UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO

Gustavo Konescki Führ

Predição de acidentes rodoviários em Santa Catarina: o papel do enriquecimento de dados no desempenho dos modelos

Florianópolis
2025

Gustavo Konescki Führ

Predição de acidentes rodoviários em Santa Catarina: o papel do enriquecimento de dados no desempenho dos modelos

Trabalho de Conclusão de Curso submetido ao curso de Bacharelado em Ciências da Computação da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Ciências da Computação.
Orientador: Prof. Renato Fileto, Dr.
Coorientador: Prof. Eduardo Camilo Inacio, Dr.

Florianópolis
2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Führ, Gustavo Konescki
Predição de acidentes rodoviários em Santa Catarina : o papel do enriquecimento de dados no desempenho dos modelos / Gustavo Konescki Führ ; orientador, Renato Fileto, coorientador, Eduardo Inacio Camilo, 2025.
85 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2025.

Inclui referências.

1. Ciências da Computação. 2. aprendizado de máquina. 3. predição de acidentes. 4. integração de dados. I. Fileto, Renato. II. Camilo, Eduardo Inacio. III. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. IV. Título.

Gustavo Konescki Führ

Predição de acidentes rodoviários em Santa Catarina: o papel do enriquecimento de dados no desempenho dos modelos

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel em Ciências da Computação e aprovado em sua forma final pelo Curso de Graduação em Ciências da Computação.

Coordenação do Curso

Prof. Renato Fileto, Dr.
Orientador

Prof. Eduardo Camilo Inacio, Dr.
Coorientador
Universidade Federal de Santa Catarina

Prof. Andre Wust Zibetti, Dr.
Universidade Federal de Santa Catarina

Prof. Jônata Tyska Carvalho, Dr.
Universidade Federal de Santa Catarina

Florianópolis, 2025.

Este trabalho é dedicado à minha família e aos meus orientadores.

AGRADECIMENTOS

Agradeço à minha família pelo suporte durante a jornada acadêmica. Agradeço também aos meus orientadores pelo apoio técnico ao longo do desenvolvimento deste trabalho.

RESUMO

Os acidentes rodoviários representam um grande problema tanto para a segurança quanto para a economia, especialmente em Santa Catarina, que está entre os estados com maior número de ocorrências no Brasil. Diante disso, este trabalho tem como objetivo desenvolver um modelo de inteligência artificial capaz de prever acidentes rodoviários em Santa Catarina, avaliando o impacto do pré-processamento e do enriquecimento dos dados no desempenho dos modelos. A tarefa foi tratada como uma classificação binária, em que se busca identificar a ocorrência ou não de acidentes em trechos de 100 metros e intervalos de horas. Neste estudo, foram coletados dados sobre acidentes registrados pela Polícia Rodoviária Federal (PRF) no trecho entre os quilômetros 100 e 239 da BR-101 em SC, abrangendo 2017 a 2024. A pesquisa incluiu uma revisão bibliográfica focalizada em modelos de predição de acidentes, o que permitiu a identificação de novas fontes de dados e domínios de atributos comuns em outras análises. Foi realizado uma análise exploratória para cada conjunto de dados afim de entender seu conteúdo e detectar possíveis erros. Na análise, constatou-se que 74,15% dos acidentes no mesmo intervalo de 100 metros possuíam pelo menos uma discrepância nos atributos da via, sugerindo inconsistências nos dados da PRF. O enriquecimento dos dados ocorreu de forma sequencial, integrando novos atributos em cada etapa. Assim, foram criados cinco conjuntos de dados distintos para o treinamento de modelos, começando com os dados brutos da PRF e terminando na versão mais enriquecida. O primeiro, PRF, utiliza os dados brutos de acidentes registrados. Em seguida, PRF/ANTT corresponde à base corrigida, com ajustes nas inconsistências. A terceira versão, PRF/ANTT+, trouxe informações adicionais de atributos da via. PRF/ANTT+/OPENMETEO, foram adicionados dados meteorológicos. Por fim, PRF/ANTT+/OPENMETEO/DNIT incorporou dados de tráfego, formando o conjunto mais completo. Foram treinados três tipos de modelos de aprendizado de máquina: Random Forest (RF), Support Vector Machine (SVM) e Multilayer Perceptron (MLP), para cada um dos cinco conjuntos. Os resultados mostraram que a correção das inconsistências nos dados foi a etapa que mais impactou positivamente o desempenho, com melhorias em todas as métricas. Por exemplo, a sensibilidade dos modelos RF, SVM e MLP aumentaram em 5%, 7% e 5%, respectivamente. enquanto a adição de novos atributos produziu ganhos mais modestos, dependendo do modelo. Como os acidentes se concentram em determinados trechos da rodovia, o atributo quilômetro acabou se tornando dominante nos modelos, e por isso os experimentos foram repetidos sem essa variável para avaliar o impacto real dos novos dados. Essa modificação ocasionou uma queda geral de desempenho, evidenciando a importância do quilômetro para o trecho estudado, embora este padrão possa não se repetir em outras regiões. Ainda assim, observou-se que a incorporação de novos atributos da via elevou a sensibilidade dos modelos RF, SVM e MLP em 2%, 4% e 3%, respectivamente.

Palavras-chave: aprendizado de máquina, integração de dados, modelos preditivos, predição de acidentes.

ABSTRACT

Road accidents represent a major problem for both safety and the economy, especially in Santa Catarina, which is among the states with the highest number of occurrences in Brazil. Therefore, this work aims to develop an artificial intelligence model capable of predicting road accidents in Santa Catarina, evaluating the impact of data preprocessing and enrichment on model performance. The task was treated as a binary classification, in which the aim is to identify whether or not accidents occurred in 100-meter stretches and hourly intervals. In this study, data on accidents recorded by the Federal Highway Police (PRF) on the stretch between kilometers 100 and 239 of the BR-101 highway in SC, covering 2017 to 2024, were collected. The research included a literature review focused on accident prediction models, which allowed the identification of new data sources and domains of attributes common in other analyses. An exploratory analysis was performed for each dataset in order to understand its content and detect possible errors. The analysis revealed that 74.15% of accidents within the same 100-meter interval had at least one discrepancy in road attributes, suggesting inconsistencies in the PRF (Federal Highway Police) data. Data enrichment occurred sequentially, integrating new attributes at each stage. Thus, five distinct datasets were created for model training, starting with the raw PRF data and ending with the most enriched version. The first, PRF, uses the raw data of recorded accidents. Next, PRF/ANTT corresponds to the corrected dataset, with adjustments for inconsistencies. The third version, PRF/ANTT+, provided additional road attribute information. PRF/ANTT+/OPENMETEO added meteorological data. Finally, PRF/ANTT+/OPENMETEO/DNIT incorporated traffic data, forming the most complete dataset. Three types of machine learning models were trained: Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), for each of the five datasets. The results showed that correcting data inconsistencies was the step that most positively impacted performance, with improvements in all metrics. For example, the sensitivity of the RF, SVM, and MLP models increased by 5%, 7%, and 5%, respectively, while the addition of new attributes produced more modest gains, depending on the model. As accidents are concentrated in certain sections of the highway, the kilometer attribute ended up becoming dominant in the models, and therefore the experiments were repeated without this variable to assess the real impact of the new data. This modification caused an overall drop in performance, highlighting the importance of the kilometer for the studied section, although this pattern may not be repeated in other regions. Still, it was observed that the incorporation of new road attributes increased the sensitivity of the RF, SVM, and MLP models by 2%, 4%, and 3%, respectively.

Keywords: machine learning, data integration, predictive models, accident prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	Artificial Neural Network
ANTT	Agência Nacional de Transporte Terrestres
AUC	Area Under the ROC Curve
BAT	Boletim de Acidentes de Trânsito
CNN	Convolutional Neural Network
DCGAN	Deep Convolutional Generative Adversarial Network
DNIT	Departamento Nacional de Infraestrutura e Transporte
DSTGCN	Deep Spatio-Temporal Graph Convolutional Network
DT	Decision Tree
ECMWF	European Centre for Medium-Range Weather Forecasts
FNN	Feedforward Neural Network
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Logistic Regression
LSTM	Long Short-Term Memory
LSTM-CNN	Long Short-Term Memory Convolutional Neural Network
MLP	Multilayer Perceptron
MSGNN	Multi-Structure Graph Neural Network
OMS	Organização Mundial da Saúde
ONU	Organização das Nações Unidas
PNCT	Plano Nacional de Controle de Tráfego
PRF	Polícia Rodoviária Federal
RCSMLP	Random Cost-Sensitive Multilayer Perceptron
RF	Random Forest
ROC	Receiver Operating Characteristic Curve
SMOTE	Synthetic Minority Over-sampling Technique
SSAE-LSTM	Stacked Sparse Autoencoder with Long Short-Term Memory
SVM	Support Vector Machine
VMDA	Volume Médio Diário Anual
WMO	World Meteorological Organization

SUMÁRIO

1 INTRODUÇÃO

Os acidentes de trânsito estão entre os maiores problemas globais, com impactos na saúde e na economia. A pesquisa da Organização Mundial da Saúde (OMS), publicada em dezembro de 2023, menciona que 1,19 milhão de pessoas morrem anualmente em acidentes de trânsito, sendo a principal causa de morte entre crianças e adultos de 5 a 29 anos (WHO, 2023). Durante uma Assembleia Geral da Organização das Nações Unidas (ONU), em 2021, a OMS definiu meta de redução de 50% em mortes e lesões no trânsito para seus países membros até 2030 (WHO, 2021). Entretanto, este objetivo está longe da realidade brasileira, onde os índices de acidentes continuam elevados. Conforme os dados da Polícia Rodoviária Federal (PRF) de 2021, o Brasil registrou 71,9 mil feridos e 5,4 mil mortos. Já em 2024, houve um crescimento de 17,5% em feridos, totalizando 84,5 mil vítimas, e 9,2% em mortos, alcançando 6,1 mil óbitos (PRF, 2024a).

Durante o governo de Juscelino Kubitscheck (1956 - 1961), o Plano de Metas incluiu a construção e pavimentação de rodovias, o que deu início a vias importantes como a BR-116. O objetivo era priorizar as rodovias em detrimento das ferrovias, a fim de conectar o território brasileiro (USP, 1956). Desde então, as rodovias são o principal meio de transporte no país, sendo responsáveis por 75% da movimentação de carga em 2024 (Transportes, 2024).

Atualmente, a PRF desempenha um papel crucial na segurança viária, monitorando mais de 75 mil quilômetros de rodovias federais (PRF, 2024b). De acordo com a Lei n.º 9.503/97 - Código de Trânsito Brasileiro, cabe à PRF “efetuar levantamento dos locais de sinistros de trânsito e dos serviços de atendimento, socorro e salvamento de vítimas”, além de “coletar dados estatísticos e elaborar estudos sobre sinistros de trânsito e suas causas, adotando ou indicando medidas operacionais preventivas e encaminhando-os ao órgão rodoviário federal” (PRF, 2024b).

Para apoiar essas atividades, a PRF dispõe de dados de acidentes desde 2006. Esses dados são extraídos do Boletim de Acidentes de Trânsito (BAT) e atualizados mensalmente em arquivos denominados *datatran*, no formato CSV. Estes documentos incluem atributos importantes sobre os acidentes, tais como data, hora, localização, condição meteorológica, número de veículos envolvidos, número de pessoas envolvidas, feridos, óbitos, entre outros (PRF, 2024a).

A Figura 1 ilustra acidentes reportados pela PRF nas rodovias de Santa Catarina no período de 01/01/2017 a 31/12/2024. Os raios dos círculos em azul em torno de locais onde aconteceram acidentes indicam as quantidades de acidentes. O trecho da BR-101 entre os quilômetros 100 (Itajaí) e 239 (Palhoça) é onde se concentra o maior número de acidentes em rodovias federais de Santa Catarina. Assim, este trecho foi selecionado como objeto de estudo neste trabalho de pesquisa. Esta escolha permite

obter uma maior quantidade de dados para treinamento e avaliação de modelos.

Figura 1 – Acidentes nas estradas de Santa Catarina.

Fonte: Autor.

1.1 MOTIVAÇÃO

De acordo com os dados abertos da PRF, no período de 2017 até o final de 2024, Santa Catarina foi o segundo estado com maior número de acidentes em rodovias federais, registrando um total de 47,4 mil ocorrências, o que representa 11,94% dos acidentes em todo o território nacional. Nesse período, aproximadamente 130,1 mil veículos e 159,4 mil pessoas estiveram envolvidos em acidentes, resultando em 72,4 mil feridos e 3,0 mil mortes. Isto equivale a uma média de um óbito por dia, evidenciando a gravidade da situação (PRF, 2024a). Adicionalmente, dados do Painel CNT de Consultas Dinâmicas sobre Acidentes Rodoviários de 2022 indicam que Santa Catarina sofreu um prejuízo estimado de 1,32 bilhão de reais em decorrência desses acidentes (CNT, 2022).

1.2 JUSTIFICATIVA

Priorizar a segurança pública envolve buscar ferramentas que apoiem a redução de acidentes rodoviários. Integrado, por exemplo, a um mapa de acidentes em Santa Catarina, o modelo poderia indicar regiões com potencial de ocorrência de acidentes, permitindo direcionar ações preventivas.

Ademais, a utilização de sistemas de apoio como os modelos preditivos também contribui para a otimização de recursos públicos, pois permite que as autoridades de trânsito concentrem esforços nas áreas mais vulneráveis. Isto possibilita um planejamento de estratégias de segurança viária mais eficiente.

Por fim, a realização desta pesquisa se torna relevante para o entendimento e avanço do estado da arte no tema. A predição de acidentes de trânsito, utilizando modelos de aprendizado de máquina, é uma abordagem recente em escala global, sendo ainda menos explorada em pesquisas realizadas no Brasil.

1.3 DELINEAMENTO DA PESQUISA PROPOSTA

1.3.1 Problema

- Os números de acidentes em Santa Catarina são altos e continuam crescendo, passando de 7.890 ocorrências em 2021 para 8.381 em 2024.
- A predição com antecedência horária de acidentes em certos trechos de rodovias pode auxiliar no planejamento de atendimento aos mesmos.

1.3.2 Pergunta de Pesquisa

É possível desenvolver um modelo de aprendizado de máquina com capacidade preditiva adequada para acidentes rodoviários em Santa Catarina, e como o enriquecimento dos dados contribui para melhorar esse desempenho?

1.3.3 Hipótese

A aplicação de técnicas de pré-processamento, especialmente com a correção e o enriquecimento dos dados da PRF, contribui positivamente para o desempenho dos modelos de aprendizado de máquina na previsão de acidentes rodoviários em Santa Catarina.

1.4 OBJETIVOS

1.4.1 Objetivo Geral

Desenvolver um modelo de predição de acidentes rodoviários em Santa Catarina, analisando o impacto do pré-processamento aplicado aos dados da PRF.

1.4.2 Objetivos Específicos

1. Efetuar uma revisão bibliográfica com foco em predição de acidentes.
2. Identificar e coletar conjuntos de dados contendo atributos relacionados na ocorrência de acidentes.
3. Realizar uma análise exploratória dos dados para compreender padrões e possíveis inconsistências.
4. Executar processos de enriquecimento nos dados para garantir melhor qualidade no treinamento dos modelos.
5. Treinar diferentes modelos preditivos, utilizando abordagens de aprendizado de máquina.
6. Avaliar o desempenho dos modelos preditivos com base em métricas apropriadas.

1.4.3 Modelagem do estudo

A predição de acidentes de trânsito, neste estudo, pode ser entendida como uma tarefa de classificação binária, 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 como tipo de pista, traçado da via, fluxo de veículos, condições climáticas, entre outros, e um conjunto de classes $C = \{Acidente, Nao\ acidente\}$, este problema pode ser modelado como uma função de classificação em que X representa o espaço de atributos dos trechos analisados:

$$f : X \rightarrow C$$

Por exemplo, considerando um trecho com pista única, curva, chuva e tráfego intenso, o modelo pode classificá-lo como *Acidente*. Enquanto num trecho de pista dupla, reta, tempo estável e baixo fluxo, o resultado esperado seria *Nao acidente*.

1.5 MÉTODO DE PESQUISA

O desenvolvimento do trabalho foi estruturado em cinco etapas principais:

1. **Entendimento de negócio:** Compreender o objetivo do projeto e os problemas a serem resolvidos.
2. **Entendimento dos dados:** Explorar as informações coletadas, identificando padrões, inconsistências e características relevantes do conjunto de dados.
3. **Preparação dos dados:** Realizar o pré-processamento necessário que possa contribuir para a melhoria do desempenho dos modelos.

4. **Modelagem:** Treinar modelos baseados em técnicas de aprendizado de máquina.
5. **Avaliação:** Comparar a capacidade de generalização dos modelos, utilizando métricas corretas.

1.6 ESTRUTURA DO TRABALHO

A estrutura deste trabalho está organizada da seguinte forma: o Capítulo 2 apresenta a base teórica, abordando os conceitos e as fontes de dados utilizadas na pesquisa. O Capítulo 3 apresenta os estudos relacionados. Em seguida, o Capítulo 4 detalha o processo de entendimento do problema, entendimento e preparação dos dados, além da construção e avaliação dos modelos. O Capítulo 5 expõe os resultados obtidos. Por fim, no Capítulo 6 são apresentadas as conclusões gerais do estudo.

2 FUNDAMENTOS

Neste capítulo, serão apresentados todos os métodos, como os modelos preditivos, técnica de balanceamento, pré-processamento e validação, métricas de avaliação.

2.1 MODELOS PREDITIVOS DE APRENDIZADO DE MÁQUINA

Neste trabalho, foram utilizados três algoritmos clássicos, geralmente usados em outros trabalhos sobre predição de acidentes: Random Forest (RF), Support Vector Machine (SVM) e Multilayer Perceptron (MLP).

Floresta Aleatória

Floresta Aleatória (em inglês, Random Forest) é um método *ensemble* que utiliza a combinação de várias árvores de decisão. O modelo aplica *bagging* para geração de amostras no conjunto de treinamento das árvores de decisão e realiza uma seleção aleatória de atributos nos nós internos das árvores, aumentando a variabilidade das árvores de decisão, o que implica em um modelo final com melhor desempenho. Em tarefas de classificação, a combinação final é feita por meio de uma votação uniforme, ou seja, a moda das predições das árvores de decisão (Tan et al., 2009).

Para obter um modelo RF com bom desempenho é essencial a escolha e ajustes de hiperparâmetros. Entre eles, o **número de estimadores** define a quantidade de árvores de decisão do classificador. Geralmente, aumentar o número de árvores melhora o desempenho do modelo, embora também aumente o tempo necessário para o treinamento. Outro hiperparâmetro crucial é o **critério de divisão** que estabelece a função usada para medir a qualidade das divisões nas árvores de decisão. Certos critérios incluem o Gini, calculado pela Equação (1), que define a impureza de um nó, e a Entropia, definida pela Equação (2), que é uma medida de desordem. Além disso, a **profundidade máxima** indica o limite na profundidade de cada árvore de decisão. Modelos com profundidade pequena podem não ser capazes de capturar padrões nos dados, resultando em subajuste (*underfitting*). Por outro lado, modelos com profundidade excessiva podem se ajustar de maneira exagerada aos dados, o que leva ao sobreajuste (*overfitting*). Por fim, o uso do ***Bootstrap***, uma técnica de amostragem com reposição, cria diferentes conjuntos de dados para cada árvore de decisão, aumentando a diversidade das amostras (Scikit-learn developers, 2024a).

$$Gini(p) = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

$$Entropia(p) = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{2}$$

onde:

- p_i é a proporção de exemplos da classe *i* em um nó;
- C é o número total de classes;

Máquina de Vetores de Suporte

Máquina de Vetores de Suporte (em inglês, Support Vector Machine) é um método que visa encontrar um hiperplano com a maior margem possível para a separação entre as classes. Para isto, o algoritmo resolve a Equação (3), utilizando multiplicadores de Lagrange, que permitem transformar o problema de otimização em sua forma dual (Tan et al., 2009).

A equação geral de um hiperplano de separação é dada por:

$$h(\boldsymbol{x}) = \mathbf{w}^T \mathbf{x} + b \tag{3}$$

onde:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

- $\mathbf{w}$ é o vetor normal ao hiperplano;
- $\mathbf{x}$ é o vetor de variáveis de entrada;
- b é um escalar que representa o viés.

Do mesmo modo, é importante a regulagem dos hiperparâmetros no modelo SVM. O **C (Parâmetro de penalização)** decide o nível de penalização por classificações incorretas nos dados de treinamento. Um valor baixo de **C** permite erros de classificação, o que pode resultar em uma melhor generalização. Já um valor alto de **C** busca classificar corretamente todos os exemplos de treinamento, podendo levar ao sobreajuste. Ademais, a função ***Kernel*** define a transformação dos dados para um espaço de maior dimensionalidade, permitindo a separação de classes que não são linearmente separáveis. As funções kernel utilizadas neste estudo estão descritas na Tabela 1, que incluem os modelos Linear, Polinomial, RBF e Sigmoide. Finalmente, o hiperparâmetro **Gamma (γ)** decreta o impacto que uma amostra de treinamento tem no hiperplano que separa as classes. Valores maiores de γ fazem com que apenas exemplos muito próximos sejam considerados, enquanto valores menores ampliam esse efeito para pontos mais distantes (Scikit-learn developers, 2024b).

Tabela 1 – Funções *kernel* utilizadas no estudo.

Nome	**Função**
Linear	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j'$
Polinomial	$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\, \mathbf{x}_i^\top \mathbf{x}_j + r)^d$
RBF	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \lVert \mathbf{x}_i - \mathbf{x}_j \rVert^2)$
Sigmoide	$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\, \mathbf{x}_i^\top \mathbf{x}_j + r)$

Fonte: Autor.

Perceptron Multicamadas

Perceptron Multicamadas (em inglês, Multi-Layer Perceptron) é uma das arquiteturas de redes neurais mais utilizadas, por possuir um bom desempenho para a maioria dos problemas e, também, por servir de base para técnicas mais avançadas. A estrutura de uma MLP é composta por uma camada de entrada, por uma ou mais camadas intermediárias e por uma camada de saída. Cada camada possui *N* neurônios conectados a todos os neurônios da próxima camada. Os neurônios da rede realizam uma soma ponderada das entradas e utilizam funções de ativação para produzir uma saída, que servirá de entrada para os neurônios da próxima camada. Quando o modelo atinge a última camada, uma função de custo é calculada para medir o erro de predição. Este erro é propagado da camada de saída até a camada inicial por meio do *backpropagation*, ajustando os pesos de cada neurônio, a fim de minimizar a função de custo. Este processo é realizado até o modelo alcançar um desempenho ou número de épocas definidos (Haykin, 2001).

Entre os hiperparâmetros do modelo MLP, temos o **tamanho das camadas ocultas**, o qual define a arquitetura interna do modelo, ou seja, o número de camadas ocultas e o número de neurônios de cada camada. Arquiteturas muito complexas podem levar ao *overfitting*, enquanto arquiteturas muito simples podem resultar em *underfitting*. Também, a **função de ativação** determina como a saída de um neurônio será calculada com base na entrada, permitindo que a rede aprenda relações não lineares nos dados. Dentre elas estão a Tanh, conhecida como tangente hiperbólica, e a ReLU, unidade linear retificada. A **taxa de aprendizado** especifica a velocidade com que a rede neural aprende. Uma taxa de aprendizado muito alta pode resultar em instabilidade no modelo, dificultando a conversão em um ponto ideal. Por outro lado, uma taxa de aprendizado muito baixa torna a conversão do modelo mais lenta. Ademais, o **otimizador** caracteriza o algoritmo utilizado para ajustar os pesos da rede neural durante o treinamento. Diferentes otimizadores, como SGD e Adam, apresentam características distintas em relação aos ajustes. Além disso, a **taxa de dropout** fixa a proporção de neurônios que são removidos aleatoriamente durante o treinamento, ajudando a prevenir o *overfitting* ao forçar a rede a não depender excessivamente de

neurônios específicos. Por último, a **penalização dos pesos** estabelece a intensidade da regularização aplicada aos pesos da rede (Scikit-learn developers, 2024a).

2.2 SUBAMOSTRAGEM ALEATÓRIA

Treinar modelos de aprendizado de máquina com dados desbalanceados induz o modelo a priorizar a classe majoritária, o que resulta em uma baixa precisão com a classe minoritária. Dados de acidentes são extremamente desbalanceados, pois possuem muito mais registros de não acidentes do que de acidentes. A técnica de **subamostragem aleatória** consiste em descartar aleatoriamente observações da classe majoritária, a fim de balancear proporcionalmente as classes (Tan et al., 2009).

2.3 TÉCNICAS DE PRÉ-PROCESSAMENTO

Codificação one-hot

Esta técnica converte dados categóricos, como textos, para um formato numérico processável para o modelo. Este método transforma cada valor distinto de um atributo em colunas binárias (0 ou 1), indicando a presença ou ausência desse valor em cada observação (Faceli et al., 2011).

Padronização

Treinar modelos de aprendizado de máquina com diferentes escalas pode levar os modelos a priorizarem atributos com escalas maiores. Esta técnica arruma este problema, modificando os dados numéricos para possuírem média zero e desvio padrão um (Faceli et al., 2011).

A fórmula para calcular o valor padronizado é dada por:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

onde:

- z: valor padronizado
- x: valor original da atributo
- μ: média dos valores do atributo
- σ: desvio padrão dos valores do atributo

2.4 VALIDAÇÃO CRUZADA

Este procedimento divide o conjunto de dados em k partes e realiza um processo de iteração k vezes, onde o modelo é treinado com $k - 1$ partes e validado com

a parte restante. Assim, a média dos resultados das métricas de avaliação fornece uma estimativa confiável do desempenho do modelo (Faceli et al., 2011).

2.5 MÉTRICAS DE AVALIAÇÃO

Sensibilidade

Proporção dos casos corretamente preditos como positivos dentre todos os casos positivos (Tan et al., 2009).

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{5}$$

onde:

- *VP* (Verdadeiros Positivos): número de casos positivos corretamente identificados;
- *FN* (Falsos Negativos): número de casos positivos que foram classificados incorretamente como negativos.

Especificidade

Proporção dos casos corretamente preditos como negativos dentre todos os casos negativos (Tan et al., 2009).

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{6}$$

onde:

- *VN* (Verdadeiro Negativos): número de casos negativos corretamente identificados;
- *FP* (Falsos Positivos): número de casos negativos que foram classificados incorretamente como positivos.

Area Under the ROC Curve (AUC)

Corresponde à área sob a Receiver Operating Characteristic Curve (ROC), que relaciona a taxa de verdadeiros positivos (sensibilidade) com a taxa de falsos positivos (1 - especificidade) sobre diferentes limites de decisão. Quanto mais próximo de 1 for o valor da AUC, melhor é a capacidade do modelo em distinguir entre as classes (Faceli et al., 2011).

3 TRABALHOS RELACIONADOS

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 seguintes repositórios acadêmicos: *ScienceDirect*, *IEEE Xplore*, *Scopus* e *Google Scholar*.

As principais palavras-chave utilizadas nas buscas foram: *"traffic accident prediction"*, *"road crash prediction"* e *"traffic safety machine learning"*.

Os critérios de inclusão adotados foram:

- Estudos que aplicam aprendizado de máquina para predição de acidentes;
- Documentação suficiente sobre os atributos utilizados;
- Trabalhos escritos em inglês.

Estudos focados exclusivamente na predição da severidade dos acidentes foram excluídos da pesquisa.

A partir desta pesquisa, foram selecionados dez artigos. Nesses estudos, observou-se quais domínios de atributos são mais frequentemente usados no treinamento de modelos. A Tabela 2 apresenta os principais domínios e os respectivos artigos que os utilizaram. Os atributos foram organizados em diferentes domínios. O tráfego reúne informações sobre volume e fluxo de veículos; o tempo físico corresponde a data e hora do acidente; o local físico descreve a posição geográfica; as condições meteorológicas incluem fatores como chuva, temperatura, etc; as características da via abrangem elementos estruturais, como número de faixas, velocidade da via, etc; e, por fim, os pontos de interesse indicam a proximidade de locais relevantes, como estabelecimentos. Observa-se que a maior parte dos estudos empregou dados de tráfego. Os outros domínios ocorrem com menor quantidade, 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 2 – Domínios de atributos mais utilizados em trabalhos relacionados.

Autor	Tráfego	Tempo físico	Local físico	Cond. meteorológicas	Características da via	Pontos de interesse
Cai et al. (2020)	✓					
Huang et al. (2020)	✓	✓	✓			
Peng et al. (2020)	✓					
Elassad et al. (2020)				✓	✓	
Yu et al. (2021)	✓	✓	✓	✓	✓	✓
Islam et al. (2021)	✓					
Zhao et al. (2022)	✓					
Tran et al. (2023)	✓	✓	✓			
Zhao et al. (2023)	✓					
Mo et al. (2024)	✓			✓	✓	
Este estudo (2025)	✓	✓	✓	✓	✓	

Fonte: Autor.

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.

Também vale ressaltar que há outros domínios que influenciam na ocorrência de acidentes de trânsito, como o fator humano e as condições do veículo, mas que não podem ser representados diretamente nos modelos. O fator humano está relacionado a dirigir embriagado, usar o celular durante a direção, condução perigosa, entre outros, enquanto as condições do veículo estão associadas à manutenção dos componentes como freios, faróis, sistema elétrico, etc.

A seguir, serão apresentados os principais trabalhos analisados na Tabela 2 com base em sua relevância para o tema de predição de acidentes, mostrando o objetivo do estudo, os conjuntos de dados e os modelos utilizados e os principais resultados.

Os Cai et al. (2020) estudaram o uso da técnica de aprendizado profundo chamada Deep Convolutional Generative Adversarial Network (DCGAN) no balanceamento entre dados de acidentes e não acidentes para o conjunto de treinamento. Este método usa duas redes neurais que competem entre si para gerar dados sintéticos realistas de acidentes. Para a pesquisa, utilizou dados de acidentes e de condições de tráfego da rodovia SR 408 de Orlando, em 2017, capturando 6.749.447 registros de não-acidentes e 625 registros de acidentes. O DCGAN foi comparado com outras duas técnicas de balanceamento: Synthetic Minority Over-sampling Technique (SMOTE) e

subamostragem aleatória. Quatro modelos de aprendizado de máquina como Logistic Regression (LR), SVM, Artificial Neural Network (ANN) e Convolutional Neural Network (CNN) foram treinados com cada técnica de balanceamento, totalizando doze modelos. Os resultados mostraram que, em modelos mais complexos como a ANN e CNN, o DCGAN obteve um desempenho bem superior em relação a outras técnicas de balanceamento, alcançando métricas de 88,8% de sensibilidade, 90,7% de especificidade e 95,6% de AUC. Também é relevante mencionar que, em modelos complexos, a técnica de subamostragem aleatória teve um desempenho bem abaixo dos outros dois métodos de sobreamostragem, devido ao baixo número de registros de acidentes para o treinamento.

O artigo de Yu et al. (2021) propõe um novo modelo para predição de acidentes chamado Deep Spatio-Temporal Graph Convolutional Network (DSTGCN). O modelo possui três partes principais: uma camada para aprender relações espaciais, outra para capturar padrões espaço-temporais, e uma última que incorpora informações externas de forma semântica. Este foi o trabalho que utilizou a maior diversidade de dados entre os trabalhos analisados. Foram utilizados dados de acidentes, tempo físico, local físico, condições de tráfego, condições meteorológicas, características da via e pontos de interesse da cidade de Beijing entre 01/08/2018 até 31/10/2018. Além do modelo proposto, foram treinados quatro modelos clássicos como Decision Tree (DT), LR, SVM e Least Absolute Shrinkage and Selection Operator (LASSO), e dois modelos do estado da arte SdAE e TARPML. O DSTGCN apresentou melhor desempenho em todas as métricas de avaliação em relação aos outros modelos, com 85,73% de F1-SCORE, 89,68% de sensibilidade e 85,08% de AUC, evidenciando uma melhor captura de padrões nos dados.

O trabalho de Peng et al. (2020) analisa diferentes meios de tratar dados de acidentes desbalanceados para dados de treinamento. Nesta pesquisa, foram investigados métodos no nível de dados, como SMOTE e subamostragem aleatória, métodos no nível de algoritmo, como MLP sensível ao custo e AdaBoost, e métodos no nível de saída, como Índice de Youden e Método de Calibração de Probabilidade. Foram utilizados dados de acidentes e condições de tráfego de rodovia em Shangai, obtendo uma amostra de 15.427.225 registros de não acidentes e 2.338 registros de acidentes. Os melhores resultados foram obtidos com os modelos Random Cost-Sensitive Multilayer Perceptron (RCSMLP) e Rusboost, que integram as estratégias nos níveis de dados, algoritmo e saída, em comparação a outros modelos clássicos. O modelo Rusboost atingiu métricas de 84,21% de sensibilidade, 81,62% de especificidade e 89% de AUC.

Mo et al. (2024) propuseram um novo modelo para predição de acidentes chamado Long Short-Term Memory Convolutional Neural Network (LSTM-CNN). Nos dados, utilizaram registros de acidentes de praças de pedágio na Flórida, com outras informações de condições de tráfego, condições meteorológicas e características da

via, o que resultou em um conjunto de 1.233.792 amostras de não-acidentes e 160 amostras de acidentes. Para lidar com o desequilíbrio nos dados de treinamento, utilizaram a técnica SMOTE e métodos de ponderação de classe. O LSTM-CNN foi avaliado em comparação a outros dois algoritmos de aprendizado de máquina como - Stacked Sparse Autoencoder with Long Short-Term Memory (SSAE-LSTM) e CatBoost - para cada técnica de balanceamento, totalizando nove combinações de modelos. O melhor modelo foi o LSTM-CNN com SMOTE, apresentando 93,7% de sensibilidade, 88,4% de especificidade e 96,3% de AUC.

Na pesquisa de Tran et al. (2023), formularam um modelo denominado Multi-Structure Graph Neural Network (MSGNN), o qual captura relações espaço-temporais entre links de uma subárea por meio de múltiplos grafos estruturados a partir de diferentes fontes de dados. Foram usados dados de acidentes, condições de tráfego e tempo e local físico de duas rodovias em Queensland. Os dados foram balanceados usando a técnica de subamostragem aleatória, totalizando 1.360 registros de não acidentes e 1.360 registros de acidentes. O desempenho do modelo proposto foi comparado em diferentes intervalos de predição, com outros seis modelos: SVM, SVM com *kernel* de grafo, RF com *kernel* de grafo, CNN, Feedforward Neural Network (FNN) e Long Short-Term Memory (LSTM). O modelo MSGNN teve o melhor desempenho com métricas de 89,6% de acurácia, 89,6% de sensibilidade, 89,5% de precisão, 89,5% de F1-SCORE, 93,4% de AUC, para o intervalo de predição de 15 minutos na rodovia de Gold Coast.

Assim como os trabalhos relatados, os demais estudos utilizados na fundamentação teórica concentram-se na investigação e desenvolvimento de técnicas de balanceamento dos dados e de modelos preditivos de aprendizado de máquina mais complexos. No entanto, nenhum destes estudos se dedica especificamente ao processo de aprimoramento, entendido aqui como a correção e o enriquecimento dos dados de acidentes dos dados, como o proposto nesta pesquisa.

4 METODOLOGIA

Foi utilizado o conjunto de dados de acidentes da PRF como o conjunto base. Através de uma pesquisa bibliográfica com foco em modelos de predição de risco de acidentes, foram identificadas novas fontes de dados com domínios de atributos utilizados em outras pesquisas. Cada conjunto foi previamente analisado, a fim de compreender seus dados e encontrar possíveis erros. O processo de enriquecimento foi de forma sequencial, ou seja, o conjunto de dados de uma etapa possui todos os atributos das etapas anteriores juntamente com os atributos do novo conjunto. Desta forma, foram gerados cinco conjuntos distintos para treinamento dos modelos, iniciando apenas com os dados da PRF e progredindo até a última versão totalmente enriquecida. Todos esses conjuntos foram treinados com diferentes modelos de aprendizado de máquina, com o objetivo de avaliar como cada etapa de enriquecimento influencia no desempenho dos modelos.

Os experimentos foram realizados utilizando a linguagem `Python` 3.9.13 em ambiente `Jupyter Notebook`. As bibliotecas principais usadas incluem `numpy` (versão 1.26.2) e `pandas` (versão 2.1.1) para manipulação de dados, `matplotlib` (versão 3.9.3) e `seaborn` (versão 0.13.2) para visualização de dados, `scikit-learn` (versão 1.6.1) e `tensorflow` (versão 2.18.0) para os modelos de aprendizado de máquina. Para o treinamento dos modelos, foi utilizada a plataforma `Kaggle`, que oferece suporte a TPU e GPU gratuitas [1].

4.1 ENTENDIMENTO DE NEGÓCIO

O entendimento do negócio neste trabalho parte do reconhecimento de que os acidentes rodoviários representam um grande problema de segurança e de gestão em Santa Catarina. O modelo pode servir de forma prática para a construção de um mapa de Santa Catarina, no qual seriam destacados os trechos da rodovia com a previção de ocorrência de acidentes. Essa visualização permitiria identificar áreas críticas e apoiar órgãos como a PRF na distribuição eficiente de recursos. Para que essa aplicação apresente resultados confiáveis, é necessário aprimorar a base de dados utilizada, incorporando atributos associados aos acidentes.

No capítulo de trabalhos relacionados, notou-se que a maioria dos estudos de predição de acidentes utilizou dados de tráfego, sendo que alguns também incorporam atributos relacionados ao tempo físico, local físico, condições meteorológicas e características da via. O conjunto de dados de acidentes da 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. Portanto, o objetivo

[1] `https://www.kaggle.com/`

é adicionar atributos dos domínios relacionados às características da via, às condições meteorológicas e ao tráfego.

Para os atributos do domínio de características da via, decidiu-se por usar inicialmente aqueles disponibilizados pela PRF para o conjunto inicial e, em seguida, enriquecer com dados mais completos fornecidos pela Agência Nacional de Transporte Terrestres (ANTT). Em relação às condições meteorológicas, o conjunto da PRF apresenta um único atributo que representa a condição do tempo no momento do acidente (como Céu claro, Chuva, entre outros). Optou-se por não utilizar este atributo, uma vez que está disponível somente nos registros de acidentes, o que inviabiliza sua aplicação nos dados de não acidentes. Assim, decidiu-se por enriquecer posteriormente com dados meteorológicos mais detalhados obtidos por meio da API da OPEN-METEO. Por fim, para complementar o conjunto com atributos do domínio de tráfego, será incorporado o conjunto de dados VMDA do Departamento Nacional de Infraestrutura e Transporte (DNIT).

4.2 ENTENDIMENTO DOS DADOS

O entendimento dos dados é uma atividade fundamental no processo da metodologia adotada. Nesta etapa, será realizada uma investigação aprofundada de cada conjunto de dados empregados no treinamento dos modelos preditivos, incluindo a descrição detalhada de cada atributo e a execução de uma análise exploratória desses dados.

4.2.1 Conjunto de dados de acidentes da PRF

Este conjunto de dados foi obtido através do portal de dados abertos da PRF e contém registros detalhados de acidentes de trânsito ocorridos nas rodovias federais brasileiras. As informações incluem data, hora, localização (UF, município, rodovia, km, latitude e longitude), tipo de acidente, número de veículos envolvidos, vítimas, óbitos, entre outros atributos (PRF, 2024a). Estes dados são a base principal para o treinamento e para a avaliação dos modelos preditivos desenvolvidos.

Foram usados arquivos da PRF referentes aos anos de 2017 a 2024. Posteriormente, os dados foram concatenados e submetidos a uma filtragem, considerando apenas os registros com a unidade da federação de Santa Catarina, rodovia BR-101 e quilômetros entre 100 e 239. Após este processo, obteve-se uma amostra final composta por 20.656 registros de acidentes.

Os atributos selecionados para o treinamento dos modelos estão descritos na Tabela 3. Foram escolhidos atributos geralmente utilizados em outras pesquisas de predição de acidentes de trânsito como: tempo físico (Yu et al., 2021; Tran et al., 2023; Huang et al., 2020), local físico (Yu et al., 2021; Huang et al., 2020) e características

da via (Elassad et al., 2020; Yu et al., 2021; Mo et al., 2024).

Tabela 3 – Atributos utilizados dos dados da PRF.

Atributos	**Descrição**
Data	Data no formato `dd/mm/aa`.
Horário	Horário no formato `hh:mm:ss`.
Km	Identificação do quilômetro da via onde ocorreu o acidente, com precisão de 0,1 km.
Município	Nome do município onde ocorreu o acidente.
Sentido da via	Sentido da via considerando o ponto de colisão. Ex: Crescente, Decrescente.
Tipo de pista	Categoria da quantidade de faixas da via principal. Ex: Simples, Dupla, Múltipla.
Traçado da via	Característica do tipo de traçado da via. Ex: Reta, Curva, Aclive, etc.
Uso do solo	Tipo de ocupação do solo. Ex: Sim (urbano), Não (rural).

Fonte: Adaptado do dicionário de dados da PRF.

Data

O atributo original *data* foi decomposto em três atributos distintos: *dia*, *mês* e *ano*. Além disso, dois atributos adicionais foram derivados: *dia da semana* que representa o dia da semana correspondente à data do acidente (segunda-feira, terça-feira, ..) e *feriado*, o qual indica se o acidente ocorreu em um feriado nacional.

A Figura 2 apresenta a quantidade de acidentes registrados para cada dia do mês. Nota-se uma distribuição relativamente homogênea, o que sugere que o dia do mês não é um forte fator preditivo para acidentes.

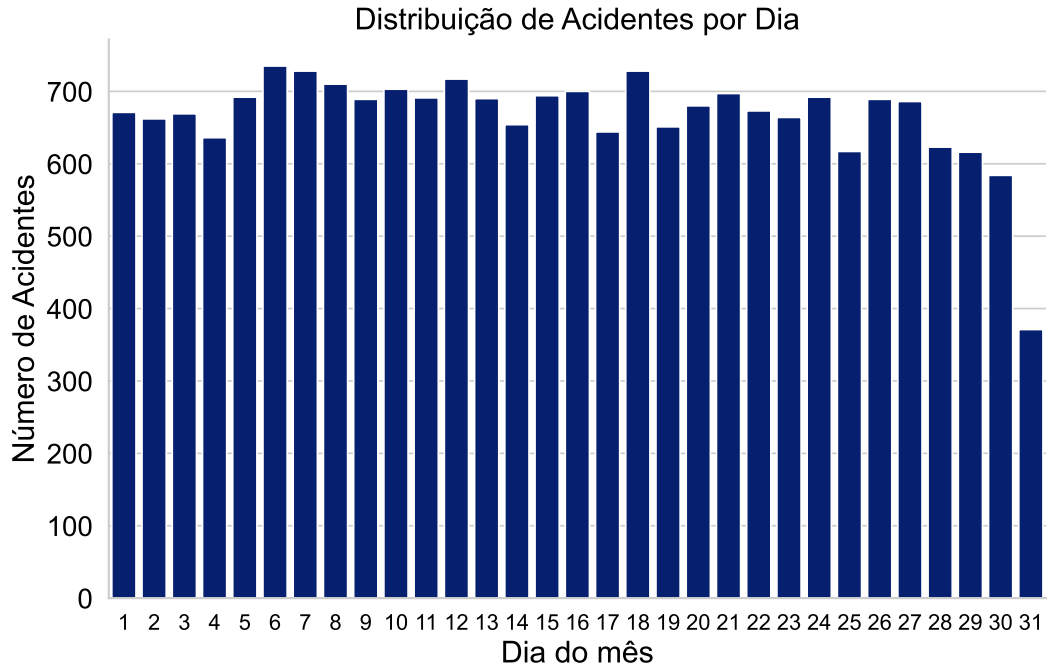
Na Figura 3, observa-se a distribuição de acidentes ao longo dos meses do ano. É possível identificar uma leve tendência de aumento nos meses de dezembro e janeiro, possivelmente relacionada ao maior fluxo de veículos devido ao período de pico turístico de Santa Catarina.

A Figura 4 mostra a evolução do número de acidentes ao longo dos anos. Observa-se uma queda em 2020, provavelmente associada à pandemia da COVID-19. Após esse período, nota-se uma retomada gradual no número de acidentes.

A distribuição da Figura 5 revela que os acidentes tendem a ocorrer com maior frequência na sexta e no sábado. Isso pode estar relacionado ao aumento do tráfego durante finais de semana.

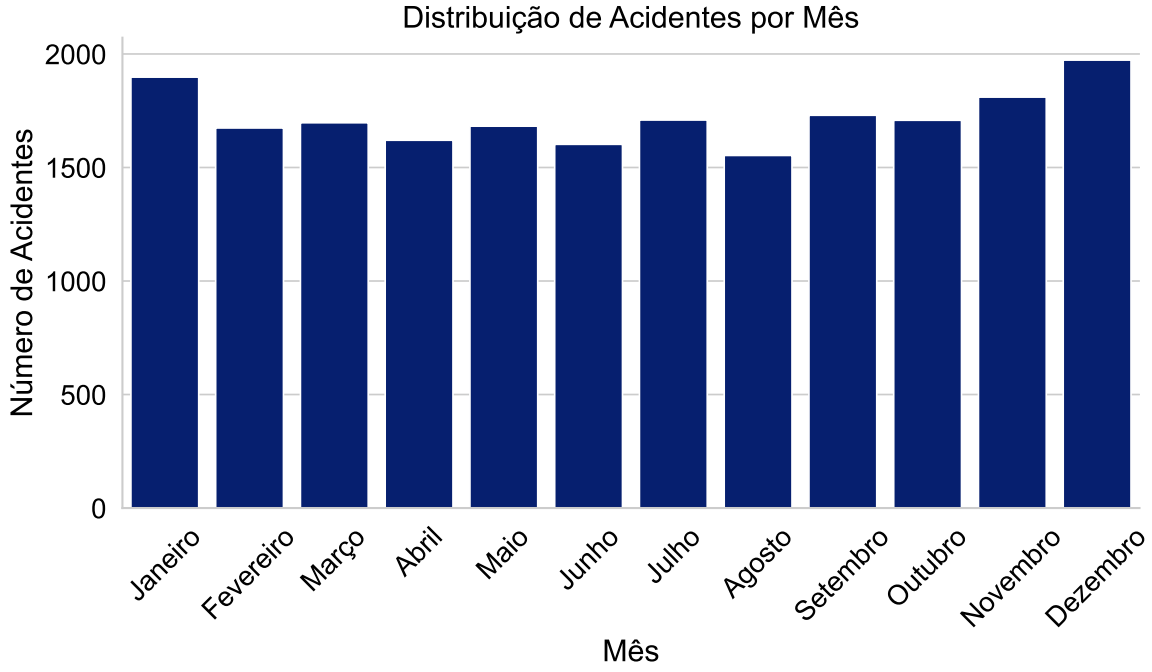
A média de acidentes é comparada entre dias comuns e feriados nacionais na Figura 6. O gráfico mostra que a média de acidentes em dias comuns e em dias com

Figura 2 – Distribuição de acidentes por dia do mês.



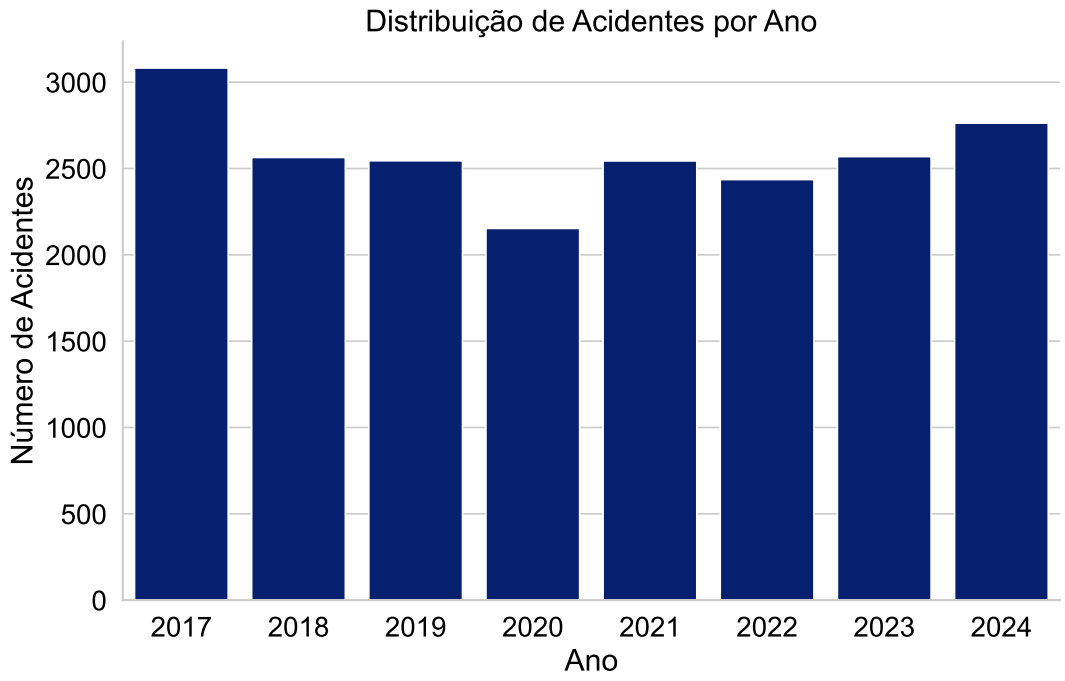
Fonte: Autor.

Figura 3 – Distribuição de acidentes por mês.



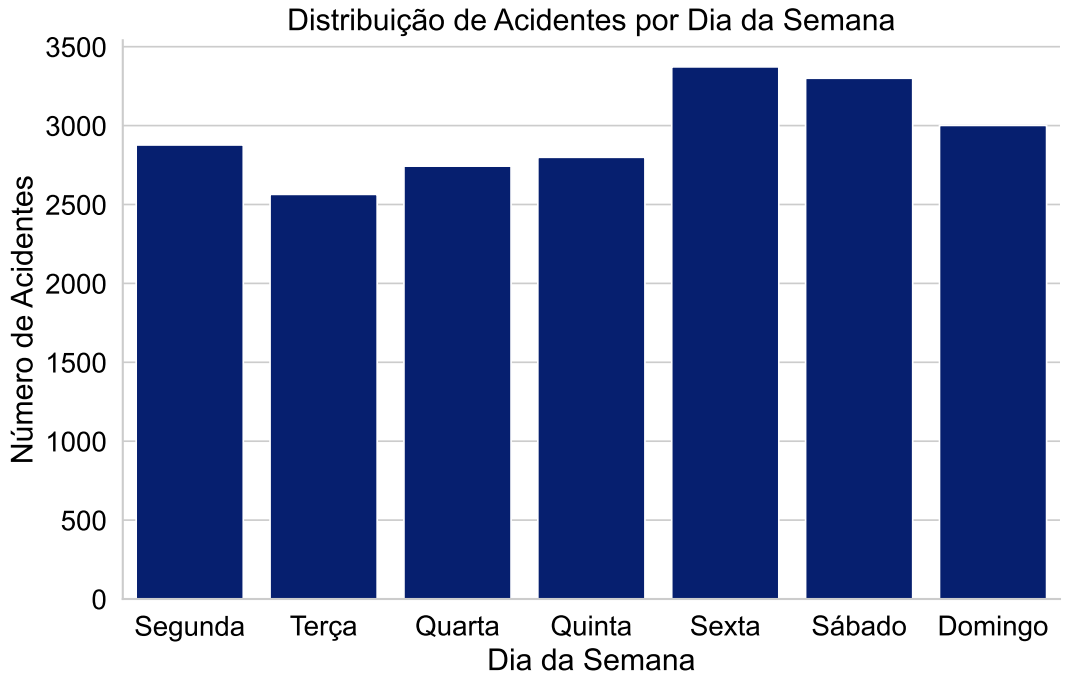
Fonte: Autor.

Figura 4 – Distribuição de acidentes por ano.



Fonte: Autor.

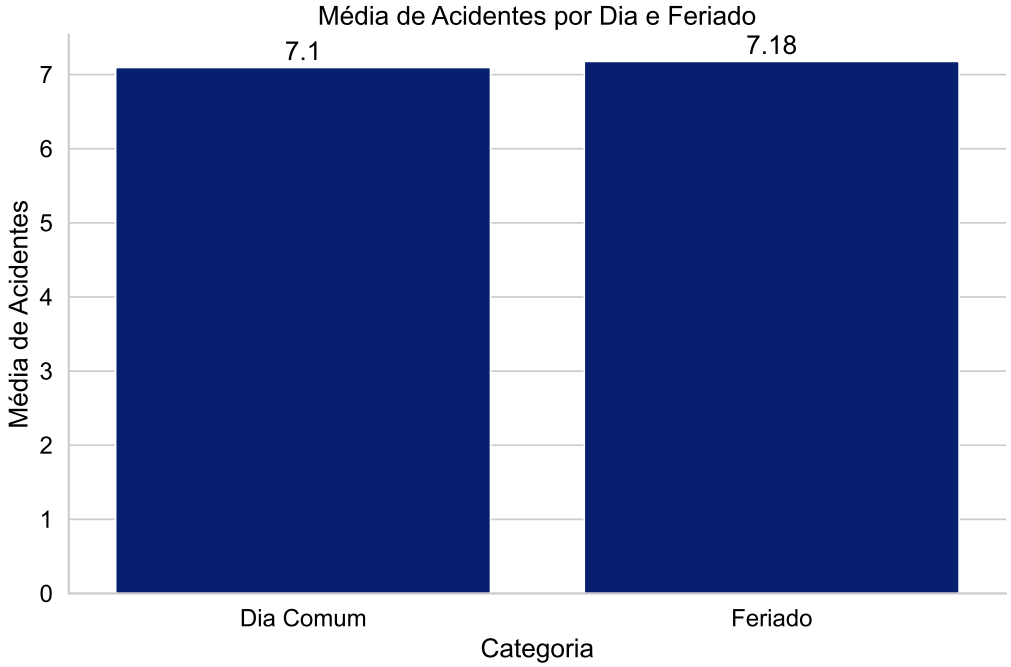
Figura 5 – Distribuição de acidentes por dia da semana.



Fonte: Autor.

feriados é muito próxima, o que propõe que o atributo pode não ser um bom indicador preditivo para os modelos.

Figura 6 – Média de acidentes em dias comuns e feriados nacionais.



Fonte: Autor.

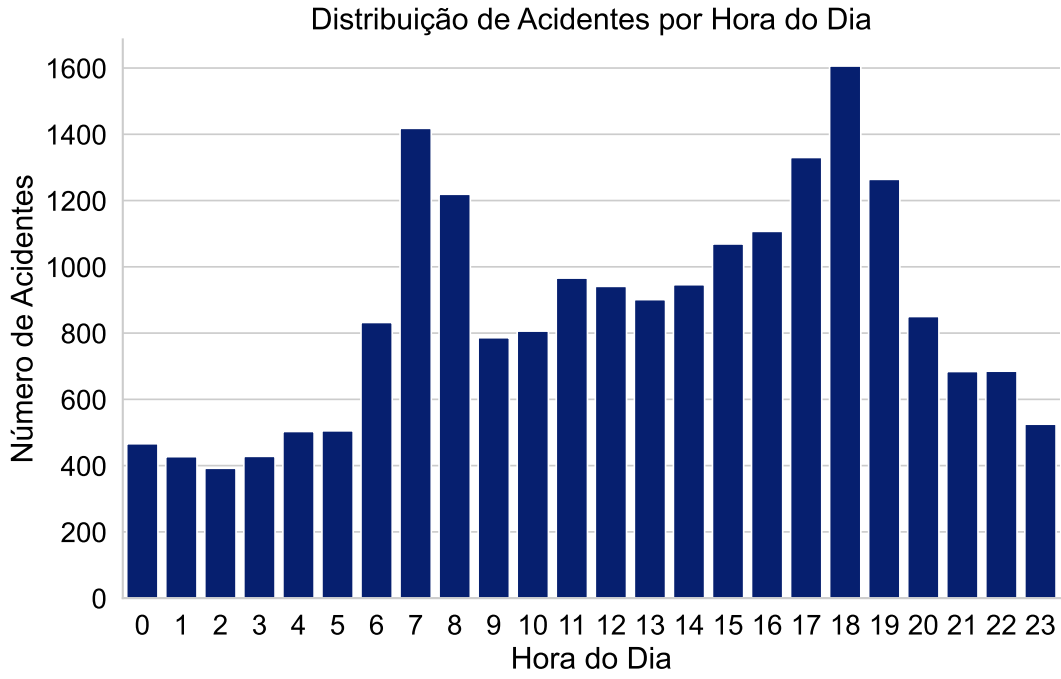
Horário

O atributo *horário* representa o instante exato do acidente, registrado no formato `hh:mm:ss`. Para fins de análise, o campo foi categorizado em faixas horárias de 1 em 1 hora. A Figura 7 mostra a distribuição dos acidentes por hora do dia. Observa-se um maior número de acidentes em horários de pico, das 7h às 8h e entre 17h e 19h, horário compatível com início e fim de expediente. Isto pode indicar um maior risco de acidentes durante o maior fluxo de veículos.

Km

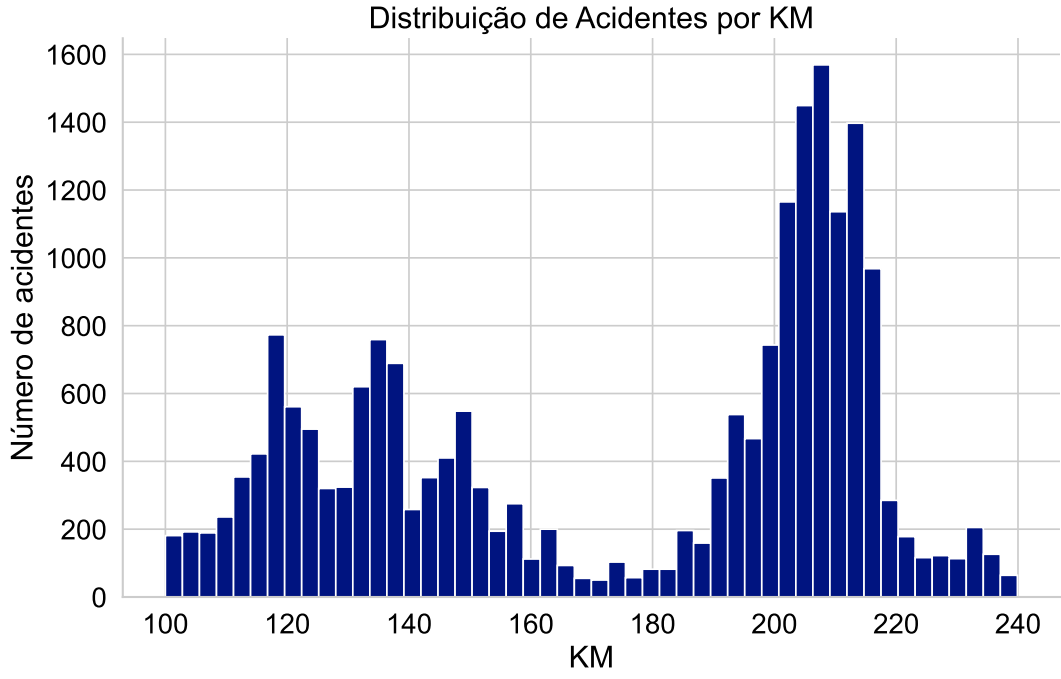
O atributo *km* indica o quilômetro da rodovia onde o acidente foi registrado, com precisão de até uma casa decimal. A Figura 8 mostra uma maior concentração de acidentes em regiões de grandes cidades, Itajaí, Balneário Camboriú e a região urbana da Grande Florianópolis, o que também sugere uma correlação entre acidentes e áreas com maior volume de tráfego.

Figura 7 – Distribuição de acidentes por horário do dia.



Fonte: Autor.

Figura 8 – Distribuição de acidentes por quilômetro.

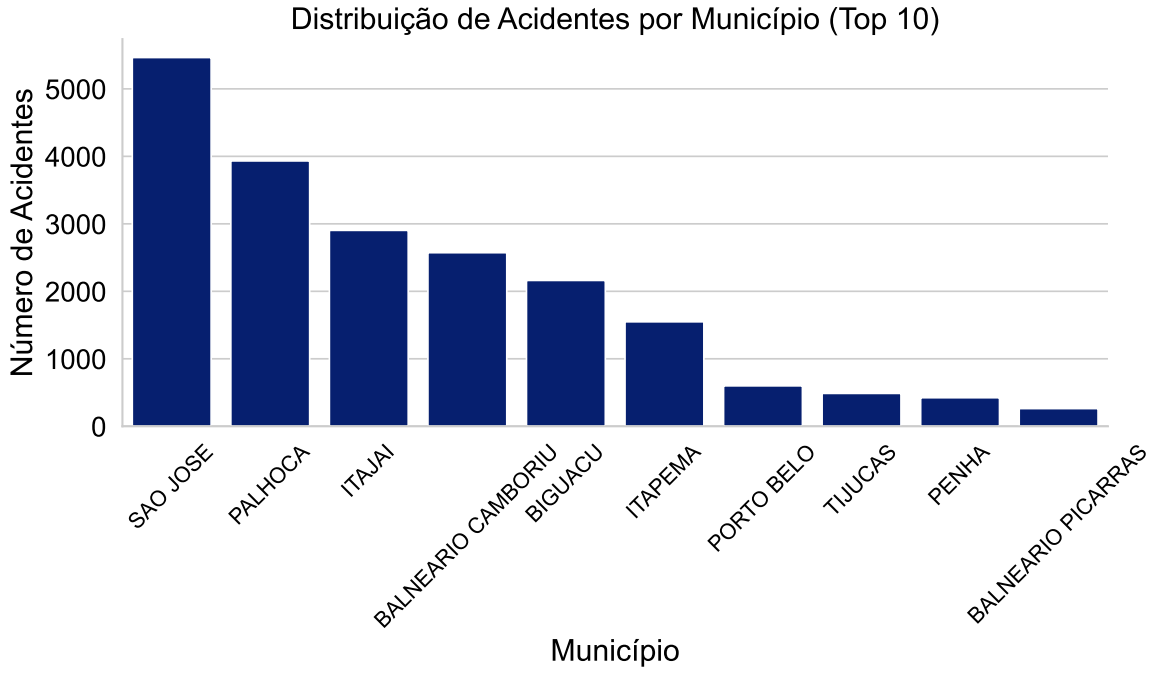


Fonte: Autor.

Município

O atributo *município* corresponde ao nome do município no qual o acidente foi registrado. A Figura 9 revela o mesmo princípio que a Figura 8, onde as maiores cidades acumulam maior número de acidentes.

Figura 9 – Distribuição de acidentes por município.



Fonte: Autor.

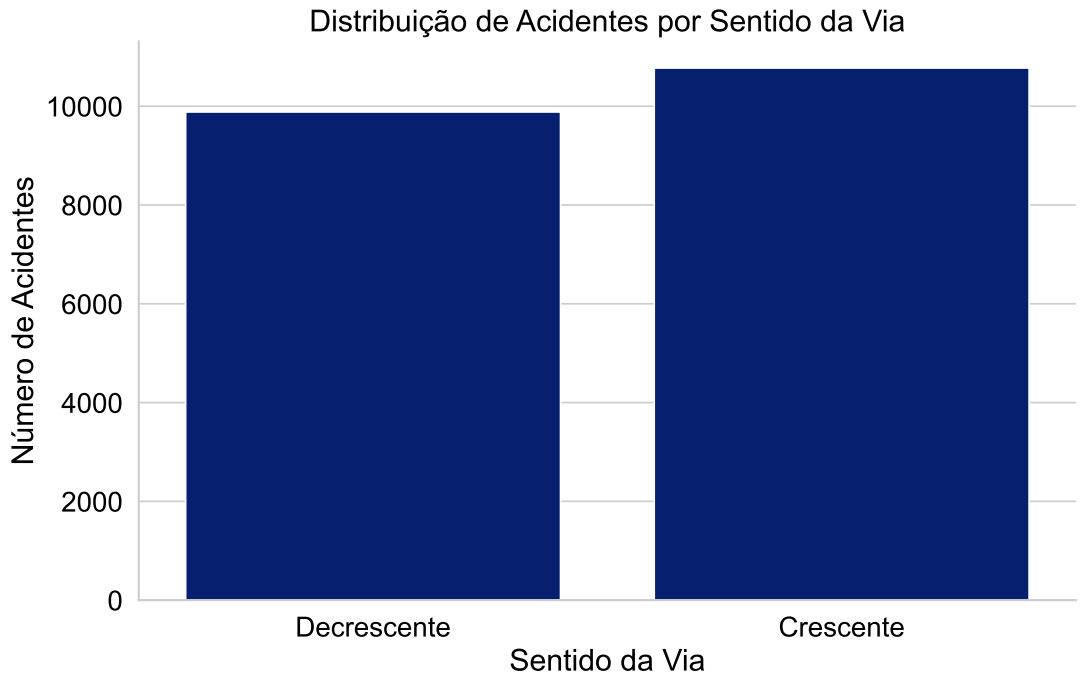
Sentido da via

Representa o sentido da via no momento do acidente, considerando a direção crescente ou decrescente do fluxo da rodovia. A Figura 10 mostra uma leve predominância de acidentes no sentido crescente da via.

Tipo de pista

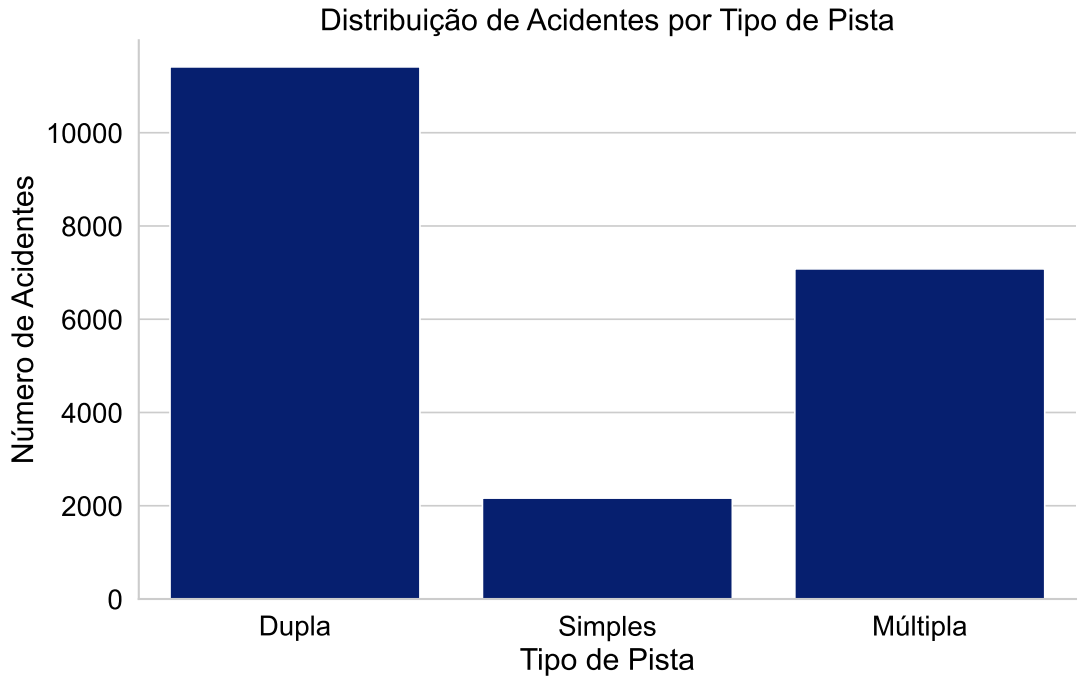
Classifica a pista quanto ao número de vias, podendo indicar se é uma via simples, dupla ou múltipla. A Figura 11 indica que a maioria dos acidentes ocorre em pistas do tipo dupla, o que pode refletir o fato de que esse tipo de via é predominante no trecho analisado da BR-101.

Figura 10 – Distribuição de acidentes por sentido da via.



Fonte: Autor.

Figura 11 – Distribuição de acidentes por tipo de pista.

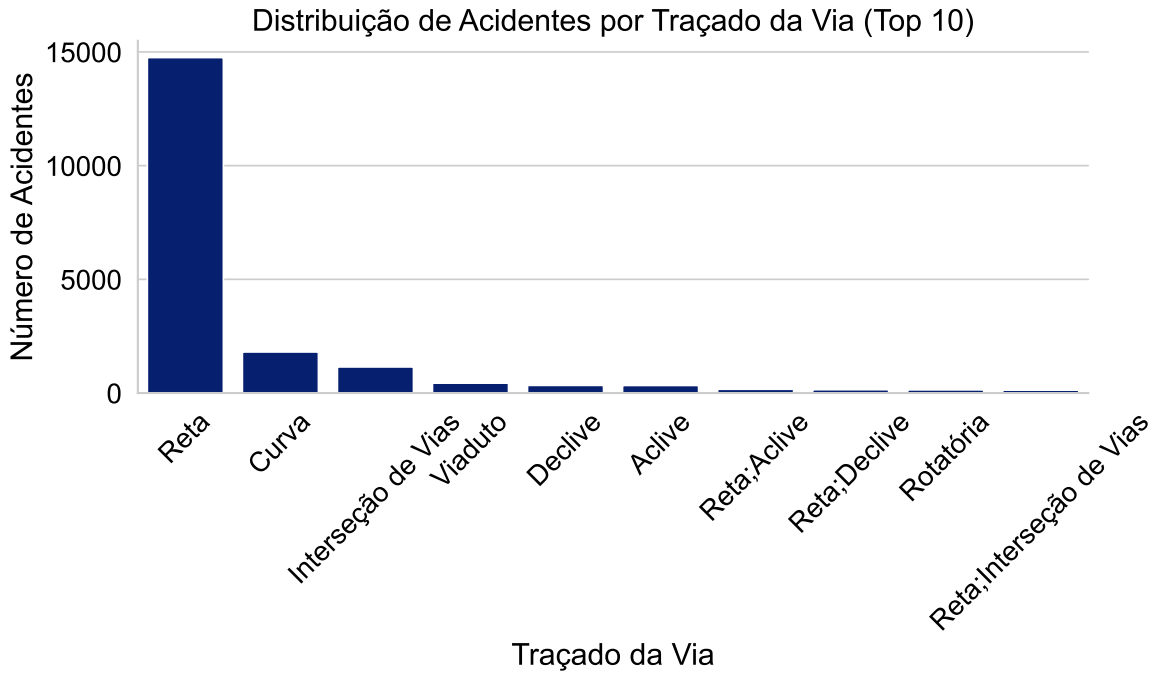


Fonte: Autor.

Traçado da via

Refere-se à característica geométrica do trecho da rodovia onde ocorreu o acidente. Ela apresenta 12 categorias distintas: Reta, Curva, Interseção de Vias, Viaduto, Declive, Aclive, Rotatória, Retorno Regulamentado, Ponte, Em Obras, Túnel e Desvio Temporário. É possível que um mesmo registro contenha múltiplas classificações separadas por ponto e vírgula, como por exemplo, Reta;Declive, indicando que o acidente aconteceu em um trecho de reta com declive. A Figura 12 exibe os dez tipos de traçado mais presentes nos acidentes do trecho analisado. Percebe-se uma superioridade de acidentes em retas, o que pode estar relacionado ao fato de que a maior parte da rodovia analisada possui esse tipo de traçado.

Figura 12 – Distribuição de acidentes por traçado da via.

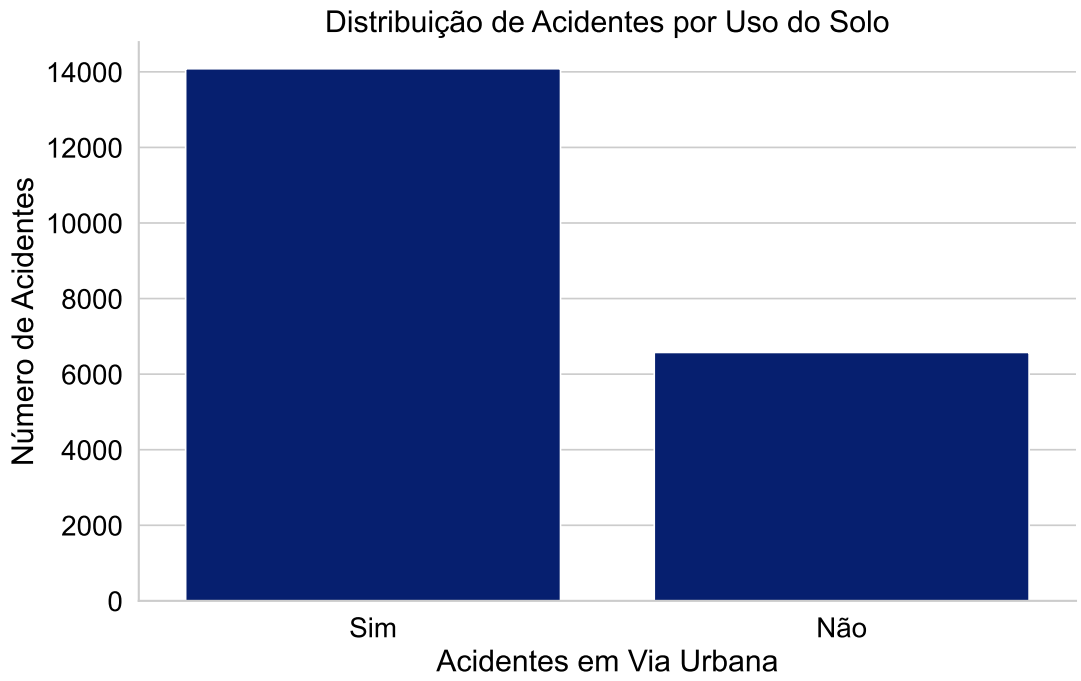


Fonte: Autor.

Uso do solo

O atributo *uso do solo* indica o tipo de ocupação onde ocorreu o acidente, podendo ser classificado como "Sim"(urbano) ou "Não"(rural). A Figura 13 apresenta o número de acidentes em áreas urbanas e rurais. Verifica-se que a maioria dos acidentes aconteceu em zonas urbanas.

Figura 13 – Distribuição dos acidentes por tipo de uso do solo.



Fonte: Autor.

Inconsistência nos dados

Além da exploração mais aprofundada de cada atributo, foi realizada uma investigação mais geral nos dados. Nesta análise, constatou-se que os dados da PRF possuem inconsistências em relação aos atributos de características da via como *tipo de pista*, *traçado da via* e *uso do solo*.

Após a 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 em características da via. A Tabela 4 ilustra tais divergências, considerando todos os registros de acidentes no quilômetro 233,0 da BR-101, sentido decrescente, entre 2017 e 2024. 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. Por esta razão, motivou a correção das informações com outras fontes de dados.

Tabela 4 – Inconsistências em registros de acidentes no km 233.0 da BR-101.

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

Fonte: Autor.

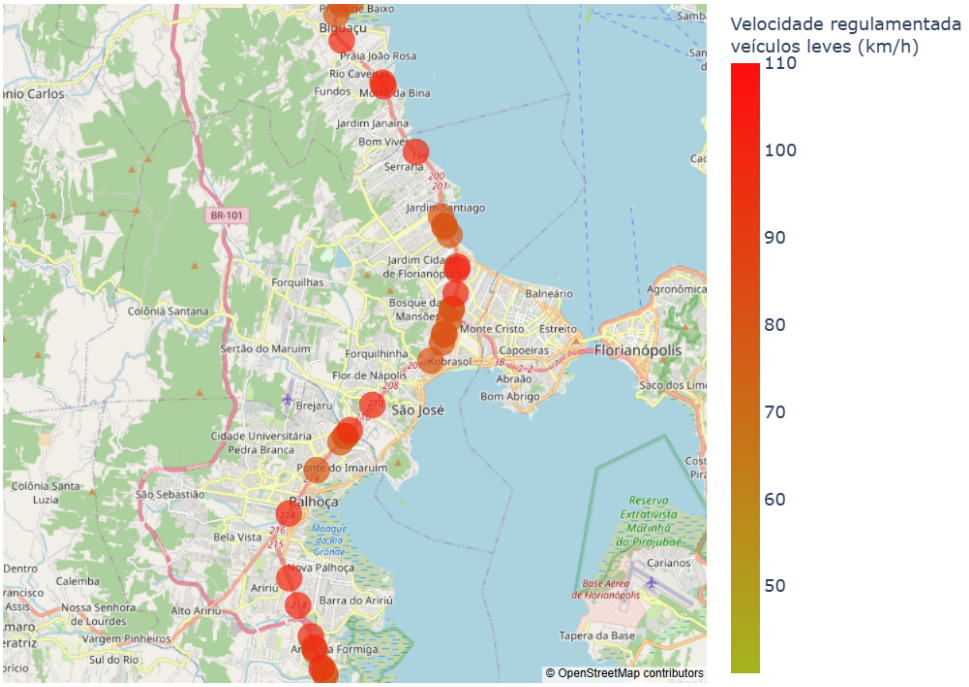
4.2.2 Conjunto de dados sobre vias da ANTT

A ANTT dispõe de quarenta coleções de dados sobre as rodovias brasileiras. Dentre todos os conjuntos disponibilizados pela ANTT, foram selecionados dados sobre valores do *km*, *município*, *número de faixas*, *traçado da via* e *uso do solo* para a correção das inconsistências nos dados da PRF. Além disso, foram escolhidos novos atributos para avaliar se sua inclusão melhora o desempenho dos modelos preditivos, tais como o *tipo de pavimento*, o *tipo de perfil do terreno*, a *velocidade regulamentada para veículos leves*, a *velocidade regulamentada para veículos pesados*, a presença de *pista marginal* e a existência de *iluminação*.

Os dados de cada atributo da ANTT estão em arquivos CSV distintos. Os valores dos atributos de velocidade e quilômetro são delimitados por latitude e longitude. A Figura 14 ilustra os valores da *velocidade regulamentada para veículos leves* na região da Grande Florianópolis, nos quais valores maiores possuem tons mais avermelhados. Por outro lado, os demais atributos são definidos por uma faixa de coordenadas que indicam os limites inicial e final de latitude e longitude. A Figura 15 exibe os valores do *traçado da via* na mesma região, em que áreas curvas estão representadas por caixas azuis, enquanto trechos retos são indicados por caixas laranjas. Observa-se a diferença entre os dois formatos de armazenamento das informações fornecidas pela ANTT. Na Figura 14, os dados são pontos (latitude e longitude); por outro lado, na Figura 15, as informações são apresentadas em áreas com limites inicial e final de latitude e longitude.

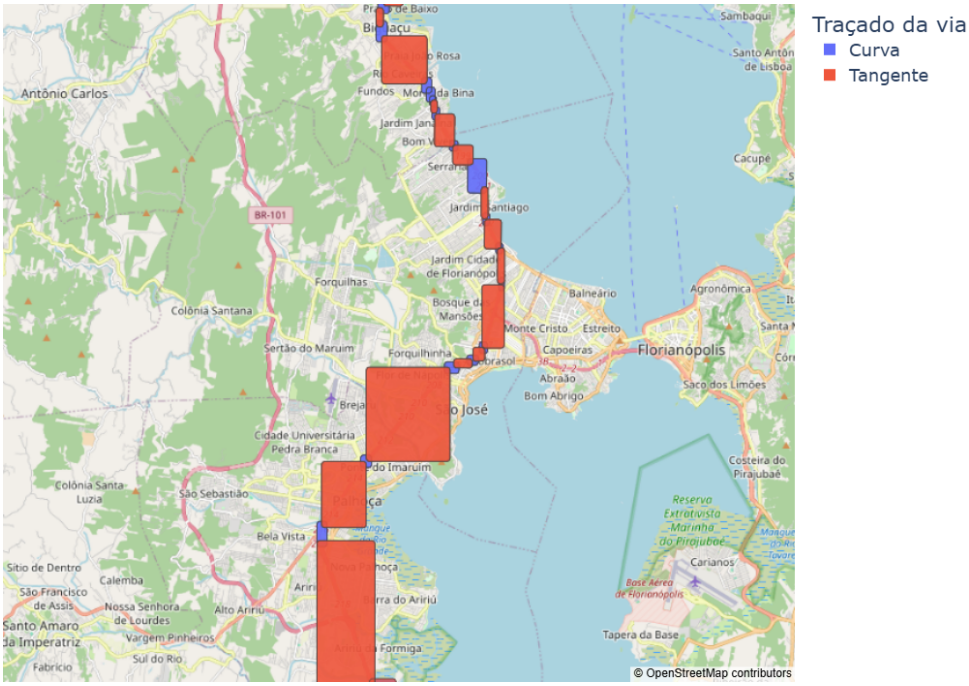
A Tabela 5 descreve os atributos escolhidos e o conjunto de dados de onde cada característica foi retirada. Os conjuntos de dados de pista marginal e iluminação não apresentam um atributo específico que represente diretamente essas características. No lugar disso, estas informações são fornecidas apenas pelas coordenadas que

Figura 14 – Velocidade regulamentada para veículos leves na região da Grande Florianópolis.



Fonte: Autor.

Figura 15 – Classificação do traçado da via na região da Grande Florianópolis.



Fonte: Autor.

indicam o trecho onde há pista marginal ou iluminação.

Tabela 5 – Atributos utilizados dos dados da ANTT.

Atributo	**Descrição**	**Conjunto de dados de origem**
Km	Representação do quilômetro mais a metragem. Ex: 317,940	Quilômetro Pista Principal
Município	Nome do município.	Município
Número de faixas	Quantidade de faixas da via principal.	Pista principal
Traçado da via	Representação do tipo de traçado. Ex: Curva ou Tangente	Traçado
Tipo de uso do solo	Representação do tipo do uso do solo. Ex: Urbano ou Rural.	Uso do Solo
Tipo do pavimento	Representação da ordem pavimento. Ex: Rígido ou Flexível.	Tipo pavimento
Tipo de perfil do terreno	Representação do tipo de perfil do terreno. Ex: Montanhoso, Plano e Ondulado	Perfil do Terreno
Velocidade regulamentada veículos leves	Representação da velocidade máxima permitida. Ex.: 40 km/h	Sinalização
Velocidade regulamentada veículos pesados	Representação da velocidade máxima permitida. Ex.: 40 km/h	Sinalização
Pista marginal	Indica se o local possui pista marginal	Pista marginal
Iluminação	Indica se o local possui iluminação	Iluminação

Fonte: Adaptado dos dicionários de dados da ANTT.

4.2.3 Conjunto de dados sobre vias do DNIT

O DNIT disponibiliza uma plataforma web de dados geográficos chamada VGeo, que contém os principais conjuntos de dados desenvolvidos pelo órgão, incluindo informações sobre rodovias federais (DNIT, 2025b).

A geometria da BR-101 entre os quilômetros 100 e 239 em Santa Catarina foi extraída a partir da base do DNIT. Estes dados estão no formato *MultiLineString*, que representa múltiplas linhas geográficas, com latitude e longitude da via. Estas informações serão fundamentais para integrar os atributos da ANTT em um único conjunto.

4.2.4 Conjunto de dados sobre condições meteorológicas da OPEN-METEO

OPEN-METEO é uma API de código aberto que oferece acesso a dados meteorológicos para qualquer local do mundo. Ele utiliza o conjunto de dados de reanálise IFS do European Centre for Medium-Range Weather Forecasts (ECMWF) para forne-

cer dados históricos sobre o clima a partir do ano de 2017 até atualmente. O conjunto de dados IFS utiliza uma combinação de observações de estações, aeronaves, boias, radares e satélites e modelos matemáticos para estimar informações em locais que não possuem aparelhos de medição. A API dispõe de 30 atributos meteorológicos, com uma resolução espacial de 9 quilômetros e uma resolução temporal em horas (Zippenfenig, 2023).

Dentre os atributos disponíveis na API da OPEN-METEO, foram selecionadas nove também utilizadas em trabalhos relacionados como Elassad et al. (2020), Yu et al. (2021) e Mo et al. (2024). A Tabela 6 descreve os atributos escolhidos, além de mostrar a unidade de medida e o intervalo de tempo de captura. Os atributos foram extraídos da API, utilizando as coordenadas dos quilômetros 100 a 239 dos dados da DNIT, com intervalos de 5 em 5 quilômetros, no período de 2017 a 2024. Os dados foram coletados em intervalos maiores em razão da resolução espacial que a API fornece.

Tabela 6 – Atributos utilizados dos dados da OPEN-METEO.

Atributos	**Descrição**	**Unidade**
Temperatura	Temperatura do ar 2 metros acima do solo.	℃
Temperatura aparente	É a 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.	mm
Umidade relativa do ar	Umidade relativa do ar 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.	m/s
Velocidade de rajada	Velocidade de rajada a 10 metros acima do solo.	m/s
Código climático	Condição meteorológica como um código numérico, seguindo o padrão WMO	Código WMO

Fonte: Adaptado do dicionário de dados da OPEN-METEO.

A Tabela 7 apresenta as estatísticas descritivas desses atributos. A temperatura média registrada foi de 20,9℃, com temperatura aparente média de 22,73℃ e ponto de orvalho médio de 17,81℃. Observa-se uma ampla variação nas temperaturas, com valores mínimos de 3,53℃ para a temperatura, 0,26℃ para a temperatura aparente e

-6,37℃ para o ponto de orvalho, enquanto os máximos atingiram 37,78℃, 45,22℃ e 29,69℃, respectivamente. Já a chuva apresentou média baixa (0,19 mm) e mediana nula, indicando predominância de dias sem chuva, apesar de eventos extremos com até 65,2 mm.

A umidade relativa possui média de 83,39% e variação entre 26,53% e 100%, enquanto a cobertura de nuvens teve média de 65,64% e desvio padrão elevado, refletindo alta variabilidade nas condições atmosféricas. As velocidades médias do vento e das rajadas foram de 9,21m/s e 23,5m/s, respectivamente, com valores máximos bastante altos (57,37m/s e 138,24m/s), evidenciando a ocorrência de ventos intensos em determinados momentos.

Tabela 7 – Estatísticas descritivas dos atributos meteorológicos.

	Temp. (℃)	**Temp. aparente** (℃)	**Ponto de orvalho** (℃)	**Chuva** (mm)	**Umidade** (%)	**C. Nuvens** (%)	**Vel. Vento** (m/s)	**Vel. Rajada** (m/s)
Média	20,90	22,73	17,81	0,19	83,39	65,64	9,21	23,58
Desvio Padrão	4,20	5,88	4,14	0,76	11,78	38,01	5,24	12,28
Mínimo	3,53	0,26	-6,37	0,00	26,53	0,00	0,00	1,08
25%	18,14	18,84	15,44	0,00	75,28	29,00	5,24	14,04
Mediana	21,07	22,81	18,24	0,00	85,77	86,00	8,14	21,60
75%	23,74	26,78	20,84	1,00	92,96	100,00	12,24	30,60
Máximo	37,78	45,22	29,69	65,20	100,00	100,00	57,37	138,24

Fonte: Autor.

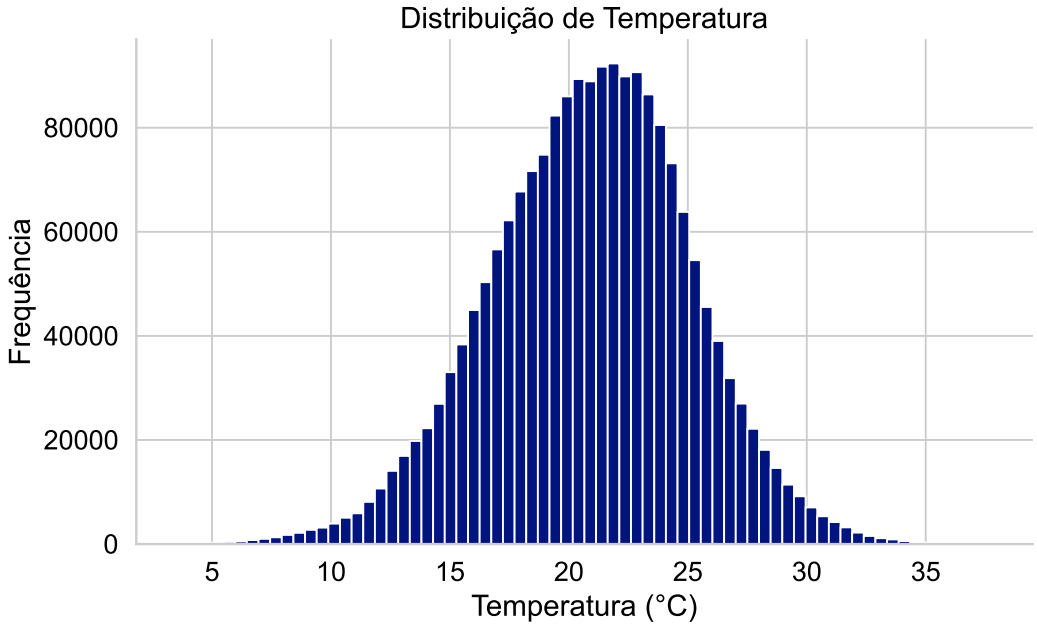
Temperatura

Representa a temperatura ambiente registrada em graus Celsius. Esta medida está relacionada à condição corporal do motorista, que em casos extremos de frio e calor, pode influenciar no desempenho da direção. A Figura 16 apresenta a distribuição desse atributo, indicando que os valores se concentram entre 12℃ e 27℃, com formato de distribuição simétrica.

Temperatura aparente

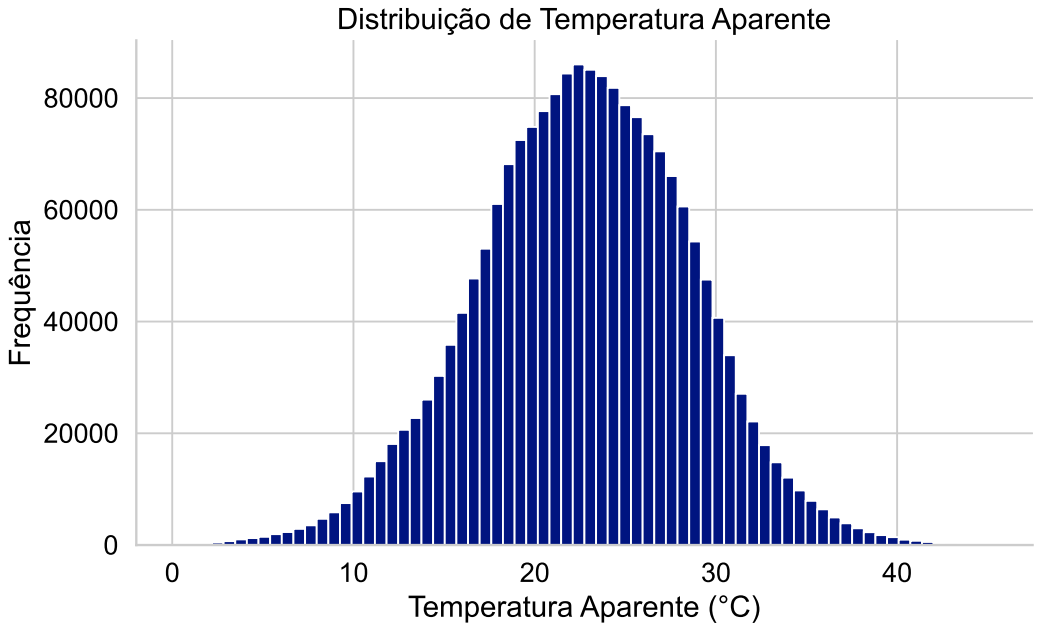
Corresponde à temperatura percebida ao toque, combinando o fator de sensação térmica, a umidade relativa e a radiação solar. A Figura 17 apresenta o histograma desse atributo, cuja distribuição é aproximadamente simétrica, refletindo sua forte correlação com o atributo *temperatura*.

Figura 16 – Histograma da temperatura.



Fonte: Autor.

Figura 17 – Histograma da temperatura aparente.

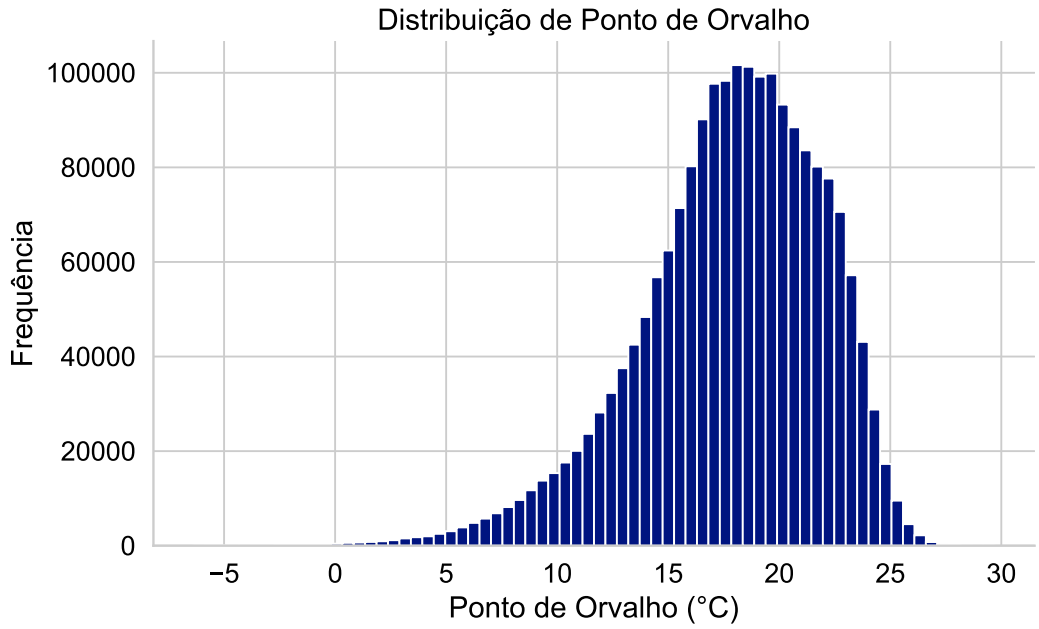


Fonte: Autor.

Ponto de orvalho

Refere-se à temperatura para que o ar deve ser resfriado para a condensação começar. Isto pode indicar condições à formação de neblina ou superfícies molhadas, dificultando a condução. A Figura 18 exibe que a distribuição do ponto de orvalho apresenta assimetria à direita, com a maior parte dos valores concentrados entre 15℃ e 22℃.

Figura 18 – Histograma do ponto de orvalho.



Fonte: Autor.

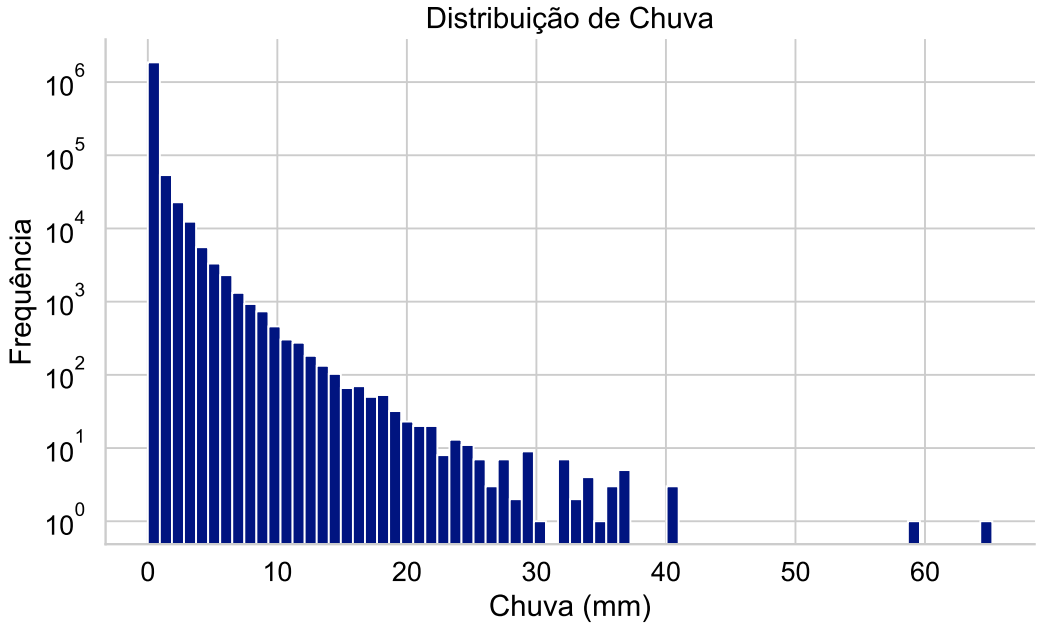
Chuva

Representa a quantidade de precipitação líquida acumulada em milímetros, podendo impactar na aderência dos pneus e na visibilidade da pista. A Figura 19 mostra o histograma desse atributo com escala logarítmica. Percebe-se que não houve precipitação e observações com mais de trinta milímetros são mais raras.

Umidade relativa do ar

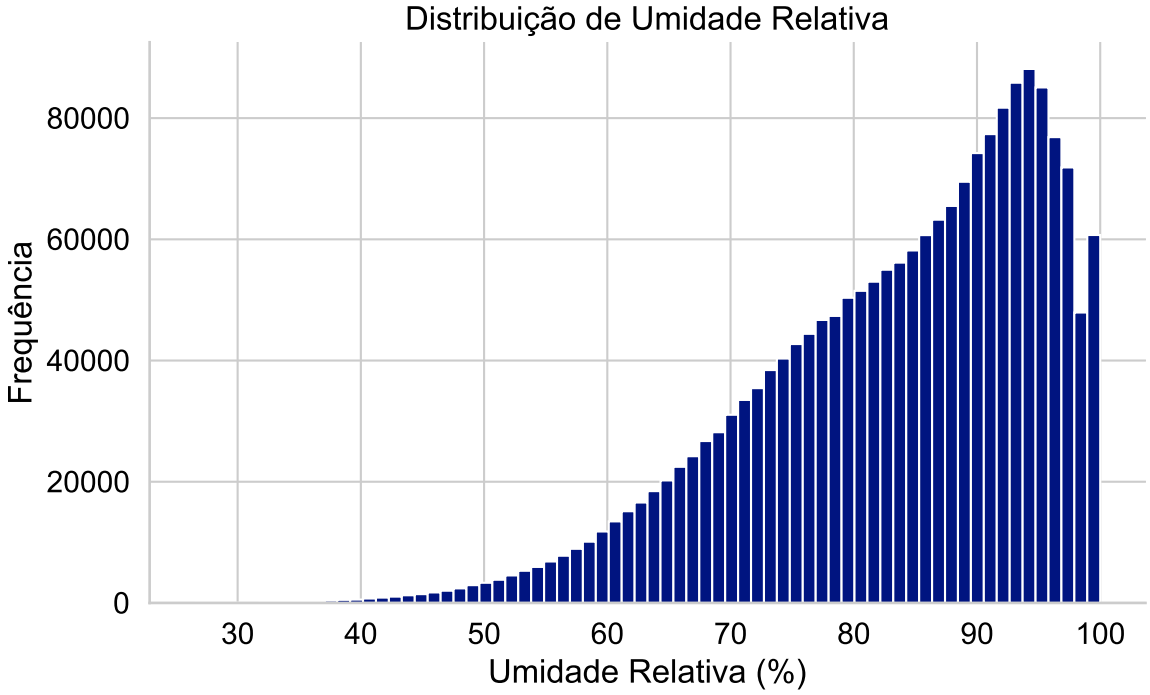
O atributo representa a porcentagem da umidade relativa do ar no momento da observação, o que pode apontar presença de neblina. A Figura 20 apresenta a distribuição deste atributo. Nota-se que a distribuição é assimétrica, com cauda para a esquerda e moda próxima a 93%, o que indica um local com alta umidade devido a ser uma região litorânea.

Figura 19 – Histograma da chuva.



Fonte: Autor.

Figura 20 – Histograma da umidade relativa do ar.

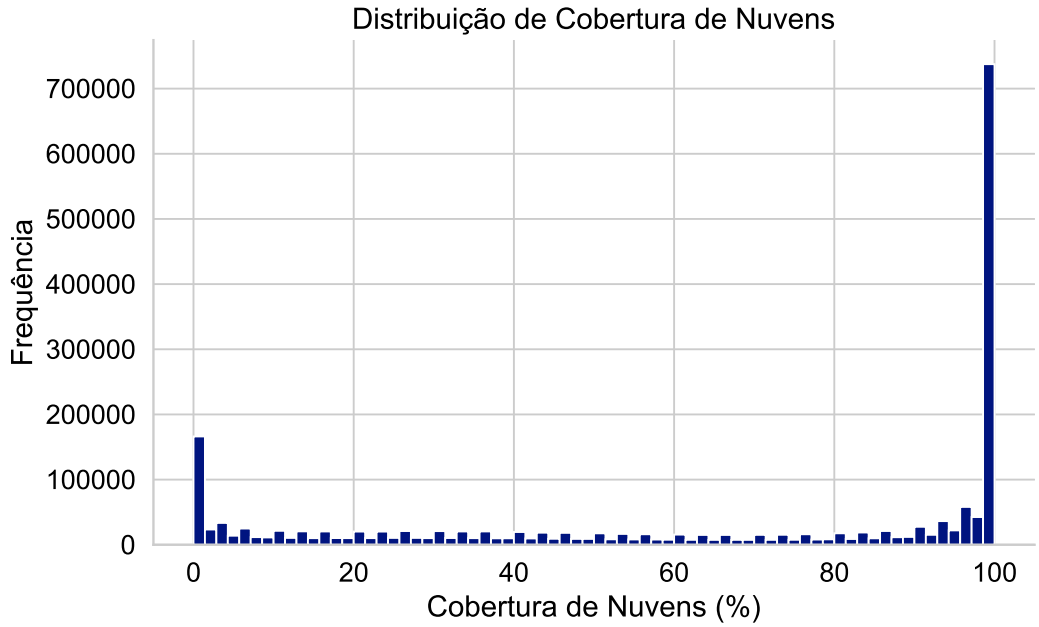


Fonte: Autor.

Cobertura das nuvens

O atributo indica o percentual de cobertura do céu por nuvens, o que reduz a incidência de luz solar e pode comprometer a visibilidade. A Figura 21 apresenta sua distribuição. O gráfico mostra uma distribuição com picos nos valores extremos, indicando que, na maior parte do tempo, o céu estava completamente limpo ou totalmente nublado, com menos ocorrências de coberturas parciais.

Figura 21 – Histograma da cobertura das nuvens.



Fonte: Autor.

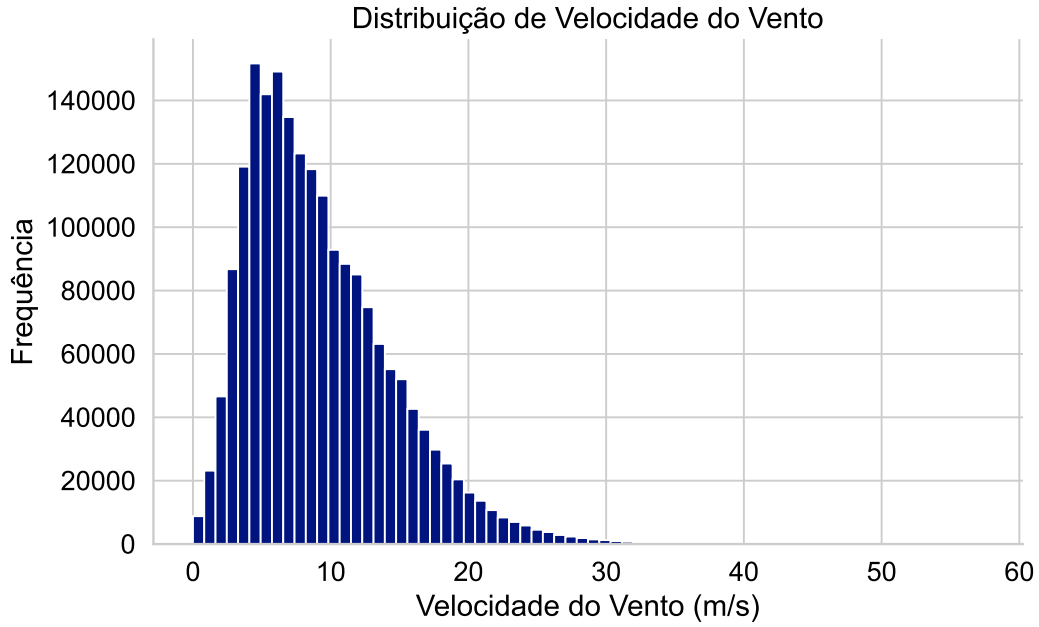
Velocidade do vento

Representa a velocidade do vento registrada em metros por segundo. Ventos fortes podem desestabilizar veículos, dificultando sua direção. A Figura 22 exibe a distribuição do atributo, na qual segue uma distribuição assimétrica com cauda para a direita. Isto assinala que, na maioria dos registros, a velocidade do vento possuía um valor ameno próximo a 8 m/s, com valores maiores menos recorrentes.

Velocidade de rajada

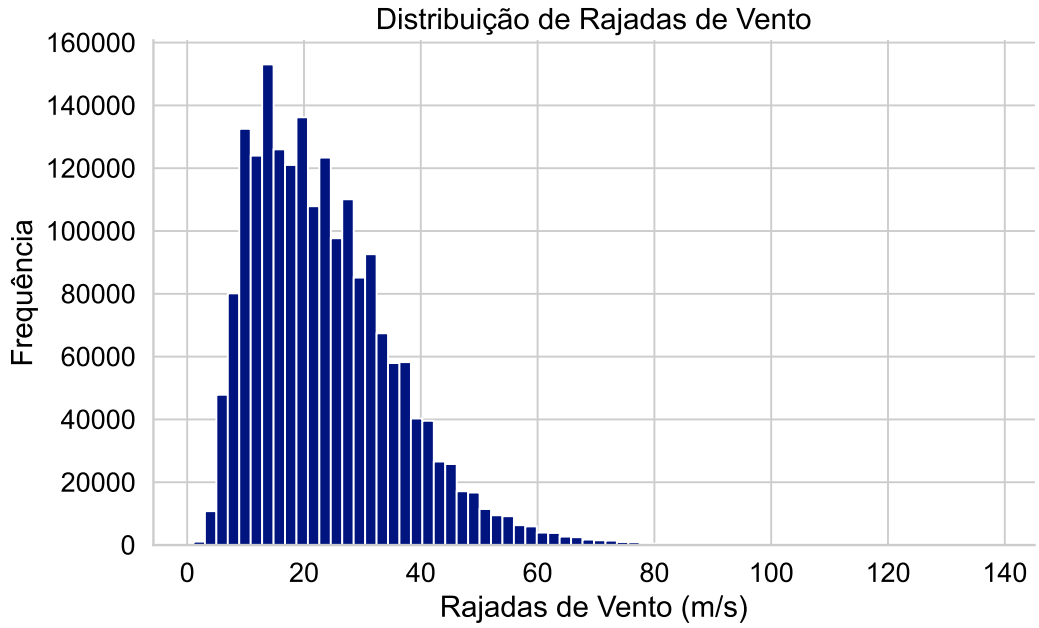
O atributo *Velocidade de rajada* é definida como a velocidade máxima do vento da hora anterior registrada em metros por segundo, impactando na condução de veículos, igualmente seu atributo influente. A distribuição do atributo pode ser observada na Figura 22.

Figura 22 – Histograma da velocidade do vento.



Fonte: Autor.

Figura 23 – Histograma da velocidade de rajada.



Fonte: Autor.

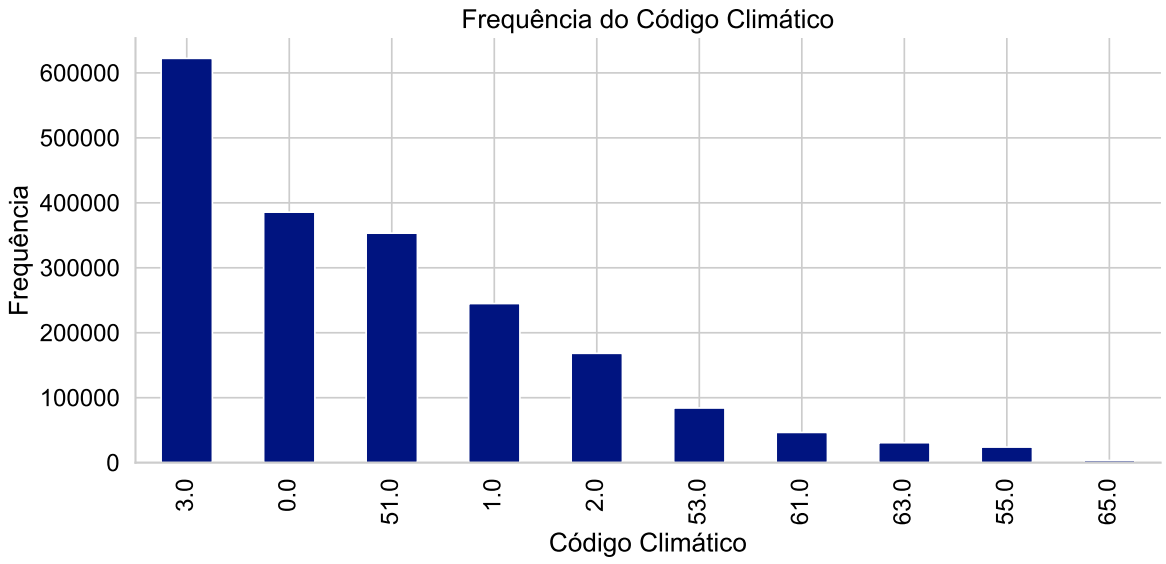
Código climático

Caracteriza-se ao código climático da World Meteorological Organization (WMO):

- 0.0 – Céu limpo
- 1.0 – Principalmente limpo
- 2.0 – Parcialmente nublado
- 3.0 – Nublado
- 51.0 – Chuvisco leve
- 53.0 – Chuvisco moderado
- 55.0 – Chuvisco intenso
- 61.0 – Chuva fraca
- 63.0 – Chuva moderada
- 65.0 – Chuva forte

A Figura 24 exibe a distribuição do atributo. A figura mostra que a maior parte das observações ocorreu sob tempo nublado, seguido por céu limpo e chuvisco leve, o que condiz com os atributos *cobertura da nuvens* e *chuva*.

Figura 24 – Distribuição do código climático.



Fonte: Autor.

4.2.5 Conjunto de dados sobre tráfego do DNIT

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, 2025a). 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, no qual seria necessário aplicar 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, 2025a). Embora o VMDA apresente granularidade diária, 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 por possibilitar a análise do fluxo em diferentes trechos da rodovia.

Os dados do VMDA foram obtidos por meio da plataforma oficial do *DNIT*, onde 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. Os valores de VMDA são apresentados em intervalos de quilômetro, por exemplo, do km 111,3 até o km 117,3. Tabela 8 descreve os atributos utilizados no treinamento dos modelos.

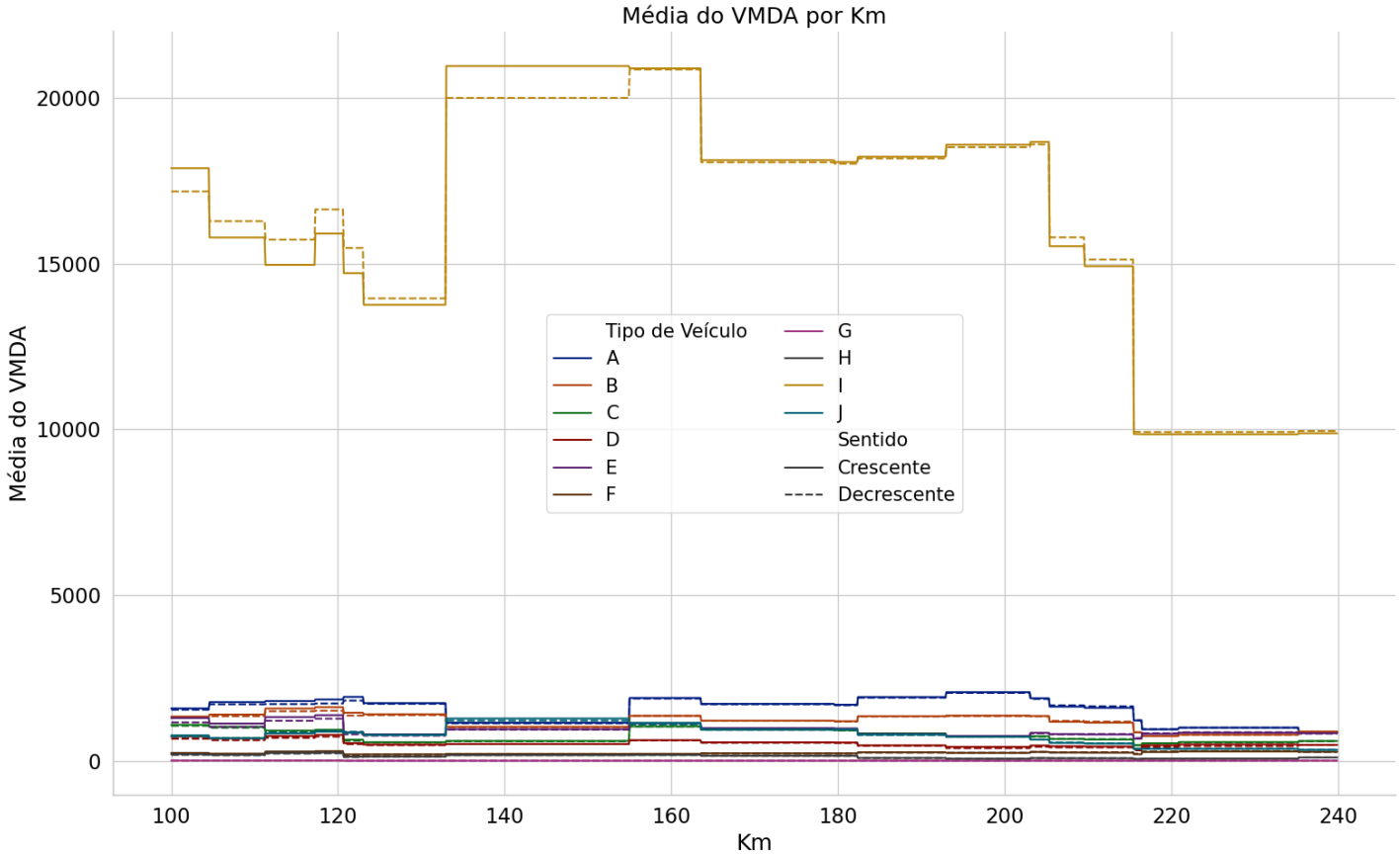
Tabela 8 – Atributos utilizados dos dados do VMDA (DNIT).

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 da DNIT.

O gráfico apresentado na Figura 25 exibe a média do VMDA ao longo dos quilômetros da rodovia por sentido da via (crescente e decrescente) e por categoria de veículo. Os dados representam a média de todos os anos disponíveis. Observa-se que a categoria I apresenta um volume significativamente superior em comparação às demais. Além disso, o gráfico mostra variações no volume de tráfego entre os diferentes quilômetros da rodovia, o que pode ser correlacionado com a ocorrência de acidentes.

Figura 25 – Distribuição VMDA por km.



Fonte: Autor.

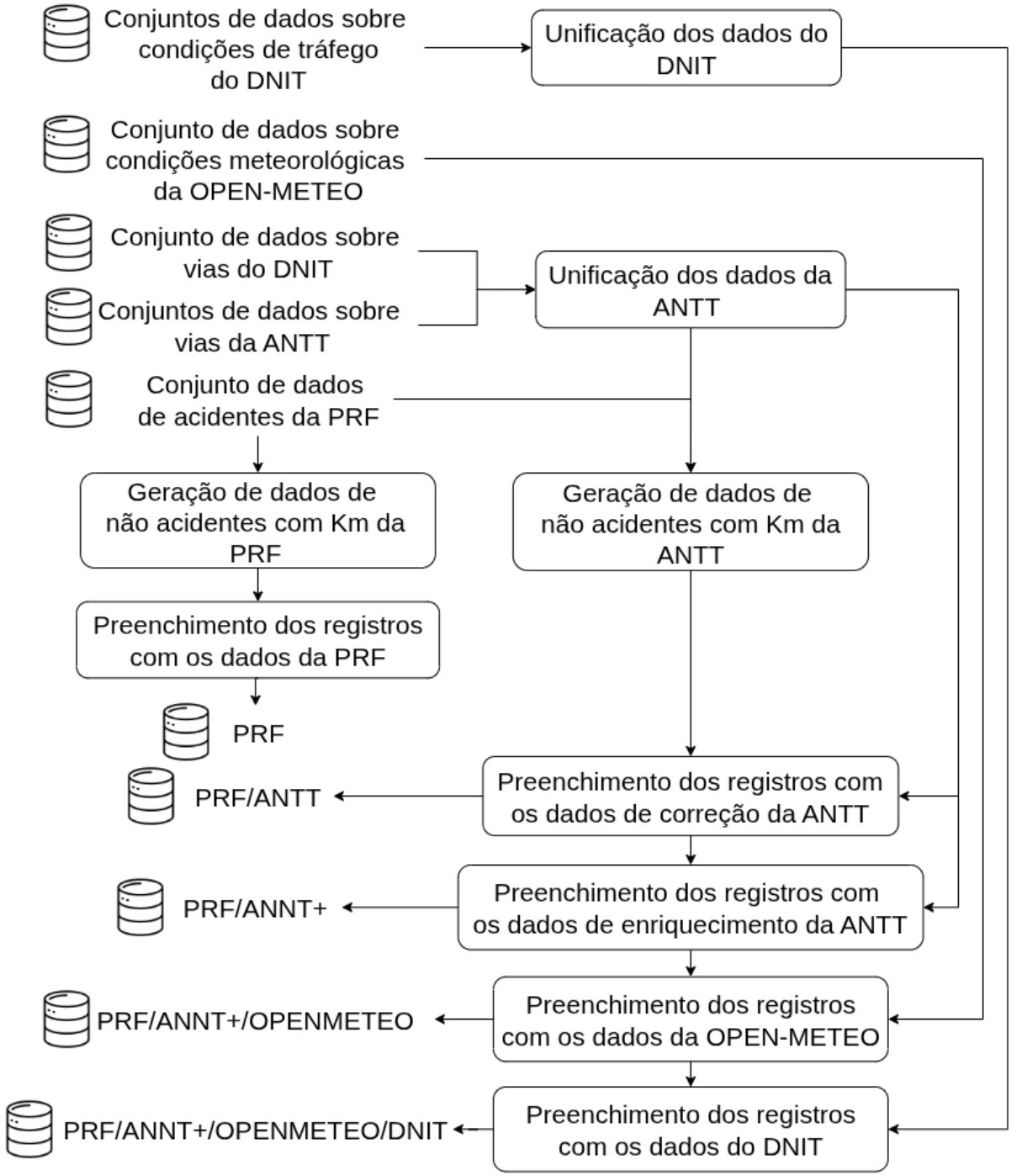
4.3 PREPARAÇÃO DOS DADOS

Nesta seção, serão apresentadas as etapas de processamento dos dados para a criação dos cinco conjuntos de dados para o treinamento e teste dos modelos preditivos. A Figura 26 mostra o fluxo geral dos dados, onde os cinco conjuntos de dados mencionados na seção anterior 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 empregam os dados de enriquecimento de maneira sequencial, ou seja, cada novo conjunto possui os dados do conjunto anterior, mais novas informações. Será explicada cada atividade desse diagrama, com a finalidade de demonstrar de forma clara como foi gerado cada novo conjunto.

4.3.1 Geração de dados de não acidentes com Km da PRF

A Figura 27 exibe o processo de geração de dados de não acidentes até a separação entre dados de treino e teste. Para treinar os modelos preditivos, foi necessário gerar amostras de não-acidentes, uma vez que os dados disponibilizados pela PRF

Figura 26 – Fluxo de dados.

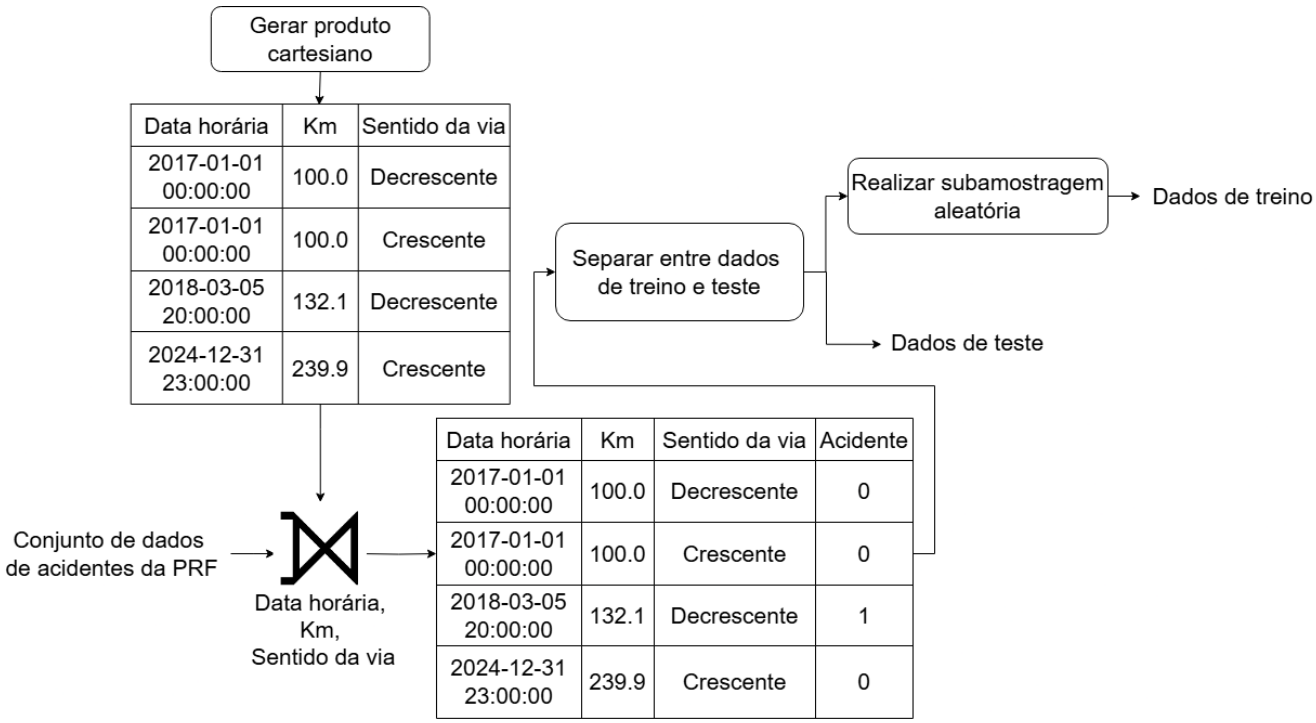


Fonte: Autor.

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 o intuito de obter todas as combinações espaço-temporais possíveis ao longo do trecho analisado. Posteriormente, foram utilizados os dados da PRF para verificar quais dessas combinações registraram acidentes, através de um *left-join* dos campos *data horária* (uma combinação do campo *data* e *hora* da PRF), *km* e *sentido da via* (mesmos campos da PRF). Ao final, obteve-se 194.955.840 registros com uma proporção de 99,99% de não acidentes e 0,01% de acidentes.

Os registros foram divididos entre dados de treino e teste. Nos dados de treinamento, foi aplicada a técnica de subamostragem aleatória nos dados de não acidentes, a fim de equilibrar a proporção de acidentes e não acidentes em 50%. Por outro lado, os dados de teste se mantiveram com a proporção preservada. 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 27 – Geração de dados de não acidentes com Km da PRF.



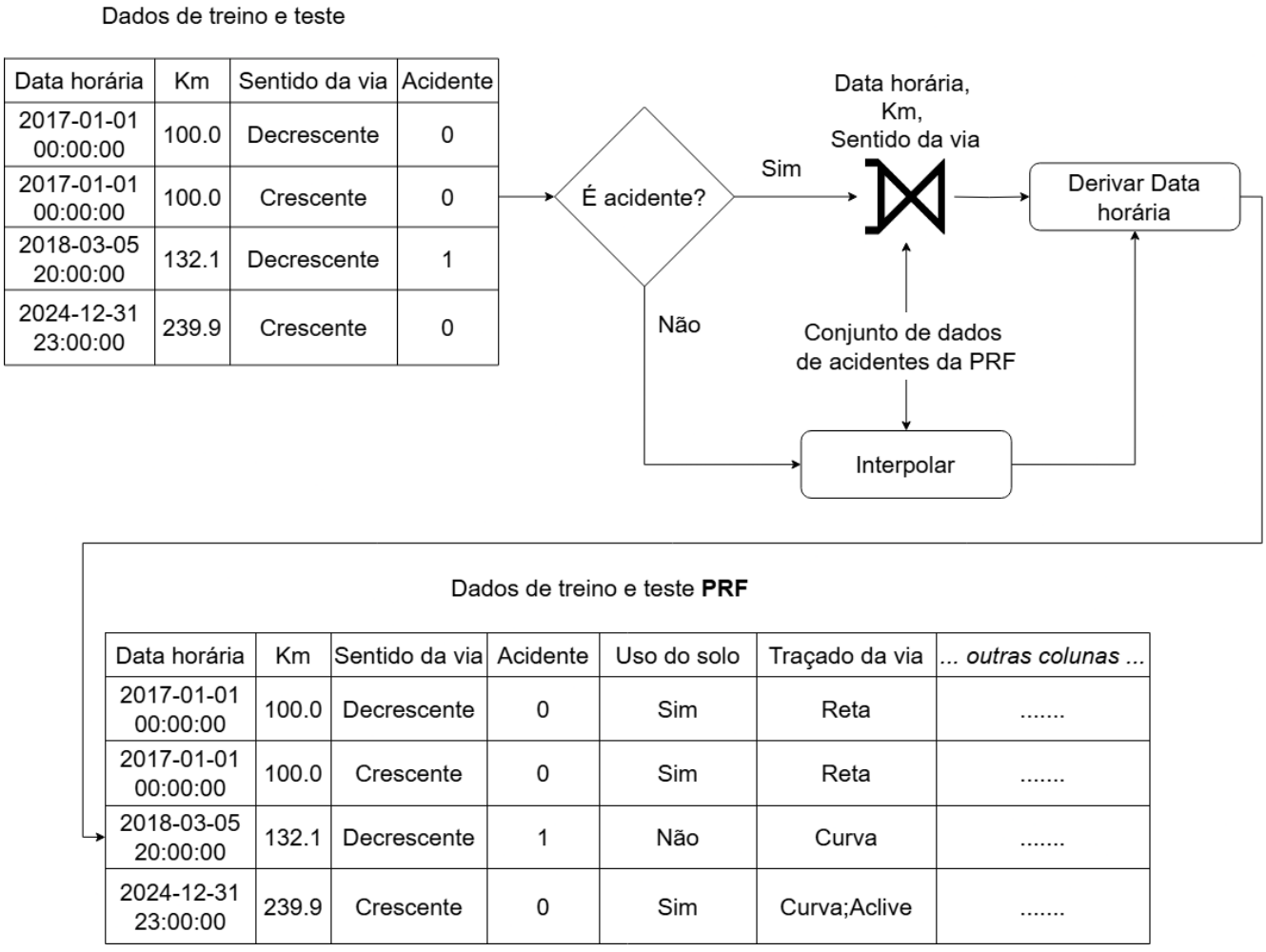
Fonte: Autor.

4.3.2 Preenchimento dos registros com os dados da PRF

A Figura 28 mostra o fluxo para completar os registros com os dados da PRF. Para preencher 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*. Em contrapartida, foi realizada uma interpolação para os dados de não acidentes, extraindo 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 *data horária*. Assim, foi criado o primeiro conjunto de dados para treinamento dos modelos denominado **PRF**.

Figura 28 – Preenchimento dos registros com os dados da PRF.



Fonte: Autor.

4.3.3 Unificação dos dados da ANTT

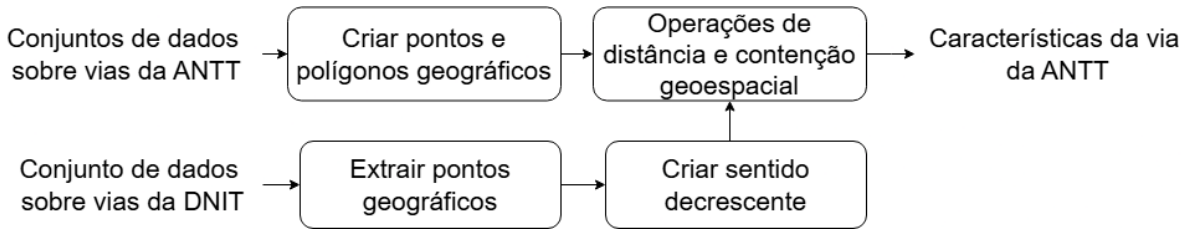
A Figura 29 apresenta o diagrama desta atividade. A partir dos dados do DNIT, foram extraídos pontos que correspondem a latitude e longitude de cada trecho de 100 metros da via. No entanto, os dados originais possuem apenas o segmento do sentido crescente da via; portanto, foi necessário criar a representação do sentido decrescente.

Para isso, os pontos do sentido crescente foram duplicados e tiveram suas posições geográficas ajustadas.

Para os conjuntos de dados da ANTT que possuem as características de velocidades e quilômetro, foram criados pontos geográficos, pois estes apresentam apenas um valor de latitude e longitude. Em relação aos outros dados, foi gerado um atributo do tipo polígono geográfico para representar a latitude e longitude final e inicial das categorias.

Para obter as propriedades das vias da ANTT nos pontos do trajeto da via do DNIT, foram realizadas operações de proximidade baseadas na distância euclidiana, a fim de identificar o ponto ou polígono de característica mais próximo de cada ponto da rodovia. Desta forma, cada ponto geográfico (representação de 100 metros) da área de interesse da BR-101 passou a possuir apenas um único valor possível para cada característica da via nos novos dados, eliminando completamente as inconsistências nos dados da PRF explicadas na Seção 4.2.1. Este conjunto de dados unificados da ANTT foi denominado **Características da via da ANTT**.

Figura 29 – Unificação dos dados da ANTT.



Fonte: Autor.

4.3.4 Geração de dados de não acidentes com Km da ANTT

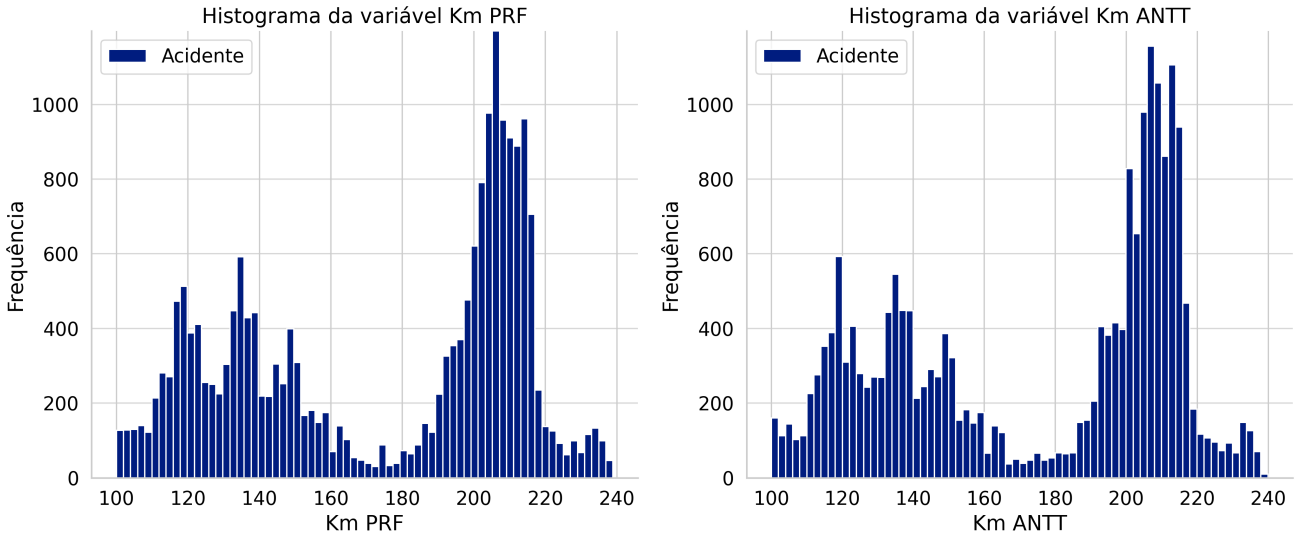
Até o ano de 2016, o atributo *km* nos registros de acidentes da PRF era anotado com granularidade em quilômetros inteiros, ou seja, se o acidente ocorreu no quilômetro 132,5, seu registro era truncado para o quilômetro 132. A partir de 2017, a PRF adotou a precisão mínima de 0,1 quilômetros e o uso de coordenadas geográficas para identificação do local do acidente.

Esta mudança motivou uma análise mais aprofundada sobre o atributo, com o intuito de verificar se o padrão de registro truncado continuou, de forma implícita, nos dados de acidentes depois de 2017. Para isto, foram realizadas operações de proximidade com base na distância euclidiana, entre as coordenadas dos registros de acidentes da PRF com as coordenadas do conjunto de dados da ANTT, com a finalidade de identificar o valor do *km* nos dados da ANTT para cada acidente.

Dos 20.656 registros de acidentes da PRF, 16.027 (77,59%) possuíam valores distintos entre os dados da PRF e da ANTT. No entanto, em 14.071 (87,73%) destas observações, a diferença observada foi de no máximo um quilômetro. A Figura 30

mostra que, embora a maioria dos valores seja discordante, a distribuição geral do atributo nos dados da PRF e ANTT apresenta grande similaridade.

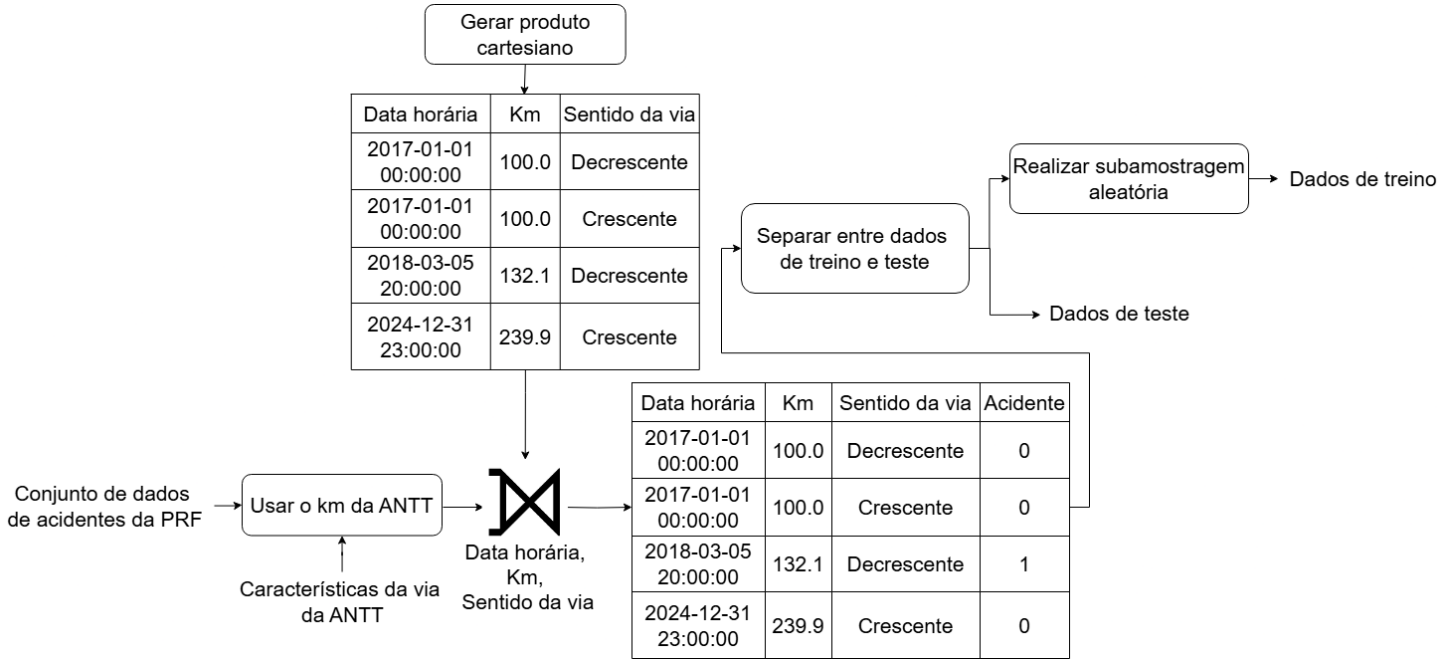
Figura 30 – Histograma de acidentes por Km da PRF e ANTT.



Fonte: Autor.

Com base nestes resultados, decidiu-se por criar uma nova versão na geração de dados de não acidentes, substituindo o *km* da PRF pelo *km* da ANTT. A Figura 31 mostra a atividade de geração de dados de não acidentes com *km* da ANTT. Nota-se que ela apresenta todas as etapas que a atividade de 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).

Figura 31 – Geração de dados de não acidentes com Km da ANTT.

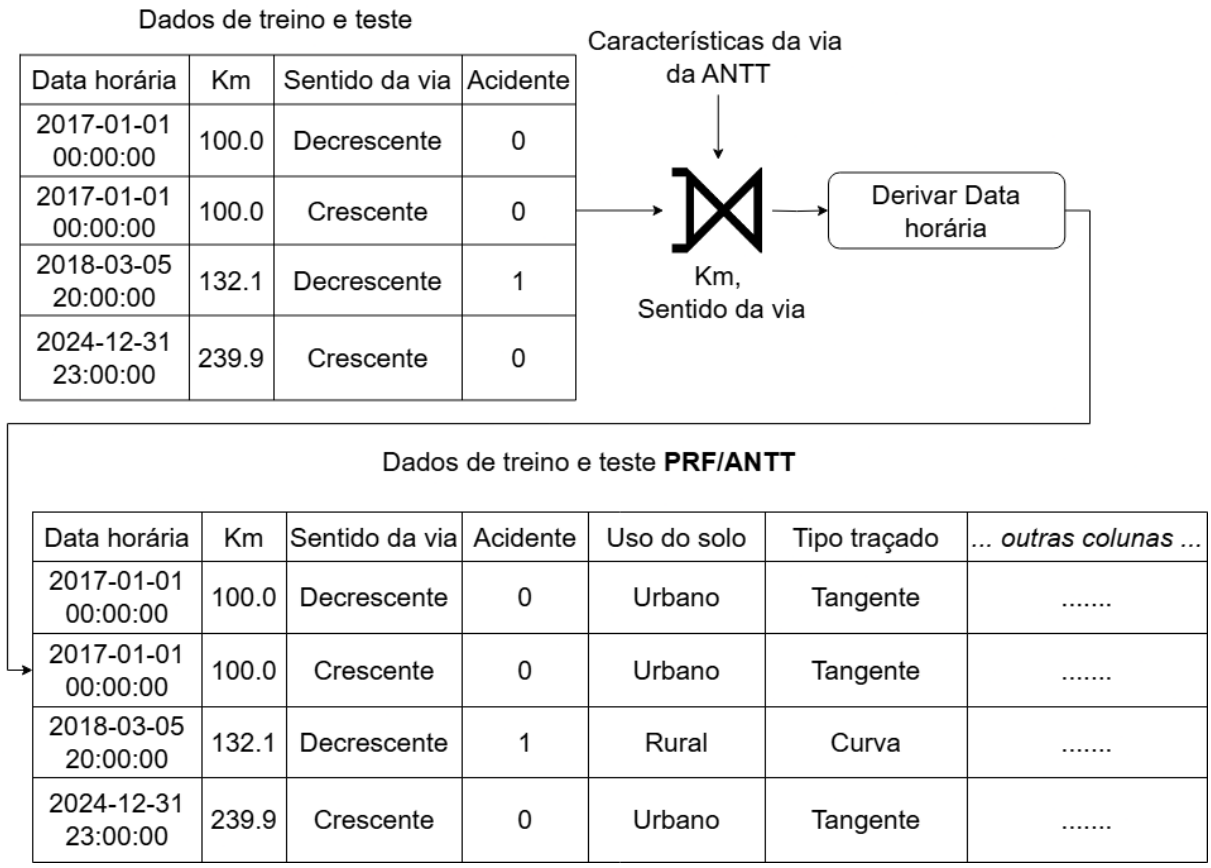


Fonte: Autor.

4.3.5 Preenchimento dos registros com os dados de correção da ANTT

A Figura 32 exibe a atividade de preencher os registros com os dados de correção da ANTT. Para completar os registros com os atributos de correção, foi realizado um *left-join* entre os registros gerados na etapa anterior com o conjunto **Características da via da ANTT**, usando os atributos *km* e *sentido da via* como chave de junção para incluir os atributos de *uso do solo*, *tipo traçado*, *número de faixas* e *município* para cada registro. O atributo de *número de faixas* da ANTT foi utilizada para substituir o atributo *tipo de pista* da PRF, enquanto os demais atributos possuem o mesmo nome em ambos os conjuntos e foram substituídas uma pela outra. Além disso, foi realizada a criação dos atributos que representam *hora*, *dia*, *dia da semana*, *mês*, *ano* e *feriado* a partir do atributo *data Horária*. Desta forma, foi criado o segundo conjunto de dados para treinamento chamado **PRF**/**ANTT**, que possui o mesmo grupo de atributos que o conjunto **PRF**, porém com a correção das inconsistências mencionadas no item 4.2.1 da PRF, utilizando os dados da ANTT.

Figura 32 – Preenchimento dos registros com os dados de correção da ANTT.



Fonte: Autor.

4.3.6 Preenchimento dos registros com os dados de enriquecimento da ANTT

A Figura 33 exibe a atividade de preencher os registros com os dados de enriquecimento da ANTT. Esta etapa recebe como entrada o conjunto **PRF**/**ANTT** gerado na etapa anterior 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, foi produzido o terceiro conjunto de treinamento apelidado de **PRF**/**ANTT+**, que possui todo o grupo de atributos fornecidos pela ANTT.

Figura 33 – Preenchimento dos registros com os dados de enriquecimento da ANTT.

Data horária	Km	Sentido da via	Acidente	Uso do solo	Tipo traçado	*... outras colunas ...*
2017-01-01 00:00:00	100.0	Decrescente	0	Urbano	Tangente	.....
2017-01-01 00:00:00	100.0	Crescente	0	Urbano	Tangente	.....
2018-03-05 20:00:00	132.1	Decrescente	1	Rural	Curva	.....
2024-12-31 23:00:00	239.9	Crescente	0	Urbano	Tangente	.....

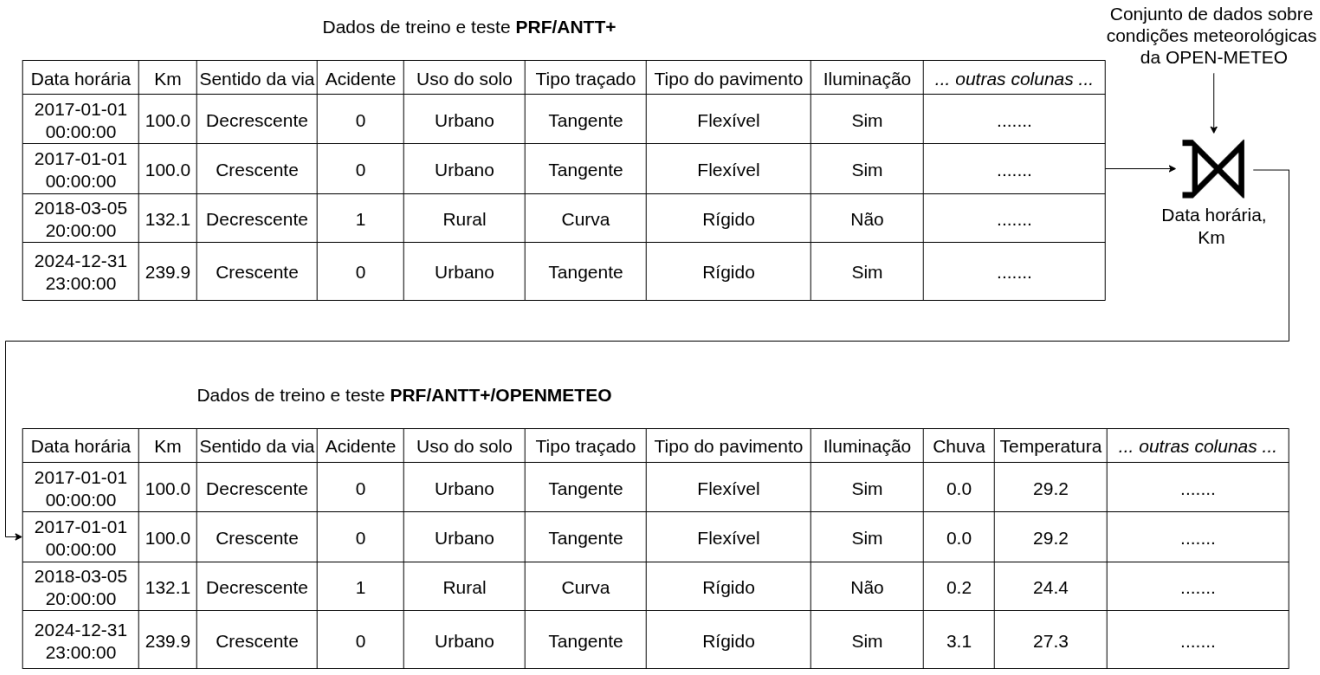
Data horária	Km	Sentido da via	Acidente	Uso do solo	Tipo traçado	Tipo do pavimento	Iluminação	*... outras colunas ...*
2017-01-01 00:00:00	100.0	Decrescente	0	Urbano	Tangente	Flexível	Sim	.....
2017-01-01 00:00:00	100.0	Crescente	0	Urbano	Tangente	Flexível	Sim	.....
2018-03-05 20:00:00	132.1	Decrescente	1	Rural	Curva	Rígido	Não	.....
2024-12-31 23:00:00	239.9	Crescente	0	Urbano	Tangente	Rígido	Sim	.....

Fonte: Autor.

4.3.7 Preenchimento dos registros com os dados da OPEN-METEO

A Figura 34 apresenta a etapa de preencher os registros com os dados da OPEN-METEO. Nesta atividade foi realizado um *left-join* diferente das fases passadas, no qual foi utilizado o atributo *data horária* exato de cada registro, associando-a ao ponto de coleta meteorológica mais próximo, com tolerância de 5 quilômetros. Deste modo, foi produzido o conjunto de treinamento denominado **PRF**/**ANTT+**/**OPENMETEO**, o qual integra dados meteorológicos aos registros.

Figura 34 – Preenchimento dos registros com os dados da OPEN-METEO.



Data horária	Km	Sentido da via	Acidente	Uso do solo	Tipo traçado	Tipo do pavimento	Iluminação	*... outras colunas ...*
2017-01-01 00:00:00	100.0	Decrescente	0	Urbano	Tangente	Flexível	Sim	.....
2017-01-01 00:00:00	100.0	Crescente	0	Urbano	Tangente	Flexível	Sim	.....
2018-03-05 20:00:00	132.1	Decrescente	1	Rural	Curva	Rígido	Não	.....
2024-12-31 23:00:00	239.9	Crescente	0	Urbano	Tangente	Rígido	Sim	.....

Data horária	Km	Sentido da via	Acidente	Uso do solo	Tipo traçado	Tipo do pavimento	Iluminação	Chuva	Temperatura	*... outras colunas ...*
2017-01-01 00:00:00	100.0	Decrescente	0	Urbano	Tangente	Flexível	Sim	0.0	29.2	.....
2017-01-01 00:00:00	100.0	Crescente	0	Urbano	Tangente	Flexível	Sim	0.0	29.2	.....
2018-03-05 20:00:00	132.1	Decrescente	1	Rural	Curva	Rígido	Não	0.2	24.4	.....
2024-12-31 23:00:00	239.9	Crescente	0	Urbano	Tangente	Rígido	Sim	3.1	27.3	.....

Fonte: Autor.

4.3.8 Unificação dos dados do DNIT

A Figura 35 mostra a unificação dos dados do DNIT. Esta etapa 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. Este conjunto de dados unificado foi denominado **VMDA do DNIT**.

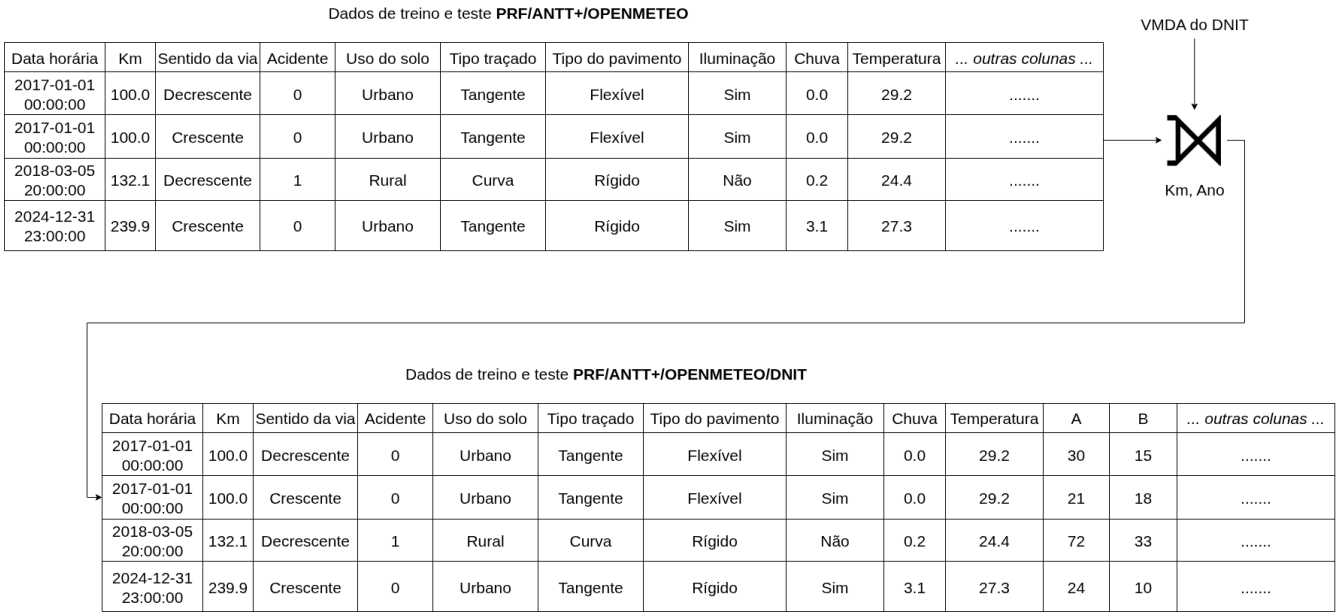
Figura 35 – Unificação dos dados do DNIT.

Fonte: Autor.

4.3.9 Preenchimento dos registros com os dados do DNIT

Por fim, no preenchimento dos registros com dados do DNIT, apresentado na Figura 36, foi realizado um *left-join* entre os registros e o conjunto **VMDA do DNIT**, utilizando os atributos *km* e *ano* como chaves de junção. 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.

Figura 36 – Preenchimento dos registros com os dados do DNIT.



Data horária	Km	Sentido da via	Acidente	Uso do solo	Tipo traçado	Tipo do pavimento	Iluminação	Chuva	Temperatura	*... outras colunas ...*
2017-01-01 00:00:00	100.0	Decrescente	0	Urbano	Tangente	Flexível	Sim	0.0	29.2	.....
2017-01-01 00:00:00	100.0	Crescente	0	Urbano	Tangente	Flexível	Sim	0.0	29.2	.....
2018-03-05 20:00:00	132.1	Decrescente	1	Rural	Curva	Rígido	Não	0.2	24.4	.....
2024-12-31 23:00:00	239.9	Crescente	0	Urbano	Tangente	Rígido	Sim	3.1	27.3	.....

Data horária	Km	Sentido da via	Acidente	Uso do solo	Tipo traçado	Tipo do pavimento	Iluminação	Chuva	Temperatura	A	B	*... outras colunas ...*
2017-01-01 00:00:00	100.0	Decrescente	0	Urbano	Tangente	Flexível	Sim	0.0	29.2	30	15	.....
2017-01-01 00:00:00	100.0	Crescente	0	Urbano	Tangente	Flexível	Sim	0.0	29.2	21	18	.....
2018-03-05 20:00:00	132.1	Decrescente	1	Rural	Curva	Rígido	Não	0.2	24.4	72	33	.....
2024-12-31 23:00:00	239.9	Crescente	0	Urbano	Tangente	Rígido	Sim	3.1	27.3	24	10	.....

Fonte: Autor.

4.4 MODELAGEM

Os atributos dos conjuntos de dados da PRF, ANTT e do *código climático* da OPEN-METEO foram convertidos em categorias utilizando a codificação *one-hot*, conforme adotado em Mo et al. (2024). Além disso, os atributos foram padronizados, também seguindo Mo et al. (2024). Foram treinados três distintos modelos de aprendizado de máquina: Floresta Aleatória, conforme utilizado em Tran et al. (2023), Huang et al. (2020) e Peng et al. (2020), Máquina de Vetores de Suporte, conforme empregado em Cai et al. (2020), Yu et al. (2021), Tran et al. (2023) e Huang et al. (2020), e Perceptron Multicamadas, aplicados em Cai et al. (2020), Huang et al. (2020) e Peng et al. (2020), para cada um dos cinco diferentes conjuntos de dados: PRF, PRF/ANTT, PRF/ANTT+ e PRF/ANTT+/OPEN-METEO, PRF/ANTT+/OPENMETEO/DNIT, totalizando quinze modelos ao todo. Estes modelos foram escolhidos devido ao seu bom desempenho para dados tabulares. Optou-se por não utilizar os modelos propostos em trabalhos relacionados, como os baseados em grafos DSTGCN de Yu et al. (2021) e MSGNN de Tran et al. (2023), pois estes exploram interseções de vias em ambientes urbanos, o que não se aplica à geometria linear da rodovia analisada. Também decidiu-se por não utilizar o modelo LSTM-CNN de Mo et al. (2024), uma vez que sua estrutura é baseada para séries temporais, enquanto este estudo não adota esse tipo de abordagem.

Os modelos foram treinados utilizando validação cruzada 5-Fold, utilizado em Elassad et al. (2020), o que permitiu uma seleção otimizada dos hiperparâmetros. A Tabela 9 apresenta todos os hiperparâmetros e valores testados, enquanto a Tabela 10 mostra os melhores valores ajustados para cada hiperparâmetro de cada modelo.

Tabela 9 – Hiperparâmetros avaliados para os modelos RF, SVM e MLP.

Modelo	**Hiperparâmetro**	**Valores avaliados**
RF	número de árvores	[100, 200, 500]
	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]
	taxa de aprendizado	[0,01, 0,001, 0,0001]
	otimizador	[Adam, SGD]
	taxa de dropout	[0, 0,1]
	regularização L2	[0, 0,0001]

Fonte: Autor.

Tabela 10 – Melhores hiperparâmetros para cada modelo e conjunto de dados.

Modelo	**Hiperparâmetro**	**PRF**	**PRF/ANTT**	**PRF/ANTT+**	**PRF/ANTT+/OPENMETEO**	**PRF/ANTT+/OPENMETEO/DNIT**
RF	número de árvores	200	500	500	500	500
	critério de divisão	Gini	Entropia	Entropia	Entropia	Entropia
	profundidade máxima	128	128	128	128	128
	bootstrap	Não	Não	Não	Não	Não
SVM	C	2	2	1	5	5
	kernel	Linear	Linear	Linear	Linear	Linear
	gamma	-	-	-	-	-
MLP	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
	taxa de aprendizado	1e-4	1e-3	1e-4	1e-4	1e-4
	otimizador	Adam	Adam	Adam	Adam	Adam
	taxa de dropout	0	0.1	0.1	0.1	0.1
	regularização L2	1e-4	1e-4	0	0	0

Fonte: Autor.

4.5 AVALIAÇÃO

Para avaliar o desempenho dos modelos, foram calculadas as métricas de especificidade, sensibilidade e AUC, também utilizadas em trabalhos com dados de teste totalmente desbalanceados como Peng et al. (2020), Islam et al. (2021), Mo et al. (2024) e Cai et al. (2020).

5 RESULTADOS

A Tabela 11 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 os 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. Nesse tipo de aplicação, é preferível priorizar a sensibilidade, pois prever um acidente que não acontece é menos grave do que deixar de prever um que de fato aconteça. Ademais, o desempenho do SVM se manteve constante em todas as métricas, já o RF sofreu reduções com a adição dos atributos.

Tabela 11 – Desempenho dos modelos para dados de validação.

Medida	Conjunto de dados	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/ANTT+/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/ANTT	0.8336 ± 0.0043	0.8233 ± 0.0062	**0.8369 ± 0.0067**
	PRF/ANTT+	0.8225 ± 0.0033	0.8250 ± 0.0055	**0.8350 ± 0.0115**
	PRF/ANTT+/OPENMETEO	0.8163 ± 0.0034	**0.8242 ± 0.0054**	0.8194 ± 0.0077
	PRF/ANTT+/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/ANTT	0.8195 ± 0.0047	0.8287 ± 0.0030	**0.8426 ± 0.0030**
	PRF/ANTT+	0.8073 ± 0.0036	0.8286 ± 0.0026	**0.8420 ± 0.0039**
	PRF/ANTT+/OPENMETEO	0.8015 ± 0.0043	0.8284 ± 0.0023	**0.8377 ± 0.0032**
	PRF/ANTT+/OPENMETEO/DNIT	0.7987 ± 0.0029	0.8315 ± 0.0020	**0.8392 ± 0.0025**

Fonte: Autor.

Para verificar a relevância estatística 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 12 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/ANTT 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 do primeiro conjunto ao último conjunto, embora essa diferença não tenha sido testada quanto à significância estatística. Já o modelo RF apresentou pequenas quedas com a adição dos atributos. A Tabela 13 mostra o desempenho dos modelos para os dados de teste. Os resultados obtidos nos dados de teste mostram que todos os modelos mantêm desempenho próximo aos dados de validação, indicando capacidade de generalização.

Tabela 12 – Teste de significância estatística (p-valor).

Medida	**Conjunto de dados**	**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	0.0305	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 x PRF/ANTT+/OPENMETEO/DNIT	0.3131	0.0728	0.4680

Fonte: Autor.

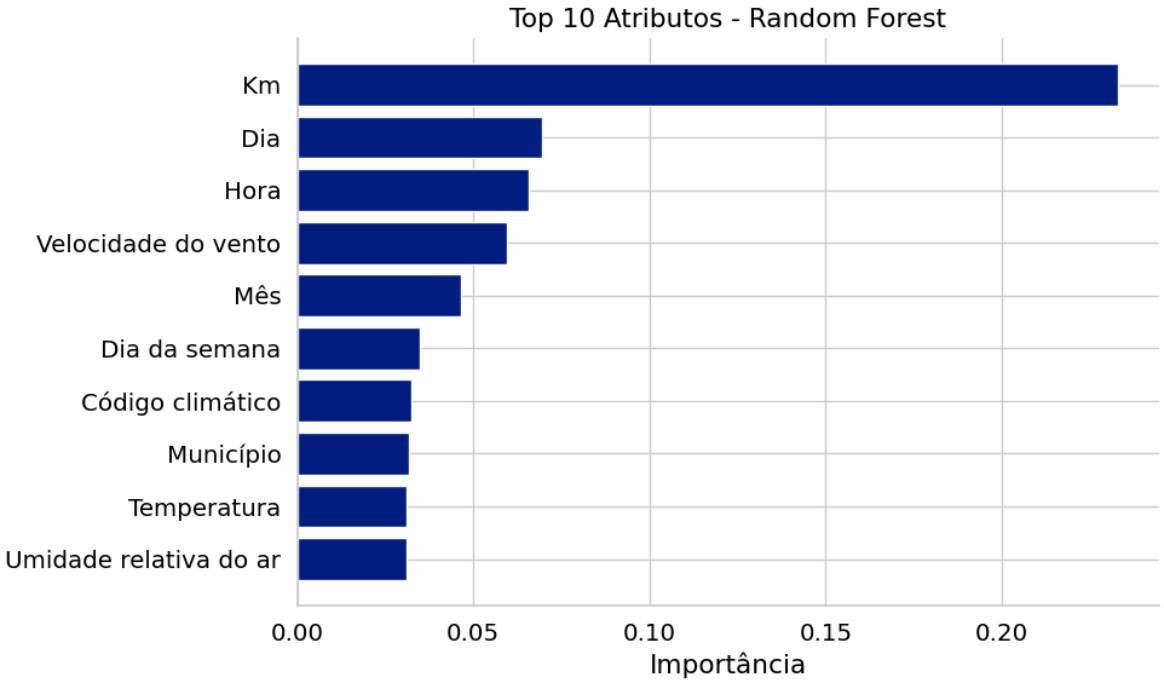
Uma possível hipótese para a ausência de ganho de desempenho com a inclusão de novos atributos é a forte concentração espacial dos acidentes em determinados trechos da rodovia, como entre os quilômetros 190 e 220 e entre os quilômetros 120 e 150, como foi mostrado na Figura 30. Este padrão pode induzir o modelo a focar excessivamente nessas regiões específicas, aumentando a importância do atributo *km* e reduzindo a influência de outros atributos. A Figura 37 apresenta os dez atributos mais relevantes para o modelo RF para o conjunto PRF/ANTT+/OPENMETEO/DNIT. Nota-se que o *km* representa 20% da importância total, enquanto os seguintes atributos situam-se na faixa de 5%.

Tabela 13 – Desempenho dos modelos para dados de teste.

Medida	Conjunto de dados	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+/OPENMETEO	0.8000	**0.8488**	**0.8488**
	PRF/ANTT+/OPENMETEO/DNIT	0.7659	0.8488	**0.8585**
Especificidade	PRF	0.7726	**0.7850**	0.7609
	PRF/ANTT	**0.8544**	0.8295	0.8482
	PRF/ANTT+	**0.8552**	0.8305	0.8412
	PRF/ANTT+/OPENMETEO	**0.8402**	0.8298	0.8286
	PRF/ANTT+/OPENMETEO/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+/OPENMETEO	0.8201	0.8393	**0.9082**
	PRF/ANTT+/OPENMETEO/DNIT	0.7931	0.8397	**0.9085**

Fonte: Autor.

Figura 37 – Top 10 atributos mais relevantes - RF.



Fonte: Autor.

Diante disso, decidiu-se refazer os experimentos com todos os modelos, excluindo o atributo *km* do treinamento, a fim de analisar o impacto real da inclusão dos novos atributos. A Tabela 14 apresenta os desempenhos dos modelos durante a validação cruzada sem o uso do *km*. Observa-se uma redução entre os desempenhos dos modelos em comparação com os valores apresentados anteriormente, evidenciando a importância do *km* para a predição no trecho analisado. No entanto, esse atributo pode não ser tão relevante em outros contextos, visto que sua forte influência pode levar o modelo a supervalorizar regiões com histórico elevado de acidentes, como determinados quilômetros, ignorando outros atributos. É possível excluir o *km* a fim de um uso mais equilibrado dos demais atributos, ainda que com leve perda de desempenho.

Adicionalmente, além do ganho de desempenho na correção dos dados da PRF, este novo experimento mostrou ganhos na inclusão de atributos relacionados às características da via, onde o modelo PRF/ANTT+ superou o desempenho do conjunto anterior. Porém, a incorporação dos dados meteorológicos e de tráfego não resultou em avanços, sugerindo que, para o trajeto de estudo, as condições estruturais das vias impactam mais nos acidentes do que fatores climáticos ou fluxo de veículos. Ademais, a ausência de dados de tráfego em um formato semelhante ao utilizado em trabalhos relacionados pode ter limitado a contribuição desta categoria.

Tabela 14 – Desempenho dos modelos para dados de validação sem o *km*.

Medida	**Conjunto de dados**	**RF**	**SVM**	**MLP**
Sensibilidade	PRF	0.6724 ± 0.0078	0.6085 ± 0.0120	**0.6797 ± 0.0082**
	PRF/ANTT	0.6838 ± 0.0060	0.6341 ± 0.0076	**0.6988 ± 0.0058**
	PRF/ANTT+	**0.7161 ± 0.0070**	0.6435 ± 0.0089	0.7117 ± 0.0073
	PRF/ANTT+/OPENMETEO	**0.7115 ± 0.0054**	0.6464 ± 0.0072	0.7096 ± 0.0043
	PRF/ANTT+/OPENMETEO/DNIT	**0.7187 ± 0.0038**	0.6551 ± 0.0046	0.7093 ± 0.0027
Especificidade	PRF	0.6633 ± 0.0025	**0.7056 ± 0.0045**	0.6600 ± 0.0044
	PRF/ANTT	0.6726 ± 0.0078	**0.7075 ± 0.0092**	0.6690 ± 0.0059
	PRF/ANTT+	0.6983 ± 0.0054	**0.7023 ± 0.0091**	0.6792 ± 0.0062
	PRF/ANTT+/OPENMETEO	0.6963 ± 0.0110	**0.7013 ± 0.0094**	0.6726 ± 0.0083
	PRF/ANTT+/OPENMETEO/DNIT	0.6877 ± 0.0074	**0.7031 ± 0.0123**	0.6773 ± 0.0067
AUC	PRF	0.7270 ± 0.0038	0.7142 ± 0.0062	**0.7301 ± 0.0048**
	PRF/ANTT	0.7385 ± 0.0053	0.7308 ± 0.0065	**0.7490 ± 0.0054**
	PRF/ANTT+	**0.7756 ± 0.0053**	0.7357 ± 0.0062	0.7613 ± 0.0055
	PRF/ANTT+/OPENMETEO	**0.7704 ± 0.0053**	0.7374 ± 0.0058	0.7581 ± 0.0041
	PRF/ANTT+/OPENMETEO/DNIT	**0.7690 ± 0.0048**	0.7410 ± 0.0054	0.7568 ± 0.0030

Fonte: Autor.

A Tabela 15 apresenta os resultados do teste de significância estatística para o novo experimento. Os maiores ganhos continuam concentrados na etapa de correção dos dados da PRF, evidenciando o impacto da melhoria da qualidade dos dados. Além disso, observa-se um aumento com a inclusão dos atributos viários da ANTT, enquanto

os demais dados não apresentaram impacto estatisticamente significativo. Observa-se na Tabela 16 o desempenho dos modelos para os dados de teste sem o atributo *km*, com métricas próximas às obtidas na validação, o que também indica boa capacidade de generalização dos modelos.

Tabela 15 – Teste de significância estatística sem o *km*.

Medida	**Conjunto de dados**	**RF**	**SVM**	**MLP**
Sensibilidade	PRF x PRF/ANTT	0.0480	0.0069	0.0052
	PRF/ANTT x PRF/ANTT+	0.0001	0.1482	0.0244
	PRF/ANTT+ x PRF/ANTT+/OPENMETEO	0.3252	0.9721	0.6321
	PRF/ANTT+/OPENMETEO x PRF/ANTT+/OPENMETEO/DNIT	0.06041	0.0588	0.9019
Especificidade	PRF x PRF/ANTT	0.0531	0.7271	0.0414
	PRF/ANTT x PRF/ANTT+	0.0006	0.4522	0.0435
	PRF/ANTT+ x PRF/ANTT+/OPENMETEO	0.7553	0.8840	0.2372
	PRF/ANTT+/OPENMETEO x PRF/ANTT+/OPENMETEO/DNIT	0.2303	0.8253	0.4025
AUC	PRF x PRF/ANTT	0.0076	0.0062	0.0007
	PRF/ANTT x PRF/ANTT+	9.68e-6	0.3065	0.0128
	PRF/ANTT+ x PRF/ANTT+/OPENMETEO	0.2081	0.6944	0.3732
	PRF/ANTT+/OPENMETEO x PRF/ANTT+/OPENMETEO/DNIT	0.7051	0.3870	0.6382

Fonte: Autor.

Tabela 16 – Desempenho dos modelos para dados de teste sem *km*.

Medida	**Conjunto de dados**	**RF**	**SVM**	**MLP**
Sensibilidade	PRF	0.6616	0.5757	**0.6868**
	PRF/ANTT	**0.6926**	0.63414	0.6829
	PRF/ANTT+	0.7107	0.6731	**0.7170**
	PRF/ANTT+/OPENMETEO	**0.7121**	0.6780	0.7024
	PRF/ANTT+/OPENMETEO/DNIT	**0.7163**	0.6926	0.6829
Especificidade	PRF	0.6718	**0.7150**	0.6684
	PRF/ANTT	0.67414	**0.7076**	0.6699
	PRF/ANTT+	0.6996	**0.7035**	0.6824
	PRF/ANTT+/OPENMETEO	0.7011	**0.7036**	0.6794
	PRF/ANTT+/OPENMETEO/DNIT	0.6984	**0.7032**	0.6821
AUC	PRF	**0.7450**	0.7126	0.7354
	PRF/ANTT	0.7462	0.7380	**0.7501**
	PRF/ANTT+	**0.7978**	0.7448	0.7653
	PRF/ANTT+/OPENMETEO	**0.7962**	0.7461	0.7546
	PRF/ANTT+/OPENMETEO/DNIT	**0.7961**	0.7513	0.7534

Fonte: Autor.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo desenvolver um modelo de predição de acidentes rodoviários em Santa Catarina, analisando o impacto do pré-processamento aplicado aos dados da PRF. 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 da PRF, reforçando a necessidade de tratar o conjunto de dados antes de desenvolver modelos preditivos. O processo de enriquecimento foi realizado em etapas, integrando aos poucos atributos de características da via, meteorológicos e de tráfego, o que permitiu avaliar o efeito incremental de cada grupo de atributos sobre o desempenho dos modelos RF, SVM e MLP.

Os resultados mostraram que a correção dos dados da PRF foi a etapa que mais impactou positivamente o desempenho dos modelos, com diferenças estatisticamente significativas entre PRF e PRF/ANTT. Etapas seguintes, incluindo a adição de atributos da ANTT, OPEN-METEO e DNIT, trouxeram ganhos pequenos ou não significativos, com o MLP mostrando aumento, SVM mantendo-se estável e RF apresentando leves quedas. Além disso, observou-se que o atributo *km* exerce forte influência sobre os modelos, concentrando grande parte da capacidade preditiva nos trechos com maior histórico de acidentes. Embora isso contribua para o desempenho no cenário analisado, pode limitar a generalização do modelo em outros contextos.

Nas limitações, destaca-se a ausência de dados de tráfego com a granularidade temporal e espacial utilizada em trabalhos relacionados, a qual 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 menor granularidade temporal. Também se destaca a possibilidade de reformular o modelo para realizar a predição da probabilidade de ocorrência de acidentes, em vez de uma classificação binária, o que permitiria uma análise mais adequada de risco de acidente. Por fim, destaca-se a importância de realizar testes complementares, além dos testes práticos, para garantir a confiabilidade do modelo antes de aplicar em ambientes reais.

REFERÊNCIAS

CAI, Qing; ABDEL-ATY, Mohamed; YUAN, Jinghui; LEE, Jaeyoung; WU, Yina. Real-time crash prediction on expressways using deep generative models. **Transportation Research Part C: Emerging Technologies**, v. 117, p. 102697, 2020. ISSN 0968-090X.

CNT. **Anuário CNT do transporte 2022**. [S.l.*:* s.n.], 2022. Acesso em: 26 set. 2024. Disponível em: https://cnt.org.br/documento/78a521c3-b71c-456b-85c8-e4ddf5e51166.

DNIT. **Departamento Nacional de Infraestrutura de Transportes - Pesquisa Nacional de Contagem de Tráfego (PNCT)**. Brasília, DF: [s.n.], 2025. https://servicos.dnit.gov.br/dadospnct. Acesso em: 7 mar. 2025.

DNIT. **Departamento Nacional de Infraestrutura de Transportes, VGeo - Sistema de Informações Geográficas do DNIT**. [S.l.*:* s.n.], 2025. https://servicos.dnit.gov.br/vgeo/ (acessado em 7 março 2025). Disponível em: https://servicos.dnit.gov.br/vgeo/.

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**, v. 205, p. 106314, 2020. ISSN 0950-7051.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André Carlos Ponce de Leon Ferreira. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro: LTC, 2011.

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**, v. 135, p. 105392, 2020. ISSN 0001-4575.

ISLAM, Zubayer; ABDEL-ATY, Mohamed; CAI, Qing; YUAN, Jinghui. Crash data augmentation using variational autoencoder. **Accident Analysis & Prevention**, v. 151, p. 105950, 2021. ISSN 0001-4575.

MO, Weiwei; LEE, Jaeyoung; ABDEL-ATY, Mohamed; MAO, Suyi; JIANG, Qianshan. Dynamic short-term crash analysis and prediction at toll plazas for proactive safety management. **Accident Analysis & Prevention**, v. 197, p. 107456, 2024. ISSN 0001-4575.

PENG, Yichuan; LI, Chongyi; WANG, Ke; GAO, Zhen; YU, Rongjie. Examining imbalanced classification algorithms in predicting real-time traffic crash risk. **Accident Analysis & Prevention**, v. 144, p. 105610, 2020. ISSN 0001-4575.

PRF, BRASIL. Polícia Rodoviária Federal. **Dados abertos da PRF**. [S.l.*:* s.n.], 2024. Acesso em: 23 set. 2024. Disponível em: https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf.

PRF, BRASIL. Polícia Rodoviária Federal. **Institucional**. [S.l.*:* s.n.], 2024. Acesso em: 23 set. 2024. Disponível em: https://www.gov.br/prf/pt-br/acesso-a-informacao/institucional#:~:text=Criada%20pelo%20Presidente%20Washington%20Lu%C3%ADs,Sistema%20Nacional%20de%20Seguran%C3%A7a%20P%C3%BAblica.

SCIKIT-LEARN DEVELOPERS. **sklearn.ensemble.RandomForestClassifier**. [S.l.], 2024. Acessado em: 27 maio 2025. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

SCIKIT-LEARN DEVELOPERS. **sklearn.svm.SVC**. [S.l.], 2024. Acessado em: 8 junho 2025. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

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

TRAN, Thanh; HE, Dan; KIM, Jiwon; HICKMAN, Mark. MSGNN: A Multi-structured Graph Neural Network model for real-time incident prediction in large traffic networks. **Transportation Research Part C: Emerging Technologies**, v. 156, p. 104354, 2023. ISSN 0968-090X.

TRANSPORTES, BRASIL. Ministério dos. **Transporte rodoviário de cargas**. [S.l.*:* s.n.], 2024. Acesso em: 23 set. 2024. Disponível em: https://www.gov.br/transportes/pt-br/assuntos/transporte-terrestre/transporte-rodoviario-de-cargas#:~:text=cerca%20de%2075%25%20de%20todas,rede%20de%20estradas%20do%20mundo.

USP. **Plano de Metas do Governo Juscelino Kubitschek**. [S.l.*:* s.n.], 1956. Acesso em: 23 set. 2024. Disponível em: https://edisciplinas.usp.br/pluginfile.php/5291773/mod_resource/content/1/Plano%20de%20Metas.pdf.

WHO. **Global Plan for the Decade of Action for Road Safety 2021-2030**. [S.l.*:* s.n.], 2021. Acesso em: 23 set. 2024. Disponível em: https://www.who.int/pt/publications/m/item/global-plan-for-the-decade-of-action-for-road-safety-2021-2030.

WHO. **Global Status Report on Road Safety 2023**. [S.l.*:* s.n.], 2023. Acesso em: 23 set. 2024. Disponível em: https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023.

YU, Le; DU, Bowen; HU, Xiao; SUN, Leilei; HAN, Liangzhe; LV, Weifeng. Deep spatio-temporal graph convolutional network for traffic accident prediction. **Neurocomputing**, v. 423, p. 135–147, 2021. ISSN 0925-2312.

ZIPPENFENIG, Patrick. **Open-Meteo.com Weather API**. [S.l.*:* s.n.], 2023. Disponível em: `https://open-meteo.com/`.

APÊNDICE A – CÓDIGO FONTE DO TRABALHO

Este apêndice apresenta o repositório contendo o código-fonte.

Repositório: `https://codigos.ufsc.br/gustavo.konescki/predicao-de-acidentes`

APÊNDICE B – PUBLICAÇÃO RELACIONADA AO TRABALHO

Este Trabalho de Conclusão de Curso resultou na seguinte publicação:

FÜHR, Gustavo Konescki; INACIO, Eduardo Camilo; FILETO, Renato. *Predição de acidentes rodoviários em Santa Catarina: impactos de aperfeiçoamentos dos dados*. In: **ESCOLA REGIONAL DE BANCO DE DADOS (ERBD)**, 20., 2025, Florianópolis/SC. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 60–69. ISSN 2595-413X. **DOI:** `https://doi.org/10.5753/erbd.2025.7274`.

O artigo completo encontra-se incluído nas próximas páginas como material complementar.

Predição de acidentes rodoviários em Santa Catarina: impactos de aperfeiçoamentos dos dados

Gustavo Konescki Führ[1], **Eduardo Camilo Inacio**[1], **Renato Fileto**[1]

[1]Dep. de Informática e Estatística (INE), Univ. Federal de Santa Catarina (UFSC)
Campus Reitor João David Ferreira Lima (Trindade), Florianópolis-SC – Brasil

gustavokf2003@gmail.com, eduardo.camilo@ufsc.br, r.fileto@ufsc.br

Resumo. *Este trabalho foca no ajuste de dados antes do treinamento de modelos para predizer acidentes rodoviários. Usa registros da Polícia Rodoviária Federal (PRF) sobre acidentes em Santa Catarina. Uma análise exploratória desses dados permitiu identificar inconsistências entre registros de acidentes em cada trecho de 100 metros. Isso motivou a correção e complementação dos dados da PRF, usando informações de outras fontes sobre as vias, antes de treinar modelos preditivos. Modelos RF, SVM e MLP foram treinados com os dados originais e melhorados. Experimentos revelaram que melhoramentos nos dados permitiram aumentar significativamente a acurácia e o F1-Score dos modelos RF e SVM, enquanto a inclusão de novas variáveis teve impacto menor.*

1. Introdução

Acidentes de trânsito estão entre os principais problemas mundiais, afetando a saúde pública e a economia. Pesquisa da Organização Mundial da Saúde (OMS)[1] publicada em dezembro de 2023, estima que 1,19 milhão de pessoas morrem anualmente em acidentes de trânsito, principal causa de morte das pessoas entre 5 e 29 anos. Segundo os dados abertos da PRF [PRF 2025], no período de 2018 a 2023, foram registrados 47.436 acidentes, em Santa Catarina, o que representa 11,94% dos acidentes em todo o território nacional, posicionando o estado como o segundo com maior número de ocorrências. Ademais, dados do Painel da Confederação Nacional do Transporte (CNT) de Consultas Dinâmicas sobre Acidentes Rodoviários[2] indicam que, apenas no ano de 2022, Santa Catarina teve prejuízo estimado em 1,32 bilhão de reais em decorrência desses acidentes.

Assim, modelos preditivos de acidentes se tornam cruciais para planejar a alocação de recursos limitados para prevenção e atendimento a emergências. Nesse contexto, pesquisadores têm explorado novas técnicas para o aprimoramento desses modelos, incluindo redes neurais baseadas em grafos, como propostas por [Yu et al. 2021, Tran et al. 2023], além de algoritmos de balanceamento de dados, como aprofundado por [Cai et al. 2020, Peng et al. 2020]. No entanto, poucas pesquisas têm se dedicado à qualidade dos dados usados no treinamento, fator essencial para robustez e bom desempenho.

Este estudo investiga a influência do aperfeiçoamento prévio dos dados empregados no treinamento de modelos preditivos de acidentes em seu desempenho. Para isso, foram utilizados registros da PRF sobre acidentes ocorridos de 2018 a 2023 entre os quilômetros 100 e 239 da BR-101 em Santa Catarina. Na análise desses dados,

[1]https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

[2]https://cnt.org.br/documento/78a521c3-b71c-456b-85c8-e4ddf5e51166

identificaram-se inconsistências nos atributos da via indicados em diferentes registros de acidentes em um mesmo trecho de 100 metros, tais como valores distintos para o tipo de pista (simples ou dupla), o traçado (curva, reta, aclive, declive, etc.) e a área cortada pelo trecho da via (urbana ou rural). Isso nos levou à formulação de métodos para corrigir e enriquecer esses dados, usando informações mais confiáveis e complementares sobre vias, de bases da Agência Nacional de Transportes Terrestres (ANTT) e do Departamento Nacional de Infraestrutura de Transportes (DNIT). Um diferencial deste trabalho é não utilizar dados de tráfego, como faz a maioria dos trabalhos relacionados. Ainda assim o desempenho obtido é competitivo, graças ao melhoramento e enriquecimento dos registros de acidentes com informação do DNIT e da ANTT.

Três algoritmos foram avaliados: Floresta Aleatória (RF), Máquina de Vetores de Suporte (SVM) e Perceptron Multicamadas (MLP). Os modelos foram treinados com os dados originais, os dados corrigidos e os dados corrigidos com variáveis adicionais, visando uma análise comparativa para verificar o desempenho. Experimentos com essas alternativas revelaram que melhoramentos nos dados permitiram aumentar a acurácia e o F1-Score dos modelos RF em cerca de 4% e 5%, respectivamente e o F1-Score do SVM em cerca de 4%. Todavia, a inclusão de novas variáveis teve impacto menor e os ganhos nos modelos MLP com melhorias nos dados foram menos significativos.

O restante desse trabalho está organizado em mais 4 seções. A Seção 2 apresenta os fundamentos do nosso estudo e discute trabalhos relacionados. A Seção 3 descreve os dados e métodos utilizados. A Seção 4 reporta e discute resultados de experimentos. Finalmente, a Seção 5 tece as conclusões e enumera alguns temas para trabalhos futuros.

2. Fundamentos

2.1. Modelos Preditivos de Aprendizado de Máquina

Modelos preditivos que utilizam técnicas de aprendizado de máquina têm-se mostrado eficientes na previsão de acidentes. Esses modelos detectam padrões em dados históricos por meio da análise de variáveis temporais, espaciais, geográficas e outras, permitindo antecipar situações de risco.

Neste trabalho, utilizamos três algoritmos para treinar classificadores, comumente usados em outros trabalhos sobre predição de acidentes rodoviários: Floresta Aleatória (*Random Forest* – RF) [Tran et al. 2023, Huang et al. 2020, Peng et al. 2020], Máquinas de Vetores de Suporte (*Support Vector Machine* – SVM) [Cai et al. 2020, Yu et al. 2021, Tran et al. 2023, Huang et al. 2020] e Perceptron multicamadas (*Multilayer Perceptron* – MLP) [Cai et al. 2020, Huang et al. 2020, Peng et al. 2020]. RF combina várias árvores de decisão para criar um preditor forte. SVM visa encontrar um hiperplano com a maior margem possível para a separação entre as classes. Além disso, o uso de funções *kernel* possibilita solucionar problemas não lineares, ao transformar o espaço de atributos em mais dimensões. MLP é uma das arquiteturas de redes neurais mais empregadas, em razão do seu alto desempenho para a maioria dos problemas e por servir de base para técnicas mais avançadas.

2.2. Técnicas de Balanceamento

Treinar modelos de aprendizado de máquina com dados desbalanceados induz o modelo a priorizar a classe majoritária, o que resulta em uma baixa precisão com a

classe minoritária. Dados de acidentes são extremamente desbalanceados, pois possuem muito mais registros de não acidentes do que de acidentes ao longo do tempo e do espaço. Nesse sentido, foi escolhida a técnica de **subamostragem aleatória** [Yu et al. 2021, Tran et al. 2023, Huang et al. 2020] para o balanceamento dos dados. Este método visa descartar aleatoriamente observações da classe majoritária, a fim de balancear proporcionalmente as classes.

2.3. Métricas de avaliação

O desempenho dos modelos de predição de acidentes foi avaliado neste trabalho através de métricas frequentemente usadas em algoritmos de classificação para dados desbalanceados. **Acurácia** mede a proporção de previsões corretas feitas por um modelo em relação ao total de previsões realizadas. **Sensibilidade (*Recall*)** é a proporção dos casos corretamente preditos como positivos dentre todos os casos positivos. **Precisão** avalia a proporção dos casos corretamente preditos positivos dentre todos os casos preditos como positivos. **F1-Score** calcula a média harmônica entre sensibilidade e precisão.

2.4. Trabalhos Relacionados

A predição de acidentes de trânsito tem sido amplamente investigada nos últimos anos, aplicando novas técnicas de balanceamento dos dados e aprendizado de máquina para melhorar o desempenho dos classificadores. Modelos baseados em grafos têm ganhado destaque por conseguirem analisar relações espaço-temporais complexas. [Yu et al. 2021] propuseram uma nova Rede Convolucional de Grafos Espaço-Temporais Profunda denominada DSTGCN que realiza operações convolucionais em grafos para aprender correlações espaciais e capturar variações dinâmicas espaço-temporais. [Tran et al. 2023] apresentaram uma Rede Neural Multi-estruturada (MSGNN) que captura relacionamentos espaço-temporais entre links de cada subárea.

Outras pesquisas destacam novas técnicas de balanceamento dos dados para a melhoria dos modelos. [Cai et al. 2020] mostraram que para modelos de aprendizado de máquina mais complexos, o uso de Rede Generativa Adversarial Convolucional Profunda (DCGAN) no balanceamento de dados resulta em aumento de desempenho dos modelos em comparação com outras técnicas. [Peng et al. 2020], além de avaliar SMOTE e sobreamostragem como estratégias de balanceamento dos dados, também examinaram táticas para tratar dados desbalanceados em nível de saída, através do Índice de Youden e do Método de Calibração de Probabilidade, e em nível de algoritmo com os modelos MLP sensível ao custo de amostragem aleatória (RCSMLP) e Rusboost. Todavia, não encontramos trabalhos focados na resolução de inconsistências de características de vias em relatos de acidentes em um mesmo trecho e correção desses problemas antes do treinamento de modelos, visando melhorar seu desempenho, como proposto na nossa pesquisa.

3. Metodologia

Esta pesquisa segue uma metodologia usual em aprendizado de máquina, com as fases de entendimento do negócio e dos dados, preparação dos dados, treinamento e avaliação de modelos. As subseções a seguir destacam seus aspectos mais relevantes.

3.1. Bases de Dados

Registros de Acidentes em SC da PRF

Para o treinamento e avaliação dos modelos de aprendizado de máquina, utilizaram-se dados sobre acidentes em Santa Catarina, registrados pela Polícia Rodoviária Federal entre 2018 e 2023. Foi estabelecido o trecho da BR-101 entre os quilômetros 100 (Vale do Itajái) a 239 (Palhoça) para os experimentos de predição, por concentrar a maior parte dos registrados. Neste período, a PRF anotou 14.792 acidentes em tal trecho. A Figura 1 ilustra acidentes reportados pela PRF nas rodovias de Santa Catarina no período. Os raios dos círculos em azul em torno de locais onde aconteceram acidentes representam as quantidades de acidentes na região.

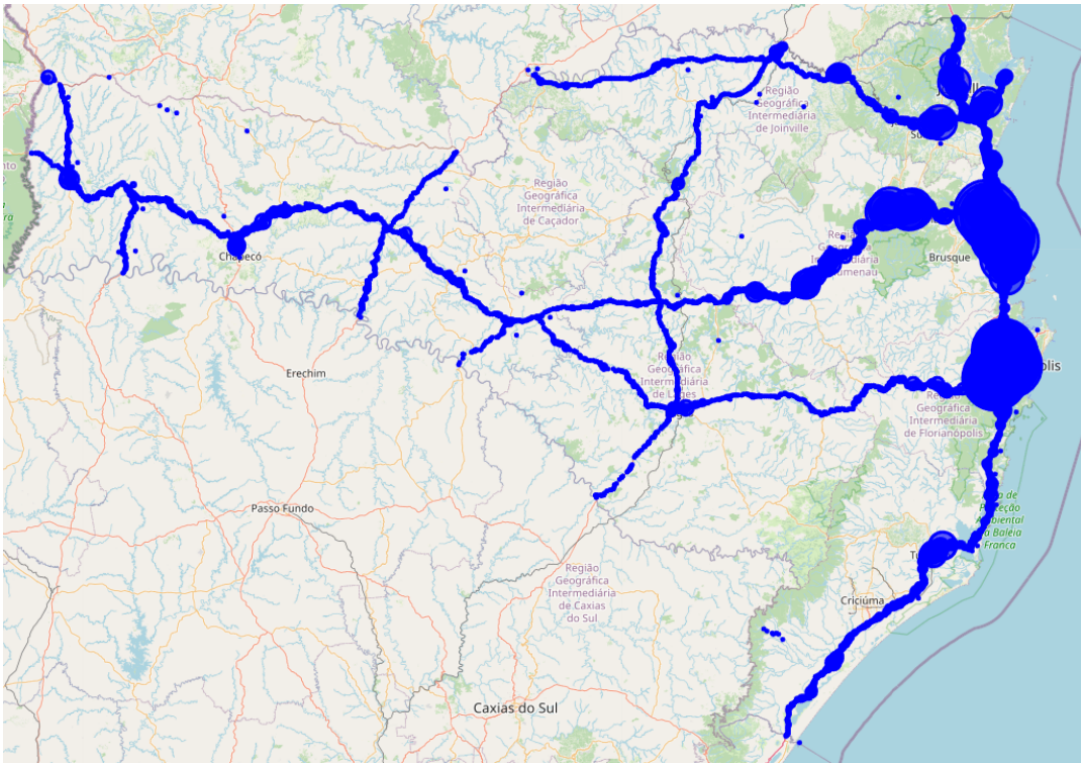


Figura 1. Acidentes registrados pela PRF em rodovias federais de SC.

Os dados da PRF têm 25 características (*features*) que cobrem os aspectos como: momento e local do acidente, características da via, condição meteorológica e número de feridos e mortos [PRF 2025]. As características selecionadas para o treinamento dos modelos estão descritas na Tabela 1. Foram escolhidas características geralmente utilizadas em outras pesquisas de predição de acidentes de trânsito como: fatores temporais [Yu et al. 2021, Huang et al. 2020], fatores espaciais ([Yu et al. 2021, Huang et al. 2020] e características da via [Yu et al. 2021, Mo et al. 2024]. A granularidade das observações foi definida em horas para o aspecto temporal e em 100 metros para o aspecto espacial, por não haver registro de múltiplos acidentes na mesma hora e local, além de pouca variação de números de acidentes em trechos consecutivos de 100 metros.

Após a análise dos dados, constatou-se que 68,46% das observações de acidentes registradas para o mesmo trecho de 100 metros apresentam pelo menos um valor divergente em características da via. A Tabela 2 ilustra tais divergências, considerando todos

os registros de acidentes no quilômetro 233,0 da BR-101, sentido decrescente, entre 2018 e 2023. Nota-se que para o mesmo trecho foram reportados diferentes valores para os atributos "tipo_pista", "tracado_via"e "uso_solo". Por este motivo, foram buscadas novas fontes de dados para corrigir e aperfeiçoar a qualidade dos atributos viários, reduzindo as inconsistências encontradas para melhor treinar modelos preditores.

Dados sobre Vias da ANTT

A Agência Nacional de Transportes Terrestres dispõe de quarenta coleções de dados sobre as rodovias brasileiras [ANTT 2025]. Destas, foram selecionados dados sobre características do quilômetro, município, número de faixas, traçado da via e uso do solo para a correção dos dados da PRF. Além disso, foram escolhidas novas variáveis para avaliar se sua inclusão melhora o desempenho dos modelos preditivos, tais como o tipo de pavimento, o tipo de perfil do terreno, a velocidade regulamentada para veículos leves, a velocidade regulamentada para veículos pesados, a presença de pista marginal e a existência de iluminação.

As características da rodovia dos dados da ANTT estão em arquivos CSV distintos, onde os valores dos atributos de velocidade e quilômetro são delimitados por latitude e longitude, enquanto os demais são definidos por uma faixa de coordenadas que indica os limites inicial e final de latitude e longitude. A Tabela 3 descreve os atributos escolhidos e o conjunto de dados de onde cada característica foi retirada. Os conjuntos de dados de pista marginal e iluminação não apresentam um atributo específico que represente diretamente essas características. No lugar disso, essas informações são fornecidas apenas pelas coordenadas que indicam o trecho onde há pista marginal ou iluminação.

Dados sobre Vias da DNIT

A geometria da BR-101, entre os quilômetros 100 e 239 em Santa Catarina, foi extraída do banco de dados geográficos do DNIT. Esses dados estão no formato *MultiLineString*, que representa múltiplas linhas geográficas, com latitude e longitude da via [DNIT 2025].

3.2. Geração de Dados de Não Acidentes

Para treinar os modelos preditivos selecionados, foi necessário criar amostras de não acidentes, pois os dados da PRF possuem apenas dados de acidentes. Para isso foram primeiramente geradas todas as combinações espaço-temporais possíveis, através do produto

Tabela 1. Características (*features*) utilizadas dos dados da PRF .

Variável	Descrição
km	Identificação do quilômetro da via onde ocorreu o acidente, com precisão de 0,1 km.
municipio	Nome do município onde ocorreu o acidente.
sentido_via	Sentido da via considerando o ponto de colisão. Ex: Crescente, Decrescente.
tipo_pista	Categoria da quantidade de faixas da via principal. Ex: Simples, Dupla, Múltipla.
tracado_via	Característica do tipo de traçado da via. Ex: Reta, Curva, Aclive, etc.
uso_solo	Tipo de ocupação do solo. Ex: Sim (urbano), Não (rural).
data_inversa	Data no formato dd/mm/aa.
horario	Horário no formato hh:mm:ss.

Tabela 2. Características da via em registros de acidentes no km 233,0 da BR-101.

km	municipio	tipo_pista	tracado_via	uso_solo
233.0	PALHOCA	Múltipla	Reta	Não
233.0	PALHOCA	Dupla	Reta	Sim
233.0	PALHOCA	Dupla	Reta	Sim
233.0	PALHOCA	Dupla	Reta;Declive	Não
233.0	PALHOCA	Dupla	Curva	Não
233.0	PALHOCA	Dupla	Curva;Declive	Não

cartesiano entre as horas no período de 2018 a 2023, os décimos de quilômetros entre os quilômetros 100 a 239 da BR-101 e os sentidos da via (crescente ou decrescente). Posteriormente, foram utilizados os dados da PRF para verificar quais dessas combinações registraram acidentes. Ao final, obteve-se 147.065.678 observações com uma proporção de 99,99% de não acidentes e 0,01% de acidentes. Por essa razão, realiza-se uma subamostragem aleatória dos dados de não acidentes no conjunto de treinamento e validação, a fim de equilibrar a proporção de acidentes e não acidentes em 50%.

Para preencher o produto cartesiano com as variáveis de características da via faltantes, foi realizado um *merge* para dados de acidentes com as variáveis registradas no momento do próprio acidente. Em contrapartida, para dados de não acidentes, foram extraídas as variáveis 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 das variáveis que representam dia, dia da semana, mês, a existência de feriado a partir do atributo "data_inversa"e hora com base no "horario".

3.3. Correção e enriquecimento dos dados sobre acidentes e não acidentes

A partir dos dados da DNIT, foram extraídos pontos que correspondem a latitude e longitude de uma certa localização da via. No entanto, os dados originais possuem apenas o segmento do sentido crescente da via. Portanto, foi necessário criar a representação do sentido decrescente, duplicando os pontos e ajustando suas posições geográficas.

Para os conjuntos de dados que possuem as características de velocidades e quilômetro, foram criados pontos geográficos, pois estes apresentam apenas um valor de latitude e longitude. Em relação aos outros dados, foi gerado um atributo do tipo polígono para representar a latitude e longitude final e inicial das categorias.

Para obter as propriedades das vias da ANTT nos pontos do trajeto da via da DNIT, foram realizadas operações de proximidade para descobrir o ponto ou polígono de característica mais próximo de um ponto da rodovia. Porém, para os conjuntos de dados de pista marginal e de iluminação, foram realizadas operações para verificar se os pontos da via estão contidos nos polígonos das características, visto que não havia dados de região sem iluminação e pista marginal. Dessa forma, cada ponto geográfico da área de interesse da BR-101 passou a possuir apenas um único valor possível para cada característica da via nos novos dados, eliminando completamente as contradições nos dados da PRF.

Tabela 3. Características utilizadas dos dados da ANTT.

Variável	Descrição	Conjunto de dados de origem
km	Representação do quilômetro mais a metragem. Ex: 317,940	Quilômetro Pista Principal
municipio	Nome do município.	Município
numero de faixas	Quantidade de faixas da via principal.	Pista principal
tracado via	Representação do tipo de traçado. Ex: Curva ou Tangente	Traçado
tipo de uso do solo	Representação do tipo do uso do solo. Ex: Urbano ou Rural.	Uso do Solo
tipo do pavimento	Representação da ordem pavimento. Ex: Rígido ou Flexível.	Tipo pavimento
tipo de perfil do terreno	Representação do tipo de perfil do terreno. Ex: Montanhoso, Plano e Ondulado	Perfil do Terreno
velocidade regulamentada veiculos leves	Representação da velocidade máxima permitida. Ex.: 40 km/h	Sinalização
velocidade regulamentada veiculos pesados	Representação da velocidade máxima permitida. Ex.: 40 km/h	Sinalização

Dados Corrigidos e Novas Variáveis

Para a criação do conjunto de dados corrigido e com variáveis adicionais, foi utilizado o mesmo conjunto de treinamento e validação. No entanto, para preencher as informações sobre as características da via, foi realizada um *merge* com o *dataset* de características criado. Além disso, foram criadas as variáveis que representam o dia, o dia da semana, o mês, a existência de feriado a partir do atributo "data_inversa"e a hora com base no atributo "horario".

3.4. Treinamento dos Modelos

As variáveis dos conjuntos de dados foram transformadas em categóricas, utilizando a técnica de One-Hot-Coding. Esse método cria uma nova coluna para cada valor possível dentro de uma categoria e atribui um valor binário (0 ou 1) para indicar a ausência ou presença desse valor. Em seguida, foram treinados três diferentes modelos de aprendizado de máquina: Floresta Aleatória, Máquina de Vetor de Suporte e Perception Multicamadas, para cada um dos três conjuntos de dados diferentes: dados da PRF, dados da PRF corrigidos e dados da PRF corrigidos com as novas variáveis, totalizando nove modelos.

Os modelos foram treinados utilizando validação cruzada 5-Fold, com base na métrica de F1-score para os dados de treinamento, o que permitiu uma seleção otimizada dos hiperparâmetros. Essa técnica divide o conjunto de dados em k partes. Em cada

iteração, o modelo é treinado com k-1 partes e validado com a parte restante. O processo é repetido k vezes e a média dos resultados fornece uma estimativa confiável da performance do modelo, minimizando o risco de sobre-ajuste. A Tabela 4 apresenta todos os hiperparâmetros e valores testados, enquanto a Tabela 5 mostra os melhores valores ajustados para cada hiperparâmetro de cada modelo.

Tabela 4. Hiperparâmetros testados para os modelos RF, SVM e MLP.

Modelo	Hiperparâmetros e Valores Testados
RF	número_de_árvores: [100, 200, 500] profundidade_máxima: [10, 50 , 100] mínimo_amostras_divisão: [2, 5, 10] mínimo_amostras_folha: [1, 2, 4] bootstrap: [Verdadeiro, Falso]
SVM	C: [0.1, 1, 10] kernel: [linear, rbf, poly, sigmoid] gamma: [scale, auto]
MLP	tamanho_camadas_ocultas: [(64), (126,64), (256, 126, 64)] ativação: [relu, tanh] taxa de aprendizado: [0.0001, 0.001, 0.01]

Tabela 5. Melhores hiperparâmetros para cada modelo e conjunto de dados.

Modelo	Hiperparâmetro	PRF	PRF corrigido	PRF corrigido + variáveis
RF	número de árvores	500	500	200
	profundidade máxima	100	100	100
	mínimo amostras divisão	2	2	2
	mínimo amostras folha	1	1	1
SVM	C	0.1	10	10
	kernel	linear	linear	linear
	gamma	scale	scale	scale
MLP	tamanho camadas ocultas	(256, 128, 64)	(256, 128, 64)	(64)
	ativação	tanh	relu	relu
	taxa de aprendizado	0.001	0.0001	0.0001

4. Resultados

A Tabela 6 exibe a média e o desvio padrão das métricas de cada modelo nos dados de validação para cada conjunto de dados. Em geral, o modelo MLP possui as melhores métricas em relação aos outros modelos, apresentando os maiores valores de acurácia, sensibilidade e F1-Score em todas as amostras de dados. O modelo SVM aplicado aos dados da PRF demonstrou um viés na seleção da variável de saída, resultando em uma alta precisão, porém com baixa sensibilidade. O modelo RF para os dados da PRF também indicou uma sensibilidade baixa, porém essa métrica apresentou melhoria com a correção dos dados.

O Teste t de Student foi aplicado para avaliar se as diferenças entre os conjuntos de dados são estatisticamente relevantes, sendo que um p-valor inferior a 0,05 indica significância na diferença entre os grupos. A Tabela 7 apresenta os valores de p-valor para a comparação entre diferentes versões dos dados dentro de cada modelo para as

métricas de acurácia e F1-score. Há uma melhoria na correção dos dados da PRF para os modelos Floresta Aleatória e Máquina de Vetores de Suporte. A adição de novas variáveis provocou uma pequena melhoria no modelo RF e uma leve redução no modelo SVM, porém ambas sem significância estatística. O Modelo MLP não apresentou um aumento significativo na correção dos dados ou na inclusão de novas características.

Tabela 6. Desempenho dos modelos para dados de validação.

Medida	Conjunto de dados	RF	SVM	MLP
Acurácia	PRF	0.7326 ± 0.0051	0.7511 ± 0.0047	0.7617 ± 0.0058
	PRF corrigido	0.7573 ± 0.0042	0.7627 ± 0.0041	**0.7653** ± 0.0045
	PRF corrigido + variáveis	0.7633 ± 0.0053	0.7564 ± 0.0063	0.7582 ± 0.0044
Sensibilidade	PRF	0.6536 ± 0.0099	0.6197 ± 0.0086	0.7600 ± 0.0136
	PRF corrigido	0.7256 ± 0.0056	0.7496 ± 0.0077	0.7597 ± 0.0060
	PRF corrigido + variáveis	0.7332 ± 0.0072	0.7463 ± 0.0108	**0.7931** ± 0.0119
Precisão	PRF	0.7764 ± 0.0043	**0.8406 ± 0.0026**	0.7626 ± 0.0061
	PRF corrigido	0.7747 ± 0.0066	0.7698 ± 0.0046	0.7683 ± 0.0050
	PRF corrigido + variáveis	0.7802 ± 0.0028	0.7616 ± 0.0047	0.7414 ± 0.0074
F1-Score	PRF	0.7097 ± 0.0055	0.7135 ± 0.0062	0.7612 ± 0.0072
	PRF corrigido	0.7493 ± 0.0045	0.7596 ± 0.0049	0.7640 ± 0.0051
	PRF corrigido + variáveis	0.7560 ± 0.0048	0.7538 ± 0.0066	**0.7663** ± 0.0039

Tabela 7. Teste de Significância Estatística (p-value).

Medida	Conjunto de dados	RF	SVM	MLP
Acurácia	PRF X PRF corrigido	7.75e-5	0.0062	0.3552
	PRF corrigido X PRF corrigido + variáveis	0.1127	0.1324	0.0532
F1-Score	PRF X PRF corrigido	3.94e-6	2.83e-6	0.5566
	PRF corrigido X PRF corrigido + variáveis	0.0808	0.2031	0.4944

5. Conclusões e Trabalhos Futuros

A criação de modelos preditores de acidentes de trânsito é fundamental para melhorar a segurança pública nas rodovias. Este estudo foca na predição de acidentes rodoviários em Santa Catarina, demonstrando que melhorias nos resultados nas métricas desses classificadores podem ser alcançadas apenas no aperfeiçoamento dos dados.

Para resumir este trabalho, foram coletados dados de acidentes da PRF entre os quilômetros 100 até 239 da BR-101 de SC, no período de 2018 a 2023. Ao analisar os dados, notou-se que 68,46% das ocorrências de acidentes no mesmo trecho de 100 metros apresentam pelo menos uma discrepância nos valores das características da via, indicando uma inconsistência nos dados. Por conseguinte, foi criado um novo conjunto de dados para corrigir as informações da PRF, além da adição de novas variáveis, utilizando a base de dados da ANTT e da DNIT. Foram treinados três diferentes modelos de aprendizado de máquina: RF, SVM e MLP, para cada um dos três conjuntos de dados diferentes: dados

da PRF, dados da PRF corrigidos e dados da PRF corrigidos com as novas variáveis, resultando em nove modelos. Os resultados obtidos para os dados de validação mostraram uma melhoria significativa na acurácia e no F1-score dos modelos RF e SVM.

Estudos mais aprofundados precisam ser feitos para avaliar a capacidade de generalização de diversos modelos. Além disso, pode-se investigar a adição de novos fatores, como condições meteorológicas, tráfego em tempo real e eventos regionais que possam influenciar a ocorrência de acidentes. Conjugando técnicas de aprendizagem de máquina com correções e enriquecimento dos dados espera-se aumentar mais os ganhos de desempenho, cuja relevância estatística sabemos que precisa também ser avaliada para cada fator, à medida que nossos estudos avançam.

Referências

[ANTT 2025] ANTT (2025). Base de dados de rodovias federais da agência nacional de transportes terrestres. `https://dados.antt.gov.br/group/rodovias?page=2` (acessado em 7 março 2025).

[Cai et al. 2020] Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., and Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. *Transportation Research Part C: Emerging Technologies*, 117:102697.

[DNIT 2025] DNIT (2025). Departamento nacional de infraestrutura de transportes, vgeo - sistema de informações geográficas do DNIT. `https://servicos.dnit.gov.br/vgeo/` (acessado em 7 março 2025).

[Huang et al. 2020] Huang, T., Wang, S., and Sharma, A. (2020). Highway crash detection and risk estimation using deep learning. *Accident Analysis Prevention*, 135:105392.

[Mo et al. 2024] Mo, W., Lee, J., Abdel-Aty, M., Mao, S., and Jiang, Q. (2024). Dynamic short-term crash analysis and prediction at toll plazas for proactive safety management. *Accident Analysis Prevention*, 197:107456.

[Peng et al. 2020] Peng, Y., Li, C., Wang, K., Gao, Z., and Yu, R. (2020). Examining imbalanced classification algorithms in predicting real-time traffic crash risk. *Accident Analysis Prevention*, 144:105610.

[PRF 2025] PRF, P. R. F. (2025). Dados abertos da PRF. `https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf` (acessado em 5 março 2025).

[Tran et al. 2023] Tran, T., He, D., Kim, J., and Hickman, M. (2023). Msgnn: A multi-structured graph neural network model for real-time incident prediction in large traffic networks. *Transportation Research Part C: Emerging Technologies*, 156:104354.

[Yu et al. 2021] Yu, L., Du, B., Hu, X., Sun, L., Han, L., and Lv, W. (2021). Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing*, 423:135–147.