# Comparação de modelos de classificação de categorias de acidentes nas rodovias federais

**Luis E. Oliveira**[1], **André P. Borges**[2]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Ponta Grossa – PR – Brasil

[2]Universidade Tecnológica Federal do Paraná (UTFPR)
Ponta Grossa – PR – Brasil

luiso.1998@alunos.utfpr.edu.br, apborges@utfpr.edu.br

***Abstract.*** *This paper is the result of applying machine learning algorithms to PRF data, with the aim of predicting a type of accident. The original information was enriched with information about tolls, speed cameras and holidays. 32 tests were carried out with different combinations of algorithms and their parameters on the data. The results suggest that more inputs would still be needed to enrich the data and achieve higher accuracy rates.*

***Resumo.*** *Este trabalho é resultado da aplicação de algoritmos de aprendizagem de máquina em dados da PRF, com objetivo de prever um tipo de acidente. As informações originais foram enriquecidas com informações sobre pedágios, radares e feriados. Foram realizados 32 testes com combinações diferentes de algoritmos e seus parâmetros nos dados. Os resultados sugerem que ainda seriam necessários mais insumos para enriquecer os dados e atingir taxas de acurácia mais elevadas.*

## 1. Introdução

Acidentes em rodovias custam à sociedade cerca de R$ 40 bilhões por ano, sendo que destes, R$ 12 bilhões são apenas de acidentes em rodovias federais [IPEA 2015]. Além dos custos diretos, existem, ainda, custos indiretos, como o reparo dos veículos objetos dos acidentes, custos com atendimento médico das vítimas, a perda, temporária ou definitiva, total ou parcial, da capacidade laborativa dos envolvidos, além de, em casos fatais, custos funerários. Ademais das cifras referente a quantidade de dinheiro gasta por conta de acidentes de trânsito, mostram-se de suma relevância, também, a perda humana decorrente deste tipo de situação. Segundo a Organização Pan-Americana da Saúde [da Saúde OPS 2018], a principal causa de morte entre os jovens de 15 a 29 anos são os acidentes de trânsito, fato que gera como consequência tanto para os entes queridos das vítimas quanto para a sociedade em si, eis que esta perde parte de sua população jovem, em capacidade produtiva.

Segundo [Fernandes and Chiavegatto 2019] com o aumento da produção, organização e categorização de dados, tornou-se possível utilizar técnicas de Mineração de Dados para resolver problemas que, sem a utilização dessas técnicas, se tornam muito difíceis ou mesmo impraticáveis, como verificação da possibilidade da ocorrência de acidentes de trânsito em determinados trechos, condições climáticas ou horários. Como exemplo,

tem-se que o sistema de previsão de risco de acidente com base em dados esparsos heterogêneos desenvolvido por [Moosavi et al. 2019] com Aprendizagem por Reforço capaz de prever de onde acidentes podem acontecer nos próximos minutos.

Acidentes de trânsito são um problema atual e relevante, razão pela qual se justifica o seu estudo. No presente trabalho, foram efetuados testes com algoritmos de Aprendizagem de Máquina, para realizar a categorização de determinado tipo de acidente. Para tanto, foram utilizados dados públicos de acidentes ocorridos em rodovias federais, disponibilizados pela Polícia Rodoviária Federal (PRF) [1] antes da ocorrência dos mesmos. Para a realização deste estudo, será utilizado o Google Colab, o qual possibilita a análise de dados utilizando a linguagem Python por serem plataformas gratuitas. Ainda, de forma sintética, o presente trabalho se propôs a realizar a avaliação, preparação e análises dos dados disponibilizado pela PRF. Sendo que nesta última etapa, foram utilizados algoritmos de categorização, com intuito de encontrar o que provê o melhor resultado, a partir dos dados disponíveis. Os melhores resultados dos testes tiveram cerca de 30% de acurácia ao categorizar os tipos de acidentes.

O artigo está organizado da seguinte forma: na seção 2 é apresentada a revisão da literatura, analisando trabalhos correlatos. Na 3 é apresentada a metodologia do trabalho, passando pela coleta, enriquecimento, padronização e aplicação dos algoritmos. Em 4 todos os casos de testes e seus resultados são explanados. Por fim, na 5 são apresentadas as considerações finais do trabalho.

## 2. Revisão da Literatura

Para atingir os objetivos da pesquisa, foi realizada uma revisão bibliográfica de artigos que abordaram Descoberta de Conhecimento e Mineração de Dados a partir dos dados de acidentes da PRF. As strings de busca utilizadas foram “Policia Rodoviária Federal” OR “PRF” AND “análise” OR “correlação” OR ”classificação”. Os resultados foram filtrados por meio da leitura dos resumos e leitura completa dos artigos, resultando em um total de 10 artigos que atenderam os critérios definidos. Os resultados encontrados foram organizados e apresentados na seção de Análise da Bibliografia (2.3), apresentando os principais conteúdos abordados pelos artigos selecionados, que embasam a pesquisa.

### 2.1. Correlação

No trabalho de [Costa et al. 2014] foi utilizado o Modelo *Apriori*, com os algoritmos J48 e PART, para buscar correlações entre os dados e a coluna “causa acidente”. Uma vez que foram geradas 38 regras de associação com confiança maior que 0,8 (em que 1 é a associação direta perfeita e 0 sem associação), concluiu-se que os algoritmos utilizados são promissores em relação à classificação das causas de acidentes.

Em [Silva et al. 2015] foram analisados trechos críticos da BR-381 entre os anos de 2008 a 2012. O trabalho realizou análises quantitativas sobre os acidentes e aplicou o Modelo *Apriori* para verificar relações entre tipos de acidente com suas causas. O trabalho concluí que o modelo *Apriori* obtém exito em identificar os fatores de contribuição dos acidentes ocorridos no trecho de rodovia estudada.

O trabalho de [Nogueira et al. 2018] analisou acidentes ocorridos em 2016, na BR-101, no trecho correspondente ao km 55 até km 90 norte, para encontrar trechos

[1] https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos

críticos utilizando o modelo *Apriori*. Devido ao filtro de tempo, BR e Km realizado no trabalho, o total de registros analisados foi de apenas 154 instâncias. Concluiu-se que os trechos críticos são compostos por segmentos contínuos de pista reta com traçado simples, possuem sinalização vertical e horizontal com faixa dupla contínua. Também verificou-se que os dias próximos e durante o final de semana (quinta, sexta, sábado e domingo) são os que têm maior ocorrência de acidentes nos locais identificados. O trabalho identificou também o ponto mais crítico da rodovia, o km "66 norte", com 17 acidentes.

A pesquisa realizada por [Campos Soares et al. 1] analisa a BR-101 por ter o mais alto volume de tráfego entre os anos de 2014 e 2016. Foram realizadas diversas análises quantitativas sobre os acidentes e aplica o modelo *Apriori* para verificar relações entre tipos de acidente com suas causas. Concluiu-se que as principais causas de acidente no trecho analisado são fatores humanos (falta de atenção, embriaguez, ultrapassagem indevida, etc.).

Em [Guerra 2019] utiliza-se dados entre os anos de 2011 e 2017 para treinamento do modelo e o ano de 2018 para teste. O trabalho analisou a correlação entre os indicadores econômicos com os acidentes. O artigo conclui que o IGP-M possui uma correlação de -88,40% com os acidentes e o IPCA -89,85%.

Em [de Oliveira Santos 2020] objetifica-se buscar padrões na base de dados da PRF nas rodovias do Rio Grande do Norte com dados de 2017 a 2019 utilizando o algoritmo *Apriori*. O trabalho finaliza apontando que, no município de Natal, entre o quilômetro 74 e 110 há maior probabilidade de haver acidentes, enquanto no município de Mossoró os quilômetros mais críticos estão entre o 37 e 73.

### 2.2. Classificação

No artigo de [de Oliveira et al. 2018] foram utilizados dados da PRF dos anos 2007 a 2015 e do ano de 2017. Aplicou-se o algoritmo *J48* e *RandomTree* para predizer se a causa de acidentes foi o condutor, chegando a um acerto na previsão de 82,7433% com algoritmo J48 e 82,8817% para o *Random Forest*.

Em [de Sousa Pereira Amorin 2019] foram classificados acidentes em graves e não-graves com a base da PRF entre os anos de 2007 e 2017, utilizando-se de vários algoritmos para encontrar o melhor classificador, com dados balanceados e desbalanceados. Também foi adicionado um atributo chamado *frequência*, calculada pela soma de acidentes que aconteceram em um quilômetro de uma rodovia, dividida pela quantidade de acidentes da base de dados, e notou-se que o mesmo não impactou tanto os classificadores. Com dados desbalanceados todas as entradas foram classificadas como Não-grave já que essa era a classe predominante. O melhor algoritmo com dados balanceados com o atributo frequência foi a Rede Neural que obteve 85% de acurácia e 87% de precisão. Enquanto o melhor algoritmo com dados balanceados e sem o atributo frequência foram o *Random-Forest + BernoulliNB* e *LogisticRegression + ExtraTrees-Classifier*. Ambas as combinações resultaram em 84,58% de acurácia e 88,14% de precisão.

No trabalho de [Vinícius Henrique de Mendonça 2020], foram utilizados dados das rodovias federais de Pernambuco, aplicando avaliações multicritérios para priorização trechos da rodovia em hierarquia de criticidade dos acidentes de trânsito. Foram criados filtros onde o usuário pode escolher os critérios de busca dos gráficos que contém as análises dos riscos de acidentalidade. O usuário adiciona seu perfil, suas preferências e

o aplicativo faz uma análise dos dados estáticos que possui, gerando uma lista de pontos onde o motorista deve ficar mais atento segundo os critérios definidos pelo motorista.

Em [Bonatto 2021] utilizou-se dados de acidentes do Rio Grande do Sul entre 2017 e 2019, totalizando 5 rodovias de pista simples com 1723 quilômetros de extensão. Buscou-se criar Modelos de Previsão de Acidentes a partir da abordagem criada nos Estados Unidos chamada *Highway Safety Manual* (HSM) - Manual de Segurança Viária. Concluí o autor que, apesar de o modelo o HSM ser o mais difundido, a sua aplicação direta e mesmo a calibração orientada pelo mesmo, utiliza um único fator de calibração que não resultaram em uma calibração satisfatória para todos os casos, logo seria melhor criar um modelo conforme os dados disponíveis em cada lugar a ser estudado.

### 2.3. Análise da Bibliografia

Nesta seção foram apresentados alguns trabalhos com temas relacionados, e utilizam a base de dados de acidentes disponibilizados pela Polícia Rodoviária Federal. Nos trabalhos que buscam correlações entre os dados, 5 dos 6 trabalhos analisados, utilizam o modelo Apriori para encontrar regras de associação entre os dados e a causa de acidentes. Já nos trabalhos que buscam classificar os dados, em 6 dos 11 trabalhos analisados, são aplicados em trechos específicos de uma ou mais rodovias. Em 2 trabalhos, criam-se ferramentas de análise estática que não fazem uso diretamente de Aprendizagem de Máquina. Por fim, em outros dois trabalhos, apesar de serem formatos de pesquisas mais próximas do proposto no presente trabalho, elas utilizam dados que caracterizam os acidentes como o "tipo de acidente" e "quantidade de mortos e feridos". Enquanto o presente trabalho se propõe a avaliar todos os trechos de rodovias disponibilizados nos dados abertos da Polícia Rodoviária Federal, sem restrições de regionalidade, em busca da correta classificação das categorias de acidentes utilizando dados que estão disponíveis antes do acidente ocorrer.

## 3. Desenvolvimento

O desenvolvimento deste trabalho consiste em alguns passos, separados em fases que seguem o processo de Descoberta de Conhecimento em Bases de Dados, descrito por [Fayyad et al. 1996]. O objetivo da pesquisa é aplicar técnicas para categorização de tipos de acidentes, utilizando informações disponibilizados pela PRF sobre as rodovias federais, utilizando apenas informações disponíveis antes de um acidente ocorrer. Também será avaliado se, com mais informações adicionadas na base de dados, os testes terão melhores resultados. No fim, espera-se que os modelos consigam classificar corretamente os tipos de acidentes em um dado trecho de rodovia e momento do dia.

#### 3.0.1. Conjunto de Dados

Para esta pesquisa foram utilizados dados de acidentes disponibilizados pela PRF[2]. No total foram utilizados 343.327 registros, os quais representam os acidentes ocorridos a partir de 2017, pois os dados anteriores têm um formato diferente, até 2021 em todo Brasil.

[2]https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos

De todo o conjunto de dados disponibilizado pela PRF serão utilizados apenas algumas das informações disponíveis, pois o objetivo é usar informações disponíveis antes da ocorrência de um acidente. Todas essas colunas do Quadro 1 serão utilizadas para criar uma catgoriação de um determinado acidente. Logo, a coluna com o valor objetivo é "tipo_acidente".

**Tabela 1. Dados utilizados**

| Variável | Descrição |
|---|---|
| data_inversa | Data da ocorrência |
| dia_semana | Dia da semana da ocorrência |
| br | Variável com valores numéricos, representando o identificador da BR do acidente |
| km | Identificação do quilômetro onde ocorreu o acidente |
| uf | Unidade da Federação |
| fase_dia | Fase do dia no momento do acidente |
| condicao_metereologica | Condição meteorológica no momento do acidente |
| tipo_acidente | Identificação do acidente |

Nem todos os dados disponibilizados serão úteis para a proposta e foram removidos do conjunto de dados original (conforme Quadro 2. Esses dados foram separados em 3 grupos: (i) dados inclusos no campo "fase_dia"; (ii) dados obtidos após a ocorrência do acidente e (iii) dados que necessitam de agentes externos para serem obtidos.

### 3.1. Pré-processamento

O primeiro item adicionado foram os dados de feriados. Para procurar uma possível relação entre feriados e os tipos de acidentes ocorridos foram usados apenas os feriados nacionais, utilizando a informação publicada no Diário Oficial do Distrito Federal entre 2017 e 2021[3].

Outra informação buscada e padronizada para inserção nos dados existentes foram os dados de quais rodovias eram pedagiadas. Para encontrar tais subsídios, foi realizada uma consulta no site da Associação Brasileira de Concessionárias de Rodovias [4]. Com os dados extraídos, ainda fora necessário realizar uma filtragem, já que a tabela trazia a mesma rodovia várias vezes em linhas diferentes, e padronização, colocando os dados nos mesmo formatos dos dados da planilha da PRF, para então inserir os novos dados no conjunto.

Os últimos dados adicionados foram referentes aos pontos de radares nas rodovias. Para encontrar tais informações fora consultado o Portal de Multas de Trânsito [5] disponibilizado pelo Departamento Nacional de Infraestrutura de Transportes — DNIT [6].

Outra etapa do pré-processamento realizado foi o tratamento de ruído dos dados. Nesta etapa foi analisado a existência de erros de preenchimento e vários registros da

[3]https://www.dodf.df.gov.br/
[4]https://melhoresrodovias.org.br/
[5]https://servicos.dnit.gov.br/multas/informacoes/equipamentos-fiscalizacao
[6]https://www.gov.br/dnit/pt-br

**Tabela 2. Dados não utilizados**

| Grupo | Variável | Descrição |
|---|---|---|
| i | horário | Horário da ocorrência |
| ii | municipio | Nome do município de ocorrência do acidente |
| ii | latitude | Latitude do local do acidente em formato geodésico decimal |
| ii | longitude | Longitude do local do acidente em formato geodésico decimal |
| ii | tipo_pista | Tipo da pista considerando a quantidade de faixas |
| iii | causa_acidente | Identificação da causa principal do acidente |
| iii | classificação_acidente | Classificação quanto à gravidade do acidente |
| iii | uso_solo | Descrição sobre as características do local do acidente |
| iii | sentido_via | Sentido da via considerando o ponto de colisão |
| iii | pessoas | Total de pessoas envolvidas na ocorrência |
| iii | mortos | Total de pessoas mortas envolvidas na ocorrência |
| iii | feridos_leves | Total de pessoas com ferimentos leves envolvidas na ocorrência |
| iii | feridos_graves | Total de pessoas com ferimentos graves envolvidas na ocorrência |
| iii | ilesos | Total de pessoas ilesas envolvidas na ocorrência |
| iii | feridos | Total de pessoas feridas envolvidas na ocorrência |
| iii | ignorados | Total de pessoas envolvidas na ocorrência e que não se soube o estado físico |
| iii | veículos | Total de veículos envolvidos na ocorrência |

mesma informação com valores diferentes. Nestes quesitos, observou-se que os dados não possuem erros de preenchimentos ou múltiplos tipos de dados na mesma coluna.

Foram removidas as linhas que não tinham o "km" e "br" preenchidas (totalizando 659 registros). Dessa forma, o conjunto de dados ficou com 342.668 linhas. Foram convertidos os dados categóricos em numéricos, utilizando a biblioteca "LabelEncoder" [7]. Enquanto para padronização da escala numérica, foi utilizada a biblioteca "StandardScaler" [8].

### 3.2. Mineração de Dados

O trabalho visa realizar classificações de um tipo de acidente. Foram utilizados os algoritmos *DecisionTree* [9], *RandomForest* [10] e *K Nearest Neighbor* [11].

[7]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
[8]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
[9]https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[10]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[11]https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

| Variável | Tipo de valores | Descrição |
|---|---|---|
| feriados | [0, 1, 2, 3] | 0 - Não feriado;<br>1 - Feriado;<br>2 - Véspera;<br>3 - Pós-feriado |
| pedagios | Numérico | Quantidade<br>de pedágios<br>na rodovia |
| radares | Numérico | Quantidade<br>de radares<br>no par br-km |

**Tabela 3. Colunas adicionadas**

## 4. Resultados e Discussão

### 4.1. Cenários dos Experimentos

Para correta sistematização do processo de execução dos algoritmos, foram levantados 32 casos de testes para os quais o conjunto de dados fora submetido. Os casos de testes executados estão representados nas tabelas 4 e 5.

Em relação aos parâmetros utilizados nos algoritmos, com o *DecisionTree* não será utilizado nenhuma alteração de parâmetros, enquanto com *RandomForest* será utilizado o parâmetro *n_estimators* com o valor 100, e com o algoritmo *K Nearest Neighbor* será realizado o teste com o parâmetro *n_neighbors* entre 1 e 20.

Foram utilizados algoritmos de balanceamento para testar qual teria o melhor impacto na base. O primeiro a ser testado fora o *RandomUnderSampler*, que nivela, de forma aleatória, as categorias com base na categoria com menos registros. Por exemplo, se tem uma categoria com 10 mil instancias e outra com apenas 100, a categoria com 10 mil instâncias também ficará com apenas 100. O Segundo algoritmo fora o *NearMiss* que tem um comportamento similar ao primeiro, com a diferença que utiliza heurísticas para determinar qual instâncias irá remover. O último algoritmo utilizado fora *OneSidedSelection*, que também diminuí a quantidade de registros, contudo não reduz ao mesmo número de registros da menor categoria. O algoritmo utiliza heurísticas para remover outliers e instancias muito parecidas.

### 4.2. Resultados Obtidos

O resultado dos testes é mostrado nas tabelas 4 e 5. Na tabela 4 os resultados dos testes que utilizaram validação cruzada são apresentados, enquanto na tabela 5 os resultados dos testes que utilizaram a separação dos dados entre treino e teste são demonstrados na mesma ordem da tabela anterior. Cada linha, em ambas as tabelas, corresponde ao mesmo caso de teste.

A primeira coluna da tabela (*) é o número do teste. A segunda coluna (Cat. $>$10k) indica se, naquele teste, foi utilizado apenas os tipos de acidentes que tinham mais de 10 mil registros (Atropelamento de Animal; Incêndio; Engavetamento; Colisão lateral mesmo sentido; Colisão com objeto; Colisão com objeto em movimento; Colisão lateral

sentido oposto; Derramamento de carga; Danos eventuais; Eventos atípicos). A terceira coluna (Colunas Adicionadas) indica se foram utilizadas as colunas adicionadas nesse trabalho (radares, pedagios e feriado). Enquanto às três ultimas colunas (DT; N & KNN; e RF), *DecisionTree*, valor de *N* e *KNN* e *RandomForest*, nessa ordem, indicam a *accuracy* obtidos pelo algoritmo no cenário descrito pelo restante das colunas.

| * | Cat. >10k | Colunas Adicionadas | Balanceamento | DT | N & KNN | RF |
|---|---|---|---|---|---|---|
| 1 | Não | Sim | Não | 0.18 | 19 - 0.26 | 0.22 |
| 2 | Não | Não | Não | 0.18 | 19 - 0.25 | 0.22 |
| 3 | Sim | Sim | Não | 0.20 | 19 - 0.28 | 0.25 |
| 4 | Sim | Não | Não | 0.20 | 19 - 0.28 | 0.25 |
| 5 | Não | Sim | RandomUnderSampler | 0.09 | 7 - 0.09 | 0.12 |
| 6 | Não | Não | RandomUnderSampler | 0.09 | 13 - 0.09 | 0.13 |
| 7 | Sim | Sim | RandomUnderSampler | 0.16 | 19 - 0.21 | 0.25 |
| 8 | Sim | Não | RandomUnderSampler | 0.16 | 19 - 0.20 | 0.19 |
| 9 | Não | Sim | NearMiss | 0.21 | 11 - 0.23 | 0.26 |
| 10 | Não | Não | NearMiss | 0.20 | 5 - 0.23 | 0.25 |
| 11 | Sim | Sim | NearMiss | 0.21 | 19 - 0.27 | 0.25 |
| 12 | Sim | Não | NearMiss | 0.21 | 19 - 0.26 | 0.25 |
| 13 | Não | Sim | OneSidedSelection | 0.23 | **19 - 0.31** | 0.29 |
| 14 | Não | Não | OneSidedSelection | 0.23 | **1 - 0.30** | 0.29 |
| 15 | Sim | Sim | OneSidedSelection | 0.24 | **1 - 0.32** | **0.31** |
| 16 | Sim | Não | OneSidedSelection | 0.24 | **1 - 0.32** | **0.30** |

**Tabela 4. Testes realizados utilizando separação de dados entre treino e teste**

Na maioria dos testes utilizando separação dos dados entre treino e teste, os valores para *n_neighbor*, em relação aos valores testados (1 a 20), foi alto, normalmente entre 11 e 19. Enquanto, nos testes utilizando validação cruzada, esse valor normalmente ficou entre 1 e 5. No primeiro grupo, testes que utilizaram validação cruzada, os resultados deram em torno de 10 pontos a mais do que os testes do segundo grupo. Os testes utilizando balanceamento deram um resultado melhor no geral em relação aos dados não balanceados.

Os cenários que obtiveram os melhores resultados foram as linhas 15 e 16 da tabela 4. Ambos os cenários utilizaram apenas os tipos de acidentes mais recorrentes, utilizaram o algoritmo de balanceamento *OneSidedSelection* e obtiveram uma *accuracy* de 32 com o algoritmo KNN e com o valor de *n* sendo 1. A única diferença entre os dois casos, foi que o cenário 15 utilizou as colunas adicionadas pelo trabalho, enquanto o teste 16 não continha essas colunas.

## 5. Considerações Finais

O presente estudo teve como objetivo classificar tipos de acidentes, utilizando informações disponibilizadas pela PRF sobre as rodovias federais, fazendo uso apenas de dados disponíveis antes de um acidente ocorrer. Também foi avaliado se, com mais informações adicionadas na base de dados, os testes teriam melhores resultados. Para isso, foi utilizada uma metodologia baseada em técnicas de aprendizado de máquina.

| * | Cat. >10k | Colunas Adicionadas | Balanceamento | DT | N & KNN | RF |
|---|---|---|---|---|---|---|
| 1 | Não | Sim | Não | 0.10 | 3 - 0.20 | 0.11 |
| 2 | Não | Não | Não | 0.07 | 3 - 0.16 | 0.08 |
| 3 | Sim | Sim | Não | 0.15 | 1 - 0.17 | 0.17 |
| 4 | Sim | Não | Não | 0.10 | 3 - 0.19 | 0.11 |
| 5 | Não | Sim | RandomUnderSampler | 0.09 | 5 - 0.07 | 0.13 |
| 6 | Não | Não | RandomUnderSampler | 0.10 | 5 - 0.07 | 0.13 |
| 7 | Sim | Sim | RandomUnderSampler | 0.15 | 1 - 0.17 | 0.17 |
| 8 | Sim | Não | RandomUnderSampler | 0.15 | 1 - 0.17 | 0.17 |
| 9 | Não | Sim | NearMiss | 0.18 | 1 - 0.15 | **0.23** |
| 10 | Não | Não | NearMiss | 0.16 | 1 - 0.16 | 0.21 |
| 11 | Sim | Sim | NearMiss | 0.16 | 1 - 0.16 | 0.19 |
| 12 | Sim | Não | NearMiss | 0.16 | 1 - 0.15 | 0.19 |
| 13 | Não | Sim | OneSidedSelection | 0.10 | **3 - 0.23** | 0.12 |
| 14 | Não | Não | OneSidedSelection | 0.10 | 3 - 0.22 | 0.12 |
| 15 | Sim | Sim | OneSidedSelection | 0.10 | **3 - 0.25** | 0.13 |
| 16 | Sim | Não | OneSidedSelection | 0.10 | **3 - 0.24** | 0.13 |

**Tabela 5. Testes realizados utilizando validação cruzada**

A taxa geral de acurácia foi baixa, fazendo com que os resultados obtidos não sejam tão satisfatórios. Isso pode ser explicado por uma série de fatores, como a falta de dados suficientes e a complexidade dos acidentes. Também notou-se que as colunas adicionadas no conjunto de dados representam uma quantidade muito pequena do total de linhas. No caso dos feriados, essa quantidade é de aproximadamente 10%, enquanto no pedágio são cerca de 1,4% e nos radares menos de 3%. Apesar do resultado não ter sido o esperado, o estudo contribuiu para a sistematização da classificação de acidentes. Para trabalhos futuros, sugere-se a utilização de um conjunto de dados maior e mais diversificado, a inclusão de mais informações sobre as rodovias, bem como a exploração de técnicas de aprendizado de máquinas mais recentes.

## Referências


Bonatto, A. Z. E. (2021). Comparação entre a transferência e o desenvolvimento de modelos de previsão de acidentes para o contexto brasileiro. *Universidade Federal do Rio Grande do Sul*, page 96.

Campos Soares, L., do Prado, H. A., Balaniuk, R., Ferneda, E., and De Bortoli, A. (1). Caracterização de acidentes rodoviários e as açoes governamentais para a redução de mortes e lesões no trânsito. *Revista Transporte y Territorio*, 0(19):182–220.

Costa, J., Bernardini, F., and Filho, J. V. (2014). A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. *AtoZ: novas práticas em informação e conhecimento*, 3(2):139–157.

da Saúde OPS, O. P.-A. (2018). Salvar vidas – pacote de medidas técnicas para a segurança no trânsito. Technical report, Organização Mundial da Saúde (OMS).

de Oliveira, M. P., Ines, J. M., da Silva Lopes, A., Castro, S. H. G., and Ferreira, W. M. (2018). Uso de mineraÇÃo de dados e tecnologia preditiva na prevenÇÃo de acidentes de trÂnsito no brasil. *Anais do XIII SIMMEC 2018 - Simpósio de Mecânica Computacional*, page 15.

de Oliveira Santos, I. J. (2020). Mineração de dados em padrões de acidentes de trânsito: o uso de dados abertos da polícia rodoviária federal no rn. *Universidade Federal do Rio Grande do Norte*, page 58.

de Sousa Pereira Amorin, B. (2019). Uso de aprendizado de máquina para classificação de risco de acidentes em rodovias. *Universidade Federal de Campina Grande*, page 106.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

Fernandes, F. T. and Chiavegatto, A. D. P. (2019). Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. *Revista Brasileira de Saúde Ocupacional [online]*, 44.

Guerra, C. I. R. (2019). Uso de data analytics para avaliar a influência econômica e social em acidentes graves de trânsito nas rodovias federais e fatores de redução de ocorrências no norte fluminense. *RDBU— Repositório Digital da Biblioteca da Unisinos*, page 38.

IPEA, P. (2015). Acidentes de trânsito nas rodovias federais brasileiras caracterização, tendências e custos para a sociedade. Technical report, Instituto de Pesquisa Econômica Aplicada (IPEA).

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., and Ramnath, R. (2019). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, page 33–42, New York, NY, USA. Association for Computing Machinery.

Nogueira, F. d. S., Lee, L., and Rissino, S. d. D. (2018). Descoberta de conhecimento na base de dados aberta da polÍcia rodoviÁria federal: IdentificaÇÃo de pontos crÍticos na rodovia br 101 no municÍpio de sÃo mateus/es. *Brazilian Journal of Production Engineering*, 4(4):70–90.

Silva, J. T. M., Maia, L. C. G., and Reis, C. V. R. (2015). O uso da descoberta de conhecimento em banco de dados nos acidentes da br-381. *XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB)*, page 22.

Vinícius Henrique de Mendonça, T. V. G. (2020). Sistema unificado de consulta e análise da acidentalidade em rodovias federais de pernambuco (sucaarf-pe). *Revista Eletrônica de Gestão Organizacional*, pages 184–197.