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.

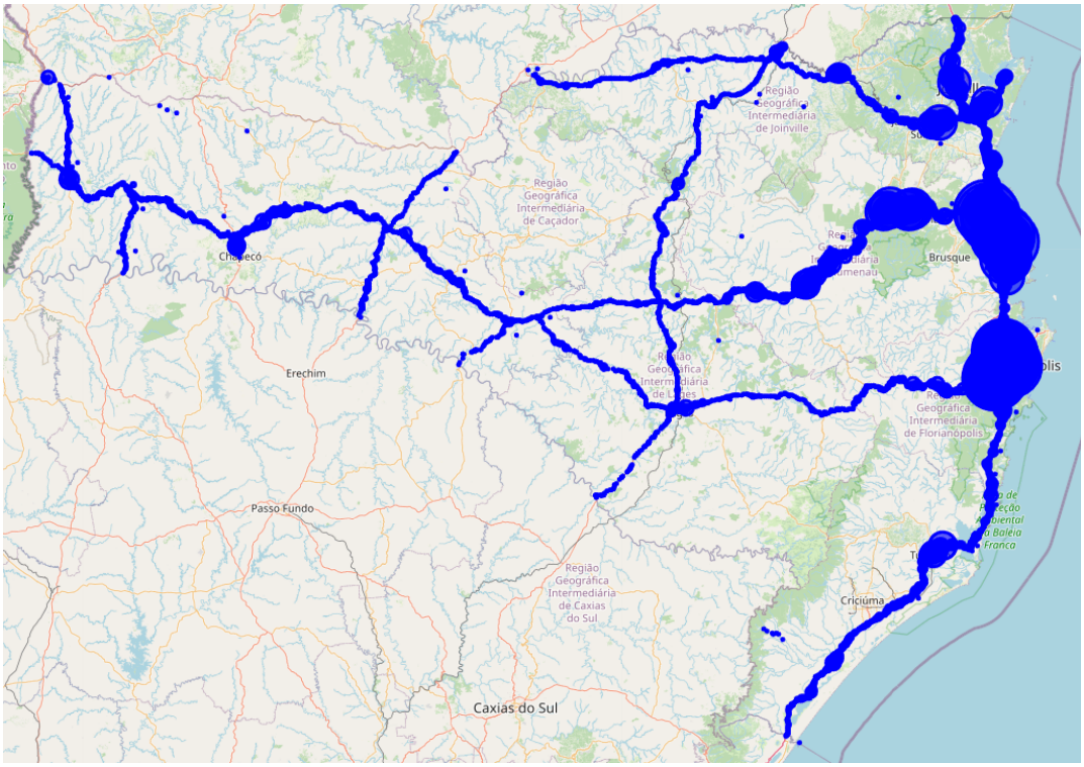

Figura 1. Acidentes registrados pela PRF em rodovias federais de SC.

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.