



Universidade Federal da Paraíba  
Centro de Ciências Sociais Aplicadas  
Graduação em Ciência de Dados para Negócios

# **Dos Dados ao Alerta: Utilizando Machine Learning para Prever e Explicar a Gravidade de Acidentes em Rodovias Federais Brasileiras**

Elaine Regina Reis Sousa

João Pessoa - PB  
2026

Elaine Regina Reis Sousa

# **Dos Dados ao Alerta: Utilizando Machine Learning para Prever e Explicar a Gravidade de Acidentes em Rodovias Federais Brasileiras**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência de Dados para Negócios do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Ciência de Dados para Negócios.

Orientador: Professor Dr. Jorge Henrique Norões Viana

**Catálogo na publicação**  
**Seção de Catalogação e Classificação**

S725d Reis Sousa, Elaine Regina.

Dos dados ao alerta: utilizando machine learning para prever e explicar a gravidade de acidentes em rodovias federais brasileiras / Elaine Regina Reis Sousa. - João Pessoa, 2026.  
70 f. : il.

Orientação: Jorge Henrique Norões Viana.  
TCC (Graduação) - UFPB/CCSA.

1. Acidentes de trânsito. 2. Aprendizado de máquina.  
3. XGBoost. 4. SHAP. 5. Segurança viária. I. Viana,  
Jorge Henrique Norões. II. Título.

UFPB/CCSA

CDU 330:004.6(043)

Elaine Regina Reis Sousa

# **Dos Dados ao Alerta: Utilizando Machine Learning para Prever e Explicar a Gravidade de Acidentes em Rodovias Federais Brasileiras**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência de Dados para Negócios do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Ciência de Dados para Negócios.

Trabalho aprovado. João Pessoa - PB, 11 de agosto de 2026:

---

**Professor Dr. Jorge Henrique Norões  
Viana**  
Orientador

---

**Prof. Dr. Antônio Vinícios Barros  
Barbosa**  
Examinador

---

**Prof. Dr. Ignácio Tavares de Araújo  
Júnior**  
Examinador

João Pessoa - PB  
2026

# Agradecimentos

Agradeço, primeiramente, a Deus, que, no decorrer desta jornada, me alcançou e me salvou. Pois d'Ele, por Ele e para Ele são todas as coisas.

À minha família, especialmente à minha mãe, Silvana; às minhas avós, Francisca e Vilani; à minha tia Francigleide, carinhosamente chamada de tia Neném; e ao meu avô Gilvan (*in memoriam*), que, mesmo diante das dificuldades, sempre fizeram todo o possível por mim, nunca deixaram que me faltasse amor e me incentivaram a buscar meus objetivos.

Aos meus amigos, conterrâneos e companheiros de apartamento, Deverton e Safira, que estão ao meu lado desde o início desta jornada.

Aos amigos com quem a Universidade me presenteou e que tornaram o caminho mais leve, especialmente Gabriela, Lorena, Natan, Ikaro e Mateus.

Às minhas irmãs em Cristo, Clara e Brenda, pelo companheirismo, pelas orações e pelo apoio ao longo desta caminhada.

Ao meu orientador, professor Jorge Henrique Norões Viana, pela excelência, pela orientação e pelo constante incentivo durante a realização deste trabalho.

A todos que, de alguma forma, fizeram parte desta trajetória, deixo minha sincera gratidão.

# Resumo

Os acidentes de trânsito em rodovias federais representam um desafio complexo para a segurança pública e para a gestão de infraestrutura no Brasil. O presente trabalho tem como objetivo desenvolver uma aplicação web interativa baseada em aprendizado de máquina (*Machine Learning* - ML) para estimar a probabilidade de uma pessoa envolvida em um acidente apresentar um desfecho grave (lesão grave ou óbito) e explicar os fatores associados à predição. A metodologia baseou-se no treinamento de um modelo preditivo utilizando o algoritmo *Extreme Gradient Boosting* (XGBoost), alimentado por dados históricos de acidentes e informações sobre a frota de veículos. Para garantir a transparência do sistema e evitar a opacidade inerente a modelos complexos (efeito "caixa preta"), integrou-se a biblioteca *SHapley Additive exPlanations* (SHAP), que fornece explicabilidade local às predições. A solução foi desenvolvida com uma arquitetura modular, separando a interface web, construída com Streamlit, dos serviços responsáveis pelo carregamento dos dados, processamento das entradas, geração das predições e interpretação dos resultados. Os resultados obtidos demonstram que o painel permite ao usuário inserir características físicas, temporais e ambientais de um trecho rodoviário e, em tempo real, estimar a probabilidade de uma pessoa envolvida em um acidente apresentar lesão grave ou evoluir para óbito, além de exibir um gráfico detalhando a contribuição de cada variável para a predição. Conclui-se que a ferramenta atua como um sistema de suporte à decisão preditivo e interpretável, possuindo potencial prático para auxiliar órgãos como a Polícia Rodoviária Federal (PRF) e o Departamento Nacional de Infraestrutura de Transportes (DNIT) na alocação preventiva de recursos de fiscalização e na priorização de melhorias na malha viária.

**Palavras-chave:** acidentes de trânsito. aprendizado de máquina. XGBoost. SHAP. segurança viária.

# Abstract

Traffic accidents on federal highways represent a complex challenge for public safety and infrastructure management in Brazil. This study aims to develop an interactive web application based on Machine Learning to estimate the probability that a person involved in an accident will experience a severe outcome (serious injury or death) and to explain the factors associated with the prediction. The methodology was based on training a predictive model using the XGBoost algorithm, fed with historical accident data and information on the vehicle fleet. To ensure system transparency and avoid the opacity inherent in complex models (the “black-box” effect), the SHAP (SHapley Additive exPlanations) library was integrated to provide local explainability for the predictions. The solution was developed using a modular architecture, separating the web interface, built with Streamlit, from the services responsible for data loading, input processing, prediction generation, and result interpretation. The results demonstrate that the dashboard allows users to enter physical, temporal, and environmental characteristics of a road segment and, in real time, estimate the probability that a person involved in an accident will sustain a serious injury or die, while also displaying a graph detailing the contribution of each variable to the prediction. It is concluded that the tool functions as a predictive and interpretable decision-support system, with practical potential to assist agencies such as the Brazilian Federal Highway Police (PRF) and the National Department of Transport Infrastructure (DNIT) in the preventive allocation of enforcement resources and the prioritization of improvements to the road network.

**Keywords:** traffic accidents. machine learning. XGBoost. SHAP. road safety.

# Lista de tabelas

|                                                                                                     |    |
|-----------------------------------------------------------------------------------------------------|----|
| Tabela 4.1 – Dicionário de variáveis . . . . .                                                      | 19 |
| Tabela 5.1 – Espaço de busca de hiperparâmetros definido para a otimização dos<br>modelos . . . . . | 28 |
| Tabela 7.1 – Hiperparâmetros otimizados e desempenho final dos modelos . . . . .                    | 49 |
| Tabela 7.2 – Comparação de impacto e posição: XGBoost (SHAP) vs. Modelo<br>Linear (PLS) . . . . .   | 57 |



# Lista de ilustrações

|                                                                                               |    |
|-----------------------------------------------------------------------------------------------|----|
| Figura 5.1 – Representação gráfica da função logística (sigmoide) . . . . .                   | 23 |
| Figura 5.2 – Representação gráfica do processo do GBM. . . . .                                | 25 |
| Figura 6.1 – Matriz de Risco: Volume Total vs. Taxa de Gravidade por Estado. . .              | 32 |
| Figura 6.2 – Matriz de Risco: Volume Total vs. Taxa de Gravidade por Tipo de Veículo. . . . . | 33 |
| Figura 6.3 – Taxa de Acidentados Graves por Fase do Dia. . . . .                              | 35 |
| Figura 6.4 – Risco por Dia da Semana. . . . .                                                 | 36 |
| Figura 6.5 – Risco por Dia da Semana e Fase do Dia. . . . .                                   | 37 |
| Figura 6.6 – Risco por Condição Meteorológica. . . . .                                        | 38 |
| Figura 6.7 – Risco por Uso do Solo. . . . .                                                   | 39 |
| Figura 6.8 – Risco por Condição Meteorológica e Uso do Solo. . . . .                          | 40 |
| Figura 6.9 – Risco por Tipo de Pista. . . . .                                                 | 42 |
| Figura 6.10–Risco por Traçado da Via. . . . .                                                 | 43 |
| Figura 6.11–Risco por Traçado da Via e Tipo de Pista. . . . .                                 | 45 |
| Figura 6.12–Importância Linear das Variáveis (PLS) . . . . .                                  | 47 |
| Figura 7.1 – Análise Global da Influência das Variáveis nas Predições do Modelo . .           | 53 |
| Figura 7.2 – Importância Global das Variáveis (SHAP) . . . . .                                | 55 |
| Figura 8.1 – Arquitetura da Aplicação . . . . .                                               | 60 |
| Figura 8.2 – Página Inicial da Aplicação . . . . .                                            | 61 |
| Figura 8.3 – Dashboard da Aplicação . . . . .                                                 | 62 |
| Figura 8.4 – Formulário da Página de Predição . . . . .                                       | 63 |
| Figura 8.5 – Exemplo de Resultados da Página de Predição . . . . .                            | 64 |

# Sumário

|       |                                                                                                             |    |
|-------|-------------------------------------------------------------------------------------------------------------|----|
| 1     | INTRODUÇÃO . . . . .                                                                                        | 11 |
| 2     | OBJETIVOS . . . . .                                                                                         | 13 |
| 3     | REFERENCIAL TEÓRICO . . . . .                                                                               | 14 |
| 3.1   | Panorama da Segurança Viária e Fatores Agravantes de Sinistros . .                                          | 14 |
| 3.2   | A Evolução da Modelagem Preditiva em Transportes: Dos Métodos<br>Estatísticos ao Machine Learning . . . . . | 15 |
| 3.3   | O Algoritmo XGBoost e sua Aplicação em Rodovias . . . . .                                                   | 16 |
| 3.4   | Inteligência Artificial Explicável (XAI) e o Framework SHAP . . . .                                         | 17 |
| 4     | DADOS . . . . .                                                                                             | 18 |
| 5     | METODOLOGIA . . . . .                                                                                       | 21 |
| 5.1   | Regressão Logística . . . . .                                                                               | 22 |
| 5.2   | XGBoost . . . . .                                                                                           | 23 |
| 5.3   | Validação Cruzada e Otimização de Hiperparâmetros . . . . .                                                 | 26 |
| 5.4   | Shap . . . . .                                                                                              | 28 |
| 5.5   | Coleta e Integração de Dados . . . . .                                                                      | 29 |
| 5.6   | Engenharia de Atributos . . . . .                                                                           | 29 |
| 6     | ANÁLISE DESCRITIVA DOS DADOS: O PANORAMA DA SEGU-<br>RANÇA VIÁRIA NO BRASIL . . . . .                       | 32 |
| 6.1   | Análise da Relação entre Volume de Acidentados e Tipo de Veículo<br>com a Taxa de Gravidade . . . . .       | 32 |
| 6.2   | Acidentados Graves por Variáveis Temporais . . . . .                                                        | 35 |
| 6.3   | Acidentados Graves por Condições Ambientais . . . . .                                                       | 38 |
| 6.4   | Acidentados Graves por Características das Vias . . . . .                                                   | 41 |
| 6.5   | Modelagem Estatística Multivariada por Partial Least Squares (PLS)                                          | 46 |
| 7     | RESULTADOS E DISCUSSÕES . . . . .                                                                           | 49 |
| 7.1   | Avaliação de Desempenho dos Modelos Preditivos . . . . .                                                    | 49 |
| 7.2   | Análise via SHAP: Fatores de Risco . . . . .                                                                | 52 |
| 7.2.1 | Análise de Contraste: Limitações do Modelo de Comparação (PLS)<br>frente ao XGBoost/SHAP . . . . .          | 56 |
| 8     | DESENVOLVIMENTO DA APLICAÇÃO . . . . .                                                                      | 59 |

|     |                                         |    |
|-----|-----------------------------------------|----|
| 8.1 | Arquitetura da Aplicação . . . . .      | 59 |
| 8.2 | Funcionalidades . . . . .               | 60 |
| 8.3 | Disponibilização da Aplicação . . . . . | 64 |
| 9   | CONSIDERAÇÕES FINAIS . . . . .          | 65 |
|     | REFERÊNCIAS . . . . .                   | 68 |

# 1 Introdução

Os acidentes de trânsito representam um dos maiores desafios à saúde pública em escala global. De acordo com a Organização Mundial da Saúde (OMS, 2024), aproximadamente 1,19 milhão de pessoas morrem anualmente em decorrência de sinistros de trânsito, que figuram entre as principais causas de mortalidade prematura no mundo. No Brasil, a situação também é alarmante. Segundo dados da OMS (2026), o país registrou, em 2021, uma taxa estimada de 15,7 mortes no trânsito por 100 mil habitantes.

Além do impacto social e emocional das perdas humanas, os acidentes acarretam custos econômicos significativos para a sociedade. Segundo Carvalho (2020), os custos relacionados a acidentes em rodovias brasileiras somam aproximadamente R\$40 bilhões ao ano, superando em quatro vezes os custos urbanos (estimados em R\$10 bilhões). Despesas com saúde, previdência, seguros e a perda de capacidade produtiva estão entre os principais componentes dessa cifra.

Esse cenário é agravado pelo descompasso entre o crescimento da frota de veículos e o investimento em infraestrutura rodoviária. De acordo com um levantamento da Confederação Nacional do Transporte (CNT, 2018) entre 2008 e 2018, a frota brasileira aumentou mais de 95,6%, enquanto a malha federal pavimentada cresceu apenas 11,3% no mesmo período.

Diante desse contexto, compreender de que forma as características estruturais das rodovias e as condições ambientais influenciam na gravidade de acidentes é fundamental para a elaboração de estratégias preventivas. Amorim et al. (2019) destacam que, embora o tema seja amplamente estudado em outros países, ainda é pouco explorado no Brasil, mesmo havendo uma base de dados oficial extensa disponibilizada pela PRF, o que evidencia a necessidade de aprofundar as pesquisas sobre os fatores de risco relacionados aos trechos rodoviários.

Um dos pilares deste trabalho é a utilização de dados granulares para alimentar modelos de aprendizagem de máquina. Para tal, foram utilizados os microdados da Polícia Rodoviária Federal (PRF) abrangendo o período de 2017 a 2025. Além disso, foi integrada a base de dados do Ministério dos Transportes referente à frota de veículos por município, introduzindo uma variável crucial de exposição ao risco e densidade veicular.

A solução final apresentada neste trabalho utiliza o algoritmo *Extreme Gradient Boosting* (XGBoost), que demonstrou superioridade técnica frente a modelos tradicionais como a Regressão Logística, alcançando uma acurácia global de 85% e um F1-Score de 0,62 para a classe de grave, frente a 67% e 0,40 do modelo linear. Para solucionar o problema da "caixa-preta" comum em modelos de aprendizado de máquina (*Machine Learning - ML*), foi implementada a técnica *SHapley Additive exPlanations* (SHAP), permitindo uma

interpretação clara do impacto de cada variável.

Quantitativamente, a análise revelou que o Tipo de Veículo é o fator de maior relevância global, registrando um impacto absoluto médio de 1,70 nas predições. Sob a ótica do aumento direto do risco, as três variáveis que mais impulsionam a gravidade dos acidentes são o Tipo de Veículo (com impacto positivo médio de +0,359), o Trecho Específico (BR + KM) (+0,126) e a Frota de Veículos (+0,105), evidenciando a dominância dos fatores veiculares, espaciais e socioeconômicos na elevação da severidade viária.

O resultado deste estudo é materializado num Painel de Previsão de Probabilidade, uma interface interativa que permite a gestores e utilizadores simularem cenários e compreenderem os fatores de risco. Assim, este trabalho pretende contribuir para a modernização das políticas de segurança viária no Brasil, fornecendo uma ferramenta de suporte à decisão baseada em evidências de dados e inteligência artificial.

## 2 Objetivos

O objetivo geral deste projeto é desenvolver e validar um modelo preditivo baseado em aprendizagem de máquina para estimar a probabilidade de uma pessoa envolvida em um acidente apresentar um desfecho grave (lesão grave ou óbito), integrando dados históricos da Polícia Rodoviária Federal (PRF) e indicadores de frota municipal, culminando na criação de um painel interativo de suporte à decisão. A fim de atingir o objetivo geral, faz-se necessário dividi-lo nos seguintes objetivos específicos:

- Coletar, integrar e tratar dados abertos da Polícia Rodoviária Federal (PRF) e do Ministério dos Transportes.
- Analisar padrões de ocorrência e gravidade dos acidentes com base em variáveis como tipo de pista, clima, horário, localização, e sazonalidade.
- Modelar a probabilidade de uma pessoa envolvida em um acidente apresentar lesão grave ou óbito.
- Desenvolver um painel de previsão (dashboard) funcional que permita ao usuário inserir características e condições associadas a um possível acidente e obter a probabilidade de uma pessoa envolvida apresentar um desfecho grave.

## 3 Referencial Teórico

### 3.1 Panorama da Segurança Viária e Fatores Agravantes de Sinistros

A compreensão dos determinantes associados à severidade dos sinistros de trânsito constitui um desafio multifatorial de ordem social, econômica e de saúde pública. No cenário nacional, as características de infraestrutura da malha viária exercem um papel preponderante na letalidade das ocorrências. De acordo com a Pesquisa CNT de Rodovias 2015, 86% das rodovias brasileiras avaliadas eram constituídas por pistas simples e de mão dupla. Em 2014, aproximadamente 90% das colisões frontais registradas nas rodovias federais ocorreram nesse tipo de pista. Embora representassem apenas 4% dos sinistros registrados, as colisões frontais foram responsáveis por 33% das mortes, evidenciando sua elevada letalidade (CNT, 2015).

Paradoxalmente, as condições operacionais que sugerem maior segurança aos condutores costumam atuar como catalisadoras de severidade. Ao avaliarem os condicionantes de risco em rodovias federais, Franceschi et al. (2022) demonstraram uma relação proporcional entre a fluidez do tráfego e a severidade dos acidentes. O estudo revela que condições meteorológicas favoráveis, como o tempo firme (ensolarado), apresentam uma chance 32,8% maior de fatalidade do que sob chuva. Além disso, em horários de menor densidade de tráfego, como a madrugada, o risco de um acidente resultar em óbito é 97,2% superior ao registrado durante o dia. Esse comportamento decorre do fato de que a fluidez viária incentiva o excesso de velocidade. Tal dinâmica amplia drasticamente a energia cinética dissipada em um eventual impacto, tornando os incidentes em condições aparentemente 'seguras' substancialmente mais letais.

Dentro do ecossistema da segurança viária, os motociclistas compõem o estrato mais crítico na categoria de Usuários Vulneráveis da Via (*Vulnerable Road Users*)<sup>1</sup>. A inexistência de barreiras de proteção metálica ou de sistemas de absorção de impacto estruturais (como airbags e zonas de deformação programada) implica que a energia do impacto é transferida de forma direta e integral para o corpo do condutor. No âmbito da saúde coletiva, Corgozinho e Montagner (2017) evidenciam a magnitude desse problema ao expor que dados institucionais apontam que 71% dos acidentes envolvendo motocicletas resultam em vítimas com necessidade de hospitalização imediata.

<sup>1</sup> O termo Usuários Vulneráveis da Via (do inglês *Vulnerable Road Users* – VRU) designa os participantes do trânsito que carecem de um habitáculo ou estrutura metálica externa de proteção (como zonas de deformação programada e *airbags*), ficando expostos diretamente à transferência integral da energia cinética em caso de colisões. Essa categoria engloba predominantemente pedestres, ciclistas e motociclistas (OECD, 1998; OMS, 2018)

Quando transposta para a malha rodoviária federal, a vulnerabilidade desse modal é fortemente potencializada pelas velocidades operacionais elevadas. No estudo seminal de Savolainen e Mannering (2007) sobre a severidade de lesões em motociclistas, os modelos probabilísticos demonstraram que o risco de lesões incapacitantes ou fatais cresce de forma geométrica à medida que aumenta a velocidade regulamentada da rodovia, diferenciando os sinistros em rodovias das ocorrências em perímetros urbanos lentos.

Recentemente, a modelagem avançada com técnicas não lineares permitiu identificar as interações complexas que cercam esses usuários. Utilizando uma arquitetura baseada no algoritmo XGBoost acoplado ao framework SHAP para investigar a fundo a severidade de acidentes motociclistas, Wisutwattanasak et al. (2024) mapearam que, em sinistros envolvendo um único veículo (como quedas e saídas de pista), o período da madrugada e fatores de distração do condutor são os principais agravantes. Contudo, em colisões com múltiplos veículos, os gráficos de dependência do SHAP evidenciaram que a proximidade e a interação física com veículos pesados configuram o maior vetor de deslocamento do risco em direção à letalidade.

## 3.2 A Evolução da Modelagem Preditiva em Transportes: Dos Métodos Estatísticos ao Machine Learning

Diante da complexidade intrínseca aos sinistros viários, os modelos lineares e de regressão tradicionais passam a enfrentar limitações severas na captura de interações multidimensionais e dados esparsos. O avanço das tecnologias de análise de dados permite hoje não apenas identificar pontos de risco, mas prever sua ocorrência com base em padrões históricos complexos. Kumar e Toshniwal (2016) empregaram técnicas de agrupamento (k-means) e mineração de regras de associação para caracterizar trechos críticos na Índia, segmentando os locais de acidentes em três categorias de frequência (alta, moderada e baixa) para extrair padrões comportamentais e de infraestrutura. Na vanguarda preditiva, Ren et al. (2018) propuseram modelos de *Deep Learning* para a predição de risco em grandes centros urbanos, demonstrando que redes neurais complexas conseguem capturar padrões espaciais e temporais não lineares de forma mais robusta que modelos estatísticos convencionais.

No contexto nacional, Amorim et al. (2019) estabelece que a aplicação de técnicas de aprendizado de máquina supervisionado e automatizado (AutoML) é uma abordagem eficaz para analisar o impacto de diversas características no risco de acidentes em rodovias federais. A autora ressalta que o estudo de acidentes é altamente dependente da localidade, o que justifica a utilização de bases de dados oficiais e extensas, como os registros da PRF. Em sua metodologia, destaca-se o uso de ferramentas de AutoML para automatizar a seleção de algoritmos e a otimização de hiperparâmetros, permitindo encontrar modelos



que superam os projetados de forma manual.

Os resultados quantitativos obtidos por Amorim et al. (2019) variaram significativamente conforme o cenário experimental adotado. Desconsiderando os cenários baseados em dados desbalanceados (Cenários A e B) por não apresentarem relevância estatística. No Cenário C, que incorporou o atributo estrutural de "Frequência de Acidentes", o modelo de Rede Neural *Multilayer Perceptron* (MLP) obteve o melhor desempenho global, alcançando 85% de acurácia, 87% de precisão, 85% de revocação e 84% de medida F. A autora destaca que o uso desse atributo derivado de frequência atuou como um fator de forte relevância, impulsionando a capacidade da rede neural em discriminar os riscos. Por outro lado, no Cenário D, que operou sem o atributo de frequência, os melhores resultados foram atingidos pelas combinações híbridas dos algoritmos *Random Forest + BernoulliNB* e *Logistic Regression + ExtraTreesClassifier*, ambos registrando marcas sólidas de 84,58% de acurácia, 88,14% de precisão, 84,58% de revocação e 84,06% de medida F.

### 3.3 O Algoritmo XGBoost e sua Aplicação em Rodovias

Expandindo as fronteiras da modelagem de rodovias, Tang et al. (2020) estabelecem que a aplicação do algoritmo XGBoost constitui uma abordagem altamente eficaz para prever o tempo de liberação de incidentes e analisar seus fatores influenciadores. Os autores ressaltam que o gerenciamento eficiente do fluxo de tráfego é dependente da análise profunda de dados históricos, justificando o uso de registros detalhados como os do *Washington Incident Tracking System* (WITS). Na engenharia metodológica de Tang et al. (2020), destaca-se a aplicação prévia do algoritmo K-means para agrupar os dados em clusters com propriedades inerentes similares antes da etapa de modelagem, estratégia que permite capturar a filosofia oculta e a heterogeneidade dos dados de forma mais precisa. Adicionalmente, os autores introduzem a Otimização Bayesiana (BOA) para automatizar o ajuste de hiperparâmetros, conferindo flexibilidade ao modelo para lidar com relações não lineares em espaços de alta dimensão.

Diferente de abordagens em que outros algoritmos encontram melhor desempenho, os experimentos de Tang et al. (2020) confirmaram a superioridade do XGBoost em relação a modelos consolidados como *Support Vector Regression* (SVR), *Random Forest* e *Adaboost*. Ao mensurar o desempenho por meio do Erro Percentual Absoluto Médio (MAPE), o modelo XGBoost alcançou índices altamente competitivos de 0,348 para o cluster de incidentes de curta duração e 0,221 para casos de longa duração. Tais resultados validam a técnica como um algoritmo escalável e de alto desempenho, capaz de superar métodos de aprendizado de máquina convencionais e fornecer previsões confiáveis para o suporte à gestão de rodovias.

Essa eficiência computacional em larga escala e a tolerância a dados esparsos correlacionam-se com a validação empírica seminal de Chen e Guestrin (2016), criadores

do algoritmo, que demonstraram a supremacia do XGBoost ao registrar que ele foi a ferramenta base de 17 das 29 soluções vencedoras da plataforma Kaggle no ano de 2015, além de demonstrar capacidade de processamento de até 1,7 bilhão de exemplos em uma única máquina, consolidando seu papel como o estado da arte para problemas de alta dimensionalidade a partir de um sistema de árvores de decisão impulsionadas por gradiente com regularização estrita.

### 3.4 Inteligência Artificial Explicável (XAI) e o Framework SHAP

Além do desempenho preditivo pura e simplesmente, surge a necessidade crítica de interpretar esses modelos complexos para fins de formulação de políticas e gestão pública, mitigando o problema da "caixa-preta" que frequentemente afasta a Inteligência Artificial da tomada de decisão governamental. Tang et al. (2020) utilizaram o cálculo tradicional de importância relativa das variáveis para identificar os determinantes primários dos incidentes, apontando quantitativamente que o tempo de resposta, o Volume Médio Diário (VMD), o tipo de incidente e o fechamento de faixas de rolamento eram os fatores mais influentes no modelo.

Todavia, a evolução metodológica proposta neste trabalho opta por avançar em relação às métricas de importância relativa locais e heurísticas. Enquanto a importância calculada por pureza de nós ou ganho (como a utilizada de forma nativa em árvores de decisão) fornece uma visão isolada e potencialmente viesada do impacto, uma vez que pode superestimar variáveis contínuas ou apresentar inconsistências graves quando a estrutura da árvore sofre pequenas alterações, a evolução recente da literatura aponta para a utilização do framework SHAP.

Fundamentado matematicamente na teoria dos jogos cooperativos e proposto originalmente por Lundberg e Lee (2017), o SHAP garante estritamente o cumprimento de três propriedades fundamentais: aditividade local, consistência e monotonicidade. Essas garantias teóricas impedem distorções na atribuição de relevância, assegurando que, se um modelo for modificado para que uma variável tenha maior impacto, sua importância calculada nunca diminuirá.

## 4 Dados

Neste trabalho, foram utilizados dados provenientes de duas fontes oficiais: a Polícia Rodoviária Federal (PRF) e o Ministério dos Transportes. A escolha dessas bases se justifica pela sua confiabilidade, abrangência temporal e espacial, além da disponibilidade em formato aberto, o que garante transparência e reprodutibilidade da pesquisa.

Os dados de acidentes foram obtidos no portal de dados abertos da PRF. O dataset selecionado está **agrupado por pessoa envolvida nos sinistros**, ou seja, cada linha corresponde a uma pessoa envolvida nos acidentes, e reúne todas as causas e tipos de acidentes, apresentando informações detalhadas sobre as ocorrências registradas nas rodovias federais brasileiras.

O recorte temporal considerado abrange o período de 2017 a 2025, contemplando 37 atributos, descritos na Tabela 4.1. esse período foi definido pela disponibilidade dos microdados da Polícia Rodoviária Federal (PRF) no formato “Agrupados por pessoa - Todas as causas e tipos de acidentes”. A escolha assegura a homogeneidade da unidade de análise, permitindo o uso de 37 atributos padronizados. Optou-se por não retroagir a períodos anteriores a 2017 para evitar inconsistências estruturais na base de dados, uma vez que a mudança na metodologia de coleta da PRF impossibilitaria uma comparação direta e entre os registros, garantindo assim maior robustez e capacidade de generalização ao modelo preditivo.

Tabela 4.1 – Dicionário de variáveis

| Variável               | Descrição                                                       |
|------------------------|-----------------------------------------------------------------|
| id                     | Identificador numérico do acidente                              |
| pesid                  | Identificador da pessoa envolvida                               |
| data_inversa           | Data do acidente (dd/mm/aaaa)                                   |
| dia_semana             | Dia da semana do acidente                                       |
| horario                | Horário do acidente (hh:mm:ss)                                  |
| uf                     | Unidade da Federação (ex.: MG, PE)                              |
| br                     | Identificador da rodovia federal (BR)                           |
| km                     | Quilômetro do acidente (ex.: 123,4 km)                          |
| municipio              | Nome do município de ocorrência                                 |
| causa_principal        | Indica se a causa foi classificada como principal               |
| causa_acidente         | Causa presumível do acidente (vestígios/provas)                 |
| ordem_tipo_acidente    | Sequência dos eventos sucessivos no acidente                    |
| tipo_acidente          | Tipo de acidente (ex.: colisão frontal, saída de pista)         |
| classificacao_acidente | Gravidade: sem vítimas, com feridos, com óbitos                 |
| fase_dia               | Fase do dia: amanhecer, pleno dia, anoitecer, etc.              |
| sentido_via            | Sentido da via: crescente ou decrescente                        |
| condicao_meteorologica | Condições climáticas (céu claro, chuva, neblina etc.)           |
| tipo_pista             | Tipo da pista (simples, dupla, múltipla)                        |
| tracado_via            | Traçado da via (reta, curva, cruzamento)                        |
| uso_solo               | Característica do local (urbano/rural)                          |
| id_veiculo             | Identificador do veículo                                        |
| tipo_veiculo           | Tipo do veículo (carro, moto, caminhão etc.)                    |
| marca                  | Marca do veículo                                                |
| ano_fabricacao_veiculo | Ano de fabricação do veículo                                    |
| tipo_envolvido         | Participação no acidente (condutor, passageiro, pedestre)       |
| estado_fisico          | Condição do envolvido (ileso, ferido leve, ferido grave, morto) |
| idade                  | Idade do envolvido                                              |
| sexo                   | Sexo do envolvido                                               |
| ileso                  | Identificação binária (1 = ileso, 0 = não)                      |
| feridos_leves          | Identificação binária (1 = sim, 0 = não)                        |
| feridos_graves         | Identificação binária (1 = sim, 0 = não)                        |
| mortos                 | Identificação binária (1 = sim, 0 = não)                        |
| latitude               | Latitude geográfica (graus decimais)                            |
| longitude              | Longitude geográfica (graus decimais)                           |
| regional               | Superintendência regional da PRF responsável                    |
| delegacia              | Delegacia da PRF responsável                                    |
| uop                    | Unidade operacional da PRF responsável                          |

Fonte: <<https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-da-prf>>

Para complementar os registros da PRF, foram utilizados dados sobre a frota de veículos por município entre os anos de 2017 e 2025 disponibilizados pelo Ministério do Transportes. A inclusão dessa variável possibilita a análise da influência da densidade veicular sobre a ocorrência e gravidade dos acidentes. Assim, a integração dessas bases proporciona uma visão abrangente que contempla tanto as características estruturais e

contextuais das rodovias, quanto às condições ambientais, oferecendo subsídios consistentes para a modelagem preditiva proposta.

## 5 Metodologia

A metodologia adotada neste trabalho tem como objetivo o desenvolvimento de uma aplicação voltada à estimar a probabilidade de uma pessoa envolvida em um acidente apresentar um desfecho grave, bem como explicar os fatores associados à predição, utilizando como base os registros anuais de acidentes disponibilizados pela Polícia Rodoviária Federal (PRF).

Para atingir esse propósito, o estudo será conduzido em etapas que compreendem a extração e integração de dados; o pré-processamento e tratamento das bases; a aplicação de técnicas de aprendizagem de máquina supervisionada; e, por fim, a visualização dos resultados em uma aplicação interativa. Este capítulo detalha o percurso metodológico adotado para o desenvolvimento dos modelos preditivos, descrevendo as etapas de preparação, transformação e modelagem dos dados de acidentes rodoviários

O Aprendizado de Máquina é uma subárea da inteligência artificial. Seu principal objetivo é permitir que sistemas computacionais sejam capazes de aprender padrões a partir de dados e utilizá-los para tomar decisões, ou realizar previsões, sem que sejam explicitamente programados para isso. De acordo com Mitchell (1997), um sistema de aprendizado de máquina pode ser considerado eficiente quando melhora seu desempenho em determinada tarefa com base na experiência adquirida. Essa definição, amplamente reconhecida no campo, destaca três elementos fundamentais: a experiência; a tarefa a ser executada; e a medida de desempenho que avalia o progresso do sistema. À medida que o sistema é exposto a mais dados (experiência), ele aprimora sua performance em relação à tarefa específica.

Os padrões descobertos por meio dessas técnicas podem ser expressos de forma estruturada, como conjuntos de regras, árvores de decisão, funções matemáticas ou redes neurais. Em muitas aplicações de mineração de dados, as descrições estruturais compreensíveis são tão importantes quanto a capacidade de prever novos exemplos, pois permitem aos usuários obter conhecimento e *insights* sobre a estrutura dos dados (WITTEN et al., 2017).

Em termos práticos, o aprendizado de máquina pode ser entendido como um conjunto de métodos e algoritmos capazes de encontrar padrões ocultos em grandes volumes de dados, com o objetivo de explicar, prever ou tomar decisões. Ele permite ao sistema aprender a partir de exemplos passados, construindo modelos que possam ser aplicados em novas situações, como prever o risco de acidentes em trechos rodoviários ainda não construídos, com base em padrões históricos de infraestrutura, clima e comportamento humano.

O Aprendizado Supervisionado é uma das abordagens mais utilizadas e investigadas

no campo do Aprendizado de Máquina. Nessa vertente, o algoritmo é treinado a partir de um conjunto de dados que contém tanto as variáveis de entrada (preditores) quanto as variáveis de saída (rótulos), permitindo que aprenda uma função de mapeamento capaz de associar corretamente novas entradas às suas respectivas saídas. O termo “supervisionado” decorre do fato de que o processo de aprendizagem é conduzido por exemplos previamente rotulados (JAMES et al., 2023).

Ainda segundo James et al. (2023), o processo envolve a estimação de uma função desconhecida  $f$  que relaciona as variáveis de entrada às de saída, utilizando para isso um conjunto de treinamento. Após o aprendizado, o modelo é submetido a dados rotulados que não foram utilizados no treinamento (dados de teste ou validação) a fim de avaliar sua capacidade de generalização e desempenho preditivo.

Em contraste, o Aprendizado Não Supervisionado é aplicado quando não há uma classe ou variável de resultado especificada nos dados. O objetivo é descobrir qualquer estrutura que seja "interessante" nos dados, permitindo que o algoritmo encontre padrões e tendências por conta própria. A técnica mais comum é o agrupamento (clustering), onde os itens são agrupados com base em sua semelhança (MATHUR, 2019).

Portanto, considerando a natureza rotulada dos dados empregados neste trabalho, optou-se pela utilização de aprendizagem supervisionada. Essa abordagem possibilita explorar o histórico de ocorrências para treinar modelos capazes de estimar a probabilidade de acidentes em trechos de rodovias, levando em conta variáveis estruturais e climáticas que influenciam a segurança viária. Para isso, o processo de modelagem foi estruturado como um experimento comparativo entre dois paradigmas distintos. A Regressão Logística foi implementada como modelo de base para avaliar a separabilidade linear dos dados. Em contrapartida, o XGBoost foi selecionado pela sua capacidade superior de capturar interações não lineares complexas.

## 5.1 Regressão Logística

A regressão logística é uma ferramenta fundamental de aprendizagem supervisionada amplamente utilizada para prever respostas qualitativas, também conhecidas como variáveis categóricas. Em vez de prever um valor numérico diretamente, a regressão logística modela a probabilidade de uma observação pertencer a uma categoria específica. Essa probabilidade, por sua vez, é usada para classificar a observação em uma das classes (JAMES et al., 2023).

O modelo logístico emprega a função logística (sigmoide) para transformar a combinação linear das variáveis preditoras em uma probabilidade que sempre varia entre 0 e 1 como ilustrado na Figura 4.1. A forma geral da função logística para a probabilidade

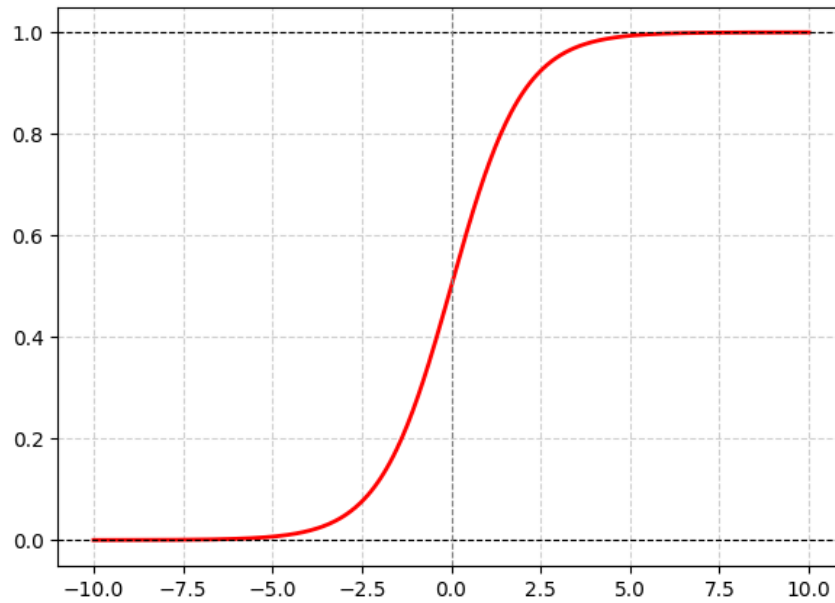
$p(X)$  de um evento ocorrer, dado um conjunto de preditores  $X$ , é :

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}} \quad (5.1)$$

Onde:

- $p(X)$  é a probabilidade de o evento de interesse ocorrer.
- $\beta_0$  é o intercepto.
- $\beta_j$  são os coeficientes de regressão para as variáveis predictoras  $X_j$ .

Figura 5.1 – Representação gráfica da função logística (sigmoide)



Fonte: Elaboração própria (2026).

## 5.2 XGBoost

O XGBoost, introduzido por Chen e Guestrin (2016), é uma implementação otimizada de *Gradient Boosting Machines* (GBM). Este último consiste na construção sequencial de múltiplos modelos de baixo desempenho individual, frequentemente árvores de decisão simples, cujo treinamento iterativo visa corrigir os erros cometidos pelos modelos anteriores. O resultado é um modelo preditivo de alto desempenho e elevada acurácia, a representação gráfica da técnica pode ser visualizada na Figura 5.2. Segundo Friedman (2001), o GBM se destaca pela capacidade de otimizar funções de perda arbitrárias por meio de métodos de descida de gradiente, o que contribui para sua ampla aplicabilidade em diferentes domínios. O método opera de forma iterativa, em que cada novo modelo é construído para corrigir os erros, ou mais precisamente, os pseudo-resíduos, gerados pelos



modelos anteriores. A ideia central consiste em, ao invés de ajustar uma única árvore de decisão complexa que poderia levar ao sobreajuste (*overfitting*), realizar o aprendizado de maneira gradual, aprimorando progressivamente as previsões a cada iteração (JAMES et al., 2023).

Segundo (JAMES et al., 2023), o processo pode ser descrito pelo seguinte algoritmo:

### 1. Inicialização:

Define-se uma função de previsão inicial  $\hat{f}(x)$ , que normalmente é igual a zero para todas as observações. Os resíduos iniciais  $r_i$  são definidos como os valores reais da variável resposta  $y_i$  para cada observação  $i$  no conjunto de treinamento:

$$\hat{f}(x) = 0 \quad (5.2)$$

$$r_i = y_i \quad (5.3)$$

### 2. Ajuste Iterativo das Árvores:

Para cada iteração  $b = 1, 2, \dots, B$  (em que  $B$  é o número total de árvores):

#### a) Ajuste de uma Árvore aos Resíduos:

Constrói-se uma nova árvore de decisão  $\hat{f}^b$  utilizando os resíduos atuais  $r_i$  como variável resposta. Essa árvore possui complexidade limitada, geralmente com um número fixo de divisões  $d$ , resultando em  $d + 1$  nós terminais.

#### b) Atualização do Modelo Geral:

Atualiza-se a função de previsão geral  $\hat{f}(x)$  adicionando uma versão ponderada da nova árvore. Essa ponderação é feita multiplicando as previsões da árvore por um parâmetro de *shrinkage*  $\lambda$  (taxa de aprendizado):

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \quad (5.4)$$

#### c) Atualização dos Resíduos:

Recalculam-se os resíduos para a próxima iteração, subtraindo a contribuição ponderada da nova árvore:

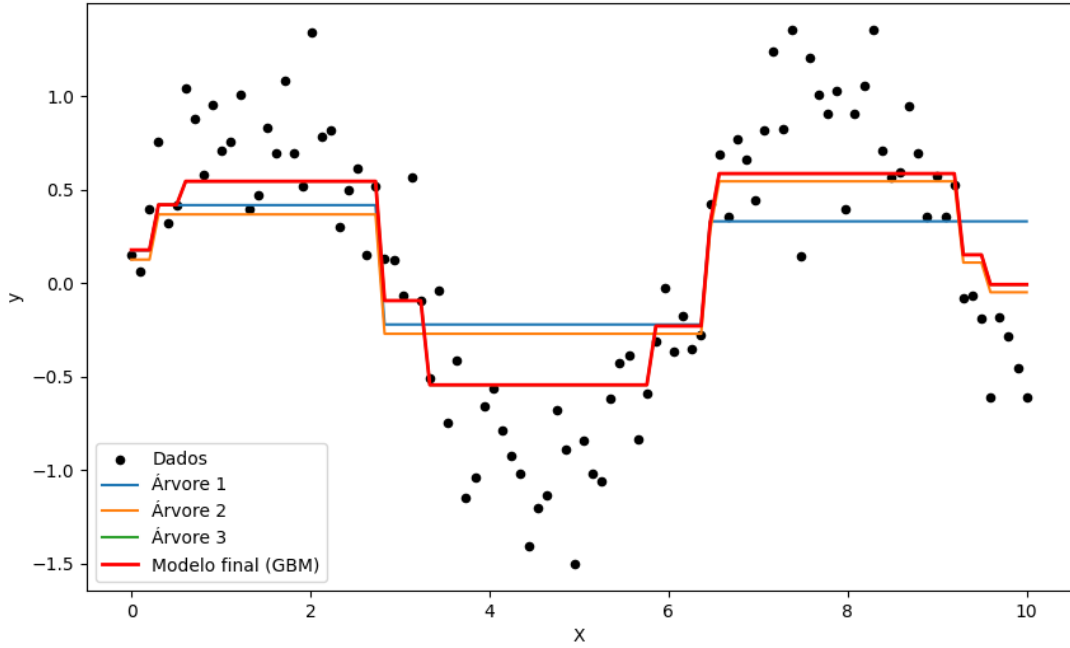
$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i) \quad (5.5)$$

### 3. Saída Final:

O modelo final é a soma ponderada de todas as árvores construídas sequencialmente:

$$\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^b(x) \quad (5.6)$$

Figura 5.2 – Representação gráfica do processo do GBM.



Fonte: Elaboração própria (2026).

Embora o XGBoost compartilhe a lógica de aprendizado aditivo do GBM, ele introduz inovações que o tornam superior em termos de desempenho e controle de erro:

### 1. A Função Objetivo com Regularização:

Enquanto o GBM tradicional foca puramente na minimização de uma função de perda  $l(y_i, \hat{y}_i)$ , o XGBoost utiliza uma função objetivo regularizada.

A diferença pode ser vista na inclusão do termo de penalidade  $\Omega$ :

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(\hat{f}^b) \quad (5.7)$$

- $\sum_{i=1}^n l(\hat{y}_i, y_i)$ : Representa a perda do treinamento (como no GBM), medindo a diferença entre a o alvo real ( $y_i$ ) e a previsão ( $\hat{y}_i$ ).
- $\sum_{b=1}^B \Omega(\hat{f}^b)$ : É o diferencial do XGBoost, onde  $\Omega$  penaliza a complexidade de cada árvore ( $\hat{f}^b$ ) para evitar o (*overfitting*).

### 2. O Termo de Penalidade $\Omega$ :

O cálculo da complexidade no XGBoost é definido matematicamente pela seguinte fórmula, que não existe no GBM padrão:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (5.8)$$

Nesta equação:

- $T$  é o número de folhas na árvore. O parâmetro  $\gamma$  controla o "custo" de se adicionar uma nova folha, servindo como um limitador de crescimento.
- $w$  são os pesos atribuídos a cada folha. O parâmetro  $\lambda$  aplica uma regularização L2 sobre esses pesos, suavizando as previsões e aumentando a robustez do modelo.

### 3. Otimização via Expansão de Taylor:

No GBM tradicional, o ajuste é feito com base nos pseudo-resíduos, que utilizam apenas a primeira derivada (gradiente) da função de perda. Já o XGBoost utiliza uma aproximação de segunda ordem da função de perda através da série de Taylor. Matematicamente, para a iteração  $t$ , o objetivo é aproximado por:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^n \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (5.9)$$

- $g_i$  (gradiente): Primeira derivada da função de perda.
- $w$  são os pesos atribuídos a cada folha. O parâmetro  $\lambda$  aplica uma regularização L2 sobre esses pesos, suavizando as previsões e aumentando a robustez do modelo.

### 4. Tratamento de Dados Esparsos (Sparsity-Awareness):

Diferente do GBM, que requer o tratamento prévio de dados ausentes, o XGBoost possui um algoritmo consciente da esparsidade. Ele aprende matematicamente uma "direção padrão" em cada nó da árvore. Se um valor está ausente, o algoritmo o direciona para o caminho que minimiza a perda, otimizando o treinamento em bases com muitos valores nulos ou variáveis categóricas esparsas.

Portanto, o diferencial do XGBoost reside na incorporação de uma função objetivo regularizada, que utiliza penalidades para controlar a complexidade das árvores e evitar o *overfitting*. Matematicamente, enquanto o GBM se limita a estatísticas de primeira ordem, o XGBoost emprega uma aproximação de segunda ordem (Série de Taylor) da função de perda, utilizando informações do gradiente e da hessiana para garantir uma convergência mais rápida e precisa em problemas complexos. Além disso, o algoritmo destaca-se pela sua consciência da esparsidade (*sparsity-aware*), sendo capaz de aprender direções padrão para lidar com valores ausentes de forma nativa e automática, o que o torna até 50 vezes mais rápido que implementações convencionais em bases de dados volumosas e incompletas como as da PRF.

## 5.3 Validação Cruzada e Otimização de Hiperparâmetros

A otimização de hiperparâmetros foi conduzida por meio de um processo de busca aleatória estruturada (*Randomized Search*), aliada a uma validação cruzada iterativa de

cinco divisões (**5-fold Cross-Validation**). Esse delineamento metodológico foi aplicado uniformemente aos modelos de Regressão Logística e XGBoost, sendo fundamental para mitigar o risco de *overfitting* e garantir que os algoritmos treinassem e validassem seus aprendizados em subconjuntos distintos, atestando a capacidade de generalização dos padrões de gravidade dos acidentes em dados inéditos. Para assegurar uma exploração abrangente e com validade estatística no espaço de soluções, estipularam-se 100 candidatos aleatórios de configuração para cada algoritmo.

No caso do classificador da Regressão Logística, utilizou-se o otimizador avançado SAGA, o qual permite a aplicação da penalização **Elastic-Net**. A vantagem do **Elastic-Net** reside na sua capacidade de combinar simultaneamente as regularizações  $L_1$  (Lasso) e  $L_2$  (Ridge), calibradas pelo hiperparâmetro `l1_ratio`. A busca também avaliou diferentes níveis da força de regularização (parâmetro `C`) e a utilização do balanceamento automático de pesos das classes (`class_weight`), uma estratégia essencial para não negligenciar os acidentes graves.

Já para o XGBoost, as distribuições de busca incluíram variáveis uniformes e discretas para calibrar os limites de complexidade arquitetônica das árvores de decisão (profundidade máxima variando de 3 a 10 níveis), as eficiências estocásticas (taxas de aprendizado e frações de amostragem de linhas e colunas), além da força de regularização para o peso das classes (`scale_pos_weight`), visando também compensar o desbalanceamento intrínseco aos dados.

Todas as avaliações de otimização utilizaram como métrica de direcionamento a maximização do F1-Score. A escolha dessa métrica justifica-se pela sua capacidade de balancear a Precisão e a Sensibilidade (*Recall*), impedindo que os modelos foquem excessivamente na classe majoritária (não acidentes graves) em detrimento da minoria de interesse (acidentes graves).

A Tabela 5.1 consolida o espaço de busca e as distribuições probabilísticas definidas para os hiperparâmetros estratégicos de cada modelo avaliado

Tabela 5.1 – Espaço de busca de hiperparâmetros definido para a otimização dos modelos

| Modelo              | Hiperparâmetro                | Espaço de Busca / Distribuição |
|---------------------|-------------------------------|--------------------------------|
| Regressão Logística | <code>solver</code>           | Fixo: ['saga']                 |
|                     | <code>penalty</code>          | Fixo: ['elasticnet']           |
|                     | <code>C</code>                | Log-uniforme (0,01 a 100)      |
|                     | <code>l1_ratio</code>         | Uniforme (0,00 a 1,00)         |
|                     | <code>class_weight</code>     | Categórica: [None, 'balanced'] |
| XGBoost             | <code>n_estimators</code>     | Inteiro (100 a 1000)           |
|                     | <code>learning_rate</code>    | Uniforme (0,01 a 0,31)         |
|                     | <code>max_depth</code>        | Inteiro (3 a 10)               |
|                     | <code>subsample</code>        | Uniforme (0,50 a 1,00)         |
|                     | <code>colsample_bytree</code> | Uniforme (0,50 a 1,00)         |
|                     | <code>scale_pos_weight</code> | Uniforme (1,00 a 11,00)        |

Fonte: Elaboração própria (2026).

## 5.4 Shap

Para mitigar a opacidade de modelos complexos como o XGBoost, utiliza-se o framework SHAP, proposto por Lundberg e Lee (2017) como uma abordagem unificada para interpretar predições de modelos. O SHAP atribui a cada característica um valor de importância para uma previsão específica, permitindo que usuários entendam o motivo por trás do resultado gerado pelo modelo.

O SHAP define um modelo de explicação  $g$  que é uma função linear de variáveis binárias, pertencendo à classe dos métodos de atribuição aditiva de características:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i \quad (5.10)$$

representa o efeito (valor de importância) atribuído a cada característica de entrada.

Sendo assim, o impacto do uso do SHAP neste trabalho reflete-se na transição de um modelo de "caixa-preta" para um sistema de suporte à decisão transparente e auditável. Enquanto algoritmos de *gradient boosting* como o XGBoost priorizam a performance preditiva, o SHAP garante que essa performance seja compreensível para gestores públicos. No contexto desta pesquisa, o impacto manifesta-se em três frentes principais:

1. **Confiabilidade das Explicações:** Ao contrário de métodos heurísticos, o SHAP oferece consistência matemática. Se o modelo aumentar a importância de uma variável (como a densidade da frota municipal) para prever um acidente grave, o valor de SHAP refletirá esse aumento de forma proporcional, garantindo que as causas apontadas pelo modelo sejam confiáveis.

2. **Identificação de Gatilhos de Risco:** O uso do SHAP permitiu identificar não apenas quais variáveis são importantes, mas como elas afetam o risco. Foi possível observar, por exemplo, o impacto marginal de condições climáticas adversas em trechos específicos, revelando padrões que métricas de importância global tradicionais ocultariam.
3. **Subsídio a Políticas Públicas:** A interpretabilidade fornecida permite que os resultados saiam do campo estatístico e entrem no campo da ação governamental. Ao explicar o motivo de um trecho de rodovia ser classificado como de alto risco, o modelo fornece a justificativa necessária para intervenções de infraestrutura ou campanhas de conscientização direcionadas, aumentando a aceitabilidade da tecnologia por parte dos órgãos fiscalizadores.

Dessa forma, a integração do SHAP eleva o rigor metodológico do trabalho, assegurando que a predição da gravidade dos acidentes seja acompanhada de uma explicação clara e tecnicamente fundamentada sobre os fatores contribuintes.

## 5.5 Coleta e Integração de Dados

A base de dados utilizada neste estudo é fruto da consolidação de duas fontes de dados abertos governamentais, abrangendo o período de 2017 a 2025. A fonte primária dos registros de sinistros foi o portal de dados abertos da Polícia Rodoviária Federal (PRF). Para cada ocorrência, foram extraídas variáveis fundamentais como Unidade da Federação (UF), rodovia (BR), quilômetro (KM), dia da semana, fase do dia, condição meteorológica, tipo de pista, traçado da via e tipo de veículo.

Para enriquecer a análise e permitir que o modelo capture o contexto socioeconômico e de exposição ao risco, os dados da PRF foram integrados aos dados de frota municipal disponibilizados pelo Ministério dos Transportes. Este processo de integração foi executado através de uma operação de junção (*left join*) utilizando como chaves de cruzamento as colunas município, uf e ano.

## 5.6 Engenharia de Atributos

A etapa de engenharia de atributos foi desenhada para transformar dados brutos em representações numéricas eficientes, utilizando um pipeline de pré-processamento padronizado.

Primeiramente, procedeu-se à definição da variável-alvo, utilizando-se como referência a coluna `estado_fisico`, que representa a condição da pessoa envolvida no acidente. Essa variável foi binarizada em duas classes: a **classe positiva (1)**, correspondente aos

casos de **desfecho grave**, compreendendo as categorias “Lesões Graves” e “Óbito”; e a **classe negativa (0)**, correspondente aos casos de **desfecho não grave**, abrangendo as categorias “Ileso” e “Lesões Leves”

Outras transformações estruturais incluíram:

- **Tratamento Espacial:** A variável de localização `km` foi limpa, arredondada para valores inteiros e concatenada à identificação da rodovia (`br`), gerando o atributo composto `br_km`. Essa abordagem permitiu agrupar geograficamente as ocorrências em trechos específicos e padronizados da malha rodoviária, descontinuando o uso da coluna isolada `km` nas etapas substituídas.
- **Geometria Viária Multirrótulo:** O atributo `tracado_via` foi tratado por meio de binarização baseada em separadores (`str.get_dummies(sep=';')`), fragmentando as características físicas da via (como curvas, retas, pontes e aclives) em colunas binárias independentes (0 ou 1), permitindo mapear cenários onde múltiplos elementos geométricos coexistem.
- **Tratamento Categórico e Numérico Dinâmico:** Para o atributo `frota`, utilizou-se o `SimpleImputer` (mediana) para tratar valores ausentes e o `StandardScaler` para normalizar a escala. Quanto aos atributos textuais (`uf`, `dia_semana` e `tipo_pista`), a estratégia de codificação foi estritamente personalizada para cada paradigma algorítmico:
  - **Para a Regressão Logística:** Aplicou-se o `OneHotEncoder` tradicional. Para mitigar a "maldição da dimensionalidade" provocada pelo espalhamento de colunas, utilizou-se o parâmetro de frequência mínima (`min_frequency=0.001`) para agrupar categorias raras em um único bin genérico, preservando apenas padrões linearmente relevantes.
  - **Para o XGBoost:** Evitou-se o uso de `OneHotEncoder` para não inflar a matriz de dados desnecessariamente. Em vez disso, as variáveis textuais foram mantidas no formato nativo de categorias do Pandas (`category`), permitindo que o algoritmo processasse esses atributos diretamente através da ativação do parâmetro interno `enable_categorical=True`, otimizando o ganho de informação nas divisões das árvores de decisão.

Após a integração, os dados passaram por um processo de limpeza para garantir a integridade dos modelos. O tratamento de valores ausentes em variáveis categóricas foi realizado via moda (`most_frequent`), e registros com inconsistências críticas de localização foram descartados.

Um desafio central foi o desbalanceamento de classes, visto que acidentes leves são substancialmente mais frequentes que acidentes graves. Para mitigar o risco de viés

em direção à classe majoritária, adotou-se uma divisão estratificada (**stratify=y**) na proporção de 80% para treino e 20% para teste. Complementarmente, no treinamento do algoritmo XGBoost, utilizou-se o ajuste de pesos para a classe minoritária, permitindo que o modelo penalizasse de forma mais rigorosa os erros de classificação em acidentes de alta gravidade.



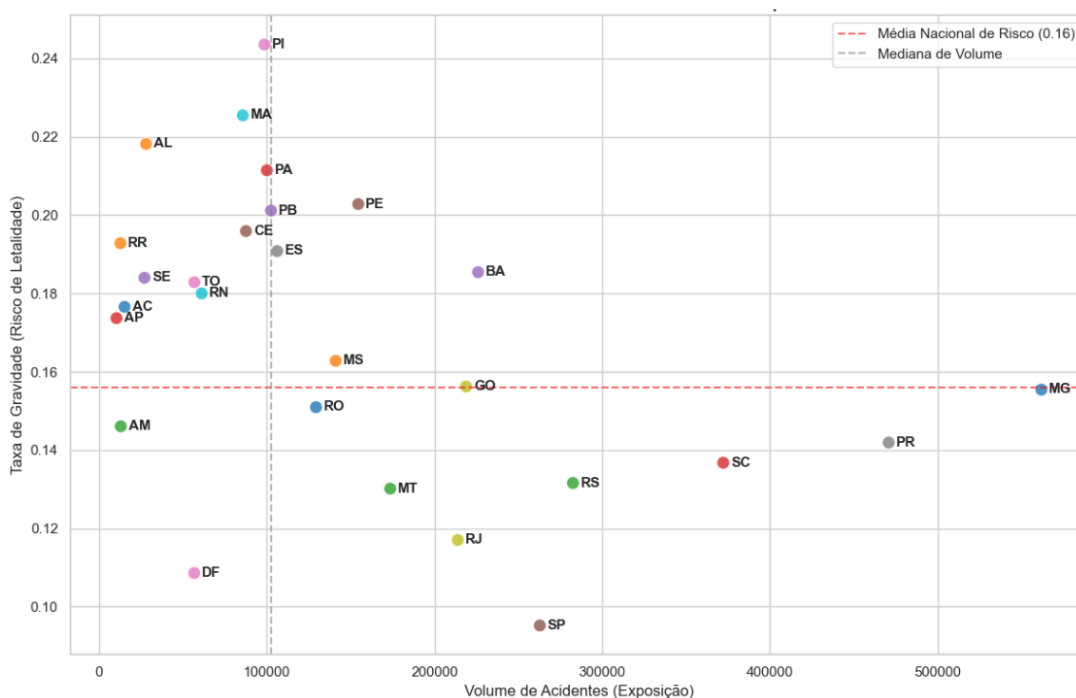
## 6 Análise Descritiva dos Dados: O Panorama da Segurança Viária no Brasil

A compreensão dos fatores que contribuem para a severidade dos acidentes nas rodovias federais brasileiras exige uma análise que transcenda a simples contagem de ocorrências. Esta seção busca detalhar e identificar padrões na malha rodoviária nacional, utilizando variáveis geográficas, veiculares e ambientais para fundamentar a necessidade de uma abordagem preditiva via aprendizado de máquina.

### 6.1 Análise da Relação entre Volume de Acidentados e Tipo de Veículo com a Taxa de Gravidade

Ao cruzar o volume total de acidentados com a taxa de gravidade por Unidade Federativa, conforme apresentado na Figura 6.1, observa-se que o risco não é distribuído de forma homogênea pelo território nacional.

Figura 6.1 – Matriz de Risco: Volume Total vs. Taxa de Gravidade por Estado.



Fonte: Elaboração própria com base em dados da PRF (2026).

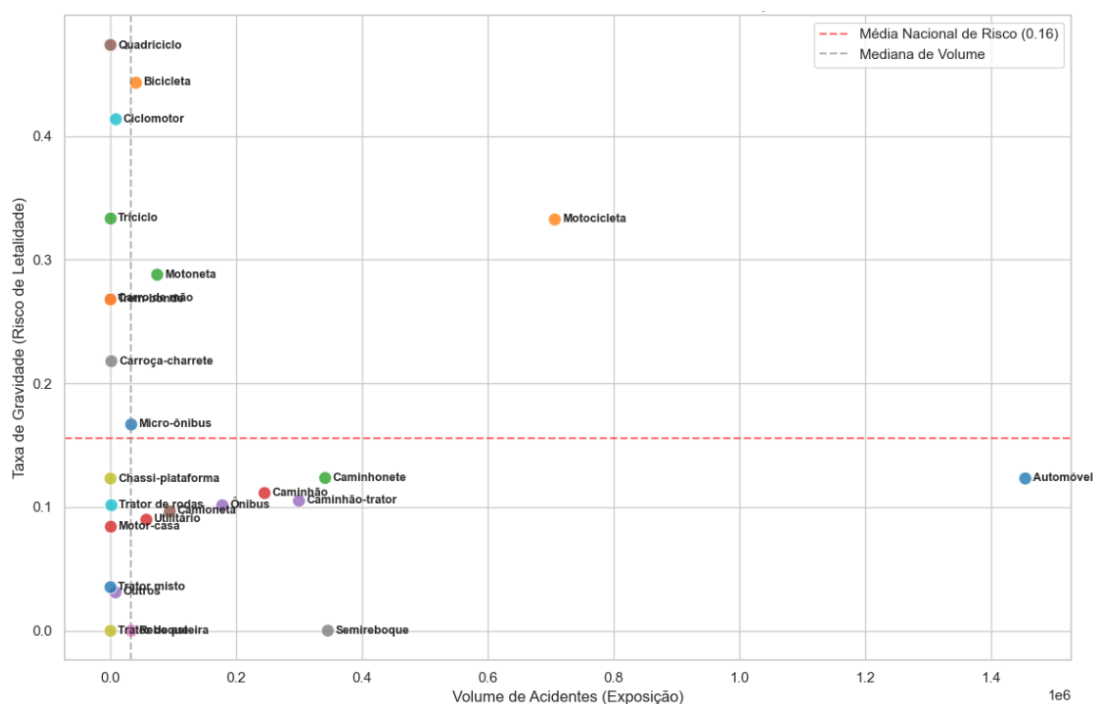
Minas Gerais (MG) posiciona-se como um outlier extremo no eixo horizontal. Segundo o Anuário da PRF (2026), Minas Gerais concentra o maior volume absoluto de

ocorrências registradas em rodovias federais sob jurisdição única, superando estados como São Paulo e Paraná devido à extensão de sua malha sob fiscalização federal. No entanto, a sua taxa de gravidade situa-se em patamares intermediários. Isso sugere que, embora a probabilidade de um acidente ocorrer seja alta devido ao fluxo intenso (exposição), a infraestrutura ou o tipo de colisão predominante não resulta necessariamente em uma letalidade superior à média nacional.

Um ponto crítico revelado pelo gráfico é o posicionamento de estados como Piauí (PI) e Maranhão (MA). Embora apresentem volumes de acidentes consideravelmente menores que os estados do Sul e Sudeste, eles registram taxas de gravidade significativamente elevadas. Este fenômeno é explicado por uma convergência de fatores infraestruturais, dinâmicas de exposição ao risco e impactos socioeconômicos. Conforme os dados da Pesquisa CNT (2023), o estado de conservação das rodovias nessas regiões é crítico: no Maranhão, 88,2% da malha rodoviária é classificada como regular, ruim ou péssima, enquanto no Piauí esse índice atinge 60,4%. Adicionalmente, a predominância de pistas simples nessas regiões favorece a ocorrência de colisões frontais, classificadas pelo Anuário da Polícia Rodoviária Federal (2026) como o tipo de acidente mais letal.

Quando a análise se desloca para as diferentes categorias de veículos, a desconexão entre a quantidade de ocorrências e a letalidade real fica ainda mais evidente, como apresentado na Figura 6.2 apresenta a relação entre o volume total de acidentes e a taxa de gravidade.

Figura 6.2 – Matriz de Risco: Volume Total vs. Taxa de Gravidade por Tipo de Veículo.



Fonte: Elaboração própria com base em dados da PRF (2026).

Os resultados demonstram que a categoria automóvel apresentou o maior volume absoluto de acidentados, ultrapassando 580 mil ocorrências. Entretanto, apesar da elevada exposição, sua taxa de gravidade permaneceu relativamente baixa (aproximadamente 0,13), indicando que, embora os automóveis estejam envolvidos em grande quantidade de acidentes, a proporção de casos letais é inferior à observada em veículos mais vulneráveis.

Por outro lado, a categoria motocicleta destacou-se por combinar elevado volume de acidentados com alta taxa de gravidade (aproximadamente 0,33), posicionando-se como um dos grupos de maior criticidade na matriz de risco. Esse resultado reforça os dados da MOURA (2025), que indicam que este modal responde por quase 40% das mortes no trânsito nacional.

Veículos de menor participação no volume total de acidentados, como bicicletas, triciclos, ciclomotores e quadriciclos, apresentaram as maiores taxas de gravidade observadas no conjunto analisado. Embora a reduzida proteção estrutural desses veículos contribua para a maior vulnerabilidade dos ocupantes, os resultados devem ser interpretados com cautela. É possível que exista um efeito de subnotificação dos acidentes de menor gravidade envolvendo esses modais, uma vez que ocorrências sem vítimas ou com danos leves frequentemente não chegam aos registros oficiais. Dessa forma, os acidentes registrados tendem a representar predominantemente eventos mais severos, elevando artificialmente a taxa de gravidade observada para essas categorias. Esse possível viés de seleção pode contribuir para a superestimação do risco relativo desses veículos quando comparados a categorias com maior cobertura de registro, como automóveis e caminhões.

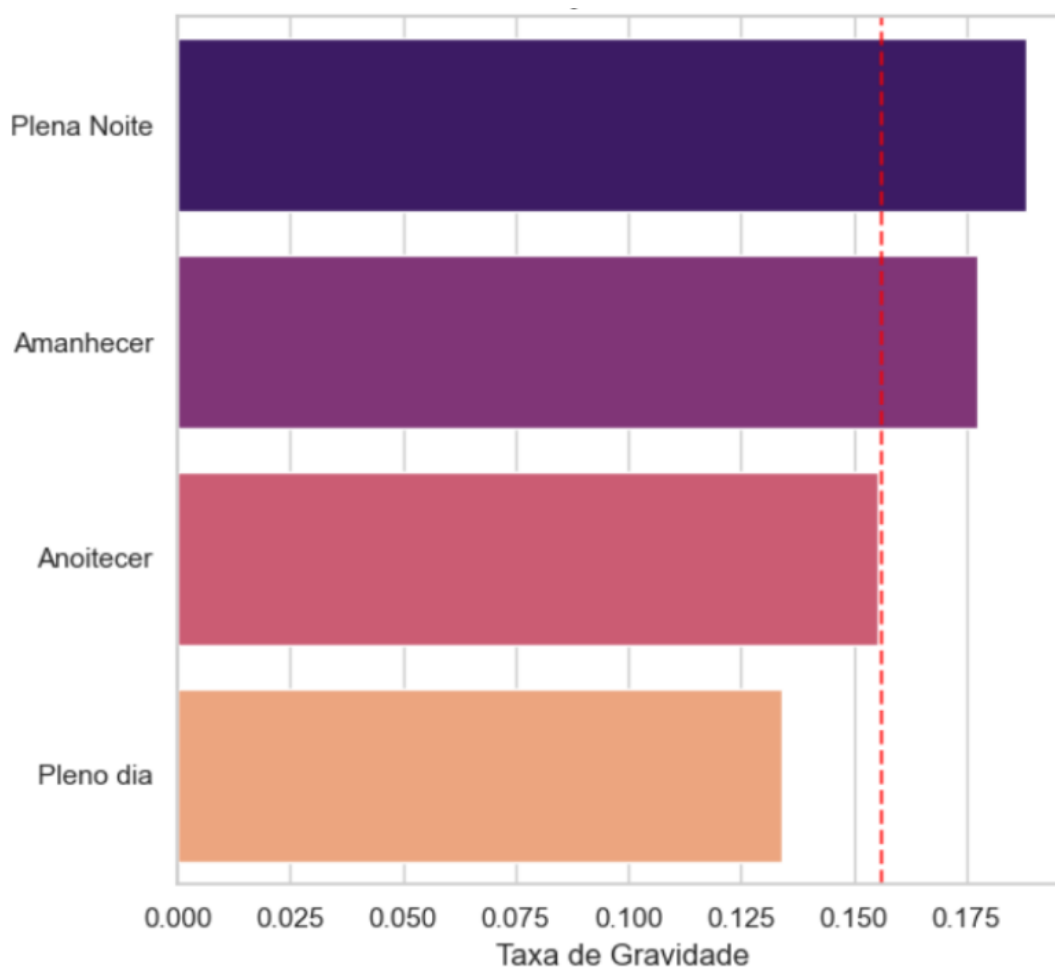
As categorias motoneta e chassi-plataforma também apresentaram taxas elevadas de gravidade, acima da média nacional de risco, ainda que com exposição relativamente moderada. Em contraste, veículos de carga e transporte pesado, como caminhão, caminhão-trator, semi-reboque e ônibus, apresentaram taxas de gravidade inferiores ou próximas à média nacional, apesar de volumes intermediários de acidentes. Observa-se ainda que determinadas categorias, como utilitário, camioneta, trator misto e trator de rodas, concentraram-se em regiões de baixa gravidade e baixo volume de acidentes, sugerindo menor criticidade relativa dentro do cenário analisado.

De forma geral, a matriz de risco permitiu identificar diferentes padrões de comportamento entre os tipos de veículos, evidenciando que a severidade dos acidentes não está necessariamente associada ao maior volume de ocorrências. Enquanto veículos de passeio concentram grande parte da exposição total, categorias mais vulneráveis, especialmente motocicletas e veículos de pequeno porte sem proteção estrutural, apresentaram maior risco relativo de letalidade.

## 6.2 Acidentados Graves por Variáveis Temporais

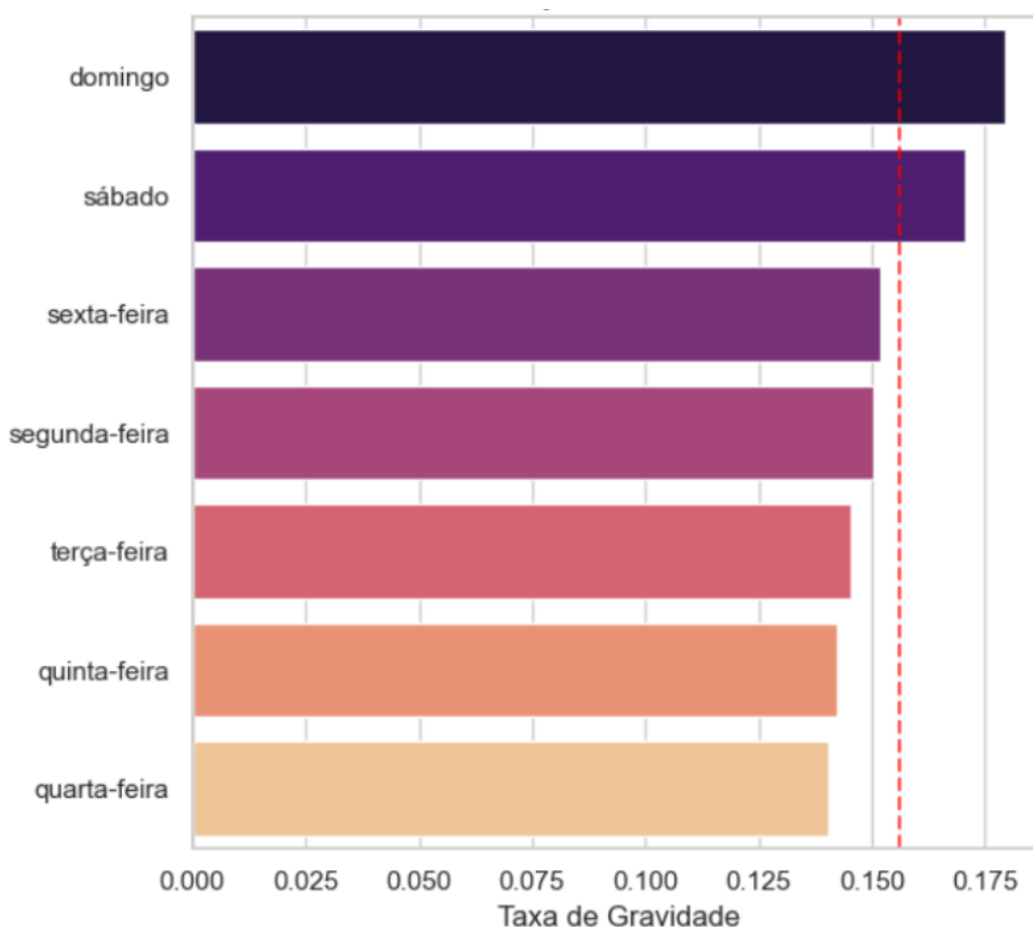
A Figura 6.3 apresenta a distribuição da taxa de gravidade dos acidentados segundo a fase do dia, enquanto a Figura 6.4 apresenta essa distribuição segundo o dia da semana. A linha tracejada representa a média geral de gravidade observada no conjunto de dados.

Figura 6.3 – Taxa de Acidentados Graves por Fase do Dia.



Fonte: Elaboração própria com base em dados da PRF (2026).

Figura 6.4 – Risco por Dia da Semana.



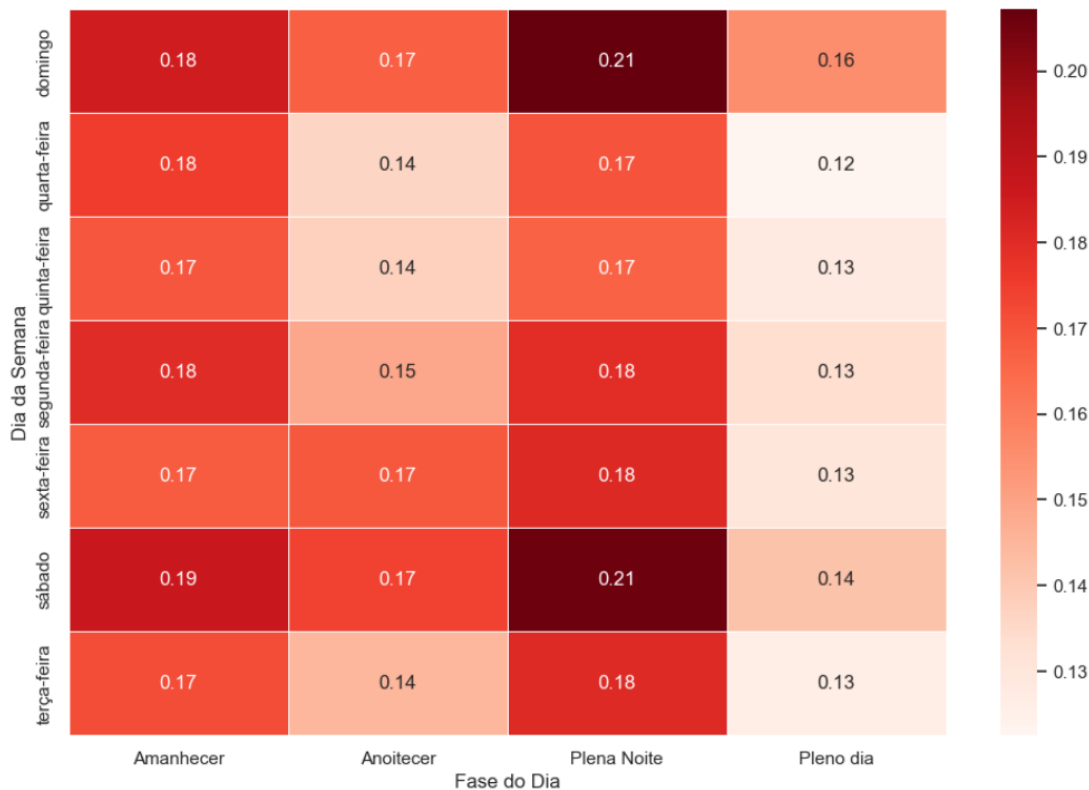
Fonte: Elaboração própria com base em dados da PRF (2026).

Os resultados demonstram que os acidentes ocorridos durante a plena noite apresentaram os maiores níveis de gravidade entre os períodos analisados, ultrapassando a média geral observada na base de dados. Os períodos de amanhecer e anoitecer também apresentaram taxas elevadas de gravidade, enquanto os acidentes registrados durante o pleno dia apresentaram os menores índices relativos.

Esse comportamento sugere forte influência das condições de visibilidade reduzida sobre a severidade das ocorrências. Além disso, fatores como fadiga dos condutores, redução da percepção de risco, menor iluminação das vias e maiores velocidades médias durante períodos noturnos podem contribuir para o aumento da letalidade dos acidentes. Na análise por dia da semana, observou-se que os acidentes registrados aos domingos apresentaram a maior taxa média de gravidade, seguidos pelos ocorridos aos sábados. Em contrapartida, os menores índices foram identificados entre quarta-feira e quinta-feira.

A Figura 6.5 apresenta a taxa média de gravidade observada para cada combinação entre os períodos do dia e os dias da semana. Tons mais intensos representam maiores níveis de gravidade.

Figura 6.5 – Risco por Dia da Semana e Fase do Dia.



Fonte: Elaboração própria com base em dados da PRF (2026).

Os resultados evidenciam que os maiores níveis de gravidade ocorreram durante o período de plena noite, especialmente aos sábados e domingos, com taxas superiores a 0,20. O maior valor observado foi registrado aos sábados durante a plena noite, atingindo aproximadamente 0,22.

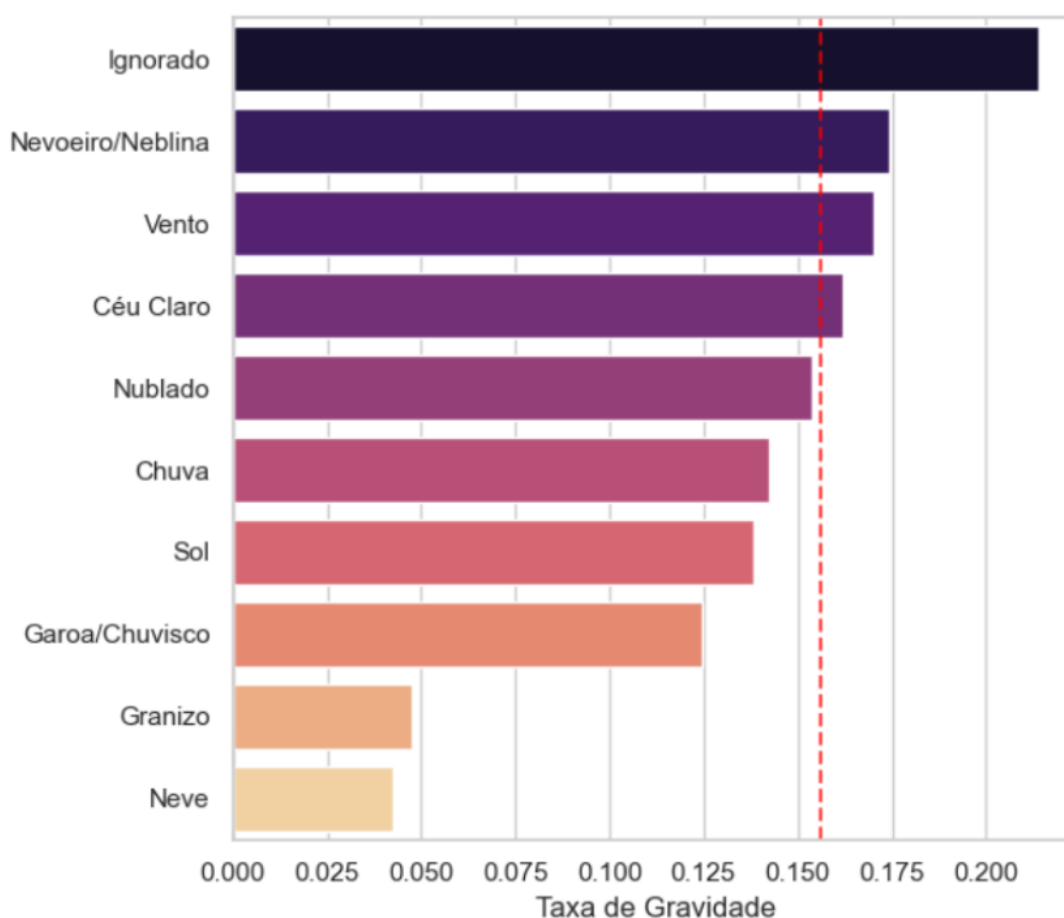
Observa-se ainda que os períodos de amanhecer também apresentaram níveis elevados de gravidade em praticamente todos os dias da semana, especialmente aos sábados e domingos. Em contrapartida, o período de pleno dia apresentou consistentemente os menores índices de gravidade, independentemente do dia analisado. Outro aspecto relevante identificado na matriz de calor é a relativa estabilidade dos padrões temporais ao longo da semana, indicando que a fase do dia exerce influência mais significativa sobre a gravidade dos acidentes do que o próprio dia da semana isoladamente.

Os resultados indicam forte associação entre períodos noturnos de finais de semana e aumento da severidade dos acidentes, possivelmente associado ao crescimento de deslocamentos recreativos, viagens de longa distância e maior incidência de comportamentos de risco, como excesso de velocidade e condução sob efeito de álcool.

### 6.3 Acidentados Graves por Condições Ambientais

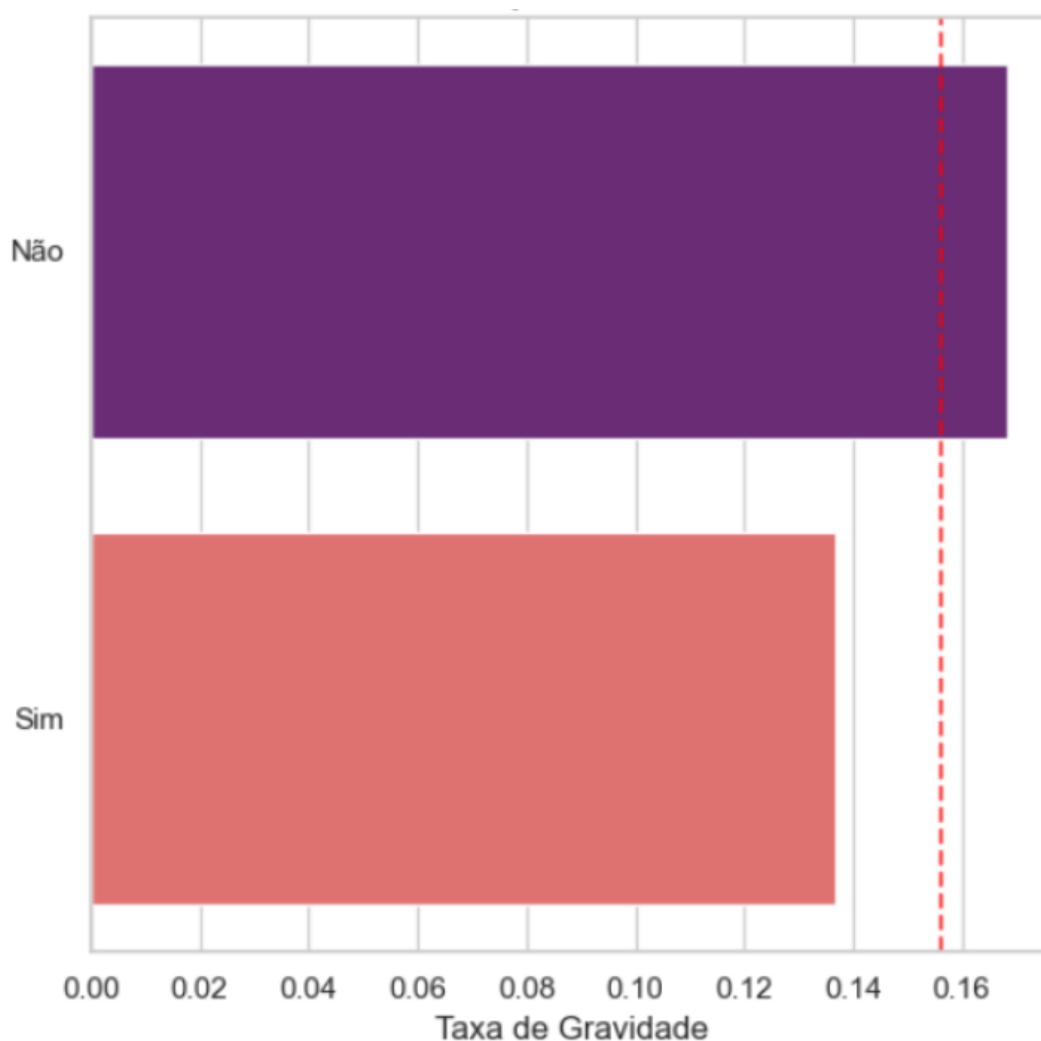
A Figura 6.6 apresenta a distribuição da taxa de acidentados graves segundo as condições meteorológicas, enquanto a Figura 6.7 apresenta essa distribuição segundo o tipo de uso do solo associado ao trecho rodoviário analisado, sendo “Não” para trechos não urbanizados e “Sim” para trechos urbanizados. A linha tracejada representa a média geral de gravidade observada no conjunto de dados.

Figura 6.6 – Risco por Condição Meteorológica.



Fonte: Elaboração própria com base em dados da PRF (2026).

Figura 6.7 – Risco por Uso do Solo.



Fonte: Elaboração própria com base em dados da PRF (2026).

Na análise das condições meteorológicas, observou-se que as maiores taxas de gravidade ocorreram nas categorias ignorado e nevoeiro/neblina, ambas acima da média geral da base analisada. A elevada severidade associada aos cenários de neblina evidencia o impacto da baixa visibilidade sobre a segurança viária, dificultando a percepção de obstáculos, a identificação da sinalização e o tempo de reação dos condutores.

As condições de céu claro também apresentaram taxas relativamente elevadas de gravidade. Esse resultado sugere que acidentes graves não estão necessariamente associados apenas a condições climáticas adversas, podendo estar relacionados a fatores comportamentais, como excesso de velocidade e redução da percepção de risco em situações de boa visibilidade.

Em contrapartida, condições de chuva, garoa/chuvisco e vento apresentaram níveis intermediários de gravidade, inferiores aos observados em cenários de neblina. Uma possível explicação para esse comportamento é a tendência de redução preventiva da velocidade



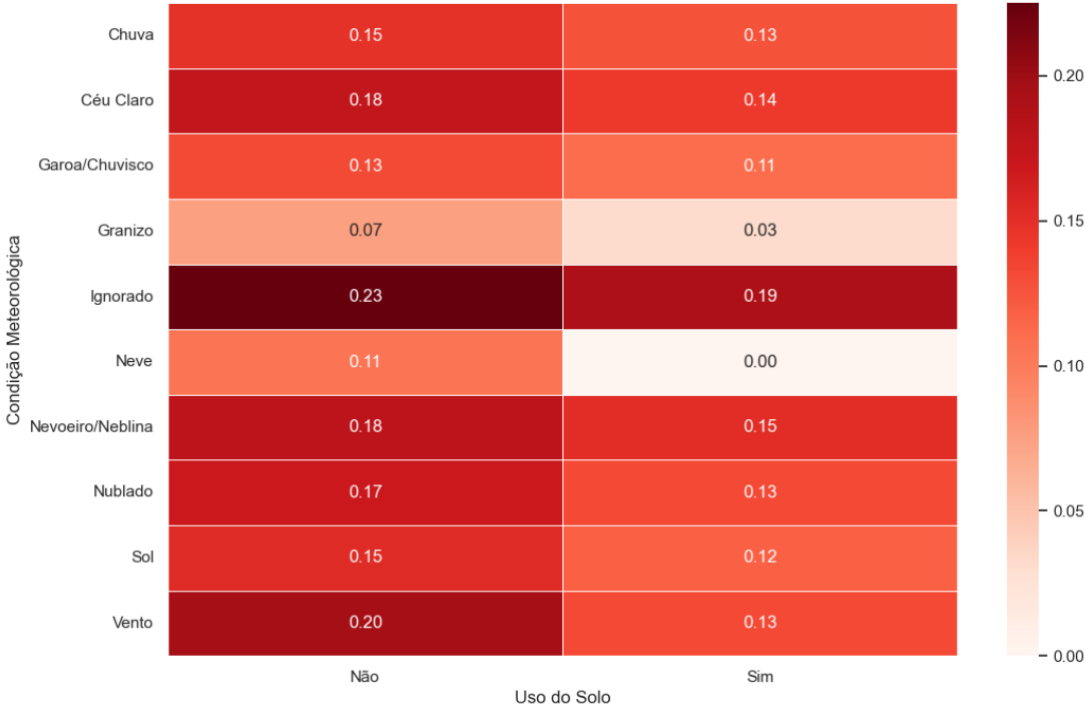
média dos veículos durante condições climáticas adversas mais perceptíveis aos condutores.

Já as categorias granizo e neve apresentaram baixas taxas de gravidade. Entretanto, esses resultados devem ser interpretados com cautela devido à baixa frequência de registros dessas condições meteorológicas na base de dados, reduzindo sua representatividade estatística.

Em relação ao uso do solo, observou-se que acidentes ocorridos em áreas classificadas como não urbanizadas apresentaram maior taxa de gravidade quando comparados às áreas urbanizadas. Esse comportamento pode estar associado às maiores velocidades operacionais praticadas em regiões rurais, além da menor presença de dispositivos de controle de tráfego e maior tempo de resposta dos serviços de emergência. Por outro lado, acidentes registrados em áreas urbanizadas apresentaram menor gravidade relativa, possivelmente em decorrência da redução da velocidade média dos veículos, maior presença de fiscalização e maior proximidade de estruturas de atendimento emergencial.

A Figura 6.8 apresenta a taxa média de gravidade observada para cada combinação entre as condições meteorológicas registradas e a classificação do uso do solo. Tons mais intensos representam maiores níveis de gravidade.

Figura 6.8 – Risco por Condição Meteorológica e Uso do Solo.



Fonte: Elaboração própria com base em dados da PRF (2026).

Os resultados demonstram que, de maneira geral, os acidentes em áreas classificadas como não urbanizadas apresentaram taxas de gravidade superiores às observadas em áreas urbanizadas para praticamente todas as condições meteorológicas analisadas. Esse comportamento reforça os resultados encontrados na análise individual da variável

uso do solo, indicando maior severidade dos acidentes em regiões rurais ou afastadas dos centros urbanos.

Entre as condições meteorológicas avaliadas, destacaram-se os registros classificados como ignorado, que apresentaram as maiores taxas de gravidade tanto em áreas urbanizadas quanto não urbanizadas, alcançando valores próximos de 0,23 e 0,20, respectivamente. Embora essa categoria deva ser interpretada com cautela devido à ausência de informação meteorológica, sua elevada gravidade pode indicar situações críticas em que o registro das condições ambientais foi comprometido em decorrência da severidade da ocorrência.

As condições de nevoeiro/neblina apresentaram uma das combinações de maior risco observadas no estudo, especialmente em áreas não urbanizadas, com taxa de gravidade próxima de 0,19. Esse resultado evidencia o impacto da baixa visibilidade sobre a segurança viária, potencializado por características frequentemente encontradas em rodovias rurais, como maiores velocidades operacionais e menor iluminação.

Observa-se também que condições de céu claro apresentaram taxas relativamente elevadas de gravidade, principalmente em áreas não urbanizadas. Esse resultado sugere que a ocorrência de acidentes graves não depende exclusivamente de condições climáticas adversas, estando também associada a fatores comportamentais e operacionais, como excesso de velocidade e redução da percepção de risco por parte dos condutores.

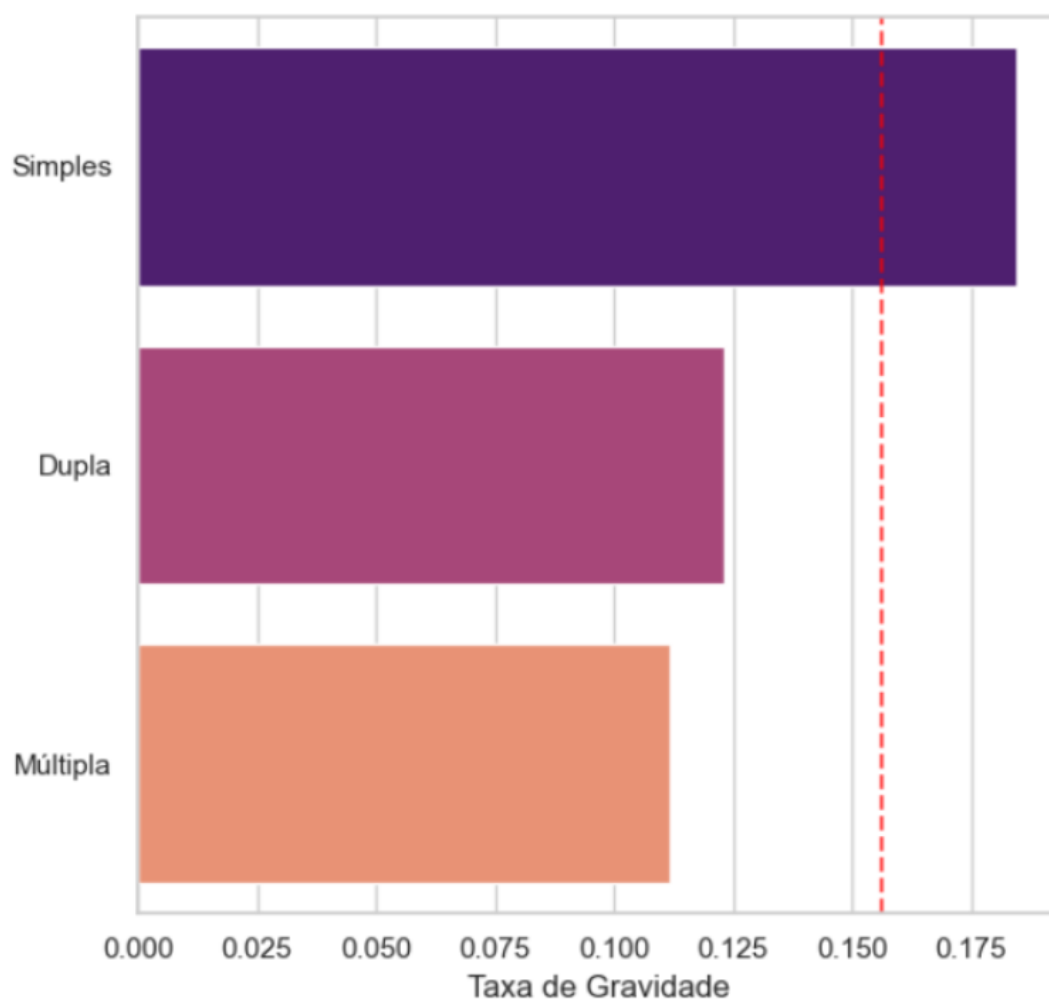
De forma geral, a análise conjunta evidencia que o uso do solo exerce influência consistente sobre a gravidade dos acidentes, independentemente da condição meteorológica observada. Os resultados indicam que a combinação entre condições de baixa visibilidade e trechos não urbanizados representa um dos cenários de maior criticidade para a segurança viária.

## 6.4 Acidentados Graves por Características das Vias

Com o objetivo de analisar a influência da infraestrutura rodoviária sobre a severidade dos acidentes, foi realizada uma avaliação da taxa de gravidade considerando variáveis relacionadas às características geométricas e operacionais das vias.

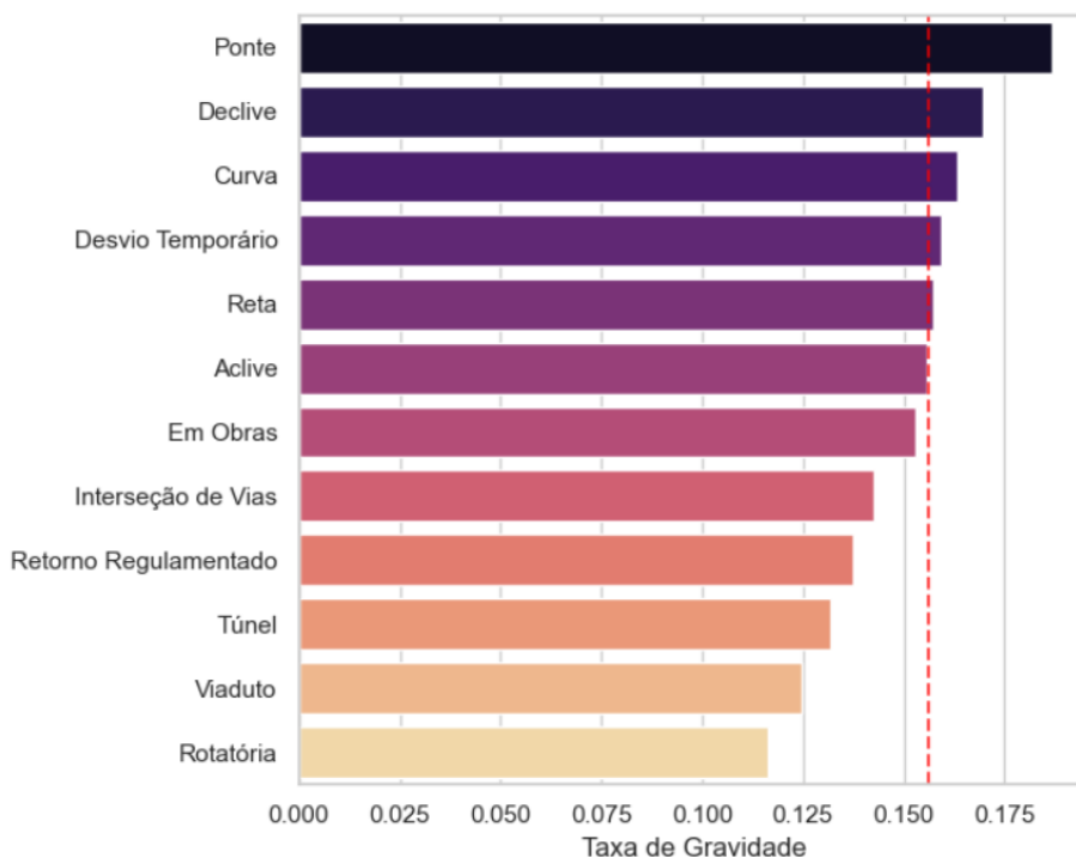
A Figura 6.9 apresenta a distribuição da taxa de acidentados segundo o tipo de pista, enquanto a Figura 6.10 apresenta essa distribuição segundo o traçado da via. A linha tracejada representa a média geral de gravidade observada no conjunto de dados.

Figura 6.9 – Risco por Tipo de Pista.



Fonte: Elaboração própria com base em dados da PRF (2026).

Figura 6.10 – Risco por Traçado da Via.



Fonte: Elaboração própria com base em dados da PRF (2026).

Na análise do tipo de pista, observou-se que rodovias classificadas como pista simples apresentaram a maior taxa de acidentes graves entre as categorias avaliadas, superando significativamente a média geral da base analisada. Em contrapartida, pistas duplas e múltiplas apresentaram níveis inferiores de severidade. Esse comportamento evidencia a influência da configuração viária sobre o risco dos acidentes. Rodovias de pista simples normalmente apresentam maior probabilidade de colisões frontais, ultrapassagens perigosas e conflitos entre fluxos opostos de veículos, fatores frequentemente associados a acidentes fatais ou de elevada gravidade. Além disso, vias duplicadas tendem a oferecer melhores condições operacionais, maior capacidade de tráfego e separação física entre sentidos opostos, reduzindo significativamente o potencial de acidentes graves.

Na análise do traçado da via, os maiores índices de gravidade foram observados em segmentos classificados como pontes, declives e curvas, todos apresentando valores superiores à média geral de gravidade. Os acidentes registrados em pontes apresentaram a maior taxa de gravidade entre todas as categorias analisadas. Esse resultado pode estar associado às limitações estruturais desses segmentos, como ausência de áreas de escape, largura reduzida da pista e menor margem para manobras evasivas em situações críticas.

Trechos em declive também apresentaram elevada severidade, possivelmente em

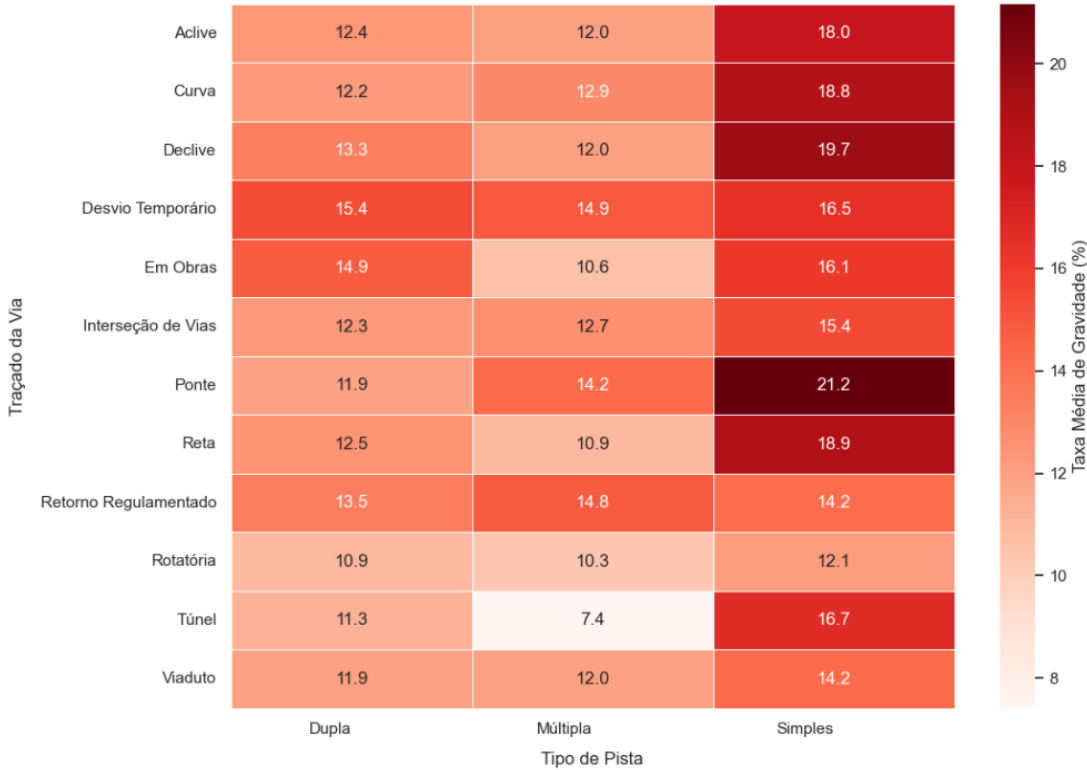
decorrência do aumento da velocidade operacional dos veículos, maior dificuldade de frenagem e redução do controle direcional, especialmente para veículos de carga. As curvas igualmente demonstraram níveis elevados de gravidade, evidenciando o impacto da geometria horizontal da via sobre a segurança operacional. Fatores como redução da visibilidade, excesso de velocidade e perda de estabilidade veicular podem contribuir para o aumento da severidade nesses segmentos.

Observou-se ainda que trechos classificados como desvio temporário e em obras apresentaram taxas elevadas de gravidade. Esses resultados sugerem que alterações provisórias nas condições operacionais da rodovia podem aumentar o risco de acidentes severos, especialmente quando associadas à sinalização inadequada, mudanças abruptas de fluxo ou redução da capacidade viária. Por outro lado, segmentos como rotatórias, viadutos e túneis apresentaram menores níveis relativos de gravidade. Uma possível explicação para esse comportamento é a redução natural da velocidade média dos veículos nesses locais, além da presença de elementos de controle operacional mais estruturados.

De maneira geral, os resultados demonstram que as características da infraestrutura rodoviária exercem influência significativa sobre a severidade dos acidentes. Elementos geométricos críticos, ausência de separação de fluxos e alterações operacionais temporárias mostraram-se associados a maiores níveis de risco.

A Figura 6.11 apresenta a taxa média de gravidade observada para cada combinação entre o tipo de pista e o traçado da via. Tons mais intensos representam maiores níveis de gravidade, indicando cenários potencialmente mais críticos para a segurança viária.

Figura 6.11 – Risco por Traçado da Via e Tipo de Pista.



Fonte: Elaboração própria com base em dados da PRF (2026).

Os resultados demonstram que as maiores taxas de acidentados graves concentram-se predominantemente em rodovias de pista simples, independentemente do traçado analisado. Esse comportamento reforça os resultados observados na análise individual, evidenciando que a ausência de separação física entre os fluxos de tráfego representa um importante fator de risco para acidentes graves. Entre as combinações analisadas, destacam-se os segmentos de ponte em pista simples e túnel em pista simples, ambos apresentando as maiores taxas médias de gravidade, superiores a 21%. Esses resultados sugerem que a combinação entre limitações estruturais do segmento e a configuração operacional da pista pode aumentar significativamente a severidade das ocorrências.

Os trechos em declive, reta e curva localizados em pistas simples também apresentaram níveis elevados de gravidade, com valores próximos ou superiores a 20%. Em declives, fatores como aumento da velocidade e maior distância de frenagem podem contribuir para o agravamento dos acidentes. Já em curvas, a redução da visibilidade e a possibilidade de perda de controle do veículo representam elementos críticos para a segurança viária.

Observa-se ainda que os segmentos classificados como desvio temporário apresentaram níveis elevados de gravidade em todos os tipos de pista analisados. Esse comportamento indica que alterações provisórias nas condições operacionais da rodovia podem gerar riscos adicionais independentemente da infraestrutura existente, especialmente quando associadas à redução da capacidade da via ou à sinalização inadequada.

Por outro lado, as menores taxas de acidentados graves foram observadas em segmentos de rotatória e em determinados trechos de pistas múltiplas, particularmente em túneis e áreas em obras. De maneira geral, as pistas múltiplas apresentaram os menores níveis médios de gravidade para a maior parte dos traçados analisados, evidenciando o papel da segregação de fluxos na redução da severidade dos acidentes.

De forma geral, a análise conjunta evidencia que a combinação entre traçados geometricamente complexos e rodovias de pista simples constitui um dos principais cenários de risco observados no estudo. Esses achados reforçam a importância de investimentos em duplicação de rodovias, adequação geométrica de segmentos críticos e melhoria da sinalização em trechos com histórico de elevada severidade dos acidentes.

## 6.5 Modelagem Estatística Multivariada por Partial Least Squares (PLS)

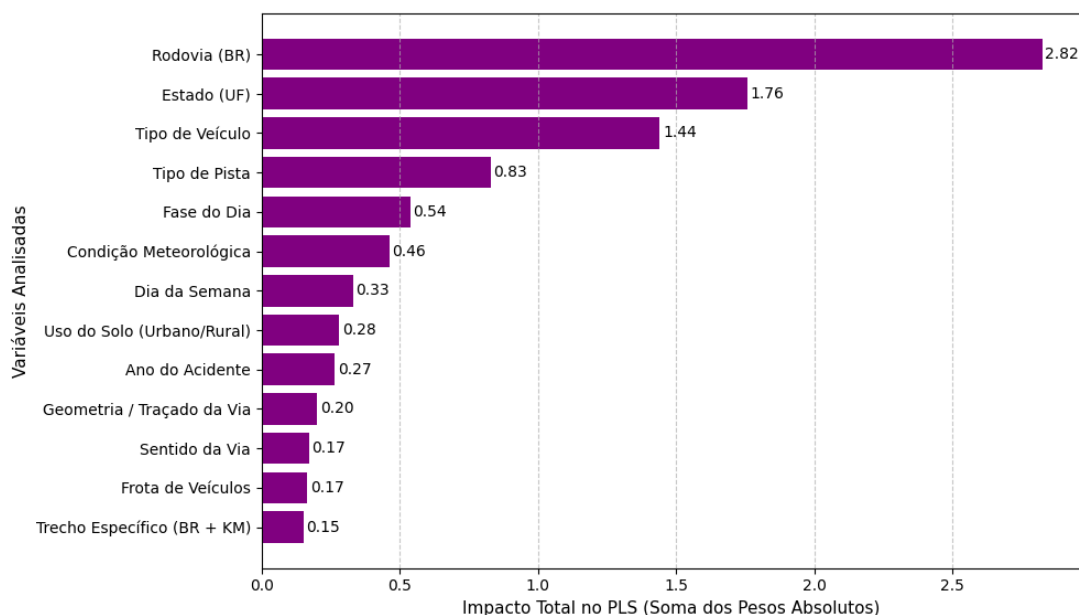
Com o objetivo de compreender a influência conjunta das variáveis explicativas sobre a variável resposta do estudo, foi aplicada a técnica de Modelagem Estatística Multivariada utilizando o método *Partial Least Squares* (PLS). Essa abordagem foi escolhida por sua capacidade de lidar com conjuntos de dados contendo múltiplas variáveis correlacionadas, permitindo identificar quais atributos apresentam maior contribuição global para a variabilidade do fenômeno analisado. Posteriormente seus resultados serão comparados aos resultados do modelo SHAP.

A robustez do PLS reside na capacidade de estruturar o impacto das variáveis por meio do critério quantitativo do *Escore VIP* (*Variable Importance in Projection*). Segundo a fundamentação matemática da técnica de Mínimos Quadrados Parciais (PLS), valores de Importância da Variável na Projeção (VIP) superiores a 1,0 indicam os atributos que possuem real relevância explicativa na sustentação do sistema (WOLD; SJOSTROM; ERIKSSON, 2001). O mapeamento desses escores VIP fornece uma linha de base estatística linear rigorosa, permitindo confrontar a relevância estrutural clássica com a sensibilidade não linear de alta dimensão capturada posteriormente pelos algoritmos de aprendizado de máquina.

O modelo PLS foi utilizado para estimar a importância relativa das variáveis originais do conjunto de dados, considerando a soma dos pesos absolutos associados aos componentes latentes gerados pelo algoritmo. Dessa forma, foi possível avaliar o impacto global de cada variável na gravidade dos acidentes.

A Figura 6.12 apresenta o resultado da importância linear global das variáveis originais obtidas pelo modelo PLS.

Figura 6.12 – Importância Linear das Variáveis (PLS) .



Fonte: Elaboração própria com base em dados da PRF e do Ministério dos Transportes (2026).

Os resultados revelam que o bloco de variáveis de cunho geoespacial detém a maior proeminência preditiva. A variável Rodovia (BR) posicionou-se no topo da hierarquia, registrando um escore VIP expressivo de 2,82, seguida pela variável Estado (UF), com 1,76. Esse comportamento evidencia que a localização geográfica e a identificação específica da rodovia federal exercem uma influência primária sobre a variável resposta, sugerindo que o perfil da malha viária regional e as características estruturais de cada rodovia ditam a linha de base da severidade dos sinistros. Logo abaixo, ainda superando o limiar crítico de significância ( $VIP > 1,0$ ), destaca-se a variável Tipo de Veículo (1,44), consolidando o fato de que as características operacionais do tráfego, notadamente a disparidade de massa e a cinemática do trauma associada aos diferentes modais, constituem um pilar explicativo essencial do modelo.

Em um segundo patamar, situam-se as variáveis de relevância moderada, cujos escores VIP posicionam-se abaixo da unidade, mas ainda exercem um papel de ajuste contextual no sistema. Nesse grupo, os aspectos de infraestrutura e visibilidade sobressaem-se, liderados por Tipo de Pista (0,83) e Fase do Dia (0,54), seguidos por Condição Meteorológica (0,46), Dia da Semana (0,33), Uso do Solo (Urbano/Rural) (0,28) e Ano do Acidente (0,27). A contribuição intermediária desses atributos indica que, embora não atuem como os macro-determinantes lineares, eles funcionam como importantes fatores condicionantes e agravantes, moldando as circunstâncias sob as quais o risco estabelecido pela localização e pelo tipo de veículo se materializa em maior gravidade.

Por fim, as variáveis Traçado da Via (0,20), Sentido da via (0,17), Frota de Veículos (0,17) e Trecho específico (KM + BR) (0,15) apresentaram escores VIP reduzidos,



situando-se na zona de baixa contribuição relativa. Esse resultado aponta que variações macro-temporais isoladas (ano) ou indicadores socioeconômicos agregados de forma linear (frota municipal) possuem capacidade explicativa residual quando confrontados com as variáveis situacionais e espaciais imediatas do sinistro.

Em síntese, a modelagem multivariada via PLS ratifica que os fatores espaciais, estruturais e de tipologia veicular dominam a explicação da severidade em detrimento de componentes puramente temporais. Este mapeamento linear robusto cumpre uma função metodológica crucial: estabelece a linha de base estatística necessária para a comparação subsequente com o framework SHAP aplicado ao modelo XGBoost. Essa transição permitirá validar a consistência dos preditores e verificar como as relações lineares aqui identificadas se comportam diante de um paradigma de aprendizado de máquina capaz de capturar interações complexas de alta dimensão e não linearidades.

## 7 Resultados e Discussões

Neste capítulo, são apresentados os resultados obtidos a partir do treinamento e avaliação dos modelos de aprendizado de máquina aplicados à predição da gravidade de acidentes de trânsito. A discussão abrange a comparação de desempenho entre as abordagens linear e não-linear, bem como a interpretabilidade XGBoost através dos valores SHAP.

### 7.1 Avaliação de Desempenho dos Modelos Preditivos

Após o término do processo de busca estocástica (*Randomized Search*), as configurações ideais foram extraídas para ambos os classificadores. A Tabela 7.1 consolida os resultados das métricas de desempenho calculadas sobre o conjunto de dados de teste, dados inéditos totalizando 813.995 observações.

Tabela 7.1 – Hiperparâmetros otimizados e desempenho final dos modelos

| Modelo              | Hiperparâmetros selecionados | Desempenho Global          | Classe 0 (Não Graves) | Classe 1 (Graves) |
|---------------------|------------------------------|----------------------------|-----------------------|-------------------|
| Regressão Logística | C: 0,0121                    | Acurácia: 67%              | Precisão: 0,92        | Precisão: 0,28    |
|                     | l1_ratio: 0,7220             | F1-Score (Macro): 0,58     | Recall: 0,66          | Recall: 0,70      |
|                     | class_weight: balanced       | F1-Score (Ponderado): 0,71 | F1-Score: 0,77        | F1-Score: 0,40    |
| XGBoost             | n_estimators: 607            | Acurácia: 85%              | Precisão: 0,96        | Precisão: 0,52    |
|                     | learning_rate: 0,2987        | F1-Score (Macro): 0,77     | Recall: 0,87          | Recall: 0,78      |
|                     | max_depth: 9                 | F1-Score (Ponderado): 0,86 | F1-Score: 0,91        | F1-Score: 0,62    |
|                     | subsample: 0,9343            |                            |                       |                   |
|                     | colsample_bytree: 0,7917     |                            |                       |                   |
|                     | scale_pos_weight: 3,8571     |                            |                       |                   |

Fonte: Elaboração própria (2026).

A calibração dos classificadores revelou comportamentos distintos e informativos sobre a estrutura matemática da base de dados, evidenciando como cada algoritmo lidou com a complexidade e o desbalanceamento das classes.

A Regressão Logística foi otimizada sob a penalização Elastic Net, que combina as propriedades de regularização L1 (Lasso) e L2 (Ridge). O comportamento dos hiperparâmetros selecionados reflete uma postura defensiva do estimador linear:

1. **Força de Regularização ( $C = 0,01209$ ) no limite inferior:** O parâmetro  $C$  representa o inverso da força de regularização ( $C = 1/\lambda$ ). Dado que o espaço de

busca estendia-se de 0,01 a 100, o valor ótimo selecionado de 0,01209 encontra-se virtualmente colado ao limite inferior. Isso indica uma penalização extremamente alta. O modelo foi forçado a encolher os coeficientes das variáveis para manter a estabilidade matemática e mitigar o *overfitting*.

2. **Dominância da Penalidade L1 (`l1_ratio` = 0,72200):** Com o índice inclinado para 1, o otimizador priorizou a penalidade Lasso. Na prática, a regularização L1 zera coeficientes de variáveis redundantes ou irrelevantes. Isso revela que, sob uma perspectiva linear, muitas características do banco de dados apresentavam alta colinearidade ou ruído, exigindo um descarte automático de atributos para garantir a convergência.
3. **Compensação de Desbalanceamento (`class_weight='balanced'`):** A escolha categórica pelo peso balanceado confirma a forte assimetria do banco de dados.

Em contrapartida à simplicidade forçada do modelo linear, o XGBoost adaptou sua arquitetura para capturar as interações não-lineares do trânsito com alta eficiência:

1. **Ação Simplificadora da Taxa de Aprendizado (`learning_rate`  $\approx$  0,30) e Profundidade Moderada (`max_depth` = 9):**
  - A Taxa de Aprendizado (`learning_rate` = 0,29865) posicionou-se no topo do espaço de busca (limite de 0,31). No contexto do gradient boosting, uma taxa de aprendizado mais elevada atua como um elemento de simplificação da estrutura do modelo, pois confere maior peso factual às correções de cada árvore individual. Isso acelera a convergência do algoritmo e evita a necessidade de um encadeamento excessivamente complexo de estimadores.
  - A Profundidade Máxima (`max_depth` = 9), embora próxima do limite superior estabelecido no espaço de busca (10), reflete uma profundidade estrutural moderada perante a elevada dimensionalidade e heterogeneidade dos dados viários. A combinação de uma taxa de aprendizado alta com a profundidade 9 permitiu ao modelo capturar interações essenciais de ordens superiores (como a relação simultânea entre tipo de veículo, traçado geométrico e localização específica) sem a necessidade de estruturas de decisão demasiadamente profundas ou propensas ao *overfitting*.
2. **Regularização Estocástica Ativa (`subsample` = 0,93430 e `colsample_bytree` = 0,79168):** Para controlar o ajuste do modelo e garantir a generalização, o algoritmo utilizou subamostragens estocásticas. Ao restringir o treinamento de cada árvore a 93,4% das observações e apenas 79,1% das colunas (variáveis) disponíveis, o otimizador impediu que variáveis altamente dominantes (como Tipo de Veículo)

monopolizassem as decisões, forçando o algoritmo a extrair padrões valiosos dos atributos secundários.

3. **Sintonia Fina do Desbalanceamento (`scale_pos_weight` = 3,85712):** Ao invés de uma compensação rígida (como o `balanced` da Regressão Logística), o XGBoost realizou uma calibração contínua de pesos, definindo que cada erro na classe de acidentes graves custaria exatamente 3,86 vezes mais do que um erro na classe não grave. Esse ajuste cirúrgico permitiu que o modelo maximizasse a identificação de acidentes severos mantendo a precisão controlada.

Avaliando o desempenho global dos classificadores, constata-se que o modelo estatístico linear de Regressão Logística obteve uma Acurácia de 67%, enquanto o algoritmo de gradient boosting XGBoost alcançou um patamar significativamente superior, registrando 85% de acertos globais. Contudo, em virtude do forte desbalanceamento amostral intrínseco aos dados de trânsito — onde a Classe 1 (Acidentes Graves) representa uma fração minoritária —, a validação do potencial de aplicação prática exige uma avaliação minuciosa das métricas isoladas por classe e das respectivas matrizes de confusão.

No que tange à Classe 1 (Acidentes Graves), que constitui o foco principal de interesse analítico deste estudo, as discrepâncias estruturais entre os modelos tornam-se ainda mais evidentes. A Regressão Logística atuou como um balizador (baseline) importante. Ao fazer uso da estratégia de penalização por pesos (`class_weight='balanced'`), o modelo alcançou uma Sensibilidade (*Recall*) razoável de 0,70, conseguindo identificar corretamente 88.267 acidentes graves. No entanto, essa sensibilidade cobrou um preço elevado em sua Precisão, que despencou para 0,28.

A matriz de confusão da Regressão Logística elucida esse comportamento: o modelo gerou expressivos 231.130 Falsos Positivos (acidentes não graves classificados erroneamente como graves). Praticamente, esse excesso de alarmes falsos compromete a aplicabilidade do modelo linear, pois saturaria recursos operacionais de resgate ou fiscalização em cenários sem gravidade real.

Em contrapartida, o modelo XGBoost superou substancialmente a abordagem linear em todas as dimensões de avaliação na classe minoritária. O algoritmo de aprendizado baseado em árvores impulsionadas atingiu uma Sensibilidade (*Recall*) de 0,78, capturando com sucesso 99.289 acidentes graves (11.022 a mais que a Regressão Logística). Paralelamente, houve uma expressiva evolução na sua precisão, que subiu para 0,52. Esse ganho reduziu os Falsos Positivos de 231.130 para apenas 92.594 ocorrências.

Sob a ótica da segurança pública e da gestão de tráfego, a redução simultânea dos Falsos Negativos (acidentes graves ignorados pelo sistema) de 38.733 na Regressão Logística para 27.711 no XGBoost é de suma relevância. O ganho generalizado de performance reflete-se diretamente no F1-Score da classe crítica, saltando de 0,40 para 0,62.

Esses resultados comprovam empiricamente que a dinâmica para a gravidade de um acidente de trânsito é caracterizada por interações complexas e estritamente não-lineares entre os fatores ambientais, humanos e viários. Consequentemente, as fronteiras adaptativas e os mecanismos estocásticos do XGBoost, devidamente ajustados e validados por divisões cruzadas de dados, consolidam o algoritmo como a ferramenta preditiva mais robusta e recomendada para subsidiar tomadas de decisão preventivas em segurança viária.

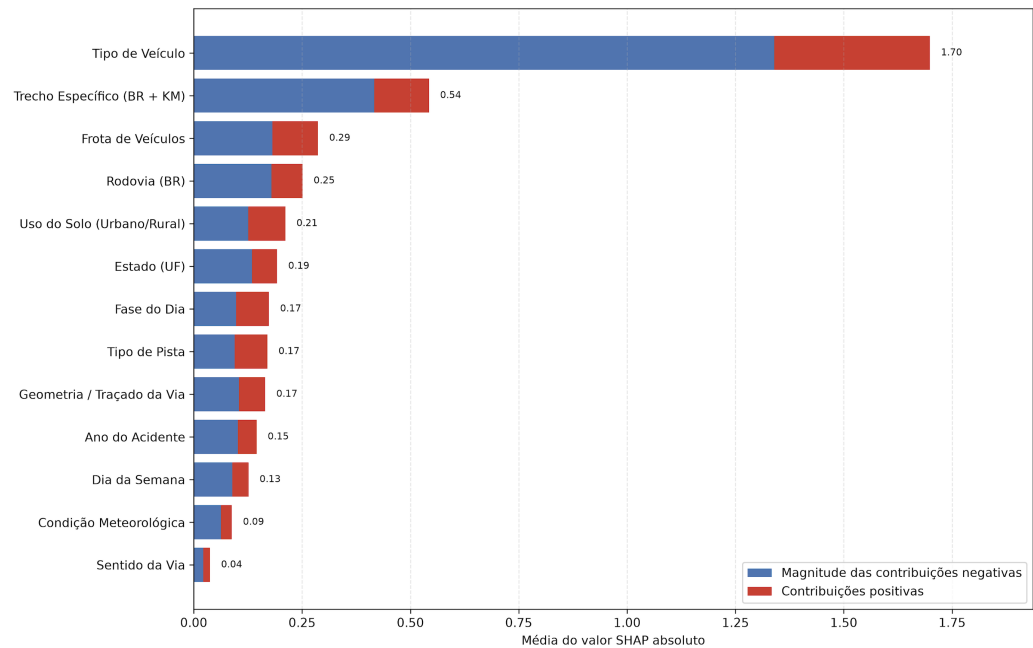
## 7.2 Análise via SHAP: Fatores de Risco

Diante dos resultados de desempenho significativamente superiores alcançados pelo XGBoost, a análise detalhada dos fatores determinantes da severidade dos sinistros de trânsito foi conduzida sob a abordagem não-linear deste algoritmo acoplado ao framework de interpretabilidade agnóstica SHAP. Os achados, padrões de comportamento e relevância de variáveis revelados por essa metodologia explicativa serão, posteriormente, comparados com os achados do capítulo 6.

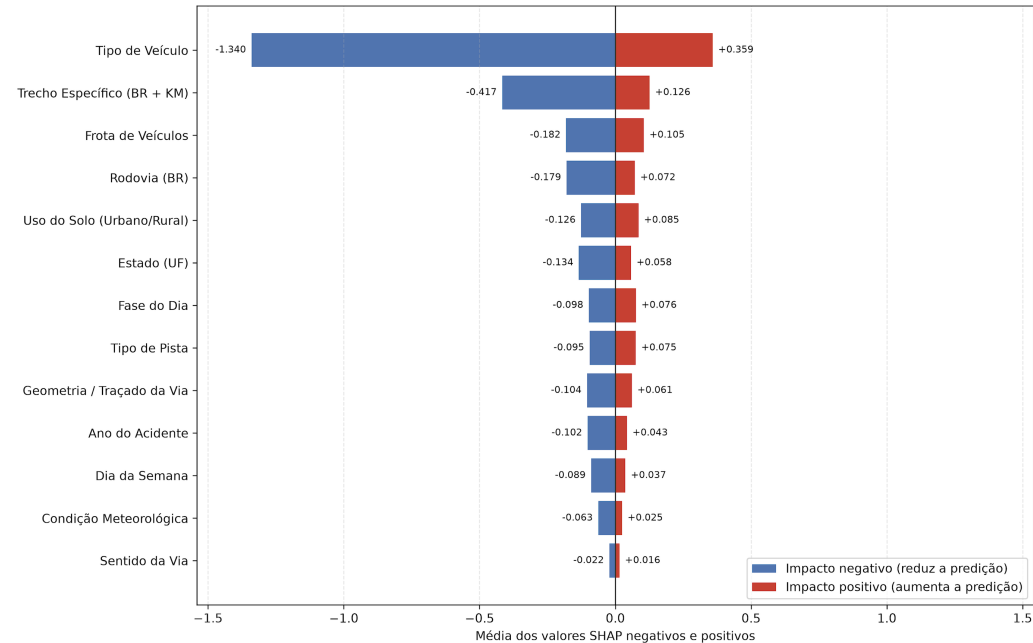
O primeiro nível de interpretação do modelo principal ocorre por meio da análise global dos valores SHAP. A Figura 7.1 apresenta duas perspectivas complementares: a importância média absoluta de cada variável, mensurada por  $mean(|SHAP\ value|)$ , e sua decomposição em contribuições médias positivas e negativas. Enquanto o primeiro gráfico estabelece a hierarquia das variáveis de acordo com a magnitude de sua influência, independentemente da direção, o segundo demonstra como essa influência se distribui entre contribuições que elevam ou reduzem a saída do modelo para a classe correspondente a lesão grave ou óbito.

Figura 7.1 – Análise Global da Influência das Variáveis nas Predições do Modelo

(a) Importância SHAP média absoluta



(b) Contribuições SHAP médias positivas e negativas



Fonte: Elaboração própria com base em dados da PRF e do Ministério dos Transportes (2026).

Sob a perspectiva da importância absoluta, o Tipo de Veículo apresentou a maior influência sobre as predições, com valor SHAP médio absoluto de 1,70, superior em mais de três vezes ao observado para o Trecho Específico (BR + KM), com 0,54. Em seguida, destacou-se a Frota de Veículos, com importância absoluta média de 0,29. Esses resultados

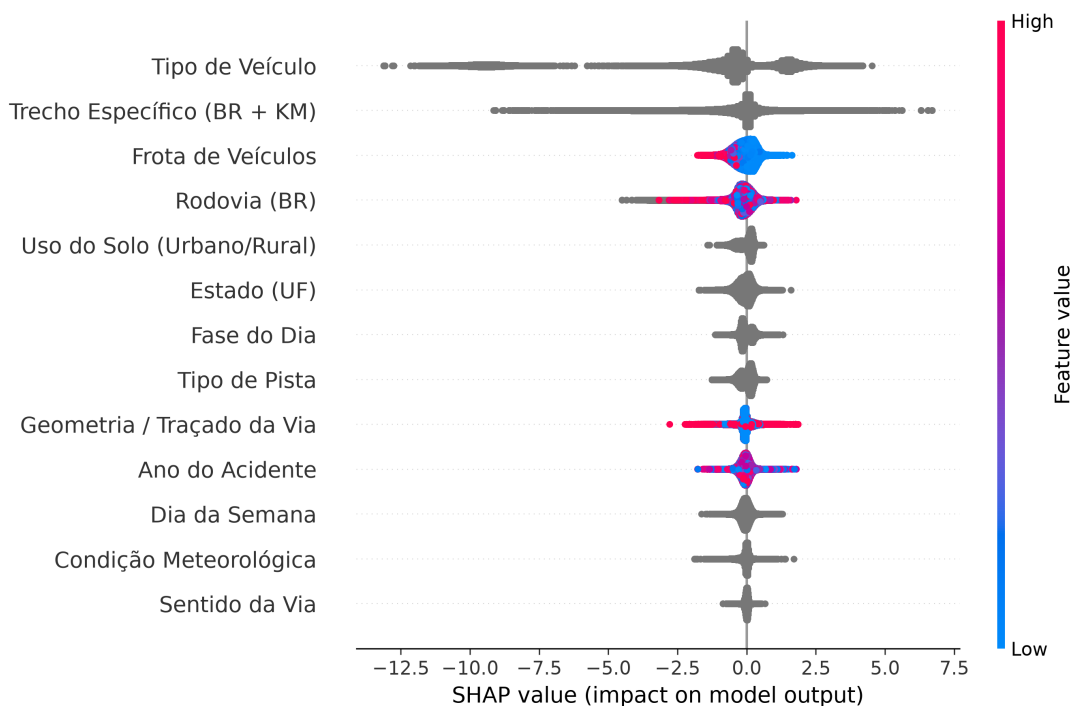
demonstram que características relacionadas ao modal de transporte, à localização do acidente e ao contexto municipal constituíram os principais fatores discriminantes utilizados pelo modelo na estimativa da gravidade.

A decomposição direcional complementa essa análise ao demonstrar que a influência dessas variáveis não ocorre em um único sentido. Considerando especificamente as contribuições positivas, as três variáveis que mais elevaram a saída do modelo para a classe de maior gravidade foram Tipo de Veículo, com contribuição positiva média de +0,359; Trecho Específico (BR + KM), com +0,126; e Frota de Veículos, com +0,105. Isso indica que determinadas categorias veiculares, localizações rodoviárias e condições associadas ao contexto municipal contribuíram mais intensamente para elevar a estimativa de que uma pessoa envolvida em um acidente apresentasse lesão grave ou evoluísse para óbito.

Entretanto, essas mesmas variáveis também apresentaram contribuições negativas médias de  $-1,340$ ,  $-0,417$  e  $-0,182$ , respectivamente. Portanto, seus efeitos dependem da categoria ou do valor observado: determinadas condições elevam a saída do modelo, enquanto outras a reduzem. O predomínio das contribuições negativas não permite classificar essas variáveis, de maneira geral, como fatores de proteção, pois o resultado também depende da distribuição das observações e do valor-base utilizado pelo SHAP.

O *Beeswarm Plot*, apresentado na Figura 7.2, aprofunda essa interpretação ao representar a distribuição dos valores SHAP de cada observação individual. Diferentemente dos gráficos globais da Figura 7.1, que sintetizam os resultados por médias, o Beeswarm Plot permite visualizar a variabilidade, a concentração e os valores extremos das contribuições.

Figura 7.2 – Importância Global das Variáveis (SHAP)



Fonte: Elaboração própria com base em dados da PRF e do Ministério dos Transportes (2026).

A distribuição dos valores para a variável Tipo de Veículo evidencia sua elevada capacidade discriminatória na predição da severidade dos acidentes. Observa-se uma distribuição assimétrica, caracterizada por uma longa cauda negativa e uma elevada concentração de observações na região positiva do eixo SHAP. Esse comportamento indica que determinadas categorias veiculares contribuem para reduzir a probabilidade prevista de acidentes fatais, enquanto outras aumentam significativamente essa probabilidade. Essa interpretação é consistente com os resultados da análise descritiva, na qual motocicletas apresentaram as maiores taxas de gravidade, ao passo que veículos de carga e transporte coletivo, como caminhões, caminhões-trator, semirreboques e ônibus, registraram taxas inferiores ou próximas à média nacional. Além disso, o resultado reforça os fundamentos apresentados no referencial teórico, segundo os quais a elevada vulnerabilidade dos motociclistas decorre da ausência de estruturas de proteção e da maior exposição às forças de impacto, especialmente em rodovias de alta velocidade.

A variável Trecho Específico (BR + KM) também apresenta ampla dispersão dos valores SHAP, distribuindo-se em ambos os lados do eixo. Esse comportamento demonstra que diferentes segmentos da malha rodoviária exercem influências distintas sobre a severidade dos acidentes. Enquanto determinados trechos contribuem para reduzir a probabilidade de ocorrência de acidentes graves, outros apresentam contribuição positiva para a severidade prevista pelo modelo. Esse resultado reforça as evidências observadas na análise exploratória, que identificou diferenças expressivas entre regiões do país, associadas



às características da infraestrutura rodoviária e às condições operacionais das vias.

Para a variável Frota de Veículos, o *Beeswarm Plot* revela um padrão consistente de distribuição das cores. Observa-se que valores reduzidos da variável (representados em azul) concentram-se predominantemente na região positiva do eixo SHAP, enquanto valores elevados (representados em vermelho) tendem a localizar-se na região negativa. Esse comportamento indica uma relação inversa entre a frota municipal de veículos e a contribuição da variável para a severidade prevista pelo modelo. Em termos interpretativos, municípios com menor frota apresentam maior tendência de contribuir para predições de acidentes graves, ao passo que municípios com maior frota tendem a exercer efeito atenuador. Esse padrão é coerente com a análise exploratória apresentada anteriormente, que evidenciou maior gravidade em determinadas regiões caracterizadas por menor densidade urbana e diferentes condições da infraestrutura rodoviária, indicando que o contexto territorial influencia a severidade dos acidentes mesmo quando considerado em conjunto com as demais variáveis do modelo.

### 7.2.1 Análise de Contraste: Limitações do Modelo de Comparação (PLS) frente ao XGBoost/SHAP

A aplicação do modelo de PLS, apresentada na Figura 6.6.1, teve como objetivo estabelecer uma base comparativa para avaliar o comportamento de uma abordagem linear em relação ao modelo principal deste estudo. A comparação entre os resultados do PLS e do XGBoost interpretado por SHAP evidenciou diferenças relevantes na forma como cada técnica representa os fatores associados à severidade dos acidentes, indicando limitações da modelagem linear para capturar as relações presentes no conjunto de dados analisado.

De acordo com a literatura sobre regressão por Mínimos Quadrados Parciais, valores de Variable Importance (VI) superiores a 1,0 indicam variáveis com contribuição relevante para a capacidade explicativa do modelo. Sob esse critério, o PLS identificou apenas três atributos como relevantes: Rodovia (BR) (VIP = 2,82), Estado (UF) (VIP = 1,76) e Tipo de Veículo (VIP = 1,44). As demais variáveis apresentaram valores inferiores ao limiar de relevância, sugerindo baixa contribuição para a explicação linear da severidade dos acidentes.

A principal divergência entre os modelos concentra-se nas variáveis Trecho Específico (BR + KM) e Frota de Veículos. Enquanto a análise SHAP posicionou o Trecho Específico como a segunda variável de maior impacto global na predição da severidade (+0,54), o PLS atribuiu a essa mesma variável um VIP de apenas 0,15, classificando-a entre as menos relevantes. Situação semelhante foi observada para a variável Frota de Veículos, que ocupou a terceira posição na análise SHAP (+0,29), mas apresentou VIP igual a 0,17 no modelo linear. A Tabela 7.2 sintetiza essas diferenças de comportamento entre os dois modelos.

Tabela 7.2 – Comparação de impacto e posição: XGBoost (SHAP) vs. Modelo Linear (PLS)

| Variável Analisada | Posição e Im-<br>pacto no XGBo-<br>ost (SHAP) | Posição e Im-<br>pacto no Modelo<br>Linear (PLS) | Comportamento<br>Capturado                                           |
|--------------------|-----------------------------------------------|--------------------------------------------------|----------------------------------------------------------------------|
| Tipo de Veículo    | 1º Lugar (Impacto<br>Global: +1,70)           | 3º Lugar (Impacto<br>Linear: 1,44)               | Consistência (Im-<br>pacto dominante<br>em ambos os espec-<br>tros)  |
| Trecho Específico  | 2º Lugar (Impacto<br>Global: +0,54)           | 13º Lugar (Impacto<br>Linear: 0,15)              | Ocultamento Li-<br>near (Relação<br>puramente local e<br>não linear) |
| Frota de Veículos  | 3º Lugar (Impacto<br>Global: +0,29)           | 12º Lugar (Impacto<br>Linear: 0,17)              | Ocultamento Li-<br>near (Interação<br>complexa socioeco-<br>nômica)  |
| Rodovia (BR)       | 4º Lugar (Impacto<br>Global: +0,25)           | 1º Lugar (Impacto<br>Linear: 2,82)               | Hipervalorização<br>Macro (Captura<br>apenas tendências<br>em bloco) |

Fonte: Elaboração própria (2026).

Essa diferença pode ser explicada pelas características matemáticas dos dois méto- dos. O PLS busca identificar componentes latentes que maximizam a covariância linear entre as variáveis explicativas e a variável resposta. Como consequência, sua capacidade de representação é favorecida quando existem relações aproximadamente lineares e de comportamento global.

Por outro lado, variáveis como Trecho Específico (BR + KM) representam infor- mações altamente localizadas. Hipoteticamente, segmentos consecutivos de uma mesma rodovia podem apresentar níveis bastante distintos de severidade em função de característi- cas específicas de cada trecho. Nessas situações, a ausência de uma tendência linear global faz com que o efeito médio da variável seja reduzido no PLS, diminuindo sua importância estatística.

Em contraste, o XGBoost, fundamentado em árvores de decisão aditivas otimizadas por gradiente, realiza sucessivas partições do espaço amostral, permitindo capturar relações não lineares, efeitos locais e interações complexas entre as variáveis. Essa característica possibilita que determinados trechos rodoviários apresentem elevada contribuição para a predição da severidade mesmo quando esse comportamento não é observado de forma homogênea em toda a malha rodoviária.

Observa-se ainda que o modelo linear atribuiu maior importância às variáveis Rodo-

via (BR) e Estado (UF), ambas associadas a informações geográficas mais agregadas. Esse comportamento sugere que o PLS representa com maior eficiência tendências macroscópicas presentes no conjunto de dados, enquanto apresenta menor sensibilidade para identificar fatores associados a segmentos específicos da infraestrutura rodoviária.

Os resultados obtidos são coerentes com a análise exploratória e com a interpretação fornecida pelo SHAP, que evidenciaram a relevância das características locais da rodovia para a severidade dos acidentes. Dessa forma, o contraste entre os modelos indica que parte importante das relações presentes nos dados possui natureza não linear, aspecto capturado de maneira mais adequada pelo XGBoost.

Sob a perspectiva da aplicação prática, essa diferença possui implicações relevantes para o apoio à tomada de decisão. Enquanto o modelo linear fornece uma visão mais agregada dos fatores associados à severidade, o XGBoost permite identificar segmentos específicos da malha rodoviária que exercem maior influência sobre o risco de acidentes graves. Essa capacidade torna o modelo mais adequado para subsidiar ações direcionadas de monitoramento, priorização de investimentos e planejamento de intervenções em trechos críticos das rodovias federais.

Em síntese, a comparação entre o XGBoost interpretado por SHAP e o modelo PLS demonstra que a severidade dos acidentes rodoviários não é explicada exclusivamente por relações lineares entre as variáveis analisadas e fatores ‘macroscópicos’. A importância atribuída pelo XGBoost a fatores como Trecho Específico (BR + KM) e Frota de Veículos evidencia a existência de padrões complexos e localizados que não foram plenamente capturados pelo modelo linear. Esses resultados corroboram a escolha do XGBoost como modelo principal deste estudo e reforçam o potencial da combinação entre aprendizado de máquina e técnicas de Inteligência Artificial Explicável para apoiar a compreensão dos fatores associados à severidade dos acidentes em rodovias federais.

## 8 Desenvolvimento da Aplicação

Para operacionalizar o modelo preditivo de Machine Learning e torná-lo acessível a usuários finais, foi desenvolvida uma aplicação web interativa. O objetivo desta ferramenta é permitir a inserção de dados contextuais, ambientais e locais, e retornar a probabilidade de ocorrência de acidentes, acompanhada de uma explicação visual sobre quais fatores mais influenciaram esse resultado.

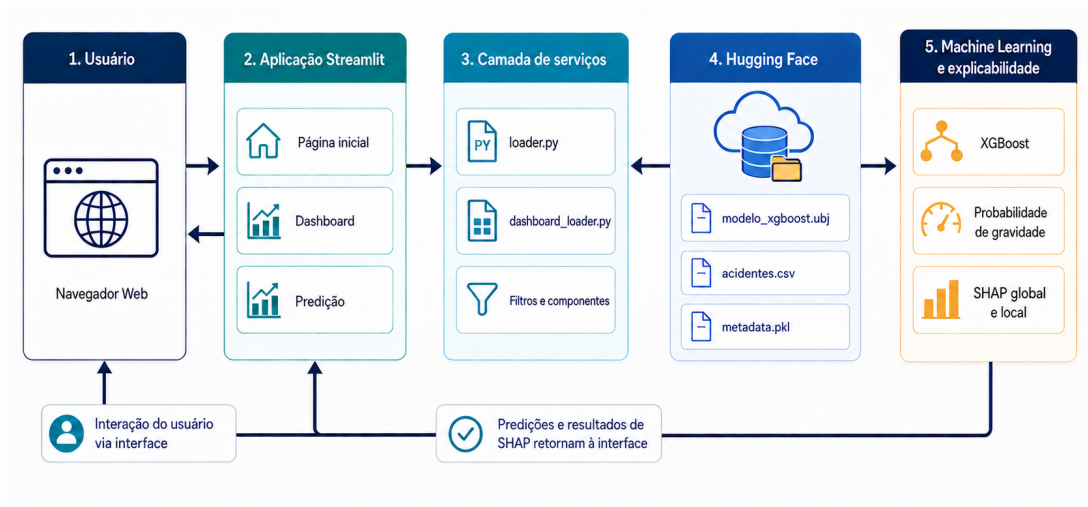
A seguir, são detalhados a arquitetura técnica adotada e as principais funcionalidades do painel.

### 8.1 Arquitetura da Aplicação

A interface foi desenvolvida utilizando o framework Streamlit, responsável pela construção das páginas, dos formulários, dos filtros e dos elementos visuais. A manipulação dos dados é realizada principalmente com as bibliotecas Pandas e NumPy, enquanto os gráficos interativos do dashboard são produzidos com a biblioteca Plotly. A arquitetura da aplicação foi estruturada de maneira modular, separando a interface visual das funções responsáveis pelo carregamento, processamento e interpretação dos dados.

O modelo XGBoost e os demais artefatos necessários à aplicação são armazenados externamente no Hugging Face Hub. Essa decisão foi adotada em razão do tamanho dos arquivos gerados durante o projeto, especialmente o modelo treinado e a base utilizada pelo dashboard, que ultrapassam os limites convencionais de armazenamento do repositório de código. Durante a inicialização da aplicação, esses arquivos são recuperados e armazenados temporariamente no ambiente de execução. A arquitetura está representada de forma visual na Figura 8.1.

Figura 8.1 – Arquitetura da Aplicação



Fonte: Elaboração própria (2026).

O fluxo da aplicação inicia-se com a interação do usuário na interface Streamlit. Dependendo da funcionalidade acessada, a aplicação carrega a base histórica para alimentar o dashboard ou encaminha as informações preenchidas no formulário para o módulo de predição. Nesse segundo fluxo, as variáveis são preparadas, submetidas ao modelo XGBoost e posteriormente encaminhadas ao módulo de explicação SHAP.

## 8.2 Funcionalidades

O painel foi desenhado para ir além de uma simples "caixa preta" preditiva, oferecendo não apenas a previsão, mas também transparência através de explicabilidade e análise dos dados históricos. As funcionalidades voltadas ao usuário são Página Inicial, Dashboard e Predição.

A Página Inicial (Figura 8.2) apresenta uma visão geral do projeto, incluindo o problema estudado, o objetivo da aplicação e as principais funcionalidades disponíveis. Essa página também introduz o significado da classificação adotada pelo modelo, diferenciando acidentes sem lesões graves daqueles que resultaram em lesões graves ou óbito. Além de contextualizar a solução, a página traz as métricas de desempenho do modelo e importância global das variáveis estimada pelo SHAP que demonstra quais variáveis possuem maior influência média absoluta no comportamento geral do modelo.

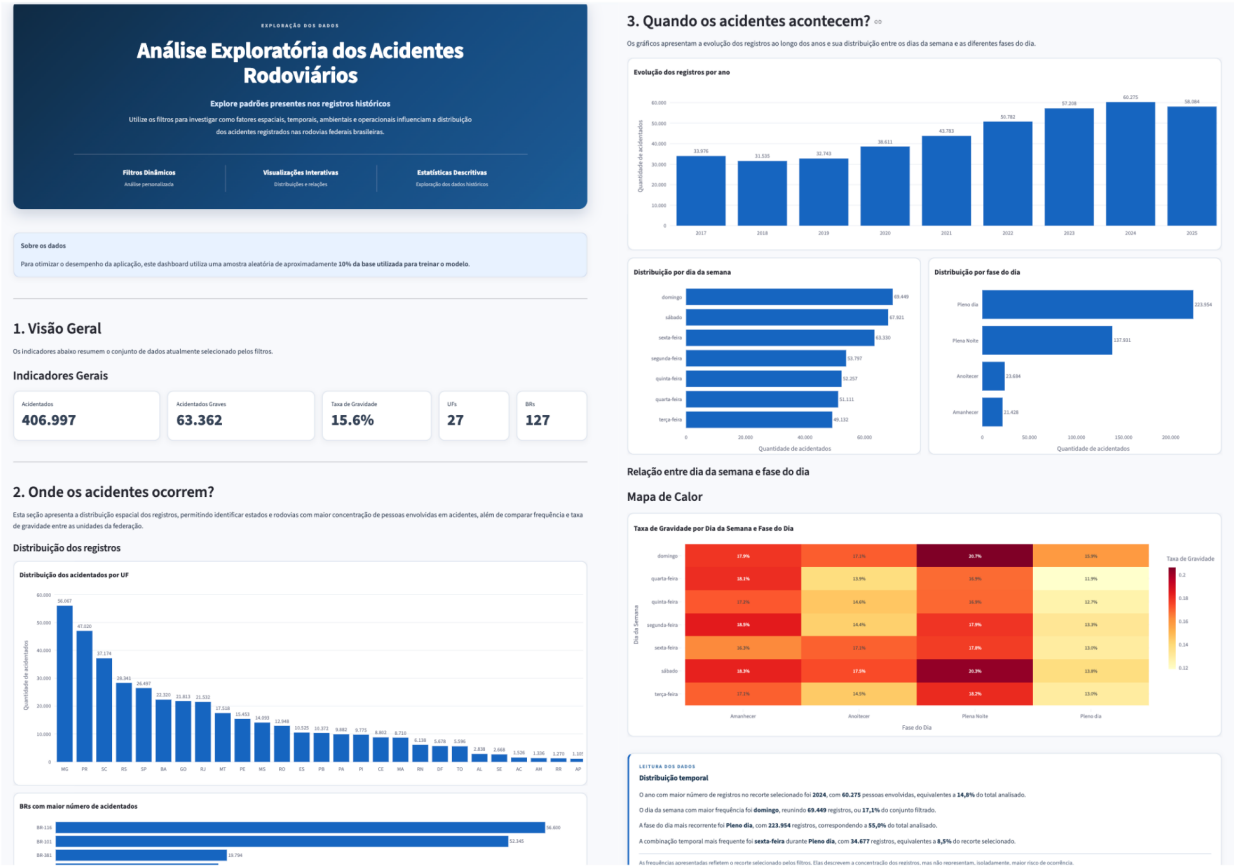
Figura 8.2 – Página Inicial da Aplicação



Fonte: Elaboração própria (2026).

O Dashboard foi desenvolvido para possibilitar a exploração dos dados históricos utilizados no projeto. A página apresenta indicadores gerais, estatísticas e visualizações interativas relacionadas aos acidentes registrados nas rodovias federais brasileiras entre 2017 e 2025. Os filtros permitem restringir os dados de acordo com diferentes características, possibilitando a análise de recortes específicos. A atualização dos indicadores e gráficos ocorre conforme as opções selecionadas, permitindo observar diferenças na frequência e na gravidade dos acidentes entre regiões, períodos e condições distintas. A Figura 8.3 apresenta parte do Dashboard desenvolvido.

Figura 8.3 – Dashboard da Aplicação



Fonte: Elaboração própria (2026).

A página Predição, representada na Figura 8.4, permite informar as características de um acidente e obter uma estimativa de sua gravidade. O formulário reúne as mesmas variáveis utilizadas durante o treinamento do modelo.

Figura 8.4 – Formulário da Página de Predição

**Predição da Gravidade de Acidentes**

Estime o risco associado a um cenário rodoviário

Informe as características do local, da via, do ambiente e do veículo para estimar a probabilidade de uma pessoa envolvida em um acidente apresentar lesões graves ou evoluir para óbito.

**XGBoost**  
Modelo preditivo

**Probabilidade**  
Estimativa individual

**SHAP**  
Explicação da previsão

Esta versão demonstrativa utiliza um modelo treinado com uma amostra de 10% dos dados disponíveis. Os resultados devem ser interpretados como uma prova de conceito.

**Simulação do cenário**

Preencha os campos abaixo com as características do cenário que deseja analisar. O resultado representa uma estimativa probabilística baseada nos padrões identificados pelo modelo nos dados históricos.

**1. Localização e contexto municipal**

Selecione a unidade federativa, a rodovia, o quilômetro e informe a frota municipal.

UF: BA Rodovia (BR): BR-110 Quilômetro: 33 Frota Municipal: 16000,00

**2. Características da via**

Tipo de pista: Dupla Perímetro Urbano: Não

Traçado da via: Curva x Declive

**3. Contexto temporal e ambiental**

Ano: 2025 Dia da semana: sábado Fase do dia: Plena Noite

Condição meteorológica: Céu Claro Sentido da via: Crescente

**4. Veículo**

Tipo de veículo: Motocicleta

Calcular probabilidade

Fonte: Elaboração própria (2026).

Após o preenchimento, a aplicação valida os valores informados e prepara o registro de entrada. Parte desse processamento envolve a construção do atributo que combina a rodovia e o quilômetro, denominado Trecho Específico (BR + KM), além da organização das variáveis categóricas e binárias no mesmo formato utilizado durante o treinamento.

Em seguida, o registro é submetido ao modelo XGBoost. Como resultado, a aplicação apresenta a classe prevista e a probabilidade estimada de o acidente resultar em lesão grave ou óbito. Além da probabilidade estimada, a aplicação apresenta uma explicação local gerada pelo framework SHAP, que mostra como cada atributo contribui para uma predição específica. Os valores SHAP positivos indicam características que contribuíram para aumentar a probabilidade estimada de gravidade, enquanto os valores negativos representam características que contribuíram para reduzi-la. A Figura 8.5 traz um exemplo dos resultados da página de predição.



Figura 8.5 – Exemplo de Resultados da Página de Predição



Fonte: Elaboração própria (2026).

## 8.3 Disponibilização da Aplicação

A aplicação foi publicada por meio do Streamlit Community Cloud, permitindo seu acesso diretamente pelo navegador, sem a necessidade de instalação local. O código-fonte foi mantido em um repositório separado daquele utilizado no treinamento, estabelecendo uma distinção entre o ambiente de desenvolvimento do modelo e o ambiente de disponibilização da interface. Os arquivos de maior volume foram hospedados no Hugging Face Hub e recuperados pela aplicação durante sua execução. Essa arquitetura possibilitou manter o repositório da interface mais leve e compatível com os limites da plataforma de versionamento. A aplicação pode ser acessada pelo endereço: <https://predicaoacidentesapp.streamlit.app/>

## 9 Considerações Finais

O presente trabalho atingiu seu objetivo principal ao desenvolver e operacionalizar um modelo preditivo baseado no algoritmo XGBoost, capaz de estimar a probabilidade de uma pessoa envolvida em um acidente em rodovias federais brasileiras apresentar lesão grave ou evoluir para óbito. Ao combinar a capacidade preditiva proporcionada pelas técnicas de Machine Learning com a transparência oferecida pela explicabilidade por meio dos valores SHAP, a ferramenta desenvolvida supera uma das principais limitações dos modelos complexos de inteligência artificial, frequentemente caracterizados como "caixas-pretas", tornando-se um sistema de apoio à decisão capaz de fornecer resultados claros, interpretáveis e tecnicamente fundamentados.

Além do desenvolvimento do modelo preditivo, este trabalho resultou na implementação de uma aplicação web interativa que integra todas as etapas necessárias para a utilização prática da solução proposta. A aplicação permite que o usuário informe características estruturais, temporais e ambientais relacionadas ao cenário do acidente, retornando em tempo real a estimativa da probabilidade de ocorrência de lesão grave ou óbito, acompanhada da explicação individual da predição por meio da técnica SHAP. Essa integração aproxima os resultados obtidos no ambiente acadêmico de um contexto de utilização prática, tornando as previsões acessíveis mesmo para usuários sem conhecimento especializado em aprendizado de máquina.

A aplicação desenvolvida apresenta potencial para auxiliar diretamente instituições responsáveis pela gestão do tráfego e da infraestrutura rodoviária nacional. No contexto da Polícia Rodoviária Federal (PRF), o modelo permite a adoção de uma abordagem preventiva para o planejamento operacional. A partir da inserção de condições esperadas para um determinado período, como fatores meteorológicos, variações sazonais do fluxo veicular e características temporais relacionadas à fase do dia, a ferramenta pode fornecer estimativas de risco que auxiliem na otimização da distribuição de viaturas e agentes. Dessa forma, cenários associados a maior probabilidade de ocorrência de lesões graves ou óbitos podem receber atenção antecipada, contribuindo para o planejamento operacional, para ações preventivas direcionadas e para a melhoria da capacidade de resposta em situações de emergência.

Da mesma forma, a ferramenta oferece suporte ao Departamento Nacional de Infraestrutura de Transportes (DNIT), especialmente por meio da interpretação dos fatores que mais influenciam as previsões do modelo. A análise dos valores SHAP permite identificar quais características da infraestrutura viária mais contribuem para o aumento das previsões de risco realizadas pelo modelo. Quando variáveis relacionadas à configuração da via, como pistas simples ou traçados em curva, apresentam contribuição significativa

para o aumento da gravidade em segmentos com elevado volume de tráfego, o modelo fornece evidências quantitativas que podem subsidiar a priorização de investimentos em obras de duplicação, correções geométricas, adequações de sinalização e demais intervenções voltadas à segurança viária.

Sob o ponto de vista metodológico, o trabalho também demonstra que a combinação entre modelos de aprendizado de máquina de alto desempenho e técnicas de Inteligência Artificial Explicável constitui uma abordagem promissora para apoiar processos decisórios em problemas complexos de segurança viária. Ao associar desempenho preditivo e interpretabilidade, a solução desenvolvida reduz a opacidade normalmente atribuída a modelos complexos e amplia a confiabilidade das informações disponibilizadas aos usuários, favorecendo sua utilização como instrumento de apoio à tomada de decisão.

Apesar dos resultados satisfatórios obtidos, o modelo apresenta limitações associadas à disponibilidade e à granularidade das bases de dados públicas utilizadas. A gravidade dos acidentes constitui um fenômeno multifatorial, influenciado por aspectos relacionados à infraestrutura, ao comportamento humano, às condições ambientais, às características dos veículos e ao contexto operacional das rodovias. Nesse sentido, o escopo da pesquisa não contempla integralmente todas as variáveis potencialmente associadas à gravidade das lesões, especialmente aquelas relacionadas ao estado de conservação da infraestrutura, ao comportamento dos condutores e à intensidade das ações de fiscalização.

Como perspectivas para trabalhos futuros, destaca-se inicialmente a incorporação de indicadores de conservação da infraestrutura rodoviária. Embora a modelagem atual considere características estruturais gerais, como tipo de pista e traçado da via, ela não avalia diretamente o nível de degradação dos trechos analisados. A inclusão de informações relacionadas à qualidade do pavimento, presença de defeitos, condições da sinalização horizontal e vertical e demais aspectos de manutenção permitiria ampliar a capacidade explicativa do modelo. Dados provenientes da Pesquisa CNT de Rodovias e de levantamentos realizados pelo próprio DNIT representam fontes promissoras para esse aprimoramento, possibilitando avaliar de forma mais precisa os impactos da conservação viária sobre a segurança do tráfego.

Outra possibilidade de evolução consiste na integração de informações geoespaciais referentes à localização de radares de velocidade e dispositivos de fiscalização eletrônica. A presença desses equipamentos pode ser interpretada como um indicativo de que determinado trecho já foi previamente identificado pelos órgãos competentes como uma área de risco elevado ou de recorrência de infrações relacionadas ao excesso de velocidade. A incorporação dessas informações permitiria investigar não apenas o efeito da fiscalização eletrônica na redução da probabilidade de acidentes, mas também possíveis fenômenos de deslocamento do risco para segmentos adjacentes não monitorados.

Em síntese, este trabalho demonstra que a utilização conjunta de técnicas de

aprendizado de máquina e Inteligência Artificial Explicável constitui uma alternativa viável para apoiar decisões relacionadas à segurança viária. Ao transformar grandes volumes de dados públicos em informações interpretáveis e disponibilizá-las por meio de uma aplicação web interativa, a solução proposta evidencia como a Ciência de Dados pode contribuir para o planejamento de ações preventivas, a priorização de investimentos em infraestrutura e o fortalecimento de políticas públicas baseadas em evidências. Embora existam oportunidades de aprimoramento mediante a incorporação de novas fontes de dados, os resultados obtidos demonstram que o sistema desenvolvido representa um avanço consistente na aplicação de técnicas analíticas para apoiar a gestão da segurança nas rodovias federais brasileiras.

# Referências

- AMORIM, B. d. S. P. et al. Uso de aprendizado de máquina para classificação de risco de acidentes em rodovias. Universidade Federal de Campina Grande, 2019.
- CARVALHO, C. H. R. d. *Custos dos acidentes de trânsito no Brasil: estimativa simplificada com base na atualização das pesquisas do Ipea sobre custos de acidentes nos aglomerados urbanos e rodovias*. Brasília, DF, 2020. Disponível em: <<https://repositorio.ipea.gov.br/items/98e21d94-320d-42bf-950b-7eb1a7f90911>>. Acesso em: 4 jun. 2026.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.
- CONFEDERAÇÃO NACIONAL DO TRANSPORTE. *Infraestrutura da maioria das rodovias brasileiras eleva risco de acidentes*. Confederação Nacional do Transporte, 2015. Publicado em 18 dez. 2015. Disponível em: <<https://cnt.org.br/agencia-cnt/infraestrutura-da-maioria-das-rodovias-brasileiras-eleva-risco-de-acidentes-cnt>>. Acesso em: 7 ago. 2026.
- CONFEDERAÇÃO NACIONAL DO TRANSPORTE. *Acidentes rodoviários e a infraestrutura*. Brasília, DF, 2018. Disponível em: <<https://cnt.org.br/acidentes-rodoviarios-os-infraestrutura>>. Acesso em: 6 ago. 2025.
- Confederação Nacional do Transporte. *Pesquisa CNT de Rodovias 2023 reforça a importância de maior investimento na malha rodoviária*. Confederação Nacional do Transporte, 2023. Disponível em: <<https://www.cnt.org.br/agencia-cnt/pesquisa-cnt-de-rodovias-2023-refora-a-importancia-de-maior-investimento-na-malha-rodoviaria>>.
- CORGOZINHO, M. M.; MONTAGNER, M. Â. Vulnerabilidade humana no contexto do trânsito motociclístico. *Saúde e Sociedade*, SciELO Public Health, v. 26, p. 545–555, 2017.
- FRANCESCHI, L. et al. Fatores relacionados à severidade de acidentes em rodovias no brasil através de um modelo de regressão logística multinomial. *Transportes*, v. 30, p. 2566, 2022.
- FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.
- JAMES, G. et al. *An Introduction to Statistical Learning with Applications in Python*. New York: Springer, 2023. Disponível em: <https://doi.org/10.1007/978-3-031-38747-0>. Disponível em: <<https://doi.org/10.1007/978-3-031-38747-0>>.
- KUMAR, S.; TOSHNIWAL, D. A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, Springer, v. 24, n. 1, p. 62–72, 2016. Disponível em: <https://doi.org/10.1007/978-3-031-38747-0>.
- LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, v. 30, 2017.

- MATHUR, P. *Machine Learning Applications Using Python: Case Studies from Healthcare, Retail, and Finance*. [S.l.]: Apress, 2019.
- MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.
- MOURA, B. d. F. *Taxa de mortes em acidentes de motocicletas cresce 12,5% no país*. Agência Brasil, 2025. Publicado em 12 maio 2025. Disponível em: <<https://agenciabrasil.ebc.com.br/direitos-humanos/noticia/2025-05/taxa-de-mortes-e-m-acidentes-de-motocicletas-cresce-125-no-pais>>.
- Organisation for Economic Co-operation and Development. *Safety of Vulnerable Road Users*. Paris: OECD Publishing, 1998.
- ORGANIZAÇÃO MUNDIAL DA SAÚDE. *Global status report on road safety 2018*. Geneva: World Health Organization, 2018.
- ORGANIZAÇÃO MUNDIAL DA SAÚDE. *Global status report on road safety 2023: country and territory profiles*. Genebra: World Health Organization, 2024. 179 p. ISBN 978-92-4-008771-2. Disponível em: <<https://www.who.int/publications/i/item/9789240087712>>. Acesso em: 6 ago. 2026.
- ORGANIZAÇÃO MUNDIAL DA SAÚDE. *Brazil: health data overview*. WHO Data, 2026. Disponível em: <<https://data.who.int/countries/076>>. Acesso em: 16 jun. 2026.
- Polícia Rodoviária Federal. *Anuário 2025*. Brasília, DF, 2026. Disponível em: <[https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/diest-arquivos/anuario-2025\\_final.html](https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/diest-arquivos/anuario-2025_final.html)>.
- REN, H. et al. A deep learning approach to the citywide traffic accident risk prediction. In: IEEE. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. [S.l.], 2018. p. 3346–3351.
- SAVOLAINEN, P.; MANNERING, F. Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes. *Accident Analysis & Prevention*, Elsevier, v. 39, n. 5, p. 955–963, 2007.
- TANG, J. et al. Traffic incident clearance time prediction and influencing factor analysis using extreme gradient boosting model. *Journal of Advanced Transportation*, Wiley Online Library, v. 2020, n. 1, p. 6401082, 2020.
- WISUTWATTANASAK, P. et al. Factors affecting single and multivehicle motorcycle crashes: insights from day and night analysis using xgboost-shap algorithm. *Big Data and Cognitive Computing*, MDPI, v. 8, n. 10, p. 128, 2024.
- WITTEN, I. H. et al. Practical machine learning tools and techniques. In: ELSEVIER PUBLISHERS AMSTERDAM, THE NETHERLANDS. *Data mining*. [S.l.], 2017. v. 2, n. 4, p. 403–413.
- WOLD, S.; SJOSTROM, M.; ERIKSSON, L. *PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58: 109–130*. 2001.

---

*Emitido em 11/08/2026*

**DOCUMENTO Nº 1/2026 - CCDN (11.00.52.09)**  
**(Nº do Documento: 1)**

**(Nº do Protocolo: NÃO PROTOCOLADO)**

*(Assinado digitalmente em 26/08/2026 10:19 )*  
ANDREA ALVES BORBA  
ASSISTENTE EM ADMINISTRACAO  
1517808

Para verificar a autenticidade deste documento entre em <https://sipac.ufpb.br/documentos/> informando seu número: **1**,  
ano: **2026**, documento (espécie): **DOCUMENTO**, data de emissão: **26/08/2026** e o código de verificação:  
**a736695cd1**