UNIVERSIDADE FEDERAL DE MINAS GERAIS

CURSO DE MESTRADO EM GEOTECNIA E TRANSPORTES

Ramon Batista de Araújo

REGRAS DE ASSOCIAÇÃO ENTRE AS CARACTERÍSTICAS DOS VEÍCULOS E OS ACIDENTES DE TRÂNSITO EM RODOVIAS FEDERAIS BRASILEIRAS ATRAVÉS DE APRENDIZADO DE MÁQUINA

Belo Horizonte

2022

Ramon Batista de Araújo

REGRAS DE ASSOCIAÇÃO ENTRE AS CARACTERÍSTICAS DOS VEÍCULOS E OS ACIDENTES DE TRÂNSITO EM RODOVIAS FEDERAIS BRASILEIRAS ATRAVÉS DE APRENDIZADO DE MÁQUINA

Versão final

Dissertação apresentada ao Curso de Mestrado em Geotecnia e Transportes da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Geotecnia e Transportes.

Área de concentração: Transportes

Orientador: Marcelo Franco Porto

Belo Horizonte
2022

Araújo, Ramon Batista de.
A663r Regras de associação entre as características dos veículos e os acidentes de trânsito em rodovias federais brasileiras através de aprendizado de máquina [recurso eletrônico]/Ramon Batista de Araújo. – 2022.
1 recurso online (115 f. : il., color.) : pdf.

Orientador: Marcelo Franco Porto.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Apêndice: f. 102-115.
Bibliografia: f. 93-101.
Exigências do sistema: Adobe Acrobat Reader.

1. Transportes - Teses. 2. Aprendizado de computador – Teses. 3. Acidentes de trânsito – Teses. 4. Algoritmos – Teses.
I. Porto, Marcelo Franco. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 656(043)

Ficha catalográfica elaborada pela Bibliotecária Letícia Alves Vieira - CRB-6/2337
Biblioteca Prof. Mário Werneck - Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
CURSO DE MESTRADO EM GEOTECNIA E TRANSPORTES

FOLHA DE APROVAÇÃO

Regras de associação entre as características dos veículos e os acidentes de trânsito em rodovias federais brasileiras por meio de aprendizado de máquina.

RAMON BATISTA DE ARAÚJO

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GEOTECNIA E TRANSPORTES, como requisito para obtenção do grau de Mestre em GEOTECNIA E TRANSPORTES, área de concentração TRANSPORTES, constituída pelos seguintes professores:

Prof. Marcelo Franco Porto - Orientador (UFMG)
Prof.ª. Renata Maria Abrantes Baracho Porto (UFMG)
Prof. Ricardo Poley Martins Ferreira (UFMG)
Prof. Sandro Laudares (PUC Minas)

Belo Horizonte, 31 de março de 2022.

Documento assinado eletronicamente por **Marcelo Franco Porto**, **Professor do Magistério Superior**, em 31/03/2022, às 11:26, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Sandro Laudares**, **Usuário Externo**, em 31/03/2022, às 11:27, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ricardo Poley Martins Ferreira**, **Professor do Magistério Superior**, em 31/03/2022, às 11:28, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Renata Maria Abrantes Baracho Porto**, **Professora do Magistério Superior**, em 01/04/2022, às 08:14, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1349320** e o código CRC **00B764B4**.

Referência: Processo nº 23072.216019/2022-41 SEI nº 1349320

AGRADECIMENTOS

Aos professores e membros do corpo docente da Universidade Federal de Minas Gerais que contribuíram durante essa jornada para minha formação. Em especial, ao amigo e professor Marcelo Franco Porto, orientador deste trabalho, por aceitar auxiliar com toda sua experiência para a conclusão deste projeto. Assim como a professora Leise Kelli de Oliveira, que teve um papel essencial no ensino de metodologia científica e análise de dados em transportes. Agradeço também aos professores da pós-graduação em ciência dos dados da Pontifícia Universidade Católica de Minas Gerais que participaram na minha formação como cientista de dados, onde aprendi a aplicar as principais técnicas de aprendizado de máquina utilizadas neste estudo. Um agradecimento também aos professores da banca avaliadora que disponibilizaram seu tempo para colaborarem com este trabalho.

Um obrigado, aos demais professores, à minha família, aos colegas e amigos que de forma direta ou indireta tornaram a realização deste trabalho possível. Por fim, agradeço a Deus à oportunidade de estar concluindo mais uma etapa na vida.

“Aprenda com o ontem. Viva o hoje. Tenha esperança para o amanhã.”

(Albert Einstein)

RESUMO

Acidentes de trânsito são considerados graves e sérios problemas de saúde pública, os quais acarretam em uma série de mortos e feridos, representando não apenas números, mas vidas humanas perdidas. Em vista disso, o impacto social, somado aos custos com expressivo número de mortos e feridos, evidencia a necessidade de uma análise mais profunda das causas de acidentes e, por este motivo, esta pesquisa teve como objetivo central identificar regras de associação entre as causas de acidentes e as características dos veículos, das estradas, dos usuários e do meio ambiente em rodovias federais brasileiras, comparando as técnicas de aprendizado de máquina *Apriori, Eclat, FP-Growth* e *FP-Max* no tratamento dos dados. Para atingir tal objetivo, a metodologia desta pesquisa empregou o uso de tabulação de dados de variáveis categóricas, utilizando-se de um método misto para coleta e transformação dos dados e análise dos resultados, por meio de um procedimento dentro de um contexto real e local em um estudo de caso. Como resultado, foi possível realizar a comparação entre os algoritmos e verificar que os algoritmos *Apriori, FP-Growth* e *Eclat* apresentam o mesmo desempenho, com índices de suporte e quantidade de características similares, onde, quanto maior a quantidade de características, menor o índice de suporte. O *FP-Max* propõe uma maior métrica de suporte para maior quantidade de características e apresentou desfecho contrário, proporcionando um resultado mais preciso. O *FP-Max* e o *Eclat* não apresentaram índices de *lift* e confiança para o banco de dados analisado. Em vista destes fatores, é possível afirmar que a colaboração de um método capaz de entender as causas dos acidentes pode auxiliar políticas públicas, de modo a ser uma boa contribuição social e científica, visto que esta pesquisa tem potencial promissor para ser utilizada como base de diversos estudos e pela Polícia Rodoviária Federal, bem como engenheiros de segurança, poder público e pelo setor privado como concessionárias de rodovias e desenvolvedores de aplicativos *mobile*.

Palavras-Chaves: Acidentes. Regras de associação. Aprendizado de máquina. Algoritmos.

ABSTRACT

Traffic accidents are pressing public health problems, which lead to a series of deaths and injuries, representing not only numbers, but lost human lives. In view of this, the social impact added to the costs with a significant number of deaths and injuries highlight the need for a deeper analysis of the causes of accident. For this reason, this research aimed at identifying association rules between the causes of accidents and the characteristics of vehicles, roads, users, and the environment on Brazilian federal highways, comparing the *Apriori, Eclat, FP-Growth,* and *FP-Max* machine learning techniques in data processing. To achieve this objective, the methodology of this research basically applied the use of categorical variables data tabulation, using a mixed method for data collection and transformation and analysis of results, through a procedure within a real and local context in a case study. In this way, through the analysis of the results, it was possible to compare the algorithms and, thus, verify that the *Apriori, FP-Growth,* and *Eclat* algorithms perform equally, with similar support and number of characteristics indexes, where the higher the number of characteristics, the lower the support index. On the other hand, the *FP-Max,* which proposes a greater support metric for a higher number of characteristics, achieved the opposite outcome, consequently providing a more accurate result. However, *FP-Max,* as well as *Eclat,* did not present *lift* and confidence indexes for the analyzed database. Therefore, taking these factors into consideration, it is possible to affirm that the collaboration of a method capable of understanding the causes of accidents can help public policies and be a strong social and scientific contribution. This research has a promising potential to be used as a basis for several studies, by the Federal Highway Police itself, safety engineers, public authorities, and also by the private sector such as highway concessionaires and *mobile* application developers.

Keywords: Accidents. Association rules. Machine Learning. Algorithms.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

ANOVA – Análise multivariada da variância

CGSIE – Coordenação Geral de Sistemas, Informações e Estatísticas

CRISP-D – *Cross-Industry Standard Process for Data Mining*

DENATRAN – Departamento Nacional de Trânsito

FENABRAVE – Federação Nacional da Distribuição de Veículos Automotores

IBGE – Instituto Brasileiro de Geografia e Estatística

MANOVA – Análise multivariada da variância

MINFRA – Ministério da Infraestrutura do Brasil

ONSV - Observatório Nacional de Segurança Viária

PRF – Polícia Rodoviária Federal

RENAVAM – Registro Nacional de Veículos Automotores

SNTT – Secretaria Nacional de Transportes Terrestres

PRF – Polícia Rodoviária Federal

ANOVA – Análise de variância

MANOVA – Análise multivariada da variância

PIB – Produto Interno Bruto

SUMÁRIO

1 INTRODUÇÃO

Os acidentes de trânsito são considerados uma questão de saúde pública, visto que consistem em uma das principais causas de mortes no mundo. De acordo com dados da Polícia Rodoviária Federal (PRF), ocorreram cerca de 96 mil acidentes nas rodovias federais brasileiras em 2016. Nestes, aproximadamente, 87 mil pessoas ficaram feridas e 6.398 mortes foram registradas, isto é, cerca de 18 mortes por dia (BRASIL, 2018). Tais acidentes acarretam custos para a economia brasileira, os quais somam até 12,3 bilhões de reais, além de custos com despesas hospitalares, atendimentos médicos, tratamentos e perda de produção, estimados em 7,9 bilhões de reais (IPEA, 2015).

O impacto social, somado aos custos com expressivo número de mortos e feridos, evidencia a necessidade de uma análise mais profunda das causas de acidentes. Então, a análise de dados se torna relevante para identificar fatores e ajudar a reduzir taxas de acidentes (KUMAR *et al.*, 2017). Neste contexto, sabe-se que um dos grandes desafios da era da informação é transformar dados e informações em conhecimento e, por isso, tem havido um importante progresso na mineração de dados e aprendizado de máquina nos últimos anos.

Conforme afirma Cunto (2008), o conhecimento extraído a partir de diferentes modelos de análise estatística dos dados de acidentes auxilia engenheiros de segurança a tomarem decisões. Atnafu & Kaur (2017) afirmam que é importante realizar uma análise cautelosa dos dados de acidentes para identificar a natureza do mesmo, tornando-se possível mitigar a dificuldade de análise em grandes quantidades, principalmente quando se tem dados de diferentes fontes e formatos e, então, contribuir para uma análise minuciosa e mais precisa. Neste sentido, técnicas como *machine learning* podem ser utilizadas para encontrar padrões ocultos e criar regras de associação entre os atributos de banco de dados de acidentes de trânsito, por exemplo, como usado por pesquisadores em diversos estudos ao redor do mundo.

Em vista disso, as principais questões desta pesquisa consistem em: (i) Existem regras de associação entre as causas dos acidentes e as características dos veículos, das estradas, dos usuários e do meio ambiente em dados de acidentes das rodovias federais brasileiras? (ii) Qual o algoritmo que melhor identifica essas associações? Essas questões, as quais concebem os objetivos deste estudo, faz com que esta pesquisa se baseie em três conceitos essenciais, determinados como pilares para determinação dos resultados, sendo eles os acidentes de trânsito e seus fatores, a técnica de *machine learning* e os algoritmos de regras de associação.

1.1 Objetivos do trabalho

Esta pesquisa busca comparar algoritmos de aprendizado de máquina para identificação das regras de associação entre as causas de acidentes, bem como as características dos veículos, das estradas, dos usuários e do meio ambiente em rodovias federais brasileiras.

Para tanto, definiram-se os objetivos específicos:

a) Criar um relatório com representação gráfica dos dados de acidentes em rodovias federais brasileiras no período de janeiro de 2017 a fevereiro 2020;
b) Analisar a independência das características dos acidentes *versus* as causas destes;
c) Obter regras de associação representadas pela tabela *IF-THEN* com índices de suporte para os algoritmos *Apriori, Eclat, FP-Growth* e *FP_Max;*
d) Realizar uma análise estatística descritiva dos resultados dos algoritmos de regras de associação;
e) Analisar a igualdade entre as médias dos índices de suportes dos diferentes algoritmos;
f) Relacionar as regras de associação pertinentes para tomadas de decisões e políticas públicas.

1.2 Estrutura da Dissertação

Esta pesquisa apresenta estrutura para apresentação dos resultados conforme o que segue.

O problema proposto no estudo é apresentado na introdução, bem como estudos relacionados, justificativa, contribuições científicas, tecnológicas e sociais, perguntas de pesquisa e objetivos, este denominado Capítulo 1.

Em sequência, o Capítulo 2 dispõe da revisão da literatura, apresentando os pilares da pesquisa, conceitos e definições necessárias para entender a metodologia e resultados deste trabalho. Além disso, ainda é apresentado um resumo dos trabalhos relacionados e das contribuições desta pesquisa para a literatura técnica.

O Capítulo 3 descreve a metodologia aplicada na pesquisa, com a qual é possível explicar como serão tratados os dados, bem como a análise exploratória e aplicação dos algoritmos de regras de associação e a obtenção dos resultados, os quais serão conforme os objetivos. Este é o capítulo em que se apresenta a fundamentação metodológica da pesquisa.

Após aplicação da metodologia, no Capítulo 4 é apresentado o desenvolvimento desta, utilizando-se dos dados de acidentes das rodovias federais brasileiras e das características dos veículos. Em seguida descreve-se a obtenção e discussão dos resultados.

Para finalizar a pesquisa, no Capítulo 5 tem-se as considerações finais e as sugestões de trabalhos futuros.

2 REVISÃO DA LITERATURA

Este capítulo consiste em uma revisão bibliográfica do tema, de modo a compreender a importância de assuntos como acidentes de trânsito no Brasil e no mundo, bem como fatores relacionados aos acidentes, tais como estradas, usuários, meio ambiente e veículos e, ainda, revisar sobre as características dos veículos brasileiros, aprendizado de máquina e os algoritmos de regras de associação os quais serão abordados na metodologia.

2.1 Acidentes de trânsito e seus fatores

Segundo Mohan *et al.* (2006), um acidente de trânsito é resultado de uma combinação de fatores relacionados às estradas, aos usuários, ao meio ambiente e a veículos. Almeida *et al.* (2013) relata que os fatores que contribuem com a causa dos acidentes de trânsito, em maior ou menor grau, são o homem, o veículo, a via e o meio ambiente, bem como o dever de cumprimento da legislação existente. Então, a combinação desses fatores pode aumentar a probabilidade de acidentes, de forma diferenciada, em determinados locais. Sendo assim, o estudo de suas associações se torna necessário para compreender e evitar acidentes.

Sabe-se que acidentes de trânsito são considerados um problema sério de saúde pública no planeta (MÁSILKOVÁ, 2017). Segundo dados do último relatório de década da Organização Mundial de Saúde, são perdidas cerca de 1,35 milhões de vidas, a cada ano, por acidentes de trânsito, sendo considerada a oitava principal causa de morte no mundo, além de ser a principal causa de morte de pessoas com idade entre 5 e 29 anos. Outro ponto importante é que, na maioria dos países, o custo de acidentes chega ser de até 3% do seu PIB (Produto Interno Bruto), segundo dados da *World Health Organization* (2018).

O enorme volume de acidentes é considerado um desafio de todos os órgãos de segurança, os quais têm por objetivo evitar que estes ocorram e, por isso, a relevância de se estudar o tema por uma abordagem multidisciplinar (GOPALAKRISHNAN, 2012). Profissionais e pesquisadores da área de transportes buscam assegurar um bom desempenho na segurança, o qual pode ser obtido por meio de alguns recursos disponíveis e dos vários componentes e instalações de transportes (CUNTO, 2008).

Para Barroso Junior *et al.* (2019), acidentes em rodovias federais brasileiras tendem a ser mais letais para indivíduos do sexo masculino, pedestres, com ocorrências na região Nordeste, aos domingos, durante a madrugada, nas curvas, nas áreas rurais e para vítimas com idades mais

elevadas. Neste sentido, Lima *et al.* (2008) afirmam que acidentes do tipo colisão frontal tendem a ser mais frequentes em pistas simples e, além destes, outro fator interessante a ser levado em consideração para as regras de associação são as condições meteorológicas.

As condições meteorológicas se tornaram um dos principais fatores que influenciam a frequência e gravidade de colisões de veículos motorizados, segundo Al-Harbi *et al.* (2012). Chung *et al*. (2005) realizaram testes estatísticos os quais mostram que a frequência média de acidentes, durante períodos de chuva, é significativamente diferente da média de frequência em outros períodos. O estudo dos autores demonstrou que a análise dos registros de acidentes revelou que as taxas de acidentes aos finais de semana são mais elevadas do que nos dias de semana. Já o estudo de Lankarani *et al*. (2014) revelou que a proporção de acidentes durante o pôr do sol foram superiores aos ocorridos ao nascer do sol, resultando em uma análise estatística de dados que demonstrou que a proporção de ferimentos e mortes foi significativamente maior ao nascer e pôr do sol do que aqueles que ocorreram durante o dia.

Além disso, segundo Karacasu & Er (2011), o fator humano também representa uma das principais causas dos acidentes de trânsito, onde a falha humana resulta de diversos fatores como educação, idade, gênero e outros. Conforme Jiménez-Mejías *et al*. (2014), embora a mortalidade por lesões relacionadas a acidentes de trânsito seja conhecida por ser maior em homens, especialmente entre motoristas jovens, a influência do gênero em cada elo da cadeia causal que leva a este resultado não é bem compreendida. Ainda, os autores puderam concluir que motoristas do sexo masculino utilizam dispositivos de segurança com menos frequência que as mulheres e, ainda, estes se envolvem em comportamentos de direção de risco com maior frequência. Apesar disto, a análise bruta não mostrou diferenças entre os sexos nos acidentes relatados, onde a análise ajustada mostrou uma tendência de que os homens tivessem relatado acidentes com menor frequência do que as mulheres, devendo ser melhor investigado, no entanto, o fator idade.

Referente aos condutores idosos, de acordo com Scialfa (1991), em comparação com os motoristas mais jovens, estes raramente se envolvem em acidentes ou violações atribuíveis à velocidade excessiva ou, ainda, em violações graves como dirigir imprudentemente ou embriagado. No entanto, os idosos são mais propensos a se envolver em acidentes de trânsito os quais englobam falha em obedecer a sinais, ceder direito de passagem ou virar com segurança. De acordo com *United States General Accounting Office* (1994), a relação entre idades dos condutores não é a mesma para todos os tipos de acidentes. Os motoristas com idade

inferior a 25 anos estão submetidos a um maior risco de acidente, seguidos por motoristas com idade acima de 65 anos.

Em função dos dados mencionados, a idade do condutor é considerada um importante fator a ser estudado nesta pesquisa e, então, essas informações serão levadas em consideração na criação de regras de associação do presente estudo. Além disso, entender as causas dos acidentes para auxiliar em políticas públicas será uma importante contribuição social desta pesquisa, visto que compreender esses conceitos é fundamental para entender e analisar o banco de dados de acidentes. Neste estudo, estão sendo analisadas as relações entre as causas dos acidentes e as características dos veículos no cenário brasileiro.

2.1.1 Os veículos e os acidentes

De acordo com Santos & Pinhão (1999), a produção de veículos no Brasil iniciou-se ao final dos anos 1950. Segundo dados do IBGE (Instituto Brasileiro de Geografia e Estatística), em 2018 a frota brasileira chegou a 100.746.553 veículos, sendo 54.175.488 automóveis em circulação (BRASIL, 2020). Além disso, segundo a Federação Nacional da Distribuição de Veículos Automotores (Fenabrave, 2020), empresa que tem como associadas as montadoras General Motors, Volkswagen, Hyundai, Fiat Chrysler, Renault, Ford, Toyota, Honda, Jeep e Nissan, somente no ano de 2019 foram vendidos, aproximadamente, 2,7 milhões de automóveis e comerciais leves. Neste sentido, vale ressaltar que esta pesquisa considera as características dos veículos constantes no Registro Nacional de Veículos Automotores (Renavam), tais como marca, idade do veículo e potência do motor.

Sabe-se que é o próprio condutor quem define a velocidade na qual o veículo se movimenta. No entanto, de acordo com Luchezi (2010), o automóvel se tornou um ícone da sociedade o qual está além do consumo por motivo de locomoção, este também passou a ser um instrumento de poder do indivíduo. Tal sentimento de poder, denominado de sensação de status por Maoski (2014), está relacionado até ao barulho do motor, da mesma forma que o comportamento imprudente dos motoristas, além do excesso de velocidade, está associado ao modo em que os fabricantes demonstram seus veículos quanto à potência do motor e ao desempenho do mesmo (PORTER, 2011). Segundo a *World Health Organization* (2018), um aumento na velocidade média está diretamente relacionado à probabilidade de ocorrência de um acidente e à gravidade das consequências deste acidente.

Outra questão importante diz respeito à idade do veículo, pois, segundo Blows *et al.* (2003), à medida em que veículos mais antigos circulam nas estradas, o risco de acidentes tem um aumento gradativo, isto é, veículos mais antigos apresentam maior risco de se envolverem em acidentes de trânsito, fato este que caracteriza um desafio para segurança de trânsito devido ao aumento da idade dos veículos que circulam nas vias. Da mesma forma, no estudo *do United States General Accounting Office* (1994), foi possível verificar que os carros mais novos apresentavam um risco ligeiramente menor de envolvimento em colisões.

Em vista disto, nota-se que um enorme volume de dados é gerado constantemente pela grande quantidade de acidentes diversos que ocorrem todos os dias, podendo estes ser de natureza heterogênea e com diferentes atributos, quantitativos e qualitativos, os quais dificultam as análises por métodos estatísticos (KUMAR & TOSHNIWAL, 2015). Essa imensa quantidade de dados possibilita o uso de novas tecnologias, as quais visam contribuir com pesquisadores e gestores a extrair conhecimento, auxiliando-os em tomadas de decisões.

Os dados apresentados conduzem ao questionamento da existência de padrões entre as causas de acidentes, as características das estradas, do condutor, do meio ambiente e das características dos veículos como idade do veículo, bom como da marca e da potência do motor. Para identificar esses padrões, utilizou-se das técnicas de *machine learning*, as quais necessitam, em um primeiro momento, de melhor compreensão de dados e, ainda, da verificação de independência entre as características e as causas dos acidentes, para tanto, empregando-se a análise de correspondência multivariada.

2.2 Análise de correspondência multivariada

Conforme determinado nos objetivos, o presente estudo pretende analisar as regras de associação entre os fatores que envolvem um acidente, tornando-se válido averiguar a dependência das variáveis para obter-se uma análise relevante dos dados antes da criação das regras de associação por aprendizado de máquina.

A análise de correspondência consiste em um procedimento para representar associações em uma tabela de frequências ou contagens. Basicamente, a metodologia emprega uma tabela de contingência, que consiste em uma tabulação de variáveis categóricas. Então, a análise transforma os dados não-métricos (qualitativos) em um nível métrico (quantitativo) e, ainda, faz redução dimensional mapeando o perceptual. Além disso, a análise de correspondência fornece uma representação multivariada de interdependência para dados qualitativos e, no caso

da análise de correspondência multivariada, envolve-se três ou mais variáveis categóricas relacionadas em um espaço comum (HAIR *et al*., 2009).

Conforme Mingoti (2007), a associação entre variáveis categóricas gera tabelas de contingência que, na maioria das vezes, são analisados através do teste qui-quadrado. Hair *et al.* (2009) afirmam que a análise de correspondência usa o qui-quadrado para padronizar os valores de frequência da tabela de contingência e formar a base para associação. Segundo Washington *et. al* (2003), o teste qui-quadrado é amplamente utilizado em análises de transporte devido à sua versatilidade e capacidade de ajustar um enorme volume de questões. De modo geral, para encontrar o resultado da análise multivariada de correspondência por meio do teste do qui-quadrado de independência das variáveis em uma tabela de contingência, adota-se as seguintes hipóteses: Hipótese nula (H_0) = Independência das variáveis e Hipótese alternativa (H_1) = Dependência das variáveis.

De acordo com Greenacre (1993), em função dos dados serem compostos por variáveis categóricas, dados dessa natureza podem ser analisados através de análise de correspondência. A mesma metodologia de análise foi utilizada por Das *et al*. (2018), os quais obtiveram resultados satisfatórios utilizando-se da análise de correspondência múltipla para determinar as principais associações entre os fatores de contribuição, em acidentes entre os anos de 2010 e 2014, em Louisiana nos Estados Unidos. Outro estudo que utilizou análise de correspondência multivariada para compreender os fatores envolvidos e seu impacto, foi realizado por Tyagi *et al.* (2018), analisando acidentes entre os anos de 2005 e 2015 no Reino Unido, os quais envolveram fatores como tempo, condições meteorológicas, idade do motorista e outros, corroborando com a análise do presente estudo. Além destes, Jalayer *et al*. (2017) também verificaram os fatores associados à ocorrência de acidentes por direção na contramão, onde foram estudados por meio de análise de correspondência múltipla no estado de Illinois nos Estados Unidos da América. Logo, nota-se a relevância científica do procedimento de análise de correspondência para verificação de fatores associados a acidentes de trânsito.

A análise de correspondência múltipla para verificar estatisticamente a independência das variáveis representa uma importante etapa deste estudo, visto que investiga se há associação entre as variáveis, deste modo, assegurando a existência de possíveis regras de associação significativas nos dados, conforme proposto nos objetivos específicos deste estudo. Além disso, ressalta-se que o objetivo geral não é apresentar quais são os fatores principais das causas de acidentes utilizando análise de correspondência múltipla e, sim, a comparação e a verificação da existência de regras de associação utilizando algoritmos de aprendizado de máquina.

2.3 Aprendizado de Máquina

Segundo Baştanlar & Ozuysal (2014), acredita-se que existe um processo que explica os dados observados, mesmo não conhecendo os detalhes de um processo como, por exemplo, o comportamento de um consumidor, o qual não é completamente aleatório. Então, técnicas como *machine learning* são ferramentas utilizadas para analisar tais dados.

Antes de aplicar técnicas de aprendizado de máquina, é preciso entender o significado desta metodologia, como funciona e porquê e onde são utilizadas. Conforme Alpaydm (2004), *machine learning* é uma programação de computadores que utiliza dados de exemplos ou experiências anteriores para resolver um determinado problema. Da mesma forma, Mohri *et al.* (2012) afirmam que aprendizado de máquina consiste em métodos computacionais que se utilizam da experiência para melhorar o desempenho ou fazer previsões precisas. Ainda segundo os autores, esses métodos combinam conceitos fundamentais da ciência da computação, matemática e estatística. Muller & Guido (2017) complementam que o aprendizado de máquina é um campo de pesquisa da estatística e ciência da computação, a qual utiliza-se de inteligência artificial, sendo conhecido aprendizado estatístico, onde a aplicação de métodos de aprendizado de máquina tornou-se muito utilizado para a resolução de problemas utilizando dados.

Para Zhang (2020), o aprendizado de máquina é um subcampo da inteligência artificial o qual constrói um modelo matemático com base em dados de amostra, a fim de fazer previsões ou decisões, sem ser explicitamente programado para realizar a tarefa e, por isso, são amplamente utilizados em diversas áreas de aplicação e, usualmente, na área de transportes, sendo esse conceito utilizado para análise de acidentes de trânsito por autores como Ali & Hamed (2018); Atnafu & Kaur (2017); Chong et al. (2005); Costa et al. (2014); Daher et al. (2016); Deekshitha et al. (2019); Figueira et al. (2017); Amorim (2019); Kumar & Toshniwal (2015); Kumar et al. (2017); Li et al. (2017); Martín et al. (2014); Meng et al. (2019); Nandurge & Dharwadkar (2017); Ozbayoglu et al. (2016); Reis et al. (2015); Shanti et al. (2011); Silva et al. (2019); Soares et al. (2018); Tayeb et al. (2015); Xi et al. (2016).

Existem diversos algoritmos e tipos de aprendizagem de máquina, os quais são divididos pela literatura em aprendizagem supervisionada, semi-supervisionada ou aprendizagem por reforço e aprendizagem não supervisionada.

- Aprendizagem Supervisionada:

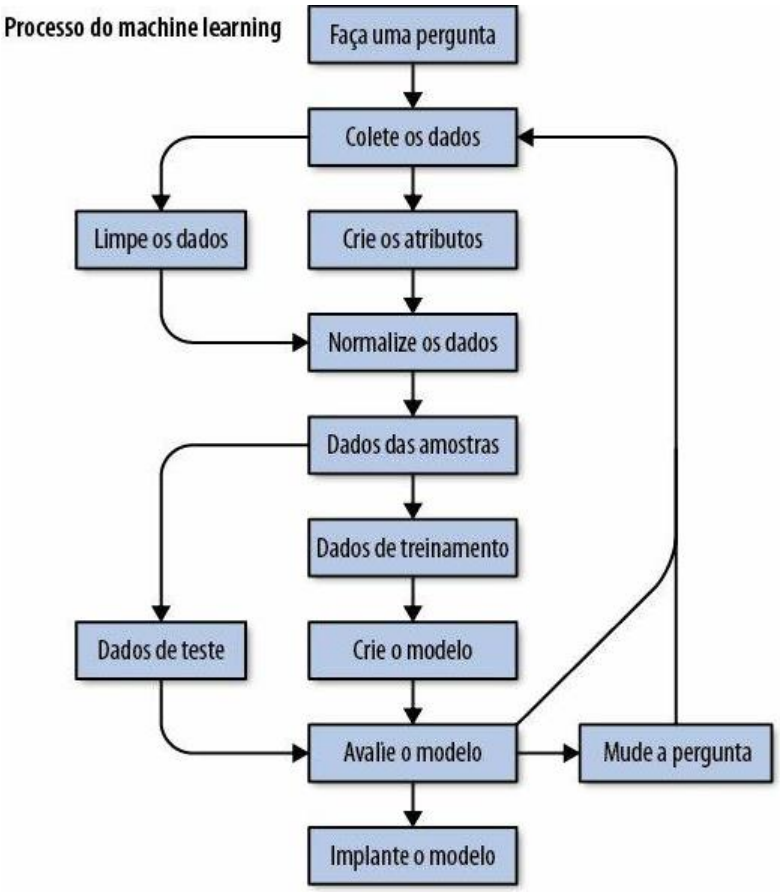
Na enciclopédia de aprendizagem de máquina, Sammut & Webb (2011) definem aprendizagem supervisionada como sendo qualquer processo de aprendizagem de máquina com a qual se aprende uma função compreendendo exemplos que têm valores de entrada e saída. De modo mais simples, Mohri *et al.* (2012) afirmam que, na aprendizagem supervisionada, o algoritmo recebe um conjunto de exemplos rotulados, chamados de dados de treinamento, aprende como resultaram os rótulos e, utilizando os dados de treino, realiza previsões nos dados de teste para todos os pontos não vistos.

Zhang (2020) exemplifica aprendizagem supervisionada como sendo um conjunto de dados rotulados, ou seja, dados com entradas e saídas conhecidas. Nesta metodologia, separa-se os conjuntos em dados de teste e dados de treino. A aprendizagem funciona como se um professor estivesse disponibilizando a um aluno um problema, isto é, o professor apresenta os dados de treino com suas soluções resolvidas, ou seja, dados de entrada e saída (rótulos), e explica ao aluno que ele deve descobrir como resolver outros problemas semelhantes, utilizando os métodos dos problemas disponibilizados. De modo geral, a Figura 2.2 exemplifica um processo de aprendizado de máquina supervisionado pela metodologia CRISP-D (*Cross-Industry Standard Process for Data Mining*) (Harrison, 2020).

De acordo com Sammut & Webb (2011), dois exemplos típicos de aprendizagem supervisionada são a aprendizagem por classificação e a regressão. Kantardzic (2011) explica a aprendizagem por classificação e regressão como sendo tarefas comuns suportadas por este tipo de aprendizagem indutiva. Para facilitar o entendimento, o autor descreve a aprendizagem supervisionada como sendo a existência de uma função de aptidão de um professor, ou algum outro método externo de estimativa do modelo proposto. O termo supervisionado denota que os valores de saída, para amostras de treinamento, são conhecidos, ou seja, fornecidos por um professor. Da mesma forma, Bishop (2006) também exemplifica os dois tipos de aprendizagem. Segundo o autor, dentro da aprendizagem supervisionada, os dados de treinamento compreendem exemplos dos vetores de entrada juntamente com seus vetores de saída correspondentes, onde têm-se dois tipos de problemas, os problemas de classificação, em que o objetivo é atribuir cada vetor de entrada a outro com um número finito de categorias discretas. Uma vez que se deseja consistir em uma ou mais variáveis contínuas, a tarefa é chamada de regressão, tais como os problemas de Regressão Linear. Como exemplos de aprendizagem supervisionada por classificação pode-se citar os algoritmos de Redes Neurais, Regressão

Logística e *Naive Bayes* e SVM. Os autores que utilizam desse método em análise de dados de acidentes são relacionados na Tabela 2.2 desse capítulo.

Figura 2.1 – Fluxo de trabalho comum no *machine learning*



Fonte: Harrison, 2020

Além deste, outro tipo de aprendizagem é a aprendizagem semi-supervisionada, também chamada de aprendizagem por reforço.

- Aprendizagem Semi-supervisionada:

Na aprendizagem semi-superivionada, conforme Monhi *et al.* (2012), o algoritmo recebe uma amostra de treinamento com dados rotulados e não rotulados e faz uma previsão para os dados não rotulados. Para Sammut & Webb (2011), o objetivo da aprendizagem semi-supervisionada é criar um modelo, através de dados rotulados e não rotulados, o qual preveja dados de teste futuros melhor do que o modelo treinado apenas com dados rotulados.

Aprendizagem semi-supervisionada tem sua importância devido à dificuldade de coletar dados rotulados, sendo necessário, então, o uso de dados não rotulados. Da mesma forma, na maioria das tarefas de reconhecimento de padrões, os seres humanos têm pequeno número de exemplos

rotulados em sua aprendizagem, o que atraiu muita atenção na comunidade de aprendizado de máquina (ZHANG, 2020). Em síntese, o aprendizado semi-supervisionado consiste em uma combinação entre a aprendizagem de máquina supervisionada e não supervisionada. A aprendizagem não supervisionada, a qual não possui rótulos de saída, será o tipo de aprendizagem abordada na metodologia deste trabalho, corroborando com as metodologias utilizadas por outros autores.

- Aprendizagem Não-Supervisionada:

Segundo Müller & Guido (2017), a aprendizagem de máquina não supervisionada inclui todos os tipos de aprendizado onde não há saída conhecida, isto é, o algoritmo extrai conhecimento desses dados apenas com os dados de entrada. Zhang (2020) relata que a principal tarefa do aprendizado de máquina não supervisionado consiste em o aluno encontrar as soluções por conta própria, ou seja, encontrar padrões, estruturas ou conhecimento em dados sem rótulos de saída. Nesta metodologia ocorre o mesmo que disponibilizar ao aluno um conjunto de padrões e requisitar que este descubra os motivos que os geraram. Conforme Morhi *et al.* (2012), na aprendizagem não supervisionada, nenhum exemplo rotulado está disponível nesse ambiente, sendo assim, pode ser difícil avaliar quantitativamente o desempenho de um aluno.

Um grande desafio no aprendizado não supervisionado consiste em avaliar se o algoritmo aprendeu algo útil, isto ocorre devido ao fato de não se identificar qual deve ser a saída correta. É muito difícil afirmar se um modelo se saiu bem frequentemente, e a única maneira de avaliar o resultado de um algoritmo não supervisionado é inspecioná-lo manualmente. Para Müller & Guido (2017), algoritmos de aprendizagem não supervisionada são usados em ambientes quando deseja entender melhor os dados.

Por conseguinte, o aprendizado de máquina utiliza dados de exemplos ou experiências anteriores para resolver um problema através de uma sequência de instruções que transformam as entradas em saídas, estas conhecidas como algoritmos. Em síntese, o aprendizado supervisionado utiliza-se de uma resposta desejada para o padrão de entrada e o aprendizado por reforço, isto quando o fator externo avalia a resposta fornecida pela rede.

Além disso, no aprendizado não supervisionado, o processo busca aprender através da estrutura dos dados, sem uma saída identificada, isto é, sem um rótulo. Alguns exemplos típicos de aprendizagem não supervisionada são clusters, mapas de auto-organização e regras de associação (SAMMUT & WEBB, 2011).

É importante ressaltar que, no presente estudo, pretende-se comparar o desempenho de algoritmos de regras de associação de aprendizagem de máquina não supervisionada na

intenção de descobrir padrões nas características dos dados de acidentes. Utiliza-se, então, algoritmos frequentemente aplicados por pesquisadores em todo mundo e outros usados em diversas áreas de estudo, conforme apresentado na introdução deste trabalho.

2.3.1 Regras de Associação

Popularizada pelo artigo de Agrawal *et al.* (1993), tendo sua origem na análise baseada em cem cestas de mercado, as regras de associação consistem em uma das ferramentas mais populares em mineração de dados (HEGLAND, 2007). Então, regras de associação representam uma das principais técnicas de mineração de dados e, ocasionalmente, esta seja a forma mais comum de descoberta de padrões por aprendizagem de máquina não supervisionados. Esta metodologia é conhecida por ser capaz de expor algo sobre os dados de um usuário, os quais ele mesmo ainda não sabia e, provavelmente, não conseguiria encontrar sem auxílio de aprendizagem de máquina (KANTARDZIC, 2011). Para Zhang (2019), o objetivo das regras de associação consiste em descobrir padrões nos dados anteriormente desconhecidos.

De acordo com Borgelt & Kruse (2002), regras de associação buscam encontrar conjuntos de atributos que são, frequentemente, ocorridos juntos, de modo que, a partir da presença de determinados atributos em um ocorrido, é possível inferir, com alta probabilidade, que certos outros atributos também estarão presentes. Conforme Hegland (2007), regras de associação são regras IF-THEN ($X \Rightarrow Y$), para um determinado conjunto de dados. Para Zhang (2019), uma regra de associação descreve um relacionamento $X \Rightarrow Y$ entre um conjunto de itens X e um único item Y. Segundo Sammut & Webb (2011), uma regra de associação tem a forma $X \Rightarrow Y$, onde X e Y são conjuntos de itens, e a interpretação significa que, se o conjunto X ocorrer, então, o conjunto Y provavelmente também ocorrerá. Facelli *et. al.* (2011) definem as regras IF como sendo os antecedentes, e a THEN as consequentes, nas quais antecedentes e consequentes são *itemsets*. Pode-se citar, como exemplo cotidiano, o caso de uma cesta de compras, onde, se em uma transação forem comprados os itens queijo, pão e manteiga, gerando suporte ao conjunto de itens formados por queijo e pão, ou seja, se o cliente comprasse apenas pão e queijo, então, provavelmente, também compraria manteiga.

Regras de associação são regras SE-ENTÃO (IF-THEN) com duas medidas que quantificam o suporte e a confiança da regra para um determinado conjunto de dados. O suporte, consiste, então, de acordo com Hegland (2007), em uma indicação da frequência com que o conjunto de

itens aparece no conjunto de dados, e a confiança consiste em uma indicação da frequência com que a regra foi considerada verdadeira.

É importante ressaltar que a importância de uma regra é, geralmente, medida pelo seu suporte, o qual consiste na porcentagem de transações as quais a regra pode ser aplicada (ou, alternativamente, à porcentagem de transações nas quais ela está correta) e sua confiança, que representa o número de casos em que a regra está correta, em relação ao número de casos em que é aplicável. Para selecionar regras interessantes do conjunto de todas as regras possíveis, um suporte mínimo é fixado.

Segundo Facelli *et al.* (2011), o conjunto de transações que suporta o *itemset* é denominado por suporte do *itemset.* Esse índice de suporte pode ser relativo ou absoluto, onde o absoluto simboliza o número total de elementos do conjunto de transações, já o suporte relativo consiste na divisão do suporte absoluto pelo número de transações, isto é, a fração de transações. Normalmente, o suporte é usado para medir a significância de um conjunto de itens em um banco de dados e avaliar as regras de associação e definir limites de seleção. Por exemplo, dada uma regra $A \rightarrow C$, onde A significa antecedente e C significa consequente, da seguinte maneira:

$$support\ (A \rightarrow C) = support\ (A\ \cup C)$$

Onde o suporte antecedente é a proporção de transações que contém o antecedente A, e o suporte consequente é a proporção de transações para o conjunto de itens do consequente C. Então, tem-se a métrica de suporte do conjunto de itens combinado A ∪ C. Para encontrar as melhores regras, utiliza-se a métrica de suporte, isto é, a fração de transações satisfazendo as regras e combinação dos conjuntos de itens, esses valores variam entre 0 e 1, de acordo com Agrawal (1993).

Seguindo os preceitos da literatura relacionada nessa pesquisa, exibida na Tabela 2.2, os valores de suporte mínimo utilizado pelos referidos autores estão descritos na Tabela 2.1.

Tabela 2.1 – Valores de suporte mínimo utilizado pelos autores dos trabalhos relacionados

Referência	Mínimo valor de suporte utilizado
Ali & Hamed (2018)	0,4
Atnafu & Kaur (2017)	-
Costa *et al.* (2014)	0,1
Daher *et al.* (2016)	0,4
Deekshitha *et al.* (2019)	-
Kumar & Toshniwal (2015)	0,3
	Continua...

Kumar *et al.* (2017)	0,2
Li *et al.* (2017)	0,4
Meng *et al.* (2019)	0,1 / 0,05
Nandurge & Dharwadkar (2017)	0,3
Reis *et al.* (2015)	-
Soares *et al.* (2018)	0,1
Tayeb *et al.* (2015)	-
Xi *et al.* (2016)	0,4

Fonte: Elaborado pelo autor

É possível verificar que os autores que utilizaram regras de associação para buscar padrões desconhecidos nos dados de acidentes rodoviários utilizam valores mínimos de suporte entre 0,05 a 0,4. Entretanto, analisar uma grande quantidade de dados com um valor de suporte baixo pode dificultar o processamento de dados.

Facelli *et al*. (2011) afirmam que seu grau de interesse é representado pela confiança das regras, e, ainda, que consiste na probabilidade de ocorrer um conjunto de termos, dado que ocorreu um outro conjunto. Da mesma forma, Agrawal (1993) exemplifica a confiança de uma regra A → C, a qual consiste na probabilidade de ver o consequente em uma transação, dado que ela também contém o antecedente. Desta forma, a confiança é 1 (máxima) para uma regra A→ C, caso o consequente e o antecedente sempre ocorrerem juntos, esta é dada pela Equação 1.

$$confiança\ (A \rightarrow C) = \frac{P(A \cup C)}{P(A)} = \frac{suporte\ (A \cup C)}{suporte\ (A)} \quad (1)$$

Outra métrica abordada pela literatura é o coeficiente de interesse (*lift*), usado para avaliar os níveis de associação. Essa métrica é dada pela divisão entre a probabilidade conjunta de duas variáveis pela sua probabilidade única, pressupondo a hipótese de independência, ou seja, em resumo, o *lift* é dado pela divisão da confiança pelo suporte. Para Brin (1997), a métrica costuma ser usada para medir a independência estatística o antecedente e consequente, ou seja, a frequência em que o antecedente e o consequente ocorrem juntos de uma regra A → C, dada pela Equação 2.

$$lift\ (A \rightarrow C) = \frac{confiança\ (A \rightarrow C)}{suporte\ (C)} \quad (2)$$

De acordo com Hegland (2007), o tamanho dos conjuntos de itens representa um fator chave na determinação do desempenho dos algoritmos de regras de associação. Para escolher um algoritmo adequado, é importante verificar por meio de histograma para o comprimento dos itens. Segundo Borgelt & Kruse (2002), o principal problema da indução de regras de

associação é que existem infinitas regras possíveis para um determinado conjunto de dados. Um exemplo disto é que, para uma cadeia de produtos de um supermercado que tenha milhares de produtos diferentes, existem bilhões ou trilhões de regras de associação possíveis. São necessários algoritmos eficientes, os quais verifiquem as regras de associação distribuindo em subconjuntos, sem perder regras importantes.

Devido à disponibilidade de algoritmos eficientes, gera-se uma grande popularidade do uso de regras de associação em mineração de dados (HEGLAND, 2007). Os principais algoritmos de regras de associação utilizados na área de acidentes rodoviários estão listados na Tabela 2.2 e, conforme objetivo da pesquisa, serão comparados no presente estudo.

2.3.2 Algoritmos de Regras de Associação

Facelli *et al.* (2011) apresentam, como principais algoritmos de regras de associação, o *Apriori* e o *FP-Growth*. No entanto, pesquisadores em todo mundo utilizam-se de outros algoritmos de regras de associação, como, por exemplo, o *Eclat*, bem como uma nova versão do *FP-Growth*, estudada por Bouakkaz *et al*. (2012), o *FP-Max*. Desta forma, conforme objetivos da pesquisa, o presente estudo vista comparar os algoritmos *Apriori*, *FP-Growth*, *FP-Max* e *Eclat*.

2.3.2.1 Apriori

O *Apriori*, introduzido por Agrawal *et al*. (1993), foi o primeiro algoritmo de regras de associação utilizado, segundo Agrawal e Srikant (1994). De acordo com Borgelt & Kruse (2002), o *Apriori* atua em duas etapas, onde, na primeira, é possível determinar os conjuntos de itens frequentes, o qual têm-se, pelo menos, o suporte mínimo fornecido, ou seja, ocorrem em pelo menos uma determinada porcentagem de todas as transações. Em seguida, na segunda etapa, geram-se as regras de associação com base no conjunto de itens frequentes descobertos na primeira etapa.

Para Kantardzic (2011), esse conjunto de itens frequentes é calculado por meio de várias interações, estas baseadas no conhecimento sobre conjuntos de itens infrequentes obtidos de iterações anteriores, assim, o algoritmo *Apriori* reduz o conjunto de itens podando os conjuntos de itens candidatos que não podem ser frequentes. Essa poda é baseada na observação de que, se um conjunto de itens for frequente, todos os seus subconjuntos também poderão ser. Antes

de entrar na segunda etapa, o algoritmo descarta todos os conjuntos de itens candidatos que possuem um subconjunto infrequente.

Para exemplificar, Facelli *et al.* (2011) explicam que o algoritmo *Apriori* inicia com a criação do conjunto F_1 de *itemsets* de tamanho 1, de modo que cada item seja membro do conjunto de itens candidatos. Os *itemsets* de tamanho k + 1 são obtidos, então, a partir dos *itemsets* de tamanho k em duas etapas, sendo, na primeira etapa, a auto combinação do conjunto Fk, quando um conjunto de candidatos com K+1 *itemsets* é gerado pela combinação Fk com ele mesmo. A união A U C dos *itemsets* A, C e Fk é gerada se eles têm o mesmo k-1-préfixo. Em seguida, na segunda etapa, etapa da poda, A U C é inserida em Fk +1 somente se todos os seus subconjuntos-k ocorrem em Fk. Após isso, conta-se os suportes de todos k+1 *itemsets* candidatos varrendo todas as transações e os suportes de todos os *itemsets* candidatos. Então, todos os *itemsets* que se tornam frequentes são inseridos em Fk+1.

Ainda, é importante ressaltar que o algoritmo *Apriori* é o principal algoritmo de regras de associação, sendo que este já foi utilizado por diversos autores na análise de acidentes rodoviários em todo planeta, tais como Ali & Hamed (2018), Atnafu & Kaur (2017), Costa *et al.* (2014), Deekshitha *et al*. (2019), Kumar & Toshniwal (2015), Li *et al*. (2017), Meng *et al*. (2019), Nandurge & Dharwadkar (2017), Reis *et al*. (2015), Soares *et al*. (2018), Tayeb *et al*. (2015), Xi *et al*. (2016). Entretanto, o *Apriori* não é o único algoritmo de regras de associação, outros algoritmos surgiram como alternativa a este.

Então, a proposta de nível de confiança do algoritmo *Apriori* implica diversas varreduras sobre o banco de dados para calcular o suporte dos *itemsets* frequentes candidatos. Como alternativa, diversos algoritmos reduziram, significantemente, essas varreduras por meio da geração de coleções de *itemsets* candidatos em uma estratégia de busca em profundidade. Outros algoritmos propostos foram o *Eclat*, desenvolvido por Zaki (2000) e o algoritmo *FP-growth*, criado por Han *et al.* (2004).

2.3.2.2 FP-Growth

Conforme já mencionado, o algoritmo *FP-Growth* (*Frequen Pattern Growth*) foi criado por Han *et al.* (2004). O *FP-Growth* consiste em um algoritmo para extrair conjuntos de itens frequentes com aplicações em aprendizagem de regras de associação, o qual surgiu como uma alternativa ao algoritmo *Apriori*.

De acordo com Han *et al.,* 2001), o algoritmo *FP-Growth* foi utilizado como um bloco de construção em *itemsets* frequentes e mineração de sequências em fluxo contínuo de dados. A estratégia de busca por profundidade e árvores de sufixo utilizadas pelo algoritmo *FP-Growth* é empregada na maioria dos algoritmos de mineração de padrões frequentes aplicados a dados de fluxo contínuo. Em particular, e o que o torna diferente do algoritmo de mineração de padrão frequente *Apriori*, o *FP-Growth* é um algoritmo que não requer geração de candidato. Internamente, ele usa uma estrutura de dados chamada *FP-Tree*, isto é, utiliza uma árvore de padrão frequente, sem gerar os conjuntos candidatos explicitamente, o que o torna particularmente atraente para grandes conjuntos de dados (RASCHKA, 2018).

Como destaque, o *Fp-Growth* se mostrou eficaz para análise de acidentes, segundo Daher *et al.* (2016) e Kumar *et al*. (2017), porém, este ainda é pouco utilizado na área de transportes, quando comparado ao algoritmo *Apriori*. Adicionalmente, surgiu uma variante que obtém conjuntos de item máximos, a *FP-Max*, tornando-a interessante para ser estudada na área de análise de acidentes.

2.3.2.3 FP-Max

Sabe-se que o algoritmo *FP-Max* é uma variante do *FP-Growth*, o qual se concentra na obtenção de conjuntos de itens máximos. Em contraste com *Apriori*, o *FP-Max* é um algoritmo de geração de padrão frequente que insere itens em uma árvore de pesquisa de padrão, permitindo um aumento linear no tempo de execução em relação ao número de itens ou entradas exclusivas.

Um conjunto de itens X é considerado máximo se X for frequente e não houver super padrão frequente contendo X. Em outras palavras, um padrão frequente X não pode ser sub padrão de um padrão frequente maior para se qualificar para a definição de item máximo.

O *FP-Max* tende a ter menor número de regras, porém, estas regras, normalmente, são apresentadas com maior número de itens, os quais apresentam melhores índices de suporte. É importante ressaltar que o *FP-Max* ainda não foi utilizado por pesquisadores na área de acidentes de trânsito e, em vista disto, está representa uma importante contribuição científica deste estudo.

2.3.2.4 Eclat

O algoritmo *Eclat* (*Equivalence Class Transformation*) foi proposto por Zaki (2000), o qual também foi desenvolvido como alternativa ao *Apriori*. Ao contrário do método *Apriori*, o método *Eclat* não se baseia no cálculo de confiança, este se baseia no cálculo das conjunções de suporte das variáveis.

Segundo Ishita & Rathod (2016), o algoritmo *Eclat* é um dos mais conhecidos algoritmo para minerar conjuntos de itens frequentes em um conjunto de transações, tendo como vantagens sobre o *Apriori* a sua velocidade e as informações de contagem de suporte, as quais são obtidas do conjunto de itens anterior, não sendo necessário verificar cada conjunto de itens novamente. De acordo com os autores, o algoritmo não obtém total proveito da propriedade a priori para reduzir o número de itens explorados. Apesar da semelhança entre os dois tipos de algoritmos, o *Eclat* foi pouco aplicado por pesquisadores na área de transporte, tendo sido utilizado por Deekshitha *et al.* (2019) para identificar fatores que influenciam acidentes.

Heaton (2016) comparou os algoritmos *Apriori*, *Eclat* e *FP-Max* aplicando em banco de dados de cestas de compras, relatando que o *Apriori* é o algoritmo de fácil entendimento, fato que o tornou mais popular. Porém, o autor conclui que os algoritmos *Eclat* e *FP-Growth* lidam com aumentos na transação máxima, tamanho e densidade do conjunto de itens frequentes, consideravelmente melhor do que o algoritmo *Apriori*. Diferentemente de outros estudos, o autor utilizou a linguagem Python para aplicação dos algoritmos, distinta da tradicional ferramenta WEKA. Heaton (2016) ainda sugere, como trabalhos futuros, que seja realizada a aplicação e comparação dos algoritmos em um conjunto de dados mais críticos. Corroborando com isto, a linguagem Python também é utilizada no presente estudo.

2.4 Ferramenta Weka e Linguagem Python

A ferramenta WEKA é amplamente utilizada para aplicar algoritmos de regras de associação tais como *Apriori* e *FP-Growth*. Nafie Ali & Mohammed Hamed (2018), por meio desta ferramenta, compararam algoritmos de regas de associação com *Apriori* e algoritmos de *clustering* utilizando da ferramenta WEKA. Segundo os autores, a interface WEKA é uma ferramenta muito útil na mineração de dados, visto que permite ao usuário escolher vários algoritmos diferentes e, ainda, compará-los para chegar aos resultados necessários.

Vários autores valeram-se de tais ferramentas nas análises de suas pesquisas. Tayeb *et al.* (2015) utilizaram a ferramenta WEKA para aplicação de regras de associação empregando o *Apriori* em acidentes em Dubai. Li *et al*. (2017), utilizaram a ferramenta analítica de dados WEKA para realizar as análises de acidentes de trânsito nos Estados Unidos. Atnafu & Kaur (2017) aplicaram algoritmos por meio da ferramenta WEKA em um banco de dados de acidentes na Índia, visando identificar a influência de fatores em acidentes rodoviários. Nandurge & Dharwadkar (2017) determinaram os principais fatores associados em acidentes de trânsito. No Brasil, utilizado por Reis *et al.* (2015), o uso da descoberta de conhecimento em banco de dados nos acidentes da BR-381. Costa *et al.* (2014) empregaram algoritmos de aprendizado supervisionado, implementados na ferramenta WEKA, em dados de acidentes dos boletins de ocorrências de rodovias federais brasileiras gerados pela Polícia Rodoviária Federal em 2012.

Dentre as muitas linguagens existentes, a linguagem Python é amplamente utilizada no campo de ciência de dados e *machine learning*, graças ao advento de suas bibliotecas (HOMEM, 2020). Dentro da linguagem existem bibliotecas para aplicação dos algoritmos, visto que a biblioteca *MLxtend* é onde se encontram os algoritmos *Apriori*, *FP-Growth* e *FP-Max*. Segundo Raschka (2018), o *MLxtend* consiste em uma biblioteca que implementa uma variedade de algoritmos e utilitários básicos para máquinas aprendizagem e mineração de dados, fornecendo, ainda, uma grande variedade de utilitários diferentes, os quais se baseiam e estendem as capacidades do Python.

Em vista desta linguagem ser muito utilizada nos dias atuais e, ainda, visando a contribuição técnico-cientifica desta pesquisa, o presente estudo pretende utilizar-se desta para comparação dos algoritmos, diferentemente da tradicional ferramenta WEKA.

2.5 MANOVA

Para fins de comparação de resultados das regras de associação de diferentes algoritmos, esse estudo pretende utilizar o método de MANOVA (*Multivariate Analysis of Variance*), assim como Kaur (2015), o qual utilizou-se de técnicas como ANOVA e MANOVA para comparar os algoritmos *Apriori* e *FP-Growth* em regras de associação para detecções de doenças hepáticas.

Em síntese, a MANOVA verifica a semelhança entre grupos multivariados e explora as relações entre as variáveis independentes e dependentes (HAIR *et al*., 2005). De acordo com Hair (2009), as técnicas de extensões multivariadas são utilizadas para avaliar a significância

estatística de diferenças entre grupos, tais como procedimentos de inferência estatística. Em MANOVA, o pesquisador tem duas variáveis estatísticas, uma para as variáveis dependentes e outra para as independentes. Nas técnicas multivariadas, a hipótese nula testada é a igualdade de vetores de médias sobre múltiplas variáveis dependentes ao longo de grupos. Sobre as medidas estatísticas de avaliação das diferenças ao longo de dimensões das variáveis dependentes, os mais comuns, em MANOVA são o lamba de *Wilks*, critério de ilai e o traço de Hotelling (*Wilks, Pillai's, Hotelling-Lawley e Roy's Greatest*). Cada uma dessas considerações é controlável em variados graus, em um planejamento MANOVA, e, ainda, fornecem ao pesquisador diversas opções para gerenciar o poder, a fim de atingir o nível desejado de poder na faixa de 0,80 ou acima disso. Para o presente estudo, têm-se como hipótese nula (H0) todas as médias de população sendo iguais, enquanto a hipótese alternativa (H1) pelo menos uma é diferente.

2.6 Trabalhos Relacionados

Estudos utilizando técnicas de *machine learning* têm sido frequentemente realizados na área de transporte, principalmente, nos últimos anos, devido ao aumento do volume de dados, o qual tem apresentado crescimento constantemente em vista do aumento no índice de acidentes. Técnicas como *machine learning* são utilizadas por pesquisadores em todo mundo para prever, classificar e encontrar associações e padrões ocultos em dados de acidentes rodoviários no Brasil e no mundo.

Chong *et al.* (2005) realizaram um estudo objetivando detectar padrões em acidentes perigosos, no qual foram desenvolvidos modelos de previsões precisos, capazes de classificar automaticamente o tipo de gravidade da lesão de vários acidentes de trânsito dos Estados Unidos nos anos de 1995 a 2000. Nesse estudo, os autores utilizaram algoritmos de Rede Neural Artificial, *Support Vector Machine* e Árvore de Decisão, obtendo sucesso na análise com o uso de paradigmas de aprendizado de máquina de Rede Neural Artificial e Árvore de Decisão. Da mesma forma, Shanti *et al.* (2011) compararam os algoritmos C4.5, CRT, CS-MC4, *Decision List*, ID3, *Naive Bayes* e *Random Trees* de modo a obter regras de classificação com objetivo de prever o modo de colisão, utilizando dados de acidentes de 2007 nos Estados Unidos, onde o algoritmo *Random Trees* obteve a melhor precisão.

Tais algoritmos de aprendizado de máquina supervisionados, utilizados para classificação e previsão, foram utilizados por outros pesquisadores como Martín *et al.* (2014), os quais fizeram

uso de Rede Neural Artificial, Rede *Bayesiana*, Árvore de Decisão, *Support Vector Machine*, Regressão e *Clustering* para identificar informações sobre pontos perigosos na rede rodoviária espanhola. Os algoritmos de Rede Neural Artificial e Árvore de Regressão também foram usados por Ozbayoglu. *et al.* (2016) para detectar automaticamente acidentes em tempo real na Turquia. Outro estudo, realizado por Atnafu & Kaur (2017), aplica algoritmos de Árvore de Regressão, J48, *Naive Bayes* e um algoritmo de regras de associação *Apriori,* em um banco de dados de acidentes na Índia, visando identificar a influência de fatores rodoviários, humanos e ambientais em acidentes, bem como a probabilidade de o acidente ser grave.

Sabe-se que algoritmo *Apriori* é um algoritmo de aprendizado de máquina não supervisionado que encontra associações, principalmente, entre variáveis categóricas, o qual foi implementado para identificar padrões em acidentes de trânsito em diversos países. Nandurge & Dharwadkar (2017) determinaram os principais fatores associados em acidentes de trânsito através dos algoritmos *Apriori, Naive Bayes* e *K-Means* para associação, agrupamento e segmentação dos dados. Já, Xi *et al.* (2016) foram capazes de determinar o tipo e a gravidade dos acidentes causados por múltiplos fatores através do algoritmo *Apriori*, mostrando eficiência do algoritmo ao lidar com a grande amostra de dados chineses existentes. Ali & Hamed (2018) compararam o desempenho dos algoritmos *Apriori* e *Cluster* utilizando a ferramenta WEKA em um banco de dados de 946 observações e 8 atributos de acidentes da Arábia Saudita, descobrindo que o *Apriori* tem melhor desempenho que o *Cluster* para identificar os fatores que causam os acidentes. Com dados da Índia, Kumar & Toshniwal (2015) identificaram os principais fatores dentre 11.574 acidentes, no período de 2009 a 2014, por meio de regras de associação *Apriori* e *K-cluster*. Em Dubai, Tayeb *et al.* (2015) compararam os algoritmos *Apriori* e *Predict Apriori* para identificar a gravidade dos acidentes em dados dos Emirados Árabes no período de 2008 a 2010, onde o *Apriori* mostrou-se mais eficaz que o *Predict Apriori*. Li *et al.* (2017) descobriram variáveis intimamente relacionadas a acidentes fatais dos Estados Unidos através dos algoritmos *Apriori*, *Naive Bayes* e *K-Means*. Outro estudo utilizando *Apriori* foi realizado na China para investigar fatores de influência de acidentes em rodovias de baixa qualidade (MENG *et al.*, 2019). Em contrapartida, um algoritmo pouco usado, o *Eclat*, foi empregado por Deekshitha *et al.* (2019) para identificar fatores que influenciam acidentes. Daher *et al.* (2016) aplicaram outro algoritmo também pouco utilizado nesse ramo de pesquisa, o *FP-Growth*, com o objetivo de identificar as principais causas de acidentes de trânsito em Nova York, através de regras de associação.

No contexto brasileiro, Silva *et al.* (2019) usaram os algoritmos *Random Florest*, *Boosted Trees* e Rede Neural Artificial para investigar a influência da priorização de variáveis no ajuste de modelos de previsão de acidentes de resposta multivariada. Além deste estudo, Figueira *et al.* (2017) analisaram dados da BR-116, entre os anos de 2012 e 2014, com auxílio de algoritmos de Árvore de Decisão para detectar acidentes de trânsito com vítimas. Já Costa *et al.* (2014) utilizaram J48, PART e *Apriori* para identificar associação entre variáveis em acidentes de trânsito em todas as rodovias federais, com dados de 2012, obtendo um índice de confiança de 0,8 como resultado da aplicação do algoritmo. Reis *et al.* (2015) utilizaram o *Apriori* em acidentes no período de 2008 a 2012, na BR-381, para encontrar os principais fatores em pista simples e dupla, por meio da ferramenta WEKA. Ainda, Soares *et al.* (2018) utilizaram o algoritmo *Apriori* e a ferramenta WEKA para identificar os principais fatores e contribuintes dos acidentes na BR-101, usando dados de 2014 a 2016. Amorim (2019) fez uso dos algoritmos *Random Florest*, *Bernoulli NB*, Rede Neural, MLP, *Logistic Regression* e *Extra Trees Classifier* de *Machine Learning* para analisar o impacto de técnicas de aprendizado de máquina supervisionado na tarefa de predição do risco de acidentes graves ou não-graves em trechos de rodovias brasileiras, porém, o autor não utilizou algoritmos de regras de associação. Amorim (2019) afirma que o assunto é muito estudado nas demais partes do mundo e que poderia, ainda, ser melhor explorado no Brasil, sugerindo como trabalho futuro o uso e a comparação de outros algoritmos de aprendizado de máquina, utilizando outros atributos e observações mais recentes do banco de dados de acidentes da Polícia Rodoviária Federal.

Neste sentido, é possível verificar que grande parte dos pesquisadores utilizaram algoritmos de aprendizado de máquina supervisionado como Rede Neural Artificial, Árvore de Decisão, *Naive Bayes*, *Support Vector Machine*, *Random Tree* e técnicas de *cluster* com intuito de prever a gravidade dos acidentes, utilizando, em sua maioria, as variáveis quantitativas. Por outro lado, o uso do algoritmo de aprendizado não supervisionado, o *Apriori*, que geralmente é utilizado em variáveis categóricas, foi analisado em várias pesquisas e se mostrou eficaz, aplicado principalmente pela ferramenta WEKA e em diversos países com características distintas das brasileiras, como China, Índia, Arábia Saudita, Estados Unidos e Emirados Árabes. Quando aplicado no Brasil, os pesquisadores utilizaram atributos que envolvem mais a infraestrutura da via e as características do condutor, abrindo oportunidades para novos estudos, incluindo, ainda, novos atributos como, por exemplo, as características do veículo, a marca do automóvel, a idade deste e a potência do motor.

Então, percebe-se que o *Apriori* é o algoritmo mais utilizado para encontrar regras de associação em acidentes de trânsito, onde este mostrou-se eficaz na área, conforme estudos apresentados. Outros algoritmos como *FP-Growth* e *Eclat* foram utilizados em outras áreas de estudo, como é o caso da pesquisa de Tate & Bewoor (2017), que compararam os algoritmos *Apriori, Tree-projection, FP-Growth* e *Eclat*, demonstrando suas vantagens e desvantagens. Kaur (2015) identificou regras de associação para detectar doenças hepáticas, usando os algoritmos *Apriori* e *FP-Growth*, comparando-os através de Análise de Variância (ANOVA) e Análise Multivariada da Variância (MANOVA). Além destes, Hunyadi (2011), utilizando a ferramenta *Rapid Miner*, comparou o número de associações resultantes em cada um dos processos gerados, através do desempenho dos algoritmos *Apriori* e *FP-Growth*, em dados de uma loja e-commerce utilizando correlação ANOVA e regressão linear.

A variante do *FP-Growth*, *FP-Max*, também foi estudada em outras áreas, conforme estudo de Bouakkaz *et al.* (2012), cuja pesquisa comparou *FP-Max* e *Apriori* em um banco de dados de portos marítimos, resultando melhor desempenho do algoritmo *FP-Max*.

Algoritmos como *Eclat, FP-Growth* e *FP-Max* foram pouco utilizados na área de estudo de acidentes de trânsito e ainda não são aplicados no cenário brasileiro, justificando-se a importância do seu estudo para o desenvolvimento da discussão científica a respeito do tema proposto. Portando, de modo a contribuir cientificamente, este estudo pretende comparar os algoritmos de regras de associação *Apriori, Eclat, FP-Growth* e *FP-Max* na análise de acidentes, utilizando a linguagem Python com a biblioteca *Mlxtend (machine learning extensions*), diferentemente da tradicional ferramenta usada nos estudos corriqueiros, a WEKA. No contexto brasileiro, pretende-se, então, complementar os trabalhos já realizados, incluindo na análise atributos como a marca do veículo, a idade e a potência do motor, os quais ainda não foram utilizadas nos estudos descritos na literatura.

A Tabela 2.2 resume os principais estudos relacionados, incluindo o país estudado, o período do banco de dados analisado e o algoritmo utilizado, de modo a atingir o primeiro objetivo específico de relacionar estudos já realizados envolvendo aprendizado de máquina e acidentes de trânsito.

Tabela 2.2 – Principais estudos relacionados sobre *machine learning* e acidentes rodoviários

Referência	**País**	**Período**	**Algoritmos**										
			Árvore de Decisão	***Apriori***	**CART**	***Clustering***	***Eclat***	***FP-Growth***	***Naive Bayes***	**Redes Neurais**	**Regressão**	**SVM**	**FP-Max**
Ali & Hamed (2018)	Arábia Saudita	2011 a 2015		*		•							
Atnafu & Kaur (2017)	Índia	2014 a 2017	•	*					•				
Chong *et al.* (2005)	Estados Unidos	1995 a 2000	•							•		•	
Costa *et al.* (2014)	Brasil	2012	•	*									
Daher *et al.* (2016)	Estados Unidos	2009 a 2013						*					
Deekshitha *et al.* (2019)	-	2014 a 2016		*			*						
Figueira *et al.* (2017)	Brasil (BR-116)	2012 a 2014			•								
Amorim (2019)	Brasil	2007 a 2017	•							•	•		
Kumar & Toshniwal (2015)	Índia	2009 a 2014		*		•							
Kumar *et al.* (2017)	Índia	2009 a 2014						*					
Li *et al.* (2017)	Estados Unidos	2007		*									
Martín *et al.* (2014)	Espanha	2008 a 2010	•			•			•	•	•	•	
Meng *et al.* (2019)	China	2012 a 2014		*									
Nandurge & Dharwadkar (2017)	Índia	2015 a 2016		*		•			•				
Ozbayoglu *et al.* (2016)	Turquia	2015								•	•		
Reis *et al.* (2015)	Brasil (BR-381)	2008 a 2012		*									
Shanti *et al.* (2011)	Estados Unidos	2007	•						•				
Silva *et al.* (2019)	Brasil (RJ e SP)	2011 a 2017	•							•			
Soares *et al.* (2018)	Brasil (BR-101)	2014 a 2016		*									
Tayeb *et al.* (2015)	Dubai	2008 a 2010		*									
Xi *et al.* (2016)	China	-		*									

Fonte: Elaborado pelo autor

Por meio da análise apresentada, observa-se que o *Apriori*, de fato, é o algoritmo mais influente na área de análise de padrões em acidentes de trânsito. Existem outros algoritmos pouco aplicados na área, mas utilizados em outras áreas de estudo e, principalmente, no Brasil, os quais também devem ser estudados. O presente estudo pretende contribuir comparando o tradicional algoritmo *Apriori* com algoritmos utilizados em outras áreas de estudo, os quais são pouco aplicados na área de transporte, sendo o *Eclat*, *FP-Growth* e *FP-Max* utilizados para análise de acidentes, empregando a linguagem Python com a biblioteca *Mlxtend (machine learning extensions*), diferentemente da tradicional ferramenta usada, a WEKA. Ainda, o estudo pretende, além de utilizar uma nova base de dados de acidentes brasileiros, incluir as características dos veículos, pois esta ainda não foi estudada por técnicas de *machine learning*.

Este capítulo apresentou os principais conceitos desta pesquisa, desde as perguntas de pesquisas e objetivos, os métodos utilizados, a apresentação dos algoritmos a serem comparados como

contribuição científica e tecnológica, a linguagem e bibliotecas utilizadas como contribuição tecnológica proposta, assim como uma síntese da literatura envolvendo os principais trabalhos relacionados ao tema, bem como as justificativas das decisões do autor nas etapas da metodologia e desenvolvimento.

Por conseguinte, a metodologia desse trabalho fundamenta-se nessa revisão da literatura, com suas contribuições cientificas, tecnológicas e sociais, aplicando-se os algoritmos *Apriori, Eclat, FP-Growth* e *FP-Max*, conforme detalhado no capítulo que segue.

3 METODOLOGIA

Esse capítulo apresenta a metodologia utilizada para que seja possível atender aos objetivos desta pesquisa, onde o objetivo principal é identificar regras de associação entre as causas de acidentes e as características dos veículos, das estradas, dos usuários e do meio ambiente em rodovias federais brasileiras, comparando as técnicas de aprendizado de máquina *Apriori*, *Eclat*, *FP-Growth* e *FP-Max* no tratamento dos dados, o qual pode ser melhor entendido por meio da fundamentação metodológica desde estudo.

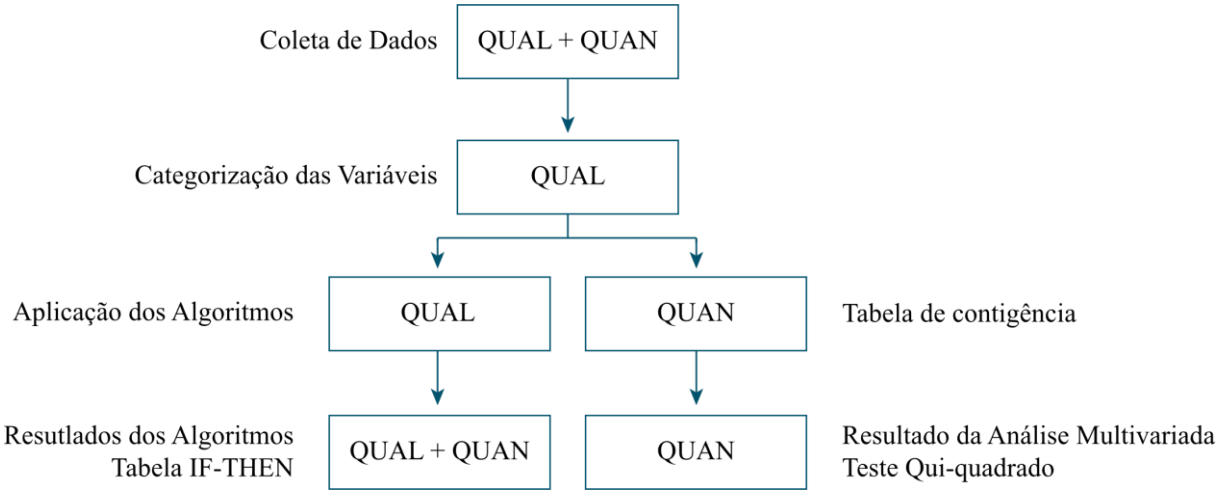
3.1 Fundamentação metodológica

O presente estudo utilizou-se de um método misto para coleta, transformação dos dados e análise dos resultados. Segundo Creswell *et. al* (2007), o método misto originou-se em 1959, onde técnicas como observações e entrevistas foram combinadas com estudos tradicionais, ou seja, combinou-se dados qualitativos com quantitativos, o que levou diversos pesquisadores de todo o mundo a desenvolver procedimentos de investigação de métodos mistos, os quais se aplicam três estratégias gerais como sequenciais, concomitantes e transformadores.

Além disso, utilizou-se de um procedimento dentro do contexto real e local, com limites dos fenômenos não definidos, isto é, um estudo de caso. Seguiu-se as etapas: definição das questões envolvidas na pesquisa; ambiente a ser estudado; procedimentos de obtenção dos dados; análise, interpretação dos resultados e descobertas, conforme Jung (2004).

Sabe-se que os procedimentos concomitantes apresentam uma análise abrangente do problema de pesquisa, uma vez que convergem dados qualitativos e quantitativos, onde dispõe de um procedimento de dados maior para analisar diferentes questões. Nessa estratégia, transformado concomitante, coleta-se dados qualitativos e quantitativos de forma simultânea e integra-se as informações na interpretação dos resultados (CRESWELL *et al.*, 2007). A Figura 3.1 apresenta a estratégia transformada concomitante de métodos mistos, adaptada para essa pesquisa.

Figura 3.1 – Método misto de análise



Fonte: Elaborado pelo autor

Os dados contendo os acidentes rodoviários e as características de veículos são dados qualitativos e quantitativos que foram coletados simultaneamente, os quais ocorreram na primeira fase do estudo. Em sequência, os dados quantitativos foram categorizados, transformando-se em qualitativos visando reduzir a amplitude dos dados quantitativos, proporcionando um melhor resultado do método de regras de associação, evitando *overfitting*, ou seja, o ajuste de um modelo muito complexo aos dados por terem naturezas específicas. É necessário categorizar os dados quantitativos para aplicação dos algoritmos de regras de associação como, por exemplo, a idade do condutor sendo estudada em faixas etárias, não em dados quantitativos contínuos.

Antes da aplicação dos algoritmos, visando compreender como é a relação entre as variáveis, cria-se uma tabela de contingência com as quantidades de observações das múltiplas variáveis categóricas, obtendo-se dados quantitativos, sendo possível efetuar uma análise estatística multivariada, atingindo o objetivo especifico (b) desse estudo.

Outro ponto importante, conforme descrito na revisão bibliográfica, é o fato de que, para os algoritmos de regras de associação, são necessários dados qualitativos para sua aplicação, gerando como resultados uma tabela qualitativa e quantitativa com dados qualitativos como *IF-THEN* e dados quantitativos como *support, confidence e lift.*, os quais foram analisados e comparados os resultados qualitativos e quantitativos de cada algoritmo, obtendo resultados qualitativos, tais como: quais são as regras mais pertinentes, isto é, aquelas com mais de um item e melhor *support*, respondendo à questão (i) desta pesquisa. Além disso, foram

comparados resultados quantitativos, como qual o algoritmo com maior número de regras *IF-THEN* e qual obteve melhores resultados de *support*, *confidence e lift*, respondendo à pergunta (ii) do estudo e atingindo o objetivo dessa pesquisa, como demonstrado no capítulo de desenvolvimento e resultados apresentado a seguir.

Por meio da revisão de literatura foi verificado que, em análise de acidentes, o algoritmo *Apriori* se mostrou eficaz em diversos países, incluindo o Brasil, conforme mostrado na Tabela 2.2 de trabalhos relacionados desse estudo. Porém, para este algoritmo, os autores pesquisaram associações em atributos que envolvem mais a infraestrutura da via e as características do condutor, ao contrário deste estudo, o qual pretende encontrar associações entre os acidentes e as características do veículo. Outros algoritmos como *Eclat, FP-Growth* e *FP-Max* ainda são pouco investigados na área de transporte e, por este motivo, serão comparados com o tradicional *Apriori*, em vista de terem se mostrado eficientes em pesquisas de outros campos de estudo. Para compreender a metodologia desse estudo, criou-se um *framework* com as etapas desenvolvidas.

3.2 *Framework* da pesquisa

Um *framework* é uma maneira de representar o conhecimento e informações úteis para a compreensão de um estudo (Minsky, 1974). Para as etapas do método deste estudo, estas estão apresentadas no *framework* da Figura 3.2.

Anterior a qualquer análise de dados, primeiramente é necessário compreender o problema de pesquisa. Neste contexto, procura-se responder às questões de pesquisa: (i) Existem regras de associação entre as causas dos acidentes e as características dos veículos, das estradas, dos usuários e do meio ambiente em dados de acidentes das rodovias federais brasileiras? (ii) Qual o algoritmo que melhor identifica essas associações? Por meio destes questionamentos, inicia-se a primeira etapa da metodologia, a obtenção dos dados.

Figura 3.2 – *Framework* da pesquisa

Fonte: Elaborado pelo autor

3.2.1 Bancos de Dados

A veracidade dos dados representa um dos pilares do *Big Data Analitycs*, o qual é muito importante para a confiabilidade dos resultados de *machine learning*. Sabe-se que uma fonte dos dados insegura pode introduzir incerteza e impactar a veracidade de um conjunto de dados. Logo, dados duvidosos e incorretos podem afetar o desempenho do aprendizado de máquina, provendo regras de associação equivocadas (L'HEUREUX *et al.*, 2017). Para essa metodologia, então, são necessários dados confiáveis de acidentes de trânsito e de características dos veículos.

Para essa metodologia, os dados são qualitativos e quantitativos, onde, no banco de dados de acidentes deve constar atributos que contenham: o momento que o acidente ocorreu, isto é, o dia, mês, ano e o horário; atributos das características das estradas, tais como o tipo e traçado da via; atributos de características do meio ambiente, tais como condições meteorológicas; atributos com características dos usuários, como idade e sexo do condutor. Ainda, referente ao banco de dados das características dos veículos, são necessários características como a marca, o ano de fabricação e potência do motor. E, por fim, são necessários para criação de regras de associação, dados com a causa do acidente.

Vale ressaltar que os dados utilizados nessa pesquisa foram obtidos de duas fontes distintas, os detalhes da obtenção dos dados estão descritos no item 4.1 do capítulo de desenvolvimento deste estudo. Com os dados obtidos, deve-se realizar uma análise precisa para a criação dos modelos de *machine learning.*

3.2.2 Análise Exploratória

Para compreender melhor os dados, realiza-se um pré-processamento e uma análise exploratória por meio da linguagem Python, como evidenciado no *framework* (Figura 2.4) deste estudo.

Dentre as muitas linguagens existentes, e diferentemente da tradicional ferramenta WEKA, utilizada por diversos estudos, colaborando tecnologicamente e cientificamente com uma outra ferramenta, opta-se pela utilização da linguagem Python. Segundo Homem (2020), a linguagem Python é amplamente utilizada no campo de ciência de dados e *machine learning*, graças ao advento de suas bibliotecas. Sendo assim, o presente estudo utiliza-se da biblioteca *Mlxtend* (*machine learning extensions*), a qual possui os algoritmos *Apriori*, *FP-Growth* e *FP-Max*. A

biblioteca *Mlxtend* não possui o algoritmo *Eclat*, como descrito no tópico 2.4 da revisão bibliográfica usando-se da biblioteca pyECLAT.

Primeiramente, antes de implementar os algoritmos e realizar uma análise exploratória, é necessário realizar um pré-processamento e limpeza dos dados, promovendo a remoção dos registros duplicados e ausentes. Após esse pré-processamento, é realizado uma análise minuciosa dos dados, verificando cada atributo, identificando os valores únicos, criando visualizações gráficas e análises estatísticas descritivas das variáveis, removendo possíveis erros na coleta de dados e gerando novas hipóteses e novos atributos. Em seguida, realiza-se a concatenação dos bancos de dados de acidentes com os de características de veículos, com isso, cria-se um relatório com as características e a visualizações gráficas da análise exploratória, sendo possível selecionar e transformar as variáveis para criação dos modelos.

3.2.3 Transformação para variável qualitativa

Com a primeira etapa da metodologia realizada, onde, até esse momento foi realizada a interpretação dos atributos dos dados abertos de acidentes, compreensão do banco de dados com as características dos veículos e exploração das variáveis, de acordo com objetivos específicos desse estudo. A etapa seguinte consiste em tratar os dados para modelagem, criando categorias, ou seja, faixas de grupos dos dados quantitativos, transformando as variáveis quantitativas e qualitativas em apenas variáveis qualitativas, como descrito na etapa de transformação da fundamentação metodológica (capítulo 3.2) de métodos mistos com estratégia transformadora concomitante, onde a integração desses dados ocorre durante a fase de análise (CRESWELL, 2007).

Com os dados limpos e os atributos tratados e transformados em variáveis categóricas, é possível realizar uma análise estatística multivariada para compreender como é a relação entre as variáveis, utilizando uma análise de correspondência.

3.2.4 Análise de correspondência Multivariada

Anterior à aplicação dos algoritmos de *machine learning*, é importante compreender como as variáveis estão relacionadas. Devido às variáveis serem qualitativas categóricas, cria-se uma tabela de contingência com a contagem das observações por variável, em seguida, realiza-se uma análise estatística multivariada pelo teste qui-quadrado na tabela de contingência.

Compreendendo se as características dos acidentes e veículos e suas causas podem ser consideradas independentes, isto é, reconhecer se a frequência das características é a mesma para todas as causas.

Após a análise estatística multivariada e a compreensão da relação dos dados, converte-se o banco de dados com variáveis qualitativas em uma matriz binária, ou seja, cada elemento da matriz indica a presença daquele valor entre todos os valores possíveis de todos os atributos, verificando se está presente ou não naquele acidente, assim como solicitado pelos algoritmos de regras de associação explicado no referencial teórico desse estudo (capítulo 2.3). Com os dados transformados, aplica-se os algoritmos *Apriori*, *Eclat*, *FP-Growth* e *FP-Max.*

3.2.5 Aplicação dos algoritmos

Ainda utilizando a linguagem Python e a biblioteca *Mlxtend* e pyECLAT constrói-se cada modelo de aprendizado de máquina dos algoritmos *Apriori*, *Eclat*, *FP-Growth* e *FP-Max*. Com a aplicação dos algoritmos, obtém-se uma tabela *IF-THEN*, isto é, regras de associação concludentes de uma condição, em outros termos, SE condição ENTÃO conclusão. Tais regras, são criadas com variáveis qualitativas, que seriam as regras de associação, e variáveis quantitativas, que são os índices de *support*, *confidence* e *lift*. Então, cria-se um relatório com representações gráficas dos resultados e compara-se os algoritmos que obtiveram maior quantidade de regras com melhores índices. Além disso, é possível, ainda, analisar os índices das regras que obtiveram maior quantidade de características, categorizando as regras de associação pertinentes para novos estudos, tomadas de decisões e políticas públicas, atendendo os objetivos específicos e respondendo às perguntas de pesquisa do estudo com base em uma fundamentação metodológica mista.

3.2.6 Comparação dos algoritmos

Ainda utilizando a linguagem Python e a biblioteca *Mlxtend* e pyECLAT, constrói-se cada modelo de aprendizado de máquina dos algoritmos *Apriori*, *Eclat*, *FP-Growth* e *FP-Max*. Com a aplicação dos algoritmos, obtém-se uma tabela *IF-THEN* com variáveis qualitativas, as quais seriam as regras de associação, e variáveis quantitativas, que são os índices de *support*, *confidence* e *lift*.

Como o algoritmo *Eclat* tem como resultado somente o índice de suporte, opta-se por utilizar esse índice para comparação entre os algoritmos. Sendo assim, cria-se um relatório com representação gráfica das estatísticas descritivas, comparando a quantidade de regras de cada algoritmo, a quantidade de itens (*IF-THEN*) e os índices de suporte por algoritmo, evidenciando aqueles que obtiveram maior número de observações, número de *IF-THEN*, e melhores índices de suporte. Além disso, executa-se uma análise de variância multivariada (MANOVA) na média de índices de suporte e tamanho de itens é possível testar estatisticamente a igualdade entre médias, conforme realizaram Kaur (2015) e Hunyadi (2011) e, com isso, satisfazendo os objetivos desse estudo e respondendo às perguntas de pesquisa.

Após isso, verifica-se quais regras obtiveram maior quantidade de características (IF) com melhores índices de suporte, categorizando as regras de associação pertinentes para tomadas de decisões e políticas públicas. Desta forma, atende-se os objetivos específicos e responde-se às perguntas de pesquisa do estudo, com base em uma fundamentação metodológica mista.

4 DESENVOLVIMENTO E RESULTADOS

Nesse capítulo, aplicou-se a metodologia descrita no capítulo anterior nos dados de acidentes das rodovias federais brasileiras e das características dos veículos, explicando todas etapas e decisões tomadas durante o tratamento e a análise dos dados. Por fim, discutiu-se os resultados obtidos, atingindo-se os objetivos e respondendo às perguntas de pesquisa. Após compreender o problema de pesquisa, aplicou-se a metodologia, obtendo-se os dados para implementação dos algoritmos *Apriori*, *Eclat*, *FP-Growth* e *FP-Max.*

4.1 Obtenção dos dados com variáveis qualitativas e quantitativas

Um ponto importante do *Big Data Analytics*, em cidades inteligentes, são os dados abertos, os quais são essenciais para o desenvolvimento destas, conforme afirmam Ahlgren *et al.* (2016). Existem evidências de que os dados abertos contribuem para melhorar a entrega do serviço público em contextos de cidades inteligentes, segundo Pereira *et al.* (2017), sendo assim, o presente estudo optou por coletar dados abertos dos acidentes de trânsito, disponibilizados pela da Polícia Rodoviária Federal do Brasil, e concatená-los com dados das características dos veículos constantes no Registro Nacional de Veículos Automotores (Renavam), concedidos pelo Ministério da Infraestrutura.

4.1.1 Dados de acidentes da Polícia Rodoviária Federal Brasileira

Os dados de acidentes foram obtidos no site da PRF (Polícia Rodoviária Federal) do Brasil, onde foram coletados os dados agrupados por pessoas, com todas as causas e tipos de acidentes dos anos de janeiro de 2017 a fevereiro de 2020. Os dados foram selecionados até fevereiro de 2020, devido ao início do estado de calamidade pública em razão da pandemia de COVID-19, aprovado pelo Congresso Nacional em 20 de março de 2020, conforme decreto legislativo nº 6 de 2020 (Brasil, 2020), onde essa mudança de rotina da população, durante um longo período, poderia afetar na criação do modelo. Optou-se, então, por esse banco de dados por ainda não ser estudado por técnicas de aprendizado de máquina, além de incluírem a marca do veículo, diferentemente dos bancos de dados dos anos anteriores.

Os bancos de dados agrupados por pessoa, com todas as causas e tipos de acidentes registrados a partir de janeiro de 2017, são acompanhados de um dicionário de variáveis de acidentes (Brasil, 2017), conforme descrito na Tabela 4.1.

Tabela 4.1 – Dicionário de variáveis dos bancos de dados de acidentes da PRF

Variável	Descrição
Id	Variável com valores numéricos, representando o identificador do acidente.
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.
horario	Horário da ocorrência no formato hh:mm:ss
uf	Unidade da Federação. Ex.: MG, PE, DF, etc.
br	Variável com valores numéricos, representando o identificador da BR do acidente.
km	Identificação do quilômetro onde ocorreu o acidente, com valor mínimo de 0,1 km e com a casa decimal separada por ponto.
municipio	Nome do município de ocorrência do acidente.
causa_principal	Identifica se a causa do acidente foi considerada como principal pelo policial.
causa_acidente	Causa presumível do acidente, baseada nos vestígios, indícios e provas colhidas no local do acidente.
ordem_tipo_acidente	Valor numérico que identifica a sequência dos eventos sucessivos que ocorreram no acidente.
tipo_acidente	Identificação do tipo de acidente.
classificação_acidente	Classificação quanto à gravidade do acidente: Sem Vítimas, Com Vítimas Feridas, Com Vítimas Fatais e Ignorado.
fase_dia	Fase do dia no momento do acidente.
sentido_via	Sentido da via considerando o ponto de colisão: Crescente e decrescente.
condição_meteorologica	Condição meteorológica no momento do acidente.
tipo_pista	Tipo da pista considerando a quantidade de faixas: dupla, simples ou múltipla.
tracado_via	Descrição do traçado da via.
uso_solo	Descrição sobre as características do local do acidente: Urbano=Sim; Rural=Não.
id_veiculo	Variável com valores numéricos, representando o identificador do veículo envolvido.
tipo_veiculo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro.
marca	Descrição da marca do veículo.
ano_fabricacao_veiculo	Ano de fabricação do veículo, formato aaaa.
tipo_envolvido	Tipo de envolvido no acidente conforme sua participação no evento.
estado_fisico	Condição do envolvido conforme a gravidade das lesões.
idade	Idade do envolvido. O código “-1” indica que não foi possível coletar tal informação.
sexo	Sexo do envolvido. O valor “inválido” indica que não foi possível coletar tal informação.

Continua...

ilesos	Valor binário que identifica se o envolvido foi classificado como ileso.
feridos_leves	Valor binário que identifica se o envolvido foi classificado como ferido leve
feridos_graves	Valor binário que identifica se o envolvido foi classificado como ferido grave.
mortos	Valor binário que identifica se o envolvido foi classificado como morto.
latitude	Latitude do local do acidente em formato geodésico decimal.
longitude	Longitude do local do acidente em formato geodésico decimal.
regional	Regional da delegacia.
delegacia	Delegacia onde foi realizada a ocorrência.
uop	Posto de operação.

Fonte: Adaptado de Brasil (2017)

Interpretar os atributos dos dados abertos de acidentes em rodovias federais brasileiras registrados pela Polícia Rodoviária Federal, é essencial para tratamento e aplicação dos algoritmos, determinado como objetivo específico desta pesquisa. Então, cada variável será explorada detalhadamente no item 4.2 deste capítulo. Esses bancos de dados de acidentes de trânsito em rodovias federais brasileiras dos anos de 2017, 2018, 2019 e 2020 foram concatenados com os dados das características dos veículos.

4.1.2 Dados das características dos veículos

No banco de dados da PRF não consta as características dos veículos, apenas o tipo de veículo, marca e ano de fabricação. Sendo assim, utilizou-se dados da potência de automóveis registrados no Renavam, disponibilizados pelo Ministério da Infraestrutura do Brasil (MINFRA), por meio de uma solicitação às informações públicas no portal Fala.br (https://sistema.ouvidorias.gov.br/). Apesar de não serem dados abertos, tal qual os dados de acidentes, qualquer cidadão pode solicitar estes, amparado pela Lei de acesso à informação, Lei nº 12.527, de 18 de novembro de 2011 (Brasil, 2011). Com esses dados, foi identificada a potência do motor para cada modelo de automóvel, conforme documento veicular do Denatran (Departamento Nacional de Trânsito), e relacionado com o modelo presente no banco de dados de acidentes.

Realizou-se, então, a solicitação dos dados pelo processo SEI nº 50650.003964/2020-85 ao MINFRA, através do portal Fala.br, no dia 05 de agosto de 2020, sendo atendida pela Coordenação Geral de Sistemas, Informações e Estatísticas (CGSIE) do Departamento Nacional de Trânsito (Denatran) da Secretaria Nacional de Transportes Terrestres (SNTT) do Ministério da Infraestrutura, enviando por *e-mail* e tendo a descrição de acordo com a Tabela 4.2.

Tabela 4.2 – Dicionário de variáveis do banco de dados das características dos veículos

Variável	Descrição
Tipo Veículo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro.
Código Marca Modelo Veículo	Variável com valores numéricos, representando o identificador da marca.
Marca Modelo	Descrição da marca do veículo.
Ano Fabricação Veículo	Ano de fabricação do veículo, formato aaaa.
Combustível Veículo	Descrição do combustível do veículo.
Potência Veículo – Frota Atual	Variável com valores numéricos, representando a potência de um veículo expressa em CV (cavalos a vapor).
Eixos Veículo – Frota Atual	Variável com valores numéricos, representando a quantidade de eixos do veículo.
Cilindradas Veículo – Frota Atual	Variável com valores numéricos, representando a capacidade voluntária do motor expressa em centímetros cúbicos.
Qtd. Veículos Frota Atual	Variável com valores numéricos, representando a quantidade de veículos na frota atual registrada no Denatran.

Fonte: Elaborado pelo autor

Em vista disso, percebe-se que os bancos de dados de acidentes e o banco de dados das características dos veículos possuem variáreis qualitativas e quantitativas, conforme determinado pelo método misto de estratégia transformadora concomitante, conforme Creswell *et al.* (2007) e descrito na metodologia do presente estudo, ou seja, os dois tipos de dados são coletados simultaneamente. Em seguida, conforme metodologia realiza-se o processamento, a análise exploratória e o tratamento preciso dos dados, conforme segue.

4.2 Análise exploratória das variáveis

Para compreender melhor os dados, realizou-se um pré-processamento e uma análise exploratória dos dados através da linguagem Python, versão 3.7.8, e da biblioteca *Mlxtend* (*machine learning extensions*) que possui os algoritmos *Apriori*, *FP-Growth* e *FP-Max* e da biblioteca pyECLAT para aplicação do algoritmo *Eclat.* Porém, antes de implementar os algoritmos e realizar uma análise exploratória, é necessário realizar um pré-processamento e limpeza dos dados.

As bibliotecas utilizadas para desenvolvimento do estudo foram a pandas e *numpy* para manipulação dos dados, a biblioteca *holidays* para definição dos feriados no brasil, as bibliotecas *seaborn* e *matplotlib* para criação de gráficos, utilizando conceitos de Data *Storytelling*, isto é, contando uma história através dos dados.

4.2.1 Pré-processamento e limpeza dos dados

A primeira etapa do pré-processamento é a importação dos dados. Em um primeiro momento, importou-se separadamente os bancos de dados dos acidentes dos anos de 2017, 2018, 2019 e 2020, descritos no tópico 4.1.1, em seguida, a importação dos dados das características dos veículos, descritos no capítulo tópico 4.1.2. Como os bancos de dados de acidentes possui o mesmo número de colunas, com os mesmos tipos e nomes, entende-se que os bancos de dados de acidentes são semelhantes e podem ser concatenados sem nenhuma obstrução. Após concatená-los, obtêm-se 1.099.273 observações e 37 variáveis. Com esse grande volume de dados, é necessário um tratamento preciso para a aplicação de *machine learning*, visto que um tratamento preciso é importante para garantir a confiabilidade dos resultados.

Com o banco de dados de acidentes importados e concatenados, iniciou-se a limpeza dos dados. Inicialmente, percebeu-se, pelo resultado das informações do banco de dados, que as colunas não possuem a mesma quantidade de registos. Optou-se pela remoção de registros ausentes e duplicados. Então, determinou-se a remoção dos registros nulos e duplicados, devido ao grande volume de registros, os quais podem ser removidos sem prejudicar o resultado dos algoritmos. Após a remoção dos valores ausentes e duplicados, o banco de dados de acidentes reduziu de 1.099.273 para 830.290 registos. Com isso, se torna necessário entender melhor cada variável para realizar uma exploração dos atributos, transformar as variáveis e aplicar os algoritmos.

4.2.2 Tratamento das variáveis do banco de dados de acidentes

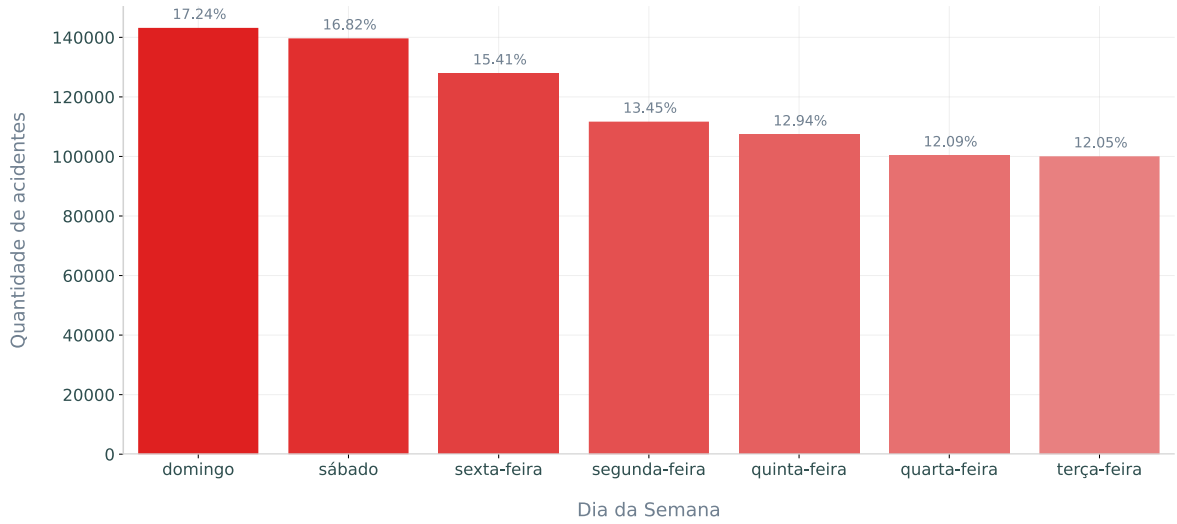
Para executar uma exploração minuciosa dos dados, primeiramente, então, é necessário entender cada variável. Realizou-se a verificação de cada um dos atributos, seguindo a ordem do dicionário dos bancos de dados, conforme descrito a seguir, onde o significado de cada atributo está especificado nas Tabela 4.1 e 4.2 deste capítulo.

As variáveis identificação do acidente e identificação da pessoa envolvida são chaves primárias dos bancos de dados, as quais não foram utilizadas nesse estudo e foram removidas.

Verificando o atributo data do acidente, que tem 1.186 valores únicos, percebe-se que o maior número de acidentes ocorreu no dia 23/12/2017, isto é, véspera do feriado de natal no Brasil, o que se leva a hipótese de que o feriado é uma possível variável significativa para regras de associação, onde nova variável categórica será criada no capítulo 4.3.

Seguindo o dicionário, a próxima variável é o dia da semana, onde percebe-se que os acidentes ocorrem em dias próximos ao final de semana, como demostrado no Gráfico 4.1.

Gráfico 4.1 – Quantidades de acidentes por dia da semana

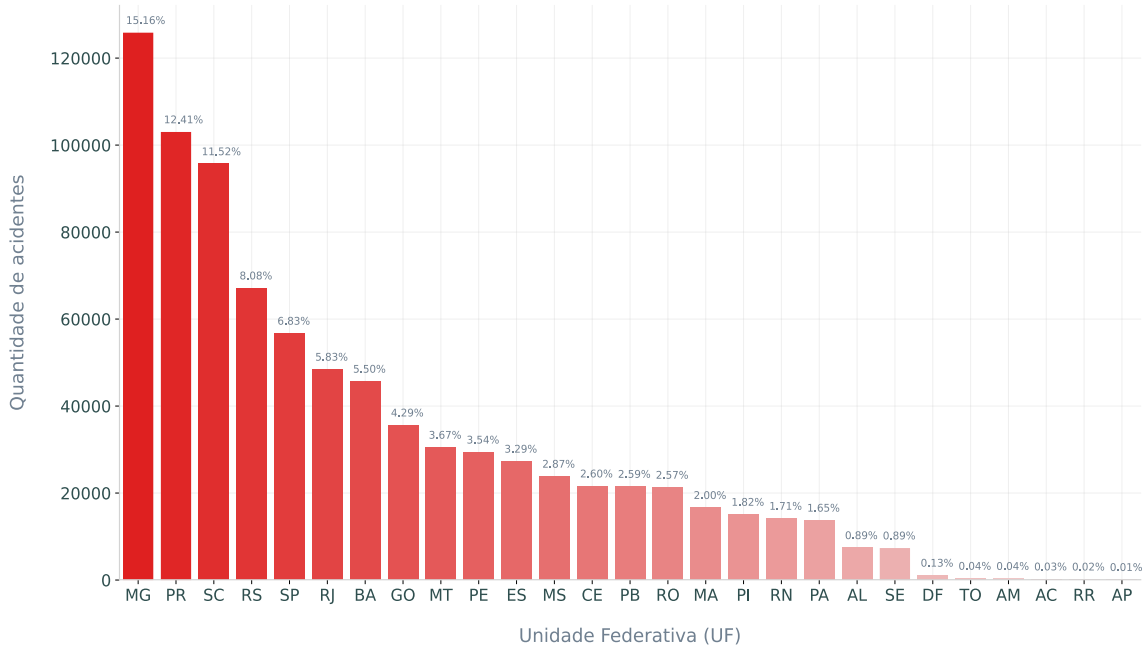


Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Um outro atributo interessante é o horário, o qual apresenta 1.430 valores únicos, onde a maioria dos acidentes ocorreram no período de 17:00 às 19:00. Esse período de apenas 2 horas do dia corresponde a, aproximadamente, 14,68% dos acidentes, os quais aconteceram, principalmente, às 18:30, horários denominados como horários de pico da tarde. Este, então, representa mais um fator com potencial de ser significativo para o algoritmo de regras de associação, o qual será categorizado no capítulo 4.3 de transformação das variáveis quantitativas em qualitativas. A variável UF (Unidade da Federação) apresenta 27 valores únicos, ou seja, aconteceram acidentes em todos os estados brasileiros. Essa análise apresenta dados interessantes, onde, por exemplo, o estado em que mais ocorreu acidente no período de janeiro de 2017 a fevereiro de 2020 em rodovias federais, foi Minas Gerais, seguido por Paraná e Santa Catarina. conforme Gráfico 4.2. Ainda, por meio dos dados é possível verificar que, apenas nos 7 estados das regiões sul e sudeste, ocorreram cerca de 63,12% dos acidentes em rodovias federais brasileiras. Mas, ocasionalmente, segundo o Relatório de Frotas de Veículos de Fevereiro de 2020 disponibilizado em Dados Abertos do Senatram (Ministério da Infraestrutura), no Brasil, o total de automóveis em fevereiro de 2020 era de 56.927.310 veículos, desse total 43.098.804 foram registrados na região Sudeste, ou seja, 75,71% dos veículos. Entretanto, não é viável analisar apenas dados da frota, pois o veículo pode estar registrado em uma região, porém, circulando e

acidentando-se em outra. Logo, deve-se analisar a quilometragem da malha rodoviária das rodovias federais por estado. Segundo o relatório de evolução da malha rodoviária do Anuário CNT (Confederação Nacional do Transporte), em 2020, o Brasil possuía um total de 73.328,70 km de malha rodoviária federal, sendo 64.022,40 km pavimentadas e 9.306,30 km não pavimentadas. Desses 73.328,70 km, apenas 24.788,10 km estão localizadas nas regiões Sul e Sudeste, isto é, 33,80% da malha rodoviária federal. Apesar de possuir apenas 33,80% da malha rodoviária, ocorrem 63,12% dos acidentes, porém, mais de 75% da frota de automóveis são registrados no Sul e no Sudeste, ou seja, é necessária uma análise de diversos outros fatores, tais como dados de contagem de tráfego em rodovias, e outros fatores, como polos socioeconômicos, para compreender esses números, caso que não é objeto desse estudo.

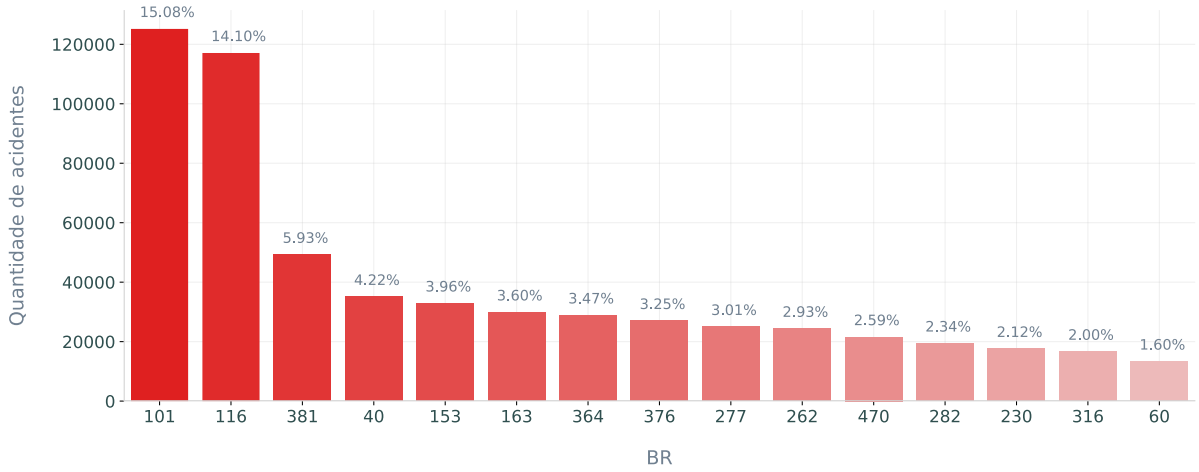
Gráfico 4.2 – Quantidade de acidentes por estado



Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Para o atributo rodovia (br), os quais obtiveram 124 valores únicos, as rodovias onde mais ocorreram acidentes desse banco de dados foram as BRs 101, 116 e 381, de acordo com Gráfico 4.3, que apresenta as 15 rodovias com maior índice de incidência para o período analisado. No entanto, vale ressaltar que a BR 101 não passa pelo estado de Minas Gerais, a qual foi a unidade federativa onde mais ocorreram acidentes.

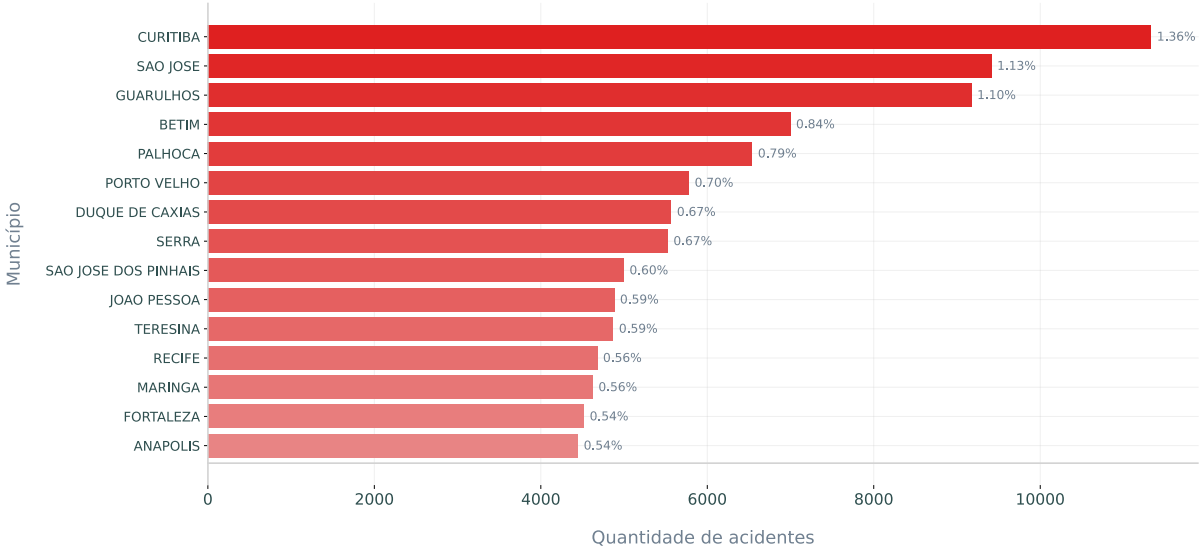
Gráfico 4.3 – As 15 rodovias com maior número de acidentes



Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Outros atributos analisados foram o km onde o acidente ocorreu (9.341 valores únicos) e o município. A variável km é uma variável com muitos valores únicos e não será utilizado na modelagem, como determinado na metodologia desse estudo. Referente ao atributo munícipio, este apresentou 1.921 dados, e, entre as cidades com maiores índices de acidentes estão Curitiba, São José, Guarulhos e Betim. Nota-se que a cidade de Guarulhos, a terceira cidade que mais ocorreu acidentes, é do estado de São Paulo, o quinto estado no gráfico de quantidade de acidentes por UF (Gráfico 4.2). Por ser uma variável com muitos valores únicos, também não será utilizada na modelagem. O Gráfico 4.4 apresenta os 15 principais municípios.

Gráfico 4.4 – Os 15 municípios com maior número de acidentes



Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Como não está sendo posta em análise a localização dos acidentes, mas sim as características das vias, não serão utilizadas variáveis como UF, br, munícipio e km. Apesar da localização não ser utilizada na aplicação dos algoritmos, essa análise exploratória das variáveis apresenta dados relevantes para outras aplicações, como descrito no capítulo de considerações finais.

Seguindo a ordem do dicionário do banco de dados (Tabela 4.1), como causas de acidentes, o banco de dados da PRF determina se aquela causa do acidente foi principal ou não, dois valores únicos. Para uma análise mais criteriosa a respeito da causa principal do acidente, optou-se por considerar somente as causas principais, removendo-se os registos com mais de uma causa, isto é, o banco de dados de acidentes contém 663.184 registros, sendo somente uma causa por registro, evitando redundância. O atributo causas do acidente possui 24 valores únicos, como exibido em ordem decrescente na Tabela 4.3.

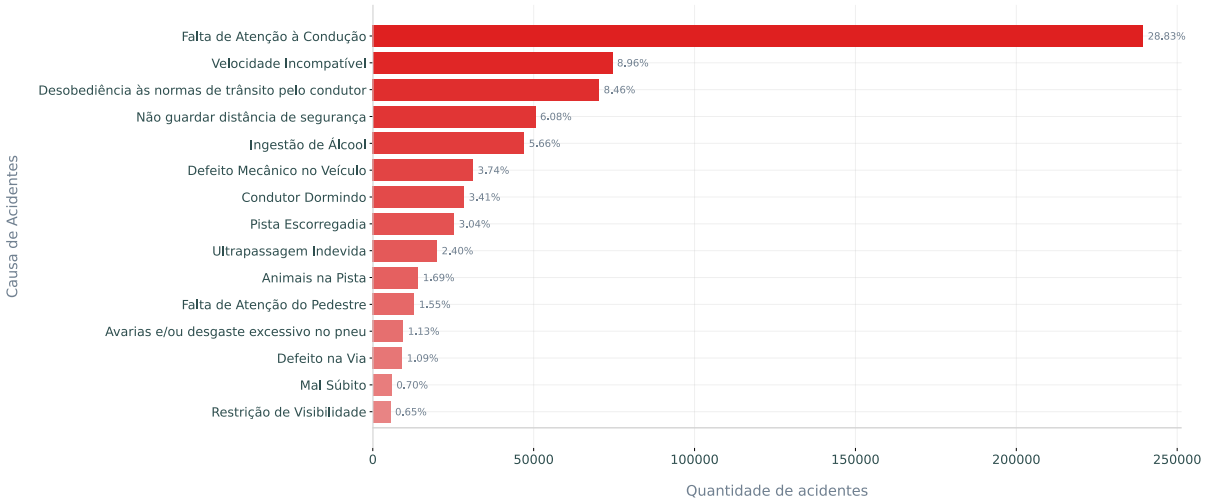
Tabela 4.3 – Causas de acidentes das rodovias federais brasileiras

Nº	**Causa principal do acidente**	**Quantidade de acidentes**
1	Falta de atenção à condução	239.349
2	Velocidade incompatível	74.414
3	Desobediência às normas de trânsito pelo condutor	70.256
4	Não guardar distância de segurança	50.489
5	Ingestão de álcool	46.996
6	Defeito mecânico no veículo	31.091
7	Condutor dormindo	28.314
8	Pista escorregadia	25.212
9	Ultrapassagem indevida	19.901
10	Animais na pista	14.070
11	Falta de atenção do pedestre	12.853
12	Avarias e/ou desgaste excessivo no pneu	9.401
13	Defeito na via	9.077
14	Mal súbito	5.839
15	Restrição de visibilidade	5.437
16	Objeto estático sobre o leito carroçável	4.795
17	Sinalização da via insuficiente ou inadequada	2.806
18	Carga excessiva e/ou mal acondicionada	2.690
19	Fenômenos da natureza	2.388
20	Agressão externa	1.969
21	Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	1.800
22	Deficiência ou não acionamento do sistema de iluminação/sinalização do veículo	1.793
23	Desobediência às normas de trânsito pelo pedestre	1.631
24	Ingestão de substâncias psicoativas	613

Fonte: Adaptado pelo autor com base nos dados da PRF de 2017 a 2020

Percebe-se a grande diferença entre as causas de acidentes por falta de atenção do condutor e as demais causas, fato relevante para o modelo, uma vez que esses dados serão associados com as características do condutor. Tem-se que as três principais causas de acidentes são a falta de atenção à condução, a velocidade incompatível e a desobediência às normas de trânsito pelo condutor. O Gráfico 4.5 apresenta as 15 principais causas.

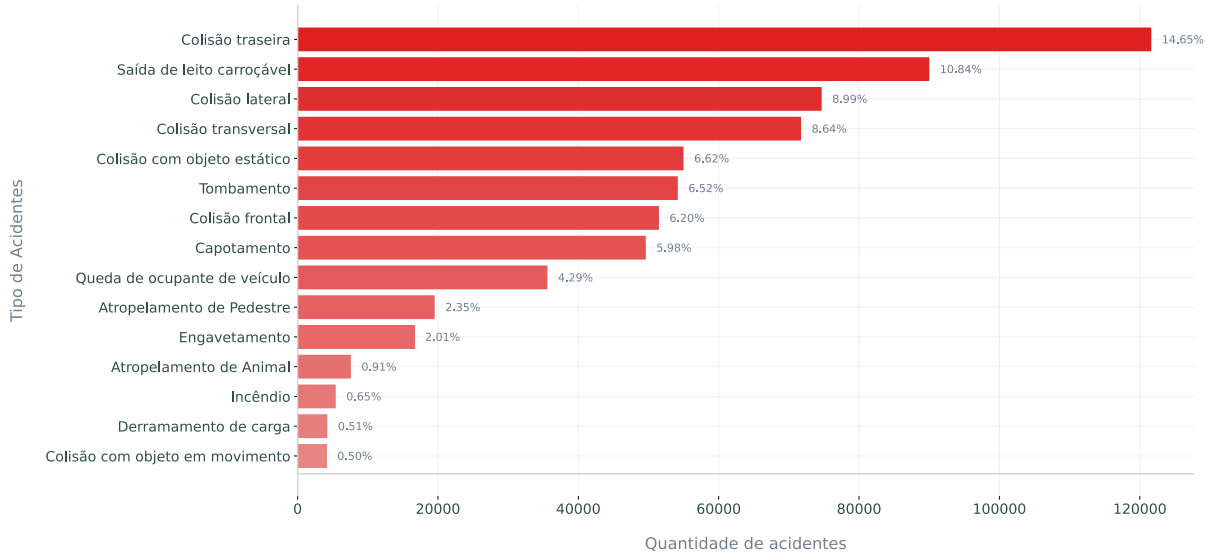
Gráfico 4.5 – As 15 maiores causas de acidentes



Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Referente ao atributo ordem do tipo de acidente, este apresenta 11 valores únicos, no entanto, não tem muita significância nesta pesquisa, bem como o tipo de acidente, o qual possui 16 tipos, sendo eles: colisão traseira, saída de leito carroçável, colisão lateral, colisão transversal, colisão com objeto estático, tombamento, colisão frontal, capotamento, queda de ocupante de veículo, atropelamento de pedestre, engavetamento, atropelamento de animal, incêndio, derramamento de carga, colisão com objeto em movimento, danos eventuais. O Gráfico 4.6 apresenta os 15 principais tipos de acidentes.

Gráfico 4.6 – Os 15 maiores tipos de acidentes



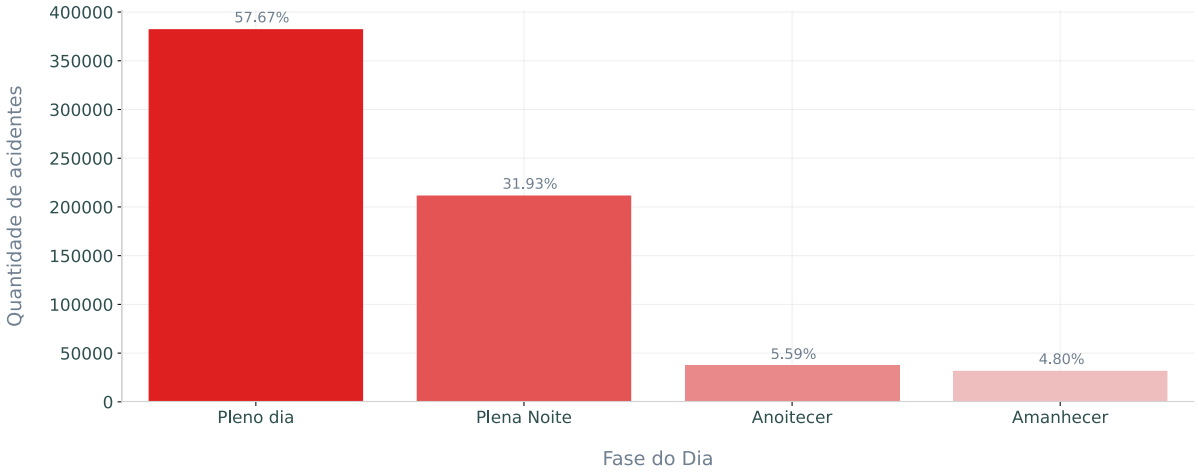
Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

O atributo tipo de acidente não é utilizado na modelagem, pois tem-se o interesse nas características anteriores ao ocorrido do acidente e o tipo de acidente é resultado deste, ou seja, atributos pós acidente.

Desses acidentes, foram 479.564 com vítimas feridas, 117.568 sem vítimas e 66.052 com vítimas fatais. Apesar do número de acidentes com vítimas fatais, felizmente, ser muito inferior ao de vítimas feridas, há um agravante que é a variável estado físico do envolvido, se é ileso, ferimentos leves, graves ou óbitos. Apesar de não se utilizar da classificação do acidente e estado físico do envolvido, essa variável demonstra mais uma vez a importância social do estudo dos acidentes de trânsito.

Já no atributo fase do dia, um atributo do meio ambiente, percebe-se que a maioria dos acidentes ocorrem durante o dia, seguido pela noite e com poucos registros ao anoitecer e amanhecer, como demonstrado no Gráfico 4.7. Um outro ponto importante é que, para criação do modelo, ou utiliza-se o atributo horário ou a fase do dia, evitando-se redundância como a fase do dia possui menor número de registros únicos, optou-se por utilizá-la. O atributo horário foi utilizado para categorizar se o momento em que o acidente ocorreu é horário de pico, como determinado no início da análise exploratória.

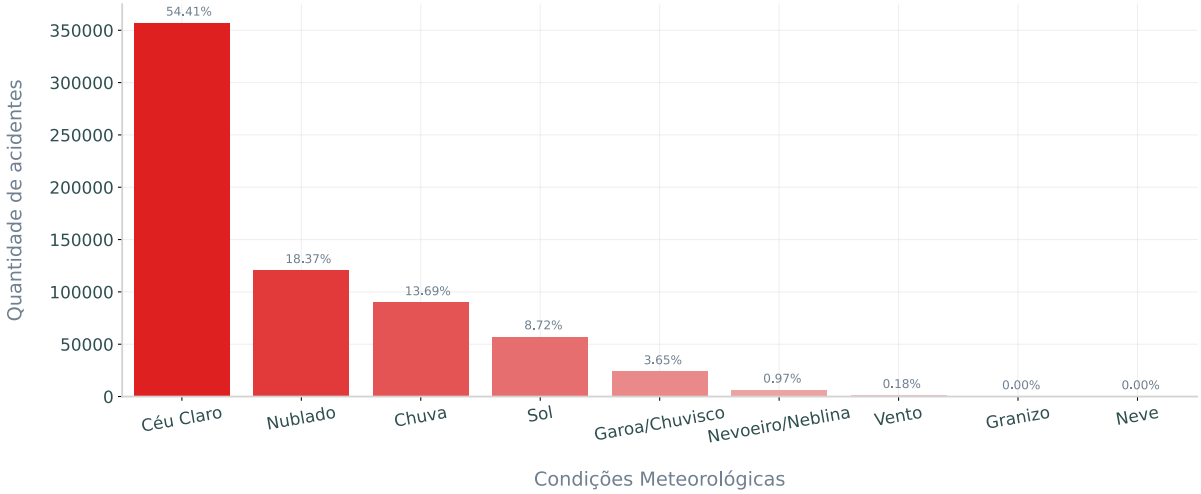
Gráfico 4.7 – Acidentes por fase do dia



Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Percebe-se que a maioria dos acidentes ocorreram em plena luz do dia, algo que pode ser relevante para as regras de associação. Outro atributo importante referente ao meio em que o acidente ocorreu são as condições meteorológicas, as quais contém 10 valores únicos. Entretanto, há observações registradas como ignorado, as quais foram removidas e, assim, as condições na qual ocorreram mais acidentes são expressas no Gráfico 4.8.

Gráfico 4.8 – Acidentes por condição meteorológica

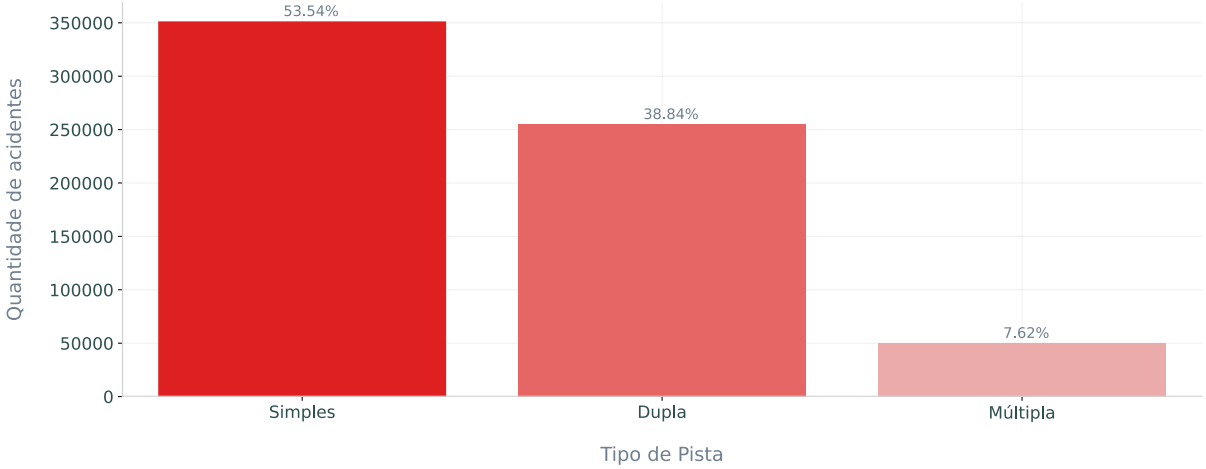


Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

O atributo sentido da via, o qual possui dois valores (crescente e decrescente), é interessante quando utilizado juntamente com a br e km, no entanto, uma vez que não serão utilizados os atributos br e km, também não utilizará o sentido da via. Já o atributo tipo de pista tem 3 valores

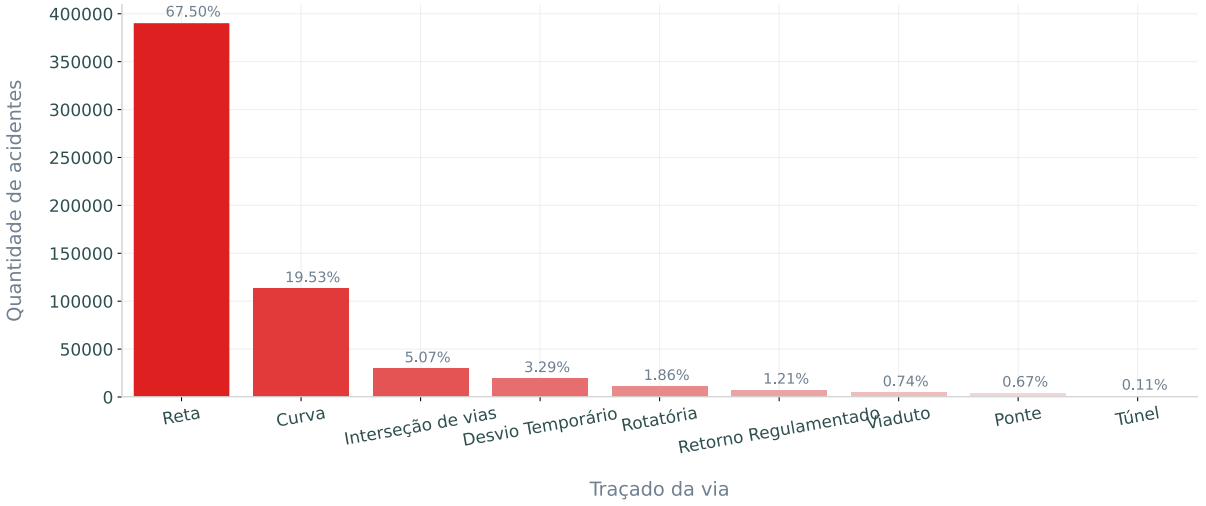
únicos (simples, dupla, múltipla) e o atributo traçado da via possui 10 valores únicos, o qual é relevante ser utilizado nos modelos desta pesquisa, visto que possui características interessantes da via, como definido no referencial teórico. Esse atributo traçado da via possui registros denominados como não informado e, por esta razão, optou-se por removê-los. Os Gráficos 4.9 e 4.10 exibem o número de acidentes por tipo de pista e traçado de via, respectivamente, onde percebe-se que o número de acidentes em pista simples é superior ao número em pista dupla e múltipla e, ainda, em traçado de via reto é bem superior aos demais.

Gráfico 4.9 – Quantidade de acidentes de trânsito por tipo de pista



Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Gráfico 4.10 – Quantidade de acidentes de trânsito por traçado da via



Fonte: Elaborado pelo autor com base nos dados da PRF de 2017 a 2020

Até o momento, após todas exclusões, o atributo uso do solo apresentou 338.816 acidentes em áreas rurais e 239.157 em áreas urbanas. Entretanto, como não é relevante a localização do

acidente, não será utilizada. Da mesma forma, a variável identificação veículo também não tem significância nesse trabalho, por ser uma variável de chave primária esta foi removida.

Então, esse banco de dados possui 21 tipos de veículos. No entanto, conforme descrito na introdução e na obtenção dos dados das características dos automóveis, esse estudo pretende utilizar apenas a potência dos automóveis, logo, é necessária a criação de um novo banco de dados, selecionando somente os registros de automóveis. À vista disso, o novo banco de dados de acidentes de automóveis possui 265.928 observações. A variável marca será explorada mais adiante, pois nesse banco de dados será utilizada apenas para concatenar com o banco de dados de potência.

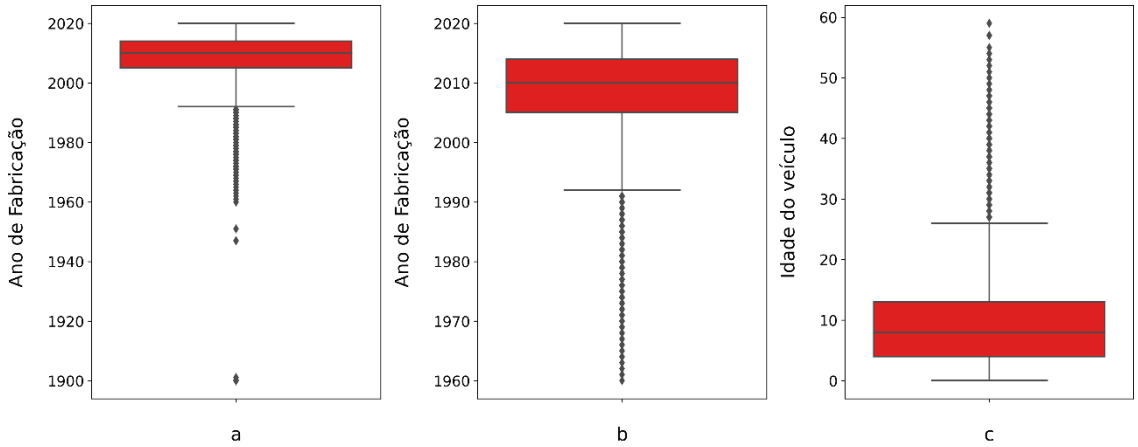
O atributo ano de fabricação de veículo foi utilizado para definir a idade do automóvel, pois entende-se que, como os registros das ocorrências são em anos diferentes, não é interessante quando o veículo foi fabricado, mas sim a idade que o veículo tinha naquele instante em que o acidente ocorreu.

Realizando uma estatística descritiva na variável do ano de fabricação do veículo, nota-se a amplitude dos dados de 1900 a 2020. Portanto, existindo presença de *outliers*, como demonstra o *boxplot* do Gráfico 4.11 (a).

Para garantir a confiabilidade dos dados, decidiu-se selecionar os veículos fabricados após o ano de 1956, ano onde foi fabricado o primeiro carro em série em solo brasileiro (ANGOLINI, 2005). Com isso, tem-se um novo *boxplot* (Gráfico 4.11) (b).

Para encontrar a idade do veículo, criou-se uma nova variável com a subtração do ano que o acidente ocorreu pelo ano de fabricação do veículo, obtendo o *boxplot*, logicamente, inverso ao do ano de fabricação após 1956 (Gráfico 4.11) (c).

Gráfico 4.11 – (a) *Boxplot* do ano de fabricação dos automóveis. (b) *Boxplot* do ano de fabricação dos automóveis de 1956 a 2020. (c) *Boxplot* da idade do veículo



Fonte: Elaborado pelo autor

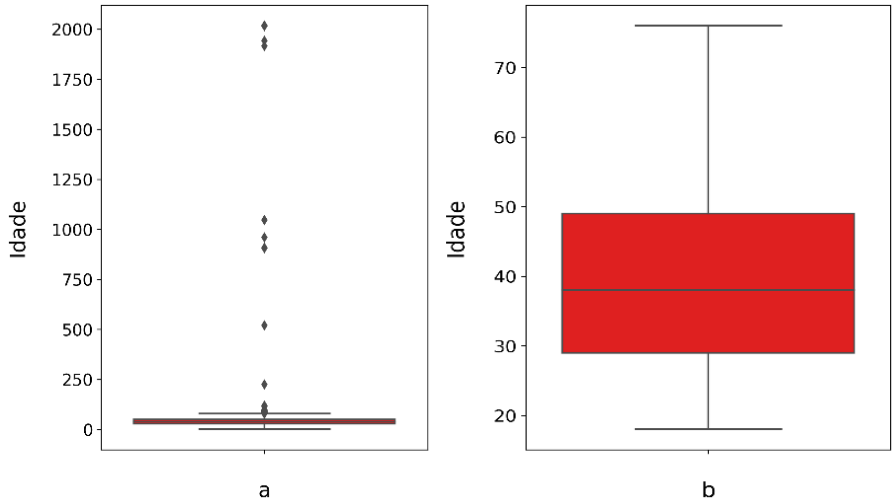
No atributo tipo de envolvido, que apresenta 4 valores únicos, sendo eles condutor, passageiro, pedestre e cavaleiro, optou-se por selecionar somente o condutor para considerar que há o mesmo número de pessoas em todos os acidentes, isto é, apenas o condutor, evitando que o modelo sofra influências. Além de que, este estudo procura saber a influência das características do condutor na gravidade do acidente e, então, diante disso, o novo banco de dados com apenas os condutores possui 171.493 registros.

Uma variável que não tem interesse para esse estudo é a estado físico do envolvido, a qual exibe a quantidade de ilesos, lesões leves, lesões graves e óbitos, pois é uma variável que tem características pós-acidente.

Como características do condutor, tem-se as variáveis idade e sexo onde, na idade do condutor, é possível perceber que há valores absurdos, como idades de 0 até 2018 anos, conforme *boxplot* a seguir (Gráfico 4.12) (a).

Nota-se que há um erro no qual deve-se ser tratado e a existência de muitos *outliers*, pois não há evidências de condutores com 0 ano e 2018 anos de vida. Com isso, determinou-se que, como foi utilizando apenas os condutores, conforme descrito no atributo tipo de envolvido, selecionou-se os registros com idades entre 18 a 76 anos, visto que 18 anos é a idade mínima permitida para dirigir legalmente no Brasil e 76 anos que a expectativa de vida no Brasil, segundo o IBGE (2020), resultando no *boxplot* do Gráfico 4.12 (b).

Gráfico 4.12 – (a) *Boxplot* das idades dos condutores. (b) *Boxplot* das idades dos condutores entre 18 e 76 anos



Fonte: Elaborado pelo autor

A variável sexo do envolvido possui 3 valores únicos, sendo 139.932 Masculino, 30749 Feminino e 11 Ignorado. Como descrito no dicionário desse banco de dados, o valor ignorado indica que não foi possível coletar a informação, sendo assim, optou-se por removê-los.

Os atributos ilesos, feridos leves, feridos graves e mortos não têm significância nesse estudo, conforme descrito anteriormente, bem como as demais variáveis como latitude, longitude, regional, delegacia e uop, pois não se pretende utilizar a localização exata e nem de onde foi realizado o boletim de ocorrência. O banco de dados dos acidentes tratado possui 169.680 observações.

Conforme descrito nos tópicos 4.1.1 e 4.1.2 desse estudo, onde nota-se que ambos os bancos de dados têm em comum as colunas marca e ano de fabricação, utilizou-se essas colunas para concatenar os bancos de dados. Com o banco de dados de acidentes tratado e preparado para receber o banco de dados das características dos veículos, é necessário realizar o tratamento dos dados do banco de dados que contém a potência do motor.

4.2.3 Tratamento dos dados das características dos veículos

O banco de dados do Renavam contém 482.312 observações e 8 atributos, no qual são descritos na Tabela 4.12. Como o intuito desse estudo é utilizar a potência do motor do veículo, eliminou-se as variáveis que não são foco desse estudo.

Nota-se, nesse banco de dados, a existência de valores ausentes. Para não prejudicar a concatenação, e seguindo o critério e procedimento do banco de dados de acidentes, foram removidos os valores ausentes e duplicados, retornando um banco de dados com 238.480 registros. Antes de concatenar os bancos de dados de características de veículos e acidentes, deve-se analisar a quantidade de veículos em frota, registrados no Renavam, de acordo com esse relatório emitido em 2020. Nessa base de dados, tem-se 54.997.729 veículos registrados em frota. Entretanto, não necessariamente essa é a frota real do país, uma vez que se pode ter veículos sem registro ou veículos com registros duplicados. Para confortar esses valores, obteve-se a base de dados de Frota Nacional de Agosto de 2020 no site do Ministério da Infraestrutura com as informações detalhadas na Tabela 4.4.

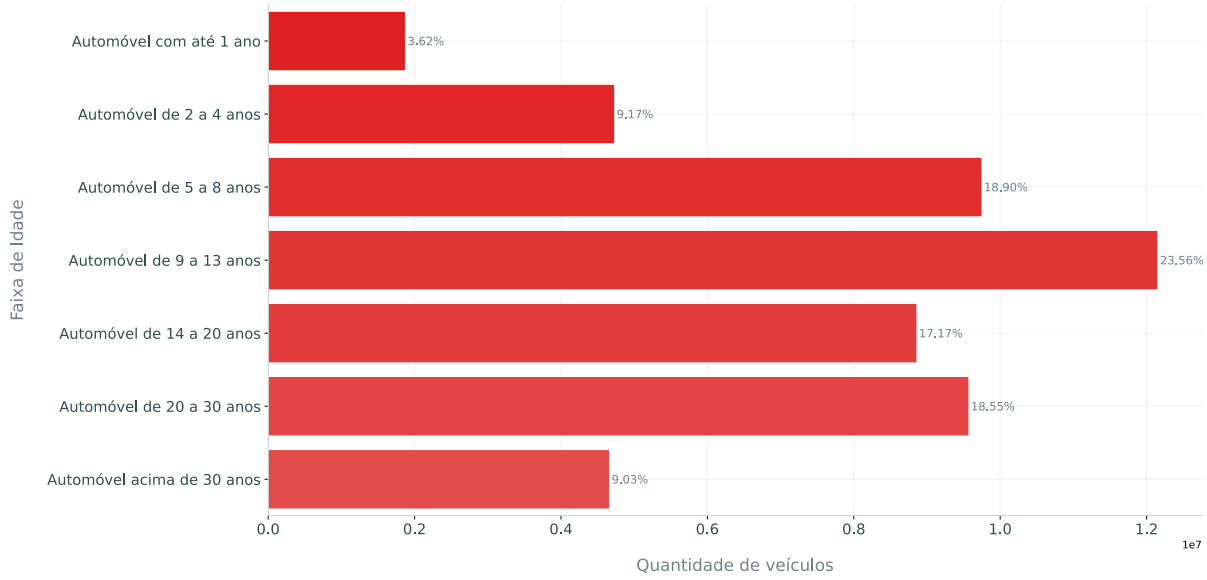
Tabela 4.4 – Quantidade de frota de automóveis (Ministério da Infraestrutura – agosto, 2020)

Grandes Regiões e Unidades da Federação	AUTOMÓVEL
Brasil	**57.424.520**
Norte	**1.867.077**
Acre	93.901
Amapá	88.243
Amazonas	420.606
Pará	648.245
Rondônia	304.817
Roraima	80.874
Tocantins	230.391
Nordeste	**7.277.071**
Alagoas	381.800
Bahia	1.942.801
Ceará	1.211.933
Maranhão	468.421
Paraíba	565.606
Pernambuco	1.387.371
Piauí	385.177
Rio Grande do Norte	586.690
Sergipe	347.272
Sudeste	**31.192.317**
Espírito Santo	1.004.462
Minas Gerais	6.533.445
Rio de Janeiro	4.680.190
São Paulo	18.974.220
Sul	**12.233.077**
Paraná	4.639.078
Rio Grande do Sul	4.486.989
Santa Catarina	3.107.010
Centro-Oeste	**4.854.978**
Distrito Federal	1.345.536
Goiás	1.945.652
Mato Grosso	785.558
Mato Grosso do Sul	778.232

Fonte: Ministério da Infraestrutura (2020)

Percebe-se uma divergência entre as informações de frota de dados abertos do Ministério da Infraestrutura de Agosto de 2020 e os dados obtidos do Renavam, porém, devido à gradualidade das informações, onde os dados do Renavam constam marca, ano de fabricação e potência do motor, esses dados foram escolhidos para serem analisados. Sendo assim, tem-se em faixa de idade veículo, que será comentada a classificação posteriormente, onde a maioria dos veículos está entre 9 a 13 anos de fabricação (Gráfico 4.13).

Gráfico 4.13 – Quantidade de veículos por idade



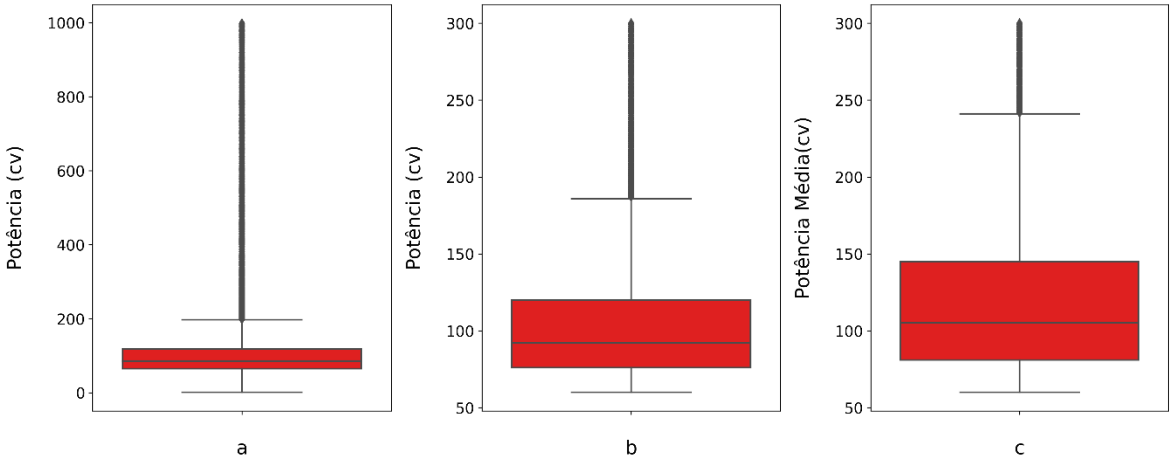
Fonte: Adaptado de Ministério de Infraestrutura (2020)

Nesse caso, não é necessário tratar a coluna de ano de fabricação, uma vez que ela será utilizada apenas para concatenar os bancos de dados, sendo necessário realizar o tratamento dos dados da potência. Realizando a estatística descritiva da potência do motor dos veículos, obtém-se o *boxplot* do Gráfico 4.14 (a).

Percebe-se a amplitude dos dados e a existência de *outliers* onde, para resolver esse problema, optou-se por selecionar as potências entre 60 a 300 cv (cavalo-vapor), valores comuns de possíveis automóveis existentes, atualmente, no Brasil, segundo Fenabrave (2020), resultando no *boxplot* do Gráfico 4.14 (b).

É importante ressaltar que existem veículos da mesma marca e modelo com potências diferentes, logo, optou-se por agrupá-los pela média da potência. O resultado pode ser visto no Gráfico 4.14 (c). Em seguida, realizou-se a preparação dos dados para concatenação com o banco de dados de acidentes.

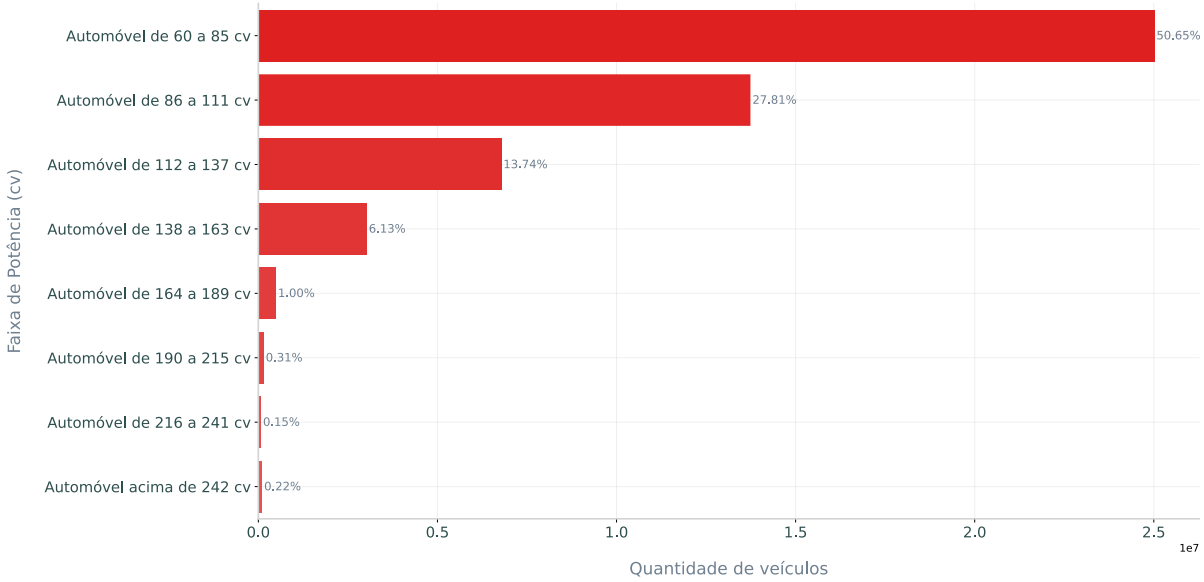
Gráfico 4.14 – (a) *Boxplot* das potências. (b) *Boxplot* das potências entre 60 a 300 cv. (c) *Boxplot* das potências médias dos veículos



Fonte: Elaborado pelo autor

Quanto à potência do motor, tem-se que a maioria dos veículos registrados no Renavam está entre automóveis de 60 a 85 cv, conforme Gráfico 4.15.

Gráfico 4.15 – Quantidade de veículos por montadora

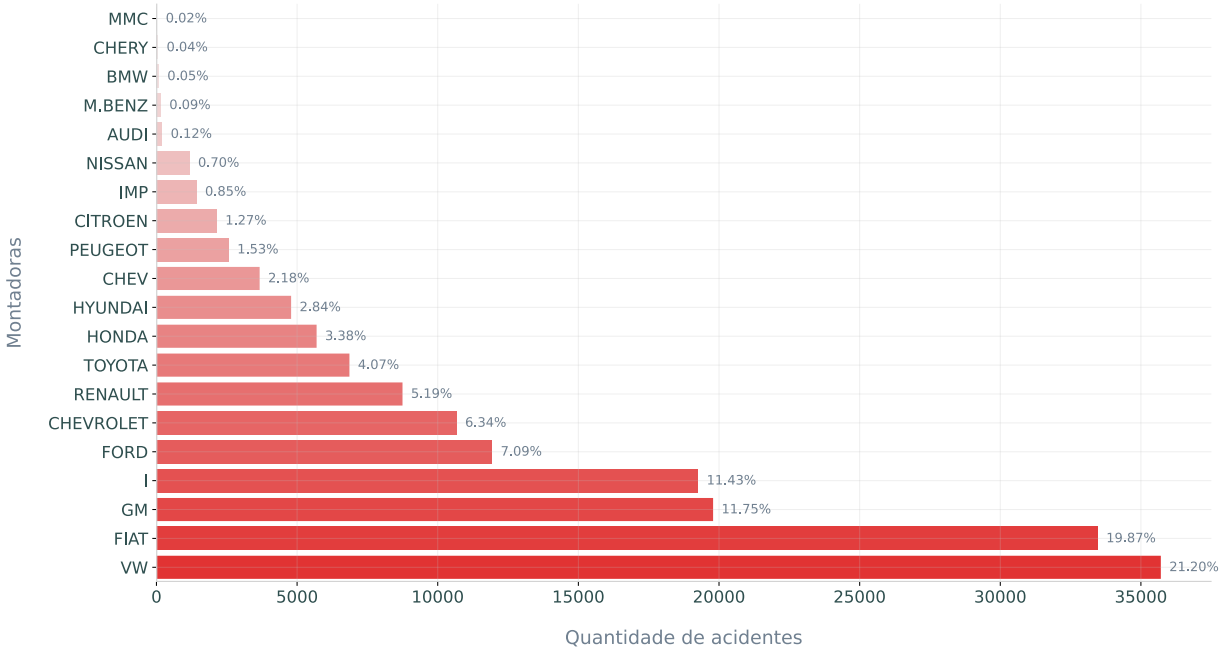


Fonte: Elaborado pelo autor

Com os bancos de dados preparados para união, concatenou-se os dois bancos de dados pela marca e ano do veículo, gerando um novo banco de dados com 171.490. Como poderia haver veículos do banco de dados de acidentes sem a referida potência no banco de dados das características dos veículos, optou-se por remover os valores ausentes, estabelecendo um novo banco de dados.

Após concatenado, optou-se, ainda, por remover o modelo do veículo, mantendo somente a marca, ou seja, a montadora, reduzindo o número de valores únicos e otimizando a variável para aplicação do modelo. Além disso, optou-se por manter somente as 20 marcas que mais se envolveram em acidentes, isto é, marcas que se envolveram em mais de 30 acidentes. O Gráfico 4.16 apresenta as 20 principais marcas que mais se envolveram em acidentes do presente banco de dados desta pesquisa.

Gráfico 4.16 – Quantidade de acidentes por montadora de automóvel



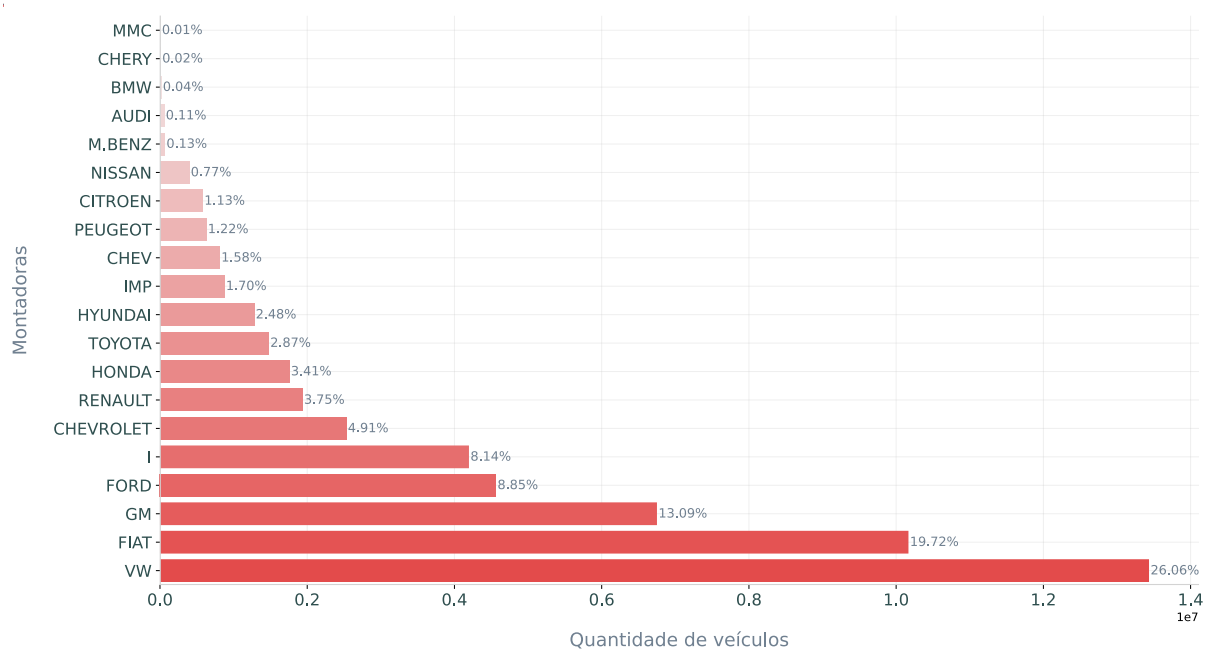
Fonte: Elaborado pelo autor

Consequentemente, as marcas nas quais mais se envolvem em acidentes são as marcas com maior quantidade de veículos registrados, conforme pode-se verificar no Gráfico 4.17.

Vale destacar que a maioria dos veículos registrados no Renavam são das montadoras Volkswagen, Fiat, GM e Ford, respectivamente, sendo VW e Fiat muito superior aos demais, o que se pode questionar, ainda, é o fato de que, se tem um maior número de veículos dessa montadora em circulação, maior é a probabilidade de ocorrer acidentes. Esse assunto será analisado posteriormente, verificando se existe alguma relação entre a causa do acidente e a montadora, principalmente causas de acidentes relacionadas a defeitos mecânicos.

Por fim, o banco de dados tratado e explorado finalizou com 126.545 observações e 12 variáveis, com as informações da Tabela 4.5.

Gráfico 4.17 – Quantidade de veículos por montadora



Fonte: Elaborado pelo autor

Tabela 4.5 – Tipos de variáveis

Nº	Variável	Tipo de variável
1	data_inversa	**quantitativa**
2	dia_semana	qualitativa
3	horario	**quantitativa**
4	causa_acidente	qualitativa
5	fase_dia	qualitativa
6	condição_meteorologica	qualitativa
7	tipo_pista	qualitativa
8	tracado_via	qualitativa
9	marca	qualitativa
		Continua...
10	idade	**quantitativa**
11	sexo	qualitativa
12	idade_veiculo	**quantitativa**
13	potencia	**quantitativa**

Fonte: Elaborado pelo autor

Após compreender e tratar cada variável dos bancos de dados de acidentes e das características dos veículos, como determinado nos objetivos específicos desta pesquisa, observa-se a existência de variáveis qualitativas e quantitativas no banco de dados após o tratamento. Conforme metodologia apresentada, a próxima etapa da pesquisa é converter as variáveis quantitativas em qualitativas, aplicando o método misto com estratégia transformadora concomitante, onde a transformação desses dados ocorre durante a fase de análise, seguindo

preceitos de Creswell *et. al.* (2007). Criam-se faixas de grupos das variáveis quantitativas, isto é, categorias para aplicação dos algoritmos de *machine learning.*

4.3 Transformação das variáveis em qualitativa

Nessa etapa da metodologia, criam-se categorias nas variáveis quantitativas, transformando-as em variáveis qualitativas. Os tipos de variáveis foram demonstrados na Tabela 4.4. São variáveis quantitativas: a data do acidente na qual são retirados os feriados, o horário no qual serão definidos os horários de pico e as variáveis idade do condutor, idade do veículo e potência do motor, onde serão criadas as faixas de grupos, categorizando-as.

Para determinar se a data em que o acidente ocorreu era feriado, utilizou-se da biblioteca *holidays* da linguagem Python. Optou-se por aderir às datas anteriores e posteriores ao feriado, ou seja, foram considerados como feriado a véspera do feriado, a data do feriado e a data posterior ao feriado.

Verificou-se, data por data, a existência daquelas datas na lista de feriados nacionais brasileiros da biblioteca, gerando uma nova variável denominada de feriado, com valores binários, sendo feriados as datas véspera, dia e pós feriados brasileiros, e não feriados os restantes das datas. A quantidade de acidentes em datas consideradas feriados e a quantidade de dias normais foram de 14.021 e 112.524, respectivamente.

Em seguida, criou-se uma nova variável, definida como horário de pico, sendo considerados como horário de pico, no Brasil, os horários entre 7 e 9 horas na parte da manhã, e na parte da tarde, entre 17 e 19 horas, segundo Resende & Souza (2009). Desta forma, horários de acidentes entre 7 e 9 horas da manhã e 17 e 19 horas da tarde foram considerados horário de pico, onde obteve-se 36.456 horários de pico e 900.089 horários normais.

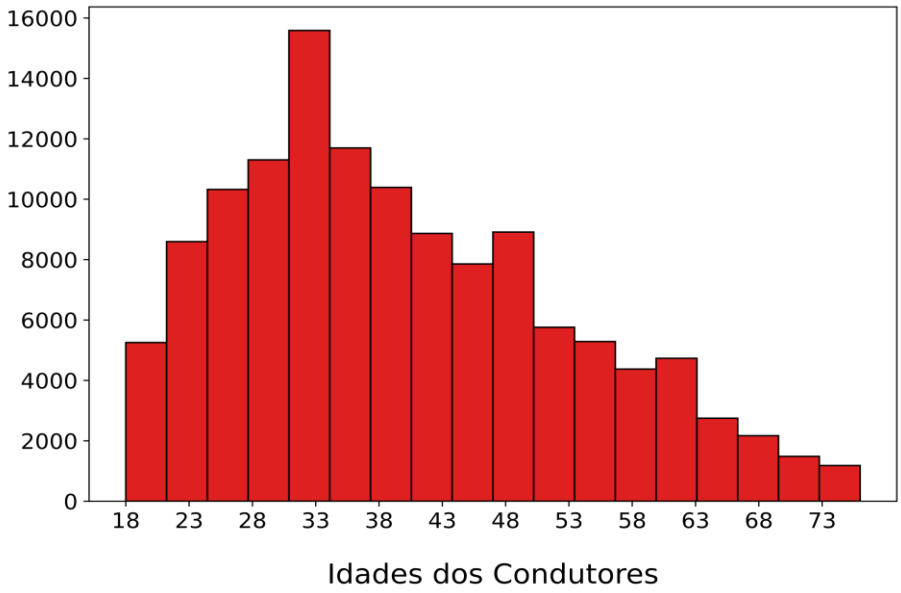
Na categorização das variáveis idade do condutor, idade do veículo e potência do motor, realizou-se uma análise estatística com histograma de cada variável. Para a variável idade do condutor, tem-se a estatística demonstrada na Tabela 4.6 e histograma do Gráfico 4.18.

Tabela 4.6 – Estatística das idades dos condutores

Estatística	Valor
Quantidade	126545
Média	39,858556
Desvio Padrão	13,112700
Mínimo	18
1º Quartil (25%)	30
2º Quartil (50%)	38
3º Quartil (75%)	49
Máximo	76

Fonte: Elaborado pelo autor

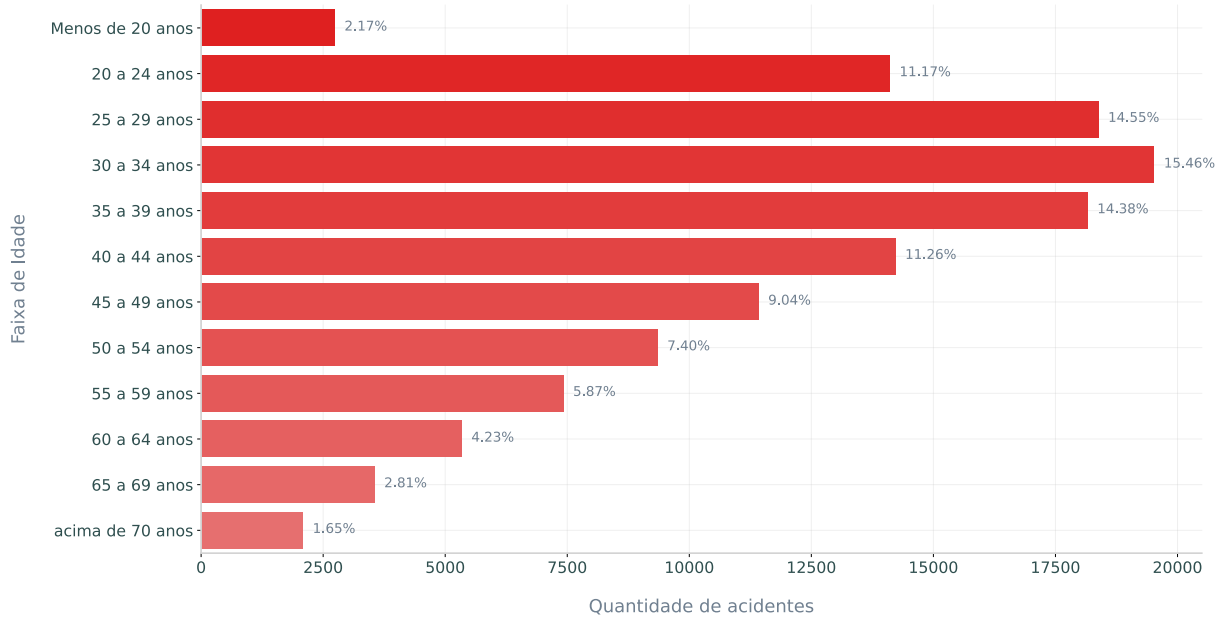
Gráfico 4.18 – Histograma da Idades dos Condutores



Fonte: Elaborado pelo autor

Através do histograma, percebe-se uma distribuição assimétrica e uma alta densidade dos dados em torno de condutores de 33 anos, assemelhando-se à pirâmide etária da população brasileira (IBGE, 2019). Por uma variável significativa para este estudo, categorizou-se utilizando os mesmos critérios de faixa etária da pirâmide etária população brasileira do IBGE. obtendo-se as faixas etárias exibidas no Gráfico 4.19.

Gráfico 4.19 – Faixa etária das idades dos condutores



Fonte: Elaborado pelo autor

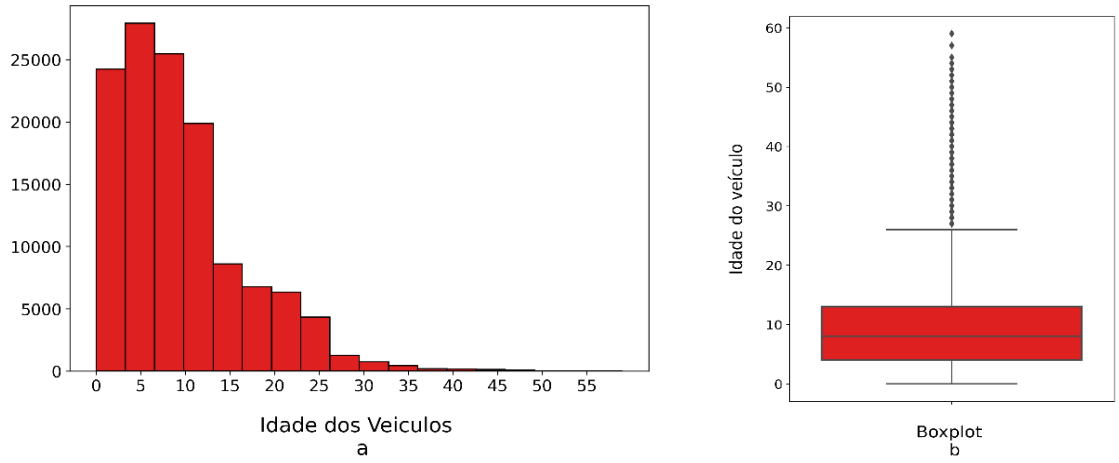
Nota-se, ainda, que a maioria dos condutores acidentados estão entre 20 a 39 anos, e levanta-se a hipótese de existência de alguma relação entre a idade do condutor e os acidentes, a qual será retomada e respondida a seguir, pois, antes disso, é necessário categorizar a idade dos veículos. Categorizando-se as variáveis idade do veículo, tem-se a estatística (Tabela 4.7), histograma e *boxplot* (Gráfico 4.20).

Tabela 4.7 – Estatística das idades dos automóveis

Estatística	**Valor**
Quantidade	126545
Média	9,4368
Desvio Padrão	7,0989
Mínimo	0
1º Quartil (25%)	4
2º Quartil (50%)	8
3º Quartil (75%)	13
Máximo	59

Fonte: Elaborado pelo autor

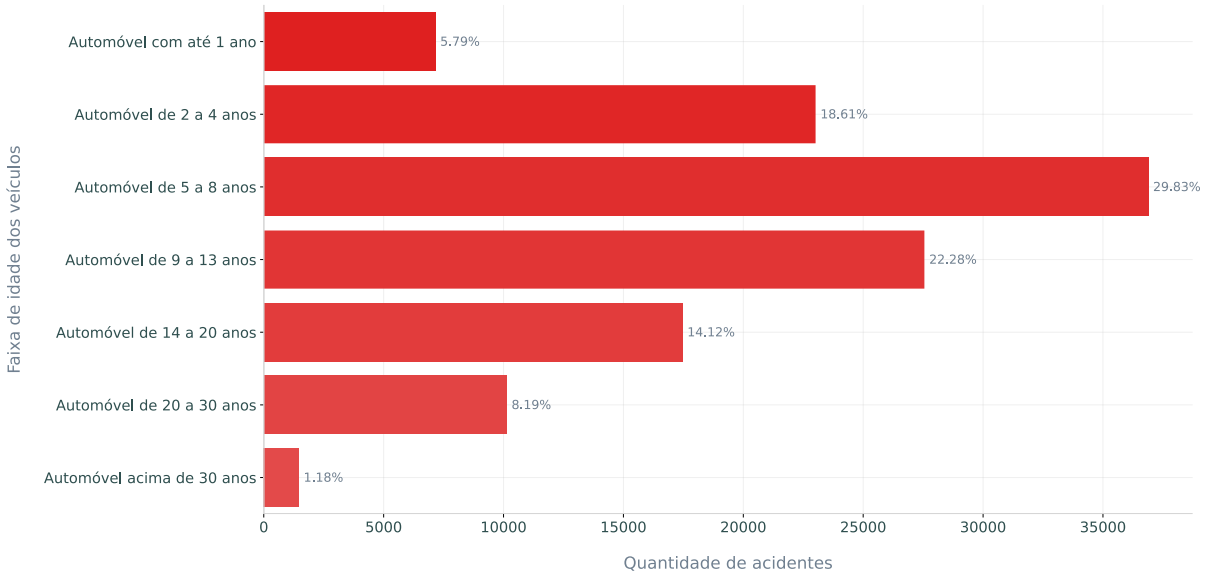
Gráfico 4.20 – (a) Histograma da idade dos veículos. (b) *Boxplot* da idade dos veículos



Fonte: Elaborado pelo autor

Para criar categorias utilizou-se da tabela estatística descritiva das idades do veículo. Como critério, foram considerados carros novos aqueles com até 1 ano, a partir disso, de 2 até o 1º Quartil (4 anos), do 1º ao 2º Quartil e do 2º ao 3 º Quartil. Após isso, considerou-se a soma da média em cada faixa até os 30 anos. Acima de 30 anos, têm-se veículos mantidos como objeto de coleção, conforme resolução 56 do Contran, de 21 maio de 1998 (CONTRAN, 1998). Portanto, têm-se as seguintes categorias apresentadas no Gráfico 4.21.

Gráfico 4.21 – Categoria das idades dos veículos



Fonte: Elaborado pelo autor

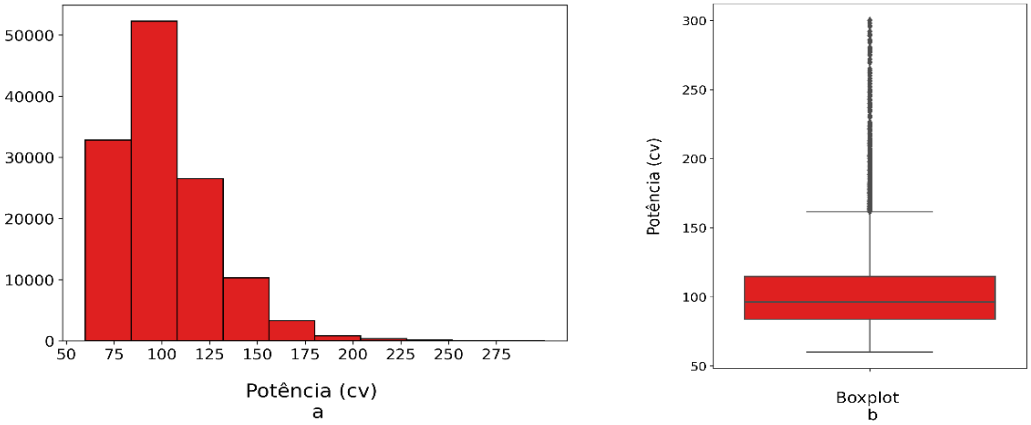
Já a variável potência do motor, possui as seguintes estatísticas descritivas na Tabela 4.8, e histograma e *boxplot* apresentadas no Gráfico 4.22.

Tabela 4.8 – Estatística das idades dos automóveis

Estatística	Valor
Quantidade	126545
Média	102,4873
Desvio Padrão	24,6535
Mínimo	60,0000
1º Quartil (25%)	83,6666
2º Quartil (50%)	96,2045
3º Quartil (75%)	114,8000
Máximo	300,0000

Fonte: Elaborado pelo autor

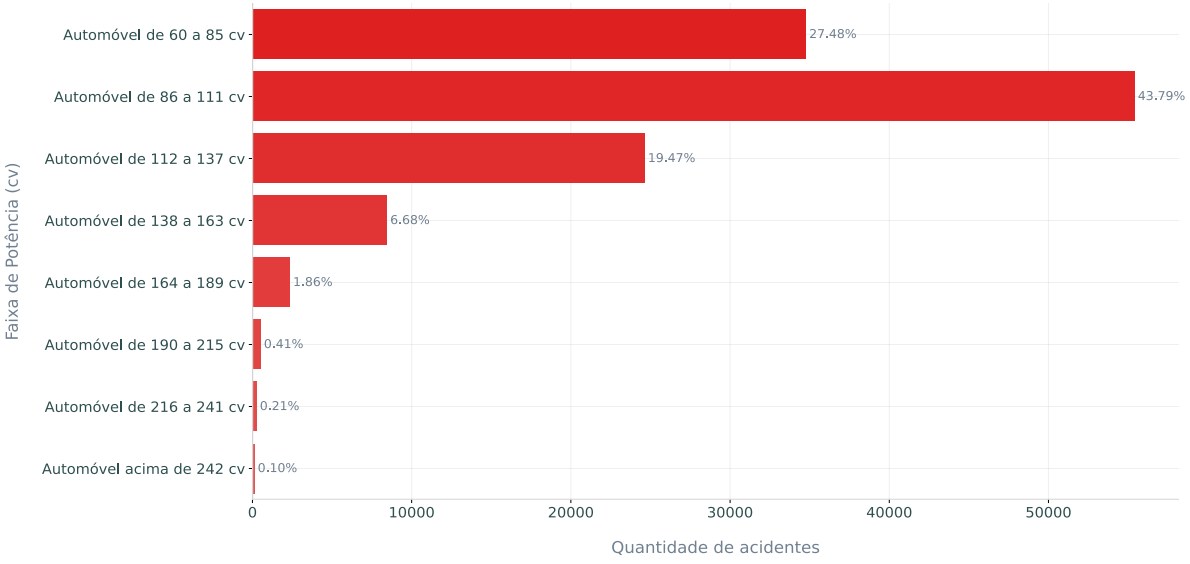
Gráfico 4.22 – (a) Histograma da potência do motor. (b) *Boxplot* da potência do motor



Fonte: Elaborado pelo autor

O critério para criação das categorias de faixas para a potência do motor foi o desvio padrão, isto é, considerado a cada 25 cv conforme exibido no Gráfico 4.23.

Gráfico 4.23 – Categoria das faixas de potência



Fonte: Elaborado pelo autor

Conforme descrito na metodologia desse trabalho, as variáveis quantitativas foram categorizadas e transformadas em qualitativas. As variáveis, após tratamento, apresentam as propriedades que constam na Tabela 4.9 e estão preparadas para criação dos modelos de regras de associação.

Tabela 4.9 – Variáveis selecionadas e categorizadas para modelagem.

Nº	Variável	Fator	Qtd. Valores Únicos	Tipo de variável
1	dia_semana	Meio Ambiente	7	qualitativa
2	Feriado	Meio Ambiente	2	qualitativa
3	hora_pico	Meio Ambiente	2	qualitativa
4	fase_dia	Meio Ambiente	4	qualitativa
5	condicao_metereologica	Meio Ambiente	8	qualitativa
6	tipo_pista	Via	3	qualitativa
7	tracado_via	Via	9	qualitativa
8	faixa_etaria	Usuário	12	qualitativa
9	sexo	Usuário	2	qualitativa
10	marca	Veículo	20	qualitativa
11	faixa_idade_veiculo	Veículo	7	qualitativa
12	faixa_pot	Veículo	8	qualitativa
13	causa_acidente	Todos	24	qualitativa
Soma valores únicos			108	

Fonte: Elaborado pelo autor

Antes de criar o modelo, analisou-se a relação entre as variáveis para verificar a independência dos fatores com a causa do acidente, isso é, para verificar se a frequência das variáveis dos fatores via, meio ambiente, usuário e veículo são as mesmas para todas as causas dos acidentes. Para isso, utilizou-se uma análise de correspondência multivariadas em uma tabela de contingência.

4.3.1 Análise de Correspondência Multivariada

Buscar um relacionamento entre as variáveis é um dos objetivos específicos desta pesquisa, para tanto, realizou-se uma análise de correspondência multivariada para compreender se as características dos acidentes e veículos, bem como as causas dos acidentes podem ser consideradas independentes. Criou-se uma tabela de contingência com as contagens das características por causa de acidente. Para criação dessa tabela, foi realizada a contagem de quantos acidentes da característica por tipo de causa de acidente, isto é, quantos acidentes com aquela característica são oriundos daquela causa de acidente, para isso utilizou-se da função

crosstab da biblioteca pandas da linguagem Python. Essa tabela de contingência se encontra no (APÊNDICE A) dessa dissertação. Com isso, obtém-se uma tabela com 84 colunas, isto é, o total de características possíveis de todas as variáveis dos fatores via, meio ambiente, usuário e veículos. Sendo uma das colunas referente aos acidentes, essa tabela apresenta 24 linhas que estão descritas na Tabela 4.3 do tópico de análise exploratória. Os 84 itens (colunas) e as 24 causas (linhas) estão listadas a seguir:

Linhas: agressão externa, animais na pista, avarias e/ou desgaste excessivo no pneu, carga excessiva e/ou mal acondicionada, condutor dormindo, defeito mecânico no veículo, defeito na via, deficiência ou não acionamento do sistema de iluminação/sinalização do veículo, desobediência às normas de trânsito pelo condutor, desobediência às normas de trânsito pelo pedestre, falta de atenção do pedestre, falta de atenção à condução, fenômenos da natureza, ingestão de substâncias psicoativas, ingestão de álcool, ingestão de álcool e/ou substâncias psicoativas pelo pedestre, mal súbito, não guardar distância de segurança, objeto estático sobre o leito carroçável, pista escorregadia, restrição de visibilidade, sinalização da via insuficiente ou inadequada, ultrapassagem indevida, velocidade incompatível.

Colunas: domingo, quarta-feira, quinta-feira, segunda-feira, sexta-feira, sábado, terça-feira, dia normal, feriado, horário normal, horário de pico, amanhecer, anoitecer, plena noite, pleno dia, chuva, céu claro, garoa/chuvisco, granizo, nevoeiro/neblina, nublado, sol, vento, dupla, múltipla, simples, curva, desvio temporário, interseção de vias, ponte, reta, retorno regulamentado, rotatória, túnel, viaduto, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 45 a 49 anos, 50 a 54 anos, 55 a 59 anos, 60 a 64 anos, 65 a 69 anos, menos de 20 anos, acima de 70 anos, feminino, masculino, AUDI, BMW, CHERY, CHEV, CHEVROLET, CITROEN, FIAT, FORD, GM, HONDA, HYUNDAI, I, IMP, M.BENZ, MMC, NISSAN, PEUGEOT, RENAULT, TOYOTA, VW, automóvel acima de 30 anos, automóvel com até 1 ano, automóvel de 14 a 20 anos, automóvel de 2 a 4 anos, automóvel de 20 a 30 anos, automóvel de 5 a 8 anos, automóvel de 9 a 13 anos, automóvel acima de 242 cv, automóvel de 112 a 137 cv, automóvel de 138 a 163 cv, automóvel de 164 a 189 cv, automóvel de 190 a 215 cv, automóvel de 216 a 241 cv, automóvel de 60 a 85 cv, automóvel de 86 a 111 cv.

Logo, novamente tem-se variáveis quantitativas (quantidade de características) para identificar a independência das variáveis criadas, realiza-se uma análise de correspondência multivariada através do teste qui-quadrado com as seguintes hipóteses: Hipótese nula (H_0) = Independência das variáveis, isto é, independente da causa as características aparecem com a mesma

frequência; e H_1 = A frequência das características dos acidentes não é a mesma para todas as causas.

Com a aplicação do teste qui-quadrado na tabela de contingência, obteve-se os seguintes resultados (Tabela 4.10).

Tabela 4.10 – Estatística do Teste qui-quadrado

Descrição	**Valor**
Qui-quadrado	86246,74
p-valor	0,0

Fonte: Elaborado pelo autor

Por meio da análise dos dados, nota-se que o p-valor é zero, logo, existem evidências para se rejeitar a hipótese nula (H_0), que é a hipótese de independência. Isso significa que, dependendo da causa do acidente, a frequência das características é diferente. Sabe-se que a frequência das características não é a mesma para todas as causas de acidentes e, então, cumpre-se o objetivo específico (b) desse trabalho.

4.4 Aplicação dos algoritmos e resultados

Sabendo que a quantidade de características dos acidentes é dependente da causa do acidente, aplica-se os algoritmos de aprendizado de máquina na tabela de variáveis categóricas, descritas na Tabela 4.8. Com a tabela de variáveis quantitativas (categóricas) transforma-se a matriz de variáveis qualitativas em uma matriz binária, ou seja, cada elemento da matriz indica a presença daquele valor entre todos os valores possíveis de todos os atributos está presente ou não naquele acidente, obtendo uma matriz de 123.413 linhas e 108 colunas, sendo 123.413 todos os acidentes do banco de dados limpo e tratado, e 108 colunas são as características dos acidentes e as suas causas, conforme soma dos valores únicos da Tabela 4.8. Tem-se um banco de dados de variáveis categóricas transformando este conjunto de dados em um formato de matriz adequado para aprendizado de máquina de regras de associação *Apriori*, *Eclat*, *FP-Growth* e *FP-Max.*

4.4.1 Apriori

O *Apriori* é o algoritmo de regras de associação mais utilizado na área, conforme já mencionado. Por isso, este foi escolhido como o primeiro algoritmo a ser aplicado nesse

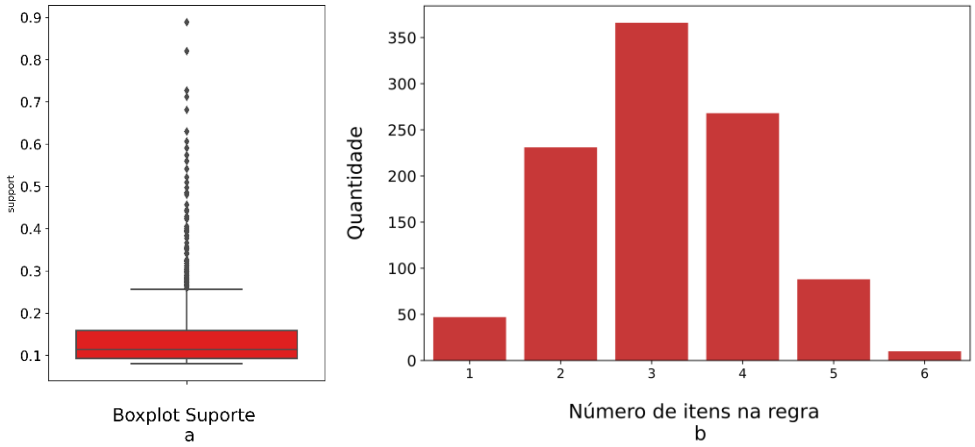
trabalho. Através da biblioteca *mlxtend* do Python, utilizou-se o valor mínimo de suporte de 0,08 por proporcionar maior quantidade de regras para comparar os algoritmos, além de ser o valor de suporte que apresentou melhores resultados no *FP-Max*, item 4.4.3 desse capítulo.
O algoritmo *Apriori* apresentou 1010 itens com a seguinte estatística de suporte, conforme demonstrado na Tabela 4.11 e no Gráfico 4.24.

Tabela 4.11 – Estatística das Regras

Estatística	Suporte	Tamanho dos itens
Quantidade	1010	1010
Média	0,1447	3,1475
Desvio Padrão	0,0908	1,0474
Mínimo	0,0800	1
1º Quartil (25%)	0,0925	2
2º Quartil (50%)	0,1141	3
3º Quartil (75%)	0,1585	4
Máximo	0,8886	6

Fonte: Elaborado pelo autor

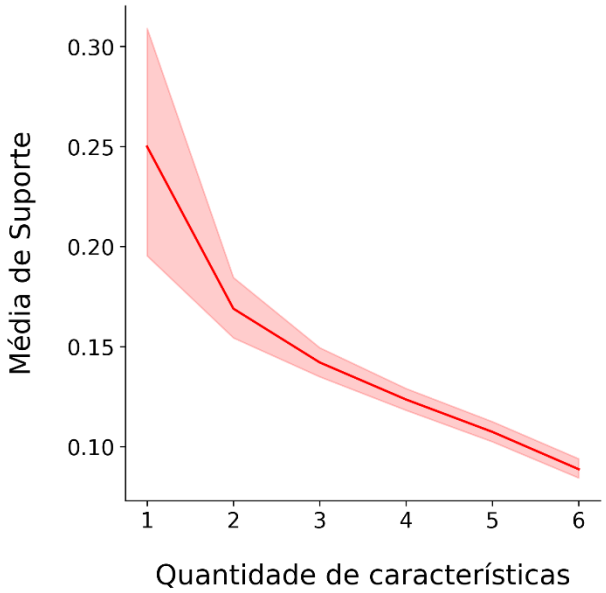
Gráfico 4.24 – (a) *Boxplot* do suporte (b) Histograma da quantidade de características



Fonte: Elaborado pelo autor

Nota-se, então, que no *Apriori* há um grande volume de itens entre os suportes 0,1 e 0,15 e média de suportes de 0,1447. A média do tamanho dos itens é de 3,1475 características, sendo o maior item com 6 características. Outro ponto importante comparado à quantidade de características por regra é que, quanto maior a quantidade de itens, menor é o suporte, conforme Gráfico 4.25, que apresenta a média e desvio padrão dos suportes por quantidade de características.

Gráfico 4.25 – Média e desvio padrão dos suportes por quantidade de características



Fonte: Elaborado pelo autor

Criando a tabela IF-THEN, nesse primeiro cenário, através da métrica de confiança, obtém-se 759 regras criadas. Um resumo das 10 regras ordenado pelos valores de *lift* estão demonstrados a seguir na Tabela 4.12.

Tabela 4.12 – Tabela IF-THEN no 1º cenário do *Apriori*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Masculino', 'Plena Noite', 'Dupla'}	{'Horário Normal'}	0,1138	0,7116	0,0944	0,8294	1,1654
{'Masculino', 'Plena Noite', 'Dupla', 'Dia Normal'}	**{'Horário Normal'}**	**0,1016**	**0,7116**	**0,0839**	**0,8257**	**1,1603**
{'Plena Noite', 'Dupla'}	{'Horário Normal'}	0,1346	0,7116	0,1106	0,8221	1,1552
{'Masculino', 'Plena Noite'}	{'Horário Normal'}	0,2802	0,7116	0,2295	0,8191	1,1510
{'Plena Noite', 'Dupla', 'Dia Normal'}	{'Horário Normal'}	0,1203	0,7116	0,0985	0,8187	1,1504
{'Masculino', 'Plena Noite', 'Automóvel de 86 a 111 cv'}	{'Horário Normal'}	0,1275	0,7116	0,1044	0,8184	1,1499
{'Masculino', 'Plena Noite', 'Dia Normal'}	{'Horário Normal'}	0,2493	0,7116	0,2032	0,8151	1,1453

Continua...

{'Masculino', 'Plena Noite', 'Automóvel de 86 a 111 cv', 'Dia Normal'}	**{'Horário Normal'}**	**0,1137**	**0,7116**	**0,0926**	**0,8144**	**1,1443**
{'Plena Noite', 'Automóvel de 86 a 111 cv'}	{'Horário Normal'}	0,1464	0,7116	0,1190	0,8128	1,1421
{'Plena Noite'}	{'Horário Normal'}	0,3261	0,7116	0,2651	0,8127	1,1420

Fonte: Elaborado pelo autor

Nota-se que, como consequências, teve somente "Horário Normal", isto ocorreu, provavelmente, devido à enorme diferença entre a quantidade de acidentes em horários normais e em horários de pico. Em geral, das 759 regras, *THEN* teve 440 'Dia Normal', 298 'Masculino' e 21 'Horário Normal', porém 'Horário Normal' apresentam maiores índices de *lift.*

Uma regra pertinente com alto índice de *lift,* é a associação entre as características sexo 'Masculino', fase do dia 'Plena Noite', tipo de pista 'Dupla', feriado ou dia normal 'Dia Normal', e hora de pico 'Horário Normal'. Uma outra regra pertinente é sexo 'Masculino', fase do dia 'Plena Noite', feriado 'Dia Normal', potência do motor 'Automóvel de 86 a 111 cv' e horário de pico 'Horário Normal'. Portanto, uma pessoa do sexo masculino dirigindo um automóvel de 86 a 111 cv, à noite, após o horário de pico tem forte associação entre os dados de acidentes em rodovias federais brasileiras.

Infelizmente, essa base de dados não apresentou nenhuma causa de acidente como consequência. Com intuito de equilibrar a quantidade de características, aplicou-se o algoritmo em mais dois cenários, um sem 'Dia Normal' e 'Horário Normal', mantendo 'Feriado' e 'Horário de Pico' e outro cenário sem as mesmas características e sem o sexo 'Masculino', obtendo-se as seguintes tabelas IF-THEN (Tabela 4.13).

- 2º Cenário (Sem 'Dia Normal' e 'Horário Normal') com mínimo de confiança de 0,08.

As 10 principais regras com maior valor de *lift* no segundo cenário.

Tabela 4.13 – Tabela IF-THEN no 2º cenário do *Apriori*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Ingestão de Álcool'}	{'Masculino'}	0,1001	0,8200	0,0907	0,906	1,1049
{'Plena Noite', 'Simples', 'Reta'}	{'Masculino'}	0,1062	0,8200	0,0937	0,8819	1,0755
{'Céu Claro', 'Plena Noite', 'Simples'}	{'Masculino'}	0,0937	0,8200	0,0824	0,8795	1,0726

Continua...

{‘Plena Noite’, ‘Simples’}	{‘Masculino’}	0,1599	0,8200	0,1401	0,8763	1,0686
{‘Plena Noite’, ‘Automóvel de 86 a 111 cv’, ‘Reta’}	{‘Masculino’}	0,1023	0,8200	0,0893	0,8737	1,0655
{‘VW’, ‘Automóvel de 86 a 111 cv’}	**{‘Masculino’}**	**0,126**	**0,8200**	**0,1109**	**0,8737**	**1,0655**
{‘VW’, ‘Simples’}	{‘Masculino’}	0,1149	0,8200	0,1003	0,8737	1,0646
{‘Plena Noite’, ‘Automóvel de 86 a 111 cv’}	{‘Masculino’}	0,1464	0,8200	0,1275	0,8710	1,0622
{‘Automóvel de 14 a 20 anos’, ‘Reta’}	**{‘Masculino’}**	**0,0963**	**0,8200**	**0,0835**	**0,8673**	**1,0576**
{‘VW’, ‘Reta’}	{‘Masculino’}	0,1435	0,8200	0,1244	0,8665	1,056

Fonte: Elaborado pelo autor

Nesse 2º cenário, percebe-se que apresentam novas características interessantes, tais como a ingestão de álcool envolvida ao sexo masculino, automóveis da marca VM, mesmo tendo valores próximos ao da marca FIAT em envolvimentos de acidentes, conforme Gráfico 4.14. Ainda sobre veículos, outra característica presente nesse cenário é o envolvimento de acidentes em reta com 'Automóvel de 14 a 20 anos' de idade e condutores do sexo masculino. Outro fator importante presente é a condição meteorológica 'Céu Claro'.

Esse cenário apresentou 103 THEN, sendo todas do sexo ‘Masculino’. Sendo assim, aplica-se o algoritmo em um 3º cenário.

- 3º Cenário (Sem ‘Dia Normal’, ‘Horário Normal’ e ‘Masculino’) com mínimo suporte de 0,05.

Neste cenário, foi necessário reduzir o valor mínimo de suporte para 0,05, sendo possível obter mais regras. Somente 4 regras foram pertinentes, como demonstra a Tabela 4.14 a seguir.

Tabela 4.14 – Tabela IF-THEN no 3º cenário do *Apriori*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Sol'}	{'Pleno dia'}	0,0812	0,5739	0,0796	0,9808	1,7089
{'Reta', 'Sol'}	**{'Pleno dia'}**	**0,0582**	**0,5739**	**0,0569**	**0,9780**	**1,7040**
{'Não guardar distância de segurança'}	**{'Reta'}**	**0,0929**	**0,6811**	**0,0788**	**0,8485**	**1,2457**
{'Pleno dia', 'Não guardar distância de segurança'}	**{'Reta'}**	**0,0657**	**0,6811**	**0,0555**	**0,8456**	**1,2415**

Fonte: Elaborado pelo autor

Já o 3º cenário, apresentou apenas 4 regras pertinentes. Nota-se, então, novas características com fortes índices de confiança, que é a associação entre a condição meteorológica ‘Sol’ e a fase do dia ‘Pleno dia’, o que corrobora com o esperado, uma vez o sol está mais presente durante o dia e não no amanhecer ou anoitecer e, muito menos, em plena noite. Fora isso, outras três regras são interessantes, pois apresentam associação entre acidentes por não guardar distância de segurança com pistas retas em pleno dia. Lembrando que, como excluiu-se as características ‘Dia Normal’, ‘Horário Normal’ e ‘Masculino’ e nesse cenário há à ausência das antônimas dessas características, logo, elas estão inclusas. Uma regra pertinente contendo essas características excluídas é {‘Dia Normal’, ‘Horário Normal’ e ‘Masculino’} + {'Pleno dia', 'Não guardar distância de segurança', 'Reta'}, isto é, uma pessoa do sexo masculino, dirigindo em um dia normal, que não é feriado, fora do horário de pico, durante a luz dia em uma pista reta, associa-se com acidentes por não guardar distância de segurança. Sendo assim, aplica-se os mesmos cenários com o algoritmo *FP-Growth*.

4.4.2 FP-Growth

Aplicando-se o algoritmo *FP-Growth* da mesma biblioteca *mlxtend* e com o mesmo valor mínimo de suporte de 0,08, obteve-se exatamente os mesmos resultados do algoritmo *Apriori* em todos os cenários, com 1010 itens nas mesmas estatísticas, para suporte e tamanho de itens, com média entre 0,1447 para suporte e 3,1475 para tamanho de itens. Assim como as mesmas tabelas IF-THEN em todos os três cenários, com regras no primeiro cenário com THEN em 440 ‘Dia Normal’, 298 ‘Masculino’ e 21 ‘Horário Normal’, no segundo e terceiro cenários apresentando as mesmas regras do algoritmo *Apriori*.
Entende-se que o algoritmo *FP-Growth* da biblioteca *mlxtend* é totalmente similar ao algoritmo *Apriori* da mesma biblioteca. Portanto, foram aplicados os algoritmos *FP-Max* e *Eclat* para fins de comparação.

4.4.3 FP-Max

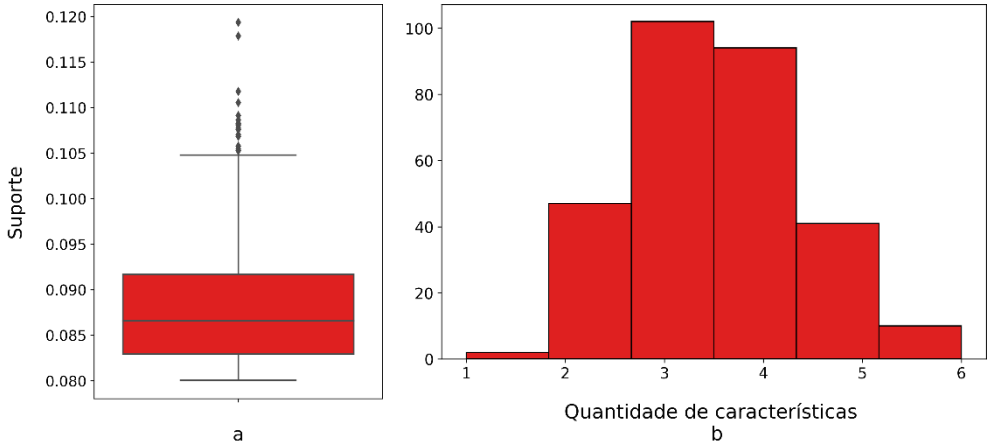
Nessa etapa, utilizou-se o mesmo valor de suporte mínimo de 0,08, onde o algoritmo *FP-Max* apresentou 296 itens, com média de 0,0886 de suporte e 3,5236 de quantidade de características. As estatísticas dos algoritmos estão apresentadas no Gráfico 4.26 e na Tabela 4.15.

Tabela 4.15 – Estatística das *FP-Max*

Estatística	Suporte	Tamanho dos itens
Quantidade	296	296
Média	0,0886	3,5236
Desvio Padrão	0,0075	1,0444
Mínimo	0,0800	1
1º Quartil (25%)	0,0829	3
2º Quartil (50%)	0,0865	3
3º Quartil (75%)	0,0916	4
Máximo	0,1193	6

Fonte: Elaborado pelo autor

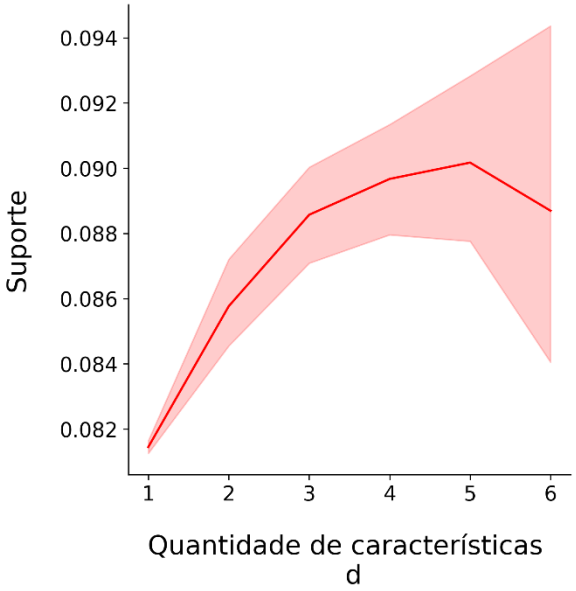
Gráfico 4.26 – (a) *Boxplot* do suporte (b) Histograma da quantidade de características



Fonte: Elaborado pelo autor

Por meio da análise dos resultados, percebe-se que os índices de suporte variam entre 0,08 e 0,11, com média de 0,08 e desvio padrão muito baixo. A média de quantidade de características por item é de 3,5236, porém, diferentemente dos algoritmos *Apriori* e *FP-Growth*, a média de suporte aumenta quando a quantidade de características é maior, como é possível verificar no gráfico da média e desvio padrão dos suportes por quantidade de características (Gráfico 4.27).

Gráfico 4.27 – Média e desvio padrão dos suportes por quantidade de características



Fonte: Elaborado pelo autor

A tabela IF-THEN foi criada pelo índice de suporte, pois não apresentou índices de *lift* e confiança, isto é, obteve valores nulos para *lift* e confiança. A tabela resultou em 394 regras criadas, onde é possível verificar um resumo das 10 regras ordenado pelos valores de suporte na Tabela 4.16.

Tabela 4.16 – Tabela IF-THEN no 1º cenário do *FP-Max*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Horário Normal', 'Masculino', 'Automóvel de 86 a 111 cv', 'Simples'}	**{'Dia Normal'}**	**-**	**-**	**0,1193**	**-**	**-**
{'Horário Normal', 'Automóvel de 86 a 111 cv', 'Simples', 'Dia Normal'}	{'Masculino'}	-	-	0,1193	-	-
{'Horário Normal', 'Dia Normal'}	{'Masculino', 'Automóvel de 86 a 111 cv', 'Simples'}	-	-	0,1193	-	-
{'Horário Normal', 'Automóvel de 86 a 111 cv'}	{'Masculino', 'Simples', 'Dia Normal'}	-	-	0,1193	-	-
{'Masculino', 'Simples', 'Dia Normal'}	{'Horário Normal', 'Automóvel de 86 a 111 cv'}	-	-	0,1193	-	-

Continua...

{'Masculino', 'Automóvel de 86 a 111 cv', 'Simples'}	{'Horário Normal', 'Dia Normal'}	-	-	0,1193	-	-
{'Automóvel de 86 a 111 cv', 'Simples', 'Dia Normal'}	{'Horário Normal', 'Masculino'}	-	-	0,1193	-	-
{'Horário Normal', 'Masculino', 'Automóvel de 86 a 111 cv', 'Dia Normal'}	{'Simples'}	-	-	0,1193	-	-
{'Horário Normal', 'Masculino', 'Simples', 'Dia Normal'}	{'Automóvel de 86 a 111 cv'}	-	-	0,1193	-	-
{'Masculino', 'Automóvel de 86 a 111 cv', 'Dia Normal'}	{'Horário Normal', 'Simples'}	-	-	0,1193	-	-

Fonte: Elaborado pelo autor

Nota-se, então, como consequências, com mais de uma característica, diferentemente do *Apriori* e *FP-Growth*. O algoritmo *FP-Max* apresentou como regras significativas 'Horário Normal', 'Masculino', 'Automóvel de 86 a 111 cv', 'Simples', 'Dia Normal', bem como o *Apriori* e *FP-Growth*. Confirmando que tais características são associações significantes desse banco de dados. Além disso, também apresentou Dia Normal e Horário normal em todas as principais associações. Então, aplicou-se o algoritmo nos dois outros cenários para comparação.

- 2º Cenário (Sem 'Dia Normal' e 'Horário Normal') com mínimo de confiança de 0,08.

As 10 principais regras com maiores valores de suporte no segundo cenário (Tabela 4.17).

Tabela 4.17 – Tabela IF-THEN no 2º cenário do *FP-Max*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Masculino'}	{'Reta', 'Céu Claro', 'Automóvel de 86 a 111 cv'}	-	-	0,1445	-	-
{'Reta', 'Céu Claro', 'Masculino'}	{'Automóvel de 86 a 111 cv'}	-	-	0,1445	-	-
{'Automóvel de 86 a 111 cv', 'Céu Claro', 'Masculino'}	**{'Reta'}**	-	-	**0,1445**	-	-
{'Reta', 'Céu Claro', 'Automóvel de 86 a 111 cv'}	{'Masculino'}	-	-	0,1445	-	-
{'Reta', 'Masculino'}	{'Céu Claro', 'Automóvel de 86 a 111 cv'}	-	-	0,1445	-	-

Continua...

{'Céu Claro', 'Masculino'}	{'Reta', 'Automóvel de 86 a 111 cv'}	-	-	0,1445	-	-
{'Automóvel de 86 a 111 cv', 'Masculino'}	{'Reta', 'Céu Claro'}	-	-	0,1445	-	-
{'Reta', 'Céu Claro'}	{'Automóvel de 86 a 111 cv', 'Masculino'}	-	-	0,1445	-	-
{'Céu Claro', 'Automóvel de 86 a 111 cv'}	{'Reta', 'Masculino'}	-	-	0,1445	-	-
{'Reta', 'Automóvel de 86 a 111 cv'}	{'Céu Claro', 'Masculino'}	-	-	0,1445	-	-

Fonte: Elaborado pelo autor

Nesse 2º cenário, com 298 regras, estão demonstradas regras interessantes, bem como *Apriori* e *FP-Growth*, que são 'Automóvel de 86 a 111 cv', 'Céu Claro', 'Masculino' 'Reta'. Entretanto, não apresentou a marca do automóvel e idade, diferente do *Apriori* e *FP-Growth*.

- 3º Cenário (Sem ‘Dia Normal’, ‘Horário Normal’ e ‘Masculino’) com mínimo suporte de 0,05.

Esse cenário apresentou 28 regras com a principal característica sendo “Falta de Atenção à Condução”, como demonstra a Tabela 4.18.

Tabela 4.18 – Tabela IF-THEN no 3º cenário do *FP-Max*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Reta', 'Céu Claro', 'Pleno dia'}	**{'Falta de Atenção à Condução'}**	**-**	**-**	**0,1069**	**-**	**-**
{'Pleno dia', 'Falta de Atenção à Condução'}	{'Reta', 'Céu Claro'}	-	-	0,1069	-	-
{'Reta', 'Pleno dia', 'Falta de Atenção à Condução'}	{'Céu Claro'}	-	-	0,1069	-	-
{'Falta de Atenção à Condução'}	{'Reta', 'Céu Claro', 'Pleno dia'}	-	-	0,1069	-	-
{'Céu Claro'}	{'Reta', 'Pleno dia', 'Falta de Atenção à Condução'}	-	-	0,1069	-	-
{'Pleno dia'}	{'Reta', 'Céu Claro', 'Falta de Atenção à Condução'}	-	-	0,1069	-	-
{'Céu Claro', 'Falta de Atenção à Condução'}	{'Reta', 'Pleno dia'}	-	-	0,1069	-	-

Continua...

{'Reta'}	{'Pleno dia', 'Céu Claro', 'Falta de Atenção à Condução'}	-	-	0,1069	-	-
{'Pleno dia', 'Céu Claro'}	{'Reta', 'Falta de Atenção à Condução'}	-	-	0,1069	-	-
{'Reta', 'Falta de Atenção à Condução'}	{'Pleno dia', 'Céu Claro'}	-	-	0,1069	-	-
{'Reta', 'Céu Claro'}	{'Pleno dia', 'Falta de Atenção à Condução'}	-	-	0,1069	-	-
{'Reta', 'Pleno dia'}	{'Céu Claro', 'Falta de Atenção à Condução'}	-	-	0,1069	-	-
{'Pleno dia', 'Céu Claro', 'Falta de Atenção à Condução'}	{'Reta'}	-	-	0,1069	-	-
{'Reta', 'Céu Claro', 'Falta de Atenção à Condução'}	{'Pleno dia'}	-	-	0,1069	-	-
{'Pleno dia', 'Simples'}	{'Reta', 'Céu Claro'}	-	-	0,1006	-	-
{'Pleno dia'}	{'Reta', 'Simples', 'Céu Claro'}	-	-	0,1006	-	-
{'Simples'}	{'Reta', 'Céu Claro', 'Pleno dia'}	-	-	0,1006	-	-
{'Reta'}	{'Pleno dia', 'Simples', 'Céu Claro'}	-	-	0,1006	-	-
{'Pleno dia', 'Céu Claro'}	{'Reta', 'Simples'}	-	-	0,1006	-	-
{'Simples', 'Céu Claro'}	{'Reta', 'Pleno dia'}	-	-	0,1006	-	-
{'Reta', 'Simples', 'Pleno dia'}	{'Céu Claro'}	-	-	0,1006	-	-
{'Reta', 'Céu Claro'}	{'Pleno dia', 'Simples'}	-	-	0,1006	-	-
{'Reta', 'Pleno dia'}	{'Simples', 'Céu Claro'}	-	-	0,1006	-	-
{'Reta', 'Simples'}	{'Pleno dia', 'Céu Claro'}	-	-	0,1006	-	-
{'Pleno dia', 'Simples', 'Céu Claro'}	{'Reta'}	-	-	0,1006	-	-
{'Reta', 'Céu Claro', 'Pleno dia'}	{'Simples'}	-	-	0,1006	-	-
{'Reta', 'Simples', 'Céu Claro'}	{'Pleno dia'}	-	-	0,1006	-	-

Continua...

{'Céu Claro'}	{'Reta', 'Simples', 'Pleno dia'}	-	-	0,1006	-	-

Fonte: Elaborado pelo autor

O algoritmo *FP-Max*, que foi pouco utilizado por autores na área, resultou em uma regra de associação interessante, onde determina a associação entre os fatores {‘Dia Normal’ ‘Horário Normal’ ‘Masculino’} + {'Reta', 'Céu Claro', 'Pleno dia'} → 'Falta de Atenção à Condução'}. Por fim, o próximo e último algoritmo aplicado foi o *Eclat*, através da biblioteca *pyeclat*.

4.4.4 Eclat

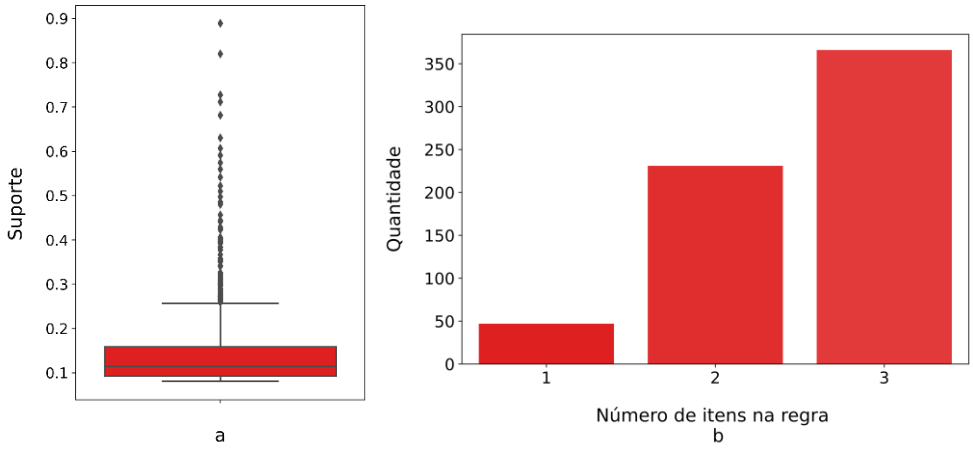
Diferente dos algoritmos *Apriori*, *FP-Growth* e *FP-Max*, o *Eclat* não pertence à biblioteca *mlxtend*. Utilizou-se a biblioteca *pyEclat* para aplicação do algoritmo, com o valor de suporte mínimo de 0,08 e valor máximo de características igual a 6. O algoritmo *Eclat* apresentou 644 itens, com média de 0,1595 de suporte e 2,4953 de quantidade de características. As estatísticas dos algoritmos estão apresentadas no Gráfico 4.28 e na Tabela 4.19.

Tabela 4.19 – Estatística das Regras.

Estatística	Suporte	Tamanho dos itens
Quantidade	644	644
Média	0,1595	2,4953
Desvio Padrão	0,1067	0,6297
Mínimo	0,0800	1
1º Quartil (25%)	0,0946	2
2º Quartil (50%)	0,1213	3
3º Quartil (75%)	0,1785	3
Máximo	0,8886	3

Fonte: Elaborado pelo autor

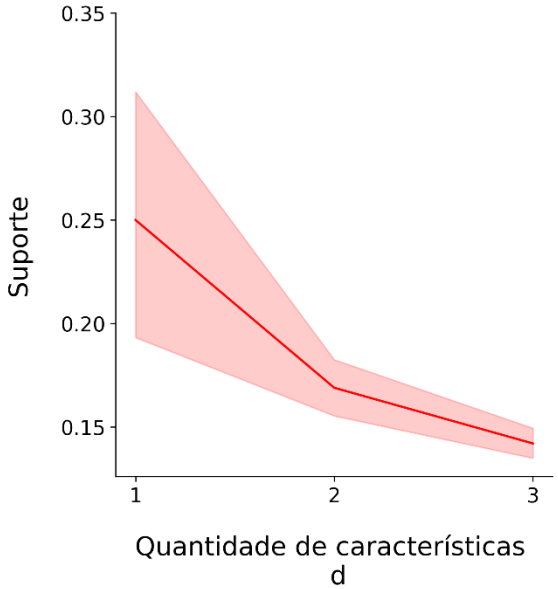
Gráfico 4.28 – (a) *Boxplot* do suporte (b) Histograma da quantidade de características



Fonte: Elaborado pelo autor

Nesse algoritmo, define-se o limite máximo de combinações e utilizou-se o valor igual a 6 para manter o valor máximo igual aos outros algoritmos que serão comparados. Observa-se que nesse algoritmo os valores máximo e mínimo de suporte foram exatamente iguais ao *Apriori* e *FP-Growth*. As destruições entre quartis foram diferentes, como pode ser visto no Gráfico 4.29, o qual tem o mesmo perfil do *Apriori* e *FP-Growth*.

Gráfico 4.29 – Média e desvio padrão dos suportes por quantidade de características



Fonte: Elaborado pelo autor

Como o *pyEclat* não retorna a tabela IF-THEN, utilizou-se do *mlxtend* com os resultados do *pyeclat* para obter a tabela IF-THEN do algoritmo *Eclat* (Tabela 4.20).

Ainda, a tabela IF-THEN foi criada pelo índice de suporte, pois não apresentou índices de *lift* e confiança, isto é, obteve valores nulos para *lift* e confiança. A tabela resultou em 394 regras criadas, onde um resumo das 10 regras ordenado pelos valores de suporte está apresentado na Tabela 4.20.

Tabela 4.20 – Tabela IF-THEN no 1º cenário do *Eclat*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{' Masculino'}	{'Dia Normal '}	-	-	0,7271	-	-
{'Dia Normal '}	{' Masculino'}	-	-	0,7271	-	-
{' Dia Normal'}	{'Horário Normal '}	-	-	0,6299	-	-
{'Horário Normal '}	{' Dia Normal'}	-	-	0,6299	-	-
{'Reta '}	{' Dia Normal'}	-	-	0,6062	-	-
{' Dia Normal'}	{'Reta '}	-	-	0,6062	-	-
{'Horário Normal '}	{' Masculino'}	-	-	0,5905	-	-
{' Masculino'}	{'Horário Normal '}	-	-	0,5905	-	-
{' Masculino'}	**{'Reta '}**	-	-	**0,5596**	-	-
{'Reta '}	**{' Masculino'}**	-	-	**0,5596**	-	-

Fonte: Elaborado pelo autor

Como IF, o *Eclat* apresentou somente itens com uma característica, mesmo assim, mostra associações interessantes como ‘Reta’ e ‘Masculino’. Porém, para melhor avaliação, utilizou-se os resultados diretos do *pyEclat* apenas com os índices de suporte, sendo os 10 principais itens com 3 características (Tabela 4.21).

Tabela 4.21 – Resumo da tabela IF-THEN no 1º cenário do *Eclat*

item	Suporte
{Masculino, Dia Normal, Horário Normal}	0,5217
{Reta, Masculino, Dia Normal}	0,4971
{Reta, Dia Normal, Horário Normal}	0,4249
{Pleno dia, Masculino, Dia Normal}	0,4042
{Reta, Masculino, Horário Normal}	0,3992
{Céu Claro, Masculino, Dia Normal}	0,3961
{Masculino, Dia Normal, Simples}	0,3578
{Pleno dia, Dia Normal, Horário Normal}	0,3551
{Céu Claro, Reta, Dia Normal}	0,3519
{Pleno dia, Reta, Dia Normal}	0,3419

Fonte: Elaborado pelo autor

Nessa tabela, nota-se as associações entre ‘Masculino’, ‘Dia Normal’ e ‘Horário Normal’, bem como nos outros algoritmos. Outros itens importantes são o ‘Céu Claro’, ‘Reta’ e ‘Dia Normal’ que têm características interessantes para análises. Sendo assim, aplicou-se o algoritmo nos dois outros cenários para comparação.

- 2º Cenário (Sem ‘Dia Normal’ e ‘Horário Normal’) com mínimo de confiança de 0,08.

As 10 principais regras com maiores valores de suporte no segundo cenário estão apresentadas na Tabela 4.22.

Tabela 4.22 – Tabela IF-THEN no 2º cenário do *Eclat*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Reta '}	{' Masculino'}	-	-	0,5596	-	-
{' Masculino'}	{'Reta '}	-	-	0,5596	-	-
{'Pleno dia '}	{' Masculino'}	-	-	0,4564	-	-
{' Masculino'}	{'Pleno dia '}	-	-	0,4564	-	-
{'Masculino '}	{' Céu Claro'}	-	-	0,4440	-	-
{' Céu Claro'}	{'Masculino '}	-	-	0,4440	-	-
{'Masculino '}	{' Simples'}	-	-	0,4056	-	-
{' Simples'}	{'Masculino '}	-	-	0,4056	-	-
{'Reta '}	{' Céu Claro'}	-	-	0,3938	-	-
{' Céu Claro'}	{'Reta '}	-	-	0,3938	-	-

Fonte: Elaborado pelo autor

No 2º cenário, o algoritmo resultou em 53 regras, associando novamente ‘Reta’, ‘Masculino’, ‘Simples’, ‘Céu Claro’. Como também apresentou somente uma característica, a Tabela 4.23 apresenta os 10 principais itens, sem aplicação do IF-THEN.

Tabela 4.23 – Resumo da tabela IF-THEN no 2º cenário do *Eclat*

item	Suporte
{Reta , Masculino , Céu Claro}	0,3236
{Reta , Pleno dia , Masculino}	0,3053
{Reta , Masculino , Simples}	0,2601
{Reta , Automóvel de 86 a 111 cv , Masculino}	0,2494
{Reta , Dupla , Masculino}	0,2411
{Pleno dia , Masculino , Céu Claro}	0,2385
{Reta , Masculino , Falta de Atenção à Condução}	0,2287
{Masculino , Simples , Céu Claro}	0,2269
{Pleno dia , Masculino , Simples}	0,2237
{Reta , Pleno dia , Céu Claro}	0,2192

Fonte: Elaborado pelo autor

Nesse caso, é possível perceber a associação entre as características pista ‘Reta’ com o sexo ‘Masculino’ e a ‘Falta de Atenção a Condução’, bem como ‘Automóvel de 86 a 111’ e ‘Céu Claro’. Como em todos os casos apresentou o sexo Masculino e, então foi aplicado o terceiro cenário.

- 3º Cenário (Sem ‘Dia Normal’, ‘Horário Normal’ e ‘Masculino’) com mínimo suporte de 0,05.

O terceiro cenário, com mínimo de suporte de 0,05, apresentou 64 regras, sendo as 10 principais listadas na Tabela 4.24.

Tabela 4.24 – Tabela IF-THEN no 3º cenário do *Eclat*

IF	THEN	Suporte IF	Suporte THEN	Suporte	Confiança	Lift
{'Céu Claro '}	{' Reta'}	-	-	0,3938	-	-
{' Reta'}	{'Céu Claro '}	-	-	0,3938	-	-
{' Reta'}	{'Pleno dia '}	-	-	0,3840	-	-
{'Pleno dia '}	{' Reta'}	-	-	0,3840	-	-
{'Reta '}	{' Simples'}	-	-	0,3095	-	-
{' Simples'}	{'Reta '}	-	-	0,3095	-	-
{'Pleno dia '}	{' Céu Claro'}	-	-	0,3013	-	-
{' Céu Claro'}	{'Pleno dia '}	-	-	0,3013	-	-
{'Reta '}	{' Automóvel de 86 a 111 cv'}	-	-	0,2996	-	-
{' Automóvel de 86 a 111 cv'}	{'Reta '}	-	-	0,2996	-	-

Fonte: Elaborado pelo autor

Como apresentou os mesmos resultados do 2º cenário, criou-se a tabela somente para uma característica (Tabela 4.25).

Tabela 4.25 – Resumo da tabela IF-THEN no 3º cenário do *Eclat*

item	Suporte
{Pleno dia, Céu Claro, Reta}	0,3236
{Céu Claro, Reta, Simples}	0,3053
{Pleno dia, Reta, Falta de Atenção à Condução}	0,2601
{Céu Claro, Reta, Automóvel de 86 a 111 cv}	0,2494
{Dupla, Pleno dia Reta}	0,2411
{Pleno dia, Reta, Simples}	0,2385
{Céu Claro, Reta, Falta de Atenção à Condução}	0,2287
{Dupla, Céu Claro, Reta}	0,2269
{Pleno dia, Reta, Automóvel de 86 a 111 cv}	0,2237
{Pleno dia, Céu Claro, Simples}	0,2192

Fonte: Elaborado pelo autor

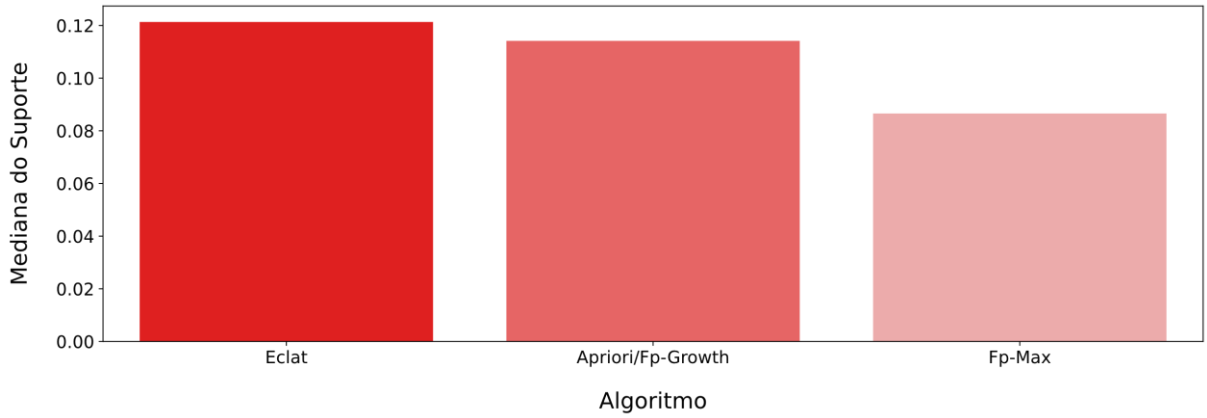
Com isso, é possível perceber a presença de associações interessantes como {Pleno dia, Reta, Falta de Atenção à Condução}, como apresentado no *Apriori*, *FP-Growth* e *FP-Max*, sendo, então, uma regra significativa que deve ser explorada por pesquisadores, engenheiros de trânsito e gestores.

Para fins de comparação entre os algoritmos, foi realizada uma estatística descritiva com a métrica de suporte e quantidade de características por item, isto se deve pelo fato que essas informações estão presente em todos os algoritmos. Além da estatística descritiva, realizou-se uma análise multivariada de variância para comparação dos algoritmos.

4.5 Comparação dos algoritmos

Nesse tópico, realizou-se a comparação dos algoritmos, como requerido no objetivo do presente estudo. A princípio, nota-se que os algoritmos *Apriori* e *FP-Growth* são os únicos que resultaram em índices significativos de *lift* e confiança. Já os algoritmos *FP-Max* e *Eclat* não apresentaram tais resultados. Para comparação, realizou-se uma estatística descritiva dos índices de suporte, onde os algoritmos *Apriori*, *FP-Growth* e *Eclat* apresentaram as melhores medianas como Gráfico 4.30.

Gráfico 4.30 – Mediana das métricas de suporte dos algoritmos



Fonte: Elaborado pelo autor

Realizou-se uma análise de variância multivariada (MANOVA) na média dos índices de suporte e tamanho de quantidade de características para testar, estatisticamente, se a diferença entre as médias é significativa. Para isso, foi necessário realizar a normalização dos dados, utilizando a biblioteca *scipy.stats.norm* do Python. Em seguida, realizou-se a verificação da normalidade

utilizando o teste de *Shapiro-Wilk*, sendo: Hipótese nula (H_0) = Os dados tem distribuição normal e a Hipótese alternativa (H_1) = Os dados não tem distribuição normal.

Como resultado do Teste *Shapiro-Wilk* obteve-se: ShapiroResult(statistic= 0.9997705221176147, pvalue=0.8119539022445679). Portanto, como o p-valor é maior que 0,05, não se rejeita a hipótese nula, ou seja, após normalizado, os dados têm distribuição normal e pode-se utilizar o modelo MANOVA para análise de variância.

Na análise de correspondência de variância, utilizou-se como variáveis respostas os índices de Suporte (*support*) e quantidade de características (*length*). Como covariável empregou-se a variável Algoritmo. Aplicando o teste MANOVA através da biblioteca *statsmodels.multivariate.manova* do Python com o seguinte *script* MANOVA.from_formula('support + length ~ Algoritmo', data=dados_normalizados), obteve-se a seguinte tabela de *multivariate linear model* (Tabela 4.26).

Tabela 4.26 – Análise Multivariada de Variância

Multivariate linear model

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.8964	2.0000	1947.0000	112.5481	0.0000
Pillai's trace	0.1036	2.0000	1947.0000	112.5481	0.0000
Hotelling-Lawley trace	0.1156	2.0000	1947.0000	112.5481	0.0000
Roy's greatest root	0.1156	2.0000	1947.0000	112.5481	0.0000

Algoritmo	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9998	2.0000	1947.0000	0.1472	0.8631
Pillai's trace	0.0002	2.0000	1947.0000	0.1472	0.8631
Hotelling-Lawley trace	0.0002	2.0000	1947.0000	0.1472	0.8631
Roy's greatest root	0.0002	2.0000	1947.0000	0.1472	0.8631

Fonte: Elaborado pelo autor

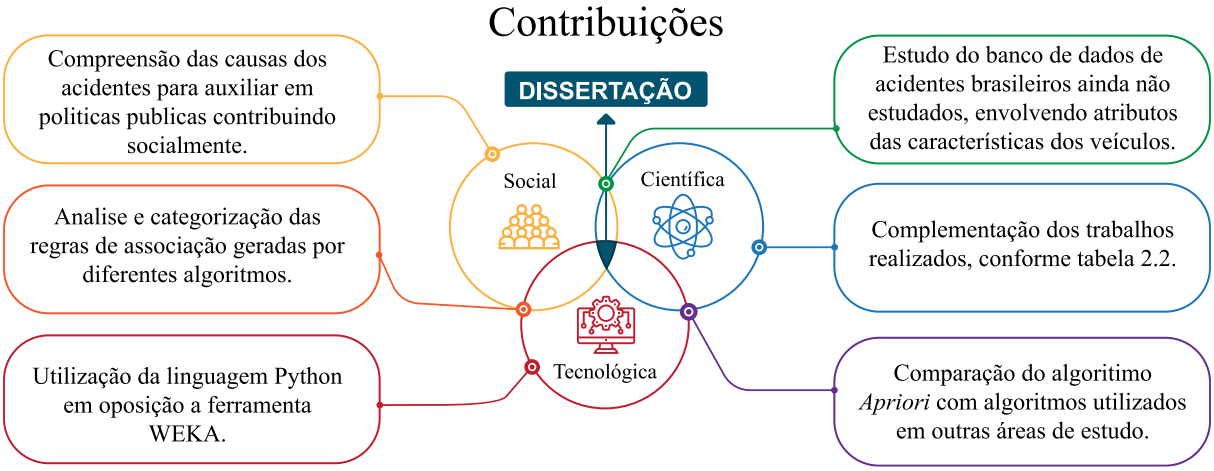
Então, por meio do teste MANOVA, tem-se que para a Hipótese nula (H_0) todas as médias de população são iguais, enquanto a hipótese alternativa (H_1) afirma que pelo menos uma é diferente. Para um nível de significância de 0,05, e ao utilizar a regra usual de decisão, ou seja, não se rejeitar H_0 se F aproximado $\leq$ F tabelado e, rejeitar H_0, se F calculado > F tabelado.

Percebe-se, então, que, por meio das estatísticas de *Wilks, Pillai's, Hotelling-Lawley* e *Roy's Greatest*, apresenta-se nível de significância P-Value = 0,8631, superior a 0,05, fazendo-se com que seja rejeitada a hipótese nula, isto é, ao menos uma das médias dos algoritmos é diferente, comprovando as análises de estatística descritivas desse estudo, onde os algoritmos de regras de associação apresentaram diferentes estatísticas de suporte e quantidade de características para os mesmos dados de acidentes. Porém, apresentaram as mesmas regras de associação relevantes entre as características dos algoritmos nos três cenários estudados, independente do algoritmo, respondendo às perguntas de pesquisa e atingindo os objetivos desse trabalho.

Com isso, esse estudo conseguiu realizar sua contribuição técnico-científica e social, comparando diferentes algoritmos utilizados em diversas áreas de estudo, através da linguagem Python em oposição a tradicional ferramenta WEKA. Além disso, analisou e categorizou regras de associação em uma base de dados de acidentes brasileiros ainda não estudados, com um diferencial, onde incluiu as características dos veículos. Logo, complementou os trabalhos relacionados (Tabela 2.2) e compreendeu as causas dos acidentes para auxiliar engenheiros de segurança a tomarem decisões e gestores na criação de políticas públicas.

Um resumo da contribuição técnico-científica e social da pesquisa está demonstrado na Figura 4.1.

Figura 4.1 – Contribuições da pesquisa



Fonte: Elaborado pelo autor

5 CONSIDERAÇÕES FINAIS

O principal objetivo deste trabalho consistiu em comparar algoritmos de aprendizado de máquina para fins de identificação das regras de associação entre causas de acidentes, características dos veículos, estradas, usuários e meio ambiente em rodovias federais brasileiras. Desta forma, buscou-se criar um relatório com representações gráficas dos dados dos acidentes em rodovias federais brasileiras no período de janeiro de 2017 a fevereiro de 2020. Além disso, esta pesquisa também visou analisar independências das características dos acidentes e suas causas, bem como obter regras de associação por meio de algoritmos de aprendizado de máquina não supervisionados e, por fim, compará-los de modo a relacionar as regras de associação pertinentes para tomadas de decisões e políticas públicas.

Com a intenção de adotar todas as definições estudadas nesse projeto, elaborou-se uma metodologia multidisciplinar baseada na revisão de literatura, utilizando-se de um método misto para coleta, transformação dos dados e análise dos resultados das técnicas de aprendizado de máquina *Apriori, Eclat, FP-Growth* e *FP-Max*, assim, foi possível responder às perguntas de pesquisa de modo a atingir os objetivos do presente estudo.

No desenvolvimento e aplicação da metodologia foram utilizados dados abertos de acidentes, os quais foram obtidos junto à PRF do Brasil, sendo que tais dados podem ser utilizados por qualquer pessoa e, por isso, tal fato consiste em um ponto importante para estudos nessa nova era de *Big Data Analytics*, uma vez que dados abertos são essenciais para cidades inteligentes e pesquisas na área de transportes, o qual visa, ainda, assegurar um bom desempenho na segurança viária. É importante ressaltar que os dados com as caraterísticas dos veículos e os dados de acidentes não são disponibilizados abertamente. Porém, a obtenção desses dados, os quais são raramente usados em pesquisas acadêmicas, bem como a orientação da aquisição dos dados do Renavam, disponibilizados pelo Ministério da Infraestrutura do Brasil (MINFRA) por meio de solicitação às informações públicas no portal Fala.br, é considerada uma importante conquista desse estudo, visto que outros pesquisadores podem utilizar o mesmo método de obtenção dos dados.

Além disso, por meio do dicionário de dados foi possível observar que os bancos de dados de acidentes e de dados das características dos veículos possuem variáreis qualitativas e quantitativas, sendo, então, necessário o método misto de estratégia transformadora concomitante. No procedimento de análise exploratória e tratamento dos dados através da linguagem Python, a qual consiste em uma linguagem *open source*, ou seja, uma linguagem de

código aberto, considerada uma das principais linguagens de cientistas de dados, podendo ser utilizada em outros estudos sem necessidade de adquirir um software não gratuito. Através da linguagem Python, foi possível realizar um pré-processamento e limpeza do enorme volume de dados obtidos, garantindo sua confiabilidade, bem como removendo registros ausentes, duplicados e desnecessários para análise e alcance dos objetivos do estudo.

Durante o processo de exploração dos atributos, para cada um deles foi gerado um relatório com representações gráficas dos dados dos acidentes, onde foi possível perceber alguns indicadores relevantes para a análise dos acidentes em rodovias federais brasileiras. Além destes indicadores, foi possível extrair novas hipóteses para trabalhos futuros na área de transporte, tecnologia e segurança viária. Ainda, outro fator importante de ser mencionado é a enorme quantidade de registros inconsistentes, isto é, um enorme volume de outliers e registros faltantes, o que leva a questionar o mecanismo e a maneira nos quais os acidentes são registrados pela PRF.

Percebe-se que a obtenção dos dados é de extrema relevância para identificar fatores e ajudar a reduzir taxas de acidentes, uma das principais causas de mortes no Brasil e no mundo. Com intuito de aprofundar no tema e ir além dos objetivos iniciais desse estudo, solicitou-se à PRF, por meio do portal Fala.br, a documentação do software e imagem das telas (*print screens*) onde é preenchido o boletim de ocorrência de acidentes em rodovias federais, visando sugerir melhorias no sistema. Conforme resposta da Diretoria de Inteligência, na manifestação nº 08198.035445/2021-95, destinadas ao DPRF – Departamento de Polícia Rodoviária Federal do Brasil, as informações geradas, adquiridas ou custodiadas no sistema, estão sob a responsabilidade da PRF, constituindo parte integrante de seu patrimônio, sendo, então, vedada a sua utilização por terceiros e as informações não são passíveis de serem fornecidas, uma vez que o sistema é de acesso restrito aos policiais rodoviários federais e o fornecimento de *prints screens* das telas do referido sistema podem comprometer as atividades desenvolvidas pelo órgão. Portanto, não foi possível entender a arquitetura do software e indicar melhorias para redução de dados inconsistentes e pontos fundamentais para garantir a disponibilidade e confiabilidade dos dados, uma vez que se percebe, baseado no procedimento metodológico e de análise desse estudo, as dificuldades que envolvem a criação de uma metodologia de aplicação de aprendizado de máquina sem dados consistentes.

Apesar das dificuldades no pré-processamento e tratamento dos dados durante o desenvolvimento e aplicação dos algoritmos, o presente estudo conseguiu obter a matriz adequada para aprendizado de máquina de regras de associação *Apriori*, *Eclat*, *FP-Growth* e

FP-Max. Desta forma, as aplicações dos algoritmos nos diversos cenários alcançaram resultados satisfatórios com valores significativos de *lift*, suporte e confiança.

Na análise dos resultados, observou-se que os algoritmos *Apriori*, *FP-Growth* e *Eclat* apresentam o mesmo desempenho, com índices de suporte e quantidade de características similares, onde, quanto maior a quantidade de características, menor o índice de suporte. Por outro lado, o *FP-Max*, que propõe uma maior métrica de suporte para maior quantidade de características, apresentou desfecho contrário, proporcionando um resultado mais preciso. O *FP-Max* e como o *Eclat* não apresentaram índices de *lift* e confiança para esse banco de dados. Ainda, além de índices de suporte e confianças significativos, os algoritmos apresentam regras de associação que levam a novos questionamentos, principalmente quanto à segurança viária, como, por exemplo, se um condutor do sexo masculino dirige um veículo em um dia que não seja feriado, fora do horário de pico, em uma reta, essas características estão associadas com acidentes onde a causa é não guardar distância de segurança. Assim como outra característica ocorre quando um condutor do sexo masculino dirige em um dia que não seja feriado, fora do horário de pico, em uma reta com céu claro em plena luz do dia está relacionado a acidentes onde a causa é a falta de atenção na condução. Apesar de essas associações serem interessantes, um ponto importante analisado por esse estudo é que as características das regras de associação que apresentam maior significância são aquelas que apresentam maior quantidade acidentes, conforme métricas observadas no relatório com representação gráfica dos acidentes. Logo, para uma análise mais precisa, sugere-se, para trabalhos futuros, a realização dessa metodologia com um maior volume de dados, onde seja possível realizar essa análise com dados balanceados, isto é, que seja possível utilizar um banco de dados onde cada variável possua a quantidade de características equivalente, como por exemplo a variável sexo, onde, em um banco de dados balanceados, tenha a mesma quantidade de observações para o sexo masculino e feminino.

Em função dos fatos mencionados e analisados, esse estudo conseguiu atingir com êxito seu objetivo principal e seus objetivos específicos, comparando algoritmos de aprendizado de máquina para identificação das regras de associação entre as causas de acidentes, bem como as características dos veículos, das estradas, dos usuários e do meio ambiente em rodovias federais brasileiras.

REFERÊNCIAS

Agrawal R. & Srikant R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

Agrawal, R., Imielienski, T., and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In: Proc. Conf on Management of Data, 207-216. New York: ACM Press

Ahlgren, B.; Hidell, M.; Ngai, E. C. H. (2016). Internet of Things for Smart Cities: Interoperability and Open Data. IEEE Internet Computing, v.20, n.6, p. 52–56. https://doi.org/10.1109/MIC.2016.124

Al-Harbi, M.; Yassin, M. F.; Shams M. B.; (2012). Stochastic modeling of the impact of meteorological conditions on road traffic accidents. Stochastic Environmental Research and Risk Assessment, v.26, n.5,. https://doi.org/10.1007/s00477-012-0584-y

Ali, F. M. N., & Hamed, A. A. M. (2018). Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents. Journal of Information and Telecommunication, v. 2 , p. 231-245. https://doi.org/10.1080/24751839.2018.1448205

Almeida R. L. F.; Bezerra Filho J. G.; Braga J. U.; Magalhães F. B.; Macedo M. C. M. & Silva K. A.; (2013). Via, homem e veículo: fatores de risco associados à gravidade dos acidentes de trânsito. Revista Saúde Pública. v.47, n.4, p. 718-731.

Alpaydin E. (2004). Introduction to Machine Learning. The MIT Press. London.

Amorim, Brunna de Sousa Pereira (2019). Uso de Aprendizado de Máquina para Classificação de Risco de Acidentes em Rodovias. Dissertação (Mestrado em Ciência da Computação), Universidade Federal de Campina Grande, Campina Grande.

Angolini, A. C. (2005). Romi-Isetta - O pequeno pioneiro. DBA Editora. Ed. 1.

Atnafu, B.; Kaur, G. (2017). Survey Paper on Analyze and Predict the Nature of Road Traffic Accidents using Data Mining Techniques in Maharashtra, India. International Journal of Engineering Technology Science and Research, v.53, n.1, p. 23–31.

https://doi.org/10.14445/22315381/IJETT-V53P206

Barroso Junior, G. T. B; Bertho, A. C. S. & Veiga, A. de C.; (2019). A letalidade dos acidentes de trânsito nas rodovias federais brasileiras. Revista Brasileira De Estudos De População, 36, 1–22. https://doi.org/10.20947/S0102-3098a0074

Başţanlar, Y., & Ozuysal, M. (2014). Introduction to Machine Learning Second Edition. In: Methods in molecular biology. Clifton v. 1107. https://doi.org/10.1007/978-1-62703-748-8_7

Bishop. C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Blows S; Ivers R. Q.; Woodward M., Connor J., Ameratunga S.

& Norton R. (2003) Vehicle year and the risk of car crash injury. Inj Prev. 2003 v. 9 n. 4 p. 353-6. doi: 10.1136/ip.9.4.353.

Borgelt, C. & Kruse, R. (2002). Induction of Association Rules: A Priori Implementation. 15th Conference on Computational Statistics. 10.1007/978-3-642-57489-4_59.

Bouakkaz, M., Ouinten, Y., Zian, B. (2012). Vertical Fragmentation of Data warehouses using the FP-Max Algorithm . In: 2012 International Conference on Innovations in Information Technology, Abu Dhabi. pp. 273-276, doi: 10.1109/INNOVATIONS.2012.6207746

Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. Brasília, p. 1. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm> Acesso em: 31 jan. 2021.

Brasil (2017). Dicionário de Variáveis - Acidentes. Dados Agregados por pessoa, com todas as causas e tipos de acidentes. Brasília.

Brasil (2020). Decreto nº 6, de 20 de março de 2020. Brasília, p. 1. Disponível em: https://www.planalto.gov.br/ccivil_03/portaria/dlg6-2020.htm. Acesso em: 31 jan. 2021

Brasil. (2018). Avaliação das políticas públicas de transportes: Segurança nas Rodovias Federais. Ministério dos Transportes Portos e Aviação, Brasília.

Brin S.; Motwani R., Ullman J. D. & Tsur S. (1997). Dynamic itemset counting and implication rules for market basket data. SIGMOD Rec. 26, 2 (June 1997), 255–264. DOI:https://doi.org/10.1145/253262.253325

Chong, M.; Abraham, A.; Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. Informatica, v. 29, p. 89-98. https://doi.org/10.31449/inf.v29i1.21

Chung, E.; Ohtani, O.; Warita, H; Kuwahara H. & Morita H. (2005). Effect of Rain on Travel Demand and Traffic Accidents. In: Proceedings of the 8th International – 2005 IEEE Conference on Intelligent Transportation Systems. Viena. p. 13-16

Confederação Nacional dos Transportes – CNT. Anuário da malha rodoviária. Disponível em https://anuariodotransporte.cnt.org.br/2020/Rodoviario/1-3-1-1-1-/Malha-rodovi%C3%A1ria-total, acessado em 10 de dezembro de 2021.

Costa, J. D. J.; Bernardini, F. C.; Viterbo Filho, J. (2014). A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. AtoZ: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 139-157. https://doi.org/10.5380/atoz.v3i2.41346

Creswell, J. W. (2007). Projeto de Pesquisa: métodos qualitativo, quantitativo e misto; tradução Luciana de Olveira da Rocha – 2 ed. Porto Alegre: Artmed.

Cunto, F. (2008). Assessing Safety Performance of Transportation Systems using Microscopic Simulation. UWSpace. https://uwspace.uwaterloo.ca/handle/10012/4111

Daher, J. R.; Chilkaka, S.; Younes, A.; Shaban, K. (2016). Association rule mining on five years of motor vehicle crashes. MATEC Web of Conferences. v. 81, n. 02017 https://doi.org/10.1051/matecconf/20168102017

Das, S.; Avelar, R.; Dixon, K. & Sun, X. (2018). Investigation on the wrong way driving crash patterns using multiple correspondence analysis. Accident; analysis and prevention. v.111. p. 43-55. 10.1016/j.aap.2017.11.016.

Deekshitha, H. R.; Sumana, K. R.; Phaneendra, D. H. D. (2019). Smart Automated Modelling using Eclat Algorithm for Traffic Accident Prediction. International Research Journal of Engineering and Technology (IRJET) v. 6, n. 5, p. 6682-6685.

https://doi.org/10.13140/RG.2.2.34583.52646

Facelli, K.; Lorena A.C; Gama. J. & Cavalho A.C.P.L.F (2011) Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro, RJ: LTC, 2011. xvi, 378 p. ISBN 9788521618805

Figueira, A. C.; Pitombo, C. S.; Oliveira, P. T. M.; S., Larocca, A. P. C. (2017). 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. https://doi.org/10.1016/j.cstp.2017.02.00

Gopalakrishnan, S. (2012). A Public Health Perspective of Road Traffic Accidents. Journal of Family Medicine and Primary Care, v.1, n. 2, p. 144-150. https://doi.org/10.4103/2249-4863.104987

Greenacre M. J. (1993). Correspondence Analysis in Practice. Academic Press, London.

HAIR, Joseph F. William C. B.; Barry J. B.; Rolph E. A. & Ronald L. T. Análise multivariada de dados. (2009) Porto Alegre: Bookman. p. 688 ISBN 9788577804023.

Han, J., Pei, J., Yin, Y. et al. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8, 53–87 (2004). https://doi.org/10.1023/B:DAMI.0000005258.31418.83

Harrison M. (2020). Machine Learning – Guia de Referência Rápida: Trabalhando com Dados Estruturados em Python. Novatec Editora; 1ª edição. São Paulo. ISBN 978-8575228173

Heaton, J. (2017). Comparing Dataset Characteristics that Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms.

Hegland, Markus. (2007). The Apriori Algorithm–a Tutorial. 11. 10.1142/9789812709066_0006.

Homem L. W. (2020). Apostila de Machine Learning. Vitória.

Hunyadi, D. (2011). Performance Comparison of Apriori and FP-Growth Algorithms in Generating Association Rules. In: The 5th European Computing Conference (ECC-11), Paris.

Ishita, R. & Rathod, A. (2016). Eclat with Large Data base Parallel Algorithm and Improve its Efficiency. In International Journal of Computer Applications. 143. 33-37. 10.5120/ijca2016910462.

Jalayer, M.; Pour-Rouholamin, M.; & Zhou, H. (2017). Wrong-Way Driving Crashes: A Multiple Correspondence Approach to Identify Contributing Factors. Traffic Injury Prevention. v.19. 10.1080/15389588.2017.1347260.

Jiménez-Mejías, E.; Amezcua-Prieto, C.; Martínez-Ruiz, V.; Dios, L.; Pablo, L. & Jiménez-Moleón, J. (2014). Gender-related differences in distances travelled, driving behaviour and traffic accidents among university students. In: Transportation Research Part F Traffic Psychology and Behaviour. v. 27. p. 81-89. 10.1016/j.trf.2014.09.008.

Jung, C. (2004). Metodologia para Pesquisa & Desenvolvimento: Aplicada a Novas Tecnologias, Produtos e Processos. Rio de Janeiro: Axcel Books do Brasil

Kantardzic M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Wiley-IEEE Press; 2nd ed. edição. ISBN 9780470890455

Karacasu, M. & Er, A. (2011). An Analysis on Distribution of Traffic Faults in Accidents, Based on Driver's Age and Gender: Eskisehir Case. In: Procedia - Social and Behavioral Sciences. v. 20. p. 776-785. 10.1016/j.sbspro.2011.08.086.

Kaur, Gagandeep. (2015). Identify and Compare discernment rules for accurate Liver Disorder detection using Apriori and FP Generation Analysis. International Journal of Computer Science and Information Technologies, v. 6, n. 3, p. 2244-2255. -

Kumar, S.; Toshniwal, D. (2015). A data mining framework to analyze road accident data. Journal of Big Data, v.2, n. 26. https://doi.org/10.1186/s40537-015-0035-y

Kumar, S.; Toshniwal, D.; Parida, M. (2017). A comparative analysis of heterogeneity in road accident data using data mining techniques. Evolving Systems, v. 8, p. 147-155. https://doi.org/10.1007/s12530-016-9165-5

L'Heureux, A.; Grolinger, K.; Elyamany, H. F.; Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. IEEE Access, v. 5, p. 7776-7797. https://10.1109/ACCESS.2017.2696365.

Lankarani, K.;Heydari, S.;Aghabeigi, M.; Moafian, G. & Hoseinzadeh, A.; Vossoughi, M. (2013). The impact of environmental factors on traffic accidents. In: Iran. Journal of injury & violence research. v. 6. https://doi.org/10.5249/jivr.v6i2.318.

Li, L., Shrestha, S., & Hu, G. (2017, June). Analysis of road traffic fatal accidents using data mining techniques. In: IEEE - 15th International Conference on Software Engineering Research, Management and Applications (SERA), Londres.

Lima, I. M. de O.; Figueiredo, J. C.; Morita P. A; Gold, P.; (2008). Fatores condicionantes da gravidade dos acidentes de trânsito nas rodovias brasileiras. Instituto de Pesquisa Econômica Aplicada (IPEA).

Luchezi, T. de F. (2010). O Automóvel como Símbolo da Sociedade Contemporânea. Anais do VI Seminário de Pesquisa em Turismo do Mercosul.

Maoski, F. (2014). Ter um carro é .... A percepção sobre o significado do carro e o comportamento do condutor. 1–96.

Martín, L., Baena, L., Garach, L., López, G., & de Oña, J. (2014). Using Data Mining Techniques to Road Safety Improvement in Spanish Roads. Procedia - Social and Behavioral Sciences, v. 160, p. 607-614. https://doi.org/10.1016/j.sbspro.2014.12.174

Másilková, M. (2017). Health and social consequences of road traffic accidents. Kontakt, v. 19, p. 43-47 https://doi.org/10.1016/j.kontakt.2017.01.007

Meng, H., Hong, Y., Ma, Y., Li, Z., Lu, J., & Siddiqui, N. A. (2019). Association Rule-Based Traffic Accident Impact Factors Analysis on Low-Grade Highways. In: 19th COTA International Conference of Transportation Professionals. p. 3549-3559, Nanjing.

Mingoti, S. A. (2007) Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Ed. UFMG.

Ministério da Infraestrutura. Frota de veículos – 2020. Disponível em https://www.gov.br/infraestrutura/pt-br/assuntos/transito/conteudo-denatran/frota-de-veiculos-2020, acessado em 10 de dezembro de 2021.

Minsky, M. L. (1974). A Framework for Representing Knowledge. Massachusetts:

Massachusetts Institute of Technology A.I. Laboratory.

Mohan, D., Tiwari, G., Khayesi, M., & Nafukho, F. M. (2006). Road traffic injury prevention training manual. In World Health Organization, Geneva.

Mohri M.; A. Rostamizadeh A. & A. Talwalkar A. (2012). Foundations of Machine Learning. The MIT Press.

Müller, A., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists.

Nandurge, P. A., & Dharwadkar, N. V. (2017). Analyzing road accident data using machine learning paradigms. In: Proceedings of the International Conference on IoT in Social, Mobile, Analytics and Cloud, Palladam. https://doi.org/10.1109/I-SMAC.2017.8058251

Ozbayoglu, M., Kucukayan, G., & Dogdu, E. (2016). A real-time autonomous highway accident detection model based on big data processing and computational intelligence. In: Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016. Washington. https://doi.org/10.1109/BigData.2016.7840798

Pereira, G. V., Macadar, M. A., Luciano, E. M., & Testa, M. G. (2017). Delivering public value through open government data initiatives in a Smart City context. Information Systems Frontiers, v. 19, p. 213–229. https://doi.org/10.1007/s10796-016-9673-7

Polícia Rodoviária Federal – PRF. Dados abertos – acidentes. Disponível em https://portal.prf.gov.br/dados-abertos-acidentes, acessado em 12 de agosto de 2020.

Porter, E. B. (2011). Handbook of Traffic Psychology. In Handbook of Traffic Psychology. Academic Press. ISBN 9780123819840. https://doi.org/10.1016/B978-0-12-381984-0.10023-2

Raschka, S. (2018) MLxtend: Providing machine learning and data science tilities and extensions to Python's scientific computing stack.

Reis, C., Silva, J., & Maia, L. (2015). O Uso Da Descoberta De Conhecimento Em Banco De Dados Nos Acidentes Da BR-381. In: XVI Encontro Nacional de Pesquisa Em Ciência Da Informação (XVI ENANCIB), João Pessoa.

Resende P. T. V. & Souza P. R. (2009). Mobilidade urbana nas grandes cidades brasileiras: Um estudo sobre os impactos do congestionamento. Nova Lima: Fundação Dom Cabral

Sammut C. & Webb G. I. (2011). Encyclopedia of Machine Learning and Data Mining. Springer US. New York p. 1335. ISBN 978-1-4899-7685-7

Scialfa, C.; Guzy, L.; Leibowitz, H.; Garvey, P. & Tyrrell, .d. (1991). Age differences in estimating velocity. Psychology and aging. v.6. p. 60-6. 10.1037//0882-7974.6.1.60.

Shanti, S., Ramani, D. R. G., Shanthi, S., & Ramani, R. G., Shanti, S., & Ramani, D. R. G. (2011). Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms. International Journal of Computer Applications, v. 35, n. 12, p. 30-37.

Silva, P. B., Andrade, M., & Ferreira, S. (2019). Priorização De Variáveis Explicativas Na Modelagem De Acidentes De Trânsito Utilizando Técnicas De Aprendizado De Máquina. In: 33º Congresso de Pesquisa e Ensino em Transporte da ANPET, Balneário Camboriú.

Soares, L. C.; Prado, H. A.; Balaniuk, R.; Ferneda, E.; Bortoli, A. (2018). Caracterização de acidentes rodoviários e as açoes governamentais para a redução de mortes e lesões no trânsito. Revista Transporte y Territorio, v. 19, p. 188-220 https://doi.org/10.34096/rtt.i19.5331

Tate, A., & Bewoor, L. (2017). Survey on frequent pattern mining algorithm for kernel trace. In: 7th IEEE International Advanced Computing Conference, Hyderabad. https://doi.org/10.1109/IACC.2017.0163

Tayeb, A. A. El, Pareek, V., & Araar, A. (2015). Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai. International Journal of Soft Computing and Engineering, v. 5. p. 1-12.

Tyagi, A.; Kumar, A.; Gandhi, A. & Mueller, K. (2018). Road Accidents in the UK (Analysis and Visualization). In: Conference: IEEE VIS 2018.

United States General Accounting Office (1994). Factors Affecting Involvement in Vehicle Crashes. In Report to Congressional Requesters. Washington.

Washington, S. P., Karlaftis, M. G., Mannering, F. L. (2003) Statistical AND Econometric

Methods FOR Transportation Data Analysis.

World Health Organization. (2018). Global status report on road safety 2018. Geneva: World Health Organization; 2018. Licence: CC BY- NC-SA 3.0 IGO

Xi, J., Zhao, Z., Li, W., & Wang, Q. (2016). A Traffic Accident Causation Analysis Method Based on AHP-Apriori. Procedia Engineering, v. 137, p. 680-687. https://doi.org/10.1016/j.proeng.2016.01.305

Zaki, M. J. (2000). Scalable algorithms for association mining. In: IEEE Transactions on Knowledge and Data Engineering, 12(3):372–390.

Zhang, X. (2020). A Matrix Algebra Approach to Artificial Intelligence. 10.1007/978-981-15-2770-8.

APÊNDICE A – Tabelas de contingência

Tabela de contingência de causas de acidentes por dias da semana

Causa do Acidente	domingo	segunda-feira	terça-feira	quarta-feira	quinta-feira	sexta-feira	sábado
Agressão Externa	44	26	28	37	33	36	55
Animais na Pista	428	356	276	276	300	329	328
Avarias e/ou desgaste excessivo no pneu	185	137	97	104	143	147	200
Carga excessiva e/ou mal acondicionada	10	26	24	28	23	31	22
Condutor Dormindo	945	582	452	410	474	531	840
Defeito Mecânico no Veículo	618	510	420	439	444	528	546
Defeito na Via	164	170	131	140	115	132	165
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	45	25	28	31	41	31	28
Desobediência às normas de trânsito pelo condutor	2103	1711	1510	1494	1625	2081	2237
Desobediência às normas de trânsito pelo pedestre	43	36	31	29	28	27	32
Falta de Atenção do Pedestre	352	262	231	256	253	329	366
Falta de Atenção à Condução	8027	6741	5780	6035	6485	7837	7912
Fenômenos da Natureza	77	58	50	51	55	73	61
Ingestão de Substâncias Psicoativas	34	14	10	11	8	18	46
Ingestão de Álcool	4151	943	654	711	925	1476	3495
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	120	34	21	27	46	42	121
Mal Súbito	151	171	202	146	146	172	158
Não guardar distância de segurança	1675	1753	1516	1416	1484	2034	1589
Objeto estático sobre o leito carroçável	113	92	99	116	108	118	118

Pista Escorregadia	773	701	493	500	537	768	752
Restrição de Visibilidade	156	126	112	108	125	114	113
Sinalização da via insuficiente ou inadequada	46	52	51	29	43	56	78
Ultrapassagem Indevida	458	346	321	316	369	442	503
Velocidade Incompatível	2383	1590	1222	1169	1306	1716	2263

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por feriado, horário de pico e fases do dia.

Causa do Acidente	Dia Normal	Feriado	Horário Normal	Horário de Pico	Amanhecer	Anoitecer	Plena Noite	Pleno dia
Agressão Externa	241	18	197	62	15	9	90	145
Animais na Pista	2013	280	1862	431	189	81	1555	468
Avarias e/ou desgaste excessivo no pneu	866	147	768	245	37	51	205	720
Carga excessiva e/ou mal acondicionada	152	12	124	40	2	3	42	117
Condutor Dormindo	3685	549	3488	746	602	81	1589	1962
Defeito Mecânico no Veículo	3076	429	2578	927	144	169	991	2201
Defeito na Via	895	122	729	288	43	54	211	709
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	202	27	154	75	7	10	167	45
Desobediência às normas de trânsito pelo condutor	11458	1303	8951	3810	404	734	4003	7620
Desobediência às normas de trânsito pelo pedestre	206	20	169	57	4	12	125	85
Falta de Atenção do Pedestre	1843	206	1492	557	60	116	1073	800
Falta de Atenção à Condução	43667	5150	33719	15098	1675	2918	13777	30447
Fenômenos da Natureza	361	64	299	126	17	31	86	291
Ingestão de Substâncias Psicoativas	120	21	115	26	4	2	71	64

Ingestão de Álcool	10908	1447	9826	2529	742	644	7625	3344
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	367	44	322	89	17	25	253	116
Mal Súbito	1031	115	846	300	38	59	226	823
Não guardar distância de segurança	10233	1234	7043	4424	266	821	2269	8111
Objeto estático sobre o leito carroçável	695	69	587	177	40	46	349	329
Pista Escorregadia	3929	595	3234	1290	218	303	984	3019
Restrição de Visibilidade	763	91	559	295	62	82	343	367
Sinalização da via insuficiente ou inadequada	317	38	265	90	14	16	139	186
Ultrapassagem Indevida	2412	343	1917	838	120	194	766	1675
Velocidade Incompatível	10237	1412	8589	3060	543	602	3316	7188

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por condições meteorológicas

Causa do Acidente	Chuva	Céu Claro	Garoa/Chuvisco	Granizo	Nevoeiro/Neblina	Nublado	Sol	Vento
Agressão Externa	14	161	8	0	1	55	19	1
Animais na Pista	134	1519	60	0	36	472	61	11
Avarias e/ou desgaste excessivo no pneu	289	437	65	0	6	144	72	0
Carga excessiva e/ou mal acondicionada	6	94	1	0	0	39	23	1
Condutor Dormindo	249	2533	116	1	78	930	318	9
Defeito Mecânico no Veículo	368	1985	104	0	18	688	337	5
Defeito na Via	247	504	30	0	4	146	83	3
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	16	140	8	0	7	54	4	0

Desobediência às normas de trânsito pelo condutor	895	7717	422	1	78	2434	1184	30
Desobediência às normas de trânsito pelo pedestre	10	151	6	0	1	42	16	0
Falta de Atenção do Pedestre	138	1248	59	0	23	436	142	3
Falta de Atenção à Condução	4678	28419	1464	1	269	9359	4524	103
Fenômenos da Natureza	310	52	16	0	10	15	17	5
Ingestão de Substâncias Psicoativas	9	84	6	0	0	33	9	0
Ingestão de Álcool	1125	7457	480	0	118	2651	490	34
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	34	272	15	0	2	78	9	1
Mal Súbito	66	714	31	0	0	210	125	0
Não guardar distância de segurança	1117	6348	341	0	41	2132	1474	14
Objeto estático sobre o leito carroçável	100	397	35	0	13	157	61	1
Pista Escorregadia	3467	253	504	0	13	260	27	0
Restrição de Visibilidade	162	367	30	0	96	112	85	2
Sinalização da via insuficiente ou inadequada	38	226	11	0	4	45	29	2
Ultrapassagem Indevida	206	1683	62	0	33	497	271	3
Velocidade Incompatível	4040	4065	791	0	92	1999	648	14

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por tipo de pista e traçado da via

Causa do Acidente	**Dupla**	**Múltipla**	**Simples**	**Curva**	**Desvio Temporário**	**Interseção de vias**	**Ponte**	**Reta**	**Retorno Regulamentado**	**Rotatória**	**Túnel**	**Viaduto**

Agressão Externa	130	35	94	46	8	3	4	193	2	1	0	2
Animais na Pista	756	48	1489	226	113	5	13	1924	6	2	1	3
Avarias e/ou desgaste excessivo no pneu	486	44	483	325	56	6	5	606	5	3	0	7
Carga excessiva e/ou mal acondicionada	84	6	74	44	17	0	0	102	0	0	0	1
Condutor Dormindo	1801	229	2204	1019	102	31	39	2990	14	22	0	17
Defeito Mecânico no Veículo	1727	336	1442	632	158	76	31	2535	14	24	0	35
Defeito na Via	308	36	673	219	45	12	11	718	6	3	2	1
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	85	13	131	19	11	10	4	176	3	4	1	1
Desobediência às normas de trânsito pelo condutor	3862	1165	7734	1354	309	2452	40	7489	325	635	8	149
Desobediência às normas de trânsito pelo pedestre	127	41	58	14	13	4	0	194	0	1	0	0
Falta de Atenção do Pedestre	1031	267	751	193	61	36	9	1712	8	18	1	11
Falta de Atenção à Condução	20823	5066	22928	5821	1187	3611	234	35393	834	1231	58	448
Fenômenos da Natureza	196	10	219	141	12	5	3	257	3	0	1	3
Ingestão de Substâncias Psicoativas	73	13	55	22	5	3	1	103	2	5	0	0
Ingestão de Álcool	4635	1163	6557	1794	301	693	115	8885	148	312	8	99
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	151	41	219	53	5	13	1	321	2	9	0	7
Mal Súbito	490	87	569	197	38	21	11	851	5	9	2	12

Não guardar distância de segurança	6278	1745	3444	626	372	239	140	9730	56	107	7	190
Objeto estático sobre o leito carroçável	427	61	276	147	35	4	11	552	0	0	1	14
Pista Escorregadia	2173	209	2142	2467	205	25	31	1741	17	16	0	22
Restrição de Visibilidade	308	57	489	125	44	49	11	574	6	23	2	20
Sinalização da via insuficiente ou inadequada	125	35	195	67	10	44	3	188	6	26	6	5
Ultrapassagem Indevida	253	42	2460	464	140	25	15	2095	7	6	0	3
Velocidade Incompatível	5795	654	5200	6052	333	193	52	4733	85	121	18	62

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por faixa etária do condutor

Causa do Acidente	**Menos de 20 anos**	**20 a 24 anos**	**25 a 29 anos**	**30 a 34 anos**	**35 a 39 anos**	**40 a 44 anos**	**45 a 49 anos**	**50 a 54 anos**	**55 a 59 anos**	**60 a 64 anos**	**65 a 69 anos**	**acima de 70 anos**
Agressão Externa	8	25	36	53	42	28	16	14	12	13	7	5
Animais na Pista	42	231	349	379	386	255	222	171	117	87	35	19
Avarias e/ou desgaste excessivo no pneu	28	147	158	167	149	93	93	67	47	29	27	8
Carga excessiva e/ou mal acondicionada	3	15	25	24	25	19	15	12	10	8	4	4
Condutor Dormindo	111	536	685	629	602	444	326	299	243	174	124	61
Defeito Mecânico no Veículo	60	390	496	528	489	441	340	265	228	116	99	53

Defeito na Via	17	91	166	164	166	137	92	66	53	30	22	13
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	4	32	25	35	28	31	21	15	17	9	8	4
Desobediência às normas de trânsito pelo condutor	318	1332	1746	1809	1746	1450	1173	1038	813	638	442	256
Desobediência às normas de trânsito pelo pedestre	5	26	38	31	33	24	20	16	14	13	4	2
Falta de Atenção do Pedestre	30	207	288	333	336	232	176	158	127	87	46	29
Falta de Atenção à Condução	1043	5272	6847	7456	6803	5387	4433	3642	3007	2288	1613	1026
Fenômenos da Natureza	13	43	58	75	70	52	32	33	21	12	11	5
Ingestão de Substâncias Psicoativas	8	23	27	25	17	19	7	5	5	3	1	1
Ingestão de Álcool	294	1588	1889	1954	1847	1479	1109	888	637	376	209	85
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	4	57	65	63	55	55	33	35	20	16	5	3
Mal Súbito	15	73	114	131	154	133	111	99	123	78	67	48
Não guardar distância de segurança	185	1187	1663	1861	1759	1316	1034	874	631	519	271	167
Objeto estático sobre o leito carroçável	12	61	115	120	99	82	95	63	51	33	19	14
Pista Escorregadia	67	489	750	786	731	494	433	293	199	148	98	36
Restrição de Visibilidade	15	99	103	130	116	93	87	59	66	38	29	19
Sinalização da via insuficiente ou inadequada	6	36	46	54	40	37	35	36	18	24	16	7
Ultrapassagem Indevida	60	292	349	385	404	342	267	219	181	122	83	51

Velocidade Incompatível	348	1588	1937	1852	1590	1259	979	793	599	345	236	123

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por sexo do condutor

Causa do Acidente	Feminino	Masculino
Agressão Externa	46	213
Animais na Pista	296	1997
Avarias e/ou desgaste excessivo no pneu	166	847
Carga excessiva e/ou mal acondicionada	20	144
Condutor Dormindo	614	3620
Defeito Mecânico no Veículo	562	2943
Defeito na Via	246	771
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	31	198
Desobediência às normas de trânsito pelo condutor	2271	10490
Desobediência às normas de trânsito pelo pedestre	37	189
Falta de Atenção do Pedestre	312	1737
Falta de Atenção à Condução	10186	38631
Fenômenos da Natureza	91	334
Ingestão de Substâncias Psicoativas	24	117
Ingestão de Álcool	1161	11194
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	34	377
Mal Súbito	263	883
Não guardar distância de segurança	2415	9052

Objeto estático sobre o leito carroçável	144	620
Pista Escorregadia	853	3671
Restrição de Visibilidade	154	700
Sinalização da via insuficiente ou inadequada	69	286
Ultrapassagem Indevida	426	2329
Velocidade Incompatível	1793	9856

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por marca do veículo (continua)

Causa do Acidente	AUDI	BMW	CHERY	CHEV	CHEVROLET	CITROEN	FIAT	FORD	GM
Agressão Externa	1	0	0	9	14	3	49	20	28
Animais na Pista	2	2	0	65	147	23	462	151	235
Avarias e/ou desgaste excessivo no pneu	1	0	0	27	50	11	235	60	137
Carga excessiva e/ou mal acondicionada	2	0	0	5	13	2	32	14	11
Condutor Dormindo	4	1	6	78	259	41	894	309	521
Defeito Mecânico no Veículo	6	1	2	54	172	56	790	284	431
Defeito na Via	0	0	0	34	88	22	224	68	83
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	0	0	0	5	12	2	48	20	32
Desobediência às normas de trânsito pelo condutor	17	4	3	257	741	153	2625	925	1619

Desobediência às normas de trânsito pelo pedestre	1	0	0	3	24	4	46	20	25
Falta de Atenção do Pedestre	1	1	2	47	124	28	425	145	243
Falta de Atenção à Condução	49	18	15	970	2949	594	10077	3615	5717
Fenômenos da Natureza	0	0	0	6	38	6	86	20	45
Ingestão de Substâncias Psicoativas	0	0	0	3	8	2	21	8	29
Ingestão de Álcool	20	3	3	160	614	117	2491	912	1985
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	0	0	0	6	19	6	84	29	62
Mal Súbito	0	0	0	19	72	11	239	95	122
Não guardar distância de segurança	7	6	3	277	754	160	2288	790	1290
Objeto estático sobre o leito carroçável	2	0	1	20	43	10	152	55	78
Pista Escorregadia	2	4	1	135	343	99	768	239	425
Restrição de Visibilidade	2	0	0	28	48	11	186	61	98
Sinalização da via insuficiente ou inadequada	0	0	0	8	20	5	85	15	33
Ultrapassagem Indevida	4	1	0	56	151	29	596	209	356
Velocidade Incompatível	20	13	1	226	696	189	2042	672	1406

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por marca do veículo

Causa do Acidente	HONDA	HYUNDAI	I	IMP	M.BENZ	MMC	NISSAN	PEUGEOT	RENAULT	TOYOTA	VW
Agressão Externa	6	7	37	1	0	0	2	10	19	12	41

Animais na Pista	60	69	291	5	6	1	13	25	127	140	469
Avarias e/ou desgaste excessivo no pneu	40	24	121	11	0	0	5	13	42	25	211
Carga excessiva e/ou mal acondicionada	6	1	22	2	0	0	1	2	10	12	29
Condutor Dormindo	119	117	449	35	4	0	35	85	199	154	924
Defeito Mecânico no Veículo	90	55	439	46	8	1	15	81	160	75	739
Defeito na Via	33	60	112	5	1	0	5	18	49	60	155
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	4	3	15	5	0	0	3	5	9	9	57
Desobediência às normas de trânsito pelo condutor	338	295	1385	130	9	0	70	178	632	497	2883
Desobediência às normas de trânsito pelo pedestre	13	3	26	2	0	0	3	4	5	11	36
Falta de Atenção do Pedestre	84	57	255	12	1	0	8	32	82	82	420
Falta de Atenção à Condução	1621	1292	5561	429	39	9	288	766	2524	1954	10330
Fenômenos da Natureza	10	13	51	3	1	0	3	8	24	17	94
Ingestão de Substâncias Psicoativas	3	3	12	3	0	0	3	5	3	2	36
Ingestão de Álcool	324	215	1253	157	13	1	47	176	468	326	3070
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	10	9	33	6	0	0	2	8	16	7	114
Mal Súbito	40	19	118	18	1	0	8	16	65	52	251
Não guardar distância de segurança	443	347	1451	62	9	7	98	157	623	467	2228

Objeto estático sobre o leito carroçável	23	13	109	4	2	0	8	13	39	38	154
Pista Escorregadia	218	180	564	19	5	1	34	72	253	194	968
Restrição de Visibilidade	23	13	100	6	3	0	2	7	34	41	191
Sinalização da via insuficiente ou inadequada	8	7	48	5	0	0	1	4	17	20	79
Ultrapassagem Indevida	77	66	263	15	1	0	18	30	107	130	646
Velocidade Incompatível	517	358	1531	106	12	2	69	205	670	546	2368

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por faixa de idade do veículo.

Causa do Acidente	**Automóvel com até 1 ano**	**Automóvel de 2 a 4 anos**	**Automóvel de 5 a 8 anos**	**Automóvel de 9 a 13 anos**	**Automóvel de 14 a 20 anos**	**Automóvel de 20 a 30 anos**	**Automóvel acima de 30 anos**
Agressão Externa	15	62	85	62	19	14	2
Animais na Pista	213	584	738	430	218	98	12
Avarias e/ou desgaste excessivo no pneu	49	158	280	231	188	92	15
Carga excessiva e/ou mal acondicionada	17	31	51	34	18	12	1
Condutor Dormindo	280	764	1253	967	632	314	24
Defeito Mecânico no Veículo	137	478	915	814	634	408	119
Defeito na Via	86	269	337	202	82	38	3
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	10	49	49	50	33	31	7
Desobediência às normas de trânsito pelo condutor	719	2084	3592	2983	1960	1228	195
Desobediência às normas de trânsito pelo pedestre	22	39	76	48	28	12	1
Falta de Atenção do Pedestre	147	438	603	449	252	144	16

Falta de Atenção à Condução	2923	9377	14743	10727	6715	3795	537
Fenômenos da Natureza	24	111	123	93	44	24	6
Ingestão de Substâncias Psicoativas	8	13	41	31	30	16	2
Ingestão de Álcool	472	1503	3056	2904	2394	1762	264
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	15	42	110	101	80	58	5
Mal Súbito	60	229	334	242	168	104	9
Não guardar distância de segurança	846	2398	3664	2533	1367	602	57
Objeto estático sobre o leito carroçável	59	175	241	146	96	40	7
Pista Escorregadia	221	1160	1522	983	438	169	31
Restrição de Visibilidade	48	166	259	171	131	66	13
Sinalização da via insuficiente ou inadequada	23	78	122	72	39	19	2
Ultrapassagem Indevida	164	489	871	623	377	210	21
Velocidade Incompatível	595	2296	3773	2602	1457	820	106

Fonte: Elaborado pelo autor

Tabela de contingência de causas de acidentes por potência do veículo

Causa do Acidente	**Automóvel de 60 a 85 cv**	**Automóvel de 86 a 111 cv**	**Automóvel de 112 a 137 cv**	**Automóvel de 138 a 163 cv**	**Automóvel de 164 a 189 cv**	**Automóvel de 190 a 215 cv**	**Automóvel de 216 a 241 cv**	**Automóvel acima de 242 cv**
Agressão Externa	67	111	52	20	4	1	4	0
Animais na Pista	598	971	476	174	57	9	6	2
Avarias e/ou desgaste excessivo no pneu	286	455	180	72	15	3	2	0
Carga excessiva e/ou mal acondicionada	49	62	35	10	5	2	1	0
Condutor Dormindo	1250	1831	803	246	81	14	5	4
Defeito Mecânico no Veículo	931	1631	668	193	61	13	6	2

Defeito na Via	293	417	214	59	24	3	6	1
Deficiência ou não Acionamento do Sistema de Iluminação/Sinalização do Veículo	53	110	44	19	0	2	1	0
Desobediência às normas de trânsito pelo condutor	3530	5787	2404	774	191	48	19	8
Desobediência às normas de trânsito pelo pedestre	63	83	51	21	4	3	1	0
Falta de Atenção do Pedestre	538	894	404	147	43	14	3	6
Falta de Atenção à Condução	13674	21462	9338	3109	892	201	92	49
Fenômenos da Natureza	112	179	98	23	12	0	1	0
Ingestão de Substâncias Psicoativas	30	79	21	7	3	1	0	0
Ingestão de Álcool	3355	5848	2244	648	186	46	23	5
Ingestão de álcool e/ou substâncias psicoativas pelo pedestre	126	186	73	18	7	1	0	0
Mal Súbito	325	503	207	85	19	2	4	1
Não guardar distância de segurança	3128	4904	2268	852	228	48	28	11
Objeto estático sobre o leito carroçável	175	332	167	63	16	5	3	3
Pista Escorregadia	1114	1928	965	390	81	24	18	4
Restrição de Visibilidade	242	346	160	78	20	6	1	1
Sinalização da via insuficiente ou inadequada	106	158	59	24	7	1	0	0
Ultrapassagem Indevida	752	1204	554	179	56	3	2	5
Velocidade Incompatível	2767	4913	2553	994	306	63	32	21

Fonte: Elaborado pelo autor