**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**

**JÚLIO CÉSAR WERNER SCHOLZ**

**YAN PIETRZAK PINHEIRO**

# PREVENDO A GRAVIDADE DE ACIDENTES RODOVIÁRIOS NO BRASIL: A INFLUÊNCIA DO AMBIENTE E CARACTERÍSTICAS DO VEÍCULO

**CURITIBA**

**2023**

**JÚLIO CÉSAR WERNER SCHOLZ**
**YAN PIETRZAK PINHEIRO**

# PREVENDO A GRAVIDADE DE ACIDENTES RODOVIÁRIOS NO BRASIL: A INFLUÊNCIA DO AMBIENTE E CARACTERÍSTICAS DO VEÍCULO

## Predicting the severity of highway accidents in Brazil: the role of environmental factors and vehicle features

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Sistemas de Informação do Curso de Bacharelado em Sistemas de Informação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Thiago Henrique da Silva

**CURITIBA**
**2023**

**JÚLIO CÉSAR WERNER SCHOLZ**
**YAN PIETRZAK PINHEIRO**

# PREVENDO A GRAVIDADE DE ACIDENTES RODOVIÁRIOS NO BRASIL: A INFLUÊNCIA DO AMBIENTE E CARACTERÍSTICAS DO VEÍCULO

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Sistemas de Informação do Curso de Bacharelado em Sistemas de Informação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 30/junho/2023

---

Thiago Henrique da Silva
Doutorado
Universidade Tecnológica Federal do Paraná

---

Anelise Munaretto Fonseca
Doutorado
Universidade Tecnológica Federal do Paraná

---

Ricardo Lüders
Doutorado
Universidade Tecnológica Federal do Paraná

**CURITIBA**
**2023**

## RESUMO

O presente trabalho problematiza os acidentes resultantes do alto uso de veículos motorizados como meio de transporte na sociedade, dado que todos os anos a vida de aproximadamente 1,3 milhões de pessoas é interrompida no mundo como resultado de um acidente de trânsito. Essa realidade tem impulsionado pesquisas para compreender e prever a gravidade desses acidentes. Nesse sentido, o estudo propõe a utilização da mineração de dados como uma abordagem para analisar e explorar conjuntos de dados de acidentes de trânsito ocorridos nas rodovias brasileiras entre os anos de 2017 e 2022, fornecidos pela Polícia Rodoviária Federal, além disso, foram inseridos dados de preços dos veículos a partir da tabela FIPE, permitindo uma análise mais abrangente que considera também o valor financeiro dos veículos. O objetivo é avaliar a capacidade preditiva de modelos de classificação em relação à gravidade dos acidentes, com foco nas características dos veículos e nos fatores do ambiente. Por meio da aplicação de algoritmos de classificação e de técnicas de explicabilidade de aprendizado de máquina, pretende-se adquirir conhecimentos relevantes relacionados a esse conjunto de dados, contribuindo para o entendimento e a prevenção de acidentes. Como resultado, os atributos relacionadas às características dos veículos impactaram mais positivamente na capacidade preditiva dos modelos quando comparados aos atributos que descrevem o ambiente e as demais variáveis.



**Palavras-chave:** análise de acidentes; mineração de dados; algoritmo de classificação; explicabilidade.

**ABSTRACT**

This work supports the problematization of the rate of traffic accidents due to the high use of automobiles for utility or transport in Brazil. Every year the lives of approximately 1.3 million people are interrupted due to traffic accidents worldwide. This fact has motivated some researchers to engage in analysis and prediction regarding the severity of accidents. In this regard, this study proposes the use of data mining as an approach to analyze and explore datasets of traffic accidents that occurred on Brazilian highways between the years 2017 and 2022 as provided by the Federal Highway Police. Additionally, vehicle price data from the FIPE table were included, allowing for a more comprehensive analysis that also considers the financial value of the vehicles. The goal is to assess the predictive capability of classification models regarding the severity of accidents, focusing on vehicle characteristics and environmental factors. Through applying classification algorithms and machine learning explainability techniques, we aim to acquire relevant knowledge related to this dataset, contributing to understanding and prevention of accidents. As a result, the attributes related to vehicle characteristics had a more positive impact on the predictive capability of the models when compared to the attributes describing the environment and other variables.

**Keywords:** analysis of accidents; data mining; classification algorithm; explicability.

## LISTA DE FIGURAS


**Figura 1 – Fluxo metodológico do estudo . . . . . 21**

**Figura 2 – Distribuição dos acidentados ao longo dos anos . . . . . 30**

**Figura 3 – Distribuição dos acidentados de acordo com o ano de fabricação do veículo . . . . . 31**

**Figura 4 – Percentual de acidentes graves de acordo com o ano de fabricação do veículo . . . . . 31**

**Figura 5 – Distribuição e percentual de acidentes graves de acordo com o tipo da pista . . . . . 32**

**Figura 6 – Distribuição da gravidade das lesões de acordo com o horário que o acidente ocorreu . . . . . 32**

**Figura 7 – Gráfico da importância das variáveis por permutação no Cenário 1 . . . 35**

**Figura 8 – Importância por permutação das variáveis da Árvore de Decisão no Cenário 2 . . . . . 36**

**Figura 9 – Importância das variáveis por permutação no Cenário 3 . . . . . 37**

**Figura 10 – Valores Shapley para Árvore de Decisão no Cenário 3 . . . . . 38**

**Figura 11 – Análise de dependência do ano de fabricação do veículo utilizando SHAP 39**

## LISTA DE TABELAS

**Tabela 1 – Técnicas utilizadas na tratativas das variáveis consideradas para análise 32**
**Tabela 2 – Performance dos modelos de classificação no Cenário 1 . . . . . 34**
**Tabela 3 – Performance dos modelos de classificação no Cenário 2 . . . . . 36**
**Tabela 4 – Performance dos modelos de classificação no Cenário 3 . . . . . 37**
**Tabela 5 – Desvio padrão do F1-score dos modelos nos três cenários . . . . . 40**
**Tabela 6 – F1-Score dos modelos de classificação nos três cenários . . . . . 40**

## LISTA DE ABREVIATURAS E SIGLAS

### Siglas

| | |
|---|---|
| DT | Árvore de Decisão, do inglês *Decision Tree* |
| EUA | Estados Unidos da América |
| FIPE | Fundação Instituto de Pesquisas Econômicas |
| IPEA | Instituto de Pesquisa Econômica Aplicada |
| KNN | K-Vizinhos Mais Próximos, do inglês *K-Nearest Neighbors* |
| MLP | Perceptron Multicamadas, do inglês *Multilayer Perceptron* |
| NB | Bayes Ingênuo, do inglês *Naive Bayes* |
| OMS | Organização Mundial da Saúde |
| PRF | Polícia Rodoviária Federal |
| RF | Floresta Aleatória, do inglês *Random Forest* |
| RENAEST | Registro Nacional de Acidentes e Estatísticas de Trânsito |
| UTFPR | Universidade Tecnológica Federal do Paraná |
| WHO | Organização Mundial da Saúde, do inglês, *World Health Organization* |

**SUMÁRIO**


**1 INTRODUÇÃO . . . . . 9**

**1.1 Objetivos . . . . . 10**

1.1.1 Objetivo geral . . . . . 10

1.1.2 Objetivos específicos . . . . . 10

**2 CONTEXTUALIZAÇÃO E REVISÃO DE LITERATURA . . . . . 11**

**2.1 Mineração de dados . . . . . 11**

**2.2 Técnicas de mineração de dados . . . . . 12**

2.2.1 Árvore de Decisão . . . . . 13

2.2.2 Floresta aleatória . . . . . 13

2.2.3 Classificador Bayes Ingênuo . . . . . 14

2.2.4 K-Vizinhos Mais Próximos . . . . . 15

2.2.5 Rede neural Perceptron multicamadas . . . . . 15

**2.3 Aprendizado de Máquina Explicável . . . . . 16**

2.3.1 Valores Shapley . . . . . 16

2.3.2 Importância por Permutação . . . . . 17

**2.4 Análise de acidentes utilizando mineração de dados . . . . . 17**

**3 METODOLOGIA . . . . . 21**

**3.1 Ambientação do tema e definição de objetivos . . . . . 22**

**3.2 Análise dos dados . . . . . 22**

**3.3 Preparação do conjunto dados . . . . . 22**

3.3.1 Tratamento . . . . . 22

3.3.2 Codificação de dados (*Data enconding*) . . . . . 23

3.3.3 Redução da dimensionalidade . . . . . 24

3.3.4 Escalonamento de dados (*Data scaling*) . . . . . 25

**3.4 Experimentação . . . . . 26**

3.4.1 Definição de cenários . . . . . 26

3.4.2 Codificação . . . . . 27

3.4.3 Medição dos resultados . . . . . 28

**4 CARACTERIZAÇÃO DOS DADOS . . . . . 30**

**5 RESULTADOS . . . . . 34**

5.1 Cenário 1 - Base . . . . . 34
5.2 Cenário 2 - Ambiente . . . . . 35
5.3 Cenário 3 - Veículos . . . . . 36
5.4 Discussão . . . . . 39
6 CONCLUSÃO . . . . . 42
REFERÊNCIAS . . . . . 44

APÊNDICES 46
APÊNDICE A – RESULTADOS COMPLEMENTARES . . . . . 48

# 1 INTRODUÇÃO

A adoção dos automóveis como meio técnico e de transporte impacta amplamente a organização da sociedade, uma vez que o automóvel se desenvolve no e para o urbano (SCHOR, 1999). Grande parte do desenvolvimento industrial e planejamento urbano foram direcionadas à constituição do sistema automobilístico. Assim, por terem se tornado uma necessidade social, os automóveis têm se constituído como parte basilar da vida contemporânea, acarretando diversas consequências.

A incorporação intensiva dos automóveis na vida cotidiana traz benefícios importantes como a agilidade no transporte humano e de cargas, além do conforto e redução de tempo de viagens, por exemplo. No entanto, há também o lado nocivo desta realidade. Com relação aos malefícios, é possível demarcar a poluição sonora e do meio ambiente, bem como o problema de maior gravidade: as mortes causadas por acidentes de trânsito.

É de conhecimento público que os acidentes de trânsito são um problema frequente da sociedade. Todos os anos, a vida de aproximadamente 1,3 milhões de pessoas é interrompida como resultado de um acidente de trânsito. Entre 20 e 50 milhões de pessoas sofrem lesões não fatais, com muitas incorrendo em deficiência como resultado de suas lesões (WHO, 2021). A WHO[1] classifica os acidentes de trânsito como a primeira causa principal de morte de crianças e jovens entre 5 e 29 anos. Segundo o portal do RENAEST[2], no ano de 2021 ocorreram 878.208 acidentes no Brasil, onde um pouco mais 1,3 milhões de pessoas estiveram envolvidas, resultando em 20.053 óbitos.

Por ser um problema mundial, a Assembleia Geral das Nações Unidas determinou alcançar uma redução de 50% no número global de mortes e ferimentos devido a acidentes de trânsito (WHO, 2021). O Brasil é um dos signatários deste esforço global e se propôs a adotar uma série de políticas e medidas, como campanhas preventivas, redução dos limites de velocidade, melhorias de infraestrutura etc. (IPEA, 2021). Concretamente, entretanto, poucos foram os esforços para alcançar esse objetivo. Segundo o IPEA são estimados em média R$ 130 bilhões de gastos com acidentes de trânsito no país, revelando, assim, que o alto número de acidentes no trânsito ainda é uma realidade.

Desta forma, iniciativas que visam a redução e prevenção dos acidentes de trânsito se fazem mais necessárias que nunca. Nesse sentido, ressalta-se que a área de análise e mineração de dados de acidentes tem se popularizado nos últimos anos, resultando em uma ampla coleção de estudos que abordam essa problemática. Justifica-se a importância dos estudos na área, e também do presente trabalho, pois se faz imprescindível compreender os fatores de risco e prever características de acidentes graves. Com essas informações é possível elaborar e implantar medidas preventivas de alta eficácia. Assim, apresenta-se que o conceito de Mine-

[1] Organização Mundial de Saúde (em inglês: World Health Organization - WHO)

[2] Disponível em: www.gov.br/infraestrutura/pt-br/ assuntos/transito/arquivos-senatran/docs/renaest Acesso em: 26/04/2022

ração de Dados está cada vez mais presente nessas atividades de descoberta de informações (CÔRTES; PORCARO; LIFSCHITZ, 2002)

No Brasil, a Polícia Rodoviária Federal vem coletando dados de acidentes desde 2007, e disponibiliza estes dados através da secção de dados abertos no site do governo a respeito dos acidentes ocorridos nas rodovias federais em todos os estados. Este conjunto de dados contempla ocorrências registradas entre os anos de 2007 até 2022. Sendo que neste período é possível observar que não houve uma redução significativa no número de mortes e uma insuficiente redução no número de feridos, mesmo com a diminuição na quantidade de pessoas envolvidas nos acidentes.

A partir do conjunto de dados da PRF, será realizada a inserção de atributos relacionados às especificações dos veículos. Para então aplicar algoritmos de classificação de dados e medir o impacto dos atributos relacionados ao veículo, aos fatores externos e do ambiente em relação à severidade dos acidentes.

Esperamos que com o resultado deste trabalho, possamos prover informações úteis a respeito das causas dos acidentes, para que essas informações possam ser utilizada por entidades públicas e privadas a fim de reduzir a severidade e o número de acidentes nas rodovias brasileiras.

## 1.1 Objetivos

Nesta seção, serão apresentados os objetivos gerais e específicos deste trabalho.

### 1.1.1 Objetivo geral

Avaliar a importância de fatores externos e do veículo na gravidade dos acidentes em rodovias federais brasileiras utilizando técnicas de aprendizado de máquina

### 1.1.2 Objetivos específicos

- Aplicar diferentes modelos de classificação com algoritmos de mineração de dados em diferentes cenários.
- Avaliar como atributos relacionados às especificações do veículo impactam na severidade das lesões nos acidentes.
- Avaliar como os fatores externos e do ambiente de um acidente impactam na severidade das lesões.

# 2 CONTEXTUALIZAÇÃO E REVISÃO DE LITERATURA

Para uma melhor compreensão do contexto no qual se situa o presente trabalho, este capítulo apresenta a mineração de dados aplicada à análise de acidentes de trânsito e uma visão geral sobre as técnicas mais relevantes. Ao final do capítulo alguns trabalhos relacionados são descritos, mostrando suas características, métodos e resultados obtidos.

## 2.1 Mineração de dados

A mineração de dados é definida por Zhu e Davidson (2007) como o processo de análise e exploração de grandes quantidades de dados com o objetivo de descobrir regras e padrões significativos que podem eventualmente se tornar conhecimento. Destaca-se também que a mineração de dados, segundo Han, Kamber e Pei (2012), também é conhecida como o processo de descobrimento de conhecimento a partir de dados (em inglês, *knowledge discovery from data* - KDD), o qual é composto pelos seguintes passos:

1. Limpeza de dados
2. Integração de dados
3. Seleção de dados
4. Transformação de dados
5. Mineração de dados
6. Avaliação de padrões
7. Apresentação do conhecimento

As quatro primeiras etapas têm como função preparar os dados brutos, a primeira é responsável pela remoção de dados inconsistentes e ruídos. A segunda etapa, se for o caso, é onde múltiplas fontes de dados serão combinadas. No terceiro, passo os dados relevantes para a tarefa de análise são recuperados do conjunto de dados. A quarta, transformação de dados, é onde os dados serão consolidados no formato mais apropriado para a tarefa de mineração.

As estratégias utilizadas para o pré-processamento de dados frequentemente envolvem a inclusão, remoção ou transformação de informações presentes no conjunto de dados de treinamento conforme citado na metodologia. A etapa de pré-processamento e transformação dos dados é fundamental (KUHN; JOHNSON, 2013), uma vez que permite a adequação do conjunto de dados à técnica de mineração utilizada, tornando-os aptos a análise subsequente e impactando positiva ou negativamente na habilidade preditiva de um modelo.

> A forma como os preditores são codificados, chamada de *feature engineering* (engenharia de características), pode ter um impacto significativo no desempenho do modelo. Por exemplo, usar combinações de preditores às vezes pode ser mais eficaz do que usar os valores individuais: a proporção de dois preditores pode ser mais eficaz do que usar dois preditores independentes. Frequentemente, a codificação mais eficaz dos dados é informada pela compreensão do problema pelo modelador e, portanto, não é derivada de nenhuma técnica matemática.(KUHN; JOHNSON, 2013, p. 27-28, tradução nossa).

Portanto, o pré-processamento de dados permite que o modelador prepare os dados de entrada para o modelo para maximizar o desempenho e a precisão do modelo. Isso pode incluir a seleção de características relevantes, a limpeza de dados ausentes ou inconsistentes, e a transformação de dados em formatos adequados para o modelo. Ao fazer isso, o modelador pode garantir que o modelo esteja recebendo as informações corretas e relevantes, resultando em um modelo mais eficaz e preciso.

A quinta etapa é a mineração de dados, neste caso considerada parte do processo de descoberta de conhecimento, é a etapa onde métodos inteligentes são aplicados para extrair padrões dos dados. E na sequência é feita uma avaliação e análise dos padrões encontrados que poderão ser apresentados e compartilhados.

## 2.2 Técnicas de mineração de dados

Na sequência são apresentados alguns algoritmos de classificação de dados. A classificação é usada para classificar cada item em um conjunto de dados em um conjunto predefinido de classes ou grupos, desta forma, os algoritmos de classificação operam construindo um modelo ou um classificador para prever rótulos categóricos (KESAVARAJ; SUKUMARAN, 2013). A tarefa de classificação pode ser resumida em 3 etapas:

- A criação do modelo;
- O treinamento do modelo;
- A validação do modelo;

O treinamento é onde o algoritmo aplicará alguma função para descrever os dados em relação às classes do problema em um processo de indução. E a validação a partir de um processo de dedução consiste na utilização do modelo treinado para averiguar sua eficácia na predição dos rótulos categóricos. Uma boa prática na etapa de validação é a utilização de dados diferentes dos utilizados no treinamento para estimar a acurácia do modelo.

Ainda segundo KESAVARAJ; SUKUMARAN, cada algoritmo de classificação é específico em seu domínio de problema, ou seja, não há algoritmo de classificação único que seja o melhor para todos os tipos de conjuntos de dados.

### 2.2.1 Árvore de Decisão

A Árvore de Decisão, do inglês *Decision Tree* (DT) é um algoritmo de aprendizado de máquina supervisionado baseado em árvore usado para resolver problemas de classificação caso a variável em estudo tenha um conjunto finito de estados ou valores possíveis, e problemas de regressão caso o conjunto seja contínuo.

O objetivo do algoritmo é prever a variável em estudo a partir de um conjunto de variáveis de entrada e seus atributos. A abordagem constrói uma estrutura em árvore mediante uma série de divisões binárias do nó raiz, por meio de ramos e passando por vários nós de decisão (nós internos) até chegar aos nós folha (ABELLAN; LOPEZ; de OÑA, 2013).

Em uma Árvore de Decisão, cada nó representa uma característica e cada ramo representa um dos estados dessa variável. Sendo assim, os três tipos de nós podem ser descritos da seguinte forma:

- Um nó raiz que não tem arestas de entrada e zero ou mais arestas de saída.
- Nós internos, cada um com exatamente uma aresta de entrada e duas ou mais arestas de saída.
- Folha ou nós terminais, cada um dos quais tem exatamente uma aresta de entrada e nenhuma borda de saída.

Os nós não terminais, que incluem a raiz e outros nós internos, contêm condições de teste de atributo para separar registros com características diferentes. Cada nó folha é criado com o valor de classe provável para a partição do conjunto de dados definido com a configuração dada pelo caminho do nó raiz até aquele nó folha (TAN; STEINBACH; KUMAR, 2013).

### 2.2.2 Floresta aleatória

Apesar da simplicidade das árvores de decisão e da facilidade de entender os resultados, seu principal problema é o sobreajuste (*overfitting*) que consiste no modelo se ajustar tão bem ao conjunto de dados de treinamento e acabar se tornando ineficaz para prever novos resultados, levando a uma perda de generalização. Para contornar esse problema, as Florestas aleatórias foram propostas.

Santos, Dias e Amado (2022) definem uma Floresta Aleatória, do inglês *Random Forest* (RF) como grupo de muitas árvores de decisão não correlacionadas que operam como um conjunto para formular uma previsão, as diferentes árvores não são correlacionadas, pois são montadas usando uma abordagem de agregação *bootstrap*. Essa abordagem, também conhecida como *bagging*, faz com que cada árvore de decisão do grupo receba amostras diferentes e aleatórias do conjunto de dados, ocasionando em árvores individuais com resultados diferentes.

Na sequência será gerada uma previsão agregada que alcançará uma melhor precisão do que qualquer uma das árvores individuais.

### 2.2.3 Classificador Bayes Ingênuo

Classificadores Bayesianos são classificadores estatísticos baseados no teorema de Bayes, o qual assume que a presença de uma característica particular em uma classe não está relacionada à presença de qualquer outra característica, essa suposição é chamada de independência condicional de classe, desta forma simplificando a computação envolvida e, nesse sentido, é considerado “ingênuo” (LEUNG, 2007).

O classificador Bayesiano Ingênuo, do inglês *Naive Bayes* (NB), atualiza a crença anterior de um evento com novas informações, o resultado é a probabilidade da classe ocorrer com os novos dados. Para classificar um registro de teste, o classificador ingênuo de Bayes calcula a probabilidade posterior para cada classe:

$$P(classe/atributos) = \frac{P(classe) * P(atributos/classe)}{P(atributos)} \quad (1)$$

Onde:

- P(classe/atributos) : Posterior Probability
- P(class) : Class Prior Probability
- P(features/class) : Likelihood
- P(features) : Predictor Prior Probability

Segundo Tan, Steinbach e Kumar (2013) Classificadores Bayesianos Ingênuos geralmente têm as seguintes características:

- Robustez a ruídos nos dados;
- São capazes de lidar com valores ausentes, pois estes são ignorados na construção do modelo;
- Robustez a atributos irrelevantes, pois se um atributo é irrelevante ele se torna quase uniformemente distribuído;
- A suposição de independência condicional pode prejudicar o desempenho do classificador se existirem atributos correlacionados;

### 2.2.4 K-Vizinhos Mais Próximos

O algoritmo K-Vizinhos Próximos, do inglês, *K-Nearest Neighbours* (KNN), é uma classificação baseada em vizinhos e é um tipo de aprendizado em instância ou aprendizado não generalizante, ou seja, o algoritmo não almeja construir um modelo que generalize o conjunto de dados, simplesmente as instâncias dos dados de treinamento são armazenadas e a computação é adiada até a avaliação das instâncias a serem classificadas (PEDREGOSA *et al.*, 2011).

TAN; STEINBACH; KUMAR definem que um classificador de vizinhos mais próximos representa cada exemplo como um ponto de dados em um espaço *d*-dimensional, onde *d* é o número de atributos. Dado um exemplo de teste, ele será classificado com base nos rótulos de seus *k*-vizinhos mais próximos, onde *k* é um número inteiro escolhido pelo usuário. Caso existam vizinhos com diferentes rótulos, o exemplo de teste será atribuído a classe de dados que tem mais representantes dentro dos vizinhos mais próximos.

O valor mais adequado para *k* varia conforme a base de dados, normalmente valores mais altos suprimem os efeitos de ruídos, mas torna os limites de classificação menos distintos (PEDREGOSA *et al.*, 2011). Como esse algoritmo depende da distância para classificação, podendo ser calculada utilizando a Distância Euclidiana ou a Distância de Pearson, se os atributos representarem unidades físicas diferentes ou estarem em escalas muito diferentes, a normalização dos dados de treinamento pode melhorar drasticamente a precisão do classificador (PIRYONESI; EL-DIRABY, 2020).

### 2.2.5 Rede neural Perceptron multicamadas

O um algoritmo de aprendizado de máquina Perceptron multicamadas, do inglês *Multi-Layer Perceptron* (MLP), é amplamente utilizado em problemas de classificação. O MLP é um tipo de rede neural artificial composta por pelo menos três camadas de neurônios (nós): uma camada de entrada, uma camada oculta e uma camada de saída. Onde cada neurônio recebe entradas ponderadas e passa por uma função de ativação não linear para produzir uma saída.

O algoritmo MLP é treinado usando um processo chamado retro propagação, que ajusta os pesos e os vieses dos neurônios para minimizar a diferença entre as saídas previstas e as saídas reais. O aprendizado ocorre justamente alterando os pesos da conexão após cada dado ser processado, com base na quantidade de erro na saída em comparação com o resultado esperado. A retro propagação é um algoritmo de otimização baseado no cálculo do gradiente, usado para atualizar os pesos e os vieses dos neurônios em todas as camadas da rede.

De acordo com Bishop e Nasrabadi (2006), o MLP é uma técnica poderosa de aprendizado de máquina, pois é capaz de modelar funções altamente não lineares, em contraste com funções lineares, que seguem uma relação linear direta, funções não lineares podem ter formas

mais complexas, curvas, interações não lineares e comportamentos não lineares em geral, ou seja, o MLP é capaz de mapear e capturar padrões não lineares nos dados.

## 2.3 Aprendizado de Máquina Explicável

No contexto de aprendizado de máquina, o conceito de 'caixa preta' refere-se a modelos ou algoritmos que são complexos e opacos em sua operação interna. Esses modelos podem produzir previsões ou decisões com alta precisão, mas sem fornecer uma compreensão clara de como essas previsões são feitas. Um exemplo de caixa preta é o modelo de floresta aleatória, pois o processo de combinação das árvores de decisão que formam o modelo não é trivial, tornando desafiador compreender exatamente cada decisão tomada.

A opacidade das caixas pretas pode ser problemática em alguns casos, especialmente em aplicações sensíveis ou críticas, onde é necessário entender as razões por trás das previsões. Portanto, técnicas de explicabilidade e interpretabilidade, como análise de importância de variáveis, visualização de modelos e explicação de instâncias, são usadas para tentar revelar o funcionamento interno dessas caixas pretas e fornecer informações sobre o processo de tomada de decisão dos algoritmos.

### 2.3.1 Valores Shapley

Os valores de Shapley ou valores SHAP, do inglês *Shapley Additive Explanations*, têm origem na teoria dos jogos cooperativos e têm sido aplicados em diversos domínios, incluindo a área de aprendizado de máquina. Neste contexto, os valores SHAP oferecem uma maneira de explicar as contribuições de cada característica ou variável individual para as predições feitas pelo modelo.

Essencialmente, os valores Shapley medem a contribuição marginal de cada característica, considerando todas as possíveis permutações das características e suas predições resultantes. Eles fornecem uma forma justa e consistente de distribuir o crédito ou a importância entre as características (LUNDBERG; LEE, 2017). Ao calcular os valores de Shapley, podemos determinar o impacto de cada característica na saída do modelo, permitindo entender o processo de tomada de decisão do modelo.

Os valores de Shapley oferecem interpretabilidade e transparência em modelos de aprendizado de máquina. Eles permitem quantificar a influência de cada característica nas predições do modelo, ajudando a identificar quais características são as mais influentes na condução dos resultados.

### 2.3.2 Importância por Permutação

A importância por permutação funciona medindo a diminuição no desempenho do modelo quando os valores de um determinado atributo são permutados aleatoriamente. Ao comparar o desempenho original do modelo com o desempenho após a permutação do atributo, podemos determinar o impacto desse atributo nas previsões do modelo. Os resultados das importâncias por permutação fornecem uma medida quantitativa da importância dos atributos. Pontuações mais altas indicam que a permutação do atributo teve um impacto negativo maior no desempenho do modelo, sugerindo que o atributo é mais importante para fazer previsões precisas.

A importância por permutação é considerada uma técnica que explica modelos de aprendizado de máquina, pois avalia diretamente o impacto de cada atributo nas previsões do modelo de maneira transparente. Não depende de detalhes complexos do modelo ou de algoritmos caixa-preta. Em vez disso, fornece uma medida clara e intuitiva da importância dos atributos que pode ser facilmente compreendida e interpretada por humanos.

## 2.4 Análise de acidentes utilizando mineração de dados

A popularização e o crescimento da mineração de dados, como área de estudo, facilitou encontrar padrões e relacionamentos não triviais em um conjunto de dados, o que antes era difícil de alcançar usando técnicas estatísticas tradicionais (YAP *et al.*, 2022). Diferente de confirmar a hipótese existente utilizando análise estatística baseada em evidências, os métodos de mineração de dados estão sendo aplicados em pesquisas de segurança no trânsito para gerar novas ideias e hipóteses não observadas e possivelmente específicas do conjunto de dados (RAIHAN; HOSSAIN; HASAN, 2017).

A natureza discreta da maioria das variáveis presentes em um conjunto de dados e principalmente a heterogeneidade dos dados são, segundo Kumar e Toshniwal (2015), uma das dificuldades encontradas ao se analisar acidentes rodoviários e de trânsito. Se essa heterogeneidade não for considerada alguma relação própria do conjunto pode permanecer oculta, posto isso, um dos desafios é segmentar de maneira adequada o conjunto levando a criação de grupos homogêneos de acidentes de trânsito.

O trabalho de Yap *et al.* (2022) pretende identificar e categorizar os aspectos dos acidentes de trânsito com base nas características dos fatores de risco e classificá-los dependendo do nível de gravidade da lesão de um acidente. Para isso aplicou-se o método de classificação por Árvore de Decisão, ideal para prever uma variável ordinal, neste caso o nível de gravidade da lesão (sem ferimentos, ferimentos leves, graves e morte).

A base de dados utilizada é composta por registros de acidentes de um estado nos Estados Unidos de 2004 a 2018. Para a construção deste modelo realizaram-se algumas etapas prévias, a primeira foi o entendimento da base de dados seguida pela limpeza e padronização

dos dados, na sequência a escolha das variáveis mais relevantes e a discretização delas se necessário.

Como critério de divisão dos nós da árvore de decisão e como métrica de pureza dos nós folha (menor o grau de incerteza das divisões) utilizou-se o coeficiente de Gini. E para a geração da árvore ótima empregou-se o processo conhecido como poda, que consiste na remoção de nós folha insignificantes de uma árvore máxima gerada anteriormente, basicamente o conjunto de dados foi separado em dois novos conjuntos, em dados de treinamento, que servem para modelar a árvore de decisão e dados de validação, que servem para avaliar o desempenho do modelo medindo a taxa de erros de classificação.

O modelo de árvore de decisão gerado neste trabalho implicou que a variável primária para dividir e prever a gravidade da lesão em um acidente é que se veículos do tipo motocicleta estivessem presentes no acidente resultando na possibilidade de acidentes com lesões mais graves, também descobriu-se que caso um pedestre estivesse envolvido no acidente existia uma maior probabilidade de lesão em comparação a não ocorrer uma lesão. É indicado que o modelo não foi capaz de gerar uma regra que previsse uma lesão grave ou fatalidade.

Os resultados revelaram que além do fator humano e fator do veículo os fatores de estrada e tipos de acidente são contribuintes para o nível de gravidade da lesão do acidente, no entanto fatores geográficos e do ambiente se mostraram menos significativos. Por outro lado, a precisão do modelo estimada em 65,78% não foi considerada satisfatória devido à natureza assimétrica da distribuição dos registros.

Já no trabalho de Labib *et al.* (2019), é possível identificar alguns pontos diferentes. Este pretende analisar e determinar a gravidade dos acidentes com base em técnicas avançadas de aprendizado de máquina. Para tal, foram utilizados Árvore de decisão, K-Vizinhos Mais Próximos, Classificador de Bayes e estímulo adaptativo.

Os autores classificam o grau de severidade em categorias, dos quase 43 mil acidentes de trânsito de Bangladesh ocorridos no período de 2001 a 2015. Primeiramente, este conjunto de dados foi divido em 30% para treino, e então, foram organizados os atributos que seriam utilizados para montar os diferentes cenários.

Foi então realizado alguns experimentos pelos autores, determinando o desempenho de cada algoritmo na capacidade de prever um acidente em quatro possíveis classes (Fatal, Grave, Lesão Simples e Colisão Motora). Os algoritmos, Bayes Ingênuo e estímulo adaptativo foram os que apresentaram uma maior acurácia, cerca de 80%. Em um segundo experimento, foi então utilizado somente as classes de acidente Fatal e Grave para a classificação. Neste, foi identificado um aumento da precisão dos algoritmos de K-Vizinhos Mais Próximos e Árvore de Decisão, e os demais obtiveram resultados muito próximos.

As variáveis do tipo de veículo e o horário dos acidentes, foram as mais importantes na capacidade preditiva dos modelos. Comprovou-se uma certa precisão e funcionalidade na previsão dos acidentes de trânsito nas abordagens propostas. Portanto, os autores recomendaram utilizar estas abordagens em um sistema para gerar alertas de situações perigosas na estrada.

Kwon, Rhee e Yoon (2015) propõem o uso de um classificador Bayes Ingênuo e de uma Árvore de Decisão para identificar a importância relativa entre os fatores de risco de um acidente em relação ao nível de gravidade das lesões. Para avaliar o desempenho de cada modelo utilizaram-se as métricas de precisão (fração de instâncias recuperadas que são relevantes) e de *recall* (fração de instâncias relevantes recuperadas) e mediu-se que a árvore de decisão obteve um desempenho superior que o Bayes Ingênuo. Uma hipótese que justificaria esse resultado é de que o algoritmo Ingênuo Bayesiano não considera a dependência entres os diversos fatores, enquanto a Árvore de Decisão considera.

A base de dados utilizada é composta por dados de acidentes coletados pela Patrulha Rodoviária da Califórnia, focando nos relatórios de acidentes ocorridos nas rodovias da Califórnia durante o período de 2004 a 2010, pois apenas em 2004 que começaram a registrar atributos relacionados as as características dos veículos envolvidos, o tipo de rodovia, a data e hora do acidente, as condições climáticas e o tipo de acidente. Como resultado descobriu-se que no conjunto de dados em estudo, quando a dependência é levada em consideração os fatores mais importantes são: tipo de colisão, falha, população do local, rodovia estadual e movimento anterior à colisão.

No estudo de Zhang *et al.* (2018) compara-se o desempenho de quatro diferentes modelos de aprendizado de máquina e de dois modelos estatísticos em categorizar corretamente o nível de gravidade de lesões em acidentes. A partir de um conjunto de dados composto por acidentes ocorridos em rodovias divergentes no estado da Flórida, nos EUA, os seguintes algoritmos de classificação foram utilizados: K-vizinhos mais próximos, Árvore de Decisão, Floresta Aleatória e máquina de vetor de suporte.

Como resultado, obteve-se que o modelo gerado pelo método de Floresta Aleatória conseguiu a maior precisão geral de previsão no conjunto de teste (53,9%), e todos os outros modelos de aprendizado de máquina foram mais precisos que os métodos estatísticos (Modelo Probit Ordenado e o Modelo Logit Multinomial). Observa-se uma menor capacidade dos modelos em prever com precisão acidentes mais graves, os autores explicam que os modelos de aprendizado de máquina obtiveram um maior desempenho em prever colisões sem lesões ou com lesões leves, pois o número de amostras desse tipo de acidente é muito maior que os acidentes fatais e graves, isto é referido como o problema de classificação multi-classe, o que também explica a precisão geral dos modelos não ser muito alta.

Ahmed *et al.* (2023) fizeram uma avaliação de diferentes modelos de aprendizado de máquina para prever a gravidade de acidentes rodoviários com base em um conjunto de dados de acidentes rodoviários da Nova Zelândia do período de 2016 a 2020. Além disso, foram analisados os resultados previstos e aplicada uma técnica de aprendizado de máquina explicável (XML, do inglês, *Explainable Machine Learning*) para avaliar a importância dos fatores que contribuem para os acidentes.

Para prever acidentes rodoviários com diferentes gravidades de lesão, foram considerados diferentes algoritmos, como *Random Forest* (RF), *Decision Jungle* (DJ), *Adaptive Boosting*

(AdaBoost), *Extreme Gradient Boosting*(XGBoost), *Light Gradient Boosting Machine* (LGBM) e *Categorical Boosting* (CatBoost). Os resultados da comparação mostraram que o modelo RF obteve a melhor classificação, com 81,45% de precisão, e 81,04% de F1-Score.

Em seguida, foi empregada a análise dos valores SHAP como técnica de XML para interpretar o desempenho do modelo RF ao nível global e local. A explicação em nível global fornece a classificação da contribuição das características para a classificação de gravidade, enquanto a explicação ao nível local explora o uso das características no modelo. Além disso, interpretações gráficas dos valores SHAP foram utilizados para investigar a relação e interação das características em relação à previsão da variável-alvo. Com base nos resultados, foi observado que a categoria de estrada e o número de veículos envolvidos em um acidente têm um impacto significativo na gravidade das lesões. As características identificadas como mais relevantes por meio da análise de SHAP foram utilizadas para retreinar os modelos de ML e medir seu desempenho. Os resultados mostraram um aumento de 6%, 5% e 8%, respectivamente, no desempenho dos modelos DJ, AdaBoost e CatBoost.

# 3 METODOLOGIA

Neste capítulo é apresentada a metodologia utilizada para a realização deste trabalho. Aqui são descritas a proposta e os métodos empregados para chegar no resultado obtido.

A proposta deste trabalho de conclusão de curso é realizar uma pesquisa empírica a partir do conjunto de dados de acidentes da PRF. Com a intenção de avaliar a influência dos atributos relacionados a especificação do veículo, fatores externos e do ambiente na severidade dos acidentes. Para isso, diferentes modelos de classificação serão construídos com estes atributos.

Para a realização deste trabalho de conclusão de curso foi realizado uma série de etapas para atingir os objetivos e garantir a qualidade do trabalho descritas na Figura 1 e nas seções subsequentes.

**Figura 1 – Fluxo metodológico do estudo**


**Fonte: Autoria própria (2023).**

## 3.1 Ambientação do tema e definição de objetivos

O primeiro passo foi a busca de dados abertos do governo para a realização da análise. No portal de dados abertos do governo, foram encontrados os dados de acidentes rodoviários fornecidos pela PRF. Com isso, em conjunto com o professor orientador, foi decidido analisar os fatores que podem impactar na severidade dos acidentes.

Após isso, com base em todos os fatores, decidimos assim focar nosso estudo nessas duas categorias de atributos: características do veículo e do ambiente, medindo qual poderia ser o impacto destes na severidade de um acidente.

Para incrementar os atributos dos veículos buscamos adquirir os dados do valor do veículo conforme a tabela FIPE. Para tal encontramos uma API e então fizemos a extração.

Nesta etapa, também realizamos uma série de pesquisas em trabalhos correlatos, buscando tanto análises direcionadas a acidentes em rodovias quanto em análise de técnicas de mineração de dados. Com isso, montamos uma bibliografia com diversos artigos relevantes ao tema para adquirir um conhecimento mais aprofundado.

## 3.2 Análise dos dados

Após contextualizar o tema e identificar as questões problemáticas, procedemos a uma análise exploratória dos dados. Utilizamos a ferramenta Microsoft Power BI para criar visualizações que nos auxiliaram na identificação das transformações necessárias, bem como na seleção dos atributos mais relevantes para cada um dos cenários analisados. Essa abordagem nos permitiu ter uma compreensão mais aprofundada dos dados e embasar nossas escolhas para as etapas subsequentes da pesquisa.

## 3.3 Preparação do conjunto dados

A fim de conduzir as análises propostas neste trabalho, foi imprescindível obter um conjunto de dados limpo. Para alcançar esse objetivo, foram realizadas algumas etapas que serão explicadas a seguir.

### 3.3.1 Tratamento

Inicialmente, a base de dados continha 985220 registros. Para realizar a análise desejada, foi realizada uma etapa de pré-processamento dos dados, na qual foram removidos registros considerados irrelevantes para a análise. Foram excluídos aqueles em que o estado físico da vítima não foi informado e os que apresentaram o tipo de envolvido no acidente como ”Testemunha” ou ”Cavaleiro” (pessoas que utilizavam animais como meio de transporte). Após a

identificação de valores nulos nos atributos de idade, sexo da pessoa, marca e modelo dos veículos, foi realizada a remoção dessas instâncias de dados, resultando em um total de 862.465 registros no conjunto final. O atributo “causa do acidente” é um exemplo de variável excluída da análise. Nesse caso, foi removido porque descreve uma característica que ocorre após o acidente e não necessariamente se relaciona com o ambiente ou com o veículo em si, mas sim com o possível comportamento do motorista responsável pelo acidente.

As variáveis marca e modelo foram as que mais apresentaram erros de digitação e informação incompleta, uma vez que, no conjunto de dados da PRF essas duas informações estavam contidas em uma única coluna no formato "marca/modelo". Diversos registros não estavam como os conformes, com o carácter separador faltando ou inserido na posição incorreta. Diversos casos foram possíveis de corrigir, no entanto casos mais extremos a saída foi atribuir a categoria ”Outros” tanto para a marca quanto para o modelo. Outro problema muito comum, era a digitação incorreta dessas informações, os casos mais comuns foram tratados pois eram fáceis de identificar.

Com as marcas e os modelos tratados, a próxima etapa foi a inserção dos dados do preço dos veículos oriundos da tabela FIPE. No total foi possível atribuir o preço a 557.429 registros, a correspondência foi feita utilizando as informações da marca, modelo e ano de fabricação de cada registro. Os 305.036 registros de veículos restantes tiveram seus preços estimados, a estratégia utilizada foi a de agrupar os registros por tipo de veículo e por ano de fabricação, e assim desses agrupamentos calcular a média de preço. Essa média então seria utilizada para preencher os preços que não haviam sido informados.

### 3.3.2 Codificação de dados (*Data enconding*)

As variáveis de interesse nesse conjunto de dados são, em sua maioria, do tipo categórica, o que pode implicar em algumas dificuldades na análise e modelagem dos dados já que algoritmos de aprendizado de máquina funcionam melhor com variáveis numéricas (XUE *et al.*, 2019). Diferentemente das variáveis numéricas, as variáveis categóricas não possuem uma ordem natural e são expressas em termos de categorias ou rótulos, o que pode limitar a análise e modelagem dos dados. Além disso, dados categóricos frequentemente apresentam desbalanceamento de categoria, o que pode levar a um viés nos resultados. No entanto, é possível aplicar técnicas de pré-processamento de dados, como a codificação de variáveis indicadoras e a codificação de rótulos.

A codificação por rótolos (*label enconding*) atribui cada categoria de uma variável a um número inteiro sequencial. Embora essa abordagem seja simples de implementar, ela pode levar a problemas na análise de dados e na construção de modelos de aprendizado de máquina (KOC; EKMEKCIOGLU; GURGUN, 2021). O *label encoding* atribui um número a cada categoria, criando uma ordem implícita. Isso pode ser problemático para variáveis categóricas que não têm uma relação ordinal intrínseca. Por exemplo, se codificarmos as categorias presentes na

variável condição meteorológica: ”sol”, ”chuva” e ”nublado” como 0, 1 e 2, respectivamente, uma ordem está sendo atribuída a essas categorias, o que pode levar o modelo a interpretar erroneamente que ”nublado” é maior que ”sol” e ”chuva”. Acarretando em resultados incorretos em modelos que assumem uma relação de ordem linear entre os valores. Outro possível problema que esse tipo de codificação pode introduzir é um viés de magnitude nos dados. Como os números atribuídos às categorias são sequenciais, o modelo pode interpretar erroneamente que os números mais altos têm uma importância ou peso maior do que os números mais baixos.

A criação de variáveis indicadoras é uma técnica usada para converter variáveis categóricas em variáveis numéricas, que são mais fáceis de serem manipuladas em algoritmos de mineração de dados. Essa função cria variáveis fictícias (também conhecidas como ”variáveis *dummy*”) para cada categoria da variável categórica original. Cada uma dessas variáveis fictícias é representada por um valor binário, indicando se a categoria está presente ou não em cada registro do conjunto de dados.

Por exemplo, no conjunto de dados em questão temos a variável categórica ”tipo_envolvido” com as categorias ”condutor”, ”passageiro” e ”pedestre”. Ao aplicar esta técnica a essa variável, seriam criadas três novas variáveis indicadoras: ”tipo_envolvido_condutor”, ”tipo_envolvido_passageiro” e ”tipo_envolvido_pedestre”. Para cada registro do conjunto de dados, o valor dessas variáveis *dummy* seria 1 ou 0, dependendo de qual foi tipo de envolvido no acidente.

Para variáveis de natureza cíclica, a codificação pode ser feita com a transformação seno-cosseno, ela é feita calculando-se os valores de seno e cosseno dos ângulos correspondentes às variáveis, que são geralmente representadas em radianos. Esses valores são usados para criar novas representações numéricas contínuas das variáveis cíclicas, permitindo que modelos de aprendizado de máquina capturem as relações e padrões circulares presentes nos dados. Essa transformação preserva a natureza cíclica das variáveis, fornecendo uma representação contínua que pode ser mais adequada para a análise e modelagem de problemas em que a ordem e a periodicidade dos dados são importantes.

Por exemplo, na tratativa da variável horário, os horários foram transformados em minutos (intervalo 0 a 1439), sem a transformação seno/cosseno, horários próximos como 23:59 e 00:00, seriam convertidos para 1439 e 0 respectivamente, e os modelos poderiam entender que existe uma diferença muito grande entre eles, o que não é verdade. Com seno e cosseno a natureza cíclica dos dados e a ordem relativa entre as categorias é preservada.

### 3.3.3 Redução da dimensionalidade

A quantidade de novas variáveis geradas pela aplicação da técnica de criação de variáveis indicadoras será sempre diretamente proporcional ao número de categorias presentes na variável original. Isso pode se tornar uma problemática em variáveis com um grande número de categorias, uma vez que o aumento no número de variáveis pode acarretar fenômeno

conhecido como "maldição da dimensionalidade". A maldição da dimensionalidade refere-se a um aumento na complexidade computacional e na dificuldade em identificar padrões relevantes, pois a densidade dos dados diminui em um espaço de alta dimensão. Isso pode resultar em problemas como sobreajuste (*overfitting*) em modelos de aprendizado de máquina e aumento do erro de generalização. Além disso, a alta dimensionalidade torna a computação mais complexa e aumenta o tempo necessário para os modelos de aprendizado de máquina processarem e analisarem os dados.

A fim de mitigar o problema da alta dimensionalidade que ocorria na transformação de variáveis com muitas categorias em múltiplas variáveis indicadoras, utilizou-se a técnica *Target Encoding*, também conhecido como *Mean Encoding*. Diferente do *Label Encoding*, que atribui um número inteiro único a cada categoria, o *Target Encoding* utiliza informações da variável de destino para atribuir valores numéricos às categorias. O objetivo é capturar a relação entre as categorias e a variável de destino, cada categoria é substituída por um valor estatístico (como a média, a mediana, a proporção) do alvo para aquela categoria específica. No presente trabalho para gerar o valor estatístico utilizamos a estratégia proposta por Micci-Barreca (2001) implementada na biblioteca de programação Scikit-Learn (PEDREGOSA *et al.*, 2011).

### 3.3.4 Escalonamento de dados (*Data scaling*)

Diferentes variáveis podem ter escalas e unidades de medida diferentes, o que pode afetar o desempenho dos modelos de aprendizado de máquina. A normalização e padronização dos valores das variáveis de um conjunto de dados, garantem que os valores estejam em uma escala específica. O escalonamento permite equilibrar as variáveis, evitando que aquelas com valores maiores dominem as com valores menores e possibilitam uma interpretação mais precisa dos coeficientes nos modelos lineares, além disso, a convergência dos algoritmos de aprendizado de máquina pode ser acelerado.

A técnica Min-Max, também conhecida como normalização Min-Max, é amplamente utilizada no escalonamento de dados. Nessa abordagem, os valores originais são transformados para uma escala específica, normalmente entre 0 e 1, preservando a relação relativa dos dados. Ao subtrair o valor mínimo ($X_{\text{min}}$) e dividir pela diferença entre o valor máximo ($X_{\text{max}}$) e mínimo da característica, os dados são ajustados para a faixa desejada. Essa técnica serve para garantir que todas as características tenham a mesma escala, facilitando a comparação e melhorando o desempenho de modelos de aprendizado de máquina. A fórmula do Min-Max é dada por:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{2}$$

Onde $X$ representa o valor de um dado, $X_{\text{min}}$ é o valor mínimo da variável e $X_{\text{max}}$ é o valor máximo da variável. $X_{\text{norm}}$ é o valor normalizado da variável, no intervalo [0,1].

A normalização é fundamental para comparar variáveis, neste trabalho a normalização se fez necessária, por exemplo, como o preço de uma moto e o preço de um caminhão, porque ela permite que essas variáveis sejam colocadas em uma escala comum. Quando lidamos com dados em diferentes escalas, compará-los diretamente pode levar a conclusões equivocadas. No exemplo mencionado, uma moto pode ser considerada cara em relação a outra moto, mas barata em comparação com um caminhão. Se os preços não forem normalizados, não seria possível realizar uma comparação adequada entre esses diferentes tipos de veículos. Então a abordagem escolhida foi a de agrupar os veículos pelo seu tipo e então normalizar o preço de cada agrupamento. Assim, motos e caminhões mais caros, por exemplo, foram atribuídos com valores próximos a 1.

## 3.4 Experimentação

Após as etapas de preparação do conjunto de dados, torna-se viável dar início à experimentação propriamente dita. Nessa fase, são realizados testes e ajustes dos modelos com o objetivo de obter resultados confiáveis e significativos. Durante esse processo, é crucial garantir a aplicação correta dos métodos experimentais, bem como a utilização apropriada das métricas de avaliação para analisar e interpretar os resultados obtidos.

### 3.4.1 Definição de cenários

Com o objetivo de comparar os fatores que podem estar influenciando em uma maior periculosidade de acidentes, foram estabelecidos três cenários distintos nos quais os algoritmos de mineração de dados seriam treinados e testados. Esses cenários foram criados com a intenção de explorar diferentes variáveis e características relacionadas aos acidentes, permitindo uma análise abrangente e identificação de possíveis padrões ou correlações. Essa abordagem multifacetada oferece uma visão mais completa e uma compreensão mais sólida sobre os fatores que contribuem para a gravidade dos acidentes.

O primeiro cenário abrange as variáveis que não são relacionadas aos veículos nem ao ambiente. As análises realizadas neste cenário fornecerão uma visão geral dos acidentes em rodovias no Brasil, e os resultados dos modelos treinados com essas características serão utilizados como base para comparações com os demais cenários. Essa abordagem permitirá estabelecer um ponto de referência fundamental para a avaliação dos demais cenários.

As variáveis utilizadas são as seguintes:

1. Dia da semana
2. Feriado
3. Horário

4. Tipo de envolvido
5. Idade do envolvido
6. Sexo do envolvido

O segundo cenário incrementa o cenário base com as variáveis relacionadas ao ambiente onde ocorreu o acidente, como dados geográficos, características da pista e condições meteorológicas. As análises, realizadas neste cenário, permitirão avaliar a influência desses fatores na previsibilidade dos modelos, identificando se há aumento ou diminuição do poder preditivo diante dessas variáveis adicionais.

As variáveis adicionadas são as seguintes:

1. Fase do dia
2. Condição meteorológica
3. Uso solo
4. Tipo da Pista
5. Traçado da via
6. Região

O terceiro cenário incorpora todas as variáveis do cenário dois e, adicionalmente, inclui as variáveis relacionadas ao veículo envolvido no acidente. Isso permite uma análise mais abrangente, considerando não apenas as características do ambiente, mas também os atributos específicos do veículo em questão. Ao incluir essas informações, seremos capazes de compreender melhor os diferentes fatores que influenciam a gravidade do acidente e suas relações.

As variáveis adicionadas são as seguintes:

1. Tipo do veículo
2. Marca do veículo
3. Ano de fabricação do veículo
4. Preço do veículo

### 3.4.2 Codificação

Com os cenários definidos, os modelos Árvore de Decisão, Floresta Aleatória, Bayes Ingênuo, Perceptron Multicamadas e K-Vizinhos Próximos apresentados no Capítulo 2, foram então construídos e treinados em cada um dos cenários, e este processo é explicado nesta subseção.

Para o treinamento e teste dos algoritmos, utilizou-se a técnica *k-fold*, com $k = 5$. Essa abordagem é utilizada para avaliar a performance de modelos de aprendizado de máquina. Nessa técnica, o conjunto de dados é dividido em *k* partes iguais, chamadas de *folds*. O modelo é treinado *k* vezes, onde em cada iteração, um dos *folds* é utilizado como conjunto de teste e os restantes são utilizados como conjunto de treinamento. Ao final, obtém-se *k* medidas de desempenho, geralmente a média ou a soma, que podem ser usadas para avaliar a performance do modelo. A técnica *k-fold* permite uma avaliação mais robusta e realista do modelo, uma vez que utiliza todos os dados para treinamento e teste, ajudando a reduzir a variância dos resultados e fornecendo uma estimativa mais precisa do desempenho do modelo em dados não vistos.

Para a escolha dos parâmetros ideais para cada classificador, a técnica de Busca Aleatória em Hiperparâmetros foi utilizada. (*RandomizedSearchCV* [1]) Esta é uma abordagem eficiente e automatizada para a busca de hiperparâmetros ideais em modelos de aprendizado de máquina. Ao contrário de uma busca exaustiva que examina todas as combinações possíveis de hiperparâmetros, o *RandomizedSearchCV* realiza uma busca aleatória dentro de um espaço pré-definido de hiperparâmetros. A técnica *RandomizedSearchCV* ajuda a encontrar um conjunto adequado de hiperparâmetros para maximizar o desempenho do modelo, resultando em melhores resultados de previsão e ajuste. Os parâmetros encontrados por está técnica foram utilizados na construção dos modelos e são apresentados no trecho de código em *Python* na Listagem 1.

### 3.4.3 Medição dos resultados

Em cada cenário os cinco algoritmos de aprendizado de máquina, mencionados na seção Técnicas de mineração de dados, foram avaliados baseados na capacidade de prever corretamente se um acidente será grave ou não. Como o conjunto de dados em análise é desbalanceado, ou seja, algumas classes têm um número significativamente maior de exemplos do que a outra, neste caso, 85.68% dos registros são de acidentes leves e apenas 14,31% são de acidentes graves.

Essa disparidade na distribuição das classes pode afetar negativamente o desempenho de modelos de aprendizado de máquina, pois eles podem ficar enviesados em direção à classe dominante, resultando em uma menor capacidade de prever corretamente a classe minoritária. Então utilizar somente a acurácia como métrica de avaliação, não é adequado já que o conjunto de dados possui um desequilíbrio de classes. Para avaliar e comparar os modelos de classificação as seguintes métricas foram utilizadas:

[1] *RandomizedSearchCV* é a nomenclatura utilizada na biblioteca de programação SciKit-Learn (PEDREGOSA *et al.*, 2011).

- Acurácia: Mede a proporção de exemplos classificados corretamente em relação ao total de exemplos. Em outras palavras, a acurácia é a medida de quão precisa é a capacidade do modelo em classificar corretamente os dados. É calculada dividindo o número de exemplos classificados corretamente pelo total de exemplos.

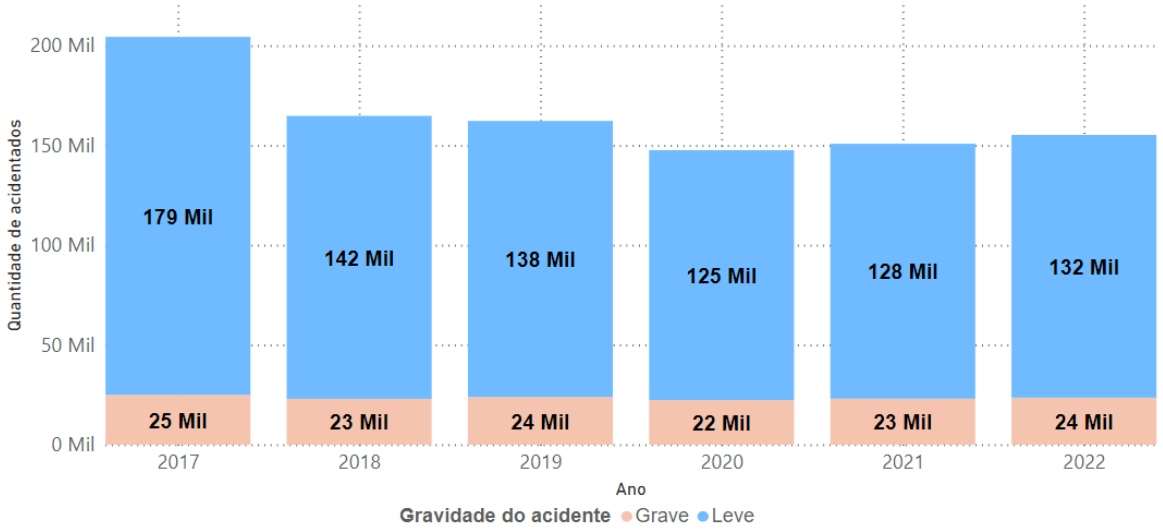
- Média do F1-Score: É a média aritmética dos valores do F1-score para cada classe individualmente. O F1-score é uma medida que combina a precisão (capacidade do modelo de classificar corretamente os exemplos positivos) e o *recall* (capacidade do modelo de identificar corretamente todos os exemplos positivos), essa média considera igualmente o desempenho de todas as classes.

- Média Geométrica (*G-Mean*): O objetivo é capturar o desempenho global do modelo levando em consideração a taxa de verdadeiros positivos de todas as classes de forma equilibrada. É baseada na média geométrica das taxas de verdadeiros positivos (TPR) de cada classe. A média geométrica é calculada multiplicando as TPRs de cada classe e, em seguida, calculando a raiz n-ésima desse produto, onde ”n” é o número de classes. O objetivo é capturar o desempenho global do modelo levando em consideração a taxa de verdadeiros positivos de todas as classes de forma equilibrada.

# 4 CARACTERIZAÇÃO DOS DADOS

O conjunto de dados utilizado neste estudo contém registros de acidentes rodoviários registrados pela polícia em rodovias federais no Brasil desde o ano de 2017 a 2022. Os dados são distribuídos publicamente por meio do portal da Polícia Rodoviária Federal, do ano de 2017 até 2022. Na Figura 2 é possível observar como estes dados estão distribuídos neste período.

**Figura 2 – Distribuição dos acidentados ao longo dos anos**



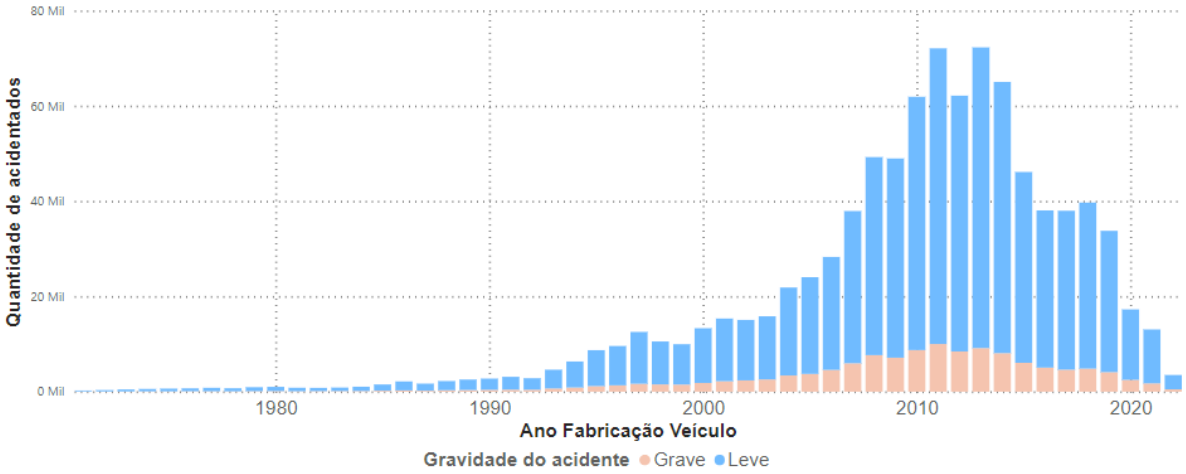

**Fonte: Autoria própria (2023).**

A Figura 3 apresenta a distribuição do ano de fabricação dos veículos e a proporção de acidentes leves e graves para cada ano. É possível observar que a maioria das pessoas acidentadas estavam envolvidas em acidentes com veículos fabricados entre os anos 2007 e 2015. Considerando que a base de dados é referentes aos acidentes ocorridos entre 2017 e 2022, a concentração de veículos mais novos é menor. Como comentado no Capítulo 3, a incidência de veículos anteriores a 1985 é muito baixa, por isso esses veículos foram agrupados na mesma categoria (1985).

Além disso, na Figura 4 é possível visualizar a proporção de pessoas com ferimentos graves entre os envolvidos levando em consideração o ano de fabricação do veículo. Existe uma tendência de queda na proporção de gravemente feridos conforme os veículos são mais novos, a partir de 2005 até 2019. Em 2020 houve um aumento no percentual e a tendência foi quebrada.

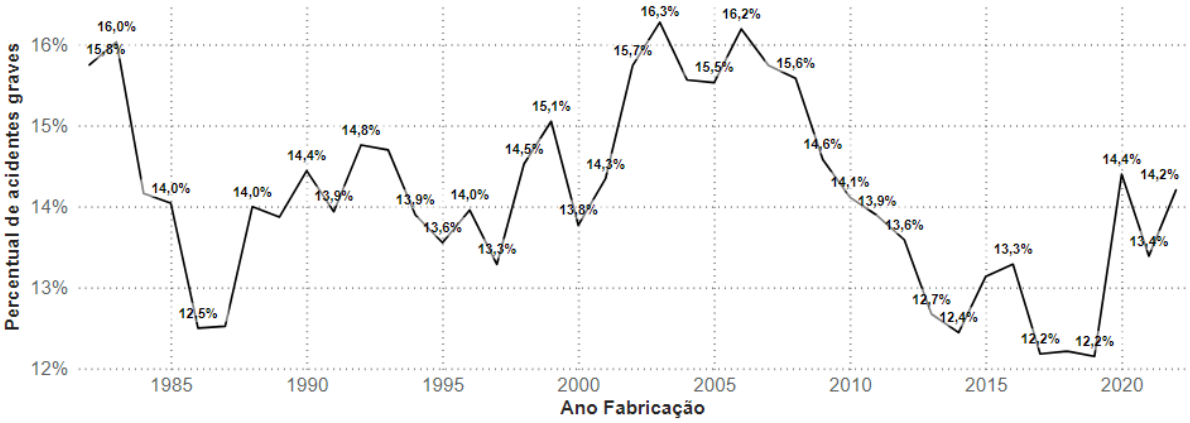
Considerando uma variável relacionado ao local em que o acidente ocorreu, a Figura 5 apresenta como os acidentes estão distribuídos diante do tipo da pista. O tipo de pista com mais acidentes, é a pista simples que também possui a maior proporção de acidentes graves. Já os tipos de pista Dupla e Múltipla, apesar de apresentarem número de ocorrência diferentes, possuem uma próxima proporção de acidentes graves.

**Figura 3 – Distribuição dos acidentados de acordo com o ano de fabricação do veículo**




**Fonte: Autoria própria (2023).**

**Figura 4 – Percentual de acidentes graves de acordo com o ano de fabricação do veículo**

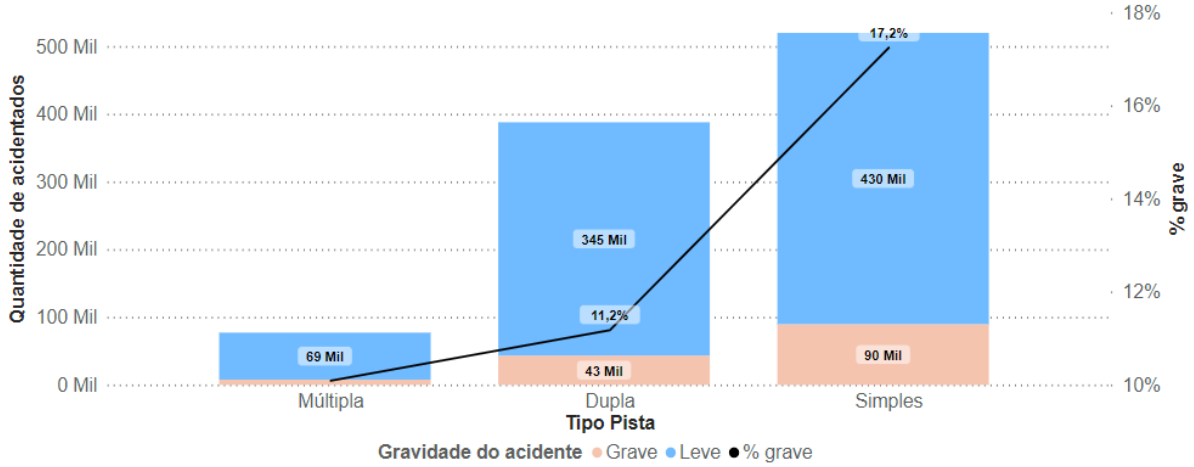



**Fonte: Autoria própria (2023).**

Já o Figura 6 apresenta a distribuição dos acidentados por horário e a porcentagem de gravemente feridos. Existem períodos onde acontecem mais acidentes e períodos onde os acidentes são mais graves. Durante os períodos entre 01:00 e 06:00, observamos uma redução no número total de acidentes, mas um aumento percentual nos acidentes graves. Em particular, o maior aumento no percentual de acidentes graves ocorre entre 01:00 e 05:00. Já durante o período das 08:00 às 17:00, a taxa de acidentes graves é mais baixa.

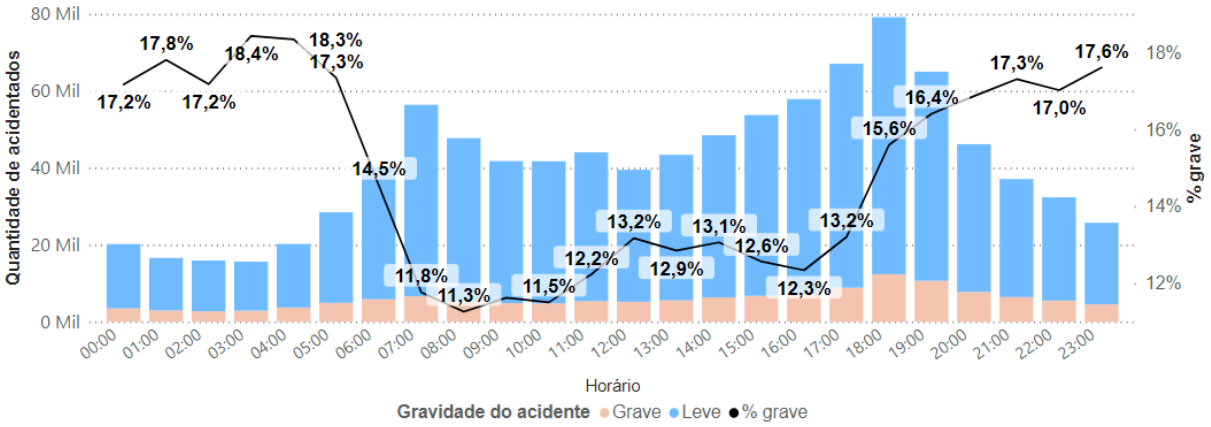
Conforme descrito, o conjunto de dados possuí uma ampla gama de variáveis e de possíveis valores, a partir das técnicas descritas no Capítulo 3 foi realizado o pré-processamento das *features* selecionadas para análise. Na Tabela 1 é descrito quais técnicas foram utilizadas para cada um dos atributos.

**Figura 5 – Distribuição e percentual de acidentes graves de acordo com o tipo da pista**




**Fonte: Autoria própria (2023).**

**Figura 6 – Distribuição da gravidade das lesões de acordo com o horário que o acidente ocorreu**




**Fonte: Autoria própria (2023).**

**Tabela 1 – Técnicas utilizadas na tratativas das variáveis consideradas para análise**

| Variável | Exemplo | Técnica Utilizada | Resultado |
|---|---|---|---|
| | | **Fatores do ambiente** | |
| Uso Solo | Rural | Binarização | 0 Ambiente Rural<br>1 Ambiente Urbano |
| Fase do dia | Pleno dia | Binarização | 0 Dia<br>1 Noite |
| Condição meteorológica | Céu Claro | Condições com menos de 0,05% das ocorrências totais foram desconsideradas<br>Target encoding | [0,1] |
| Tipo de pista | Dupla | Target encoding | [0,1] |
| Região | | Inferida a partir do unidade federativa<br>Target Encoding | [0,1] |

| | | | |
|---|---|---|---|
| Traçado da via | Reta | Target encoding | [0,1] |
| | | **Fatores do veículo** | |
| Marca do veículo | Fiat | Marcas com menos de 500 registros foram agrupadas na categoria Outras<br>Target encoding | [0,1] |
| Tipo do veículo | Automóvel | Tipos de veículo com menos de 0,05% das ocorrências totais foram desconsideradas<br>Target encoding | [0,1] |
| Ano de fabricação do veículo | 2015 | Veículos mais antigos que 1980 foram agrupados<br>Normalização dos valores | [0,1] |
| Preço do veículo | R$ 25.000 | Valores faltantes foram inferidos<br>Normalização dos valores | Preço [0,1]<br>Nova coluna binária indicando se o preço foi inferido |
| | | **Fatores diversos** | |
| Dia semana | Terça | Target Encoding | [0,1] |
| Feriado | Sim | Binarização | 0 não é feriado<br>1 é feriado |
| Horário | 21:10:35 | Transformação horas para minutos<br>Transformação seno e cosseno | 2 novas colunas<br>Horário Seno [0,1]<br>Horário Cosseno [0,1] |
| Tipo de envolvido | Condutor | Remoção de categorias<br>Codificação em variáveis indicadoras | 3 novas colunas<br>Condutor 0 ou 1<br>Passageiro 0 ou 1<br>Pedestre 0 ou 1 |
| Sexo | M | Remoção de categorias<br>Binarização | 0 Feminino<br>1 Masculino |

# 5 RESULTADOS

Nesta seção, é analisado de forma objetiva os resultados encontrados com relação aos objetivos estabelecidos. Para isso, foram desenvolvidas subseções específicas para cada cenário, onde os resultados são descritos e analisados. Além disso, foi adicionado uma seção de discussão para uma análise abrangente dos resultados, com a identificação de relações, padrões e tendências relevantes.

## 5.1 Cenário 1 - Base

Visando de analisar os resultados deste cenário a Tabela 2 apresenta o desempenho dos modelos. Nela é possível observar a confirmação que o conjunto de dados é desbalanceado já que a acurácia média dos modelos é alta (84,2%) e o F1-score médio e a média Geométrica, que consideram o desbalanceamento dos dados, são baixas em comparação a acurácia que não considera o desequilíbrio.

Para este cenário, o modelo que obteve um melhor desempenho segundo o F1-Score e a *G-Mean* foram os baseados em Árvore: Árvore de Decisão e Floresta Aleatória, os demais algoritmos obtiveram resultados muito semelhantes entre si.

**Tabela 2 – Performance dos modelos de classificação no Cenário 1**

| Modelos | Acurácia | Média F1-score | Média Geométrica |
|---|---|---|---|
| Árvore de Decisão (DT) | 82% | 52% | 31% |
| Floresta Aleatória (RF) | 83% | 52% | 30% |
| Bayes Ingênuo (NB) | 86% | 52% | 25% |
| Perceptron Multicamadas (MLP) | 86% | 51% | 24% |
| K-Vizinhos Próximos (KNN) | 86% | 51% | 23% |

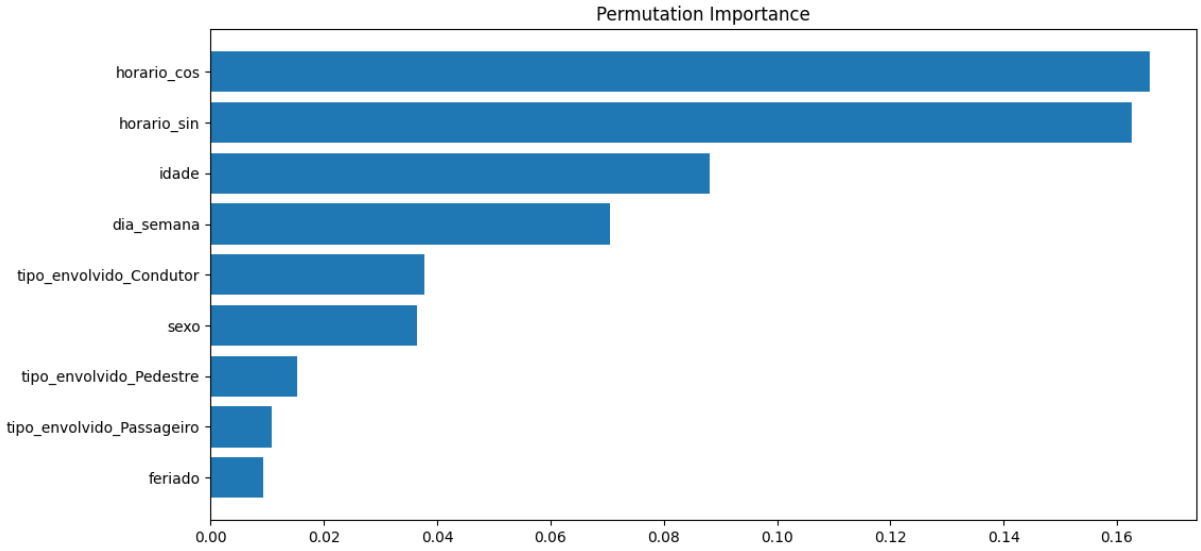
**Fonte: Autoria própria (2023).**

A importância das variáveis por permutação para o cenário 1 está representado no gráfico contido na Figura 7, nele é possível observar que o horário foi a variável mais importante para prever se um acidente será grave ou não. É importante ressaltar que a transformação seno cosseno que foi aplicada na variável horário não é ideal para modelos baseados em árvore pois eles fazem uma divisão com base em um único atributo de cada vez e os atributos de seno/cosseno devem ser considerados simultaneamente para identificar corretamente os pontos no tempo dentro de um período. E como uma informação está sendo representada em duas características, matematicamente mais peso será atribuído a ela do ponto de vista do algoritmo.

No entanto, a importância do horário é muito superior a todas as outras variáveis, indicando que pode existir uma relação entre gravidade do acidente e o horário em que ele ocorre. A segunda variável com mais importância é a idade da pessoa acidentada, e a variável que

menos gerou impacto na performance do modelo é aquela que indica se o acidente aconteceu em um feriado.

**Figura 7 – Gráfico da importância das variáveis por permutação no Cenário 1**




**Fonte: Autoria própria (2023).**

## 5.2 Cenário 2 - Ambiente

Ao analisarmos as métricas de performance contidas na Tabela 3, é possível notar que a Árvore de Decisão se diferencia dos demais modelos, uma vez que a acurácia sofreu uma redução percentual de aproximadamente 7,3%, porém, as métricas que levam em consideração o desbalanceamento das variáveis-alvo aumentaram, principalmente a média geométrica. Isso pode indicar uma pequena melhora na capacidade da árvore de decisão em identificar acidentes graves se comparada ao mesmo modelo no cenário anterior.

Por outro lado, os demais modelos não obtiveram resultados significativos, principalmente considerando a média do F1-score que apenas aumentou em 2% para o modelo de Floresta Aleatória, e para os algoritmos de classificação NB, MLP e KNN não houve uma diferença notável, o que indica que existe uma dificuldade para esses modelos em capturar possíveis relações no conjunto de dados.

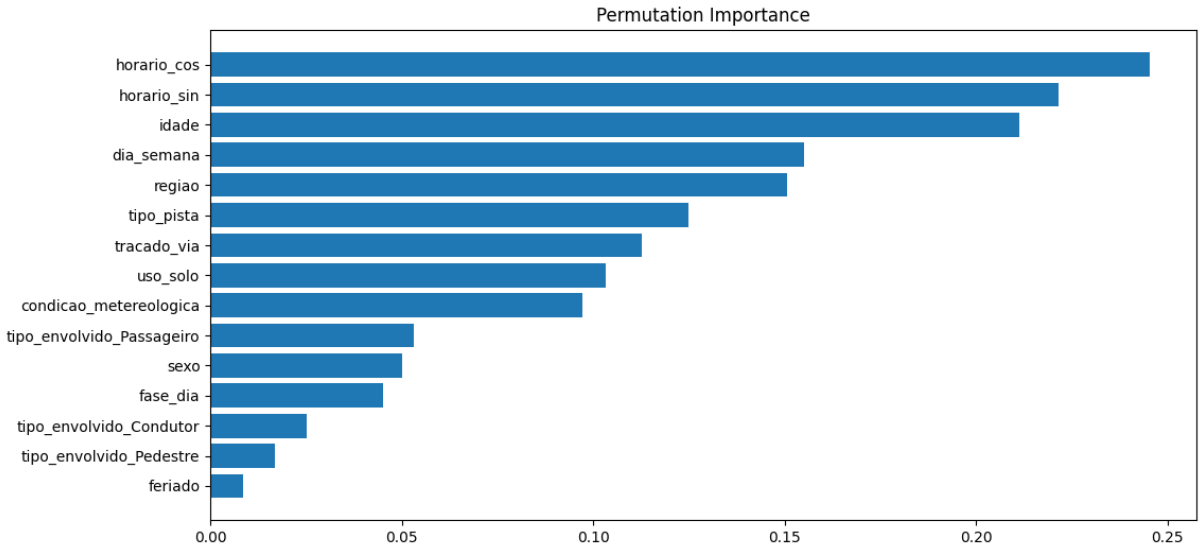
O gráfico contido na Figura 8, indica possíveis motivos para a melhora no desempenho da árvore de decisão. O padrão apresentado no cenário 1, onde as variáveis ”horário”, ”idade” e ”dia da semana” ocupam as primeiras posições em importância, se repetiu neste segundo cenário, no entanto, as 6 novas *features* relacionadas ao ambiente ocuparam as 6 próximas posições mais importantes. Em especial a região e o tipo de pista foram as que mais contribuíram para a capacidade preditiva do modelo, em contrapartida, a variável adicionada neste cenário, que obteve o menor valor de importância foi a "fase do dia".

**Tabela 3 – Performance dos modelos de classificação no Cenário 2**

| Modelos | Acurácia | Média F1-score | Média Geométrica |
|---|---|---|---|
| Árvore de Decisão (DT) | 76% | 54% | 44% |
| Floresta Aleatória (RF) | 85% | 54% | 32% |
| Bayes Ingênuo (NB) | 86% | 52% | 26% |
| Perceptron Multicamadas (MLP) | 86% | 51% | 23% |
| K-Vizinhos Próximos (KNN) | 85% | 52% | 26% |

**Fonte: Autoria própria (2023).**

**Figura 8 – Importância por permutação das variáveis da Árvore de Decisão no Cenário 2**




**Fonte: Autoria própria (2023).**

## 5.3 Cenário 3 - Veículos

A Tabela 4 contém os resultados dos modelos de classificação utilizados para analisar este cenário. É possível verificar que houve um aumento da média F1-Score e da Média Geométrica em todos os modelos, o que pode ser dito que há uma relação entre estes atributos e a gravidade dos acidentes. O modelo Floresta Aleatória e Árvore de Decisão apresentaram um acréscimo considerável na capacidade preditiva, já quanto a média do F1-score o modelo de Bayes Ingênuo apresentou o melhor resultado: 61%, indicando um aumento percentual (variação relativa) de 17.31% em relação aos cenários 1 e 2. Considerando a média Geométrica os modelos DT e NB obtiveram o melhor desempenho.

É possível observar na Figura 9 alguns possíveis motivos que levaram a este aumento, com 3 atributos do veículo dentre os 5 atributos mais influenciadores do conjunto, dando um grande destaque principalmente no tipo de veículo.

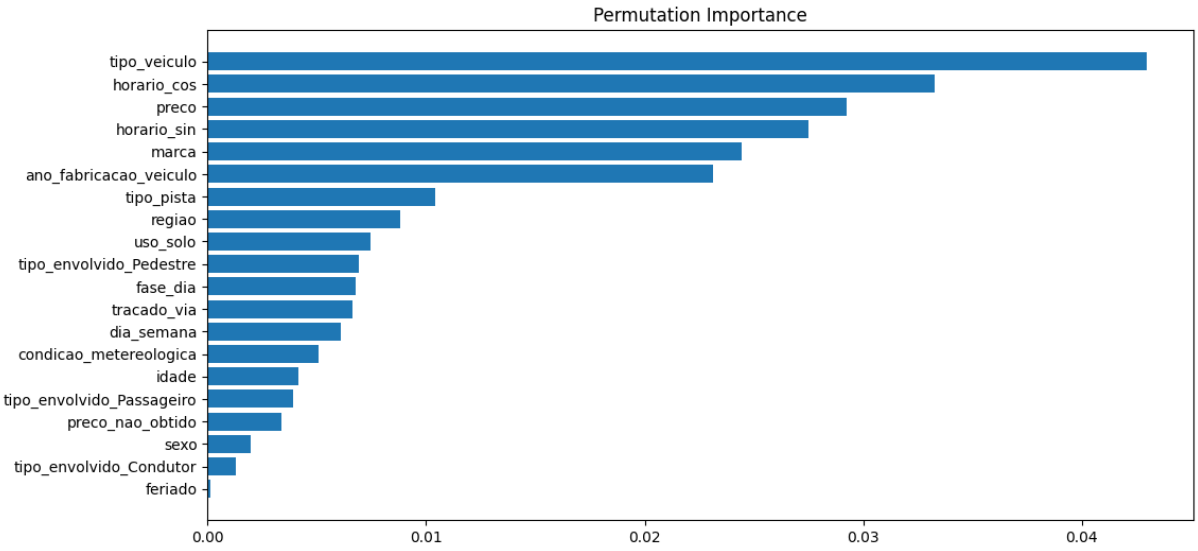
Como explicado por Broeck *et al.* (2021) "o cálculo dos valores Shapley se torna intratável até mesmo para o modelo probabilístico mais simples que não assume independência entre

**Tabela 4 – Performance dos modelos de classificação no Cenário 3**

| Modelos | Acurácia | Média F1-score | Média Geométrica |
|---|---|---|---|
| Árvore de Decisão (DT) | 79% | 59% | 52% |
| Floresta Aleatória (RF) | 86% | 59% | 41% |
| Bayes Ingênuo (NB) | 84% | 61% | 49% |
| Perceptron Multicamadas (MLP) | 86% | 55% | 31% |
| K-Vizinhos Próximos (KNN) | 85% | 57% | 37% |

**Fonte: Autoria própria (2023).**

**Figura 9 – Importância das variáveis por permutação no Cenário 3**



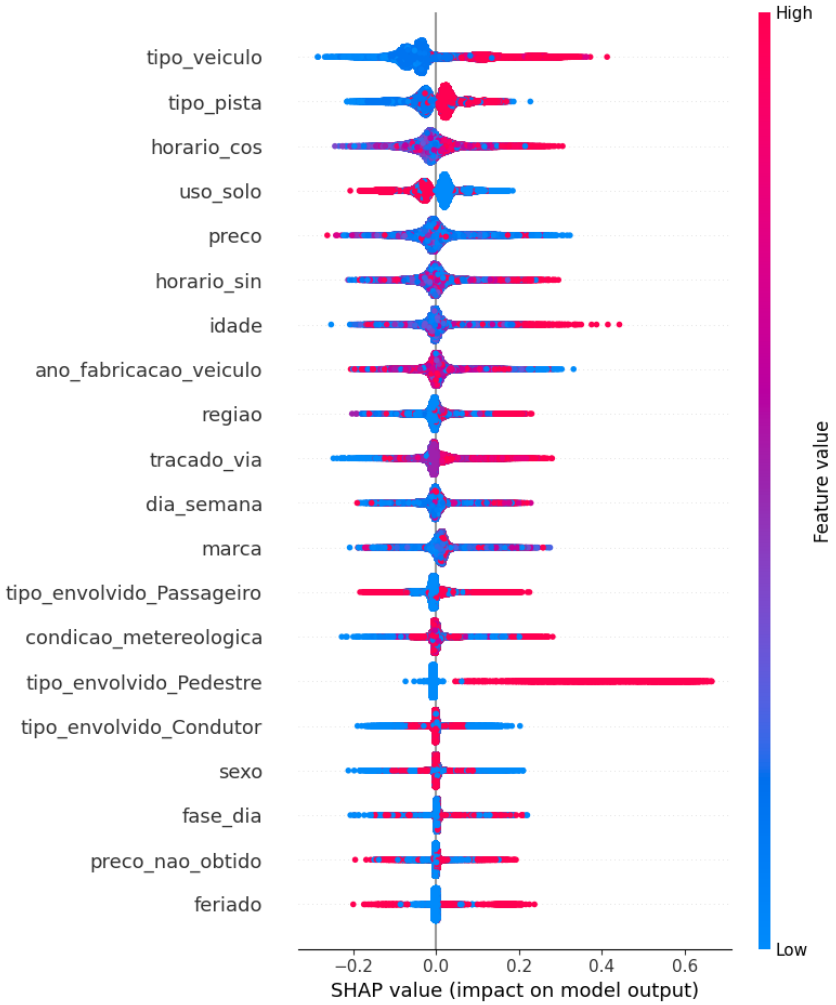

**Fonte: Autoria própria (2023).**

as características: o Bayes Ingênuo". Ou seja, os valores de Shapley são tipicamente utilizados para avaliar as contribuições individuais de características em cenários de teoria dos jogos cooperativos. No caso do Naive Bayes, trata-se de um modelo probabilístico que estima a probabilidade de uma classe dada um conjunto de características. O NB calcula a probabilidade de cada característica de forma independente, assumindo independência condicional, e as combina para fazer previsões. Assim, o conceito de valores de Shapley, que avalia as contribuições conjuntas de variáveis, não é diretamente aplicável, por esse motivo o segundo modelo com as melhores métricas foi escolhido para análise, neste caso, a Árvore de Decisão.

Na Figura 10 é possível verificar como ocorreu a distribuição da importância baseada nos valores SHAP para a Árvore de Decisão. Quanto maior o valor SHAP (mais a direita no eixo X), mais impactante para decidir se um acidente será grave, e o contrário, valores negativos, indicam se um acidente será leve. A ordem das *features* no eixo Y corresponde ao valor absoluto médio dos valores SHAP para cada variável. Esta ordem representa o impacto médio de cada variável em decidir a gravidade de um acidente. Neste caso a variável ”tipo de veículo” foi a mais

impactante na capacidade preditiva do modelo e a variável que indica se um acidente aconteceu em um feriado foi a que menos gerou impacto.

**Figura 10 – Valores Shapley para Árvore de Decisão no Cenário 3**

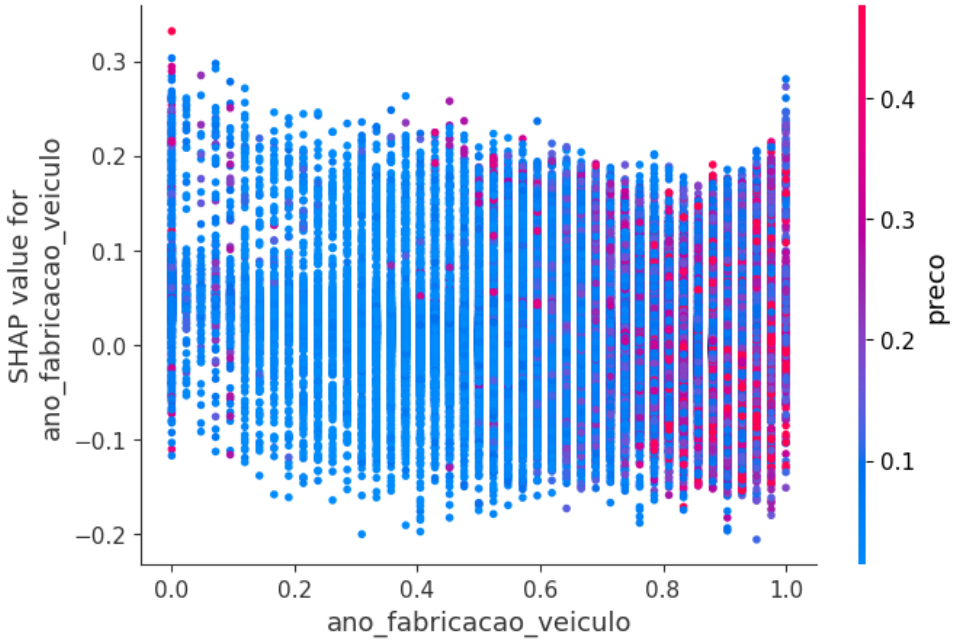



**Fonte: Autoria própria (2023).**

Ao analisar a variável "tipo de envolvido no acidente", observa-se que o modelo possui uma tendência extremamente alta de classificar o acidente como grave quando envolve um pedestre. Como o número de ocorrências com pedestres corresponde a apenas 1,57% do total de acidentados, isso explica por que essa variável não assumiu uma posição de maior importância na análise dos valores SHAP.

O preço e ano de fabricação são as variáveis seguintes na ordem de importância e que são relacionadas aos veículos. No gráfico de dispersão representado na Figura 11, é possível visualizar o efeito da variável ”ano de fabricação” nas previsões feitas pelo modelo. É um conhecimento geral que veículos mais recentes geralmente possuem um valor monetário mais alto, e essa tendência é evidenciada na Figura 11. Verifica-se também que a medida que o ano de fabricação dos veículos diminui, observa-se uma sutil redução nos valores SHAP, o que sugere que veículos mais novos possuem uma contribuição menor para a previsão de acidentes graves e uma maior para acidentes leves.

**Figura 11 – Análise de dependência do ano de fabricação do veículo utilizando SHAP**




**Fonte: Autoria própria (2023).**

Na Figura 10, também é possível observar uma concentração de instâncias de dados com o ano de fabricação do veículo menor, apresentando valores SHAP acima de 0,2. Isso sugere que os veículos antigos estão desempenhando um papel significativo na previsão de acidentes graves. Por outro lado, os veículos fabricados recentemente estão mais distribuídos no intervalo de valores SHAP entre -0,2 e 0,2.

## 5.4 Discussão

A análise por cenários que incrementam variáveis, realizada neste trabalho, permitiu uma compreensão mais detalhada do impacto de cada conjunto de fatores adicionados separadamente, fornecendo perspectivas sobre o impacto dessas características no desempenho dos modelos e permitindo identificar a relevância de cada fator para a tarefa de previsão. É importante ressaltar, que ao dividir a análise em cenários separados, pode-se perder informações sobre possíveis interações entre as variáveis, por esse motivo os cenários foram construídos de forma incremental e sem a remoção de variáveis.

Nos três cenários avaliados, observou-se um desbalanceamento no conjunto de dados, uma vez que a acurácia média dos modelos foi de 84,07%, enquanto o F1-score médio e a média geométrica, que levam em consideração o desbalanceamento, apresentaram resultados mais baixos. Os algoritmos de Árvore de Decisão e Floresta Aleatória demonstraram melhor desempenho em termos de F1-score e média geométrica, enquanto os demais algoritmos tiveram resultados semelhantes entre si, com exceção do Bayes Ingênuo, que obteve o melhor desempenho no cenário 3, com um F1-score médio de 61%. Todos os modelos foram avaliados

**Tabela 5 – Desvio padrão do F1-score dos modelos nos três cenários**

| Cenários | Modelos | | | | |
|---|---|---|---|---|---|
| | Árvore de Decisão | Floresta Aleatória | Bayes Ingênuo | Perceptron Multicamadas | K-Vizinhos Próximos |
| Cenário 1 Desvio Padrão do F1-Score | 0,19% | 0,13% | 0,45% | 0,40% | 0,39% |
| Cenário 2 Desvio Padrão do F1-Score | 0,056% | 0,11% | 0,48% | 0,73% | 0,21% |
| Cenário 3 Desvio Padrão do F1-Score | 0,06% | 0,45% | 0,28% | 0,45% | 0,22% |

**Fonte: Autoria própria (2023).**

**Tabela 6 – F1-Score dos modelos de classificação nos três cenários**

| **Modelos** | **Média F1-score** | | |
|---|---|---|---|
| | **Cenário 1** | **Cenário 2** | **Cenário 3** |
| Árvore de Decisão | 52% | 54% | 59% |
| Floresta Aleátoria | 52% | 54% | 59% |
| Bayes Ingênuo | 52% | 52% | 61% |
| Perceptron Multi Camadas | 51% | 51% | 55% |
| K-Vizinhos Próximos | 51% | 52% | 57% |

**Fonte: Autoria própria (2023).**

utilizando *k-fold* e mostraram consistência e estabilidade nos diferentes *folds*, apresentando um desvio padrão da média do F1-score inferior a 0,5% em todos os cenários, como indicado na Tabela 5.

A análise por cenários realizada neste trabalho e a Tabela 6 indicam que as variáveis relacionadas ao ambiente contribuem de forma pouco significativa para a capacidade preditiva, quando comparadas às características dos veículos adicionadas no Cenário 3. Nesse terceiro cenário, todos os modelos apresentaram melhor desempenho, indicando uma relação entre esses atributos e a gravidade dos acidentes. O classificador Bayes Ingênuo obteve um aumento significativo no F1-score, com um incremento percentual de 17,31% em relação ao Cenário 1 e 2, enquanto a Floresta Aleatória e a Árvore de Decisão registraram um aumento relativo no F1-score de 13,46% em comparação com o cenário 1 e 9,26% em relação ao cenário 2.

A Tabela 6 também demonstra que os classificadores K-Vizinhos Próximos e Perceptron Multicamadas não conseguiram capturar nenhuma importância das variáveis do ambiente adicionadas no Cenário 2, já que a capacidade preditiva se manteve basicamente inalterada em relação ao primeiro cenário. Somente com a adição dos fatores dos veículos que a melhora nos resultados de classificação foi alcançada, no entanto, o KNN e o MLP não desempenharam melhor que os modelos baseado em árvore (DT e RF) e o classificador probabilístico de Bayes.

Ao analisar e comparar os resultados dos diferentes cenários, é possível identificar quais fatores são mais relevantes para a tarefa de previsão. A variável ”tipo de veículo” foi identificada como a principal para classificar a gravidade de um acidente. O tipo de veículo mais importante

para a capacidade preditiva da DT, é o tipo ”Motocicleta”, esse resultado não é exclusivo do conjunto de dados em estudo, pois esse padrão também foi ser observado na análise feita no trabalho de Yap *et al.* (2022) que analisou uma base de dados de acidentes dos Estados Unidos de 2004 a 2018.

As visualizações geradas a partir da importância das variáveis, indicaram uma grande relevância do fator horário. Uma possível explicação pode ser observada na Figura 6, onde são apresentados os horários com um percentual mais alto de acidentes graves em comparação aos demais. Durante o período diurno, em geral, a taxa de acidentes graves é menor, enquanto durante a noite essa taxa aumenta. Esse efeito também pode ser observado na distribuição das instâncias na *feature* fase do dia na Figura 10. Neste caso, os pontos vermelhos representam a fase do dia como noite e os azuis dia. Nota-se um predomínio de valores altos (vermelhos), principalmente na parte positiva do eixo X, indicando um impacto significativo na previsão de acidentes graves.

É evidente, a partir da Figura 10, que as variáveis ”tipo de pista” e ”uso do solo”, desempenham um papel de alta importância na capacidade classificatória da Árvore de Decisão no Cenário 3. No entanto, analisando as métricas do Cenário 2 e o gráfico de importância elas não são tão impactantes, este aumento ocorreu somente quando adicionado os atributos relacionados aos veículos. Dessa forma é possível identificar relação de dependência entre essas variáveis e aos demais fatores do terceiro cenário, pois foi nessa etapa que a capacidade preditiva dos modelos aumentou de forma considerável.

Como evidenciado anteriormente, para o fator ”uso solo”, existe uma tendência preditiva em classificar acidentes graves se eles acontecem em um ambiente não urbano. Essa tendência também foi observada no estudo que analisou uma base de dados de acidentes ocorridos na Nova Zelândia. No Brasil, considerando que as rodovias que atravessam áreas urbanas são submetidas a mais restrições de segurança do que as rodovias não urbanas (CARMO; JUNIOR, 2019), como limites de velocidade e sinalização mais frequente, é possível que essa seja uma das razões que explicam a tendência observada na Árvore de Decisão.

# 6 CONCLUSÃO

A crescente utilização dos automóveis como principal meio de locomoção na vida cotidiana dos indivíduos aliada à urgência de reduzir o número de acidentes graves que acometem milhares de famílias diariamente, foram os principais motivos que nos levaram a iniciar este trabalho. Assim como na aplicabilidade de mineração de dados para realizar análises extremamente ricas com dados. A partir disso, decidimos por juntar estas duas áreas e realizar uma análise mais aprofundada em como os atributos do veículo e ambiente estavam contribuindo para a gravidade do acidente.

A análise dos atributos do veículo e ambiente em um conjunto de dados de acidentes é uma ferramenta importante para a identificação de padrões que contribuem para acidentes graves. A compreensão desses fatores pode guiar políticas de prevenção de acidentes de trânsito e melhorias na infraestrutura das estradas.

A proposta do presente trabalho foi de realizar um estudo dos fatores relacionados ao veículo e ambiente na severidade dos acidentes e que pode ser contemplado ao decorrer do estudo, trazendo assim algumas contribuições ao tema proposto.

No decorrer do projeto, foram encontradas algumas dificuldades, sendo a maior delas trabalhar com um conjunto de dados gerado com base em um formulário preenchido com campos livres, o que acarretou alguns dados incoerentes que tiveram que ser extensivamente tratados e até removidos da análise em casos mais extremos.

Além disso, o fato dos dados estarem desbalanceados também afetou na maneira que o nosso trabalho foi desenvolvido. Pois é normal termos uma quantidade de dados muito maior de acidentes leves do que graves, e no desenvolvimento do trabalho nenhum novo registro foi gerado a fim de balancear estes dados. Para contornar tal problema, testamos algumas abordagens para conseguir tratar os cenários da maneira correta e obter resultados coerentes.

Após realizar a limpeza dos dados e preparação do conjunto para as análises dos atributos, foi possível começar a visualizar resultados conforme a proposta do trabalho. No final, foi possível observar que o ambiente possuí uma relevância para os algoritmos de explicabilidade, porém não melhoraram significativamente a capacidade preditiva dos modelos. No entanto, os atributos relacionados ao veículo conseguiram se destacaram em ambas as situações, resultando em 84% de acurácia e média do F1-score de 61% para o classificador Bayes Ingênuo e para árvore de Decisão apesar da acurácia ser de 79% a *G-Mean* atingiu 52% e o F1-score 59%.

A princípio, a melhora nas métricas observada ao longo dos cenários pode não parecer muito expressiva. No entanto, é importante ressaltar que essa melhora ocorreu justamente nas métricas que levam em consideração o desbalanceamento presente no conjunto de dados. Em outras palavras, à medida que novas variáveis foram adicionadas, os modelos foram capazes de capturar padrões relevantes que contribuíram para o aumento da previsão de acidentes graves, que representa a classe minoritária e mais crítica. Embora a prevenção de acidentes como um

todo seja um objetivo importante, direcionar recursos e esforços para a previsão e prevenção de acidentes graves pode trazer benefícios substanciais para a sociedade, incluindo a redução de mortes, a redução de custos tanto públicos como privados e a melhoria geral da segurança rodoviária.

A metodologia utilizada demonstrou que a análise dividida em cenários pode contribuir de forma mais eficaz para o entendimento do impacto de cada fator relacionado a acidentes. A criação de novas características foi uma etapa fundamental para ampliar as explicações pré-existentes das causas de acidentes. Dessa forma, trabalhos futuros podem dar continuidade à análise das ocorrências em rodovias brasileiras, uma vez que há inúmeros atributos relacionados a essas ocorrências que não estavam presentes no conjunto de dados da PRF e podem ser coletados e analisados no futuro. Levando em consideração que a localização geográfica onde o acidente ocorreu possui certa importância, uma possibilidade interessante seria realizar uma análise mais específica das regiões do Brasil e utilizar a Latitude e Longitude do ponto em que o acidente ocorreu, que são atributos presentes na base de dados da PRF. Uma outra possibilidade seria fazer o levantamento de onde estão localizadas as sinalizações de transito ou equipamentos de fiscalização e analisar se existe uma relação com a localidade dos acidentes.

A utilização da técnica de explicabilidade de valores SHAP é extremamente valiosa e oferece uma explicação visual das relações na capacidade preditiva do modelo. No entanto, neste trabalho, restringimos o seu uso apenas à árvore de Decisão devido ao alto custo computacional e ao tempo significativo necessário para realizar os cálculos. Trabalhos futuros têm a possibilidade de ampliar o uso dessa técnica para explicar outros modelos de aprendizado de máquina. Dessa forma, seria possível obter *insights* abrangentes e interpretáveis sobre a contribuição de cada variável em diferentes algoritmos de predição. Isso permitiria uma compreensão mais aprofundada e robusta dos fatores que influenciam a capacidade de previsão desses modelos. A utilização de técnicas de mineração de dados mais atuais, como Redes Neurais Profundas e Algoritmos Evolucionários pode ser uma abordagem interessante para analisar os acidentes nas rodovias.

# REFERÊNCIAS


ABELLAN, J.; LOPEZ, G.; de OÑA, J. Analysis of traffic accident severity using decision rules via decision trees. **Expert Systems with Applications**, v. 40, n. 15, p. 6047–6054, 2013. ISSN 0957-4174. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417413003138.

AHMED, S. *et al.* A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. **Transportation Research Interdisciplinary Perspectives**, v. 19, p. 100814, 2023. ISSN 2590-1982. Disponível em: https://www.sciencedirect.com/science/article/pii/S2590198223000611.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [*S.l.*]: Springer, 2006. v. 4.

BROECK, G. V. d. *et al.* On the tractability of shap explanations. **Journal of Artificial Intelligence Research**, May 2021.

CARMO, C. L. d.; JUNIOR, A. A. R. Segurança em rodovias inseridas em áreas urbanas na região sul do Brasil. **urbe. Revista Brasileira de Gestão Urbana**, Pontifícia Universidade Católica do Paraná, v. 11, p. e20170182, 2019. ISSN 2175-3369. Disponível em: https://doi.org/10.1590/2175-3369.011.e20180182.

HAN, J.; KAMBER, M.; PEI, J. 1 - introduction. *In*: HAN, J.; KAMBER, M.; PEI, J. (Ed.). **Data Mining (Third Edition)**. Third edition. Boston: Morgan Kaufmann, 2012, (The Morgan Kaufmann Series in Data Management Systems). p. 1–38. ISBN 978-0-12-381479-1. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780123814791000010.

IPEA, I. d. P. E. A. Por uma agência nacional de prevenção e investigação de acidentes de transportes. IPEA, 2021. Disponível em: http://dx.doi.org/10.38116/ntdiset81.

KESAVARAJ, G.; SUKUMARAN, S. A study on classification techniques in data mining. *In*: **2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)**. [*S.l.*: *s.n.*], 2013. p. 1–7.

KOC, K.; EKMEKCIOGLU, O.; GURGUN, A. Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. **Automation in Construction**, v. 131, p. 103896, 2021. ISSN 0926-5805. Disponível em: https://www.sciencedirect.com/science/article/pii/S0926580521003472.

KUHN, M.; JOHNSON, K. Data pre-processing. *In*: ____. **Applied Predictive Modeling**. New York, NY: Springer New York, 2013. p. 27–59. ISBN 978-1-4614-6849-3. Disponível em: https://doi.org/10.1007/978-1-4614-6849-3_3.

KUMAR, S.; TOSHNIWAL, D. A data mining framework to analyze road accident data. **Journal of Big Data, Springer ISSN: 2196-1115**, v. 2, 11 2015.

KWON, O. H.; RHEE, W.; YOON, Y. Application of classification algorithms for analysis of road safety risk factor dependencies. **Accident Analysis & Prevention**, v. 75, p. 1–15, 2015. ISSN 0001-4575. Disponível em: https://www.sciencedirect.com/science/article/pii/S0001457514003273.

LABIB, M. F. *et al.* Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. p. 1–5, 2019.

LEUNG, K. M. Naive bayesian classifier. **Polytechnic University Department of Computer Science/Finance and Risk Engineering**, v. 2007, p. 123–156, 2007.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *In*: GUYON, I. *et al.* (Ed.). **Advances in Neural Information Processing Systems 30**. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

MICCI-BARRECA, D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. **SIGKDD Explor. Newsl.**, Association for Computing Machinery, New York, NY, USA, v. 3, n. 1, p. 27–32, jul 2001. ISSN 1931-0145. Disponível em: https://doi.org/10.1145/507533.507538.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PIRYONESI, S. M.; EL-DIRABY, T. E. Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. **Journal of Transportation Engineering, Part B: Pavements**, v. 146, n. 2, p. 04020022, 2020.

RAIHAN, M. A.; HOSSAIN, M.; HASAN, T. Data mining in road crash analysis: the context of developing countries. **International Journal of Injury Control and Safety Promotion**, v. 25, p. 1–12, 05 2017.

RENAEST. **Registro Nacional de Acidentes E Estatísticas de Trânsito**. Secretaria Nacional de Trânsito, 2022. Disponível em: https://www.gov.br/infraestrutura/pt-br/assuntos/transito/arquivos-senatran/docs/renaest.

SANTOS, K.; DIAS, J. P.; AMADO, C. A literature review of machine learning algorithms for crash injury severity prediction. **Journal of Safety Research**, v. 80, p. 254–269, 2022. ISSN 0022-4375. Disponível em: https://www.sciencedirect.com/science/article/pii/S0022437521001584.

SCHOR, T. O automóvel e o desgaste social. **São Paulo em perspectiva**, SciELO Brasil, v. 13, p. 107–116, 1999.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining: Pearson New International Edition PDF eBook**. [*S.l.*]: Pearson Education, 2013. ISBN 9781292038551.

WHO, W. H. O. **Road traffic injuries**. World Health Organization, 2021. Disponível em: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

XUE, P. *et al.* Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms. **Energy**, v. 188, p. 116085, 2019. ISSN 0360-5442. Disponível em: https://www.sciencedirect.com/science/article/pii/S0360544219317803.

YAP, L. *et al.* A data mining approach to analyse crash injury severity level. **Journal of Engineering Science and Technology**, p. 1–14, 02 2022.

ZHANG, J. *et al.* Comparing prediction performance for crash injury severity among various machine learning and statistical methods. **IEEE Access**, v. 6, p. 60079–60087, 2018.

ZHU, X.; DAVIDSON, I. Knowledge discovery and data mining: Challenges and realities. 01 2007.

# APÊNDICES

# APÊNDICE A – Resultados complementares

**Listagem 1 – Parâmetros utilizados para os modelos de classificação**

```
def get_model(md):
    if md == 'NB':
        model = GaussianNB(var_smoothing=2.848035868435799e-06)
    elif md == 'KNN':
        model = KNeighborsClassifier(n_neighbors=9, n_jobs=-1)
    elif md == 'MLP':
        model = MLPClassifier(hidden_layer_sizes=(20, 20, 20),
                              activation='logistic',learning_rate='adaptive')
    elif md == 'DT':
        model = DecisionTreeClassifier()
    elif md == 'RF':
        model = RandomForestClassifier(n_jobs=-1)

    return model
```

**Fonte: Autoria própria (2023).**