

UNIVERSIDADE FEDERAL DO PARANÁ

NICOLAS GREGORI DE MORAES RODRIGUES

MODELO PREDITIVO PARA SINISTROS DE TRÂNSITO COM VÍTIMAS FATAIS
NA BR-116: UMA ANÁLISE DE DADOS ABERTOS DA POLÍCIA RODoviÁRIA
FEDERAL

CURITIBA

2024

NICOLAS GREGORI DE MORAES RODRIGUES

MODELO PREDITIVO PARA SINISTROS DE TRÂNSITO COM VÍTIMAS FATAIS
NA BR-116: UMA ANÁLISE DE DADOS ABERTOS DA POLÍCIA RODOVIÁRIA
FEDERAL

Trabalho de conclusão de curso apresentado ao curso de Graduação em Gestão da Informação, Departamento de Ciência e Gestão da Informação do Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Gestão da Informação.

Orientadora: Profa. Dra. Denise Fukumi Tsunoda

CURITIBA

2024

AGRADECIMENTOS

Aos meus amigos de trilha, Daniel, Leonardo Marques, Leonardo Mahmoud e Marcos, por cada aventura e perrengue que passamos juntos.

Às minhas amigas da faculdade, Helena e Luana, pelo apoio acadêmico e por deixarem esse caminho mais leve.

Aos meus professores, que com paciência e dedicação compartilharam seu conhecimento e me inspiraram a ir além.

Aos meus familiares. Meus pais, Kelen e Victor, por seu amor incondicional, ensinamentos e pelo apoio que nunca me faltou. A minha irmã Nicole, nem sempre nos damos bem, mas unidos vencemos qualquer um.

E, principalmente, à minha namorada Ana, que esteve comigo em cada etapa, compartilhando alegrias, desafios e conquistas e que me motivou a dar o primeiro passo nesta jornada. Seu amor é o que dá cor e significado aos meus dias.

RESUMO

Este trabalho de conclusão de curso tem como objetivo analisar dados da base da Polícia Rodoviária Federal, abrangendo os anos de 2020 a 2024, utilizando o algoritmo *Random Forest* para identificar os principais fatores que contribuem para a ocorrência de acidentes com vítimas fatais na rodovia BR-116. A relevância desta rodovia deve-se ao seu extenso trajeto, tem origem no Ceará e termina na fronteira do Rio Grande do Sul com o Uruguai e interliga importantes polos econômicos, além de estar entre as rodovias com maior número de acidentes e fatalidades no Brasil. O estudo envolveu o treinamento de modelos de aprendizado de máquina para extração de conhecimento a partir dos dados. As etapas de pesquisa realizadas incluíram limpeza e normalização dos dados, análise estatística descritiva, definição do método, ajuste de hiperparâmetros e aplicação do processo de descoberta de conhecimento em bases de dados (KDD). A pesquisa, de natureza aplicada, abordagem quantitativa e caráter descritivo, utilizou o *software Weka* para a etapa de mineração de dados. Foram selecionados atributos relevantes ao problema de pesquisa durante o pré-processamento, seguidos pela transformação dos dados para aplicação do algoritmo *Random Forest*. Os hiperparâmetros foram ajustados visando aprimorar o desempenho do modelo, medido pelo coeficiente de correlação na base de testes. Ao final, a análise permitiu identificar os fatores que mais influenciam a ocorrência de vítimas fatais nos acidentes, destacando o fator humano como principal determinante. O estudo demonstrou o potencial do aprendizado de máquina para auxiliar na compreensão e prevenção de acidentes em rodovias federais.

Palavras-chave: Sinistro de trânsito, mineração de dados, dados abertos

Abstract

The reasearch aims to analyze data from the Federal Highway Police database, covering the years 2020 to 2024, using the Random Forest algorithm to identify the main factors contributing to fatal accidents on the BR-116 highway. The significance of this highway lies in its extensive route, originating in Ceará and ending at the border of Rio Grande do Sul with Uruguay, connecting important economic hubs. It is also among the highways with the highest number of accidents and fatalities in Brazil. The study involved training machine learning models to extract knowledge from the data. The research steps included data cleaning and normalization, descriptive statistical analysis, method definition, hyperparameter tuning, and applying the Knowledge Discovery in Databases (KDD) process. This applied, quantitative, and descriptive research used the Weka software for the data mining stage. Relevant attributes to the research problem were selected during preprocessing, followed by data transformation for the application of the Random Forest algorithm. Hyperparameters were fine-tuned to improve the model's performance, measured by the correlation coefficient on the test dataset. In the end, the analysis identified the factors that most influence the occurrence of fatal accidents, highlighting the human factor as the main determinant. The study demonstrated the potential of machine learning to aid in understanding and preventing accidents on federal highways..

Keywords: Traffic accident, data mining, open data

LISTA DE GRÁFICOS

Gráfico 1 - Distribuição de sinistros entre os meses	32
Gráfico 2 - Distribuição de sinistros entre os dias da semana.....	33
Gráfico 3 - Distribuição de sinistros entre as horas do dia	33
Gráfico 4 - Taxa de acidentes fatais entre as horas do dia	34

LISTA DE QUADROS

Quadro 1 - Descrição dos atributos	22
Quadro 2 – Levantamento Bibliográfico	27
Quadro 3 – Atributos quanto as características do sinistro	35

LISTA DE TABELAS

Tabela 1 - Características das Bases de Dados	23
Tabela 2 - Características das Bases de Dados Filtrada para BR-116	23
Tabela 3 – Atributos alterados.....	30
Tabela 4 – Atributos referentes ao local de sinistro.....	34
Tabela 5 – Atributo fase_dia.....	36
Tabela 6 – Atributo sentido_via	37
Tabela 7 – Atributo condicao_metereologica	37
Tabela 8 – Atributo tipo_pista.....	38
Tabela 9 – Atributo traçado da via.....	38
Tabela 10 – Testes <i>numIterations</i> com <i>bagSizePercent</i>	39
Tabela 11 – Testes <i>numIterations</i> dobrado	40
Tabela 12 – Testes <i>numIterations</i> com <i>bagSizePercent</i> e <i>numFeatures</i>	40
Tabela 13 – Testes <i>seed</i>	41

LISTA DE ABREVIATURAS E SIGLAS

ABTLP	Associação Brasileira de Transporte e Logística de Produtos Perigosos
Conj.	Conjunto
KDD	<i>Knowledge Discovery in Databases</i>
NBR	Normativa Brasileira
ODS	Objetivos de Desenvolvimento Sustentável
PNATRANS	Plano Nacional de Redução de Mortes e Lesões no Trânsito
PRF	Polícia Rodoviária Federal
PVS-ANTT	Programa Vias Seguras

SUMÁRIO

1 INTRODUÇÃO	10
1.1 PROBLEMATIZAÇÃO	10
1.2 OBJETIVOS	11
1.2.1 Objetivo Geral	11
1.2.2 Objetivos Específicos	11
1.3 JUSTIFICATIVA	12
1.3.1 Justificativa Social	12
1.3.2 Justificativa Econômica	12
1.3.3 Justificativa para o Curso de Gestão da Informação	13
1.3.4 Justificativa Pessoal	13
1.4 ESTRUTURA DO DOCUMENTO	13
2 FUNDAMENTAÇÃO TEÓRICA	15
2.1 SEGURANÇA VIÁRIA E ACIDENTES DE TRÂNSITO	15
2.2 GESTÃO DA INFORMAÇÃO, ANÁLISE DE DADOS E MODELOS PREDITIVOS	
16	
2.3 FATORES HUMANOS E COMPORTAMENTAIS	18
2.4 INFRAESTRUTURA E CONDIÇÕES DA VIA	18
3 ENCAMINHAMENTO METODOLÓGICO	20
3.1 MATERIAIS	21
3.1.1 Bases de Dados	21
3.1.2 Descrição do Algoritmo	24
3.2 TRABALHOS RELACIONADOS	27
4 RESULTADOS E ANÁLISES	29
4.1 ETAPA DE NORMALIZAÇÃO E ANÁLISES DESCRITIVAS	30
4.2 MODELO PREDITIVO	39
5 CONSIDERAÇÕES FINAIS	43
REFERÊNCIAS	45

1 INTRODUÇÃO

Segundo Pinheiro, Silva e Salomão (2018) a BR-116 é a mais extensa e uma das mais importantes rodovias do Brasil, interligando cidades e atravessando dez estados brasileiros. Ela se estende desde o município de Fortaleza no Ceará até Jaguarão, situado na divisa do Rio Grande do Sul com o Uruguai.

A rodovia possui um dos maiores índices de acidentes em estradas brasileiras. Como demonstrado a lista disponibilizada por Moreira (2024), onde três dos dez trechos com mais acidentes registrados pertencem a esta via. Desta forma a análise dos padrões existentes em acidentes na BR-116 é fundamental para entender os principais fatores que influenciam nos eventos, além de poder embasar políticas de segurança e prevenção.

Neste contexto, é possível utilizar a gestão da informação para analisar os dados abertos publicados pelo governo. A *Knowledge Discovery in Databases* (KDD) serve como uma maneira de compreender os dados e prever as condições mais propícias a fatalidades. Seleção, pré-processamento, transformação, mineração, avaliação e interpretação dos dados servem para transformar grandes volumes de dados brutos em conhecimentos úteis para a tomada de decisão.

No estudo, foram utilizados dados do ano de 2020 até 2024, para investigar os padrões e identificar fatores associados aos acidentes. A análise dos dados abertos possibilita uma visão das características dos acidentes, como horários de maior incidência, condições da pista e dados do clima. Desta forma é apresentado uma compreensão dos acidentes na BR-116 tendo como principal insumo a base de boletins de acidentes de trânsito da Polícia Rodoviária Federal (PRF).

1.1 PROBLEMATIZAÇÃO

Segundo os estudos apresentados na matéria da Associação Brasileira de Transporte e Logística de Produtos Perigosos as Rodovias (ABTLP, 2020), a BR-101 e a BR-116 representam o primeiro e segundo lugar na lista das mais perigosas do Brasil. Em razão de concentrarem os maiores números de acidentes no país. Juntas somam um total de 25,3% dos acidentes com vítimas ocorridos nas estradas nacionais durante ano de 2020.

Analisando os dados da PRF (2024) no ano de 2023 os números de acidentes com feridos sobem para 34,1%, sendo um aumento de aproximadamente 9,8% no número destes eventos. Os valores podem estar inflados por conta do recorte temporal dos estudos. Como divulgado em 2024 pelo Instituto Brasileiro de Geografia e Estatística (IBGE) houve um aumento de 71,5% na quantidade de viagens do ano de 2021 até 2023.

A motivação do estudo na BR-116, se deu em razão da alta quantidade de fatalidades registradas em 2023. Ao criar uma lista ordenada de forma decrescente a rodovia estudada aparece em primeiro lugar dentre as que possuem maior número de vítimas fatais por número de pessoas envolvidas nos acidentes. Como as estradas citadas anteriormente contemplam o primeiro requisito, onde foram registrados os maiores números de acidentes no país, a BR-116 possui maior taxa de mortos por envolvidos.

Com isso é possível estabelecer o questionamento base do estudo - na seguinte pergunta de pesquisa - Quais fatores podem ser identificados nos dados disponibilizados pela Polícia Rodoviária Federal como influentes nos sinistros da BR-116, que possam ser usados para identificar maior chance de vítimas fatais?

1.2 OBJETIVOS

Os objetivos são separados em objetivo geral e objetivos específicos.

1.2.1 Objetivo Geral

Treinar um modelo de aprendizado de máquina para identificar os atributos mais relevantes na ocorrência de vítimas fatais nos sinistros ocorridos na BR-116, com a base de dados abertos disponibilizados pela Polícia Rodoviária Federal de 2020 até 2023.

1.2.2 Objetivos Específicos

- Avaliar os dados para o mapeamento de fatores externos que influenciam em acidentes fatais na rodovia;

- Analisar a distribuição de acidentes fatais na BR-116 com ênfase em fatores alteráveis para a redução das mortes, como infraestrutura, sinalização ou educação dos motoristas;
- Descrever o funcionamento do método de aprendizado de máquina *Random Forest*;
- Treinar o modelo de aprendizado de máquina com alteração de hiperparâmetros para o melhorar a predição nos dados da BR-116.

1.3 JUSTIFICATIVA

As justificativas são separadas nos aspectos sociais, econômicos, para gestão da informação e pessoais.

1.3.1 Justificativa Social

A análise dos padrões de acidentes na BR-116 tem um impacto direto na segurança dos usuários da rodovia e na sociedade. O estudo feito por Lobato (2018 p. 29) demonstra um aumento de trinta mil ocorrências nos trechos de rodovias apresentados entre 2008 e 2013. Estes acidentes afetam a população brasileira, como demonstra a redação do Jornal da USP (2018) por meio de entrevistas com professores da área saúde.

Esse trabalho permite identificar fatores críticos que podem subsidiar campanhas educativas e medidas preventivas, contribuindo para a redução de acidentes e, por consequência, para a proteção da vida e da integridade física das pessoas.

1.3.2 Justificativa Econômica

A BR-116 é uma via essencial para o transporte de bens e pessoas. Segundo a página Juntos no Caminho esta é a rodovia mais influente no transporte de cargas, ligando o Sul ao Norte do país. Já o Portal Ipiranga destaca a importância da rodovia na ligação entre São Paulo e o Rio de Janeiro, dois polos econômicos ativos segundo o portal.

Pelo informativo da ABTLP (2020) os custos de acidentes em rodovias federais chegaram a R\$ 10,22 bilhões no ano de 2020, sendo os estados mais afetados é Minas Gerais (com R\$ 1,36 bilhões), Paraná (R\$ 1,10 bilhões) e Santa Catarina (R\$ 1,03 bilhões).

A análise deste trabalho aprofundada nos conhecimentos fornecidos pode auxiliar na criação de estratégias de prevenção, reduzindo esses custos. Melhorias em segurança podem aumentar a eficiência na circulação de pessoas e mercadorias, beneficiando a economia nacional.

1.3.3 Justificativa para o Curso de Gestão da Informação

Este trabalho é uma aplicação prática dos princípios da Gestão da Informação, pois envolve a coleta, tratamento, análise e interpretação de dados abertos. Além disso, demonstra a importância do profissional de Gestão da Informação na criação de interpretações relevantes partindo de dados. Assim como, exemplifica que os conhecimentos adquiridos no curso podem ser aplicados para resolver problemas reais.

1.3.4 Justificativa Pessoal

A realização deste estudo é uma oportunidade de desenvolver habilidades práticas em análise de dados, utilização de algoritmos de aprendizado de máquina e interpretação de resultados. O trabalho permite desenvolver uma compreensão mais ampla sobre segurança viária e políticas de prevenção de acidentes, temas de grande relevância. A experiência adquirida com este projeto proporciona preparo para atuação em áreas estratégicas na análise e gestão de dados.

1.4 ESTRUTURA DO DOCUMENTO

O documento inicia-se com uma introdução, onde é apresentado um resumo da rodovia estudada, a problematização do trabalho focada nos acidentes com vítimas fatais, o objetivo geral e específicos do estudo e as justificativas para a realização dele, divididas entre social, econômico, para o curso de Graduação em Gestão da Informação e pessoais.

Na fundamentação teórica são apresentados conceitos quanto a segurança viária e acidentes de trânsito, gestão da informação, análise de dados e modelos preditivos, fatores humanos e comportamentais e infraestrutura e condições da via.

O encaminhamento metodológico apresenta as delimitações tomadas para realização do trabalho, listando os materiais (bases de dados e algoritmo) junto da pesquisa de trabalhos relacionados.

A normalização e aplicação do algoritmo são etapas cruciais para a realização do estudo. Foi realizado os seguintes passos: normalização da base de dados unificada, estatística descritiva dos atributos presentes na base e, por fim, a aplicação do algoritmo buscando correlações fortes.

Nas considerações finais está apresentado os conhecimentos adquiridos com a análise das bases de dados, junto da análise do algoritmo, para confrontar com os objetivos, apresentados na introdução.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentado alguns dos principais conceitos para fundamentar a pesquisa.

2.1 SEGURANÇA VIÁRIA E ACIDENTES DE TRÂNSITO

A segurança viária é tida pelo Ministério do Transporte (2020) como “conjunto de métodos, ações e normas existentes necessários para a circulação segura de pessoas e veículos nas ruas e rodovias, com a finalidade de prevenir e reduzir o risco de acidentes”. O conceito é complementado por Loureiro e Nodari (2001) na construção de uma rodovia o engenheiro deve considerar os fatores que preveem a ocorrência de acidentes. Os autores relatam que no período entre 1960 e 1970 durante a construção das rodovias se buscava mitigar os danos das colisões, mas atualmente os esforços se voltam para a prevenção destes eventos.

Para Cunto e Loureiro os acidentes são eventos acarretados por uma série de fatores, não podendo ser identificados em uma situação do cotidiano. A Norma Brasileira (NBR) 10.697 (2018, p1) dispensa o termo “acidente de trânsito”, utilizando no lugar deste o termo “sinistro de trânsito”. O sinistro é descrito como “todo evento que resulte em danos materiais ou corporais, prejuízo ao trânsito, às vias ou ao meio ambiente em que pelo menos uma das partes está em movimento em vias terrestres ou áreas abertas ao público”. Posteriormente estes eventos são classificados de acordo com os dados corporais, divididos como “sinistro de trânsito sem vítima”, “sinistro de trânsito com vítima não fatal” e “sinistro de trânsito com vítima fatal”.

Conforme o estudo feito por Carvalho (2020) os sinistros chegam a custar R\$ 40 bilhões por ano no Brasil por conta do acidente e aproximadamente dez bilhões em produtos danificados. Fora as quarenta e cinco mil pessoas que perdem suas vidas por ano no Brasil em sinistros de trânsito com vítimas fatais.

Existem várias frentes de ações para a redução destes eventos, até mesmo nos Objetivos de Desenvolvimento Sustentável (ODS) e em projetos governamentais. No entanto, não é apresentado uma definição de rodovias seguras, desta maneira não é possível afirmar quais características são necessárias para definir uma estrada como segura. Ademais, segundo as Nações Unidas a segurança é quantificada por

meio de amostras temporais e redução dos riscos, como na meta da ODS número 3.6 - “reduzir na metade as mortes em acidentes de trânsito em todo o mundo”.

No Brasil temos o Programa Vias Seguras (PVS-ANTT) é pautado nas 12 Metas Globais de Desempenho para a Segurança no Trânsito focadas nos fatores que influenciam em acidentes e na letalidade, junto das medidas que os países podem adotar para reduzir estes dados.

O PVS-ANTT (2023) trata-se de um programa que liga as Metas Globais de Desempenho para a Segurança no Trânsito com o quadro dos seis pilares do Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS). Os pilares do PVS-ANTT são – melhor da gestão das rodovias, veículos mais, usuários mais seguros e uma melhor resposta a acidentes. Os dois pilares restantes do PNATRANS são - vias mais seguras, normalização e fiscalização do tráfego. Estes pilares classificam as 154 ações do programa em vias federais, estaduais e municipais. No caso de emprego da classificação mencionada, este trabalho estaria pautado nos pilares de usuários mais seguros e vias mais seguras, em razão dos dados utilizados caracterizarem estas duas entidades.

2.2 GESTÃO DA INFORMAÇÃO, ANÁLISE DE DADOS E MODELOS PREDITIVOS

A gestão da informação é a aplicação de princípios administrativos à aquisição, organização controle, disseminação e uso da informação para a operacionalização efetiva de organizações de todos os tipos (WILSON, 1997, apud TARAPANOFF, 2006, p. 21). A utilização de métodos estatísticos, condensação das informações para apresentá-las e embasar uma argumentação em informações valorosas são algumas tarefas atribuídas para o gestor da informação. Toda essa cadeia de ações é apresentada neste documento. Desde a coleta de dados em fontes confiáveis, o tratamento destes e posterior uso do método de predição para ser compreendido e gerar conhecimento.

A análise de dados aplicada à segurança viária é uma abordagem essencial para conhecer, monitorar e mitigar os fatores que contribuem para a prevenção de sinistros de trânsito e suas consequências. Essa utiliza técnicas estatísticas, modelagem preditiva e aprendizado de máquina para processar grandes volumes de dados, como registros de acidentes, buscando padrões de tráfego, condições

meteorológicas e características da infraestrutura viária que contribuam para certos eventos.

Para isto é aplicada a *Knowledge Discovery in Databases* (KDD) sendo possível identificar tendências, mapear áreas de maior risco e desenvolver intervenções eficazes, como campanhas de conscientização, melhoramento de vias urbanas e rurais e políticas públicas.

De acordo com Souza (2023) o KDD possui como etapas a seleção dos dados, pré-processamento, transformação dos dados, mineração de dados, avaliação e interpretação e utilização do conhecimento descoberto. Ao transformar dados brutos em conhecimento e registrando a informação, a análise de dados potencializa a tomada de decisões baseadas em evidências, promovendo um trânsito mais seguro e sustentável para todos os usuários.

Os modelos preditivos utilizam técnicas estatísticas e computacionais, como regressões, árvores de decisão e algoritmos de aprendizado de máquina. As regressões, como a linear e a logística, são amplamente usadas para identificar relações entre variáveis, como as condições climáticas na ocorrência de acidentes. De acordo com a IBM estes modelos se dividem em três categorias pelo aprendizado supervisionado, não supervisionado, semissupervisionado ou de reforço.

As árvores de decisão oferecem uma abordagem de categorização dos dados para a tomada de decisões baseadas em critérios hierárquicos. O método possibilita a utilização de grandes e mais complexos conjuntos de dados, ajustando-se automaticamente para melhorar a precisão das previsões. Quando aplicado, ajuda a identificar fatores de risco e prever tendências nos acidentes. Neste trabalho foi utilizado o *Random Forest*, sendo um algoritmo do tipo supervisionado.

2.3 FATORES HUMANOS E COMPORTAMENTAIS

Na NBR 10697 (2018, p. 6), o fator humano é definido como o “comportamento do homem como condutor, passageiro ou pedestre, responsável pela ocorrência do sinistro”. Esse elemento é amplamente reconhecido como um dos mais críticos para a segurança viária, uma vez que erros, decisões inadequadas ou comportamentos arriscados frequentemente antecedem os acidentes.

De acordo com dados do Ministério dos Transportes (2018), entre 2007 e 2016, 53,7% das mortes em acidentes de trânsito foram causadas por desrespeito às leis de trânsito ou falta de atenção. Em comparação, fatores como animais na pista, falhas mecânicas e defeitos nas vias somaram apenas cerca de 4,5% das fatalidades no mesmo período.

Existem ações educativas desempenhadas por órgãos do governo, como o Detran-PR que visam alertar os riscos do trânsito às pessoas. A Agência Estadual de Notícias do Paraná (2024) procura cumprir o objetivo de, como descrito pela mesma, “conscientizar a população sobre segurança no trânsito e destacar a visibilidade dos motociclistas”.

2.4 INFRAESTRUTURA E CONDIÇÕES DA VIA

Segundo a NBR10697 (2020) o fator viário em sinistros é descrito como “características ou deficiência na via ou sua sinalização, responsável pela ocorrência do sinistro”. A Agência CNT Transporte (2021) aponta um aumento de até 52,9% nas mortes a cada cem acidentes em trechos com sinalização considerada péssima ou ineficiente e um aumento de 53,5% em locais com problema nas na pintura das faixas centrais.

Mesmo com essas afirmações a influência da infraestrutura viária e da sinalização não é facilmente mensurável nos acidentes, agravado pelo pouco acúmulo de conhecimento no assunto (PEO 2002, apud Nodari 2003 p. 27). Uma infraestrutura bem projetada, com pavimentação adequada, iluminação eficiente e sistemas de drenagem, reduz os riscos de acidentes, especialmente em condições adversas, como chuva ou baixa visibilidade ou no caso de motoristas menos experientes.

Além disso, a implantação de dispositivos de segurança leva em consideração diversos fatores. Conforme o guia divulgado pelo Departamento Nacional de

Infraestrutura de Transportes (1998, p. 2) as ações tomadas pelos órgãos responsáveis pelo Gerenciamento da Segurança Viária podem ser preventivas, padronizando a infraestrutura em todas as rodovias, ou corretivas, devendo ser adotado em problemas locais específicos.

Por mais que ocasionem acidentes, os fatores viários não são tão presentes ou significativos no recorte estudado. Na base de sinistros da PRF 10% dos registros na BR-116 são atribuídos a condição da rodovia, diminuindo para 7% para eventos com vítimas fatais.

A rodovia apresentada acima, e estudada neste trabalho, é do tipo Longitudinal, corta o país de norte a sul, indo do Ceará até o Rio Grande do Sul. Possuindo 4610 quilômetros. Embora a maioria dos acidentes ocorrerem na região sudeste do Brasil. Na cidade de São Paulo a estrada atravessa por dentro, sem contornar a grande metrópole. Já nas estradas mais ao sul acontecem eventos climáticos mais extremos, como ciclones, chuvas fortes e, por consequência, deslizamentos, várias ações são tomadas para adequar melhor as vias para este clima.

A infraestrutura viária e a sinalização são indispensáveis para a segurança no trânsito, especialmente em rodovias extensas e diversificadas como a BR-116. Embora ainda sendo obscura quanta a influência destes fatores nos sinistros, sua contribuição para a ocorrência de acidentes não deve ser subestimada. Entretanto, é possível reduzir riscos, garantir maior fluidez no tráfego e preservar vidas, reafirmando o papel essencial da infraestrutura na segurança rodoviária.

3 ENCAMINHAMENTO METODOLÓGICO

Esta é uma pesquisa com abordagem quantitativa, natureza aplicada, objetivo descritivo e de modalidade e procedimentos aplicados. Para isto foi feito a descrição do modelo de aprendizado de máquina e das bases de dados utilizadas.

O detalhamento do modelo utilizado permite compreender como os dados devem ser processados e o motivo da técnica adotada conseguir identificar padrões nos acidentes. Com isso, há o aumento da confiabilidade do estudo, facilidade na análise de desempenho, contribuição para a compreensão e validade dos resultados deste trabalho.

Para a análise dos padrões, foram utilizadas as bases de dados abertos disponibilizadas pela PRF. As quatro bases, de 2020 a 2023 (separadas por ano), incluem informações sobre os acidentes e são atualizadas mensalmente. As bases devem ser descritas para uma compreensão do significado de seus atributos, de maneira a ser possível a normalização.

Ao utilizar as bases de dados fornecidas, separadas por ano de registro, é preciso realizar um processo de integração desses conjuntos de dados e posterior normalização. Essa nova base de dados permite melhores análises, garantindo que todos os fatores relevantes sejam considerados de maneira consistente e facilita a identificação de padrões sem vieses que possam aparecer em um ano específico. Além de descomplicar a aplicação dos modelos de aprendizagem de máquina.

Os dados provenientes de diferentes fontes costumam apresentar escalas, unidades de medida e formatos variados, o que pode prejudicar no aprendizado de máquina. A normalização visa ajustar os valores para uma mesma escala ou padrão, reduzindo o impacto de grandes variações que poderiam distorcer os resultados. Dessa forma, é possível a comparação entre variáveis e melhora a convergência dos algoritmos.

A estatística descritiva visa melhor compreensão da estrutura e as características dos dados, identificar padrões iniciais, detectar inconsistências ou outliers, e verificar a qualidade das informações disponíveis. Estas ações são cruciais, pois garantem que o modelo está encontrando relações mais complexas, que justificam a sua utilização no lugar de um painel de.

Com as bases normalizadas e integradas o algoritmo é aplicado, testando diferentes hiperparâmetros cada um tendo suas características específicas para o

treinamento do método. A aplicação deste permite investigar como as variáveis influenciam a ocorrência de acidentes fatais. Por meio do uso da técnica supervisionada é visado encontrar correlações e tendências que não seriam facilmente visíveis em análises tradicionais.

Com base nas análises é preciso discorrer sobre os resultados encontrados. O principal questionamento envolve a possibilidade, em caso de acidente, prever se haverá vítimas fatais. Se for possível realizar está predição, questionar quais são os fatores que mais influenciam nas ocorrências e se é viável realizar estas predições

3.1 MATERIAIS

Neste trabalho foram necessários aplicativos de computador para normalização, análise e métodos preditivos nos dados abertos da PRF. Segundo o próprio site do governo as bases são atualizadas mensalmente com base nas informações registradas pelo Boletim de Acidentes de Trânsito (BAT). Os dados são abertos ao público pelo governo brasileiro com base na Lei de Acesso à Informação (LAI) e pelo compromisso assumido na criação do Plano de Ação Nacional de Governo Aberto.

Utilizadas as bases de 2020 a 2023 unificadas e normalizadas para realização do treinamento do algoritmo. O resultado do treinamento foi testado na base de 2024 para a validação do treinamento ser padronizada quanto aos dados de teste.

3.1.1 Bases de Dados

Para criar a base de dados é preciso compreender as bases existentes, unificá-las e normalizar a nova base. A base de dados dos acidentes criada pela PRF é composta por trinta atributos descritos por nome, tipo de dados registrados e objetivo do atributo na tabela Descrição dos atributos apresentados no Quadro 1.

Quadro 1 - Descrição dos atributos

Atributo	Tipo	Objetivo
id	Numérico	Número de identificação
data_inversa	Data	Data
dia_semana	Textual	Data
horario	Tempo	Hora
uf	Textual	Estado
br	Textual	Estrada
km	Numérico	Local mais específico da rodovia
municipio	Textual	Cidade do acidente
causa_acidente	Textual	Possível causa do acidente
tipo_acidente	Textual	Característica principal do acidente
classificacao_acidente	Textual	Se há vítimas
fase_dia	Textual	Fase em que ocorreu o acidente
sentido_via	Textual	Ocorrido na subida ou decida
condicao_metereologica	Textual	Condições climáticas
tipo_pista	Textual	Número de pistas e orientação
tracado_via	Textual	Traçado da pista
uso_solo	Textual	Uso solo do veículo
peessoas	Numérico	Número de pessoas envolvidas
mortos	Numérico	Óbitos decorrentes do acidente
feridos_leves	Numérico	Feridos que não são considerados graves
feridos_graves	Numérico	Feridos com hospitalização superior a 24 h
ilesos	Numérico	Sem ferimentos
ignorados	Numérico	Sem dados
feridos	Numérico	Números de feridos no acidente
veiculos	Numérico	Número de veículos envolvidos
latitude	Numérico	Latitude específica do acidente
longitude	Numérico	Longitude específica do acidente
regional	Textual	Região estadual
delegacia	Textual	Delegacia responsável
uop	Textual	Centro da PRF responsável

Fonte – O autor (2024)

Os atributos descritos acima estão presentes nos bancos de dados do período estudado. Os dados que se alteram entre as bases são os registros de ocorrências, ou seja, as instâncias presentes em cada um. Diante destas diferenças estão registradas na tabela Características das Bases de Dados (Tabela 1)

Tabela 1 - Características das Bases de Dados

Base de Dados	Instâncias	Envolvidos	Mortos	Mortos/Envolvidos
datatran2020	63.585	157.302	5.292	0,0336
datatran2021	64.567	160.832	5.397	0,0336
datatran2022	64.606	165.290	5.441	0,0329
datatran2023	67.766	177.215	5.627	0,0318
Total 2020-2023	260.524	660.639	21.757	0,0329
datatran2024	53.928	139.667	4.515	0,0323
Total Geral	314.452	800.306	26.272	0,0328

Fonte: O autor (2024) com base nos dados da PRF

Conforme observado na Tabela 1, é possível notar uma pequena diminuição na quantia de mortos por envolvidos em acidentes, mas que logo aumenta no ano de 2024. No entanto, pode ocorrer por conta da falta de dados dos últimos meses do ano apresentado, este indo até o mês de setembro. Os dados são recolhidos de pelo menos 118 rodovias acompanhadas pela PRF nestes anos. Ao realizar um filtro nos dados para apresentar apenas as informações da BR-116 é gerada a Tabela 2.

Tabela 2 - Características das Bases de Dados Filtrada para BR-116

Base de Dados	Instancias	Envolvidos	Mortos	Mortos/Envolvidos
datatran2020	9.155	22.100	690	0,0312
datatran2021	9.704	23.504	690	0,0294
datatran2022	9.830	24.424	640	0,0262
datatran2023	10.804	27.813	741	0,0266
Total 2020-2023	39.493	97.841	2.761	0,0282
datatran2024	8.459	21.545	597	0,0277
Total Geral	47.952	119.386	3.358	0,0281

Fonte: O autor (2024)

Após a aplicação do filtro ocorre uma diminuição da média de vítimas fatais, o que pode indicar um maior volume de dados sobre acidentes nesta BR, ou seja, uma recorrência maior nos registros de acidentes não fatais na via. Já que representa mais de 15% das instancias dos dados gerais de janeiro de 2020 até setembro de 2024.

Os dados do período de 2020 até 2023 foram utilizados para o treinamento, pois esses anos apresentam registros contínuos de acidentes em todos os meses, garantindo uma base de dados mais completa e abrangente. Por outro lado, os dados referentes ao ano de 2024 foram reservados para a fase de teste, servindo como uma base independente para avaliar a capacidade de generalização dos modelos. Ao utilizar o ano de 2024 como conjunto de teste é possível verificar se os padrões e previsões identificados pelos algoritmos são aplicáveis a dados recentes. Esta abordagem garante que o modelo possa ser aplicado a cenários futuros.

3.1.2 Descrição do Algoritmo

Para o treinamento de máquina foi utilizado o *Weka* na versão 3.8.6 para este trabalho. O método de predição escolhido foi o *Random Forest* dado sua facilidade de uso junto da eficiência e precisão do modelo como dito pela Escola Britânica de Artes Criativas e Tecnologia (2024). Este aplicado realizando alterações em hiperparâmetros para melhorar o aprendizado.

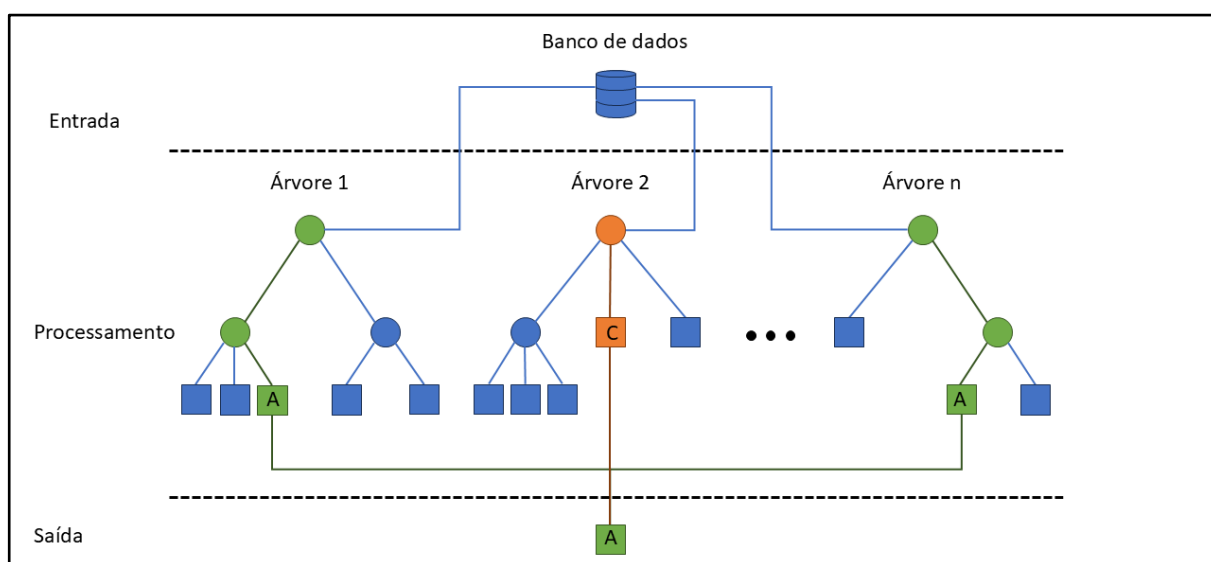
O aplicativo *Weka* foi criado pela Universidade de Waikato, na Nova Zelândia, tem seu nome derivado de uma espécie de pássaro. Foi programado por Eibe Frank, Mark Hall, Peter Reutemann e Len Trigg, junto de outros desenvolvedores que tiveram uma menor participação no projeto (*Weka Wiki*). Os autores disponibilizam ferramentas gratuitas para mineração de dados, compatíveis com a maioria das plataformas e aceita diversos formatos de bases de dados (Softonic, 2023).

A descrição dos algoritmos serve para esclarecer o processo pelo qual os dados serão analisados, garantindo a transparência e a compreensão das técnicas utilizadas no estudo. Além disso, esse detalhamento ajuda a justificar a escolha de cada algoritmo, destacando suas características, vantagens e limitações, o que facilita a interpretação dos resultados e permite a replicação do estudo por outros pesquisadores.

O *Random Forest* cria diversas árvores de decisão para realizar a predição dos dados, cada árvore é feita com uma partição diferente da base de dados. As partições da base são realizadas de forma aleatória, isto é feito para evitar um hiper ajuste das árvores para a base. No final, com base na média ou maioria dos resultados, é calculada a saída do algoritmo para cada registro da base de teste.

Se aplicado um banco de dados em um algoritmo já treinado para prever se um evento tem resultado A ou C. O método irá prever o tributo meta dos registros conforme o resultado da maioria das árvores geradas no treinamento, demonstrado na Ilustração 1.

Ilustração 1 – Funcionamento do *Random Forest*



Fonte: O autor (2024)

Com base em Breiman (2001) temos alguns aspectos importantes de precisão do algoritmo:

- Convergência: é o processo de criação da saída com base nas árvores criadas durante o treinamento;
- Função de Margem: é a média de votos a mais que a classe correta recebe em relação as demais, quanto maior a margem melhor é a predição;
- Erro de Generalização: a classe correta recebe menos votos que as demais
- Força e Correlação: o desempenho é medido nestas variáveis.
 - Força: quanto mais as árvores acertam o resultado correto, maior é a força na predição.

- Correlação entre Classificadores: similaridade de previsões entre árvores, pode ser gerar vieses se forem muito correlacionadas;

Ainda seguindo a explicação de Breiman (2001), no caso de classes binárias, que poderão ser utilizadas neste estudo, ocorre uma simplificação da função de margem. Em que para ser considerado uma predição forte é preciso de uma taxa de acerto maior que 50%, que seria um valor alcançável por predições fracas ou até mesmo uma predição aleatória.

É possível alterar algumas configurações de treinamento, chamadas de hiperparâmetros em mineração de dados, que influenciam diretamente na maneira com que os dados serão utilizados e na composição final das árvores de decisão geradas. Cada algoritmo possui um conjunto de propriedades configuráveis ligadas aos seus métodos de aprendizagem. Estes conjuntos, se bem ajustados, melhora os resultados da predição e consequentemente do trabalho. A documentação do *Weka* apresenta a seguinte descrição dos hiperparâmetros disponíveis para configuração do método:

- *seed*: A semente aleatória para o treinamento;
- *representCopiesUsingWeights*: Os pesos das instâncias devem ser explícitos ou as instâncias devem ser apenas copiadas;
- *storeOutOfBagPredictions*: Se deve separar previsões de fora do treinamento para realizar os testes;
- *numExecutionSlots*: O número de *threads* a serem utilizados para o conjunto;
- *bagSizePercent*: Quantidade de instâncias nos conjuntos para o treinamento, calculado em porcentagem;
- *numDecimalPlaces*: O número de casas decimais a ser utilizado;
- *batchSize*: Tamanho do lote em cada iteração;
- *printClassifiers*: Se deve apresentar os classificadores individuais;
- *numIterations*: Quantidade de árvores a serem criadas;
- *debug*: Se verdadeiro, são apresentadas informações adicionais;
- *outputOutOfBagComplexityStatistics*: Se verdadeiro, exibi estatísticas baseadas em complexidade, quando a avaliação *storeOutOfBagPredictions* é verdadeira;
- *classifier*: Classificador base;
- *breakTiesRandomly*: Resolver empates aleatoriamente entre atributos;

- *doNotCheckCapabilities*: Se configurado, testa as capacidades do classificador antes de construí-lo, aumenta tempo de processamento;
- *maxDepth*: A profundidade máxima da árvore, o método se beneficia de árvores mais profundas;
- *computeAttributeImportance*: classifica os atributos por média de impureza;
- *calcOutOfBag*: Se calcula o erro fora da bolsa;
- *numFeatures*: Número de atributos escolhidos aleatoriamente.

O *Random Forest* no *Weka* é uma técnica de aprendizado de máquina que constrói um conjunto de árvores de decisão para realizar tarefas de classificação e regressão. A força reside na sua capacidade de criar um modelo, utilizando a ideia de "votação" entre múltiplas árvores para melhorar a generalização. Além disso, o uso de amostras e a seleção aleatória de atributos em cada divisão aumentam a diversidade entre as árvores. O algoritmo também oferece ferramentas para avaliar a importância dos atributos, ajudando a entender quais características têm maior impacto na previsão. Apesar de construir árvores de decisão o aplicativo não possui um elemento gráfico para visualização destas. Embora seja mais pesado para treinamento do que modelos que geram árvores individuais, é uma excelente escolha para problemas complexos.

3.2 TRABALHOS RELACIONADOS

Para iniciar a pesquisa em bases de dados foram escolhidos três repositórios, sendo estes a biblioteca de trabalhos de conclusão de curso de gestão da informação (TCC - GI), BRAPCI e Scielo. Também foram adotadas as seguintes estratégias de pesquisa "machine learning" e "aprendizado de máquina". Com o objetivo de retornar as publicações que tenham métodos de análise utilizados na área. As informações recuperadas estão apresentadas abaixo no Quadro 2.

Quadro 2 – Levantamento Bibliográfico

LEVANTAMENTO BIBLIOGRÁFICO						
DATA	BASE DE DADOS	CAMPO DE BUSCA	ESTRATÉGIA	FILTRO	1FILTRO	RESULTADOS
12/11/ 2024	TCC - GI	todos os índices	machine learning	-	-	2
12/11/ 2024	TCC - GI	todos os índices	aprendizado de máquina	-	-	4
12/11/ 2024	BRAPCI	todos os índices	machine learning	-	-	59
12/11/ 2024	BRAPCI	todos os índices	aprendizado de máquina	-	-	36
12/11/ 2024	SCIELO	todos os índices	machine learning	-	-	439
12/11/ 2024	SCIELO	todos os índices	aprendizado de máquina	-	-	66

Fonte: O autor (2024)

Como uma forma de conseguir resultados mais específicos foram testadas estratégias compostas nas bases de dados. Nas estratégias haviam sido usados os termos apresentado anteriormente em conjunto com os termos “rodoviário”, “rodovias” e “BR”, desta forma só foi possível recuperar apenas um trabalho em evento que utilizava mineração de dados para descobrir regras quanto a acidentes na BR-341 realizado por alunos da Fundação Mineira de Educação e Cultura (FUMEC).

O trabalho apresentado em evento descreve regras quanto a acidentes sem vítimas, com vítimas feridas e com vítimas fatais. O que já permite apresentar algumas ideias gerais do que é possível encontrar na BR-116.

De acordo com Cristian, Jersone e Luiz (2015, p. 19),

Os acidentes em pista simples para os sexos feminino e masculino são normalmente associados a colisões frontais em curva, devido ao traçado complexo, velocidade incompatível para o trecho, condições meteorológicas desfavoráveis[...] em pista dupla, são observados, para o sexo masculino, os fatores de contribuição como: colisão traseira associada à falta de atenção em via reta e céu claro. Observam-se, também, acidentes com motocicletas e acidentes envolvendo atropelamento de pessoas. Esses fatores de contribuição foram também encontrados para o sexo feminino, porém, os acidentes do sexo feminino não tiveram a motocicleta como fator de contribuição.

No documento também é descrito que há um aumento no número de acidentes entre os anos de 2009 e 2012, porém com o número de vítimas se reduzindo no mesmo período. Indicando uma melhora nos equipamentos de segurança dos veículos que transitam nas estradas neste espaço de tempo.

Para a avaliar os hiperparâmetros e o algoritmo como um todo foi utilizado o coeficiente de correlação em conjunto com o erro médio absoluto.

4 RESULTADOS E ANÁLISES

Com a compreensão da base de dados e seus atributos é realizado a normalização, com a adequação a base ao algoritmo, de modo a facilitar o treinamento de máquina.

A exploração no banco de dados normalizado permite a compreensão do conteúdo e as características dos dados. Esta etapa envolve a inspeção das variáveis, a identificação de padrões, tendências e possíveis inconsistências.

Como evidenciado nas tabelas há uma similaridade entre os dados da BR-116 e os acidentes mais comuns. De maneira a identificar semelhanças com os estudos e dados apresentados pelas ODS e PRF associados à estas ocorrências, como condições de comportamento do motorista, pedestres e condições da rodovia.

Ao término da análise preliminar é esperado que pelo menos uma das configurações de hiperparâmetros apresente bom desempenho na previsão de acidentes, permitindo estimar a probabilidade de incidentes em diferentes cenários e com dados que possam ser coletados na observação de uma ocorrência.

4.1 ETAPA DE NORMALIZAÇÃO E ANÁLISES DESCRITIVAS

Para a aplicação do algoritmo é feita uma remoção de atributos, com o intuito é retirar dados que identifiquem um acidente específico ou que não possam ser alterados por processos prévios, por exemplo a quantidade de veículos envolvidos ou o número de identificação do sinistro. Ademais, são removidos os atributos que estão apresentando dados redundantes ou específicos, como “município”, “latitude” e “longitude”, mantendo o atributo “km”. O atributo “br” é igual para todas as instâncias, também sendo removido. Depois da exclusão dos atributos ficou a seguinte relação, tendo algumas alterações apresentadas a seguir na Tabela 3.

Tabela 3 – Atributos alterados

Atributo	Tipo	Alterações
data_inversa	Numérico	Removido o ano, separado em dia e mês, ambos textuais
dia_semana	Textual	
horario	Textual	Removido os minutos
uf	Textual	
km	Textual	Arredondado para baixo. Exemplo: 1,9 vira 1
causa_acidente	Textual	
tipo_acidente	Textual	
fase_dia	Textual	
sentido_via	Textual	
condicao_metereologica	Textual	
tipo_pista	Textual	
tracado_via	Binário	Normalizado por <i>One-hot Encoding</i>
mortos	Binário	Valor 0 para sem vítimas fatais e 1 para com vítimas fatais

Fonte: O autor (2024)

Não houve normalização na maioria dos atributos, pois o método escolhido consegue realizar o treinamento nos dados textuais da base sem comprometer seu desempenho. Entretanto, para uma maior adequação dos dados de data fora removido a informação do ano da ocorrência, junto da separação em dois novos

atributos, “dia” e “mês”, ambos contendo o tipo numérico. Logo primeiro de janeiro de 2020 ficou representado como “dia” = 1 e “mês” = 1. No atributo “horario” foi removido os minutos e segundos presentes na base original, deixando apenas as horas representadas como o tipo numérico. A hora 12:45:00, vira apenas, “horario” = 12. O atributo “tracado_via” também é alterado utilizando *One-hot Encoding*.

O “tracado_via” apresenta uma descrição do caminho da rodovia. As opções possíveis são: reta, curva, declive, aclive, interseção de vias, viaduto, ponte, rotatória, retorno regulamentado, desvio temporário, túnel e em obras. A estratégia de normalização foi escolhida por conta da soma destas características em uma mesma instância, podendo ter um trecho de, por exemplo, “curva; túnel; em obras”. Este método de tratamento de dados separa um atributo em vários, cada um representando uma das opções possíveis do atributo original.

Partindo disto, as opções de registro possíveis são 1, o trecho possui a característica, e 0, o trecho não possui a característica, tornando-o do tipo binário. Desta forma, voltando para o exemplo, um trecho sendo “curva; túnel; em obras”, se torna equivalente a um trecho “curva; em obras; túnel”. Além de possibilitar a filtragem e processamento individualmente de cada atributo. Assim a “curva; túnel; em obras” pode ser agrupada com a “reta; em obras” por ambas estarem em obras.

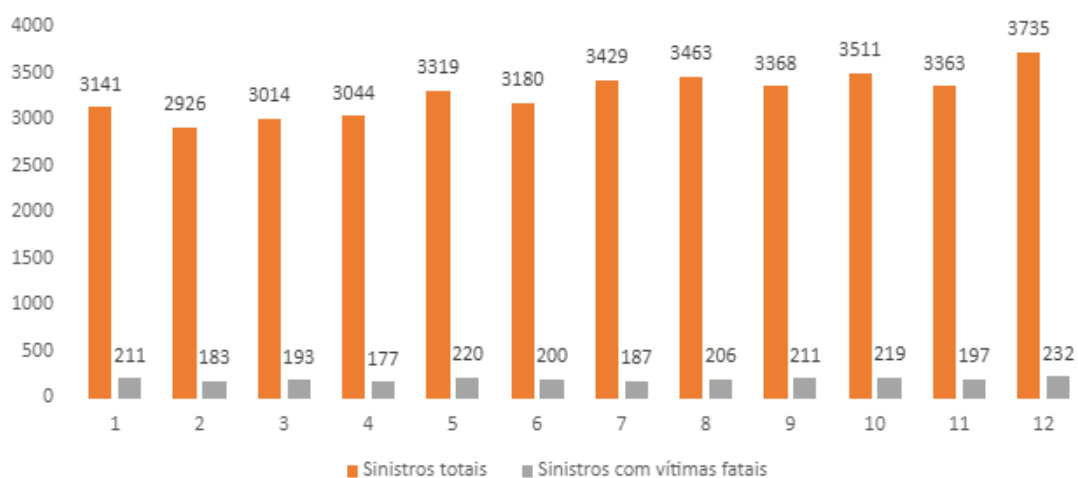
Por fim, o atributo “mortos” foi alterado para binário de modo a facilitar a validação do método. Uma vez que o objetivo do trabalho não é acertar o valor exato de vítimas fatais, mas sim, demonstrar padrões de características aparecem em acidentes com vítimas fatais para evitá-los futuramente.

Ao final da normalização da base de dados fica com vinte e cinco atributos, mantendo todas as trinta e nove mil novecentos e noventa e quatro instâncias da base original. Todos registros foram mantidos por estarem completos e não possuem valores vazios ou incoerentes. Para facilitar a descrição e apenas para criar uma explicação mais precisa desta nova base, ela foi separada em quatro. Atributos referentes a tempo (“dia”, “mes”, “dia_semana”, “horario”), atributos de localização (“uf”, “km”), características do sinistro (“causa_acidente”, “tipo_acidente”) e condições do ambiente (os demais atributos com exceção de “mortos”).

Para os atributos referentes ao tempo, o dia segue uma distribuição normal ao longo do mês, indicando que os sinistros estão uniformemente distribuídos, sem grandes concentrações em datas específicas. Para o atributo mês exibe um padrão particular: meses com 30 dias apresentam uma quantidade menor de sinistros

comparados aos meses com 31 dias, enquanto fevereiro, com menos dias, registra o menor número de ocorrências. Em contraste com isto, acontece um aumento na quantidade de sinistros durante o ano, tendo dezembro como pico de registros. Possivelmente associado ao maior fluxo de veículos devido a férias e festividades de fim de ano. Esta distribuição pode ser observada também nos sinistros com vítima fatal demonstrado no Gráfico 1.

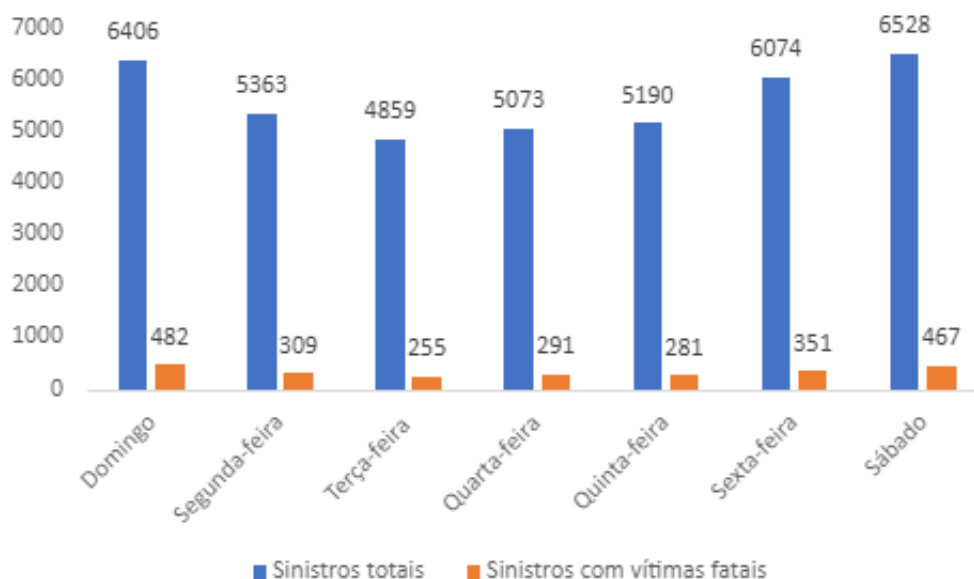
Gráfico 1 - Distribuição de sinistros entre os meses



Fonte: O autor (2024)

Outro atributo com tendências bem aparentes é o dia_semana, como apresentado no Gráfico 2. Este demonstra uma correlação forte entre de sinistros totais e sinistros com vítimas fatais, observando que durante os dias úteis entre 5 e 6% dos acidentes tenham vítimas fatais, podendo chegar a 7 e 8% nos finais de semana.

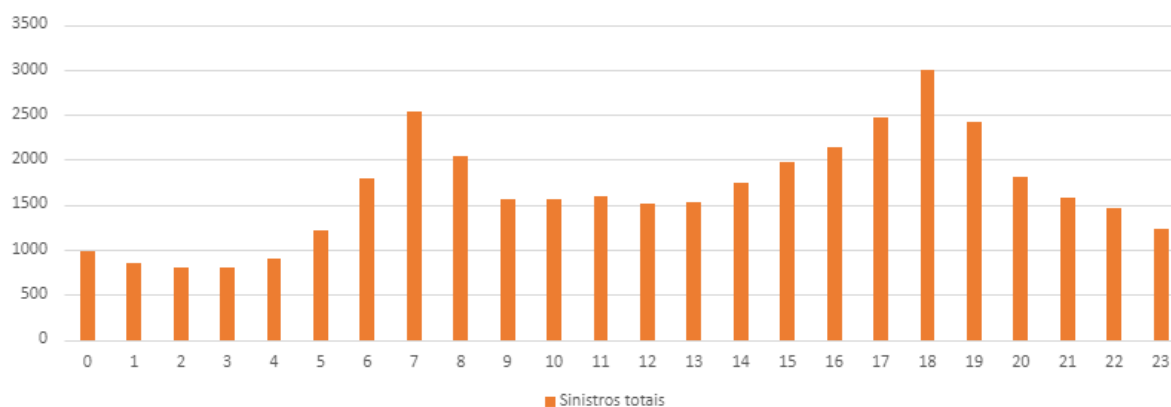
Gráfico 2 - Distribuição de sinistros entre os dias da semana



Fonte: O autor (2024)

O “horario” revela dois picos diários: um menor às 7h, associado ao início da rotina diária, e outro maior às 18h, coincidente com o horário de término das atividades de trabalho, indicando uma correlação entre a movimentação nos horários de pico e a ocorrência de sinistros, observável no Gráfico 3.

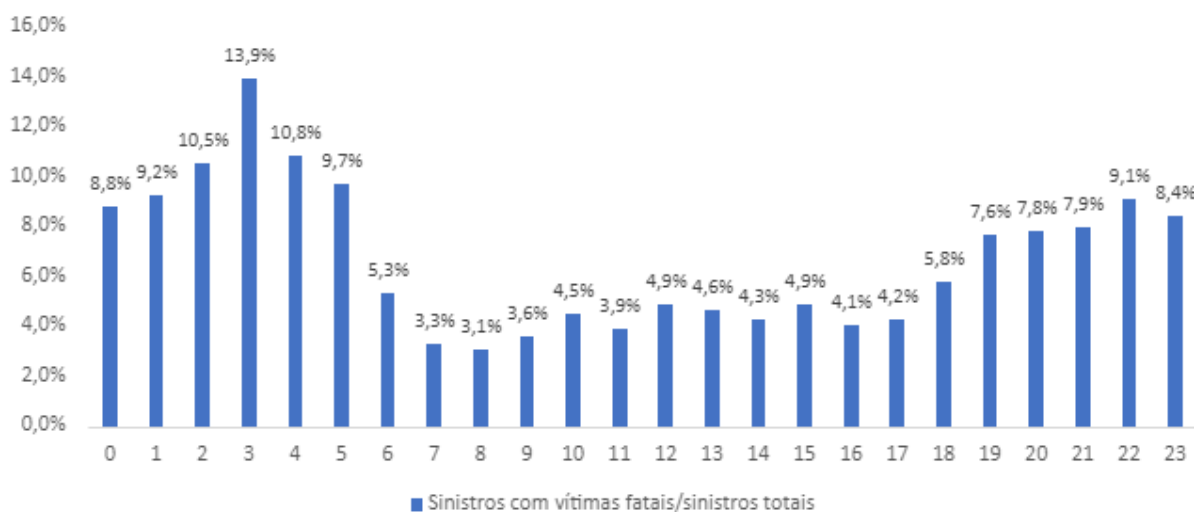
Gráfico 3 - Distribuição de sinistros entre as horas do dia



Fonte: O autor (2024)

Comparando o Gráfico 3 com o Gráfico 4 há uma diferença quanto a correlação com sinistros com vítimas fatais, onde entre 2 e 4 da manhã a porcentagem chega a ultrapassar os 10%, tendo seu pico próximo aos 14% às 3 horas.

Gráfico 4 - Taxa de acidentes fatais entre as horas do dia



Fonte: O autor (2024)

As taxas mais altas durante o período de 19 horas da noite e as 5 horas da manhã podem indicar maior incidência de fatalidades por conta do cansaço do motorista, ou, pelo menor tráfego de veículos, permitindo que quem trafegue no intervalo atinja maiores velocidades. Pode haver até mesmo uma correlação entre estes dados e as causas de acidentes mais comuns ligadas a fatalidades, sendo velocidade incompatível, não reação do condutor ou reação tardia do condutor.

Os atributos referentes ao local não devem ser analisados de forma isolada. Em virtude, dos estados possuírem diferenças entre as extensões da rodovia que os percorre e a contagem do km, que está diretamente ligada ao estado em que a rodovia se encontra, sempre iniciando no km = 0. Por exemplo, o Paraná chega até o km 211 enquanto a Bahia atinge o 933, logo o km 133 pode se referir tanto a um trecho no Paraná quanto na Bahia. Essa característica reforça a necessidade de integrar ambos os dados para análises mais precisas, evitando confusões que poderiam comprometer a compreensão dos dados.

Tabela 4 – Atributos referentes ao local de sinistro

Estado	Sinistros totais	Sinistros fatais	km máximo	Fatais/sinistros totais	Fatais/km máximo
CE	2.500	202	545	8%	0,3706
PB	26	6	13	23%	0,4615
PE	147	18	91	12%	0,1978
BA	3.395	383	933	11%	0,4105

MG	4.655	403	816	9%	0,4939
RJ	5.807	331	339	6%	0,9764
SP	12.547	631	569	5%	1,1090
PR	4.310	209	211	5%	0,9905
SC	1.261	93	310	7%	0,3000
RS	4.845	160	659	3%	0,2428

Fonte: O autor (2024)

Com base na Tabela 4, observa-se que Pernambuco e Paraíba não possuem um total de acidentes relevante a ponto de garantir que valores fora do padrão sejam diluídos na amostra estudada. No entanto, o estado de São Paulo é em disparado o que mais possui registros totais e por km de estrada. A Bahia, por outro lado, é o estado com maior taxa de acidentes fatais em relação ao total de acidentes.

Nas características dos acidentes há dois atributos possíveis, diferenciados no Quadro 3 conforme seus valores únicos, valores mais repetidos, menos repetidos, característica com maior número de acidentes fatais e característica que tem maior chances de ocasionar um acidente fatal.

Quadro 3 – Atributos quanto as características do sinistro

	causa_acidente	tipo_acidente
Valores únicos	89	20
Mais repetidos	Reação tardia ou ineficiente do condutor	Colisão traseira
Menos repetidos	"Pedestre - Ingestão de álcool/ substâncias psicoativas" e "Sinalização encoberta"	Danos eventuais
Causam mais acidentes fatais	Velocidade Incompatível	Atropelamento de Pedestre
Maior chance de fatalidade	Desobediência às normas de trânsito pelo pedestre	Atropelamento de Pedestre

Fonte: O autor (2024)

No Quadro 3 é possível observar informações sobre causas de acidentes, a mais comum se dá por conta de descuido do motorista. De acordo com a causa de mais fatalidade, sendo a velocidade incompatível com a via. Outro dado aparente no quadro é a fragilidade do pedestre em comparação com os passageiros, em que por mais que os motoristas sejam a principal causa, se o pedestre desobedece às normas de trânsito, este sim, tem a maior chance de fatalidade. Ressaltando este

comportamento temos que os tipos de acidentes fatais mais comuns e os que possuem maior chance de fatalidade são atropelamento de pedestres.

Na BR-116 observa-se que fatores humanos foram responsáveis por 83% dos acidentes em rodovias federais e 91% das ocorrências com vítimas fatais. Entre as principais causas de acidentes fatais, destacam-se a velocidade incompatível, a ausência de reação e reação tardia do motorista. A velocidade incompatível é a infração mais comum no Brasil registrada pela PRF (2024), as violações mais registradas em 2023 foram excesso de velocidade (50.690 ocorrências), ultrapassagens proibidas (6.553) e ausência do cinto de segurança (5.645).

Por mais que o uso de substâncias como drogas lícitas ou ilícitas alterarem a percepção de mundo do motorista, embora importante, ocupou apenas a oitava posição entre os fatores associados a sinistros de trânsito, contribuindo para 584 acidentes. Quando filtrados para sinistros com vítimas fatais onde houve a ingestão de álcool, esse número caiu para 27 ocorrências, posicionando o consumo de álcool em décimo quarto lugar entre as causas mais frequentes de acidentes fatais.

Por fim, temos os atributos referentes a condição do ambiente. Começando por fase_dia. As características quanta a número de sinistros e quantidade de sinistros com vítimas fatais de cada um é apresentado na Tabela 5.

Tabela 5 – Atributo fase_dia		
fase_dia	n° sinistros	n° sinistros com vítimas fatais
Pleno dia	20.887	848
Plena Noite	14.459	1.310
Amanhecer	2.037	183
Anoitecer	2.110	95

Fonte: O autor (2024)

O atributo fase_dia consolida as informações do atributo horário ao agrupar os registros em períodos mais amplos. Porém, estes períodos podem acontecer em horários diferentes, portando este não fora retirado da base de dados. Enquanto o atributo horário detalha picos específicos, como os registrados às 7h e 18h, a fase do

dia permite observar tendências gerais relacionadas a posição do sol ou a sua ausência.

Tabela 6 – Atributo sentido_via

sentido_via	n° sinistros	n° sinistros com vítimas fatais
Decrescente	19.199	1.163
Crescente	20.294	1.273

Fonte: O autor (2024)

O sentido da via, relatado na Tabela 6, traz a informação do sentido em que o trecho se encontra. Quando este vai para o norte é apresentado como decrescente, quando para o sul, crescente.

As condições climáticas são descritas na Tabela 7.

Tabela 7 – Atributo condicao_metereologica

condicao_metereologica	n° sinistros	n° sinistros com vítimas fatais
Céu Claro	22.267	1.400
Nublado	7.860	510
Sol	2.297	95
Chuva	4.115	236
Nevoeiro/Neblina	428	48
Garoa/Chuveiro	2.006	112
Ignorado	463	32
Vento	54	3
Neve	1	0
Granizo	2	0

Fonte: O autor (2024)

No quadro pode-se observar que o céu nublado e nevoeiro/neblina são fatores que levam a uma maior chance de sinistros com vítima fatal, seguidos por chuva e garoa/chuveiro. Ainda que céu claro seja o dado mais registrado, a probabilidade de acontecer uma fatalidade é baixa.

Ao considerarmos o tipo de pista, nota-se que quanto mais pistas estiverem disponíveis para tráfego, menos há a possibilidade de sinistros com vítimas conforme demonstrado na Tabela 8.

Tabela 8 – Atributo tipo_pista

tipo_pista	n° sinistros	n° sinistros com vítimas fatais
Simples	14.265	1.214
Múltipla	6.228	255
Dupla	19.000	967

Fonte: O autor (2024)

O traçado da via, apresentado na Tabela 9, é o único atributo que a soma de n° de sinistros não fica em 39.493, já que o mesmo trecho pode adquirir mais de uma das características.

Tabela 9 – Atributo traçado da via

Traçado da pista	n° sinistros	n° sinistros com vítimas fatais
Reta	27.725	1.770
Curva	7.612	493
Declive	2.560	218
Aclive	2.115	166
Interseção de Vias	1.958	73
Viaduto	493	15
Ponte	246	24
Rotatória	547	8
Retorno Regulamentado	595	8
Desvio Temporário	115	2
Túnel	11	1
Em Obras	282	10

Fonte: O autor (2024)

Os traçados de via como declives, aclives e pontes estão entre os mais perigosos como apresentado na tabela, algumas das possíveis causas seriam: Os declives favorecem o ganho de velocidade, muitas vezes dificultando a frenagem.

Enquanto, os aclives, por sua vez, podem limitar a visibilidade dos motoristas, ocultando obstáculos, ou, em casos de veículos mais pesados, exerce uma desaceleração instigando veículos mais leve a realizar uma ultrapassagem. Já as pontes apresentam riscos específicos, como largura limitada, ausência de acostamento. Esses traçados requerem maior atenção dos condutores.

4.2 MODELO PREDITIVO

Ao término da análise de todos os atributos é aplicado o treinamento de máquina, com a base normalizada e os dados de teste prontos foi preciso apenas carregá-los no programa *Weka*. No primeiro teste, com os ajustes padrões do aplicativo, houve um uso excessivo da memória RAM do computador utilizado. Então, após alguns testes no *numIterations*, foi encontrado o valor de 60 por conta da limitação, o coeficiente de correlação ficou em 0,6651. Para comparar com modelos que possuem menos árvores são utilizados *numIterations* em 20, 30, 40 e 50, obtendo coeficientes de 0,553, 0,6062, 0,613 e 0,6421. O que demonstra que quanto maior a quantidade de árvores presentes no *Random Forest* melhor é a sua predição. Ajustando o *bagSizePercent* para 10 foi possível utilizar maiores valores no atributo anterior, como descrito abaixo na Tabela 10.

Tabela 10 – Testes *numIterations* com *bagSizePercent*

Hiperparâmetros	Conj. 1	Conj. 2	Conj.3	Conj. 4	Conj. 5	Conj. 6
<i>bagSizePercent</i>	100	100	10	10	7	7
<i>numIterations</i>	60	50	100	120	100	94
Coeficiente de correlação	0,6651	0,5621	0,7442	0,743	0,7745	0,7779

Fonte: O autor (2024)

Os dados apresentados mostram diferentes configurações de hiperparâmetros para um modelo de aprendizado de máquina e seus respectivos coeficientes de correlação. O maior coeficiente de correlação (0,7779) foi alcançado no Conj. 6, com *bagSizePercent* de 7% e *numIterations* de 94, enquanto o pior desempenho (0,5621) ocorreu no Conj. 2, usando 100% no *bagSizePercent* e 50 iterações. Observa-se que valores menores de *bagSizePercent*, especialmente em torno de 7%, tendem a melhorar os resultados, embora isso não seja absoluto, como

mostram os Conj. 3 e 4 (10%). Além disso, ao realizar um ajuste fino do método é notável que mais iterações não garantem melhor desempenho, nos valores de 94 e 100, ainda que o coeficiente de correlação seja parecido (variando 0,0034 entre o maior e o menor) o treinamento com menor número de iterações obteve um melhor resultado. Mesmo assim como o método tem melhor desempenho com mais árvores foi testado dobrar o número de iterações e foram obtidos os dados apresentados na Tabela 11.

Tabela 11 – Testes *numIterations* dobrado

Hiperparâmetros	Conj. 1	Conj. 2	Conj.3	Conj. 4	Conj. 5	Conj. 6
<i>bagSizePercent</i>	6	7	7	7	7	7
<i>numIterations</i>	200	200	205	195	199	198
Coeficiente de correlação	0,7835	0,7886	0,787	0,7867	0,7886	0,7903

Fonte: O autor (2024)

O desempenho entre as diferentes configurações de hiperparâmetros testadas teve uma menor variação comparando com os conjuntos anteriores. O melhor resultado foi alcançado no Conj. 6 (0,7903), utilizando *bagSizePercent* de 7% e *numIterations* de 198, seguido por desempenhos semelhantes no Conj. 2 e Conj. 5 (0,7886 cada). Isso sugere que, para esta configuração, o aumento do número de iterações para valores entre 195 e 205 não gerou diferenças significativas, indicando um certo ajuste fino nesse intervalo. Por outro lado, o uso de *bagSizePercent* de 6% no Conj. 1 apresentou o menor coeficiente (0,7835), reforçando que a configuração de 7% foi mais eficiente para maximizar a correlação. Para o próximo passo foi adicionado o *numFeatures* aos conjuntos e alterado nos demais valores para encontrar ótimos locais temos os resultados descritos na Tabela 12.

Tabela 12 – Testes *numIterations* com *bagSizePercent* e *numFeatures*

Hiperparâmetros	Conj. 1	Conj. 2	Conj.3	Conj. 4	Conj. 5	Conj. 6
<i>bagSizePercent</i>	7	6	6	6	6	6
<i>numFeatures</i>	15	10	11	11	11	11
<i>numIterations</i>	150	200	200	210	211	209
Coeficiente de correlação	0,6981	0,7356	0,7555	0,7574	0,7573	0,7569

Fonte: O autor (2024)

Os resultados mostram o impacto de diferentes configurações de hiperparâmetros no desempenho do modelo, conforme indicado pelos coeficientes de correlação. O melhor desempenho (0,7574) foi obtido no Conj. 4, com *bagSizePercent* de 6%, *numFeatures* de 11 e *numIterations* de 210. Observa-se que configurações com *bagSizePercent* em torno de 6% e *numFeatures* entre 10 e 11 apresentaram resultados consistentes, com coeficientes superiores a 0,73 em todos os casos. Além disso, o aumento no número de iterações para valores entre 200 e 210 parece estar associado a um leve ganho de desempenho, embora os Conjuntos 4, 5 e 6 demonstrem um possível valor ótimo.

Utilizando os dois melhores conjuntos de dados apresentados, foi introduzido a *seed*. Mesmo sendo uma semente aleatória apresenta melhora no coeficiente de correlação conforme a Tabela 13.

Tabela 13 – Testes *seed*

Hiperparâmetros	Conj. 1	Conj. 2	Conj.3	Conj. 4	Conj. 5	Conj. 6
<i>bagSizePercent</i>	7	7	7	6	6	6
<i>numFeatures</i>	0	0	0	11	11	11
<i>numIterations</i>	198	198	198	210	210	210
<i>seed</i>	2	3	84	2	3	84
Coeficiente de correlação	0,793	0,7412	0,7973	0,7423	0,758	0,7709

Fonte: O autor (2024)

O impacto das configurações de *seed* e *bagSizePercent* no desempenho do modelo. Os melhores coeficientes de correlação foram obtidos no Conj. 3 (0,7973) e no Conj. 1 (0,793), ambos com *bagSizePercent* de 7%, *numFeatures* fixado em 0, *numIterations* de 198 e diferentes valores de *seed* (84 e 2, respectivamente). Por outro lado, configurações com *bagSizePercent* de 6%, combinadas com *numFeatures* de 11 e *numIterations* de 210, apresentaram coeficientes de correlação inferiores, embora ainda robustos, com valores entre 0,7423 e 0,7709. Essa informação sugere que a redução do *bagSizePercent* para 6% e o aumento das iterações para 210 nem sempre resultam em melhor desempenho. O impacto da *seed* aparece em todas as configurações analisadas.

Ao término da análise envolvendo diversas configurações de hiperparâmetros, o maior coeficiente de correlação obtido foi de 0,7973, evidenciando o potencial de desempenho do modelo nos testes realizados. Esse resultado destaca a importância do ajuste parâmetros, como *bagSizePercent*, *numFeatures*, *numIterations* e *seed*, pois variações sutis entre as combinações podem influenciar o desempenho. A identificação dessa configuração reforça a necessidade da busca por hiperparâmetros otimizados, garantindo que o modelo alcance sua máxima capacidade preditiva e uma associação linear mais forte entre as variáveis preditas e reais.

Pela organização do método a sequência dos 5 atributos que melhor organizam os dados é dada por em obras, túnel, tipo_acidente, fase_dia, sentido_via, indo do mais relevante para o menos. Destes apenas tipo_acidente e fase_dia que descrevem características que tendem a ocorrências com vítimas fatais. Enquanto os atributos em obras e túnel descrevem situações onde a tendência de fatalidades fica próxima a zero. Já os 5 atributos menos considerados na predição são dia, interseção de vias, aclave, declive e curva, indo do menos relevante para o mais. Nestes são notáveis situações onde não há tanta interferência das características no resultado dos sinistros.

5 CONSIDERAÇÕES FINAIS

Desta forma, entende-se que o questionamento da pesquisa, a saber: “quais fatores podem ser identificados nos dados disponibilizados pela Polícia Rodoviária Federal como influentes nos sinistros da BR-116 que possam ser usados para predição de vítimas fatais nas ocorrências?”, foi respondido utilizando-se os dados que evidenciam o fator humano como o maior obstáculo para a segurança no trânsito, especialmente devido à imprudência e à desobediência às normas de circulação. Atitudes como excesso de velocidade, ultrapassagens proibidas e desatenção durante a condução dos veículos são responsáveis por grande parte das fatalidades nas estradas. Não é apenas o comportamento dos motoristas que levam a fatalidades na rodovia. Imprudência dos pedestres é um fator que levam a mais acidentes com vítimas fatais, por estes serem os elos mais fracos neste meio.

O estudo serviu para destacar a necessidade de reforçar a conscientização, a fiscalização e a educação para promover comportamentos mais responsáveis e seguros. Além da necessidade de investimento para a segurança dos pedestres que circulam nas vias.

A avaliação de dados para o mapeamento de fatores externos que influenciam em acidentes fatais na rodovia, sendo feita utilizando o próprio programa do *Weka*, onde os atributos foram classificados de acordo com o quanto estes organizam a base e reduzem a entropia dos dados.

A análise da distribuição de acidentes fatais na BR-116 com ênfase em fatores alteráveis para a redução das mortes, como infraestrutura, sinalização ou educação dos motoristas. A distribuição de acidentes fatais acompanha algumas tendências normais, porém fatalidades aparecem em maior ou menor proporção, principalmente de acordo com o comportamento do motorista ou pedestre. Foi possível chegar nesta conclusão por meio da estatística descritiva realizada nos atributos.

Ao descrever o funcionamento do método de aprendizado de máquina *Random Forest* e explicar sua eficiência. Neste trabalho foi possível descrever o funcionamento do método de aprendizado de máquina, levando em conta explicações de programadores e na documentação do próprio *Weka*. Fora a explicação as configurações do treinamento de máquina presentes no aplicativo, de modo a conseguir alterá-las para uma melhora no treinamento e teste do algoritmo.

O treinamento de modelo de aprendizado de máquina com alteração de hiperparâmetros para o melhoramento de performance. As alterações feitas em *bagSizePercent*, *numFeatures*, *numIterations* e *seed* tiveram um papel importante para a melhora da performance do *Random Forest*. Tendo um excelente ganho de desempenho em relação aos parâmetros padrões. Isto foi possível testando diferentes combinações para as variáveis e avaliando o resultado dos testes.

O objetivo do trabalho é o treinamento do modelo de aprendizado de máquina *Random Forest* tendo como atributo meta se há ou não vítimas fatais nos sinistros ocorridos na BR-116 com base nos dados abertos disponibilizados pela Polícia Rodoviária Federal. Este foi atingindo com êxito, o ajuste de hiperparâmetros junto dos resultados de coeficiente de correlação foram satisfatórios, estando próximos de 0,8. O que demonstra uma boa correlação, entre os atributos tidos como mais influentes na predição e a ocorrência de fatalidades.

Para trabalhos futuros a aplicação de outros métodos de predição ou a junção de mais dados na base servirão para incrementar a predição e encontrar novos padrões nos acidentes. Além da implementação de processos mais consolidados em órgãos governamentais. Afim de criar uma cultura orientada a dados na gestão pública, melhorando a resolução de problemas com a tomada de decisões pautadas em informação.

REFERÊNCIAS

AGÊNCIA CNT TRANSPORTE. Infraestrutura das vias tem relação com a segurança no trânsito. **Conferência nacional de transporte**, mai. 2021. Disponível em: <https://cnt.org.br/agencia-cnt/infraestrutura-das-vias-tem-relacao-com-a-seguranca-no-transito>. Acesso em: 5 dez. 2024.

AGÊNCIA ESTADUAL DE NOTÍCIAS. Trânsito. **Detran-PR promove ações educativas para reduzir acidentes com motociclistas**, set. 2024. Disponível em: <https://www.aen.pr.gov.br/Noticia/Detran-PR-promove-acoes-educativas-para-reduzir-acidentes-com-motociclistas>. Acesso em: 5 dez. 2024.

AGÊNCIA NACIONAL DE TRANSPORTES TERRESTRES. **Programa Vias Seguras**. 2023. Disponível em: <https://www.gov.br/antt/pt-br/acesso-a-informacao/institucional/programa-prorev/programa-vias-seguras>. Acesso em: 20 nov. 2024.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 10697**: Pesquisa de sinistros de trânsito - Terminologia. 2020. Disponível em: <https://www.abramet.com.br/repo/public/commons/ABNT%20NBR10697%202020%20Acidentes%20de%20Transito%20Terminologia.pdf>. Acesso em: 27 nov. 2024.

Associação Brasileira de Transporte e Logística de Produtos Perigosos. **BR-116 é a rodovia que mais mata no Brasil, diz estudo**. Informativo ABTLP, 2020. Disponível em: <https://www.abtlp.org.br/index.php/br-116-e-rodovia-que-mais-mata-no-brasil-diz-estudo/>. Acesso em: 16 nov. 2024.

BREIMAN L. Random Forest. **Machine Learning**. Universidade da Califórnia, v 45, p 5–32, out. 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Acesso em: 17 nov. 2024.

CUNTO F.J.C. LOURERIRO C.F.G. O uso da microssimulação na avaliação do desempenho da segurança viária. **Transportes**, São Paulo, v. 19, n. 3 p. 5-11, 2011. Disponível em: <http://repositorio.ufc.br/handle/riufc/58742>. Acesso em: 3 dez. 2024.

IBM. Machine Learning. **O que é aprendizado de máquina (ML)?**. Disponível em: <https://www.ibm.com/br-pt/topics/machine-learning>. Acesso em: 5 dez. 2024.

JUNTOS NO CAMINHO. Disponível em: <https://juntosnocaminho.com.br/rodovias-mais-movimentadas-do-brasil/>. Acesso em 20 fev. 2024.

MINISTÉRIO DO TRÂNSPORTE. Departamento nacional de estradas de rodagem diretoria de desenvolvimento tecnológico. **Guia de redução de acidentes com base em medidas de engenharia de baixo custo**. Rio de Janeiro. 1998. Disponível em: https://www.gov.br/dnit/pt-br/assuntos/planejamento-e-pesquisa/ipr/coletanea-de-manuais/vigentes/703_guia_de_reducao_de_acidentes.pdf. Acesso em: 20 dez. 2024.

MINISTÉRIO DO TRANSPORTE. Transporte terrestre. **Segurança viária**. 2020. Disponível em: <https://www.gov.br/transportes/pt-br/assuntos/transporte-terrestre/rodovias-federais/seguranca-viaria>. Acesso em: 3 dez. 2024.

MOREIRA V. As 10 rodovias mais perigosas do Brasil. **MobiAuto**, nov. 2023. Disponível em: <https://www.mobiauto.com.br/revista/as-10-rodovias-mais-perigosas-do-brasil/3639>. Acesso em: 5 dez. 2024.

NODARI C.T. LINDAU L.A. Auditoria da segurança viária. **Transportes**, v. 9 n. 2, 2001. Disponível em: <https://revistatransportes.org.br/anpet/article/view/170>. Acesso em: 3 dez. 2024.

NODARI C.T. **Método de avaliação da segurança potencial de segmentos rodoviários rurais de pista simples**. 2003. 210 f. Tese (Doutorado em Engenharia de Produção) – Universidade Federal do Rio Grande do Sul. Porto Alegre 2003. Disponível em: <https://lume.ufrgs.br/handle/10183/3675>. Acesso em: 20 dez. 2024.

PINHEIRO, A. V. S.; SILVA, M. F. de S.; SALOMÃO, P. E. Análise de Patologias no Pavimento Flexível da BR116 Trecho Sudeste: Teófilo Otoni – Itambacuri (Minas Gerais / Brasil). **Revista Multidisciplinar do Nordeste Mineiro**, [S. l.], v. 1, n. 1, 2023. Disponível em: <https://revista.unipacto.com.br/index.php/multidisciplinar/article/view/830>. Acesso em: 19 dez. 2024.

POLÍCIA RODOVIÁRIA FEDERAL. Dados Abertos da PRF. **Boletim de Acidente de Trânsito**. 2024. Disponível em: <https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-da-prf>. Acesso em: 17 nov. 2024.

POLÍCIA RODOVIÁRIA FEDERAL. Segurança viária .**PRF aponta alta de mortes e lesões nas rodovias federais paranaenses em 2024**. 2024. Disponível em: <https://www.gov.br/prf/pt-br/noticias/estaduais/parana/2024/maio/prf-aponta-alta-de-mortes-e-lesoes-nas-rodovias-federais-paranaenses-em-2024>. Acesso em: 27 nov. 2024.

REIS C.V.T. SILVA J.T.M. MAIA L.C.G. O Uso da Descoberta de Conhecimento em Banco de Dados nos Acidentes da Br-381. **Encontro Nacional de Pesquisa em Ciência da Informação**, 16. João Pessoa. 2015. Disponível em: <https://brapci.inf.br/#/v/187669>. Acesso em: 16 nov. 2024.

RIBEIRO DE CARVALHO C.H. Custos dos Acidentes de Trânsito no Brasil. **Instituto de Pesquisa Econômica Aplicada**, Brasília, 2020. Texto para discussão. Disponível em: <https://www.ipea.gov.br/atlasviolencia/arquivos/artigos/7018-td2565.pdf>. Acesso em: 23 nov. 2024.

SOFTONIC. Weka. Disponível em: <https://weka.softonic.com.br>. Acesso em: 20 dez. 2024.

SOUZA. A. Knowledge Discovery in Databases (KDD): Revelando Perspectivas a partir de Dados. **Medium**, jun. 2023. Disponível em: <https://medium.com/blog-do->

zouza/knowledge-discovery-in-databases-kdd-462ea2775715. Acesso em: 3 dez. 2024.

SOUZA A.A. **Análise dos acidentes nas rodovias federais concedidas no Brasil**. 2018. 127 f. Dissertação (Mestrado em Geotecnia e Transportes) - Universidade Federal de Minas Gerais. Belo Horizonte. 2018. Disponível em: <http://hdl.handle.net/1843/RAOA-BATJRD>. Acesso em 20 fev. 2024

SOUZA E.D. DIAS E.W. NASSIF M.E. A GESTÃO DA INFORMAÇÃO E DO CONHECIMENTO NA CIÊNCIA DA INFORMAÇÃO: perspectivas Teóricas e Práticas Organizacionais **inf. & Soc.** João Pessoa, v.21, n.1, p. 55-70, jan./abr. 2011. Disponível em: https://www.researchgate.net/profile/Edivanio-Souza/publication/293150407_Information_and_knowledge_management_on_information_science_Theoretical_perspectives_and_organizational_practices/links/5ccc5ca492851c4eab80f1e0/Information-and-knowledge-management-on-information-science-Theoretical-perspectives-and-organizational-practices.pdf. Acesso em: 5 dez. 2024.

ORGANIZAÇÃO PAN-AMERICANA DE SAÚDE. **Segurança no trânsito**. Disponível em: <https://www.paho.org/pt/topicos/seguranca-no-transito>. Acesso em: 28 nov. 2024.

PORTAL IPIRANGA. **Conheça as 7 Principais Rodovias do Brasil**. Disponível em: <https://portal.ipiranga.wps/portal/pt-br/ipiranga/blog-ipiranga/categorias/dicas-para-motoristas/conheca-as-7-principais-rodovias-do-brasil#:~:text=A%20BR-116%20é%20mais,Bittencourt%20e%20Rodovia%20Presidente%20Dutra>. Acesso em: 20 dez. 2024.

TARAPANOFF, Kira (Org.). Inteligência, informação e conhecimento [em corporações]. Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia: Ibict; UNESCO, 2006. 453 p. Disponível em: <http://livroaberto.ibict.br/handle/1/465>. Acesso em: 5 dez. 2024.

WEKA. Waikato Environment for Knowledge Analysis. Universidade de Waikato. Programa de Computador. Disponível em: <https://ml.cms.waikato.ac.nz/weka/downloading.html>. Acesso em: 8 jul. 2024.

WEKA WIKI. Development. Disponível em: <https://waikato.github.io/weka-wiki/development/>. Acesso em: 20 dez. 2024.