

Predição de acidentes de trânsito em Santa Catarina: avaliando o impacto do pré-processamento dos dados no desempenho dos modelos.

Traffic accident prediction in Santa Catarina: evaluating the impact of data preprocessing on model performance.

Predicción de accidentes de tránsito en Santa Catarina: evaluación del impacto del preprocesamiento de datos en el desempeño del modelo.

Gustavo Konescki Führ¹
Eduardo Camilo Inacio²

Resumo: Este estudo teve como objetivo investigar o impacto do pré-processamento de dados no desempenho de modelos preditivos de acidentes de trânsito. A questão central da pesquisa foi verificar a possibilidade de obter ganhos de desempenho apenas com o enriquecimento dos dados. Para isso, a pesquisa utilizou dados da PRF sobre acidentes entre os quilômetros 100 e 239 da rodovia BR-101 em SC, de 2017 a 2024. Primeiramente, constatou-se inconsistências nos dados, o que motivou sua correção e o enriquecimento com dados meteorológicos e de tráfego, gerando cinco conjuntos de dados para treinamento. Treinaram-se três modelos de aprendizado de máquina: RF, SVM e MLP, para cada conjunto de dados, totalizando quinze modelos. Os resultados mostraram que a correção de inconsistências nos dados aumentou o desempenho dos modelos, enquanto a adição de novos atributos produziu ganhos mais modestos, dependendo do modelo.

Palavras-chave: Aprendizado de Máquina. Predição de Acidentes de Trânsito. Integração de Dados.

Abstract: This study aimed to investigate the impact of data preprocessing on the performance of traffic accident predictive models. The central research question was whether it would be possible to achieve performance gains solely through data enrichment. To this end, the study used PRF data on accidents between kilometers 100 and 239 of highway BR-101 in Santa Catarina, from 2017 to 2024. First, inconsistencies were identified in the data, which led to their correction and enrichment with meteorological and traffic data, generating five training datasets. Three machine learning models were trained for each dataset: RF, SVM, and MLP, for a total of fifteen models. The results showed that correcting data inconsistencies increased model performance, while adding new attributes produced more modest gains, depending on the model.

Keywords: Machine Learning. Traffic Accident Prediction. Data Integration.

¹Graduando em Ciência da Computação. Universidade Federal de Santa Catarina. <https://orcid.org/0009-0005-4444-4193>. E-mail: gustavokf2003@gmail.com

²Doutor. Universidade Federal de Santa Catarina. <https://orcid.org/0000-0003-3424-8269>. E-mail: eduardo.camilo@ufsc.br.

Resumen: Este estudio tuvo como objetivo investigar el impacto del preprocesamiento de datos en el rendimiento de los modelos predictivos de accidentes de tránsito. La pregunta de investigación es: ¿sería posible lograr mejoras de rendimiento únicamente mediante el enriquecimiento de datos? Para ello, el estudio utilizó datos PRF sobre accidentes entre los kilómetros 100 y 239 de la carretera BR-101 en Santa Catarina, de 2017 a 2024. En primer lugar, se identificaron inconsistencias en los datos, lo que condujo a su corrección y enriquecimiento con datos meteorológicos y de tráfico, generando cinco conjuntos de datos de entrenamiento. Se entrenaron tres modelos de aprendizaje automático para cada conjunto de datos: RF, SVM y MLP, para un total de quince modelos. Los resultados mostraron que la corrección de las inconsistencias de los datos mejoró el rendimiento del modelo, mientras que la adición de nuevos atributos produjo mejoras más modestas, dependiendo del modelo.

Palabras clave: Aprendizaje automático. Predicción de accidentes de tráfico. Integración de datos.

Submetido 07/10/2025

Acceto 12/03/2026

Publicado 10/04/2026

Considerações Iniciais

Os acidentes de trânsito estão entre os maiores problemas globais, com impactos na saúde e na economia. A pesquisa da Organização Mundial da Saúde (WHO), divulgada em dezembro de 2023, mostra que os acidentes de trânsito causam aproximadamente 1,19 milhão de mortes por ano, sendo a principal causa de óbitos entre crianças e adultos de 5 a 29 anos (WHO, 2023). Em 2021, a WHO estabeleceu o objetivo de diminuir pela metade o número de mortes e lesões no trânsito até 2030 entre seus países integrantes (WHO, 2021). Entretanto, na realidade brasileira, o número de acidentes continua elevado. De acordo com os dados de 2021 da Polícia Rodoviária Federal (PRF), o Brasil registrou 71,9 mil feridos e 5,4 mil mortos. Já em 2024, houve um crescimento de 17,5% no total de feridos, alcançando 84,5 mil, e 9,2% no número de vítimas fatais, atingindo 6,1 mil (PRF, 2024).

Ademais, a situação dos acidentes rodoviários em Santa Catarina é preocupante. Segundo a PRF, o estado foi o segundo com o maior número de acidentes em rodovias federais no Brasil entre o início de 2017 e o fim de 2024, somando aproximadamente 47,4 mil registros, correspondendo a 11,94% dos acidentes no território nacional. Neste período, cerca de 130,1 mil veículos e 159,4 mil pessoas se envolveram em acidentes, resultando em 72,4 mil feridos e 3,0 mil mortes (PRF, 2024). No que tange aos custos, dados do Painel CNT de Consultas Dinâmicas sobre Acidentes Rodoviários apontam que o estado obteve um gasto estimado de 1,32 bilhão de reais causado por acidentes no ano de 2022 (CNT, 2022).

Priorizar a segurança pública envolve buscar ferramentas que viabilizem a redução de acidentes rodoviários. Integrado a um mapa de acidentes em Santa Catarina, o modelo de inteligência artificial poderia indicar regiões com maior risco de ocorrência, permitindo o direcionamento de ações preventivas. Além disso, o uso de modelos preditivos contribui para a otimização dos recursos públicos, já que possibilita concentrar esforços nas áreas mais vulneráveis e planejar estratégias de segurança viária de forma mais eficiente.

A predição de acidentes de trânsito, neste estudo, pode ser entendida como uma tarefa de classificação, em que os rótulos correspondem à ocorrência ou não de um acidente em um trecho de via de 100 metros e em um intervalo temporal de horas. Dado um vetor de atributos $x \in \mathbb{I}$ (características como tipo de pista, traçado, fluxo de veículos, condições climáticas, entre outros) e um conjunto de classes $C = \{\text{Acidente}, \text{Não acidente}\}$, este problema pode ser modelado como uma função: $F : x \in \mathbb{I} \rightarrow C$. Num exemplo, considerando um trecho com

pista simples, curva acentuada, chuva e tráfego intenso, o modelo pode classificá-lo como $F(x) = \text{Acidente}$. Enquanto num trecho de pista dupla, reta, tempo estável e baixo fluxo, o resultado esperado seria: $F(x) = \text{Nao acidente}$.

Este trabalho analisa o impacto do pré-processamento dos dados utilizados no treinamento de modelos preditivos de acidentes sobre o seu desempenho. Para a investigação, foram utilizados registros de acidentes da PRF. Por meio de pesquisa bibliográfica com foco em modelos de predição de acidentes, identificaram-se novas fontes de dados com domínios de atributos utilizados em outras pesquisas. Na análise dos dados, constatou-se que os dados da PRF possuem inconsistências em relação aos atributos *tipo de pista*, *traçado da via* e *uso do solo*. Ainda, foram identificadas possíveis tendências de arredondamento do atributo *km* para valores inteiros, devido a forma de registro de anos anteriores. Isso motivou a correção desses atributos com outras fontes de dados da Agência Nacional de Transportes Terrestres (ANTT) e do Departamento Nacional de Infraestrutura de Transportes (DNIT).

Após a correção desses dados, realizou-se o processo de aprimoramento dos dados por meio das fontes identificadas na pesquisa bibliográfica. Incorporaram-se características da via da ANTT, atributos meteorológicos da API OPEN-METEO e dados de tráfego do DNIT. O enriquecimento foi sequencial, ou seja, o conjunto de dados de uma etapa possui todos os atributos das etapas anteriores juntamente com os atributos de um novo conjunto de dados. Dessa forma, foram gerados cinco conjuntos de dados distintos para o treinamento dos modelos, iniciando apenas com os dados da PRF sem correção e progredindo até uma versão totalmente enriquecida. Esses conjuntos foram treinados com diferentes métodos de aprendizado de máquina, visando avaliar como cada etapa de enriquecimento influencia no desempenho dos modelos.

Metodologia

Trata-se de pesquisa quantitativa, de natureza aplicada e com objetivos explicativos, voltada à análise do impacto do enriquecimento de dados no desempenho de modelos de predição de acidentes. Quanto aos procedimentos, caracteriza-se como experimental, com avaliação comparativa de modelos em diferentes cenários (Gil, 2008). As etapas de pesquisa envolveram: pesquisa bibliográfica por trabalhos relacionados, identificação de base de dados

relevantes, análise e pré-processamento dos dados, treinamento de algoritmos de aprendizado de máquina e avaliação dos modelos por métricas de desempenho.

Para identificar os atributos mais relacionados a acidentes de trânsito, realizou-se uma pesquisa bibliográfica em artigos publicados entre 2019 e 2024. A pesquisa foi conduzida entre os meses de maio e junho de 2025, utilizando os repositórios acadêmicos: ScienceDirect, IEEE Xplore, Scopus e Google Scholar. As principais palavras-chave utilizadas nas buscas foram: "traffic accident prediction", "road crash prediction models" e "traffic safety machine learning". Foram aplicados os seguintes critérios de inclusão:

- Estudos que aplicam aprendizado de máquina para predição de acidentes.
- Documentação suficiente sobre os atributos utilizados.

Estudos focados na predição de severidade do acidente foram excluídos da pesquisa.

A partir desta pesquisa, foram selecionados dez artigos. Além disso, foram adicionados dois artigos em português por utilizarem dados brasileiros: Kraut e Sapia (2022) e Ferreira *et al.* (2024). Nesses estudos, observou-se quais domínios de atributos são mais frequentemente utilizados no treinamento de modelos. A Tabela 1 apresenta os principais domínios e os respectivos artigos que os utilizaram. Observa-se que a maior parte dos estudos empregou dados de tráfego. Os outros domínios ocorrem com menor frequência, porém de maneira distribuída entre os estudos, com uma única diferença no domínio de pontos de interesse, a qual aparece em apenas um estudo.

Tabela 1 - Domínios de atributos mais utilizados em trabalhos relacionados.

Autor	Tráfego	Tempo físico	Local físico	Condições meteorológicas	Características da via	Pontos de interesse
Cai <i>et al.</i> (2020)	✓					
Huang <i>et al.</i> (2020)	✓	✓	✓			
Peng <i>et al.</i> (2020)	✓					



Elassad <i>et al.</i> (2020)				✓	✓	
Yu <i>et al.</i> (2021)	✓	✓	✓	✓	✓	✓
Islam <i>et al.</i> (2021)	✓					
Zhao <i>et al.</i> (2022)	✓					
Kraut e Sapia (2022)		✓	✓		✓	
Tran <i>et al.</i> (2023)	✓	✓	✓			
Zhao <i>et al.</i> (2023)	✓					
Ferreira <i>et al.</i> (2024).	✓	✓	✓		✓	
Mo <i>et al.</i> (2024)	✓			✓	✓	

Fonte: Autores (2025)

O conjunto da PRF contém registros de acidentes de trânsito ocorridos nas rodovias federais brasileiras. As informações incluem domínio de tempo (data, hora), domínio de local (unidade de federação, município, rodovia, km), domínio de características da via (tipo de pista, tipo de traçado, uso do solo), além de atributos como tipo de acidente, número de veículos envolvidos, vítimas, óbitos, entre outros (PRF, 2024).

O conjunto de dados de acidentes PRF já possui atributos que pertencem aos domínios de tempo físico e local físico, além de conter atributos limitados sobre condições meteorológicas e características da via. Desta forma, realizou-se uma busca por fontes de dados relacionadas a características da via, condições meteorológicas e tráfego que abrangem o espaço-tempo do estudo.

Uma das principais fontes de tráfego do Brasil é o Plano Nacional de Controle de Tráfego (PNCT) fornecido pelo DNIT. O PNCT apresenta o volume de veículos por categoria em 320 pontos espalhados pelas rodovias federais, registrando a contagem horária de cada categoria de veículo (DNIT, 2025b). No entanto, existem poucos pontos de coleta que estão localizados na região de interesse da rodovia. Diante da limitada disponibilidade de informações que coincidam com o intervalo espacial e temporal da análise, decidiu-se por não utilizar este conjunto de dados, visto que sua integração exigiria a aplicação de técnicas sofisticadas de interpolação para suprir as lacunas existentes.

Outra fonte relevante de informações é o Volume Médio Diário Anual (VMDA) de veículos que trafegam em um determinado trecho de rodovia, disponibilizado pelo DNIT (DNIT, 2025b). Embora o VMDA apresente uma estimativa diária média, diferente da granularidade horária adotada neste estudo e não possua variações semanais ou mensais de tráfego por se tratar de uma média anual, optou-se por utilizá-lo, por ser o conjunto de dados mais adequado disponível e possibilitar análise do fluxo em diferentes trechos da rodovia.

Na maioria dos estudos relacionados, foram observadas correlações entre a variação do volume de tráfego em dois postos de contagem próximos dentro de intervalos de 5 a 15 minutos anteriores à ocorrência de acidentes de trânsito. Contudo, não foi possível identificar nenhuma fonte de dados disponível que permitisse realizar esse tipo de análise.

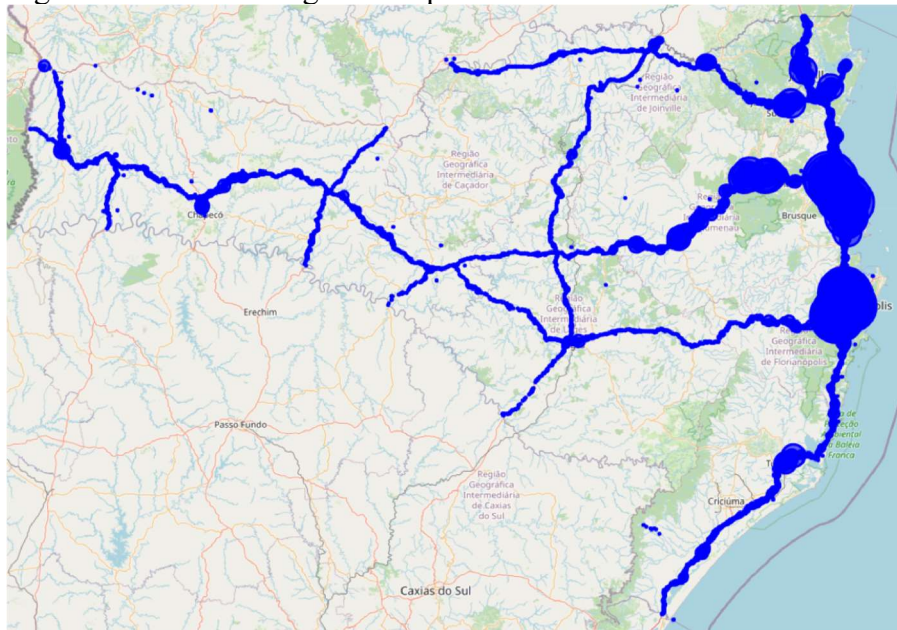
No caso das informações sobre as condições meteorológicas, identificou-se a API Open-Meteo como fonte adequada. Esta API de código aberto disponibiliza dados meteorológicos para qualquer região do planeta, fazendo uso do conjunto de dados de reanálise IFS do Centro Europeu de Previsões Meteorológicas a Médio Prazo (ECMWF), oferecendo acesso a dados históricos sobre o clima a partir do ano de 2017 até o presente. A API fornece 30 atributos meteorológicos, com uma resolução espacial de 9 quilômetros e uma resolução temporal em horas (Zippenfenig, 2023).

Para o domínio de características da via, utilizaram-se conjuntos de dados da ANTT e do DNIT. A ANTT dispõe de quarenta coleções de dados sobre as rodovias brasileiras. Dentre elas, há conjuntos de dados contendo informações para cada rodovia federal sobre o número de faixas, o perfil topográfico, a velocidade regulamentada para veículos leves, entre outros (ANTT, 2025). Já o DNIT disponibiliza uma plataforma web de dados geográficos chamada VGeo, que contém informações sobre rodovias federais (DNIT, 2025a).

Após a seleção dos conjuntos de dados, serão descritos em detalhe os atributos escolhidos e como foram empregados na construção dos conjuntos, o que viabiliza uma melhor compreensão sobre o papel de cada fonte na modelagem.

Foram utilizados dados sobre acidentes em Santa Catarina, registrados pela PRF entre 2017 e 2024. Foi estabelecido o trecho da BR-101 entre os quilômetros 100 (Itajaí) e 239 (Palhoça) para os experimentos de predição, devido à alta densidade de ocorrências registradas nessa extensão. A Figura 1 ilustra acidentes reportados pela PRF no período. Os raios dos círculos em azul em torno de locais onde aconteceram acidentes representam as quantidades de acidentes na região.

Figura 1 - Acidentes registrados pela PRF em rodovias federais de SC.



Fonte: Autores (2025)

Foram usados arquivos dataatran do site da PRF, que foram concatenados e submetidos a uma filtragem, considerando apenas os registros localizados na área de interesse. Obteve-se uma amostra final composta por 20.656 registros de acidentes. Os atributos selecionados para o treinamento dos modelos estão descritos na Tabela 2.

Tabela 2 - Descrição dos atributos utilizados dos dados da PRF.

Atributo	Descrição
Km	Identificação do quilômetro onde ocorreu o acidente, com precisão de 0,1 km.
Município	Nome do município onde ocorreu o acidente.
Sentido da via	Considerando o ponto de colisão. Ex: Crescente, Decrescente.
Tipo de pista	Quantidade de faixas da via principal. Ex: Simples, Dupla, Múltipla.
Traçado da via	Característica do tipo de traçado. Ex: Reta, Curva, Aclive, etc.
Uso do solo	Tipo de ocupação. Ex: Sim (urbano), Não (rural).
Data inversa	Formato dd/mm/aa.
Horário	Formato hh:mm:ss.

Fonte: Adaptado do dicionário de dados da PRF (2025)

Na análise dos dados, constatou-se que 74,15% das observações de acidentes registradas para o mesmo trecho de 100 metros apresentam pelo menos um valor divergente nos atributos *sentido da via*, *tipo de pista* e *uso do solo*. Estas divergências indicam inconsistência nos dados. A Tabela 3 ilustra tais divergências, considerando registros de acidentes no quilômetro 233,0. Nota-se que, para o mesmo trecho, foram reportados diferentes valores para os atributos *tipo de pista*, *traçado da via* e *uso do solo*, ou seja, o mesmo local foi considerado em registros diferentes como pista dupla ou múltipla; com traçado sendo curva, reta, curva com declive ou reta com declive; e indicado como uma região urbana ou rural. O treinamento de modelos de aprendizado de máquina com essas inconsistências pode levar a preditores com fraco desempenho, o que motivou a correção das informações com outras fontes de dados.

Tabela 3 - Características da via em registros de acidentes no km 233.0 da BR-101 em SC.

Km	Município	Tipo de pista	Traçado da via	Uso do solo
233.0	PALHOCA	Dupla	Curva; Declive	Não
233.0	PALHOCA	Múltipla	Reta	Não
233.0	PALHOCA	Dupla	Reta	Sim
233.0	PALHOCA	Dupla	Reta; Declive	Não
233.0	PALHOCA	Dupla	Curva; Declive	Não
233.0	PALHOCA	Dupla	Curva	Não

Fonte: Autores (2025)

Dentre os conjuntos disponibilizados pela ANTT, foram selecionados valores de quilometragem, *município*, *número de faixas*, *traçado da via* e *uso do solo* para a correção das inconsistências nos dados da PRF. Adicionalmente, foram escolhidos novos atributos de características da via para avaliar se sua inclusão melhora o desempenho dos modelos preditivos, tais como *tipo de pavimento*, *tipo de perfil do terreno*, *velocidade regulamentada para veículos leves*, *velocidade regulamentada para veículos pesados*, *presença de pista marginal* e a existência de *iluminação*. A Tabela 4 descreve os atributos escolhidos.

Tabela 4 - Descrição dos atributos utilizados dos dados da ANTT.

Atributo	Descrição
Km	Identificação do quilômetro da rodovia em que o registro está localizado.
Município	Nome do município correspondente ao trecho da rodovia
Número de faixas	Quantidade de faixas disponíveis na via
Traçado da via	Característica do tipo de traçado. Ex: Tangente, Curva.
Tipo de uso do solo	Tipo de ocupação. Ex: Urbano, Rural.
Tipo do pavimento	Material da superfície da via. Ex: Rígido, Flexível.
Tipo de perfil do terreno	Relevo onde a rodovia está inserida. Ex: Montanhoso, Plano,

	Ondulado
Velocidade regulamentada veículos leves	Limite de velocidade para automóveis. Ex: 40 km/h
Velocidade regulamentada veículos pesados	Limite de velocidade para caminhões e ônibus. Ex: 40 km/h
Pista marginal	Indicação da presença de pista marginal.
Iluminação	Indicação da existência de iluminação

Fonte: Adaptado dos dicionários de dados da ANTT (2025)

A geometria da via foi extraída a partir da base do DNIT. Esses dados estão no formato MultiLineString, que representa múltiplas linhas geográficas com latitude e longitude da via. Tais informações foram fundamentais para integrar os atributos da ANTT em um único conjunto. Contudo, notou-se que representam apenas a geometria do sentido crescente da rodovia. Portanto, foi necessária a criação da representação correspondente ao sentido decrescente. Dos atributos disponíveis na API OPEN-METEO, foram selecionados nove. A Tabela 5 apresenta esses atributos. A extração foi realizada utilizando as coordenadas geográficas correspondentes aos dados do DNIT, com espaçamento de cinco quilômetros entre os pontos. A escolha por intervalos maiores se deve à resolução espacial que a API possui.

Tabela 5 - Descrição dos atributos utilizados dos dados da API OPEN-METEO.

Atributo	Descrição
Temperatura	Temperatura do ar 2 metros acima do solo.
Temperatura aparente	Temperatura percebida que combina a sensação térmica, umidade relativa e radiação solar.
Ponto de orvalho	Temperatura do ponto de orvalho a 2 metros acima do solo.
Chuva	Precipitação líquida da hora anterior, incluindo chuviscos locais e chuva de sistemas de grande escala.
Umidade relativa do ar	Umidade relativa do ar a 2 metros acima do solo.
Cobertura das nuvens	Cobertura total de nuvens como fração de área.

Velocidade do vento	Velocidade do vento a 10 metros acima do solo.
Velocidade de rajada	Velocidade de rajada a 10 metros acima do solo.
Código climático	Condição meteorológica como um código numérico.

Fonte: Adaptado do dicionário de dados da OPEN-METEO (2025)

Os dados de tráfego do VMDA foram obtidos por meio da plataforma oficial do DNIT, na qual cada ano de registro encontra-se em arquivo separado. Para o ano de 2024, como os dados oficiais ainda não estavam disponíveis, utilizou-se o último ano de registro disponível (2023) como estimativa. A Tabela 6 descreve os atributos utilizados no treinamento dos modelos.

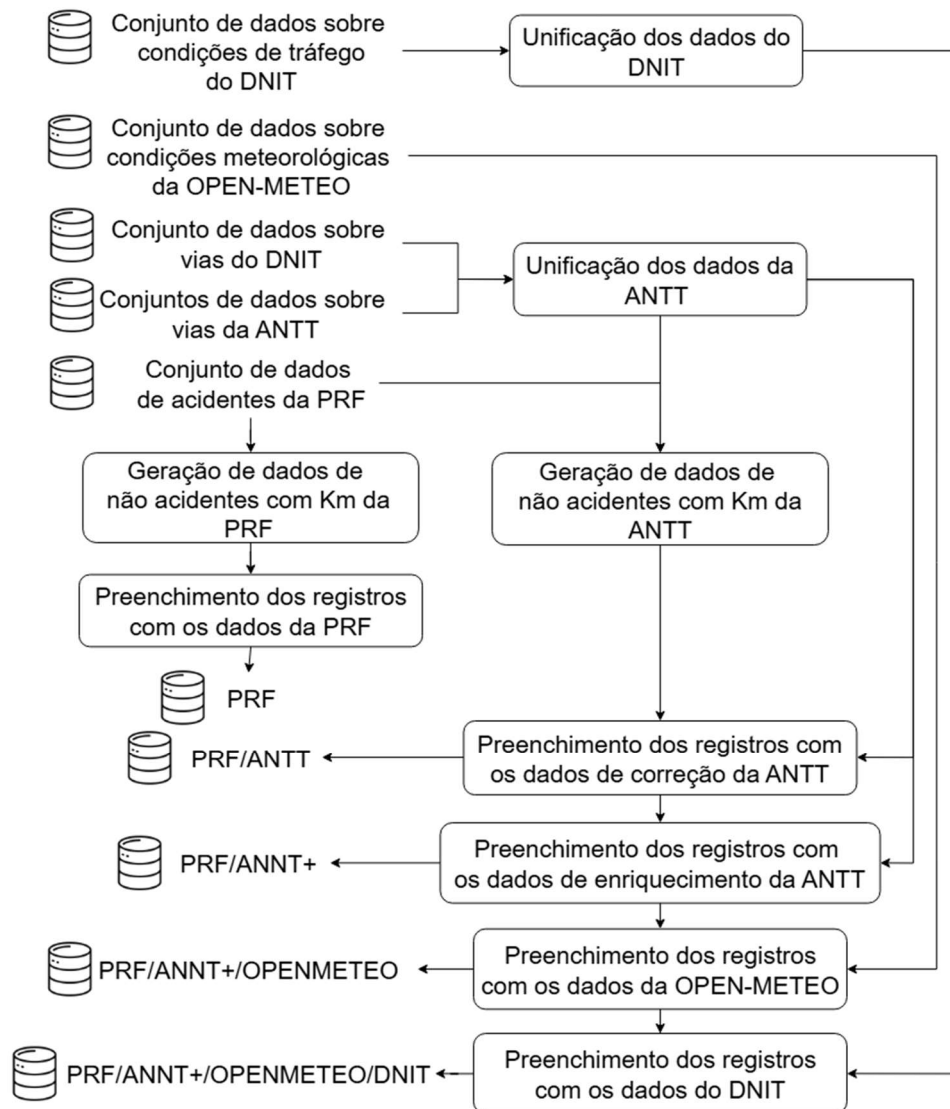
Tabela 6 - Descrição dos atributos utilizados dos dados do DNIT.

Atributo	Descrição
A	VMDA de ônibus ou caminhões de 2 eixos
B	VMDA de ônibus ou caminhões de 3 eixos
C	VMDA de ônibus ou caminhões de 4 eixos
D	VMDA de ônibus ou caminhões de 5 eixos
E	VMDA de ônibus ou caminhões de 6 eixos
F	VMDA de ônibus ou caminhões de 7 eixos
G	VMDA de ônibus ou caminhões de 8 eixos
H	VMDA de ônibus ou caminhões de 9 eixos
I	VMDA de carros, vans, etc
J	VMDA de motocicletas

Fonte: Adaptado do dicionário de dados do DNIT (2025)

A Figura 2 mostra o fluxo geral dos dados. Os cinco conjuntos de dados mencionados anteriormente são utilizados em atividades, representadas por blocos, para a construção dos cinco conjuntos usados no treinamento e teste dos modelos. Nota-se que o conjunto PRF utiliza apenas dados do conjunto de acidentes da PRF, enquanto os demais conjuntos usam dados de enriquecimento de maneira sequencial, ou seja, cada novo conjunto possui dados do conjunto anterior, agregando incrementalmente novas informações.

Figura 2 - Fluxo geral dos dados.

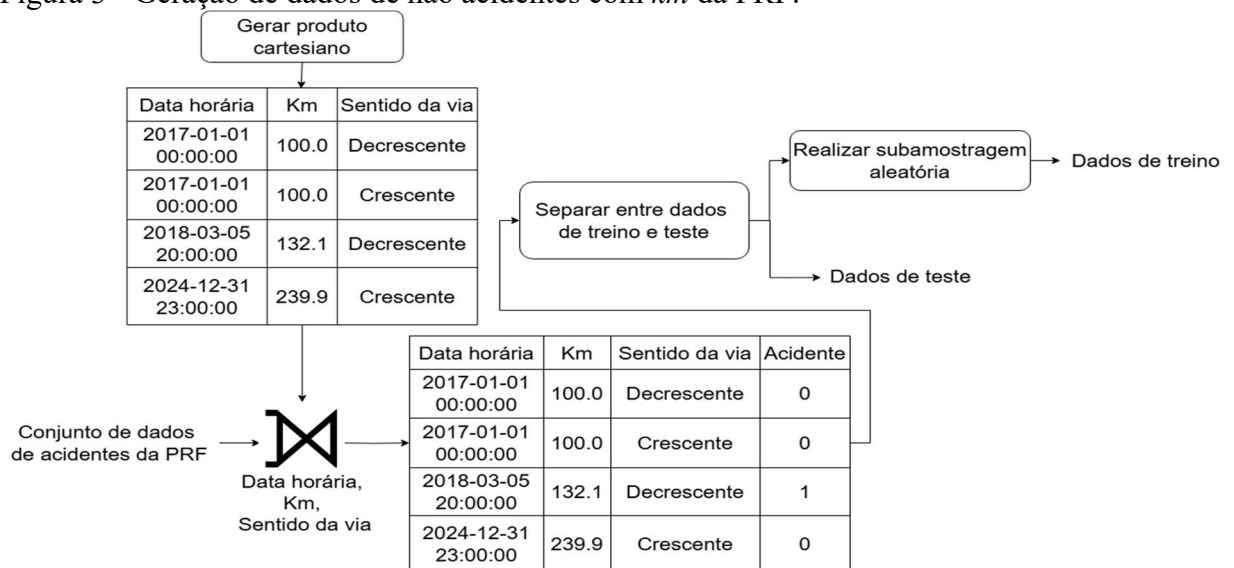


Fonte: Autores (2025)

A Figura 3 mostra o processo de geração de dados de não-acidentes com o atributo *km* da PRF. Foi necessário gerar amostras de não-acidentes, uma vez que os dados disponibilizados pela PRF contêm apenas registros de acidentes. Foi construído um produto cartesiano entre três dimensões: as horas no período de 2017 a 2024, cada 100 metros entre os quilômetros 100 a 239 da BR-101 e os sentidos da via (crescente ou decrescente), com intuito de obter todas as combinações espaço-temporais possíveis ao longo do trecho analisado. Realizou-se um left-join com os dados da PRF, utilizando como chaves os campos de *data horária* (gerados a partir da junção dos campos *data_inversa* e *hora*), *km* e *sentido da via*. Foi possível identificar quais combinações registraram ocorrências de acidentes e, por exclusão, obter exemplos de não-acidentes a serem utilizados.

Obteve-se 194.955.840 registros, onde apenas 0,01% correspondiam a acidentes, enquanto 99,99% representavam sem acidentes, revelando um forte desbalanceamento. Os dados foram divididos em conjuntos de treinamento e teste. No conjunto de treinamento, aplicou-se o método de subamostragem aleatória das amostras de não acidentes para equilibrar a proporção de acidentes e não acidentes em 50%. No conjunto de teste manteve-se a proporção original, a fim de representar a distribuição real de acidentes. No final, foram obtidos 40.622 registros de treinamento, sendo 20.311 acidentes e 20.311 não acidentes, e 1.949.559 registros de teste com 205 acidentes e 1.949.354 não acidentes.

Figura 3 - Geração de dados de não acidentes com *km* da PRF.



Fonte: Autores (2025)

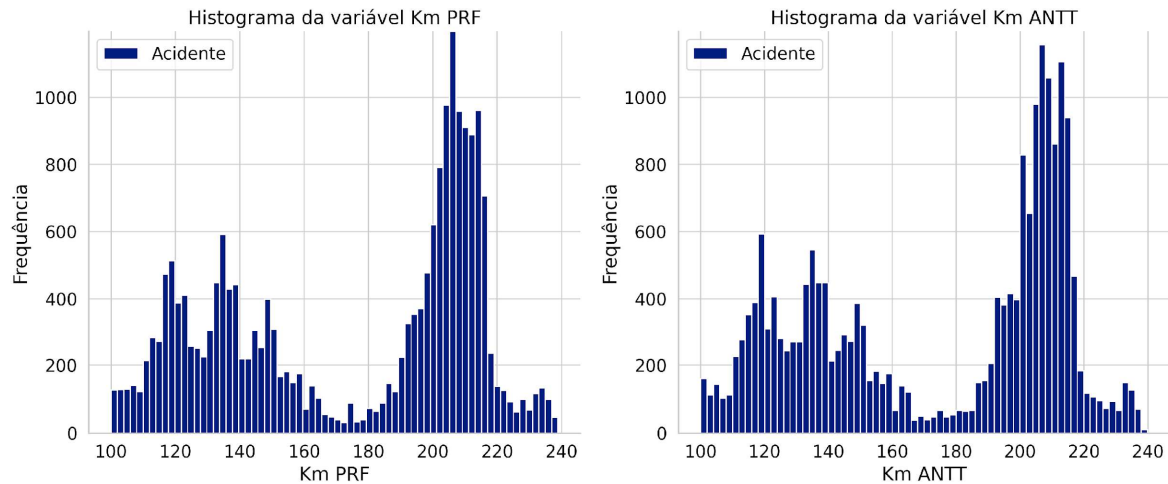
No preenchimento dos registros com os dados da PRF, para completar o produto cartesiano com os atributos faltantes de *traçado da via*, *tipo de pista*, *uso do solo* e *município*, foi realizado um left-join para dados de acidentes com os atributos registrados no momento do próprio acidente, através da *data horária*, *km* e *sentido da via*. Para dados de não-acidentes, foi realizada uma interpolação, extraíndo os atributos da observação com a data mais próxima disponível para o respectivo trecho mais próximo, considerando uma tolerância de 200 metros. Por fim, foi realizada a criação dos atributos que representam *hora*, *dia*, *dia da semana*, *mês*, *ano* e *feriado* a partir do atributo da *data horária*. Assim, foi criado o primeiro conjunto de dados para treinamento dos modelos, denominado PRF.

Na etapa de unificação dos dados da ANTT, a partir dos dados do DNIT, foram extraídas as coordenadas geográficas referentes a cada segmento de 100 metros da rodovia. Como os dados originais possuem apenas o segmento do sentido crescente da via, foi necessário duplicar os pontos e ajustar suas posições geográficas para criar a representação do sentido decrescente. A integração entre os dados de vias da ANTT e o traçado da BR-101 foi realizada por meio de operações de proximidade baseadas na distância euclidiana, a fim de identificar a característica geográfica mais próxima de cada ponto da rodovia. Este conjunto de dados unificados da ANTT foi denominado “Características da via da ANTT”.

Até o ano de 2016, o atributo *km* nos registros de acidentes da PRF era anotado com granularidade em quilômetros inteiros. A partir de 2017, a PRF passou a adotar uma precisão mínima de 0,1 quilômetros, além do uso de coordenadas geográficas para identificar o local exato do acidente. Esta alteração motivou uma investigação mais detalhada sobre o atributo, com o objetivo de verificar se, após a atualização no padrão de coleta, o registro truncado continuava presente de forma implícita. Foram efetuados métodos de proximidade com base na distância euclidiana, relacionando coordenadas dos registros de acidentes da PRF com coordenadas do conjunto de dados da ANTT, a fim de identificar o valor correspondente do atributo *km* na base da ANTT para cada acidente registrado.

Dos 20.656 registros de acidentes da PRF, 16.027 (77,59%) apresentaram valores diferentes de *Km* entre os dados da PRF e da ANTT. Entretanto, em 14.071 (87,73%) destes casos, a diferença observada situou-se dentro da margem de um quilômetro. O Gráfico 1 ilustra que a distribuição geral do atributo nos dados da PRF e da ANTT apresenta grande similaridade, embora a maioria dos valores seja discordante.

Gráfico 1 - Histograma de acidentes por *km* da PRF e ANTT.



Fonte: Autores (2025)

Com base nestes resultados, optou-se por criar uma nova versão do processo de geração de dados de não-acidentes, utilizando o atributo *km* proveniente da ANTT, em substituição ao atributo original da PRF. Esta nova tarefa apresenta todas as etapas que a geração de dados de não-acidentes com *km* da PRF possui, porém, a operação de left-join é realizada com campos *data-horária* (uma combinação do campo *data* e *hora* da PRF), *km* (da ANTT) e *sentido da via* (da PRF).

No preenchimento dos registros com dados de correção da ANTT, foi realizado um left-join entre os registros gerados na etapa anterior e o conjunto “Características da via da ANTT”, utilizando os atributos *km* e *sentido da via* como chaves de junção para inclusão dos atributos *uso do solo*, *tipo de traçado*, *número de faixas* e *município* em cada registro. O atributo *número de faixas* da ANTT foi utilizado para substituir o atributo *tipo de pista* da PRF, enquanto os demais atributos possuem o mesmo nome em ambos os conjuntos e foram substituídos um pelo outro. Foram gerados atributos temporais a partir do campo da *data horária*, como *hora*, *dia*, *dia da semana*, *mês*, *ano* e *feriado*, criando o conjunto de dados chamado PRF/ANTT.

A atividade de preencher os registros com os dados de enriquecimento da ANTT recebe como entrada o conjunto PRF/ANTT e adiciona o restante dos atributos do conjunto de “Características da via da ANTT” por meio de um left-join com os atributos *km* e *sentido da via*. Assim, produziu-se o terceiro conjunto de treinamento, apelidado de PRF/ANTT+.

No preenchimento dos registros com dados da OPEN-METEO, a junção foi realizada com base no atributo *data horária*, associando-a ao ponto de coleta meteorológica mais próximo, com tolerância de 5 quilômetros. Como resultado, foi criado o conjunto de dados denominado PRF/ANTT+/OPENMETEO.

A unificação dos dados do VMDA iniciou com a padronização dos nomes das colunas em todos os conjuntos. Os conjuntos foram concatenados em um único conjunto e foi realizada a extração para obter a região de interesse. Realizou-se uma expansão dos registros, obtendo todos os segmentos de 100 metros e todos os anos, com o respectivo volume de tráfego para cada categoria de veículo. Por fim, no preenchimento dos registros com dados do DNIT foi realizado um left-join com os atributos *km* e *ano*. Assim, foi gerado o último conjunto chamado PRF/ANTT+/OPENMETEO/DNIT, o qual obtém a coleção de atributos de todos os conjuntos de dados adquiridos.

Para a modelagem, os atributos presentes nos conjuntos de dados da PRF, da ANTT e do *código climático* da OPEN-METEO foram convertidos em categorias utilizando a codificação one-hot, técnica que converte variáveis categóricas, como textos, em um formato numérico compreensível para os modelos. Já os demais atributos da OPEN-METEO, atributos de tráfego do DNIT e o atributo *ano* foram submetidos à padronização, processo que modifica os dados numéricos para possuírem média zero e desvio padrão um, garantindo o equilíbrio das escalas entre atributos distintos.

Em seguida, foram treinados três algoritmos distintos de aprendizado de máquina: *Random Forest* (RF), *Support Vector Machines* (SVM) e *Multilayer Perceptron* (MLP). A RF é um método de *ensemble* baseado em múltiplas árvores de decisão com amostragem por *bagging* e seleção aleatória de atributos, promovendo diversidade e robustez no modelo (Tan; Steinbach; Kumar, 2009, p. 289). A SVM identifica o hiperplano ótimo de separação entre classes, visando maior capacidade de generalização (Tan; Steinbach; Kumar, 2009, p. 255). Já a MLP é uma arquitetura de rede neural com múltiplas camadas densamente conectadas que utiliza retropropagação do erro para ajuste dos pesos (Haykin, 2001, p. 183). Estes modelos foram escolhidos por sua recorrência em estudos relacionados e devido ao bom desempenho para dados tabulares. Cada modelo foi treinado e avaliado com os cinco conjuntos de dados gerados: PRF, PRF/ANTT, PRF/ANTT+, PRF/ANTT+/OPENMETEO, PRF/ANTT+/OPENMETEO/DNIT, resultando em um total de quinze modelos. Para garantir

uma seleção adequada dos hiperparâmetros, foi utilizada validação cruzada (*Cross-Validation*), que permite avaliar o desempenho dos modelos através de um processo iterativo, obtendo uma estimativa mais confiável dos resultados. A Tabela 7 apresenta todos os hiperparâmetros e valores testados.

Tabela 7 - Hiperparâmetros testados para os modelos RF, SVM e MLP.

Modelo	Hiperparâmetro	Valores avaliados
RF	critério de divisão	[Gini, Entropia]
	profundidade máxima	[32, 64, 128]
	bootstrap	[Sim, Não]
SVM	C	[1, 2, 5]
	kernel	[Linear, Polinomial, RBF, Sigmoide]
	gamma	[Scale, Auto]
MLP	tamanho de camadas ocultas	[(512, 256, 128, 64), (256, 128, 64), (128, 64)]
	função de ativação	[ReLU, Tanh]
	otimizador	[Adam, SGD]
	taxa de dropout	[0, 0.1]
	regularização L2	[0, 0.0001]

Fonte: Autores (2025)

Para avaliar o desempenho dos modelos sobre os dados, foram calculadas as métricas de especificidade, sensibilidade e a área sob a curva ROC (AUC). A sensibilidade (taxa de verdadeiros positivos) mede a proporção dos casos corretamente preditos como positivos dentre todos os casos positivos. A especificidade (taxa de verdadeiros negativos) calcula a proporção dos casos corretamente preditos como negativos dentre todos os casos negativos. A AUC relaciona a taxa de verdadeiros positivos com a taxa de falsos positivos (1 - especificidade). Após a etapa de validação cruzada, foram selecionados os melhores valores para cada hiperparâmetro com base no desempenho dos modelos em cada conjunto de dados. A Tabela 8 mostra os melhores valores ajustados para cada hiperparâmetro de cada modelo.

Tabela 8 - Melhores hiperparâmetros para cada hiperparâmetro e conjunto de dados.

Hiperparâmetro	PRF	PRF/ANTT	PRF/ANTT+	PRF/ANTT+/OPENMETEO	PRF/ANTT+/OPENMETEO/DNIT
critério de divisão	Gini	Entropia	Entropia	Entropia	Entropia
profundidade máxima	128	128	128	128	128
bootstrap	Não	Não	Não	Não	Não
C	2	2	1	2	5
kernel	Linear	Linear	Linear	Linear	Linear
gamma	-	-	-	-	-
tamanho de camadas ocultas	(256, 128, 64)	(512, 256, 128, 64)	(512, 256, 128, 64)	(256, 128, 64)	(256, 128, 64)
função de ativação	ReLU	ReLU	ReLU	ReLU	ReLU
otimizador	Adam	Adam	Adam	Adam	Adam
taxa de dropout	0	0.1	0.1	0.1	0.1
regularização L2	0.0001	0.0001	0.0001	0	0

Fonte: Autores (2025)

De maneira geral, os hiperparâmetros se mantiveram constantes durante o enriquecimento dos dados. No próximo tópico, serão apresentados os resultados obtidos nos experimentos realizados com os diferentes conjuntos de dados e modelos, avaliando o impacto do enriquecimento sequencial dos dados no desempenho dos modelos.

Análise dos dados e Resultados

A Tabela 9 exibe a média e o desvio padrão das métricas de cada modelo nos dados de validação para cada conjunto de dados durante o processo de validação cruzada. A correção dos dados da PRF foi a que mais impactou positivamente nos resultados. Entre os conjuntos PRF e PRF/ANTT, todos os modelos apresentaram ganhos em todas as métricas, o que mostra a importância de uma base de dados consistente. Por outro lado, a inclusão de atributos adicionais de características de vias da ANTT, de condições meteorológicas da OPEN-METEO e de tráfego do DNIT resultou em apenas pequenos ganhos ou até mesmo pequenas quedas de desempenho em alguns modelos. No caso do MLP, observa-se que a cada conjunto adicional a sensibilidade aumentou progressivamente, enquanto a especificidade apresentou pequenas reduções, indicando que o modelo se beneficiou na detecção de acidentes. O desempenho do SVM se manteve constante em todas as métricas, em contrapartida, o RF sofreu reduções com a adição dos atributos.

Uma possível hipótese para esse comportamento é a forte concentração espacial dos acidentes em determinados trechos da rodovia, como entre os quilômetros 190 e 220 e entre os quilômetros 120 e 150, como mostrado no Gráfico 1. Este padrão pode induzir o modelo a focar excessivamente nessas regiões específicas, aumentando a importância do atributo *km* e reduzindo a influência de outros atributos. Ademais, a ausência de dados de tráfego em um formato semelhante ao utilizado em trabalhos relacionados pode ter limitado a contribuição desta categoria.

Tabela 9 - Desempenho dos modelos para os dados de validação.

Medida	Conjunto	RF	SVM	MLP
Sensibilidade	PRF	0.7167 ± 0.0051	0.7576 ± 0.0067	0.7585 ± 0.0082
	PRF/ANTT	0.8053 ± 0.0077	0.8340 ± 0.0057	0.8483 ± 0.0028
	PRF/ANTT+	0.7921 ± 0.0073	0.8322 ± 0.0051	0.8490 ± 0.0093
	PRF/ANTT+/ OPENMETEO	0.7867 ± 0.0073	0.8325 ± 0.0040	0.8560 ± 0.0033



	PRF/ANNT+/OPENMETEO/DNIT	0.7974 ± 0.0041	0.8373 ± 0.0043	0.8613 ± 0.0042
Especificidade	PRF	0.7827 ± 0.0029	0.7806 ± 0.0054	0.7798 ± 0.0048
	PRF/ANNT	0.8336 ± 0.0043	0.8233 ± 0.0062	0.8369 ± 0.0067
	PRF/ANNT+	0.8225 ± 0.0033	0.8250 ± 0.0055	0.8350 ± 0.0115
	PRF/ANNT+/OPENMETEO	0.8163 ± 0.0034	0.8242 ± 0.0054	0.8194 ± 0.0077
	PRF/ANNT+/OPENMETEO/DNIT	0.8000 ± 0.0041	0.8257 ± 0.0057	0.8172 ± 0.0070
AUC	PRF	0.7497 ± 0.0034	0.7691 ± 0.0053	0.7692 ± 0.0041
	PRF/ANNT	0.8195 ± 0.0047	0.8287 ± 0.0030	0.8426 ± 0.0030
	PRF/ANNT+	0.8073 ± 0.0036	0.8286 ± 0.0026	0.8420 ± 0.0039
	PRF/ANNT+/OPENMETEO	0.8015 ± 0.0043	0.8284 ± 0.0023	0.8377 ± 0.0032
	PRF/ANNT+/OPENMETEO/DNIT	0.7987 ± 0.0029	0.8315 ± 0.0020	0.8392 ± 0.0025

Fonte: Autores (2025)

Para verificar a robustez entre as métricas, foi aplicado o teste t de Student entre os conjuntos de dados, sendo que um p-valor inferior a 0,05 indica significância na diferença entre os grupos. A Tabela 10 apresenta os valores de p-valor para a comparação entre diferentes versões dos dados dentro de cada modelo para as métricas. Os resultados mostram que entre PRF e PRF/ANNT houve um ganho estatisticamente significativo em todos os modelos e métricas. Nas etapas seguintes, não foram observados ganhos significativos entre conjuntos adjacentes para os modelos SVM e MLP. No caso do MLP, nota-se uma tendência de aumento

em relação ao primeiro conjunto ao último conjunto, embora essa diferença não tenha sido testada quanto à significância estatística. O modelo RF, apresentou pequenas quedas com a adição dos atributos.

Tabela 10 - Teste de significância estatística (p-valor)

Medida	Conjuntos	RF	SVM	MLP
Sensibilidade	PRF x PRF/ANTT	5.85e-8	1.21e-7	2.99e-8
	PRF/ANTT x PRF/ANTT+	0.0386	0.6488	0.8755
	PRF/ANTT+ x PRF/ANTT+/OPENMETEO	0.3271	0.9409	0.1988
	PRF/ANTT+/OPENMETEO x PRF/ANTT+/OPENMETEO /DNIT	0.0342	0.1403	0.0834
Especificidade	PRF x PRF/ANTT	4.53e-8	6.50e-6	6.90e-7
	PRF/ANTT x PRF/ANTT+	0.0032	0.7031	0.7847
	PRF/ANTT+ x PRF/ANTT+/OPENMETEO	0.03051	0.8516	0.0538
	PRF/ANTT+/OPENMETEO x PRF/ANTT+/OPENMETEO /DNIT	0.0003	0.7173	0.6820
AUC	PRF x PRF/ANTT	9.43e-9	4.76e-8	2.20e-9
	PRF/ANTT x PRF/ANTT+	0.0032	0.9653	0.8299
	PRF/ANTT+ x PRF/ANTT+/OPENMETEO	0.0693	0.8900	0.1244
	PRF/ANTT+/OPENMETEO	0.3131	0.0728	0.4680

x
PRF/ANTT+/OPENMETEO
/DNIT

Fonte: Autores (2025)

É fundamental medir o desempenho dos modelos em dados não utilizados durante o treinamento, verificando se os algoritmos conseguem manter o desempenho para dados nunca vistos. A Tabela 11 apresenta o desempenho dos modelos nos dados de teste, permitindo avaliar sua capacidade de generalização.

Tabela 11 - Desempenho dos modelos para os dados de teste

Medida	Conjunto	RF	SVM	MLP
Sensibilidade	PRF	0.7475	0.7778	0.7525
	PRF/ANTT	0.8000	0.8488	0.8098
	PRF/ANTT+	0.8098	0.8488	0.8293
	PRF/ANTT+/OPEN METEO	0.8000	0.8488	0.8488
	PRF/ANTT+/OPEN METEO/DNIT	0.7659	0.8488	<u>0.8585</u>
Especificidade	PRF	0.7726	0.7850	0.7609
	PRF/ANTT	0.8544	0.8295	0.8482
	PRF/ANTT+	<u>0.8552</u>	0.8305	0.8412
	PRF/ANTT+/OPEN METEO	0.8402	0.8298	0.8286
	PRF/ANTT+/OPEN METEO/DNIT	0.8204	0.8307	0.8231

AUC	PRF	0.7650	0.7814	0.8265
	PRF/ANTT	0.8272	0.8392	0.8998
	PRF/ANTT+	0.8325	0.8396	0.9008
	PRF/ANTT+/OPEN METEO	0.8201	0.8393	0.9082
	PRF/ANTT+/OPEN METEO/DNIT	0.7931	0.8397	<u>0.9085</u>

Fonte: Autores (2025)

Os resultados obtidos nos dados de teste mostram que todos os modelos mantêm desempenho próximo com os dados de validação, indicando capacidade de generalização. Isto sugere que os modelos não sofreram sobreajuste nos dados de treinamento e conseguem manter seu desempenho para dados não vistos.

Considerações Finais

Este estudo teve como objetivo avaliar o impacto do enriquecimento dos dados na predição de acidentes rodoviários em Santa Catarina. Para isso, foram analisados registros de acidentes da PRF na BR-101/SC e identificadas novas fontes de dados a partir de revisão bibliográfica sobre predição de acidentes. Ao longo do estudo, foram identificadas inconsistências nos registros de acidentes da PRF, reforçando a necessidade de tratar o conjunto de dados antes de desenvolver modelos preditivos. O processo de enriquecimento foi realizado em etapas, permitindo avaliar o efeito incremental de cada grupo de atributos sobre o desempenho dos modelos RF, SVM e MLP.

Quanto à pergunta de pesquisa — se seria possível obter ganhos de desempenho apenas com o enriquecimento dos dados — os experimentos confirmaram a hipótese positivamente. O objetivo proposto foi alcançado, destacando-se que a etapa de correção dos dados da PRF foi a que gerou o maior impacto positivo, com ganhos estatisticamente significativos. As fases subsequentes trouxeram contribuições marginais: a rede MLP apresentou crescimento gradual, o SVM manteve-se estável e a RF exibiu leves declínios.

De modo geral, os resultados obtidos reforçam a importância do pré-processamento dos dados como etapa fundamental na construção de um modelo preditivo. Enquanto uma parte da literatura se concentra na seleção de algoritmos, este estudo apresentou que a qualidade dos dados de entrada também é um fator relevante.

Entre as dificuldades enfrentadas, destaca-se a ausência de dados de tráfego com a granularidade temporal e espacial utilizada em trabalhos relacionados. Essa limitação foi parcialmente contornada pelo uso do VMDA do DNIT, que fornece apenas médias anuais e não captura variações semanais ou horárias.

Como trabalhos futuros, sugere-se investigar arquiteturas mais avançadas de aprendizado profundo, capazes de capturar relações espaço-temporais complexas. Além disso, recomenda-se a incorporação de novas fontes de dados que não foram identificadas neste estudo, como informações de tráfego em maior granularidade temporal.

Agradecimentos

Os autores agradecem ao professor Renato Fileto, da Universidade Federal de Santa Catarina (UFSC), pelo apoio e contribuições ao longo do desenvolvimento da pesquisa, e aos revisores, pelos valiosos comentários.

Referências

AGÊNCIA NACIONAL DE TRANSPORTES TERRESTRES (ANTT). **Base de dados de rodovias federais**. Brasília: ANTT, 2025. Disponível em: <<https://dados.antt.gov.br/group/rodovias?page=2>>. Acesso em: 27 set. 2025.

CAI, Qing et al. Real-time crash prediction on expressways using deep generative models. **Transportation Research Part C: Emerging Technologies**, Amsterdã, v. 117, p. 102697, 2020. ISSN 0968-090X. Disponível em: <<https://doi.org/10.1016/j.trc.2020.102697>>. Acesso em: 27 set. 2025.

CONFEDERAÇÃO NACIONAL DO TRANSPORTE (CNT). **Anuário CNT do transporte 2022**. Brasília: CNT, 2022. Disponível em: <<https://cnt.org.br/documento/78a521c3-b71c-456b-85c8-e4ddf5e51166>>. Acesso em: 6 jul. 2025.

DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES (DNIT). **VGeo – Sistema de Informações Geográficas do DNIT**. Brasília: DNIT, 2025. Disponível em: <<https://servicos.dnit.gov.br/vgeo/>>. Acesso em: 7 mar. 2025.

DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES (DNIT). **Pesquisa Nacional de Contagem de Tráfego (PNCT)**. Brasília: DNIT, 2025. Disponível em: <<https://servicos.dnit.gov.br/dadospnct>>. Acesso em: 7 mar. 2025.

ELASSAD, Zouhair; MOUSANNIF, Hajar; AL MOATASSIME, Hassan. A proactive decision support system for predicting traffic crash events: A critical analysis of imbalanced class distribution. **Knowledge-Based Systems**, Amsterdã, v. 205, p. 106314, 2020. ISSN 0950-7051. Disponível em: <<https://doi.org/10.1016/j.knosys.2020.106314>>. Acesso em: 27 set. 2025.

FERREIRA, Martines et al. Utilização da inteligência artificial na predição de acidentes de trânsito na cidade de Campo Grande–MS. **Revista Camalotes**, Campo Grande, v. 3, n. 2, p. 1-11, 2024. ISSN 2965-7498. Disponível em: <<https://doi.org/10.62559/recam.v3i2.96>>. Acesso em: 18 dez. 2025.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6.ed. São Paulo: Atlas, 2008.

HAYKIN, Simon S. **Redes neurais: princípios e prática**. 2.ed. Porto Alegre: Bookman, 2001.

HUANG, Tingting; WANG, Shuo; SHARMA, Anuj. Highway crash detection and risk estimation using deep learning. **Accident Analysis & Prevention**, Amsterdã, v. 135, p. 105392, 2020. ISSN 0001-4575. Disponível em: <<https://doi.org/10.1016/j.aap.2019.105392>>. Acesso em: 27 set. 2025.

ISLAM, Zubayer et al. Crash data augmentation using variational autoencoder. **Accident Analysis & Prevention**, Amsterdã, v. 151, p. 105950, 2021. ISSN 0001-4575. Disponível em: <<https://doi.org/10.1016/j.aap.2020.105950>>. Acesso em: 27 set. 2025.

KRAUT, Caio; SAPIA, Helton Molina. Aprendizado de máquina utilizando agrupamento e regressão na previsão de locais de acidentes de trânsito em zonas urbanas. **Colloquium Exactarum**, São Paulo, v. 14, n. 1, p. 1–11, 2022. ISSN: 2178-8332. Disponível em: <<https://journal.unoeste.br/index.php/ce/article/view/4280>>. Acesso em: 18 dez. 2025.

MO, Weiwei et al. Dynamic short-term crash analysis and prediction at toll plazas for proactive safety management. **Accident Analysis & Prevention**, Amsterdã, v. 197, p. 107456, 2024. ISSN 0001-4575. Disponível em: <<https://doi.org/10.1016/j.aap.2024.107456>>. Acesso em: 27 set. 2025.

PENG, Yichuan et al. Examining imbalanced classification algorithms in predicting real-time traffic crash risk. **Accident Analysis & Prevention**, Amsterdã, v. 144, p. 105610, 2020. ISSN 0001-4575. Disponível em: <<https://doi.org/10.1016/j.aap.2020.105610>>. Acesso em: 27 set. 2025.

POLÍCIA RODOVIÁRIA FEDERAL (PRF). **Dados abertos da PRF**. Brasília: PRF, 2024. Disponível em: <<https://www.gov.br/prf/pt-br/aceso-ainformacao/dados-abertos/dados-abertos-da-prf>>. Acesso em: 22 jul. 2025.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining: Mineração de Dados**. Rio de Janeiro: Ciência Moderna, 2009.

TRAN, Thanh et al. MSGNN: A Multi-structured Graph Neural Network model for real-time incident prediction in large traffic networks. **Transportation Research Part C: Emerging Technologies**, Amsterdã, v. 156, p. 104354, 2023. ISSN 0968-090X. Disponível em: <<https://doi.org/10.1016/j.trc.2023.104354>>. Acesso em: 27 set. 2025.

WORLD HEALTH ORGANIZATION (WHO). **Global plan for the decade of action for road safety 2021–2030**. Geneva: WHO, 2021. Disponível em: <<https://www.who.int/pt/publications/m/item/global-plan-for-the-decade-of-action-for-road-safety-2021-2030>>. Acesso em: 23 jul. 2025.

WORLD HEALTH ORGANIZATION (WHO). **Global Status Report on Road Safety 2023**. Geneva: WHO, 2023. Disponível em: <<https://www.who.int/teams/social-determinants-ofhealth/safety-and-mobility/global-status-report-on-road-safety-2023>>. Acesso em: 26 jul. 2025.

YU, Le et al. Deep spatio-temporal graph convolutional network for traffic accident prediction. **Neurocomputing**, Amsterdã, v. 423, p. 135–147, 2021. ISSN 0925-2312. Disponível em: <<https://doi.org/10.1016/j.neucom.2020.09.043>>. Acesso em: 27 set. 2025.

ZHAO, Chao et al. Unsupervised anomaly detection based method of risk evaluation for road traffic accident. **Applied Intelligence**, Dordrecht, v. 53, p. 369–384, 2023. Disponível em: <<https://doi.org/10.1007/s10489-022-03501-8>>. Acesso em: 27 set. 2025.

ZHAO, Liping et al. Highway Traffic Crash Risk Prediction Method considering Temporal Correlation Characteristics. **Journal of Advanced Transportation**, Londres, 2023. Disponível em: <<https://doi.org/10.1155/2023/9695433>>. Acesso em: 27 set. 2025.

ZIPPENFENIG, Patrick. **Open-Meteo.com Weather API**. 2023. Disponível em: <<https://open-meteo.com/>> . Acesso em: 26 jul. 2025.