participar de racha, Reação tardia ou ineficiente do condutor, Retorno proibido, Suicídio (presumido), Trafegar com motocicleta (ou similar) entre as faixas, Transitar na contramão, Transitar no Acostamento, Ultrapassagem Indevida, Velocidade Incompatível, Acesso irregular, Transitar na calçada, Condutor deixou de manter distância do veículo da frente, Condutor Dormindo, Transtornos Mentais (exceto suicídio), Condutor não acionou o farol baixo durante o dia em rodovias de pista simples, Deixar de acionar o farol da motocicleta (ou similar). |
| Pista | Curva acentuada, Declive acentuado, Deficiência do Sistema de Iluminação/Sinalização, Demais falhas na via, Desvio temporário, Faixas de trânsito com largura insuficiente, Falta de acostamento, Falta de elemento de contenção que evite a saída do leito carroçável, Iluminação deficiente, Objeto estático sobre o leito carroçável, Obras na pista, Obstrução na via, Pista em desnível, Pista esburacada, Pista Escorregadia, Restrição de visibilidade em curvas horizontais, Restrição de visibilidade em curvas verticais, Sinalização encoberta, Sinalização mal posicionada, Sistema de drenagem ineficiente, Acúmulo de água sobre o pavimento, Acostamento em desnível, Redutor de velocidade em desacordo, Acúmulo de areia ou detritos sobre o pavimento, Acúmulo de óleo sobre o pavimento, Afundamento ou ondulação no pavimento, Área urbana sem a presença de local apropriado para a travessia de pedestres, Ausência de sinalização, Pista Escorregadia. |
| Veículo | Demais falhas mecânicas ou elétricas, Faróis desregulados, Problema com o freio, Modificação proibida, Problema na suspensão, Avarias e/ou desgaste excessivo no pneu, Carga excessiva e/ou mal acondicionada, Problema na suspensão. |
| Outros | Demais Fenômenos da natureza, Entrada inopinada do pedestre, Fumaça, Chuva, Ingestão de álcool e/ou substâncias psicoativas pelo pedestre, Ingestão de álcool ou de substâncias psicoativas pelo pedestre, Neblina, Pedestre andava na pista, Pedestre cruzava a pista fora da faixa, Animais na Pista. |

Source: from the authors (2023).

Para avaliar os modelos, foi utilizada a validação cruzada com 10 dobras, essencial para garantir a generalização a novos dados e fornecer uma estimativa confiável de desempenho em cenários reais. O processo consistiu em treinar o modelo em nove dobras e avaliá-lo na décima, repetindo esse procedimento 10 vezes para assegurar uma representação equitativa de todo o conjunto de dados. Os modelos foram ajustados com os parâmetros padrão dos pacotes: no RF, foram analisadas as variáveis mais importantes; no SVM, foi utilizado um kernel linear; e, no KNN, foram testados 10 valores de k para otimizar a acurácia. O resultado final de cada modelo corresponde à média das métricas de desempenho obtidas em cada dobra ao longo das iterações.

Para problemas de classificação binária, a avaliação da melhor solução pode ser definida com base na matriz de confusão, onde a linha representa a classe prevista e a coluna a classe real. Nesta matriz, Tabela 3 verdadeiro positivo (vp) e verdadeiro negativo (vn) referem-se ao número de instâncias corretamente classificadas, enquanto falso positivo (fp) e falso negativo (fn) denotam as instâncias classificadas incorretamente. A partir dessa matriz, diversas métricas podem ser geradas para avaliar o desempenho do classificador, sendo que algumas métricas foram estendidas para problemas de classificação multiclasse (HOSSIN e SULAIMAN 2015).

Table 3: Confusion Matrix for Binary Classification.

| | Classe Positiva Real | Classe Negativa Real |
|---|---|---|
| Classe Positiva Prevista | Verdadeiro Positivo (vp) | Falso Negativo (fn) |
| Classe Negativa Prevista | Falso Positivo (fp) | Verdadeiro Negativo (vn) |

Source: According to Hossin and Sulaiman (2015).

A avaliação do desempenho do melhor modelo entre os ajustados foi conduzido através de uma análise detalhada de medidas estatísticas, tais como acurácia, acurácia balanceada, precisão, *recall* e especificidade detalhadas na Tabela 4. Essas métricas proporcionarão uma visão abrangente das capacidades preditivas de cada algoritmo, permitindo a seleção do modelo mais adequado às características específicas dos dados (HOSSIN e SULAIMAN 2015; RAINIO; TEUHO; KLÉN, 2024; XU *et al.* 2023). A abordagem de múltiplos algoritmos proporciona uma análise comparativa sólida, permitindo-nos não apenas identificar o modelo mais eficaz, mas também compreender como diferentes algoritmos respondem a nuances específicas dos dados, contribuindo para *insights* mais profundos e confiáveis na modelagem preditiva dos acidentes.

Table 4: Performance Metrics for Machine Learning Models.

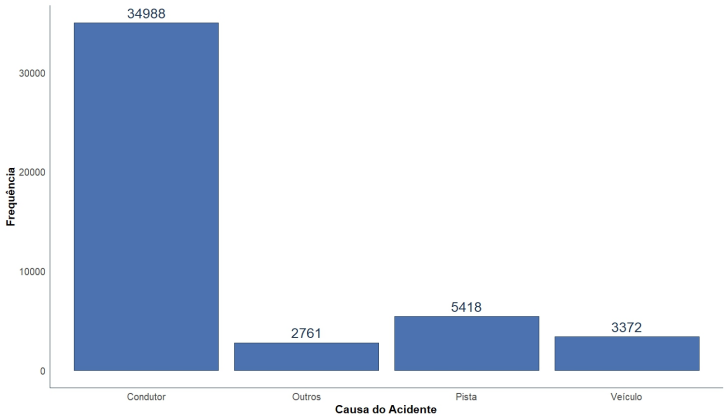
| Métricas | Fórmula | Foco da Avaliação |
|---|---|---|
| Acurácia | $\frac{\text{vp}+\text{vn}}{\text{vp}+\text{fp}+\text{vn}+\text{fn}}$ | Em geral, a métrica de acurácia mede a proporção de previsões corretas sobre o número total de instâncias avaliadas |
| Acurácia balanceada | $\frac{1}{2}\left(\frac{\text{vp}}{\text{vp}+\text{fn}}+\frac{\text{vn}}{\text{vn}+\text{fp}}\right)$ | Avalia a acurácia média em ambas as classes positivas e negativas, considerando o desequilíbrio das classes. Útil para conjuntos de dados desbalanceados, pois combina sensibilidade e especificidade para fornecer uma compreensão mais detalhada do desempenho do modelo. |
| Especificidade | $\frac{\text{vn}}{\text{vn}+\text{fp}}$ | Esta métrica é usada para medir a fração de padrões negativos que são corretamente classificados |
| F1 Score | $2\left(\frac{\text{Precisão}\cdot \textit{Recall}}{\text{Precisão}+\textit{Recall}}\right)$ | Equilibra precisão e *recall*, proporcionando uma única métrica de desempenho geral. |
| *Kappa* | $\frac{Acc-p_e}{1-p_e}$ | *Kappa* ajusta a taxa de concordância pelo acaso, oferecendo uma medida robusta de concordância para variáveis categóricas. |
| Precisão | $\frac{\text{vp}}{\text{vp}+\text{fp}}$ | A precisão é usada para medir os padrões positivos que são corretamente previstos a partir do total de padrões previstos em uma classe positiva. |
| *Recall* | $\frac{\text{vp}}{\text{vp}+\text{fn}}$ | O *recall* é usado para medir a fração de padrões positivos que são corretamente classificados |

Source: Modified according to Hossin e Sulaiman (2015); Rainio *et al.*, (2024); XU *et al.*, (2023).
Legend: em que, *Acc* é a acurácia do modelo, e $p_e = \frac{(\text{vp}+\text{fn})(\text{vp}+\text{fp})+(\text{vn}+\text{fp})(\text{vn}+\text{fn})}{(\text{vp}+\text{vn}+\text{fp}+\text{fn})^2}$.

## Resultados

A análise descritiva marca o início de nosso processo de investigação, oferecendo uma perspectiva abrangente dos dados em questão. Nesse sentido, elaboramos gráficos que sintetizam o comportamento dos dados. Conforme evidenciado na Figura 1, torna-se claro que as principais causas de acidentes estão vinculadas a fatores relacionados ao condutor. Isso ressalta a necessidade de uma análise mais aprofundada para compreender os padrões e fatores específicos envolvidos nessas ocorrências.

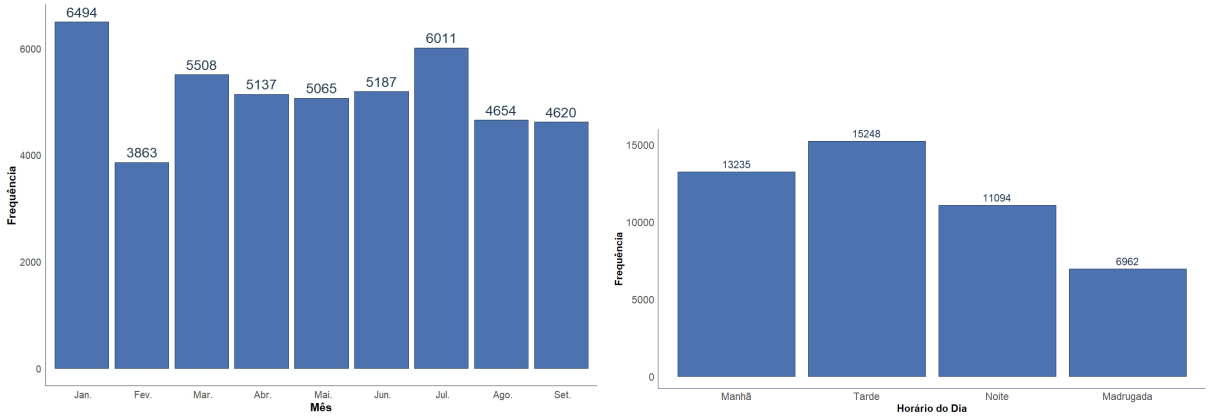
Figure 1: Causes of accidents.




Source: from the authors (2023).

Pode-se observar na Figura 2a que os meses de janeiro e julho destacam-se como os períodos que registram o maior número de ocorrências. Essa tendência pode ser atribuída aos períodos de férias, nos quais o fluxo de veículos aumenta nas vias de Minas Gerais devido às viagens e passeios. A análise mais aprofundada da Figura 2a pode chamar a atenção para os padrões temporais dessas ocorrências e contribuir para a implementação de medidas preventivas durante esses períodos específicos, seja por meio de conscientização ou aumento de fiscalização e policiamento.

Figure 2: Month and Period of the Day of the Causes of Accidents.




(a) (b)

Source: from the authors (2023).

Ademais, os períodos da manhã e tarde também registram maiores índices, como apresentado na Figura 2b. Além da questão do maior fluxo, através deste padrão pode-se realizar uma análise mais aprofundada para compreender os fatores específicos que contribuem para essa

tendência. A identificação das circunstâncias associadas a esses picos pode direcionar estratégias preventivas e intervenções específicas para otimizar a segurança viária nesses momentos

A partir de então, segue-se com o ajuste de cada um dos modelos. O primeiro modelo ajustado aos dados de acidentes de trânsito foi a *Random Forest* (RF). Para realizar esse ajuste, utilizamos os conjuntos de treino e teste mencionados na metodologia. Uma vez que já tínhamos criado esses conjuntos previamente, nossa preocupação principal foi a validação cruzada do modelo. Portanto, os resultados obtidos pelo RF podem ser visualizados na Tabela 5.

Table 5: Model Performance RF.

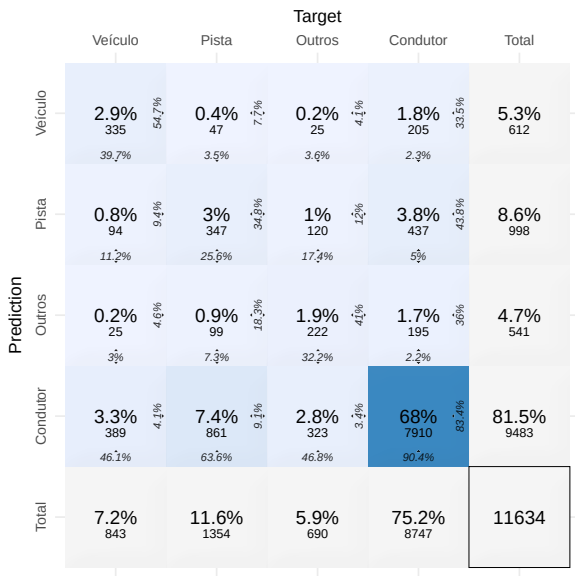
| Métrica de Avaliação | RF |
|---|---|
| Acurácia | 0,76 |
| Acurácia Balanceada | 0,65 |
| Especificidade | 0,58 |
| *Kappa* | 0,35 |
| Tempo (s) | 537,98 |

Source: from the authors (2023).

Destacamos que a RF obteve uma acurácia de 76%, mas a acurácia balanceada de 65% indica dificuldades com classes minoritárias. A especificidade média ponderada foi de 58% indica um desempenho razoável na previsão de negativos, enquanto o índice Kappa de 35% aponta uma concordância moderada, sugerindo necessidade de melhorias em dados desbalanceados. O tempo de execução foi de 537,98 segundos, uma consideração importante para a eficiência computacional em aplicações práticas. Em resumo, a RF demonstrou bom desempenho geral, com ênfase na acurácia, mas com oportunidades de refinamento em termos de tratamento de classes minoritárias e eficiência temporal.

A diagonal principal da Matriz de Confusão apresentada na Figura 3 destaca as classificações corretas, enquanto os elementos fora da diagonal indicam as classificações equivocadas.

Figure 3: Confusion matrix *Random Forest*.




Source: from the authors (2023).

De acordo com a Tabela 6, a RF demonstrou um desempenho robusto nos conjuntos de treino e teste para a classe "Condutor". Destaca-se pela elevada pontuação F1, precisão e *recall*. No entanto, para a classe "Outros", o desempenho moderado no treino não se refletiu

totalmente no teste, sugerindo uma possível dificuldade de generalização. A classe "Pista" apresentou desafios tanto no treino quanto no teste, com uma significativa queda nas métricas no conjunto de teste. Quanto à classe "Veículo", o modelo obteve bom desempenho no treino, mas enfrentou desafios na generalização durante o teste, evidenciados por uma redução nas métricas, especialmente no *recall*. Essa análise, baseada nos resultados apresentados na Tabela 6, fornece insights cruciais para ajustes específicos e possíveis melhorias futuras do modelo.

Table 6: Evaluation of the Random Forest (RF) adjusted for the training and test set.

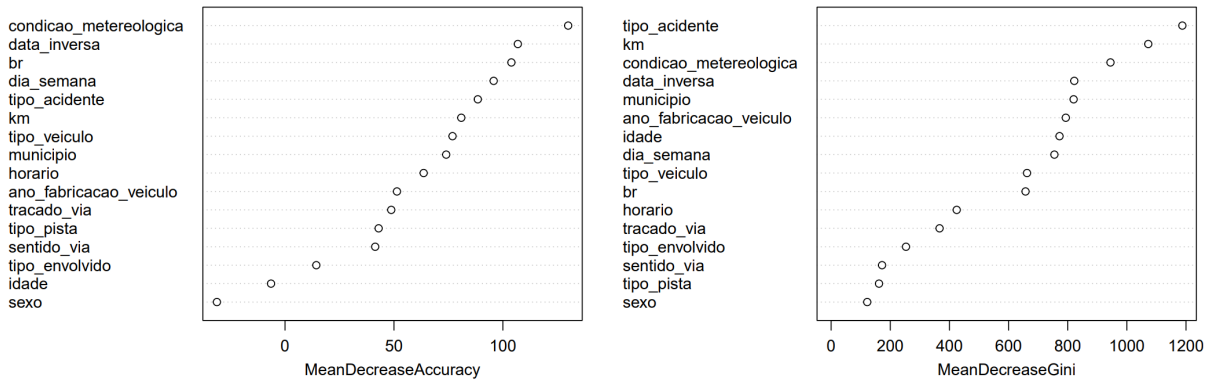
| Random Forest | Treino | | | Teste | | |
|---|---|---|---|---|---|---|
| | F1 | Precisão | Recall | F1 | Precisão | Recall |
| Condutor | 0,93 | 0,90 | 0,96 | 0,87 | 0,83 | 0,90 |
| Outros | 0,67 | 0,75 | 0,61 | 0,36 | 0,41 | 0,32 |
| Pista | 0,59 | 0,68 | 0,52 | 0,30 | 0,35 | 0,26 |
| Veículo | 0,72 | 0,81 | 0,65 | 0,46 | 0,55 | 0,40 |

Source: Authors (2023).

A RF apresenta uma funcionalidade poderosa, a capacidade de avaliar a importância de cada variável. Essa importância é medida pelo impacto nas métricas de acurácia ao permutar cada variável explanatória, fornecendo insights valiosos sobre quais características mais influenciam no desempenho do modelo. Essa análise é geralmente representada graficamente para uma compreensão visual e intuitiva da contribuição relativa de cada variável.

A análise das métricas "MeanDecreaseAccuracy"e "MeanDecreaseGini" (Figura 4) revelou informações valiosas sobre a importância das variáveis no modelo RF. Ao considerar o "MeanDecreaseAccuracy", observamos que a variável mais crucial para a precisão do modelo é a condição meteorológica. Isso significa que a exclusão dessa variável teria um impacto mais significativo na acurácia do modelo, indicando sua importância na tomada de decisões preditivas.

Figure 4: The importance of attributes in classifying the causes of RF accidents.

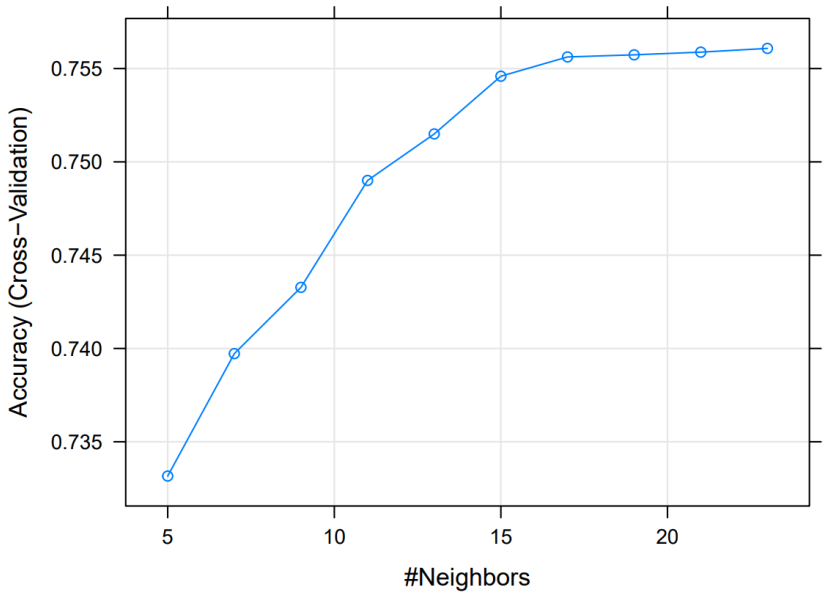



Source: from the authors (2023).

Por outro lado, ao avaliar o "MeanDecreaseGini", destacou-se a variável tipo de acidente como a mais relevante para a redução da impureza nos nós da floresta. Isso sugere que a exclusão da variável tipo de acidente resultaria em uma diminuição menos eficaz na impureza dos dados, indicando sua importância na segmentação e classificação dos exemplos.

O segundo modelo a ser ajustado foi o *K-Nearest Neighbors* (K-NN) aos dados de acidentes de trânsito. Para realizar este ajuste, foram utilizados os conjuntos de treino e teste citados na metodologia.

Figure 5: Analysis of the Distribution of Accuracy in K-NN Cross-Validation.




Source: from the authors (2023).

Diferentes valores do parâmetro $k$ foram testados durante o ajuste do modelo, e a métrica de acurácia foi utilizada para avaliar o desempenho. O valor ótimo para $k$, determinado pela maior acurácia durante a validação cruzada, foi encontrado em $k = 23$ (Figura 5), resultando em uma acurácia de 75,62% e um índice *Kappa* de 14,64%. Essa escolha de $k$ reflete a capacidade do modelo K-NN em generalizar bem para novos dados, proporcionando uma acurácia robusta durante o processo de validação.

A Tabela 7 oferece uma visão abrangente do desempenho do modelo K-NN ao ser avaliado por várias métricas obtidos através da matriz de confusão apresentado na Figura 6. A acurácia do modelo atingiu 75%, indicando a proporção de previsões corretas em relação ao total de observações. No entanto, ao considerar a acurácia balanceada, que leva em conta o desempenho para cada classe, o valor alcançado foi de 55%.

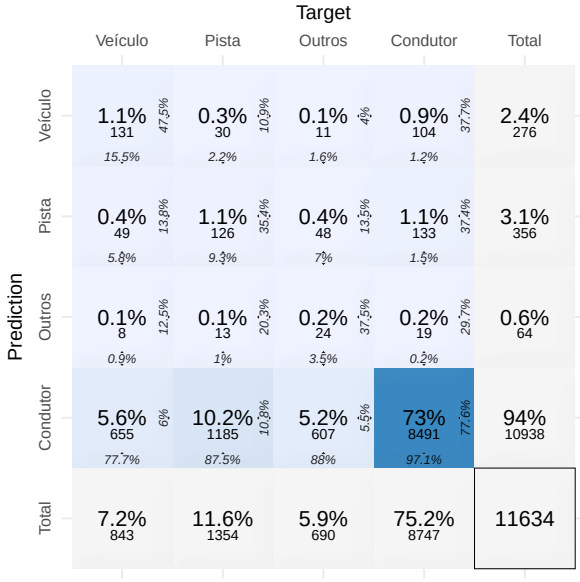
Table 7: Performance of the K-NN Model.

| Métrica de Avaliação | K-NN |
|---|---|
| Acurácia | 0,75 |
| Acurácia Balanceada | 0,55 |
| Especificidade | 0,36 |
| Kappa | 0,14 |
| Tempo (s) | 10519,18 |

Source: from the authors (2023).

A especificidade média ponderada, medida que expressa a habilidade do modelo em identificar corretamente os verdadeiros negativos, apresentou um resultado de 36%. O índice *Kappa*, que avalia a concordância além da chance aleatória, foi registrado em 14%. É importante observar que o tempo de execução do modelo K-NN foi considerável, totalizando 10.519,18 segundos. Essa métrica ganha relevância ao ser comparada com o tempo obtido no modelo RF, proporcionando insights sobre a eficiência computacional de ambos os modelos.

Figure 6: K-NN model confusion matrix.




Source: from the authors (2023).

A avaliação do modelo K-NN revela um desempenho variado nas métricas de F1, precisão e *recall* para cada classe (Tabela 8). Destaca-se um bom desempenho na classe "Condutor", tanto no conjunto de treino quanto no de teste, indicando um equilíbrio entre precisão e *recall*. Entretanto, observa-se um desafio significativo nas classes "Outros", "Pista"e "Veículo", com métricas de F1, precisão e *recall* mais baixas. Comparando com os resultados obtidos pela RF no conjunto de treino e teste, percebe-se que a RF apresentou resultados geralmente melhores, alcançando F1, precisão e *recall* superiores para essas classes em comparação com o K-NN.

Table 8: Evaluation of the K-NN adjusted for the training and test set.

| K-NN | Treino | | | Teste | | |
|---|---|---|---|---|---|---|
| | F1 | Precisão | Recall | F1 | Precisão | Recall |
| Condutor | 0,87 | 0,78 | 0,98 | 0,86 | 0,78 | 0,97 |
| Outros | 0,10 | 0,49 | 0,06 | 0,06 | 0,38 | 0,03 |
| Pista | 0,20 | 0,48 | 0,13 | 0,15 | 0,35 | 0,09 |
| Veículo | 0,27 | 0,57 | 0,18 | 0,23 | 0,47 | 0,16 |

Source: from the authors (2023).

Essa análise sugere que, no contexto específico desse problema de classificação, a RF pode ser mais robusta ao lidar com a diversidade das classes. No entanto, é importante considerar outros fatores, como o tempo de treinamento e a interpretabilidade, ao escolher o modelo mais adequado para implementação em um cenário prático.

O terceiro modelo aplicado para a análise dos dados de acidentes de trânsito foi o *Support Vector Machine* (SVM). O SVM é uma técnica de aprendizado de máquina que busca encontrar um hiperplano de separação ótimo entre diferentes classes, maximizando a margem entre os pontos de dados.

O SVM linear foi aplicado aos dados de acidentes de trânsito, e os resultados são apresentados na Tabela 9. A métrica de acurácia alcançou 79%, indicando um desempenho sólido na classificação das diferentes classes. A acurácia balanceada, que considera o desempenho em todas as classes, atingiu 67%. Ao analisar a especificidade do modelo, observamos um valor de

58%, indicando a capacidade de identificar corretamente as instâncias negativas. O índice *kappa*, que leva em consideração a concordância além do acaso, foi de 41%, reforçando a robustez do modelo. É relevante notar que o SVM linear demandou um tempo de processamento de 676,88 segundos.

Table 9: Model Performance SVM.

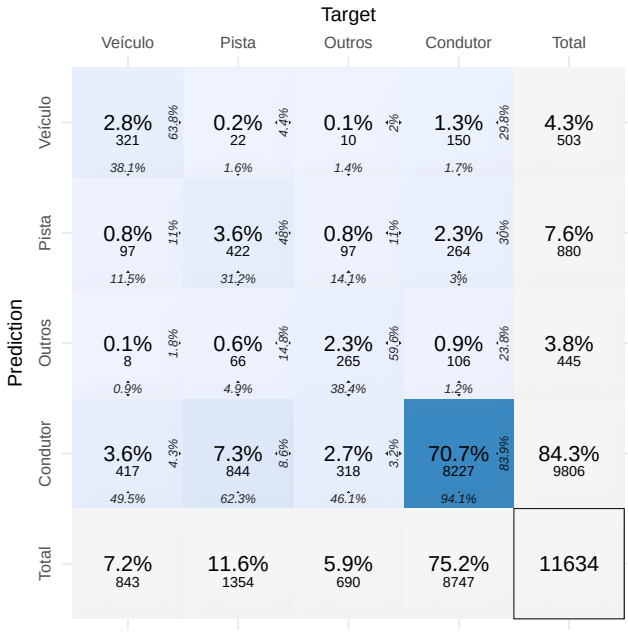
| Métrica de Avaliação | SVM |
|---|---|
| Acurácia | 0,79 |
| Acurácia Balanceada | 0,67 |
| Especificidade | 0,58 |
| Kappa | 0,41 |
| Tempo (s) | 676,88 |

Source: from the authors (2023).

A quantidade de *Support Vectors*, que são os pontos de dados mais importantes para a determinação do hiperplano de separação, foi identificada como 17015. Essa informação é crucial para compreender a complexidade do modelo.

A Figura 7 apresenta a Matriz de Confusão para o modelo SVM.

Figure 7: SVM model confusion matrix.




Source: from the authors (2023).

Ao analisar a Tabela 10 de avaliação do SVM para treino e teste, observamos que o modelo manteve um desempenho robusto em ambas as situações para a classe "Condutor", por exemplo, o F1-score, precisão e *recall* foram consistentemente altos tanto no treino quanto no teste, indicando uma capacidade consistente de classificação. No entanto, para as demais classes o desempenho no conjunto de teste foi inferior ao desempenho no conjunto de treino indicando uma tendência de overfitting.

Table 10: Evaluation of the SVM adjusted for the training and test set.

| SVM | Treino | | | Teste | | |
|---|---|---|---|---|---|---|
| | F1 | Precisão | Recall | F1 | Precisão | Recall |
| Condutor | 0,90 | 0,85 | 0,96 | 0,89 | 0,84 | 0,94 |
| Outros | 0,57 | 0,72 | 0,46 | 0,47 | 0,60 | 0,38 |
| Pista | 0,47 | 0,60 | 0,39 | 0,38 | 0,48 | 0,31 |
| Veículo | 0,57 | 0,76 | 0,46 | 0,48 | 0,64 | 0,38 |

Source: from the authors (2023).

Os resultados obtidos a partir da aplicação dos diferentes modelos de *Machine Learning* são apresentados na Tabela 11. Cada modelo foi avaliado com base em diversas métricas, incluindo acurácia, acurácia balanceada, especificidade, índice *kappa* e tempo de execução.

Table 11: Results of Machine Learning Models applied to the traffic accident dataset in Minas Gerais in 2023.

| Modelo | Acurácia | Acurácia balanceada | Especificidade | *Kappa* | Tempo(s) |
|---|---|---|---|---|---|
| RF | 0,76 | 0,65 | 0,58 | 0,35 | 537,98 |
| k-NN | 0,75 | 0,55 | 0,36 | 0,14 | 10519,18 |
| SVM | 0,79 | 0,67 | 0,58 | 0,41 | 676,8 |

Source: from the authors (2023).

O modelo RF demonstrou uma acurácia de 76%, acurácia balanceada de 65%, especificidade de 58%, índice *kappa* de 35%, e um tempo de execução de 537,98 segundos. Esses resultados sugerem um desempenho geral razoável do modelo, com boa capacidade de classificação, especialmente em termos de acurácia.

Já o modelo K-NN apresentou uma acurácia de 75%, acurácia balanceada de 55%, especificidade de 36%, índice *kappa* de 14%, e um tempo de execução mais elevado, totalizando 10.519,18 segundos. Esses resultados indicam um desempenho inferior em comparação com o Random Forest, com destaque para uma menor especificidade e índice kappa.

Por fim, o modelo SVM obteve uma acurácia de 79%, acurácia balanceada de 67%, especificidade de 58%, índice *kappa* de 41%, e um tempo de execução de 676,8 segundos. Esses resultados sugerem que o SVM teve um desempenho superior em relação aos outros modelos, alcançando uma acurácia e um índice *kappa* mais elevados.

## Conclusão

A obtenção de dados históricos consistentes e confiáveis sobre acidentes é fundamental para diagnósticos precisos e para a criação de modelos robustos que contribuem para a redução de acidentes e projeções futuras. Contudo, enfrentamos um desbalanceamento considerável entre as classes na variável avaliada, que foi a causa dos acidentes, e imprecisões no preenchimento dos dados.

Diante desse cenário, este trabalho explorou e comparou diferentes algoritmos de *Machine Learning* na classificação de acidentes de trânsito, enfatizando a importância das etapas de pré-processamento para os resultados preditivos. A aplicação de modelos de ML proporcionou *insights* valiosos. A RF demonstrou um desempenho geral razoável, destacando-se pela acurácia. O modelo k-NN apresentou resultados inferiores, especialmente em termos de especificidade e do índice *kappa*, além de demandar um tempo significativamente maior para a execução. Por outro lado, o modelo SVM se destacou, alcançando uma acurácia superior e um índice *kappa* mais elevado em comparação com os demais.

Em conclusão, entre os algoritmos comparados, o SVM apresentou os melhores resultados, oferecendo um bom equilíbrio entre precisão e tempo de processamento. Isso ajuda a fornecer dados robustos que podem auxiliar órgãos públicos, como a Polícia Rodoviária Federal, no preenchimento do Boletim de Acidente de Trânsito, contribuindo para o desenvolvimento de políticas de segurança e prevenção de acidentes no estado de Minas Gerais. Assim, almejamos reduzir a incidência e os impactos negativos dos acidentes de trânsito na população.

## Agradecimentos

A Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG pela concessão da bolsa de doutorado ao primeiro autor e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES pela concessão da bolsa de doutorado ao segundo autor.

## Referências

DETRAN-MG. **Registro Nacional de Acidentes e Estatísticas de Trânsito**. 2024. Disponível em: https://transito.mg.gov.br/sobre-1/estatisticas/registro-nacional-de-acidentes-de-transito. Acesso em: 15 jan. 2024.

HOSSIN, M. N.; SULAIMAN A. Review on Evaluation Metrics for Data Classification Evaluations. **International Journal Of Data Mining & Knowledge Management Process**, [S.L.], v. 5, n. 2, p. 01-11, 31 mar. 2015. Academy and Industry Research Collaboration Center (AIRCC). http://dx.doi.org/10.5121/ijdkp.2015.5201.

LIAW A, WIENER M (2022). randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.7-1.1. https://cran.r-project.org/web/packages/randomForest/index.html.

MALAQUIAS, E. O.; TOSTA, M. C. R.; CHAVES, G. L. D.; RIBEIRO, G. M. ACIDENTES EM RODOVIAS BRASILEIRAS: um estudo com técnicas de machine learning para classificar a causa das ocorrências. **In: 35° CONGRESSO DE PESQUISA E ENSINO EM TRANSPORTE DA ANPET**, 35., 2021, online. [S. I.]: Anpet, 2021. p. 2322-2334.

MEGNIDIO-TCHOUKOUEGNO, M; ADEDEJI, J. A. Machine Learning for Road Traffic Accident Improvement and Environmental Resource Management in the Transportation Sector. **Sustainability**, [S.L.], v. 15, n. 3, p. 2014, 20 jan. 2023. MDPI AG. http://dx.doi.org/10.3390/su15032014.

MEYER, D.; DIMITRADOU, E.; HORNIK, K.; WEINGESSEL, A.; LEISCH, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-16, 2023. https://CRAN.R-project.org/package=e1071.

MINAS GERAIS. SECRETARIA DE SAÚDE DE MINAS GERAIS. . **Acidentes por Transporte Terrestre**. Disponível em: http://vigilancia.saude.mg.gov.br/index.php/paineis-tematicos/. Acesso em: 15 jan. 2024.

PRF (Polícia Rodoviária Federal). Dados Abertos da PRF. 2023. Disponível em: https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf. Acesso em: 12 jan. 2023.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2023. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RAINIO, O.; TEUHO, J.; KLÉN, R. Evaluation metrics and statistical tests for machine learning. **Scientific Reports**, [S.L.], v. 14, n. 1, p. 255-269, 13 mar. 2024. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-024-56706-x.

RIPLEY, B.; VENABLES, W. class: Functions for Classification. R package version 7.3-22, 2023.https://cran.r-project.org/web/packages/class/index.html.

World Health Organization. **Despite notable progress, road safety remains urgent global issue**. 2023. https://www.who.int/news/item/13-12-2023-despite-notable-progress-road-safety-remains-urgent-global-issue. Acesso: 10 jan. 2024.

World Health Organization. **Global status report on road safety 2023**. 2023. Disponível em: https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023. Acesso em: 10 jan. 2024.

World Health Organization. **Road safety Brazil 2023 country profil**e. 2023. Disponível em: https://www.who.int/publications/m/item/road-safety-bra-2023-country-profile. Acesso em: 10 jan. 2024.

XU, Haojie *et al.* PFD-Assisted Sampling PLL With Seamless PFD-SPD Switching Scheme and Supply-Insensitive RO. **Ieee Microwave And Wireless Technology Letters**, [S.L.], v. 33, n. 10, p. 1474-1477, out. 2023. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/lmwt.2023.3307733.