UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

ROMÁRIO DE JESUS MENEZES

MODELAGEM PREDITIVA DA OCORRÊNCIA DE ACIDENTES GRAVES NA BR-101 EM SERGIPE UTILIZANDO ÁRVORES DE DECISÃO

São Cristóvão – SE

2025

ROMÁRIO DE JESUS MENEZES

MODELAGEM PREDITIVA DA OCORRÊNCIA DE ACIDENTES GRAVES NA BR-101 EM SERGIPE UTILIZANDO ÁRVORES DE DECISÃO

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharelado em Ciências Atuariais.

Orientador (a): Prof. Dr. Luiz Henrique Dore

São Cristóvão – SE

2025

ROMÁRIO DE JESUS MENEZES

MODELAGEM PREDITIVA DA OCORRÊNCIA DE ACIDENTES GRAVES NA BR-101 EM SERGIPE UTILIZANDO ÁRVORES DE DECISÃO

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como um dos pré-requisitos para obtenção do grau de Bacharelado em Ciências Atuarias.

Aprovado em 02/04/2025, Nota Final 09.

Banca Examinadora:

Documento assinado digitalmente
gov.br
LUIZ HENRIQUE GAMA DORE DE ARAUJO
Data: 09/04/2025 08:58:32-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Luiz Henrique Dore

Orientador

Documento assinado digitalmente
gov.br
ALLAN ROBERT DA SILVA
Data: 08/04/2025 21:34:25-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Allan Robert da Silva

1° Examinador

Documento assinado digitalmente
gov.br
CLEBER MARTINS XAVIER
Data: 08/04/2025 21:11:22-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cleber Martins Xavier

2° Examinador

AGRADECIMENTOS

Primeiramente, gostaria de expressar minha profunda gratidão a Deus, por me dar força e sabedoria para superar os desafios e alcançar este marco em minha vida.

Agradeço à minha família, que sempre esteve ao meu lado, a minha mãe, uma mulher guerreira que sempre lutou pelo meu melhor, me apoiando em cada passo desta jornada. Você é a minha rocha e a minha inspiração.

Aos meus amigos, que foram mais do que colegas de classe, vocês foram a minha família longe de casa. Vocês estiveram comigo em cada passo desta jornada, compartilhando risadas, incontáveis noites de estudo. Cada um de vocês contribuiu para a minha experiência e crescimento de uma maneira única e valiosa. Agradeço a todos vocês por sua amizade, apoio e por todos os momentos memoráveis que compartilhamos juntos. Vocês tornaram esta jornada muito mais rica e significativa.

Aos meus amigos que estão fora do ambiente acadêmico, Tiago, Thiago Antônio, Natália, Alex, Matheus, Vinicius, Guilherme. Vocês me proporcionaram momentos de riso e alegria, servindo como um lembrete constante do mundo além dos livros e das salas de aula. Cada conversa, cada encontro, cada memória compartilhada com vocês me ajudou a manter a perspectiva e a equilibrar minha vida. Vocês me mostraram o valor da amizade verdadeira e do apoio incondicional. Agradeço a todos vocês por estarem sempre lá, por me fazerem rir quando eu mais precisava e por serem uma parte inestimável da minha vida.

Por fim, mas não menos importante, agradeço aos meus professores. Vocês me desafiaram, me orientaram e me inspiraram. Vocês não apenas me ensinaram matérias, mas também lições valiosas de vida.

Obrigado!

RESUMO

Este estudo tem como objetivo analisar os fatores que mais influenciam a sobrevivência de pessoas envolvidas em acidentes de trânsito na BR-101, no estado de Sergipe, a partir dos dados fornecidos pela Polícia Rodoviária Federal (PRF). O conjunto de dados utilizado contém informações detalhadas sobre os acidentes, incluindo condições climáticas, estado da via, tipo e quantidade de pistas, horário dos acidentes, tipo de colisão e gravidade, entre outros aspectos relevantes. A análise foca na classificação dos acidentes quanto à sobrevivência das vítimas, diferenciando os casos “Sem Vítimas” e “Com Vítimas”. O objetivo principal é identificar como fatores ambientais e estruturais afetam a probabilidade de sobrevivência. Foram utilizados modelos como *CART* (*Classification and Regression Trees*), *Bagging* e *Random Forest*. Os resultados mostraram que, embora o modelo CART tenha alcançado a maior acurácia 0,676, ele apresentou baixo desempenho na identificação de acidentes graves, com *recall* de apenas 0,098. A versão balanceada do *CART* melhorou o recall 0,329, mas reduziu a acurácia. Já os métodos de agregação, como *Bagging* e *Random Forest*, demonstraram melhor capacidade de identificar casos graves, com recalls superiores a 0,57 e F1-Scores em torno de 0,45, destacando-se como as abordagens mais eficazes para o problema. A partir dessa abordagem, o estudo pretende fornecer uma análise detalhada dos acidentes na BR-101 em Sergipe, auxiliando gestores públicos na implementação de políticas de segurança viária e na definição de medidas preventivas prioritárias para reduzir a gravidade dos acidentes e aumentar a segurança nas rodovias.

Palavra-chave: Árvores de decisão, Acidentes de trânsito, Sobrevivência, Fatores de risco

Abstract

This study aims to analyze the factors that most influence the survival of people involved in traffic accidents on BR-101, in the state of Sergipe, based on data provided by the Federal Highway Police (PRF). The dataset used contains detailed information about the accidents, including weather conditions, road conditions, type and number of lanes, time of accidents, type of collision and severity, among other relevant aspects. The analysis focuses on the classification of accidents according to the survival of victims, differentiating between “No Victims” and “With Victims” cases. The main objective is to identify how environmental and structural factors affect the probability of survival. Models such as CART (Classification and Regression Trees), Bagging and Random Forest were used. The results showed that, although the CART model achieved the highest accuracy of 0.676, it performed poorly in identifying serious accidents, with a recall of only 0.098. The balanced version of CART improved the recall by 0.329, but reduced the accuracy. Aggregation methods, such as Bagging and Random Forest, demonstrated a better ability to identify serious cases, with recalls above 0.57 and F1-Scores around 0.45, standing out as the most effective approaches to the problem. Based on this approach, the study aims to provide a detailed analysis of accidents on BR-101 in Sergipe, assisting public managers in implementing road safety policies and defining priority preventive measures to reduce the severity of accidents and increase highway safety.

Keyword: Decision trees, Traffic accidents, Survival, Risk factors

LISTA DE ILUSTRAÇÃO

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

ABCR: Associação Brasileira de Concessionárias de Rodovias.

ASMETRO: Sistema de saúde e seguradoras de automóveis e motocicletas.

CART: Classification and Regression Trees.

CTB: Código de Trânsito Brasileiro.

CP: Parâmetro de complexidade.

DNIT: Departamento Nacional de Infraestrutura de Transportes.

KDD: Knowledge-Discovery in Databases

PRF: Polícia Rodoviária Federal.

RF: Random Forest.

SIM: Sistema de Informação sobre Mortalidade.

X-*VAL RELA TIVE ERROR*: Erro Relativo de Validação Cruzada.

1. INTRODUÇÃO

A lei que entrou em vigor em 23 de setembro de 1996, do art. 1° da Lei 9.503 no Código de Trânsito Brasileiro (CTB), define que todos os brasileiros têm o direito de gozar, em condições seguras, do trânsito tanto como motorista ou como pedestre. Sendo dever dos órgãos e entidades componentes do Sistema Nacional de Trânsito, adotar medidas com esse fim. Assim como o legislador achou por bem estabelecer para o entendimento comum a definição do que é considerado trânsito:

> *"Considera-se trânsito a utilização das vias por pessoas, veículos e animais, isolados ou em grupos, conduzidos ou não, para fins de circulação, parada, estacionamento e operação de carga ou descarga" (BRASIL, 1996, p. 1).*

Contudo, apesar da lei e dos esforços institucionais para garantir os direitos dos cidadãos, existem fatores que contribuem para a ocorrência desses sinistros, tais como a imprudência de trânsito, bebidas alcoólicas, legislações vigentes para a melhoria do tráfego de veículos, fiscalização, entre outros obstáculos que afetam a segurança das vias, tornando-as dificultosas para deslocamentos (Ferraz et al., 2012).

Acidentes nas rodovias brasileiras têm aumentado por três anos consecutivos desde 2021 (SCHUINSKI, 2024), mesmo com medidas para a diminuição de acidentes, bem como fiscalização, modernização em equipamentos e novas tecnologias implementadas pelo poder público. Porém, a falha humana é um dos fatores mais recorrentes para acidentes nas estradas brasileiras (G1, 2023), e o que mais agrava o aumento dos números de ocorrência de acidentes são os períodos de festas, mais precisamente os de finais de ano (Portaldotransito, 2024).

Os acidentes de trânsito em solo brasileiro causam um elevado número de mortes. Nos anos anteriores a 2017, o Brasil detinha números alarmantes, totalizando 150 mil os números de mortos e feridos anuais, colocando o país na quinta posição no ranking de maior mortalidade por lesões por consequência de acidentes de trânsito (Costa et al. 2017). Segundo os indicadores obtidos nos registros de ocorrências da Polícia Rodoviária Federal (PRF) que são disponibilizados, um dos principais fatores responsáveis por esses acidentes é: comportamento inadequado dos motoristas, o qual é ainda agravado pelas precárias condições das vias, as condições meteorológicas dentre outras causas (BRASIL, 2011).

Este trabalho tem como principal objeto a investigação e identificação da análise

de possíveis fatores particulares de risco de acidente na BR-101 no estado de Sergipe. A árvore de decisão é uma técnica de aprendizado de máquina amplamente utilizada para a tomada de decisões e análise de dados, que pode ser aplicada para identificar padrões e fatores de risco, como os associados a acidentes de trânsito. Ao aplicar essa ferramenta ao contexto de acidentes na rodovia de Sergipe, é possível segmentar variáveis e a partir dela construir um modelo visual de decisões com base em características específicas. Em uma árvore de decisão, cada "nó" representa uma condição ou pergunta sobre uma variável, enquanto as "ramificações" indicam os resultados possíveis com base nas respostas.

Nessa análise, as variáveis a serem exploradas referentes à classificação do acidente são: condição das vias, dia da semana, sentido da via, tipo de pista, uso do solo, fase do dia, condição meteorológica. Assim, a proposta deste estudo é examinar os fatores de risco relacionados a acidentes de rodovias, visando identificar padrões que possam auxiliar na criação de estratégias de prevenção e redução desses incidentes. Os acidentes de rodovias constituem um sério problema de saúde pública, causando perdas humanas, além de impactos econômicos e sociais.

Assim, a importância social e econômica dos acidentes nas rodovias, que ceifam vidas e causam ferimentos diariamente, torna necessário que essa questão seja abordada com cuidado e que sejam elaboradas soluções direcionadas pelo governo. Esses acontecimentos impactam a saúde pública, como também a econômica, por gerar altos custos no sistema de saúde e seguradoras de automóveis e motocicletas (Asmetro, 2021). Por isso, buscar contribuir para o debate por considerar o tema de suma importância social, econômica.

2. OBJETIVOS

O presente capítulo visa estabelecer com clareza os objetivos da pesquisa a ser desenvolvida. Os objetivos serão introduzidos em duas sessões: geral e específica, proporcionando uma compreensão abrangente dos propósitos deste trabalho.

2.1 Geral

Analisar os fatores de risco associados aos acidentes da BR-101 no Estado de Sergipe, utilizando modelos preditivos baseados em árvores de decisão, *randomforest*, *bagging* e *ROC* para identificar padrões relacionados à ocorrência de classificação de acidentes, e analisar quais acidentes tiveram vítimas graves ou não na BR-101.

2.2 Objetivos Específicos

1. Construir base de dados sobre os acidentes na BR-101 no Estado de Sergipe.
2. Minerar dados para extrair padrões que possibilitem a aplicação das técnicas de árvore de decisão.
3. Identificar, a partir da árvore de decisão, as variáveis independentes que mais influenciam na variável dependente. Explorar como o período do acidente, condições meteorológicas e das pistas influenciamna classificação dos acidentes, sendo classificados como: graves ou não graves.

3. JUSTIFICATIVA

A escolha do tema desta pesquisa é justificada pela necessidade de fornecer uma previsão de acidentes de trânsito no Nordeste brasileiro, com foco na segurança viária, um assunto de importância em âmbito das políticas públicas, gestão de tráfego e econômia. Dados disponibilizados pela Polícia Rodoviária Estadual (BPRV) informam que, no ano de 2024, houve uma redução de 4,88% no número de acidentes nas rodovias estaduais de Sergipe comparado ao ano anterior. Em consonância com esse dado, o mesmo período registrou um aumento de 20% no número de vítimas fatais e de 3,04% no de feridos. Esses números mostram que, embora a quantidade de ocorrências tenha diminuído, a gravidade dos acidentes aumentou. Entre as principais causas, destacam-se o consumo de álcool e o excesso de velocidade, mas variáveis como condições climáticas, tipo de pista e dia da semana também podem influenciar esses eventos(BRPv, 2024).

Os recursos disponíveis para intervenções em segurança de trânsito são limitados, o que torna inviável investir em todos os trechos rodoviários. Identificar esses trechos que mais oferecem riscos é uma tarefa complexa. Para isso, este trabalho propõe o uso de uma abordagem estatística baseada em árvores de decisão, uma técnica capaz de mapear relações de variáveis como: ano, mês, dia da semana, fase do dia, condição meteorológica, sentido da via, tipo de pista, traçado da via e uso do solo. Diferentemente de métodos tradicionais, que podem não captar interações mais intrincadas, a árvore permite identificar subgrupos de risco.

A aplicação desse modelo preditivo contribui para o avanço do conhecimento na área de análise de dados aplicada à segurança viária. Por meio de uma abordagem acessível, é possível modelar padrão, apoiando a formulação de políticas públicas e estratégias de intervenção. Além disso, este estudo busca preencher lacunas no entendimento atual sobre acidentes rodoviários no Nordeste brasileiro, uma região onde os fatores geográficos, comportamentais e estruturais ainda são pouco explorados de forma integrada.

4. REVISÃO LITERÁRIA

Neste capítulo, vamos analisar algumas etapas. Primeiro, apresentar uma pequena explicação geral à literatura sobre acidentes nas rodovias. Na seção 4.3 exibe-se alguns trabalhos que mostram a árvore de decisão como uma ferramenta para ajudar na resolução desse problema. As árvores de decisão são estruturas que ajudam na tomada de decisão, sua abordagem visual ajuda a mapear opções viáveis e resultados, assim facilitando na escolha da melhor solução.

4.1 ACIDENTES DE RODOVIAS NO BRASIL

Para ICMR (2009), a segurança do trânsito envolve o desenvolvimento de transportes que sejam sustentáveis, nesse caso, nas rodovias onde acidentes causam mortes em vários países, causando destruição nas famílias e impactos econômicos (SATRIA; CASTRO, 2016). GAN et al (2020) expõem que os acidentes de trânsito produzem efeitos negativos para a sociedade, isso demonstra que eles não são adequados para ter um desenvolvimento no sistema de transporte.

Para Adanu et al. (2018), os acidentes de trânsito referem-se à ocorrência de prejuízos a indivíduos ou bens, resultantes da interação de fatores dinâmicos, como pessoas, veículos, vias e o ambiente ao redor. Logo, os indicadores extraídos no registro de ocorrências da Polícia Rodoviária Federal (PRF), que são disponibilizados para atender à Lei de Acesso à Informação (12.527/2011), demonstram que um dos fatores que mais causam acidentes está ligado aos condutores, às más condições das vias, à falta de sinalização e à falta de atenção.

Para Chong et al. (2005), o uso de técnicas de aprendizado de máquina para modelagem dos dados sobre acidentes de rodovias ajuda a compreender as características, como condições meteorológicas, as vias e as variáveis como tipo de pista que estavam relacionadas aos acidentes. Além disso, neste estudo, as árvores de decisão são utilizadas como um método estatístico muito comum na mineração de dados para seu modo de abordagem com muitas informações e apresentam formas para dar explicações mais aparentes e intuitivas. Logo, tendo em vista a maneira gráfica como as árvores de decisão se apresentam no processo decisório e as afinidades entre os diferentes fatores que se classificam em um acidente, como condição das vias, dia da semana, sentido da via, tipo de pista, uso do solo, fase do dia, condição meteorológica, classificação de acidente.

4.2 FATORES DE RISCO ASSOCIADOS AOS ACIDENTES DE RODOVIAS

Conforme Silva (2020, p. 58), a desatenção ao volante, como o uso de celular e ultrapassagens indevidas, está entre as principais causas de colisões fatais. Acidentes nas rodovias do Brasil são especialmente causados por diversos fatores, sendo a imprudência, o álcool e o excesso de velocidade um dos principais fatores para isso acontecer. As condições das pistas também são um fator bastante preocupante.

De acordo com Nogueira (2019, p. 75), a fiscalização ainda é insuficiente para coibir essa prática, resultando em milhares de vítimas todos os anos. Ferreira (2017) critica as rodovias que não têm sinalização e têm pavimentação precária, elevando os riscos, especialmente em áreas de grande fluxo de veículos pesados.

Segundo Lima et al. (2021, p. 30), a educação no trânsito, aliada a políticas públicas eficazes, pode reduzir significativamente os acidentes. Para diminuir esses riscos, especialistas desses casos fazem defesa em aumentar a fiscalização, aumentar a campanha de conscientização e investir em infraestrutura.

4.2.1 O PAPEL DAS RODOVIAS FEDERAIS (BRS) NA OCORRÊNCIA DE ACIDENTES

As rodovias no Brasil desempenham um papel crucial na mobilidade do país, fazendo a conexão de Estados e deixando mais fácil o transporte de cargas grandes e de pessoas que se deslocam. Porém, nessas vias infelizmente ocorre um número de acidentes bastante elevado, e muitas das vezes resulta com vítimas fatais. São vários fatores que acarretam para isso ocorrer, como a carência de reformas em rodovias e o alto fluxo de veículos pesados, unido à falta de responsabilidade dos condutores. O DNIT (Departamento Nacional de Infraestrutura de Transportes) mostrou um relatório bastante preocupante sobre “mais de 50% dos acidentes graves no Brasil ocorrem em rodovias federais, com destaque para colisões frontais e saídas de pista”(IPEA, 2018).

A conservação das rodovias é um dos principais fatores que aumentam o risco de acidentes. Segundo Nogueira (2020, p. 85), a precariedade da infraestrutura rodoviária brasileira eleva o número de acidentes, especialmente em trechos onde há baixa visibilidade e falta de manutenção asfáltica. As condições das rodovias deixam tudo fica mais difícil para o motorista que precisa ter bastante cuidado, caso ocorra uma

desatenção, aumenta o risco de perder o controle dos veículos, principalmente sob condições climáticas.

Além do mais, não basta a infraestrutura ser precária, como também muitos veículos pesados nas rodovias contribuem para ocorrer mais acidentes. Caminhões e carretas constituem uma parte considerável do tráfego nas BRs, o que traz desafios adicionais, como o aumento do tempo necessário para frenagens e a realização de ultrapassagens arriscadas. Conforme Lima (2019, p. 47), as rodovias brasileiras foram projetadas para um volume de tráfego menor do que o atual, levando a estradas sobrecarregadas e mais propensas a colisões entre veículos de diferentes tamanhos.

4.3 ÁRVORES DE DECISÃO COMO FERRAMENTA DE ANÁLISE PREDITIVA

As árvores de decisão são bastante usadas para análise preditiva de acidentes, pois elas ajudam a classificar dados e identificar padrões que irão contribuir para a ocorrência de sinistros. Para Silva (2021), a técnica usada é baseada no algoritmo de aprendizado supervisionado, a estrutura de dados, assim facilitando a interpretação dos fatores de risco. Conforme Nogueira (2019, p. 75), as árvores são eficazes para prevenir as variáveis associadas a colisões fatais, assim permitindo ações mais preventivas.

Essa abordagem analisa múltiplas variáveis simultaneamente, como condições climáticas, horário, tipo de via e comportamento do condutor. Estudos indicam que a combinação de árvores de decisão com bancos de dados georreferenciados melhora a precisão na identificação de pontos críticos de acidentes (LIMA et al., 2020). Com isso, órgãos de trânsito podem direcionar investimentos em infraestrutura e fiscalização para áreas mais vulneráveis.

5. METODOLOGIA

Este capítulo apresenta os dados coletados e as tecnologias utilizadas na pesquisa. Adotamos uma abordagem qualitativa, descritiva e exploratória, com o objetivo de analisar os acidentes graves ou não na BR-101. O estudo foi conduzido por meio de uma combinação de técnicas de mineração de dados, modelos preditivos e árvores de decisão.

5.1 Tipo de Pesquisa

Esta pesquisa é documental, pois ela se baseia na coleta de dados, organização e análise, provenientes de registros oficiais relacionados a acidentes de rodovias. No contexto deste estudo, os dados documentais serão obtidos a partir de fontes oficiais, como:

Registros da Polícia Rodoviária Federal (PRF), que disponibilizam informações detalhadas sobre acidentes de trânsito em rodovias federais, incluindo número de vítimas fatais, ilesos, veículos envolvidos e localização por BR;

A aplicação de documentos oficiais permite uma análise detalhada da realidade dos acidentes das rodovias. Além disso, adota uma abordagem sistemática, conseguindo ter uma visão mais clara, mas também coletar dados confiáveis que ajudam a construir modelos preditivos. Assim, eles vão seguindo procedimentos de organização, limpeza e análise estatística para garantir a integridade e a validade dos resultados obtidos.

5.2 Coleta de Dados

A fonte dos dados explorados neste trabalho foram os registros de acidentes disponibilizados abertamente pela Polícia Rodoviária Federal (PRF), em seu site na internet: www.prf.gov.br. Segundo o próprio portal da página, os dados citados estão isentos de qualquer restrição de licenças ou mecanismo de controle. Dessa forma, os dados abertos são convenientes para pesquisas.

Ademais, os dados sobre acidentes se encontram agrupado: O dicionário das variáveis está no (apêndice 9, p. 53)

Período do acidente, condições meteorológicas e das pistas influenciam na classificação dos acidentes, sendo classificados como: graves ou não graves. Acidentes agrupados por ocorrência: cada instância representa a ocorrência de único acidente, dados sobre as circunstâncias do acontecimento estão presentes, mas não há campos

sobre os dados pessoais dos envolvidos, apenas quantificações quanto ao número de feridos, mortos, etc.

Optou-se por utilizar nessa pesquisa o segundo agrupamento de dados, uma vez que as características dos indivíduos relacionados aos acidentes foram julgadas como importantes fatores de interesse. Além disso, dados associados às circunstâncias das ocorrências também estão presentes, isso possibilita uma análise mais completa no tocante aos relacionamentos entre fatores extrínsecos e intrínsecos aos acidentados.

5.3 Procedimentos Metodológicos

O processo de análise seguirá as seguintes etapas:

1. **Seleção e organização dos dados**: Importação dos conjuntos de dados, limpeza e pré-processamento, eliminando inconsistências e tratando valores ausentes;
2. **Análise exploratória dos dados (EDA)**: uso de estatísticas descritivas e visualização de dados para identificar padrões preliminares e *outliers*;
3. **Modelagem preditiva**: Aplicação do algoritmo de árvores de decisão (como *DecisionTreeClassifier*no Rstudio) para identificar as variáveis mais relevantes associadas à ocorrência e gravidade dos acidentes;
4. **Avaliação do modelo**: Validação dos resultados por meio de métricas como acurácia, precisão, recall e matriz de confusão, além da interpretação dos resultados obtidos.

5.4 Processos de Descoberta de Padrões em Bancos de Dados

De acordo com Fayyad et al. (1996), a mineração de dados constitui uma etapa do processo de conhecimento em Bancos de Dados (*Knowledge Discovery in Databases*), também denominado KDD. O KDD é um processo de extração de informações úteis, dividido em cinco etapas: seleção de dados, pré-processamento, transformação, mineração e avaliação. Conforme Cardoso e Machado (2008), essa técnica permite a obtenção de conhecimento em bases de dados, possibilitando a descoberta de conhecimento implícito no agrupamento de dados.

A análise de grandes quantidades de dados se torna inviável para qualquer pessoa sem o auxílio de ferramentas computacionais apropriadas. Portanto, é imprescindível a utilização de ferramentas que possibilitem a análise, a interpretação e a

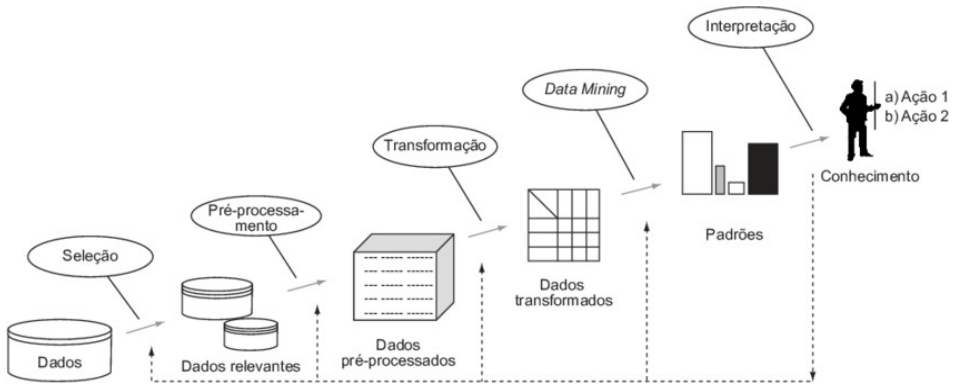
relação entre dados,de modo a elaborar e escolher as estratégias de ação mais eficazes. Para Goldschmidt (2015), a área surge como descoberta de conhecimento em banco de dados, que inclui a Mineração de Dados (*Data Mining*) como uma de suas fases.

Soczek (2014) afirma, com os progressos tecnológicos observados desde o início da década de 2000, relacionados à execução de bancos de dados, inteligência computacional, o uso de procedimentos como a Mineração de Dados se tornou cada vez mais necessário. Este tipo de abordagem é fundamental para descobrir informações necessárias que contribuam na tomada de decisões em situações de insegurança.

Goldschmidt (2015) também diz que a Mineração de Dados é o procedimento que constrói uma estrutura de informação que é vista, examinada e compreendida já no pós-processamento. O autor relata que, neste momento, os resultados atingidos são examinados e, com esses métodos, a fim de analisar os dados, são definidos com a ajuda de especialistas. É fundamental distinguir que a mineração de dados faz parte de um projeto maior, seu descobrimento de aprendizagem em estruturas de dados, igualmente citado como KDD. Fayyad (1996) et al. citam que, nessa edição, são encontrados autores que apontam Data Mining e KDD ao mesmo tempo como sinônimos.

Na figura 1, é possível observar como, para iniciar um processo de KDD, é preciso compreender sua aplicação e seus objetivos. Este processo é dividido em cinco fases: Seleção, pré-processamento, transformação, mineração e interpretação

Figura 1 - Fases da descoberta de conhecimento em bases de dados



Fonte: Adaptado de Fayyad*et al.* (1996)

Seleção: Definição dos dados para a análise. Pré-processamento: Limpeza e organização dos dados, eliminando inconsistências e tratando valores ausentes. Transformação:

Conversão dos dados para um formato apropriado, como normalização ou criação de novas variáveis. Mineração: Aplicação de técnicas analíticas, como algoritmos de aprendizado de máquina, para descobrir padrões e relações nos dados. Interpretação: Análise e validação dos resultados obtidos, extraindo insights valiosos para a tomada de decisões.

5.4.1 Escolha e Preparação dos Dados

Tan (2009) declara que o primeiro nível do seguimento KDD é a processamento dos dados. A fim de identificar e corrigir as informações para futuras análises. Primeiramente, é básico dispor de uma interpretação robusta do quadro de trabalho para adotar as estruturas de dados que serão empregadas na procura de informações. De acordo com Halmenschlager (2002), em certas circunstâncias, o angariamento e o grupo de dados podem se cristalizar em um obstáculo, demandando frequentemente ajustes complexos e a incorporação de conjuntos de dados relacionados. Isso pode sobrecarregar os sistemas de mineração, que não conseguem aplicar vários arquivos ao mesmo tempo.

Portanto, Hall et al (2009) faz uma sugestão de que os modelos preditivos devem incluir elementos e variáveis preditivas que atuam como objetivos para compreender as especificidades de razões que podem levar a acidentes. A imprecisão ou os erros nas informações podem prejudicar os resultados obtidos e exigir correções antes do início da análise. Outro ponto importante é a redução na dimensão. Isso inclui a exclusão de recursos não necessários e melhora o processo analítico.

5.4.2 Pré-processamento

Tan et. al. (2009) explica que, após a seleção dos dados, ocorre às etapas de pré-processamento e modificação, utilizando diversas estratégias e metodologias interligadas, o que torna os dados mais adequados para a mineração.

Castanheira (2008) diz que nesta etapa abrange intervenções de como fazer uma abordagem caso perceba a falta de dados em alguns campos, a limpeza de dados como a verificação de inconsistências, diminuição da quantidade de campos em cada registro feito, o preenchimento ou a exclusão de valores inexistentes, assim removendo os dados que são duplicados.

Logo, a limpeza é uma etapa eficaz do pré-processamento, ela faz a eliminação de valores duplicados e faltantes, reparando as falhas e utilizando os valores que estão ausentes. Caso existam dados faltantes, podem-se empregar metodologias como a imputação estatística (utilização da média, mediana ou moda), a eliminação de registros imprecisos ou a substituição por valores estimados.

5.4.3 Mineração de dados (Data Mining)

Soczek (2014) aborda como, a partir do início dos anos 2000, os avanços em execuções de bases de dados, inteligência artificial e redes de comunicação permitiram um uso sem precedentes de ferramentas, como a mineração de dados. Essa metodologia se tornou essencial para a pesquisa de informações cruciais, ajudando a orientar a tomada de decisões em situações de incerteza.

Para Han (2006) as grandes quantidades de dados existir sem ter uma ferramenta disponível para a análise podem ser observadas como "dados ricos, mas informações pobres", assim sendo é impossível um humano analisar uma quantidade absurda de dados sem o uso de uma ferramenta computacional para ajudar. Hand (2001) diz que o interesse em verificar os dados aumenta com a probabilidade de extrair informações que serão úteis para seus possíveis donos, esse seria um fator para preocupar a mineração.

Mineração de Dados é uma área da disciplina de Banco de Dados que aplica técnicas e algoritmos para extrair informações significativas de bases de dados com grande densidade. Assim, trata-se de uma das abordagens para adquirirmos conhecimento a partir dessas bases, possibilitando a descoberta de saberes que estão implícitos na coleta de dados (CARDOSO; MACHADO, 2008).

Os algoritmos utilizados na mineração de dados envolvem técnicas como estatística, aprendizado de máquina, reconhecimento de padrões, inteligência artificial, recuperação de informações, processamento de sinais e análise espacial ou temporal dos dados. Essa prática é considerada um dos mais promissores avanços interdisciplinares nas tecnologias da informação (LAROSE, 2005).

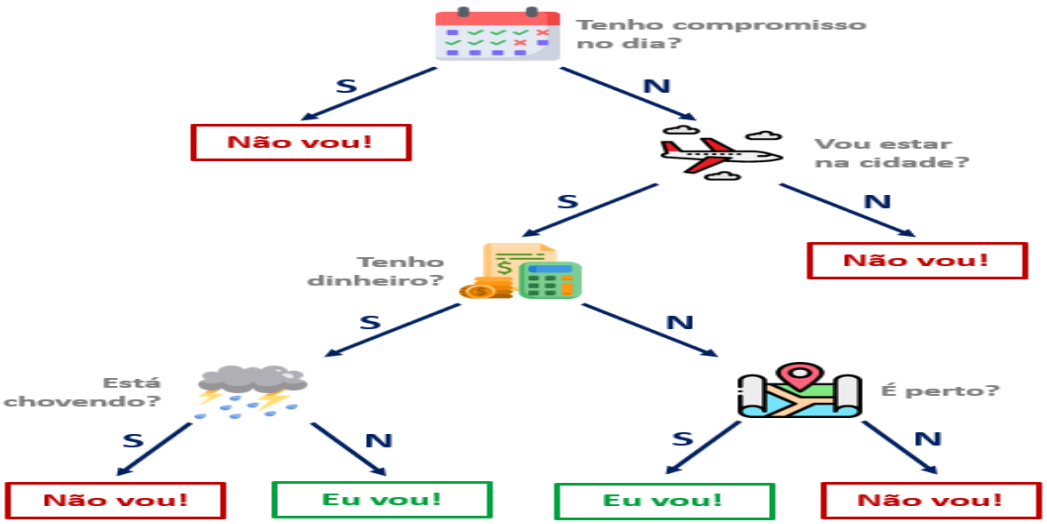
5.5 ÁRVORES DE DECISÃO PARA MINERAÇÃO DE DADOS

A árvore de decisão é um modelo utilizado para a classificação ou regressão, composto por um conjunto de nós e arcos, também chamados de ramos

(FURNKRANZ, 2012). Este modelo tem a forma de uma árvore, na qual cada nó interno representa um teste em uma característica de uma instância, enquanto os arcos simbolizam o resultado desse teste. Os nós externos, conhecidos como nós terminais ou folhas, correspondem às classes de classificação. Para classificar uma instância específica, a árvore é percorrida de maneira vertical, seguindo os arcos associados aos nós cujas características atendem aos critérios estabelecidos até se chegar a um nó-folha que contém a nova classificação da instância.

A figura 2 ilustra uma árvore de decisão, onde cada elipse marca um critério de avaliação aplicado a um conjunto de dados. Os retângulos representam as classificações finais, isto é, as decisões tomadas. O processo inicia-se na raiz e segue pela estrutura através dos testes sequenciais. Com base nas respostas recebidas, o caminho avança por diferentes ramos até alcançar um nó-folha que determina a classificação final.

Figura 2: Representação visual da árvore de decisão



Fonte: Heitor Catunda (2024).

5.5.1 CART

O Cart foi introduzido pelos estatísticos Leo Breiman, Jerone Friedman, Richard Olsen e Charles Stone em seu trabalho intitulado "*Classification andRegressionTrees*", publicado em 1984. O estudo em questão tem importância na área de aprendizado de máquina, sendo citado na literatura de mineração de dados. A sua principal função é a de criar árvores de decisão com dimensões menores e um melhor desempenho. O Cart permite a manipulação de atributos preditivos, sejam eles categóricos ou quantitativos,

para o particionamento que envolve atributos categóricos, todas as combinações possíveis para formar dois subconjuntos são testadas.

A árvore é construída por meio de um processo de divisão binária, onde cada nó é dividido em dois subconjuntos. À medida que a árvore é percorrida da raiz até as folhas, são feitas perguntas simples de sim ou não. Quando a recursividade ocorre no subconjunto gerado, quando não é mais viável realizar novas divisões da árvore. Ela é baseada no critério Gini e Entropia quando trabalha com dados qualitativos.

5.5.2 IMPUREZA DE GINI

A impureza do Gini é a probabilidade de classificar incorretamente o ponto de dados aleatório no conjunto de dados se ele for rotulado com base na distribuição de classe do conjunto de dados. O *CART* utiliza o Gini Index para medir o índice de impureza *do dataset* a ser analisado, este cálculo é determinado por:

$$GI = 1 - \sum_{i}^{\square} (p_i)^2$$

GI = impureza de Gini

pi = proporção de elementos da classe i no conjunto de dados

Se todas as amostras pertencem a uma única classe (pi=1 para uma classe e 0 para as outras), então GI=0, indicando pureza máxima. Se as classes estão distribuídas igualmente (por exemplo, 50% de uma classe e 50% de outra), a impureza de Gini é máxima.

5.5.3 COMPLEXIDADE ALGORITMO *CART*

A complexidade de uma árvore de decisão está diretamente relacionada à sua profundidade e ao número de nós. Durante o treinamento, o algoritmo divide os dados recursivamente, resultando em um tempo computacional médio de O (n log n), mas que pode chegar a $O(n^2)$ no pior caso, quando os cortes são desbalanceados. Já a predição tem complexidade O(d), onde d é a profundidade da árvore. Overfitting ocorre quando um modelo se torna excessivamente ajustado aos dados de treinamento, assimilando também os ruídos e particularidades que não aparecem em dados reais ou futuros. Como

consequência, ele apresenta um desempenho excelente nos dados de treino, mas falha ao lidar com novos dados (teste). Para evitar o *overfitting* e melhorar a eficiência, técnicas como poda e limitação da profundidade são utilizadas (Breiman, L., 1984).

Número máximo de nós internos:

$N = 2^d - 1$, onde d é a profundidade da árvore.

Complexidade no pior caso:

O(N^2), quando os cortes são desbalanceados.

Complexidade média (balanceada):

O(*nlogn*), onde n é o número de amostras no conjunto de dados.

Na validação cruzada, o conjunto de dados é dividido em k subconjuntos ou dobras. O modelo é treinado em k-1 dessas dobras e testado nas dobras restante. Esse processo é repetido várias vezes, com cada dobra sendo usado uma vez como conjunto de teste.

5.5.5 RAZÃO DE GANHO

Para Lin (2011), o atributo que proporciona uma máxima razão de ganho é selecionado como a raiz da árvore de decisão. A seguir, procura-se ordenar qual nó se resolverá, desse modo a árvore será arquitetada de forma que pode ser repetida inúmeras vezes na raiz até as folhas. Cada nó interior estima-se uma característica; cada ramo da árvore representa um valor dessa propriedade e cada folha contribui com uma classificação (SILBERSCHATZ et al., 2006).

Com base nisso, a Razão de Ganho é dada pela Eq.:

$$\text{RazãoGanho (S, A)} = \frac{Ganho(S, A)}{Gini(S)}$$

Ganho(S,A) = o quanto a impureza diminui ao dividir por A.

Gini(S) = mede a impureza do conjunto antes da divisão.

5.5.6 Processo de Poda

A poda é uma etapa crucial no processo de construção de árvores de decisão, pois busca restringir as dimensões da árvore ao eliminar segmentos que não agregam valor à precisão da classificação. Com isso, resulta em uma estrutura mais simples e

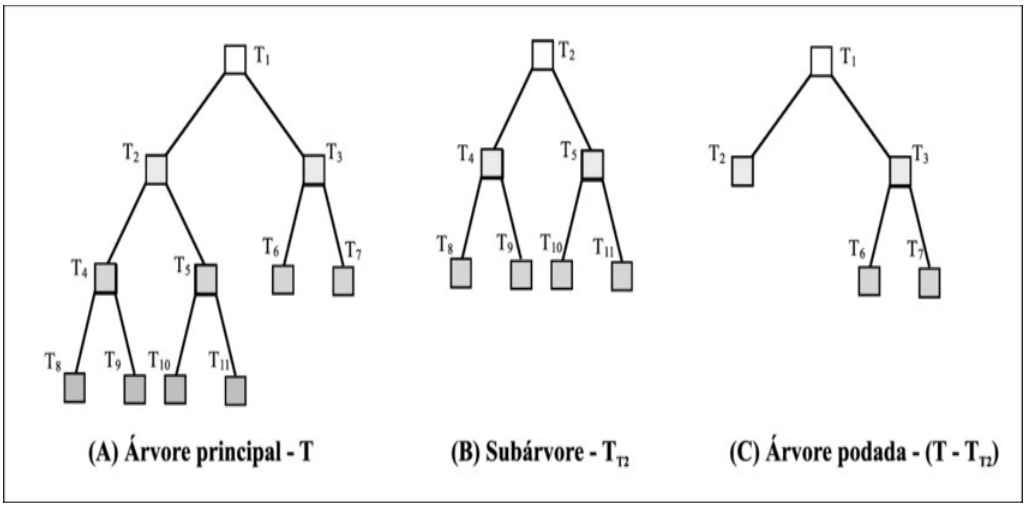
compreensível, além de melhorar seu desempenho. Para lidar com o problema de sobrecarga e precisão na classificação, as técnicas de poda podem ser implementadas de duas maneiras: a primeira é conhecida como pré-poda, que interrompe a construção da árvore ao atingir um critério de parada. A segunda, chamada de pós-poda, é realizada somente após a finalização da árvore, ou seja, quando todos os exemplos do conjunto já foram distribuídos ao longo dela, com o objetivo de reduzir suas dimensões até atingirem níveis ideais. Ambas as abordagens de poda têm como objetivo eliminar partes da árvore que não são essenciais para a precisão da classificação, resultando em estruturas menos complexas. Enquanto a pré-poda evita a criação de uma árvore que será destruída posteriormente, a pós-poda é considerada mais confiável, pois utiliza todos os exemplos durante sua construção.

SGARBI (2007) fala que, após a conclusão do processo de construção da árvore, inicia-se a fase “poda”. Esta fase envolve a remoção de partes da árvore que não contribuem para a correta classificação dos dados, tornando a árvore menos complexa e mais compreensível. Isso pode ser feito utilizando duas abordagens diferentes (SGARBI, 2007; CASTANHEIRA, 2008):

> *1. Durante o treinamento de dados, é usado um método conhecido como poda de redução de erros. Não dividir mais o conjunto de treinamento de dados.*
>
> *2. Pós-poda é o processo de remoção de estruturas de uma árvore após a construção.*

De acordo com Castanheira (2008), a metodologia empregada foi a poda posterior através do algoritmo J48, implementação do algoritmo C4.5. Embora essa abordagem tenha um custo computacional superior em relação à poda com redução de erros, ela proporciona as melhorias na árvore.O critério poda envolve a análise da proporção de erros em cada subárvore, começando pelas folhas. Para Quinlan (1993), cada subárvore é examinada para determinar se a mudança de uma subárvore para outra resulta em uma proporção menor de erros. Caso isso ocorra, a poda ocorre naquela subárvore. Como é demonstrado na figura 3:

Figura 3: Exemplo de podar. (A) Árvore completa; (B) subárvore; e (C) árvore final após a poda.



Fonte: (Modificado de Breiman et al., 1984).

5.6 O MÉTODO DE BAGGING

O método de *bootstrapaggregating*, frequentemente chamado de *bagging*, foi introduzido por Breiman em 1996 e tem sido amplamente utilizado na área de aprendizado de máquina. Esse procedimento envolve a criação de conjuntos de dados por meio da amostragem *bootstrap*, que são então submetidos a um determinado processo de interesse, com os resultados sendo combinados em um único resultado. As técnicas mais comuns usadas são classificações, como as árvores de decisão. O *bagging*é conhecido por sua capacidade de reduzir a variância das previsões, resultando em uma melhoria na precisão daquilo que se busca prever.

O *bagging* é uma abordagem geral que não se limita apenas às árvores de classificação e regressão, e seu objetivo é diminuir a variância de um modelo de aprendizado. Uma forma intuitiva de reduzir a variância seria criar múltiplos conjuntos de treinamento e, a partir de cada um deles, treinar regressões cujas previsões poderiam ser combinadas por meio da média, formando assim um *ensemble* de regressores (daí a denominação “método de *ensemble*”). Contudo, essa alternativa costuma ser inviável,

visto que normalmente não dispomos de vários conjuntos de dados de treinamento. A ideia central do*bagging*é criar amostras *bootstrap* a partir do conjunto de dados original. No contexto do *bagging* aplicado a um problema de regressão, geramos B amostras *bootstrap*, as quais servirão como dados de treinamento para cada regressor a ser treinado.

Assim, treinamos os regressores$\hat{\psi}$1,...,$\hat{\psi}$B e o regressor agregado é a média:

$$\hat{\psi}\, bag(x)=\frac{1}{b}\sum_{i=1}^{B} \square\, \hat{\psi}\, i(x)3$$

$\hat{\psi}\, bag$ é a previsão agregada para uma nova entrada x utilizando o modelo bagging.

B é o número de modelos (ou modelos base) no conjunto. Normalmente, B é um número grande (por exemplo, 100 ou mais árvores em um modelo Random Forest).

$\hat{\psi}\, i(x)$é a previsão do modelo i-ésimo para a entrada x. Cada modelo i é treinado em uma amostra diferente do conjunto de dados (gerada pelo processo de bootstrap).

Ao treinar regressores com o algoritmo *CART*, não devemos podar as árvores resultantes. Construímos árvores muito grandes, que, individualmente, apresentam alta variância e um viés reduzido. O método de *bagging* é responsável por minimizar essa variância. No caso de árvores de classificação, o *bagging* opera de forma semelhante. A principal diferença é que o classificador agregado $\hat{\psi}$ bag faz previsões baseadas no "voto da maioria" dos classificadores $\hat{\psi}$1, . . . ,$\hat{\psi}$B, que foram treinados com B amostras de *bootstrap*. O *bagging* proporciona significativos ganhos em desempenho preditivo ao combinar milhares de árvores, resultando em um único regressor ou classificador

5.6.1 RANDOM FOREST

Para Breiman (2001), a floresta aleatória é uma mistura de árvores de decisão, onde a árvore vai depender dos valores do vetor aleatório, sendo independente e distribuídos entre a árvore. Nesse contexto, após alcançar um determinado número de árvores, elas emitem uma predição para cada classe do problema, de acordo com o vetor de entrada. Portanto, a classe que obtiver a maior quantidade de predições é selecionada.

A *randomforest* (ou Floresta Randômica) é uma técnica de aprendizado de máquina versátil e amplamente utilizada, reconhecida por sua facilidade de uso e a capacidade de gerar bons resultados, frequentemente sem a necessidade de ajustes nos

hiperparâmetros. Este método pode ser aplicado tanto em tarefas de classificação quanto de regressão, além de mitigar problemas frequentes encontrados em outras árvores de decisão, como o *overfitting* no conjunto de treinamento. O algoritmo de Floresta Randômica, como seu nome sugere, elabora uma coleção de árvores de decisão de forma aleatória. Esta "floresta" consiste em uma combinação de várias árvores de decisão, geralmente treinadas por meio do método de *bagging*. A ideia fundamental do *bagging* é que a união de vários modelos de aprendizado pode resultar em uma performance geral superior. Assim, o algoritmo de florestas aleatórias desenvolve múltiplas árvores de decisão e as integra para alcançar previsões mais precisas e estáveis.

5.7 ROC

Em 1966, *Green* e *Swets*introduziram a análise ROC, que está associada à teoria de detecção de sinais, com suas pesquisas direcionadas à psicologia. Segundo essa teoria, os observadores escolhem uma regra de otimização para um sinal específico, visando maximizar as respostas corretas na análise (*Green andSwets*, 1966). A curva característica de operação do receptor (*ReceiverOperatingCharacteristic*, ROC) é utilizada para mensurar a precisão de medições contínuas em relação à previsão de um resultado binário. Na área da medicina, as curvas ROC possuem um histórico extenso de aplicação na avaliação de testes diagnósticos, principalmente na radiologia e em diagnósticos em geral. Além disso, as curvas ROC têm sido amplamente empregadas na teoria de detecção de sinais. As representações gráficas das curvas ROC empíricas são geralmente exibidas em um gráfico bidimensional, representando um plano unitário com medidas de probabilidade variando entre 0 e 1. Na coordenada das abcissas, é chamada de Fração de falsos positivos (FFP), enquanto na coordenada das ordenadas temos a Fração de verdadeiros positivos (FVP). Em um processo de classificação, existem quatro tipos distintos de ocorrências: quando a ocorrência é positiva e é classificada como tal, temos os Verdadeiros Positivos (VP). Se essa mesma ocorrência é classificada como negativa, surgem os Falsos Negativos (FN). Quando a ocorrência é negativa e é classificada corretamente como negativa, obtemos os Verdadeiros Negativos (VN). Por outro lado, se a ocorrência negativa é classificada como positiva, temos os Falsos Positivos (FP). que são expressas por:

$$FFP = \frac{FP}{VN + FP} \quad \text{e} \quad FVP = \frac{VP}{VP + FN}$$

Na área do diagnóstico médico, é comum que essas variáveis sejam denominadas como sensibilidade (eixo y) e 1-especificidade (eixo x). Além disso, existem duas outras frações que se referem aos verdadeiros negativos (FVN) e aos falsos negativos, conhecidos como Fração de Falsos Negativos (FFN).

$$FVN = \frac{VN}{VN + FP} \quad \text{e} \quad FFN = \frac{FN}{VP + FN}$$

Na análise ROC, quando tanto os verdadeiros positivos quanto os verdadeiros negativos atingem 1 (100%), considera-se que se está diante de um teste perfeito, sem qualquer tipo de erro (Pepe, 2011).

Acurância mede a proporção de acertos do modelo (tanto positivos quanto negativos) sobre o total de previsões feitas.

Precisão mede a proporção de acertos entre todas as vezes que o modelo previu positivo.

Recall mede a proporção de acertos entre todos os casos realmente positivos.

Tabela 1: Matriz de confusão.

		CLASSE REAL	
		POSITIVO (D)	NEGATIVO (D)
Classe Prevista	POSITIVO ¿	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	NEGATIVO ¿	Falso Negativo (FN)	Verdadeiro Negativo (VN)

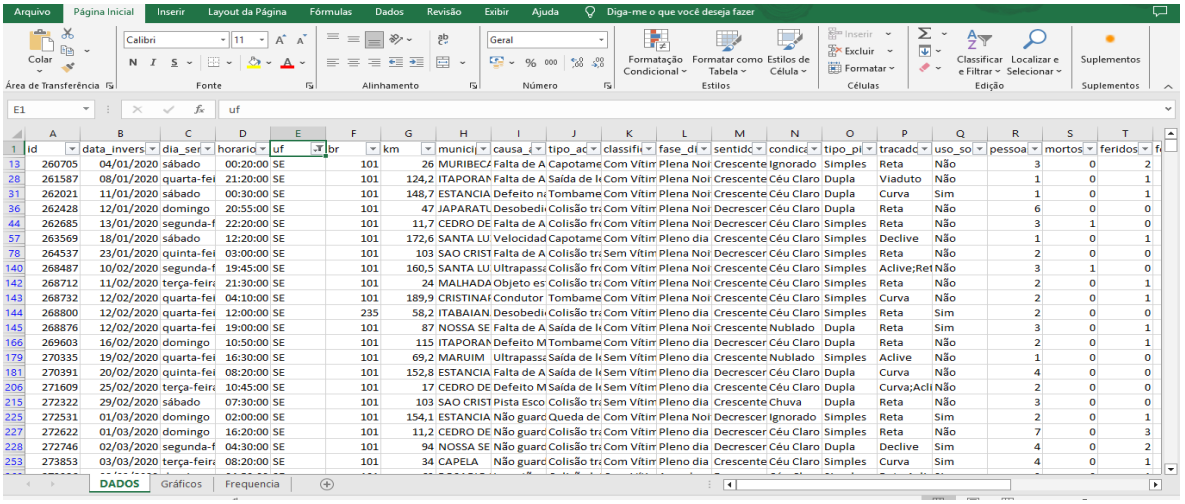
Fonte: (Fawcett, 2003)

5.8 Ferramentas de Análise de Dados

Os cálculos realizados nesse trabalho foram feitos utilizando a linguagem R, através da plataforma computacional própria e do jupyter notebook, na plataforma Anaconda (R, 2024; Distribuição de software Anaconda, 2020).

Excel para organização preliminar de dados, fazendo análise descritiva e observando o tamanho do banco de dados.

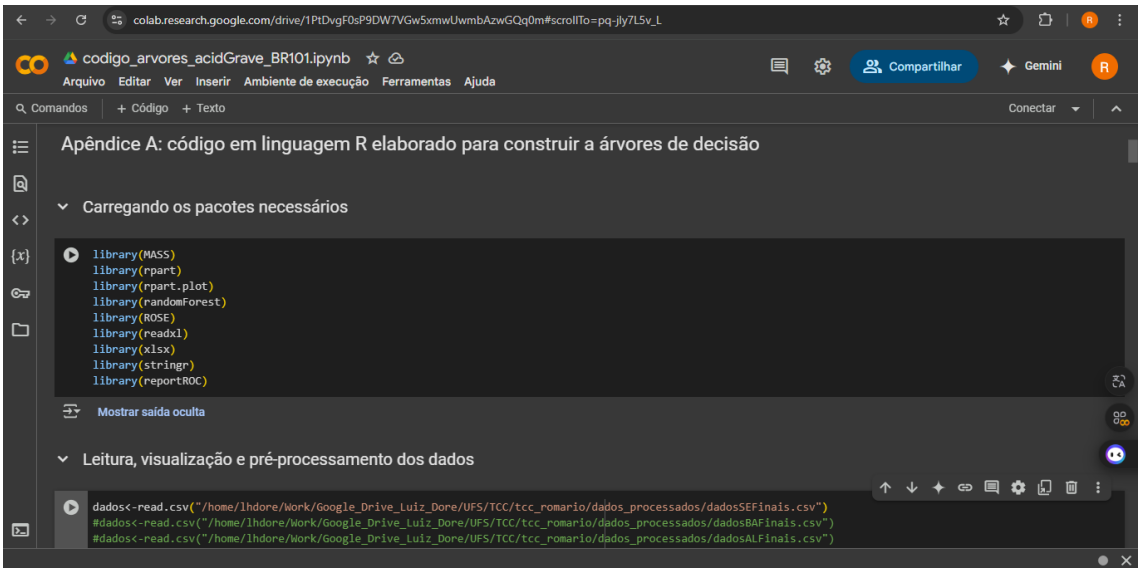
Figura 4: Banco de dados PRF no Excel.



Fonte: autor (2025)

Com uma interface clara e fácil de usar, esta ferramenta organiza de maneira eficiente blocos de código e gráficos em formato de documento, permitindo também a adição de comentários, imagens, vídeos e até equações matemáticas, proporcionando uma descrição mais completa. Isso torna os resultados e os métodos empregados muito mais apresentáveis. A figura ilustra a interface do jupyter Notebook. É importante ressaltar que esse ambiente é compatível com outras linguagens de programação como a linguagem R, por exemplo *rpart e randomforest* (modelagem preditiva com árvores de decisão), *rpart.plot*(Facilita a visualização de árvores de decisão geradas com *rpart*), rose (Pacote para lidar com desbalanceamento de classes em aprendizado de máquina) e *MASS*(Fornece funções para estatística aplicada moderna, incluindo métodos para regressão robusta).

Figura 5: Base sendo rodada no Google Colab.



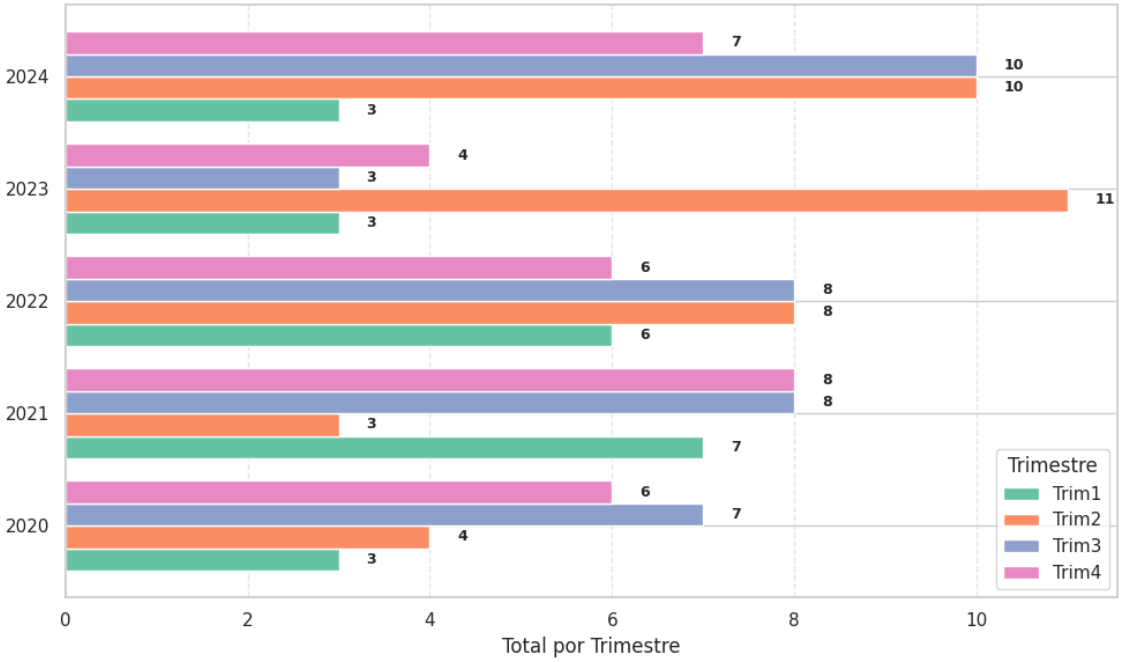
Fonte: autor (2025)

6. RESULTADOS

6.1 Análise descritiva

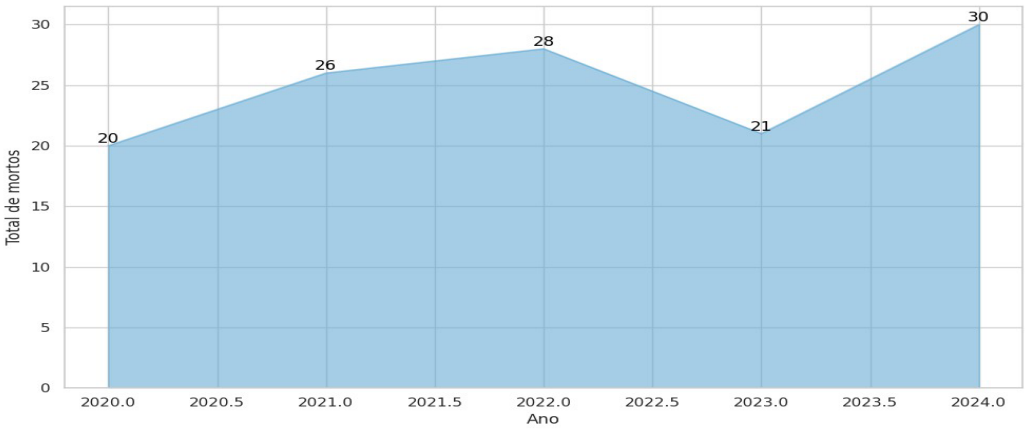
Os resultados desta pesquisa foram baseados em dados obtidos através das bases da Polícia Rodoviária Federal (PRF) para o período de 2020 a 2024, abrangendo os estados da Bahia, Sergipe e Alagoas. A análise envolveu o levantamento de informações sobre o número total de acidentes, vítimas fatais, feridos (ilesos e graves), veículos envolvidos e as rodovias federais mais afetadas. São 9 variaveis independentes e 1 variavel alvo sendo dependente.

Figura 6: Gráfico de barras mostrando a Br-101 com mais mortes durante os trimestres por ano.



A figura 6 é um gráfico de barras horizontais com os totais de mortos por trimestre, observa-se que os anos que tem mais quedas em mortes na BR-101 foi entre 2020 e 2023, enquanto 2020, 2021 e 2024 o número de mortos subiu.

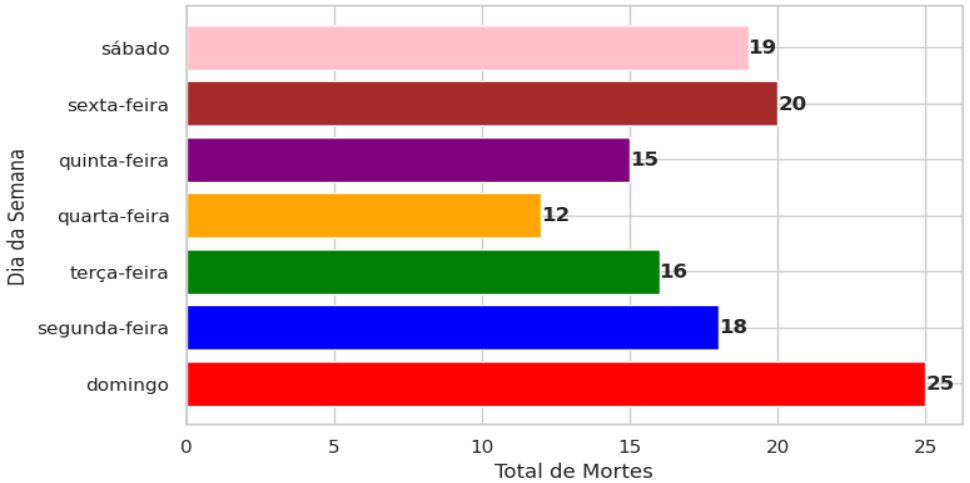
Figura 7: Gráfico de área mostrando a Br-101 com mais mortes durante os anos.



Fonte: Elaboração própria (2025)

A figura 7 em Sergipe Em Sergipe, o número de mortes nas rodovias federais (BRs) apresentou um aumento significativo nos últimos anos. Em 2020, houve 20 mortes em comparação com 2021 que teve 26, esse número cresceu mais na BR-101 indo para 28 mortes em 2022. No entanto, apesar da redução no número de mortes em 2023, com aproximadamente 21 mortes, em 2024 esse número saltou para 30. Uma das rodovias mais afetadas é a BR-101.

Figura 8: gráfico de barras dos dias da semana com mais mortes em 2020 a 2024

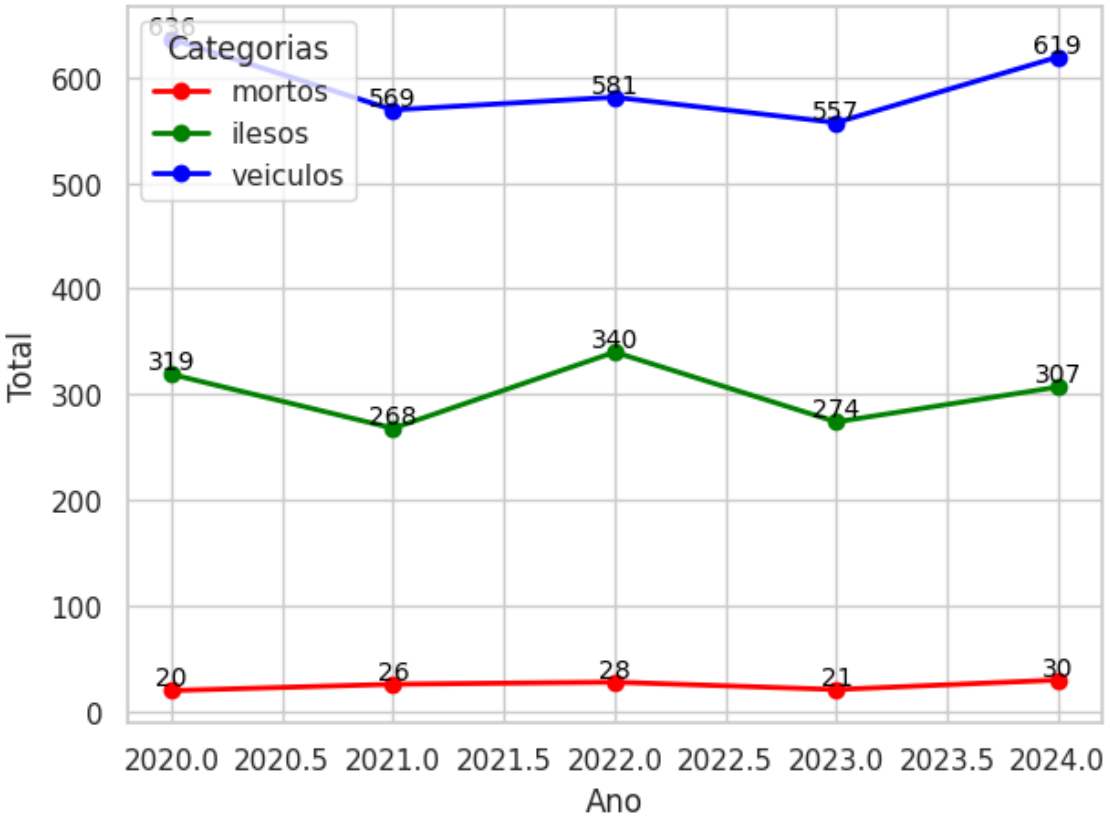


Fonte: Elaboração própria (2025)

A figura 8 mostra a soma de mortes nos dias da semana, ele demonstra que mais ocorrências acontecem no Sexta, Sábado e Domingo, isso indica uma maior incidência de acidentes fatais nos finais de semana e na sexta-feira. Isso se deve ao maior volume

de tráfego, aumento de consumo de bebida alcoólica e velocidade excessiva.

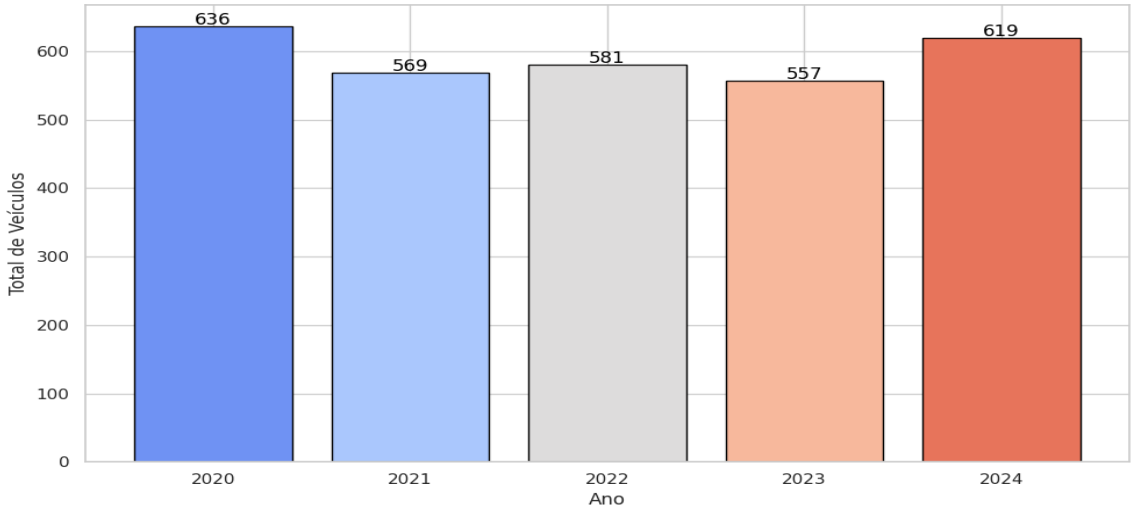
Figura 9: o gráfico de linha que mostra a soma de ilesos, veículos envolvidos em acidentes e a soma de mortos no Estado de Sergipe entre 2020 e 2024.



Fonte: Elaboração própria (2025)

A figura 9 mostra a evolução de veículos envolvidos em acidentes na BR-101. Em 2020, teve 636 acidentes e, logo nos anos de 2021 a 2023, esses números diminuíram e, em 2024, esse número aumentou novamente. O total de ilesos em 2020, 2022 e 2024 tem um número bastante significativo e em 2021 e 2023 esses números caem, já o de mortos em 2021, 2022 e 2024 tiveram maiores índices enquanto 2020 e 2023 números menores.

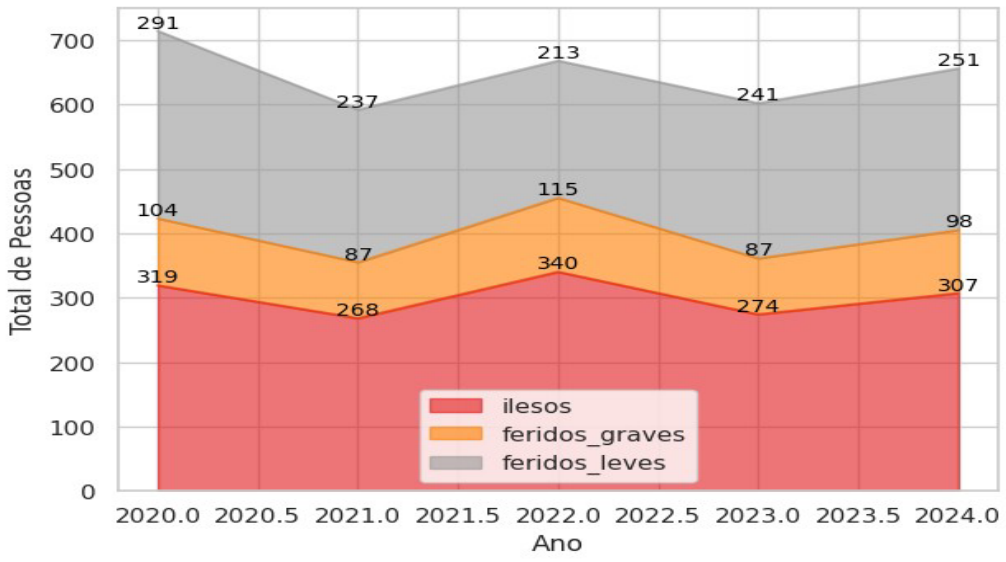
Figura 10: Gráfico de coluna soma de veículos envolvidos em acidentes entre 2020 e 2024.



Fonte: Elaboração própria (2025)

A figura 10 apresenta o total de veículos registrado em Sergipe (SE). Observa-se que o número de veículos envolvidos em acidente em 2020 era de 636, em 2021 número começou a diminuir para 569 e em 2022 aumentou mais, em 2023 deu uma leve queda e em 2024 esse número saltou para 619 veículos envolvidos.

Figura 11: Gráfico área de ilesos e feridos graves e leves em Sergipe nos respectivos anos.



Fonte: Elaboração própria (2025)

A figura 11 demonstra a soma de ilesos e feridos graves e leves em Sergipe de 2020 até 2024, como pode observa em 2020 tem 319 pessoas que saíram ilesos enquanto 104 tiveram feridos graves e 291 feridos leves, entre 2021 e 2023 tem uma taxa menor de ocorrência por ilesos e 2022 e 2024 essa taxa aumenta. Enquanto a taxa de feridos graves e leves no ano de 2021 e 2023 tem uma pequena queda e depois cresce entre 2022 e 2024.

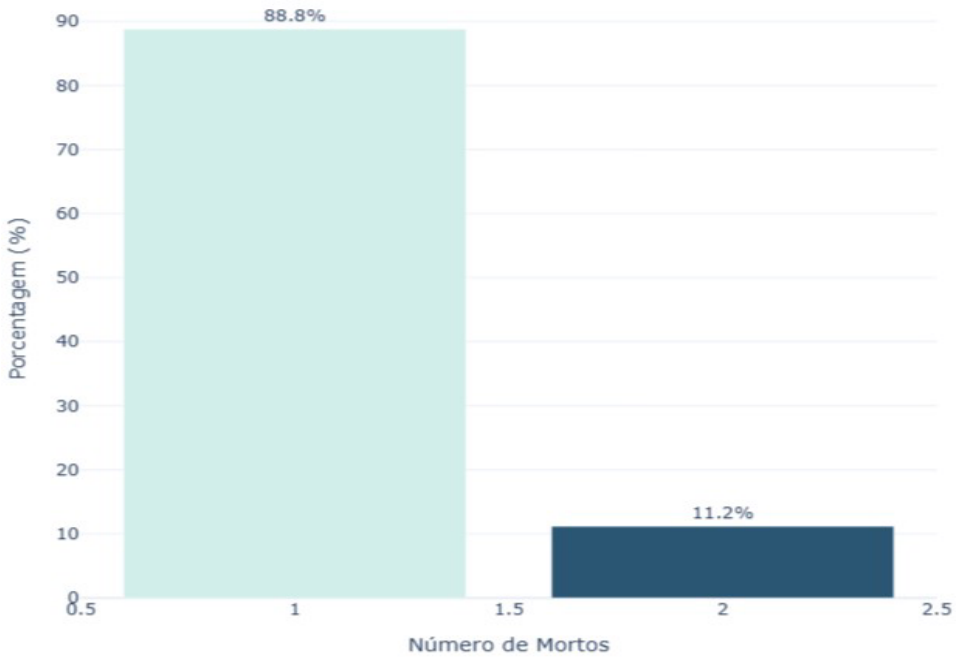
Tabela 2: Tabela de frequência

SE	101	Frequência
0	0	0,00%
1	111	88,8%
2	11	11.2%

Elaboração própria (2025)

Tabela 2 de frequência do número de mortos citados nos dados, como pode ser analisado 1 morte é citada 111 vezes enquanto 2 mortes é citada 11 vezes.

Figura 12: Gráfico de coluna mostrando o percentual de mortes.



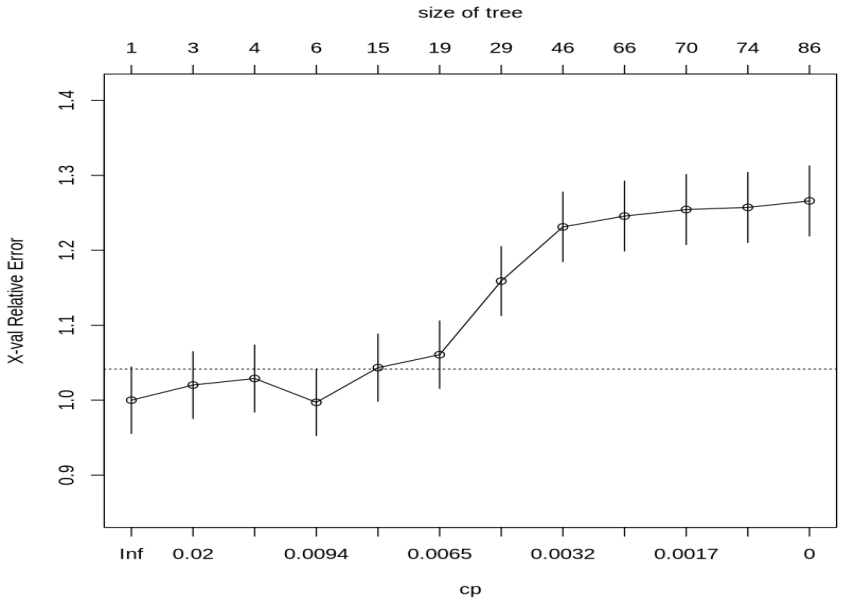
Fonte: Elaboração própria (2025)

Visualmente, a figura 12 mostra barras laranja que representam a frequência

absoluta e uma linha cinza que representa o percentual. Assim, 1 morte é relatada 111 vezes, sendo 88,8% dos dados. Podemos observar que a Categoria 1 domina tanto em frequência quanto em percentual, já 2 mortes são relatadas 11 vezes, sendo 11,2% e sendo a menor quantidade de eventos.

6.2 Resultados e Discussões

Figura 13: Relação entre o parâmetro de complexidade (CP) e o erro relativo à validação cruzada para uma árvore de decisão.



Fonte: Elaboração própria (2025)

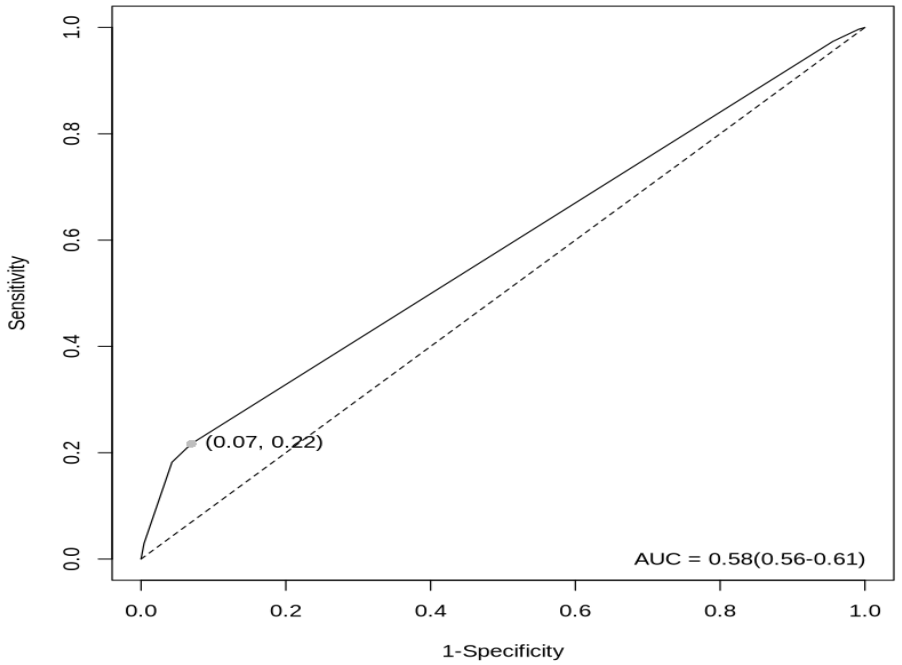
A figura 13 do Eixo X (CP - *ComplexityParameter*) - está representando o parâmetro de complexidade da árvore, os valores maiores de CP resultam em árvores menores e simples, enquanto os menores CP resultam em árvores maiores e mais complexas. Eixo Y (X-*val Rela tive Error* - Erro Relativo de Validação Cruzada) - está medido o erro da árvore em validação cruzada, isso quer dizer que quanto menor o valor, melhor a árvore generaliza os dados.

Linha com círculos, esses pontos representam uma árvore treinada com um valor específico de CP, o erro diminui até certo ponto, depois começa a aumentar, indicando overfitting. Barras verticais (intervalos de erro) indicam a incerteza do erro da validação cruzada. Se dois pontos estiverem sobrepostos, a diferença entre eles pode não ser significativa. O CP com o melhor desempenho está próximo de 0.0094 ou 0.0065, pois

ele minimiza o erro.

6.2.1 DETERMINANDO A PROBABILIDADE DE CORTE PELA CURVA *ROC*

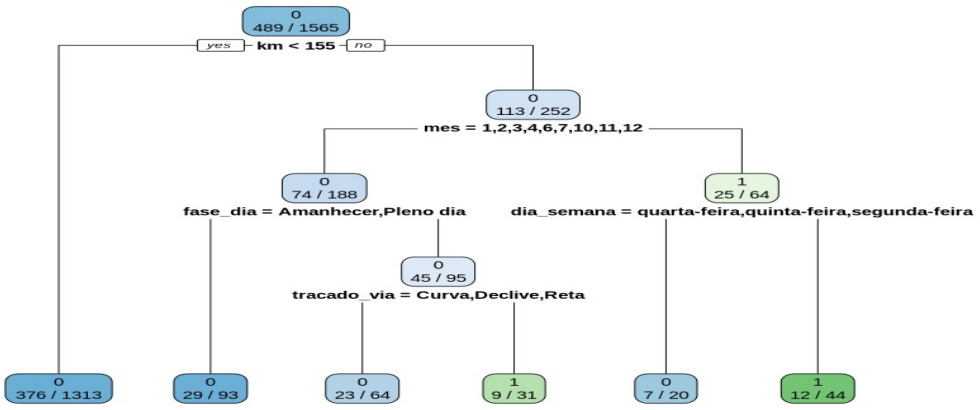
Figura 14: Análise da curva ROC onde passa a probabilidade de corte utilizada nos 4 modelos.



Fonte: Elaboração própria (2025)

A análise da curva *ROC* retornou uma probabilidade de corte, que está sendo utilizada de forma consistente nos quatro procedimentos analisados: *CART, CART balanceado*, *Random Forest (RF) e Bagging.* A análise da curva *ROC* permitiu determinar a probabilidade de corte mais adequada para a classificação dos acidentes graves. O ponto de corte identificado foi de 0.329, esse valor foi incorporado como referência central na avaliação de desempenho do modelo, sendo utilizado em todas as etapas do estudo para a classificação dos acidentes. Dessa forma, a escolha desse *threshold*(limite) possibilitou um equilíbrio entre sensibilidade e especificidade, garantindo uma melhor distinção entre acidentes graves e não graves.

Figura 15: Árvore de decisão podada.



Font

e: Elaboração própria (2025)

A árvore de decisão apresentada na figura 15 busca classificar eventos, sendo acidentes graves, com base em diferentes variáveis, como quilometragem da rodovia, mês do ano, fase do dia, dia da semana e traçado da via. A estrutura da árvore indica padrões importantes que podem contribuir para a compreensão dos fatores de risco.O primeiro fator determinante identificado é a quilometragem (km) da rodovia. Se o valor for menor que 155, a maior parte dos casos pertence à classe 0, indicando um menor risco de acidentes graves. Já quando km ≥ 155, a análise se torna mais detalhada e envolve outras variáveis para uma classificação mais precisa.Entre os fatores que contribuem para um aumento do risco, observa-se que em determinados meses do ano (janeiro, fevereiro, março, abril, junho, julho, outubro, novembro e dezembro) a ocorrência de acidentes graves pode ser mais expressiva. Dentro desse contexto, a fase do dia também se mostra relevante: quando os acidentes acontecem ao amanhecer ou pleno dia, o traçado da via desempenha um papel fundamental. Vias com curvas, declives ou retas tendem a ter um maior número de registros de acidentes graves, especialmente quando comparadas a outros tipos de traçados.

Além disso, outro aspecto importante identificado na análise é a influência do dia da semana. Às segundas, quartas e quintas-feiras, há um aumento da incidência de casos classificados como 1 (acidentes graves). Isso pode estar relacionado a um maior

fluxo de veículos nesses dias, seja por razões laborais ou condições específicas do tráfego.

Com base nessa árvore de decisão, é possível identificar áreas e condições onde o risco de acidentes graves é mais elevado, permitindo ações preventivas mais eficazes. Medidas como reforço da sinalização, fiscalização mais intensa em determinados trechos e horários e campanhas educativas podem ser implementadas para mitigar esses riscos e aumentar a segurança nas rodovias.

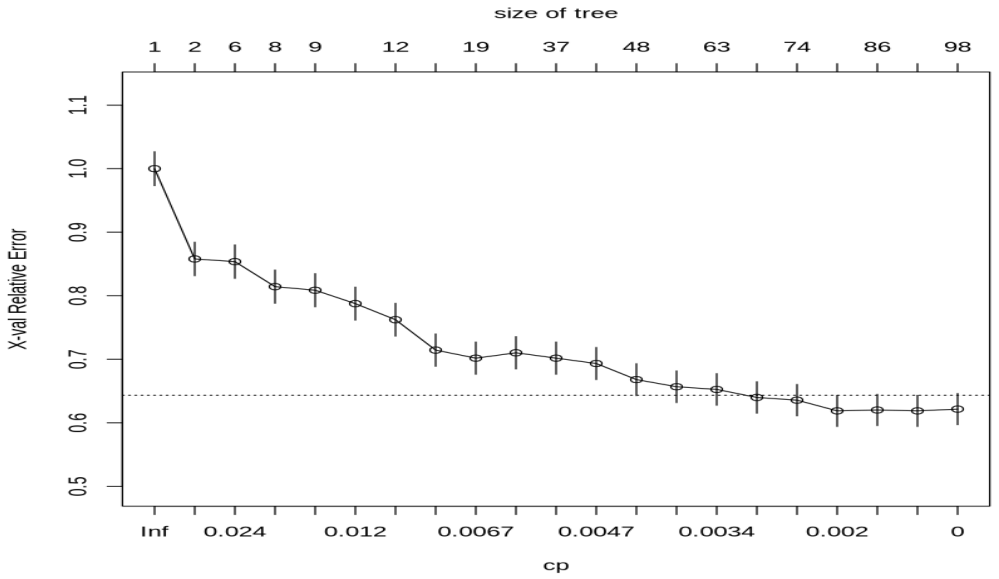
6.2.2 DADOS BALANCEADOS

Os dados apresentados indicam a quantidade de ocorrências em duas categorias: casos não graves (classe 0) e casos graves (classe 1). Os valores são:

- 750 casos não graves
- 711 casos graves

Essa distribuição mostra que as duas classes estão quase equilibradas, com uma diferença de apenas 39 ocorrências entre elas. Em termos percentuais, a classe 0 representa cerca de 51,3% dos casos, enquanto a classe 1 representa 48,7%. Essa proximidade entre as classes caracteriza um conjunto de dados balanceado.

Figura 16: Relação entre o parâmetro de complexidade (CP) e o erro relativo de validação cruzada depois do balanceamento para uma árvore de decisão.

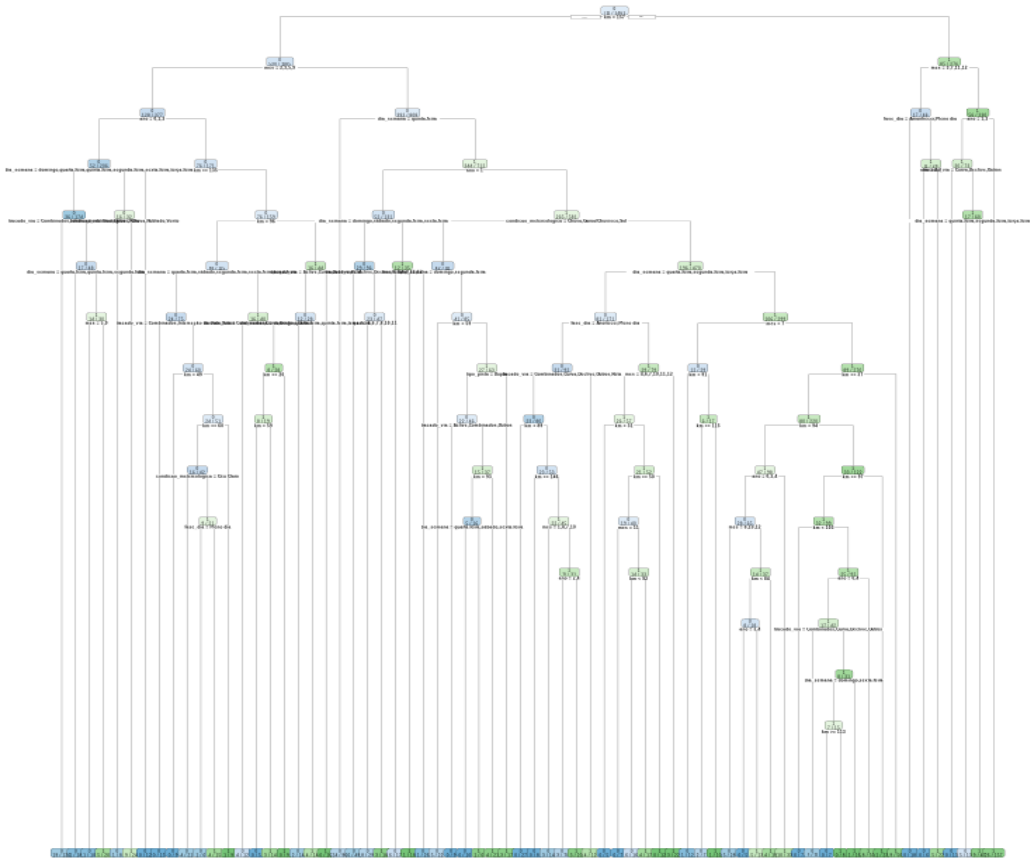


Fonte: Elaboração própria (2025)

A figura 16 exibe a relação entre o parâmetro de complexidade (cp) e o erro relativo da árvore de decisão, adquirido pela validação cruzada. O eixo x simula os

valores de cp, enquanto o eixo y exibe o erro relativo. Na parte superior, temos o número de folhas (nós terminais) da árvore correspondente a cada valor de cp. Assim, mostra que a árvore de decisão melhora seu desempenho conforme cresce, mas após um certo ponto, o ganho se estabiliza. Um bom ponto de corte para cp estaria na região onde o erro é minimizado sem um crescimento excessivo da árvore, garantindo um modelo equilibrado entre precisão e generalização.

Figura 17: Árvore de decisão depois de ter aplicado o balanceamento dos dados.



Fonte: Elaboração própria (2025)

A figura 17 Depois de balancear os dados, a árvore de decisão provavelmente mudou de algumas maneiras, a árvore se tornou mais complexa, com mais perguntas e mais caminhos a seguir. Ela teve decisões mais justas, onde a árvore provavelmente está tomando decisões mais justas, sem favorecer um tipo de decisão em detrimento de outro. Teve melhor desempenho geral do modelo, melhorou especialmente na capacidade de tomar decisões sobre os tipos de dados que antes eram menos comuns.

A figura 17 mostra uma árvore de decisão com muitos ramos e nós, indicando

que é um modelo complexo. A coloração alternada entre azul e verde sugere que a árvore está lidando com dois tipos principais de decisões. A distribuição mais uniforme das cores na parte inferior da imagem indica que o balanceamento dos dados teve umefeito positivo, resultando em decisões mais equilibradas.

6.2.3 TREINAMENTO DA ÁRVORE UTILIZANDO *BAGGING*.

No método de treinamento de balanceamento da árvore, foi utilizado *o bagging*, nesse exemplo foram utilizadas 5000 árvores, garantindo uma árvore robusta. O *mtry* faz uma determinação no número de variáveis selecionadas de forma aleatória para dividir os nós. No caso do modelo *bagging*, todas as variáveis disponíveis devem ser usadas em cada divisão, foi utilizado o valor de mtry igual ao número total de variáveis do conjunto de dados, ou seja, todas as variáveis foram consideradas como candidatas em cada divisão da árvore.. Usando o *nodesize,* demonstra-se o tamanho mínimo dos nós nos terminais, enquanto o *cutoff* define os pontos de corte para classificação das classes.

No *bagging*, criam várias amostras *bootstrap* a partir do conjunto de dados do treinamento, assim como o treinamento de uma árvore de decisão para cada amostra. Sendo contrário à floresta aleatória (*Random Forest*), onde nela apenas um subconjunto das variáveis é usado em cada divisão dos nós, no *bagging* todas as variáveis são usadas no processo.

Quando o modelo é treinado revelando a estimativa do erro fora da amostra (*Out-of-Bag error, OOB*) foide 44,89%, indicando um desempenho moderado na classificação. A matriz de confusão apresentou os seguintes resultados:

Tabela 03: Matriz de confusão com o modelo *Bagging*

	Não graves	Graves	*class.error*
Não graves	420	330	0.4400000
Graves	162	184	0.4682081

Fonte: Elaboração própria (2025)

A matriz teve 420 observações da classe de não graves que foram classificadas corretamente, enquanto 330 não foram classificadas corretamente. Na classe de graves, 184 observações foram previstas corretamente, e 162 foram incorretamente atribuídas a classe de não graves. O erro estimado para a classe 0 foi de 44%, enquanto a classe de graves foi de 46,82%.

6.2.4 TREINAMENTO DA ÁRVORE UTILIZANDO FLORESTAS ALEATÓRIAS.

Na floresta aleatória, ajustado com os seguintes parâmetros: 5.000 replicações de árvore, número de variáveis por divisão foi 3 com a seguinte fórmula:

$$\sqrt{p} = 3$$

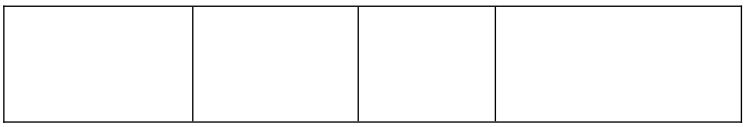
$$p = 3^2 = 9$$

Para cada amostra *bootstrap* selecionada, uma árvore é construída. Entretanto, cada vez que a divisão de um nó vai ser avaliada, apenas uma amostra das variáveis, selecionada aleatoriamente, é considerada. Portanto, no caso do método das Florestas Aleatórias, o argumento *mtry* pode assumir qualquer valor entre 1 e o número de variáveis disponíveis menos 1.

Quando o modelo é treinado revelando a estimativa do erro fora da amostra (*Out-of-Bag error, OOB*) *foid* e 46,62%, indicando um desempenho moderado na classificação. A matriz de confusão apresentou os seguintes resultados:

Tabela 04: Matriz de confusão usando *Random Forest*.

	Não graves	Graves	*class.error*
Não graves	393	357	0.4760000
Graves	154	192	0.4450867



Fonte: Elaboração própria (2025)

A matriz teve 393 observações da classe de não graves que foram classificadas corretamente, enquanto 357 não foram classificadas corretamente. Na classe de graves, 192 observações foram previstas corretamente, e 154 foram incorretamente atribuídas a classe de não graves. O erro estimado para a classe de não graves foi de 47,6%, enquanto a classe de graves foi de 44,51%.

6.2.5 COMPARANDO OS DESEMPENHOS DOS MODELOS.

Tabela 05: desempenho com modelo *CART, CART* Balanceado, *Bagging e RF*.

	CART	*CARTBalanceado*	*Bag*	*RF*
Acurácia	0.6759062	0.5714286	0.5842217	0.5458422
Precisão	0.3783784	0.3092105	0.3796296	0.3529412
Recall	0.0979021	0.3286713	0.5734266	0.5874126
F1-Score	0.1555556	0.3186441	0.4568245	0.4409449

Elaboração própria (2025)

O *recall* mede a capacidade do modelo de identificar corretamente os acidentes graves. Em outras palavras, ele representa a probabilidade de um acidente grave ser corretamente classificado como grave pelo modelo. Quando uma árvore de decisão é ajustada usando o algoritmo *CART* em um conjunto de dados desbalanceado, o *recall* fica muito baixo, alcançando apenas 10%. Isso significa que apenas 10% dos acidentes graves são corretamente identificados como graves. Após o balanceamento dos dados, o desempenho melhora, elevando o recall para 33% (ou seja, o modelo passa a reconhecer corretamente 33% dos acidentes graves). Além disso, ao substituir a árvore de decisão pelo modelo de *Random Forest* (*RF*), há uma melhoria significativa no *recall*, que sobe para 58,74%. Isso indica que o modelo *Random Forest* consegue identificar uma

proporção muito maior de acidentes graves corretamente.

O *Bagging* conseguiu um desempenho equilibrado, aumentando tanto a precisão quanto o *recall*, o que resultou no melhor *F1-Score* entre os modelos. Isso significa que ele conseguiu detectar mais acidentes graves sem aumentar tanto os falsos positivos. Já o *Random Forest* se destacou pelo maior *recall*, sendo o modelo que melhor identificou acidentes graves. No entanto, sua precisão e acurácia foram um pouco menores do que as do *Bagging*, indicando que ele também classificou mais acidentes leves como graves.

Ambas as técnicas combinam múltiplas árvores de decisão, reduzindo a variância do modelo e melhorando sua capacidade de detectar corretamente os acidentes mais severos. Dessa forma, o uso de *Bagging* e *Random Forest* torna a classificação de acidentes mais confiável, equilibrando a sensibilidade na identificação dos casos mais críticos com a redução de erros.

7 Conclusão

Este estudo apresentou como objetivo avaliar os fatorescomo período do acidente, condições meteorológicas e condições das pistas influenciam a gravidade dos acidentes

na BR-101 em Sergipe e estimar diversos modelos de aprendizado de máquina para classificar os eventos. Foram usadas técnicas como *CART, Bagging e Random Forest*, Sendo possível ver que o balanceamento dos dados e o uso de modelos mais robustos têm impacto na capacidade de identificar acidentes graves corretamente. Os resultados obtidos demonstraram que a árvore de decisão simples (*CART*) teve dificuldades em considerar corretamente os acidentes graves, principalmente com dados desbalanceados, onde o *recall* foi baixo com apenas 10%. Portanto, depois de aplicar técnicas de balanceamento, houve uma melhora no *recall*, mostrando que é importante tratar os dados antes de fazer modelagem. Os modelos usados como o Random Forest tiveram destaque pelo maior recall 58,74% de acertos ao identificar acidentes graves. O método *Bagging* teve equilíbrio na precisão e *recall*, tendo melhor *F1-Score* indicando um desempenho bom na classificação dos acidentes. Assim, os resultados podem contribuir para os órgãos responsáveis pela segurança, tendo uma reformulação em políticas públicas para reduzir os índices de acidentes graves. Logo, adotar medidas preventivas como sinalização e fiscalização no trecho da BR-101 onde há mais acidentes críticos, pode ser um artifício para potencializar essas abordagens.

7.1 Limitações da pesquisa

Alguns aspectos envolvidos na pesquisa tornaram-se fatores limitantes, dentre os quais destacam-se:

a) Limitação na disponibilidade e qualidade dos dados, o que pode ter influenciado a precisão dos modelos.

b) Desequilíbrio das classes nos dados originais, tornando necessária a aplicação de técnicas de balanceamento.

c) Restrição na escolha de variáveis, podendo haver fatores não considerados que impactam a gravidade dos acidentes.

d) Uso de um conjunto específico de algoritmos, sem a exploração de outros métodos que poderiam melhorar o desempenho.

e) Dependência de métricas padrões de avaliação, sem considerar possíveis análises qualitativas dos erros do modelo.

f) Restrições computacionais que limitaram o ajuste fino dos hiper parâmetros e o uso de modelos mais complexos.

7.2 Sugestões para trabalhos futuros

Em face das limitações da pesquisa e dos resultados obtidos, são sugestões para trabalhos futuros:

a) Explorar fontes de dados mais amplas e detalhadas para melhorar a qualidade das previsões.

b) Testar outras técnicas de balanceamento e seleção de variáveis para otimizar a detecção deacidentesgraves.

c) Implementar modelos mais avançados, como redes neurais ou algoritmos híbridos, para aprimorar a classificação.

d) Realizar análises interpretativas dos erros do modelo, investigando padrões nos falsos positivos e falsos negativos.

8 REFERÊNCIAS BIBLIOGRÁFICAS

ADANU, E. K.; JONES, S.; ODERO, K. Identification of factors associated with road crashes among functionally classified transport modes in namibia. ***Scientific African, Elsevier,*** *v. 7, p. e00312, 2020.*

BRASIL. Constituição (2011). Lei no 12527, de 18 de novembro de 2011. Acesso A Informação. Brasília, DF.

BREIMAN, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Chapman & Hall.

BREIMAN, Leo; FRIEDMAN, Jerome H.; OLSHEN, Richard A.; STONE, Charles J. Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984.

BREIMAN, Leo. Bagging predictors. Machine Learning, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, Leo. Statistical modeling: The two cultures. Statistical Science, v. 16, n. 3, p. 199–231, 2001.

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. *Knowledge. Management using data mining: a case studyofthe Federal UniversityofLavras.* **Revista de Administração Pública**, v. 42, n. 3, p. 495–528, junho, 2008.

CASTANHEIRA, L.G. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões,** Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, Minas Gerais, 2008.

CASTANHEIRA, Luciana Gomes. **Aplicação de Técnicas de Mineração de Dados *em Problemas de Classificação de Padrões***. *Belo Horizonte: UFMG, 2008.*

CHONG, M. ; ABRAHAM, A. ; PAPRZYCKI, M. ***Traffic accident analysis using machine learning paradigms****. Informática, v. 29, n. 1, 2005.*

COSTA Medeiros, Wilma Maria, et al. "Perfil epidemiológico das vítimas de acidentes de trânsito atendidas num serviço público de emergência da região metropolitana de Natal/RN." HOLOS 7 (2017): 213-224.

Disponível em: <https://www.hashtagtreinamentos.com/arvore-decisao-ciencia-dados>. Acesso em: 14 mar. 2025.

Distribuição de software Anaconda . Software de computador. Vers. 2-2.4.0. Anaconda, nov. 2016. Web. https://anaconda.com.

DNIT – Departamento Nacional de Infraestrutura de Transportes. **Relatório de Segurança Viária nas Rodovias Federais**. Brasília: DNIT, 2021.

F5 NEWS - SERGIPE ATUALIZADO. BPRV Registra redução de acidentes nas rodovias estaduais de Sergipe em 2024. Disponível em:

<https://www.f5news.com.br/cotidiano/bprv-registra-reducao-de-acidentes-nas-rodovias-estaduais-de-sergipe-em-2024.html>. Acesso em: 14 mar. 2025.

Falha humana é uma das principais causas de acidentes em estradas brasileiras. Disponível em: <https://g1.globo.com/bom-dia-brasil/noticia/2023/03/17/falha-humana-e-uma-das-principais-causas-de-acidentes-em-estradas-brasileiras.ghtml>. Acessoem: 14 mar. 2025.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. ***Attempt to isolate mast-cell precursors based on the differential sensitivity to UV-B and X-irradiation.*** *Toxic Substances Journal, v. 13, n. 2, p. 85–95, 1994.*

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. ***From Data Mining to Knowledge Discovery in Databases.*** *p. 18, 1996.*

FERREIRA, A. **A infraestrutura viária e seu impacto nos acidentes de trânsito.**Revista de Engenharia de Transportes, v. 14, n. 2, p. 50-65, 2017.

FERRAZ, Antônio Clóvis Pinto et al. **Segurança viária**. São Carlos: Suprema Gráfica e editora, 2012.

FURNKRANZ, J.; GAMBERGER, D.; LAVRAC, N. ***Foundations of Rule Learning****. [S.l.]: Springer-Verlag Berlin, 2012.*

GAN, J.; LI, L.; ZHANG, D.; YI, Z.; XIANG, Q. ***An alternative method for traffic accident severity prediction:*** *using deep forests algorithm. Journalofadvancedtransportation, Hindawi, v. 2020.*

GERAL, S. Acidentes de trânsito custam R$ 132 bilhões para o Brasil. Disponível em: <https://asmetro.org.br/portalsn/2021/05/21/acidentes-de-transito-custam-r-132-bilhoes-para-o-brasil/>. Acesso em: 14 mar. 2025.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. ***Data mining:*** conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro: Elsevier, 2015.

GREEN, D. M., &Swets, J. A. (1966**).***Signal Detection Theory and Psychophysics.* Wiley.

HALL, M. A.; WITTEN, I. H.; FRANK, E. The WEK. ***A Data Mining Software:*** *An Update. SIGKDD Explorations, 2009.*

HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. Principles of Data Mining. Cambridge: MIT Press, 2001.

HALMENSCHLAGER, Carine. **Um Algoritmo para indução de árvores e regras de decisão.** 112 f. Tese (Mestrado em Ciências da Computação) — Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação, Porto Alegre, RS, 2002.

HAN, J.; KAMBER, M. ***Data Mining:*** *concepts and techniques. [S.l.: s.n.], 2006.*

HAN, J.; KAMBER, M.; PEI, J. ***Data Mining:*** *Concepts and Techniques. 3. ed. Morgan Kaufmann, 2011.*

ICMR. ***Development of a Feasibility Module for Road Traffic Injuries Surveillance****. v. 39, 43-47 p. 11, 34. 2009.*

LAROSE, D. T. ***Discovering Knowledge in Data:*** *an introduction to data mining, 2005.*

LIMA, R. **A influência do fluxo de veículos pesados nos índices de acidentes nas rodovias brasileiras.** Trabalho de Conclusão de Curso (Graduação em Engenharia de Transportes) – Universidade Federal do Paraná, Curitiba, 2019.

LIMA, R.; SILVA, J.; CARVALHO, M. Educação no trânsito e segurança viária. **Revista Brasileira de Mobilidade Urbana**, v. 10, n. 3, p. 20-35, 2021.

LIMA, R.; SILVA, J.; CARVALHO, M. Uso de aprendizado de máquina na previsão de acidentes de trânsito. **Revista Brasileira de Inteligência Artificial**, v. 22, n. 1, p. 45-60, 2020.

NOGUEIRA, F. Árvores de decisão na análise de acidentes rodoviários. **Revista de Engenharia de Transportes**, v. 16, n. 3, p. 70-85, 2019.

NOGUEIRA, F. Infraestrutura rodoviária e segurança viária: Um estudo das BRs brasileiras. **Revista de Mobilidade Urbana**, v. 14, n. 3, p. 75-90, 2020.

O que é uma DecisionTree? Ibm.com, 30 Jan. 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/decision-trees>. Acessoem: 14 mar. 2025.

QUINLAN, J.R.C. 4.5: ***Programming for machine learning. San Mateo:*** *Morgan Kauffmann, 302p. 1993.*

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

SATRIA, R.; CASTRO, M. ***Gis tools for analyzing accidents and road design: a review. Transportation Research Procedia****, Elsevier, v. 18, p. 242–247, 2016.*

SCHUINSKI, R. M. **Mortes no trânsito sobem pelo 3° ano seguido;** o que explica a nova alta? Disponível em: <https://www.uol.com.br/carros/noticias/redacao/2024/03/21/o-que-explica-nova-alta-de-mortes-no-transito-do-brasil.htm>. Acesso em: 14 mar. 2025.

SGARBI, J.A. **Domótica Inteligente:** Automação Residencial Baseada em Comportamento. Centro Universitário da FEI, São Bernardo do Campo, 2007.

SILBERSCHATZ, A. et al. **Sistema de Banco de Dados**. p. 781, p. 485-501, 5. ed. – Rio de Janeiro: Elsevier, 2006.

SILVA, T. **Estatísticas de acidentes de trânsito no Brasil.** Relatório Técnico do Ministério da Saúde, Brasília, 2020.

SILVA, T.; SOUZA, L. Técnicas de mineração de dados para prevenção de acidentes. **Revista de Mobilidade e Segurança Viária**, v. 12, n. 2, p. 80-95, 2021.

SOCZEK, F. C.; ORLOVSKI, R. Mineração de Dados: Conceitos e aplicação de

algoritmos em uma Base de Dados na área da saúde. **Revista Científica Semana Acadêmica**, v. 01, p. 01, 2014.

SUPER USER. **Ipea - Instituto de Pesquisa Econômica Aplicada**. Disponível em: <https://www.ipea.gov.br/portal/mestrado-profissional-em-politicas-publicas-e-desenvolvimentodesafios/index.php?option=com_content&view=article&id=3211&catid=29&Itemid=34>. Acesso em: 6 apr. 2025.

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

9 APÊNDICE

DICIONÁRIO DE VARIÁVEIS
(DADOS DO BAT – A PARTIR DE 2017)

ID VARIÁVEL	NOME DA VARIÁVEL	DESCRIÇÃO
1	*id*	Variável com valores numéricos, representando o identificador do acidente.
2	*data_inversa*	Data da ocorrência no formato dd/mm/aaaa.
3	*dia_semana*	Dia da semana da ocorrência. Ex.: Segunda, Terça, etc.
4	*horário*	Horário da ocorrência no formato hh:mm:ss.
5	*uf*	Unidade da Federação. Ex.: MG, PE, DF, etc.
6	*br*	Variável com valores numéricos, representando o identificador da BR do acidente.
7	*km*	Identificação do quilômetro onde ocorreu o acidente, com valor mínimo de 0,1 km e com a casa decimal separada por ponto.
8	*municipio*	Nome do município de ocorrência do acidente
9	*causa_acidente*	Identificação da causa principal do acidente. Neste conjunto de dados são excluídos os acidentes com a variável causa principal igual a “Não”.
10	*tipo_acidente*	Identificação do tipo de acidente. Ex.: Colisão frontal, Saída de pista, etc. Neste conjunto de dados são excluídos os tipos de acidentes com ordem maior ou igual a dois. A ordem do acidente demonstra a sequência cronológica dos tipos presentes

		na mesma ocorrência.
11	*classificação_acidente*	Classificação quanto à gravidade do acidente: Sem Vítimas, Com Vítimas Feridas, Com Vítimas Fatais e Ignorado.
12	*fase_dia*	Fase do dia no momento do acidente. Ex. Amanhecer, Pleno dia, etc.
13	*sentido_via*	Sentido da via considerando o ponto de colisão: Crescente e decrescente.
14	*condição_meteorologica*	Condição meteorológica no momento do acidente: Céu claro, chuva, vento etc.
15	*tipo_pista*	Tipo da pista considerando a quantidade de faixas: Dupla, simples ou múltipla.
16	*tracado_via*	Descrição do traçado da via.
17	*uso_solo*	Descrição sobre as características do local do acidente: Urbano=Sim;Rural=Não.
18	*pessoas*	Total de pessoas envolvidas na ocorrência.
19	*mortos*	Total de pessoas mortas envolvidas na ocorrência.
20	*feridos_leves*	Total de pessoas com ferimentos leves envolvidas na ocorrência.
21	*feridos_graves*	Total de pessoas com ferimentos graves envolvidas na ocorrência.
22	*ilesos*	Total de pessoas ilesas envolvidas na ocorrência.
23	*ignorados*	Total de pessoas envolvidas na ocorrência

		e que não se soube o estado físico.
24	*feridos*	Total de pessoas feridas envolvidas na ocorrência (é a soma dos feridos leves com os graves).
25	*veiculos*	Total de veículos envolvidos na ocorrência.
26	*latitude*	Latitude do local do acidente em formato geodésico decimal.
27	*longitude*	Longitude do local do acidente em formato geodésico decimal.
28	*regional*	Superintendência regional da PRF cujo acidente ocorreu dentro dos limites de sua circunscrição. Atenção nem sempre a UF da regional coincide com a UF do acidente. Ex: A circunscrição da SPRF- DF grande parte está localizada na UF "GO".
29	*delegacia*	delegacia da PRF cujo acidente ocorreu dentro dos limites de sua circunscrição.
30	*uop*	UOP= unidade operacional. Unidade operacional da PRF cujo acidente ocorreu dentro dos limites de sua circunscrição.