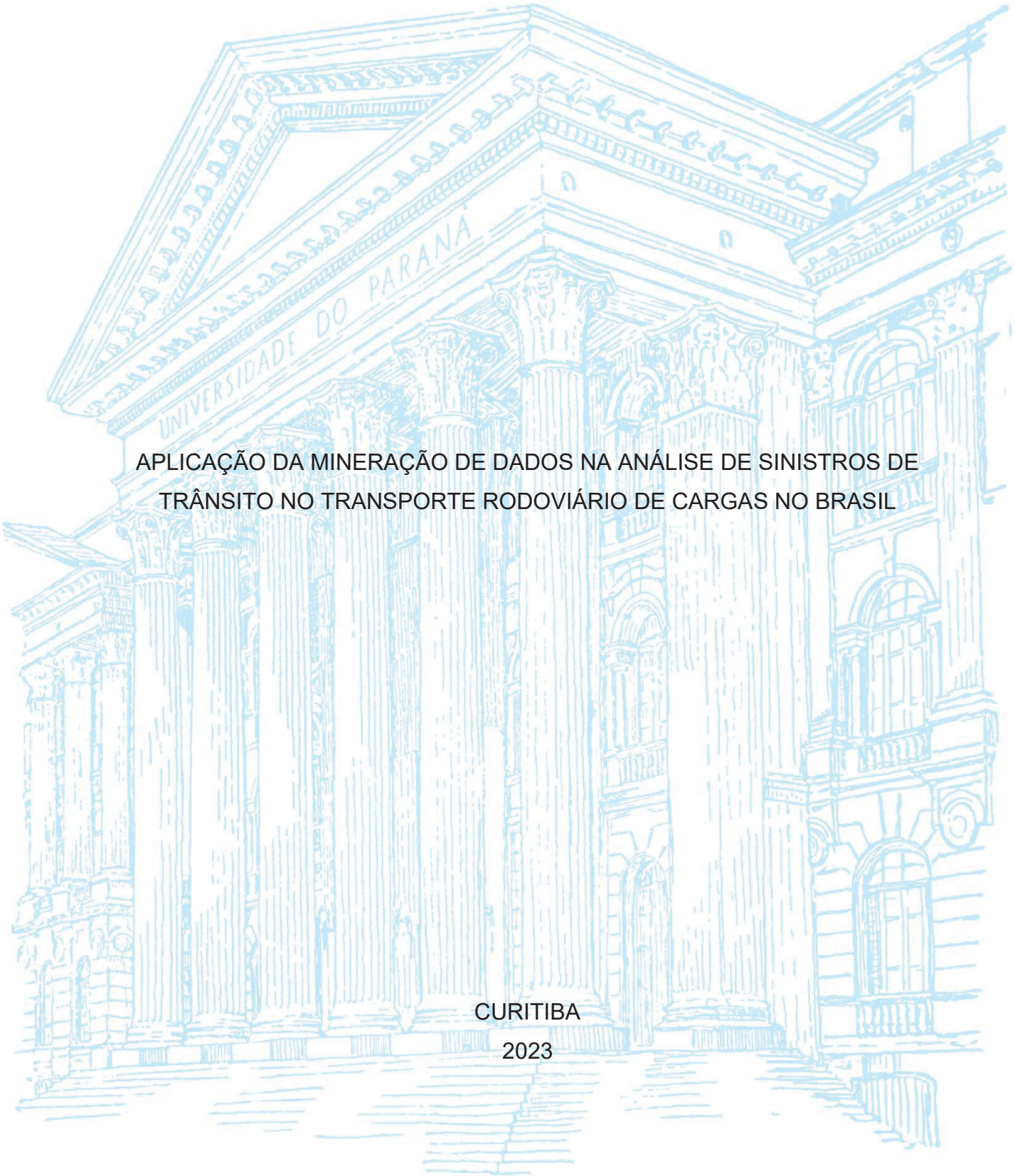
UNIVERSIDADE FEDERAL DO PARANÁ

MARILYN DE SOUZA CYGANCZUK



APLICAÇÃO DA MINERAÇÃO DE DADOS NA ANÁLISE DE SINISTROS DE TRÂNSITO NO TRANSPORTE RODOVIÁRIO DE CARGAS NO BRASIL

CURITIBA

2023

MARILYN DE SOUZA CYGANCZUK

APLICAÇÃO DA MINERAÇÃO DE DADOS NA ANÁLISE DE SINISTROS DE TRÂNSITO NO TRANSPORTE RODOVIÁRIO DE CARGAS NO BRASIL

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito para a obtenção do título de Doutor em Gestão da Informação.

Orientador: Prof. Dr. José Simão de Paula Pinto.

Coorientador: Prof. Dr. Jorge Tiago Bastos.

CURITIBA

2023

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS

Cyganczuk, Marilyn de Souza
Aplicação da mineração de dados na análise de sinistros de trânsito no transporte rodoviário de cargas no Brasil / Marilyn de Souza Cyganczuk. – Curitiba, 2023.
1 recurso on-line : PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Gestão da Informação.
Orientador: Prof. Dr. José Simão de Paula Pinto.
Coorientador: Prof. Dr. Jorge Tiago Bastos.

1. Gestão da informação. 2. Transporte de carga. 3. Mineração de dados. 4. COVID-19 (Doença). I. Pinto, José Simão de Paula. II. Bastos, Jorge Tiago. III. Universidade Federal do Paraná. Programa de Pós-Graduação em Gestão da Informação. IV. Título.

Bibliotecária: Maria Lidiane Herculano Graciosa CRB-9/2008

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA INFORMAÇÃO - 40001016058P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **MARILYN DE SOUZA CYGANCZUK** intitulada: **APLICAÇÃO DA MINERAÇÃO DE DADOS NA ANÁLISE DE SINISTROS DE TRÂNSITO NO TRANSPORTE RODOVIÁRIO DE CARGAS NO BRASIL**, sob orientação do Prof. Dr. JOSÉ SIMÃO DE PAULA PINTO, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutora está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 04 de Abril de 2023.

Assinatura Eletrônica
24/04/2023 09:01:32,0
JOSÉ SIMÃO DE PAULA PINTO
Presidente da Banca Examinadora

Assinatura Eletrônica
19/04/2023 18:22:25,0
CASSIUS TADEU SCARPIN
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
27/04/2023 15:16:06,0
MICHELLE ANDRADE
Avaliador Externo (UNIVERSIDADE DE BRASÍLIA)

Assinatura Eletrônica
19/04/2023 20:24:43,0
RONAN ASSUMPCAO SILVA
Avaliador Interno (INSTITUTO FEDERAL DO PARANÁ)

DEDICATÓRIA

Aos meus pais e filho.

AGRADECIMENTOS

Em primeiro lugar a Deus, maior fonte de força espiritual.

Ao meu orientador, Prof. Dr. José Simão de Paula Pinto, por aceitar orientar esse projeto de tese. Agradeço por todo apoio e orientação durante esses cinco anos e por todas as discussões que ajudaram a melhorar o desenvolvimento desse trabalho.

Ao meu coorientador, Prof. Dr. Jorge Tiago Bastos, pelas contribuições que foram importantes para o desenvolvimento do meu projeto.

Aos meus pais Dorvina Burger de Souza Cyganczuk e Miguel Cyganczuk, que sempre me incentivaram a estudar, pelo amor, carinho, educação e exemplos na construção do que sou hoje.

Ao meu esposo Jefferson Steidel dos Santos, meu amor e companheiro, pela compreensão, paciência e amor.

Ao meu filho Fernando Steidel Cyganczuk, que com seu amor e carinho me estimulou nos momentos difíceis.

À minha família e amigos que sempre me apoiaram durante o período de doutorado.

À coordenação, professores e amigos de curso do Programa de Pós-Graduação em Gestão da Informação da Universidade Federal do Paraná.

À Simone Batista, por todo suporte no PPGGI.

Aos membros da banca, pela disposição em avaliar meu trabalho.

A todas as pessoas que direta ou indiretamente influenciaram no desenvolvimento deste trabalho.

RESUMO

O Brasil está entre os países com maior número de mortes por sinistros de trânsito do mundo. Considerando o cenário das rodovias federais brasileiras, os veículos de carga são o terceiro tipo de veículo mais comum envolvido. Ainda, a pandemia de COVID-19 teve um impacto significativo na mobilidade humana em todo o mundo. Em relação ao estudo deste problema, os bancos de dados de sinistros de trânsito contêm uma série de informações capazes de orientar a tomada de decisão dos gestores para melhorar a segurança no trânsito. Para investigar tais bases, dispõe-se de metodologias como a *Knowledge Discovery in Databases* (KDD), ou seja, descoberta de conhecimento das bases de dados. A Mineração de Dados (MD), umas das etapas do KDD, pode ser vista como uma técnica para auxiliar nos processos de extração e busca das informações, sendo possível encontrar nos dados armazenados informações úteis que podem não ser perceptíveis em sua forma natural. O objetivo desta tese é testar técnicas de mineração de dados para a análise de dados de sinistros de trânsito, bem como comparar padrões de sinistros encontrados na literatura envolvendo o transporte rodoviário de cargas com os padrões de sinistros ocorridos em rodovias federais do Brasil utilizando ferramentas de mineração de dados, a partir dos dados disponibilizados pela Polícia Rodoviária Federal (PRF), no período de 2017 a 2021 e investigar os possíveis impactos da pandemia de COVID-19 nos sinistros de trânsito, visando contribuir no processo decisório dos gestores de organizações públicas e privadas. Metodologicamente, foi realizada uma comparação de algoritmos de mineração de dados, avaliando o desempenho de cada técnica de mineração e a compação da literatura com sinistros no Brasil. O estudo revela ser possível extrair fatores que influenciam nos sinistros de trânsito como os fatores humano, da via e do ambiente, corroborando com os resultados encontrados na literatura. Ao comparar os quatro algoritmos, o estudo mostrou que o algoritmo J48 se apresentou como um classificador satisfatório nos testes realizados.

Palavras-chave: Transporte Rodoviário de Cargas. Mineração de Dados. COVID-19.

ABSTRACT

Brazil is among the countries with the highest number of deaths from traffic accidents in the world. Considering the scenario of Brazilian federal highways, freight vehicles are the third most common type of vehicle involved. Still, the COVID-19 pandemic has had a significant impact on human mobility around the world. Regarding the study of this problem, the databases of traffic accidents contain a series of information capable of guiding the decision making of managers to improve traffic safety. To investigate such databases, there are methodologies such as Knowledge Discovery in Databases (KDD), that is, knowledge discovery of databases. Data Mining (DM), one of the stages of KDD, can be seen as a technique to assist in the processes of extracting and searching for information, making it possible to find useful information in the stored data that may not be perceptible in its natural form. The objective of this thesis is to test data mining techniques for the analysis of traffic claims data, as well as to compare claims patterns found in the literature involving road freight transport with the claims patterns occurred on federal highways in Brazil using tools of data mining, based on data provided by the Federal Highway Police (PRF), from 2017 to 2021 and to investigate the possible impacts of the COVID-19 pandemic on traffic claims, aiming to contribute to the decision-making process of managers of public organizations and private. Methodologically, a comparison of data mining algorithms was performed, evaluating the performance of each mining technique and comparing the literature with claims in Brazil. The study reveals that it is possible to extract factors that influence traffic accidents, such as human, road and environmental factors, corroborating the results found in the literature. When comparing the four algorithms, the study showed that the J48 algorithm presented itself as a satisfactory classifier in the tests performed.

Keywords: Road Cargo Transportation. Data Mining. COVID-19.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE SIGLAS

CNT	- Confederação Nacional do Transporte
GI	- Gestão da Informação
KDD	- Knowledge Discovery in Databases
MD	- Mineração de Dados
PRF	- Polícia Rodoviária Federal
TRC	- Transporte Rodoviário de Cargas

SUMÁRIO

1. INTRODUÇÃO

De acordo com a NBR-10697/2020 da Associação Brasileira de Normas Técnicas (ABNT, 2020), sinistro de trânsito é todo evento que resulte em dano ao veículo ou à sua carga e/ou em lesões a pessoas e/ou animais, e que possa trazer dano material ou prejuízos ao trânsito, à via ou ao meio ambiente, em que pelo menos uma das partes está em movimento nas vias terrestres ou em áreas abertas ao público. A mudança do termo acidente para sinistro deve-se ao fato do termo acidente trazer a conotação de algo imprevisível e incontrolável. De acordo com a Associação Brasileira de Medicina do Tráfego (Abramet), mais de 90% dos sinistros de trânsito registrados no Brasil têm como causa o fator humano, assim os sinistros não acontecem por acaso, podendo, em muitos casos, serem evitados (ABRAMET, 2021).

Segundo o relatório da Organização Mundial da Saúde (OMS), *o Global status report on road safety 2018,* o número de mortes anuais no trânsito atingiu 1,35 milhão, sendo que os sinistros de trânsito são a principal causa de morte entre pessoas de 5 a 29 anos (WHO, 2018). Os sinistros de trânsito podem ocorrer devido a diversos fatores como: comportamento humano, habilidades dos motoristas, condição meteorológica, condição do veículo ou da via, entre outros. É necessário determinar a prevalência de fatores ou combinação de fatores para avaliar as medidas adotadas e sugerir novas contramedidas que acabariam por melhorar os registros de segurança (ZANNE; GROZNIK; TWRDY, 2013).

Tranchitella *et al.* (2021), apontam o cumprimento inadequado das leis de trânsito, o aumento da frota de veículos e o consumo de bebidas alcoólicas como fatores que podem contribuir para a ocorrência de sinistros de trânsito.

O elevado número de sinistros de trânsito ocorridos em rodovias federais envolvendo veículos de carga constitui-se como um dos desafios principais às políticas públicas no setor de transportes e para outros setores como a saúde e segurança pública (SOUZA; PAIVA; REIMÃO, 2005). De acordo com a Confederação Nacional do Transporte - CNT (2022), o sistema de transportes de cargas é fundamental para a movimentação da economia do país, pois 64,90% de toda a carga transportada no Brasil em 2021 usou o sistema modal rodoviário.

Entre todos os tipos de veículos, os caminhões são os que mais contribuem para os sinistros de trânsito, lesões e mortes devido à sua alta proporção entre a população rodoviária, bem como seu tamanho, peso e outras características únicas (ZHU; SRINIVASAN, 2011; HE *et al.*, 2019). Segundo Moradpour e Long (2019), os sinistros de trânsito são preocupações importantes para os tomadores de decisão de transporte.

Os sinistros de trânsito não são uma exceção ao que os dados digitais podem oferecer. Consequentemente, a análise de relatórios de sinistros de trânsito pode influenciar os planos de instituições governamentais e privadas em relação à segurança no trânsito. Diversos modelos analíticos são aplicados para investigar os efeitos de uma variedade de fatores potenciais que influenciam o nível de lesões devido aos sinistros de trânsito. Da mesma forma, modelos foram utilizados para prever o nível de lesões por sinistro, sendo que as técnicas de mineração de dados são consideradas bem-sucedidas na abordagem de análise de previsão de sinistros (ALKHEDER; ALRUKAIBI; AIASH, 2020). A mineração de dados auxilia nas tomadas de decisões, tornando as organizações mais fortes perante o mercado extremamente competitivo e auxiliando no gerenciamento de riscos (DE SOUZA JUNIOR; VILLELA, 2019).

Para diminuir o número de sinistros de trânsito é necessário o conhecimento de suas causas, isso é possível por meio da coleta e gerenciamento das informações disponíveis sobre os sinistros. Além disso, é necessário identificar os fatores mais importantes e padrões significativos que afetam a ocorrência e a gravidade dos sinistros. Nos últimos anos, técnicas de mineração de dados têm sido utilizadas em estudos que analisam acidentes de trânsito obtendo resultados satisfatórios, como os estudos de Tao *et al.* (2016); Mafi, Abdelrazig e Doczy (2018); John e Shaiba (2019); Alkheder, Alrukaibi e Aiash (2020); Chen *et al.* (2020); Lee *et al.* (2020), entre outros.

Emergências como a pandemia de COVID-19, impõe uma série de desafios à saúde pública, dificultando a tomada de decisão por parte das autoridades em suas diferentes esferas de governo (CARDOSO *et al.,* 2020; COLONNA; INTINI, 2020).

Diversos países responderam impondo bloqueios a fim de reduzir a propagação da infecção e potenciais mortes, sendo uma das restrições à limitação da mobilidade pessoal, visto que o tráfego aéreo e terrestre foram dois

dos principais vetores para a disseminação do COVID-19 (KRAEMER *et al.*, 2020; LAU *et al.*, 2020; PEERI *et al.*, 2020; SALADIÉ; BUSTAMANTE; GUTIÉRREZ, 2020; WU; LEUNG; LEUNG, 2020a).

Diante do exposto, este estudo tem como objetivo testar técnicas de mineração de dados para a análise de dados de sinistros de trânsito, bem como identificar padrões de sinistros envolvendo o transporte rodoviário de cargas em rodovias federais do Brasil, a partir dos dados armazenados no banco de dados da Polícia Rodoviária Federal, utilizando e comparando técnicas e algoritmos de mineração de dados e investigar se houveram impactos da pandemia de COVID-19 nos sinistros de trânsito, visando auxiliar no processo decisório dos gestores de organizações públicas e privadas.

1.1 PROBLEMA DE PESQUISA

Os sinistros de trânsito se configuram como grave problema de saúde pública de um país, no caso o Brasil, visto que as estatísticas mostram que o trânsito brasileiro é o terceiro mais violento do mundo, ficando atrás apenas da China e Índia, segundo dados divulgados pela Organização Mundial da Saúde através do relatório Global Status Report on Road Safety (WHO, 2018).

Analisar os dados de sinistros de trânsito, tentando extrair algum padrão e encontrar os principais fatores que estejam causando estes sinistros, é uma iniciativa que pode auxiliar o processo de tomada de decisão para que haja uma redução de sinistros nas rodovias federais brasileiras. Para o processo de descoberta de padrões, podem ser utilizadas técnicas de mineração de dados (REZENDE, 2003).

A menor circulação de veículos, como efeito imediato das medidas preventivas de enfrentamento à pandemia, poderia ter um impacto positivo na ocorrência de sinistros de trânsito e nas mortes a eles associadas. Entretanto, notícias veiculadas pela mídia apresentam uma redução de sinistros em termos gerais, porém, há referência a ocorrência de eventos com maior gravidade e maior número de infrações de motoristas (SANTOS *et al.*, 2020).

No estudo de Katrakazas *et al.* (2020), foi realizada uma pesquisa com objetivo de quantificar o efeito da pandemia de COVID-19 no comportamento de direção e na segurança no trânsito e identificar mudanças no comportamento de

direção causadas por restrições da COVID-19 na Grécia e Arábia Saudita. Foi demonstrado que os volumes de tráfego, reduzidos devido às medidas de limitação de circulação, levaram a um aumento nas velocidades de 6-11%; destacando-se as acelerações bruscas mais frequentes e eventos de frenagem bruscos (aumento de até 12%), bem como uso de telefone celular (aumento de até 42%) durante março e abril de 2020. Pesquisas recentes apresentadas por Vingilis *et al.* (2020) também descobriram que os jovens motoristas do sexo masculino estavam mais dispostos a ultrapassar o limite de velocidade e a frequência de tais excessos foi maior durante esse período da pandemia.

Nessa perspectiva, a mineração de dados pode ajudar na tomada de decisões descobrindo conexões e associações escondidas, incluindo informações do período de pandemia e prevendo tendências futuras do transporte rodoviário de cargas (TRC) com relação aos sinistros de trânsito. Assim, tem-se o seguinte problema de pesquisa:

Como a utilização de técnicas de mineração de dados pode subsidiar a tomada de decisão na sinistralidade e para os direcionamentos estratégicos, como a situação da pandemia, nas organizações públicas e privadas de transporte de cargas no Brasil?

1.2 OBJETIVOS

Neste item são apresentados o objetivo geral e os objetivos específicos da presente pesquisa.

1.2.1 Objetivo Geral

O objetivo desta tese é testar técnicas de mineração de dados para a análise de dados de sinistros de trânsito, bem como comparar padrões de sinistros encontrados na literatura envolvendo o transporte rodoviário de cargas com os padrões de sinistros ocorridos em rodovias federais do Brasil utilizando ferramentas de mineração de dados, a partir dos dados disponibilizados pela Polícia Rodoviária Federal (PRF), no período de 2017 a 2021 e investigar os possíveis impactos da pandemia de COVID-19 nos sinistros de trânsito, visando

contribuir no processo decisório dos gestores de organizações públicas e privadas.

1.2.2 Objetivos Específicos

Este estudo, de forma a atender o seu objetivo geral, também pretende atingir os seguintes objetivos específicos:

1) Verificar na literatura quais são os principais atributos utilizados nos estudos de sinistros de trânsito;
2) Selecionar atributos para a base de dados, contendo as informações da Polícia Rodoviária Federal (PRF) e os principais atributos encontrados na literatura;
3) Verificar na literatura quais os fatores que contribuem para a ocorrência dos sinistros de trânsito envolvendo o transporte rodoviário de cargas;
4) Verificar quais são as técnicas de mineração de dados mais utilizadas nos estudos de sinistros de trânsito;
5) Utilizar técnicas e algoritmos de mineração de dados para realizar os experimentos com os atributos selecionados;
6) Avaliar o desempenho de cada técnica de mineração encontrada na revisão bibliográfica, comparando seus resultados para encontrar o que desempenha o melhor resultado para identificar padrões de sinistros envolvendo o transporte rodoviário de cargas;
7) Verificar se houve influência da pandemia de COVID-19 nos sinistros de trânsito.

1.3 JUSTIFICATIVA

Este estudo apresenta como foco a temática relacionada à aplicação de técnicas e algoritmos de mineração de dados como ferramentas de suporte a tomada de decisão envolvendo sinistros de trânsito no transporte rodoviário de cargas.

O fluxo de informações eficaz e eficiente detém um efeito multiplicador, acelerando todos os setores organizacionais, tornando-se a força motora do

desenvolvimento econômico, social e tecnológico (CHIUSOLI; REZENDE, 2019).

Conforme Al-Turaik *et al.* (2016), em muitos países os dados sobre sinistros rodoviários são estudados para explorar os fatores que levam aos sinistros. Pesquisadores e autoridades buscam identificar padrões e relações entre fatores de risco e os níveis de gravidade das lesões. Entre os atributos importantes que geralmente são estudados nos dados de sinistros de trânsito estão: características do motorista, características do veículo, condição das estradas, fatores climáticos e ambientais.

O estudo apresentado possui ainda quatro justificativas elencadas a seguir:

Acadêmica: A Assembleia Geral da Organização das Nações Unidas, no dia 02 de março de 2010, proclamou oficialmente o período de 2011 a 2020 como a Década Mundial de Ação pela Segurança no Trânsito a fim de estimular esforços em todo o mundo para conter e reverter à tendência crescente de fatalidades e ferimentos graves em sinistros no trânsito no planeta, com o principal objetivo de reduzir pela metade o número de fatalidades no trânsito mundial (BRASIL, 2019). Porém, conforme dados apresentados na Terceira Conferência Ministerial Global sobre Segurança Viária, em 2020, atualmente as lesões no trânsito são a principal causa de morte entre crianças e jovens adultos, na faixa de 5 a 29 anos, além de que os países gastam, em média, 3% do seu Produto Interno Bruto - PIB, soma de todas as riquezas produzidas em um ano, com as vítimas do trânsito (BASTOS *et al.*, 2020).

Segundo dados do IPEA de 2019, os sinistros de trânsito no Brasil matam cerca de 45 mil pessoas por ano e deixam mais de 300 mil pessoas com lesões graves. Observou-se que os sinistros em rodovias custam à sociedade brasileira cerca de R$ 40 bilhões por ano, enquanto nas áreas urbanas, em torno de R$ 10 bilhões, sendo que o custo relativo à perda de produção responde pela maior fatia desses valores, seguido pelos custos hospitalares. O estudo ainda aponta: quanto maior a gravidade do acidente, maiores os custos associados a ele, sobretudo quando há vítimas fatais envolvidas, elevando substancialmente o custo final (CARVALHO, 2020).

Em comparação com outros tipos de veículos, o tamanho maior e o centro de gravidade mais alto dos caminhões resultam em distâncias de

frenagem mais longas e consequências mais graves quando envolvidos em sinistros. Além disso, os sinistros com caminhões têm impactos econômicos mais elevados por causa dos danos a cargas de alto valor e atrasos em viagens. Assim, pesquisas que identifiquem fatores influentes em sinistros com caminhões facilitariam o desenvolvimento de contramedidas que poderiam reduzir o número e a gravidade dos sinistros envolvendo caminhões (ZHOU; ZHANG, 2019).

Fortalecendo essa ideia, os autores Blazquez *et al.* (2018) apontam que os sinistros envolvendo caminhões detêm o fluxo de carga, causando interrupções na cadeia de abastecimento e gerando atrasos operacionais onerosos. Os custos associados a colisões com caminhões demonstram a necessidade de aumentar a segurança nas operações de transporte rodoviário.

A análise efetuada e o conhecimento adquirido a respeito deste assunto poderão servir de guia aos órgãos gestores do sistema rodoviário no planejamento de ações adequadas objetivando reduzir significativamente os índices de sinistros envolvendo veículos de carga. Alguns estudos similares foram realizados envolvendo sinistros com veículos em geral priorizando os veículos de passeio e motocicletas, como o estudo de Galvão; De Fátima Marin (2010) o qual identificaram, por meio da aplicação da tecnologia de mineração de dados, regras sobre sinistros de transporte ocorridos no município de Cuiabá, Mato Grosso, no ano de 2006, neste estudo foi aplicada a tecnologia de minerar dados, por meio do algoritmo *APriori* e o *software* utilizado foi o WEKA.

Na pesquisa de Jesus Costa, Bernardini e Viterbo Filho (2014), os autores analisaram a viabilidade da aplicação do processo de mineração de dados sobre os dados fornecidos pela Polícia Rodoviária Federal (PRF) em 2012 (boletins de ocorrência), a fim de identificar associações entre variáveis relacionadas aos sinistros de trânsito em todas as rodovias federais, utilizou-se para tanto os algoritmos *Apriori*, J48 e PART e o *software* WEKA, a partir dos resultados, foi possível observar que alguns resultados obtidos com os algoritmos J48 e PART são promissores em relação à classificação das causas de sinistros.

De acordo com os valores obtidos na pesquisa, ao se utilizar o algoritmo *Apriori*, foram geradas 38 regras de associação com confiança maior que 0,8. O estudo ainda aponta que extrair algum padrão dos sinistros

rodoviários e encontrar os principais fatores que estejam causando esses sinistros pode auxiliar o processo de tomada de decisão, assim como futuros planejamentos, para que haja uma redução de sinistros nas rodovias federais brasileiras (DE JESUS COSTA; BERNARDINI; VITERBO FILHO, 2014).

Além disso, a chegada da Resolução nº 808, de 15 de dezembro de 2020, do Conselho Nacional de Trânsito (CONTRAN), que dispõe sobre o Registro Nacional e Estatística de Trânsito (RENAEST), sendo um sistema de registro, gestão e controle de dados e informações sobre sinistros e estatísticas de trânsito, coletados pelos órgãos que compõem o Sistema Nacional de Trânsito (SNT) e pelos demais órgãos e entidades que efetuam os registros de sinistros de trânsito, que apuram suas circunstâncias, ou prestam atendimento às suas vítimas. Os dados obtidos são consolidados em base nacional, organizada e mantida pelo órgão máximo de trânsito da União. Apesar de existirem dados e informações complementados por outros sistemas, como o de Registro Nacional de Veículos Automotores (RENAVAM), Registro Nacional de Carteira de Habilitação (RENACH) e Registro Nacional de Infrações de Trânsito (RENAINF), o RENAEST, tem como base os dados registrados sobre sinistros de trânsito, através do Boletim de Ocorrência de Acidente de Trânsito (BRASIL, 2020).

A relevância do tema foi observada a partir de um levantamento em bases de artigos publicados em revistas acadêmicas como o *Web Of Science*, *Science Direct* e SciELO, com o período de 10 anos (2011-2021), considerando as palavras individuais “*traffic accident*” e utilizando o conector OR com o termo “*traffic safety*” e o conector AND com o termo “*data mining*” e “*truck*” no resumo, palavra-chave ou título, a partir do qual tornou-se perceptível a carência de estudos envolvendo a comparação de diversas técnicas e algoritmos de mineração de dados nos sinistros de trânsito envolvendo o TRC, apontando a necessidade do presente estudo. Além da ausência de estudos utilizando os dados da PRF no Paraná para uma análise de sinistros do período antes e durante a pandemia de COVID-19.

Segundo Vingilis *et al.* (2020), em meio à pandemia de COVID-19, os motoristas de caminhão estão passando por mudanças em seu trabalho que podem estar afetando sua saúde e segurança. As atuais restrições de viagem para o público em geral significam que as estradas estão menos congestionadas

e, portanto, possivelmente mais seguras para trabalhadores essenciais, como os motoristas de caminhão. No entanto, algumas reportagens sugerem que os motoristas podem estar viajando em velocidades mais altas. A pesquisa ainda indica um aumento nas compras online nos Estados Unidos, Canadá, Reino Unido e na Alemanha para todos os produtos desde a pandemia. No entanto, a grande demanda inicial de caminhoneiros para reabastecimento de mercadorias no início da pandemia foi seguida por demandas de cargas imprevisíveis devido à desaceleração e agora à reabertura da economia.

Outro ponto mencionado no estudo é que antes do COVID-19, os caminhoneiros tinham falta de áreas de descanso públicas e locais de beira de estrada para refeições e banheiros; essa situação foi agravada pela pandemia e isso pode afetar o risco de colisão (VINGILIS *et al.*, 2020).

Assim, o trabalho ainda busca contribuir na verificação da influência da pandemia de COVID-19 nos sinistros de trânsito envolvendo o transporte rodoviário de cargas, através de uma comparação dos dados de sinistros antes da pandemia e durante a pandemia de COVID-19.

Pessoal: aprimorar os conhecimentos acerca dos temas envolvidos neste projeto.

Programa: abrange as três áreas da linha de pesquisa, uma vez que o tema aborda o uso de ferramentas computacionais (tecnologia) para transformar dados em conhecimento (informação) para favorecer a tomada de decisão (gestão). Além disso, o uso de ferramentas de mineração de dados auxilia no apoio à decisão (DE JESUS COSTA; BERNARDINI; VITERBO FILHO, 2014). Dessa forma, o estudo busca contribuir analisando padrões dos sinistros e encontrando os principais fatores que estejam causando esses sinistros para poder auxiliar o processo de tomada de decisão.

No processo de busca por melhores resultados, surgem processos analíticos que auxiliam na tomada de decisão, como o processo de mineração de dados (FERNANDES; CHIAVEGATTO FILHO, 2019), assim, por meio do seu uso é possível que os gestores das organizações públicas e privadas de transporte de cargas consigam mitigar seus riscos de envolvimento em sinistros e otimizar seus custos, aumentando as vantagens competitivas da organização. Além disso, as ferramentas de mineração de dados fazem parte do estudo da linha de pesquisa do programa: Informação e Tecnologia.

Social: A ocorrência do sinistro causa enormes danos humanos, econômicos e perdas sociais. Dado este fato, a segurança no trânsito tem sido um problema sério que está intimamente relacionada à saúde e desenvolvimento que exige que as organizações tomem medidas abrangentes eficazes (WU *et al.*, 2020b).

Mais impactante que o custo econômico dos sinistros é o custo humano e social como: o sofrimento físico e psicológico das vítimas, o sofrimento psicológico dos familiares e pessoas com ligação com as vítimas, doenças de natureza psicológica que acometem vítimas e pessoas próximas (depressão, fobias, etc.), perda de qualidade de vida de muitas das vítimas e de seus familiares, desestruturação econômica de famílias, distanciamento de entes queridos em razão do tratamento hospitalar e de reabilitação, entre outros (FERRAZ *et al.*, 2012).

1.4 ESTRUTURA DO TRABALHO

Este trabalho compõe-se de seis capítulos, incluindo esta introdução. No segundo capítulo, aborda-se o referencial teórico, compreendendo os conceitos e principais abordagens e estudos sobre o tema. No terceiro capítulo, apresenta-se a metodologia a ser utilizada para o desenvolvimento deste estudo. O quarto capítulo traz os resultados obtidos neste estudo e avaliação dos resultados. O quinto capítulo traz uma discussão dos resultados. Finalmente, no sexto capítulo são feitas as considerações finais do estudo, limitações encontradas e sugestões para trabalhos futuros.

2. REFERENCIAL TEÓRICO

Neste Capítulo, será apresentado o referencial teórico a fim de possibilitar uma melhor compreensão do tema e do problema de pesquisa.

2.1 GESTÃO DA INFORMAÇÃO

A informação é considerada um recurso organizacional que merece ser administrado (ALVES; DUARTE, 2015). Para Davenport (2002, p. 84), “grandes

volumes de informação entram e saem das organizações sem que ninguém tenha plena consciência de seu impacto, valor ou custo". Assim, o gerenciamento da informação é fundamental para a obtenção do sucesso, das oportunidades e da manutenção de vantagem competitiva (ALVES; DUARTE, 2015).

A gestão da informação (GI) é desenvolvida por meio de processos, que, de acordo com McGee e Prusak (1994), são um conjunto de tarefas que devem estar conectadas entre si e com o ambiente em que estão inseridas. Nesse sentido, Choo (2003a) corrobora que a GI é considerada uma rede de processos que objetivam adquirir, criar, organizar, distribuir e usar a informação.

De acordo com Hoffmann (2016), a GI nas organizações torna-se importante para que ocorra alinhamento das estratégias da organização em decorrência do crescente volume de informações formais e informais, a complexidade dos diversos processos e a grande velocidade das mudanças, sejam elas tecnológicas, econômicas, sociais ou culturais.

Segundo Choo (2003b), teoricamente, todas as decisões deveriam ser tomadas racionalmente, baseadas em informações completas e alternativas plausíveis. No mundo ideal, as decisões deveriam ser tomadas com a análise de todas as alternativas possíveis. Porém, na prática, isso não ocorre, pois existem choques de interesses, falta de informações, entre outras.

A informação tornou-se um importante ativo para as organizações, todavia, em alguns ambientes, ela ainda não é vista com o devido valor, tão pouco é gerida como outros recursos organizacionais (CÂNDIDO; VALE, 2018).

Existem diversas técnicas de tomada de decisão como: árvores de decisão, matriz de tomada de decisão, método de simulação de Monte Carlo, raciocínio baseado em casos, algoritmos genéticos e aprendizagem de máquina (técnica máquinas de vetores suporte - SVM).

Árvore de decisão é a forma mais simples dos modelos de decisão usados rotineiramente. Ela é valorizada pela sua transparência e excelente capacidade de descrever as opções alternativas (SOÁREZ; SOARES; NOVAES, 2014).

A árvore de decisão é uma ferramenta visual que descreve graficamente os três principais componentes de um problema de decisão: o modelo propriamente dito, as probabilidades de ocorrência dos vários eventos que estão sendo modelados e os valores dos desfechos que existem no final de cada

percurso (CHAPMAN; SONNENBERG, 2003). Na subseção 2.6.2.1 as árvores de decisão serão abordadas mais profundamente.

A matriz de decisão fornece os parâmetros iniciais para a aplicação dos métodos de decisão multicritério para determinar a melhor alternativa. Consiste em colher informações que constituam elementos de ponderação que levarão à possível solução de interesse, em linhas e colunas de uma matriz. Para cada intersecção de linha *versus* coluna é aplicada uma pontuação (LOPEZ, 2017). A abordagem multicritério de apoio à decisão conta com a vantagem de substituir escolhas intuitivas por decisões justificadas que permitem a participação dos atores envolvidos em diferentes etapas da construção da matriz de decisão (SANTOYO, 2011).

Segundo Ramos (2017), os métodos de Monte Carlo são uma poderosa ferramenta de simulação estocástica com ampla aplicação em engenharia, já que permitem lidar com problemas com comportamento altamente não linear e que envolvem um grande número de variáveis aleatórias com diferentes distribuições de probabilidade. O método é um tipo especial de simulação utilizada em modelos envolvendo eventos probabilísticos. Esse método é denominado de Monte Carlo porque utiliza um processo aleatório, tal como um lançamento de dados ou o girar de uma roleta, para selecionar os valores de cada variável em cada tentativa (MORSE, 1986; CORRAR, 1993).

O raciocínio baseado em casos (RBC), é uma abordagem de Inteligência Artificial para resoluções de problemas e aquisição de conhecimento, que tem como princípio que “problemas semelhantes, possuem soluções semelhantes” (AAMODT; PLAZA, 1994). Diante desse princípio, a técnica consiste na presença de um novo problema, adaptar soluções de problemas similares resolvidos no passado, para encontrar uma solução adequada para o problema em questão (NASCIMENTO, 2018).

Dessa forma, o RBC funciona semelhante ao modelo cognitivo humano que se baseia em experiências passadas para gerar hipóteses de soluções para novas situações similares (VON WANGENHEIN; VON WANGENHEIN; RATEKE, 2013).

A técnica RBC diferencia-se das demais técnicas de Inteligência Artificial, principalmente, pela forma como o conhecimento é empregado e representado. As demais técnicas, por exemplo, a Lógica *Fuzzy* utiliza o

conhecimento representado por meio de um conjunto em regras, modelos, quadros, roteiros, entre outras. Já na técnica de RBC, esse conhecimento é representado através de casos concretos (VON WANGENHEIM; VON WANGENHEIM; RATEKE, 2013).

Os Algoritmos Genéticos são uma heurística de otimização criada proposta por John Holland em 1960, e baseia-se na teoria evolucionária sugerida por Charles Darwin. Mitchell (1998) descreve o processo do AG iniciando com uma população aleatória (soluções), que será submetida sucessivas vezes a operações que evoluirão até uma geração que irá atender a alguns critérios de avaliação. Nesse processo evolucionário, cada indivíduo (cromossomo) da população é uma possível solução do problema. Esses indivíduos são avaliados em uma função que visa atestar o quão aptos eles estão como uma solução viável. Os indivíduos com maior aptidão comporão uma nova população, na qual serão aplicados os operadores genéticos de mutação e cruzamento (DA SILVA; DE OLIVEIRA, 2020).

De acordo com Goldberg (2006) os "Algoritmos Genéticos são algoritmos de busca baseados na mecânica da seleção natural e da genética natural". Baseando-se na seleção do mais forte, as cordas binárias que estruturam esse algoritmo se recombinam de modo a gerar novos indivíduos (cordas). As informações passadas dos agentes são levadas em conta para a aprendizagem.

A Aprendizagem de Máquina (ML, *Machine Learning*) é um campo da inteligência computacional que estuda o uso de técnicas computacionais para automaticamente detectar padrões em dados e utilizá-los para fazer predições e tomar decisões. A técnica SVM e uma técnica de ML empregada em problemas de regressão e de classificação, sendo caracterizada como uma técnica de aprendizado supervisionado, pois se utiliza de um conjunto de dados cujas saídas são previamente conhecidas para detectar padrões e produzir um modelo capaz de deduzir as saídas corretas para novos dados. Tal técnica é fundamentada na Teoria de Aprendizagem Estatística e foi desenvolvida por Vladimir Vapnik, Bernhard Boser, Isabelle Guyon e Corrina Cortes (BOSER; GUYON; VAPNIK, 1992; CORTES; VAPNIK, 1999).

2.2 TRANSPORTE RODOVIÁRIO DE CARGAS

Nas últimas décadas o país observou uma intensificação do seu sistema de produção e do crescimento econômico, aumentando a demanda por transporte, a qual tem sido suprida majoritariamente pelo transporte rodoviário em função da reduzida oferta de infraestrutura ferroviária e hidroviária no país (PÉRA *et al.*, 2021)

Conforme TABELA 1, observa-se a participação dos diferentes modais na movimentação anual de cargas em 2020, sendo o transporte rodoviário o principal modal de transporte utilizado no Brasil (BRASIL, 2021).

TABELA 1 - MATRIZ DO TRANSPORTE DE CARGAS – MOVIMENTAÇÃO ANUAL

Modal	Bilhões (TKU*)	Participação (%)
Rodoviário	1.548,00	64,86
Aquaviário	375,2	15,72
Ferroviário	356,80	14,95
Dutoviário	106,10	4,45
Aéreo	0,6	0,02
Total	2.386,70	100,0

FONTE: PNL (2021).
NOTA:*TKU – Toneladas transportadas por quilômetro útil.

Observa-se na TABELA 1, de acordo com o Plano Nacional de Logística e Transporte: PNL 2035 (BRASIL, 2021), que o modal rodoviário transporta a maior quantidade de cargas nacionais, 1.548 bilhões de toneladas transportadas por quilômetro útil, representando 64,86% de participação em comparações com os outros modais.

A competição no segmento de transporte rodoviário de cargas tem se intensificado nas últimas décadas. As empresas brasileiras de transporte de cargas têm um desafio de manter continuamente o incremento da eficiência e produtividade para manter a competitividade (BLANCO; PAIVA; WANKE, 2014).

Além disso, o Brasil apresenta uma frota de mais de dois milhões de caminhões, com uma idade média bastante elevada de 10,11 anos para veículos de empresa e de 20,3 anos para veículos de caminhoneiros autônomos e com um alto grau de heterogeneidade em termos de tipos de equipamentos e de combinações veiculares utilizadas, bem como de prestadores de serviços:

empresas transportadoras de cargas (que podem atuar tanto com frota própria quanto com agregados) com 1.343.498 veículos, cooperativas de transporte de carga com 28.954 veículos e motoristas autônomos com 836.988 veículos, totalizando 2.209.440 veículos de transporte de carga em 2021 (ANTT, 2022).

A preponderância rodoviária no sistema de transporte e logística brasileiro, mensurada em termos do volume transportado, valor adicionado e empregos com carteira assinada registrados no segmento, é uma expressão de sua inserção disseminada em praticamente todos os setores produtivos, mercados de destino e tipos de viagem - de curta, média ou longa distância. Ou seja, os ônibus e caminhões são usados no Brasil para prestar variados tipos de serviços de transporte, embora sejam mais eficientes e competitivos nos trajetos curtos e médios e no serviço logístico porta a porta. Isso, em razão de sua menor capacidade de carregamento por veículo, de sua maior velocidade e do maior acesso a regiões mais remotas (CNT, 2019a).

2.3 TRANSPORTE RODOVIÁRIO DE CARGAS E GESTÃO DA INFORMAÇÃO

A sobrevivência das organizações depende da habilidade de gerir as informações e gerar conhecimento aos executivos ou gestores nas tomadas de decisão. A partir do entendimento dos fluxos informacionais, é possível desenvolver ações de melhorias que diretamente se relacionam ao sucesso do processo decisório e, consequentemente, ao sucesso organizacional (MENDONÇA; VARVAKIS, 2018).

Ainda segundo Mendonça e Varvakis (2018), estudar os fluxos informacionais, com enfoque no uso da informação para tomada de decisão, contribui para entender como, onde e para qual finalidade a informação é utilizada nas organizações.

A dificuldade de gerir dados e informações que sejam analisadas e interpretadas de forma adequada e eficaz, objetivando auxiliar um processo de tomada de decisão, são comuns nos ambientes organizacionais. Muitas empresas obtêm diversos dados sobre o seu negócio e mercado que está inserido, no entanto, não consegue transformar em informações relevantes e

estratégicas para melhores decisões (DA SILVA; ALMEIDA SILVA; SIMOES GOMES, 2016).

Os obstáculos enfrentados pelos tomadores de decisão nas empresas de transporte de carga estão relacionados ao processamento limitado de informações ou dados insuficientes (SHI *et al.*, 2019).

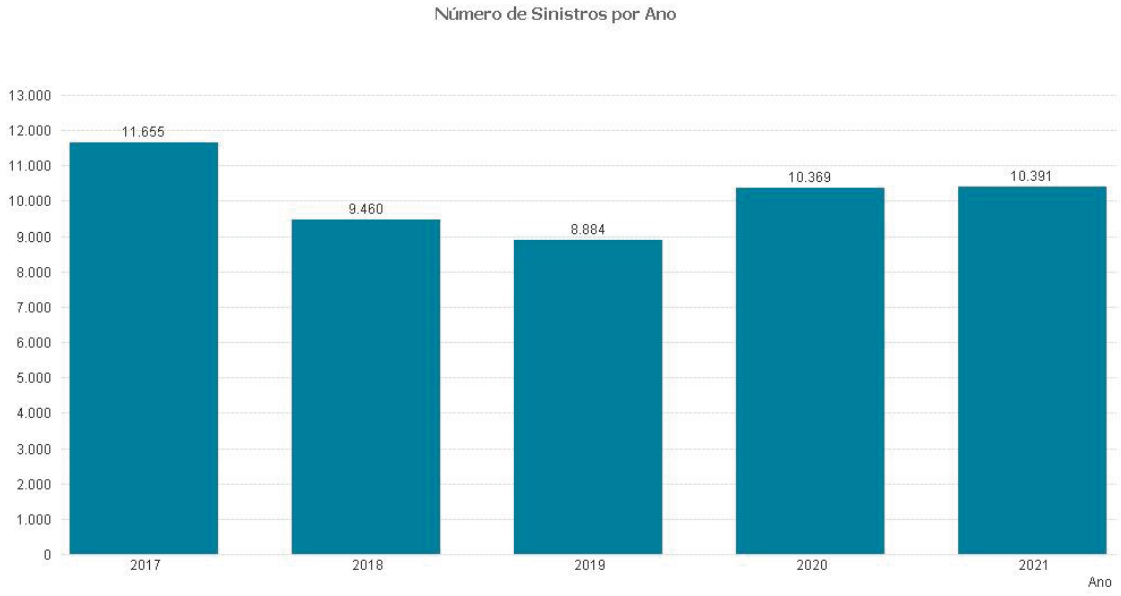
Devido ao grande fluxo de informações envolvendo o transporte de cargas, são necessários sistemas de informação que transferem eficientemente o fluxo de informações de um ponto para outro permitindo analisar o fluxo de transporte, bem como melhorar a metodologia de cálculos do planejamento de transporte e previsões futuras. Por exemplo, informações sobre localização e hora de chegada da carga oferece uma oportunidade para preparar antecipadamente o local para descarregamento/carregamento (JARAŠŪNIENĖ; BATARLIENĖ; VAIČIŪTĖ, 2016).

Entretanto, o estudo de De Abreu e Schinaider (2020), apresenta tecnologias que concedem suporte a tomada de decisão e o gerenciamento do transporte de cargas, como os aplicativos CargoX, Fretefy, Quero Frete, TruckPad, entre outros, que viabilizam o compartilhamento de cargas disponíveis para realização dos serviços de TRC no modelo da economia compartilhada, facilitando a conexão entre motoristas que estão disponíveis com seus ativos ociosos a embarcadores demandantes de cargas que precisam atender seus clientes; influenciando diretamente na produtividade, assim como na subutilização dos veículos, caracterizando menos perdas operacionais e consumo de combustível.

2.4 TRANSPORTE RODOVIÁRIO DE CARGAS E SINISTROS

O transporte rodoviário é o principal modal de transporte utilizado no Brasil (CNT, 2019). O GRÁFICO 1 apresenta os sinistros que envolveram pelo menos um caminhão no período de 2017 a 2021. Segundo dados da Polícia Rodoviária Federal (PRF) (BRASIL, 2022), os sinistros envolvendo TRC representam uma alta proporção do total de sinistros de trânsito em rodovias federais, apresentando números abaixo dos veículos de passeio e motocicletas apenas, sendo, portanto, o terceiro tipo de veículo que mais se envolve em sinistros.

GRÁFICO 1 - NÚMERO DE SINISTROS RODOVIÁRIOS ENVOLVENDO PELO MENOS UM CAMINHÃO



Fonte: A autora com base nos dados da PRF (2022).

Conforme GRÁFICO 1, o número de sinistros envolvendo pelo menos um caminhão a partir de 2017 diminuiu ao longo dos anos de 2018 e 2019, aumentando em 2020 e 2021 em comparação com 2018 e 2019. Em 2021, foram registrados 10.391 sinistros nas rodovias federais brasileiras com o envolvimento de pelo menos um caminhão. Foram registradas 1.911 vítimas fatais nos sinistros envolvendo veículos de carga, o que representa 9,13% de todas as ocorrências de óbitos envolvendo outros tipos de veículos em sinistros em 2022 (BRASIL, 2022).

Segundo a terminologia do Atendimento de Sinistros de Trânsito da PRF (2018), a PRF adota os seguintes critérios para a classificação do estado físico das pessoas envolvidas no sinistro:

- Ileso é a pessoa que não apresenta nenhuma queixa de dor, sinal ou sintoma de lesões provenientes do sinistro, mesmo que apresente alterações psicológicas ou que seja encaminhada para atendimento hospitalar;
- Lesão leve (feridos leves) é considerada a lesão em pessoa que, por consequência do sinistro, apresenta ao menos um sinal ou sintoma como: queixa de dores em geral, relacionadas à dinâmica do sinistro,

pequenos cortes, contusões e escoriações (inclusive as provocadas por cinto de segurança), queimaduras de 1º grau. (até 10% da superfície corporal), fratura dos dentes, pequenas hemorragias externas, pequenas entorses, luxações e/ou fraturas fechadas e/ou abertas dos dedos;

- Lesões graves (feridos graves) são definidas como lesão em pessoa que, por consequência do sinistro, não foi classificada como leve ou não tenha como resultado o óbito;
- Morto é definido como a pessoa em óbito no local (com sinais evidentes de morte ou com a condição de morto constatada por profissional legalmente habilitado) em consequência de sinistro de trânsito.

Em comparação com outros tipos de veículos, como automóveis e motocicletas, o tamanho, o peso bruto e o centro de gravidade mais alto dos caminhões resultam em consequências mais graves quando envolvidos em sinistros (CHANG; MANNERING, 1999; MOKHTAR; PERVEZ, 2012; ZHOU; ZHANG; SHU; YAN, 2019).

Segundo dados do Instituto de Pesquisa Econômica Aplicada IPEA, no ano de 2014 o custo total dos sinistros de trânsito nas rodovias federais, estaduais e municipais atingiu o valor aproximado de R$ 40 bilhões, com um custo médio de R$ 647 mil por sinistro fatal (CARVALHO, 2020). Santana *et al.* (2013) apontam que, embora o TRC seja um setor estratégico para o Brasil, apresenta diversos problemas estruturais, com alto custo social, incluindo alta mortalidade por sinistros de trabalho com motoristas de caminhões.

De acordo com os dados da PRF, em 2014, houve 167.247 sinistros de trânsito nas rodovias federais brasileiras, com 8.233 mortes e 26.182 feridos graves. Esses sinistros geraram um custo para sociedade de R$ 12,8 bilhões, sendo que 62% desses custos estavam associados às vítimas dos sinistros, como cuidados com a saúde e perda de produção devido às lesões ou morte, e 37,4% associados aos veículos, como danos materiais e perda de cargas, além dos procedimentos de remoção dos veículos acidentados. Considerando somente os custos associados aos caminhões envolvidos em sinistros, houve um custo de aproximadamente R$ 135 mil em média por veículo considerando tanto sinistros sem vítimas, com vítimas e vítimas fatais (CARVALHO, 2020).

Já em 2018, de acordo com o Relatório Geral da CNT (2019), estima-se que o prejuízo total para a economia brasileira foi de pelo menos R$ 9,73 bilhões, sendo 57,7% gerados por sinistros com vítimas feridas; 37,7% por sinistros com vítimas fatais; e 4,6% por sinistros sem vítimas.

Segundo Zaloshnja e Miller (2004), as colisões envolvendo caminhões grandes impõem uma variedade de custos ao veículo e ao seu motorista, outros motoristas direta ou indiretamente envolvidos na colisão e à sociedade como um todo. Além de custos como danos materiais, serviços de emergência e atrasos em viagens, lesões e mortes apresentam custos significativos.

Conforme o Relatório Anual da Seguradora Líder (2020), os dados apontam que 12.039 indenizações foram pagas em sinistros envolvendo caminhões em 2020 no Brasil. Estas indenizações englobam despesas médicas, quando há invalidez permanente ou morte dos envolvidos no sinistro (LÍDER, 2021).

O bom desempenho do motorista na condução segura de um veículo depende também das condições e das características da via, associadas ao pavimento, à geometria da via e às sinalizações horizontal e vertical. Essas características, somadas às especificidades dos veículos, aos fatores comportamentais dos motoristas e às condições climáticas, influenciam diretamente o grau de conforto e segurança de um sistema rodoviário e, consequentemente, a propensão à ocorrência de sinistros (CNT, 2019).

Um dos fatores mais críticos que afetam os resultados de segurança no trânsito é a infraestrutura rodoviária e o meio ambiente, como por exemplo, a pavimentação da rodovia, desenho geométrico da via, controle de tráfego, iluminação, condições climáticas, entre outras (ELVIK *et al.*, 2009).

Em 2019, um total de 67.106 quilômetros de rodovias federais (públicas e concedidas) foram avaliados pela CNT em todo o país. Desse total, 34.298 quilômetros (51,11%) apresentam algum tipo de problema, sendo classificados como regular 25.399 quilômetros, ruim 7.269 quilômetros e péssimo 1.630 km. Para esta avaliação, foram feitas análises das características do pavimento, da sinalização e da geometria da via. O estudo apontou que rodovias de baixa qualidade aumentam o risco de sinistros e demandam altos investimentos imediatos seja para manutenção e restauração, seja, em casos mais críticos, para a reconstrução. A cada ano, a Pesquisa CNT de Rodovias vem apontando

problemas de qualidade nas rodovias brasileiras e, devido à falta de investimentos, não são percebidas melhoras ao longo dos anos. Essa deficiência na maior parte da extensão das rodovias avaliadas faz com que o custo operacional aumente, tornando ainda mais cara à prestação do serviço de transporte de carga, que é feito em sua maioria por rodovias (CNT, 2019b).

De acordo com o Relatório Gerencial da CNT (2019b), as características da malha rodoviária, incluindo o seu estado de conservação, a qualidade do pavimento e a sua manutenção contínua, influenciam diretamente a segurança, os custos e a eficiência energética do transporte, refletindo também no meio ambiente e na saúde dos trabalhadores do setor e da população. O setor de transporte no Brasil possui uma participação de 22,8% nas emissões de dióxido de carbono (CO_2) do país, sendo que 89,9% das emissões desse setor (ou 20,5% das emissões do Brasil) advêm do modal rodoviário, contribuindo para o agravamento do aquecimento global e para os efeitos nocivos na saúde pública. Esse cenário se deve ao grande consumo de combustíveis, em especial os de origem fóssil. Assim, as inadequações na pavimentação das rodovias, por exemplo, trechos com buracos, trincas, ondulações e erosões, são alguns dos fatores que levam ao aumento do consumo de combustível nos veículos e, consequentemente, das emissões atmosféricas (CNT, 2019b).

Segundo o Balanço Energético Nacional - BEN 2019, o setor de transporte ocupa o primeiro lugar no que diz respeito ao uso de energia no Brasil, com participação de 32,7% no consumo total de energia do país em 2018, ultrapassando a produção industrial. O maior responsável por esse quadro é o modal rodoviário, com consumo de 93,3% da energia destinada a toda atividade transportadora. O diesel derivado de petróleo se destaca como a principal fonte de energia utilizada, representando 45,3% do seu consumo energético, seguido da gasolina (27,6%) e do etanol (20,1%). Ressalta-se, ainda, que os combustíveis fósseis, não renováveis e poluentes, correspondem a mais de 75% do consumo de energia no transporte rodoviário nacional (CNT, 2019b).

Estima-se que o estado péssimo do pavimento pode quase dobrar o custo operacional do TRC, uma vez que o adicional de custos pode chegar a até 91,5%. Assim, considerando os resultados da Pesquisa CNT de Rodovias 2019, calcula-se que o país gasta, em média, 28,5% a mais do que deveria para

transportar seus insumos, bens de produção e bens de consumo por rodovias, apenas em razão de problemas no pavimento (CNT, 2019b).

Tradicionalmente, atribui-se a um sinistro fatores humanos, veiculares e ambientais de modo que em cada um desses domínios existem uma série de variáveis capazes de aumentar ou reduzir as chances de um sinistro ocorrer. No fator humano, a incidência de comportamentos de risco influencia diretamente na ocorrência de sinistros; no fator veicular, as condições de conservação, manutenção e disponibilidade de itens de segurança dos veículos também são determinantes e no fator ambiental, tanto as condições climáticas quanto a qualidade da infraestrutura viária têm papel preponderante na ocorrência de sinistros (BASTOS *et al.*, 2020).

Dentre os fatores contribuintes aos sinistros com caminhões, a fadiga e a sonolência dos motoristas de caminhão também são apresentadas, tanto pela literatura científica internacional como nacional, como um dos fatores preponderantes na causalidade de sinistros, causados pelo tempo de direção (FRAGOSO; GARCIA, 2019).

A legislação do Código de Trânsito Brasileiro (CTB), especificamente nos seus art. 67-A, 67-B, 67-C, 67-D, 67-E dispõe sobre normas, aplicadas aos motoristas rodoviários de cargas e de passageiros, para o período de descanso e de direção e o art. 99, define que somente poderão transitar pelas vias terrestres veículos cujo peso e dimensões atenderem aos limites estabelecidos pelo Conselho Nacional de Trânsito (CONTRAN), delimitando que o excesso de peso será mensurado por equipamento de pesagem ou pela verificação de documento fiscal, bem como, possibilitando o estabelecimento de percentuais de tolerância sobre os limites de Peso Bruto Total (PBT) e Peso Bruto por Eixo. Com relação ao peso, a Resolução nº 210/2006 estabelece os limites de peso e dimensões para veículos que transitem por vias terrestres. A Resolução nº 525/2015, também regulamenta e dispõe sobre a fiscalização do tempo de direção do motorista profissional e a Lei nº 13.103/2015 disciplina a jornada de trabalho e também o tempo de direção do motorista profissional.

Importante ressaltar, que a atividade de motorista profissional de carga ou de passageiros ultrapassa os arredores da empresa, colocando o trabalhador em contato direto com os usuários das vias. Desse modo, as legislações visam garantir maior segurança aos usuários das vias e redução nos números de

sinistros causados também por dependentes de substâncias psicoativas (KONZEN, 2017). A Resolução nº 691/2017, dispõe sobre o exame toxicológico em motoristas profissionais.

A PRF coleta e divulga, desde 2007, dados de sinistros que aconteceram nas rodovias federais do país. Esses dados estão disponíveis no site da PRF através de um conjunto de arquivos .CSV contendo informações de cada sinistro, separados por ano. De 2007 a 2017, mais de 1,6 milhão de sinistros foram registrados nas rodovias brasileiras (BRASIL, 2021).

Existem outras bases de dados referentes aos sinistros de trânsito no Brasil. O seguro obrigatório DPVAT, destinado a indenizar vítimas de sinistros de trânsito ocorridos em todo o território nacional, sejam pedestres, passageiros ou motoristas, brasileiros ou estrangeiros é administrado pela Seguradora Líder. A seguradora realiza análises sobre os sinistros de trânsito, apresentando relatórios anuais, informando sobre as indenizações, frota de veículos e perfil das vítimas de trânsito por estado, com informações sobre tipo do envolvido no sinistro, idade, horário do sinistro e tipo do veículo envolvido (LÍDER, 2021).

A Associação Brasileira de Concessionárias de Rodovias (ABCR) representa o setor de concessões de rodovias, esta também apresenta informações dos sinistros com base nos dados da PRF. A base contém informações do total de sinistros, número de pessoas envolvidas, número de pessoas sem lesão, número de vítimas feridas, número de mortos e tipo de veículos envolvidos nos sinistros de 12 estados: Bahia, Espírito Santo, Goiás, Minas Gerais, Mato Grosso do Sul, Mato Grosso, Paraná, Pernambuco, Rio de Janeiro, Rio Grande do Sul, Santa Catarina e São Paulo (ABCR, 2020).

Informações sobre sinistros de trânsito em rodovias concedidas também podem ser obtidas no Centro Nacional de Supervisão Operacional (CNSO) da Agência Nacional de Transportes Terrestres (ANTT). O sistema apresenta um monitoramento de acidentes, onde pode-se selecionar o ano e o trecho de concessão para obter os seguintes dados: total de sinistros, número de vítimas ilesas, número de vítimas leves, número de vítimas moderadas, número de vítimas graves e óbitos naquele trecho. O número de mortes em sinistros ocorridos nas rodovias reguladas pela ANTT e administradas pelas 21 concessionárias apresentou queda de 13,2% entre 2015 e 2019 (ANTT, 2020).

Há também o Infosiga SP, para auxiliar na elaboração de políticas públicas relacionadas à segurança no trânsito, ele contém um banco de dados que reúne informações de sinistros no Estado de São Paulo de diversas fontes, como Polícia Civil, Polícia Militar e Polícia Rodoviária Federal. Atualizado mensalmente, o Infosiga SP fornece dados de faixa etária e sexo da vítima, tipo do veículo envolvido e perfil do sinistro. Ele ainda possui o Infomapa SP, que traz a posição geográfica das ocorrências com vítimas fatais no estado. Nele é possível ver a localização dos sinistros com automóveis, motocicletas, pedestres, ônibus, caminhões, bicicletas e outros que causaram mortes, com indicações da faixa etária da(s) vítima(s), o período em que aconteceu o acidente (manhã, tarde, noite e madrugada) e o tipo de ocorrência (INFOSIGA, 2021).

Melhorar a segurança no trânsito e os riscos de sinistros é um dos objetivos mais importantes para os formuladores de políticas de transporte na sociedade contemporânea. A NBR ISO 39001/2015 (Sistemas de Gestão de Segurança Viária) da Associação Brasileira de Normas Técnicas (ABNT) (2015) introduziu as diretrizes de atividades baseadas na segurança destinadas a reduzir os sinistros rodoviários, de acordo com os Sistemas de Gestão da Qualidade (ISO 9000). Essas diretrizes destinam-se a gerentes de infraestrutura, administradores e entidades privadas e definem um gerenciamento padrão para redução do risco de rodovias (CONCA; RIDELLA; SAPORI, 2016).

Segundo a NBR ISO 39001/2015 (Sistemas de Gestão Segurança Viária SV – Requisitos com orientações para uso) da Associação Brasileira de Normas Técnicas (ABNT, 2015), a norma tem como objetivo especificar os requisitos para um sistema de gestão que permita a uma organização, que interage com o sistema viário, reduzir mortes e ferimentos graves em sinistros de trânsito.

Conforme Martins e Garcez (2019), os registros de sinistros de trânsito das rodovias federais do Brasil estão armazenados nos bancos de dados disponibilizados pela Policia Rodoviária Federal (PRF). Neste banco de dados contém fatos sobre as ocorrências e as pessoas envolvidas nos sinistros. Ainda segundo os autores, visto a grande quantidade de informações sobre os sinistros, tem-se como resultado que o banco de dados contém centenas de linhas onde cada linha representa um registro sobre determinada ocorrência de sinistro ou uma pessoa envolvida. Porém, é importante destacar que além do grande volume de registros, eles estão em forma de dados, ou seja, são vistos

como fatos crus e para que passem a ter um significado é necessário transformá-los em informação através da organização e refinação dos mesmos, passando a ter valor adicional além dos fatos individuais.

2.5 ANÁLISE DE DADOS UTILIZANDO MINERAÇÃO DE DADOS

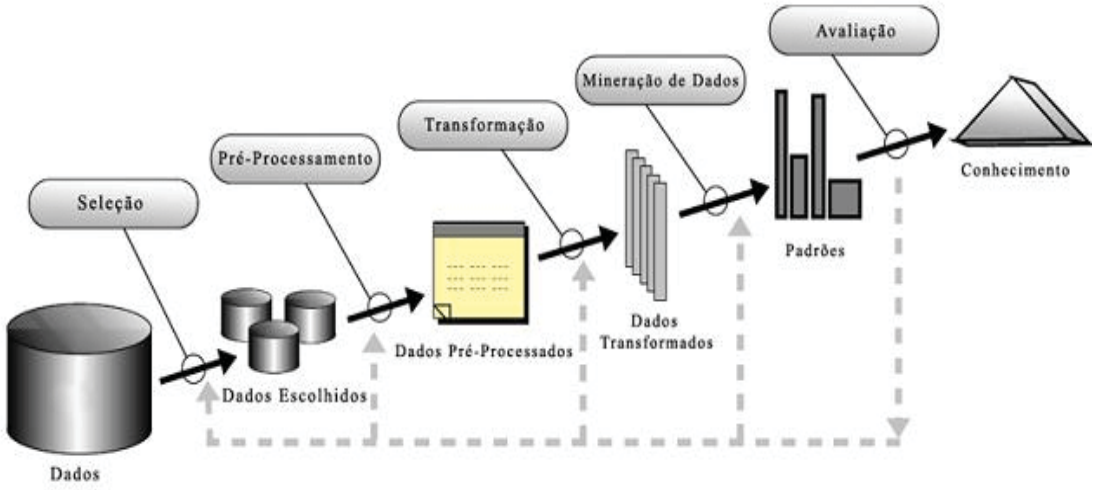
Há uma quantidade significativa de dados armazenados nos bancos de dados, assim, é necessário encontrar técnicas para extrair informações e conhecimentos, explorando esses dados armazenados para uso na solução de problemas e tomada de decisão. A mineração de dados (MD) é um processo analítico que combina inteligência artificial, estatística e aprendizado de máquina (NAFIE ALI; MOHAMED, 2018).

A MD e o Aprendizado de Máquina são tópicos da inteligência artificial que se concentram na descoberta e previsão de padrões com base nos dados coletados (WITTEN *et al.,* 2016).

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de Dados) é uma tentativa de solucionar o problema causado pela chamada “era da informação”: a sobrecarga de dados. Nesse contexto, a descoberta de conhecimento em bases de dados (KDD) pode ser definida como o processo de extração de informação a partir de dados registrados em uma base de dados, um conhecimento implícito, previamente desconhecido, potencialmente útil e compreensível (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O conceito mais utilizado para o termo KDD é de Fayyad, Piatetsky-Shapiro e Smyth (1996), que o define como um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão, sendo a MD uma das etapas desse processo, como demonstrado na FIGURA 1.

FIGURA 1 - FASES DO PROCESSO DE KDD



Fonte: Adaptado de FAYYAD, PIATETSKYSHAPIRO E SMYTH (1996).

Segundo Corcovia e Alves (2019), as etapas de KDD são descritas como:

1. Seleção: Criação de um conjunto de dados alvo, selecionar um conjunto de dados, ou focar num subconjunto, onde a descoberta deve ser realizada. Selecionar ou segmentar os dados de acordo com critérios definidos;
2. Pré-processamento: Operações básicas tais como, remoção de ruídos quando necessário, manipular campo de dados ausentes, formatação de dados de forma a adequá-los à ferramenta de mineração;
3. Transformação: Redução de dados e projeção, com a utilização de características úteis para representar os dados dependendo do objetivo da tarefa, com o objetivo de reduzir o número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados;
4. Mineração de Dados: Escolha e execução do algoritmo de aprendizagem de acordo com a tarefa a ser cumprida. É a verdadeira extração dos padrões de comportamento dos dados;
5. Interpretação - Avaliação: Interpretação dos resultados, com possível retorno aos passos anteriores; consolidação; incorporação e documentação do conhecimento e comunicação aos interessados; identificado os padrões estes são interpretados, e os mesmos darão suporte a tomada de decisões.

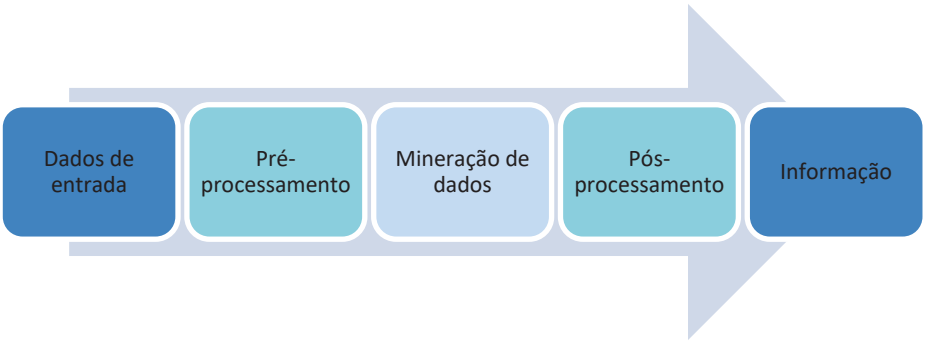
De acordo com Han, Pei e Kamber (2011), após o pré-processamento, os métodos de MD poderão ser aplicados. Os padrões descobertos pelos

métodos são extraídos e pode haver a necessidade da aplicação de um pós-processamento:

- Avaliação dos padrões encontrados: verifica os padrões que são realmente interessantes para a descoberta de conhecimento;
- Apresentação do conhecimento: utiliza técnicas de visualização e representação de conhecimento para apresentar o conhecimento extraído.

Calil *et al.* (2008) corroboram estabelecendo que as fases do KDD podem ser agrupadas em três grandes grupos: pré-processamento, mineração de dados e pós-processamento. De acordo com os autores, o pré-processamento inclui todas as etapas que consideram a preparação da base de dados, cujos dados serão fornecidos como entrada para o(s) algoritmo(s) de MD. Ainda de acordo com os autores, o pós-processamento contempla a depuração e/ou síntese dos padrões descobertos. Tan, Steinbach e Kumar (2009) acrescentam que o processo consiste em uma série de passos de transformação, desde o pré-processamento dos dados, passando pela MD até o pós-processamento. A FIGURA 2 exemplifica o processo descrito.

FIGURA 2 - PROCESSO DA MINERAÇÃO DE DADOS



FONTE: Adaptado de TAN, STEINBACH E KUMAR (2009).

Nas próximas subseções serão abordados os três grandes grupos: pré-processamento, mineração de dados e pós-processamento.

2.5.1 Pré-Processamento

A etapa de pré-processamento compreende as funções que se relacionam a captação, à organização e ao tratamento de dados, cujo objetivo é preparar os dados para os algoritmos da etapa seguinte (mineração de dados) (GOLDSCHMIDT; PASSOS, 2005).

Um passo necessário para a realização da MD é o pré-processamento dos dados. Han, Pei e Kamber (2011) definiram algumas etapas:

- Seleção dos dados: recupera os dados relevantes ao domínio desejado;
- Tratamento dos dados: remove o ruído e dados inconsistentes;
- Enriquecimento dos dados: complementa os dados existentes com outros de diferentes bancos de dados;
- Transformação dos dados: codifica os dados de acordo com as entradas dos métodos utilizados na mineração.

Segundo Castro e Ferrari (2016) são três problemas que podem ocorrer com os dados e que comprometem a qualidade dos mesmos: a incompletude, a inconsistência e o ruído.

A incompletude significa que os dados não estão completos, faltando um atributo importante para determinar a entidade (por exemplo, entidade estudante sem o atributo nome do estudante) ou um valor para um atributo de um determinado objeto (por exemplo, a ausência do objeto idade do estudante “18 anos” no atributo idade do estudante). Entretanto, nem sempre essa ausência é notada, a não ser quando um especialista no domínio do problema analisa a base e percebe a falta (CASTRO; FERRARI, 2016).

A inconsistência ocorre quando um valor está fora do domínio do atributo ou apresenta grande discrepância em relação aos outros dados. O especialista percebe a inconsistência quando compara os dados de um conjunto de atributos (por exemplo, para a entidade estudante tem-se os atributos e os objetos a seguir: nome: “Antônio Santos”, idade: 05 anos, estado civil: casado, ao comparar os dados percebe-se que é impossível um estudante de 05 anos ser casado, ou seja, um dos atributos está inconsistente). A inconsistência pode ser entendida como discrepância. Esse problema geralmente ocorre nos casos de integração de dados que estejam, inicialmente, em bases de dados divergentes (CASTRO; FERRARI, 2016).

O ruído, dentro do contexto do KDD, está relacionado as modificações dos valores originais, incidem em erros de medidas ou em valores muito diferentes de outros valores da mesma base de dados, sendo chamados de *outliers*. São complexos de serem notados porque para percebê-los é necessário conhecer os domínios esperados dentro de um atributo ou qual a distribuição deveria ocorrer para os valores dos atributos, pois ele ocorre com valores não esperados (SILVA; PERES; BOSCARIOLI, 2016).

Um dado ruidoso é aquele que apresenta alguma variação em relação ao seu valor sem ruído e, portanto, ruídos na base de dados podem levar a inconsistências. Dependendo do nível de ruído, nem sempre é possível saber se ele está ou não presente em um dado (CASTRO; FERRARI, 2016).

De acordo com Castro e Ferrari (2016), os dados são valores quantitativos ou qualitativos associados a alguns atributos. Com relação a estrutura, os autores afirmam que eles podem ser: estruturados, semiestruturados ou não estruturados.

Os dados estruturados são aqueles que estão dentro de estruturas tabulares. São organizados em tabelas (entidade) com colunas (atributo) e linhas (instância). Normalmente resultam de processos de geração de dados de sistemas transacionais (SILVA; PERES; BOSCARIOLI, 2016). Segundo Castro e Ferrari (2016), uma base de dados é estruturada quando os dados residem em campos fixos em um arquivo – por exemplo, uma tabela, uma planilha ou um banco de dados. Uma das vantagens dos dados estruturados apresentada pelos autores é a facilidade de armazenagem, acesso e análise.

Os dados semiestruturados são os dados estruturados de forma menos rígida, não possuem a estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Geralmente, nos dados semiestruturados, são usados marcadores (por exemplo, *tags*) para identificar certos elementos dos dados, mas a estrutura não é rígida (BUNEMAN, 1997; CASTRO; FERRARI, 2016).

Os dados não estruturados são definidos como aqueles que não possuem um modelo de dados, que não estão organizados de uma maneira predefinida ou que não residem em locais definidos. Os autores complementam que geralmente se referem a textos livres, imagens, vídeos, sons, páginas web,

arquivos PDF, entre outros. Assim, costumam ser de difícil estruturação, acesso e análise (CASTRO; FERRARI, 2016; SILVA; PERES; BOSCARIOLI, 2016).

Outro conceito importante do pré-processamento é o atributo. O atributo corresponde à característica de um objeto; é um subconjunto de dados que qualifica e modela entidades para que seja possível efetuar uma análise com maior precisão. O valor de um atributo de um dado objeto é uma medida de quantidade daquele atributo, podendo ser numérica ou categórica. Os atributos numéricos podem assumir quaisquer valores numéricos por exemplo, valores discretos (inteiros) ou contínuos (reais) ao passo que as quantidades categóricas assumem valores correspondentes a símbolos distintos (CASTRO; FERRARI, 2016; SILVA; PERES; BOSCARIOLI, 2016).

O atributo categórico pode ser binário, nominal ou ordinal. O atributo binário é aquele que pode assumir apenas dois valores possíveis, por exemplo, “0” ou “1”. O atributo nominal é definido como aquele cujos valores possuem símbolos ou rótulos distintos, por exemplo: o atributo fumante (fumante ou não fumante), não existe ordem entre suas categorias. O atributo ordinal é definido como aquele que permite ordenar suas categorias, embora não necessariamente haja uma noção explícita de distância entre as categorias, por exemplo: o atributo altura dos estudantes (baixo, médio e alto) (SILVA, 2015; CASTRO; FERRARI, 2016).

Conforme Castro e Ferrari (2016), o pré-processamento dos dados está associado a três fatores que quase sempre estão interligados:

1) Entender os problemas existentes na base de dados (incompletude, inconsistência e ruído);
2) Ter identificado o problema que está tentando resolver com as bases de dados;
3) Compreender as características das técnicas de mineração de dados a serem utilizadas na MD.

2.5.2 Mineração de Dados

Para Simoudis (1996), a MD é o processo de extração de informações válidas, compreensíveis e úteis e previamente desconhecidas, a partir de grandes bases de dados e utilizadas para a tomada de decisões cruciais. Já para

Dunham (2003), minerar dados é encontrar informações escondidas em um banco de dados.

A MD, também conhecida como *data mining*, é a área que busca novos padrões e relacionamentos interessantes em uma grande quantidade de dados. A MD é definida como o processo de descoberta de novas correlações significativas, padrões e tendências, trabalhando em quantidades abundantes de dados armazenados em depósitos (CHAWLA; SHARMA, 2016, GONZALEZ, *et al*., 2016).

Tan, Steinbach e Kumar (2009) consideram que “a mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados”. A partir desta conceituação, podem-se destacar quatro elementos essenciais: processo, descoberta automática, informações úteis e grandes depósitos de dados. Como processo, significa que a MD é um método ou sistema com regras específicas. A descoberta automática implica a obtenção de resultados por meios puramente mecânicos, sendo por isso, em algumas vezes, resultados imprevisíveis à cogitação humana. O elemento informações úteis significa que os resultados obtidos devem ser proveitosos para a tomada de decisões. Por último, o elemento depósito de dados implica, *a priori*, a existência de um sistema gerenciador de banco de dados para o armazenamento, indexação e processamento de consultas (TAN; STEINBACH; KUMAR, 2009).

Ainda segundo os autores, a MD é associada à influência de áreas como: Estatística (amostragem, estimativa e teste de hipóteses), Inteligência Artificial (algoritmos de busca, técnicas de modelagem e teorias de aprendizagem), Ciência da Informação (recuperação da informação), entre outras (TAN; STEINBACH; KUMAR 2009).

Uma base de dados pode ser definida como uma coleção de objetos que podem ser eventos, observações ou registros. Estes objetos de dados são caracterizados por valores em um conjunto de características pré-determinadas chamadas de atributos (TAN; STEINBACH; KUMAR 2009; WITTEN; FRANK; HALL, 2011).

Os valores dos atributos em um determinado registro são uma representação simbólica ou numérica das características reais do objeto. Diferentes tipos de atributos são utilizados para verificar a consistência entre as

propriedades dos valores medidos e as propriedades do atributo (TAN; STEINBACH; KUMAR, 2009).

Os atributos podem ser classificados como quantitativos ou qualitativos. Os atributos quantitativos podem ser representados por valores reais ou valores inteiros possuindo propriedades de ordem e distância entre os valores. Por sua vez, os atributos qualitativos possuem como valores símbolos ou nomes (rótulos) que mesmo sendo representados com números, não possuem a maioria das propriedades dos atributos numéricos, considerando que estes valores podem ser iguais ou diferentes entre si (KANTARDZIC, 2011).

Os tipos de atributos podem ainda serem divididos em atributos nominais e ordinais (no caso dos atributos qualitativos) e atributos intervalares e proporcionais (no caso dos atributos quantitativos) (TAN; STEINBACH; KUMAR, 2009).

Atributos nominais nos fornecem apenas informações para distinção de um objeto a outro não possuindo uma ordem significativa. Cada valor representa uma categoria, código ou estado sendo possível apenas realizar operações de igualdade e desigualdade (TAN; STEINBACH; KUMAR, 2009; KANTARDZIC, 2011).

Atributos Ordinais possuem valores que fornecem informações suficientes para ordenar os objetos, sem existir relação de distância entre os valores, não sendo possível dizer quão diferente um valor é de outro (TAN; STEINBACH; KUMAR, 2009; WITTEN; FRANK; HALL, 2011).

Os atributos intervalares são representados em unidades fixas e de igual tamanho permitindo valores positivos e negativos. As diferenças entre os valores são significativas, sendo possível a realização de operação de soma e subtração, ou seja, existe uma unidade de medida. Além disso, o valor zero não é considerado um “zero-real”, já que o zero não é definido como ausência de característica (TAN; STEINBACH; KUMAR, 2009; HAN; PEI; KAMBER, 2011).

Já os atributos proporcionais, tanto a proporção quanto a diferença entre os valores são significativas, possuindo a representação do “zero-real”. É possível realizar a ordenação dos valores assim como o cálculo da diferença entre eles (TAN; STEINBACH; KUMAR, 2009).

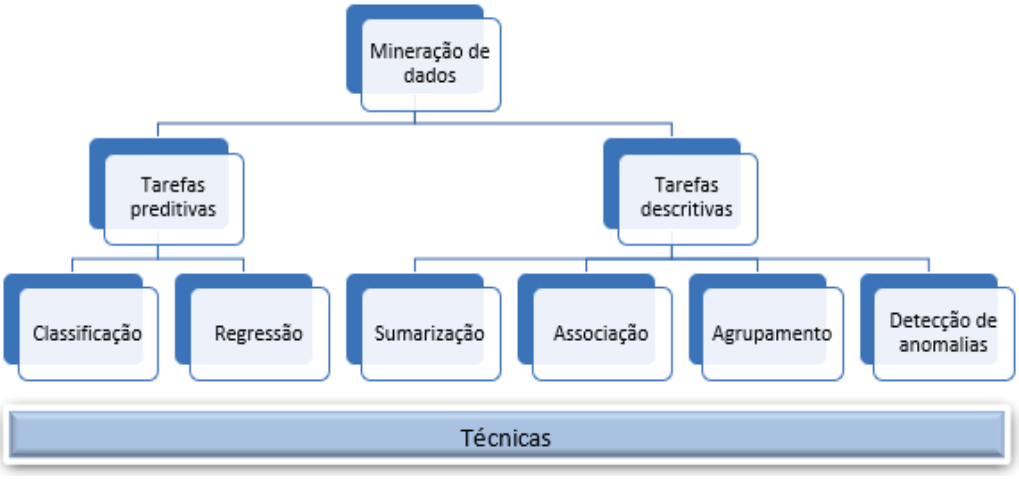
Para especificar os tipos de padrões a serem encontrados em um modelo de descoberta de conhecimento são utilizadas funcionalidades da

mineração de dados. De forma geral, as tarefas de mineração de dados podem ser classificadas em duas categorias principais: descritiva e preditiva (HAN, 2005).

As tarefas preditivas, que são baseadas em algoritmos supervisionados, têm como objetivo prever o valor de um determinado atributo baseado nos valores de outros atributos. O atributo a ser previsto é comumente conhecido como a variável dependente ou alvo, enquanto que os atributos para fazer a previsão são conhecidos como as variáveis independentes ou explicativas. Já as tarefas descritivas, que são baseadas em algoritmos não supervisionados, derivam padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumam os relacionamentos subjacentes nos dados. As tarefas descritivas da mineração de dados são muitas vezes exploratórias em sua natureza e frequentemente requerem técnicas de pós-processamento para validar e explicar os resultados (TAN; STEINBACH; KUMAR, 2009).

As tarefas preditivas podem ser implementadas por meio de técnicas de classificação ou regressão, ao passo que as tarefas descritivas contemplam as técnicas de sumarização, associação, agrupamento e detecção de anomalias (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996; LAROSE, 2005; TAN, STEINBACH e KUMAR, 2009). A FIGURA 3 demonstra as tarefas e técnicas da mineração de dados de acordo com a abordagem de Fayyad, Piatetsky-Shapiro e Smyth (1996); Larose (2005) e Tan, Steinbach e Kumar (2009).

FIGURA 3 - PRINCIPAIS TAREFAS E TÉCNICAS DA MINERAÇÃO DE DADOS



FONTE: A autora (2021).

Nas próximas subseções serão analisadas as principais técnicas de mineração de dados e seus respectivos algoritmos. Inicialmente serão analisadas as tarefas preditivas (classificação e regressão) e posteriormente as tarefas descritivas (sumarização, associação, agrupamento e detecção de anomalias).

2.5.2.1 Classificação

Segundo Hodeghatta e Nayak (2016) a classificação é uma subcategoria das tarefas preditivas, cujo objetivo é determinar a qual classe determinado registro pertence. O processo é realizado em duas etapas, treinamento e teste. O conjunto de dados é dividido em duas partes, normalmente 70% dos dados para treinamento e 30% dos dados para teste.

No treinamento, um modelo é construído a partir da análise de dados provenientes de uma amostra de dados de treinamento e o conjunto de atributos que definem a variável classe. Os dados do treinamento são uma amostra do banco de dados e o atributo de classe já é conhecido. Já no teste, o modelo gerado na primeira etapa (treinamento) é aplicado em uma segunda amostra de dados, dados de teste, onde a variável de classe é predita e comparada com o valor existente no banco de dados, apurando assim, a acurácia do modelo (HODEGHATTA; NAYAK, 2016).

De acordo com Castro e Ferrari (2016), o treinamento consiste em usar os dados de treinamento para ajustar os parâmetros livres do modelo, de tal forma que o seu desempenho atinja determinado nível de qualidade, avaliado pela aplicação do modelo gerado aos dados de teste.

Na classificação, determina-se a relação entre a variável classe e as entradas ou variáveis explicativas. Normalmente, os modelos representam as regras de classificação ou fórmulas matemáticas. Depois que essas regras são criadas, o modelo pode ser usado para prever a classe de outros conjuntos de dados onde a determinada classe é desconhecida (HODEGHATTA; NAYAK, 2016).

De acordo com Castro e Ferrari (2016), classificar um objeto significa atribuir a ele um rótulo, chamado classe, de acordo com a categoria à que ele

pertence. Assim, um algoritmo de classificação é usado na construção de um modelo de classificação, também chamado classificador, o qual é construído com base em um conjunto de treinamento com dados rotulados. Os autores destacam que há uma grande variedade de algoritmos de classificação na literatura, sendo possível separá-los de acordo com a sua estrutura em: baseados em conhecimento; baseados em árvore; conexionistas; baseados em distância; baseados em função; e probabilísticos.

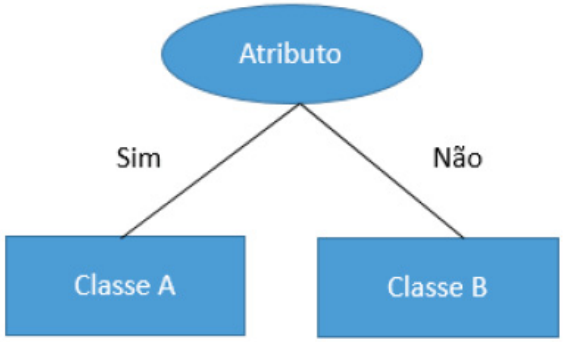
Ainda de acordo com os autores, o modelo baseado em conhecimento opera por meio de um conjunto de regras usadas para atribuir determinada classe a um objeto caso ele satisfaça condições predefinidas conforme FIGURA 4 (CASTRO; FERRARI, 2016).

FIGURA 4 - MODELO BASEADO EM CONHECIMENTO

FONTE: Adaptado de CASTRO E FERRARI (2016)

Conforme FIGURA 5, no modelo baseado em árvores, o nó raiz e os nós intermediários das árvores representam testes sobre um atributo, os ramos representam os resultados desses testes e os nós folhas, os rótulos de classe (COPPIN, 2010; CASTRO; FERRARI, 2016).

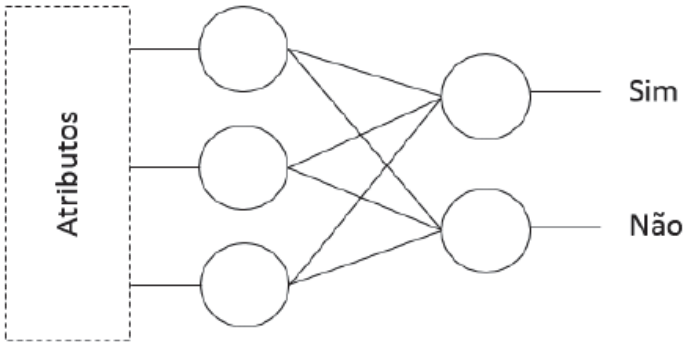
FIGURA 5 - MODELO BASEADO EM ÁRVORES



FONTE: Adaptado de CASTRO E FERRARI (2016)

Os modelos conexionistas são definidos pelos autores como aqueles baseados em redes de unidades (nós) interconectadas. Os autores afirmam que os sistemas conexionistas são um tipo de grafo e, embora haja diferentes sistemas conexionistas, os mais comuns são as Redes Neurais Artificiais (RUSSELL; NORVIG, 2013; CASTRO; FERRARI, 2016). A FIGURA 6 representa o modelo.

FIGURA 6 - MODELO CONEXIONISTA

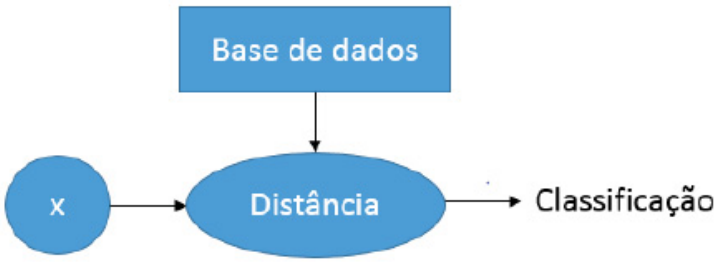


FONTE: CASTRO E FERRARI (2016)

Segundo Castro e Ferrari (2016), nos modelos baseados em distância, representado na FIGURA 7, o processo de classificação se dá calculando a

distância entre o objeto cuja classe se deseja conhecer e um ou mais objetos rotulados. Os autores apontam que a classe do objeto desconhecido passa a ser a mesma daqueles objetos que estão a uma menor distância dele.

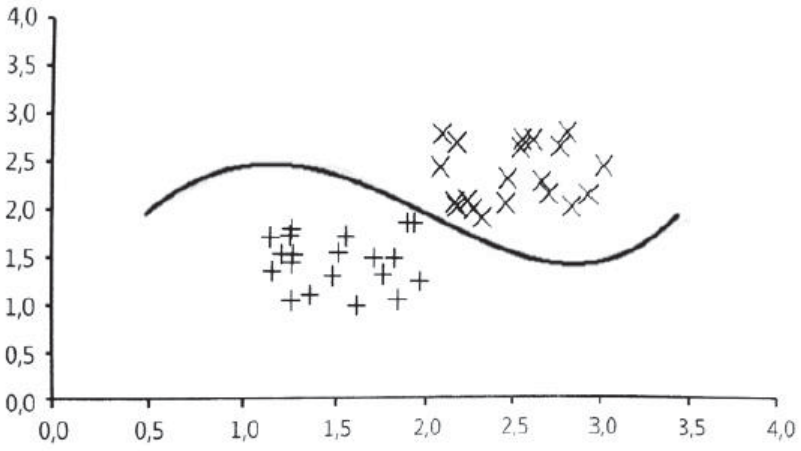
FIGURA 7 - MODELO BASEADO EM DISTÂNCIA



FONTE: Adaptado de CASTRO E FERRARI (2016)

De acordo com Castro e Ferrari (2016), os modelos baseados em função são paramétricos baseados em funções predefinidas e cujos parâmetros são ajustados durante o processo de treinamento. Os autores afirmam que após o treinamento, um novo objeto de classe desconhecida é apresentado à função, cujo valor é calculado e que representa, de alguma forma, a classe desse objeto. A FIGURA 8 representa este modelo.

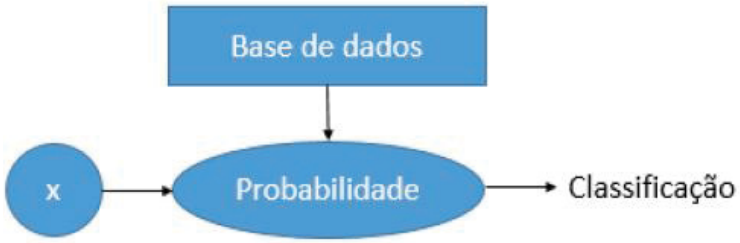
FIGURA 8 - MODELO BASEADO EM FUNÇÃO



FONTE: CASTRO E FERRARI (2016)

Por fim, segundo Castro e Ferrari (2016), o modelo probabilístico permite atribuir uma probabilidade de um objeto pertencer a uma ou mais classes possíveis, conforme demonstra a FIGURA 9.

FIGURA 9 - MODELO PROBABILÍSTICO



FONTE: Adaptado de CASTRO E FERRARI (2016)

Entre os principais métodos (algoritmos) de classificação na literatura, Castro e Ferrari (2016) destacam: classificador k-NN; árvores de decisão; regras de classificação; classificador *one-rule* (1R); e classificador *Naive Bayes*.

O classificador k-NN (k-*nearest neighbors*) corresponde ao método dos k-vizinhos mais próximos, é um dos classificadores não paramétricos baseados em distância. Ele opera da seguinte maneira: dado um objeto x_0 cuja classe se deseja inferir, encontra-se os k objetos x_i i = 1, ..., k da base que estejam mais próximos a x_0 e, depois, se classifica o objeto x_0 como pertencente à classe da maioria dos k-vizinhos mais próximos (CASTRO; FERRARI, 2016). Existem várias funções para calcular a distância entre duas características, tais como, distância *Manhattan*, distância euclidiana, distância de cosseno ou correlação (EBRAHIMPOUR; KOUZANI, 2007).

O único parâmetro a ser definido no k-NN é o valor de k, ou seja, o número de vizinhos mais próximos a serem considerados para se definir a classe de x_0. Embora não exista uma regra para se definir k, valores grandes reduzem o efeito dos ruídos na classificação, mas tornam fronteiras de classe menos definidas. Geralmente, o valor de k é definido por alguma heurística ou por

tentativa e erro (CASTRO; FERRARI, 2016). O Algoritmo é descrito na FIGURA 10 como um pseudocódigo do k-NN.

FIGURA 10 - PSEUDOCÓDIGO DO ALGORITMO K-NN

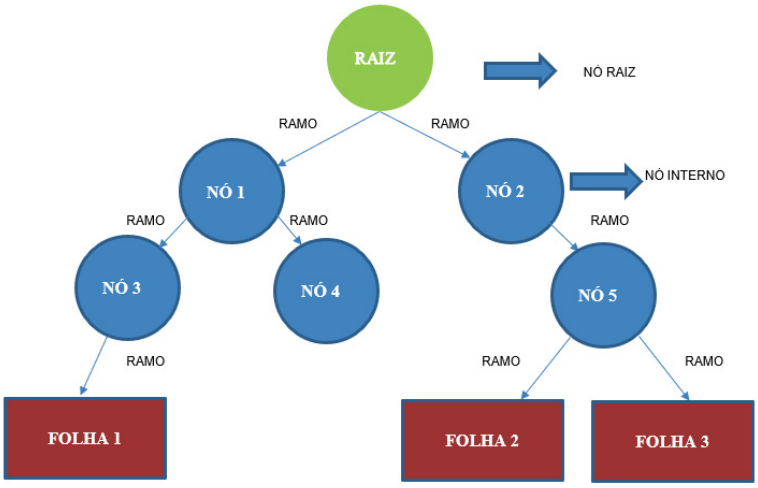
```
Entrada
    k : número de vizinhos
    data : base de dados com n objetos e m atributos (n x m)
    classe : vetor contendo a classe de cada objeto da base (n x 1)
    obj : objeto que deve ser classificado (1 x m)
Saída
    C : rótulo indicativo da classe do objeto
Passos
    // Calcular a distância entre base de dados e o objeto D(n x 1))
    D = dist(data,obj);
    Obj = Ø;
    // Determinar os k objetos mais próximos
    Para i=1:k Faça
    {
        aux = D[1];
        pos = 1;
        Para j=2:n Faça
        {
            Se (aux > D[j]) e (j ∩ obj == Ø) Então
            {
                aux = D[j];
                pos = j;
            }
        }
        Objs.Add(pos);
    }
    // Pegar a classe dos k objetos mais próximos
    Ck = classe[objs];

    // Determinar a classe mais frequente
    C = moda(Ck);
```

FONTE: CASTRO E FERRARI (2016)

Uma árvore de decisão (*decision tree*) é uma estrutura em forma de árvore na qual cada nó interno corresponde a um teste de um atributo, cada ramo representa um resultado do teste e os nós folhas representam classes ou distribuições de classes. O nó mais elevado da árvore é conhecido como nó raiz, e cada caminho da raiz até um nó folha corresponde a uma regra de classificação (CASTRO; FERRARI, 2016; YADAV & CHANDEL, 2015). Os elementos da árvore de decisão podem ser visualizados na FIGURA 11.

FIGURA 11 - ELEMENTOS DE UMA ÁRVORE DE DECISÃO



FONTE: Adaptado de QUINLAN (1993).

Árvores de decisão são representações estruturadas de um conjunto de dados. Um nó marca uma decisão a ser tomada a partir de várias alternativas e cada nó terminal indica uma determinada classificação (AZIZ; SANAA; HASSANIEN, 2017). Uma árvore utiliza uma estratégia de dividir para conquistar, onde um problema complexo é decomposto em subproblemas mais simples e recursivamente a mesma estratégia é aplicada a cada subproblema (QUINLAN, 1993).

Segundo Quinlan (1993), as arvores de decisão classificam instâncias, classificando-os da raiz da árvore para baixo até algum nó folha, que prevê a classificação da instância. Cada nó da árvore especifica um teste de algum atributo e cada ramo descendente, a partir desse nó, corresponde a um dos valores possíveis para este atributo. Um exemplo é classificar, iniciando no nó raiz da árvore, testando o atributo especificado por este nó, em seguida, movendo-se para baixo, para o galho da árvore correspondente ao valor do atributo no exemplo dado. Este processo é repetido para a sub-árvore com raiz no novo nó.

A FIGURA 11 representa uma arvore de decisão binária. Ela consiste de nodos e ramos. Cada nodo representa um simples teste ou decisão. No caso de uma árvore binária, a decisão pode ser verdadeira ou falsa. O nodo inicial é comumente referido como nodo raiz. Dependendo do resultado do teste, a árvore poderá se ramificar à esquerda ou à direita em direção a outro nodo. Por fim, o nodo terminal, conhecido também como nodo folha, representado na forma de quadrados, é alcançado, e uma decisão é realizada a uma classe designada. Na

construção de árvore é normalmente utilizada a convenção de realizar decisões verdadeiras em ramos da direita e decisões falsas no ramo da esquerda (DE LIMA, 2017).

As árvores de decisão visualizam o domínio da tarefa como uma classificação. A estrutura subjacente consiste em uma coleção de atributos ou propriedades que são utilizadas para descrever casos individuais, cada caso pertencente a um exato conjunto de classes. Os atributos podem ser contínuos ou discretos. O valor de um caso de um atributo contínuo é sempre um número real, enquanto seu valor de um atributo discreto é um pequeno conjunto de valores possíveis para esse atributo (QUINLAN, 1993).

Uma árvore utiliza uma estratégia de dividir para conquistar, onde um problema complexo é decomposto em subproblemas mais simples e recursivamente a mesma estratégia é aplicada a cada subproblema (QUINLAN, 1993).

Para construir a árvore de decisão, foram desenvolvidos alguns algoritmos, como: ID3 (QUINLAN, 1979), CART (BREIMAN *et al.,* 1984), C4.5 (QUINLAN, 1993), abrangendo outros. Os principais passos do algoritmo para construção de uma árvore podem ser descritos na FIGURA 12. Obtendo como entrada para a função GeraÁrvore um conjunto de dados D. No passo 3, o algoritmo avalia o critério de parada. Se mais divisões do conjunto de dados são necessárias, é escolhido o atributo que maximiza alguma medida de impureza, descrito no passo 5. No passo 7, a função GeraÁrvore é recursivamente aplicada a cada partição do conjunto de dados D (FACELI *et al.,* 2011).

FIGURA 12 - ALGORITMO PARA A CONSTRUÇÃO DE ÁRVORE DE DECISÃO

1. Algoritmo para construção de uma Árvore de Decisão

Entrada: Um conjunto de treinamento **D** = {(xi,yi), i=1,…,n}
Saída: Árvore de Decisão
1 /* Função GeraÁrvore (D) * /;
2 **se** critério de parada **(D)** = Verdadeiro **então**
3 **Retorna:** um nó folha rotulado com a constante que minimiza a função perda;
4 **fim**
5 Escolha o atributo que maximiza o critério de divisão em **D**;
6 **para cada** partição dos exemplos **D**, baseado nos valores do atributo escolhido **faça**
7 Induz uma subárvore Árvorei = GeraÁrvore **(Di)**;
8 **fim**
9 **Retorna:** Árvore contendo um nó de decisão baseado no atributo escolhido, e descendentes Árvorei;

Fonte: FACELI *et al.* (2011).

Uma árvore pode ser induzida pela sucessiva seleção e subdivisão de atributos. Estes atributos podem ser escolhidos de forma randômica, e eventualmente uma árvore pode ser formada de nodos terminais a partir do momento em que cada nodo possui membros de apenas uma única classe. Assim, a taxa de erro aparente é minimizada e normalmente é zero. A tarefa da indução é desenvolver regras de classificação que podem determinar a classe de um objeto através dos valores de seus atributos. Os objetos são descritos em termos de uma coleção de atributos. Cada atributo mede alguma característica importante do objeto; e cada objeto, no domínio da aplicação, pertence a um conjunto de classes mutuamente exclusivas onde a classe deste objeto é conhecida (QUINLAN, 1986).

Um classificador baseado em regra é uma forma de classificação de informações que utiliza um conjunto de regras “se... então” (TAN, STEINBACH E KUMAR, 2009). Para Castro e Ferrari (2016), as regras de classificação constituem uma alternativa popular às arvores de decisão. Ainda de acordo com os autores, o antecedente de uma regra é uma série de testes similares àqueles feitos nos nós da árvore de decisão e o consequente da regra fornece a classe ou as classes (ou a distribuição de probabilidades sobre as classes) aplicáveis aos objetos cobertos por aquela regra. Os autores afirmam que é fácil ler um conjunto de regras diretamente de uma árvore de decisão: uma regra é gerada para cada nó folha da árvore; o antecedente da regra inclui uma condição para

cada nó do caminho da raiz à folha; e o consequente da regra é a classe especificada pela folha.

Segundo Witten e Frank (2002), uma alternativa às árvores de decisão para a construção de regras de classificação é usar um algoritmo de cobertura (*covering algorithm*), que busca uma maneira de cobrir todos os objetos de cada classe e, ao mesmo tempo, exclui os objetos fora da classe. Assim, os métodos de cobertura resultam num conjunto de regras em vez de uma árvore. Os algoritmos de cobertura adicionam testes à regra em construção com o objetivo de criar regras que tenham máxima acurácia. Os autores destacam um dos algoritmos pioneiros para indução de regras de classificação, denominado PRISM, o qual toma como entrada um conjunto de objetos e fornece como saída um conjunto de regras de classificação.

De acordo com Castro e Ferrari (2016), seja $C_i, i = 1, \ldots, m$, a i-ésima classe existente na base de dados e a_j, o j-ésimo par atributo-valor. O algoritmo pode ser resumido como a seguir. Para cada classe C_i, os seguintes passos: calcular a probabilidade de ocorrência, $p(C_i|a_j)$, ou seja, a probabilidade de que a_j seja classificado como pertencente à classe C_i; selecionar o valor a_j tal que $p(C_i|a_j)$ seja máxima e crie um subconjunto dos dados de treinamento contendo todos os objetos com valor a_j ; repetir os passos anteriores para esse subconjunto até que ele contenha apenas objetos da classe C_i. A regra induzida é a conjunção de todos os pares atributo-valor usada na criação do subconjunto homogêneo; remover da base de treinamento todos os objetos cobertos por essa regra; repetir os passos anteriores até que todos os objetos da classe C_i tenham sido removidos.

O classificador uma-regra (*one-rule* - 1R) é definido por Castro e Ferrari (2016) como uma “forma fácil de encontrar regras de classificação que testam um único atributo da base de dados”. Além de simples, o algoritmo tem baixo custo computacional e muitas vezes são capazes de descobrir boas regras que caracterizam a estrutura dos dados, podendo fornecer altos valores de acurácia. De acordo com os autores, a ideia geral do algoritmo é a seguinte, conforme FIGURA 13:

- São construídas regras que testam um único atributo, ramificando-o, sendo que cada ramo corresponde a diferentes valores do atributo;

- A melhor classificação de cada ramo é aquela que usa a classe que ocorre com mais frequência nos dados de treinamento;
- Assim, a taxa de erro das regras pode ser facilmente determinada por meio da contagem do número de erros que ocorre para os dados de treinamento, ou seja, do número de objetos que não possuem a maioria nas classes.

FIGURA 13 - ALGORITMO DE ÁRVORE DE DECISÃO

```
Entrada
    data: base de dados com n objetos e m atributos (n x m)
    classe: vetor contendo a classe de cada objeto da base (n x 1)
Saída
    R: regras de classificação
Passos
    regras = ø;

    Para i = 1:m Faça
    {
        // Pegue os valores do atributo que está sendo analisado
        aux = ordenar( data[1:n][i] );
        v.Add( aux[1] );
        p = 1;
        Para j=2:aux.size() Faça
        {
            Se (aux[j] <> v[p]) Então
            {
                v.Add( aux[j] );
                p = p + 1;
            }
    }

    Para j=1:v.size() faça
    {
        // Selecione os objetos que possuem o valor v(j) no atributo
        idx = ø;
        Para l=1:n Faça
            Se (data[l][i] == v(j)) Então
                idx.Add(l);

        // Determine a classe mais frequente entre estes objetos
        c = moda(classe[idx]);

        // Quantidade de objetos que possuem o par atributo-valor
        qObj = idx.size();

        // Quantidade de objetos que possuem a regra atr-val-classe
        aux = ø;
        Para l=1:classe[idx].size Faça
            Se (classe[idx[l]] == c) Então
                aux.Add(idx[l]);

        qRegra = aux.size();
        // Adicione na coleção a regra encontrada
        // atributo: i, valor: v[j], classe: c
        // Erro: qObj – qRegra
        regras.Add(i,v[j],c,qObj – qRegra);
    }
}
E = zeros(m,1);
// Para cada atributo somar o erro das regras
Para i=1:regras.size() faça
    E[ regras[i][1] ] = E[ regras[i][1] ] + regras[i][4];

// Retornar conjunto de regras do atributo com o menor erro
erro = E[1];
atr = 1;
Para i=2:m Faça
    Se (erro > E[i]) Então
    {
        atr = i;
        erro = E[i];
    }

R = ø;
Para i=1:regras.size() Faça
    Se (regra[i][1] == atr) Então

        R.Add( regra[i] );
```

FONTE: Adaptado de CASTRO E FERRARI (2016)

O classificador *Naïve Bayes* ou bayesiano consiste de uma abordagem estatística para resolver problemas de classificação de padrões, fundamentados no teorema de Bayes. Essa abordagem é baseada na quantificação das comparações entre as várias decisões utilizando a probabilidade e o custo de tais decisões, admitindo que os problemas de decisão sejam postos em termos probabilísticos e que estes valores são conhecidos (DUDA *et al.*, 2018).

Eles também apresentam alta acurácia e velocidade de processamento quando aplicados a grandes bases de dados. Os classificadores *Naïve Bayes* assumem que o efeito do valor de um atributo em uma dada classe é independente dos valores dos outros atributos. Essa premissa denominada independência condicional da classe, tem como objetivo simplificar os cálculos e, por causa dela, o algoritmo é denominado *Naïve* (HAN; PEI; KAMBER, 2011)

O classificador bayesiano, ou Naïve Bayes, opera da seguinte maneira segundo Castro e Ferrari (2016):

- Cada objeto é representado por um vetor de características m-dimensional $x = (x_1, x_2, \ldots, x_m)$, o qual representa uma medição sobre cada um dos m atributos $A_1, A_2, \ldots, A_m$;
- Assume-se que a base de dados possui c classes, $C_1, C_2, \ldots, C_c$. Dado um objeto x com classe desconhecida, o classificador deve ser usado para predizer a classe à qual esse objeto pertence com base na maior probabilidade *a posteriori* encontrada, dado x – ou seja, o classificador bayesiano especifica uma classe C_i para o objeto x se, e somente se:

$$P(C_i|x) > P(C_j|x), \forall j \neq i \tag{2.1}$$

Portanto, o algoritmo maximiza $P(C_i|x)$. A classe C_i para a qual $P(C_i|x)$ é maximizada é denominada hipótese *a posteriori* máxima. Pelo Teorema de Bayes:

$$P(C_i|x) = P(x|C_i)\,P(C_j)/\,P(x) \tag{2.2}$$

- Como $P(x)$ é constante para todas as classes, somente $P(x|C_i)P(C_i)$ precisa ser maximizado. Se as probabilidades *a priori* não são

conhecidas, assume-se que as classes possuem a mesma probabilidade $P(C_1) = P(C_2) = \ldots = P(C_c)$, e o objetivo torna-se maximizar $P(x|C_i)$. Note que as probabilidades *a priori* devem ser estimadas por $P(C_i) = S_i/S$, onde S_i é o número de objetos de treinamento da classe C_i e S, o número total de objetos;

- Para conjuntos de dados com muitos objetos, o cálculo de $P(x|C_i)$ torna-se computacionalmente caro e, por isso, a premissa da independência condicional de classe é assumida para os atributos, de modo que:

$$P(x|C_i) = \prod_{k=1}^{m} P(x_k|C_i) \tag{2.3}$$

As probabilidades $P(x_1|C_i), P(x_2|C_i), \ldots, P(x_m|C_i)$ podem ser estimadas a partir dos objetos de entrada $P(x_i|C_i)$, onde:

- Se A_k é categórico, então $P(x_k|C_j) = S_{ik}/S_i$, onde S_{ik} é o número de objetos da classe C_i com valor x_k para A_k e S_i, o número de objetos de treinamento pertencentes à classe C_i.
- Se A_k é contínuo, então o atributo assume tipicamente uma distribuição de probabilidade gaussiana, de forma que:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi\sigma_{C_i}}} \exp\left(-\frac{(x_k-\mu_{C_i})^2}{2{\sigma_{C_i}}^2}\right) \tag{2.4}$$

onde $g(x_k, \mu_{C_i}, \sigma_{C_i})$ é a função densidade gaussiana ou normal para o atributo A_k, enquanto μ_{C_i} e σ_{C_i} são a média e o desvio padrão, respectivamente, do atributo A_k para as amostras da classe C_i.

- Para classificar um objeto x de classe desconhecida, $P(x|C_i)P(C_i)$ é avaliado para cada classe C_i. O objeto x é especificado à classe C_i se, e somente se:

$$P(x|C_i)P(C_i) > P(x|C_j)P(C_j), \forall\, j \neq i \tag{2.5}$$

O pseudocódigo do algoritmo *Naïve Bayes* é representado na FIGURA 14.

FIGURA 14 - PSEUDOCÓDIGO DO ALGORITMO NAÏVE BAYES

```
Entrada
   data : base de dados com n objetos e m atributos (n x m)
   classe : vetor contendo a classe de cada objeto da base (n x 1)
   obj : objeto que deve ser classificado (1 x m)
Saída
   Pobj : probabilidade de cada classe para obj
Passos
   Pobj = ø;

   // Calcular as probabilidades para os trios: atribulo-valor-classe
   P = CalcularProbabilidades(data,classe);

   // Calcular as probabilidades de cada classe
   Pc = CalcularProbabilidades(classe);

   // Determinar os k objetos mais próximos
   Para cada c em classe faça
   {
      aux = 1;

      // Produtório das probabilidades de cada par atributo-valor para
      // classe c de acordo com os valores dos atributos de obj
      Para i=1:m faça
         aux = aux * P[i,obj[i],c];

      aux = aux * Pc[c];

      // Probabilidade de obj pertencer à classe c
      Pobj.Add(aux,c);
   }
```

FONTE: CASTRO E FERRARI (2016)

2.5.2.2 Regressão

Como a classificação, a regressão também é uma técnica de mineração de dados. Assim, também possuem duas etapas, treinamento e teste. No treinamento é gerado o modelo de regressão baseado na variável alvo e as variáveis explicativas. E no teste, é avaliada a predição da classe predita (variável alvo). Assim, é calculada, entre outros, a matriz de confusão e acurácia para validação do modelo. Além disso, a regressão determina a relação entre a variável alvo e as variáveis explicativas, a diferença é que ao invés de variáveis categóricas, a variável alvo é contínua, significando que a resposta pode assumir uma gama de valores infinitos (OZDEMIR, 2016).

Uma diferença importante entre a classificação e regressão, encontra-se na avaliação da saída. No caso dos classificadores, essa avaliação é baseada em algumas medidas de acurácia do classificador, ou seja, a quantidade de

objetos classificados corretamente. No caso da regressão, a quantidade costuma ser medida calculando-se uma distância ou um erro entre a saída do estimador e a saída desejada (CASTRO; FERRARI, 2016).

De acordo com os autores, a tarefa de classificação pode ser vista como um caso particular da regressão, no qual a saída é discreta. Assim, praticamente todos os algoritmos de regressão podem ser usados para classificação, mas a recíproca não é verdadeira.

Entre os principais métodos de regressão encontrados na literatura, Castro e Ferrari (2016) destacam: regressão linear; regressão polinomial; rede neural do tipo *Adaline*; rede neural do tipo *Multi-Layer Perceptron* (MLP); e redes neurais do tipo Função de Base Radial (RBF).

A regressão linear é um método de predição numérica que consiste em encontrar uma relação linear, entre preditores e uma variável de resposta, estabelecendo uma relação de causa efeito entre elas (OZDEMIR, 2016). Os modelos de regressão linear são métodos estatísticos capazes de modelar a relação entre uma variável dependente e uma ou mais variáveis independentes (CASTRO; FERRARI, 2016).

Segundo Castro e Ferrari (2016), se a forma do relacionamento funcional entre as variáveis dependentes e independentes é conhecida, mas podem existir parâmetros cujos valores são desconhecidos e podem ser estimados a partir do conjunto de treinamento, então a regressão é dita paramétrica. Se não há conhecimento prévio sobre a forma da função que está sendo estimada, diz-se que ela é não paramétrica.

Dado um conjunto de objetos $\{(x_j, y_j)\} j = 1, \ldots, n$, um modelo de regressão linear assume uma relação linear entre as variáveis de entrada (independentes), representada pelo vetor x_j, e a variável de saída (dependente), representada por y_j. Essa relação é modelada por um erro ε_j, que adiciona ruído à relação linear entre as variáveis. O modelo de regressão pode então ser escrito como:

$$y_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_L x_{jL} + \varepsilon_j = x_j^T . \beta + \varepsilon_j, \; j = 1, \ldots, n \qquad (2.6)$$

A Equação 2.6 corresponde à combinação linear do vetor de coeficientes $\beta \in \Re^1$, com o vetor de entradas $x_j \in \Re^1$ (objeto), mais o ruído ε_j.

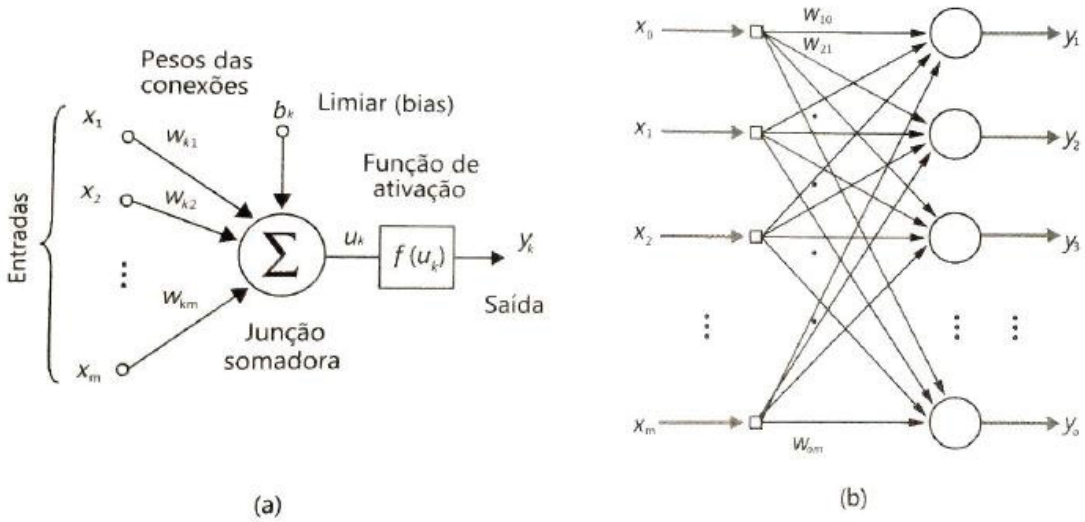
Já a regressão polinomial é um modelo de regressão no qual a relação entre as variáveis independentes e a variável dependente pode ser não linear e tem a forma de um polinômio de grau n (CASTRO; FERRARI, 2016), conforme equação 2.7.

$$y_j = \alpha_0 + \alpha_1 x_j + \alpha_2 x_j^2 + \cdots + \alpha_p x_j^p + \varepsilon_j, j = 1, \ldots, n \tag{2.7}$$

A rede neural do tipo *Adaline*, inicialmente foi chamado de (*ADAptative LINear Element*) e posteriormente de (*ADAptative LInear NEuron*), foi proposta por Windrow e Hoff nos anos de 1960. As redes são bem simples e possuem a arquitetura do neurônio idêntica à do *Perceptron*, tendo como diferencial a regra de aprendizado, neste caso chamado de regra delta ou o algoritmo dos quadrados mínimos (FURTADO, 2019).

O *Perceptron* consiste em uma rede neural com uma única camada de pesos, ou seja, um conjunto de neurônios de entrada e um conjunto de neurônios de saída, com pesos sinápticos e bias ajustáveis. Se os padrões de entrada forem linearmente separáveis, o algoritmo de treinamento do *Perceptron* possui convergência garantida, ou seja, é capaz de encontrar um conjunto de pesos que classifica corretamente os dados. Os pesos dos neurônios que compõem o *Perceptron* serão tais que as superfícies de decisão produzidas pela rede neural estarão apropriadamente posicionadas no espaço. Entretanto, a rede *Adaline* se diferencia por apresentar neurônios usando função de ativação linear em vez de função de sinal, conforme FIGURA 15 (HAYKIN, 2007; CASTRO; FERRARI, 2016).

FIGURA 15 - NEURÔNIO (A) E REDE NEURAL DO TIPO PERCEPTRON E ADALINE (B)



ONTE: CASTRO E FERRARI (2016)

A rede neural *Perceptron* ocupa um papel importante no desenvolvimento das redes neurais, uma vez que, foi a primeira rede a ser descrita via algoritmo (HAYKIN, 2007). A FIGURA 16 mostra uma representação diagramática do *perceptron* proposto por Rosenblatt (1957), onde $(x_1, x_2, \dots, x_m)$ são as entradas, $(w_1, w_2, \dots, w_m)$ são os pesos sinápticos, o bias b, $f(u_k)$ é a função de ativação e u_k é o campo local induzido. A partir da equação 2.8, o campo local induzido do neurônio pode ser encontrado.

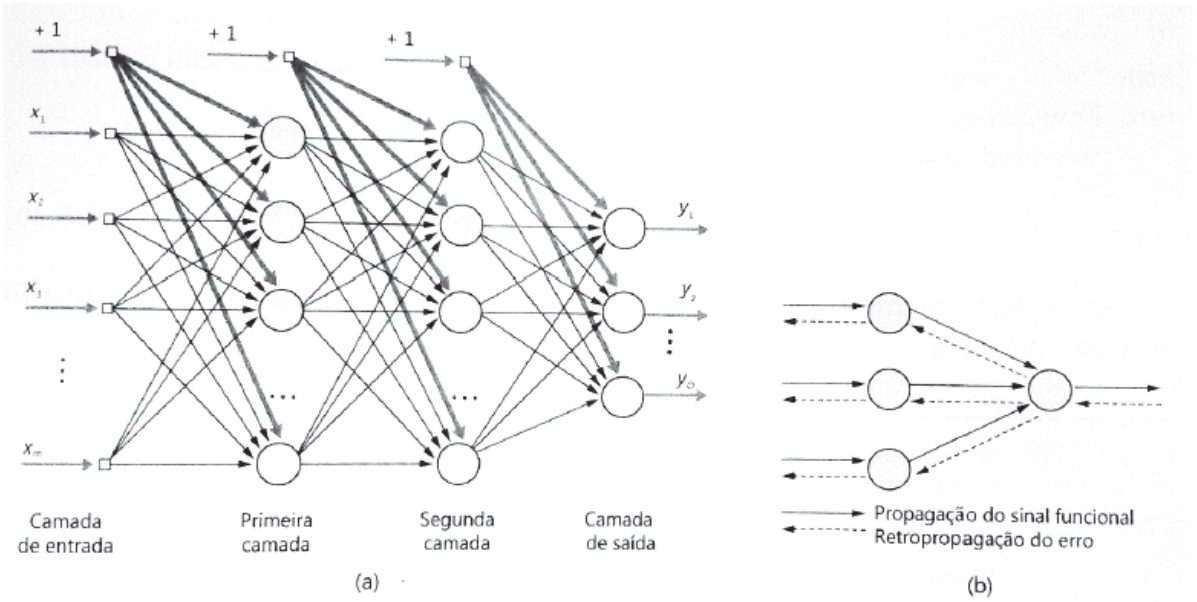
$$u_k = \sum_{i=1}^{M} w_i x_i + b \tag{2.8}$$

A função de ativação representa o efeito que a entrada e o estado atual da ativação exercem no próximo estado do neurônio (HAYKIN, 2007).

A rede neural do tipo *Multi-Layer Perceptron (MLP)* ou *Perceptron* de múltiplas camadas é uma rede do tipo *Perceptron* com pelo menos uma camada intermediária. Trata-se de uma generalização do *Perceptron* simples e da rede *Adaline*. O treinamento da rede MPL foi feito originalmente utilizando-se um algoritmo denominado retropropagação do erro, conhecido como *backpropagation*. Esse algoritmo consiste em: propagação positiva do sinal funcional, durante a qual todos os pesos da rede são mantidos fixos; e retropropagação do erro, durante a qual os pesos da rede são ajustados com base no erro (CASTRO; FERRARI, 2016).

Segundo Haykin (2007), a rede neural do tipo *Multi-Layer Perceptron (MLP)* consiste em uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída, conforme FIGURA 16. Os neurônios que pertencem a camada de entrada são denominados unidades de entrada. Esses neurônios propagam os valores das entradas para as camadas seguintes sem modificação. As camadas intermediárias, escondidas ou ocultas transmitem as informações por meio das conexões entre as camadas de entrada e saída. Finalmente, a camada de saída transmite a resposta da rede neural à entrada aplicada na camada de entrada.

FIGURA 16 - REDE NEURAL DE MÚLTIPLAS CAMADAS (A), SENTIDO DE PROPAGAÇÃO DO SINAL DE ENTRADA E RETROPROPAGAÇÃO DO ERRO (B).



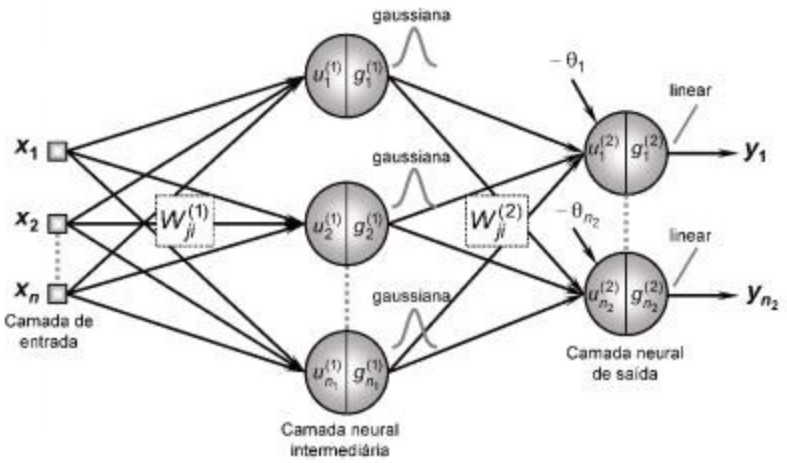
FONTE: CASTRO E FERRARI (2016)

Nas redes neurais do tipo Função de Base Radial (RBF, do inglês *Radial Basis Function*), o projeto das redes com múltiplas camadas a propagação positiva do sinal é visto como um problema de aproximação de função em um espaço multidimensional. A camada de entrada é responsável por conectar a rede a seu ambiente, a camada intermediária aplica uma transformação não linear do espaço de entrada para o espaço intermediário e a camada de saída é linear (CASTRO; FERRARI, 2016). Os dados são propagados da camada de

entrada, seguindo para a camada oculta, até chegarem à camada de saída, sem retroalimentação da rede (FAUSETT, 2006).

A construção tradicional de uma RBF apresenta 3 camadas totalmente distintas, que podem ser observadas na FIGURA 17. A camada de entrada é constituída por unidades sensoriais que interligam a rede com o ambiente externo. A camada oculta faz transformações não lineares no espaço de entrada, sendo caracterizada pela utilização de funções de base radial. E a camada de saída é responsável pela resposta final da rede modelada. Os pesos sinápticos da camada intermediária, também chamados de centros, representam os centros dos subgrupos de dados categorizados e são utilizados como centros das funções de bases radiais no processo de treinamento (HAYKIN, 2007).

FIGURA 17 - ARQUITETURA DE UMA RBF



FONTE: SILVA *et al.* (2010)

2.5.2.3 Sumarização

A sumarização é a abstração ou generalização dos dados. Nas técnicas de sumarização, busca-se extrair informações condensadas a partir da massa de dados, resultando em um conjunto menor que fornece uma visão mais geral dos dados; tal visão pode potencialmente consistir em informação útil e relevante para tomada de decisão (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

2.5.2.4 Regras de Associação

As técnicas de associação têm como objetivo encontrar relações ou dependências entre atributos de um objeto a partir da análise de conjuntos de dados; essas dependências são geralmente expressas por meio de regras, que indicam relações de causa e efeito entre atributos de um objeto (CIOS *et al.*, 2007).

Segundo Vasconcelos e Carvalho (2018), as regras de associação têm como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação, ou seja, encontrar relacionamentos ou padrões frequentes entre conjuntos de dados. Ou ainda, uma regra de associação é uma expressão da forma A → B, onde A e B são *itemsets*. Imaginando que A e B são conjuntos de produtos, a ideia por trás destra regra é: pessoas que consomem o produto A têm a tendência de também consumir o produto B.

De acordo com Castro e Ferrari (2016), a mineração de regras de associação é uma técnica usada na construção de relações sob a forma de regras entre itens de uma base de dados transacional. Os autores explicam que diferentemente do agrupamento, que busca relações de similaridade entre objetos, as regras de associação buscam relações entre os atributos dos objetos, ou seja, os itens que compõem a base. Assim, os autores afirmam que o objetivo é encontrar regras fortes de acordo com alguma medida do grau de interesse da regra.

As regras de associação não são diferentes das regras de classificação, exceto pelo fato de que elas podem ser usadas para predizer qualquer atributo, não apenas a classe e não são planejadas para serem usadas em conjunto. Além disso, essa característica que dá a liberdade de predizer combinações de atributos também. Diferentes regras de associação expressam diferentes regularidades da base de dados e geralmente são usadas para estimar relações distintas entre itens (HARRINGTON, 2012; CASTRO E FERRARI, 2016).

Os autores complementam que como uma grande quantidade de regras de associação pode ser derivada a partir de uma base de dados, mesmo que pequena, normalmente se objetiva a derivação de regra que suporta um grande número de transações e que possui uma confiança razoável para as transações

as quais elas são aplicáveis. Os autores afirmam que esses requisitos estão associados a dois conceitos centrais em mineração de regras de associação: suporte e confiança. Melanda (2004) define tais aspectos como:

- Suporte: Quantifica a incidência de um *itemset* em um conjunto de dados, ou seja, indica a frequência com que um *itemset* ocorre em relação ao total de transações de uma base de dados;
- Confiança: Indica a frequência com que determinados *itemsets* ocorrem juntos em relação ao número total de transações em que o primeiro *itemset* especificado ocorre.

Os conceitos de suporte e confiança permitem definir o problema geral de mineração de regras de associação: “Dado um conjunto de transações, o problema de minerar regras de associação corresponde a encontrar todas as regras que satisfaçam um valor mínimo predefinido de suporte (*minsup*) e um valor predefinido de confiança (*minconf*)” (CASTRO; FERRARI, 2016).

Entre os principais algoritmos de descoberta de regras de associação encontrados na literatura, Castro e Ferrari (2016) destacam: o algoritmo *Apriori* e o algoritmo *FP-Growth*.

Apriori é um algoritmo que extrai de regras de alta confiança, assim, encontrando relações entre os dados. Esse algoritmo utiliza uma abordagem iterativa que consiste dos k−*itemsets* serem usados para explorar os (k+1)−*itemsets* (HAN; PEI; KAMBER, 2011).

O algoritmo *Apriori* é o método mais conhecido para a mineração de regras de associação e emprega busca em profundidade e geram conjuntos de itens candidatos de k elementos a partir de conjuntos de itens com k – 1 elementos. Os itens candidatos não frequentes são eliminados, e toda a base de dados é rastreada e os conjuntos de itens frequentes são obtidos a partir dos conjuntos de itens candidatos. A estratégia adotada pelo algoritmo consiste em decompor o problema em duas sub tarefas: geração do conjunto de itens frequentes e geração das regras. A geração do conjunto de itens frequentes consiste em encontrar todos os conjuntos cujo suporte seja maior que o *minsup* especificado. Já a geração de regras corresponde ao uso dos conjuntos de itens frequentes para gerar as regras desejadas. A ideia geral é que se, por exemplo, ABCD e AB são frequentes, então é possível determinar se a regra AB → CD é válida calculando a razão confiança = suporte (ABCD / suporte AB). Se a

confiança for maior ou igual a *minconf*, então a regra é válida (CASTRO; FERRARI, 2016). Na FIGURA 18 é mostrado o algoritmo *Apriori.*

FIGURA 18 - ALGORITMO *APRIORI*

Algoritmo 1: Algoritmo Apriori

```
F1 ←{Conjuntos de itens freqüentes de tamanho 1} /* Na
     primeira passagem k = 1                          */
1  para k = 2; Fk−1 != vazio; k++ faça
   |  /* Na segunda passagem k = 2                     */
2  |  Ck ← apriori-gen (Fk−1) /* Novos candidatos     */
3  |  para todo transação t ∈ T faça
4  |  |  Ct ← subconjunto (Ck, t) /* Candidatos contidos
   |  |       em t                                    */
5  |  |  para todo candidato c ∈ Ct faça
6  |  |  |  c.contagem++
7  |  |  fim
8  |  |  Fk ← {c ∈ Ck|c.contagem ≥ MinSup}
9  |  fim
10 fim
11 Resposta F ← Reunião de todos os Fk
```

FONTE: VASCONCELOS E CARVALHO (2018)

De acordo com Castro e Ferrari (2016) e Nandi, Pereira e Felippe (2015), o algoritmo FP-*Growth* (*Frequente Pattern Growth),* é baseado em uma estrutura em árvore de prefixos para os padrões frequentes, denominada FP-*Tree* (*Frequent Pattern Tree),* a qual armazena de forma comprimida a informação sobre os padrões frequentes.

A essência do algoritmo está baseada em três aspectos primeiro, a compressão da base de dados em uma estrutura em árvore (*FP-Tree*) cujos nós possuem apenas itens frequentes de comprimento unitário (F_1) e organizada de modo que aqueles nós que ocorrem mais frequentemente tenham maiores chances de compartilhar nós do que os de baixa frequência. Segundo os autores, o uso de um algoritmo de mineração da árvore que evita a geração de uma grande quantidade de conjuntos candidatos. Esse algoritmo inicia com um padrão frequente de comprimento 1 (padrão sufixo inicial), avalia apenas o conjunto de itens frequentes que co-ocorrem com o padrão sufixo, constrói sua

FP-Tree e executa uma mineração recursiva na árvore. Por último, o uso de um método particional para decompor a tarefa de mineração em subtarefas menores, reduzindo significativamente o espaço de busca (CASTRO; FERRARI, 2016).

Segundo os autores, uma árvore de itens frequentes, denominada *FP-Tree*, é uma estrutura em árvore que possui um nó raiz, rotulado como *null*, um conjunto de subárvores de itens prefixos como filhos da raiz e uma tabela de cabeçalho dos itens frequentes. Os autores ainda destacam que cada nó na subárvore de itens possui três campos: *nome_do_item*, *contagem* (número de transações representadas pela porção do caminho que chega àquele nó) e *ligação_do_nó* (conectando-o ao próximo nó da árvore que possua o mesmo nome ou à raiz, caso não haja outro nó). Por fim, cada valor na tabela de cabeçalho de itens frequentes possui dois campos: *nome_do_item* e *cabeçalho da ligação_do_ nó*, o qual aponta para o primeiro nó na *FP-Tree* que possui o mesmo *nome_do_item*.

Com base nessas definições, o processo de construção da *FP-Tree* pode ser definido da seguinte forma, conforme Castro e Ferrari (2016):

- Determinação da lista de itens frequentes: dado o parâmetro de entrada relativo ao suporte mínimo, *minsup*, é necessário fazer a leitura da base de dados e determinar o conjunto de itens frequentes de tamanho 1, F_1, e seus respectivos suportes. Ordene F_1 em ordem decrescente e denomine de L a lista de itens frequentes;
- Construção da árvore: é necessário criar um nó raiz para a *FP-Tree*, chamada simplesmente de *Tree*, e rotular como *null*. Para cada transação t_i, recomenda-se selecionar e ordenar o conjunto de itens frequentes da transação t_i de acordo com a ordem de L. Seja [p|P] a lista de itens frequentes ordenados, onde p é o primeiro elemento e P, o restante da lista. A função é chamada de *InsertTree* ([p|P], *Tree*).
- Função *InsertTree*([p|P],*Tree*): se a árvore possui um descendente D, tal que *D.nome_do_item* = *p.nome_do_item*, então a contagem de *D* é em 1; senão, é necessário criar um nó *D* e atribuir 1 à sua contagem, de modo que seu nó pai seja ligado à *Tree* e a todos os nós com o mesmo *nome_do_item*. Deve-se chamar a função *InsertTree*(P,D) recursivamente enquanto P for não vazio.

O algoritmo permite que a *FP-Tree* seja construída com apenas duas leituras da base de dados. A primeira determina e ordena o conjunto de itens frequentes, ao passo que a segunda constrói a árvore (CASTRO; FERRARI, 2016).

2.5.2.5 Regras de Agrupamento

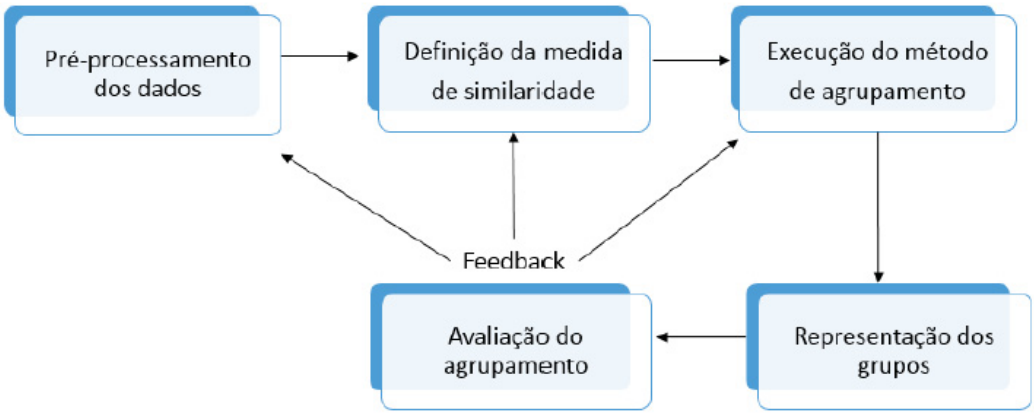
Técnicas de agrupamento buscam identificar grupos de objetos a partir dos seus atributos. Nesse contexto, os grupos também podem ser conhecidos como *clusters*. Esses grupos são formados de modo a maximizar a similaridade entre objetos dentro de um mesmo grupo e minimizá-la entre objetos de grupos diferentes (LAROSE, 2005).

Nisbet, Elder e Miner (2009) definem a Clusterização/Agrupamento, como a detecção de subgrupos semelhantes entre uma grande variedade de casos e atribui essas observações aos seus subgrupos ou *clusters*. “Tão importante quanto identificar tais grupos é a necessidade de determinar como esses grupos são diferentes” (NISBET, ELDER & MINER, 2009).

A análise de grupos é definida por Castro e Ferrari (2016) como a organização de um conjunto de objetos (normalmente representados por vetores de características, ou seja, pontos em um espaço multidimensional) em grupos baseada na similaridade entre eles. Ainda de acordo com os autores, dito de outra forma, agrupar objetos é o processo de particionar um conjunto de dados em subconjuntos (grupos) de forma que os objetos em cada grupo (idealmente) compartilhem características comuns, em geral proximidade em relação a alguma medida de similaridade ou distância.

Castro e Ferrari (2016) dividem o processo de agrupamento de dados em cinco etapas, conforme FIGURA 19.

FIGURA 19 - PROCESSO DE AGRUPAMENTO DE DADOS



FONTE: Adaptado de CASTRO E FERRARI (2016)

De acordo com Castro e Ferrari (2016), a etapa de pré-processamento de dados, que consiste na preparação da base para o agrupamento, pode envolver todas as etapas típicas de pré-processamento de dados, como limpeza, integração, redução, transformação e discretização. A etapa de definição da medida de similaridades (proximidade) ou dissimilaridade (distância) entre objetos é utilizada, segundo os autores, durante o agrupamento propriamente dito. Na etapa de execução os autores afirmam que os métodos de agrupamento podem ser divididos em hierárquicos ou particionais. Os métodos hierárquicos criam uma decomposição hierárquica dos dados, enquanto os métodos particionais, dado um conjunto n objetos, um método particional constrói k partições de dados, sendo que cada partição representa um cluster $(k \leq n)$.

A representação dos grupos é definida como o processo de extrair uma representação simples e compacta dos grupos obtidos a partir do agrupamento da base. Os autores destacam que as formas típicas de representação dos grupos são: protótipos, estruturas em grafos, estruturas em árvore e rotulação. Por fim, a etapa de avaliação do agrupamento depende, de acordo com os autores, do contexto e dos objetivos da análise. Os autores afirmam que a saída do algoritmo de agrupamento pode ser avaliada com relação à qualidade do agrupamento, o que pode ser feito por uma medida de avalição externa – isto é, os grupos encontrados são comparados com uma estrutura de agrupamento conhecida *a priori* ou uma avaliação interna, ou seja, tenta-se determinar se a

estrutura encontrada pelo algoritmo é apropriada aos dados (CASTRO; FERRARI, 2016).

Entre os principais algoritmos de agrupamento presentes na literatura, Castro e Ferrari (2016) destacam: algoritmo k-médias (*k-means*); algoritmo k-medoides (*k-medoids*); algoritmo *fuzzy* k-médias; árvore geradora mínima; DBScan; *single-linkage*; e *complete-linkage*.

O método k-médias é um dos algoritmos não hierárquicos mais conhecidos. É um método iterativo que tem como objetivo encontrar a melhor divisão de n dados em k agrupamentos, de forma que a distância total entre o centro de um grupo e seus respectivos dados, somada para todos os grupos, seja minimizada (NALDI, 2010).

Segundo Naldi (2010), considerando o conjunto de objetos $X = \{p_1, p_2, \dots, p_n\}$, sendo n o número de objetos a ser agrupado em k agrupamentos, o algoritmo inicia com a determinação dos k centros iniciais e então forma os k grupos seguindo uma medida de similaridade e associando cada objeto ao centro mais próximo, neste caso a medida de similaridade costuma ser a distância Euclidiana, formando os clusters X_i. Após o primeiro agrupamento, os centros são recalculados para que as distâncias entre ele e os objetos do grupo sejam as menores possíveis, em seguida a divisão é refeita com base nos novos centros e os objetos são rearranjados. Este processo deve ser feito até que os objetos se estabilizem e não mudem mais de agrupamentos, ou até que seja atingido um número pré-definido de iterações.

Para definir o algoritmo k-medoides, Castro e Ferrari (2016) definem antes um “medoide” como o objeto com a menor dissimilaridade média a todos os outros objetos, ou seja, é o objeto mais centralmente localizado no grupo. Assim, os autores destacam que o algoritmo k-medoides é um método de agrupamento relacionado ao k-médias, mas que usa um objeto da base como protótipo em lugar de um centroide. Os autores afirmam que ambos são algoritmos particionais que visam minimizar o erro quadrático entre os objetos de um grupo e seu protótipo (no caso do k-médias, o centroide do grupo, e do k-medoides, o medoide do grupo). Para os autores, uma diferença importante entre esses métodos é que o k-medoides escolhe objetos da própria base como os centros do grupo, ao passo que o k-médias calcula o centro do grupo a partir dos objetos neles contidos. Além disso, o algoritmo k-medoides é mais robusto a

ruído e a valores discrepantes do que o k-médias, pois o centro do grupo será necessariamente um objeto da base e, portanto, ruídos e valores discrepantes podem não influenciar tão fortemente a definição do centro.

O método *fuzzy* k-médias é definido como uma extensão do algoritmo k-médias na qual cada objeto possui um grau de pertinência em relação aos grupos da base. Neste algoritmo, um objeto pode pertencer a mais de um grupo, porém, com variados graus de pertinência (CASTRO; FERRARI, 2016).

Segundo Castro e Ferrari (2016), para cada objeto x da base há um valor $u_k(x)$ correspondente a seu grau de pertinência ao grupo k. Por convenção, a soma dos graus de pertinência de um objeto a todos os grupos da base deve ser 1, que é o valor máximo possível de pertinência de um objeto a um grupo, conforme a função 2.9.

$$\sum_{i=1}^{k} u_i(x) = 1, \qquad i = 1, 2, \ldots, k \tag{2.9}$$

O centroide c_i de cada grupo i é a média de todos os objetos do grupo, ponderada pelos seus respectivos graus de pertinência ao grupo, conforme a função 2.10.

$$\mathrm{c_i} = \frac{\sum_{\mathrm{x}} \mathrm{u_i(x)^m x}}{\sum_{\mathrm{x}} \mathrm{u_i(x)^m}} \tag{2.10}$$

onde m é conhecido como parâmetro de fuzzificação.

Os valores de pertinência são então normalizados e *fuzzificados* pelo parâmetro real $m > 1$, tal que sua soma seja 1, conforme função 2.11:

$$u_i(x) = \frac{1}{\sum_j \left(\frac{d(c_i, x)}{d(c_j, x)} \right)^{\frac{2}{(m-1)}}} \tag{2.11}$$

onde $d(.,.)$ é uma das medidas de distâncias discutidas anteriormente; $1 < m \leq \infty$, é um coeficiente que pondera quanto o grau de pertinência influencia a medida de distância definida. Os valores de m geralmente utilizados estão no intervalo [1,30] e pode-se defini-lo apenas experimentalmente. Para a maioria dos casos, $1 < m \leq 3$(CASTRO; FERRARI, 2016).

O funcionamento do algoritmo *fuzzy* k-médias é muito similar ao k-médias e pode ser descrito como no Algoritmo representado na FIGURA 20.

FIGURA 20 - PSEUDOCÓDIGO DO ALGORITMO *FUZZY* K-MÉDIAS

```
Entrada
        k : número de grupos
        data : base de dados com n objetos e m atributos (n x m)
        it_max : número máximo de iterações
        cfm : coeficiente de fuzzyficação m
Saída
        U : matriz de pertinência dos objetos aos grupos (n x k)
        C : matriz dos centroides (k x m)
Passos
        // Gerar aleatoriamente a matriz de pertinência inicial
        U = rand(n, k);

        // Normalizar a matriz para que o somatório de cada linha seja 1
        Para i = 1 : n Faça
        {
           soma = 0;
           Para j = 1 : k Faça
              soma = soma + U[i][j];

           Para j = 1 : k Faça
              U[i][j] = U[i][j] / soma;
        }

        // Inicializar variáveis de controle
        it = 0; // número de iterações
        J = 0;
        J_nova = 1;

        // Atualização da matriz de pertinência
        Enquanto (it < it_max) ou (J <> J_nova) Faça
        {
            // Definir a posição dos centroides utilizando Equação 4.25
            C = Calcular_Centroides(data,U,cfm);

           // Calcular a nova matriz de pertinência utilizando Equação 4.26
           U = Calcular_Pertinencia(data,C,cfm);

           // Calcular valor da função objetivo utilizando Equação 4.27
           J = J_nova;
           J_nova = Funcao_Objetivo(data,C,U,cfm);

           // Atualizar variáveis de controle
           it = it + 1;
        }
```

FONTE: Adaptado de CASTRO E FERRARI (2016)

A árvore geradora mínima (*Minimal Spanning Tree* – MST), de acordo com Sedgewick (2003), baseia-se em teoria dos grafos, onde deve atender aos seguintes requisitos: uma árvore é uma árvore geradora se ela é um subgrafo

que contém todos os nós do grafo; uma árvore geradora mínima de um grafo é uma árvore com peso mínimo, onde o peso de uma árvore é definido como a soma dos pesos de suas arestas; e um caminho minimax entre um parte de nós é aquele que minimiza o custo (peso máximo do caminho) sobre todos os caminhos. De acordo com os autores, a MST sempre percorre os caminhos minimax, forçando a conexão entre dois nós próximos antes de sair em busca de outro nó.

Segundo Castro e Ferrari (2016), vários algoritmos podem ser utilizados para gerar a MST para determinado conjunto de dados. O algoritmo representado na FIGURA 21 determina o subconjunto de arestas da árvore geradora mínima e é conhecido como algoritmo de *Prim*. Os autores destacam que este algoritmo realiza uma busca gulosa sobre um grafo conectado, ponderado e não direcionado. O algoritmo de *Prim* aumenta continuamente o tamanho da árvore, partindo de um único nó até que a árvore cubra todos os nós do grafo.

FIGURA 21 - PSEUDOCÓDIGO DO ALGORITMO DE PRIM, USADO PARA GERAR A MST.

```
Entrada
        D : matriz de distância (n x n)
Saída
        mst : subgrafo da árvore geradora mínima (n − 1 x 3)
Passos
        mst = ø;
        // Conjunto dos nós já conectados à mst
        con = 1;
        // Conjunto dos nós não conectados à mst
        dis = [2:n];

        // Construção da MST
        Para i = 1 : n − 1 Faça
        {
                // Procurar o menor valor de conexão entre os nós dos conjuntos
                // conectados e desconectados, retornando posições da matriz D
                aux = 0;
                Para i = 1 : con.size() Faça
                Para j = 1 : dis.size() Faça
                        Se (aux > D[con[i]][dis[j]]) Então
                        {
                                idx_con = con[i];
                                idx_dis = dis[j];
                                aux = D[con[i]][dis[j]];
                        }

                // Guardar a aresta da MST (com seu comprimento)
                Mst[i][1] = con[idx_con];
                Mst[i][2] = dis[idx_dis];
                Mst[i][3] = D[con[idx_con]][dis[idx_dis]];

                // Transferir o nó do conjunto desconectado para o conectado
                con.Add( dis[idx_dis] );

                // Remover o nó desconectado do conjunto desconectado
                dis.Rem( idx_dis );
        }
```

FONTE: Adaptado de CASTRO E FERRARI (2016)

O algoritmo DBScan (*Density Based Spatial Clustering of Applications with Noise*) é baseado em densidade. Diferentemente dos algoritmos *K-Means*, o DBScan não recebe *a priori* a quantidade de *clusters* que deve ser encontrada. Ao invés disso, o algoritmo considera o número mínimo de pontos que um *cluster* pode ter (parâmetro *MinPts*) e o valor da distância máxima do raio entre um ponto (objeto) e outro, para serem considerados pontos do mesmo *cluster*. O DBScan trabalha muito bem com *outliers*, de modo que são identificados pelo algoritmo e informados no resultado (ESTER *et al.*, 1996).

O *single-linkage* é definido por Castro e Ferrari (2016) como um método aglomerativo hierárquico no qual novos grupos são criados unindo os grupos mais semelhantes. Os autores afirmam que o agrupamento inicial é formado apenas por *singletons (*grupos formados por apenas um objeto), e a cada interação do método um novo grupo é formado por meio da união dos dois grupos mais similares da iteração anterior. Os autores destacam que nesse método a distância (proximidade) entre o novo grupo e os demais é determinada como a menor distância entre os elementos do novo grupo e os grupos remanescentes. Matematicamente, a função de ligação, à distância $D(g_1, g_2)$ entre os grupos g_1 e g_2, é descrita pela expressão 2.12.

$$D(g_1, g_2) = \min\left(d(x, y)\right) \tag{2.12}$$

onde $d(x, y)$ é a distância entre os elementos x e y, e g_1 e g_2 são dois grupos.

Por fim, o algoritmo *complete-linkage*, opera de maneira similar ao *single-linkage*, mas a distância do novo grupo aos demais é calculada como a distância máxima entre os elementos do novo grupo aos grupos restantes. Matematicamente, a função de ligação, a distância $D(g_1, g_2)$ entre os grupos g_1 e g_2, é descrita pela expressão 2.13.

$$D(g_1, g_2) = max\left(d(x, y)\right) \tag{2.13}$$

onde $d(x, y)$ é a distância entre os elementos x e y, e g_1 e g_2 são dois grupos.

2.5.2.6 Detecção de anomalias

Segundo Castro e Ferrari (2016), a detecção de anomalias, também conhecida como detecção de *outliers*, é definida como um valor discrepante, ou seja, um valor que se localiza significativamente distante dos valores considerados normais. Os autores destacam que uma anomalia não é necessariamente um erro ou um ruído, podendo caracterizar um valor ou uma classe bem definida, porém de baixa ocorrência, às vezes indesejada, ou que reside fora de agrupamentos ou classes típicas.

O objetivo da detecção de anomalias é a busca de padrões de dados não conformes com o comportamento esperado. Essas não conformidades podem ser informações discordantes, exceções ou peculiaridades (CHANDOLA; BANERJEE; KUMAR, 2009).

Os tipos de anomalias, de acordo com Chandola, Banerjee e Kumar (2009), podem ser classificados como:

- Anomalias ponto: se um dado individual de uma instância pode ser considerado como anormal, então a instância é denominada como um ponto anormal.
- Anomalias contextuais: se uma instância é anormal num contexto específico, de atributos contextuais ou comportamentais. Os atributos contextuais são usados para determinar o contexto para a instância. Já os atributos comportamentais definem as características não contextualizadas para a instância.
- Anomalias coletivas: se uma parte dos dados da instância são anormais, todo o resto é anormal. Os dados individuais da instância anormal podem não ser anomalias, mas juntos se tornam anormais.

As técnicas de detecção de anomalias podem operar nos seguintes modos (CHANDOLA; BANERJEE; KUMAR, 2009):

- Detecção de anomalias supervisionada: o treinamento supervisionado, utilizando uma base de dados com os dados rotulados para as instâncias normais e anormais.
- Detecção de anomalias semi-supervisionadas: assume os dados de treinamento, rotulando as instâncias somente para a classe normal.
- Detecção de anomalias não supervisionadas: não requer dados de treinamento, e são amplamente aplicáveis, pois assumem que as instâncias normais sejam mais frequentes que as anomalias nos dados de testes.

Entre os principais métodos de detecção de anomalias presentes na literatura, Castro e Ferrari (2016) destacam: métodos estatísticos paramétricos e não paramétricos; e métodos algorítmicos baseados em proximidades, redes neurais artificiais e em aprendizagem de máquina.

Segundo Castro e Ferrari (2016), os métodos estatísticos para detecção de anomalias normalmente geram um modelo probabilístico de dados e testam

se determinado objeto foi gerado por tal modelo ou não. Assim, os autores afirmam que essas técnicas são essencialmente baseadas em modelo, ou seja, assumem ou estimam um modelo estatístico que captura a distribuição dos objetos da base e avalia os objetos em relação a quão bem eles se ajustam ao modelo. Os autores destacam que se a probabilidade de certo objeto ter sido gerado por esse modelo for muito baixa, então ele é rotulado como uma anomalia.

Os métodos paramétricos assumem que os dados são gerados por uma distribuição conhecida e, na maioria das vezes, ajustam um modelo específico aos dados; portanto, a fase de treinamento envolve estimar os parâmetros do modelo para uma base de dados. Já os métodos não paramétricos são definidos como aqueles que não assumem uma distribuição predefinida dos dados nem um modelo específico que deverá ser ajustado aos dados (CASTRO; FERRARI, 2016).

Os métodos baseados em proximidades são, segundo os autores, normalmente simples de implementar e não assumem nenhuma premissa sobre a distribuição dos objetos da base, podendo ser aplicados tanto de forma não supervisionada quanto supervisionada, e o princípio básico da operação desses métodos é o cálculo de alguma medida de similaridade ou distância entre pares de objetos da base.

2.5.2.6.1 Redes Neurais

Redes Neurais Artificiais são modelos computacionais que buscam representar uma estrutura neural de organismos inteligentes. Uma Rede Neural possui células (neurônios) de processamento distribuídas trabalhando em paralelo, conectadas através de ligações diretas, cuja principal função é distribuir padrões de ativação, de maneira similar ao cérebro humano (ZHU, 2017).

Redes Neurais são organizadas em camadas, a camada de entrada contém nós responsáveis por receber os dados de entrada, após isso, a informação é transmitida para as camadas escondidas que irão fazer a maior parte do processamento, e por fim, o resultado é apresentado na camada de saída (LIU *et al.*, 2017).

2.5.2.6.2 Aprendizagem de máquina e Máquinas de Vetores de Suporte

A aprendizagem de máquina (*Machine Learning*) é definida por Castro e Ferrari (2016) como uma área de pesquisa que visa desenvolver programas computacionais capazes de automaticamente melhorar seu desempenho pela experiência.

Aprendizagem de máquina é um subconjunto da inteligência artificial, onde o sistema é configurado para aprender e pensar como um ser humano. A informação inicial é dada ao sistema, onde o algoritmo pode aprender os dados e sua classificação, o objetivo final do sistema é fazer as suas próprias decisões no futuro (semelhante a um ser humano). Trabalha com certo grau de probabilidade, com base nos dados que são analisados e as decisões que o sistema irá adotar, seu núcleo está treinado para fazer previsões e sua capacidade é de prever eventos futuros com base em eventos passados (NGUYEN; ARMITAGE, 2008).

As Máquinas de Vetores de Suporte (SVMs) constituem uma técnica de aprendizado de máquina e foram introduzidas para resolver problemas de reconhecimento de padrões sendo estendidas também para problemas de regressão e aprendizagem de máquinas (BURGES, 1998).

As Máquinas de Vetores de Suporte são baseadas num tipo de aprendizado chamado de aprendizado supervisionado. Este aprendizado consiste de três componentes (HAYKIN, 2007):

- Ambiente: conjunto de vetores de entrada x.
- Supervisor: O supervisor fornece para a máquina as entradas, juntamente com as saídas associadas a cada uma delas, ou seja, fornece a resposta d para cada vetor de entrada x recebido de acordo com uma função $f(x)$ desconhecida. Dessa forma, ações podem ser tomadas a fim de valorizar os acertos e punir os erros obtidos pela máquina, possibilitando que o processo de aprendizagem se efetue com sucesso.
- Máquina ou algoritmo de aprendizagem: capaz de implementar funções de mapeamento de entrada-saída da forma $y = f(x, r)$, onde y é a resposta produzida pela máquina e r é um conjunto de parâmetros usados como pesos aos valores do vetor x.

Os dados do conjunto de treinamento devem ser estatisticamente representativos para que a máquina possa reconhecer possíveis padrões posteriores não apresentados inicialmente, propriedade conhecida como generalização (VAPNIK, 1999).

Além desta grande amostra de dados, é necessário que as funções $d = F(x,r)$ tenham comportamento determinístico, ou seja, para certo conjunto de entrada X, e um conjunto de parâmetros r, a saída deve ser sempre a mesma. O objetivo da máquina de aprendizado é escolher uma função $f(x,r)$ que seja capaz de mapear a relação de x e y, onde r são os parâmetros desta relação. As funções usadas para aprender este mapeamento são conhecidas como funções indicadoras em problemas de classificação e de funções de aproximação em problemas de regressão (VAPNIK, 1999).

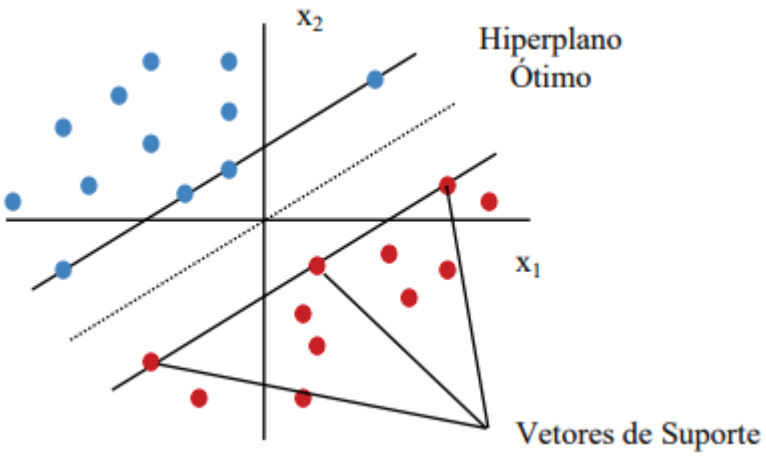
Para escolher a melhor função que se ajuste ao conjunto de treinamento é necessária uma medida de discrepância $L(y,F(x,r))$, que indica a diferença da saída desejada d e da saída obtida y. Para problemas de classificação binária, com somente duas classes, são usadas funções de discrepância como:

$$L\big(y,F(x,r)\big) = \begin{Bmatrix} 0, se\ y = f(x,r) \\ 1, se\ y \neq f(x,r) \end{Bmatrix} \tag{2.14}$$

Seu processo de aprendizado pode ser usado na resolução de problemas de classificação e de regressão. Num contexto de classificação binária, por exemplo, a ideia principal da SVM é construir um hiperplano como superfície de separação ótima entre exemplos positivos e exemplos negativos (HAYKIN, 2007).

O hiperplano de separação encontrado por uma SVM mostrado na FIGURA 22 é ótimo, pois só existe um ponto que minimiza a função de custo quadrática existente no problema de otimização característico da SVM.

FIGURA 22 - EXEMPLO DE UMA SUPERFÍCIE ÓTIMA DE SEPARAÇÃO ENTRE DUAS CLASSES.



FONTE: Adaptado de HAYKIN (2007)

Sendo um conjunto de treinamento $\{(x1, y1), (x2, y2), \ldots, (xn, yn)\}$ onde y tem valor 1 ou -1, indicando a que classe x pertence, e x é um vetor de p dimensões normalizado (com valores 0, 1 ou -1) a SVM divide o hiperplano de acordo com a seguinte equação:

$$W^T x + b = 0 \tag{2.15}$$

em que o vetor w é um vetor perpendicular ao hiperplano, x é o vetor de entrada e b é uma variável que permite que a margem do hiperplano seja maximizada, pois sem esta variável o hiperplano obrigatoriamente passaria pela origem (LORENA; DE CARVALHO, 2007; ALMEIDA, 2010).

Considerando a maior margem de separação, é necessário dar atenção aos hiperplanos paralelos ao hiperplano ótimo mais próximo aos vetores de suporte de cada classe. Estes hiperplanos podem ser descritos pelas equações:

$$W^T x + b = 1 \tag{2.16}$$

$$W^T x - b = -1 \tag{2.17}$$

Segundo Lorena & De Carvalho (2007) e Almeida (2010), sendo o conjunto de dados de aprendizado linearmente separável, podemos selecionar estes hiperplanos maximizando a distância entre eles de modo que não haja pontos no intervalo destes pontos. A geometria mostra que a distância entre

estes dois hiperplanos é $2/|w|$, portando se quer minimizar o valor de $|w|$, garantindo que para cada vetor de entrada x_i:

$$\begin{cases} W^T x_i + b \geq 1 \; se \; y_i = 1 \\ W^T x_i + b \leq 1 \; se \; y_i = -1 \end{cases} \quad (2.18)$$

Podendo ser reescrito como:

$$y_i[W^T x_i + b] \geq 1 \quad (2.19)$$

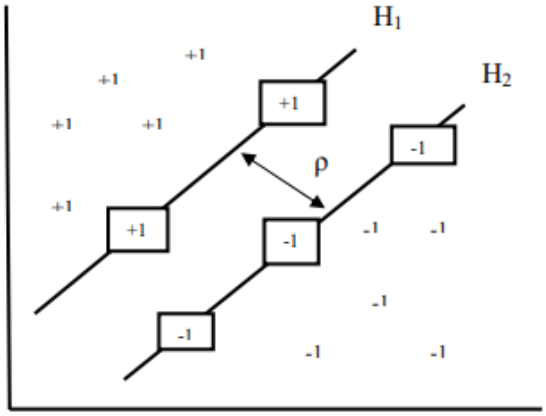
2.5.2.6.2.1 Hiperplano Ótimo para Padrões Linearmente Separáveis

O objetivo de uma SVM para padrões linearmente separáveis é construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima. Considerando uma amostra de treinamento $\{(x_i, d_{i)}\}_{i=1}^{N}$, em que x_i é o padrão de entrada para o i-ésimo exemplo e d_i é a resposta desejada correspondente, também chamado saída-alvo. Inicialmente, assume-se que estes padrões representam duas classes distintas “linearmente separáveis”. A equação de uma superfície de decisão na forma de um hiperplano que realiza esta separação é dada pela Equação 2.20:

$$\begin{aligned} W^T x + b \geq 0 \;\; para \; d_i = 1 \\ W^T x + b < 0 \;\; para \; d_i = -1 \end{aligned} \quad (2.20)$$

A margem de separação (representada por ρ) é a distância entre o hiperplano definido na Equação 2.15 e o ponto de dado mais próximo, isto para um vetor peso w e bias b específicos. O objetivo de uma máquina de vetor de suporte é encontrar o hiperplano particular para o qual a margem de separação ρ é máxima. Sob esta condição, a superfície de decisão é referida como um Hiperplano ótimo, conforme FIGURA 23.

FIGURA 23 - HIPERPLANO ÓTIMO PARA PADRÕES LINEARMENTE SEPARÁVEIS.



FONTE: ALMEIDA (2010)

Considerando ainda que w_0 e b_0 representam os valores ótimos do vetor peso e do bias, respectivamente. O hiperplano ótimo representa uma superfície de decisão linear multidimensional no espaço de entrada e é definido pela equação (2.21):

$$W^T x + b_0 = 0 \tag{2.21}$$

Segundo Duda, Hart e Stock (1973), a função discriminante que fornece uma medida algébrica da distância de x até o hiperplano é dada pela Equação (2.22):

$$g(x) = W_0^T x + b_0 \tag{2.22}$$

Onde x é representador por:

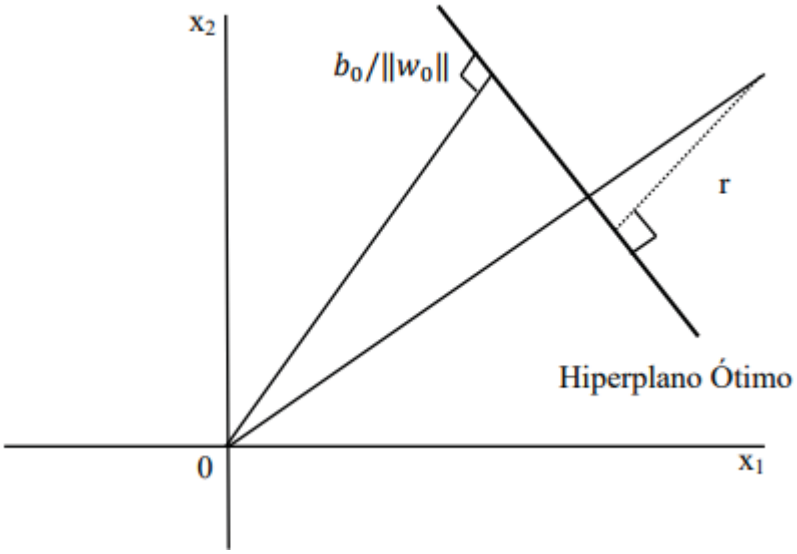
$$x = x_p + \mathrm{r} \frac{W_0}{\|W_0\|} \tag{2.23}$$

em que x_pé a projeção normal de x sobre o hiperplano ótimo, e r é a distância algébrica desejada; r é positivo se x estiver no lado positivo do hiperplano ótimo e negativo se x estiver no lado negativo. Uma vez que por definição $g(x_p) = 0$, então resulta que:

$$g(x) = W_0^T x + b_0 = r\|W_0\| \tag{2.24}$$

De acordo com HAYKIN (2007), observa-se na FIGURA 24, que a distância da origem (i.e., x=0) até o hiperplano ótimo é dada por $\frac{b_0}{\|W_0\|}$. Se $b_0 > 0$, a origem está no lado positivo do hiperplano ótimo; se $b_0 < 0$ ela está no lado negativo. Se $b_0 = 0$, o hiperplano ótimo passa na origem.

FIGURA 24 - DISTÂNCIAS ALGÉBRICAS DE UM PONTO ATÉ O HIPERPLANO ÓTIMO PARA UM CASO BIDIMENSIONAL



FONTE: HAYKIN (2007)

Assim, (W_0, b_0), satisfazem as restrições da equação (2.20), onde os pontos de dados particulares (W_i, b_i) são chamados de vetores de suporte, isto, é são aqueles pontos de dados que se encontram mais próximos da superfície de decisão e são, portanto, os mais difíceis de classificar. Estes vetores de suporte tem influência direta na localização ótima da superfície de decisão, por isso o nome Máquina de Vetor de Suporte - SVM.

Assim, o hiperplano ótimo é único no sentido de que o vetor peso W_0 fornece a máxima separação possível entre exemplos positivos e negativos. Esta condição é alcançada minimizando a norma euclidiana do vetor peso w. A margem de separação é então definida pela Equação (2.25):

$$\rho = \frac{2}{\|W_0\|} \tag{2.25}$$

2.5.2.6.2.2 Hiperplano Ótimo para Padrões Não-Linearmente Separáveis

Segundo HAYKIN (2007), para um conjunto de treinamento onde os padrões são não-separáveis, pode-se encontrar um hiperplano ótimo que minimize a probabilidade de erro de classificação, sendo essa probabilidade calculada através da média sobre o conjunto de treinamento. O hiperplano de separação entre classes é suave se um ponto de dado (x_i, d_i) violar a seguinte condição dada pela equação (2.26):

$$d_i(W^T x_i + b) \geq 1 \quad para\ i = 1, 2, \dots, N \tag{2.26}$$

Para generalizar a situação descrita acima, é inserida uma variável escalar e não negativa $\xi = (\xi_1, \xi_2, \dots \xi_i)$ chamada de variável de folga que é incluída na equação que define o hiperplano de separação, dado por:

$$d_i(W^T x_i + b) \geq 1 - \xi_1 \quad para\ i = 1, 2, \dots, N \tag{2.27}$$

As variáveis de folga ξ_i medem o desvio de cada amostra de sua condição ideal de separabilidade de padrões. Para $0 < \xi < 1$ o ponto de dado se encontra dentro da região de separação, mas no lado correto da superfície de decisão. Para $\xi_i > 1$, o ponto de dado está localizado no lado incorreto do hiperplano de separação (ALMEIDA, 2010; LORENA & DE CARVALHO, 2007).

De acordo com Almeida (2010), os vetores de suporte são as amostras que satisfazem a igualdade presente na Equação (2.25), isto é, são as amostras que estão mais próximas do hiperplano. Se um exemplo $\xi_i > 0$ for deixado fora do conjunto de treinamento, a superfície de decisão não muda. Deste modo, os vetores de suporte são definidos do mesmo modo tanto para o caso linearmente separável como para o caso não linearmente separável.

Da mesma forma, o objetivo é encontrar um hiperplano de separação para o qual o erro de classificação é minimizado. Isto pode ser feito minimizando o funcional dado pela Equação (2.28):

$$\Phi(\xi) = \sum_{i=1}^{N} I(\xi i - 1) \tag{2.28}$$

Em relação ao vetor peso w, a restrição da Equação (2.30) e a restrição em relação à $\|W\|^2$:

$$\|W\|^2 \leq \frac{1}{\rho} \tag{2.29}$$

A função I (ξ) é uma função indicadora e definida por:

$$I(\xi) = \begin{cases} 0 \; se \; \xi \leq 0 \\ 1 \; se \; \xi > 0 \end{cases} \tag{2.30}$$

A minimização de Φ (ξ) em relação a w é um problema de otimização. Para solucionar esse problema deve ser feita uma aproximação dada por:

$$\Phi(\xi) = \sum_{i=1}^{N} \xi i \tag{2.31}$$

O funcional a ser minimizado em relação ao vetor peso w é dado pela equação (2.32), que satisfaz a equação (2.33):

$$\Phi(w, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi i \tag{2.32}$$

$$\begin{aligned} d_i(w^T x_i + b) &\geq 1 - \xi i \qquad && i = 1,2 \ldots, N \\ \xi_i &\geq 0 && \text{para todo i} \end{aligned} \tag{2.33}$$

A minimização de w está relacionada com a minimização da dimensão VC, o valor da dimensão VC equivale ao maior número de exemplos de treinamento que podem ser aprendidos pela máquina sem erros, sendo calculado da seguinte forma: VC = 2 quando o problema pode ser separado por

uma reta, VC = 3 quando o problema pode ser separado por um plano, VC = 4, quando o problema é separado por um hiperplano e assim por diante. A dimensão VC é $n + 1$ sendo n a dimensão do espaço vetorial em questão.

O termo $C \sum_{i=1}^{N} \varepsilon_i$da equação 2.32 faz com que o hiperplano de separação ótimo se torne menos sensível à presença de exemplos ruins no conjunto de treinamento. O parâmetro C pode ser considerado como um parâmetro de regularização, que controla o compromisso entre a complexidade da máquina e o número de erros de treinamento. Este parâmetro é escolhido pelo usuário e normalmente é determinado experimentalmente, através do desempenho do algoritmo via dados de validação, ou de forma analítica estimando a dimensão VC (ALMEIDA, 2010).

Usando o método dos multiplicadores de Lagrange, pode-se formular o problema dual para padrões não separáveis, que maximizam a função objetivo, dada pela equação (2.34), sujeito as restrições da equação (2.35):

$$Q(\propto) = \sum_{i=1}^{N} \propto_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \propto_i \propto_j d_i d_j x_i^T x_j \tag{2.34}$$

$$\sum_{i=1}^{N} \propto_i d_i = 0 \quad \text{e} \quad 0 \leq \propto_i \leq C \quad \text{para i =1,2, ..., N} \tag{2.35}$$

De acordo com Haykin (2007), a função objetivo $Q(\propto)$ a ser maximizada é a mesma para os casos de padrões linearmente separáveis e não separáveis. O caso não separável difere do caso separável pelo fato de que a restrição $\propto_i \geq 0$ é substituído pela restrição mais rigorosa $0 \leq \propto_i \leq C$. Exceto por esta modificação, a otimização restrita para o caso não separável e os cálculos dos valores ótimos do vetor peso w e do bias b procedem do mesmo modo como no caso linearmente separável. A solução ótima para o vetor peso w é dada por:

$$w_0 = \sum_{i=1}^{N_S} \propto_{o,i} d_i x_i \tag{2.36}$$

em que N_S é o número de vetores de suporte. Enquanto que b pode ser determinado a partir de α, e pelas novas condições de Karush-Kuhn-Tucker (HAYKIN, 2007):

$$\propto_i [d_i(w^T x_i + b) - 1 + \xi_i] = 0 \qquad \text{i} = 1, 2, \ldots, \text{N}$$

$$\mu_i \xi_i = 0 \qquad i = 1, 2, \ldots, N \tag{2.37}$$

Os μ_i são multiplicadores de Lagrange que foram introduzidos para forçar a não negatividade das variáveis de folga ξ_i para todo i. No cálculo do ponto mínimo (ponto de sela), a derivativa da função Lagrangeana com respeito às variáveis ξ_i é zero, sendo assim:

$$\alpha_i + \mu_i = C \tag{2.38}$$

Por fim, pode-se determinar o bias ótimo b_0 utilizando qualquer ponto de dado (x_i, d_i) do conjunto de treinamento na Equação (2.27), para o qual tem-se $0 \leq \propto_i \leq C \; e \; \xi_i = 0.$

A superfície de decisão da SVM, que no espaço de características é sempre linear, normalmente é não-linear no espaço de entrada. Conforme mencionado anteriormente, a ideia de uma Máquina de Vetor de Suporte depende de duas operações matemáticas que podem ser resumidas como: o mapeamento não-linear de um vetor de entrada para um espaço de características de alta dimensionalidade e a construção de um hiperplano ótimo para separar as características descobertas no primeiro passo. Esta não linearidade da superfície de separação é obtida pelas funções de *Kernel* (ou núcleo do produto interno) e depende da solução de problemas que representam exemplos não-linearmente separáveis. Um *Kernel* é uma função que recebe dois pontos x_i e x_j do espaço de entradas e calcula o produto escalar desses dados no espaço de características (HERBRICH, 2001). dado pela equação (2.39):

$$k(x_i, x_j) = \Phi(x_i).\Phi(x_j) \tag{2.39}$$

Para garantir a convexidade do problema de otimização e que o *Kernel* apresente mapeamento nos quais seja possível o cálculo de produtos escalares, utiliza-se a função *kernel* que segue as condições estabelecidas pelo teorema de Mercer. Os *Kernels* que satisfazem a condição de Mercer são caracterizados por dar origem a matrizes positivas semi-definidas k, em que cada elemento k_{ij}é definido por $k_{ij} = k(x_i, x_j)\ para\ todo\ i, j = 1, \ldots, n$ (HERBRICH, 2001).

As funções Kernel mais utilizadas são dadas no QUADRO 1.

QUADRO 1 - EXEMPLOS DE FUNÇÕES *KERNEL*

Kernel	**Expressão**	**Parâmetros**
Polinomial	$((x^i)^t x^j + k)^p$	p e k
Gaussiano	$e^{-\frac{\|x^i - x^j\|^2}{2\sigma^2}}$	σ
Sigmoidal	$\tanh\,(m(x^i)^t x^j + k$	m e k

FONTE: A autora (2022).

Os vetores suportes são os dados x^{i} cujos multiplicadores de Lagrange α_{i} possuem valores não nulos. Estes são os valores que contribuem para a construção do hiperplano ótimo. Uma característica das soluções do SVM é a esparsidade dos Multiplicadores de Lagrange, ou seja, apenas uma fração dos pontos será vetor suporte (CRISTIANINI; SHAWE-TAYLOR, 2000).

Para classificar os padrões do conjunto de treinamento como vetores suporte verificam-se os valores dos respectivos Multiplicadores de Lagrange conforme as condições de Karush-Kuhn-Tucker (KKT), ou seja:

Se $\alpha_i = 0, y_i \cdot f(x^i) > 1$ então x^i é considerado um vetor comum, que se situa do lado correto, a região da sua classe.

Se $0 < \alpha_i < C, y_i \cdot f(x^i) = 1$ então x^i é um vetor suporte, situa-se sobre a margem da região da sua classe. É conhecido como vetor suporte não-*bound* (VS-VB) (CRISTIANINI; SHAWE-TAYLOR, 2000).

A mineração de dados surgiu com o objetivo principal de dar suporte à tomada de decisões nas organizações. Portanto, a aplicação de técnicas de mineração de dados em sistemas de descoberta de conhecimento em banco de

dados busca a descoberta de regras e padrões em dados que trarão o conhecimento suficiente e adequado para aquelas pessoas responsáveis pela tomada de decisões (DIAS, 2001).

Segundo Castro e Ferrari (2016) os *softwares* mais utilizados para dar suporte às tarefas de mineração de dados são: *Weka; Matlab, R, Wolfram Mathematica; RapidMiner; SAS; SSPS; Orange; Mahout; Elki; e Libsvm*.

2.5.3 Pós-Processamento

A etapa de pós-processamento abrange o tratamento e o entendimento dos conhecimentos obtidos na mineração de dados. Tal tratamento tem como objetivo viabilizar a avaliação da utilidade do conhecimento descoberto (FAYYAD; PIATESTKY; SHAPIRO, 1996).

Hussain *et al.* (2000) apresentam um método que identifica, a partir de um conjunto de padrões descoberto, um subconjunto de regras que representam regras de exceção. A FIGURA 25 mostra a estrutura geral das regras de exceção, considerando uma regra de “bom senso”, ou “senso comum”, e uma regra de referência. Conforme FIGURA 25, A e B são conjuntos não-vazios de pares de atributo valor, e C representa a classe predita pela regra. O símbolo “¬” denota a negação lógica. É importante observar que uma regra de exceção é uma especialização de uma regra de senso comum, e uma regra de exceção prediz uma classe distinta da classe prevista pela regra de senso comum. Este método assume que regras de senso comum representam padrões conhecidos pelo usuário, tendo em vista que aquelas regras têm uma grande cobertura, ao contrário das regras de exceção, que, em geral, são desconhecidas, uma vez que elas têm baixa cobertura. Sendo assim, as regras de exceção tendem a ser surpreendentes, dado o fato de representarem uma contradição em relação à regra de senso comum. É importante observar que a regra de referência auxilia na explicação da causa da regra de exceção (CALIL *et al.*, 2008; Hussain *et al.*, 2000).

FIGURA 25 - ESTRUTURA DAS REGRAS DE EXCEÇÃO

$A \rightarrow C$ regra de senso comum (alta cobertura e alta precisão)
$A, B \rightarrow \neg C$ regra de exceção (baixa cobertura, alta precisão)
$B \rightarrow \neg C$ regra de referência (baixa cobertura e/ou baixa precisão)

FONTE: HUSSAIN *et al.* (2000)

Segundo Tan, Steinbach e Kumar (2009), o pós-processamento deve assegurar que somente resultados válidos e úteis sejam incorporados ao sistema de apoio à decisão.

2.6 ESTUDOS CORRELATOS

Com objetivo de analisar os estudos que abordam os temas apresentados anteriormente, foram realizadas duas revisões sistemáticas que consistiram de quatro etapas: busca nas bases de dados científicas, seleção por meio de leitura técnica, mapeamento dos estudos correlatos e elaboração de uma tabela com os principais estudos. Na primeira etapa foram feitas buscas nas bases científicas *Web Of Science*, *Science Direct* e SciELO de artigos nos idiomas inglês, espanhol ou português, com o período de 12 anos (2011-2022), considerando as palavras individuais "*traffic accident"* e utilizando o conector OR com o termo "*traffic safety*" e o conector AND com o termo "*data mining*" no resumo, palavra-chave ou título, onde foram identificados, respectivamente, 59, 126 e 0 artigos, totalizando 185 artigos, sendo destes 9 duplicados, resultando em 176 artigos.

Durante a segunda etapa, foi realizada uma leitura flutuante dos resumos dos artigos, sendo descartados 121 artigos por não terem relação com o tema. Após análise dos resumos, 52 artigos foram selecionados para leitura integral e análise do conteúdo, a qual a seleção se deu por meio da conferência de que as palavras buscadas estavam no modelo ou hipóteses de pesquisa considerando os atributos utilizados nos estudos. Com esta seleção, a quantidade de artigos foi reduzida para 48. Através da classificação JCR (*Journal Citation Reports*) considerando o impacto dos periódicos, foi utilizado o *metodi in ordinatio* (Pagani *et al.*, 2015) para a análise dos principais estudos. Informações referentes aos 48 estudos encontram-se no APÊNDICE 1.

Nesta revisão da literatura, foi verificado que as metodologias mais utilizadas em estudos de sinistros de trânsito e mineração de dados nos últimos 12 anos foram: Árvores de Decisão, Naive Bayes, Redes Neurais e Máquinas de Vetores de Suporte.

Na segunda revisão foram feitas buscas nas bases científicas *Web Of Science*, *Science Direct* e SciELO de artigos nos idiomas inglês, espanhol ou português, com o período de 22 anos (2001-2022), considerando as palavras individuais "*traffic accident*" e utilizando o conector OR com o termo "*traffic safety*" e o conector AND com o termo "*data mining*" e "*truck*" no resumo, palavra-chave ou título, onde foram identificados, respectivamente, 2, 99 e 0 artigos, totalizando 101 artigos.

Durante a segunda etapa, foi realizada uma leitura flutuante dos resumos dos artigos, sendo descartados 63 artigos por não terem relação com o tema. Após análise dos resumos, 38 artigos foram selecionados para leitura integral e análise do conteúdo, a qual a seleção se deu por meio da conferência de que as palavras buscadas estavam no modelo ou hipóteses de pesquisa considerando os atributos/variáveis utilizados nos estudos e considerando sinistros envolvendo caminhões. Com esta seleção, a quantidade de artigos foi reduzida para 12. Nesta análise foram verificados os atributos mais utilizados nos estudos e os padrões de sinistros envolvendo o transporte rodoviário de cargas. Informações referentes aos 12 estudos encontram-se no APÊNDICE 2.

O estudo de Hakkanen e Summala (2001) mostrou que o condutor caminhoneiro não foi o principal responsável pelos sinistros de trânsito ocorridos na Finlândia no período de 1991 a 1997, 83% dos sinistros foram originados por outros usuários como condutores de veículos e pedestres. A pesquisa ainda apontou que os fatores que se mostraram preponderantes nos sinistros foram a idade do motorista (motoristas mais jovens), a fase do dia (o período da noite foi preponderante), o histórico de sinistros destes condutores e a existência de doenças crônicas nos condutores de veículos de carga. A colisão frontal foi tipo de sinistro mais comum no período da pesquisa, ocorrendo principalmente em pista simples, visto que na Finlândia, o sistema rodoviário principal consiste neste tipo de pista.

No estudo de Forkenbrock e Hanley (2003), utilizou dados do Instituto de Pesquisa em Transporte da Universidade de Michigan considerando os

sinistros no período de 1995 a 1998. O trabalho mostrou que os caminhões com mais reboques têm maior probabilidade de se envolver em colisões fatais nas seguintes condições: escuridão; neve, lama ou gelo na superfície da estrada; envolvimento de três ou mais veículos e em estradas com limites de velocidade entre 105 a 120 km/h. Assim os fatores que mais impactaram nos sinistros foram com relação à fase do dia, condição da superfície na estrada, número de veículos envolvidos e o limite de velocidade na via.

As taxas de sinistros envolvendo veículos de carga em Taiwan no período de 1994 a 1998 variaram significativamente com fatores como idade do motorista, tipo de veículo envolvido e as condições da via como obstruções, defeitos na superfície da estrada e visibilidade, conforme aponta a pesquisa de Tsai e Su (2004). Os resultados do estudo indicaram que as taxas de sinistros com veículos de carga em Taiwan foram altas nos cenários envolvendo caminhões e veículos e os sinistros mais graves ocorreram envolvendo motoristas mais idosos (TSAI; SU, 2004). Já o estudo de Chang e Chien (2013) realizado em Taiwan entre 2005 e 2006, mostrou que consumo de bebidas alcoólicas, o não uso do cinto de segurança, o tipo de veículo envolvido, o tipo de colisão, a circunstância, o número de veículos envolvidos e o local do sinistro foram os principais determinantes de gravidade nas lesões envolvendo sinistros com transporte rodoviário de cargas.

A pesquisa de Khorashadi *et al.* (2005), realizada com dados de sinistros da Califórnia no período de 1997 a 2000, mostrou que sinistros em ambientes rurais envolvendo combinações de caminhão trator e reboque, a probabilidade de ferimentos grave/fatais aos motoristas aumentou cerca de 26% em relação aos sinistros envolvendo caminhões com apenas um reboque. Em áreas urbanas, esta mesma probabilidade aumentou quase 700%. Em sinistros em que o uso de álcool ou drogas foi identificado como a causa principal do sinistro, a probabilidade de lesão grave/fatal aumentou cerca de 250% nas áreas rurais e quase 800% nas áreas urbanas. Além disso, o estudo apontou como principal tipo de sinistro a colisão lateral e principalmente com veículos de carga com ano de fabricação anterior a 1981. Comparativamente o estudo de Lemp *et al.* (2011), realizado com dados de 2001 a 2003 nos EUA, mostrou que a probabilidade de fatalidades e lesões graves aumenta conforme o número de reboques, mas diminui considerando o comprimento do caminhão e a classificação de peso

bruto do veículo. O estudo ainda conclui que a probabilidade de fatalidade aumenta quando os sinistros apresentam as seguintes características: iluminação fraca na via, a superfície da estrada estiver com neve ou gelo e a condição climática de neblina.

No estudo de Islam e Hernandez (2013) no Texas - EUA, no período de 2005 a 2008, o nível de gravidade da lesão foi considerado altamente influenciado por uma série de interações complexas relacionadas a fatores humanos, veiculares e ambientais da estrada, como por exemplo a idade do condutor de caminhão, excesso de velocidade na via, não uso de cinto de segurança, o traçado da via ser uma curva, via sem iluminação e no período de verão. Análogo o estudo de Islam *et al.* (2014) no Alabama, considerando o período de 2010 a 2012, a pesquisa mostrou que as influências de uma variedade de variáveis na gravidade das lesões foram diferentes, considerando fatores com relação ao motorista, via, veículo e ambiente.

Já a pesquisa de Pahukula, Hernandez e Unnikrishnan (2015) realizada no Texas no período de 2006 a 2010, apontou que a fase do dia, o fluxo de tráfego, condições de luz, condições da superfície da estrada, época do ano e porcentagem de caminhões na estrada foram consideradas como as principais variáveis influenciadoras de sinistros graves.

Por fim, o estudo de Zheng, Lu e Lantz (2018), apresenta os seguintes fatores como contribuintes principais em sinistros na Dakota do Norte e Colorado no período de 2010-2016: atributos da empresa de transporte rodoviário (por exemplo, tamanho da empresa), valores de inspeção de segurança, status de comércio da empresa de transporte (por exemplo, interestadual ou intraestadual), hora do dia, idade do motorista, primeiros eventos prejudiciais e condição de registro estão significativamente associados com a gravidade de lesões por sinistro.

Pela primeira vez, provou-se que as características da empresa de transporte rodoviário de cargas e do motorista têm um impacto significativo na gravidade das lesões causadas por acidentes de caminhão. Alguns dos resultados deste estudo ainda corroboram com a gravidade de sinistros como: a superfície de estrada molhada, má visualização (condições escuras ou de pouca luz ou nevoeiro/más condições meteorológicas), forte vento, peso bruto do caminhão (mais de 11 mil kg) e colisões com veículos opostos. Condutores

jovens (menos de 25 anos) e condutores idosos (acima de 75 anos) são os grupos com maior probabilidade de se envolver em acidentes que resultam em mortes. Além disso, o nível de gravidade do acidente de caminhão aumenta quando mais veículos estão envolvidos nos sinistros (ZHENG; LU; LANTZ, 2018).

Outra descoberta interessante neste estudo é o aumento da probabilidade de sinistros fatais quando o tempo está bom ou quando a superfície da estrada não apresenta condições adversas, talvez porque as condições adversas tornem as pessoas mais atentas aos riscos potenciais (ZHENG; LU; LANTZ, 2018).

Em uma pesquisa realizada na Suécia, com informações de sinistros que ocorreram entre 1995 e 2004, verificou-se que colisões com veículos de transporte rodoviário de cargas geralmente ocorriam durante o dia, em dias úteis, no inverno e em estradas com pista dupla a uma velocidade de aproximadamente 70 a 90 km/h (BJÖRNSTIG; BJÖRNSTIG; ERIKSSON, 2008). Este trabalho apresentou como fatores importantes nestes sinistros a fase do dia, o dia da semana, período do ano, tipo de pista e limite de velocidade na via. Fatores como ambiente, geometria da estrada e características do tráfego se mostraram com menos importância na gravidade dos sinistros.

Com relação à prevenção de sinistros graves, o estudo de Zhu e Srinivasan (2011), realizado com dados de sinistros em rodovias nos EUA de 2001 a 2003, mostrou que o uso de airbags e cintos de segurança trazem benefícios para a segurança dos motoristas e passageiros. O estudo salientou que o uso de drogas ilícitas e desatenção do motorista de caminhão aumentam a gravidade da lesão nos sinistros. Além disso, no caso de colisões frontais ou ocorridas em cruzamentos (maioria apontada no estudo) os fatores dia da semana (finais de semana) e a fase do dia (noturno) tiveram um maior impacto.

Os sinistros de trânsito são uma das maiores ameaças à saúde pública do mundo (PAKGOHAR *et al.,* 2011). Diferentes estudos na literatura utilizam diversas técnicas de mineração de dados para analisar sinistros rodoviários. Os modelos variam da aplicação de técnicas de classificação, como árvores de decisão, técnicas de agrupamento e regras de associação. A maioria dos estudos concentra-se na gravidade de um sinistro para encontrar padrões associados a ele (JOHN; SHAIBA, 2019). Conforme revisão sistemática, os

atributos mais presentes nos 12 artigos são nesta ordem: idade do condutor, condição meteorológica, tipo do veículo envolvido, tipo do acidente, condição da superfície da estrada, tipo de pista, fase do dia, sexo do condutor, dia da semana, traçado da via, hora, limite de velocidade, tipo de solo, local, ano veículo, tempo de habilitação, uso do cinto de segurança, teste embriaguez ao volante, número de envolvidos, número de veículos envolvidos, data e condição de pavimentação.

2.7 MEDIDAS DE DESEMPENHO DOS ALGORITMOS UTILIZADOS NO ESTUDO

Para este estudo foram consideradas as quatro técnicas mais utilizadas nos estudos correlatos: árvores de decisão, redes neurais, *Naive Bayes* e máquina de vetores de suporte. Foi utilizado o *software* WEKA, juntamente com os algoritmos J48 (árvore de decisão), *Multilayer Perceptron* (MLP – redes neurais), *Naive Bayes* e *Sequential Minimal Optimization* (SMO – máquina de vetores de suporte).

Para a análise do desempenho dos algoritmos, foram considerados os índices de desempenho: acurácia, curva ROC e estatística *Kappa*.

Acurácia: mede a proximidade entre o valor obtido na classificação e o valor verdadeiro da medição, conforme equação 2.40:

$$\frac{\mathrm{VP+VN}}{\mathrm{VP+VN+FP+FN}} \tag{2.40}$$

onde:

VP = verdadeiro positivo

VN = verdadeiro negativo

FP = falso positivo

FN = falso negativo (THARWAT, 2021).

Segundo Tharwat (2021), essa medida avalia quão efetivo um algoritmo é, mediante a probabilidade deste realizar predições corretas. A acurácia

corresponde às instâncias corretamente classificadas, sendo a melhor precisão (acurácia) 1 (um) e a pior 0 (zero).

Curva ROC: Segundo Tharwat (2021), área ROC (do inglês, *Receiver Operating Characteristic*) é um gráfico que visualiza troca entre taxa de verdadeiros positivos (TVP) e taxa de falsos positivos (TFP). Segundo o autor, para cada limiar, calculam-se TVP e TFP e se plota num gráfico. Os melhores classificadores têm maior curva para a esquerda. A área sob a curva, conhecida como ROC *AUC score* (*area under the curve*), é usada como uma medida de qualidade, ou seja, um número que determina se a curva ROC é boa. Hosmer e Lemeshow (2013) sugerem uma regra geral para a classificação da capacidade de discriminação, dependendo do valor da curva ROC:

- ≤ 0,5 Não tem poder discriminativo
- 0,5–0,7 Discriminação fraca
- 0,7–0,8 Discriminação aceitável
- 0,8–0,9 Discriminação boa
- ≥ 0,9 Discriminação excelente.

Estatística *Kappa*: avalia o nível de concordância e ligação dos dados dentro de uma base de dados sendo que, se o número estatístico ficar próximo do 0 (zero), significa uma maior discordância das informações, e ficando o mais próximo do 1 (um) indica maior ligação e concordância.

Segundo Lantz (2019), o valor *Kappa* aceitável é aquele acima de 0,80. Já os intervalos definidos por Landis e Koch (1977) são:

- ≤ 0 Nenhuma concordância
- 0,01–0,2 Leve concordância
- 0,21–0,4 Concordância Regular
- 0,41–0,6 Concordância Moderada
- 0,61–0,8 Concordância Substancial
- 0,81-1 Concordância quase perfeita.

2.7.1 J48

O algoritmo J48 pode ser utilizado tanto com atributos contínuos e discretos, quanto com valores categóricos e valores ausentes. O tratamento de

atributos contínuos envolve a consideração de todos os valores presentes no conjunto de treinamento, fazendo com que sejam ordenados de forma crescente e, após esta ordenação, seja selecionado o valor que favorecerá a redução da entropia (RAMYA *et al.,* 2015).

Os cálculos para escolha do atributo referente ao nó raiz são realizados através da redução de entropia: Cálculo $Info(S)$ para identificar a classe no conjunto de treinamento S:

$$Info(S) = -\sum_{i=1}^{k} \left\{ \left[\frac{freq(C_i,S)}{|S|} \right] \log_2 \left[\frac{freq(C_j,S)}{|S|} \right] \right\} \quad (2.41)$$

Onde $|\mathrm{S}|$é o número de casos no conjunto de treinamento; $\mathrm{C_i}$ é a classe: $\mathrm{i} = 1, 2, 3, \ldots, \mathrm{k}, \mathrm{k}$ =número de classes; $\mathrm{freq}\ (\mathrm{C_i}, \mathrm{S})$ = número de casos em $\mathrm{C_i}$. Após há o cálculo do valor da informação esperada, $\mathrm{Info_x(S)}$, para o atributo x da partição S. Onde n é o número de valores possíveis que o atributo pode assumir, sendo o número de nós-filhos, N é o número total de objetos do nó-pai e $\mathrm{N(S_i)}$ é o número de exemplos associados ao nó filho $\mathrm{S_i}$ (QUINLAN, 1993).

$$Info_x(S) = -\sum_{i=1}^{n} \left[\left(\frac{|S_i|}{|S|} Info(S_i) \right] \right. \quad (2.42)$$

Assim, o ganho da informação é dado pela equação (2.42):

$$Ganho(X) = Info(S) - Info_x(S) \quad (2.43)$$

Segundo Quinlan (1993), o uso do critério de ganho de informação para escolha do nó raiz da árvore favorece dados com grandes variações nos valores, podendo representar um viés. Assim, a razão de ganho de informação, representada pela equação (2.43), em que o denominador normaliza o conjunto de amostra de atributos que apresentam grandes variações, pode superar tal limitação ao suavizar favorecimentos que por ventura venham a acontecer e certificar que a melhor escolha tenha sido feita.

$$RG = \frac{Ganho(X)}{\sum_{i=1}^{k} \left[\frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \right]} \quad (2.44)$$

Neste processo, foram verificadas duas fases de construção da árvore de decisão, o crescimento, composto pelas fases de treinamento e teste; e a poda. Na fase de crescimento os dados são divididos em grupos, podendo ser direcionados para treinamento e aprendizado da estrutura, e ainda para o teste que idêntica a capacidade preditiva da árvore (LAROSE; LAROSE, 2014). A sequência de passos para construção e poda da árvore está abordada no pseudocódigo do algoritmo conforme FIGURA 26 (CAMILO; SILVA, 2009).

FIGURA 26 - PSEUDOCÓDIGO DO ALGORITMO J48

```
Input: um conjunto de dados D
begin
    Árvore={};
    if D é "puro" OU existe outro critério de parada then
        encerrar;
    foreach atributo a ∈ D do
        Calcular ganho de informação;
    end
    a_melhor = Melhor atributo de acordo com o calculo do ganho de informação;
    Árvore = Cria um nó baseado no a_melhor;
    D_v = Divide o subconjunto de D baseado no a_melhor;
    foreach D_v do
        Árvore_v = J48(D_v);
        Fixe a Árvore_v no galho obtido no passo anterior;
    end
end
return [Árvore]
```

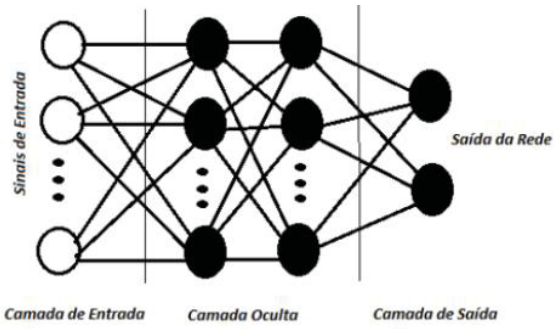
FONTE: CAMILO e SILVA (2009)

2.7.2 *Multilayer Perceptron*

A rede Perceptron de Multicamadas (MLP, do inglês *Multilayer Perceptron*) é construída a partir de um conjunto de nós fonte, os quais formam a camada de entrada da rede, uma ou mais camadas ocultas de nós computacionais *(perceptrons)* e uma camada de saída, também de nós computacionais. Com exceção da camada de entrada, todas as outras camadas realizam processamento (SOUZA, 2012).

Na FIGURA 27, verifica-se um tipo de rede neural unidirecional distribuída em camadas. Em cada camada é possível ter vários neurônios. As camadas se comunicam camada a camada até atingir a última camada, a FIGURA 27 apresenta uma camada de entrada com n fontes de neurônios, duas camadas ocultas com n neurônios computacionais, e uma camada de saída com dois neurônios.

FIGURA 27 - MODELO DE REDE MLP



FONTE: SOUZA (2012).

Segundo Souza (2012), ao projetar uma MLP é necessário considerar dois aspectos, primeiro determinar o número de camadas ocultas, e segundo o número de neurônios em cada camada. A rede ainda pode ser classificada em dois grupos: parcialmente conectadas e totalmente conectadas. Em uma MLP totalmente conectada, cada neurônio de uma camada é conectado a todos os neurônios da camada anterior e a todos os neurônios das camadas posteriores. Em uma MLP parcialmente conectada, algumas conexões entre neurônios não acontecem.

Entre a camada de entrada e a camada de saída, pode-se ter uma ou mais camadas ocultas. As camadas ocultas proporcionam complexidade e possibilidade de resolver problemas não linearmente separáveis. Assim, uma das principais características de uma rede MLP é sua capacidade de resolver problemas não lineares, para isso, é necessário que a função de ativação dos neurônios pertencentes às camadas ocultas seja não linear. Em geral, a função de ativação é sigmoide (SOUZA, 2012).

Segundo Haykin (2007), uma rede do tipo MLP tem três características que a distinguem dos demais tipos de rede: a primeira, para cada neurônio da rede há uma função de ativação não linear. Ao contrário do que ocorre com o *Perceptron* proposto por Rosenblatt, cuja função de ativação possui curvatura abrupta, nas redes do tipo MLP essa curvatura é diferençável ao longo de todo o domínio. Segundo, a rede contém uma ou mais camadas ocultas, que são diferentes da camada de entrada e de saída. Os neurônios dessas camadas ocultas são responsáveis pela capacidade de aprendizagem de problemas complexos, e por último, existe um alto grau de conectividade entre os neurônios. Isso significa que um neurônio de qualquer camada da rede está conectado a

todos os neurônios da camada anterior. Uma simples mudança topológica, como a inclusão ou a exclusão de um neurônio em qualquer camada, implica mudança na população das conexões sinápticas ou de seus pesos.

Ainda de acordo com o autor, a MLP é uma rede progressista, ou seja *feedforward*, sendo observada quando as saídas dos neurônios em qualquer camada particular conectam-se unicamente aos neurônios da camada seguinte, sendo entradas para os neurônios seguintes. Como consequência, a entrada se propaga através da rede, camada a camada, em um sentido progressivo.

De acordo com Haykin (2007), a rede MLP tem sido aplicada na solução de diversos problemas, através do treinamento supervisionado com o algoritmo de retropropagação do erro (do inglês, *backpropagation*). Este treinamento consiste em dois passos, um passo para frente e um passo para trás, respectivamente a propagação da entrada da rede as camadas posteriores e retropropagação do erro que ocorre no sentido contrário da camada de saída as camadas ocultas. Neste processo é realizado o ajuste nos pesos sinápticos, conforme equações (2.44) e (2.45).

$$\rho_j(n) = d_j(n) - y_j(n), \cdot\cdot \tag{2.45}$$

onde $\rho_j(n)$ é a diferença entre o valor desejado para determinada entrada e saída gerada pela rede da equação (2.45):

$$\varepsilon(n) = \frac{1}{2}\sum_{j\epsilon C} p_j^2\,(n), \tag{2.46}$$

onde C é o conjunto de todos os neurônios que pertencem a camada de saída da rede e $\varepsilon(n)$ é um valor instantâneo do erro.

Basicamente esse processo de aprendizagem consiste em encontrar um conjunto dos pesos W que minimize o custo da função de erro. A eficiência deste processo é observada por medição dos erros apresentados pela rede, realizada com novas entradas ainda não apresentadas a rede. Esse procedimento mede o desempenho de generalização da rede. No algoritmo *backpropagation*, a saída produzida pela rede é comparada a uma resposta desejada e melhorada a cada interação. De forma genérica, esse método possui uma função erro na saída da

rede. Após seu cálculo, é realizada a propagação em sentido contrário (retropropagação). Nesse procedimento os pesos sinápticos das camadas ocultas são atualizados. Por fim, esse processo consiste em ajustar os pesos sinápticos a fim de minimizar o erro entre a saída produzida pela rede em relação à respectiva saída desejada (SOUZA, 2012).

2.7.3 *Naive Bayes*

O classificador *Naïve Bayes* é determinado por cada dado de amostra, ou dado de treinamento é representado por um vetor de dimensão $n, X = (x_1, x_2, \ldots, x_n)$ e cada atributo é representado por $A_1, A_2, \ldots, A_K$.

Supondo que existam m classes, $C_1, C_2, \ldots, C_m$. Dado uma amostra, X(exemplo, sem classe), o classificador irá nos informar que X pertence à classe que tem a maior probabilidade *a posteriori*, condicionada em X. Assim, o classificador *Naive Bayes* associa uma amostra não conhecida X para uma classe C_i, se, e somente se, $P(C_i|X) > P(C_j|X)\ para\ 1 \leq j \leq m, j \neq i$. Assim, $P(C_i|X)$ é maximizado. A classe C_i para a qual $P(C_i|X)$ é maximizada é chamada de hipótese máxima *a posteriori* (HAN; KAMBER; PEI, 2012).

De acordo com o Teorema de Bayes, conforme equação (2.46):

$$P(C_i|X) = \frac{PC_i.P(X|C_i)}{P(X)} \quad (2.47)$$

Como $P(X)$ é constante para todas as classes, somente $P(X|C_i)P(C_i)$ precisa ser maximizado. No aprendizado não supervisionado, onde a probabilidade *a priori* da classe C_i, $P(C_i)$, não é conhecida, então isto é assumido que as probabilidades das classes são iguais, isto é, $P(C_1) = P(C_2) = \cdots = P(C_m)$, e aí se deve maximizar $P(X|C_i)$ (HAN; KAMBER; PEI, 2012).

No aprendizado supervisionado a probabilidade *a priori* pode ser estimada por $P(C_i) = \frac{S_i}{S}$, onde S_i é o número de amostras que possuem a classe C_i, e S é o total de amostras utilizadas para o treinamento.

Ainda de acordo com os autores, dado o conjunto de dados com muitos atributos, tem-se um tempo computacional elevado para calcular $P(X|C_i)$. No

sentido de reduzir este tempo computacional para o cálculo, o algoritmo *Naive Bayes* assume a independência condicional da classe. Com isto presume-se que os atributos são condicionalmente independentes uns dos outros, dado a classe da amostra, isto é, não há relação de dependência entre os atributos, conforme equação (2.47):

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{2.48}$$

Ou seja:

$$P(X|C_i) = P(x_1|C_i) \times P\big(x_2\big|C_i\big) \times \ldots \times (x_n|C_i)\big) \tag{2.49}$$

Onde x_k se refere ao valor do atributo de A_k da tupla X. Para cada atributo, observa-se se o atributo é categórico ou de valor contínuo. Se A_k for categórico, então $P(x_k|C_i)$ é o número de tuplas da classe C_i em D que possui o valor x_k para A_k, dividido por $|C_{i,D}|$, o número de tuplas da classe C_i em D (HAN; KAMBER; PEI, 2012).

Porém, se os atributos possuírem valor contínuo, existe duas formas de calcular a probabilidade condicional das classes. Uma dela é particionar cada atributo contínuo e então substituir estes valores por seus intervalos discretos correspondentes, assim transformando os atributos contínuos em atributos ordinais. A probabilidade condicional de $P(x_k|C_i)$ é verificada pelo cálculo da fração de registros de treinamento pertencentes à classe C_i que estão inclusos no intervalo que corresponde a x_k. O erro de análise depende da estratégia de particionamento empregada e do número de intervalos discretos. Se o número de intervalos for grande, haverá poucos registros de treinamento em cada intervalo para que se obtenha uma avaliação confiável. Em contrapartida, se houver um número pequeno de intervalos, alguns intervalos podem associar registros de diferentes classes, desta forma perdendo os limites de decisão corretos (TAN; STEINBACH; KUMAR, 2009).

Segundo Han, Kamber e Pei (2012), por meio da distribuição Gaussiana com uma média μ e desvio padrão σ, é considerada outra forma de calcular a probabilidade condicional, através da equação (2.49):

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.50)$$

Assim,

$$P(x_k|C_i) = g(x_k, \mu_{Ci}, \sigma_{Ci}) \quad (2.51)$$

Onde μ_{Ci} é a média e o σ_{Ci} é o desvio padrão dos valores do atributo A_k para as tuplas de treinamento da classe C_i. Após ligam-se essas duas quantidades à equação (2.49), junto com x_k, para estimar $P(x_k|C_i)$. Para prever o rótulo de classe X, $P(C_i)P(X|C_i)$ é avaliado para cada classe C_i. O classificador então prevê que o rótulo de classe da tupla X é a classe C_i, se e somente se,

$$P(C_i)P(X|C_i) > P(C_j)P(X|C_j) para\ 1 \leq j \leq m, j \neq i \quad (2.52)$$

Então, o conjunto de teste é associado para a classe $\mathrm{C_i}$ para o qual $\mathrm{P(X|C_i)P(C_i)}$ é máximo. O classificador *Naïve Bayes* se torna o mais preciso em comparação a outros classificadores quando sua suposição de independência condicional entre as classes for verdadeira. Contudo, na prática, podem existir dependências entre as variáveis (HAN; KAMBER; PEI, 2012).

2.7.4 *Sequential Minimal Optimization* (SMO)

O *Sequential Minimal Optimization* (SMO) é um algoritmo heurístico que utiliza apenas duas variáveis em cada iteração. Com isso apresenta uma solução analítica na iteração, além de não haver a necessidade de resolver o problema quadrático. Trabalhando com dois Multiplicadores de Lagrange de cada vez e mantendo os outros fixos, a condição principal para o SMO é $\sum_{i=1}^{\mathrm{l}} \alpha_i\, y_i = 0$, ou representado pela equação (2.52):

$$\sum_{i\in A} \alpha_i y_i = \sum_{j\in B} \alpha_j y_j \quad (2.53)$$

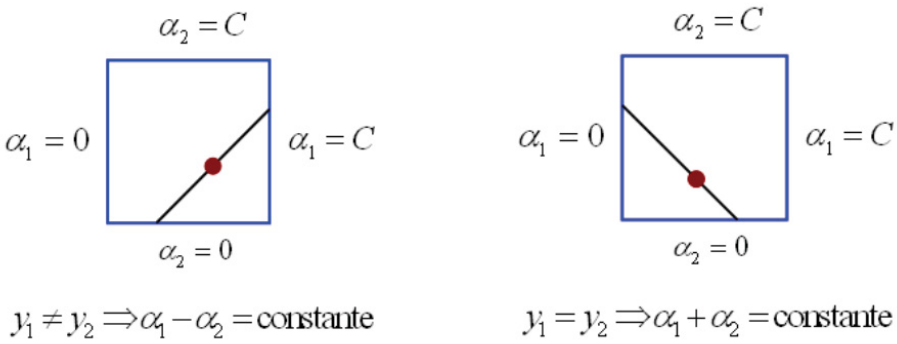
Ou seja, essa condição obriga que quando um Multiplicador é atualizado o outro deve ser ajustado para manter a condição verdadeira. A escolha dos dois pontos é feita a partir de uma heurística, já a atualização dos valores é feita de forma analítica (CRISTIANINI e SHAWE-TAYLOR, 2000).

A fim de não violar a condição de $\sum_{i=1}^{l} \alpha_i \, y_i = 0$, os novos Multiplicadores de Lagrange devem respeitar a equação (2.53).

$$\alpha_1^{new} y_1 + \alpha_2^{new} y_2 = constante = \alpha_1^{old} y_1 + \alpha_2^{old} y_2 \qquad (2.54)$$

Segundo Ales (2008), o algoritmo primeiramente encontra o valor para α_2^{new}, e utiliza-o para obter o valor de α_1^{new}. Para que os novos valores sejam viáveis para a resolução do problema deve-se respeitar as restrições: $0 \leq \alpha_1, \alpha_2 \leq C$, conforme FIGURA 28.

FIGURA 28 - LIMITAÇÕES DOS MULTIPLICADORES DE LAGRANGE



FONTE: PLATT (1998).

Podendo ser restrito pela condição da equação (2.54):

$$L \leq \alpha_2^{new} \leq H \qquad (2.55)$$

Esses limitantes dependem das classes a que os pontos pertencem, ou seja, se $y_1 \neq y_2$, então $L = max\{0, \alpha_2^{old} - \alpha_1^{old}\}$ e $H = min\{C, C - \alpha_1^{old} + \alpha_2^{old}\}$. Se $y_1 = y_2$, então $L = max\{0, \alpha_1^{old} + \alpha_1^{old} - C\}$ e $H = min\{C, \alpha_1^{old} + \alpha_2^{old}\}$.

Onde α_1^{new} é o novo valor do Multiplicador de Lagrange do ponto x^i e α_i^{old} é o valor anterior. O valor da função em x^i que denota a função atual determinada pelos valores dos Multiplicadores de Lagrange e por b no estágio atual da aprendizagem é dado por:

$$f(x^i) = \sum_{j=1}^{l} \alpha_j \, y_j K(x^j, x^i) + b \tag{2.56}$$

O valor E_i determina a diferença de $f(x^i)$ e o padrão y_i a que pertence o ponto x^i , ou seja, é a distancia do ponto ao hiperplano atual dado pela atualização dos Multiplicadores de Lagrange, é dado pela equação (2.56):

$$E_i = f(x^i) - y_i = \left[\sum_{j=1}^{l} \alpha_j \, y_j K(x^j, x^i) + b\right] - y_i \tag{2.57}$$

A quantidade adicional exigida é a segunda derivada da função objetivo ao longo da linha diagonal, que pode ser expressa por κ , definido por:

$$\kappa = K(x^1, x^1) + K(x^2, x^2) - 2K(x^1, x^2) = \|\emptyset(x^1) - \emptyset(x^2)\|^2 \tag{2.58}$$

onde x^1 e x^2 são os pontos associados a α^1 e α^2, respectivamente. O máximo valor da função objetivo será obtido com o valor pela equação (2.58):

$$\alpha_2^{new} \begin{cases} H, se\ \alpha_2^{aux} > H \\ \alpha_2^{aux}, se\ L \leq \alpha_2^{aux} \leq H \\ L, se\ \alpha_2^{aux} < L \end{cases} \tag{2.59}$$

Sendo α_2^{new} um valor truncado, ou seja, limitado por $L \leq \alpha_2^{aux} \leq H$, dado por:

$$\alpha_2^{aux} = \alpha_2^{old} + \frac{y_2(E_1 - E_2)}{\kappa} \tag{2.60}$$

Através do valor de α_2^{new}, encontra-se α_1^{new} dado pela equação (2.60):

$$\alpha_1^{new} = \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new}) \tag{2.61}$$

O SMO usa dois critérios para selecionar dois pontos ativos, ou seja, $0 < \alpha < C$, para garantir que a função objetivo aproveite um grande acréscimo, na atualização dos valores.

Têm-se duas heurísticas, uma para a escolha de α_1 e outra para escolha de α_2. Na primeira heurística, o ponto x^2 é escolhido entre os pontos que violam as condições de *Karush - KuhnTucker* (KKT). O algoritmo percorre todo o conjunto de dados de treinamento que violam as condições de KKT e seleciona um para atualizar, isso é feito através de um dos testes $E_2.y_2 < -tol$ e $\alpha_2 < C$ ou $E_2.y_2 > tol$ e $\alpha_2 > 0$. Quando tal ponto é encontrado, utiliza-se a segunda heurística para selecionar o ponto x^1, que deve ser escolhido de tal maneira que seja atualizado com x^2, causando uma grande mudança, que deve resultar em um grande acréscimo na função objetivo dual.

Para encontrar um bom ponto sem muitos cálculos, uma heurística rápida escolhe x^1, maximizando o valor dado por $|E_1 - E_2|$, se E_2 é positivo, o SMO escolhe o x^1 com o menor E_1, e se E_2 é negativo, então o SMO escolhe x^1 com o maior E_1.

Se esta escolha falhar em obter um acréscimo significante na função objetivo dual, o SMO experimenta cada ponto x^1 que tenha valores de α diferente dos limites, ou seja, $0 < \alpha < C$, começando aleatoriamente. Se ainda não houver progresso significativo, o algoritmo busca por todo o conjunto de dados de treinamento para encontrar um ponto x^1 adequado.

Se ocorrer alteração de valores, a heurística retorna para escolher outros pontos x^2 e x^1, até que todos os pontos estejam obedecendo as condições de KKT. Caso contrário termina o processo. A lista dos erros de todos os pontos de treinamento é mantida na memória para reduzir contas adicionais. A solução satisfaz as condições de complementaridade de Karush-Kuhn-Tucker que para classificar com a máxima margem entre os conjuntos deve obedecer $\alpha_i\left[y_i\left(\langle w.x^i\rangle + b\right) - 1\right] = 0$ para $i = 1, 2, \ldots, l$, ou seja, $\alpha_i\left[y_i f(x^i) - 1\right] = 0$.

Cristianini e Shawe-Taylor (2000) descrevem três critérios de parada, que são detalhados a seguir: monitoramento do valor da função, especificamente o valor do crescimento a cada passo, ou seja, o treinamento para quando a taxa de crescimento da função objetivo for menor que certa tolerância, por exemplo,

10^{-9}. Monitoramento das condições de KKT para o problema primal e por fim, outra maneira para caracterizar a solução é por meio do gap entre as funções objetivas: dual e primal, porém, esse critério de parada é válido apenas quando se tem um hiperplano linear.

Resolvendo o SMO, o valor de b , do vetor w e dos erros E_i são calculados separadamente. Após cada iteração, em que as condições de KKT são satisfeitas para ambos os pontos x^1 e x^2, os valores podem ser atualizados analisando sempre o valor atual com o anterior da função $f(x^i) = \sum_{j=1}^{l} \alpha_j\, y_j K(x^j, x^i) + b$.

O valor de b para o ponto x^1 é definido por b_1, o qual deve forçar a saída do SVM para y_1 quando a entrada for o ponto x^1, dado por:

$$b_1 = -\left[E_1 + y_1\left(\alpha_1^{new} - \alpha_1^{old}\right)K(x^1, x^1) + y_2\left(\alpha_2^{new} - \alpha_2^{old}\right)K(x^1, x^2) + b^{old}\right] \quad (2.62)$$

O valor de b para o ponto x^2 é definido por b_2, o qual deve forçar a saída do SVM para y_2 quando a entrada for o ponto x^2, dado pela equação (2.62).

$$b_2 = -\left[E_2 + y_1\left(\alpha_1^{new} - \alpha_1^{old}\right)K(x^1, x^2) + y_2\left(\alpha_2^{new} - \alpha_2^{old}\right)K(x^2, x^2) + b^{old}\right] \quad (2.63)$$

Se os valores b_1 e b_2 são iguais, então este será o novo valor de b, ou seja, $\mathrm{b}^{\mathrm{new}} = \mathrm{b}_1 = \mathrm{b}_2$. Caso contrário, o intervalo entre b_1 e b_2 são todos os thresholds que são consistentes com as condições de KKT, portanto, o SMO escolhe o *threshold* que está no meio do intervalo, ou seja:

$$b^{new} = \frac{b_1 + b_2}{2} \quad (2.64)$$

Quando os dados são separados linearmente, pode-se atualizar o valor do vetor w pela equação (2.64).

$$w^{new} = w^{old} + y_1(\alpha_1^{new} - \alpha_1^{old})\,\vec{x}^1 + y_2\,(\alpha_2^{new} - \alpha_2^{old})\,\vec{x}^2 \quad (2.65)$$

Os erros $\mathrm{E_i}$ são atualizados a cada iteração pela equação (2.65).

$$E_i^{new} = E_i^{old} + y_1\left(\alpha_1^{new} - \alpha_1^{old}\right)K\left(x^1, x^i\right) + y_2\left(\alpha_2^{new} - \alpha_2^{old}\right)K\left(x^2, x^i\right) + b^{new} - b^{old} \quad (2.66)$$

O objetivo da heurística SMO é obter os valores dos Multiplicadores de Lagrange para que estes tenham os erros tendendo à zero. A atualização da função objetivo pode ser feita pelo *gap*, que é a diferença entre a função objetivo atual e a anterior dada pela equação (2.66).

$$\begin{aligned} &\mathrm{gap}\left(\alpha_1^{\mathrm{new}} - \alpha_1^{\mathrm{old}}\right) + \left(\alpha_2^{\mathrm{new}} - \alpha_2^{\mathrm{old}}\right) - \left[\sum_{j=1}^{2}\sum_{i=1}^{1} y_j\, y_i\left(\alpha_j^{\mathrm{new}} - \alpha_j^{\mathrm{old}}\right)K\left(x^j, x^i\right)\right] + \\ &-\frac{1}{2}y_1^2\left(\alpha_1^{\mathrm{new}} - \alpha_1^{\mathrm{old}}\right)K\left(x^1, x^1\right) - \frac{1}{2}y_2^2\left(\alpha_2^{\mathrm{new}} - \alpha_2^{\mathrm{old}}\right)K\left(x^2, x^2\right) + \\ &-y_1 y_2\left[\alpha_1^{new}\alpha_2^{new} - \alpha_1^{old}\alpha_2^{old}\right]K(x^1, x^2) \end{aligned} \quad (2.67)$$

O SMO possui uma rotina de laços (*loops*), forçando a heurística a procurar por todos os pontos de treinamento os quais estão infringindo as condições de KKT.

3. METODOLOGIA

Neste capítulo, será descrita a metodologia de pesquisa utilizada neste estudo a fim de alcançar os objetivos propostos.

3.1 CLASSIFICAÇÃO DA PESQUISA

Quanto à natureza, esta pesquisa se caracteriza como aplicada, sendo realizada com o propósito de resolver um problema concreto, neste caso a mitigação de riscos de sinistros no transporte rodoviário de cargas.

Foi realizada uma pesquisa bibliográfica, na qual foram levantados os assuntos relacionados ao tema pesquisado, buscando evidenciar os aspectos relevantes à mineração de dados e aos atributos envolvidos nos sinistros de trânsito envolvendo o transporte rodoviário de cargas.

Para a coleta e análise de dados foram adotados os seguintes procedimentos:

- Levantamento dos dados de sinistros envolvendo o transporte rodoviário de cargas no Paraná utilizando a base dos dados da Polícia Rodoviária Federal;
- Levantamento bibliográfico para verificar quais atributos e fatores contribuem para a ocorrência de sinistros de trânsito envolvendo transporte rodoviário de cargas para seleção de atributos na base de dados da Polícia Rodoviária Federal;
- Tratamento dos dados de forma a adequá-los às ferramentas de mineração de dados;
- Realização de testes com as quatro técnicas de mineração de dados mais utilizadas constantes nos estudos com base na revisão bibliográfica;
- Avaliação do desempenho de cada técnica utilizada, comparando seus resultados para encontrar a que desempenha o melhor resultado;
- Verificação dos padrões de associação dos atributos dos sinistros de trânsito;
- Identificação dos fatores que mais contribuem para a incidência de sinistros envolvendo o transporte de cargas;

- Comparação dos dados de sinistros antes da pandemia e durante a pandemia de COVID-19;
- Apresentação dos resultados obtidos e interpretação dos resultados;

O QUADRO 2 apresenta uma síntese da coleta de dados, relacionando os objetivos específicos deste estudo com os instrumentos de coleta de dados correspondentes.

QUADRO 2 - SÍNTESE DOS OBJETIVOS ESPECÍFICOS RELACIONADO COM O INSTRUMENTO DE COLETA DE DADOS

Objetivo específico	Instrumento de coleta de dados
1 - Verificar na literatura quais são os principais atributos utilizados nos estudos de sinistros de trânsito;	Levantamento bibliográfico.
2 – Selecionar atributos para a base de dados, contendo as informações da Polícia Rodoviária Federal (PRF) e os principais atributos encontrados na literatura;	Levantamento bibliográfico e dados do banco de dados da PRF.
3 – Verificar na literatura quais os fatores que contribuem para a ocorrência dos sinistros de trânsito envolvendo o transporte rodoviário de cargas;	Levantamento bibliográfico.
4 - Verificar através do levantamento bibliográfico quais são as técnicas de mineração de dados mais utilizadas;	Levantamento bibliográfico.
5 - Utilizar técnicas e algoritmos de mineração de dados para realizar os experimentos com os atributos selecionados;	As técnicas mais utilizadas são as encontradas na revisão da literatura
6 - Avaliar o desempenho de cada técnica de mineração;	Calculo da acurácia, curva ROC e *Kappa*
7 - Verificar se houve influência da pandemia de COVID-19 nos sinistros de trânsito;	Comparação dos dados de 2017 a 2019 e de 2020 a 2021

FONTE: A Autora (2022).

3.2 IMPORTAÇÃO DOS DADOS

Os dados sobre sinistros de trânsito são consolidados e produzidos a partir de agentes públicos de trânsito. No estudo em questão, foram inseridos somente dados disponíveis do site da Polícia Rodoviária Federal. Na base de dados do Departamento da Polícia Rodoviária Federal foi realizado o *download* das planilhas em formato .CSV dos anos 2017, 2018, 2019, 2020 e 2021 (PRF, 2022). As planilhas contêm informações de sinistros de trânsito de todas as rodovias federais do Brasil, porém no estudo foi utilizados ao dados apenas do estado do Paraná.

3.2.1 Levantamento dos dados das variáveis envolvidas nos sinistros de trânsito

Os dados do banco de dados da Polícia Rodoviária Federal estão agrupados de duas formas: por ocorrência e por pessoa envolvida no sinistro: primeiramente por ocorrência com dados do período de 2007 à 2021, considerando cada ocorrência (sinistro) registrada pela PRF e segundo agrupados por pessoa envolvida no sinistro, para cada pessoa envolvida são registrados as informações do sinistro considerando todas as causas e tipos de sinistros com dados do período de 2017 e 2021. Entre os anos de 2007 e 2016, os registros de sinistros eram realizados por meio do sistema BR-Brasil, de modo que o policial responsável pela ocorrência inseria os dados referentes ao envolvidos, ao local, aos veículos e à dinâmica do sinistro. Em 2017, o sistema BR-Brasil foi descontinuado e a PRF passou a utilizar um novo sistema para registro das ocorrências de sinistros de trânsito.

Para esta pesquisa, foi realizada a junção desses dois bancos de dados (agrupado por ocorrência e por pessoa envolvida) por meio do número de identificação do sinistro, resultando em uma só planilha organizada segundo os atributos do acidente. Os dados utilizados são referentes ao estado do Paraná. Neste banco de dados, a Polícia Rodoviária Federal utiliza 35 variáveis para caracterizar o sinistro de trânsito descrito no QUADRO 3:

QUADRO 3 - DESCRIÇÃO DOS ATRIBUTOS UTILIZADOS PELA PRF (continua)

Nome da Variável	Descrição
Id	Variável com valores numéricos, representando o identificador do sinistro.
Pesid	Variável com valores numéricos, representando o identificador da pessoa envolvida.
Data_inversa	Data da ocorrência no formato dd/mm/aaaa.
Dia_semana	Dia da semana da ocorrência (segunda, terça, quarta, quinta, sexta, sábado ou domingo).
Horário	Horário da ocorrência no formato hh:mm:ss.
UF	Unidade da Federação. Ex.: MG, PE, DF, PR.
BR	Variável com valores numéricos, representando o identificador da BR do sinistro.
KM	Identificação do quilômetro onde ocorreu o sinistro, com valor mínimo de 0,1 km e com a casa decimal separada por ponto.
Município	Nome do município de ocorrência do sinistro.
Causa_sinistro	Causa presumível do sinistro, baseada nos vestígios, indícios e provas colhidas no local do sinistro.
Ordem_tipo_sinistro	Valor numérico que identifica a sequência dos eventos sucessivos que ocorreram no sinistro.
Tipo_sinistro	Identificação do tipo do sinistro. Ex. colisão frontal, colisão traseira.
Classificação_sinistro	Classificação quanto à gravidade do sinistro: sem vítimas, com vítimas feridas, com vítimas fatais e ignorado.
Fase_dia	Fase do dia no momento do sinistro: amanhecer, anoitecer, pleno dia e plena Noite.
Sentido_via	Sentido da via considerando o ponto de colisão: crescente e decrescente.
Condição_meteorológica	Condição meteorológica no momento do sinistro: céu claro, chuva, garoa/chuvisco, granizo, ignorado, neve, nevoeiro/neblina, nublado, sol e vento.

QUADRO 3 - DESCRIÇÃO DOS ATRIBUTOS UTILIZADOS PELA PRF (conclusão)

Nome da Variável	Descrição
Tipo_pista	Tipo de pista considerando a quantidade de faixas: simples, dupla e múltipla.
Traçado_via	Descrição do traçado da via: curva, desvio temporário, intersecção de vias, não informado, ponte, reta, retorno regulamentado, rotatória, túnel e viaduto.
Uso_solo	Descrição sobre as características do local do sinistro: Urbano= Sim e Rural = não.
Id_veículo	Variável com valores numéricos, representando o identificador do veículo envolvido.
Tipo_veículo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro: automóvel, bicicleta, caminhão, caminhão-trator, caminhonete, camioneta, carro de mão, carroça-charrete, ciclomotor, micro-ônibus, motocicleta, motoneta, não informado, ônibus, quadriciclo, reboque, semirreboque, trator de esteira, trator de rodas, trator misto, trem-bonde, triciclo, utilitário e outros.
Marca	Descrição da marca do veículo.
Ano_fabricação_veículo	Ano de fabricação do veículo no formato aaaa.
Tipo_envolvido	Tipo de envolvido no sinistro conforme sua participação no evento: cavaleiro, condutor, não informado, passageiro, pedestre e testemunha.
Estado_físico	Condição do envolvido conforme a gravidade das lesões: ileso, lesões leves, lesões graves e óbito.
Idade	Idade do envolvido.
Sexo	Sexo do envolvido.
Ilesos	Valor binário (quantidade) que identifica se o envolvido foi classificado como ileso/ Número de pessoal ilesas no sinistro.
Feridos_leves	Valor binário (quantidade) que identifica se o envolvido foi classificado como ferido leve/ Número de feridos leves no sinistro.
Feridos_graves	Valor binário (quantidade) que identifica se o envolvido foi classificado como ferido grave/ Número de feridos graves no sinistro.
Mortos	Valor binário (quantidade) que identifica se o envolvido foi classificado como morto/ Número de óbitos no sinistro.
Latitude	Latitude do local do sinistro em formato geodésico decimal.
Longitude	Longitude do local do sinistro em formato geodésico decimal.
Pessoas	Total de pessoas envolvidas na ocorrência.
Veículos	Total de veículos envolvidos na ocorrência.

Fonte: Dicionário das variáveis da Polícia Rodoviária Federal (2017).

Neste estudo, foram selecionados os atributos mais relevantes para o entendimento do problema da sinistralidade viária segundo o referencial teórico pesquisado, conforme descrito na seção a seguir.

3.2.2 Tratamento dos dados e seleção dos atributos

Simultaneamente às atividades de extração e importação foi realizado o tratamento dos dados ou pré-processamento. Nesta fase foram corrigidas

inconsistências, pois os conteúdos dos atributos podiam estar incompletos, redundantes, ruidosos ou esparsos (COSTA *et al.*, 2019).

Adicionalmente, a seleção de atributos também pode ser uma técnica utilizada com o intuito de reduzir a quantidade dos dados, facilitando a aplicação de algoritmos de mineração. Esta redução visa eliminar atributos que não agregam informações para a análise, produzindo assim uma representação mais compacta, mais facilmente interpretável do objetivo a ser alcançado, direcionando a atenção do usuário sobre os atributos mais relevantes (WITTEN; FRANK; HALL, 2011).

Existe uma grande quantidade de informações sobre sinistros de trânsito extraídas do banco de dados da PRF e algumas das variáveis podem ocultar ou confundir o efeito de outras mais significativas. Portanto, pesquisadores na área de sinistros de trânsito, especificamente no domínio da gravidade de sinistros de trânsito, concentraram suas pesquisas na tentativa de identificar as variáveis mais significativas que contribuem para a ocorrência de uma gravidade específica de lesão em um sinistro de trânsito (MUJALLI ; OÑA, 2011).

A natureza dos atributos da base da PRF mostrou que era necessário reduzir sua quantidade. A seleção dos atributos foi realizada com base no referencial teórico, no qual foram apontados os atributos mais utilizados nas pesquisas dos últimos 22 anos, utilizando o critério da relevância que o atributo poderia ter na mineração dos dados.

Inicialmente, foram considerados os atributos escritos na forma de código, sem a descrição completa do atributo nas análises, devido ao nome ser extenso.

O atributo “causa_acidente” foi retirado por descrever um possível comportamento do condutor que causou o acidente a partir de evidências disponíveis e/ou suposições, sem um critério objetivo. Possíveis valores da causa do acidente são: falta de atenção do motorista, direção sob efeito do álcool, ultrapassagem indevida, entre outras. Quanto à causa do acidente, as categorias “falta de atenção” e “outras” totalizaram mais de 50% das causas, o que levou à hipótese de que o registro desse atributo tenha um elevado nível de subjetividade, concentrando possivelmente registros cuja causa dos acidentes não pôde ser identificada com precisão pelo policial rodoviário (BARROSO JUNIOR; BERTHO; VEIGA, 2019).

Foram retirados também os seguintes atributos: Id (identificação do acidente), Pesid (identificação da pessoa envolvida), ordem_tipo_sinistro (ordem do sinistro), Id_veículo (identificação do veículo), marca do veículo, latitude, longitude, UF, BR, Km, município e sentido da via.

A identificação do acidente consiste em um atributo para fins de registro, de modo que foi utilizado apenas para combinar os atributos contidos nas bases agrupadas por ocorrência e por pessoa envolvida. Do mesmo modo, a identificação da pessoa corresponde a um atributo para fins de registro de cada pessoa envolvida no sinistro. O atributo ordem do sinistro foi retirado por representar uma situação de sinistros sucessivos (por exemplo, uma colisão frontal seguida de uma colisão traseira), que representa apenas 6% das situações. Os atributos de identificação e marca do veículo foram retirados tendo em vista que o primeiro é utilizado apenas para fins de registro do veículo e o segundo está fora do escopo deste trabalho, pois não será conduzida nenhuma análise por marca do veículo. Os atributos latitude, longitude, UF, BR, km, município e sentido da via não foram utilizados tendo em vista que o escopo do trabalho não inclui análises geográficas, a fim de manter uma amostra agregada e representativa de todo o território nacional para a identificação dos padrões dos sinistros envolvendo veículos pesados de carga.

No caso do atributo tipo do veículo, foi realizada uma filtragem, sendo que não foram utilizados os sinistros que não apresentavam pelo menos um veículo de carga envolvido, assim, foram selecionados os veículos de transporte de carga pesada com a seguinte participação nos sinistros: caminhão-trator (51,14%), caminhão (48,15%), semirreboque (0,51%) e reboque (0,20%).

Na França, as bases de dados de sinistros de trânsito do Observatório Nacional Interministerial de Segurança Viária (ONISR), contém as mesmas variáveis do banco de dados da PRF e as seguintes variáveis: indicação de via com ciclovia, largura da via utilizada para tráfego de veículos, condição da superfície/pista, proximidade de escola, ponto de colisão, manobra principal antes do sinistro, o espaço ocupado no veículo pelo usuário no momento do sinistro, motivo da viagem no momento do sinistro, existência de equipamento de segurança, uso do equipamento de segurança, localização do pedestre, ação do pedestre, o pedestre acidentado estava sozinho ou não.

Em comparação, nos Estados Unidos, os dados de sinistros de trânsito da *Federal Motor Carrier Safety Administration* (FMCSA) do Departamento de Transportes dos Estados Unidos são parecidos com os dados da PRF e contêm ainda as seguintes variáveis: tipo de carroceria do caminhão, tipo de carga e classificação do peso do caminhão. Isso mostra que o banco de dados da PRF poderia incluir mais dados que são considerados importantes nos sinistros, como por exemplo, o uso de cinto de segurança, teste de embriaguez ao volante, iluminação da estrada, limite de velocidade na via e classificação do peso do caminhão.

Do conjunto formado pelos 35 atributos originais importados, foram considerados relevantes para a identificação e análise dos padrões de sinistros de trânsito 21 atributos, conforme os estudos apontados no referencial teórico nos 12 estudos: idade do condutor, condição meteorológica, tipo do veículo envolvido, tipo do envolvido, tipo do acidente, tipo de pista, fase do dia, sexo do condutor, dia da semana, traçado da via, hora, tipo de solo, ano veículo, classificação do sinistro, estado físico dos envolvidos, número de envolvidos e número de veículos envolvidos, data, além de outros 3 atributos como número de mortos, feridos leves e feridos graves, descritos no QUADRO 4.

QUADRO 4 - ATRIBUTOS SELECIONADOS PARA A MINERAÇÃO

Atributo	Descrição dos Atributos

1	Data_inversa	Data da ocorrência no formato dd/mm/aaaa., porém para o estudo foi considerado o mês de ocorrência.
2	Dia_semana	Dia da semana da ocorrência (segunda, terça, quarta, quinta, sexta, sábado ou domingo).
3	Horário	Horário da ocorrência no formato hh:mm:ss., porém foi agregado em períodos de horários.
4	Tipo_sinistro	Identificação do tipo do sinistro. Ex. colisão frontal, colisão traseira.
5	Classificação_sinistro	Classificação quanto à gravidade do sinistro: sem vítimas, com vítimas feridas e com vítimas fatais.
6	Fase_dia	Fase do dia no momento do sinistro: amanhecer, anoitecer, pleno dia e plena noite.
7	Condição_meteorológica	Condição meteorológica no momento do sinistro: céu claro, chuva, garoa/chuvisco, granizo, ignorado, neve, nevoeiro/neblina, nublado, sol e vento.
8	Tipo_pista	Tipo de pista considerando a quantidade de faixas: simples, dupla e múltipla.
9	Traçado_via	Descrição do traçado da via: curva, desvio temporário, intersecção de vias, ponte, reta, retorno regulamentado, rotatória, túnel e viaduto.
10	Uso_solo	Descrição sobre as características do local do sinistro: Urbano= sim e Rural = não.
11	Pessoas	Total de pessoas envolvidas na ocorrência.
12	Mortos	Número de pessoas mortas no sinistro
13	Feridos_leves	Número de feridos leves no sinistro
14	Feridos_graves	Número de feridos graves no sinistro
15	Veículos_envolvidos	Número de veículos envolvidos no sinistro
16	Tipo_envolvido	Tipo de envolvido no sinistro conforme sua participação no evento: cavaleiro, condutor, passageiro, pedestre e testemunha.
17	Estado_físico	Condição do envolvido conforme a gravidade das lesões: ileso, lesões leves, lesões graves e óbito.
18	Idade	Idade do envolvido.
19	Sexo	Sexo do envolvido.
20	Tipo_veículo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro: automóvel, bicicleta, caminhão, caminhão-trator, caminhonete, camioneta, carro de mão, carroça-charrete, ciclomotor, micro-ônibus, motocicleta, motoneta, não informado, ônibus, quadriciclo, reboque, semirreboque, trator de esteira, trator de rodas, trator misto, trem-bonde, triciclo, utilitário e outros.
21	Ano_fabricação_veículo	Ano que foi fabricado o veículo envolvido no sinistro

FONTE: A Autora (2021).

Um sinistro de trânsito pode ser analisado considerando-se três momentos, o pré-sinistro, o durante sinistro e o pós-sinistro, assim, para a realização da investigação dos fatores determinantes e consequências do referido evento destaca-se a Matriz de Haddon (HADDON JR, 1980). Nos estudos de Ferraz *et al.* (2012) e Panitz (1999) são relacionadas as principais ações associadas a cada um dos três elementos que compõem o sistema de trânsito, no sentido de evitar os sinistros (período pré-sinistro), de minimizar as consequências dos sinistros no instante em que ocorrem (momento do sinistro)

e de minimizar os efeitos após os sinistros (período pós-sinistro, de atendimento às vítimas e tratamento médico-hospitalar). Uma versão adaptada dessa matriz é mostrada no QUADRO 5, na qual estão relacionados os atributos pré-sinistro, durante sinistro e pós-sinistro, que compõem o banco de dados da PRF relacionados com os fatores humanos, veículo e com relação ao ambiente.

QUADRO 5 - ADAPTAÇÃO DA MATRIZ DE *HADDON* APLICADA A SINISTROS DE TRÂNSITO

		Fatores		
		Humano	Veículo	Ambiente
Fase do evento	Pré-sinistro	Pessoas envolvidas no sinistro (tipo do envolvido, idade e sexo)	Veículos envolvidos no sinistro (tipo do veículo e ano de fabricação do veículo)	Infraestrutura (tipo de pista, traçado da via e uso do solo). Condição climática e condição do dia (data, horário, dia da semana, hora, fase do dia e condição meteorológica)
	Durante sinistro			Tipologia (tipo do sinistro)
	Pós-sinistro	Pessoas (classificação do sinistro e estado físico)		

FONTE: A Autora (2021).

Conforme a escolha dos atributos foi obtida uma análise exploratória par a par das variáveis, sendo observado: quais variáveis do pré-sinistro apresentam maior influência sobre a variável do durante sinistro; quais variáveis do durante sinistro apresentam maior influência sobre a variável do pós-sinistro e quais variáveis do pré-sinistro apresentam maior influência sobre a variável do pós-sinistro. E com o objetivo de apoio a decisão, por exemplo, é possível verificar quais combinações de fatores do pré e durante sinistro são capazes de produzir sinistros mais graves ou qual combinação de fatores do pré-sinistro são capazes de produzir determinados tipos de sinistros.

Após o tratamento da base dos dados a ser utilizada para a mineração, iniciou-se a fase de escolha das opções de classificação de dados. Há um painel no WEKA chamado *Test options* (Opções de Teste), onde é possível escolher algumas configurações para o classificador que será utilizado. Estas opções determinam pontos importantes de como será o comportamento do algoritmo e

de como a base de dados será testada. Existem quatro modos de teste: *Use training set, Supplied test set, cross-validation e Percentage split.*

Neste trabalho será utilizado o *Cross-validation,* a estimativa de validação cruzada é uma forma de avaliar como o modelo se comporta diante de variações nas amostras de treinamento, ajudando a evitar um dos problemas da influência da divisão dos dados na métrica. Os testes são feitos com dados que o modelo não viu anteriormente (BOUCKAERT et al., 2020).

Este modo de teste realiza um laço de repetição de *i* iterações, sendo *i* o número de *folds* (número de pares de subconjuntos treinamento-teste) fornecido como entrada. A cada iteração deste laço é criado um subconjunto de treinamento, e sobre este, é aplicado o algoritmo de classificação previamente escolhido. Além disso, é criado também um subconjunto de teste. Existe ainda outro laço de repetição dentro do anterior que repete *j* vezes, sendo *j* o número de instâncias do subconjunto de teste criado. A cada iteração do segundo laço, é processada uma predição do classificador sobre a instância de teste corrente. Em outras palavras, o conhecimento obtido é testado em cada instância do conjunto de teste e os resultados de cada um destes testes alimentam uma série de dados estatísticos, dentre os quais se destaca a acurácia do conhecimento obtido. Com objetivo de uma avaliação mais precisa, por meio do desvio padrão, foram realizados 10 (dez) testes.

Os classificadores no WEKA foram desenvolvidos para prever uma única "classe" atributo, que é utilizada para a predição. Por padrão, o atributo classe é considerado como sendo o último atributo nos dados. Caso seja treinado um classificador para prever um atributo diferente, na caixa abaixo da caixa de opções de teste terá uma lista suspensa de atributos para escolher. Neste estudo, os atributos classe utilizados para classificação são o tipo do acidente, o estado físico das vítimas e a classificação do acidente, sendo variáveis durante e pós-sinistro. Conforme QUADRO 5, foram analisados os sinistros do pré-sinistro com o durante sinistro, os dados do pré-sinistro e pós-sinistro e os dados do durante sinistro com pós-sinistro. O *software Microsoft Excel* foi utilizado para avaliação da metodologia utilizada por meio de estatística descritiva elaborada a partir dos atributos identificados como mais importantes no processo de mineração de dados aplicado.

Posteriormente será realizada a análise dos sinistros de trânsito no Paraná, especificamente em 2020, ano que marca o início da pandemia da COVID-19, ocasião em que decretos governamentais instituíram medidas de "*lockdown*" no estado e o ano de 2021, e compará-los com anos anteriores (2017 2018 e 2019). Pretende-se, pois, verificar se houve alguma influência ou impacto da pandemia na classificação dos sinistros envolvendo o transporte rodoviário de cargas.

4 RESULTADOS

Primeiramente são apresentados os resultados dos testes considerando a relação entre as etapas “pré-sinistro x durante sinistro”, posteriormente a análise entre as etapas do “pré-sinistro x pós-sinistro” e, finalmente, o estudo da relação entre as etapas” durante sinistro x pós-sinistro”. Finalizando, foram realizados testes considerando o período anterior à pandemia de COVID-19 (2017, 2018 e 2019) e os dados do período durante pandemia (2020 e 2021).

4.1 RESULTADOS PRÉ-SINISTRO X DURANTE SINISTRO

Os resultados obtidos por meio dos testes do pré-sinistro x durante sinistro foram organizados nas TABELAS 2, 3, 4 e 5 de acordo com o melhor e o pior resultado de cada algoritmo de classificação considerando o número de instâncias (ocorrências de sinistros de trânsito envolvendo transporte rodoviário e cargas) classificadas corretamente, curva ROC, estatística *Kappa* e a acurácia.

TABELA 2 - RESULTADOS UTILIZANDO ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO OS TIPOS DOS SINISTROS ATROPELAMENTOS.

Algoritmo	**Número de instâncias**	**Instâncias classificadas corretamente**	**Curva ROC**	**Estatística *Kappa***	**Acurácia**	**Tempo de execução (s)**
SMO	1501	1455	0,732	0,8351	96,93%	0,15
Naive Bayes	1501	1452	0,934	0,8179	96,73%	0,02
Multilayer Perceptron	1501	1451	0,933	0,8027	96,66%	21,96
J48	1501	1445	0,856	0,7662	96,26%	0,01

FONTE: A autora (2022).

Considerando-se os experimentos realizados nesta pesquisa, foram selecionados aqueles que apresentaram as melhores medidas de desempenho,

sendo que os principais critérios de seleção foram a Curva ROC e o coeficiente de concordância *Kappa*, pois são medidas de desempenho com elevadas precisão (habilidade do modelo em prever corretamente as classes) e robustez (habilidade do modelo para avaliar ou prever corretamente, utilizando dados ruidosos e com viés).

Observou-se na TABELA 2 que os maiores coeficientes *Kappa* foram encontrados nos testes com o algoritmo de Máquina de Vetores de Suporte (SMO) e de Naive Bayes, porém em todos os testes realizados e em todos os parâmetros o *Kappa* obteve classificação acima de 0,75, aceitável segundo Landis e Koch (1977) com uma concordância substancial. Em relação à Curva ROC, todos os algoritmos apresentaram discriminação aceitável conforme apontado nos intervalos descritos no estudo de Tharwat (2021).

Os tipos de sinistro foram englobados em quatro grandes grupos: atropelamento (representando os tipos de sinistros atropelamento de animal e de pedestre), colisões (correspondendo a colisão frontal, colisão lateral, colisão com objeto em movimento, colisão traseira, colisão transversal e engavetamento), saída de pista (englobando os sinistros por capotamento, colisão com objeto estático, queda de ocupante do veículo, saída de leito carroçável e tombamento) e outros (derramamento de carga, danos eventuais, eventos atípicos e incêndio). Primeiramente foi analisado o tipo de sinistro atropelamento, em seguida os outros tipos de sinistro.

Considerando o tipo do sinistro atropelamento, que engloba atropelamento de animais e de pedestres, para a classificação apresentada na TABELA 2, o algoritmo de Máquina de Vetores de Suporte (SMO) apresentou-se como um classificador de dados satisfatório, os testes realizados trouxeram resultados positivos, além de ter apresentado agilidade na execução.

Considerando o tipo de sinistro atropelamento de animal, conforme QUADRO 6, utilizando os algoritmos *Multilayer Perceptron*, J48, *Naive Bayes* e SMO, os fatores com relação à via e ao ambiente apresentaram um impacto maior neste tipo de sinistro, onde o traçado da via (fator via) foi apontado nos algoritmos *Naive Bayes*, apresentando uma probabilidade maior de ocorrência, e no algoritmo SMO. Já o uso do solo (fator via) se mostrou preponderante no algoritmo *Multilayer Perceptron* e J48. Com relação ao fator ambiente, a condição meteorológica no momento da ocorrência do sinistro se mostrou

preponderante em relação aos outros atributos considerados, principalmente utilizando os algoritmos *Naive Bayes* e SMO.

QUADRO 6 - RESULTADO DOS ALGORITMOS CONSIDERANDO OS TIPOS DE SINISTROS ATROPELAMENTO DE ANIMAIS E DE PEDESTRES.

			Atributo de "resultado" Durante							
		Algoritmos utilizados	**Multilayer Perceptron**		**J48**		**Naive Bayes**		**SMO**	
Atributo de entrada Pré	**Domínio do atributo**	**Atributo**	AA	AP	AA	AP	AA	AP	AA	AP
	Fator humano	Sexo								
		Idade								
	Fator veículo	Tipo veículo								
		Ano de fabricação			X	X				
	Fator via	Tipo de pista				X				
		Traçado da via		**X**			**X**	**X**	**X**	**X**
		Uso do solo	X		X	X				
	Fator ambiente	Horário								
		Dia da semana	X							
		Fase do dia			X	X				
		Condição meteorológica					**X**	**X**	**X**	**X**
		Data (mês)		X			X	X		

Legenda: AA (Atropelamento de animal), AP (Atropelamento de pedestre)
Fonte: A autora (2022).

Utilizando estatística descritiva para validação dos resultados, considerando o fator veículo, a análise resultou em 32 atropelamentos de animal (47,05%) envolvendo veículos com idade de frota de 12 a 22 anos, sendo a média 18,40 anos. No fator via, com relação ao traçado da via, 42 atropelamentos de animal, ou seja, 61,76% ocorreram em uma reta. Outro atributo analisado em relação à via foi o uso do solo, para o qual 63 atropelamentos de animal (92,64%) aconteceram em solo rural (fora da área urbana). Considerando o fator ambiente, verificou-se que 40 atropelamentos de animal (58,82%) ocorreram em plena noite entre o período das 23h00min às 03h00min. Já considerando a condição meteorológica, 39 atropelamentos de

animal (57,35%) ocorreram com céu claro e 22 (32,35%) com o céu nublado. Outubro foi o mês com maior ocorrência de atropelamentos de animal, com 25% das ocorrências, e 57,35% dos sinistros envolvendo atropelamento animal ocorreram no final de semana.

O mesmo ocorreu considerando o sinistro atropelamento de pedestre, porém considerando o fator via, o traçado apareceu com um peso significativo nos algoritmos *Multilayer Perceptron*, *Naive Bayes* e SMO. Já no fator ambiente, a condição meteorológica foi indicada como preponderante nos algoritmos *Naive Bayes* e SMO, e o mês que ocorreu este tipo de sinistro foi apontado como preponderante nos algoritmos *Multilayer Perceptron* e *Naive Bayes*. Assim, os fatores via e ambiente apresentaram um maior impacto nos tipos de sinistro atropelamento de animal e de pedestre.

Em análise complementar, utilizando estatística descritiva, dos resultados relacionados aos 1.433 atropelamentos de pedestre, a idade da frota (de caminhões) apresentou-se entre 2 a 12 anos, com média de 6,20 anos, representando 48,15% dos casos. Considerando o fator via, o traçado da via indicou que 976 (68,11%) sinistros ocorreram em uma reta. Outro atributo analisado, o tipo de pista, mostrou que a maioria dos sinistros envolvendo atropelamento de pedestre ocorreu em pista dupla 877 (61,20%) e 467 (32,59%) em pista simples. Já considerando o uso do solo houve uma divisão de 724 (50,52%) sinistros ocorreram em solo rural e 709 (49,48%) em solo urbano.

Verificou-se no fator ambiente que o atributo fase do dia apresentou 723 atropelamentos de pedestre que ocorreram em plena noite (50,45%), entre 23h00min às 05h00min. A condição meteorológica apontou 798 (55,69%) atropelamentos de pedestre com céu claro e 346 (24,14%) ocorreram com céu nublado. No mês de julho ocorreram mais atropelamentos de pedestres (13,96%) comparados aos outros meses.

A síntese dos resultados referentes aos atropelamentos pode ser verificada no QUADRO 7.

QUADRO 7 - SÍNTESE DOS RESULTADOS CONSIDERANDO OS TIPOS DE ACIDENTE ATROPELAMENTO DE ANIMAIS E DE PEDESTRES.

Atributos	AA	AP

Ano de fabricação	Idade da frota de 12 a 22 anos	Idade da frota de 2 a 12 anos
Tipo de pista	-	Maioria ocorreu em pista dupla
Traçado da via	Maioria ocorreu em uma reta	Maioria ocorreu em uma reta
Uso do solo	Maioria em solo rural	Aproximadamente metade em solo rural e metade em solo urbano
Dia da semana	Maioria ocorreu no domingo	-
Fase do dia	Maioria ocorreu em plena noite	Maioria ocorreu em plena noite
Condição meteorológica	Maioria ocorreu com céu claro	Maioria ocorreu com céu claro
Data (mês)	Outubro com maior ocorrência	Julho com maior ocorrência

Legenda: AA (Atropelamento de animal), AP (Atropelamento de pedestre)
Fonte: A autora (2022).

Considerando o tipo de sinistro colisões, que engloba os sinistros: colisão frontal, colisão lateral, colisão com objeto em movimento, colisão traseira, colisão transversal e engavetamento, observou-se na TABELA 3 que o maior coeficiente *Kappa* foi encontrado no teste com o algoritmo J48 (concordância substancial de acordo com o estudo de Landis e Koch (1977)), além da Curva ROC ter discriminação excelente (acima de 0,90) conforme estudo de Hosmer e Lemeshow (2013). Os testes considerando os outros algoritmos não foram satisfatórios, abaixo de 60% de acurácia e estatística *Kappa* apresentando discordância de informações conforme intervalos definidos no estudo de Landis e Koch (1977).

TABELA 3 - RESULTADOS DOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO OS TIPOS DOS SINISTROS ENVOLVENDO COLISÕES.

Algoritmo	Número de instâncias	Instâncias classificadas corretamente	Curva ROC	Estatística *Kappa*	Acurácia	Tempo de execução (s)
J48	25995	18953	0,914	0,6283	72,91%	0,41
Multilayer Perceptron	25995	14732	0,780	0,3881	56,67%	375,08
Naive Bayes	25995	12159	0,700	0,2058	46,77%	0,01
SMO	25995	11032	0,639	0,1058	42,43%	2505,73

FONTE: A autora (2022).

Considerando o tipo do sinistro colisões, para a classificação apresentada na TABELA 3, o algoritmo de J48 apresentou-se como um classificador de dados satisfatório, o teste realizado trouxe resultado aceitável, com mais de 72% de acurácia, além de ter apresentado agilidade na execução.

Conforme QUADRO 8, considerando os tipos de sinistro colisão e o algoritmo que apresentou melhor resultado, o J48, os fatores com relação à via e ao ambiente apresentaram um impacto maior nos tipos de sinistro colisão frontal, colisão lateral, colisão com objeto em movimento, colisão traseira, colisão transversal e engavetamento, onde o tipo de pista e o traçado da via (fator via) apresentou um peso maior no algoritmo J48, apresentando uma probabilidade maior de ocorrência. Já o uso do solo (fator via) se mostrou preponderante nos sinistros envolvendo colisão lateral e colisão traseira. Com relação ao fator ambiente, o horário no momento da ocorrência do sinistro se mostrou preponderante em relação aos outros atributos considerados. Já o atributo dia da semana apareceu com um peso significativo nos sinistros envolvendo colisão frontal, colisão traseira e transversal. A condição meteorológica e a fase do dia se mostraram preponderantes nos sinistros envolvendo colisão lateral. Assim, os fatores via e ambiente apresentaram um maior impacto nos tipos de sinistros classificados como colisões.

A partir de análise adicional no editor de planilhas, verificou-se 2.414 colisões frontais, com relação ao fator humano, atributo idade, 1.993 (82,56%) sinistros envolvendo colisões frontais ocorreram com condutores de faixa etária

de 25 a 59 anos conforme classificação de faixa etária da OMS (CARVALHO; PEDROSA, 2015). Considerando o fator via, o traçado da via indicou que 1.186 (49,13%) colisões frontais ocorreram em uma reta e 648 (26,84%) em uma curva. Outro atributo analisado, o tipo de pista, mostrou que a maioria das colisões frontais ocorreu em pista simples 2.145 (88,85%).

No fator ambiente, o atributo fase do dia apresentou 1.297 colisões frontais que ocorreram em pleno dia (53,73%), entre 05h00min e 16h00min. O período da semana apontado com maior probabilidade de colisões frontais foi no final de semana, com 26,84%.

Analisando as colisões laterais, houve 4.268 sinistros. Considerando o fator via, o traçado da via indicou que 2.314 (54,21%) colisões laterais ocorreram em uma reta e 747 (17,50%) em uma curva. Para outro atributo analisado, o tipo de pista, tem-se que a maioria das colisões laterais ocorreu em pista dupla 2.261 (52,97%). Já considerando o uso do solo, 2.824 (66,16%) colisões laterais ocorreram em solo rural.

No fator ambiental, o atributo fase do dia apresentou 2.662 colisões laterais que ocorreram em pleno dia (62,37%), entre 05h00min às 16h00min. A condição meteorológica apontou 2.297 (53,82%) colisões laterais com céu claro e 771 (18,06%) ocorreram com céu nublado.

Verificou-se nas 278 colisões com objeto em movimento, considerando o fator via, o traçado da via indicou que 116 (41,72%) colisões com objeto em movimento ocorreram em uma reta e 98 (35,25%) em uma curva. Para o atributo tipo de pista, tem-se que a maioria das colisões com objeto em movimento ocorreu em pista dupla 145 (52,15%). Já no fator ambiente, o atributo fase do dia apresentou 176 colisões com objeto em movimento que ocorreram em pleno dia (63,31%), entre 05h00min às 16h00min.

Analisando as colisões traseiras, houve 6.604 sinistros, para os quais a idade da frota predominante foi de 1 a 12 anos, com média de 7,30 anos, com 50,19% dos caminhões nesta categoria. Considerando o fator via, o traçado da via indicou que 3.758 (56,90%) colisões traseiras ocorreram em uma reta. Para o atributo tipo de pista, tem-se que a maioria dos sinistros envolvendo colisão traseira ocorreu em pista dupla 4.175 (63,22%). Já considerando o uso do solo, 4.592 (69,53%) colisões traseiras ocorreram em solo rural.

No fator ambiente, o atributo fase do dia apresentou 3.769 colisões traseiras que ocorreram em pleno dia (57,07%), entre 05h00min às 16h00min. O dia da semana indicado com maior probabilidade de colisões traseiras foi na quinta-feira com 21,05%.

Considerando as colisões transversais, houve 2.152 sinistros, considerando o fator via, o traçado da via indicou que 967 (44,93%) colisões transversais ocorreram em uma reta e 389 (18,08%) em intersecção de vias. Para outro atributo analisado, o tipo de pista, a maioria das colisões transversais ocorreu em pista simples 1.305 (60,64%).

No fator ambiente, o atributo fase do dia apresentou 1.214 colisões transversais que ocorreram em pleno dia (56,41%), entre 05h00min às 16h00min. O período da semana apresentado com maior probabilidade de colisões transversais foi no final de semana com 23,42%.

Analisando as colisões por engavetamento, houve 882 sinistros, considerando o fator via, o traçado da via indicou que 352 (39,90%) sinistros ocorreram em uma reta e 161 (18,25%) em desvio temporário. Outro atributo analisado, o tipo de pista, mostrou que a maioria dos sinistros envolvendo engavetamento ocorreu em pista dupla 569 (64,51%).

Considerando o fator ambiente, o atributo fase do dia apresentou 3769 sinistros que ocorreram em pleno dia (51,11%), entre 05h00min às 16h00min.

QUADRO 8 - RESULTADO DOS ALGORITMOS UTILIZADOS CONSIDERANDO OS TIPOS DE ACIDENTE COLISÃO FRONTAL, COLISÃO LATERAL, COLISÃO COM OBJETO EM MOVIMENTO, COLISÃO TRASEIRA, COLISÃO TRANSVERSAL E ENGAVETAMENTO.

		Atributo de "resultado" Durante																								
		Algoritmos utilizados	**Multilayer Perceptron**						**J48**						**Naive Bayes**						**SMO**					
Atributo de entrada Pré	**Domínio do atributo**	**Atributo**	CF	CL	COM	CT	CTR	E	CF	CL	COM	CT	CTR	E	CF	CL	COM	CT	CTR	E	CF	CL	COM	CT	CTR	E
	Fator humano	Sexo																								
		Idade							X																	
		Tipo envolvido					X									X	X				X		X		X	X
	Fator veículo	Tipo veículo																								
		Ano de fabricação	X		X							X														
	Fator via	Tipo de pista		X		X			**X**	**X**	**X**	**X**	**X**	**X**							X		X			X
		Traçado da via			X		X	X	**X**	**X**	**X**	**X**	**X**	**X**	X	X	X	X	X	X	X	X	X	X	X	X
		Uso do solo			X					X		X														
	Fator ambiente	Horário							**X**	**X**	**X**	**X**	**X**	**X**												
		Dia da semana			X				X			X	X													
		Fase do dia								X																
		Condição meteorológica	X	X		X		X		X					X	X	X	X	X	X	X	X				
		Data (mês)	X			X	X																X			

Legenda: CF (Colisão frontal), CL (Colisão lateral), COM (Colisão objeto em movimento), CT (Colisão traseira), CTR (Colisão transversal), E (Engavetamento).
Fonte: A autora (2022).

A síntese dos resultados referentes às colisões pode ser verificada no QUADRO 9.

QUADRO 9 - SÍNTESE DOS RESULTADOS CONSIDERANDO OS TIPOS DE ACIDENTE COLISÕES.

Atributos	CF	CL	COM	CT	CTR	E
Idade	Maioria ocorreu com condutores de faixa etária de 25 a 59 anos	-	-	-	-	-
Ano de fabricação	-	-	-	A idade da frota neste tipo de colisão é de 1 a 12 anos	-	-
Tipo de pista	Maioria ocorreu em pista simples	Maioria ocorreu em pista dupla	Maioria ocorreu em pista dupla	Maioria ocorreu em pista dupla	Maioria ocorreu em pista simples	Maioria ocorreu em pista dupla
Traçado da via	Maioria ocorreu em uma reta e curva	Maioria ocorreu em uma reta e curva	Maioria ocorreu em uma reta e curva	Maioria ocorreu em uma reta	Maioria ocorreu em uma reta e intersecção de vias	Maioria ocorreu em uma reta e desvio temporário
Uso do solo	-	Maioria em solo rural	-	Maioria em solo rural	-	-
Dia da semana	Maior probabilidade de ocorrência no final de semana	-	-	Maior probabilidade de ocorrência na quinta-feira	Maior probabilidade de ocorrência no final de semana	-
Fase do dia – horário	Maioria ocorreu em pleno dia das 05h00 até as 16h00	Maioria ocorreu em pleno dia das 05h00 até as 16h00	Maioria ocorreu em pleno dia das 05h00 até as 16h00	Maioria ocorreu em pleno dia das 05h00 até as 16h00	Maioria ocorreu em pleno dia das 05h00 até as 16h00	Maioria ocorreu em pleno dia das 05h00 até as 16h00
Condição meteorológica	-	Maioria ocorreu com céu claro e nublado	-	-	-	-

Legenda: CF (Colisão frontal), CL (Colisão lateral), COM (Colisão objeto em movimento), CT (Colisão traseira), CTR (Colisão transversal), E (Engavetamento).
Fonte: A autora (2022).

Considerando o tipo de sinistro saída de pista, que engloba os sinistros: capotamento, colisão com objeto estático, queda de ocupante de veículo, saída

de leito carroçável e tombamento, observou-se na TABELA 4 que o maior coeficiente *Kappa* foi encontrado no teste com o algoritmo J48, porém nenhum algoritmo apresentou resultado satisfatório (todos abaixo de 70% de acurácia, estatística Kappa próximo a 0 e Curva ROC com discriminação fraca). Assim, não há atributos que possam impactar nestes tipos de sinistros, considerando os algoritmos aplicados.

TABELA 4 - RESULTADOS APRESENTADOS PELOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO O TIPO DE SINISTRO SAÍDA DE PISTA.

Algoritmo	**Número de instâncias**	**Instâncias classificadas corretamente**	**Curva ROC**	**Estatística *Kappa***	**Acurácia**	**Tempo de execução (s)**
J48	20255	10569	0,805	0,3348	52,17%	0,37
Multilayer Perceptron	20255	9366	0,686	0,2230	46,24%	376,14
Naive Bayes	20255	8263	0,630	0,1229	40,79%	0,02
SMO	20255	7807	0,575	0,0999	38,54%	1128,89

FONTE: A autora (2022).

Com relação ao tipo de sinistro outros, que engloba os sinistros: derramamento de carga, danos eventuais, eventos atípicos e incêndio, observaram-se na TABELA 5 que os maiores coeficientes *Kappa* foram encontrados nos testes com os algoritmos *Multilayer Perceptron* e J48, os algoritmos *Naive Bayes* e SMO não apresentaram resultados satisfatórios (acurácia abaixo de 70% e estatística *Kappa* próximo a 0 representando nenhuma concordância).

TABELA 5 - RESULTADOS APRESENTADOS PELOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO OS TIPOS DOS SINISTROS: DERRAMAMENTO DE CARGA, DANOS EVENTUAIS, EVENTOS ATÍPICOS E INCÊNDIO.

Algoritmo	Número de instâncias	Instâncias classificadas corretamente	Curva ROC	Estatística *Kappa*	Acurácia	Tempo de execução (s)
Multilayer Perceptron	3008	2592	0,903	0,7899	86,17%	48,91
J48	3008	2476	0,870	0,7399	82,31%	0,02
Naive Bayes	3008	2053	0,722	0,114	68,25%	0,02
SMO	3008	2006	0,586	0,0184	66,68%	69,21

FONTE: A autora (2022).

Considerando o tipo do sinistro outros, para a classificação apresentada na TABELA 5, os algoritmos *Multilayer Perceptron* e J48 foram indicados como classificadores de dados satisfatórios. Os testes realizados trouxeram resultados positivos, com mais de 80% de acurácia, além de ter apresentado agilidade na execução. A estatística *Kappa* se mostrou com concordância substancial e a Curva ROC apresentou uma discriminação boa à excelente segundo critérios estabelecidos nos estudos de Landis e Koch (1977), Hosmer e Lemeshow (2013).

Conforme QUADRO 10, considerando os algoritmos que apresentaram melhores resultados, *Multilayer Perceptron* e J48, os fatores com relação ao veículo, à via e ao ambiente apresentaram um impacto maior nos tipos de sinistros: derramamento de carga, danos eventuais, eventos atípicos e incêndio. O ano de fabricação do veículo (fator veículo) apresentou maior impacto nos sinistros envolvendo derramamento de carga, danos eventuais e incêndio. Já o atributo tipo de pista (fator via), indicado pelo algoritmo *Multilayer Perceptron*, apresentou um impacto maior nos sinistros envolvendo derramamento de carga, eventos atípicos e incêndio. Já o traçado da via (fator via) se mostrou preponderante nos sinistros envolvendo derramamento de carga e danos eventuais. Com relação ao fator ambiente, considerando o algoritmo *Multilayer Perceptron,* a condição meteorológica se mostrou preponderante para os

sinistros derramamento de carga, eventos atípicos e incêndio o horário da ocorrência do sinistro se mostrou preponderante em relação aos outros atributos considerados, conforme indicado a partir da aplicação do algoritmo J48. Já os atributos fase do dia e dia da semana resultaram com um impacto significativo nos sinistros envolvendo derramamento de carga e incêndio. Assim, o fator veículo, via e ambiente apresentaram um maior impacto nos tipos de sinistros classificados como outros.

QUADRO 10 - RESULTADO DOS ALGORITMOS UTILIZADOS CONSIDERANDO OS TIPOS DE ACIDENTE DERRAMAMENTO DE CARGA, DANOS EVENTUAIS, EVENTOS ATÍPICOS E INCÊNDIO.

		Atributo de "resultado" Durante																
		Algoritmos utilizados	**Multilayer Perceptron**				**J48**				**Naive Bayes**				**SMO**			
Atributo de entrada Pré	**Domínio do atributo**	**Atributo**	DC	DE	EA	I	DC	DE	EA	I	DC	DE	EA	I	DC	DE	EA	I
	Fator humano	Sexo			X													
		Idade																
		Tipo envolvido																
	Fator veículo	Tipo veículo																
		Ano de fabricação	**X**	**X**			**X**	**X**		**X**								
	Fator via	Tipo de pista	**X**		**X**	**X**	**X**											
		Traçado da via		X			X				X	X	X	X	X		X	X
		Uso do solo																
	Fator ambiente	Horário						X	X	X								
		Dia da semana				X					X							
		Fase do dia	X															
		Condição meteorológica	**X**		**X**	**X**	**X**				X	X	X	X	X	X		X
		Data (mês)							X		X	X	X	X			X	

Legenda: DC (Derramamento de carga), DE (Danos eventuais), EA (Eventos atípicos) e I (Incêndio).
Fonte: A autora (2022).

Aplicando-se a análise lógica utilizando o *software Microsoft Excel,* foram investigados os 1.106 sinistros envolvendo derramamento de carga. Considerando o fator veículo, a idade predominante da frota foi de 1 a 12 anos, com 48,46% dos sinistros envolvendo derramamento de carga. Considerando o fator via, o traçado da via indicou que 483 (43,67%) sinistros envolvendo derramamento de carga ocorreram em uma curva. Outro atributo analisado, o tipo de pista, mostrou que a maioria dos sinistros envolvendo derramamento de carga ocorreu em pista simples 624 (56,42%). Considerando o fator ambiental, o atributo condição meteorológica apontou 571 (51,62%) sinistros com céu claro.

Considerando o tipo de sinistro danos eventuais, houve 154 sinistros. Analisando o fator veículo, a idade predominante da frota foi de 12 a 22 anos, com 50% dos veículos de transporte de carga. Já a análise do fator via apontou que o traçado da via associado à maior ocorrência de sinistros envolvendo danos eventuais foi em uma reta (42,20%). Já o fator ambiental apresentou 85 sinistros envolvendo danos eventuais (55,19%) em pleno dia entre 05h00min às 16h00min.

Para o tipo de sinistro eventos atípicos, o fator humano com relação ao sexo se mostrou importante com 32 sinistros (94,12%) envolvendo o condutor do sexo masculino. No fator via, o tipo de pista mostrou que a maioria dos sinistros envolvendo eventos atípicos ocorreu em pista dupla com 18 sinistros (52,94%). Já a análise do fator ambiente mostrou que o horário que mais ocorreu este tipo de sinistro foi entre 05h00min às 16h00min, com 52,94%. Para outro atributo analisado, a condição meteorológica, tem-se que 16 sinistros (47,05%) ocorreram com o céu claro e 29,41% ocorreram com o céu nublado. O mês de maior ocorrência foi dezembro, com 25% dos sinistros envolvendo eventos atípicos.

Analisando o sinistro incêndio, ocorreram 389 sinistros. O fator veículo, considerando o ano de fabricação, apontou a idade da frota de 12 a 22 anos como predominante, com média de 16,70 anos, com 50,90% dos sinistros. A análise do tipo de pista (fator via) mostrou que a maioria dos sinistros envolvendo incêndio ocorreu em pista simples com 197 (50,64%) sinistros. Já a análise do fator ambiente mostrou que o horário que mais ocorreu este tipo de sinistro foi entre 05h00min às 16h00min, com 56,04%. A análise do atributo condição meteorológica apontou que 209 sinistros (53,73%) ocorreram com o céu claro. O dia da semana também foi apontado como atributo importante considerando o sinistro incêndio, sendo 96 (24,68%) sinistros ocorreram no final de semana e 20,05% ocorreram na quarta-feira.

A síntese dos resultados referentes aos tipos de sinistros outros pode ser verificada no QUADRO 11.

QUADRO 11 - SÍNTESE DOS RESULTADOS CONSIDERANDO OS TIPOS DE ACIDENTE OUTROS.

Atributos	DC	DE	EA	I
Sexo	-	-	O sexo preponderante foi o masculino	-
Ano de fabricação	A idade da frota é de 1 a 12 anos	A idade da frota é de 2 a 22 anos	-	A idade da frota é de 2 a 22 anos
Tipo de pista	Maioria ocorreu em pista simples	-	Maioria ocorreu em pista dupla	Maioria ocorreu em pista simples
Traçado da via	Maioria ocorreu em curva	Maioria ocorreu em uma reta	-	-
Dia da semana	-	-	-	Maioria ocorreu no final de semana e quarta – feira
Fase do dia – horário	-	Maioria ocorreu em pleno dia das 05h00 até as 16h00	Maioria ocorreu em pleno dia das 05h00 até as 16h00	Maioria ocorreu em pleno dia das 05h00 até as 16h00
Condição meteorológica	Maioria ocorreu com céu claro	-	Maioria ocorreu com céu claro e nublado	Maioria ocorreu com céu claro
Data (mês)	-	-	Maioria ocorreu no mês de dezembro	-

Legenda: DC (Derramamento de carga), DE (Danos eventuais), EA (Eventos atípicos) e I (Incêndio)
Fonte: A autora (2022).

Considerando esses resultados, em geral, verifica-se que o fator via e ambiente se mostraram mais preponderantes nos algoritmos utilizados e que as variáveis do pré-sinistro que apresentam maior influência sobre a variável do durante sinistro foram o tipo de pista e o traçado da via dentro do fator via, e o horário, dia da semana e condição meteorológica considerando o fator ambiente.

4.2 RESULTADOS PRÉ-SINISTRO X PÓS-SINISTRO

Para a análise do pré-sinistro com o pós-sinistro, os atributos classe utilizados para classificação foram o estado físico das vítimas e a classificação do acidente. Os resultados obtidos através dos testes foram organizados nas TABELAS 6 e 7 de acordo com o melhor e o pior resultado de cada algoritmo de classificação.

TABELA 6 - RESULTADOS APRESENTADOS PELOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO O ESTADO FÍSICO DAS VÍTIMAS: ILESO, LESÕES LEVES, LESÕES GRAVES E ÓBITO.

Algoritmo	Número de instâncias	Instâncias classificadas corretamente	Curva ROC	Estatística *Kappa*	Acurácia	Tempo de execução (s)
J48	50759	46781	0,969	0,8756	92,16%	0,91
Multilayer Perceptron	50759	44886	0,975	0,8133	88,42%	656,38
SMO	50759	43314	0,915	0,7573	85,33%	527,31
Naive Bayes	50759	43237	0,967	0,7545	85,18%	0,04

FONTE: A autora (2022).

Considerando o estado físico das vítimas de sinistro (ileso, lesões graves, lesões leves e óbito), observou-se na TABELA 6 que os maiores coeficientes *Kappa* foram encontrados nos testes com o algoritmo J48 e o *Multilayer Perceptron*, porém em todos os testes realizados em todos os parâmetros o *Kappa* obteve classificação acima de 0.61, apresentando uma concordância substancial, conforme intervalos definidos na pesquisa de Landis e Koch (1977). Assim os algoritmos J48 e *Multilayer Perceptron* apresentaram-se como classificadores de dados aceitáveis. Os testes realizados trouxeram resultados positivos além de ter apresentado agilidade na execução, no caso do algoritmo J48. A Curva ROC também apresentou discriminação excelente em todos os testes.

A análise do estado físico das vítimas é em relação a todos os envolvidos no sinistro, como condutores, passageiros e pedestres.

QUADRO 12 - RESULTADO DOS ALGORITMOS UTILIZADOS CONSIDERANDO O ESTADO FÍSICO DAS VÍTIMAS

		Atributo de "resultado" Pós															
Atributo de entrada Pré		**Algoritmos utilizados**	***Multilayer Perceptron***				**J48**				***Naive Bayes***				**SMO**			
	Domínio do atributo	**Atributo**	I	LL	LG	O	I	LL	LG	O	I	LL	LG	O	I	LL	LG	O
	Fator humano	Sexo		**X**		**X**	**X**	**X**	**X**	**X**							**X**	
		Idade					X	X	X	X								
		Tipo envolvido		**X**	**X**	**X**	**X**	**X**	**X**	**X**					**X**	**X**	**X**	**X**
	Fator veículo	Tipo veículo																
		Ano de fabricação																
	Fator via	Tipo de pista								X							X	X
		Traçado da via					**X**		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**		
		Uso do solo						X										
	Fator ambiente	Horário																
		Dia da semana																
		Fase do dia					X											
		Condição meteorológica		X							X	X	X	X				
		Data (mês)	X		X						X	X	X	X				

Legenda: I (Ileso), LL (Lesões Leves), LG (Lesões Graves) e O (Óbito)
Fonte: A autora (2022).

Considerando o estado físico das vítimas de sinistro (ileso, lesões graves, lesões leves e óbito), conforme QUADRO 12, utilizando os algoritmos *Multilayer Perceptron*, J48, *Naive Bayes* e SMO, os fatores: humano, relacionados à via e ao ambiente apresentaram um impacto maior nesses estados físicos das vítimas, de modo que os atributos sexo, idade e o tipo do envolvido apresentaram um peso maior nos algoritmos *Multilayer Perceptron*, J48 e SMO. Já o traçado da via (fator via) apresentou um impacto maior nos algoritmos J48, *Naive Bayes* e SMO. Já o tipo de pista (fator via) se mostrou preponderante no algoritmo SMO. Com relação ao fator ambiente, a condição meteorológica e a data se mostraram preponderantes em relação aos outros atributos considerados, principalmente utilizando os algoritmos *Multilayer Perceptron e Naive Bayes*. Ainda conforme a QUADRO 12, pode-se verificar que o estado físico das vítimas foi influenciado predominantemente pelos fatores humano, via e ambiente.

A partir de análise adicional no editor de planilhas*,* verificou-se 19.509 sinistros com ilesos. Com relação ao fator humano, no caso do atributo idade, 16.488 (84,51%) sinistros envolvendo ilesos ocorreram com pessoas de faixa etária de 25 a 59 anos, conforme classificação de faixa etária da OMS (CARVALHO; PEDROSA, 2015). Considerando o sexo, a maioria dos ilesos foi do sexo masculino com 18.879 (96,77%) e 18.164 (93,10%) participaram como condutores.

Considerando o fator via, o traçado da via indicou que 9.847 (50,47%) dos sinistros envolvendo ilesos ocorreram em uma reta. No fator ambiente, o atributo fase do dia apresentou 11.470 sinistros que ocorreram em pleno dia (58,79%), entre 05h00min às 16h00min. A condição meteorológica apontou que 9.729 (49,87%) sinistros com ilesos ocorreram com céu claro e 3.706 (19,00%) ocorreu com céu nublado. O mês com maior ocorrência de sinistros com ilesos foi março, com 1.863 sinistros.

Utilizando estatística descritiva para validação dos resultados, considerando que houve 1.926 sinistros com lesões graves, a análise do fator humano apresentou 1.552 (80,58%) sinistros com vítimas na faixa etária de 25 a 59 anos, sendo 1.769 (91,85%) do sexo masculino. No fator via, com relação ao traçado da via, 716 sinistros com lesões graves, ou seja, 37,17% ocorrem em uma reta e 520 (27,00%) sinistros ocorreram em uma curva. Outro atributo analisado em relação à via foi o tipo de pista, para o qual 987 sinistros com lesões graves (51,25%) aconteceram em pista simples. Considerando o fator ambiente, verificou-se que 1.104 sinistros (57,32%) ocorreram

com céu claro. Outubro foi o mês com maior probabilidade de ocorrência de sinistros envolvendo lesões graves (9,00%).

Analisando as lesões leves, houve 5.350 sinistros. Com relação ao fator humano, atributo idade, 4.429 (82,78%) sinistros envolvendo sinistros com lesões leves ocorreram com vítimas de faixa etária de 25 a 59 anos conforme classificação de faixa etária da OMS (CARVALHO; PEDROSA, 2015). Considerando o sexo, a maioria das vítimas com lesões leves são do sexo masculino com 4.930 (92,15%).

Já o fator via, o atributo traçado da via indicou que 1.806 (33,76%) dos sinistros envolvendo vítimas com lesões leves ocorreram em uma reta e 1.779 (33,25%) ocorreram em uma curva. Já considerando o uso do solo, 4.474 sinistros, ou seja, 83,63% ocorreram em solo rural. No fator ambiente, a condição meteorológica apontou que 2.693 (50,34%) sinistros com vítimas com lesões leves ocorreram com céu claro e 896 (16,75%) ocorreram com céu nublado. O mês com maior ocorrência de sinistros com lesões leves foi agosto, com 516 sinistros.

Considerando os sinistros em que houve óbito, um total de 1.049 sinistros, analisando o fator humano, 721 (68,73%) sinistros com óbitos ocorreram com vítimas de faixa etária de 25 a 59 anos e o sexo predominante foi o masculino, com 939 (89,51%) sinistros. No fator via, o atributo traçado da via indicou que 415 (39,56%) dos sinistros envolvendo óbitos ocorreram em uma reta e 312 (29,74%) ocorreram em uma curva. Já considerando o tipo de pista, 517 sinistros, ou seja, 49,28% ocorreram em pista simples. No fator ambiente, a análise da condição meteorológica apontou que 539 (51,38%) sinistros com óbitos ocorreram com céu claro. O mês com maior probabilidade de ocorrência de óbitos foi junho, com 135 sinistros.

A síntese dos resultados referentes ao estado físico das vítimas pode ser verificada no QUADRO 13.

QUADRO 13 - SÍNTESE DOS RESULTADOS CONSIDERANDO O ESTADO FÍSICO DAS VÍTIMAS

Atributos	I	LL	LG	O
Sexo	O sexo preponderante foi o masculino	O sexo preponderante foi o masculino	O sexo preponderante foi o masculino	O sexo preponderante foi o masculino
Idade	A faixa etária foi de 25 a 59 anos	A faixa etária foi de 25 a 59 anos	A faixa etária foi de 25 a 59 anos	A faixa etária foi de 25 a 59 anos
Tipo de pista	-	-	Maioria ocorreu em pista simples	Maioria ocorreu em pista simples
Traçado da via	Maioria ocorreu em reta	Maioria ocorreu em reta e curva	Maioria ocorreu em uma reta e curva	Maioria ocorreu em reta e curva
Uso solo	-	Maioria em solo rural	-	-
Condição meteorológica	Maioria ocorreu com céu claro e nublado	Maioria ocorreu com céu claro e nublado	-	Maioria ocorreu com céu claro
Data (mês)	Maioria ocorreu no mês de março	Maioria ocorreu no mês de agosto	Maioria ocorreu no mês de outubro	Maioria ocorreu no mês de junho

Legenda: I (Ileso), LL (Lesões Leves), LG (Lesões Graves) e O (Óbito).
Fonte: A autora (2022).

Considerando a classificação do acidente, para a classificação apresentada na TABELA 7, os algoritmos de J48 e *Multilayer Perceptron* apresentam-se como adequados classificadores de dados, o teste realizado trouxe resultado positivo, apesar de o coeficiente *Kappa* ser menor que 0,61 no algoritmo *Multilayer Perceptron*, o algoritmo apresentou mais de 70% de acurácia e Curva ROC com discriminação boa.

TABELA 7 - RESULTADOS APRESENTADOS PELOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE: SEM VÍTIMAS, COM VÍTIMAS FERIDAS E VÍTIMAS FATAIS.

Algoritmo	Número de instâncias	Instâncias classificadas corretamente	Curva ROC	Estatística *Kappa*	Acurácia	Tempo de execução (s)
J48	50759	42764	0,907	0,7557	84,24%	2,07
Multilayer Perceptron	50759	38055	0,841	0,5323	74,97%	874,94
Naive Bayes	50759	31308	0,705	0,1517	61,67%	0,09
SMO	50759	30751	0,569	0,0745	60,58%	2560,7

FONTE: A autora (2022).

Conforme QUADRO 14, considerando a classificação do sinistro e os algoritmos que apresentaram melhores resultados, os fatores com relação à via e ao

ambiente resultaram com um impacto maior nos tipos de sinistros que apresentam vítimas feridas e vítimas fatais, para os quais o tipo de pista e o traçado da via (fator via) apresentaram um peso maior no algoritmo J48. Já o traçado da via (fator via) se mostrou preponderante também no algoritmo *Multilayer Perceptron*. No algoritmo J48, com relação ao fator ambiente, a condição meteorológica se mostrou preponderante em relação aos outros atributos considerados.

Quando não há vitimas no sinistro, os algoritmos J48 e *Multilayer Perceptron* apresentaram resultados similares, de forma que o tipo do envolvido (fator humano), o traçado da via (fator via) e a condição meteorológica (fator ambiente) se mostraram preponderantes. Assim, os fatores via e ambiente apresentaram um maior impacto nos sinistros envolvendo vítimas fatais e com ferimentos. Nos sinistros sem vítimas, os fatores humano, via e ambiente se mostraram com maior impacto.

QUADRO 14 - RESULTADO DOS ALGORITMOS UTILIZADOS CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE: SEM VÍTIMAS, COM VÍTIMAS FERIDAS E VÍTIMAS FATAIS.

		Atributo de "resultado" Pós						
		Algoritmos utilizados	**J48**			***Multilayer Perceptron***		
Atributo de entrada Pré	**Domínio do atributo**	**Atributo**	SV	CVF	VF	SV	CVF	VF
	Fator humano	Sexo				X		
		Idade						
		Tipo envolvido	X			X		
	Fator veículo	Tipo veículo						
		Ano de fabricação						X
	Fator via	Tipo de pista	X	X	X			
		Traçado da via	**X**	**X**	**X**	**X**	**X**	**X**
		Uso do solo					X	
	Fator ambiente	Horário						
		Dia da semana						
		Fase do dia						
		Condição meteorológica	**X**	**X**	**X**		**X**	
		Data (mês)						

Legenda: SV (Sem vítimas), CVF (Com vítimas feridas) e VF (Vítimas fatais).
Fonte: A autora (2022).

Foi realizada também uma análise utilizando estatística descritiva para os 18.927 sinistros envolvendo vítimas feridas. No fator via, com relação ao traçado da

via, 8.815 sinistros, ou seja, 46,57% ocorrem em uma reta. Para o atributo uso do solo, 13.543 sinistros envolvendo vítimas feridas (71,55%) acontecem em solo rural e 9.221 sinistros foram em pista dupla (48,72%). Considerando o fator ambiente, verificou-se a condição meteorológica, 10.081 sinistros envolvendo vítimas feridas (53,26%) ocorreram com céu claro e 3.169 (16,74%) com o céu nublado.

Analisando sinistros com vítimas fatais, houve 5.282 sinistros. A idade da frota foi de 1 a 12 anos, com média de 8,30 anos, com 51,06%. Considerando o fator via, o traçado da via indicou que 2.489 (47,12%) sinistros fatais ocorreram em uma reta e 1.289 (24,40%) em curva. Outro atributo analisado, o tipo de pista, mostrou que a maioria dos sinistros envolvendo vítimas fatais, ocorreu em pista simples 3.107 (58,82%). Considerando o fator ambiental, a condição meteorológica apontou que 2.136 (40,44%) sinistros com vítimas fatais ocorreram com céu nublado.

Analisando sinistros sem vítimas, houve 8.076 sinistros, sendo que 7.304 (90,44%) sinistros tiveram o condutor como principal envolvido. No fator humano, verificou-se ainda que o sexo preponderante foi o masculino, em 87,88% dos sinistros. Considerando o fator via, o traçado da via indicou que 3.837 (47,51%) sinistros ocorreram em uma reta e 2.063 (25,54%) em curva. Para outro atributo analisado, o tipo de pista, tem-se que a maioria dos sinistros sem vítimas, ocorreu em pista dupla 4.574 (56,64%). No fator ambiental, a análise da condição meteorológica apontou 3.703 (45,85%) sinistros sem vítima ocorreram com o com céu claro.

A síntese dos resultados referentes à classificação do sinistro pode ser verificada no QUADRO 15.

QUADRO 15 - SÍNTESE DOS RESULTADOS CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE

Atributos	SV	CVF	VF

Sexo	O sexo preponderante foi o masculino	-	-
Tipo do envolvido	Condutor foi o principal envolvido	-	-
Ano de fabricação	-	-	A frota tem de 1 a 12 anos
Tipo de pista	Maioria ocorreu em pista dupla	Maioria ocorreu em pista dupla	Maioria ocorreu em pista simples
Traçado da via	Maioria ocorreu em reta e curva	Maioria ocorreu em reta	Maioria ocorreu em uma reta e curva
Uso solo	Maioria em solo rural	-	-
Condição meteorológica	Maioria ocorreu com céu claro	Maioria ocorreu com céu claro e nublado	Maioria ocorreu com céu nublado

Legenda: SV (Sem vítimas), CVF (Com vítimas feridas) e VF (Vítimas fatais).
Fonte: A autora (2022).

Considerando esses resultados, em geral, verifica-se que o fator humano, via e ambiente se mostraram mais preponderantes nos algoritmos utilizados e que as variáveis do pré-sinistro que apresentam maior impacto sobre a variável do pós-sinistro foram o sexo, a idade e o tipo de envolvido (fator humano), tipo de pista e traçado da via (fator via) e condição meteorológica (fator ambiente).

4.3 RESULTADOS DURANTE SINISTRO X PÓS-SINISTRO

Para a análise da relação “durante sinistro x pós-sinistro”, os atributos classe utilizados para classificação foram a classificação do acidente e o estado físico das vítimas, respectivamente. Os resultados obtidos através dos testes foram organizados nas TABELAS 8 e 9 de acordo com o melhor e o pior resultado de cada algoritmo de classificação.

TABELA 8 - RESULTADOS DOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO O ESTADO FÍSICO DAS VÍTIMAS: ILESO, LESÕES LEVES, LESÕES GRAVES E ÓBITO COM O TIPO DO SINISTRO.

Algoritmo	Número de instâncias	Instâncias classificadas corretamente	Curva ROC	Estatística *Kappa*	Acurácia	Tempo de execução (s)
J48	50759	23299	0,574	0,0514	45,90%	0,01
Naive Bayes	50759	23278	0,574	0,0504	45,85%	0,01
SMO	50759	23278	0,527	0,0504	45,85%	93,3
Multilayer Perceptron	50759	22322	0,561	0,0276	43,97%	87,2

FONTE: A autora (2022).

Considerando o estado físico das vítimas de sinistro (ileso, lesões graves, lesões leves, óbito e não informado), observou-se na TABELA 8 que nenhum algoritmo se apresentou como ótimo classificador de dados. Os testes realizados trouxeram resultados negativos, com baixo coeficiente *Kappa* (próximo à zero), Curva ROC com discriminação fraca e taxa de acurácia menor de 50%.

Da mesma forma, considerando a classificação do acidente, para a classificação apresentada na TABELA 9, nenhum algoritmo apresentou resultado satisfatório.

TABELA 9 - RESULTADOS DOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE: SEM VÍTIMAS, COM VÍTIMAS FERIDAS E VÍTIMAS FATAIS COM O TIPO DO SINISTRO.

Algoritmo	Número de instâncias	Instâncias classificadas corretamente	Curva ROC	Estatística *Kappa*	Acurácia	Tempo de execução (s)
Multilayer Perceptron	50759	29941	0,613	0,0503	58,98%	67,74
J48	50759	29937	0,613	0,0505	58,97%	0,01
Naive Bayes	50759	29937	0,613	0,0505	58,97%	0,01
SMO	50759	29937	0,557	0,0505	58,97%	147,63

FONTE: A autora (2022).

Assim, conforme TABELA 8 e 9, considerando os algoritmos utilizados nesta pesquisa, não foi possível concluir alguma relação dos atributos durante e pós-sinistro.

4.4 ANÁLISE PRÉ E DURANTE PANDEMIA COVID-19

Os resultados obtidos a partir dos testes com os dados anteriores à pandemia de COVID-19 considerando os anos de 2017, 2018 e 2019 e os dados durante pandemia, 2020 e 2021, foram organizados nas TABELAS 10 e 11, respectivamente, de acordo com o melhor e o pior resultado de cada algoritmo de classificação considerando o número de instâncias classificadas corretamente, curva ROC, estatística *Kappa* e a acurácia.

TABELA 10 - RESULTADOS APRESENTADOS PELOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE: SEM VÍTIMAS, COM VÍTIMAS FERIDAS E VÍTIMAS FATAIS NOS ANOS 2017, 2018 E 2019.

Algoritmo	**Número de instâncias**	**Instâncias classificadas corretamente**	**Curva ROC**	**Estatística *Kappa***	**Acurácia**	**Tempo de execução (s)**
J48	29999	24005	0,878	0,7731	80,07%	0,79
Multilayer Perceptron	29999	20093	0,734	0,5256	66,98%	457,75
SMO	29999	16790	0,531	0,3419	55,97%	633,22
Naive Bayes	29999	15879	0,609	0,1255	52,96%	0,07

FONTE: A autora (2022).

Considerando a classificação do acidente nos anos de 2017, 2018 e 2019, para a classificação apresentada na TABELA 10, o algoritmo J48 apresentou-se como um classificador de dados satisfatório, com índice de Curva ROC com discriminação boa, estatística Kappa com concordância substancial e acurácia acima de 80%. Conforme QUADRO 16, considerando a classificação do sinistro e o algoritmo que apresentou melhor resultado, os fatores humanos e com relação à via e ao ambiente resultaram com um impacto maior nos tipos de sinistros que apresentam vítimas feridas e vítimas

fatais, de modo que a idade (fator humano), o tipo de pista e o traçado da via (fator via) apresentaram um impacto maior no algoritmo J48. Com relação ao fator ambiente, a condição meteorológica se mostrou preponderante em relação aos outros atributos considerados.

Quando não há vitimas no sinistro, o algoritmo J48 apontou que o traçado da via (fator via) e a condição meteorológica (fator ambiente) se mostraram preponderantes.

QUADRO 16 - RESULTADO DOS ALGORITMOS UTILIZADOS CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE: SEM VÍTIMA, COM VÍTIMAS FERIDAS E VÍTIMAS FATAIS NOS ANOS 2017, 2018 E 2019.

	Algoritmo utilizado	**J48**		
Domínio do atributo	**Atributo**	SV	CVF	VF
Fator humano	Sexo			
	Idade		X	X
	Tipo envolvido			
Fator veículo	Tipo veículo			
	Ano de fabricação			
Fator via	Tipo de pista	**X**	**X**	**X**
	Traçado da via	**X**	**X**	**X**
	Uso do solo			
Fator ambiente	Horário			
	Dia da semana			
	Fase do dia			X
	Condição meteorológica	**X**	**X**	**X**
	Data (mês)			

Legenda: SV (Sem vítimas), CVF (Com vítimas feridas) e VF (Vítimas fatais).
Fonte: A autora (2022).

Assim, os fatores humano, via e ambiente apresentaram um maior impacto nos sinistros envolvendo vítimas fatais e com ferimentos. Nos sinistros sem vítimas, os fatores via e ambiente se mostraram mais importantes.

Foi realizada também uma análise utilizando planilhas sobre os 16.778 sinistros envolvendo vítimas feridas nos anos de 2017, 2018 e 2019, a faixa etária de 25 a 59 anos foi predominante com 7.900 sinistros, ou seja, 47,08%. No fator via, com relação ao traçado da via, 7.494 sinistros (44,67%) ocorreram em uma reta. Para o atributo em relação à via, o tipo de pista, 8.336 sinistros envolvendo vítimas feridas (49,68%) aconteceram em pista simples. Considerando o fator ambiente, verificou-se a condição meteorológica, de modo que 8.438 sinistros envolvendo vítimas feridas (50,29%) ocorreram com céu claro e 2.972 (17,71%) sinistros com o tempo nublado.

Analisando sinistros com vítimas fatais nos anos de 2017, 2018 e 2019, houve 3.507 sinistros. A faixa etária mais atingida foi de 25 a 59 anos com 43,65%. Considerando o fator via, a análise do traçado da via indicou que 1.664 (47,45%) sinistros fatais ocorreram em uma reta e 783 (22,33%) em curva. Outro atributo analisado, o tipo de pista, mostrou que a maioria dos sinistros envolvendo vítimas fatais ocorreu em pista simples, 61,45%. Considerando o fator ambiental, a maioria dos registros com vítimas fatais ocorreu em pleno dia (49,33%); já em relação à condição meteorológica, 1.722 (49,10%) sinistros com vítimas fatais ocorreram com céu claro.

Analisando sinistros sem vítimas nos anos de 2017 a 2019, houve 9.714 sinistros. Considerando o fator via, o traçado da via indicou que 4.501 (46,33%) sinistros ocorreram em uma reta e 2.629 (27,05%) em curva. Outro atributo analisado, o tipo de pista, mostrou que a maioria dos sinistros sem vítimas ocorreu em pista dupla 5.521 (56,83%). No fator ambiental, a análise da condição meteorológica apontou 4.370 (45,00%) sinistros sem vítima ocorreram com o com céu claro e 1.921 (19,78%) com céu nublado.

Considerando a classificação do acidente nos anos de 2020 e 2021, para a classificação apresentada na TABELA 11, os algoritmos de J48 e *Multilayer Perceptron* apresentam-se como adequados classificadores de dados. Os testes realizados trouxeram resultados aceitáveis, apesar de o coeficiente Kappa ser menor que 0,75 no algoritmo *Multilayer Perceptron*, ele apresentou mais de 80% de acurácia e a Curva ROC apresentou discriminação boa.

TABELA 11 - RESULTADOS APRESENTADOS PELOS ALGORITMOS DE CLASSIFICAÇÃO CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE: SEM VÍTIMAS, COM VÍTIMAS FERIDAS E VÍTIMAS FATAIS NOS ANOS 2020 E 2021.

Algoritmo	Número de instâncias	Instâncias classificadas corretamente	Curva ROC	Estatística *Kappa*	Acurácia	Tempo de execução (s)
J48	20760	17850	0,913	0,7731	85,98%	0,24
Multilayer Perceptron	20760	17562	0,878	0,7027	84,59%	303,79
Naive Bayes	20760	13086	0,639	0,2317	63,03%	0,01
SMO	20760	12870	0,551	0,1662	62,00%	309,11

FONTE: A autora (2022).

Considerando a classificação do sinistro e os algoritmos que apresentaram melhores resultados, os fatores com relação à via e ao ambiente resultaram com um impacto maior nos tipos de sinistros que apresentam vítimas feridas e vítimas fatais, de modo que o tipo de pista e o traçado da via (fator via) apresentaram um impacto maior no algoritmo J48. Já o traçado da via (fator via) se mostrou preponderante também no algoritmo *Multilayer Perceptron*. Com relação ao fator ambiente, a condição meteorológica se mostrou preponderante em relação aos outros atributos considerados, no algoritmo J48.

Quando não há vitimas no sinistro, os algoritmos J48 e *Multilayer Perceptron* apresentaram resultados similares, de forma que o tipo do envolvido (fator humano), o traçado da via (fator via) e a condição meteorológica (fator ambiente) se mostraram preponderantes.

QUADRO 17 - RESULTADOS DOS ALGORITMOS UTILIZADOS CONSIDERANDO A CLASSIFICAÇÃO DO ACIDENTE: SEM VÍTIMA, COM VÍTIMAS FERIDAS E VÍTIMAS FATAIS NOS ANOS DE 2020 E 2021.

		Atributo de "resultado" Durante						
		Algoritmos utilizados	**J48**			***Multilayer Perceptron***		
Atributo de entrada Pré	**Domínio do atributo**	**Atributo**	SV	CVF	VF	SV	CVF	VF
	Fator humano	Sexo						
		Idade						
		Tipo envolvido				X		
	Fator veículo	Tipo veículo						
		Ano de fabricação						
	Fator via	Tipo de pista	**X**	**X**	**X**			
		Traçado da via	**X**	**X**	**X**	**X**	**X**	**X**
		Uso do solo						
	Fator ambiente	Horário						
		Dia da semana						
		Fase do dia						
		Condição meteorológica	**X**	**X**	**X**	**X**		
		Data (mês)						

Legenda: SV (Sem vítimas), CVF (Com vítimas feridas) e VF (Vítimas fatais).
Fonte: A autora (2022).

Analisando os sinistros com vítimas feridas nos anos de 2020 e 2021, houve 12.868 sinistros. No fator via, com relação ao traçado da via, 5.945 sinistros (46,20%) ocorreram em uma reta. Para o atributo tipo de pista (fator via), 6.902 sinistros envolvendo vítimas feridas (53,64%) aconteceram em pista simples. Considerando o fator ambiente, verificou-se a condição meteorológica, de modo que 7.249 sinistros envolvendo vítimas feridas (56,33%) ocorreram com céu claro e 3.481 (27,05%) dos sinistros ocorreram em condição de chuva.

Já considerando os sinistros com vítimas fatais nos anos de 2020 e 2021, houve 3.765 sinistros. Analisando o fator via, a análise do traçado da via indicou que 1.725 (45,80%) sinistros fatais ocorreram em uma reta e 989 (26,26%) em curva. Outro atributo analisado, o tipo de pista, mostrou que a maioria dos sinistros envolvendo vítimas fatais, ocorreu em pista simples, 2.161 sinistros (57,38%). Considerando o fator ambiental, em relação à condição meteorológica, 2.027 (53,82%) sinistros com vítimas fatais, entre 2020 e 2021, ocorreram com céu claro.

Analisando sinistros sem vítimas nos anos de 2020 e 2021, houve 4.127 sinistros. Considerando o fator via, o traçado da via indicou que 1.884 (45,65%) sinistros ocorreram em uma reta e 1.040 (25,20%) em uma curva. Outro atributo

analisado, o tipo de pista, mostrou que a maioria dos sinistros sem vítimas, ocorreu em pista dupla 2.385 (57,79%). No fator ambiental, a condição meteorológica apontou 2.121 (51,39%) sinistros sem vítima ocorreram com o com céu claro e 1.435 (34,77%) com chuva.

A síntese dos resultados referentes à análise pré e durante pandemia pode ser verificada no QUADRO 18.

QUADRO 18 - SÍNTESE DOS RESULTADOS CONSIDERANDO A ANÁLISE PRÉ E DURANTE PANDEMIA DE COVID-19

Antes da pandemia (2017 a 2019)			
Atributos	**CVF**	**VF**	**SV**
Idade	A faixa etária foi de 25 a 59 anos	A faixa etária foi de 25 a 59 anos	-
Tipo do envolvido	-	-	-
Tipo de pista	Maioria ocorreu em pista simples	Maioria ocorreu em pista simples	Maioria ocorreu em pista dupla
Traçado da via	Maioria ocorreu em reta	Maioria ocorreu em uma reta e curva	Maioria ocorreu em reta e curva
Fase do dia	-	Maioria ocorreu em pleno dia	-
Condição meteorológica	Maioria ocorreu com céu claro e nublado	Maioria ocorreu com céu claro	Maioria ocorreu com céu claro e nublado
Durante a pandemia (2020 e 2021)			
Atributos	**CVF**	**VF**	**SV**
Idade	-	-	-
Tipo do envolvido	-	-	Condutor foi o principal envolvido
Tipo de pista	Maioria ocorreu em pista simples	Maioria ocorreu em pista simples	Maioria ocorreu em pista dupla
Traçado da via	Maioria ocorreu em reta	Maioria ocorreu em uma reta e curva	Maioria ocorreu em reta e curva
Fase do dia	-	-	-
Condição meteorológica	Maioria ocorreu com céu claro e chuva	Maioria ocorreu com céu claro	Maioria ocorreu com céu claro e chuva

Legenda: SV (Sem vítimas), CVF (Com vítimas feridas) e VF (Vítimas fatais).
Fonte: A autora (2022).

Em relação à totalidade dos sinistros com vítimas feridas e fatais ocorridos no Paraná, houve um aumento quando comparados os anos de 2019 e 2020, ano que marca o início da pandemia. Inclusive o ano de 2021 teve aumento de sinistros envolvendo vítimas feridas e fatais em comparação com 2020.

Quando comparado os fatores que influenciaram nos sinistros de trânsito nos anos antes da pandemia, ou seja, 2017, 2018 e 2019, eles indicaram os fatores humanos, da via e do ambiente como preponderantes, já nos anos de 2020 e 2021 foram presentes os fatores da via e do ambiente.

5 DISCUSSÃO DOS RESULTADOS

Ao comparar os quatro algoritmos, identificou-se que o algoritmo J48 se apresentou como um classificador ótimo em sete dos dez testes realizados. Já o algoritmo *Multilayer Perceptron* se apresentou com bons resultados em cinco testes. Os algoritmos *Naive Bayes* e SMO não apresentaram resultados satisfatórios na maioria dos testes. Para a análise de registros de sinistros de trânsito, o algoritmo J48 se mostrou, ao longo do estudo, mais apropriado.

No que diz respeito à discussão dos resultados no âmbito da segurança viária, nesta seção serão destacadas as principais diferenças identificadas nos quadros que sintetizam os atributos dos sinistros que apresentaram maior impacto nas relações estabelecidas, conforme:

- Quadros 7, 9 e 11 para a relação entre atributos “pré-sinistro x durante sinistro”;
- Quadros 13 e 15 para a relação entre atributos “pré-sinistro x pós-sinistro”;
- Quadro 18 para a relação entre atributos “pré-sinistro x durante sinistro” nos períodos antes da pandemia (2017 a 2019) e durante a pandemia (2020 e 2021).

Foram aqui consideradas apenas as relações para as quais foi obtido resultado satisfatório para o desempenho de no mínimo um dos algoritmos testados, conforme classificações dos estudos de Tharwat (2021) para a análise da acurácia e Curva ROC, Hosmer e Lemeshow (2013) para análise da Curva ROC e os estudos de Landis e Koch (1977) e Lantz (2019) para a classificação da estatística *Kappa*. Dessa forma, os resultados para a relação entre os atributos “durante sinistro x pós-sinistro” não serão discutidos, tendo em vista que não foi obtido desempenho satisfatório dos algoritmos.

O presente trabalho teve como objetivo geral testar técnicas de mineração de dados para a análise de dados de sinistros de trânsito, bem como comparar padrões de sinistros encontrados na literatura envolvendo o transporte rodoviário de cargas com os padrões de sinistros ocorridos em rodovias federais do Brasil utilizando ferramentas de mineração de dados, a partir dos dados disponibilizados pela Polícia Rodoviária Federal (PRF), no período de 2017 a 2021 e investigar os possíveis impactos da pandemia de COVID-19 nos sinistros de trânsito, visando contribuir no processo decisório dos gestores de organizações públicas e privadas. Primeiramente foram identificados e mapeados os artigos que abordam o tema e as suas relações

apresentadas, por meio da revisão bibliográfica. Após a realização do mapeamento dos fatores que contribuíram para os sinistros envolvendo o transporte rodoviário de cargas, foi possível realizar comparações com os fatores que contribuíram para os sinistros ocorridos em rodovias federais do Paraná.

No âmbito atropelamentos de animais e pedestres e outros tipos de sinistros, não foram encontrados registros na literatura envolvendo o transporte rodoviário de cargas para a comparação e análise.

Os resultados envolvendo colisões (colisões frontais, laterais, transversais, traseiras, com objeto em movimento e engavetamentos) indicaram que a idade do condutor se associou a uma maior propensão e gravidade nos sinistros envolvendo o transporte rodoviário de cargas, principalmente para condutores jovens. Outro ponto é a ocorrência deste tipo de sinistro em pista simples, em concordância com os estudos de Hakkanen e Summala (2001) e Zheng, Lu e Lantz (2018). Além disso, ao contrário de estudos anteriores (Hakkanen e Summala, 2001; Forkenbrock e Hanley, 2003; Zhu e Srinivasan, 2011), a maioria dos sinistros envolvendo colisões ocorreu durante o dia, os estudos de Björnstig, Björnstig e Eriksson (2008) e Zheng, Lu e Lantz (2018) indicaram este aspecto.

As colisões (colisões frontais, laterais, transversais, traseiras, com objeto em movimento e engavetamentos) foram apontadas nos resultados com maior probabilidade de ocorrência em condições climáticas de céu claro e nublado, convergente com o estudo de Zheng, Lu e Lantz (2018) e divergente dos estudos de Khorashadi *et al.* (2005) e Tsai e Su (2004) que apontaram a principal condição climática de neblina para sinistros envolvendo colisões. A colisão frontal foi apontada nos resultados com maior probabilidade de ocorrência nos finais de semana, confirmando o estudo de Zhu e Srinivasan (2011).

Os resultados também indicam que a idade do condutor contribuiu significativamente para a gravidade dos sinistros envolvendo o transporte rodoviário de cargas, conforme estudos (Tsai e Su, 2004; Islam e Hernandez, 2013; Islam *et al.* 2014 e Zheng, Lu e Lantz, 2018). O céu claro ou nublado foi apontado nos resultados como um dos fatores de sinistros graves e com vítimas fatais, corroborando com o estudo de Lemp *et al.* (2011).

O traçado da via do tipo reta e curva foi identificado nos resultados de sinistros mais graves como fator significativo. Esta condição também se manifestou no estudo de Islam e Hernandez (2013). Além disso, a época do ano que mais ocorreram

sinistros graves e óbitos foram nos meses de outubro e junho respectivamente, conforme também apontado no estudo de Björnstig, Björnstig e Eriksson (2008); diferente dos resultados dos estudos de Islam e Hernandez (2013) e Pahukula, Hernandez e Unnikrishnan (2015), que mostram maior gravidade nos sinistros nos meses de verão.

Sinistros com vítimas graves ou fatais envolvendo o transporte rodoviário de cargas tiveram também como fatores significativos: condição da superfície da estrada (Forkenbrock e Hanley, 2003; Lemp *et al.*, 2011; Pahukula, Hernandez e Unnikrishnan, 2015; Zheng, Lu e Lantz, 2018) ; número de veículos envolvidos (Forkenbrock e Hanley, 2003; Chang e Chien, 2013; Pahukula, Hernandez e Unnikrishnan, 2015; Zheng, Lu e Lantz, 2018); o limite de velocidade na via (Forkenbrock e Hanley, 2003; Björnstig; Björnstig; Eriksson, 2008; Islam e Hernandez, 2013), consumo de bebidas alcoólicas e drogas (Khorashadi *et al.*, 2005; Zhu e Srinivasan, 2011; Chang e Chien, 2013), o uso do cinto de segurança (Chang e Chien, 2013; Zhu e Srinivasan, 2011), Islam e Hernandez, 2013), o tipo de veículo envolvido (Tsai e Su, 2004; Chang e Chien, 2013); obstruções na via (Tsai e Su, 2004), número de reboques dos veículos de carga (Khorashadi *et al.*, 2005; Lemp *et al.*, 2011); dimensões e peso do veículo de carga (Lemp *et al.*, 2011; Zheng, Lu e Lantz, 2018), iluminação na via (Lemp *et al.*, 2011; Islam e Hernandez, 2013; Pahukula, Hernandez e Unnikrishnan, 2015; Zheng, Lu e Lantz, 2018), fluxo de tráfego (Pahukula, Hernandez e Unnikrishnan, 2015), características da empresa de transporte rodoviário de cargas (Zheng, Lu e Lantz, 2018) e desatenção do motoristas (Zhu e Srinivasan, 2011).

Além disso, o histórico de sinistros dos condutores e a existência de doenças crônicas nos condutores de veículos de carga conforme apontados no estudo de Hakkane e Summala (2001) são fatores que não puderam ser analisados devido à falta destas informações no banco de dados da PRF.

6 CONSIDERAÇÕES FINAIS

Este estudo utilizou algumas técnicas, no âmbito da Gestão da Informação, orientadas à classificação de dados: árvore de decisão, rede neural, máquina de vetores de suporte e *Naive Bayes*, para modelar dados de sinistros de trânsito, com dados da PRF no estado do Paraná. Os resultados, portanto, devem ser interpretados considerando este recorte geográfico da pesquisa.

A contribuição deste trabalho está na comparação dos padrões de sinistros de trânsito encontrados na literatura envolvendo o transporte rodoviário de cargas com os padrões de sinistros ocorridos em rodovias federais do Brasil utilizando ferramentas de mineração de dados.

É relevante salientar que a aplicação de tais metodologias torna-se uma opção interessante para a gestão da informação da segurança viária. Neste trabalho, quatro categorias de fatores foram consideradas como potenciais determinantes dos sinistros de trânsito: humano, veículo, via e ambiente. Com várias combinações de categorias, os achados do presente estudo revelam que os fatores via e ambiente foram mais preponderantes que os fatores humano e do veículo.

Considerando os resultados apresentados, verificou-se que o fator via e ambiente se mostraram mais preponderantes nos algoritmos utilizados e que as variáveis do pré-sinistro que apresentam maior influência sobre a variável do durante sinistro são o tipo de pista, traçado da via, no fator via, e considerando o fator ambiente são as variáveis: horário, dia da semana e condição meteorológica.

Na análise pré-sinistro x pós-sinistro, verificou-se que os fatores humano, via e ambiente se mostraram mais preponderantes nos algoritmos utilizados e que as variáveis do pré-sinistro que apresentam maior influência sobre a variável do pós-sinistro são o sexo, a idade e o tipo de envolvido (fator humano), tipo de pista e traçado da via, sendo o fator via, e considerando o fator ambiente é a variável condição meteorológica. Nenhuma relação foi encontrada na análise realizada com os dados do durante sinistro x pós-sinistro.

Ao comparar os quatro algoritmos, identificou-se que o algoritmo J48 se apresentou como um classificador satisfatório em 7 dos 10 testes realizados.

Já na comparação de fatores que influenciaram os sinistros envolvendo o transporte rodoviário de cargas que constam na literatura com os fatores encontrados

neste trabalho, verificou-se que os fatores humano, via e ambiente se mostraram mais preponderantes, confirmando, de maneira geral, o resultado da pesquisa.

Por fim, o presente trabalho realizou uma análise da pandemia da COVID-19, que gerou grandes repercussões em todo o mundo, dentre elas mudanças nos padrões de mobilidade urbana e nos sinistros de trânsito. Em relação à totalidade dos sinistros com vítimas feridas e fatais ocorridos no Paraná, houve um aumento quando comparados os anos de 2019, 2020 e 2021. Entretanto, não foram identificadas diferenças significativas nas relações entre os atributos dos sinistros ocorridos antes e durante a pandemia.

A utilização de técnicas de mineração de dados se mostrou útil para uma análise de caráter mais exploratório sobre grandes bases de dados, constituindo-se, portanto de uma etapa de análise preliminar às etapas mais detalhadas e que demandam a consideração de aspectos técnicos pertinentes à área de especialidade dos dados; neste caso, a segurança viária e seu caráter interdisciplinar. Dessa forma, entende-se que o uso das técnicas de mineração de dados apresenta um grande potencial de contribuição para análises de grandes bases de dados de sinistros, principalmente com a implementação em curso do Registro Nacional de Acidentes e Estatísticas de Trânsito (Renaest). Contudo, a aplicação de técnicas de mineração de dados não dispensa a visão crítica do analista especialista no tema sob investigação.

Como sugestão para avanços a partir deste, trabalhos futuros podem ser realizados incluindo outras técnicas e algoritmos de mineração de dados. Outra sugestão, seria a implementação em outros estados do Brasil, visto que o estudo se limitou aos dados de sinistros das rodovias federais do estado do Paraná. Além da utilização de outras bases de dados, com dados da saúde e informações das rodovias.

6.1 LIMITAÇÕES DO ESTUDO

Dentre as limitações deste estudo, incluem-se as relativas à abrangência dos achados e aos dados, visto que foram utilizados dados da PRF, que constituem apenas sinistros em rodovias federais. As rodovias estaduais e municipais não foram objeto de análise, embora constituam a maior parte da malha rodoviária do país. A partir dos dados apresentados, ficou evidenciada a dificuldade em analisar um padrão para a ocorrência dos sinistros com vítimas e sem vítimas, pois a base de dados da

PRF não contempla mais informações que poderiam ser importantes na análise, como por exemplo, a velocidade na via, peso do caminhão, uso do cinto de segurança no momento do sinistro, entre outras. Além disso, muitas vítimas consideradas como feridos graves podem ter evoluído ao óbito nos 30 dias depois do sinistro, o que não pode ser identificado na base de dados da PRF.

Outra limitação foi em relação à análise dos dados envolvendo o estado físico das vítimas. Foram englobados todos os envolvidos no sinistro e não apenas os dados dos condutores de caminhões. Porém, o estado físico dessas vítimas depende também de outros fatores que influenciam no sinistro como, por exemplo, a velocidade do veículo no momento do sinistro, uso do cinto de segurança, limite de velocidade na via, peso do caminhão, entre outros.

As limitações relacionadas à análise sob a ótica da segurança viária foram relacionadas principalmente à ausência de dados de volume de tráfego (exposição) associados aos locais onde ocorreram os sinistros e/ou às condições de circulação (como fase do dia e condição meteorológica, por exemplo). Da mesma forma, não foram quantificadas as extensões de trechos retos e curvos – esforço que estaria fora do escopo desta pesquisa. Também não foi considerada a caracterização da população de condutores, os quais podem concentrar-se em determinada faixa etária e sexo. Esta associação seria bastante difícil, tendo em vista que circulam nas rodovias federais do Paraná não apenas condutores paranaenses, mas também condutores de outras unidades da federação.

REFERÊNCIAS

AAMODT, A.; PLAZA, E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. **AI communications**, v. 7, n. 1, p. 39–59, 1994.

ABELLÁN, J.; LÓPEZ, G.; DE OÑA, J. Analysis of traffic accident severity using decision rules via decision trees. **Expert Systems with Applications**, v. 40, n. 15, p. 6047-6054, 2013.

AGÊNCIA NACIONAL DE TRANSPORTES TERRESTRES (ANTT). **CNSO Rodovias.** 2020. Disponível em <https://cnsoestrategico.antt.gov.br/SASReportViewer/?reportUri=%2Freports%2Freports%2F8138c8c6-7edd-4dee-9c26 669805a3a839&page=vi173&sso_guest=true&printEnabled=false&shareEnabled=false&informationEnabled=false&commentsEnabled=false&alertsEnabled=false&reportViewOnly=true&reportContextBar=false>. Acesso em: 11 mai. 2020.

AGÊNCIA NACIONAL DE TRANSPORTES TERRESTRES (ANTT). **CNSO Rodovias.** 2022. Disponível em < https://portal.antt.gov.br/resultado/-/asset_publisher/m2By5inRuGGs/content/id/1851367>. Acesso em: 18 jan. 2022.

ALES, V. T. **O algoritmo sequential minimal optimisation para resolução do problema de support vector machine: uma técnica para reconhecimento de padrões**. 2008. Tese de Doutorado. Master's thesis, Dissertação (Mestrado em Ciências)–Universidade Federal do Paraná, Curitiba.

ALKHEDER, S.; ALRUKAIBI, F.; AIASH, A. Risk analysis of traffic accidents' severities: An application of three data mining models. **ISA transactions**, v. 106, p. 213-220, 2020.

ALKHEDER, S.; TAAMNEH, M.; TAAMNEH, S. Severity prediction of traffic accident using an artificial neural network. **Journal of Forecasting**, v. 36, n. 1, p. 100-108, 2017.

ALMEIDA, N. C. **Sistema inteligente para diagnóstico de patologias na laringe utilizando máquinas de vetor de suporte.** 2010. Dissertação de Mestrado. (Mestrado em Engenharia Elétrica e da Computação)-Universidade Federal do Rio Grande do Norte, Natal.

AL-TURAIKI, I.; ALOUMI, M.; ALOUMI, N.; ALGHAMDI, K. Modeling traffic accidents in Saudi Arabia using classification techniques. In: **2016 4th Saudi International Conference on Information Technology (Big Data Analysis) (KACSTIT)**. IEEE, 2016. p. 1-5. 2016.

ALVES, C. A.; DUARTE, E. N. A relação entre a Ciência da Informação e a Ciência da Administração. **TransInformação**, v. 27, n. 1, p. 37-46, 2015.

AN, S., ZHANG, T., ZHANG, X., & WANG, J. Unrecorded accidents detection on highways based on temporal data mining. **Mathematical Problems in Engineering**, v. 2014, 2014.

ASSOCIAÇÃO BRASILEIRA DE CONCESSIONÁRIAS DE RODOVIAS (ABCR). **Setor em números – Acidentes de Tráfego.** 2020. Disponível em < https://abcr.org.br/institucional/biblioteca/relatorios>. Acesso em: 11 mai. 2020.

ASSOCIAÇÃO BRASILEIRA DE MEDICINA DE TRÁFEGO. ABRAMET. **ABNT muda terminologia e adota a expressão sinistro de trânsito para qualificar incidentes no tráfego.** 2021. Disponível em: <https://www.abramet.com.br/noticias/abnt-muda-terminologia-e-adota-a-expressao-sinistro-de-transito-para-qualificar-incidentes-no-trafego/>. Acesso em: 22 abr. 2021.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR ISO 9001:2015. Sistemas de Gestão da Qualidade –Requisitos.** Rio de Janeiro, p.10. 2015.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 10697:2020 Pesquisa de Sinistros de Trânsito – Terminologia.** Rio de Janeiro, p. 01. 2020.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 39001: 2015. Sistemas de Gestão Segurança Viária SV – Requisitos com orientações para uso.** Rio de Janeiro, p. 01. 2015.

AZIZ, A. S. A.; SANAA, E. L.; HASSANIEN, A. E. Comparison of classification techniques applied for network intrusion detection and classification. **Journal of Applied Logic**, v. 24, p.109-118, 2017.

BARROSO JUNIOR, G. T.; BERTHO, A. C. S.; VEIGA, A. C. A letalidade dos acidentes de trânsito nas rodovias federais brasileiras em 2016. **Revista Brasileira de Estudos de População**, v. 36, 2019.

BASTOS, J. T.; GARONCE, F. V.; SANTOS, P. A. B.; IGARASHI, A. V.; ANDRADE, G. A. M. Desempenho brasileiro na década de ação pela segurança no trânsito: análise, perspectivas e indicadores 2011- 2020. Brasília: Viva editora, 2020.

BLANCO, E. E.; PAIVA, E. L.; WANKE, P. F. Efficiency drivers in the Brazilian trucking industry: a longitudinal study from 2002-2010. **International Journal of Physical Distribution & Logistics Management**, 2014.

BLAZQUEZ, C. A; PICARTE, B.; CALDERÓN, J. F.; LOSADA, F. Spatial autocorrelation analysis of cargo trucks on highway crashes in Chile. **Accident Analysis & Prevention**, v. 120, p. 195-210, 2018.

BJÖRNSTIG, U.; BJÖRNSTIG, J.; ERIKSSON, A. Passenger car collision fatalities–with special emphasis on collisions with heavy vehicles. **Accident Analysis & Prevention**, v. 40, n. 1, p. 158-166, 2008.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the fifth annual workshop on Computational learning theory.** 1992. p. 144-152.

BRASIL. Lei nº. 9.503, de 23 de setembro de 1997. Institui o Código de Trânsito Brasileiro. **Diário Oficial da União**, Brasília, DF, 24 set. 1997. Seção 1, P. 21201.

BRASIL. Lei nº 13.103, de 02 de março de 2015. Dispõe sobre o exercício da profissão de motorista; altera a Consolidação das Leis do Trabalho - CLT, aprovada pelo Decreto-Lei no 5.452, de 1o de maio de 1943, e as Leis nos 9.503, de 23 de setembro de 1997 - Código de Trânsito Brasileiro, e 11.442, de 5 de janeiro de 2007 (empresas e transportadores autônomos de carga), para disciplinar a jornada de trabalho e o tempo de direção do motorista profissional; altera a Lei no 7.408, de 25 de novembro de 1985; revoga dispositivos da Lei no 12.619, de 30 de abril de 2012; e dá outras providências. **Diário Oficial da União.** Brasília, DF, 03 mar. 2015. Seção 1, p. 1.

BRASIL. Ministério da Infraestrutura. **Década Mundial de Ações Para a Segurança do Trânsito - 2011/2020: Juntos Podemos Salvar Milhões de Vidas**. 2019. Disponível em: < https://www.gov.br/infraestrutura/pt-br/assuntos/transito/conteudo-denatran/semana-nacional-de-transito-2011-denatran#:~:text=A%20Assembl%C3%A9ia%20Geral%20da%20Organiza%C3%A7%C3%A3o,fatalidades%20e%20ferimentos%20graves%20em > Acesso em: 11 mai. 2021.

BRASIL. Ministério da Infraestrutura. **Resolução n° 525 de 29 de abril de 2015**. Disponível em: <https://www.gov.br/infraestrutura/pt-br/assuntos/transito/conteudo-denatran/resolucoes-contran> Acesso em: 11 mai. 2021.

BRASIL. Ministério da Infraestrutura. **Resolução n° 691 de 28 de setembro de 2017**. Disponível em: <https://www.gov.br/infraestrutura/pt-br/assuntos/transito/conteudo-denatran/resolucoes-contran> Acesso em: 11 mai. 2021.

BRASIL. Ministério da Infraestrutura. **Resolução CONTRAN nº 808, de 15 de dezembro de 2020**. Publicado em: 24/12/2020, Edição: 246, Seção: 1, Página: 122, Ministério da Infraestrutura/Conselho Nacional de Trânsito. Disponível em <https://www.in.gov.br/web/dou/-/resolucao-contran-n-808-de-15-de-dezembro-de-2020- 296172888> Acesso em 12 dez. 2022.

BRASIL. Ministério da Infraestrutura. **Plano Nacional de Logística e Transporte: PNL 2035.** Brasília, 2021. Disponível em: < https://www.gov.br/infraestrutura/pt-br/assuntos/politica-e-planejamento/politica-e-planejamento/RelatorioExecutivoPNL_2035final.pdf> Acesso em: 25 jan.2023.

BRASIL. Polícia Rodoviária Federal. PRF. **Dados Abertos.** 2022. Disponível em: < https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos> Acesso em: 11 Set. 2022.

BREIMAN, L., FRIEDMAN, J., STONE, C. J., & OLSHEN, R. A. **Classification and regression trees**. CRC press, 1984.

BUENDIA, R., CANDEFJORD, S., FAGERLIND, H., BÁLINT, A., & SJÖQVIST, B. A. On scene injury severity prediction (OSISP) algorithm for car occupants. **Accident Analysis & Prevention**, v. 81, p. 211-217, 2015

BUNEMAN, P. Semistructured data. In: **Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems**. 1997. p. 117-121.

BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, v. 2, n. 2, p. 121-167, 1998.

CAI, Q. Cause analysis of traffic accidents on urban roads based on an improved association rule mining algorithm. **IEEE Access**, v. 8, p. 75607-75615, 2020.

CALIL, L. A. D. A.; CARVALHO, D. R.; SANTOS, C. B. D.; & VAZ, M. S. M. G. **Mineração de dados e pós-processamento em padrões descobertos.** Publ. UEPG Ci. Exatas Terra, Ci. Agr. Eng., Ponta Grossa, p. 207-215, 2008.

CAMILO, C. O.; SILVA, J. C.MMineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1-29, 2009.

CANDEFJORD, S., BUENDIA, R., FAGERLIND, H., BALINT, A., WEGE, C., & SJÖQVIST, B. A. On-scene injury severity prediction (OSISP) algorithm for truck occupants. **Traffic injury prevention**, v. 16, n. sup2, p. S190-S196, 2015.

CÂNDIDO, A. C; VALE, M. A. Práticas de gestão da informação e inovação aberta em um polo tecnológico brasileiro. **Perspectivas em Ciência da Informação**, v. 23, n. 4, p. 184-204, 2018.

CARDOSO, P. V.; DA SILVA SEABRA, V.; BASTOS, I. B.; COSTA, E. D. C. P. A importância da análise espacial para tomada de decisão: um olhar sobre a pandemia de COVID-19. **Revista Tamoios**, v. 16, n. 1, 2020.

CARVALHO, C. H. R. **Custos dos acidentes de trânsito no Brasil: estimativa simplificada com base na atualização das pesquisas do Ipea sobre custos de acidentes nos aglomerados urbanos e rodovias. 2020.** Disponível em: <https://www.ipea.gov.br/atlasviolencia/arquivos/artigos/7018-td2565.pdf> Acesso em: 22 de abril de 2021.

CARVALHO, A. P.; PEDROSA, E. M. Satisfação dos usuários com o acolhimento implantado em uma unidade de saúde da família. **Revista Enfermagem Digital Cuidado e Promoção da Saúde-2015 Jan-Jun**, v. 1, n. 1, p. 37-42, 2015.

CARVALHO, M. M.; RABECHINI, R. **Fundamentos em gestão de projetos: construindo competências para gerenciar projetos.** Atlas. São Paulo. 2019.

CASTRO, L. N.; FERRARI, D. G. Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva 2016.

CASTRO, Y.; KIM, Y. J. Data mining on road safety: factor assessment on vehicle accidents using classification models. **International Journal of Crashworthiness**, v. 21, n. 2, p. 104-111, 2016.

CHAGAS, D. M. **Estudo sobre fatores contribuintes de acidentes de trânsito urbano.** Dissertação (Mestrado em Engenharia de Produção) — Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, v. 41, n. 3, p. 1-58, 2009.

CHANG, L.Y.; CHIEN, J. T. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. **Safety science**, v. 51, n. 1, p. 17-22, 2013.

CHANG, L. Y.; MANNERING, F. Analysis of injury severity and vehicle occupancy in truck-and non-truck-involved accidents. **Accident Analysis & Prevention**, v. 31, n. 5, p. 579-592, 1999.

CHAPMAN, G. B.; SONNENBERG, F. A. **Decision making in health care: theory, psychology, and applications**. Cambridge University Press, 2003.

CHAWLA, S.; SHARMA, R. Application of Data Mining in Bioinformatics. **International Journal of Engineering Science and Computing**, v. 6, n. 6, p. 7426-9. jun. 2016.

CHEN, L., HUANG, S., YANG, C., & CHEN, Q. Analyzing factors that influence expressway traffic crashes based on association rules: using the Shaoyang–Xinhuang section of the Shanghai–Kunming expressway as an example. **Journal of transportation engineering, Part A: Systems**, v. 146, n. 9, p. 05020007, 2020.

CHIUSOLI, C. L.; REZENDE, D. A. Sistema de informações municipais como apoio à tomada de decisões dos cidadãos. **Navus-Revista de Gestão e Tecnologia**, v. 9, n. 3, p. 124-142, 2019.

CHOO, C. W. **Gestão de informação para a organização inteligente: a arte de explorar o meio ambiente.** Lisboa: Caminho. 2003a.

CHOO, C. W. **A organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões**. São Paulo: Senac São Paulo, 2003b.

CIOS, K. J.; PEDRYCZ, W.; SWINIARSKI, R. W.; & KURGAN, L. A. **Data mining: a knowledge discovery approach**. Springer Science & Business Media, 2007.

CNI. Confederação Nacional da Indústria. **Sinistros com Veículos de Carga e Obras da União em Rodovias Federais.** Disponível em: <https://www.portaldaindustria.com.br/publicacoes/2020/6/sinistros-com-veiculos-de-carga-e-obras-da-uniao-em-rodovias-federais/>. Acesso em: 24 fev. 2021. 2020.

CNT. Confederação Nacional do Transporte. **Anuário do transporte 2021.** Disponível em: < https://anuariodotransporte.cnt.org.br/2021/>. Acesso em: 24 nov. 2022. 2022.

CNT. Confederação Nacional do Transporte. **Sinistros Rodoviários Estatísticas Envolvendo Caminhões.** 2019. Disponível em: <https://cnt.org.br/sinistros-rodoviarios-caminhoes>. Acesso em: 24 fev. 2021. 2019.

CNT. Confederação Nacional do Transporte. **Transporte em Números.** 2019. Disponível em: <https://www.cnt.org.br/analises-transporte>. Acesso em: 24 fev. 2021. 2019a.

CNT. Confederação Nacional do Transporte. **Pesquisa CNT de Rodovias 2019. Relatório Gerencial.** 2019. Disponível em: <https://www.cnt.org.br/analises-transporte>. Acesso em: 24 fev. 2021. 2019b.

COLONNA, P.; INTINI, P. Compensation effect between deaths from COVID-19 and crashes: the Italian case. **Transportation research interdisciplinary perspectives**, v. 6, p. 100170, 2020.

CONCA, A.; RIDELLA, C.; SAPORI, E. A risk assessment for road transportation of dangerous goods: a routing solution. **Transportation Research Procedia**, v. 14, p. 2890-2899, 2016.

COPPIN, B. **Inteligência artificial.** Rio de Janeiro, RJ: LTC, 2010.

CORCOVIA, L. O.; ALVES, R. S. APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS. **Revista Interface Tecnológica**, v. 16, n.1, p. 90-101, 2019.

CORRAR, L. J. O modelo econômico da empresa em condições de incerteza aplicação do método de simulação de Monte Carlo. **Caderno de estudos**, p. 01-11, 1993.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, v. 20, n. 3, p. 273-297, 1995.

COSTA, C. N.; COUTINHO, J. V.; DE MAGALHÃES, L. H.; & ARBEX, M. A. Descoberta de conhecimento em bases de dados. **Revista Eletrônica: Faculdade Santos Dumont**, v. 2, 2019.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines and other kernel-based learning methods**. Cambridge university press, 2000.

DA CRUZ FIGUEIRA, A., PITOMBO, C. S., & LAROCCA, A. P. C. Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil. **Case studies on transport policy**, v. 5, n. 2, p. 200-207, 2017.

DA SILVA, R. A.; ALMEIDA SILVA, F. C.; SIMOES GOMES, C. F. Using Business Intelligence (Bi) In Making Support Sistma Strategic Decision. **Revista Geintec-Gestão Inovação e Tecnologias**, v. 6, n. 1, p. 2780-2798, 2016.

DA SILVA, A. B.; DE OLIVEIRA Í. M. Soluções Mobile, Algoritmos Genéticos e o Problema do Caixeiro-Viajante: Um Estudo Bibliométrico. **InterSciencePlace**, v. 15, n. 3, 2020.

DAVENPORT, T. H. **Ecologia da informação: porque só a tecnologia não basta para o sucesso na era da informação**. 4. ed. São Paulo: Futura, 2002.

DE ABREU, A. M.; SCHINAIDER, A. D. A ECONOMIA COMPARTILHADA NO TRANSPORTE RODOVIÁRIO DE CARGAS. **Facit Business and Technology Journal**, v. 1, n. 18, 2020.

DE JESUS COSTA, J.; BERNARDINI, F. C.; VITERBO FILHO, J. A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. **AtoZ: novas práticas em informação e conhecimento**, v. 3, n. 2, p. 139-157, 2014.

DE LA VEGA, D. S.; VIEIRA, J. G. V.; TOSO, E. A. V.; FARIA, R. N. A decision on the truckload and less-than-truckload problem: An approach based on MCDA. **International Journal of Production Economics**, v. 195, p. 132-145, 2018.

DE LIMA, J. F. **Modelo Fuzzy Para Avaliação De Imóveis Utilizando Árvore de Decisão**. Dissertação de Mestrado (Programa de Pós-Graduação em Engenharia de Processos) - Universidade Federal do Pará, Pará, 2017.

DE MELO, W. A., ALARCÃO, A. C. J., DE OLIVEIRA, A. P. R., PELLOSO, S. M., & CARVALHO, M. D. D. B. Age-related risk factors with nonfatal traffic accidents in urban areas in Maringá, Paraná, Brazil. **Traffic injury prevention**, v. 18, n. 2, p. 157-163, 2017.

DE OÑA, J.; LÓPEZ, G.; ABELLÁN, J. Extracting decision rules from police accident reports through decision trees. **Accident Analysis & Prevention**, v. 50, p. 1151-1160, 2013.

DE OÑA, J.; MUJALLI, R. O.; CALVO, F. J. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. **Accident Analysis & Prevention**, v. 43, n. 1, p. 402-411, 2011.

DE SOUZA JUNIOR, C. A.; VILLELA, H. F. Um estudo sobre Publicações Relacionadas a Mineração de Dados. **Computação & Sociedade**, v. 1, n. 1, 2019.

DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES. DNIT. **Resolução n° 210 de 13 de novembro de 2006.** Disponível em: <http://189.9.128.64/rodovias/operacoes-rodoviarias/sistema-de-gerenciamento-de-autorizacao-especial-de-transito-siaet/RESOLUCAO2102006CONTRANCONSOLIDADA.rtf/view>. Acesso em: 11 mai 2021.

DIAS, M. M. **Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados**. Tese (Doutorado) - Curso de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2001.

DOĞRU, N.E.J.D.E.T.; SUBAŞI, A. Comparison of clustering techniques for traffic accident detection. **Turkish Journal of Electrical Engineering & Computer Sciences**, v. 23, n. Sup 1, p. 2124-2137, 2015.

DUDA, P.; RUTKOWSKI, L.; JAWORSKI, M.; RUTKOWSKA, D. On the Parzen kernel-based probability density function learning procedures over time-varying streaming data with applications to pattern classification. **IEEE transactions on cybernetics**, v. 50, n. 4, p. 1683-1696, 2018.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification and scene analysis**. New York: Wiley, 1973.

DUNHAM, M. H. **Data mining: introductory and advanced topics.** Upper Saddle River, p. 1–6, 2003.

EBRAHIMPOUR, H.; KOUZANI, A. Face recognition using bagging KNN. In: **International conference on signal processing and communication systems (ICSPCS'2007) Austrália, gold coast**. sn, 2007. p. 17-19.

EL MAZOURI, F. Z.; ABOUNAIMA, M. C.; ZENKOUAR, K. Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France. **Journal of Big Data**, v. 6, p. 1-30, 2019.

ELVIK, R.; VAA, T.; HOYE, A.; SORENSEN, M. **The handbook of road safety measures.** Emerald Group Publishing, 2009.

ERNSTBERGER, A., JOERIS, A., DAIGL, M., KISS, M., ANGERPOINTNER, K., NERLICH, M., & SCHMUCKER, U. Decrease of morbidity in road traffic accidents in a high income country–an analysis of 24,405 accidents in a 21 year period. **Injury**, v. 46, p. S135-S143, 2015.

ESTER, M.; KRIEGEL, H. P.; SANDER, J.; & XU, X. **A density-based algorithm for discovering clusters in large spatial databases with noise.** In: kdd. 1996. p. 226-231.

FACELI, K.; LORENA, A. C.; GAMA, J.; & CARVALHO, A. C. P. D. L. **Inteligência Artificial: Uma abordagem de aprendizado de máquina.** Rio de Janeiro: LTC, 2011.

FARHANGI, F., SADEGHI-NIARAKI, A., NAHVI, A., & RAZAVI-TERMEH, S. V. Spatial modelling of accidents risk caused by driver drowsiness with data mining algorithms. **Geocarto International**, v. 37, n. 9, p. 2698-2716, 2022.

FAUSETT, L. V. **Fundamentals of neural networks: architectures, algorithms and applications**. Pearson Education India, 2006.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; & SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v.17, n. 3, p. 37-37, 1996.

FERNANDES, F. T.; CHIAVEGATTO FILHO, A. D. P. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. **Revista Brasileira de Saúde Ocupacional**, v. 44, 2019.

FERRAZ, A. C. P.; RAIA JR, A.; BEZERRA, B.; BASTOS, T.; RODRIGUES, K. **Segurança viária.** São Carlos: Suprema Gráfica e Editora, 2012.

FORKENBROCK, D. J.; HANLEY, P. F. Fatal crash involvement by multiple-trailer trucks. **Transportation Research Part A: Policy and Practice**, v. 37, n. 5, p. 419-433, 2003.

FRAGOSO, A.; GARCIA, E. G. Transporte rodoviário de carga: sinistros de trabalho fatais e fiscalização trabalhista. **Revista Brasileira de Saúde Ocupacional**, v. 44, p. 1-12, 2019.

FURTADO, M. I. V. **Redes neurais artificiais: uma abordagem para sala de aula.** Atena Editora, 2019.

GALVÃO, N. D.; DE FÁTIMA MARIN, H. Características das vítimas de sinistro de trânsito por meio da técnica da mineração de dados. **Journal of Health Informatics**, v. 2, n. 4, 2010.

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

GJERDE, H., NORMANN, P. T., CHRISTOPHERSEN, A. S., SAMUELSEN, S. O., & MØRLAND, J. Alcohol, psychoactive drugs and fatal road traffic accidents in Norway: a case–control study. **Accident Analysis & Prevention**, v. 43, n. 3, p. 1197-1203, 2011.

GOLDBERG, D. E. **Genetic algorithms.** [S.l.]: Pearson Education India, 2006.

GOLDSCHMIDT, R. R.; PASSOS, E. **Data Mining: Um Guia Prático.** Rio de Janeiro: Campus, 2005.

GONZALEZ, G. H.; TAHSIN, T.; GOODALE, B. C.; GREENE, A. C.; & GREENE, C. S. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. **Briefings in Bioinformatics**, v. 17, n. 1, p. 33-42, 2016.

GREGORIADES, A.; MOUSKOS, K. C. Black spots identification through a Bayesian Networks quantification of accident risk index. **Transportation Research Part C: Emerging Technologies**, v. 28, p. 28-43, 2013.

GRISELDA, L., & JOAQUÍN, A. Using decision trees to extract decision rules from police reports on road accidents. **Procedia-social and behavioral sciences**, v. 53, p. 106-114, 2012.

GUPTA, M.; SOLANKI, V. K.; SINGH, V. K. A novel framework to use association rule mining for classification of traffic accident severity. **Ingeniería solidaria**, v. 13, n. 21, p. 37-44, 2017.

HADDON JR., W. Advances in the epidemiology of injuries as a basis for public policy. **Public health reports**, v. 95, n. 5, p. 411, 1980.

HÄKKÄNEN, H.; SUMMALA, H. Fatal traffic accidents among trailer truck drivers and accident causes as viewed by other truck drivers. **Accident Analysis & Prevention**, v. 33, n. 2, p. 187-196, 2001.

HAN, J.“**Data Mining: Concepts and Techniques**” Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2005.

HAN, J.; KAMBER, M.; PEI, J. Clustering analysis. **Data Mining: Concept and Technique, MK imprint of Elsevier, New York**, p. 478-490, 2012.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. Elsevier, 2011.

HARRINGTON, P. **Aprendizado de máquina em ação** . Simon e Schuster, 2012.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2007.

HE, Y. Q.; RONG, Y. L.; LIU, Z. P.; DU, S. P. Traffic influence degree of urban traffic accident based on speed ratio. **Journal of highway and transportation research and development (English edition)**, v. 13, n. 3, p. 96-102, 2019.

HERBRICH, R. **Learning kernel classifiers: theory and algorithms**. MIT press, 2001.

HODEGHATTA, U. R.; NAYAK, U. **Business analytics using R-a practical approach**. Apress, 2016.

HOFFMANN, W. A. M. Gestão do conhecimento e da informação em organizações baseados em inteligência competitiva. **Ciência da Informação**, v. 45, n. 3, 2016.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, Rodney X. **Applied Logistic Regression.** John Wiley & Filhos, 2013.

HUSSAIN, F.; LIU, H.; SUZUKI, E.; LU, H. Exception rule mining with a relative interestingness measure. In: **Pacific-Asia Conference on Knowledge Discovery and Data Mining**. Springer, Berlin, Heidelberg, 2000. p. 86-97.

INFOSIGA. **Sistema de Informações Gerenciais de Acidentes de Trânsito do Estado de São Paulo.** São Paulo, SP. 2020. Disponível em: <http://painelderesultados.infosiga.sp.gov.br/dados.web/ViewPage.do?name=obitos_publico&contextId=8a80809939587c0901395881fc2b0004>. Acesso em: 11 mai. 2021.

ISLAM, M.; HERNANDEZ, S. Large truck–involved crashes: exploratory injury severity analysis. **Journal of Transportation Engineering**, v. 139, n. 6, p. 596-604, 2013.

ISLAM, S.; JONES, S. L.; DYE, D. Comprehensive analysis of single-and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. **Accident Analysis & Prevention**, v. 67, p. 148-158, 2014.

JARAŠŪNIENĖ, A.; BATARLIENĖ, N.; VAIČIŪTĖ, K. Application and Management of Information Technologies in Multimodal Transportation. **Procedia Engineering**, v. 134, p. 309-315, 2016.

JOHN, M.; SHAIBA, H. Apriori-Based Algorithm for Dubai Road Accident Analysis. **Procedia Computer Science**, v. 163, p. 218-227, 2019.

KALMEGH, S. R. Effective classification of Indian News using Lazy classifier IB1And IBk from weka. **Journal of information and computing science**, v. 6, p. 160-168, 2019.

KANTARDZIC, M. **Data mining: concepts, models, methods, and algorithms**. John Wiley & Sons, 2011.

KATRAKAZAS, C.; MICHELARAKI, E.; SEKADAKIS, M.; YANNIS, G. A descriptive analysis of the effect of the COVID-19 pandemic on driving behavior and road safety. **Transportation research interdisciplinary perspectives**, v. 7, p. 100186, 2020.

KHORASHADI, A.; NIEMEIER, D.; SHANKAR, V.; MANNERING, F. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. **Accident Analysis & Prevention**, v. 37, n. 5, p. 910-921, 2005.

KOGANI, M., ALMASI, S. A., ANSARI-MOGADDAM, A., DALVAND, S., OKATI-ALIABAD, H., TABATABAEE, S. M., & ALMASI, S. Z. Relationship between using cell phone and the risk of accident with motor vehicles: An analytical cross-sectional study. **Chinese journal of traumatology**, v. 23, n. 06, p. 319-323, 2020.

KONZEN, R. **A nova jornada de trabalho do motorista profissional regulamentada pela Lei n. 13.103 de 02 de março de 2015 e suas**

particularidades. Monografia (Disciplina Trabalho de Curso I) - Curso de Direito da Universidade de Santa Cruz do Sul, UNISC, Santa Cruz do Sul, 2017.

KORAMATI, S., MAJUMDAR, B. B., PANI, A., & SAHU, P. K. A registry-based investigation of road traffic fatality risk factors using police data: A case study of Hyderabad, India. **Safety science**, v. 153, p. 105805, 2022.

KRAEMER, M. U.; YANG, C. H.; GUTIERREZ, B.; WU, C. H.; KLEIN, B., PIGOTT, D. M.; ... & SCARPINO, S. The effect of human mobility and control measures on the COVID-19 epidemic in China. **Science**, v. 368, n. 6490, p. 493-497, 2020.

KUMAR, S.; TIWARI, P.; DENIS, K. V. Augmenting classifiers performance through clustering: A comparative study on road accident data. **International Journal of Information Retrieval Research (IJIRR)**, v. 8, n. 1, p. 57-68, 2018.

KWON, O. H.; RHEE, W.; YOON, Y. Application of classification algorithms for analysis of road safety risk factor dependencies. **Accident Analysis & Prevention**, v. 75, p. 1-15, 2015.

LANDIS, J. R.; KOCH, G. G. The Measurement of Observer Agreement for Categorical Data. **Biometria**, pág. 159-174, 1977.

LANTZ, B. **Machine learning with R: expert techniques for predictive modeling**. Editora Packt Ltda, 2019.

LAROSE, D. T. **Discovering Knowledge in Data:** An Introduction to Data mining. New Jersey: John Willey & Sons. Inc., Publication, 2005.

LAROSE, D.T.; LAROSE, C. D. **Discovering knowledge in data: an introduction to data mining**. John Wiley & Sons, 2014.

LARSON, E. W.; GRAY, C. F. **Project management: The managerial process**. 2011.

LAU, H.; KHOSRAWIPOUR, V.; KOCBACH, P.; MIKOLAJCZYK, A.; SCHUBERT, J.; BANIA, J.; KHOSRAWIPOUR, T. The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. **Journal of travel medicine**, v. 27, n. 3, p. 37, 2020.

LEE, J.; YOON, T.; KWON, S.; LEE, J. Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. **Applied Sciences**, v. 10, n. 1, p. 129, 2020.

LEMP, J. D.; KOCKELMAN, K. M.; UNNIKRISHNAN, A. Analysis of large truck crash severity using heteroskedastic ordered probit models. **Accident Analysis & Prevention**, v. 43, n. 1, p. 370-380, 2011.

LI, J.; HE, J.; LIU, Z.; ZHANG, H.; ZHANG, C. Traffic accident analysis based on C4. 5 algorithm in WEKA. In: **MATEC Web of Conferences**. EDP Sciences, 2019. p. 01035.

LIN, L.; WANG, Q.; SADEK, A. W. Data mining and complex network algorithms for traffic accident analysis. **Transportation Research Record**, v. 2460, n. 1, p. 128-136, 2014.

LÍDER, Seguradora. **Relatório Anual 2020.** 2021. Disponível em < https://www.seguradoralider.com.br/Documents/RelatorioAnual/Relatorio%20Anual%20-%202020%20v3.pdf?#zoom=65%>. Acesso em: 11 mai. 2021.

LIU, W.; WANG, Z.; LIU, X.; ZENG, N.; LIU, Y.; & ALSAADI, F. E. A survey of deep neural network architectures and their applications. **Neurocomputing**, v. 234, p. 11-26, 2017.

LOPEZ, E. R. A. **Localização de aterro sanitário baseado em modelo de decisão multicritério**. 2017. Dissertação (Programa de Pós-Graduação em Engenharia de Produção). Universidade Federal de Pernambuco. 2017.

LORENA, A. C.; DE CARVALHO, A. C. P. L. F. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43-67, 2007.

MAFI, S.; ABDELRAZIG, Y.; DOCZY, R. Machine learning methods to analyze injury severity of drivers from different age and gender groups. **Transportation research record**, v. 2672, n. 38, p. 171-183, 2018.

MARTINS, M. A.; GARCEZ, T. V. **Priorização das causas de acidentes de trânsito em rodovias federais de Pernambuco: uma abordagem multicritério e multiperíodo.** In: XXXIX Encontro Nacional de Engenharia de Produção, 2019, Santos/SP. Anais do XXXIX Encontro Nacional de Engenharia de Produção, 2019.

MBAKWE, A. C., SAKA, A. A., CHOI, K., & LEE, Y. J. Alternative method of highway traffic safety analysis for developing countries using delphi technique and Bayesian network. **Accident Analysis & Prevention**, v. 93, p. 135-146, 2016.

MCGEE, J. V.; PRUSAK, L. **Gerenciamento estratégico da informação.** Campus, Rio de Janeiro, 1994.

MELANDA, E. A. **Pós-processamento de regras de associaçao**. 2004. Tese (Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional). Universidade de São Paulo. 2004.

MENDONÇA, T. C.; VARVAKIS, G. Análise do uso da informação para tomada de decisão gerencial em gestão de pessoas: estudo de caso em uma instituição bancária. **Perspectivas em Ciência da Informação**, v. 23, n. 1, p. 104-119, 2018.

MITCHELL, M. **An Introduction to Genetic Algorithms.** MIT Press, Cambridge, Massachussetts, 1998.

MOKHTAR, W.; PERVEZ, N. **Underbody drag for pickup trucks.** In: 30th AIAA Applied Aerodynamics Conference**.** p. 3210. 2012.

MORADPOUR, S.; LONG, S. Using combined multi-criteria decision-making and data mining methods for work zone safety: a case analysis. **Case studies on transport policy**, v. 7, n. 2, p. 178-184, 2019.

MORSE, W. J. **Cost Accounting: Processing, Evaluating, and Using Cost Data**. Third Edition. Addison Wesley Publishing Company. 1986.

MOUSSA, G. S.; OWAIS, M.; DABBOUR, E. Variance-based global sensitivity analysis for rear-end crash investigation using deep learning. **Accident Analysis & Prevention**, v. 165, p. 106514, 2022.

MUJALLI, R. O.; LÓPEZ, G.; GARACH, L. Bayes classifiers for imbalanced traffic accidents datasets. **Accident Analysis & Prevention**, v. 88, p. 37-51, 2016.

MUJALLI, R. O.; OÑA, J. A. method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. **Journal of safety research**, v. 42, n. 5, p. 317-326, 2011.

NAFIE ALI, F. M.; MOHAMED, A. A. Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents. **Journal of Information and Telecommunication**, v. 2, n. 3, p. 231-245, 2018.

NALDI, M. C. **Técnicas de combinação para agrupamento centralizado e distribuído de dados.** 2010. 276f. Tese (Doutorado em Ciências Matemáticas e de Computação) - Universidade de São Paulo. São Paulo. 2010.

NANDI, J. C. B.; PEREIRA, R. M.; FELIPPE, G. O Algoritmo de Associação Frequent Pattern-Growth na Shell Orion Data Mining Engine. **Anais SULCOMP**, v. 7, 2015.

NASCIMENTO, P. B. **Recomendação de ação pedagógica no ensino de introdução à programação por meio de raciocínio baseado em casos.** 2018. Dissertação (Programa de Pós-Graduação em Informática). Universidade Federal do Amazonas. 2018.

NGUYEN, T. T.; ARMITAGE, G. A survey of techniques for internet traffic classification using machine learning. **IEEE communications surveys & tutorials**, v. 10, n. 4, p. 56–76, 2008.

NISBET, R.; ELDER, J.; MINER, G. **Handbook of statistical analysis and data mining applications**. Academic press, Oxford: Elsevier. 2009.

NOUR, M. K., NASEER, A., ALKAZEMI, B., & JAMIL, M. A. Road traffic accidents injury data analytics. **International Journal of Advanced Computer Science and Applications**, v. 11, n. 12, 2020.

OZDEMIR, S. **Principles of Data Science.** Packt Publishing, chapter 1, pp. 1-20. 2016.

PAHUKULA, J.; HERNANDEZ, S.; UNNIKRISHNAN, A. A time of day analysis of crashes involving large trucks in urban areas. **Accident Analysis & Prevention**, v. 75, p. 155-163, 2015.

PAKGOHAR, A.; TABRIZI, R. S.; KHALILI, M.; & ESMAEILI, A. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. **Procedia Computer Science**, v. 3, p. 764-769, 2011.

PANITZ, M. A. **A segurança viária e o fator humano: verificação da presença de álcool-direção no sistema de transporte rodoviário do RGS.** Dissertação (Mestrado) - Curso de Pós-graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1999.

PEERI, N. C.; SHRESTHA, N.; RAHMAN, M. S.; ZAKI, R.; TAN, Z.; BIBI, S.; ... & HAQUE, U. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? **International journal of epidemiology**, v. 49, n. 3, p. 717-726, 2020

PÉRA, T.G.; ROCHA, F.V.; BASTIANI, F.P.; SANTOS, R.M.; COSTA, E.V.; JOÃO, A.M.; CAIXETA-FILHO, J.V. **Análise do Excesso de Peso Entre Eixos no Transporte Rodoviário de Cargas.** Série Logística do Agronegócio – Oportunidades e Desafios, V.5, 39 p., Grupo ESALQ-LOG/USP, Piracicaba, Brasil, 2021.

POLÍCIA RODOVIÁRIA FEDERAL. PRF. **M-015: atendimento de acidentes de Trânsito.** Brasília: Polícia Rodoviária Federal: Versão 04. 2018.

QUINLAN, J. R. **Discovering rules by induction from large collections of** examples. Expert systems in the micro electronics age, Edinburgh University Press, 1979.

_____. **Induction of decision trees.** Machine learning, v. 1, n. 1, p. 81-106, 1986.

_____. **C4.5: programs for machine learning.** São Mateo: Morgan Kauffman, 1993.

RAMOS, S. Simulação de Monte Carlo na avaliação de incertezas de medição. **ESTATÍSTICO 2017 JÚNIOR**, p. 30. 2017.

RAMYA, M. C., LOKESH, V., MANJUNATH, T. N., & HEGADI, R. S. A Predictive model construction for mulberry crop productivity. **Procedia Computer Science**, v. 45, p. 156-165, 2015.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações.** Editora Manole Ltda., 2003.

ROVŠEK, V.; BATISTA, M.; BOGUNOVIĆ, B. Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree.

Transport, v. 32, n. 3, p. 272-281, 2017.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial.** 3. ed. Rio de Janeiro: Elsevier, 2013.

SALADIÉ, Ò.; BUSTAMANTE, E.; GUTIÉRREZ, A. COVID-19 lockdown and reduction of traffic accidents in Tarragona province, Spain. **Transportation research interdisciplinary perspectives**, v. 8, p. 100218, 2020.

SANTANA, V.; MOURA, M. C. P.; PEDRA, F.; CORRÊA, H.; VENÂNCIO, J.; & BELINO, L. Morbimortalidade por sinistros de trabalho em motoristas do transporte de carga, 2006-2012. **Boletim Epidemiológico Sinistros de Trabalho**, v. 3, n. 6, p. 1-4, 2013.

SANTOS, M. M. DE O.; COSTA, M. M. DA; BIANCHI, A. S. Mobilidade e Saúde: qual impacto da pandemia de COVID-19 no trânsito? **CadernoS de PsicologiaS**, Curitiba, n. 1, 2020. Disponível em: <https://cadernosdepsicologias.crppr.org.br/mobilidade-e-saude-qual-impacto-da-pandemia-de-COVID-19-no-transito>. Acesso em: 21/05/2021.

SANTOYO, A. H. **Bases teórico metodológicas para la valoración económica de bienes y servicios ambientales a partir de técnicas de decisión multicriterio.** Estudio de caso: Parque Nacional Viñales, Pinar del Río, República de Cuba. 2011.
SEDGEWICK, R. **Algorithms in java, part 5: Graph algorithms**. Addison-Wesley Professional, 2003.

SELVI, H. Z.; CAGLAR, B. Using cluster analysis methods for multivariate mapping of traffic accidents. **Open Geosciences**, v. 10, n. 1, p. 772-781, 2018.

SHI, Y.; ARTHANARI, T.; LIU, X.; YANG, B. Sustainable transportation management: Integrated modeling and support. **Journal of Cleaner Production**, v. 212, p. 1381-1395, 2019.

SHOKOHYAR, S.; TAATI, E.; ZOLFAGHARI, S. The effect of drivers' demographic characteristics on road accidents in different seasons using data mining. **Promet-Traffic&Transportation**, v. 29, n. 6, p. 555-567, 2017.

SILVA, L. Tomada de decisão baseada em dados (dddm) e Aplicações em Informática em Educação. **Jornada de Atualização em Informática na Educação**, v. 4, n. 1, p. 21-46, 2015.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados.** Elsevier Brasil, 2016.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais artificiais para engenharia e ciências aplicadas: curso prático. São Paulo: Artliber, 2010.

SIMOUDIS, E. **Reality check for data mining.** IEEE Expert, p. 26–33, 1996.

SOÁREZ, P. C.; SOARES, M. O.; NOVAES, H. M. D. Modelos de decisão para avaliações econômicas de tecnologias em saúde. **Ciência & Saúde Coletiva**, v. 19, n. 10, p. 4209-4222, 2014.

SOEANU, A.; DEBBABI, M.; ALHADIDI, D.; MAKKAWI, M.; ALLOUCHE, M.; BÉLANGER, M.; & LÉCHEVIN, N. Transportation risk analysis using probabilistic model checking. **Expert Systems with Applications**, v. 42, n. 9, p. 4410-4421, 2015.

SOUZA, F. A. A. **Análise de desempenho da rede neural artificial do tipo multilayer perceptron na era multicore.** 2012. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte.

SOUZA, J. C.; PAIVA T.; REIMÃO, R. Sleep habits, sleepiness and accidents among truck drivers. **Arquivos de Neuro-Psiquiatria**, v. 63, p. 925-930, 2005.

SUN, Z., XING, Y., WANG, J., GU, X., LU, H., & CHEN, Y. Exploring injury severity of vulnerable road user involved crashes across seasons: A hybrid method integrating random parameter logit model and Bayesian network. **Safety science**, v. 150, p. 105682, 2022.

TAAMNEH, M.; ALKHEDER, S.; TAAMNEH, S. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. **Journal of Transportation Safety & Security**, v. 9, n. 2, p. 146-166, 2017.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining: mineração de dados**. Ciência Moderna, Rio de Janeiro, 2009.

TAO, G.; SONG, H.; LIU, J.; ZOU, J.; CHEN, Y. A traffic accident morphology diagnostic model based on a rough set decision tree. **Transportation Planning and Technology**, v. 39, n. 8, p. 751-758, 2016.

THARWAT, A. Classification assessment methods. **Applied computing and informatics**, v. 17, n. 1, pág. 168-192, 2021.

TRANCHITELLA, F. B.; SANTOS, R.S.D.; EL BACHA, J.J.S.H.; SOBRADO, J.V.; SANTOS, M.B.S.D.; COLOMBO SOUZA, P. Mortalidade por Acidentes de Transporte no Município de São Paulo: 2005-2015. **Acta Ortopédica Brasileira**, v. 29, p. 193-196, 2021.

TSAI, M. C.; SU, C. C. Scenario analysis of freight vehicle accident risks in Taiwan. **Accident analysis & prevention**, v. 36, n. 4, p. 683-690, 2004.

VAPNIK, V. **The nature of statistical learning theory**. Springer science & business media, 1999.

VARGAS, R. V. **Gerenciamento de Projetos: Estabelecendo diferenciais competitivos.** Brasport, Rio de Janeiro, 2009.

VASCONCELOS, L. M. R.; CARVALHO, C. L. Aplicação de regras de associação para mineração de dados na web. **Revista Telfract**, v. 1, n. 1, 2018.

VINGILIS, E.; BEIRNESS, D.; BOASE, P.; BYRNE, P.; JOHNSON, J.; JONAH, B.; ... & WIESENTHAL, D. L. Coronavirus disease 2019: What could be the effects on Road safety? **Accident Analysis & Prevention**, v. 144, p. 105687, 2020.

VON WANGENHEIM, C. G.; VON WANGENHEIM, A.; RATEKE, T. **Raciocínio baseado em casos.** Editora Manole Ltda, 2013.

WARD, H.; LYONS, R.; THOREAU, R. Road safety research report no. 69, under-reporting of road casualties-phase 1. **Department for Transport (DfT)**, 2006.

WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. **Acm Sigmod Record**, v. 31, n. 1, pág. 76-77, 2002.

WITTEN, I. H., FRANK, E. e HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques.** Morgan Kaufmann Publishers Inc, 3ed, San Francisco CA USA, 2011.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; & PAL, C. J. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2016.

WORLD HEALTH ORGANIZATION (WHO). **Global status report on road safety 2018**. World Health Organization, 2018. Disponível em: <https://www.who.int/publications/i/item/9789241565684> Acesso em: 22 de abril de 2021.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B; PHILIP, S. Y.; ZHOU, Z. H.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and information systems**, v. 14, n. 1, p. 1-37, 2008.

WU, J. T.; LEUNG, K.; LEUNG, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. **The Lancet**, v. 395, n. 10225, p. 689-697, 2020a.

WU, W.; JIANG, S.; LIU, R.; JIN, W.; MA, C. Economic development, demographic characteristics, road network and traffic accidents in Zhongshan, China: gradient boosting decision tree model. **Transportmetrica A: transport science**, v. 16, n. 3, p. 359-387, 2020b.

XIAOBING, X.; CHAO, B.; FENG, C. An insight into traffic safety management system platform based on cloud computing. **Procedia-Social and Behavioral Sciences**, v. 96, p. 2643-2646, 2013.

YADAV, A. K.; CHANDEL, S. S. Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model. **Renewable Energy,** v. 75, p. 675-693, 2015.

YANG, Y., HE, K., WANG, Y. P., YUAN, Z. Z., YIN, Y. H., & GUO, M. Z. Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods. **Physica A: Statistical Mechanics and Its Applications**, v. 595, p. 127083, 2022.

ZALOSHNJA, E.; MILLER, T. R. Costs of large truck-involved crashes in the United States. **Accident Analysis & Prevention**, v. 36, n. 5, p. 801-808, 2004.

ZANNE, M.; GROZNIK, A.; TWRDY, E. Assessment of traffic safety among young people aged 15 to 24 in Slovenia. **Promet-Traffic&Transportation**, v. 25, n. 2, p. 147-156, 2013.

ZHANG, C.; SHU, Y.; YAN, L. A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo, China. **IEEE Access**, v. 7, p. 140197-140205, 2019.

ZHANG, G., YAU, K. K., ZHANG, X., & LI, Y. Traffic accidents involving fatigue driving and their extent of casualties. **Accident Analysis & Prevention**, v. 87, p. 34-42, 2016.

ZHENG, Z.; LU, P.; LANTZ, B. Commercial truck crash injury severity analysis using gradient boosting data mining model. **Journal of safety research**, v. 65, p. 115-124, 2018.

ZHOU, T.; ZHANG, J. Analysis of commercial truck drivers' potentially dangerous driving behaviors based on 11-month digital tachograph data and multilevel modeling approach. **Accident Analysis & Prevention**, v. 132, p. 105256, 2019.

ZHU, A. Artificial Neural Networks. The International Encyclopedia of Geography. **IEEE Computer Graphics and Applications**, v. 21, n. 6, p. 34-47, 2017.

ZHU, X.; SRINIVASAN, S. Modeling occupant-level injury severity: An application to large-truck crashes. **Accident Analysis & Prevention**, v. 43, n. 4, p. 1427-1437, 2011.

APÊNDICE 3 - ESTUDOS DE SINISTROS DE TRÂNSITO E MINERAÇÃO DE DADOS

Posição	Título	Autores	*Metodi In ordinatio*	Metodologia utilizada
1	Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks.	DE ONA, J., LÓPEZ, G., MUJALLI, R., & CALVO, F. J. (2013)	269,993	Rede Bayesiana e Agrupamento de classes latentes.
2	Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks.	DE OÑA, J.; MUJALLI, R. O.; CALVO, F. J. (2011)	265,993	Rede Bayesiana
3	Traffic accidents involving fatigue driving and their extent of casualties.	ZHANG, G.; YAU, K. K.; ZHANG, X.; LI, Y. (2016)	220,993	Regressão logística
4	Analysis of traffic accident severity using decision rules via decision trees.	ABELLÁN, J.; LÓPEZ, G.; DE OÑA, J. (2013)	209,954	Árvores de decisão
5	Alcohol, psychoactive drugs and fatal road traffic accidents in Norway: a case–control study.	GJERDE, H.; NORMANN, P. T.; CHRISTOPHERSEN, A. S.; SAMUELSEN, S. O.; MORLAND, J. (2011)	193,993	Regressão logística
6	Severity prediction of traffic accident using an artificial neural network.	ALKHEDER, S.; TAAMNEH, M.; TAAMNEH, S. (2017)	133,306	Rede Neural Artificial
7	Extracting decision rules from police accident reports through decision trees.	DE OÑA, J.; LÓPEZ, G.; ABELLÁN, J. (2013)	127,993	Árvores de decisão
8	The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach.	PAKGOHAR, A.; TABRIZI, R. S.; KHALILI, M.; & ESMAEILI, A. (2011)	123	Classificação e Árvores de Decisão
9	Black spots identification through a Bayesian Networks quantification of accident risk index.	GREGORIADES, A.; MOUSKOS, K. C. (2013)	119,089	Rede Bayesiana
10	Bayes classifiers for imbalanced traffic accidents datasets.	MUJALLI, R. O.; LÓPEZ, G.; GARACH, L. (2016)	116,993	Rede Bayesiana

Posição	Título	Autores	*Metodi In ordinatio*	Metodologia utilizada
11	Application of classification algorithms for analysis of road safety risk factor dependencies.	KWON, O. H; RHEE, W.; YOON, Y. (2015)	103,993	*Naive Bayes* e Árvore de decisão
12	A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks.	MUJALLI, R. O.; DE OÑA, J. (2011)	88,487	Rede Bayesiana
13	Machine learning methods to analyze injury severity of drivers from different age and gender groups.	MAFI, S.; ABDELRAZIG, Y.; DOCZY, R. (2018)	80,019	Aprendizagem de máquina - Máquinas de vetor suporte
14	Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study.	LEE, J.; YOON, T.; KWON, S.; LEE, J. (2020)	77,838	Árvore de decisão e rede neural.
15	Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates.	TAAMNEH, M.; ALKHEDER, S.; TAAMNEH, S. (2017)	77,003	Técnica de classificação: Árvore de Decisão, Regras de Indução, Naive Bayes e rede neural Perceptron Multicamadas (MLP)
16	Economic development, demographic characteristics, road network and traffic accidents in Zhongshan, China: gradient boosting decision tree model.	WU, W.; JIANG, S.; LIU, R.; JIN, W.; MA, C. (2020)	72,277	Árvore de decisão - *Gradient boosting*
17	Alternative method of highway traffic safety analysis for developing countries using Delphi technique and Bayesian network.	MBAKWE, A. C.; SAKA, A. A.; CHOI, K.; LEE, Y. J. (2016)	71,993	Técnica Delphi e Rede Bayesiana
18	Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree.	ROVŠEK, V.; BATISTA, M.; BOGUNOVIĆ, B. (2017)	69,469	Classification and Regression Tree (CART)
19	Data mining on road safety: factor assessment on vehicle accidents using classification models.	CASTRO, Y.; KIM, Y. J. (2016)	66,055	Rede Bayesiana, Árvores de Decisão e Redes Neurais.

Posição	Título	Autores	*Metodi In ordinatio*	Metodologia utilizada
20	Decrease of morbidity in road traffic accidents in a high income country–an analysis of 24,405 accidents in a 21 year period.	ERNSTBERGER, A.; JOERIS, A., DAIGL, M.; KISS, M.; ANGERPOINTNER, K.; NERLICH, M.; SCHMUCKER, U. (2015)	63,586	Regressão multivariada
21	An alternative method for traffic accident severity prediction: using deep forests algorithm.	GAN, J.; LI, L.; ZHANG, D.; YI, Z.; XIANG, Q. (2020)	61,249	Árvore de decisão
22	Data mining and complex network algorithms for traffic accident analysis.	LIN, L.; WANG, Q.; SADEK, A. W. (2014)	56,56	Algoritmo de detecção de comunidade e regras de associação
23	Spatial modelling of accidents risk caused by driver drowsiness with data mining algorithms.	FARHANGI, F., SADEGHI-NIARAKI, A., NAHVI, A., & RAZAVI-TERMEH, S. V. (2022)	55,889	Algoritmos de Árvore de Decisão, Floresta Aleatória e Máquina de vetores de suporte
24	Risk analysis of traffic accidents' severities: An application of three data mining models.	ALKHEDER, S.; ALRUKAIBI, F.; AIASH, A.	55,468	Árvore de decisão, Redes Bayesianas e Máquinas de Vetores de Suporte
25	On scene injury severity prediction (OSISP) algorithm for car occupants.	BUENDIA, R.; CANDEFJORD, S.; FAGERLIND, H.; BÁLINT, A.; SJOQVIST, B. A.	52,993	Regressão logística
26	Cause Analysis of Traffic Accidents on Urban Roads Based on an Improved Association Rule Mining Algorithm.	CAI, Q. (2020)	51,367	Regras de Associação - Algoritmo *Apriori*
27	Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents.	NAFIE ALI, F. M.; MOHAMED HAMED, A. A. (2018)	49	Algoritmo *Apriori e Clustering*

Posição	Título	Autores	*Metodi In ordinatio*	Metodologia utilizada
28	Analyzing factors that influence expressway traffic crashes based on association rules: using the Shaoyang–Xinhuang section of the Shanghai–Kunming expressway as an example.	CHEN, L., HUANG, S., YANG, C., & CHEN, Q. (2020)	48,774	Regras de Associação
29	Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil.	DA CRUZ FIGUEIRA, A. C.; PITOMBO, C. S.; LAROCCA, A. P. C. (2017)	48	Árvores de decisão
30	Apriori-based algorithm for Dubai road accident analysis.	JOHN, M.; SHAIBA, H. (2019)	46	Algoritmo *Apriori*
31	Relationship between using cell phone and the risk of accident with motor vehicles: An analytical cross-sectional study.	KOGANI, M.; ALMASI, S. A.; ANSARI-MOGADDAM, A.; DALVAND, S., OKATI-ALIABAD, H.; TABATABAEE, S. M.; ALMASI, S. Z. (2020)	46	Regressão logística
32	Road Traffic Accidents Injury Data Analytics.	NOUR, M. K., NASEER, A., ALKAZEMI, B., & JAMIL, M. A. (2020)	41	Árvores de decisão, Máquinas de Vetores de Suporte, Redes Neurais e Extreme Gradient Boosting (XGBoost)
33	The Effect of Drivers' Demographic Characteristics on Road Accidents in Different Seasons Using Data Mining.	SHOKOHYAR, S.; TAATI, E.; ZOLFAGHARI, S. (2017)	40	Algoritmos Kohonen, K-Means e Two-step
34	Augmenting classifiers performance through clustering: A comparative study on road accident data.	KUMAR, S.; TIWARI, P.; DENIS, K. V. (2018)	40	Máquinas de Vetores de Suporte, Floresta Aleatória, Naive Bayes, Técnicas de agrupamento de classes latentes (LCC) e agrupamento BIRCH.

Posição	Título	Autores	*Metodi In ordinatio*	Metodologia utilizada
35	Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France.	EL MAZOURI, F. Z.; ABOUNAIMA, M. C.; ZENKOUAR, K. (2019)	38	Regras de Associação. Algoritmo *Apriori*
36	Age-related risk factors with nonfatal traffic accidents in urban areas in Maringá, Paraná, Brazil.	DE MELO W. A.; ALARCÃO, A. C. J.; DE OLIVEIRA, A. P. R.; PELLOSO, S. M.; CARVALHO, M. D. D. B. (2017)	36,491	Regressão logística
37	A novel framework to use association rule mining for classification of traffic accident severity.	GUPTA, M.; SOLANKI, V. K.; SINGH, V. K. (2017)	35	Regras de Associação
38	Comparison of clustering techniques for traffic accident detection.	DOĞRU, N.E.J.D.E.T.; SUBAŞI, A. (2015)	35	Algoritmos: *Density-Based Spatial Clustering Of Applications With Noise (DBSCAN), Expectation Maximization (EM), Hierarchical Clustering (HC) e K-means,*
39	Using cluster analysis methods for multivariate mapping of traffic accidents.	SELVI, H. Z.; CAGLAR, B. (2018)	34	*K-means, K-medoids e Agglomerative and Divisive Hierarchical Clustering (AGNES)*
40	Using decision trees to extract decision rules from police reports on road accidents.	GRISELDA, L.; DE OÑA, J.; JOAQUÍN, A. (2012)	33	Árvores de decisão
41	A traffic accident morphology diagnostic model based on a rough set decision tree.	TAO, G.; SONG, H.; LIU, J.; ZOU, J.; CHEN, Y. (2016)	31,845	Árvore de decisão
42	Unrecorded accidents detection on highways based on temporal data mining.	AN, S., ZHANG, T., ZHANG, X., & WANG, J. (2014)	27,305	Mineração de dados em séries temporais
43	On-scene injury severity prediction (OSISP) algorithm for truck occupants.	CANDEFJORD, S.; BUENDIA, R.; FAGERLIND, H.; BÁLINT, A.; WEGE, C.; SJOQVIST, B. A. (2015)	23,491	Regressão logística

Posição	Título	Autores	*Metodi In ordinatio*	Metodologia utilizada
44	Assessment of traffic safety among young people aged 15 to 24 in Slovenia.	ZANNE, M.; GROZNIK, A.; TWRDY, E. (2013)	8,001	Modelagem de Dependência e *Clustering*
45	Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods.	YANG, Y., HE, K., WANG, Y. P., YUAN, Z. Z., YIN, Y. H., & GUO, M. Z. (2022)	7,98	Rede bayesiana
46	Variance-based global sensitivity analysis for rear-end crash investigation using deep learning.	MOUSSA, G. S., OWAIS, M., & DABBOUR, E. (2022)	7,51	Rede neural
47	A registry-based investigation of road traffic fatality risk factors using police data: A case study of Hyderabad, India.	KORAMATI, S., MAJUMDAR, B. B., PANI, A., & SAHU, P. K. (2022)	6,78	Regras de Associação
48	Exploring injury severity of vulnerable road user involved crashes across seasons: A hybrid method integrating random parameter logit model and Bayesian network.	SUN, Z., XING, Y., WANG, J., GU, X., LU, H., & CHEN, Y. (2022)	5,99	Modelo logit e rede bayesiana

APÊNDICE 4 - ESTUDOS DE SINISTROS DE TRÂNSITO ENVOLVENDO CAMINHÕES

Autores	Fonte de Dados Utilizada	Metodologia	Variáveis/atributos utilizados	Principais Conclusões
HÄKKÄNEN, H.; SUMMALA, H. (2001)	Dados obtidos de relatórios de casos de todos os sinistros fatais em que motoristas de caminhão reboque estiveram envolvidos durante o período de 1991–1997 na Finlândia.	Comparação dos dados coletados e questionário realizado com 251 caminhoneiros	Idade do motorista de caminhão; Tempo de experiência do motorista de caminhão; Número de sinistros em que se envolveu nos últimos 5 anos; Horário do sinistro; Tempo de descanso antes do sinistro; Ocorrência de doença crônica.	A pesquisa apontou que motoristas mais jovens e dirigir durante a noite foram preditores significativos de serem os principais responsáveis por sinistros. Além disso, a probabilidade de ocorrência do sinistro aumentava em mais de três vezes se o motorista tivesse uma doença crônica. A direção prolongada antes do sinistro, histórico de sinistros ou infrações de trânsito não tiveram efeito significativo nos sinistros.
FORKENBROC, D. J.; HANLEY, P. F. (2003)	Caminhões envolvidos em sinistros fatais. Dados do Instituto de Pesquisa em Transporte da Universidade de Michigan (UMTRI) considerando os anos de 1995-1998.	Análise de classificação múltipla e detector automático de interação	Número de reboques; Fase do dia; Número de veículos envolvidos; Limite de velocidade da via; Condição pista; Sentido da via; Condição do ambiente (urbano ou rural); Condição da superfície da estrada.	O estudo mostrou que os caminhões com vários reboques têm maior probabilidade de se envolver em colisões fatais nas seguintes condições: escuridão; neve, lama ou gelo na superfície da estrada; envolvimento de três ou mais veículos e em estradas com limites de velocidade entre 65 a 75 mph.
TSAI, M. C.; SU, C. C. (2004)	Relatórios de sinistros policiais em Taiwan, ocorridos entre 1994 e 1998 envolvendo caminhões.	Modelo Linear Generalizado	Idade dos motoristas; Tipo de veículos envolvidos; Tipos de estradas.	Os resultados empíricos indicam que as taxas de sinistros com veículos de carga em Taiwan foram altas nos cenários envolvendo caminhões e veículos e os sinistros mais graves ocorreram envolvendo motoristas mais idosos.

Autores	Fonte de Dados Utilizada	Metodologia	Variáveis/atributos utilizados	Principais Conclusões
KHORASHADI, A.; NIEMEIER, D.; SHANKAR, V.; MANNERING, F. (2005)	Sinistros envolvendo caminhões que ocorreram no período de 1997 e 2000. Dados do Sistema de Vigilância e Análise de Sinistros de Trânsito, do Departamento de Transporte da Califórnia (Caltrans) e do Sistema Integrado de Registros de Trânsito, mantido pela Patrulha Rodoviária da Califórnia.	Modelo de Regressão Logística *Logit* Multinomial	Número de veículos envolvidos; Características do caminhoneiro; Tipo de veículo envolvido; Tipo do sinistro; Características do motorista do carro envolvido; Características do passageiro do carro; Condição climática; Fase do dia; Número de pistas; Uso do solo: urbano ou rural; Ano de fabricação do caminhão; Idade do motorista; Estado físico dos envolvidos; Traçado da via.	A pesquisa mostrou que sinistros em ambientes rurais envolvendo combinações de caminhão trator e reboque, a probabilidade de ferimentos grave/fatais aos motoristas aumentou cerca de 26% em relação aos sinistros envolvendo caminhões de única unidade. Em áreas urbanas, esta mesma probabilidade aumentou quase 700%. Em sinistros em que o uso de álcool ou drogas foi identificado como a causa principal do sinistro, a probabilidade de lesão grave/fatal aumentou cerca de 250% nas áreas rurais e quase 800% nas áreas urbanas.
BJÖRNSTIG, U.; BJÖRNSTIG, J.; ERIKSSON, A. (2008)	Sinistros rodoviários no período de 1995 e 2004. Dados da Polícia Oficial da Suécia.	Análise estatística Qui-quadrado utilizando SPSS 13.0	Tipo do sinistro; Tipo do veículo envolvido; Tipo de pista; Mês que ocorreu o sinistro; Idade dos envolvidos; Sexo dos envolvidos; Posição dos passageiros; Condição climática; Dias da semana.	O estudo verificou que colisões com veículos pesados geralmente ocorriam durante o dia, em dias úteis, no inverno e em estradas com duas pistas a uma velocidade de aproximadamente 70 e 90 km/h.
ZHU, X.; SRINIVASAN , S. (2011)	Colisões envolvendo caminhões que ocorreram entre abril de 2001 e dezembro de 2003 nos EUA. Dados da Federal Motor Carrier Safety Administration Departamento de Transporte dos EUA.	Modelo *probit* ordenado	Características do caminhoneiro; Tipo de veículo envolvido; Características do motorista do veículo envolvido; Características do passageiro do veículo envolvido; Condição climática; Condição pista; Tipo de pista.	O estudo verificou que o uso de drogas ilícitas e desatenção do motorista de caminhão aumentam a gravidade da lesão, enquanto o uso de airbags e cintos de segurança trazem benefícios para a segurança dos motoristas e passageiros.

Autores	Fonte de Dados Utilizada	Metodologia	Variáveis/atributos utilizados	Principais Conclusões
LEMP, J. D.; KOCKELMAN, K. M.; UNNIKRISHNAN, A. (2011).	Colisões envolvendo caminhões grandes que ocorreram entre abril de 2001 e dezembro de 2003 nos EUA. Dados da Federal Motor Carrier Safety Administration Departamento de Transporte dos EUA e National Highway Traffic Safety Administration.	Modelo *probit* ordenado padrão e heterocedásticos	Tipo de veículos envolvidos; Condição climática; Condição da pista; Indicador de velocidade dos veículos envolvidos; Tipo de pista; Características do motorista de caminhão; Características do veículo e do motorista envolvido; Características do caminhão; Traçado da via; Fase do dia; Número de pessoas envolvidas; Uso do solo: Urbano e Rural.	Verificou-se na pesquisa que a probabilidade de fatalidades e lesões graves aumenta com o número de reboques, mas diminui considerando o comprimento do caminhão e a classificação de peso bruto do veículo.
CHANG, L. Y.; CHIEN, J. T. (2013)	Dados de sinistros de caminhão em rodovias nacionais no período entre 2005 e 2006. Dados do Banco de Dados Nacional de Sinistros de Trânsito de Taiwan.	CART - Árvore de classificação e regressão Modelo não paramétrico	Sexo dos envolvidos; Tipo de veículo envolvido; Sistema de contenção; Condição de sobriedade; Tipo de acidente; Circunstâncias do sinistro; Tipo de pista; Traçado da via; Hora do sinistro; Condição da superfície da estrada; Condição climática; Fase do dia; Dia da Semana; Número de veículos envolvidos; Localização do sinistro.	Os resultados do estudo mostraram que beber e dirigir, o não uso do cinto de segurança, tipo de veículo, tipo de colisão, circunstância, número de veículos envolvidos e local do sinistro foram os principais determinantes dos resultados de gravidade das lesões envolvendo sinistros de caminhão.
ISLAM, M.; HERNANDEZ, S. (2013)	Sinistros envolvendo caminhões no período de 2005-2008. Dados do Sistema Nacional de Amostragem Automotiva dos EUA.	Modelo *probit* ordenado	Características humanas; Características dos veículos; Tipo de sinistro; Condição climática; Traçado da via; Fase do dia; Condição da pista.	Segundo o estudo, o nível de gravidade da lesão foi considerado altamente influenciado por uma série de interações complexas relacionadas a fatores humanos, veiculares e ambientais da estrada.

Autores	Fonte de Dados Utilizada	Metodologia	Variáveis/atributos utilizados	Principais Conclusões
ISLAM, S.; JONES, S. L.; DYE, D. (2014)	Sinistros que ocorreram no período de 2010-2012. Dados do Centro de Segurança Pública Avançada da Universidade do Alabama.	Modelo *logit* misto	Idade dos motoristas envolvidos; Ano do caminhão; Hora do sinistro; Condição da superfície da pista; Peso do caminhão; Características do caminhão; Características do sinistro; Uso do solo: Urbano ou Rural; Tipo do sinistro; Traçado da via; Tipo de pista; Condição climática; Fase do dia; Velocidade dos veículos envolvidos; Sentido da via.	A pesquisa mostrou que as influências de uma variedade de variáveis na gravidade das lesões foram diferentes, resultando em sinistros urbanos e rurais envolvendo caminhões grandes e um único veículo ou com vários veículos.
PAHUKULA, J.; HERNANDEZ, S.; UNNIKRISHNAN, A. (2015)	Sinistros envolvendo caminhões no Texas no período de 2006-2010 fornecidos pelo Sistema de Informação de Registro de Sinistro do Texas.	Modelo de Regressão *Logit* com Fatores Aleatórios	Fase do dia, tipo do acidente, idade do condutor, sexo do condutor, mês, objeto atingido, condição da superfície da estrada, condições meteorológicas, manobra do veículo antes do acidente, alinhamento da estrada, uso de cinto de segurança, classificação do sinistro, porcentagem de caminhões na rodovia.	De acordo com o estudo, os diferentes períodos de tempo têm, de fato, diferentes fatores que contribuem para a gravidade de cada lesão, destacando ainda mais a importância de examinar os sinistros com base na fase do dia. Fluxo de tráfego, condições de luz, condições da superfície da estrada, época do ano e porcentagem de caminhões na estrada foram consideradas como as principais variáveis influenciadoras.

Autores	Fonte de Dados Utilizada	Metodologia	Variáveis/atributos utilizados	Principais Conclusões
ZHENG, Z.; LU, P.; LANTZ, B. (2018)	Sinistros envolvendo caminhões na Dakota do Norte e Colorado no período de 2010-2016 fornecidos pela Federal Motor Carrier Safety Administration Departamento de Transporte dos EUA.	Algoritmo de mineração de dados - *Gradient boosting*	Características da empresa transportadora (número total de caminhões, valor da inspeção, data de registro e localização); Características da colisão (dia da semana, hora do dia e número de lesões); Características do ambiente (tipo de estrada, condição de iluminação, condição da superfície da estrada e condições climáticas); Características dos motoristas envolvidos (idade, classe da carteira de motorista); Características do caminhão (tipo de carga, configuração, tipos de carroceria e peso bruto do veículo).	Segundo o estudo, vários fatores, como atributos da empresa de transporte rodoviário (por exemplo, tamanho da empresa), valores de inspeção de segurança, status de comércio da empresa de transporte (por exemplo, interestadual ou intraestadual), hora do dia, idade do motorista, primeiros eventos prejudiciais e condição de registro estão significativamente associados com a gravidade de lesões por sinistro.